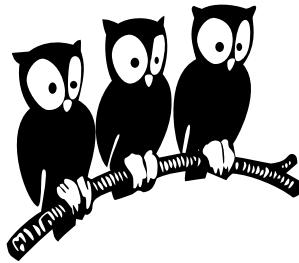


Département de Physique  
École Normale Supérieure

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 vendredi 26 avril 2013 à 22 h 01

---

# Remerciements

---

...

**thèse**: version du vendredi 26 avril 2013 à 22 h 01

---

# Table des matières

---

<b>Liste des figures</b>	vii
<b>Avant-propos</b>	1
<b>Chapitre 1 - Introduction générale.</b>	3
1.1 La différenciation cellulaire . . . . .	4
1.2 Les réseaux de régulation génétique . . . . .	5
1.3 Décrire les interactions au seins des réseaux. . . . .	6
<b>Chapitre 2 - Modèles de fixation des Facteurs de Transcription à l'ADN.</b>	15
2.1 Les modèles de fixation . . . . .	17
2.2 Description des données biologiques . . . . .	18
2.3 Présentation de l'algorithme . . . . .	18
2.4 Performance des modèles . . . . .	18
2.5 Analyse des corrélations . . . . .	18
2.6 Comparaison avec des données <i>in vitro</i> . . . . .	18
<b>Chapitre 3 - <i>Imogene</i> : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle</b>	21
3.1 . . . . .	23
<b>Chapitre 4 - Étude de la différenciation épidermale chez la drosophile</b>	25
4.1 . . . . .	27
<b>Chapitre 5 - Étude de la différenciation musculaire chez la souris</b>	29
5.1 . . . . .	31
<b>Chapitre 6 - Chapitre d'exemples</b>	33
6.1 Titre de la section . . . . .	35
<b>Conclusion</b>	36
<b>Bibliographie</b>	39



---

# Liste des figures

---

<b>Introduction générale.</b>	<b>3</b>
1.1 Le paysage de la différenciation cellulaire . . . . .	4
1.2 Vision géométrique des transitions entre types cellulaires . . . . .	5
1.3 Différents exemples de reprogrammation cellulaire . . . . .	5
1.4 Les antagonismes entre facteurs de transcription sculptent le paysage des destins cellulaires . . . . .	6
1.5 Les différentes composantes de la régulation transcriptionnelle . . . . .	7
1.6 Construction et utilisation du modèle PWM . . . . .	8
1.7 Les états épigénétiques des CRMs . . . . .	9
1.8 Approches pour la prédiction des CRMs . . . . .	9
1.9 Méthodes de validation des CRMs . . . . .	10
1.10 Étapes d'une expérience de ChIP-seq . . . . .	11
1.11 Différents CRMs conduisent à différents patterns d'expression . . . . .	12
<b>Modèles de fixation des Facteurs de Transcription à l'ADN.</b>	<b>15</b>
2.1 Description graphique de l'algorithme. . . . .	19
<i>Imogene : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle</i>	<b>21</b>
<b>Étude de la différenciation épidermale chez la drosophile</b>	<b>25</b>
<b>Étude de la différenciation musculaire chez la souris</b>	<b>29</b>
<b>Chapitre d'exemples</b>	<b>33</b>
6.1 Caption courte, pour la liste des figures. . . . .	35



---

# 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.

**thèse**: version du vendredi 26 avril 2013 à 22 h 01

Avant-propos

---

---

# Chapitre 1

## Introduction générale.

---

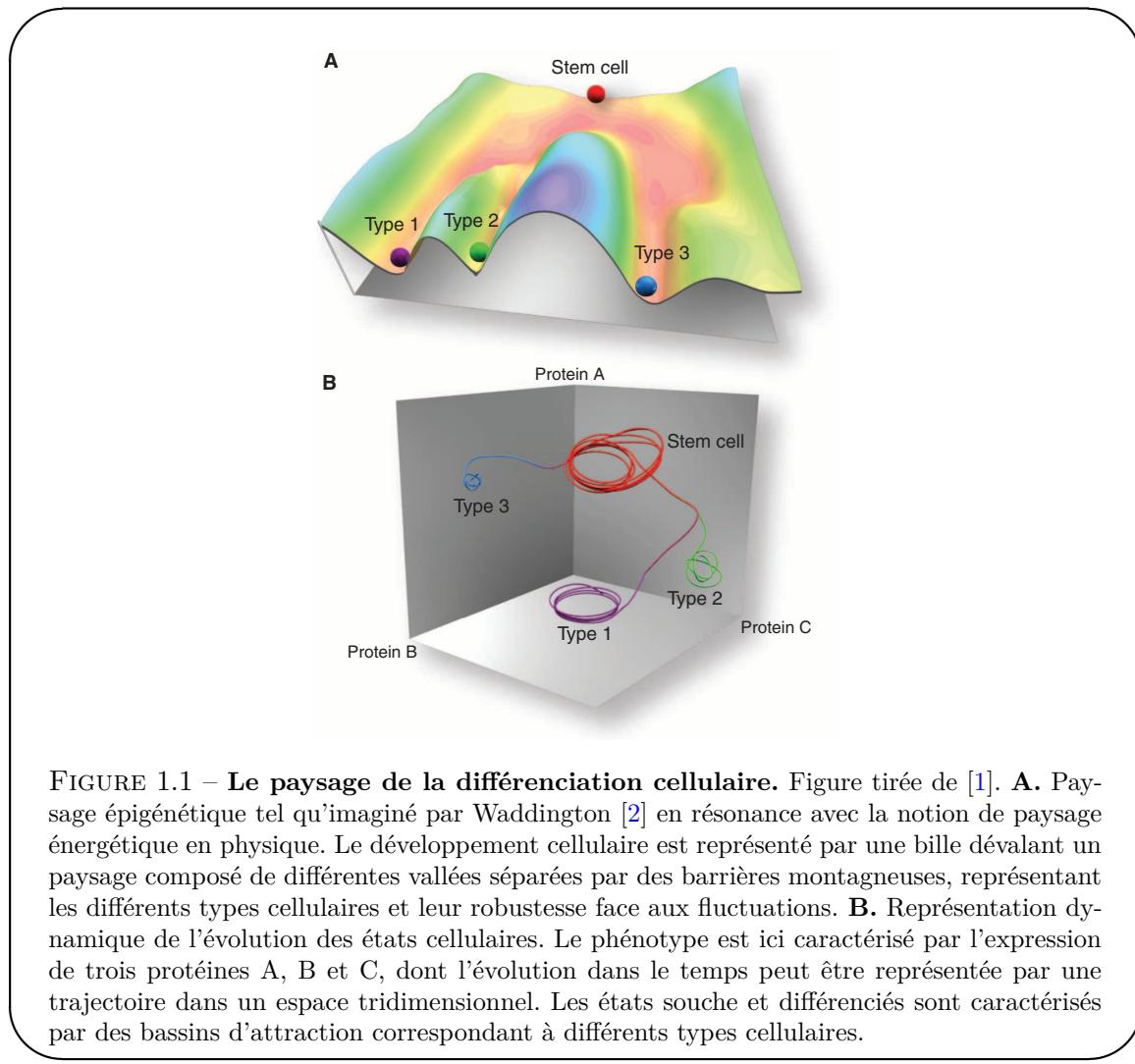
<b>1.1</b>	<b>La différenciation cellulaire</b>	<b>4</b>
1.1.1	Qu'est-ce que le phénotype d'une cellule ?	4
1.1.2	Les cellules se spécialisent au cours du développement	4
1.1.3	Les cellules sont reprogrammables	4
<b>1.2</b>	<b>Les réseaux de régulation génétique</b>	<b>5</b>
1.2.1	Vision cybernétique de la cellule	5
1.2.2	Divers modes de régulation	5
<b>1.3</b>	<b>Décrire les interactions au sein des réseaux.</b>	<b>6</b>
1.3.1	Le modèle PWM	6

aller rapidement sur nouvelles techniques. statistiques du genome

## 1.1 La différenciation cellulaire

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

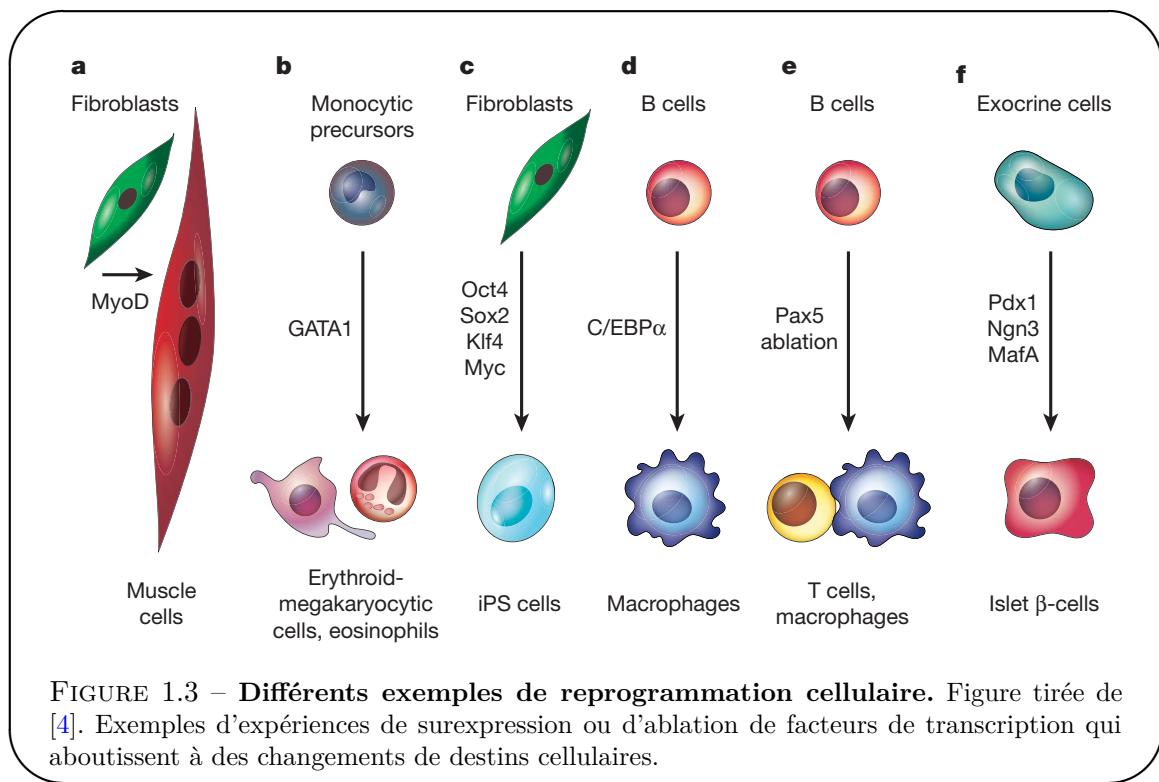
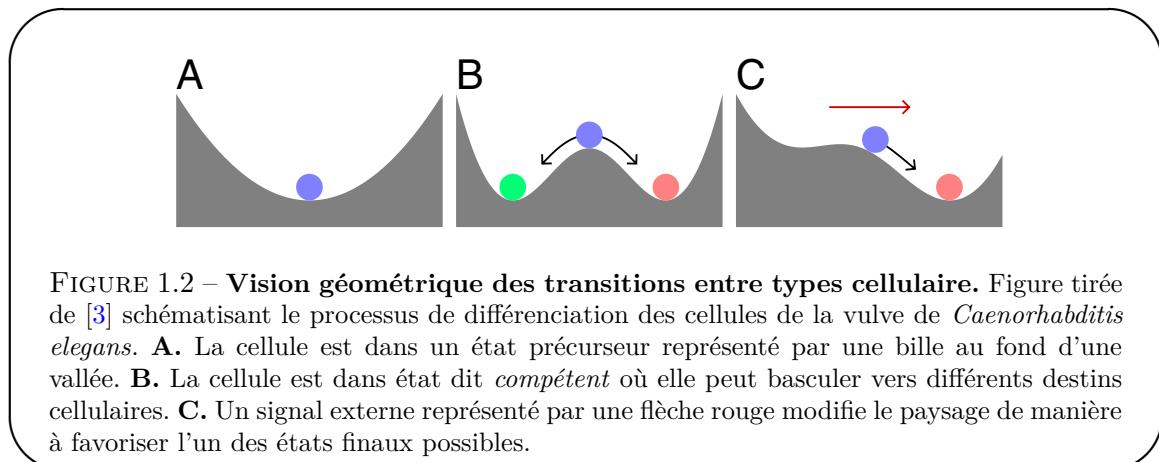
### 1.1.2 Les cellules se spécialisent au cours du développement



### 1.1.3 Les cellules sont reprogrammables

Fibroblastes, IPS : seulement un ou quelques facteurs suffisent à changer le phénotype d'une cellule.

→ comment interpréter ces résultats ?



## 1.2 Les réseaux de régulation génétique

### 1.2.1 Vision cybernétique de la cellule

Monod, Jacob. Promoteurs.

### 1.2.2 Divers modes de régulation

Enhancers, épigénétique, post-transcriptionnelle, etc.

→ quels outils pour exhiber cette circuiterie ?

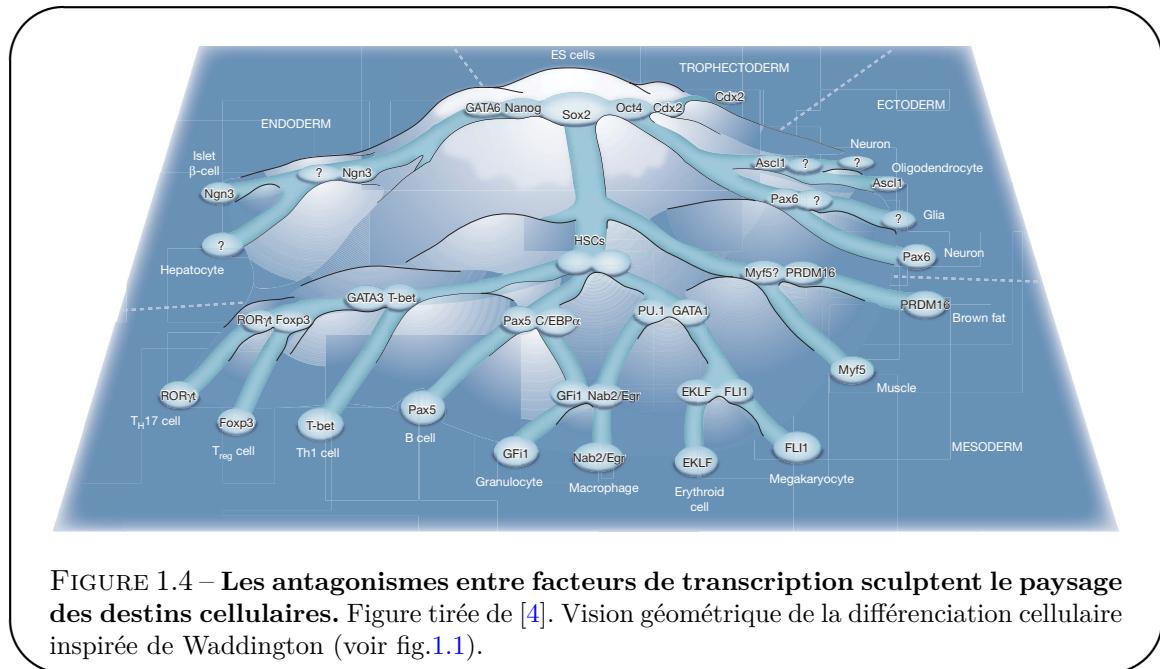


FIGURE 1.4 – Les antagonismes entre facteurs de transcription sculptent le paysage des destins cellulaires. Figure tirée de [4]. Vision géométrique de la différenciation cellulaire inspirée de Waddington (voir fig.1.1).

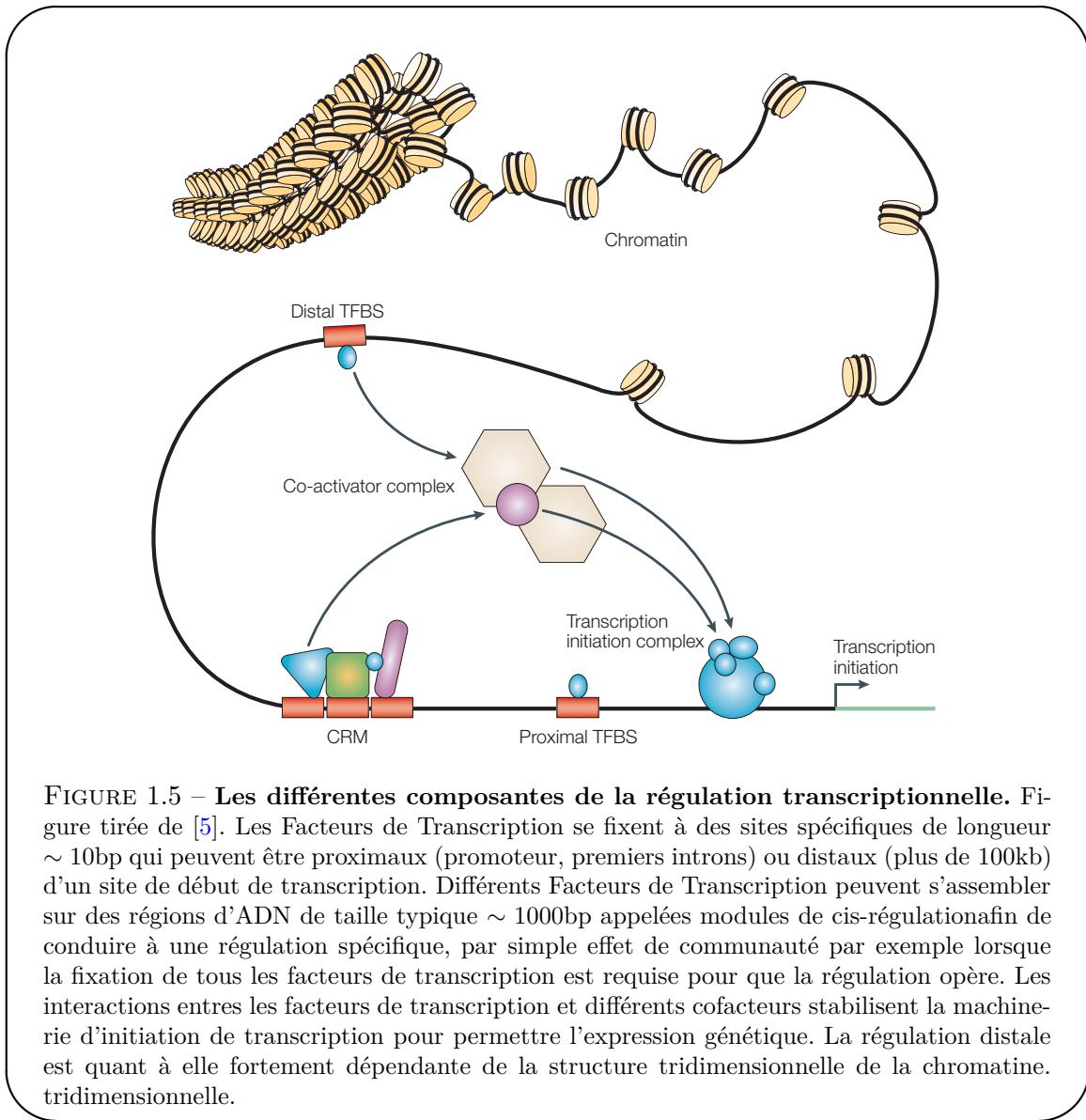
### 1.3 Décrire les interactions au seins des réseaux.

- Régulation transcriptionnelle. Facteurs de transcription. Diffusion, fixation.
- Modèles mathématiques (PWM, biophysiques).
- Coopérativité, fonctions logiques, coefficients de Hill (Uri Alon).
- Données biologiques grande échelle : ChIP-seq etc... Bioinformatique.
- ENCODE, Taipale, etc.
- Evolution de la régulation Odom, Sinha.

#### 1.3.1 Le 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.

Visualisation sur UCSC.



**FIGURE 1.5 – Les différentes composantes de la régulation transcriptionnelle.** Figure tirée de [5]. Les Facteurs de Transcription se fixent à des sites spécifiques de longueur  $\sim 10\text{bp}$  qui peuvent être proximaux (promoteur, premiers introns) ou distaux (plus de  $100\text{kb}$ ) d'un site de début de transcription. Différents Facteurs de Transcription peuvent s'assembler sur des régions d'ADN de taille typique  $\sim 1000\text{bp}$  appelées modules de cis-régulation afin de conduire à une régulation spécifique, par simple effet de communauté par exemple lorsque la fixation de tous les facteurs de transcription est requise pour que la régulation opère. Les interactions entre les facteurs de transcription et différents cofacteurs stabilisent la machine d'initiation de transcription pour permettre l'expression génétique. La régulation distale est quant à elle fortement dépendante de la structure tridimensionnelle de la chromatine. tridimensionnelle.

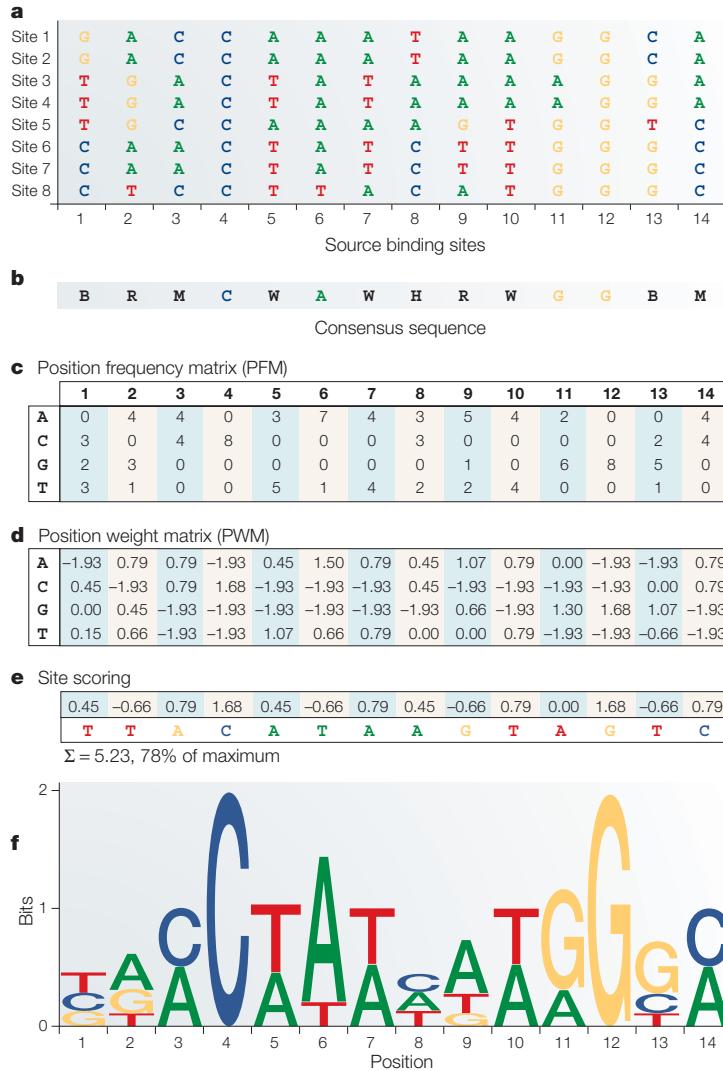


FIGURE 1.6 – Construction et utilisation du modèle PWM. Figure tirée de [5]. (a) Supposons connus un certain nombre de sites de fixation d'un facteur de transcription (dans ce cas MEF2). (b) Séquence consensus correspondante utilisant les symboles IUPAC. (c) Une matrice de fréquence est construite, indiquant 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 correspondants. (f) La PWM peut être représentée sous forme de logo [6]. Dans cette représentation, la hauteur d'une colonne représente le contenu en information ou information relative moyenne d'une position, et la taille des bases reflète leur fréquence observée.

## 1.3. Décrire les interactions au sein des réseaux.

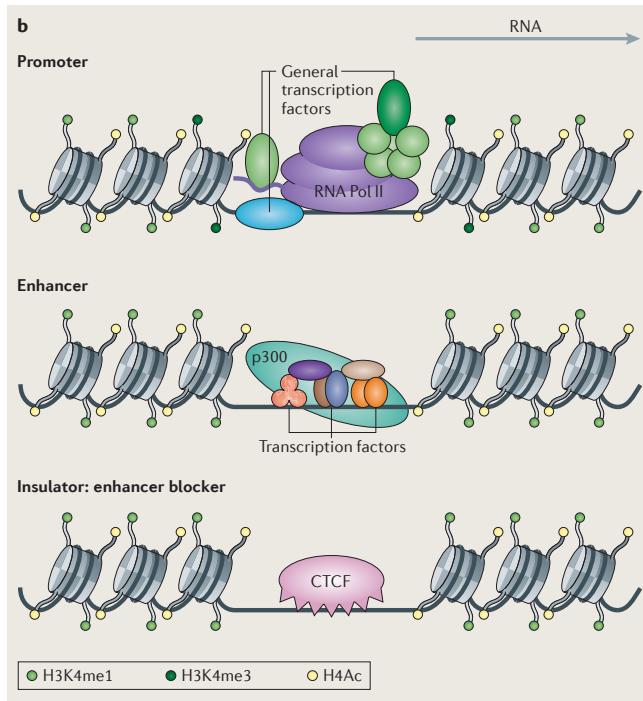


FIGURE 1.7 – Les états épigénétiques des CRMs. Figure tirée de [7].

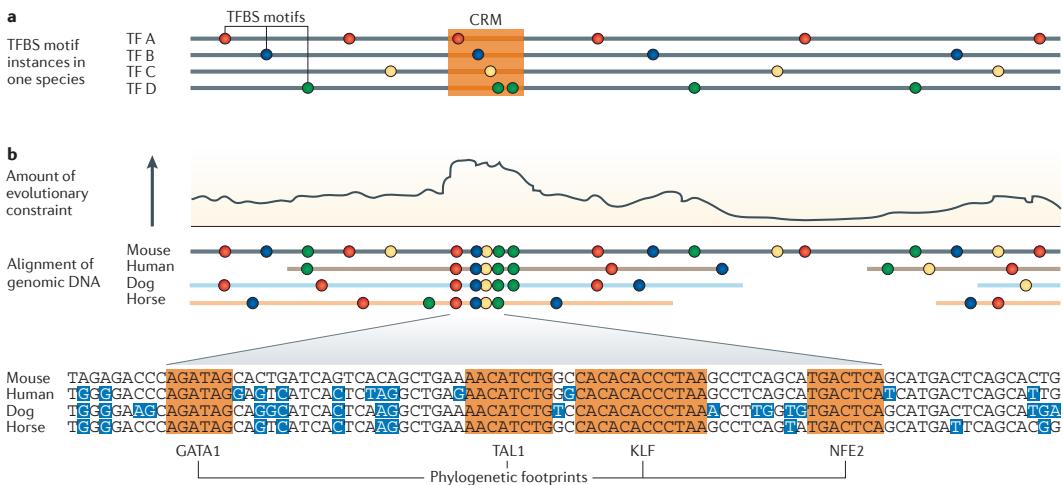


FIGURE 1.8 – Approches pour la prédition des CRMs. Figure tirée de [7].

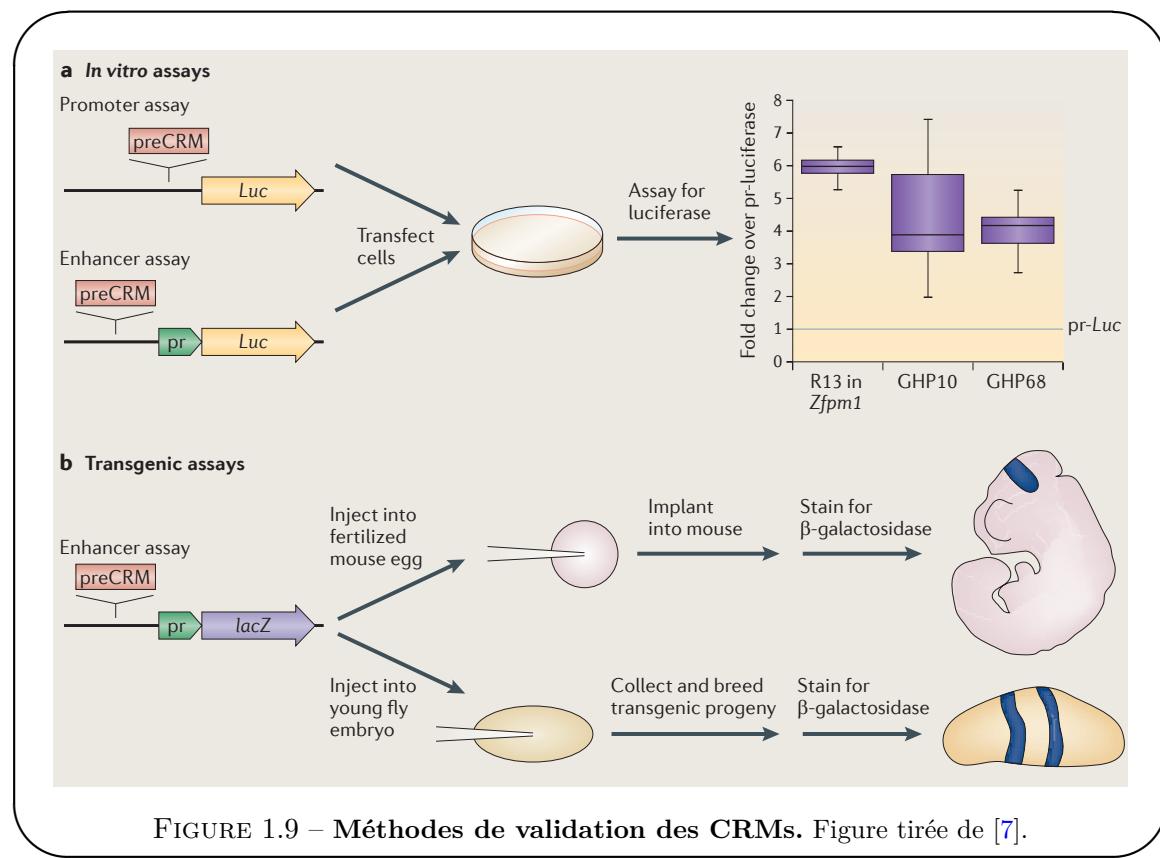


FIGURE 1.9 – Méthodes de validation des CRMs. Figure tirée de [7].

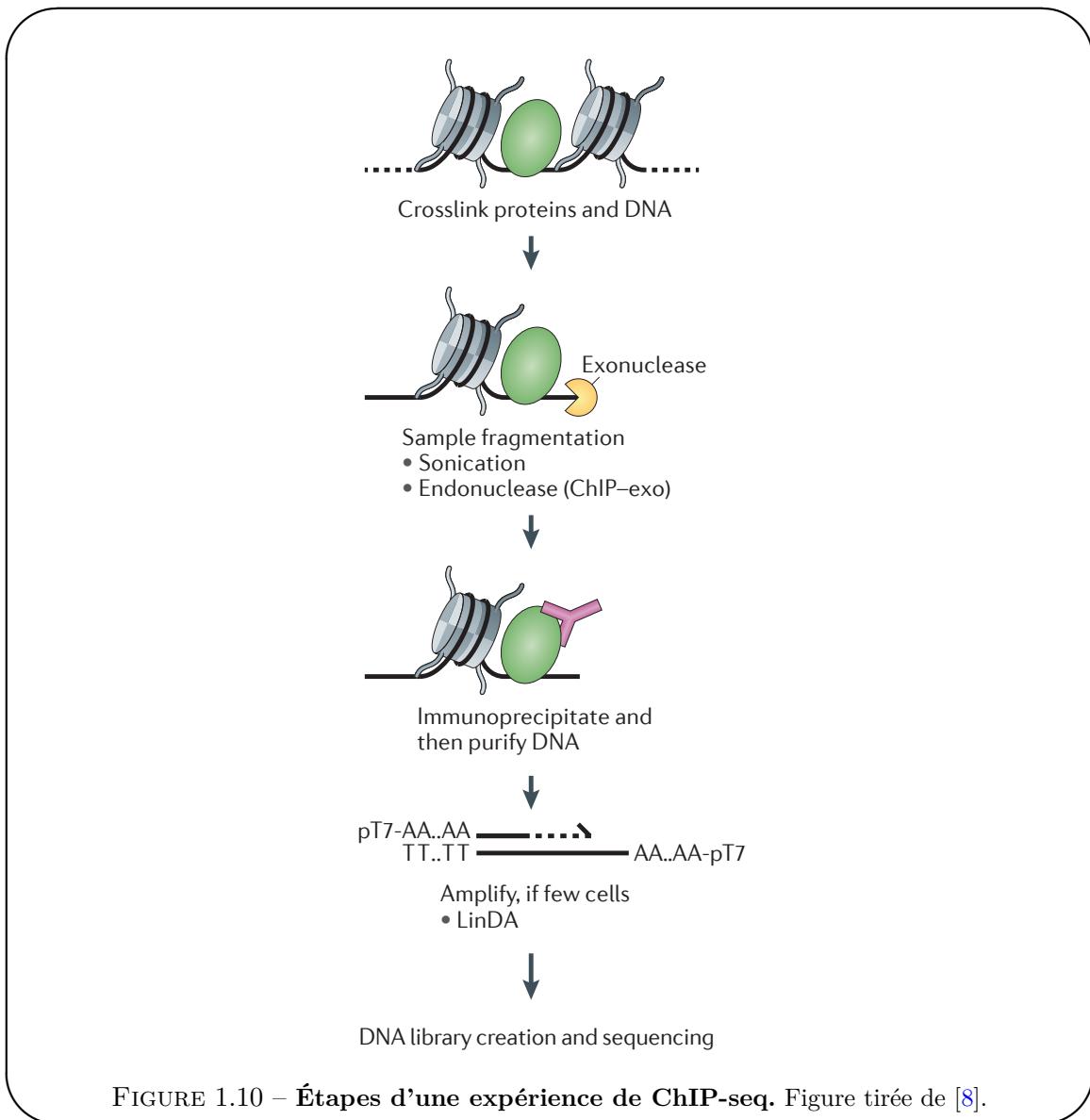


FIGURE 1.10 – Étapes d'une expérience de ChIP-seq. Figure tirée de [8].

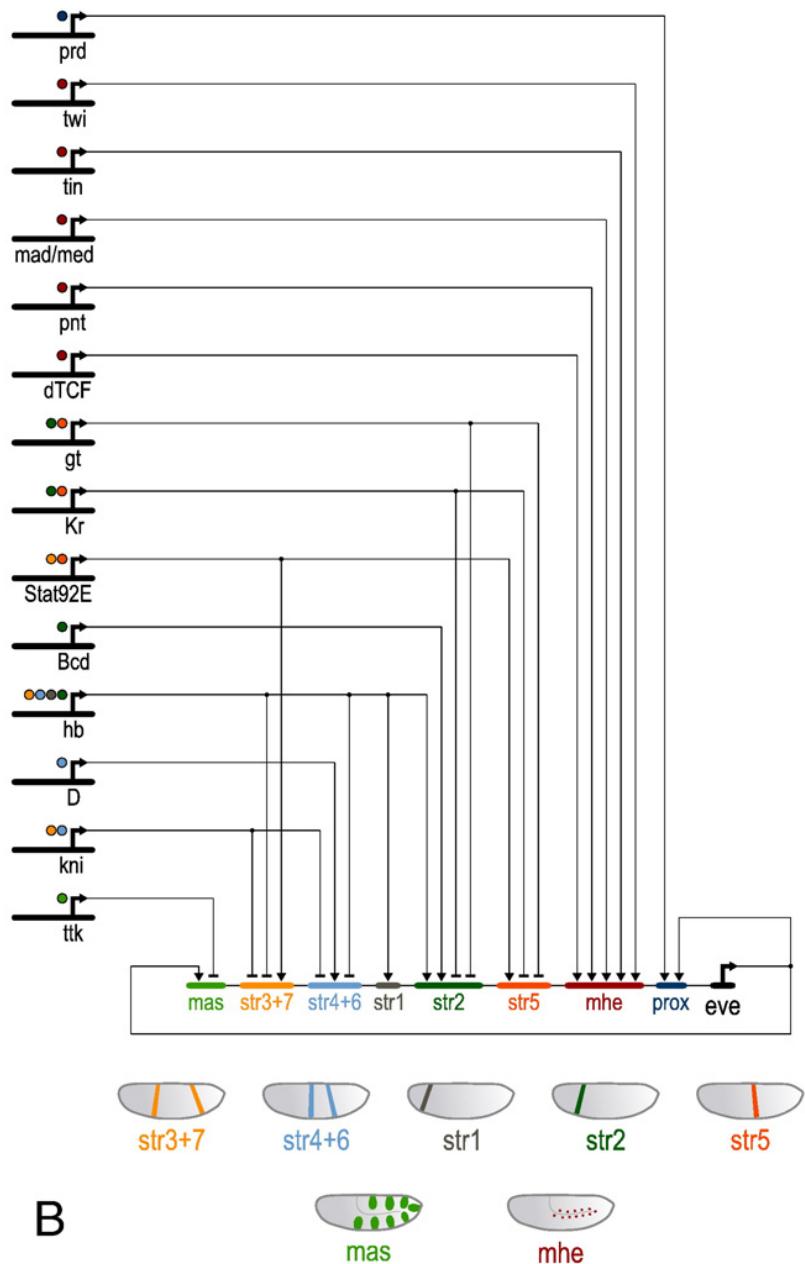


FIGURE 1.11 – Différents CRMs conduisent à différents patterns d’expression.  
Figure tirée de [9].

### **1.3. Décrire les interactions au sein des réseaux.**

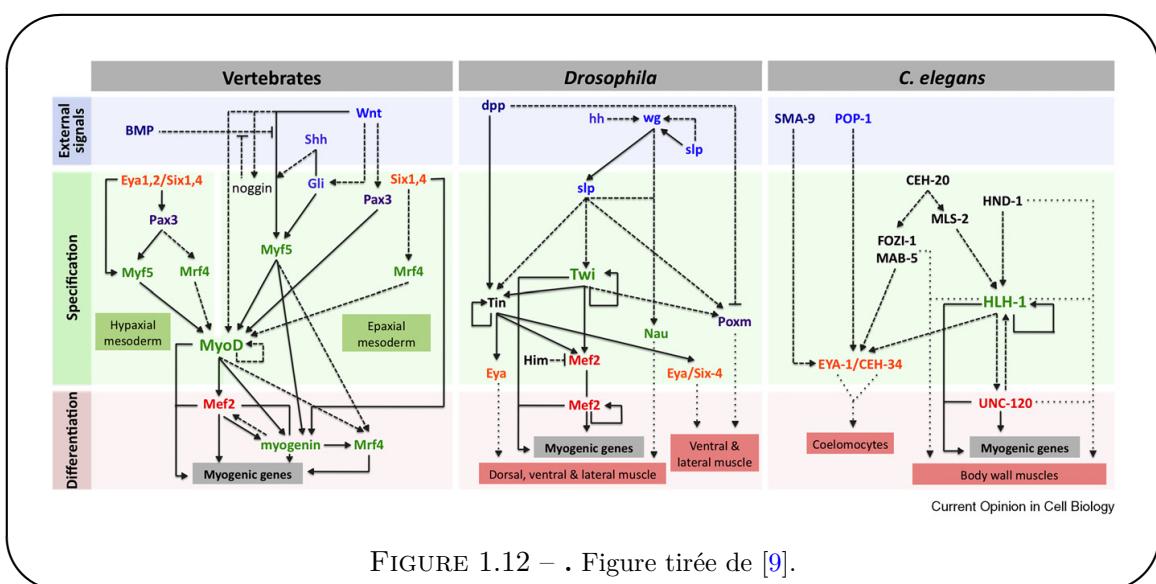


FIGURE 1.12 – . Figure tirée de [9].



---

## Chapitre 2

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

---

<b>2.1</b>	<b>Les modèles de fixation . . . . .</b>	<b>17</b>
2.1.1	Modèles de maximum d'entropie . . . . .	17
2.1.2	Modèles de mélange . . . . .	18
<b>2.2</b>	<b>Description des données biologiques . . . . .</b>	<b>18</b>
2.2.1	Les données ChIP . . . . .	18
2.2.2	Statistique « background » des séquences . . . . .	18
<b>2.3</b>	<b>Présentation de l'algorithme . . . . .</b>	<b>18</b>
<b>2.4</b>	<b>Performance des modèles . . . . .</b>	<b>18</b>
<b>2.5</b>	<b>Analyse des corrélations . . . . .</b>	<b>18</b>
2.5.1	Quantification par l'Information Directe . . . . .	18
2.5.2	Description par des patterns de Hopfield . . . . .	18
<b>2.6</b>	<b>Comparaison avec des données <i>in vitro</i> . . . . .</b>	<b>18</b>
2.6.1	Conclusion de la section 2.6 . . . . .	19

## 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 [10, 11] sur l'ensemble des distributions de probabilités des états étant données les contraintes imposées. Cette fonctionnelle s'écrit [12]

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

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  $\lambda$  et les  $\beta_\alpha$  sont les multiplicateurs de Lagrange correspondant respectivement à la contrainte de normalisation de la distribution de probabilité et aux différentes observables  $\mathcal{O}_\alpha$ . 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 [13, 14] et à partir des données du projet ENCODE [15] accessibles à partir du site internet de UCSC<sup>1</sup>, 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 [16].

### 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*

---

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

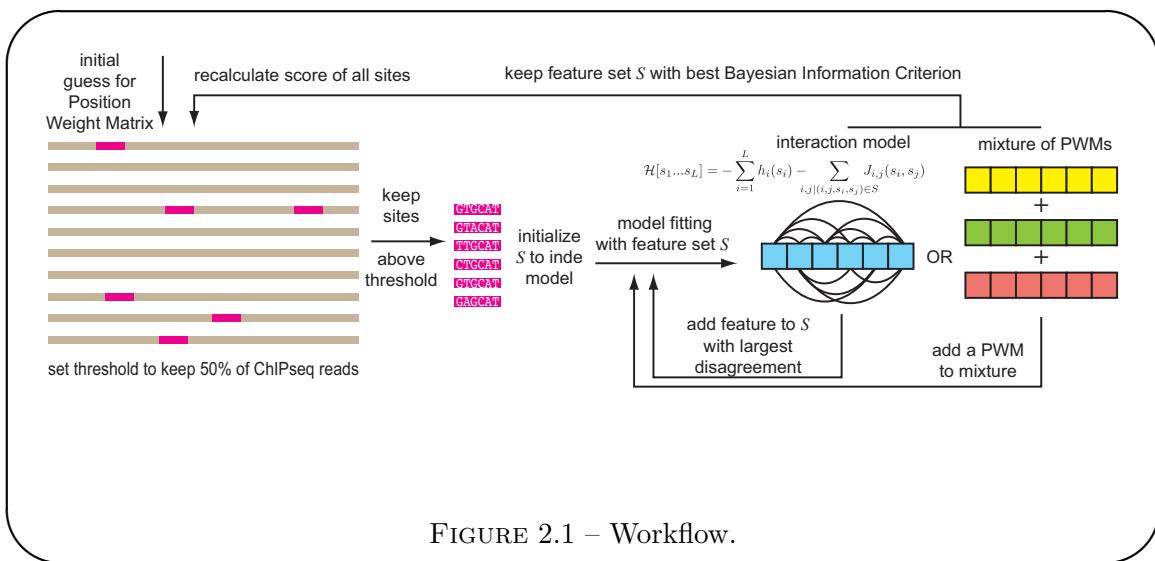


FIGURE 2.1 – Workflow.

### 2.6.1 Conclusion de la section 2.6



---

## Chapitre 3

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

---

3.1	.....	23
-----	-------	----

## 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

---

4.1	.....	27
-----	-------	----

## **Introduction du chapitre 4**

## 4.1

## Conclusion du chapitre 4

---

## Chapitre 5

# Étude de la différenciation musculaire chez la souris

---

<b>5.1</b>	.....	31
------------	-------	----

## Introduction du chapitre 5

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

## 5.1

## Conclusion du chapitre 5

---

# Chapitre 6

## Chapitre d'exemples

---

<b>6.1</b>	<b>Titre de la section</b>	<b>35</b>
6.1.1	Titre de la sous-section	35
6.1.2	Conclusion de la section 6.1	35

## **Introduction du chapitre 6**

## 6.1 Titre de la section

FIGURE 6.1 – Caption longue, pour mettre sous la figure.

### 6.1.1 Titre de la sous-section

- Titre de la sous-sous-section
- Titre de la sous-sous-section

$$\hat{H} = \int d^3\vec{r} \int_0^\infty d\omega \hbar\omega \hat{\vec{f}}^\dagger(\vec{r}, \omega) \cdot \hat{\vec{f}}(\vec{r}, \omega) + \sum_{\alpha=i,f} \hbar\omega_\alpha \hat{\xi}_\alpha + \hat{H}_Z \quad (6.1)$$

- le premier terme blabla
- le deuxième terme bliblou
- enfin, le dernier terme blubly

FIGURE 6.2

$$\begin{cases} \vec{H}_i &= H_0 \vec{u}_y e^{i(\alpha_i x - \gamma_i z)} \\ \vec{H}_r &= r_m H_0 \vec{u}_y e^{i(\alpha_i x + \gamma_i z)} \\ \vec{H}_t &= t_m H_0 \vec{u}_y e^{i(\alpha_i x - \gamma_t z)} \end{cases}$$

$$\Gamma_{i \rightarrow f} = \frac{27}{64} \frac{n_{th} + 1}{\tau_0} \left( \frac{c}{\omega} \right)^3 \frac{1}{d^4} \frac{2}{\mu_0 \omega} \operatorname{Re}(Z_S) \quad (6.2)$$

#### Remarque

Remarque en footnotesize.

#### Application numérique

$$\lambda_V(x, y) \simeq \lambda_L \sqrt{\frac{\mu_0 \varepsilon}{B_0(x, y) + \mu_0 \varepsilon}}.$$

$\lambda_L$

### 6.1.2 Conclusion de la section 6.1



---

# **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. FURUSAWA et K. KANEKO, “A Dynamical-Systems View of Stem Cell Biology”, *Science* **338**, n° 6104, 215–217 (Oct 2012). (Page [4](#).)
- [2] 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). (Page [4](#).)
- [3] F. CORSON et E. D. SIGGIA, “Geometry, epistasis, and developmental patterning”, *Proc Natl Acad Sci USA* **109**, n° 15, 5568–75 (Apr 2012). (Page [5](#).)
- [4] T. GRAF et T. ENVER, “Forcing cells to change lineages”, *Nature* **462**, n° 7273, 587–94 (Dec 2009). (Pages [5](#) et [6](#).)
- [5] W. W. WASSERMAN et A. SANDELIN, “Applied bioinformatics for the identification of regulatory elements”, *Nature Reviews Genetics* **5**, n° 4, 276–87 (Apr 2004). (Pages [7](#) et [8](#).)
- [6] L. M. GIOCOCOMO, M.-B. MOSER et E. I. MOSER, “Computational models of grid cells”, *Neuron* **71**, n° 4, 589–603 (Aug 2011). (Page [8](#).)
- [7] L. HARTWELL, J. HOPFIELD, S. LEIBLER et A. MURRAY, “From molecular to modular cell biology”, *Nature* **402**, n° 6761, 47 (1999). (Pages [9](#) et [10](#).)
- [8] 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 [11](#).)
- [9] 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 [12](#) et [13](#).)
- [10] E. JAYNES, “Information theory and statistical mechanics. II”, *Physical review* **108**, n° 2, 171 (1957). (Page [17](#).)
- [11] C. SHANNON, “A Mathematical Theory of Communication”, *Bell Syst Tech J* **27**, n° 4, 623–656 (Jan 1948). (Page [17](#).)
- [12] A. SIGAL, R. MILO, A. COHEN, N. GEVA-ZATORSKY, Y. KLEIN, Y. LIRON, N. ROSEN-FELD, 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 [17](#).)
- [13] 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 [18](#).)
- [14] 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.

## Bibliographie

---

- 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 18.)
- [15] E. P. CONSORTIUM, “A user’s guide to the encyclopedia of DNA elements (ENCODE)”, *Plos Biol* **9**, n° 4, e1001046 (Apr 2011). (Page 18.)
- [16] A. F. A. SMIT, R. HUBLEY et P. GREEN, “RepeatMasker Open-3.0”, (1996-2010). (Page 18.)

## 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 vendredi 26 avril 2013 à 22 h 01