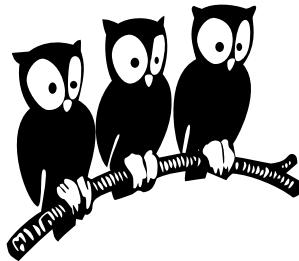


Département de Physique
École Normale Supérieure

Laboratoire de Physique Statistique



THÈSE de DOCTORAT de l'UNIVERSITÉ PARIS 7

Spécialité : Physique Théorique

présentée par

Marc SANTOLINI

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

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

Soutenue le ZZ septembre 2013
devant le jury composé de :

M.	Vincent HAKIM	Directeur de thèse
M.	Martin Weigt	Rapporteur
M.	ZZZ	Examinateur
M.	ZZZ	Président du jury
M.	ZZZ	Rapporteur
M.	Pascal Maire	Membre invité

Remerciements

...

thèse:version du mardi 28 mai 2013 à 17 h 13]

Table des matières

Liste des figures	vii
Principales abréviations utilisées	ix
Avant-propos	1
Chapitre 1 - Introduction générale.	3
	3
1.1 Le phénotype cellulaire	4
1.2 Les réseaux de régulation génétique	7
1.3 Modèles mathématiques des interactions protéine-ADN	14
1.4 Mesures expérimentales des interactions protéine-ADN	19
1.5 Les modules de cis-régulation	26
1.6 Banques de données	33
Chapitre 2 - Modèles de fixation des Facteurs de Transcription à l'ADN.	35
	35
2.1 Les modèles de fixation	37
2.2 Description des données biologiques	38
2.3 Présentation de l'algorithme	38
2.4 Performance des modèles	38
2.5 Analyse des corrélations	38
2.6 Comparaison avec des données <i>in vitro</i>	38
Chapitre 3 - <i>Imogene</i> : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle	41
	41
3.1	43
Chapitre 4 - Étude de la différenciation épidermale chez la drosophile	45
	45
4.1	47

Chapitre 5 - Étude de la différenciation musculaire chez la souris	49
49	
5.1	51
Conclusion	53
Bibliographie	55

Liste des figures

Introduction générale.	3
	3
1.1 Le paysage de la différenciation cellulaire	5
1.2 Spécification spatio-temporelle du type cellulaire	6
1.3 Différents exemples de reprogrammation cellulaire	7
1.4 Vision cybernétique du traitement de l'information par la cellule	8
1.5 Un réseau de régulation génétique type	9
1.6 Caractéristiques de l'épigénome	10
1.7 Exemples de motifs dans les réseaux de régulation génétique	12
1.8 Évolution du réseau transcriptionnel : l'exemple de la régulation myogénique.	13
1.9 Différents états du facteur de transcription	15
1.10 Construction et utilisation du modèle PWM	17
1.11 Étapes d'une expérience de ChIP-seq	23
1.12 Résolution des expériences ChIP-on-chip et ChIP-seq	24
1.13 Expérience d'empreinte à la DNase I chez la levure : vers une résolution au nucléotide près	25
1.14 Différents CRMs conduisent à différents patterns d'expression	27
1.16 Les états épigénétiques des CRMs	29
1.17 Approches pour la prédiction des CRMs	30
1.21 Méthodes de validation des CRMs	34
 Modèles de fixation des Facteurs de Transcription à l'ADN.	35
	35
2.1 Description graphique de l'algorithme.	39
 <i>Imogene</i> : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle	41
	41
 Étude de la différenciation épidermale chez la drosophile	45
	45
 Étude de la différenciation musculaire chez la souris	49
	49

Principales abréviations utilisées

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

Chapitre 1

Introduction générale.

1.1 Le phénotype cellulaire	4
1.1.1 Qu'est-ce que le phénotype d'une cellule ?	4
1.1.2 La différenciation cellulaire	4
1.1.3 La cellule dans l'organisme : une spécification spatio-temporelle	6
1.1.4 La reprogrammation cellulaire	6
1.2 Les réseaux de régulation génétique	7
1.2.1 Vision cybernétique de la cellule	7
1.2.2 Divers modes de régulation	8
1.2.3 Câblage du réseau et fonction	11
1.2.4 Évolution des réseaux génétiques	11
1.3 Modèles mathématiques des interactions protéine-ADN	14
1.3.1 Modes de recherche du site de fixation par le TF	14
1.3.2 Modèle PWM	15
1.3.3 Modèle biophysique	16
1.3.4 Modèle thermodynamique	18
1.4 Mesures expérimentales des interactions protéine-ADN	19
1.4.1 Approches <i>in vitro</i> : MITOMI, SPR, PBM, CSI, SELEX, et HT-SELEX	20
1.4.2 Approche clonale : la technique de simple hybride	22
1.4.3 Approches <i>in vivo</i> : ChIP-on-chip, ChIP-seq, DNase	23
1.5 Les modules de cis-régulation	26
1.5.1 Modules et fonctions logiques	26
1.5.2 Encodage de patterns spatiaux	26
1.5.3 Différents états des CRMs	27
1.5.4 Prédiction des CRMs	28
1.5.5 Grammaire des enhancers : enhanceosome vs billboard	29
1.5.6 Évolution des enhancers	30
1.5.7 Les « shadow enhancers »	32
1.5.8 Validation expérimentale	33
1.6 Banques de données	33
1.6.1 Séquences génomiques et alignements	33
1.6.2 Annotations (TSSs, repeats...)	33
1.6.3 Jaspar et Transfac	33
1.6.4 Visualisation sur UCSC	33
1.6.5 Le projet ENCODE	33

1.1 Le phénotype cellulaire

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

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

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

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

Ces différentes propriétés observables caractérisent le *phénotype* cellulaire (littéralement « exhiber un type » en grec). Ce phénotype est le résultat de la modulation par des facteurs environnementaux de l'expression génétique qui détermine le contenu en protéines de la cellule.

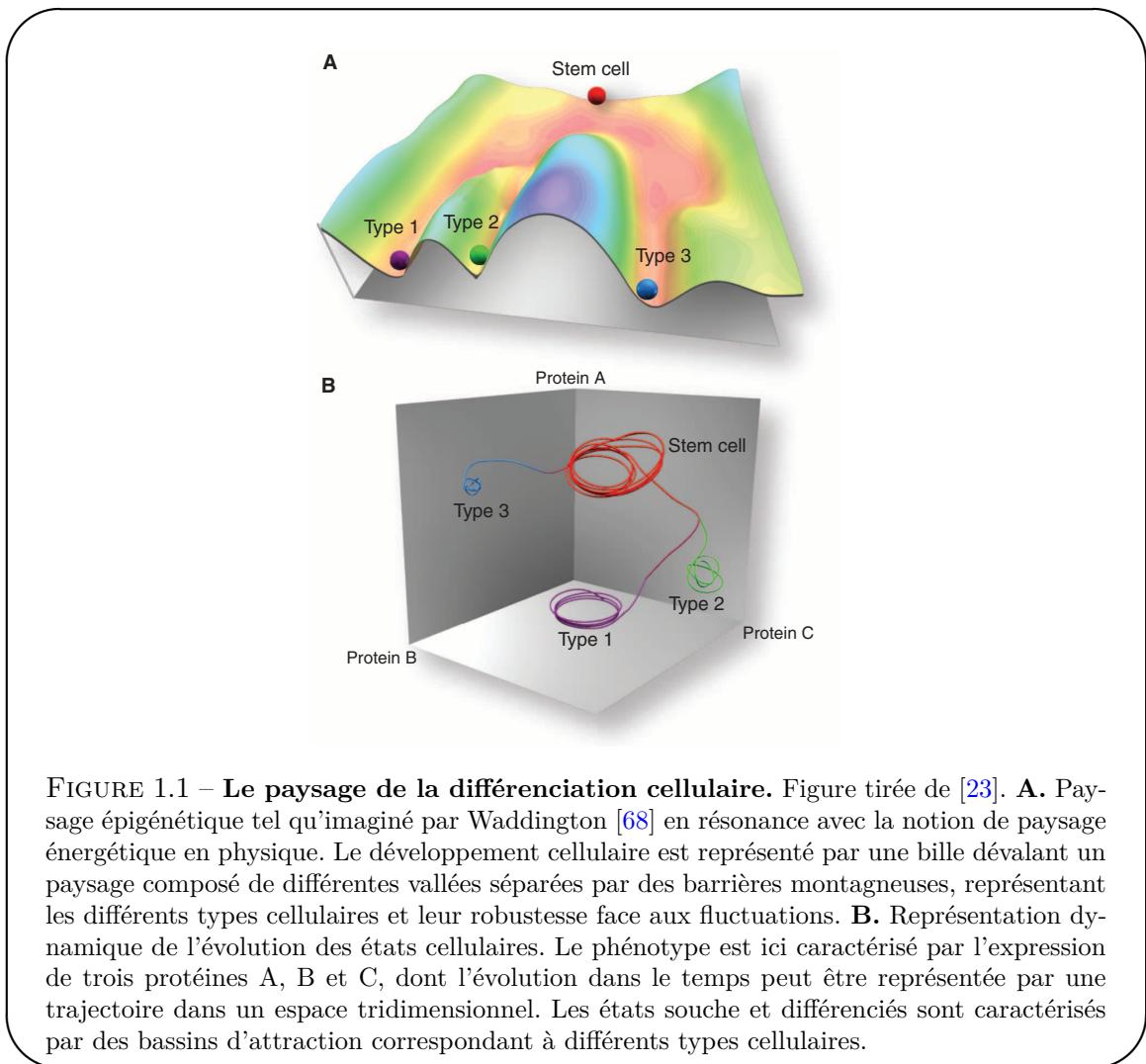
1.1.2 La différenciation cellulaire

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

Dans un écrit aujourd'hui célèbre datant de 1957 [68], 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 proces-

1. Il existe cependant quelques cas connus d'organismes multicellulaires procaryotes, par exemple chez les bactéries magnétotactiques [37].

1.1. Le phénotype cellulaire



sus de différenciation cellulaire est comparé à une bille dévalant une pente et dont la trajectoire suit les multiples embranchements de vallées escarpées, chacune représentant un état de développement différent. Les vallées sont séparées par des pics dont la hauteur reflète la difficulté de passer d'un état à un autre, et les destinations finales possibles de la bille correspondent aux différents types cellulaires.

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

comme un « attracteur » de la dynamique. Une fois au sein d'un attracteur, l'état cellulaire est robuste aux perturbations du niveau d'expression génétique des différentes composantes. Les attracteurs peuvent alors être vu comme des types cellulaires distincts correspondant aux différentes vallées de la représentation de Waddington [36].

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

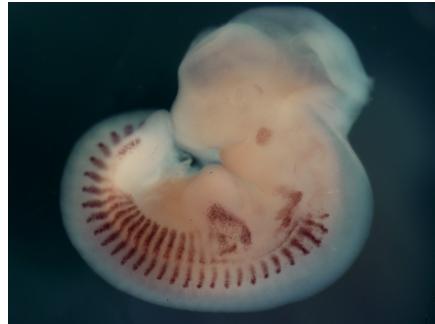
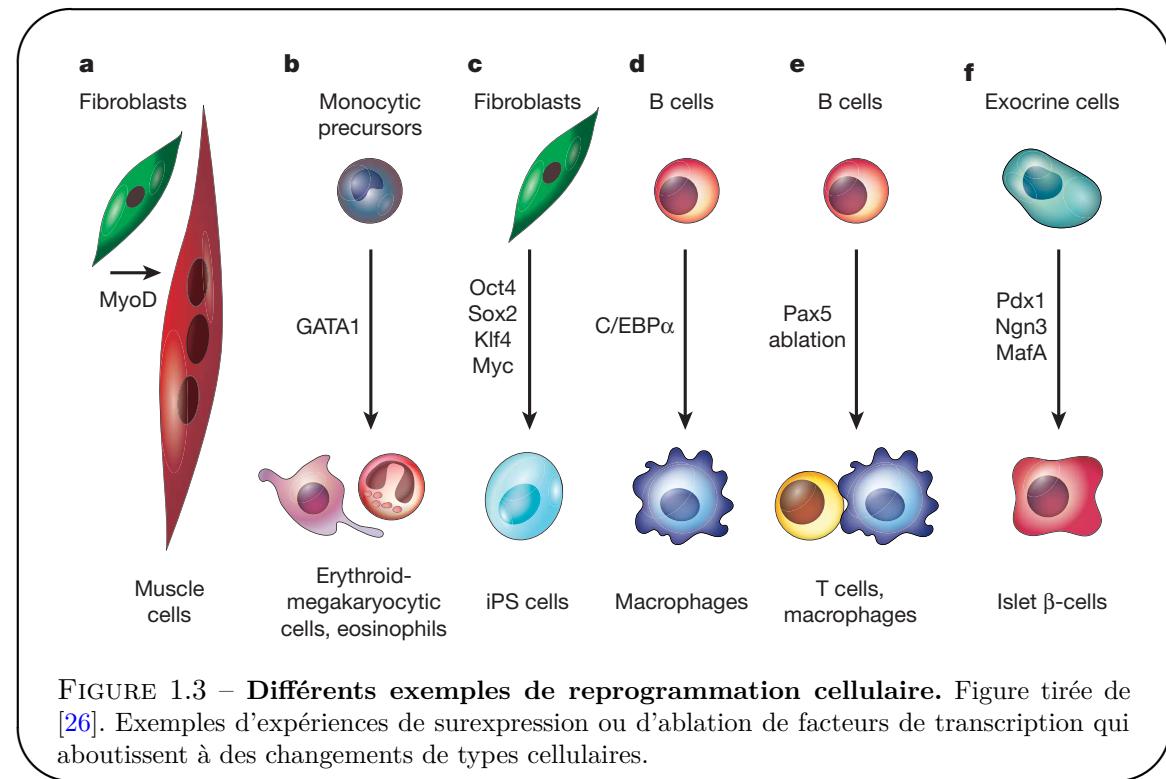


FIGURE 1.2 – Spécification spatio-temporelle du type cellulaire. Hybridization *in situ* de *Myog*, marqueur des cellules musculaires squelettiques différencierées, chez un embryon de souris de 11,5 jours. Le pattern de spécification des cellules myogéniques est clairement visible au niveau des futures vertèbres.

Un fait remarquable à propos de la différenciation cellulaire est que celle-ci opère à un rythme précis et dans un contexte cellulaire bien défini. Aussi, les trajectoires dans l'espace d'expression génétique que nous avons présentées précédemment sont fonction de l'espace – la position de la cellule dans l'organisme, qui détermine en particulier la concentration des signaux qu'elle reçoit – et du temps – les étapes de développement se succédant de manière irréversible –. Ainsi, la différenciation des cellules observe certains *patterns* spatio-temporels bien définis : par exemple, dans le cas de la formation des muscles, le marqueur des cellules du muscle squelettique *Myog* est exprimé chez la souris dès 8 jours embryonnaires au niveau des somites, les futures vertèbres de la souris adulte (voir fig 1.2).

1.1.4 La reprogrammation cellulaire

Dans les paragraphes précédents, nous avons présenté la vision classique selon laquelle des cellules souches totipotentes se différencient en des cellules de moins en moins plastiques, jusqu'à atteindre un état différencié stable. Néanmoins, depuis plusieurs décénies, différentes expériences ont exhibé la plasticité des états différenciés. Par exemple, Blau et al. ont montré en 1985 que des programmes d'expression génétiques dormants peuvent être exprimés de manière dominante dans des cellules différencierées par la fusion de différents types cellulaires [12]. Puis différents travaux ont montré qu'il était possible de convertir des lignées de cellules en introduisant certaines protéines régulatrices de la transcription, ou Facteurs de Transcription (TFs) [18, 40] (voir fig 1.3). Parallèlement, des expériences réalisées chez plusieurs espèces ont montré que le transfert de noyaux de cellules différencierées embryonnaires ou adultes dans un oeuf énucléé peut mener à la formation d'un organisme complet, montrant de manière univoque que l'identité des cellules différencierées peut être complètement renversée [28]. Enfin, l'avancée la plus récente dans



ce domaine a été la démonstration que des cellules somatiques différencierées peuvent être reprogrammées en cellules souches puripotentes par simple introduction d'un cocktail de 4 facteurs de transcription [64] (fig 1.3C).

1.2 Les réseaux de régulation génétique

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

1.2.1 Vision cybernétique de la cellule

Le paradigme qui règne sur la biologie moléculaire depuis plus d'un demi siècle est celui des réseaux génétiques. L'expression est gènes est en effet régulée par des protéines, les facteurs de transcription, qui sont elles-mêmes exprimées par d'autres gènes, créant ainsi des interactions entre gènes. Par ailleurs, les protéines peuvent réguler l'activité d'autres protéines, et certains ARN issus de la transcription de gènes non codants opèrent aussi de manière primordiale dans la régulation de l'activité génétique, le tout formant un réseau complexe d'interactions. La compréhension de ce réseau et des fonctions qu'il englobe forme le socle de la discipline de biologie des systèmes. Dans ce cadre, la cellule est vue comme une machine interprétant différents signaux reçus en entrée et qui, une fois traités par le réseau interne de régulation, réagit en sortie en modifiant son état ou son comportement (fig 1.4). L'intérêt d'une telle description mécanistique est qu'elle permet d'opérer quantifications mathématiques et prédictions, ce qui l'a rendue extrêmement fertile au cours des dernières décennies [48].

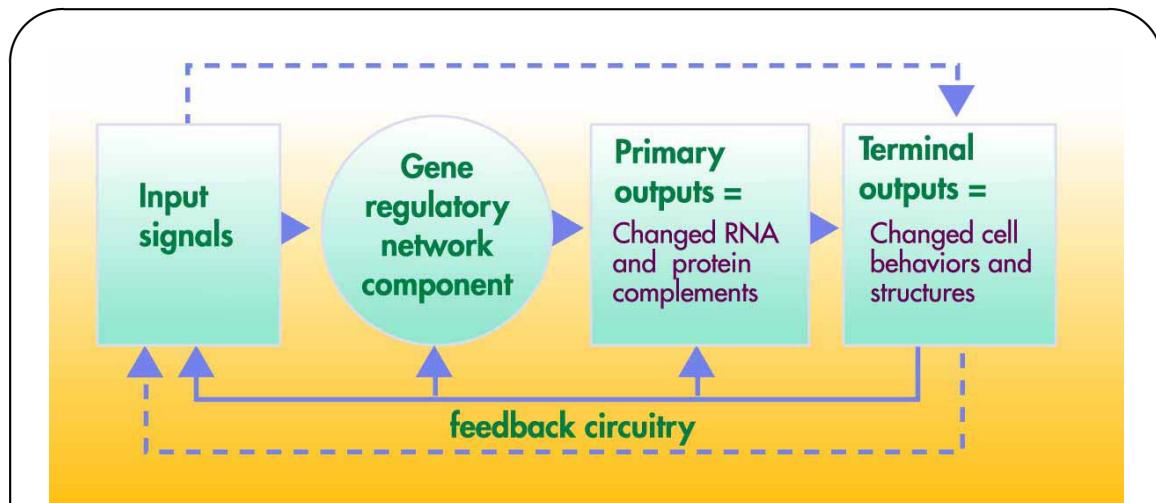


FIGURE 1.4 – Vision cybernétique du traitement de l’information par la cellule.
 Figure tirée du programme “Genomes to life” du Département de l’Énergie des États-Unis datant de 2001 [1] schématisant un réseau de régulation cellulaire comme un système de traitement entrée/sortie, possédant trois composantes fondamentales : (1) un système de réception et de transduction des signaux d’entrées qui peuvent être intra- ou extra-cellulaires (plusieurs signaux pouvant affecter un même gène cible), (2) un “composant central” (*core component*) composé du réseau de régulation génétique traitant les signaux, et (3) de l’expression moléculaire des ARNs et protéines des gènes cibles observée en sortie. Le processus résulte en la modification du phénotype ou de la fonction de la cellule. Des boucles de régulation (*feedback*) assurent le contrôle et la stabilité des différentes étapes.

1.2.2 Divers modes de régulation

Les modes de régulation qui permettent à la cellule d’interpréter des signaux et de changer d’état sont nombreux. Nous allons nous concentrer ici sur ceux internes aux réseaux génétiques, et affectant au final la production de protéines ou d’ARNs et donc l’état cellulaire (fig. 1.5).

- **Régulation génétique**

Tout d’abord, un réseau d’expression génétique est caractérisé par un jeu d’interactions entre différents gènes. Ces interactions se font par l’intermédiaire de protéines régulatrices appelées facteurs de transcription ou TFs, qui sont au nombre de ~ 1400 chez l’homme [67], 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.

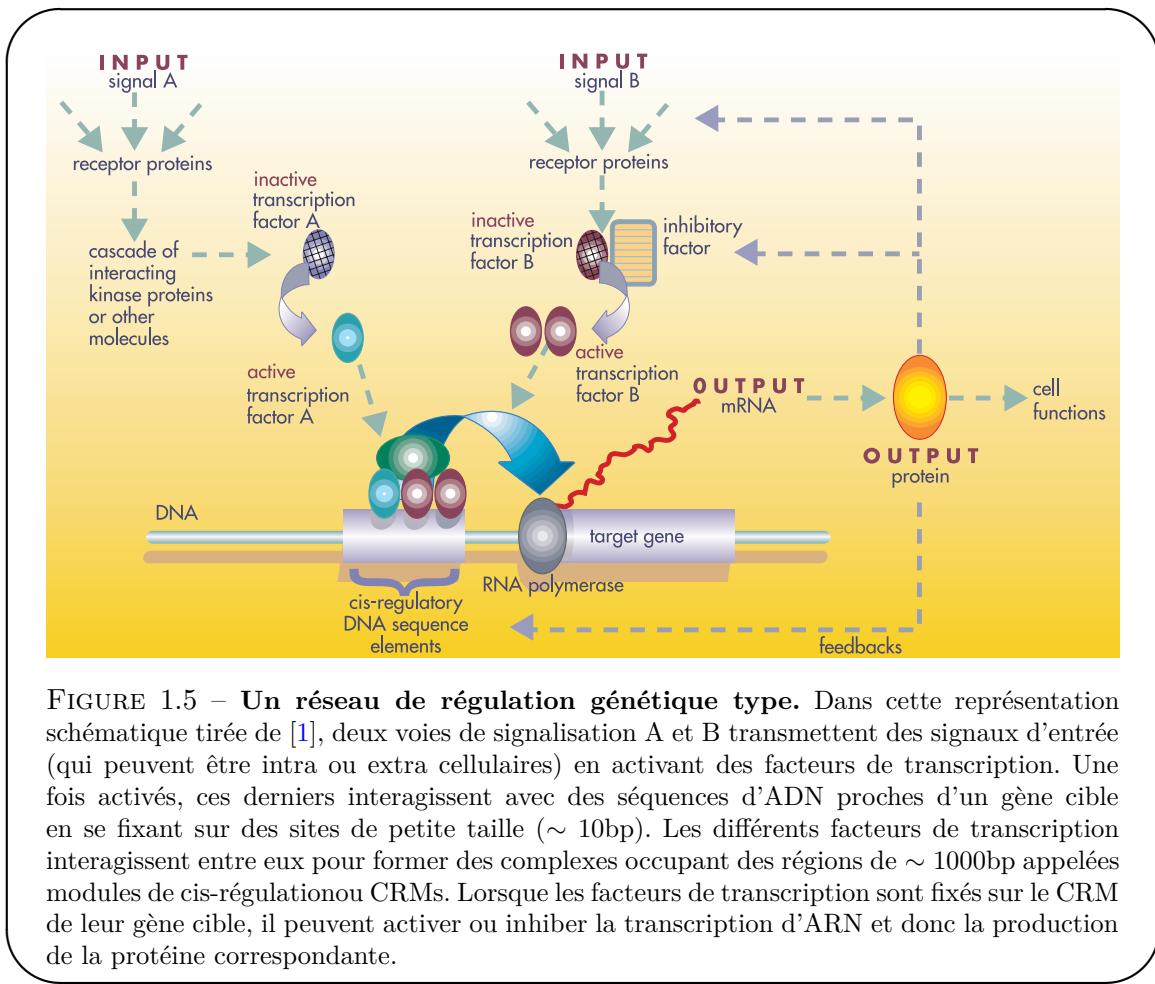


FIGURE 1.5 – Un réseau de régulation génétique type. Dans cette représentation schématique tirée de [1], deux voies de signalisation A et B transmettent des signaux d'entrée (qui peuvent être intra ou extra cellulaires) en activant des facteurs de transcription. Une fois activés, ces derniers interagissent avec des séquences d'ADN proches d'un gène cible en se fixant sur des sites de petite taille ($\sim 10\text{bp}$). Les différents facteurs de transcription interagissent entre eux pour former des complexes occupant des régions de $\sim 1000\text{bp}$ appelées modules de cis-régulation ou CRMs. Lorsque les facteurs de transcription sont fixés sur le CRM de leur gène cible, il peuvent activer ou inhiber la transcription d'ARN et donc la production de la protéine correspondante.

• Régulation épigénétique

Outre la régulation génétique, due à l'action de protéines issues de séquences codantes et se fixant sur des séquences d'ADN, régulation qui est donc entièrement encodée dans le génome et transmise à la descendance, il existe un autre mode de régulation de la transcription des gènes qui permet notamment d'acquérir une modification d'expression génétique transmise à la descendance sans qu'il y ait modification du code génétique : on parle de régulation épigénétique. Cette régulation passe notamment par la modification des propriétés chimiques de l'ADN et des histones sur lequel il s'enroule pour former la chromatine. Ainsi, la méthylation des dimères CpG de l'ADN² au niveau des régions riches en CG, ou îlots CpG, situées près de nombreux promoteurs et habituellement dépourvues de ces marques conduit à une inactivation du gène cible [10]. Par ailleurs, la méthylation des histones au niveau des résidus lysines entraîne la fermeture de la chromatine, empêchant l'expression du ou des gène(s) situés à leur niveau, alors que l'acétylation des mêmes lysines entraîne au contraire une ouverture de la chromatine, favorisant ainsi la transcription génétique [27]. Ce mode de régulation sera développé plus en détails en section 1.5.3.

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

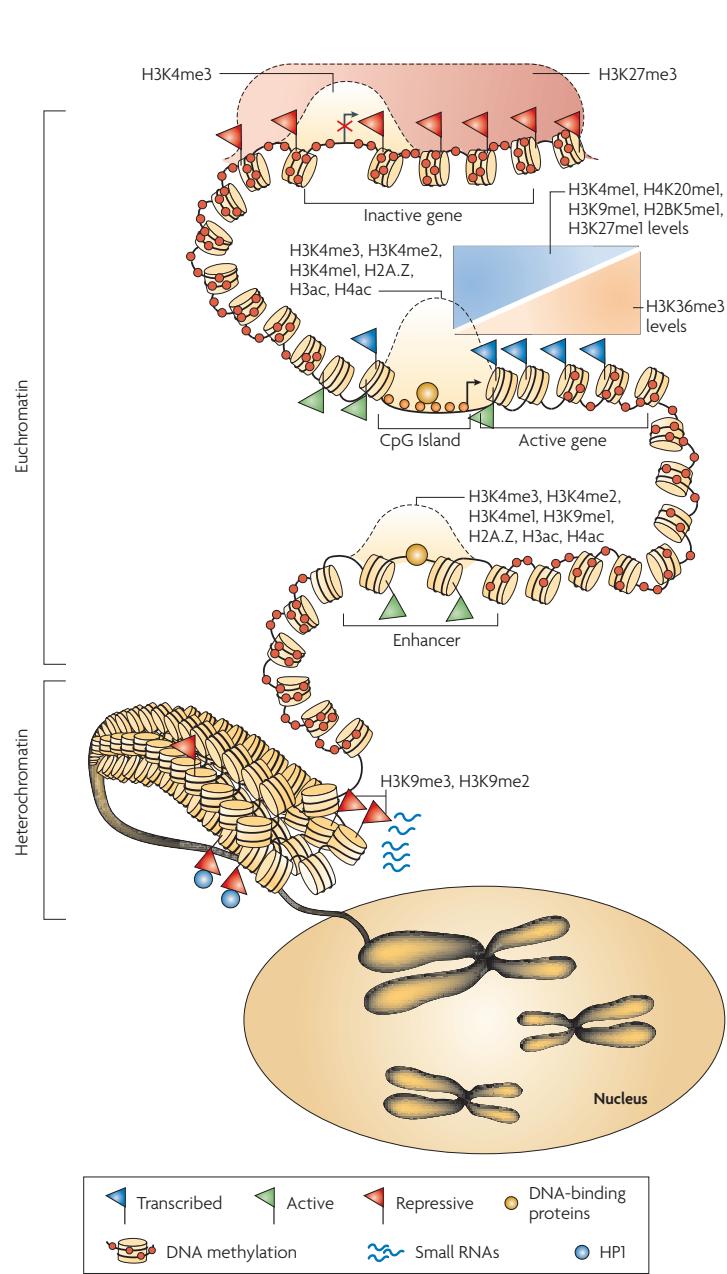


FIGURE 1.6 – Caractéristiques de l'épigénome. Figure tirée de [53]. Les chromosomes sont partagés entre régions accessibles d'euchromatine et régions difficilement accessibles d'hétérochromatine. Les régions hétérochromatiques sont marquées par de la di- et triméthylation de la lysine 9 de l'histone H3 (H3K9me2 et H3K9me3). La méthylation de l'ADN est pervasive à travers le génome et est seulement absente dans les régions telles que les îlots CpG, les promoteurs et les CRMs. La modification H3K27me3 couvre de larges régions englobant des gènes inactifs. Les marques H3K4me3, H3K4me2, H3K4me1 et l'acétylation des histones marquent les TSSs des gènes actifs. Les marques H3K4, H3K9, H3K27, H4K20 et H2BK5 marquent les régions transcris activement à proximité de la région 5' des gènes (en aval), alors que la marque H3K36 marque les gènes transcrits dans leur région 3' (en amont).

- **Régulation post-transcriptionnelle**

Les modifications post-transcriptionnelles affectent les ARNs issus de la transcription des gènes. Ces modifications peuvent être causées par des ARNs doubles brins ou dsRNA (*double-stranded RNAs*) qui, une fois clivés par la protéine Dicer, forment des petits peptides de 22 nts appelés siRNAs (*small interfering RNAs*) qui recrutent le complexe protéique RISC (*RNA-induced silencing complex*) et ciblent spécifiquement des ARNm [29, 30]. Cette méthode est connue sous le nom d'interférence ARN (RNAi) et est aujourd'hui couramment utilisée pour inhiber l'expression d'un gène. De manière similaire, les microARNs ou miRNAs sont des ARNs de ~ 23 nts issus d'ARNs plus longs appelés « épingles à cheveux » ou *hairpins* qui s'associent à la protéine *Argonaute* du complexe RISC pour entraîner la dégradation spécifique d'ARNms [5].

- **Régulation post-traductionnelle**

Les modifications post-traductionnelles affectent les protéines issues de la traduction des ARNs. Ces modifications passent par une modification chimique des protéines, typiquement la phosphorylation, ou comme nous l'avons vu pour la régulation épigénétique, la méthylation ou l'acétylation. Ces modifications peuvent avoir pour effet de changer l'activité de la protéine, que ce soit en modifiant son activité enzymatique ou en déclenchant sa relocalisation nucléaire. Par ailleurs, il existe aussi des modifications de structure de la protéine, comme c'est le cas du facteur de transcription *Shavenbaby* chez la Drosophile : dans sa forme native, cette protéine inhibe la transcription de ses gènes cible ; cependant ses résidus terminaux peuvent être clivés par des petits peptides de 11 à 32 acides aminés encodés par le gène *Pri*, rendant la protéine transcriptionnellement active [39].

1.2.3 Câblage du réseau et fonction

Maintenant que nous avons vu la nature des interactions au sein des réseaux génétiques, nous pouvons nous pencher sur leur structure. Celle-ci est en effet loin d'être due au hasard. Ainsi, plusieurs études, réalisées chez divers organismes de la bactérie à l'homme, ont révélé que les réseaux de transcription contiennent un petit ensemble de motifs de régulation récurrents, appelés motifs de réseaux [2, 55, 46] (fig. 1.7). Ces motifs peuvent être vus comme les pièces élémentaires servant à la construction de réseaux fonctionnels. De tels motifs furent d'abord détectés de manière systématique chez la bactérie *Escherichia coli* en remarquant qu'ils apparaissaient dans le réseau de transcription bien plus souvent qu'on ne l'attendrait dans un réseau aléatoire [55]. Les mêmes motifs ont ensuite été trouvés chez la levure [46, 42] et chez l'homme [49]. 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 [3]. 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

Au cours de l'évolution, les réseaux de régulation génétique changent : modification des constituants, recâblage du réseau, duplication d'éléments... Néanmoins, certaines modifications sont plus défavorisées du point de vue évolutif que des autres. Par exemple, la modification d'un régulateur, par exemple une mutation d'un certain acide aminé d'un facteur de transcription,

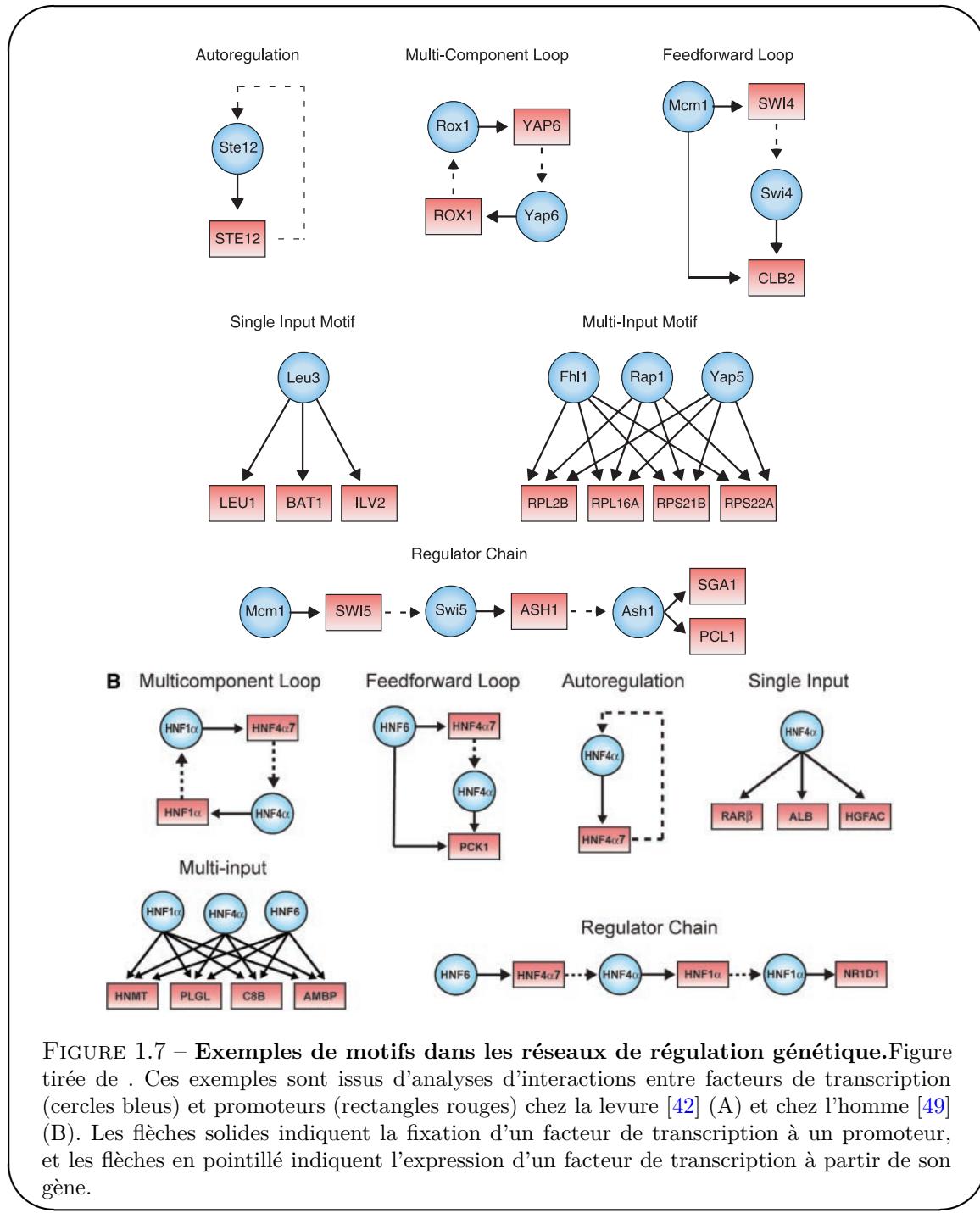


FIGURE 1.7 – Exemples de motifs dans les réseaux de régulation génétique. Figure tirée de . Ces exemples sont issus d'analyses d'interactions entre facteurs de transcription (cercles bleus) et promoteurs (rectangles rouges) chez la levure [42] (A) et chez l'homme [49] (B). Les flèches solides indiquent la fixation d'un facteur de transcription à un promoteur, et les flèches en pointillé indiquent l'expression d'un facteur de transcription à partir de son gène.

aura des conséquences sur l'ensemble des éléments régulés par ce facteur de transcription. Par contre, la modification d'un site de reconnaissance de ce facteur de transcription sur l'ADN n'aura qu'une portée locale sur la régulation du gène associé. Par ailleurs, certains motifs du réseau, comme les boucles d'autorégulation ou les boucles feedforward, peuvent avoir une grande importance fonctionnelle, favorisant leur conservation.

À titre d'exemple, prenons le cas du réseau de différenciation du muscle squelettique présenté en figure 1.8, que nous étudierons plus en détail dans le chapitre 5 de ce manuscrit. Au coeur de

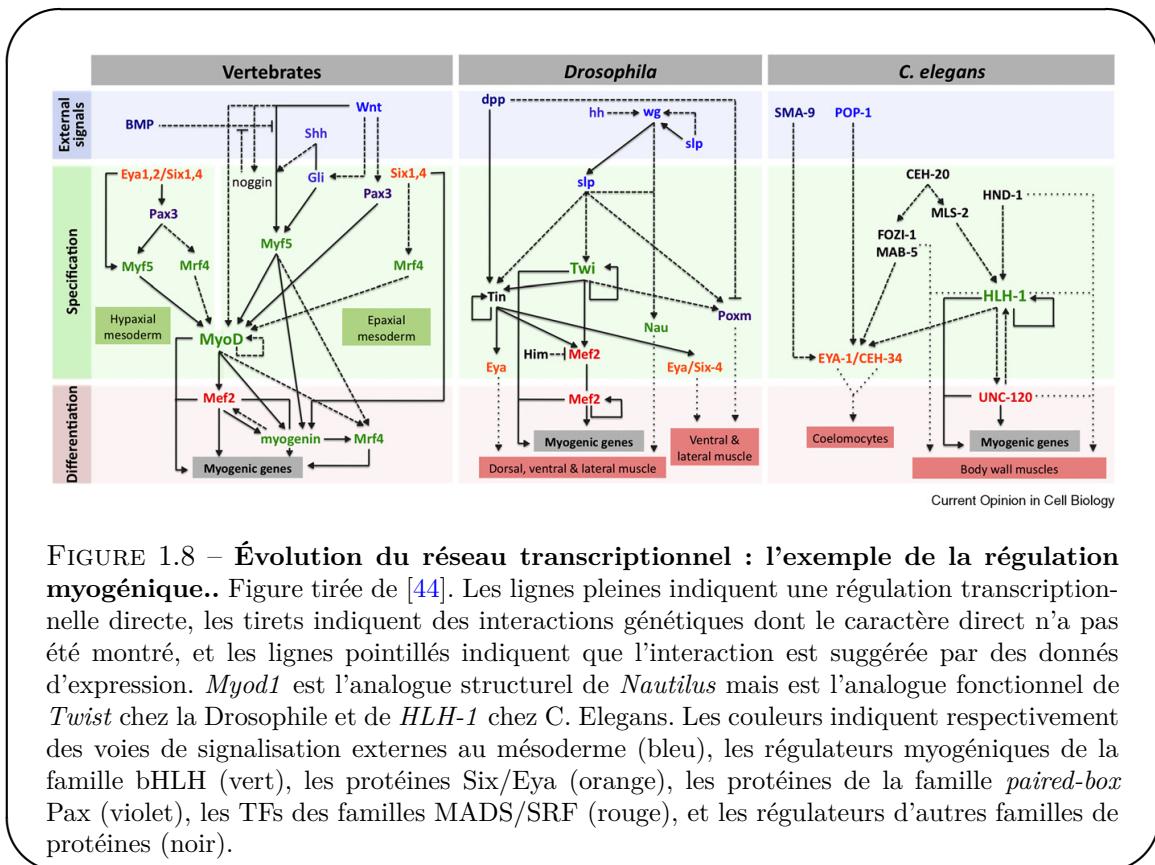


FIGURE 1.8 – Évolution du réseau transcriptionnel : l'exemple de la régulation myogénique.. Figure tirée de [44]. Les lignes pleines indiquent une régulation transcriptionnelle directe, les tirets indiquent des interactions génétiques dont le caractère direct n'a pas été montré, et les lignes pointillées indiquent que l'interaction est suggérée par des données d'expression. *Myod1* est l'analogue structurel de *Nautilus* mais est l'analogue fonctionnel de *Twist* chez la Drosophile et de *HLH-1* chez *C. Elegans*. Les couleurs indiquent respectivement des voies de signalisation externes au mésoderme (bleu), les régulateurs myogéniques de la famille bHLH (vert), les protéines Six/Eya (orange), les protéines de la famille *paired-box* Pax (violet), les TFs des familles MADS/SRF (rouge), et les régulateurs d'autres familles de protéines (noir).

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 [71]. 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 [6]. 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*.

Ainsi le cœur du réseau est conservé dans la forme (topologie), même s'il y a des divergences

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

Nous voyons donc que l'évolution d'un réseau génétique possède de multiples facettes : conservation de motifs de réseau fonctionnellement importants (dans notre exemple, la boucle feed-forward au coeur du réseau régissant l'aspect temporel de l'expression des cibles), recâblage des interactions pour traiter différents signaux d'entrée... Par ailleurs, il apparaît que plus qu'à des TFs particuliers, c'est à des familles de TFs que nous avons affaire. Aussi un même rôle au sein du réseau peut-il être rempli par différents membres d'une même famille, comme c'est le cas pour *Myod1* et *Twist*. Ceci s'explique par le fait que les membres d'une même famille partagent des propriétés d'interaction avec l'ADN semblables. Ces interactions sont à la source du fonctionnement du réseau, et nous allons maintenant présenter plus en avant leurs propriétés.

1.3 Modèles mathématiques des interactions protéine-ADN

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

1.3.1 Modes de recherche du site de fixation par le TF

Un facteur de transcription peut être dans plusieurs états : en diffusion tridimensionnelle, auquel cas il est dit "libre", ou bien fixé sur l'ADN. Dans ce dernier cas, il interagit avec l'ADN selon deux modes : une attraction non spécifique d'énergie E_{ns} indépendante de la position sur r l'ADN, et une interaction spécifique $E_s(r)$ qui dépend de la séquence de taille $l \sim 10$ à la position r . L'interaction non spécifique est due à l'interaction électrostatique entre la protéine chargée positivement et l'ADN chargé négativement, alors que l'interaction spécifique implique des liaisons hydrogènes entre le domaine de fixation de la protéine et les nucléotides du site de fixation. La protéine passe d'un mode à l'autre en changeant de conformation. Au final, le facteur de transcription peut être dans trois états thermodynamiques représentés en figure 1.9 : en diffusion tridimensionnelle libre, fixé non spécifiquement (diffusion unidimensionnelle le long de la structure d'ADN), et fixé spécifiquement sur l'ADN. Ces trois modes contribuent à la cinétique de la recherche d'un site fonctionnelle [8, 73, 72]. Ainsi, l'attraction non spécifique conduit la protéine à passer à peu près autant de tant fixé sur l'ADN qu'en diffusion libre. La recherche de site de reconnaissance est donc un processus mixte de diffusion unidimensionnelle sur l'ADN et de diffusion tridimensionnelle dans le milieu. Lorsqu'il est fixé sur l'ADN, le facteur

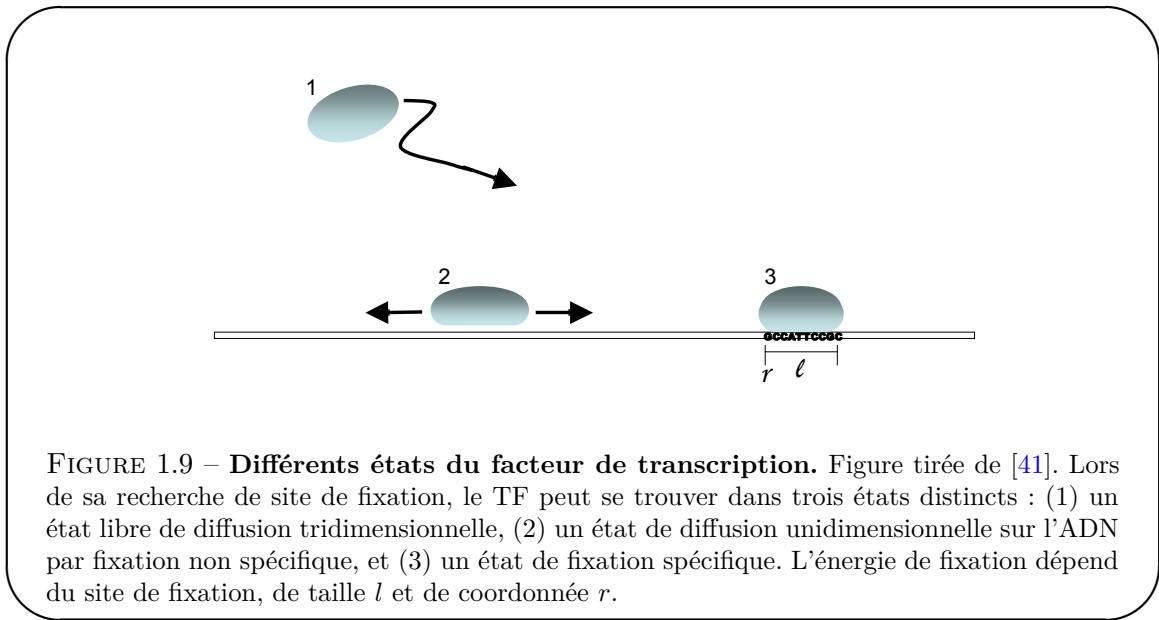


FIGURE 1.9 – Différents états du facteur de transcription. Figure tirée de [41]. Lors de sa recherche de site de fixation, le TF peut se trouver dans trois états distincts : (1) un état libre de diffusion tridimensionnelle, (2) un état de diffusion unidimensionnelle sur l'ADN par fixation non spécifique, et (3) un état de fixation spécifique. L'énergie de fixation dépend du site de fixation, de taille l et de coordonnée r .

diffuse dans un paysage d'énergie E_{ns} plat lorsqu'il est dans sa conformation de fixation non spécifique, ou dans un paysage d'énergie $E_s(r)$ dans sa conformation de fixation spécifique. Cela permet au facteur d'échantillonner les sites de faible énergie $E_s(r)$ tout en évitant les barrières de haute énergie en passant en mode de recherche non spécifique. Ce processus s'avère au final très efficace [24, 58]. Les temps de recherche sont typiquement inférieurs à une minute, ce qui est petit devant les processus de régulation de la cellule qui se déroulent au mieux sur quelques minutes. Il est donc pertinent de décrire l'effet d'un site de fixation sur la régulation d'un gène cible par la probabilité qu'il a de fixer un TF à l'équilibre thermodynamique.

1.3.2 Modèle PWM

Présenté en 1987 par Berg et von Hippel [7], le modèle PWM est le modèle le plus simple décrivant l'énergie de fixation spécifique entre un facteur de transcription et un site de fixation sur l'ADN. Ce modèle repose sur plusieurs hypothèses. Tout d'abord, il y a l'hypothèse importante que les sites de fixation des TFs sur l'ADN ont été sélectionnés au cours de l'évolution pour leur propriété de sites de reconnaissance, qu'elle que soit la concentration du TF dans la cellule. En d'autres termes, le processus de sélection discrimine les sites de fixation sur la seule base de leur énergie de fixation à un TF donné : les sites ayant une énergie fixation dans une certaine gamme sont retenus, les autres rejettés. Par ailleurs, au sein de cette gamme d'énergie « utile », toutes les séquences sont équiprobables. Enfin, la dernière hypothèse est que chaque nucléotide d'un site de fixation contribue de manière indépendante, c'est-à-dire additive à l'énergie totale du site. Cette hypothèse permet de simplifier le problème en gardant le nombre de paramètres petit. L'argument de Berg et von Hippel est que ce problème est analogue à celui de physique statistique consistant à déduire les taux d'occupation des niveaux d'énergie de particules indépendantes sachant que l'énergie totale doit avoir une certaine valeur moyenne E . La solution de ce problème est donnée par la formule de Boltzmann reliant énergie et taux d'occupation :

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

où $f_{i,b}$ est la probabilité d'observer la base b à la position i du site de fixation, $E_{i,b}$ est

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

La connaissance des fréquences des bases permet de définir une autre quantité utile caractérisant la variabilité des séquences de fixation, l'information relative des sites par rapport à une séquence d'ADN aléatoire [61] :

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

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

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

L'énergie totale d'un site S_i est alors

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

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

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

1.3.3 Modèle biophysique

Dans le paragraphe précédent, nous avons vu que le modèle PWM est basé sur une hypothèse forte, celle que les sites de fixation ont été sélectionnés sur la base de leur seule affinité ou énergie envers un TF. Néanmoins, à aucun moment n'intervient la concentration du TF dans la

1.3. Modèles mathématiques des interactions protéine-ADN

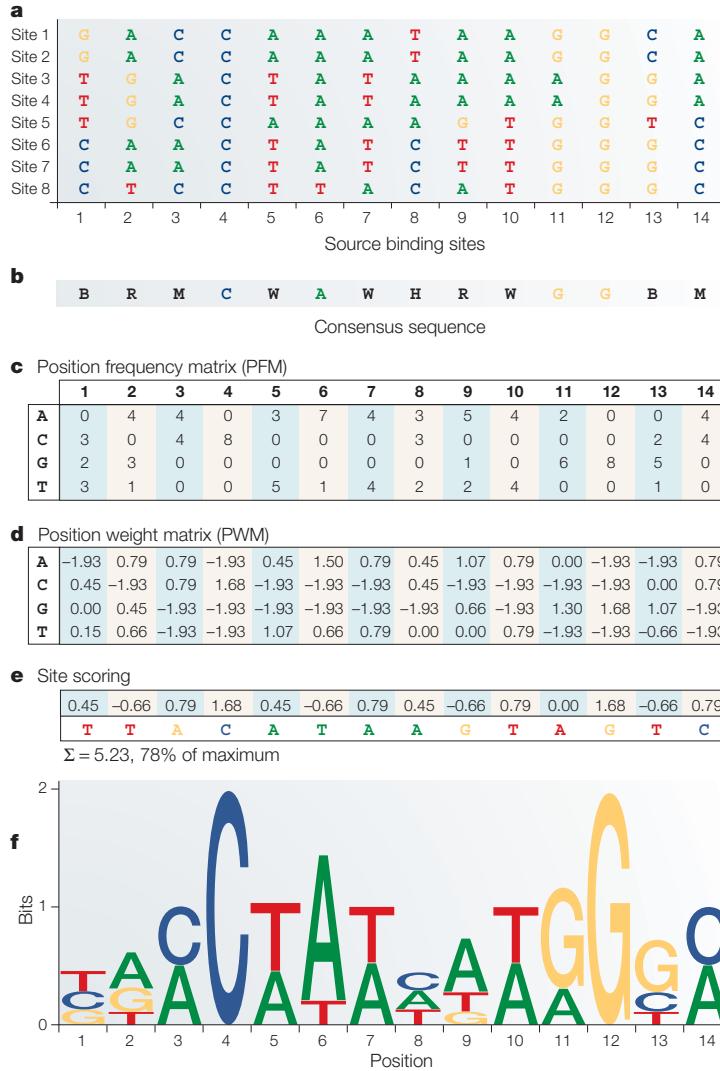


FIGURE 1.10 – Construction et utilisation du modèle PWM. Figure tirée de [70]. (a) Supposons connus un certain nombre de sites de fixation d'un facteur de transcription (dans ce cas MEF2). (b) Séquence consensus correspondante utilisant les symboles IUPAC. (c) Une matrice de fréquence est construite, indiquant pour chaque nucléotide sa multiplicité à une position donnée dans le site. (d) La PWM est simplement construite en prenant le logarithme relatif des fréquences PWMs par rapport aux fréquences *a priori* des nucléotides. (e) Le score (ou énergie) d'une séquence d'ADN donnée est calculé en additionnant les poids PWMs correspondant. (f) La PWM peut être représentée sous forme de logo [25]. Dans cette représentation, la hauteur d'une colonne représente le contenu en information ou information relative moyenne d'une position, et la taille des bases reflète leur fréquence observée.

cellule, dont dépend pourtant la probabilité de fixation. C'est ce que tente de capturer le modèle biophysique [24, 19, 77].

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

$$TF + S_i \rightleftharpoons TF : S_i \quad (1.5)$$

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

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

où $E_i = -\ln(K_i)$ est l'énergie libre standard de fixation (souvent notée ΔG), et $\mu = \ln[TF]$ est le potentiel chimique, ces deux quantités étant exprimées en kT . Ici nous avons considéré qu'il n'y avait qu'un seul site de fixation. De manière générale, le site est en compétition avec le fond génomique, ce qui ajoute une contribution à μ (voir description thermodynamique). À l'instar du modèle PWM, l'énergie E_i est généralement prise comme étant une fonction additive des énergies individuelles des différentes bases du site. Ainsi, lorsque le TF est à faible concentration ($\mu \rightarrow -\infty$), le modèle biophysique écrit en équation 1.7 se réduit au modèle PWM.

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 [24]. On pourra par ailleurs se référer à l'excellente revue [41]. Considérons le cas simple d'un seul facteur de transcription interagissant avec un génome de taille $L \gg 1$ ne contenant qu'un seul site fonctionnel, le reste de la séquence étant aléatoire. La protéine se fixe à l'ADN avec une probabilité 1/2. Lorsqu'elle est fixée, elle est à l'équilibre entre le mode spécifique et le mode non spécifique. Nous désirons savoir avec quelle probabilité elle est fixée de manière spécifique. La fonction de partition, énumérant tous les poids de Boltzmann associés aux différents états accessibles au TF fixé, s'écrit :

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

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

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

où Z_0 est la fonction de partition d'une séquence aléatoire, et nous avons introduit l'énergie E_i définie par

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

Dans le cas d'un site de reconnaissance, $E_s(i) \gg E_{ns}$ de sorte que $E_i \simeq E_s(i)$ [24] (ZZ Check ZZ). La probabilité que le facteur soit fixé sur le site fonctionnel s'écrit finalement :

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

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

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

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

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

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

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

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

La valeur de μ est liée à la fois au nombre de TFs ainsi qu'à la possibilité de se fixer dans le fond génomique. Elle est bien approximée par [24]

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

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

1.4 Mesures expérimentales des interactions protéine-ADN

Ces dernières années, des avancées technologiques considérables ont permis d'une part d'établir des modèles de fixation spécifique pour de nombreux TFs, d'autre part de localiser leurs sites de fixation dans le génome. Ces avancées ont eu lieu autant sur le plan *in vitro*, utilisant protéines purifiées et séquences nucléiques artificielles pour déduire l'affinité protéine-ADN, que sur le plan *in vivo*, mesurant l'interaction de la protéine avec l'ADN génomique [63].

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 [45]. Cette technique repose sur l'utilisation d'un système microfluidique composé de chambres dans lesquelles un fluide dont on peut facilement modifier la composition circule dans des canaux d'un diamètre de l'ordre de $1\mu\text{m}$ dont le microenvironnement est ainsi finement contrôlé. Dans ce cas, le fluide contient des gènes synthétique codant pour le TF ainsi que du matériel permettant la synthèse de la protéine au sein de la chambre, évitant l'étape de purification du TF. Chaque chambre du système contient des anticorps attachés à la surface permettant de capturer le TF et une certaine concentration d'une séquence d'ADN spécifique contenant une marque fluorescente. Le système contient ainsi des centaines de séquences d'ADN différentes, chacune étant présente à différentes concentrations. Lorsque le TF est fixé par les anticorps, il recrute des séquences d'ADN selon leur affinité. Celles qui ne se fixent pas sont lavées. Au final, les séquences fixées produisent un signal de fluorescence. La comparaison des signaux pour différentes concentrations d'ADN donne accès au rapport des constantes d'équilibre K_{eq} (eq. 1.6). La comparaison avec une séquence référence dont la constante K_{eq} est connue permet alors de déterminer le K_{eq} absolu pour chaque séquence de fixation.

En utilisant 17 systèmes de ce type, ils ont ainsi pu mesurer l'affinité de 4 TFs de type bHLH à 464 séquences d'ADN différentes : les séquences consensus et des séquences ayant une, deux, trois ou quatre mutations. À titre de comparaison, ils ont construit une PWM à partir des séquences contenant une seule mutation, puis ont prédit les énergies attendues des séquences à plusieurs mutations. La prédiction de la PWM s'est avérée bonne dans seulement 56% des cas pour les séquences à deux mutations, 10% pour les séquences à 3 mutations et 0% des cas pour les séquences à 4 mutations, montrant les limites de ce modèle indépendant confronté à des données d'interactions d'ordre supérieur. Ainsi, un modèle plus raffiné prenant en compte l'énergie d'interaction non spécifique et incluant des interactions entre nucléotides voisins permet de rendre compte des valeurs observées [62]. Nous reviendrons amplement sur la nécessité de prendre en compte les interactions entre paires de nucléotides lors de l'interaction spécifique entre TF et ADN dans le chapitre 2.

- **Approche physique : la microscopie SPR**

La méthode de résonance des plasmons de surface (SPR en anglais) est habituellement utilisée pour étudier l'interaction d'une protéine avec un ligand (qui peut être une autre protéine), mais elle peut aussi être utilisée pour mesurer les interactions entre une protéine et quelques centaines de séquences d'ADN différentes [56, 14]. Le principe de la microscopie SPR est que l'angle de réflexion de la lumière sur une fine surface d'or, par exemple, dépend de la masse de molécules attachées de l'autre côté de sa surface. Si de l'ADN est attaché à la surface, la fixation du TF induit un changement d'angle de reflection lumineuse mesurable au cours du temps. Ainsi, la cinétique de fixation du TF jusqu'à l'atteinte de l'équilibre est accessible. On peut de même étudier la dissociation du TF lors du lavage de la surface. Ces mesures donnent directement accès aux taux d'association k_{on} et de dissociation k_{off} que la simple mesure de la constante d'équilibre $K_{eq} = k_{on}/k_{off}$ ne permet habituellement pas de déterminer.

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

L'analyse de fixation des protéines par puce à ADN (*Protein-Binding Microarray* ou PBM) est une technologie haut débit qui a été développée au cours des 10 dernières années [9]. Les puces sont composées de 44,000 puits auxquels sont fixés des brins d'ADN. Une puce contient tous les sites de fixation de 10bp possibles, groupés dans les puits en fonction de leur coeur de 8bp (32,768 séquences en comptant les deux brins) avec possibilité de distinguer les bases flanquantes dans certains cas spécifiés. Un TF, purifié à partir de cellules ou synthétisé *in vitro*, est ajouté à la puce, qui est ensuite lavée pour se débarrasser des fixations non spécifiques. La quantité de protéine fixée à un puits donné est déterminée grâce à un anticorps fluorescent contre la protéine. L'enrichissement en protéine est calculé relativement au bruit de fond (anticorps non spécifique par exemple). Il est alors possible d'utiliser ces mesures pour bâtir une PWM du TF (voir par exemple [38]).

Une autre méthode utilise aussi des puces à ADN incluant toutes les séquences de 10 bases possibles : c'est l'identification de site apparenté (*Cognate Site Identifier* ou CSI) [69]. Une différence technique avec les PBMs est que l'ADN est d'abord synthétisé en simple brin puis se replie en double brin pour former le site de fixation, évitant ainsi de devoir générer l'ADN double brin à partir de précurseurs. Par ailleurs, le TF est en compétition avec un marqueur fluorescent qui peut se fixer à l'ADN : il n'est donc pas nécessaire d'utiliser un marquage spécifique sur le TF ou sur un anticorps, ce qui rend la procédure plus généralisable. Finalement, la spécificité du TF est représentée par un “paysage de spécificité” qui encapsule l'information de fluorescence de l'ensemble des variations par rapport à une séquence consensus dans une représentation simple [15].

- **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* [50, 66, 11, 75]. Une bibliothèque de sites de fixation potentiels est d'abord générée en synthétisant des séquences d'ADN aléatoires ou en utilisant des séquences génomiques. Les bouts de ces séquences contiennent des précurseurs permettant l'amplification exponentielle par PCR. Le TF purifié est ajouté aux sites et les séquences fixées sont séparées des séquences non fixées, par exemple par retard sur gel. Après un cycle de sélection, les séquences récupérées sont enrichies en séquences de basse affinité pour le TF, car celles-ci sont simplement initialement bien plus abondantes que les séquences de haute affinité. Afin d'augmenter la proportion de séquence de grande affinité, les séquences filtrées sont amplifiées avant d'être filtrées à nouveau, ceci sur plusieurs cycles. À la fin de ce processus, les séquences sélectionnées sont clonées et séquencées, résultant en un nombre typique de moins de ~ 100 séquences indépendantes [22]. 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 [52], permettant de réaliser des modèles plus précis [47].

Depuis la méthode SELEX a été mise au point, des avancées considérables ont été faites dans les techniques de séquençage, permettant l'obtention de millions de séquences à la fois : on parle de séquence haut-débit (*high-throughput*) ou encore séquençage massivement parallèle. L'utilisation de ces nouvelles techniques dans l'expérience SELEX a mené à la méthode HT-SELEX [47], aussi appelée Bind-n-Seq [79]. Il est alors possible d'estimer un modèle d'énergie à

partir des fréquences d'observation des différentes séquences dès le premier cycle [47]. 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 bp comme c'est souvent le cas chez la bactérie. Cette technique a récemment été poussée encore plus loin [34]. En utilisant des protéines marquées, les auteurs ont réalisé un HT-SELEX à partir d'extraits cellulaires, et en utilisant un code barre aux séquences d'ADN de chaque expérience, ils ont pu analyser les sites de fixation pour plusieurs TFs en parallèle. Ils ont récemment pu utiliser cette technique pour obtenir des modèles de spécificité pour 411 TFs humains, la plus grande étude de ce genre réalisée à ce jour [35].

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 dans lesquelles le facteur de transcription se fixe à la proie expriment *HIS3*, croissent et forment des colonies, d'où la notion de gène de sélection. Par ailleurs, la stringence de la sélection peut être modulée en ajoutant au milieu différentes concentrations de 3-amino-triazole (3-AT), un inhibiteur de *HIS3*. De cette façon l'affinité du site de fixation peut être estimée plus finement.

Dans les études de ce type, les sites de fixation présents au sein des colonies sont séquencés individuellement, ce qui permet d'obtenir environ 50 séquences pour une expérience de sélection donnée []. Néanmoins, il semble possible d'utiliser les nouvelles technologies de séquençage pour récupérer l'ensemble des sites de fixation des bactéries présentes sur une plaque [63]. 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

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

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

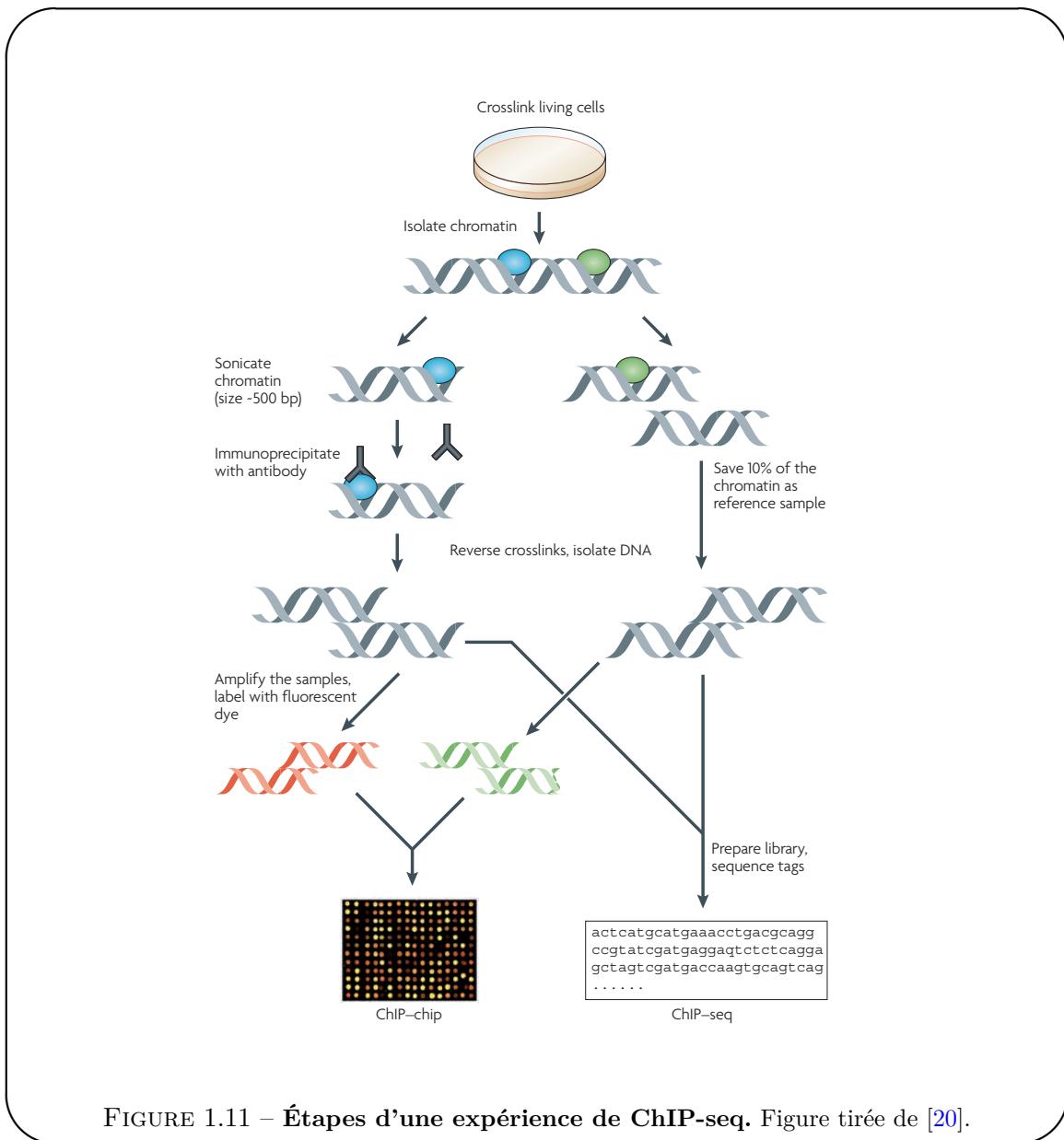


FIGURE 1.11 – Étapes d'une expérience de ChIP-seq. Figure tirée de [20].

La technique d'immunoprécipitation de la chromatine (ChIP) (fig. 1.11) consiste dans un premier temps à induire la réticulation (*crosslink*) des protéines se liant à l'ADN avec l'ADN en

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

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

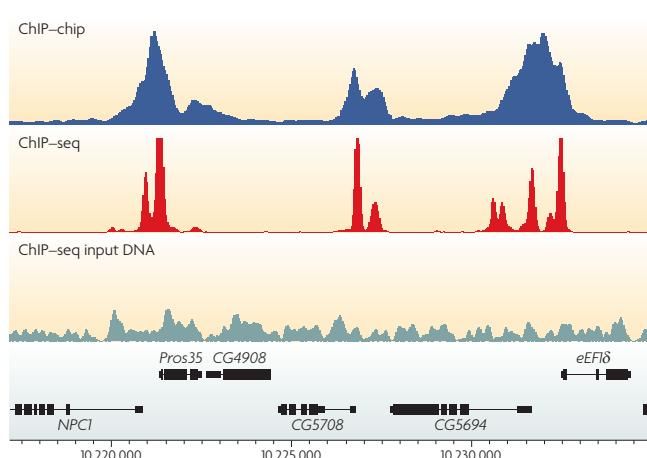


FIGURE 1.12 – Résolution des expériences ChIP-on-chip et ChIP-seq. Figure tirée de [51]. Exemples de profils de fixation de la protéine à chromodomaine Chromator générés à partir d'expériences de ChIP-on-chip (intensité relative par rapport au contrôle, bleu) et de ChIP-seq (densité de séquences, rouge) dans la lignée cellulaire S2 de *Drosophila Melanogaster*. On peut noter la plus grande résolution de l'expérience ChIP-seq pour déterminer les sites de liaison. L'ADN utilisé en *input* de l'expérience de ChIP-seq et servant de contrôle est montré en gris.

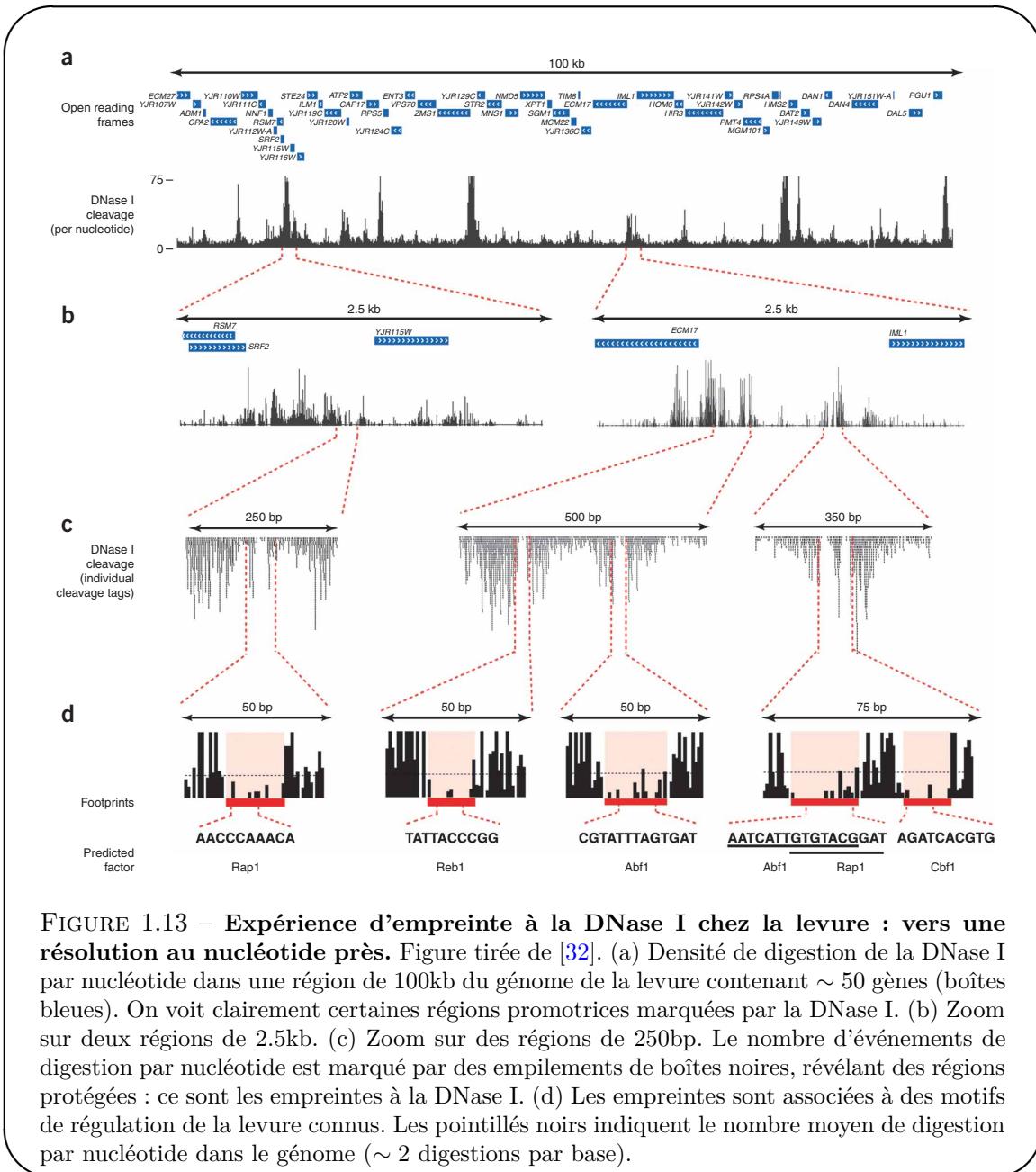
Dans le cas du ChIP-seq, l'échantillon immunoprécipité est analysé par séquençage à haut débit, résultant en une librairie de *reads* d'une longueur typique variant entre 27 et 50bp issus des bouts des séquences. Ces *reads* sont ensuite alignés sur un génome de référence. À chaque position du génome correspond ainsi un certain nombre de séquences précipitées et d'*input*. En comparant ce nombre au nombre moyen dans le locus et à l'*input*, il est possible d'identifier des pics correspondant à la fixation du facteur (voir par exemple le programme d'appel de pics ChIP-seq MACS [76]).

Dans les deux cas, il faut noter que l'on a affaire à la fixation *moyenne* du facteur sur

1.4. Mesures expérimentales des interactions protéine-ADN

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

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



Contrairement aux techniques précédentes, l'empreinte à la DNase 1 ne repose pas sur l'étude d'un facteur de transcription précis, mais permet au contraire d'avoir un ensemble de sites de fixation dans le génome pour un type cellulaire donné, avec une précision au nucléotide près. Cette méthode repose sur le fait que la fixation stable des facteurs de transcription à l'ADN n'est possible que si la région est pauvre en nucléosomes, les protéines autour desquelles s'enroule l'ADN : on parle de région de chromatine ouverte. Ces régions sont préférentiellement digérées par l'endonucléase DNase I. Étant donné que la majorité de l'ADN est enroulé autour de nucléosomes, les sites hypersensibles à la digestion par DNase I (*DNaseI-hypersensitive* ou DHS) correspondent essentiellement à des régions de chromatine ouverte ayant des rôles de régulation génétique : promoteurs, enhancers...

En combinant la technique de DHS avec le séquençage à haut débit, l'expérience de DNase-seq permet d'identifier tous les types de région de régulation à l'échelle du génome [65]. 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 *DNaseI 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

1.5.1 Modules et fonctions logiques

1.5.2 Encodage de patterns spatiaux

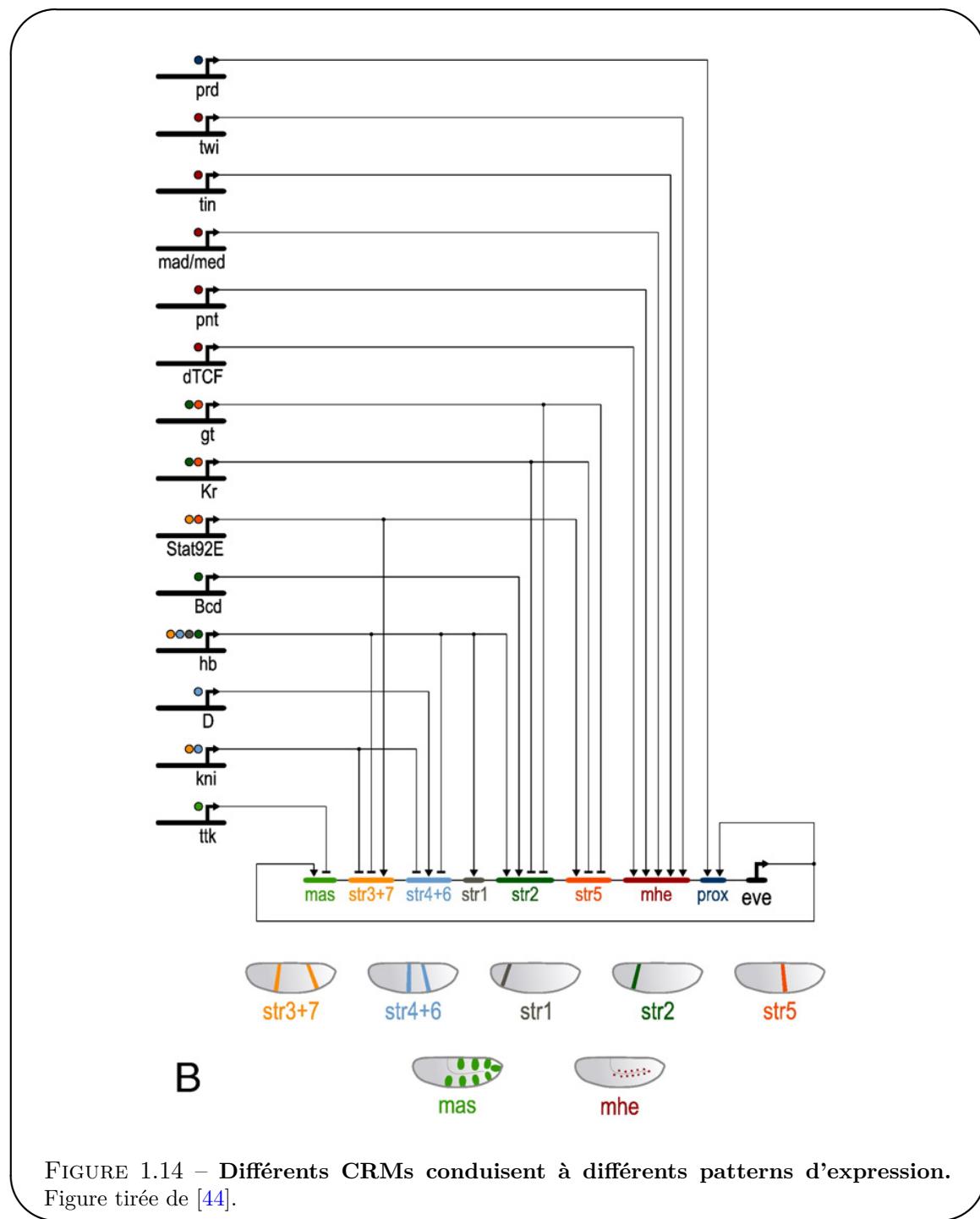


FIGURE 1.14 – Différents CRMs conduisent à différents patterns d'expression.
Figure tirée de [44].

1.5.3 Différents états des CRMs

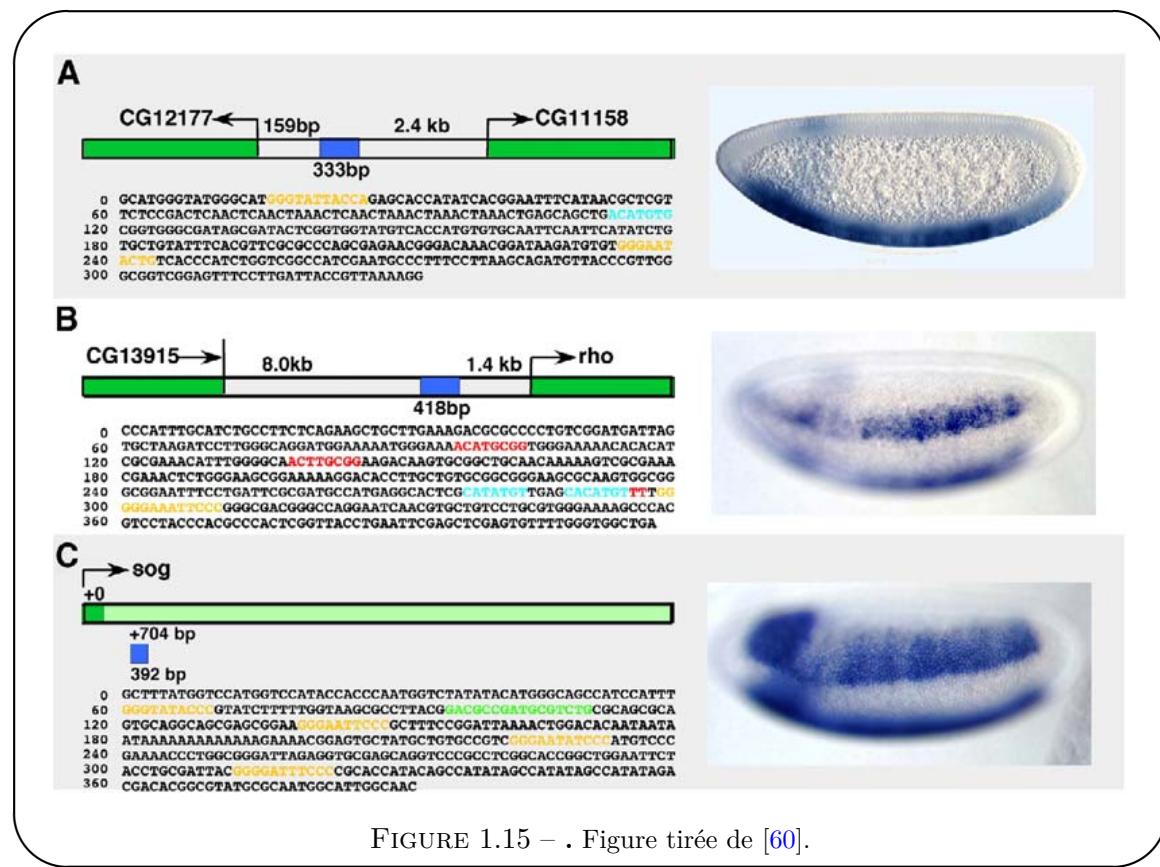


FIGURE 1.15 – . Figure tirée de [60].

1.5.4 Prédiction des CRMs

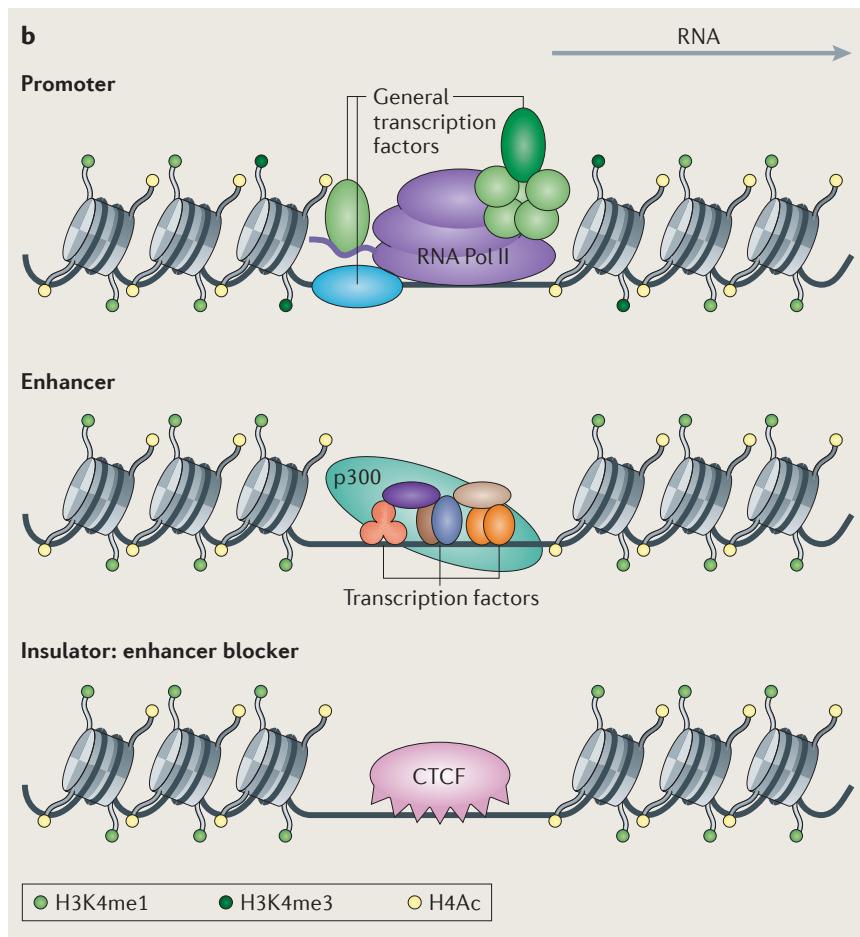


FIGURE 1.16 – Les états épigénétiques des CRMs. Figure tirée de [31].

1.5.5 Grammaire des enhancers : enhanceosome vs billboard

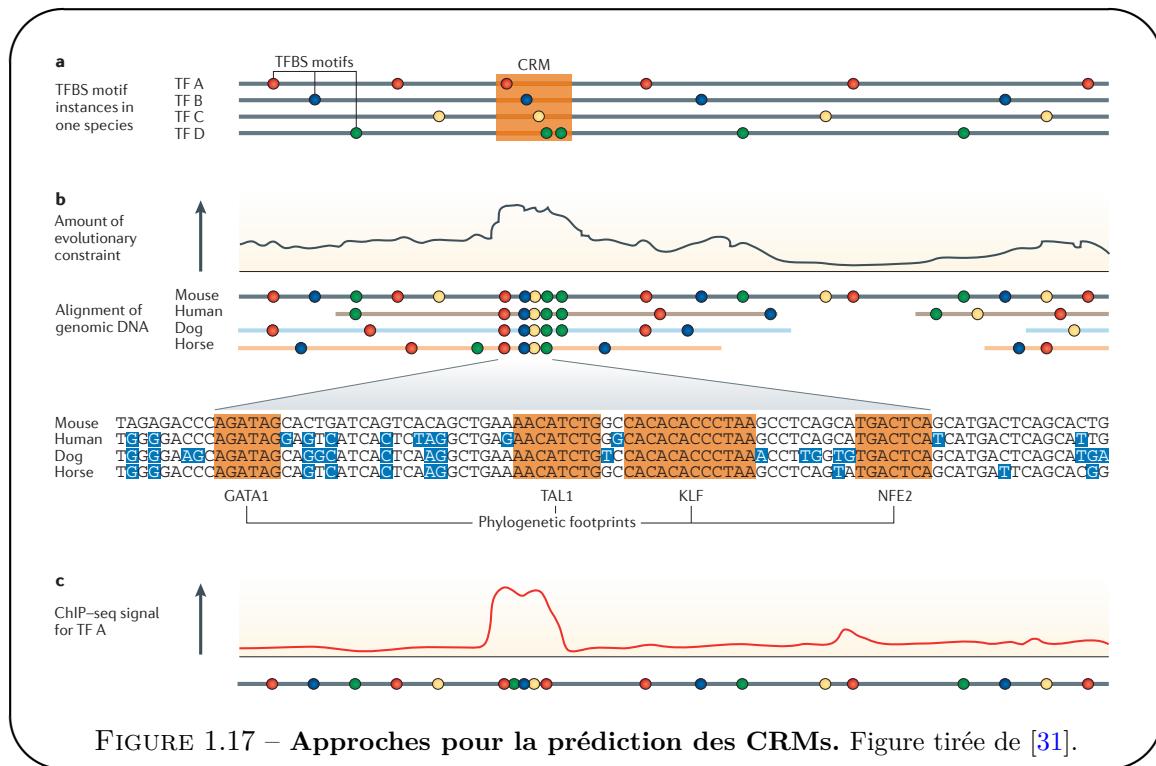


FIGURE 1.17 – Approches pour la prédition des CRMs. Figure tirée de [31].

1.5.6 Évolution des enhancers

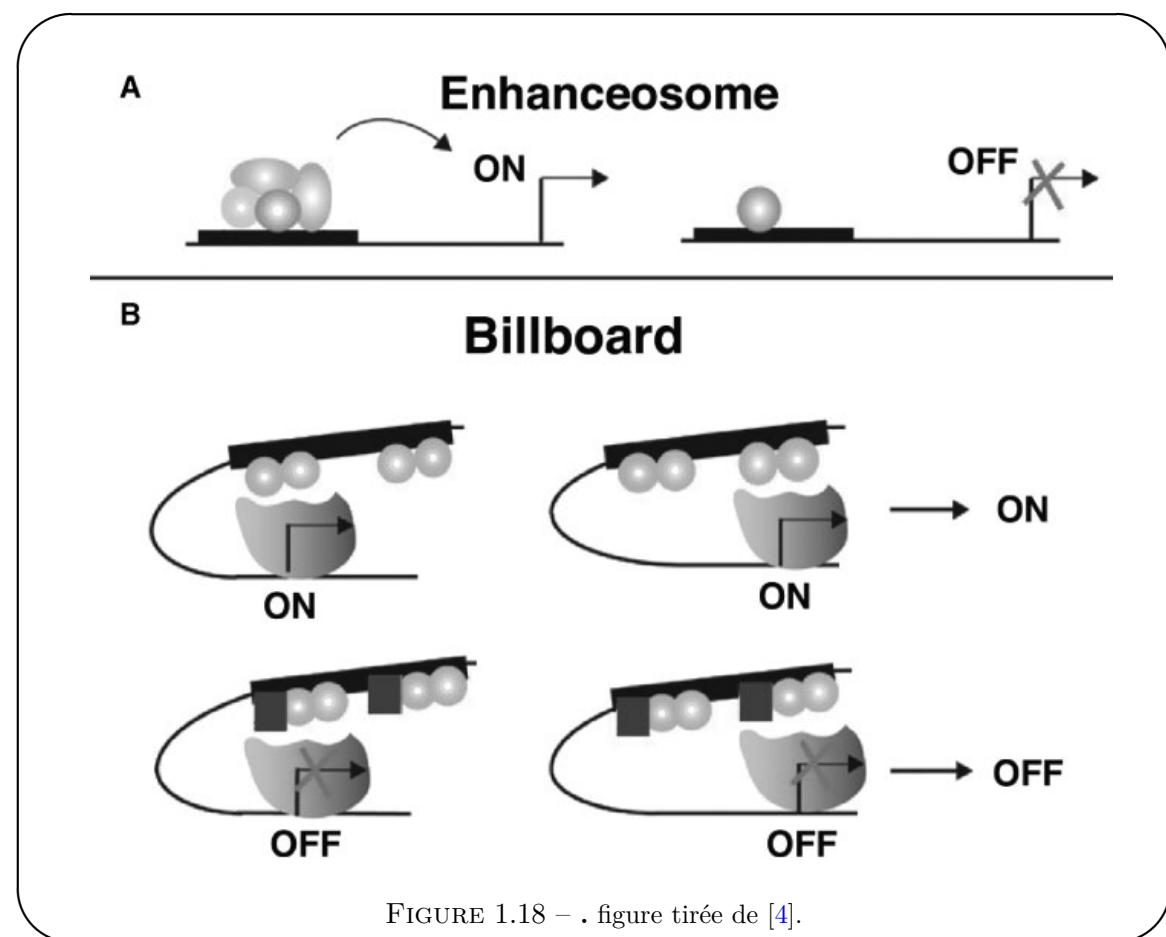


FIGURE 1.18 – . figure tirée de [4].

[21]

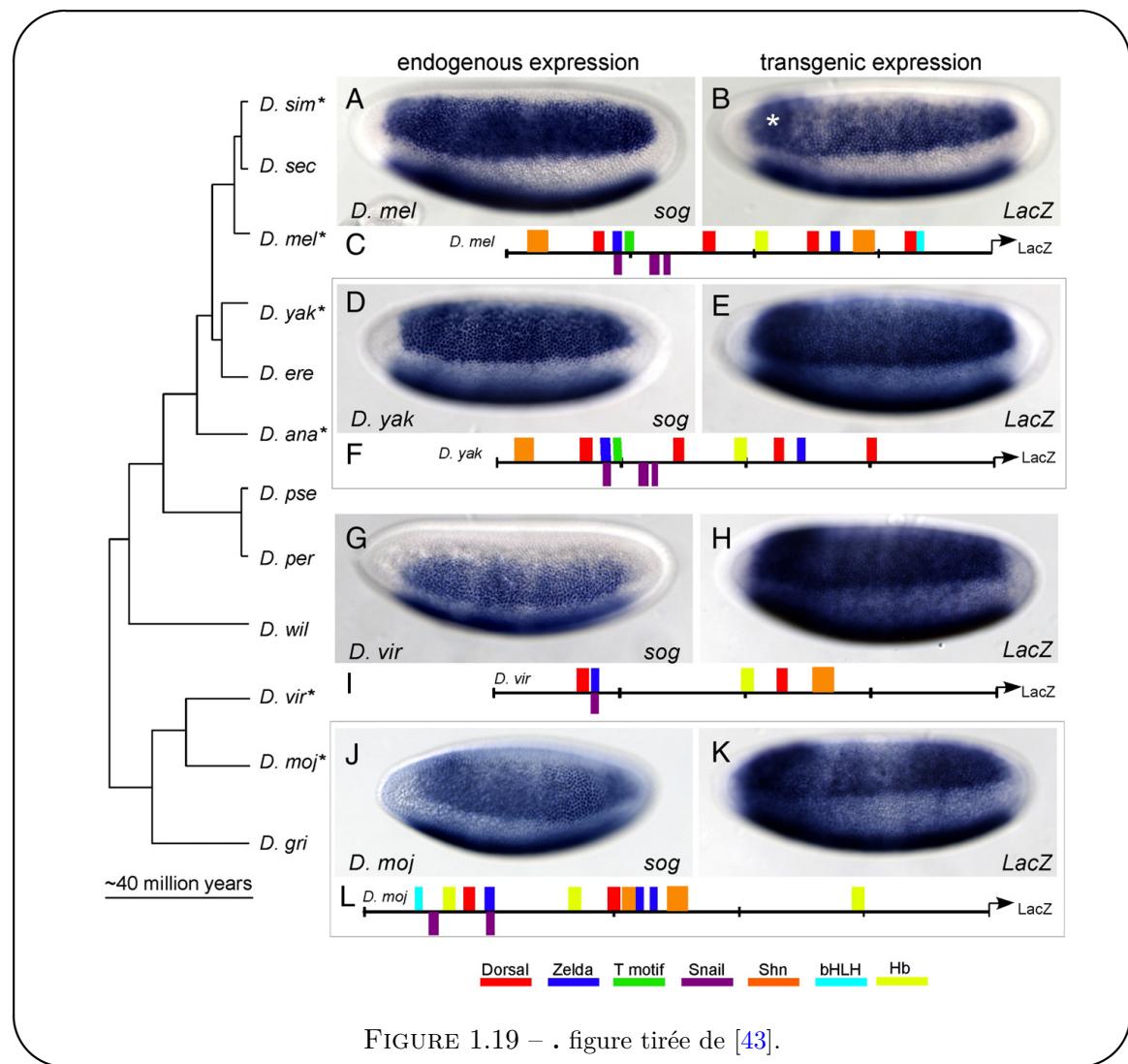


FIGURE 1.19 – . figure tirée de [43].

1.5.7 Les « shadow enhancers »

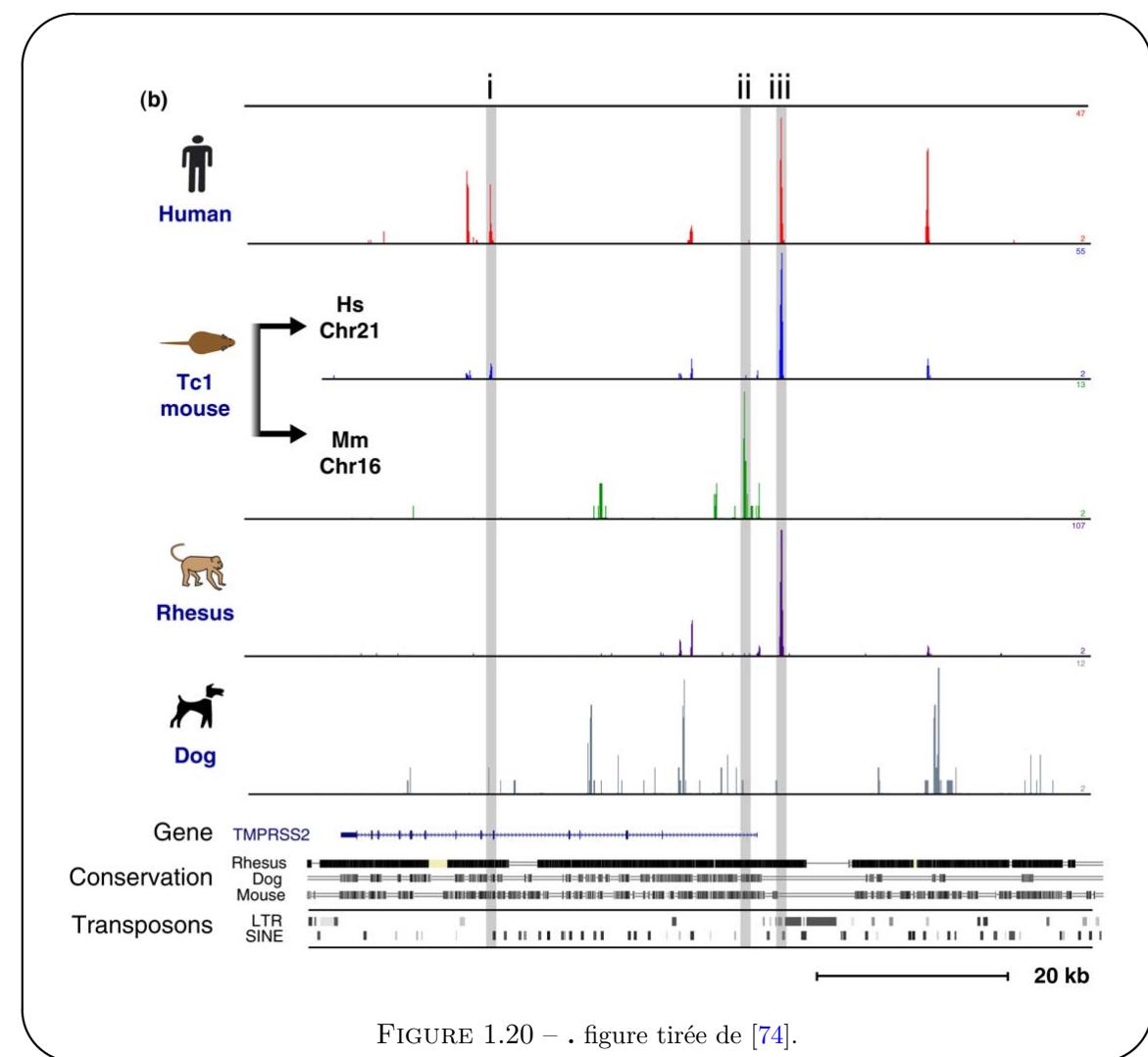


FIGURE 1.20 – . figure tirée de [74].

1.5.8 Validation expérimentale

1.6 Banques de données

1.6.1 Séquences génomiques et alignements

statistiques du genome (lognormal)

1.6.2 Annotations (TSSs, repeats...)

1.6.3 Jaspar et Transfac

1.6.4 Visualisation sur UCSC

1.6.5 Le projet ENCODE

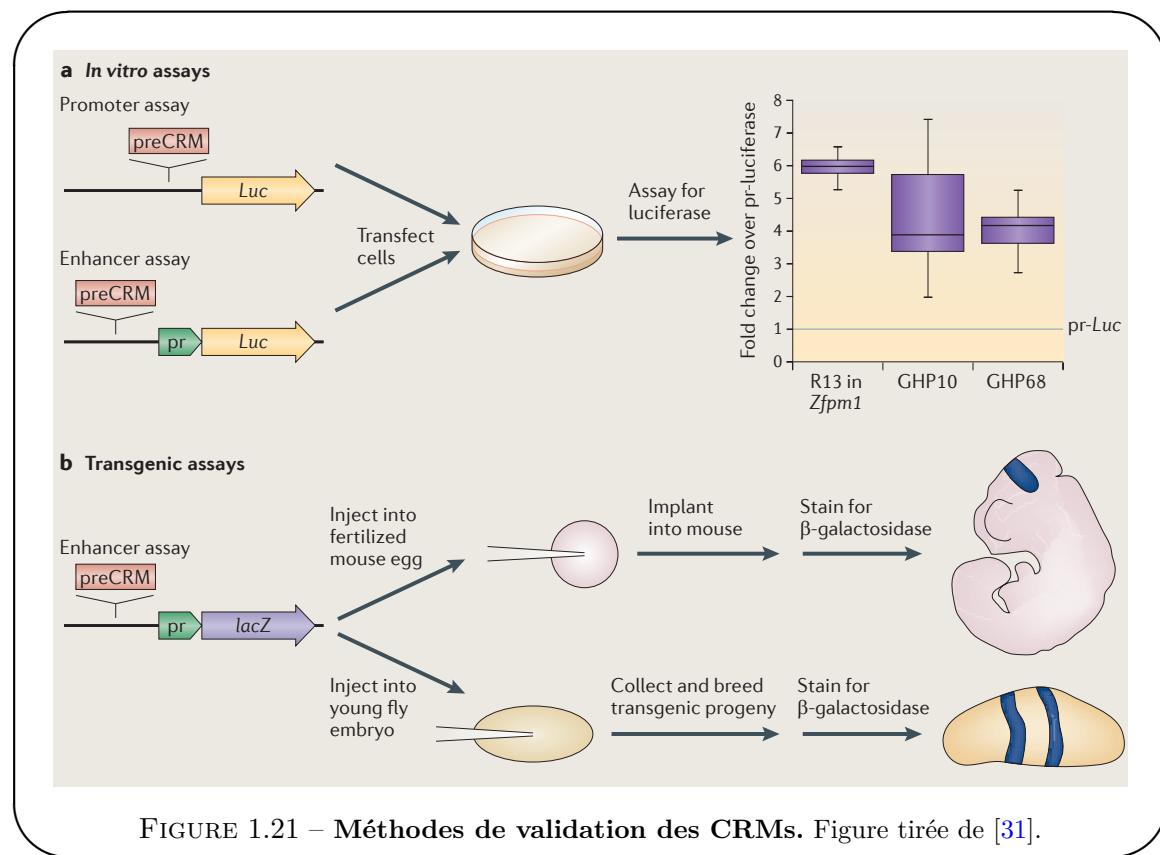


FIGURE 1.21 – Méthodes de validation des CRMs. Figure tirée de [31].

Chapitre 2

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

3.1	43
------------	-------	----

Introduction du chapitre 2

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

• L'énergie de fixation. Les Facteurs de Transcription peuvent s'accrocher à l'ADN. La fixation est décrite par une énergie qui peut se décomposer en deux composantes. L'une est indépendante de la séquence et prend en considération la courbure de l'ADN etc. L'autre dépend de la séquence. Cette dernière peut être décrite par divers modèles de fixation.

- **Description des modèles existants.**
- Différentes données biologiques utilisées : PBM, SELEX, ChIP.
- Différences in vitro et in vivo.

2.1 Les modèles de fixation

2.1.1 Modèles de maximum d'entropie

La théorie de l'information offre un cadre conceptuel permettant de déterminer les probabilités d'un ensemble d'états étant données plusieurs contraintes mesurables, ou *observables*. L'étape clé consiste à maximiser une fonctionnelle connue sous le nom d'entropie [33, 54] sur l'ensemble des distributions de probabilités des états étant données les contraintes imposées. Cette fonctionnelle s'écrit [57]

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

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

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

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

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

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

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

La solution peut finalement se mettre sous la forme

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

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

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

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

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

- Le modèle PWM
 - Le modèle de corrélation de paires
- Fixation de jauge.

2.1.2 Modèles de mélange

2.2 Description des données biologiques

2.2.1 Les données ChIP

Les données que nous utilisons proviennent d'expériences ChIP-on-chip réalisées chez la mouche (*Drosophila Melanogaster*) et d'expériences ChIP-seq réalisées chez la souris (*Mus Musculus*). Ces données ont été récupérées à partir de la littérature [78, 16] et à partir des données du projet ENCODE [17] accessibles à partir du site internet de UCSC³, pour un total de 27 Facteurs de Transcription. Parmi eux, il y a 5 Facteurs de Transcription impliqués dans le développement de la mouche : Bap, Bin, Mef2, Tin, Twi, 11 Facteurs de Transcription régulant les cellules souches chez les mammifères : c-Myc, E2f1, Esrrb, Klf4, Nanog, n-Myc, Oct4, Sox2, Stat3, Tcfcp2l1, Zfx, et 11 facteurs impliqués dans la myogenèse chez les mammifères : Cebpb, E2f4, Fosl1, Max, MyoD, Myog, Nrsf, Smad1, Srf, Tcf3, Usf1. Au total, il y a entre 678 et 38292 pics de ChIP, avec une taille moyenne de 280bp.

Les séquences d'ADN peuvent contenir un certain nombre de séquences « polluantes » peu informatives issues de rétrotransposons ou de duplication excessives de dinucléotides. Ces séquences répétées, ou *repeats*, sont en grand nombre et peuvent donc biaiser la statistique lors de la recherche de sites de fixation. Pour éviter ce biais, ces séquences ont été masquées à l'aide du logiciel RepeatMasker [59].

2.2.2 Statistique « background » des séquences

Présence de corrélations.

2.3 Présentation de l'algorithme

Descente de gradient.

2.4 Performance des modèles

2.5 Analyse des corrélations

2.5.1 Quantification par l'Information Directe

2.5.2 Description par des patterns de Hopfield

2.6 Comparaison avec des données *in vitro*

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

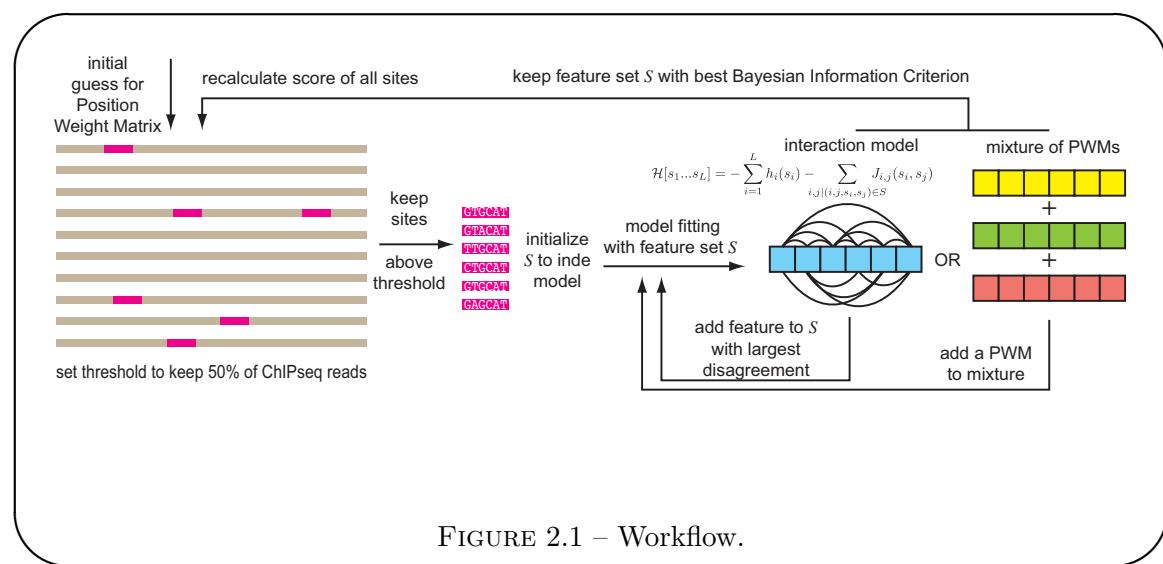
2.6. Comparaison avec des données *in vitro*

FIGURE 2.1 – Workflow.

2.6.1 Conclusion de la section 2.6

[**thèse**: version du mardi 28 mai 2013 à 17 h 13]

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

Chapitre 3

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

5.1	51
-----	-------	----

Introduction du chapitre 3

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

3.1

Chapitre 4

Étude de la différenciation épidermale chez la drosophile

Introduction du chapitre 4

4.1

Conclusion du chapitre 4

Chapitre 5

Étude de la différenciation musculaire chez la souris

Introduction du chapitre 5

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

5.1

Conclusion du chapitre 5

Conclusion

Résumé

Perspectives

Bibliographie

Dans la version pdf, les numéros de page sont des liens qui renvoient à l'occurrence de la citation dans le texte.

- [1] 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, Apr 2001. (Pages 8 et 9.)
- [2] ALON, U. An introduction to systems biology : Design principles of biological circuits (mathematical and computational biology series vol 10), 2007. (Page 11.)
- [3] ALON, U. Network motifs : theory and experimental approaches. *Nat Rev Genet* 8, 6 (2007), 450–461. (Page 11.)
- [4] ARNSTI, D. N., AND KULKARNI, M. M. Transcriptional enhancers : Intelligent enhan- ceosomes or flexible billboards ? *J Cell Biochem* 94, 5 (Apr 2005), 890–8. (Page 31.)
- [5] BARTEL, D. P. Micornas : target recognition and regulatory functions. *Cell* 136, 2 (Jan 2009), 215–33. (Page 11.)
- [6] BAYLIES, M. K., BATE, M., AND RUIZ GOMEZ, M. Myogenesis : a view from drosophila. *Cell* 93, 6 (Jun 1998), 921–7. (Page 13.)
- [7] BERG, O., AND VON HIPPEL, P. Selection of dna binding sites by regulatory proteins : Statistical-mechanical theory and application to operators and promoters. *Journal of molecular biology* 193, 4 (1987), 723–743. (Page 15.)
- [8] BERG, O. G., WINTER, R. B., AND VON HIPPEL, P. H. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. models and theory. *Biochemistry* 20, 24 (Nov 1981), 6929–48. (Page 14.)
- [9] BERGER, M. F., PHILIPPAKIS, A. A., QURESHI, A. M., HE, F. S., ESTEP, 3RD, P. W., AND BULYK, M. L. Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* 24, 11 (Nov 2006), 1429–35. (Page 21.)
- [10] BIRD, A. Dna methylation patterns and epigenetic memory. *Genes Dev* 16, 1 (Jan 2002), 6–21. (Page 9.)
- [11] BLACKWELL, T. K., AND WEINTRAUB, H. Differences and similarities in dna-binding preferences of myod and e2a protein complexes revealed by binding site selection. *Science* 250, 4984 (Nov 1990), 1104–10. (Page 21.)
- [12] BLAU, H. M., PAVLATH, G. K., HARDEMAN, E. C., CHIU, C. P., SILBERSTEIN, L., WEBSTER, S. G., MILLER, S. C., AND WEBSTER, C. Plasticity of the differentiated state. *Science* 230, 4727 (Nov 1985), 758–66. (Page 6.)
- [13] BRAZMA, A., PARKINSON, H., SCHLITT, T., AND SHOJATALAB, M. A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays. http://www.ebi.ac.uk/microarray/biology_intro.html, Oct 2001. (Page 4.)

Bibliographie

- [14] CAMPBELL, C. T., AND KIM, G. Spr microscopy and its applications to high-throughput analyses of biomolecular binding events and their kinetics. *Biomaterials* 28, 15 (May 2007), 2380–92. (Page 20.)
- [15] CARLSON, C. D., WARREN, C. L., HAUSCHILD, K. E., OZERS, M. S., QADIR, N., BHIMSARIA, D., LEE, Y., CERRINA, F., AND ANSARI, A. Z. Specificity landscapes of dna binding molecules elucidate biological function. *Proc Natl Acad Sci U S A* 107, 10 (Mar 2010), 4544–9. (Page 21.)
- [16] CHEN, X., XU, H., YUAN, P., FANG, F., HUSS, M., VEGA, V. B., WONG, E., ORLOV, Y. L., ZHANG, W., JIANG, J., LOH, Y.-H., YEO, H. C., YEO, Z. X., NARANG, V., GOVINDARAJAN, K. R., LEONG, B., SHAHAB, A., RUAN, Y., BOURQUE, G., SUNG, W.-K., CLARKE, N. D., WEI, C.-L., AND NG, H.-H. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133, 6 (Jun 2008), 1106–17. (Page 38.)
- [17] CONSORTIUM, E. P. A user's guide to the encyclopedia of dna elements (encode). *Plos Biol* 9, 4 (Apr 2011), e1001046. (Page 38.)
- [18] DAVIS, R. L., WEINTRAUB, H., AND LASSAR, A. B. Expression of a single transfected cdna converts fibroblasts to myoblasts. *Cell* 51, 6 (1987), 987–1000. (Page 6.)
- [19] DJORDJEVIC, M., SENGUPTA, A. M., AND SHRAIMAN, B. I. A biophysical approach to transcription factor binding site discovery. *Genome Res* 13, 11 (Nov 2003), 2381–90. (Page 17.)
- [20] FARNHAM, P. J. Insights from genomic profiling of transcription factors. *Nat Rev Genet* 10, 9 (Sep 2009), 605–16. (Page 23.)
- [21] FESCHOTTE, C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 9, 5 (May 2008), 397–405. (Page 31.)
- [22] FIELDS, D. S., HE, Y., AL-UZRI, A. Y., AND STORMO, G. D. Quantitative specificity of the mnt repressor. *J Mol Biol* 271, 2 (Aug 1997), 178–94. (Page 21.)
- [23] FURUSAWA, C., AND KANEKO, K. A dynamical-systems view of stem cell biology. *Science* 338, 6104 (Oct 2012), 215–217. (Page 5.)
- [24] GERLAND, U., MOROZ, J., AND HWA, T. 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 (2002), 12015. (Pages 15, 17, 18 et 19.)
- [25] GIOCOMO, L. M., MOSER, M.-B., AND MOSER, E. I. Computational models of grid cells. *Neuron* 71, 4 (Aug 2011), 589–603. (Page 17.)
- [26] GRAF, T., AND ENVER, T. Forcing cells to change lineages. *Nature* 462, 7273 (Dec 2009), 587–94. (Page 7.)
- [27] GREER, E. L., AND SHI, Y. Histone methylation : a dynamic mark in health, disease and inheritance. *Nat Rev Genet* 13, 5 (May 2012), 343–57. (Page 9.)
- [28] GURDON, J. B., AND MELTON, D. A. Nuclear reprogramming in cells. *Science* 322, 5909 (Dec 2008), 1811–5. (Page 6.)
- [29] HAMMOND, S. M., CAUDY, A. A., AND HANNON, G. J. Post-transcriptional gene silencing by double-stranded rna. *Nat Rev Genet* 2, 2 (Feb 2001), 110–9. (Page 11.)
- [30] HANNON, G. J. Rna interference. *Nature* 418, 6894 (Jul 2002), 244–51. (Page 11.)
- [31] HARTWELL, L., HOPFIELD, J., LEIBLER, S., AND MURRAY, A. From molecular to modular cell biology. *Nature* 402, 6761 (1999), 47. (Pages 29, 30 et 34.)

- [32] 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. Global mapping of protein-dna interactions in vivo by digital genomic footprinting. *Nat Methods* 6, 4 (Apr 2009), 283–9. (Page 25.)
- [33] JAYNES, E. Information theory and statistical mechanics. ii. *Physical review* 108, 2 (1957), 171. (Page 37.)
- [34] 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. Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. *Genome Res* 20, 6 (Jun 2010), 861–73. (Page 22.)
- [35] 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. Dna-binding specificities of human transcription factors. *Cell* 152, 1-2 (Jan 2013), 327–39. (Page 22.)
- [36] KAUFMANN, S. The origins of order, 1993. (Page 6.)
- [37] KEIM, C. N., MARTINS, J. L., ABREU, F., ROSADO, A. S., DE BARROS, H. L., BOROVJAK, R., LINS, U., AND FARINA, M. Multicellular life cycle of magnetotactic prokaryotes. *FEMS Microbiol Lett* 240, 2 (Nov 2004), 203–8. (Page 4.)
- [38] KINNEY, J. B., TKACIK, G., AND CALLAN, C. G. Precise physical models of protein-dna interaction from high-throughput data. *Proc Natl Acad Sci USA* 104, 2 (Jan 2007), 501–6. (Page 21.)
- [39] KONDO, T., PLAZA, S., ZANET, J., BENRABAH, E., VALENTI, P., HASHIMOTO, Y., KOBAYASHI, S., PAYRE, F., AND KAGEYAMA, Y. Small peptides switch the transcriptional activity of shavenbaby during drosophila embryogenesis. *Science* 329, 5989 (Jul 2010), 336–9. (Page 11.)
- [40] KULESSA, H., FRAMPTON, J., AND GRAF, T. Gata-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes Dev* 9, 10 (May 1995), 1250–62. (Page 6.)
- [41] LÄSSIG, M. From biophysics to evolutionary genetics : statistical aspects of gene regulation. *BMC Bioinformatics* 8, Suppl 6 (2007), S7. (Pages 15 et 18.)
- [42] LEE, T., RINALDI, N., ROBERT, F., ODOM, D., BAR-JOSEPH, Z., GERBER, G., HANNETT, N., HARBISON, C., THOMPSON, C., AND SIMON, I. Transcriptional regulatory networks in *saccharomyces cerevisiae*. *Science* 298, 5594 (2002), 799. (Pages 11 et 12.)
- [43] LIBERMAN, L. M., AND STATHOPOULOS, A. Design flexibility in cis-regulatory control of gene expression : Synthetic and comparative evidence. *Developmental Biology* 327, 2 (Mar 2009), 578–589. (Page 32.)
- [44] LIU, Y.-H., JAKOBSEN, J. S., VALENTIN, G., AMARANTOS, I., GILMOUR, D. T., AND FURLONG, E. E. M. A systematic analysis of tinman function reveals eya and jak-stat signaling as essential regulators of muscle development. *Developmental Cell* 16, 2 (Feb 2009), 280–91. (Pages 13 et 27.)
- [45] MAERKL, S., AND QUAKE, S. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315, 5809 (2007), 233. (Page 20.)

Bibliographie

- [46] MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D., AND ALON, U. Network motifs : simple building blocks of complex networks. *Science* 298, 5594 (Oct 2002), 824–7. (Page 11.)
- [47] NAGARAJ, V. H., O'FLANAGAN, R. A., AND SENGUPTA, A. M. Better estimation of protein-dna interaction parameters improve prediction of functional sites. *BMC Biotechnol* 8, 1 (Jan 2008), 94. (Pages 21 et 22.)
- [48] NURSE, P., AND HAYLES, J. The cell in an era of systems biology. *Cell* (Jan 2011). (Page 7.)
- [49] 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. Control of pancreas and liver gene expression by hnf transcription factors. *Science* 303, 5662 (Feb 2004), 1378–81. (Pages 11 et 12.)
- [50] OLIPHANT, A. R., BRANDL, C. J., AND STRUHL, K. 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 (Jul 1989), 2944–9. (Page 21.)
- [51] PARK, P. J. Chip-seq : advantages and challenges of a maturing technology. *Nat Rev Genet* 10, 10 (Oct 2009), 669–80. (Page 24.)
- [52] ROULET, E., BUSSO, S., CAMARGO, A. A., SIMPSON, A. J. G., MERMOD, N., AND BUCHER, P. High-throughput selex sage method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol* 20, 8 (Aug 2002), 831–5. (Page 21.)
- [53] SCHONES, D. E., AND ZHAO, K. Genome-wide approaches to studying chromatin modifications. *Nat Rev Genet* 9, 3 (Mar 2008), 179–91. (Page 10.)
- [54] SHANNON, C. A mathematical theory of communication. *Bell Syst Tech J* 27, 4 (Jan 1948), 623–656. (Page 37.)
- [55] SHEN-ORR, S., MILO, R., MANGAN, S., AND ALON, U. Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics* 31, 1 (2002), 64–68. (Page 11.)
- [56] SHUMAKER-PARRY, J. S., AEBERSOLD, R., AND CAMPBELL, C. T. 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 (Apr 2004), 2071–82. (Page 20.)
- [57] SIGAL, A., MILO, R., COHEN, A., GEVA-ZATORSKY, N., KLEIN, Y., LIRON, Y., ROSENFELD, N., DANON, T., PERZOV, N., AND ALON, U. Variability and memory of protein levels in human cells. *Nature* 444, 7119 (Nov 2006), 643–646. (Page 37.)
- [58] SLUTSKY, M., AND MIRNY, L. A. Kinetics of protein-dna interaction : facilitated target location in sequence-dependent potential. *Biophys J* 87, 6 (Dec 2004), 4021–35. (Page 15.)
- [59] SMIT, A. F. A., HUBLEY, R., AND GREEN, P. Repeatmasker open-3.0. <http://www.repeatmasker.org>, 1996-2010. (Page 38.)
- [60] STATHOPOULOS, A., AND LEVINE, M. Genomic regulatory networks and animal development. *Dev Cell* 9, 4 (Oct 2005), 449–62. (Page 28.)
- [61] STORMO, G., AND FIELDS, D. Specificity, free energy and information content in protein-dna interactions. *Trends in biochemical sciences* 23, 3 (1998), 109–113. (Page 16.)
- [62] STORMO, G. D., AND ZHAO, Y. Putting numbers on the network connections. *Bioessays* 29, 8 (Aug 2007), 717–21. (Page 20.)

- [63] STORMO, G. D., AND ZHAO, Y. Determining the specificity of protein-dna interactions. *Nature Reviews Genetics* 11, 11 (Nov 2010), 751–60. (Pages 19 et 22.)
- [64] TAKAHASHI, K., AND YAMANAKA, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126, 4 (Aug 2006), 663–76. (Page 7.)
- [65] THURMAN, R. E., RYNES, E., HUMBERT, R., VIERSTRA, J., MAURANO, M. T., HAU-GEN, E., SHEFFIELD, N. C., STERGACHIS, A. B., WANG, H., VERNOT, B., GARG, K., JOHN, S., SANDSTROM, R., BATES, D., BOATMAN, L., CANFIELD, T. K., DIEGEL, M., DUNN, D., EBERSOL, A. K., FRUM, T., GISTE, E., JOHNSON, A. K., JOHNSON, E. M., KUTYAVIN, T., LAJOIE, B., LEE, B.-K., LEE, K., LONDON, D., LOTAKIS, D., NEPH, S., NERI, F., NGUYEN, E. D., QU, H., REYNOLDS, A. P., ROACH, V., SAFI, A., SANCHEZ, M. E., SANYAL, A., SHAFER, A., SIMON, J. M., SONG, L., VONG, S., WEAVER, M., YAN, Y., ZHANG, Z., ZHANG, Z., LENHARD, B., TEWARI, M., DORSCHNER, M. O., HANSEN, R. S., NAVAS, P. A., STAMATOYANNOPOULOS, G., IYER, V. R., LIEB, J. D., SUNYAEV, S. R., AKEY, J. M., SABO, P. J., KAUL, R., FUREY, T. S., DEKKER, J., CRAWFORD, G. E., AND STAMATOYANNOPOULOS, J. A. The accessible chromatin landscape of the human genome. *Nature* 489, 7414 (Sep 2012), 75–82. (Page 26.)
- [66] TUERK, C., AND GOLD, L. Systematic evolution of ligands by exponential enrichment : Rna ligands to bacteriophage t4 dna polymerase. *Science* 249, 4968 (Aug 1990), 505–10. (Page 21.)
- [67] VAQUERIZAS, J. M., KUMMERFELD, S. K., TEICHMANN, S. A., AND LUSCOMBE, N. M. A census of human transcription factors : function, expression and evolution. *Nat Rev Genet* 10, 4 (Apr 2009), 252–63. (Page 8.)
- [68] WADDINGTON, C. H., ET AL. 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.* (1957), ix+–262. (Pages 4 et 5.)
- [69] WARREN, C. L., KRATOCHVIL, N. C. S., HAUSCHILD, K. E., FOISTER, S., BREZINSKI, M. L., DERVAN, P. B., PHILLIPS, JR, G. N., AND ANSARI, A. Z. Defining the sequence-recognition profile of dna-binding molecules. *Proc Natl Acad Sci U S A* 103, 4 (Jan 2006), 867–72. (Page 21.)
- [70] WASSERMAN, W. W., AND SANDELIN, A. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics* 5, 4 (Apr 2004), 276–87. (Page 17.)
- [71] WEINTRAUB, H., TAPSCOTT, S. J., DAVIS, R. L., THAYER, M. J., ADAM, M. A., LASSAR, A. B., AND MILLER, A. D. 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 (Jul 1989), 5434–8. (Page 13.)
- [72] WINTER, R. B., BERG, O. G., AND VON HIPPEL, P. H. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. the escherichia coli lac repressor–operator interaction : kinetic measurements and conclusions. *Biochemistry* 20, 24 (Nov 1981), 6961–77. (Page 14.)
- [73] WINTER, R. B., AND VON HIPPEL, P. H. Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. the escherichia coli repressor–operator interaction : equilibrium measurements. *Biochemistry* 20, 24 (Nov 1981), 6948–60. (Page 14.)

- [74] WITTKOPP, P. J., AND KALAY, G. Cis-regulatory elements : molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics* 13, 1 (Dec 2011), 59–69. (Page 33.)
- [75] WRIGHT, W. E., BINDER, M., AND FUNK, W. Cyclic amplification and selection of targets (casting) for the myogenin consensus binding site. *Mol Cell Biol* 11, 8 (Aug 1991), 4104–10. (Page 21.)
- [76] 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. Model-based analysis of chip-seq (macs). *Genome Biol* 9, 9 (2008), R137. (Page 24.)
- [77] ZHAO, Y., GRANAS, D., AND STORMO, G. D. Inferring binding energies from selected binding sites. *PLoS Comput Biol* 5, 12 (Dec 2009), e1000590. (Page 17.)
- [78] ZINZEN, R., GIRARDOT, C., GAGNEUR, J., BRAUN, M., AND FURLONG, E. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* 462, 7269 (2009), 65–70. (Page 38.)
- [79] ZYKOVICH, A., KORF, I., AND SEGAL, D. J. Bind-n-seq : high-throughput analysis of in vitro protein-dna interactions using massively parallel sequencing. *Nucleic Acids Res* 37, 22 (Dec 2009), e151. (Page 21.)

Résumé

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

Abstract

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

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

thèse:version du mardi 28 mai 2013 à 17 h 13]