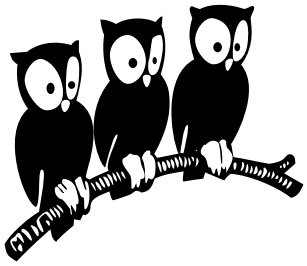
Département de Physique
École Normale Supérieure

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

From cis-regulatory DNA motifs to tissue specific expression

Soutenue le ZZ septembre 2013
devant le jury composé de :

M.	**Vincent HAKIM**	.....	Directeur de thèse
M.	**ZZZ**	.....	Rapporteur
M.	**ZZZ**	.....	Examinateur
M.	**ZZZ**	.....	Président du jury
M.	**ZZZ**	.....	Rapporteur
M.	**ZZZ**	.....	Membre invité
M.	**ZZZ**	.....	Membre invité

Remerciements

. . .

Table des matières

Liste des figures

Introduction

Les réseaux de régulation génétique

- **Bref historique**
  Monod, Jacob. Promoteurs.

- **Divers modes de régulation**

  → Promoteurs

  → Enhancers

  → Épigénétique

Interactions entre les facteurs de transcription et l'ADN.

Quelques remarques sur la version pdf du manuscrit

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

Modèles d'accrochage des Facteurs de Transcription à l'ADN.

Introduction du chapitre 1

1.1 Modèles de maximum d'entropie

1.1.1 Le modèle PWM

?? Nous suivons la méthodologie de [1, 2]. The pairwise model distribution is the one that maximizes the entropy given the one and two point correlations. In a general way, the model distribution P_m maximizes

const

1.1.2 Conclusion de la section 1.1

Chapitre 2

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

Introduction du chapitre 2

2.1

Chapitre 3

Étude de la différenciation épidermale chez la drosophile

Introduction du chapitre 3

3.1

Conclusion du chapitre 3

Chapitre 4

Étude de la différenciation musculaire chez la souris

Introduction du chapitre 4

4.1

Conclusion du chapitre 4

Chapitre 5

Chapitre d'exemples

Introduction du chapitre 5

5.1 Titre de la section

FIGURE 5.1 – Caption longue, pour mettre sous la figure.

5.1.1 Titre de la sous-section

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

$$\widehat{H} = \int d^3\overrightarrow{r} \int_0^\infty d\omega \hbar\omega \overrightarrow{\widehat{f}}^\dagger(\overrightarrow{r},\omega) \cdot \overrightarrow{\widehat{f}}(\overrightarrow{r},\omega) + \sum_{\alpha=i,f} \hbar\omega_\alpha \widehat{\xi}_\alpha + \widehat{H}_Z \tag{5.1}$$

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

FIGURE 5.2

$$\begin{cases} \overrightarrow{H_i} &= H_0 \overrightarrow{u_y} e^{i(\alpha_i x - \gamma_i z)} \\ \overrightarrow{H_r} &= r_m H_0 \overrightarrow{u_y} e^{i(\alpha_i x + \gamma_i z)} \\ \overrightarrow{H_t} &= t_m H_0 \overrightarrow{u_y} e^{i(\alpha_i x - \gamma_t z)} \end{cases}$$

$$\Gamma_{i\to f} = \frac{27}{64} \frac{n_{th}+1}{\tau_0} \left(\frac{c}{\omega}\right)^3 \frac{1}{d^4} \frac{2}{\mu_0 \omega} \mathrm{Re}\left(Z_S\right) \tag{5.2}$$

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

5.1.2 Conclusion de la section 5.1

Conclusion

Résumé

Perspectives

Liste des Annexes

Annexe A

Titre

Bibliographie

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

[1] T. Mora et W. Bialek, "Are biological systems poised at criticality ?", *J Stat Phys* **144**, n° 2, 268–302 (2011). (Page 5.)

[2] E. Jaynes, "Information theory and statistical mechanics. II", *Physical review* **108**, n° 2, 171 (1957). (Page 5.)

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.