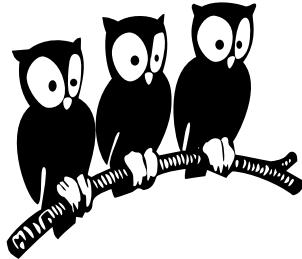


Six1 4Six1,4 Six4

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

thèse:version du lundi 6 mai 2013 à 16 h 44

Remerciements

...

thèse:version du lundi 6 mai 2013 à 16 h 44

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	16
1.5 Les modules de cis-régulation	17
1.6 Banques de données	25
Chapitre 2 - Modèles de fixation des Facteurs de Transcription à l'ADN.	27
	27
2.1 Les modèles de fixation	29
2.2 Description des données biologiques	30
2.3 Présentation de l'algorithme	30
2.4 Performance des modèles	30
2.5 Analyse des corrélations	30
2.6 Comparaison avec des données <i>in vitro</i>	30
Chapitre 3 - <i>Imogene</i> : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle	33
	33
3.1	35
Chapitre 4 - Étude de la différenciation épidermale chez la drosophile	37
	37
4.1	39

Chapitre 5 - Étude de la différenciation musculaire chez la souris	41
5.1	41
Conclusion	43
Bibliographie	45
	47

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
1.9 Construction et utilisation du modèle PWM	15
1.10 Étapes d'une expérience de ChIP-seq	16
1.11 Différents CRMs conduisent à différents patterns d'expression	17
1.13 Les états épigénétiques des CRMs	19
1.14 Approches pour la prédiction des CRMs	20
1.18 Méthodes de validation des CRMs	25
 Modèles de fixation des Facteurs de Transcription à l'ADN.	 27
	27
2.1 Description graphique de l'algorithme.	31
 <i>Imogene : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle</i>	 33
	33
 Étude de la différenciation épidermale chez la drosophile	 37
	37
 Étude de la différenciation musculaire chez la souris	 41
	41

Principales abréviations utilisées

ARNm	ARN messager
bp	Paire de base
CRM	Module de cis-régulation (<i>Cis-Regulatory Module</i>)
ISH	Hybridation <i>in situ</i> (<i>In-Situ Hybridization</i>)
kb	kilobases (1000bp)
nt	Nucléotide
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	12
1.2.4 Évolution des réseaux génétiques	13
1.3 Modèles mathématiques des interactions protéine-ADN	14
1.3.1 Modèle biophysique	14
1.3.2 Modèle thermodynamique	14
1.3.3 Modèle PWM	14
1.4 Mesures expérimentales des interactions protéine-ADN	16
1.4.1 Approches <i>in vitro</i> : PBM, SELEX, HT-SELEX	16
1.4.2 Approches <i>in vivo</i> : ChIP-on-chip, ChIP-seq, DNase	16
1.5 Les modules de cis-régulation	17
1.5.1 Modules et fonctions logiques	17
1.5.2 Encodage de patterns spatiaux	17
1.5.3 Différents états des CRMs	19
1.5.4 Prédiction des CRMs	20
1.5.5 Grammaire des enhancers : enhanceosome vs billboard	21
1.5.6 Évolution des enhancers	22
1.5.7 Les « shadow enhancers »	24
1.5.8 Validation expérimentale	25
1.6 Banques de données	25
1.6.1 Séquences génomiques et alignements	25
1.6.2 Annotations (TSSs, repeats...)	25
1.6.3 Jaspar et Transfac	25
1.6.4 Visualisation sur UCSC	25
1.6.5 Le projet ENCODE	25

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 [2]. 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 [4], 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 [1].

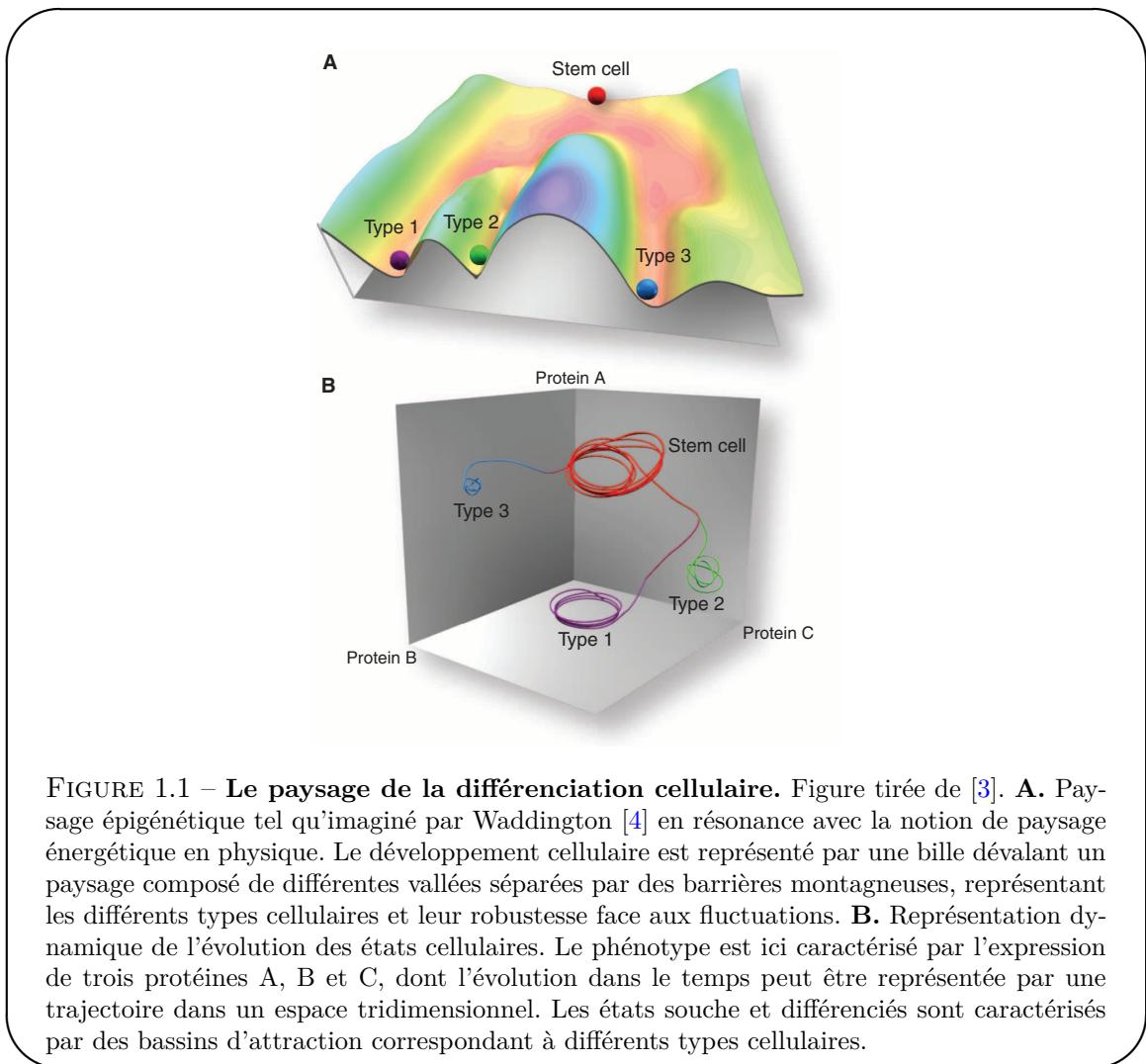


FIGURE 1.1 – Le paysage de la différenciation cellulaire. Figure tirée de [3]. **A.** Paysage épigénétique tel qu’imaginé par Waddington [4] en résonance avec la notion de paysage énergétique en physique. Le développement cellulaire est représenté par une bille dévalant un paysage composé de différentes vallées séparées par des barrières montagneuses, représentant les différents types cellulaires et leur robustesse face aux fluctuations. **B.** Représentation dynamique de l’évolution des états cellulaires. Le phénotype est ici caractérisé par l’expression de trois protéines A, B et C, dont l’évolution dans le temps peut être représentée par une trajectoire dans un espace tridimensionnel. Les états souche et différenciés sont caractérisés par des bassins d’attraction correspondant à différents types cellulaires.

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 [5].

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

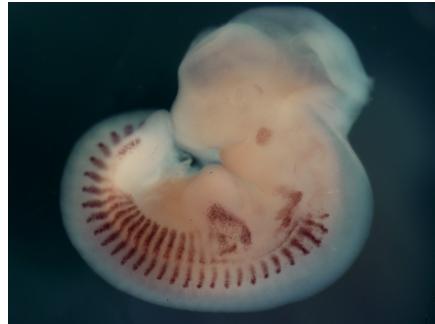
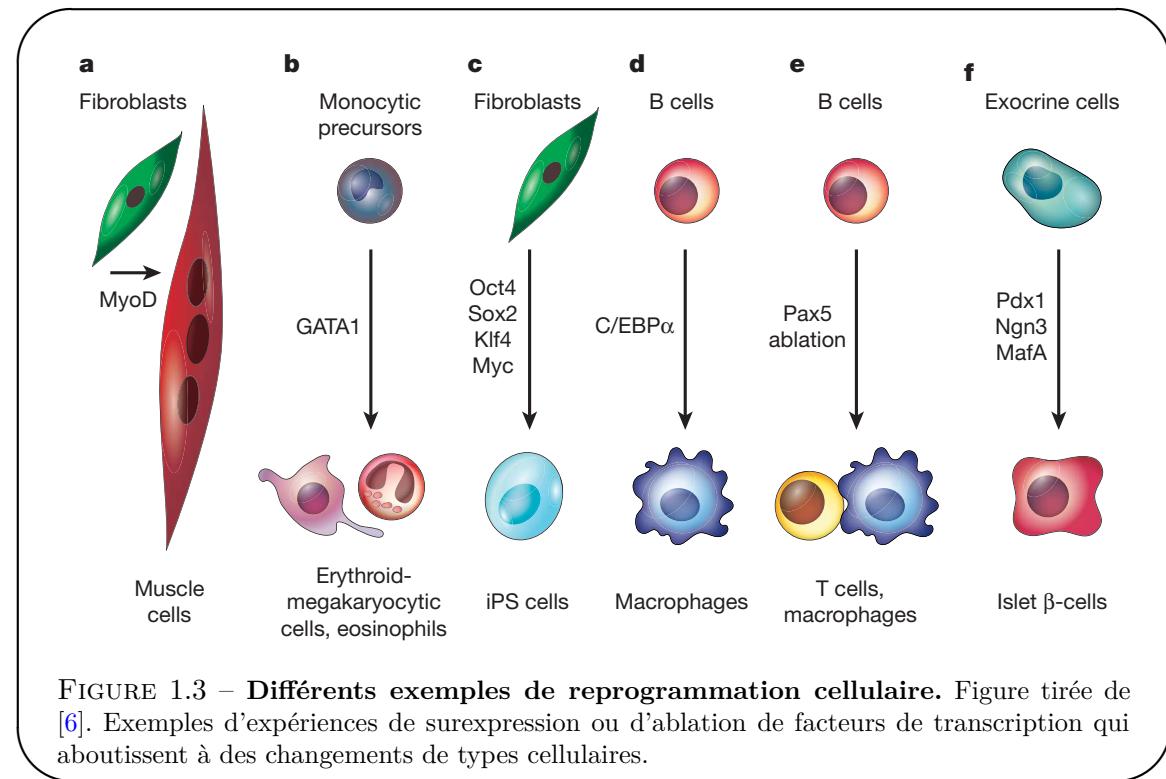


FIGURE 1.2 – Spécification spatio-temporelle du type cellulaire. Hybridization *in situ* de *Myog*, marqueur des cellules musculaires squelettiques différencierées, chez un embryon de souris de 11,5 jours. Le pattern de spécification des cellules myogéniques est clairement visible au niveau des futures 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 décennies, 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 [7]. 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) [8, 9] (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 [10]. 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 [11] (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 [13].

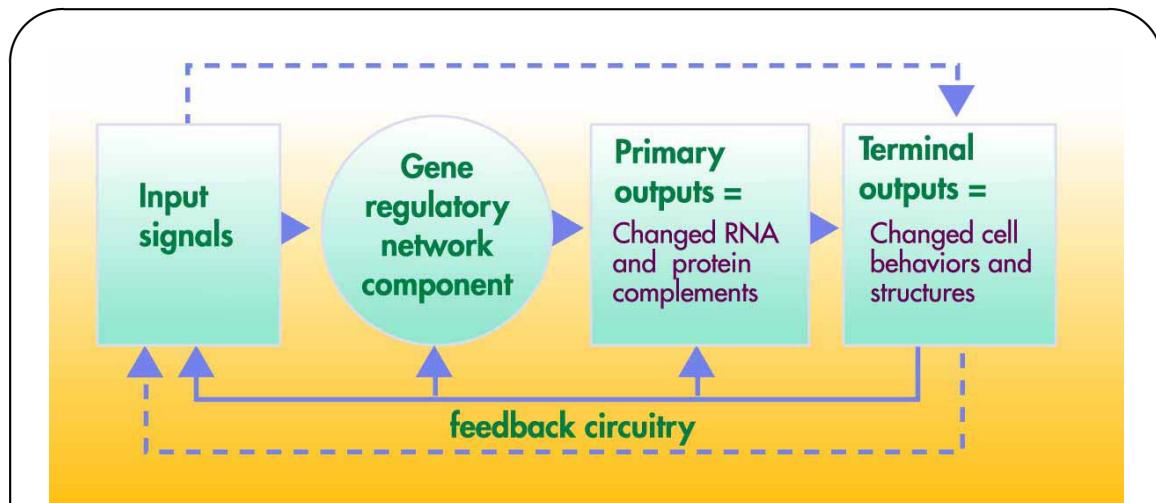


FIGURE 1.4 – Vision cybernétique du traitement de l’information par la cellule. Figure tirée du programme “Genomes to life” du Département de l’Énergie des États-Unis datant de 2001 [12] schématisant un réseau de régulation cellulaire comme un système de traitement entrée/sortie, possédant trois composantes fondamentales : (1) un système de réception et de transduction des signaux d’entrée 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 ~ 830 chez l’homme [14]. Les gènes qui les expriment représentent donc $\sim 3\%$ 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 $\sim 10\text{bp}$ et interagissent avec la machinerie transcriptionnelle au niveau du promoteur du gène cible. Les TFs peuvent se fixer sur le promoteur même, comme c’est souvent le cas chez la bactérie, ou dans des régions distales allant jusqu’à plusieurs centaines de kb, comme on trouve plus couramment chez les organismes complexes. Par ailleurs, différents TFs peuvent se combiner sur certaines régions de régulation contenant de multiples sites de fixation pour former des complexes protéiques. Ces régions, appelées (CRMs) ou plus communément *enhancers*, sont d’une taille typique de $\sim 1000\text{bp}$ et ont la particularité de conduire à une expression spatio-temporelle très spécifique du gène cible. Ces différents points seront amplement développés en section 1.5.

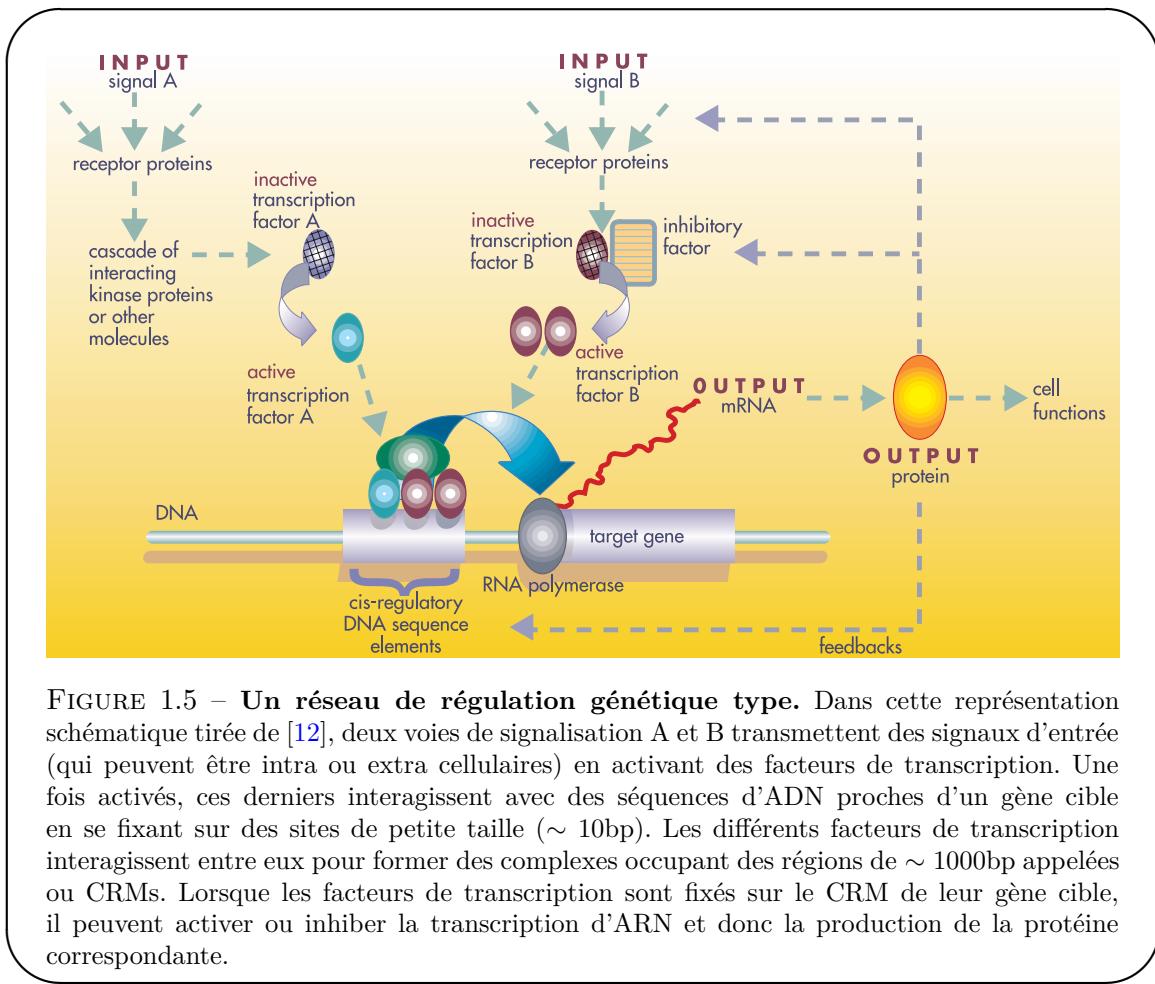


FIGURE 1.5 – Un réseau de régulation génétique type. Dans cette représentation schématique tirée de [12], deux voies de signalisation A et B transmettent des signaux d'entrée (qui peuvent être intra ou extra cellulaires) en activant des facteurs de transcription. Une fois activés, ces derniers interagissent 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 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 [16]. 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 [17]. 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

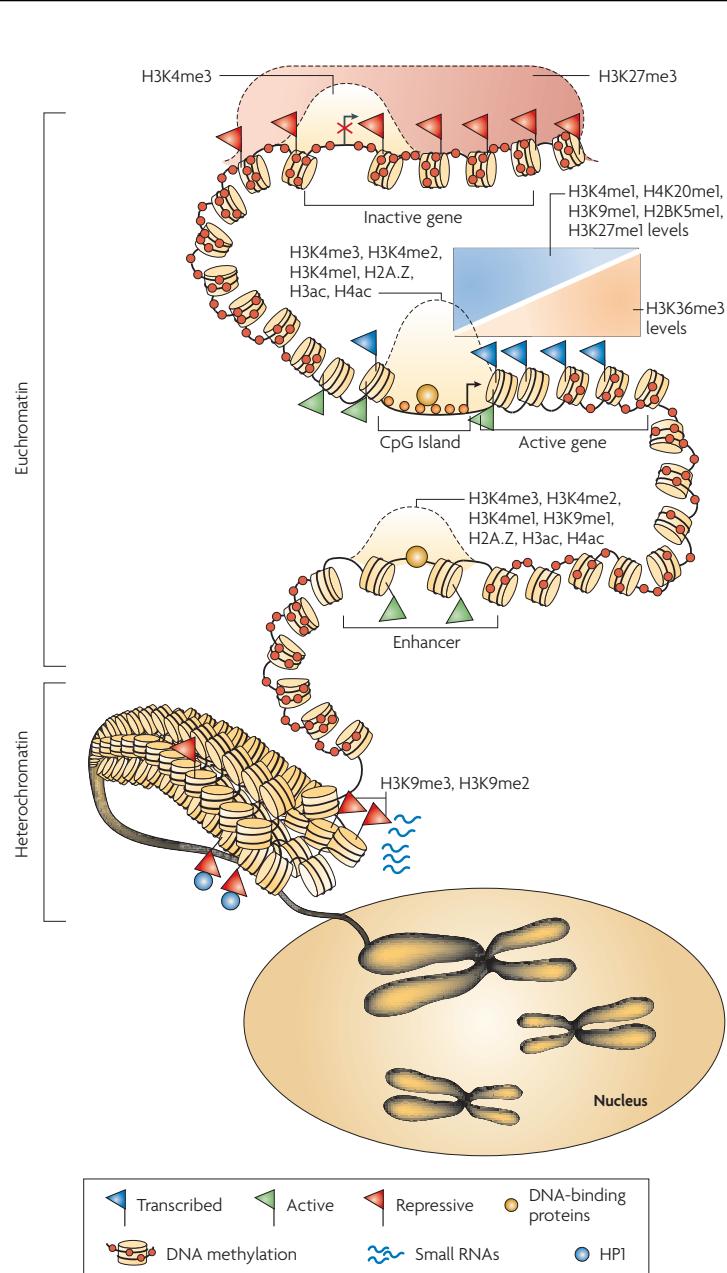


FIGURE 1.6 – Caractéristiques de l'épigénome. Figure tirée de [15]. Les chromosomes sont partagés entre régions accessibles d'euchromatine et régions difficilement accessibles d'hétérochromatine. Les régions hétérochromatiques sont marquées par de la di- et triméthylation de la lysine 9 de l'histone 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 [18, 19]. Cette méthode est connue sous le nom d'interférence ARN (RNAi) et est aujourd'hui couramment utilisée pour inhiber l'expression d'un gène. De manière similaire, les microARNs ou miRNAs sont des ARNs de ~ 23 nts issus d'ARNs plus longs appelés « épingle à cheveux » ou *hairpins* qui s'associent à la protéine *Argonaute* du complexe RISC pour entraîner la dégradation spécifique d'ARNms [20].

- **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 [21].

1.2.3 Câblage du réseau et fonction

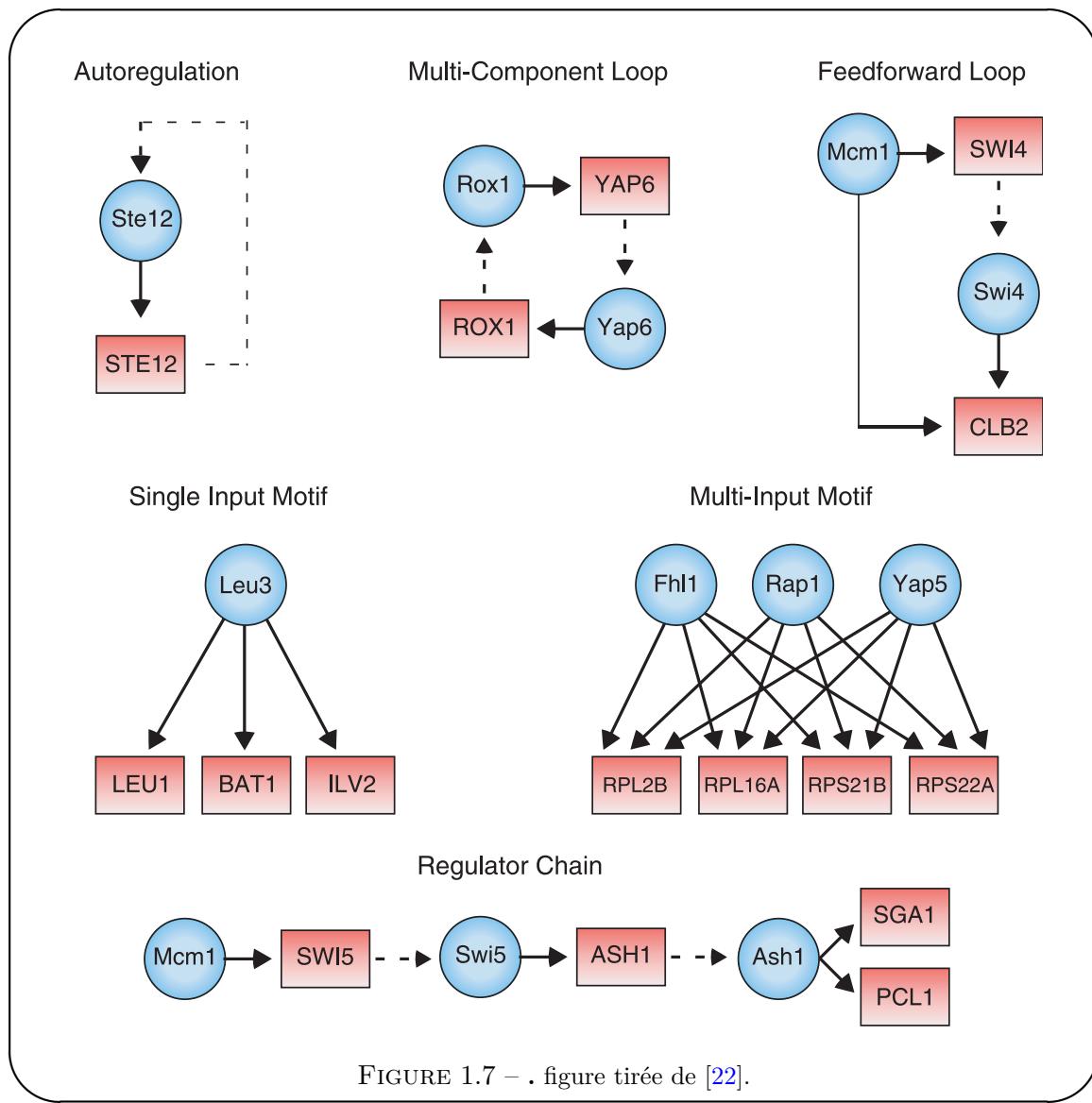


FIGURE 1.7 – . figure tirée de [22].

1.2.4 Évolution des réseaux génétiques

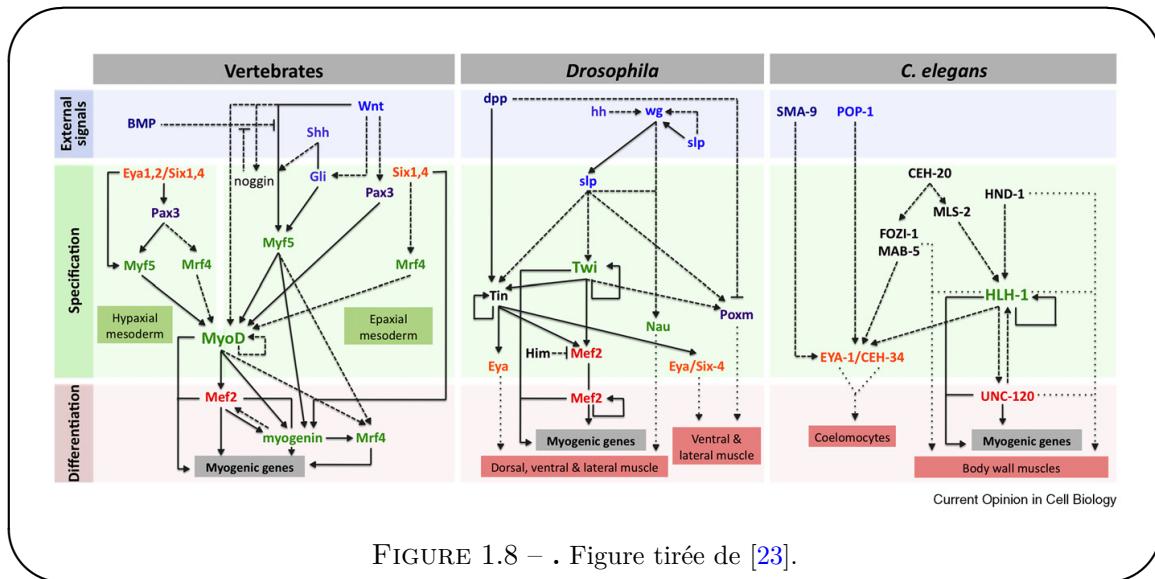


FIGURE 1.8 – . Figure tirée de [23].

1.3 Modèles mathématiques des interactions protéine-ADN

1.3.1 Modèle biophysique

1.3.2 Modèle thermodynamique

1.3.3 Modèle PWM

Le modèle PWM est le modèle le plus simple décrivant l'énergie de fixation entre un facteur de transcription et un site de fixation sur l'ADN. Ce modèle est basé sur l'hypothèse que l'énergie de fixation à un site est la somme des énergies de fixation à chaque nucléotide. Si la concentration du facteur de transcription est faible, ce modèle se réduit à un modèle biophysique.

1.3. Modèles mathématiques des interactions protéine-ADN

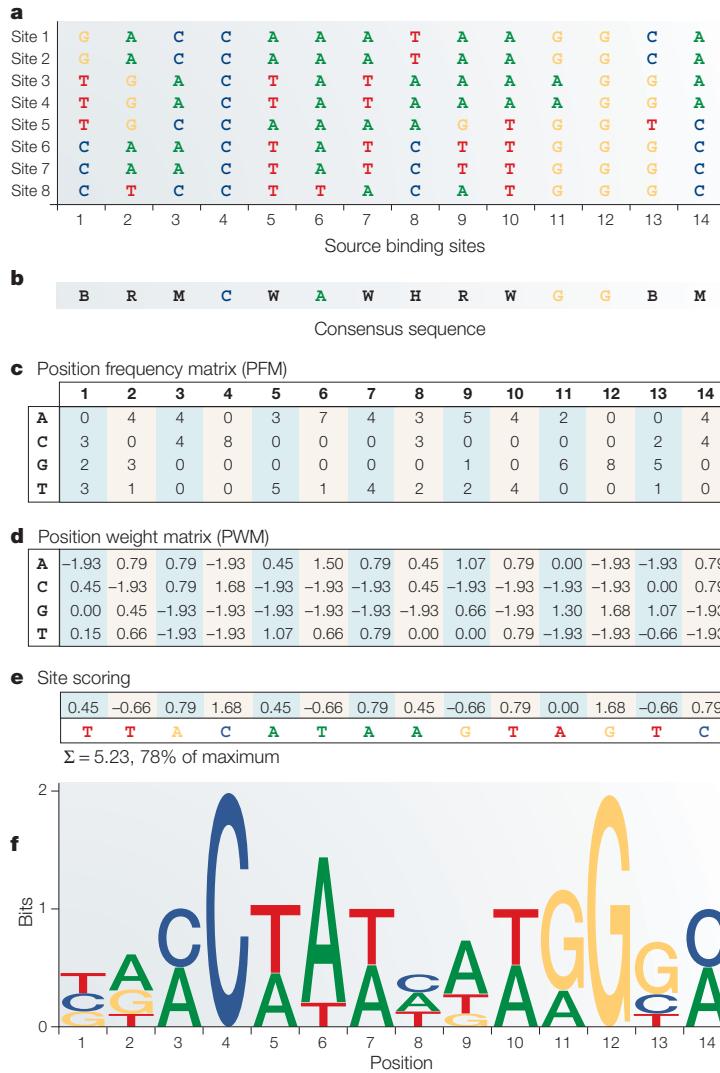


FIGURE 1.9 – Construction et utilisation du modèle PWM. Figure tirée de [24]. (a) Supposons connus un certain nombre de sites de fixation d'un facteur de transcription (dans ce cas MEF2). (b) Séquence consensus correspondante utilisant les symboles IUPAC. (c) Une matrice de fréquence est construite, 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 *background* 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.

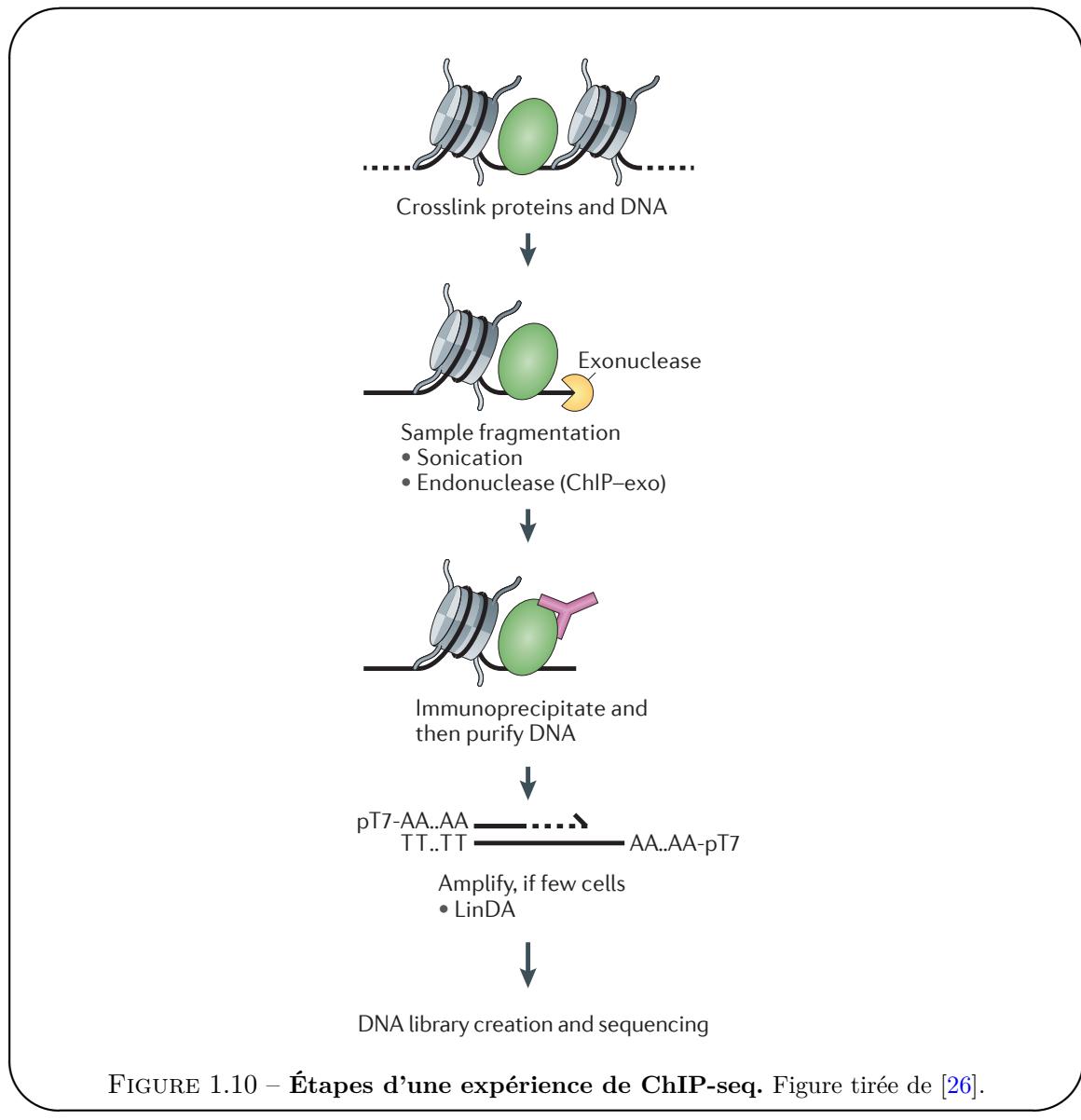
1.4 Mesures expérimentales des interactions protéine-ADN

1.4.1 Approches *in vitro* : PBM, SELEX, HT-SELEX

- PBM
- SELEX
- HT-SELEX

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

- ChIP-on-chip
- ChIP-seq

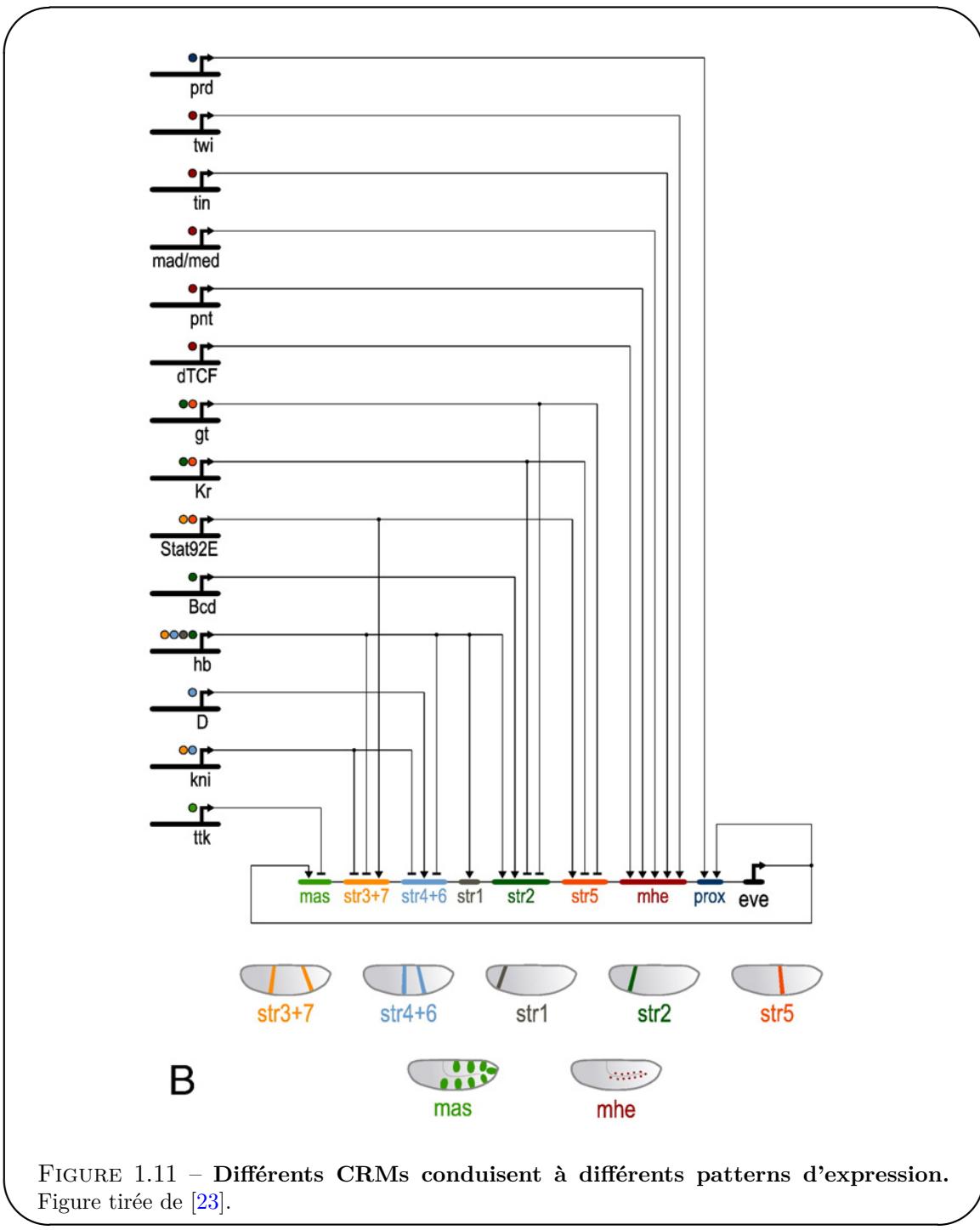


- DNase

1.5 Les modules de cis-régulation

1.5.1 Modules et fonctions logiques

1.5.2 Encodage de patterns spatiaux



Chapitre 1. Introduction générale.

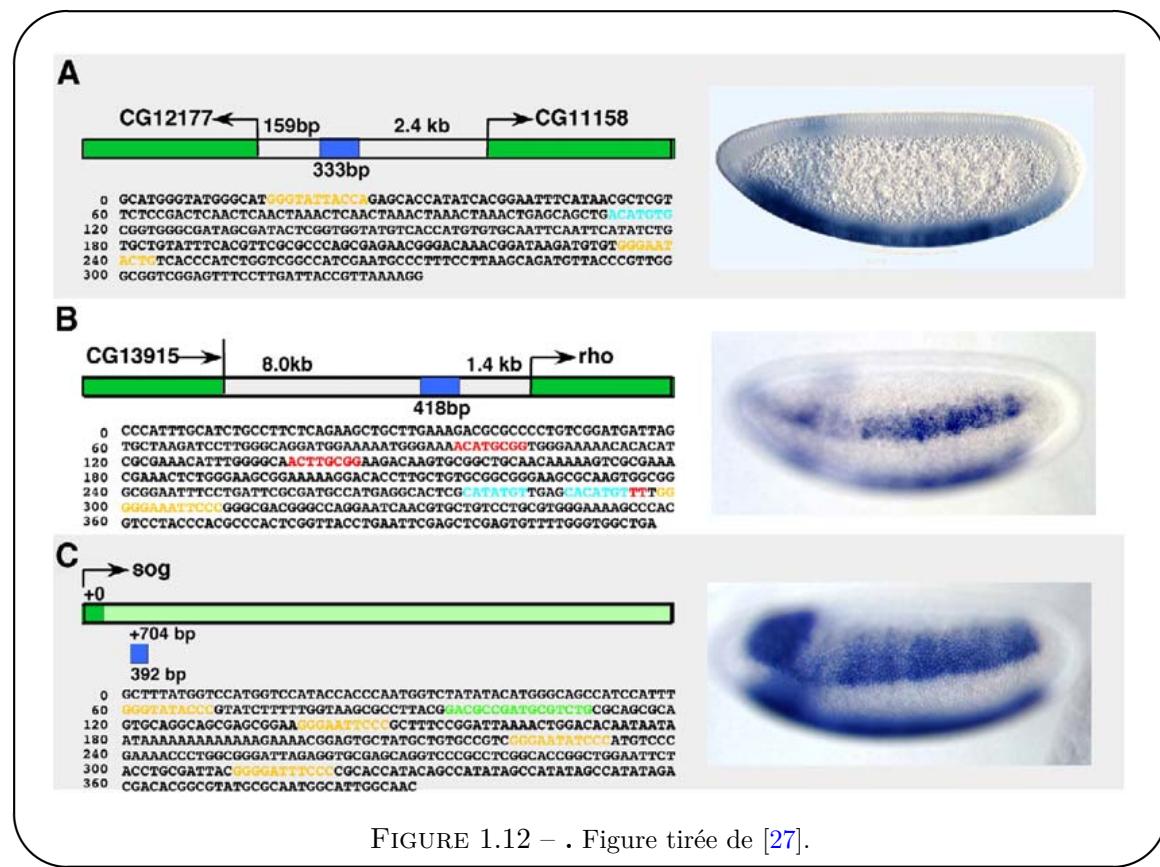


FIGURE 1.12 – . Figure tirée de [27].

1.5.3 Différents états des CRMs

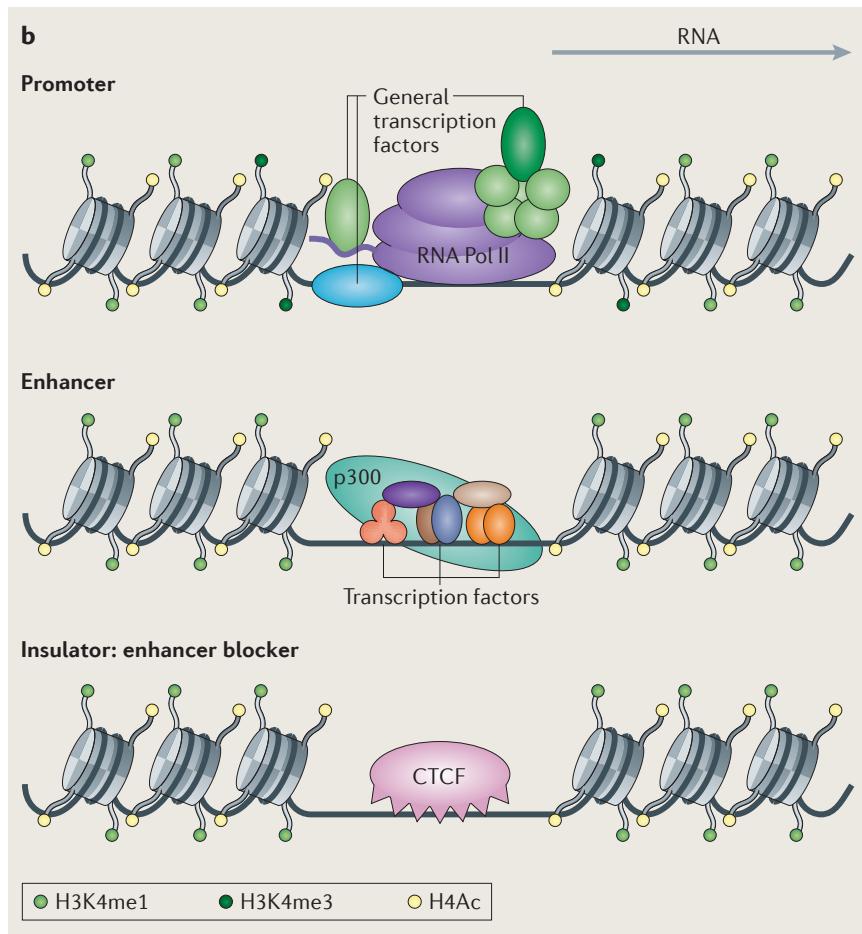


FIGURE 1.13 – Les états épigénétiques des CRMs. Figure tirée de [28].

1.5.4 Prédition des CRMs

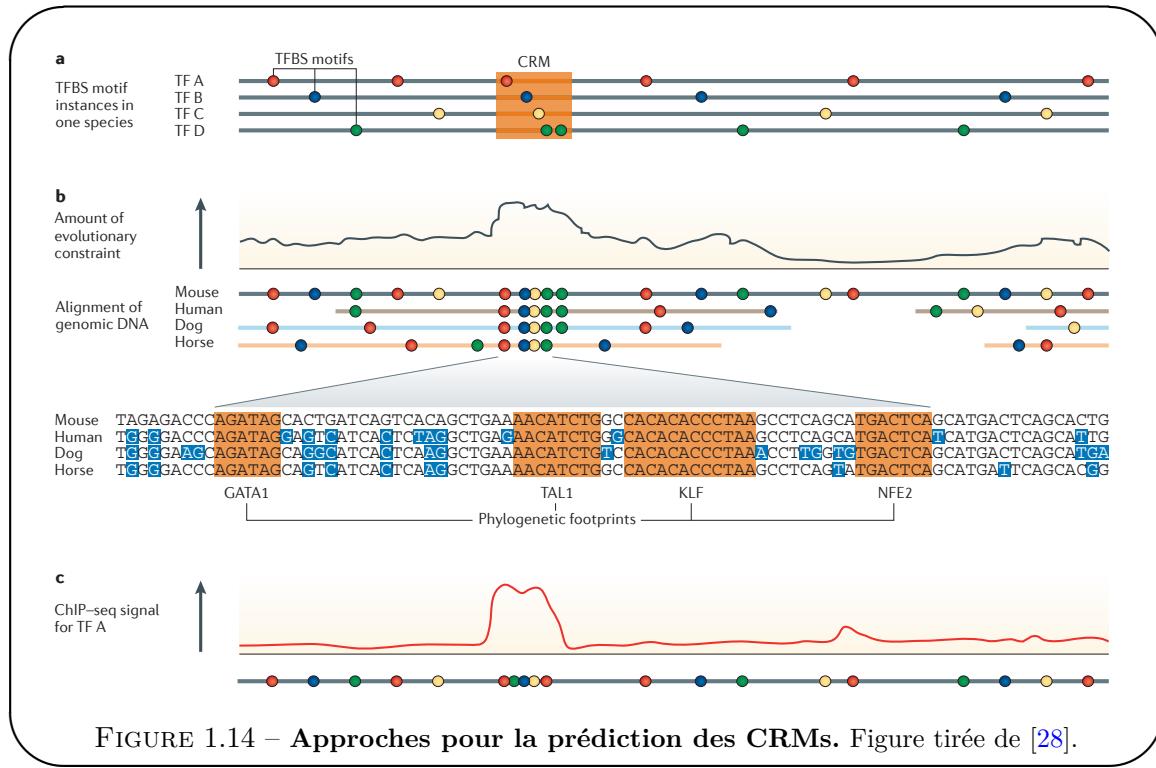
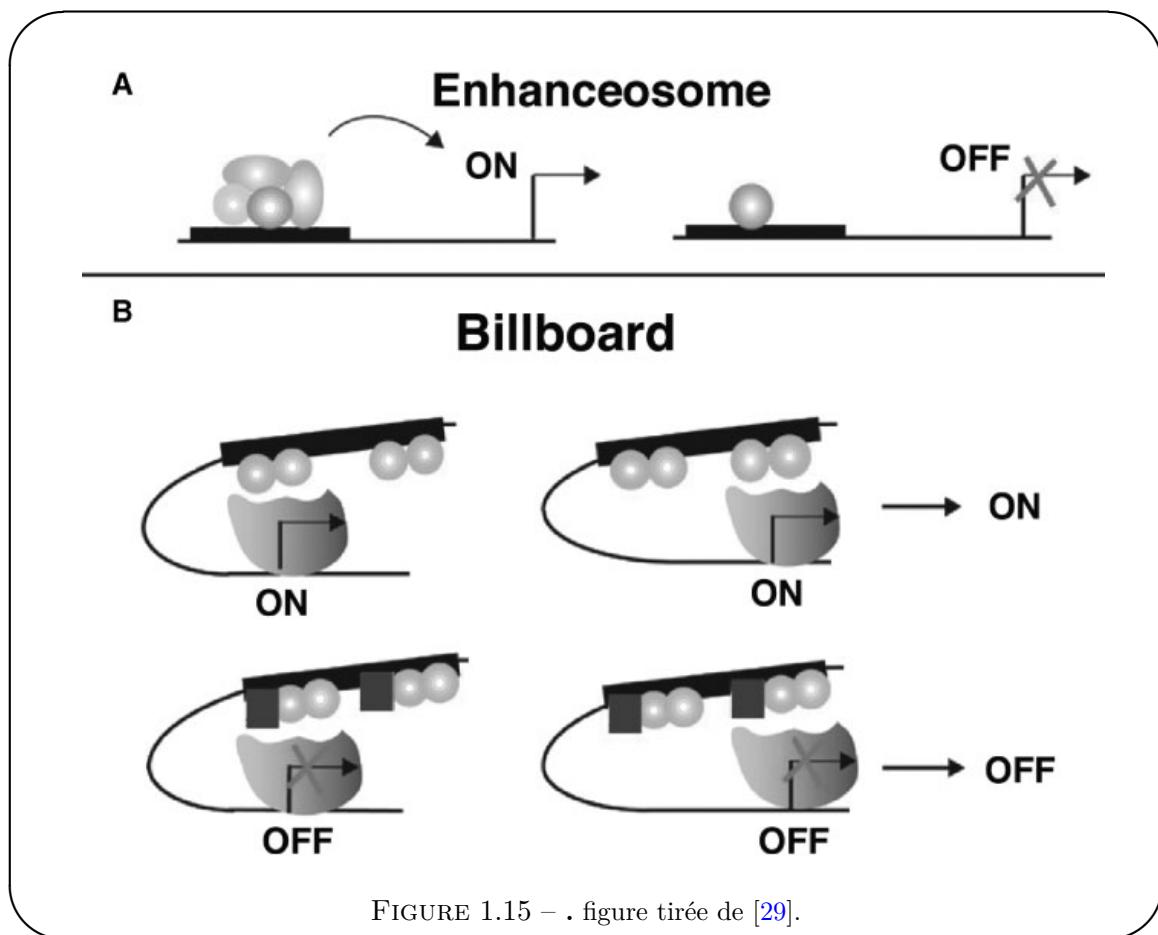


FIGURE 1.14 – Approches pour la prédition des CRMs. Figure tirée de [28].

1.5.5 Grammaire des enhancers : enhanceosome vs billboard



1.5.6 Évolution des enhancers

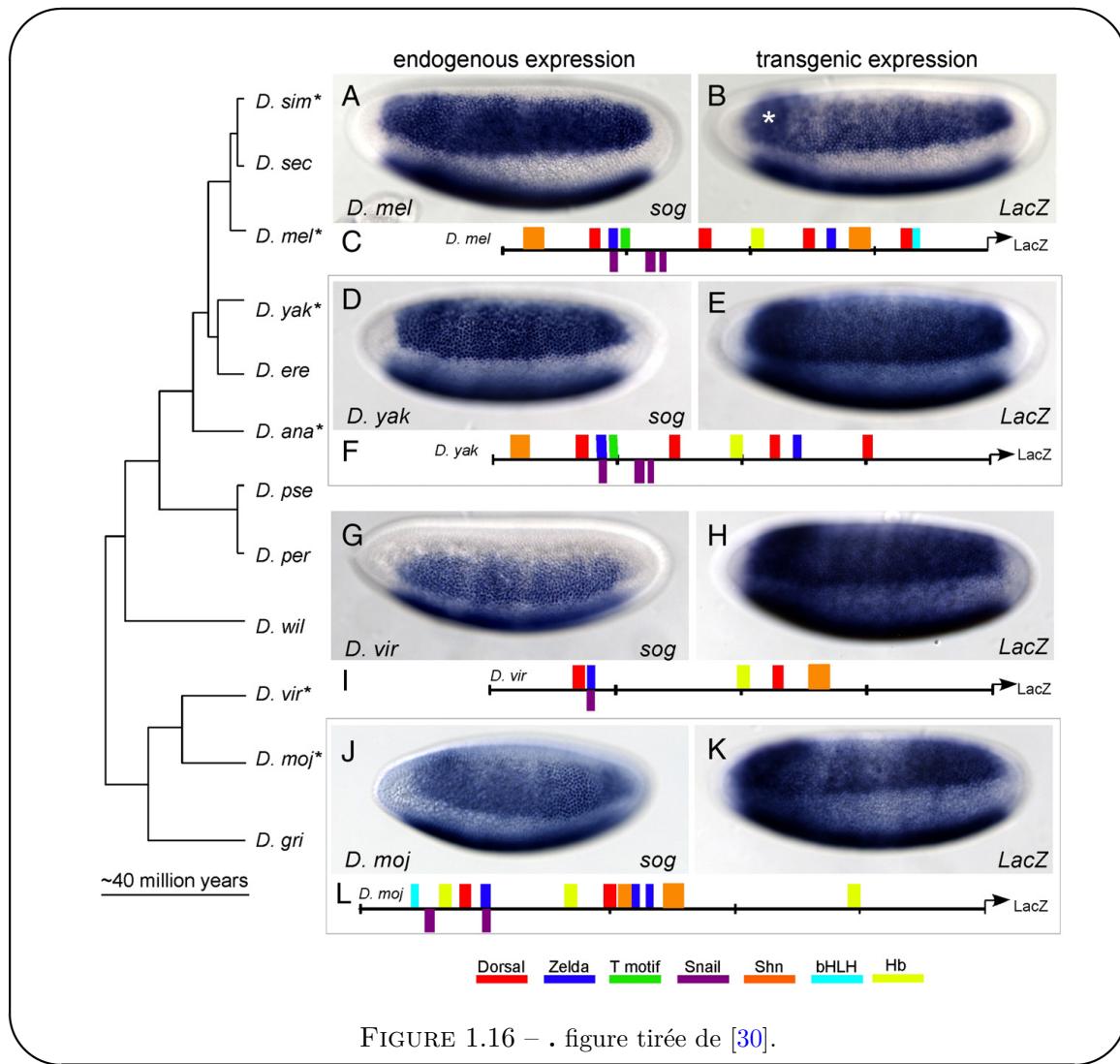


FIGURE 1.16 – . figure tirée de [30].

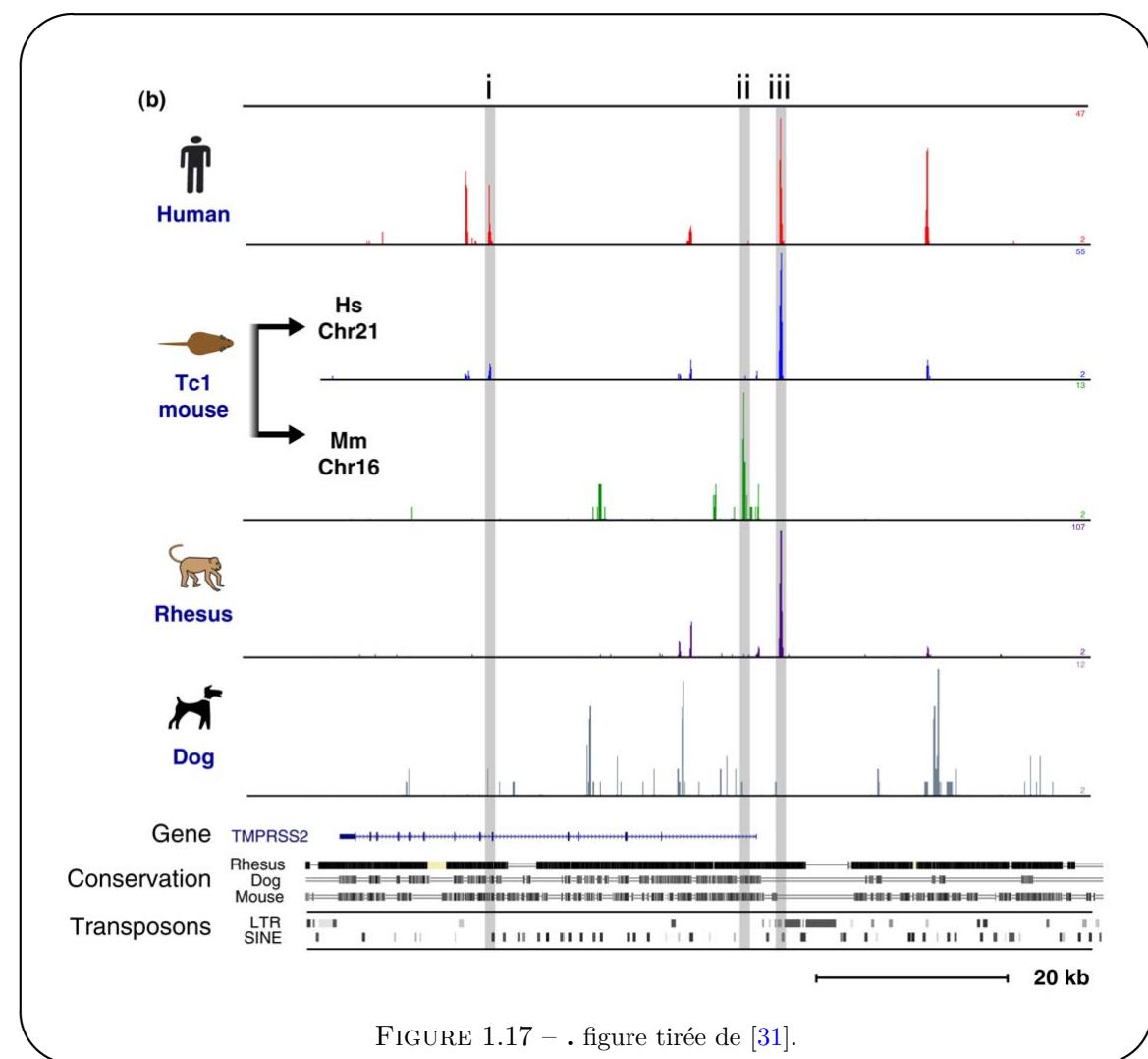


FIGURE 1.17 – . figure tirée de [31].

[32]

1.5.7 Les « shadow enhancers »

1.5.8 Validation expérimentale

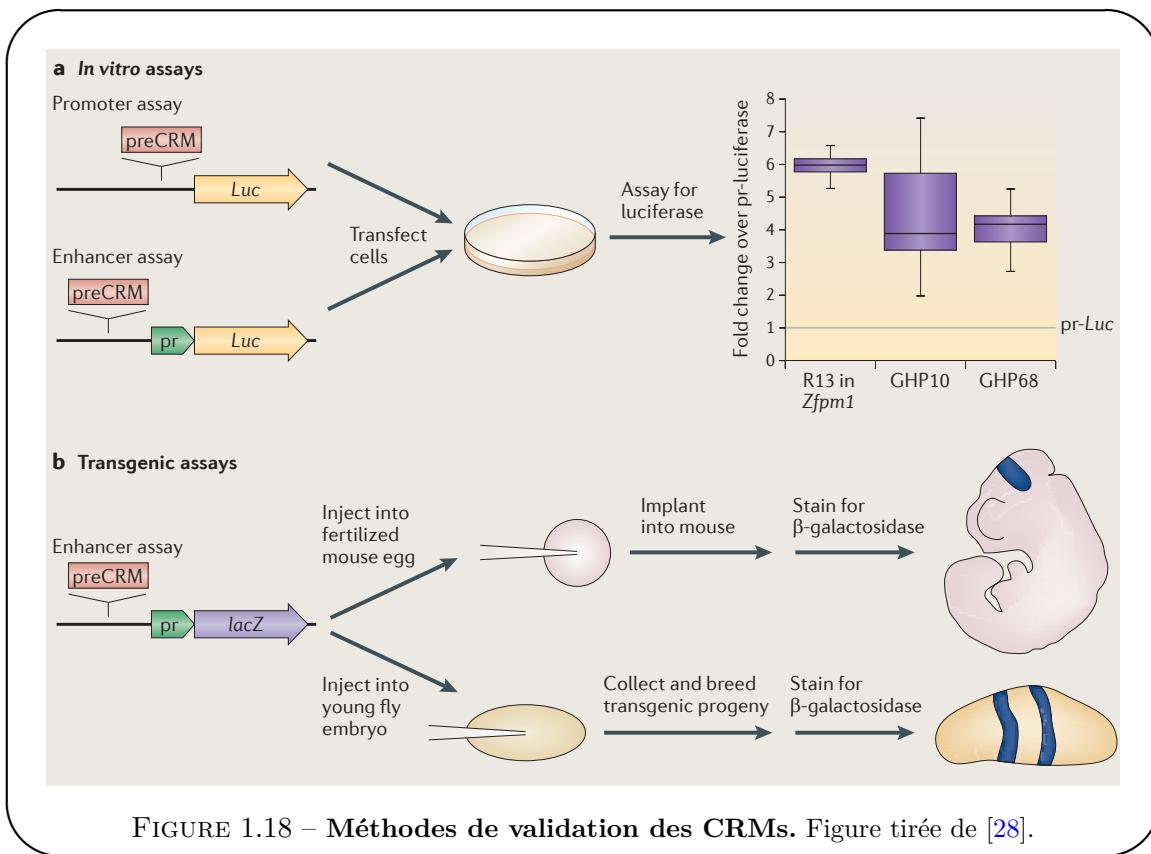


FIGURE 1.18 – Méthodes de validation des CRMs. Figure tirée de [28].

1.6 Banques de données

1.6.1 Séquences génomiques et alignements

statistiques du genome (lognormal)

1.6.2 Annotations (TSSs, repeats...)

1.6.3 Jaspar et Transfac

1.6.4 Visualisation sur UCSC

1.6.5 Le projet ENCODE

Chapitre 2

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

3.1	35
------------	-------	----

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, 34] sur l'ensemble des distributions de probabilités des états étant données les contraintes imposées. Cette fonctionnelle s'écrit [35]

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

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 [36, 37] et à partir des données du projet ENCODE [38] 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 [39].

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/>

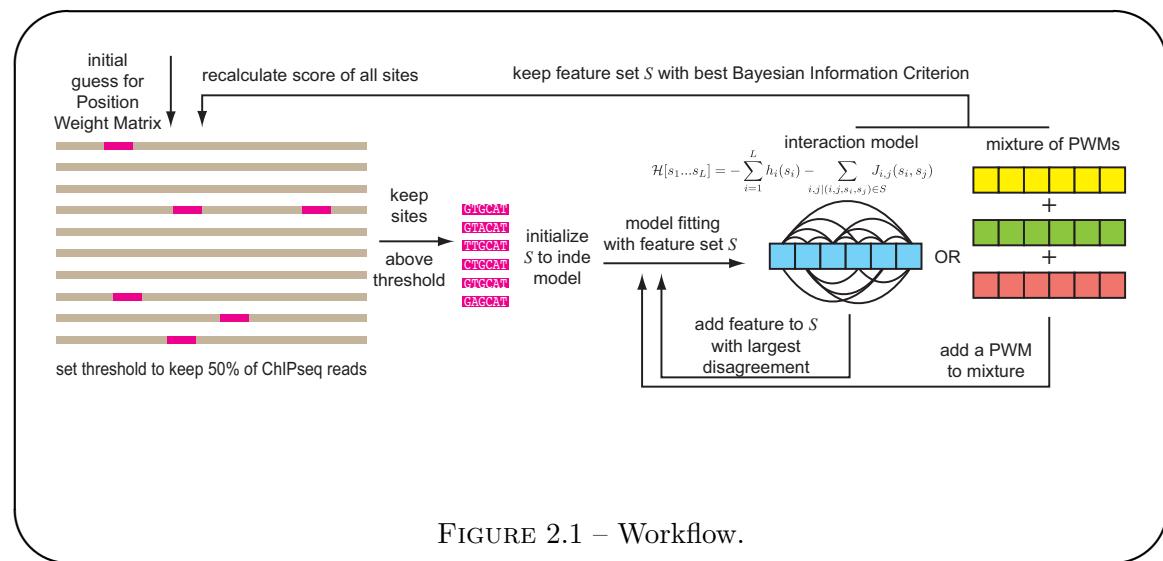
2.6. Comparaison avec des données *in vitro*

FIGURE 2.1 – Workflow.

2.6.1 Conclusion de la section 2.6

Chapitre 3

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

5.1	43
-----	-------	----

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] C. N. KEIM, J. L. MARTINS, F. ABREU, A. S. ROSADO, H. L. DE BARROS, R. BOROJEVIC, U. LINS et M. FARINA, "Multicellular life cycle of magnetotactic prokaryotes", *FEMS Microbiol Lett* **240**, n° 2, 203–8 (Nov 2004). (Page 4.)
- [2] A. BRAZMA, H. PARKINSON, T. SCHLITT et M. SHOJATALAB, "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.)
- [3] C. FURUSAWA et K. KANEKO, "A Dynamical-Systems View of Stem Cell Biology", *Science* **338**, n° 6104, 215–217 (Oct 2012). (Page 5.)
- [4] C. H. WADDINGTON 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.* pages ix+–262 (1957). (Pages 4 et 5.)
- [5] S. KAUFMANN, "The origins of order", (1993). (Page 6.)
- [6] T. GRAF et T. ENVER, "Forcing cells to change lineages", *Nature* **462**, n° 7273, 587–94 (Dec 2009). (Page 7.)
- [7] H. M. BLAU, G. K. PAVLATH, E. C. HARDEMAN, C. P. CHIU, L. SILBERSTEIN, S. G. WEBSTER, S. C. MILLER et C. WEBSTER, "Plasticity of the differentiated state", *Science* **230**, n° 4727, 758–66 (Nov 1985). :1985vn :1985vn :1985vn :1985vn
- [8] R. L. DAVIS, H. WEINTRAUB et A. B. LASSAR, "Expression of a single transfected cDNA converts fibroblasts to myoblasts", *Cell* **51**, n° 6, 987–1000 (1987). (Page 6.)
- [9] H. KULESSA, J. FRAMPTON et T. GRAF, "GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts", *Genes Dev* **9**, n° 10, 1250–62 (May 1995). :1995ys :1995ys :1995ys :1995ys
- [10] J. B. GURDON et D. A. MELTON, "Nuclear reprogramming in cells", *Science* **322**, n° 5909, 1811–5 (Dec 2008). :2008zr :2008zr :2008zr :2008zr
- [11] K. TAKAHASHI et S. YAMANAKA, "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors", *Cell* **126**, n° 4, 663–76 (Aug 2006). :2006kx :2006kx :2006kx :2006kx
- [12] O. OF BIOLOGICAL, E. RESEARCH et O. OF ADVANCED SCIENTIFIC COMPUTING RESEARCH OF THE U.S. DEPARTMENT OF ENERGY, "Genomes to life : accelerating biological discovery", http://genomicscience.energy.gov/roadmap/GTLcomplete_web.pdf (Apr 2001). (Pages 8 et 9.)
- [13] P. NURSE..., "The Cell in an Era of Systems Biology", *Cell* (Jan 2011). (Page 7.)
- [14] A. JOLMA, J. YAN, T. WHITINGTON, J. TOIVONEN, K. R. NITTA, P. RASTAS, E. MORGUNOVA, M. ENGE, M. TAIPALE, G. WEI, K. PALIN, J. M. VAQUERIZAS, R. VINCENTELLI, N. M. LUSCOMBE, T. R. HUGHES, P. LEMAIRE, E. UKKONEN, T. KIVIOJA et J. TAIPALE,

Bibliographie

- “DNA-Binding Specificities of Human Transcription Factors”, *Cell* **152**, n° 1-2, 327–39 (Jan 2013). (Page 8.)
- [15] D. E. SCHONES et K. ZHAO, “Genome-wide approaches to studying chromatin modifications”, *Nat Rev Genet* **9**, n° 3, 179–91 (Mar 2008). (Page 10.)
- [16] A. BIRD, “DNA methylation patterns and epigenetic memory”, *Genes Dev* **16**, n° 1, 6–21 (Jan 2002). (Page 9.)
- [17] E. L. GREER et Y. SHI, “Histone methylation : a dynamic mark in health, disease and inheritance”, *Nat Rev Genet* **13**, n° 5, 343–57 (May 2012). (Page 9.)
- [18] S. M. HAMMOND, A. A. CAUDY et G. J. HANNON, “Post-transcriptional gene silencing by double-stranded RNA”, *Nat Rev Genet* **2**, n° 2, 110–9 (Feb 2001). (Page 11.)
- [19] G. J. HANNON, “RNA interference”, *Nature* **418**, n° 6894, 244–51 (Jul 2002). (Page 11.)
- [20] D. P. BARTEL, “MicroRNAs : target recognition and regulatory functions”, *Cell* **136**, n° 2, 215–33 (Jan 2009). (Page 11.)
- [21] T. KONDO, S. PLAZA, J. ZANET, E. BENRABAH, P. VALENTI, Y. HASHIMOTO, S. KOBAYASHI, F. PAYRE et Y. KAGEYAMA, “Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis”, *Science* **329**, n° 5989, 336–9 (Jul 2010). (Page 11.)
- [22] T. LEE, N. RINALDI, F. ROBERT, D. ODOM, Z. BAR-JOSEPH, G. GERBER, N. HANNETT, C. HARBISON, C. THOMPSON et I. SIMON, “Transcriptional regulatory networks in *Saccharomyces cerevisiae*”, *Science* **298**, n° 5594, 799 (2002). (Page 12.)
- [23] Y.-H. LIU, J. S. JAKOBSEN, G. VALENTIN, I. AMARANTOS, D. T. GILMOUR et E. E. M. FURLONG, “A systematic analysis of Tinman function reveals Eya and JAK-STAT signaling as essential regulators of muscle development”, *Developmental Cell* **16**, n° 2, 280–91 (Feb 2009). (Pages 13 et 17.)
- [24] W. W. WASSERMAN et A. SANDELIN, “Applied bioinformatics for the identification of regulatory elements”, *Nature Reviews Genetics* **5**, n° 4, 276–87 (Apr 2004). (Page 15.)
- [25] L. M. GIOCOMO, M.-B. MOSER et E. I. MOSER, “Computational models of grid cells”, *Neuron* **71**, n° 4, 589–603 (Aug 2011). (Page 15.)
- [26] T. S. FUREY, “ChIP-seq and beyond : new and improved methodologies to detect and characterize protein-DNA interactions”, *Nature Reviews Genetics* **13**, n° 12, 840–52 (Dec 2012). (Page 16.)
- [27] A. STATHOPOULOS et M. LEVINE, “Genomic regulatory networks and animal development”, *Dev Cell* **9**, n° 4, 449–62 (Oct 2005). (Page 18.)
- [28] L. HARTWELL, J. HOPFIELD, S. LEIBLER et A. MURRAY, “From molecular to modular cell biology”, *Nature* **402**, n° 6761, 47 (1999). (Pages 19, 20 et 25.)
- [29] D. N. ARNSTI et M. M. KULKARNI, “Transcriptional enhancers : Intelligent enhanceosomes or flexible billboards ?”, *J Cell Biochem* **94**, n° 5, 890–8 (Apr 2005). (Page 21.)
- [30] L. M. LIBERMAN et A. STATHOPOULOS, “Design flexibility in cis-regulatory control of gene expression : Synthetic and comparative evidence”, *Developmental Biology* **327**, n° 2, 578–589 (Mar 2009). (Page 22.)
- [31] P. J. WITTKOPP et G. KALAY, “Cis-regulatory elements : molecular mechanisms and evolutionary processes underlying divergence”, *Nature Reviews Genetics* **13**, n° 1, 59–69 (Dec 2011). (Page 23.)
- [32] C. FESCHOTTE, “Transposable elements and the evolution of regulatory networks”, *Nat Rev Genet* **9**, n° 5, 397–405 (May 2008). (Page 23.)

- [33] E. JAYNES, “Information theory and statistical mechanics. II”, *Physical review* **108**, n° 2, 171 (1957). (Page 29.)
- [34] C. SHANNON, “A Mathematical Theory of Communication”, *Bell Syst Tech J* **27**, n° 4, 623–656 (Jan 1948). (Page 29.)
- [35] A. SIGAL, R. MILO, A. COHEN, N. GEVA-ZATORSKY, Y. KLEIN, Y. LIRON, N. ROSENFELD, T. DANON, N. PERZOV et U. ALON, “Variability and memory of protein levels in human cells”, *Nature* **444**, n° 7119, 643–646 (Nov 2006). (Page 29.)
- [36] R. ZINZEN, C. GIRARDOT, J. GAGNEUR, M. BRAUN et E. FURLONG, “Combinatorial binding predicts spatio-temporal cis-regulatory activity”, *Nature* **462**, n° 7269, 65–70 (2009). (Page 30.)
- [37] X. CHEN, H. XU, P. YUAN, F. FANG, M. HUSS, V. B. VEGA, E. WONG, Y. L. ORLOV, W. ZHANG, J. JIANG, Y.-H. LOH, H. C. YEO, Z. X. YEO, V. NARANG, K. R. GOVINDARAJAN, B. LEONG, A. SHAHAB, Y. RUAN, G. BOURQUE, W.-K. SUNG, N. D. CLARKE, C.-L. WEI et H.-H. NG, “Integration of external signaling pathways with the core transcriptional network in embryonic stem cells”, *Cell* **133**, n° 6, 1106–17 (Jun 2008). (Page 30.)
- [38] E. P. CONSORTIUM, “A user’s guide to the encyclopedia of DNA elements (ENCODE)”, *Plos Biol* **9**, n° 4, e1001046 (Apr 2011). (Page 30.)
- [39] A. F. A. SMIT, R. HUBLEY et P. GREEN, “RepeatMasker Open-3.0”, <http://www.repeatmasker.org> (1996-2010). (Page 30.)

Résumé

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

Abstract

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

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

thèse: version du lundi 6 mai 2013 à 16 h 44