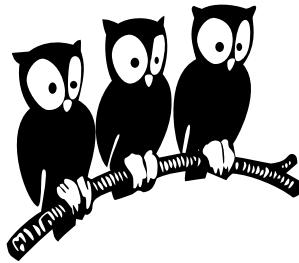


Département de Physique
École Normale Supérieure

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

...

Table des matières

| | |
|--|-----|
| Liste des figures | vii |
| Avant-propos | 1 |
| Chapitre 1 - Introduction générale. | 3 |
| 1.1 Le phénotype cellulaire | 4 |
| 1.2 Les réseaux de régulation génétique | 7 |
| 1.3 Modèles mathématiques des interactions protéine-ADN | 10 |
| 1.4 Mesures expérimentales des interactions protéine-ADN | 12 |
| 1.5 Les modules de cis-régulation | 13 |
| 1.6 Banques de données | 20 |
| Chapitre 2 - Modèles de fixation des Facteurs de Transcription à l'ADN. | 21 |
| 2.1 Les modèles de fixation | 23 |
| 2.2 Description des données biologiques | 24 |
| 2.3 Présentation de l'algorithme | 24 |
| 2.4 Performance des modèles | 24 |
| 2.5 Analyse des corrélations | 24 |
| 2.6 Comparaison avec des données <i>in vitro</i> | 24 |
| Chapitre 3 - <i>Imogene</i> : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle | 27 |
| 3.1 | 29 |
| Chapitre 4 - Étude de la différenciation épidermale chez la drosophile | 31 |
| 4.1 | 33 |
| Chapitre 5 - Étude de la différenciation musculaire chez la souris | 35 |
| 5.1 | 37 |
| Chapitre 6 - Chapitre d'exemples | 39 |
| 6.1 Titre de la section | 41 |
| Conclusion | 42 |
| Bibliographie | 45 |

Liste des figures

| | |
|---|-----------|
| Introduction générale. | 3 |
| 1.1 Le paysage de la différenciation cellulaire | 5 |
| 1.2 Différents exemples de reprogrammation cellulaire | 6 |
| 1.4 Les différentes composantes de la régulation transcriptionnelle | 7 |
| 1.7 Construction et utilisation du modèle PWM | 11 |
| 1.8 Étapes d'une expérience de ChIP-seq | 12 |
| 1.9 Différents CRMs conduisent à différents patterns d'expression | 13 |
| 1.11 Les états épigénétiques des CRMs | 15 |
| 1.12 Approches pour la prédiction des CRMs | 16 |
| 1.16 Méthodes de validation des CRMs | 20 |
| | |
| Modèles de fixation des Facteurs de Transcription à l'ADN. | 21 |
| 2.1 Description graphique de l'algorithme. | 25 |
| | |
| <i>Imogene</i> : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle | 27 |
| | |
| Étude de la différenciation épidermale chez la drosophile | 31 |
| | |
| Étude de la différenciation musculaire chez la souris | 35 |
| | |
| Chapitre d'exemples | 39 |
| 6.1 Caption courte, pour la liste des figures. | 41 |

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 | 5 |
| 1.1.3 La reprogrammation cellulaire | 6 |
| 1.1.4 La cellule dans l'organisme : une spécification spatio-temporelle | 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 | 7 |
| 1.2.3 Câblage du réseau et fonction | 8 |
| 1.2.4 Évolution des réseaux génétiques | 9 |
| 1.3 Modèles mathématiques des interactions protéine-ADN | 10 |
| 1.3.1 Modèle biophysique | 10 |
| 1.3.2 Modèle thermodynamique | 10 |
| 1.3.3 Modèle PWM | 10 |
| 1.4 Mesures expérimentales des interactions protéine-ADN | 12 |
| 1.4.1 Approches <i>in vitro</i> : PBM, SELEX, HT-SELEX | 12 |
| 1.4.2 Approches <i>in vivo</i> : ChIP-on-chip, ChIP-seq, DNase | 12 |
| 1.5 Les modules de cis-régulation | 13 |
| 1.5.1 Modules et fonctions logiques | 13 |
| 1.5.2 Encodage de patterns spatiaux | 13 |
| 1.5.3 Différents états des CRMs | 15 |
| 1.5.4 Prédiction des CRMs | 16 |
| 1.5.5 Grammaire des enhancers : enhanceosome vs billboard | 17 |
| 1.5.6 Évolution des enhancers | 18 |
| 1.5.7 Validation expérimentale | 20 |
| 1.6 Banques de données | 20 |
| 1.6.1 Séquences génomiques et alignements | 20 |
| 1.6.2 Annotations (TSSs, repeats...) | 20 |
| 1.6.3 Jaspar et Transfac | 20 |
| 1.6.4 Visualisation sur UCSC | 20 |
| 1.6.5 Le projet ENCODE | 20 |

aller rapidement sur nouvelles techniques. statistiques du genome

1.1 Le phénotype cellulaire

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

1.1.2 La différenciation cellulaire

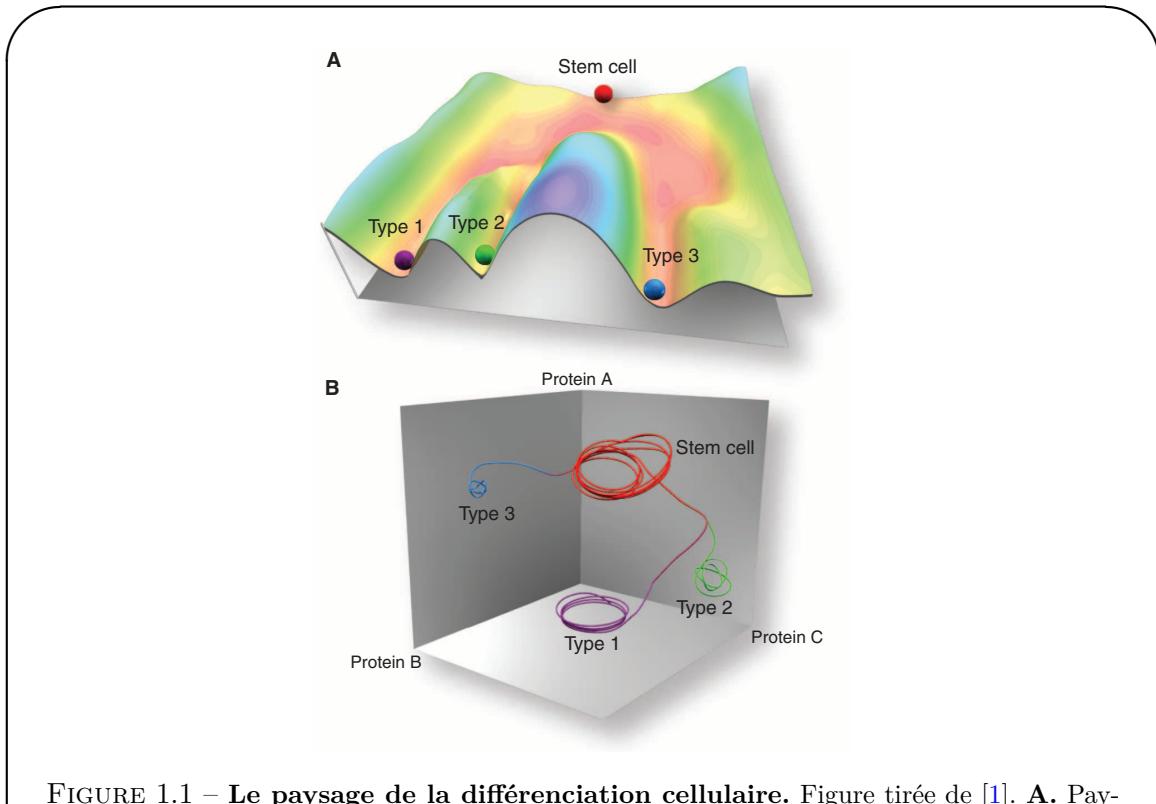


FIGURE 1.1 – Le paysage de la différenciation cellulaire. Figure tirée de [1]. **A.** Paysage épigénétique tel qu’imaginé par Waddington [2] en résonance avec la notion de paysage énergétique en physique. Le développement cellulaire est représenté par une bille dévalant un paysage composé de différentes 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.

1.1.3 La reprogrammation cellulaire

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

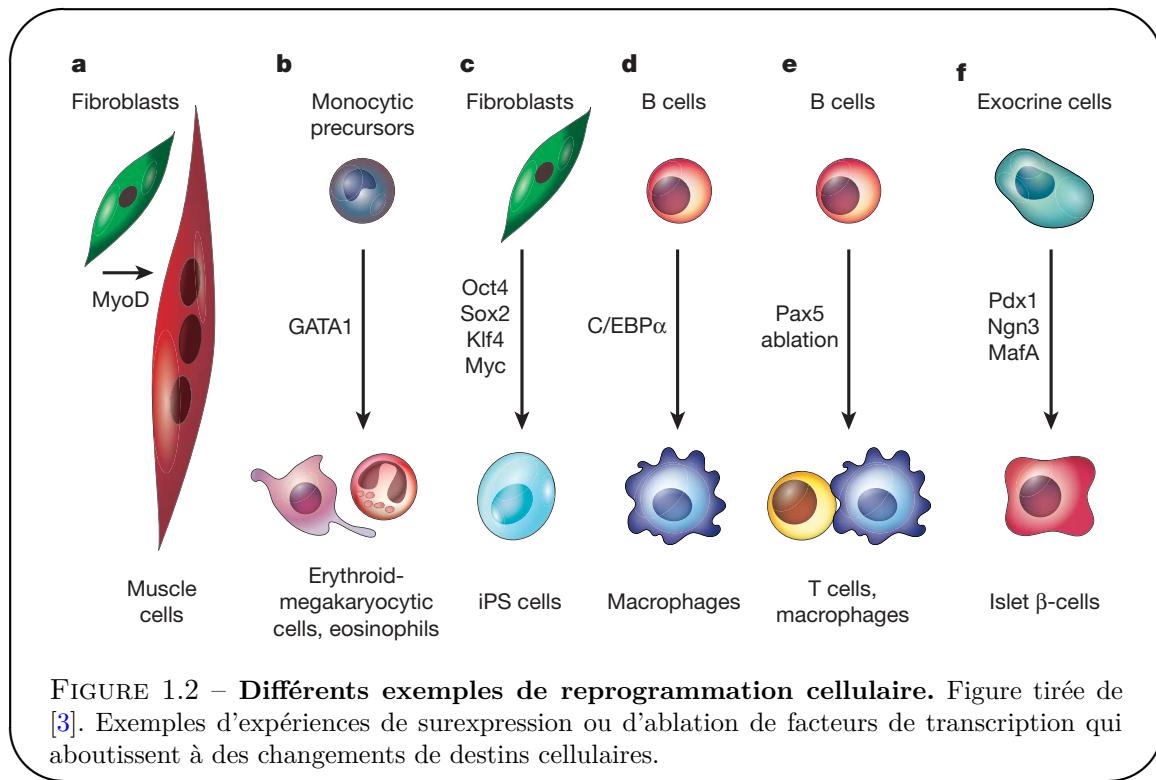


FIGURE 1.2 – Différents exemples de reprogrammation cellulaire. Figure tirée de [3]. Exemples d'expériences de surexpression ou d'ablation de facteurs de transcription qui aboutissent à des changements de destins cellulaires.

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



FIGURE 1.3 – .

1.2 Les réseaux de régulation génétique

1.2.1 Vision cybernétique de la cellule

1.2.2 Divers modes de régulation

- Régulation génétique

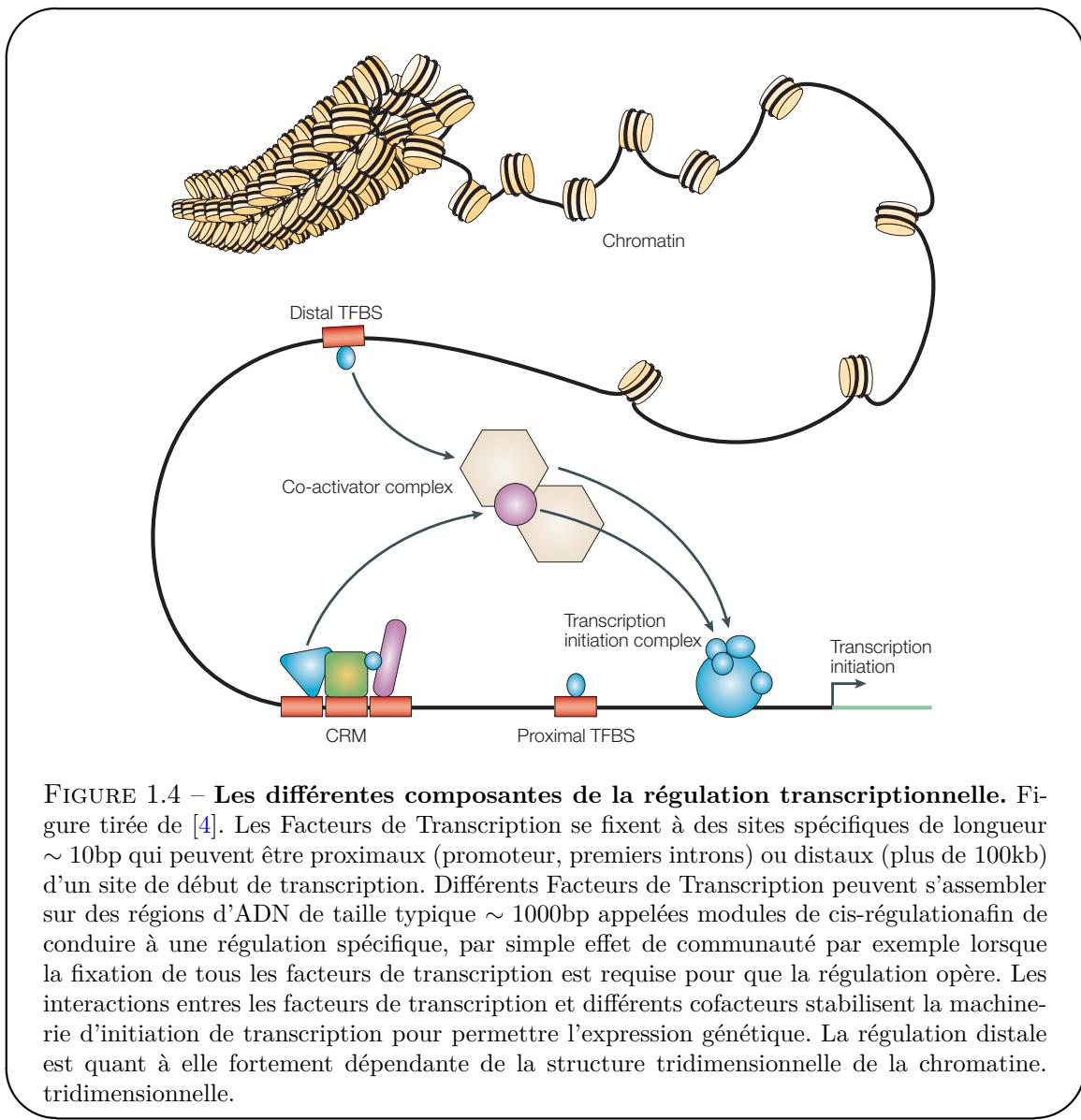


FIGURE 1.4 – Les différentes composantes de la régulation transcriptionnelle. Figure tirée de [4]. Les Facteurs de Transcription se fixent à des sites spécifiques de longueur ~ 10bp qui peuvent être proximaux (promoteur, premiers introns) ou distaux (plus de 100kb) d'un site de début de transcription. Différents Facteurs de Transcription peuvent s'assembler sur des régions d'ADN de taille typique ~ 1000bp 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.

- Régulation épigénétique
- Régulation post-transcriptionnelle
- Régulation post-traductionnelle

1.2.3 Câblage du réseau et fonction

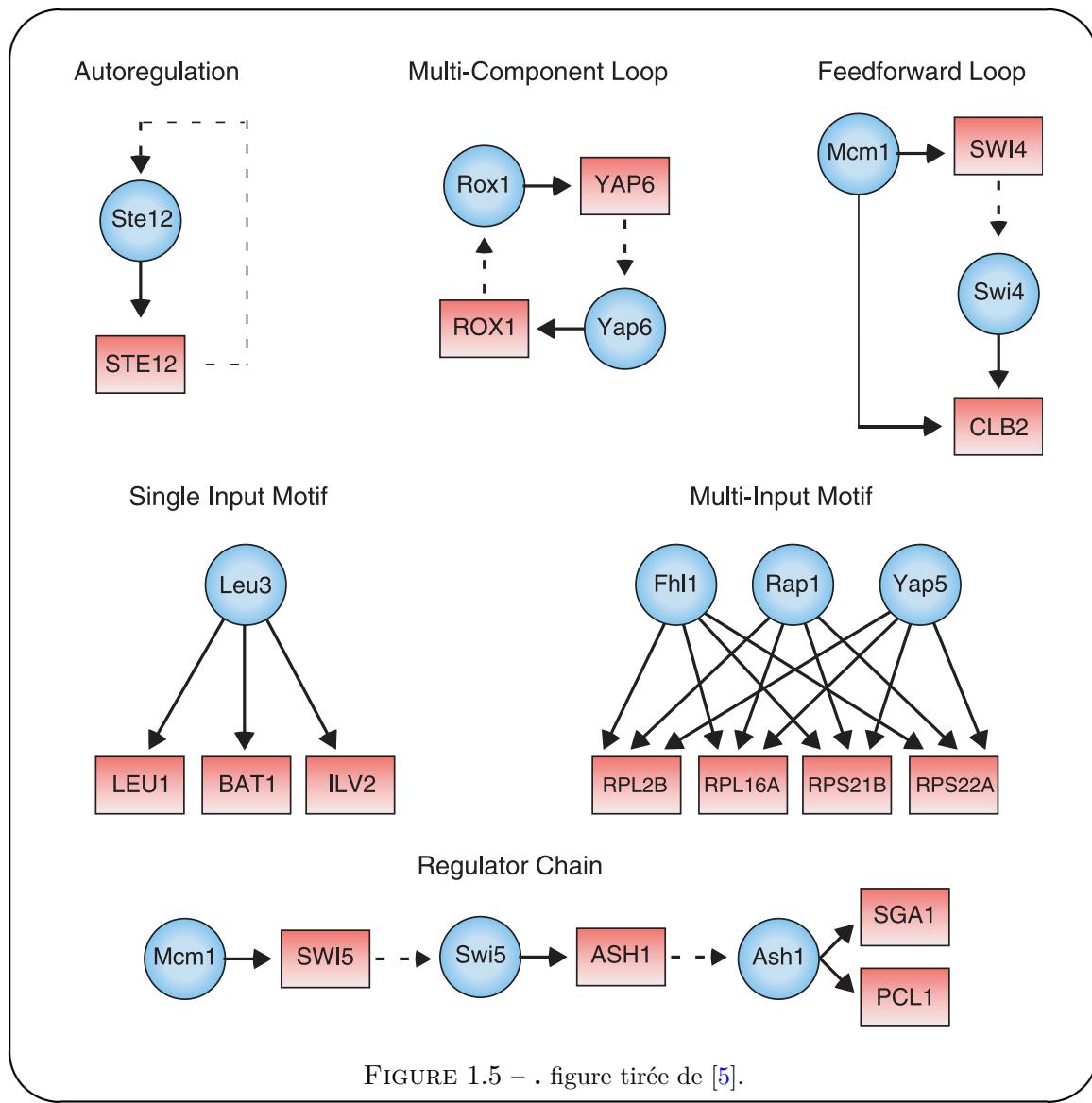


FIGURE 1.5 – . figure tirée de [5].

1.2.4 Évolution des réseaux génétiques

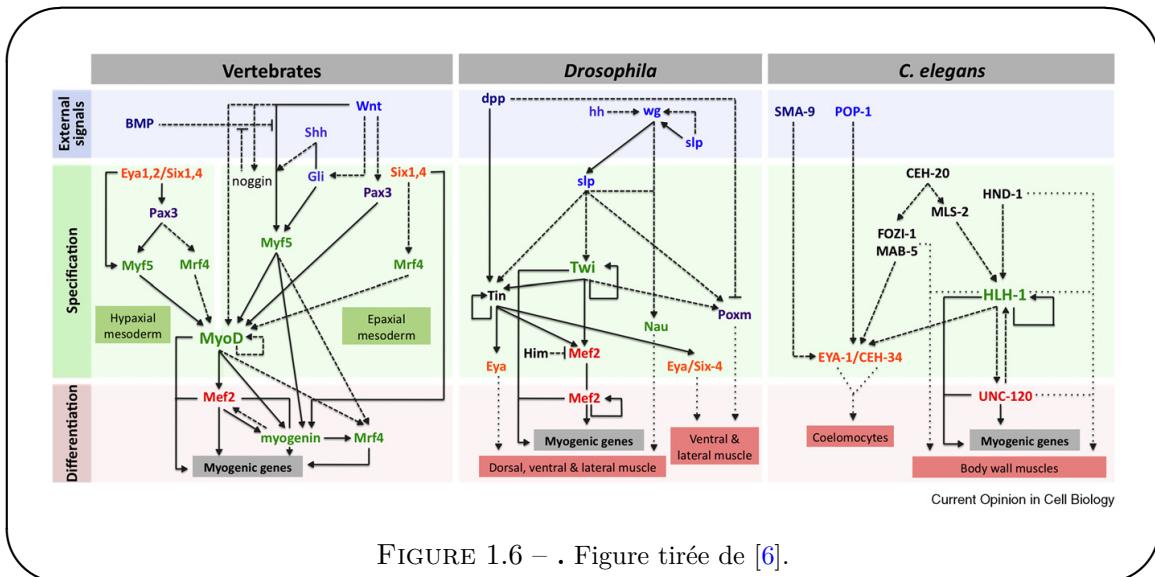


FIGURE 1.6 – . Figure tirée de [6].

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

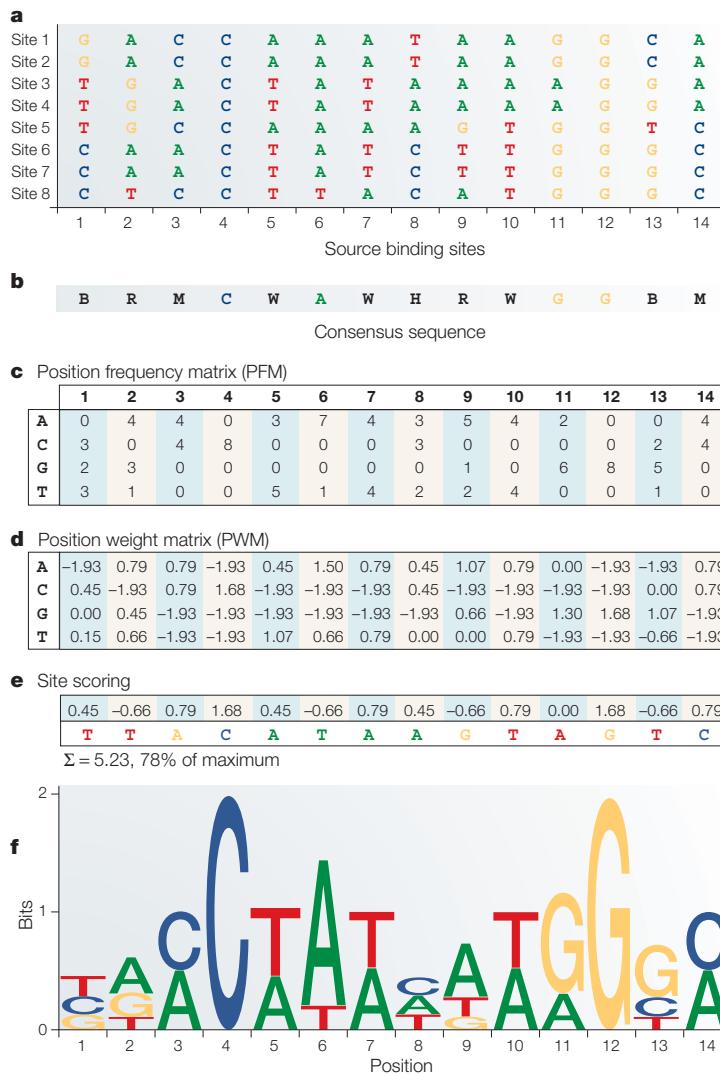


FIGURE 1.7 – Construction et utilisation du modèle PWM. Figure tirée de [4]. (a) Supposons connus un certain nombre de sites de fixation d'un facteur de transcription (dans ce cas MEF2). (b) Séquence consensus correspondante utilisant les symboles IUPAC. (c) Une matrice de fréquence est construite, indiquant 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 [7]. 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

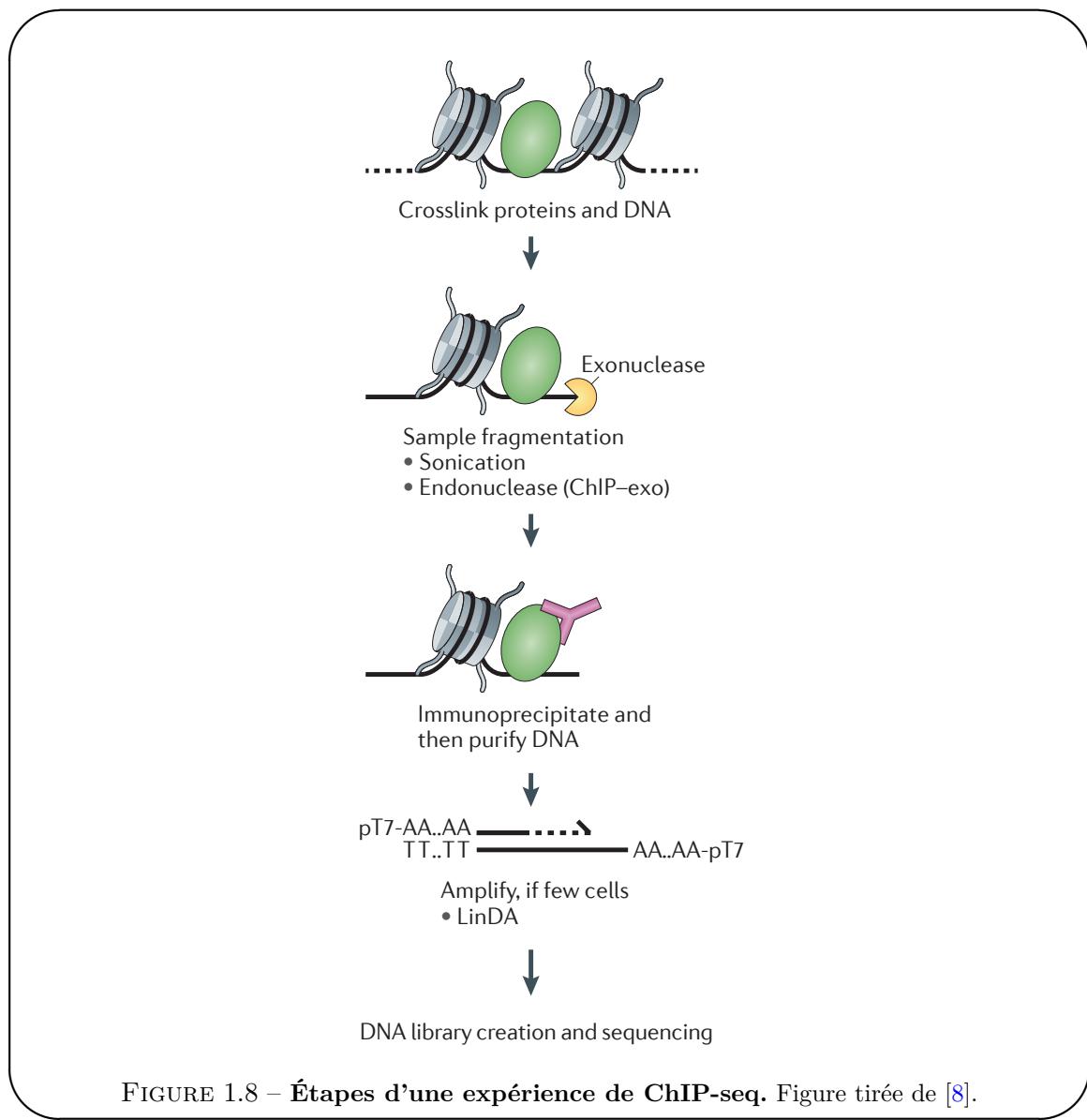


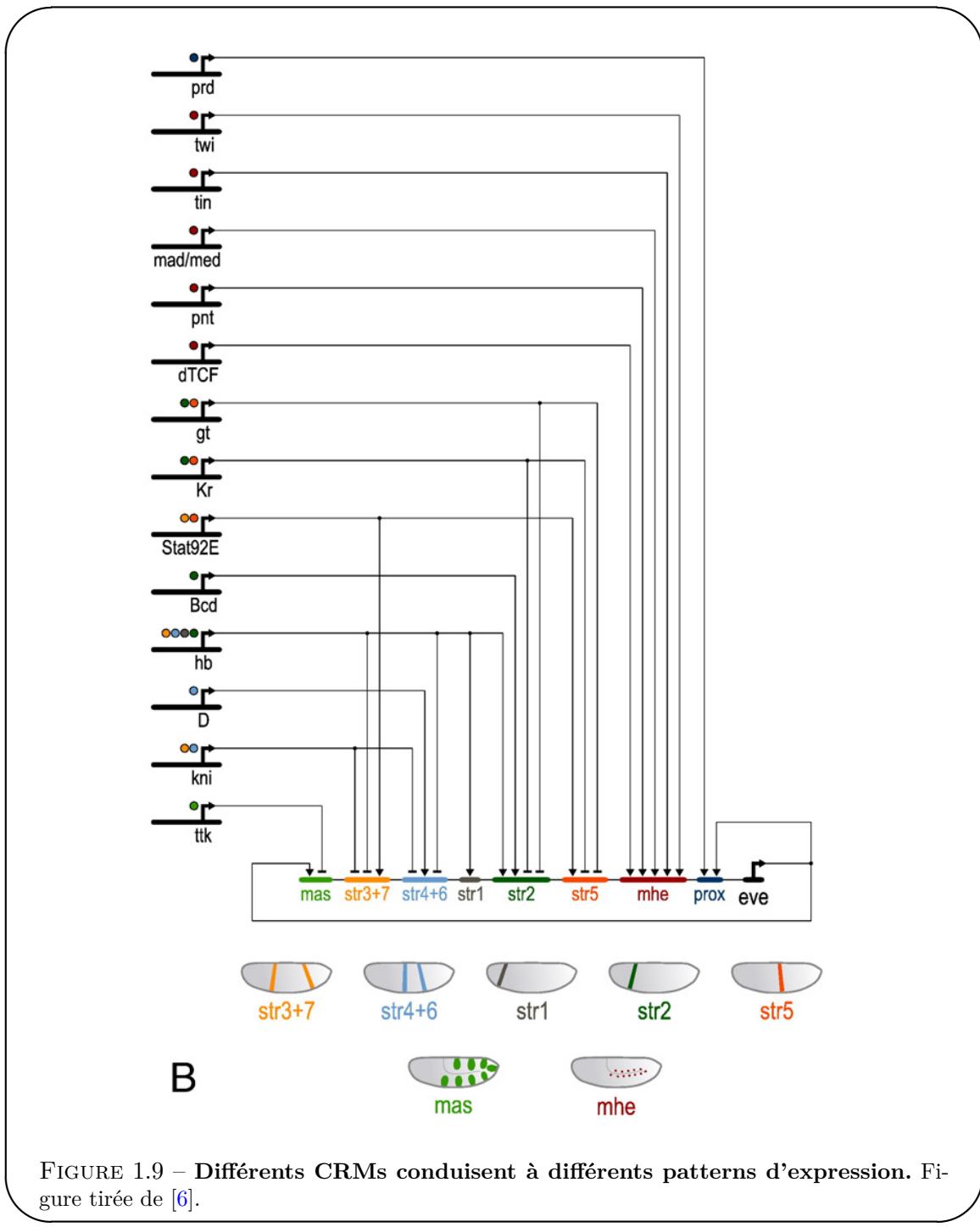
FIGURE 1.8 – Étapes d'une expérience de ChIP-seq. Figure tirée de [8].

- DNase

1.5 Les modules de cis-régulation

1.5.1 Modules et fonctions logiques

1.5.2 Encodage de patterns spatiaux



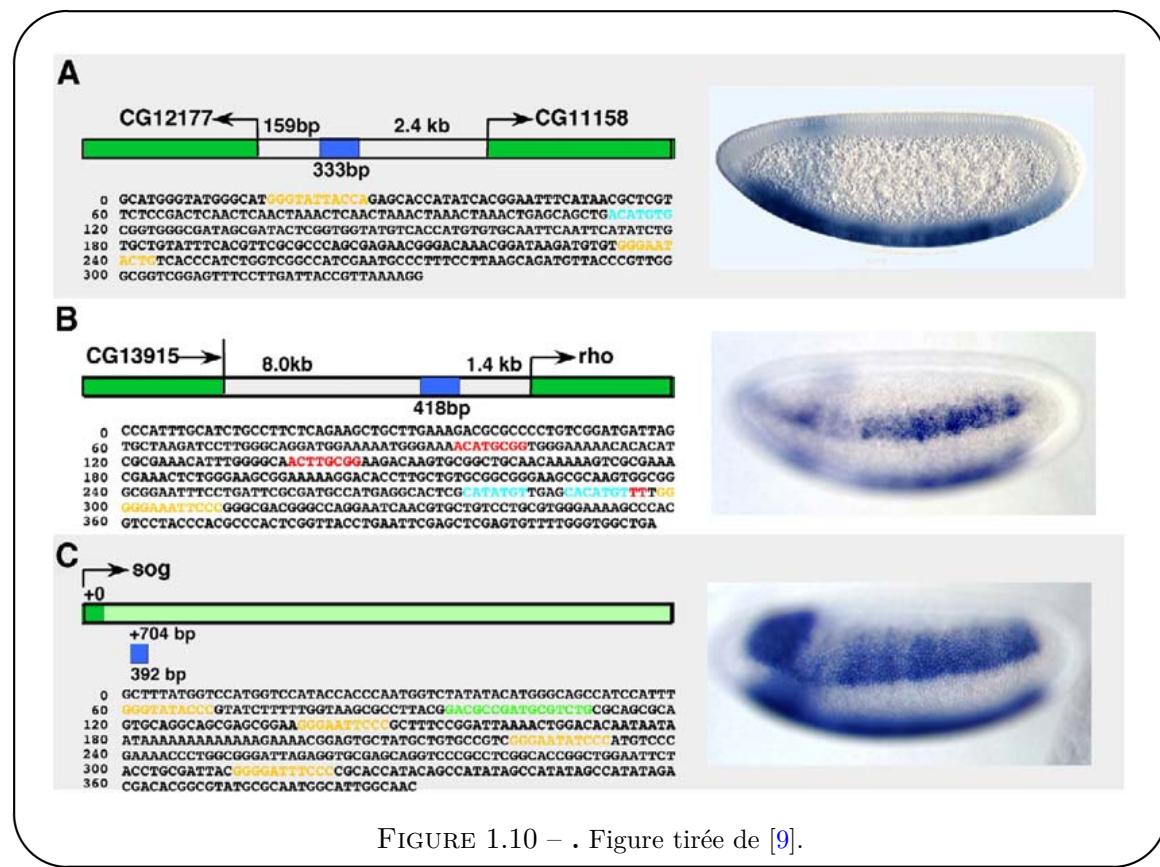


FIGURE 1.10 – . Figure tirée de [9].

1.5.3 Différents états des CRMs

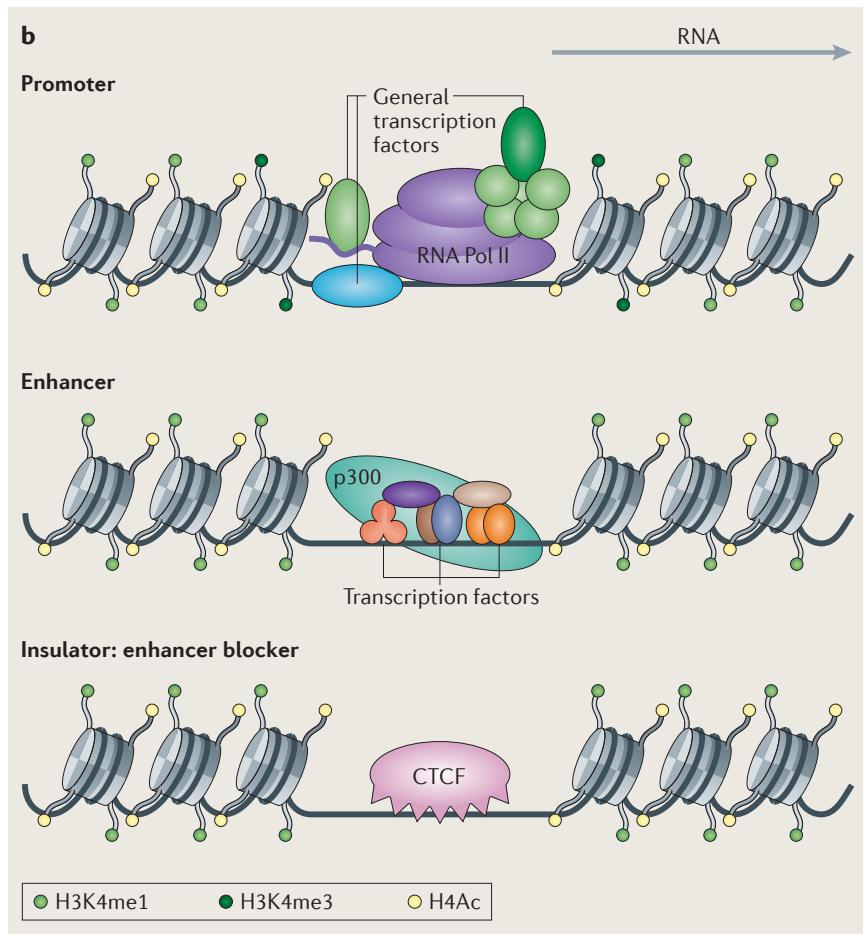


FIGURE 1.11 – Les états épigénétiques des CRMs. Figure tirée de [10].

1.5.4 Prédition des CRMs

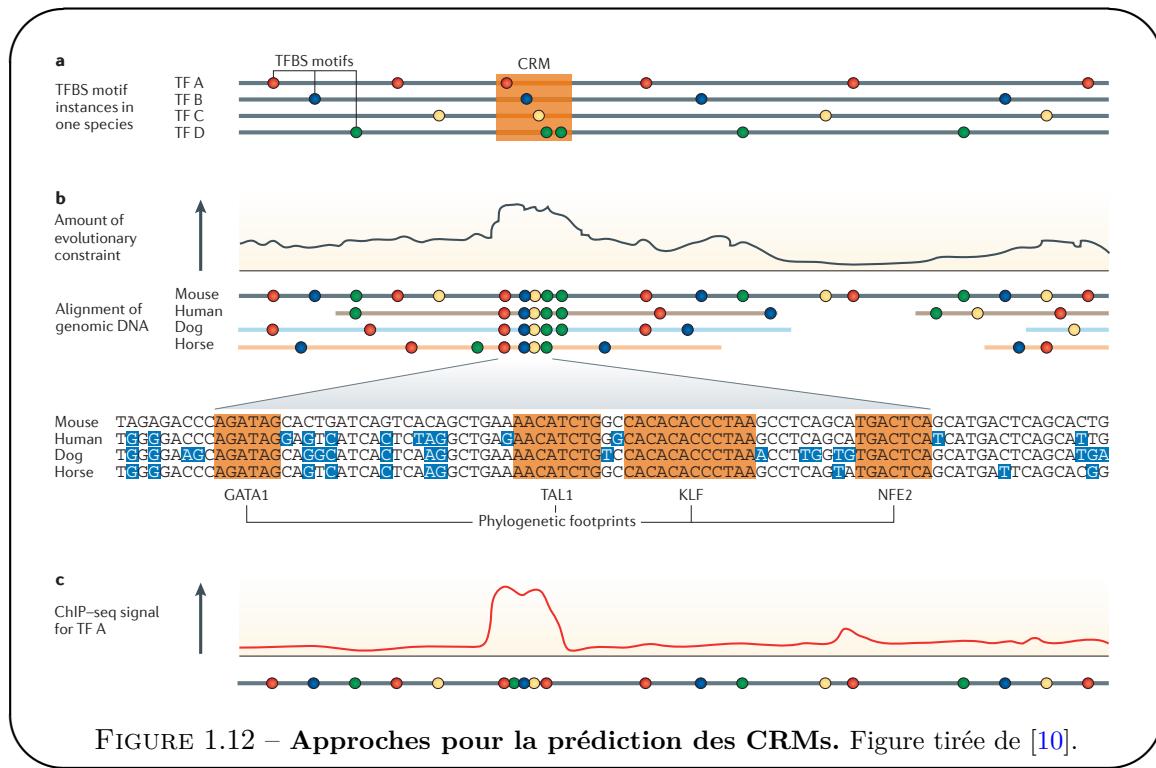
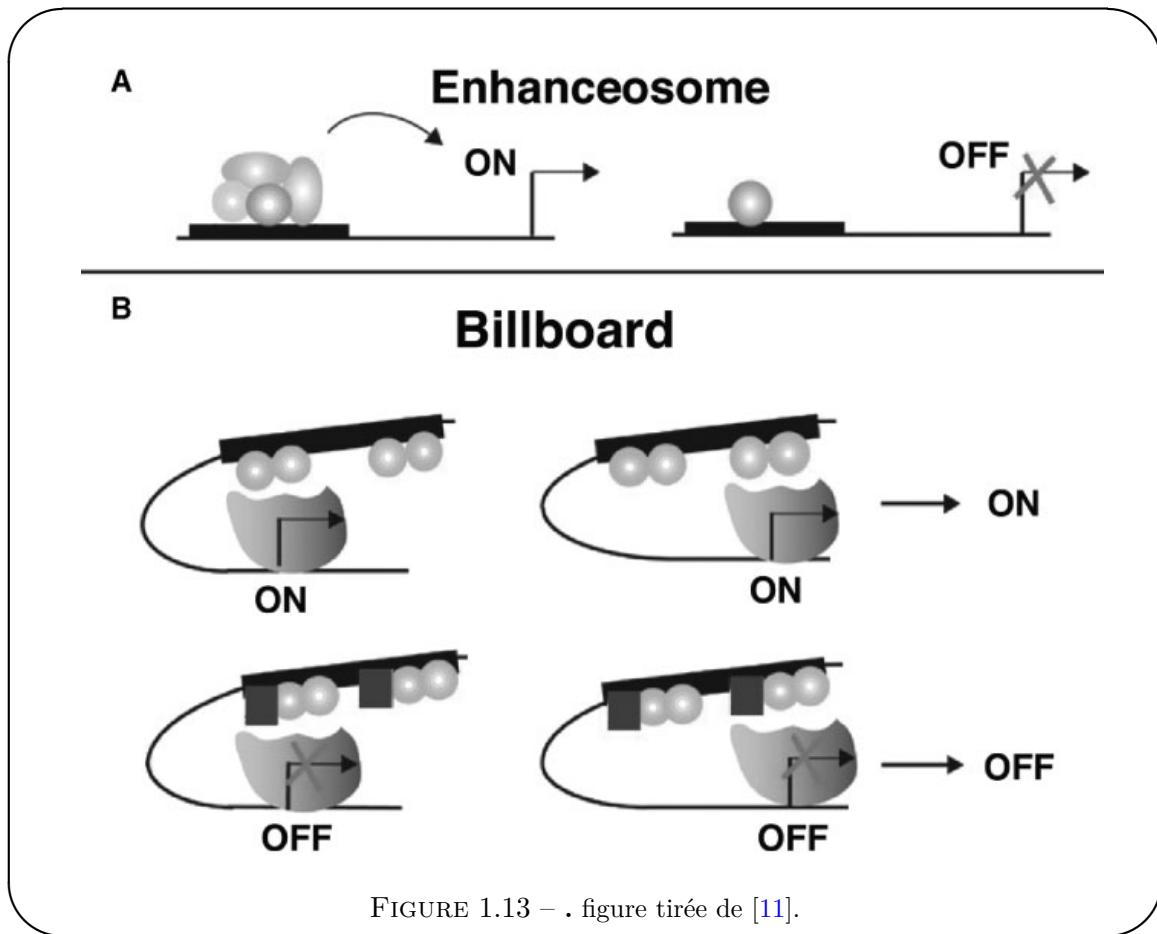


FIGURE 1.12 – Approches pour la prédition des CRMs. Figure tirée de [10].

1.5.5 Grammaire des enhancers : enhanceosome vs billboard



1.5.6 Évolution des enhancers

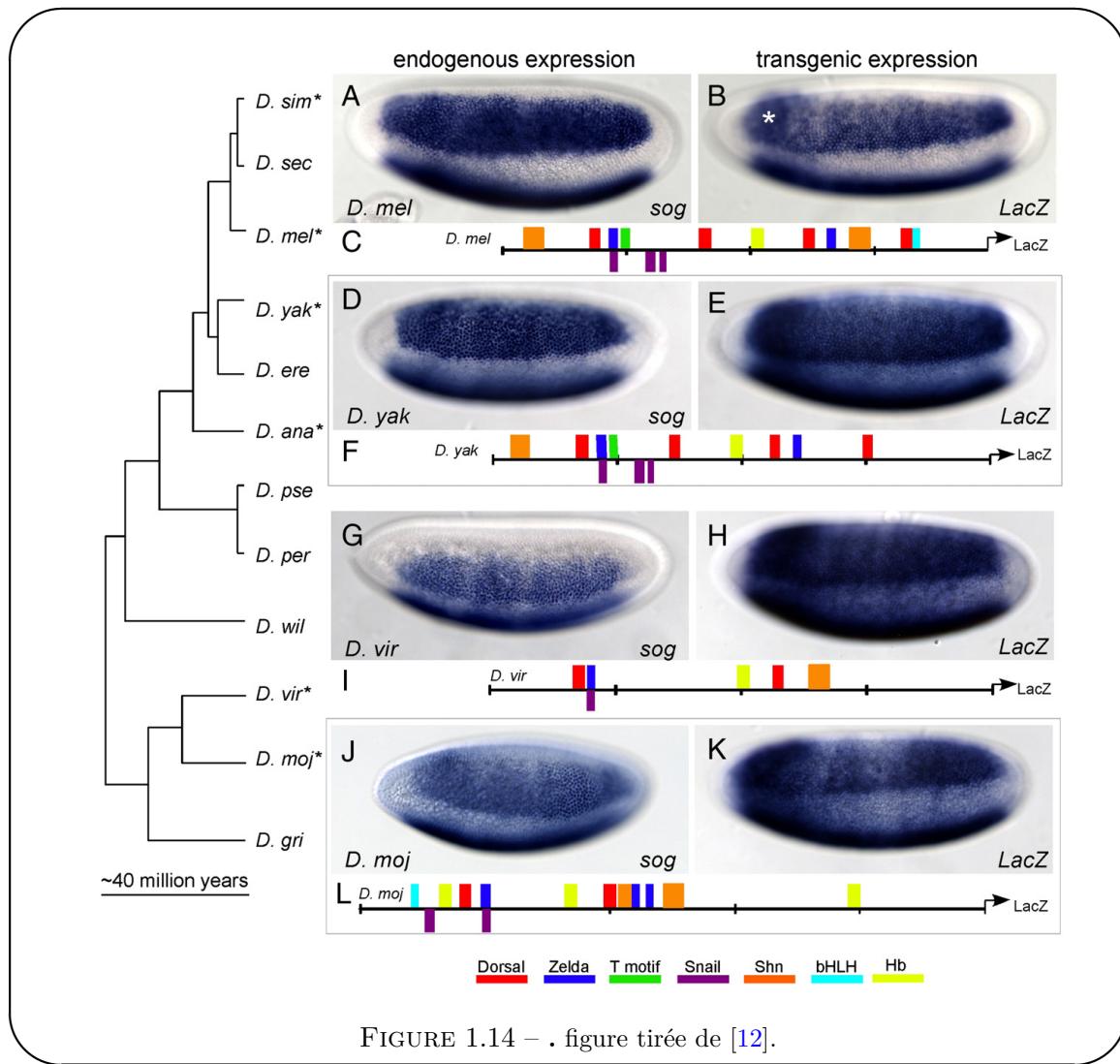
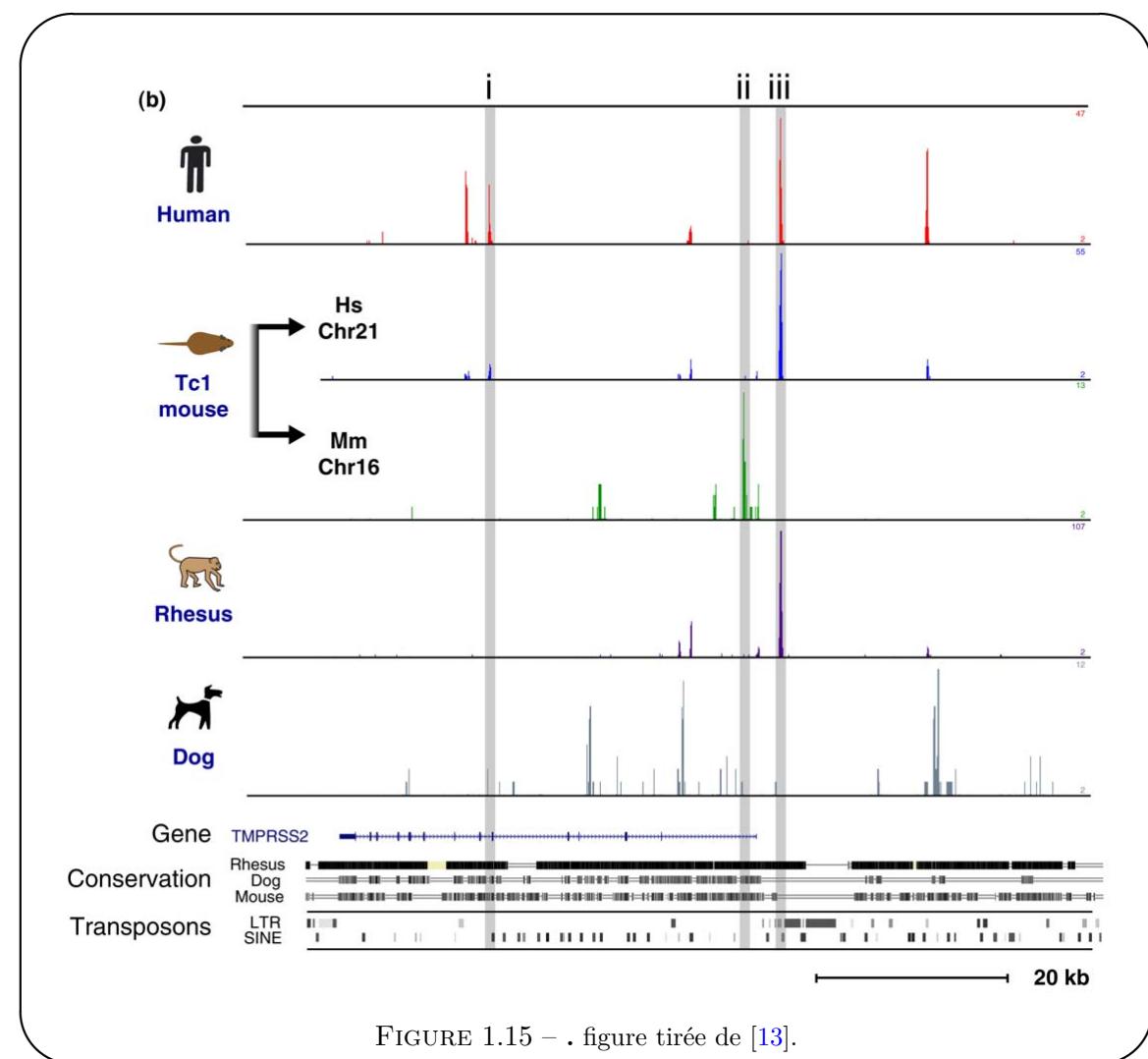


FIGURE 1.14 – . figure tirée de [12].



1.5.7 Validation expérimentale

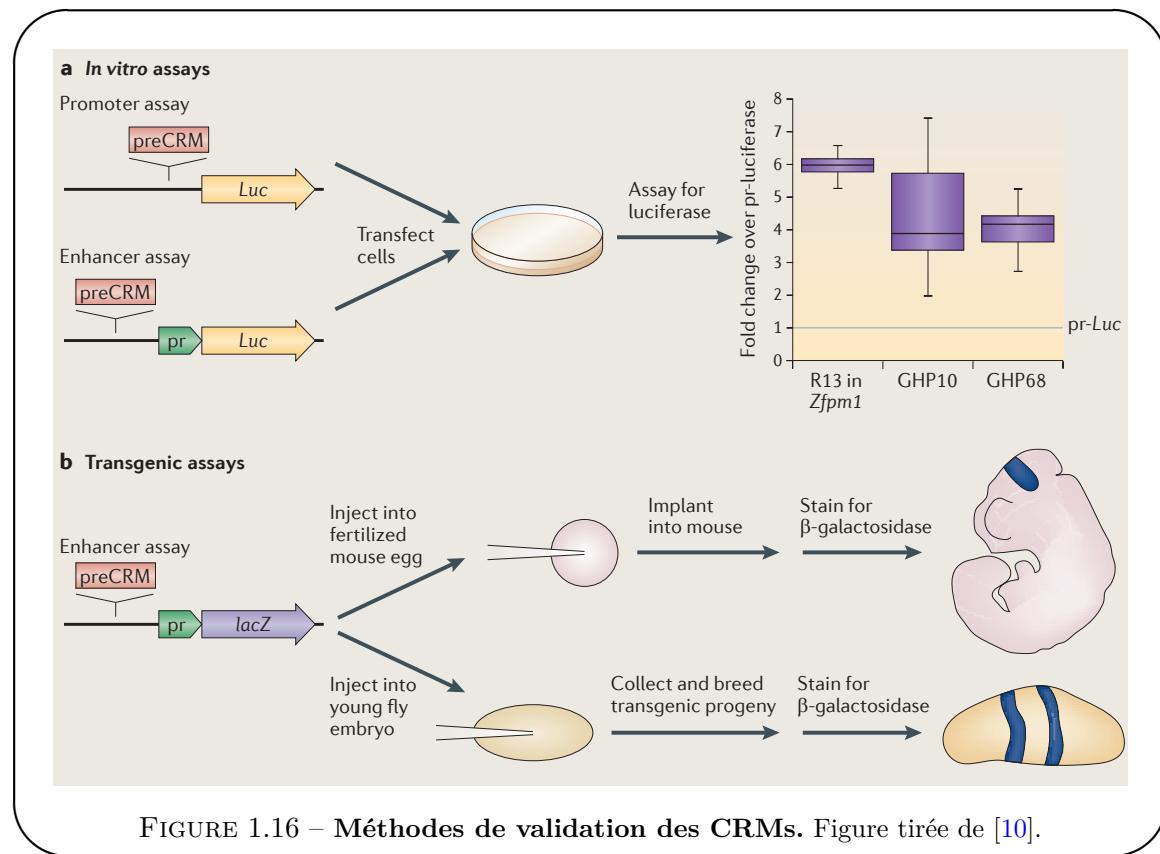


FIGURE 1.16 – Méthodes de validation des CRMs. Figure tirée de [10].

1.6 Banques de données

- 1.6.1 Séquences génomiques et alignements
- 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.

| | | |
|------------|---|-----------|
| 2.1 | Les modèles de fixation | 23 |
| 2.1.1 | Modèles de maximum d'entropie | 23 |
| 2.1.2 | Modèles de mélange | 24 |
| 2.2 | Description des données biologiques | 24 |
| 2.2.1 | Les données ChIP | 24 |
| 2.2.2 | Statistique « background » des séquences | 24 |
| 2.3 | Présentation de l'algorithme | 24 |
| 2.4 | Performance des modèles | 24 |
| 2.5 | Analyse des corrélations | 24 |
| 2.5.1 | Quantification par l'Information Directe | 24 |
| 2.5.2 | Description par des patterns de Hopfield | 24 |
| 2.6 | Comparaison avec des données <i>in vitro</i> | 24 |
| 2.6.1 | Conclusion de la section 2.6 | 25 |

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 [14, 15] sur l'ensemble des distributions de probabilités des états étant données les contraintes imposées. Cette fonctionnelle s'écrit [16]

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

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 [17, 18] et à partir des données du projet ENCODE [19] accessibles à partir du site internet de UCSC¹, 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 [20].

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/>

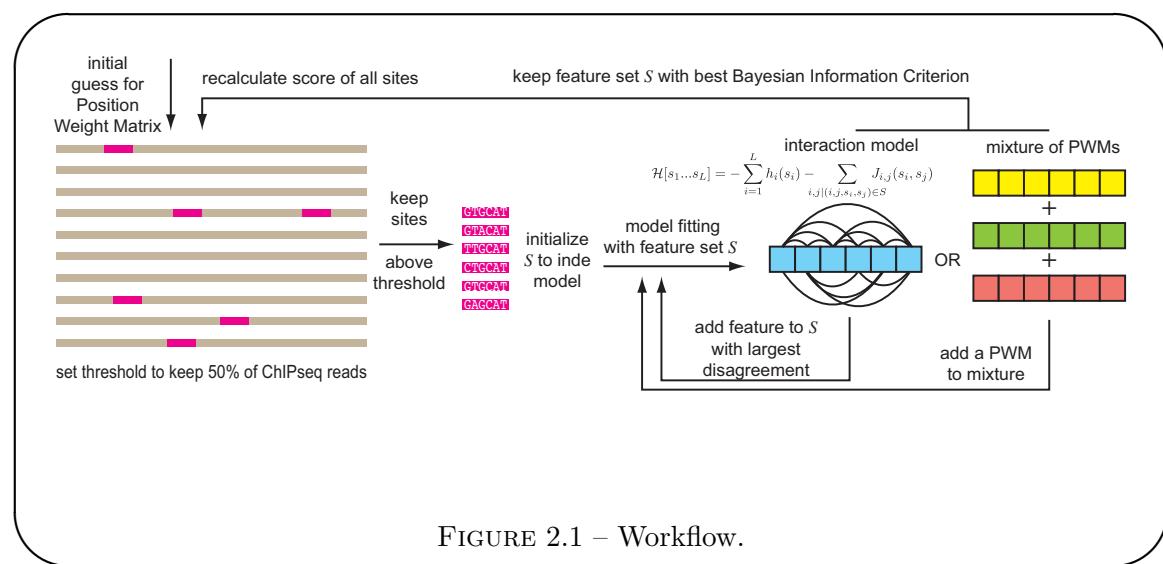
2.6. Comparaison avec des données *in vitro*

FIGURE 2.1 – Workflow.

2.6.1 Conclusion de la section 2.6

[**thèse**: version du lundi 29 avril 2013 à 19 h 21]

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

| | | |
|-----|-------|----|
| 3.1 | | 29 |
|-----|-------|----|

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 | | 33 |
|-----|-------|----|

Introduction du chapitre 4

4.1

Conclusion du chapitre 4

Chapitre 5

Étude de la différenciation musculaire chez la souris

| | | |
|------------|-------|-----------|
| 5.1 | | 37 |
|------------|-------|-----------|

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

| | | |
|------------|------------------------------|-----------|
| 6.1 | Titre de la section | 41 |
| 6.1.1 | Titre de la sous-section | 41 |
| 6.1.2 | Conclusion de la section 6.1 | 41 |

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

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

λ_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 [5](#).)
- [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 [5](#).)
- [3] T. GRAF et T. ENVER, “Forcing cells to change lineages”, *Nature* **462**, n° 7273, 587–94 (Dec 2009). (Page [6](#).)
- [4] 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 [11](#).)
- [5] 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 [8](#).)
- [6] 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 [9](#) et [13](#).)
- [7] L. M. GIOCINO, M.-B. MOSER et E. I. MOSER, “Computational models of grid cells”, *Neuron* **71**, n° 4, 589–603 (Aug 2011). (Page [11](#).)
- [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 [12](#).)
- [9] A. STATHOPOULOS et M. LEVINE, “Genomic regulatory networks and animal development”, *Dev Cell* **9**, n° 4, 449–62 (Oct 2005). (Page [14](#).)
- [10] L. HARTWELL, J. HOPFIELD, S. LEIBLER et A. MURRAY, “From molecular to modular cell biology”, *Nature* **402**, n° 6761, 47 (1999). (Pages [15](#), [16](#) et [20](#).)
- [11] D. N. ARNOSTI et M. M. KULKARNI, “Transcriptional enhancers : Intelligent enhancerosomes or flexible billboards ?”, *J Cell Biochem* **94**, n° 5, 890–8 (Apr 2005). (Page [17](#).)
- [12] 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 [18](#).)
- [13] 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 [19](#).)

Bibliographie

- [14] E. JAYNES, “Information theory and statistical mechanics. II”, *Physical review* **108**, n° 2, 171 (1957). (Page [23](#).)
- [15] C. SHANNON, “A Mathematical Theory of Communication”, *Bell Syst Tech J* **27**, n° 4, 623–656 (Jan 1948). (Page [23](#).)
- [16] 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 [23](#).)
- [17] 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 [24](#).)
- [18] 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 [24](#).)
- [19] E. P. CONSORTIUM, “A user’s guide to the encyclopedia of DNA elements (ENCODE)”, *Plos Biol* **9**, n° 4, e1001046 (Apr 2011). (Page [24](#).)
- [20] A. F. A. SMIT, R. HUBLEY et P. GREEN, “RepeatMasker Open-3.0”, (1996-2010). (Page [24](#).)

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 29 avril 2013 à 19 h 21]