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.	**Emmanuel BARILLOT**	Examinateur
M.	**Vincent HAKIM**	Directeur de thèse
M.	**Pascal MAIRE**	Examinateur
M.	**Massimo VERGASSOLA**	Rapporteur
M.	**Martin WEIGT**	Rapporteur
M.	**Alain ZIDER**	Examinateur

Remerciements

...

Table des matières

Liste des figures vii

Principales abbréviations utilisées ix

Avant-propos 1

Chapitre 1 - Introduction générale. 3

1.1 Le phénotype cellulaire 5
1.2 Les réseaux de régulation génétique 10
1.3 Les interactions protéine-ADN : modèles mathématiques 20
1.4 Les interactions protéine-ADN : mesures expérimentales 27
1.5 Les modules de cis-régulation (CRMs) 37
1.6 Prédiction et validation des CRMs 50
1.7 Bases de données 59

Chapitre 2 - Modèles de fixation des Facteurs de Transcription à l'ADN. 65

2.1 Observations de corrélations au sein des TFBS 66
2.2 Modèles existants permettant de décrire la statistique des TFBS 67
2.3 Modèles de maximum d'entropie 72
2.4 Article 77
2.5 Analyse thermodynamique des modèles 109
2.6 Conclusion et perspectives 112

Chapitre 3 - *Imogene* : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle 115

3.1 Les approches existantes pour la recherche de motifs et de modules de régulation 116
3.2 Article 127
3.3 Calcul de la moyenne de la postérieure par une méthode MCMC 156

Annexe A - Statistiques génomique 165

Bibliographie 168

Liste des figures

Introduction générale. 3

1.1 Le paysage de la différenciation cellulaire 6
1.2 Spécification spatio-temporelle du type cellulaire 8
1.3 Différents exemples de reprogrammation cellulaire 9
1.4 Vision cybernétique du traitement de l'information par la cellule 11
1.5 Un réseau de régulation génétique type 12
1.6 Caractéristiques de l'épigénome 15
1.7 Exemples de motifs dans les réseaux de régulation génétique 16
1.8 Évolution du réseau transcriptionnel : l'exemple de la régulation myogénique. . 18
1.9 Différents états du facteur de transcription 20
1.10 Construction et utilisation du modèle PWM 23
1.11 Étapes d'une expérience de ChIP-on-chip et ChIP-seq 33
1.12 Résolution des expériences ChIP-on-chip et ChIP-seq 35
1.13 Expérience d'empreinte à la DNase I chez la levure : vers une résolution au nucléotide près 36
1.14 Les différents types de CRMs et leurs marques épigénétiques 38
1.15 Différents *enhancers* conduisent à différents patterns d'expression 40
1.16 Deux modèles d'enhancers : enhanceosome et billboard 42
1.17 L'enhanceosome de l'interferon-β 43
1.18 Flexibilité du code de cis-régulation au cours de l'évolution chez les *Drosophiles* 44
1.19 Évolution de la fixation de HNF4α chez les mammifères 46
1.20 « Shadow enhancer » du gène de segmentation *Hunchback* 48
1.21 De l'enhancer au super-enhancer 49
1.22 Différentes approches pour la prédiction des CRMs 51
1.23 Méthodes de validation des CRMs par transfection et transgenèse 56
1.24 Impact physiologique de la délétion et de la mutation d'un enhancer 57
1.25 Évolution du coût de séquençage 59
1.26 Visualisation de données ChIP-seq *via* le site UCSC 62

1.27 Les différentes données obtenues par le projet ENCODE 63

Modèles de fixation des Facteurs de Transcription à l'ADN. 65

2.1 Différents modèles pour décrire les corrélations entre nucléotides dans les sites de fixation de facteurs de transcription 68
2.2 Illustration d'un système dont on veut maximiser l'entropie 72
2.3 Chaleur spécifique pour différents TFs 111
2.4 Histogrammes des valeurs des champs h et couplages J 113

***Imogene* : un algorithme d'identification de motifs et de modules de régulation transcriptionnelle 115**

3.1 Illustration de l'approche espérance-maximisation 117
3.2 Conditions initiales 160
3.3 Corrélations entre les échantillons 161
3.4 Estimation de la convergence 163
3.5 Comparaison des approches par MCMC et par descente de gradient 164

Statistiques génomique 165

A.1 Distribution des tailles intergéniques et introniques chez différentes espèces . . 166

Principales abbréviations utilisées

ARNm	ARN messager
bHLH	*basic Helix-Loop-Helix*
bp	Paire de base
ChIP	Immunoprécipitation de la chromatine (*Chromatin immunoprecipitation*)
CRM	Module de cis-régulation (*Cis-Regulatory Module*)
DHS	Hypersensible à la DNase I (*DNaseI-hypersensistive*)
ESC	Cellule souche embryonnaire (*Embryonic Stem Cell*)
ISH	Hybridation *in situ* (*In-Situ* Hybridization)
kb	kilobases (1000bp)
MRF	Facteur de régulation myogénique (*Myogenic Regulatory Factor*)
nt	Nucléotide
PCR	Réaction en chaîne par polymérase (*Polymerase Chain Reaction*)
PWM	Matrice de poids (*Position Weight Matrix*)
TF	Facteur de transcription (*Transcription Factor*)
TFBS	Site de fixation d'un facteur de transcription (*Transcription Factor Binding Site*)
TSS	Site d'initiation de la transcription (*Transcription Start Site*)

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 5

1.1.1 Qu'est-ce que le phénotype d'une cellule ? 5

1.1.2 La différenciation cellulaire 6

1.1.3 La cellule dans l'organisme : une spécification spatio-temporelle 8

1.1.4 La reprogrammation cellulaire 8

1.2 Les réseaux de régulation génétique 10

1.2.1 Vision cybernétique de la cellule 10

1.2.2 Divers modes de régulation 10

1.2.3 Câblage du réseau et fonction 17

1.2.4 Évolution des réseaux génétiques 17

1.3 Les interactions protéine-ADN : modèles mathématiques 20

1.3.1 Modes de recherche du site de fixation par le TF 21

1.3.2 Modèle PWM 21

1.3.3 Modèle biophysique 24

1.3.4 Modèle thermodynamique 25

1.4 Les interactions protéine-ADN : mesures expérimentales 27

1.4.1 Approches *in vitro* : MITOMI, SPR, PBM, CSI, SELEX, et HT-SELEX 28

1.4.2 Approche clonale : la technique de simple hybride 31

1.4.3 Approches *in vivo* : ChIP-on-chip, ChIP-seq, DNAse I 32

1.5 Les modules de cis-régulation (CRMs) 37

1.5.1 Les différents types de CRMs 37

1.5.2 Grammaire des enhancers : enhanceosome vs billboard 41

1.5.3 Évolution des enhancers 43

1.5.4 Les « shadow enhancers » 47

1.5.5 Par delà les enhancers : les « super-enhancers » 48

1.6 Prédiction et validation des CRMs 50

1.6.1 Méthodes utilisant la concentration en sites de fixation 50
1.6.2 Méthodes utilisant la phylogénie 52
1.6.3 Méthodes utilisant les marques épigénétiques et de ChIP-seq pour des TFs 54
1.6.4 Validation expérimentale 55
1.6.5 Implication des CRMs dans les maladies humaines 57
1.7 Bases de données 59
1.7.1 Obtention de données génomiques 59
1.7.2 Obtention de données sur les TFs 61
1.7.3 Outils de visualisation 61
1.7.4 Le projet ENCODE 63

1.1 Le phénotype cellulaire

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

Les organismes vivants sont constitués de cellules de l'ordre de quelques microns, facilement observables à l'aide d'un simple microscope optique. Chaque cellule contient 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 auxquels nous nous intéressons dans cette thèse. Les cellules qui les constituent sont majoritairement eucaryotes, c'est-à-dire qu'elles possèdent un noyau renfermant le matériel génétique. [1]

Bien que possédant toutes le même matériel génétique (à quelques variations près), les cellules d'un organisme apparaissent d'emblée comme hétérogènes, que ce soit dans leur forme ou dans leurs constituants. Par exemple, chez l'homme, les neurones sont composés d'un corps d'une dizaine de microns de diamètre contenant le noyau, de milliers de prolongements dont certains peuvent atteindre une taille de plus de 1.5 mètre, et sont riches en ions sodium et potassium, tandis que les fibres musculaires squelettiques sont de forme longue et tubulaire, possèdent plusieurs dizaines de noyaux et expriment actine et myosine.

Cette diversité semble néanmoins limitée. Aussi, parmi les $\sim 6\ 10^{13}$ cellules du corps humain, on peut distinguer ~ 320 différents types cellulaires (Brazma et al., 2001). Bien entendu, ce nombre dépend du seuil de similarité choisi : deux cellules d'un même type n'expriment pas *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 de 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 (étymologiquement « exhiber un type » en grec). Nous allons le voir, ce phénotype est le résultat de la modulation par des facteurs environnementaux de l'expression génétique qui conditionne le contenu en protéines de la cellule.

1. Il existe cependant quelques cas connus d'organismes multicellulaires procaryotes, dont les cellules ne possèdent pas de noyau, par exemple chez les bactéries magnétotactiques (Keim et al., 2004).

1.1.2 La différenciation cellulaire

FIGURE 1.1 – **Le paysage de la différenciation cellulaire.**
Figure tirée de (Furusawa and Kaneko, 2012). **A.** Paysage épigénétique tel qu'imaginé par Waddington (Waddington et al., 1957) 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 difficilement franchissables, 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.

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. Schématiquement, au cours du développement d'un organisme, des cellules non différenciées ou souches empruntent un chemin unidirectionnel de différenciation qui restreint peu à peu le nombre de types cellulaires qu'elles peuvent potentiellement devenir, passant de l'état souche totipotent à des états pluripotents successifs avant la différenciation

finale. Ainsi, la formation des cellules somatiques, qui sont les cellules d'un organisme n'étant ni souches ni germinales (les cellules qui donnent naissance aux gamètes ou cellules sexuelles), est le résultat d'un processus de différenciation initial lors du développement embryonnaire au cours duquel les cellules issues de l'œuf 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 et les os (mésoderme), ou encore la peau et le système nerveux (ectoderme).

Dans un écrit aujourd'hui célèbre datant de 1957 (Waddington et al., 1957), 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 processus 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 autres 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 par leur rôle primordial dans la définition de l'état cellulaire, l'expression des protéines (et donc des gènes qui les produisent) domine généralement 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, induisent un changement de l'état cellulaire, ce qui se traduit par une trajectoire dans l'espace des 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 (Kaufmann, 1993).

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

FIGURE 1.2 – **Spécification spatio-temporelle du type cellulaire.**
Hybridation *in situ* de l'ARN du gène *Myog*, marqueur de la différenciation des progéniteurs du muscle squelettique, chez des embryons de souris âgés de 9.5, 10.5 et 11.5 jours (de gauche à droite), observés sous un même grossissement de 10. Le motif (*pattern*) de spécification du muscle squelettique est clairement visible au niveau des somites, lieu des futures vertèbres. Images tirées de la base de donnée Embrys (http://embrys.jp).

Au sein de l'organisme, la différenciation cellulaire 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 de son environnement – et du temps – le stade de développement de l'organisme –. Il est ainsi possible d'observer chez l'embryon certains motifs ou *patterns* spatio-temporels d'expression génétique correspondant à des organes en formation et révélés par par hybridation *in situ* de l'ARN de certains gènes spécifiques d'un type cellulaire. Par exemple, dans le cas de la formation des muscles squelettiques, le gène de différenciation terminale *Myog* est exprimé chez la souris dès 8 jours embryonnaires au niveau des somites, segments correspondant aux futures vertèbres de la souris adulte, puis commence à être exprimé au niveau des bourgeons de membres à 11.5 jours (voir fig 1.2).

1.1.4 La reprogrammation cellulaire

Depuis plusieurs décennies, différentes expériences ont exhibé la plasticité des états différenciés, élargissant ainsi considérablement la vision classique selon laquelle des cellules souches totipotentes se différencient de manière irréversible en des cellules de moins en moins plastiques, jusqu'à atteindre un état différencié stable. Par exemple, (Blau et al., 1985) ont mon-

FIGURE 1.3 – **Différents exemples de reprogrammation cellulaire.**
Figure tirée de (Graf and Enver, 2009), montrant plusieurs exemples d'expériences de surexpression ou d'ablation de facteurs de transcription aboutissant à des changements de types cellulaires.

tré en 1985 que des programmes d'expression génétique dormants peuvent être exprimés de manière dominante dans des cellules différenciées par la fusion de différents types cellulaires : ainsi, la fusion de cellules musculaires avec des cellules non musculaires permettait l'activation des gènes de type musculaire dans le type cellulaire non musculaire. Puis différents travaux ont montré qu'il était possible de convertir des lignées de cellules différenciées en un autre type cellulaire en introduisant certaines protéines régulatrices de la transcription, ou Facteurs de Transcription (TFs) (Davis et al., 1987; Kulessa et al., 1995) : on parle alors de transdifférenciation, dont l'un des exemples canoniques est la différenciation de cellules de la peau ou fibroblastes en cellules musculaires par l'introduction du facteur de différenciation myogénique MyoD (voir fig 1.3). Parallèlement, des expériences réalisées chez plusieurs espèces de mammifères ont montré que le transfert de noyaux de cellules différenciées 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 différenciées peut être complètement renversée (Gurdon and Melton, 2008). Enfin, l'avancée la plus récente dans ce domaine a été la démonstration que des cellules somatiques différenciées peuvent être reprogrammées en

cellules souches puripotentes par simple introduction d'un « cocktail » de 4 facteurs de transcription : Oct4, Sox2, c-Myc et Klf4 (Takahashi and Yamanaka, 2006) (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 des gènes est en effet régulée par des protéines, les facteurs de transcription, qui sont eux-mêmes issus de l'expression d'autres gènes, créant ainsi un réseau d'interactions entre gènes. Certaines protéines peuvent par ailleurs directement réguler l'activité d'autres protéines, et certains ARNs issus de la transcription de gènes non codants jouent aussi un rôle fondamental dans la régulation de l'activité génétique, le tout formant un réseau complexe d'interactions à différents niveaux. La compréhension de ce réseau et des fonctions qui en résultent forme le socle de la biologie des systèmes. Dans ce cadre, la cellule est vue comme une unité de traitement d'information qui interprète différents signaux reçus en entrée, les traite par un réseau interne de régulation, et 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 (Nurse and Hayles, 2011).

1.2.2 Divers modes de régulation

Les modes de régulation qui permettent à la cellule d'interpréter des signaux afin de changer d'état sont nombreux. Nous allons nous concentrer ici sur ceux affectant la production d'ARNs ou de protéines (fig. 1.5).

- **Régulation génétique**

Tout d'abord, un réseau d'expression génétique est caractérisé par un jeu d'interactions entre différents gènes. Ici, nous centrons notre attention sur les gènes codant pour des pro-

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 (U.S. Department of Energy, 2001) schématisant un réseau de régulation cellulaire comme un système de traitement entrée/sortie, possédant trois composantes fondamentales : (1) un système de réception et de transduction des signaux d'entrées qui peuvent être intra- ou extra-cellulaires (plusieurs signaux pouvant affecter un même gène cible), (2) un « composant central » (*core component*) composé du réseau de régulation génétique traitant les signaux, et (3) d'un signal de sortie consistant en l'expression moléculaire des ARNs et protéines des gènes cibles. Le processus résulte en la modification du phénotype de la cellule. Des boucles de régulation (*feedback*) assurent le contrôle et la stabilité des différentes étapes.

FIGURE 1.5 – **Un réseau de régulation génétique type.**
Dans cette représentation schématique tirée du rapport du U.S. Department of Energy (2001), deux voies de signalisation A et B transmettent des signaux d'entrée (qui peuvent être intra ou extra cellulaires) en rendant des facteurs de transcription actifs. 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 ($\sim$ 10bp). Les différents facteurs de transcription interagissent entre eux pour former des complexes occupant des régions de $\sim$ 1000bp appelées modules de cis-régulation ou CRMs (voir section 1.5). 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.

téines, mais on pourrait inclure les gènes codant pour des ARNs non traduits, ceux-ci pouvant être impliqués dans la régulation. Dans ce réseau, les interactions se font par l'intermédiaire de protéines régulatrices appelées facteurs de transcription ou TFs, qui sont au nombre de $\sim$ 1400 chez l'homme (Vaquerizas et al., 2009), soit 6% des protéines encodées. 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 $\sim$ 10bp et interagissent avec la machinerie transcriptionnelle au niveau du promoteur du gène cible. Les TFs peuvent se fixer sur le promoteur même, comme c'est souvent le cas chez la bactérie, ou dans des régions distales allant jusqu'à plusieurs centaines de kb, comme on trouve plus couramment chez les organismes complexes. Par ailleurs, différents TFs peuvent se combiner sur certaines régions de régulation contenant de multiples sites de fixation pour former des complexes protéiques. Ces régions, appelées modules de cis-régulation (CRMs) ou plus communément *enhancers*, sont d'une taille typique de $\sim$ 1000bp et ont la particularité de conduire à une expression spatio-temporelle très spécifique du gène cible. Ces différents points seront amplement développés en section 1.5.

- **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 : c'est la 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 (fig. 1.6) [2]. Ainsi, la méthylation des dimères CpG de l'ADN [3] 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 (Bird, 2002). Par ailleurs, la méthylation des histones au niveau des résidus lysines entraîne la fermeture de la chromatine, empêchant l'expression du ou des

2. Il est à noter que certains emploient le terme épigénétique pour qualifier la fixation des TFs sur l'ADN. Ici, le terme épigénétique réfère seulement aux modifications chimiques affectant les histones et l'ADN, et donc l'accessibilité du génome

3. 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

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 (Greer and Shi, 2012). Ce mode de régulation sera développé plus en détail en section 1.5.1.

- **Régulation post-transcriptionnelle**

Les modifications post-transcriptionnelles affectent les ARNs issus de la transcription des gènes. Ces modifications peuvent être causées par des microARNs ou miRNAs qui sont des ARNs de $\sim$ 23 nts issus d'ARNs se repliant en structure double brin de type « épingles à cheveux » ou *hairpins*. Les miRNAs s'associent à la protéine *Argonaute* du complexe RISC (*RNA-induced silencing complex*) pour entraîner la dégradation spécifique d'ARNms (Bartel, 2009). De manière similaire, certains *hairpins* de taille plus importante sont clivés par la protéine Dicer pour former plusieurs petits ARNs de taille similaire aux miRNAs : ce sont les siRNAs (*small interfering RNAs)*). Ceux-ci recrutent aussi le complexe protéique RISC et ciblent spécifiquement des ARNm (Hammond et al., 2001; Hannon, 2002). Ce phénomène est connu sous le nom d'interférence ARN (RNAi) et a donné lieu à une méthode aujourd'hui couramment utilisée pour inhiber l'expression d'un gène.

- **Régulation post-traductionnelle**

Les modifications post-traductionnelles affectent les protéines issues de la traduction des ARNs. Elles 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. 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 (Kondo et al., 2010).

FIGURE 1.6 – **Caractéristiques de l'épigénome.**

Figure tirée de (Schones and Zhao, 2008). Les chromosomes sont partagés entre régions accessibles d'euchromatine et régions difficilement accessibles d'hétérochromatine. Les régions hétérochromatiques sont marquées par la di- et triméthylation de la lysine 9 de l'histone H3 (H3K9me2 et H3K9me3). La méthylation de l'ADN est répandue à travers tout le génome, mais est absente de certaines régions comme 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 transcrites activement à proximité de la région 5′ des gènes (en amont), alors que la marque H3K36 marque les gènes transcrits dans leur région 3′ (en aval).

FIGURE 1.7 – **Exemples de motifs dans les réseaux de régulation génétique.**
Ces exemples sont issus d'analyses d'interactions entre facteurs de transcription (cercles bleus) et promoteurs (rectangles rouges) (A) chez la levure (Lee et al., 2002) et (B) chez l'homme (Odom et al., 2004). Les flèches solides indiquent la fixation d'un facteur de transcription à un promoteur, et les flèches en pointillé indiquent l'expression d'un facteur de transcription à partir de son gène.

1.2.3 Câblage du réseau et fonction

Maintenant que nous avons vu la nature des interactions au sein des réseaux génétiques, nous pouvons nous pencher sur leur structure. Notamment, plusieurs études réalisées chez divers organismes de la bactérie à l'homme ont révélé que les réseaux de transcription contiennent un petit ensemble de motifs de régulation récurrents, appelés motifs de réseaux (Alon, 2007a; Shen-Orr et al., 2002; Milo et al., 2002) (fig. 1.7). De tels motifs furent d'abord détectés de manière systématique chez la bactérie *Escherichia coli* en remarquant qu'ils apparaissaient dans le réseau de transcription bien plus souvent qu'on ne l'attendrait dans un réseau aléatoire (Shen-Orr et al., 2002). Les mêmes motifs ont ensuite été trouvés chez la levure (Milo et al., 2002; Lee et al., 2002) et chez l'homme (Odom et al., 2004). Une explication possible de la récurrence de ces motifs est liée aux fonctions qu'ils remplissent. Par exemple, la boucle d'autorégulation négative, qui est trouvée chez la moitié des répresseurs d'*Escherichia coli*, possède deux fonctions : l'une est de parvenir rapidement à un état d'équilibre en utilisant un promoteur fort, l'autre est de servir de tampon au bruit d'expression (Alon, 2007b). Un autre motif récurrent est la boucle feedforward. Celle-ci consiste en 3 gènes : un régulateur X, qui régule Y, tous deux régulant Z. Dans le cas où des interactions sont des activations et que X et Y sont requis pour activer Z, cette boucle peut servir de tampon au bruit d'expression de X, évitant que des fluctuations de son niveau d'expression n'entraîne par erreur l'activation de Z.

1.2.4 Évolution des réseaux génétiques

L'importance des motifs est rendue plus claire lorsque l'on s'intéresse à l'évolution des réseaux. En effet, au cours de l'évolution, les réseaux de régulation génétique changent : modification des constituants, recâblage du réseau, duplication d'éléments... Néanmoins, certaines modifications sont plus défavorisées du point de vue évolutif que des autres. Ainsi, les motifs tels que les boucles d'autorégulation ou les boucles feedforward, de par leur importance fonctionnelle, auront tendance à être conservés. Pour ce qui est des éléments du réseau, la modification d'un régulateur, par exemple la mutation d'un acide aminé au sein d'un facteur de transcription, aura des conséquences sur l'ensemble des éléments régulés par ce facteur de transcription et pourra donc être fortement délétère. Par contre, la modification d'un site de reconnaissance de ce facteur de transcription sur l'ADN n'aura qu'une portée locale sur la régulation du gène associé.

FIGURE 1.8 – **Évolution du réseau transcriptionnel : l'exemple de la régulation myogénique..** Figure tirée de (Liu et al., 2009). Les lignes pleines indiquent une régulation transcriptionnelle directe, les tirets indiquent des interactions génétiques dont le caractère direct n'a pas été montré, et les lignes pointillés indiquent que l'interaction est suggérée par des donnés d'expression. *Myod1* est l'analogue structurel de *Nautilus* mais est l'analogue fonctionnel de *Twist* chez la Drosophile et de *HLH-1* chez *C. elegans*. Les couleurs indiquent respectivement des voies de signalisation externes au mésoderme (bleu), les régulateurs myogéniques de la famille bHLH (vert), les protéines Six/Eya (orange), les protéines de la famille *paired-box* Pax (violet), les TFs des familles MADS/SRF (rouge), et les régulateurs d'autres familles de protéines (noir).

À titre d'exemple, prenons le cas du réseau de différenciation du muscle squelettique présenté en figure 1.8, que nous étudierons plus en détail dans le chapitre **??** de ce manuscrit. Au coeur de ce réseau génétique se trouvent les facteurs de régulation myogéniques ou MRFs, des facteurs de transcription de type bHLH qui ont la capacité de convertir des cellules non mesodermales, c'est-à-dire n'étant pas destinées à devenir des progéniteurs musculaires, en cellules ayant des propriétés musculaires (Weintraub et al., 1989). Ces facteurs sont dits « régulateurs maîtres » de la différenciation musculaire. Chez les vertébrés il y a quatre MRFs : *Myf5*, *Mrf4*, *Myod1*, qui ont des roles redondant dans la spécification des progéniteurs musculaires, et *Myog*, qui conduit à la différenciation terminale. Chez la Drosophile c'est le TF *Twist* qui semble être le principal MRF, mais contrairement aux MRFs des vertébrés, son rôle ne s'arrête pas au contrôle de la différenciation musculaire mais est plus général dans le développement

du mésoderme (Baylies et al., 1998). C'est cependant le gène *Nautilus* qui possède la séquence d'acides aminés la plus proche de celle des MRFs vertébrés. Ce dernier permet la spécification des progéniteurs myogéniques, et son expression est restreinte au développement musculaire. Néanmoins, les mutants *nautilus* sont viables et son rôle semble mineur comparé aux MRFs vertébrés. Enfin, chez le ver *Caenorhabditis elegans*, c'est l'orthologue de *Myod1*, *hlh-1*, qui tient rôle de MRF.

Malgré ces différences (nombre de MRFs, membre de la famille bHLH tenant ce rôle), on retrouve dans les trois cas une boucle feedforward conservée au niveau de la régulation des cibles des MRFs (fig. 1.8). Ainsi, MyoD régule l'expression de Mef2 et l'activité de MAPK p38 en même temps que l'expression de plusieurs cibles initiales, et par la suite MyoD et phospho-Mef2 co-régulent des gènes plus tardifs. Ce mécanisme permet ainsi de réguler l'aspect temporel de l'expression génétique. Chez la Drosophile, le même motif est observé avec Twist et Mef2 et chez *C. elegans* avec HLH-1 et le TF UNC-129, de la même famille que Mef2.

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

Nous voyons donc que l'évolution d'un réseau génétique possède de multiples facettes : conservation de motifs de réseau fonctionnellement importants (dans notre exemple, la boucle feedforward au coeur du réseau régissant l'aspect temporel de l'expression des cibles), recâblage des interactions pour traiter différents signaux d'entrée... Par ailleurs, il apparaît que

plus qu'à des TFs particuliers, c'est à des familles de TFs que nous avons affaire. Aussi un même rôle au sein du réseau peut-il être rempli par différents membres d'une même famille, comme c'est le cas pour *Myod1* et *Twist*. Ceci s'explique par le fait que les membres d'une même famille partagent des propriétés d'interaction avec l'ADN semblables. Ces interactions sont à la source du fonctionnement du réseau, et nous allons maintenant présenter plus en avant leurs propriétés.

1.3 Les interactions protéine-ADN : modèles mathématiques

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

FIGURE 1.9 – **Différents états du facteur de transcription.** Figure tirée de (Lässig, 2007). Lors de sa recherche de site de fixation, le TF peut se trouver dans trois états distincts : (1) un état libre de diffusion tridimensionnelle, (2) un état de diffusion unidimensionnelle sur l'ADN par fixation non spécifique, et (3) un état de fixation spécifique. L'énergie de fixation dépend du site de fixation, de taille l et de coordonnée r.

1.3.1 Modes de recherche du site de fixation par le TF

Un facteur de transcription peut être dans plusieurs états : en diffusion tridimensionnelle, auquel cas il est dit « libre », ou bien fixé sur l'ADN. Dans ce dernier cas, il interagit avec l'ADN selon deux modes : une attraction non spécifique d'énergie E_{ns} indépendante de la position sur l'ADN, et une interaction spécifique $E_s(r)$ qui dépend de la séquence de taille $l \sim 10$ à la position r sur l'ADN. L'interaction non spécifique est due à l'interaction électrostatique entre la protéine chargée positivement et l'ADN chargé négativement, alors que l'interaction spécifique implique des liaisons hydrogènes entre le domaine de fixation de la protéine et les nucléotides du site de fixation. Le facteur de transcription peut ainsi se trouver dans trois états thermodynamiques représentés en figure 1.9 : en diffusion tridimensionnelle libre, fixé non spécifiquement (diffusion unidimensionnelle le long de la structure d'ADN), et fixé spécifiquement sur l'ADN. Ces trois modes contribuent à la cinétique de la recherche d'un site fonctionnel (Berg et al., 1981; Winter and von Hippel, 1981; Winter et al., 1981). Ainsi, l'attraction non spécifique conduit la protéine à passer à peu près autant de temps fixé sur l'ADN qu'en diffusion libre. La recherche de site de reconnaissance est donc un processus mixte de diffusion unidimensionnelle sur l'ADN et de diffusion tridimensionnelle dans le milieu. Lorsqu'il est fixé sur l'ADN, le facteur diffuse dans un paysage d'énergie E_{ns} plat lorsqu'il est dans sa conformation de fixation non spécifique, ou dans un paysage d'énergie $E_s(r)$ dans sa conformation de fixation spécifique. Cela permet au facteur d'échantillonner les sites de faible énergie $E_s(r)$ tout en évitant d'être bloqué par les barrières de haute énergie en passant en mode de recherche non spécifique. Ce processus s'avère au final très efficace : les temps de recherche sont typiquement inférieurs à une minute, ce qui est petit devant les processus de régulation de la cellule qui se déroulent au mieux sur quelques minutes (Gerland et al., 2002; Slutsky and Mirny, 2004). Il est donc pertinent de décrire l'effet d'un site de fixation sur la régulation d'un gène cible par la probabilité qu'il a de fixer un facteur de transcription à l'équilibre thermodynamique.

1.3.2 Modèle PWM

Présenté en 1987 par Berg et von Hippel (Berg and von Hippel, 1987), le modèle PWM est le modèle le plus simple décrivant l'énergie de fixation spécifique entre un facteur de transcription et un site de fixation sur l'ADN. Ce modèle repose sur plusieurs hypothèses.

Tout d'abord, il y a l'hypothèse importante que les sites de fixation des TFs sur l'ADN ont été sélectionnés au cours de l'évolution pour leur propriété de sites de reconnaissance, quelle que soit la concentration du TF dans la cellule. En d'autres termes, le processus de sélection discrimine les sites de fixation sur la seule base de leur énergie de fixation à un TF donné : les sites ayant une énergie de fixation dans une certaine gamme sont retenus, les autres rejetés. Par ailleurs, au sein de cette gamme d'énérgie « utile », toutes les séquences sont équiprobables. Enfin, la dernière hypothèse est que chaque nucléotide d'un site de fixation contribue de manière indépendante, c'est-à-dire additive à l'énergie totale du site. Cette hypothèse permet de simplifier le problème en gardant le nombre de paramètres petit.

L'argument de Berg et von Hippel est que ce problème est analogue à celui de physique statistique consistant à déduire les taux d'occupation des niveaux d'énergie de particules indépendantes sachant que l'énergie totale doit avoir une certaine valeur moyenne E. La solution de ce problème est donnée par la formule de Boltzmann reliant énergie et taux d'occupation :

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

où $f_{i,b}$ est la probabilité d'observer la base b à la position i du site de fixation, $E_{i,b}$ est l'énergie associée (en $k_B T$), $\mathcal{Z}_i$ est la fonction partition qui permet de normaliser la distribution à la position i, et λ est un facteur sans dimension, analogue du β de la thermodynamique, et lié au processus de sélection. Dans la suite, nous intégrerons ce facteur à l'énergie.
La connaissance des fréquences des bases permet de définir une autre quantité utile caractérisant la variabilité des séquences de fixation, l'information relative des sites par rapport à une séquence d'ADN aléatoire (Stormo and Fields, 1998) :

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

où L est la taille du site de fixation et π_b correspond à la probabilité *a priori* d'observer la base b dans le génome. Il est usuel de définir l'énergie relativement au fond génomique :

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

L'énergie totale d'un site S_i est alors

a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

b

1	2	3	4	5	6	7	8	9	10	11	12	13	14
B	R	M	C	W	A	W	H	R	W	G	G	B	M

Consensus sequence

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

d Position weight matrix (PWM)

A	−1.93	0.79	0.79	−1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	−1.93	−1.93	0.79
C	0.45	−1.93	0.79	1.68	−1.93	−1.93	−1.93	0.45	−1.93	−1.93	−1.93	−1.93	0.00	0.79
G	0.00	0.45	−1.93	−1.93	−1.93	−1.93	−1.93	−1.93	0.66	−1.93	1.30	1.68	1.07	−1.93
T	0.15	0.66	−1.93	−1.93	1.07	0.66	0.79	0.00	0.00	0.79	−1.93	−1.93	−0.66	−1.93

e Site scoring

0.45	−0.66	0.79	1.68	0.45	−0.66	0.79	0.45	−0.66	0.79	0.00	1.68	−0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

$\Sigma = 5.23$, 78% of maximum

f

FIGURE 1.10 – **Construction et utilisation du modèle PWM.** Figure tirée de (Wasserman and Sandelin, 2004). (a) Supposons connus un certain nombre de sites de fixation d'un facteur de transcription (dans ce cas MEF2). (b) Séquence consensus correspondante utilisant les symboles IUPAC. (c) Une matrice de fréquence est construite, indiquant pour chaque nucléotide sa multiplicité à une position donnée dans le site. (d) La PWM est simplement construite en prenant le logarithme relatif des fréquences PWMs par rapport aux fréquences *a priori* des nucléotides. (e) Le score (ou énergie) d'une séquence d'ADN donnée est calculé en additionnant les poids PWM correspondants. (f) La PWM peut être représentée sous forme de logo (Giocomo et al., 2011). 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.

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

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

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

1.3.3 Modèle biophysique

Le modèle PWM est basé sur une hypothèse forte, celle que les sites de fixation ont été sélectionnés sur la base de leur seule affinité ou énergie envers un TF. Néanmoins, à aucun moment n'intervient la concentration du TF dans la cellule, dont dépend pourtant la probabilité de fixation. C'est ce que tente de capturer le modèle biophysique (Gerland et al., 2002; Djordjevic et al., 2003; Zhao et al., 2009).

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

$$TF + S_i \rightleftharpoons TF : S_i \tag{1.5}$$

où $TF : S_i$ dénote le complexe entre le TF et le site S_i. La constante d'équilibre de cette réaction s'écrit selon la loi d'action de masse :

$$K_i = \frac{[TF : S_i]}{[TF][S_i]} \tag{1.6}$$

Le site peut être dans deux états : occupé par le TF où libre. Aussi, la probabilité que le TF soit fixé au site s'écrit simplement

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

où $E_i = -kT \ln(K_i)$ est l'énergie libre standard de fixation (souvent notée ΔG), $\mu = kT \ln[TF]$ est le potentiel chimique, k est la constante de Boltzmann, T la température et $\beta = 1/kT$. Ici nous avons considéré qu'il n'y avait qu'un seul site de fixation. De manière générale, le site est en compétition avec le fond génomique, ce qui ajoute une contribution à μ (voir section 1.3.4). À l'instar du modèle PWM, l'énergie E_i est généralement prise comme étant une fonction additive des énergies individuelles des différentes bases du site. Ainsi, lorsque le TF est à faible concentration ($\mu \to -\infty$), le modèle biophysique écrit en équation 1.7 se réduit au modèle PWM.

1.3.4 Modèle thermodynamique

La description biophysique peut être réécrite en termes thermodynamiques en utilisant des raisonnements simples sur le nombre d'états possibles et leur énergie (et donc poids de Boltzmann) associée. Nous adoptons ici l'approche de (Gerland et al., 2002). On pourra par ailleurs se référer à l'excellente revue (Lässig, 2007). Considérons le cas simple d'un seul facteur de transcription interagissant avec un génome de taille $L \gg 1$ ne contenant qu'un seul site fonctionnel, le reste de la séquence étant aléatoire. Nous l'avons vu, l'expérience montre que la protéine se fixe à l'ADN avec une probabilité $1/2$. Lorsqu'elle est fixée, elle est à l'équilibre entre le mode spécifique et le mode non spécifique. Nous désirons savoir avec quelle probabilité elle est fixée de manière spécifique. La fonction de partition, énumérant tous les poids de Boltzmann associés aux différents états accessibles au TF fixé s'écrit :

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

Notons i la position du site fonctionnel. On peut écrire :

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

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

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

Dans le cas d'un site de reconnaissance, $E_{ns} \gg E_s(i)$ de sorte que $E_i \simeq E_s(i)$ (Gerland et al., 2002). La probabilité que le facteur soit fixé sur le site fonctionnel s'écrit finalement :

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

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

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

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

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

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

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

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

La valeur de μ est liée à la fois au nombre de TFs ainsi qu'à la possibilité de se fixer dans le fond génomique. Elle est fixée implicitement par l'équation :

$$n = \sum_{r=1}^{L} \frac{1}{1 + e^{\beta(E_r - \mu)}} \tag{1.15}$$

qui signifie simplement que le nombre de TFs n dans le système est égal à la somme sur tous les sites de fixation possibles pondérée par la probabilité que le TF y soit fixé. Lorsque $\mu \to -\infty$ et que la fonction de Fermi peut être approximée par la loi de Boltzmann, l'équation peut s'inverser et l'on trouve (Aurell et al., 2007)

$$\mu = F_0 + kT \log n \tag{1.16}$$

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

1.4 Les interactions protéine-ADN : mesures expérimentales

Ces dernières années, des avancées technologiques considérables ont permis d'une part d'établir des modèles de fixation spécifique pour de nombreux TFs, d'autre part de localiser leurs sites de fixation dans le génome. Ces avancées ont eu lieu autant sur le plan *in vitro*, utilisant protéines purifiées et séquences nucléiques artificielles pour déduire l'affinité protéine-ADN, que sur le plan *in vivo*, mesurant l'interaction de la protéine avec l'ADN génomique (Stormo and Zhao, 2010).

1.4.1 Approches *in vitro* : MITOMI, SPR, PBM, CSI, SELEX, et HT-SELEX

- **Approche microfluidique : MITOMI**

En 2007, Maerkl et Quake ont mis au point une technique appelée MITOMI (Mechanically Induced Trapping Of Molecular Interactions) permettant une mesure directe de l'affinité d'un TF à des centaines de séquences d'ADN à la fois (Maerkl and Quake, 2007). Cette technique repose sur l'utilisation d'un système microfluidique composé de chambres dans lesquelles un fluide dont on peut facilement modifier la composition circule dans des canaux d'un diamètre de l'ordre de 1μm dont le microenvironnement est finement contrôlé. Le fluide contient des gènes synthétiques codant pour le TF ainsi que du matériel permettant la synthèse de la protéine directementau sein de la chambre, ce qui évite de purifier préalablement le TF. Chaque chambre du système contient des anticorps fixés à la surface permettant de capturer le TF et une certaine concentration d'une séquence d'ADN spécifique contenant une marque fluorescente. Le système contient ainsi des centaines de séquences d'ADN différentes, chacune étant présente à différentes concentrations. Lorsque le TF est fixé par les anticorps, il recrute des séquences d'ADN selon leur affinité. Celles qui ne se fixent pas sont lavées. Au final, les séquences fixées produisent un signal de fluorescence. La comparaison des signaux pour différentes concentrations d'ADN donne accès au rapport des constantes d'équilibre K_{eq} (eq. 1.6). La comparaison avec une séquence référence dont la constante K_{eq} est connue permet alors de déterminer le K_{eq} absolu pour chaque séquence de fixation.

En utilisant 17 systèmes de ce type, ils ont ainsi pu mesurer l'affinité de 4 TFs de type bHLH à 464 séquences d'ADN différentes : les séquences consensus et des séquences ayant une, deux, trois ou quatre mutations. À titre de comparaison, ils ont construit une PWM à partir des séquences contenant une seule mutation, puis ont prédit les énergies attendues des séquences à plusieurs mutations. La prédiction de la PWM s'est avérée bonne dans seulement 56% des cas pour les séquences à deux mutations, 10% pour les séquences à 3 mutations et 0% des cas pour les séquences à 4 mutations, montrant les limites de ce modèle indépendant confronté à des données d'interactions d'ordre supérieur. Un modèle plus raffiné prenant en compte l'énergie d'interaction non spécifique et incluant des interactions entre nucléotides voisins permet néanmoins de rendre compte des valeurs observées (Stormo and Zhao, 2007). Nous reviendrons sur la nécessité de prendre en compte les interactions entre paires de nu-

cléotides lors de l'interaction spécifique entre TF et ADN dans le chapitre 2.

- **Approche physique : la microscopie SPR**

La méthode de résonance des plasmons de surface (*Surface Plasmon Resonance* ou SPR) est habituellement utilisée pour étudier l'interaction d'une protéine avec un ligand (qui peut être une autre protéine), mais elle peut aussi être utilisée pour mesurer les interactions entre une protéine et quelques centaines de séquences d'ADN différentes (Shumaker-Parry et al., 2004; Campbell and Kim, 2007). Le principe de la microscopie SPR est que l'angle de réflection de la lumière sur une fine surface d'or, par exemple, dépend de la masse de molécules fixées de l'autre côté de sa surface. Si de l'ADN est lié à la surface, la fixation du TF induit un changement de masse et donc d'angle de reflection lumineuse mesurable au cours du temps. Ainsi, la cinétique de fixation du TF jusqu'à l'atteinte de l'équilibre est accessible. On peut de même étudier la dissociation du TF lors du lavage de la surface. Ces mesures donnent directement accès aux taux d'association k_{on} et de dissociation k_{off} que la simple mesure de la constante d'équilibre $K_{eq} = k_{on}/k_{off}$ ne permet habituellement pas de déterminer.

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

L'analyse de fixation des protéines par puce à ADN (*Protein-Binding Microarray* ou PBM) est une technologie haut débit qui a été développée au cours des 10 dernières années (Berger et al., 2006). Les puces sont composées de $44,000$ puits auxquels sont liés des brins d'ADN. Une puce contient tous les sites de fixation de 8bp possibles ($4^8/2 = 32,768$ séquences en prenant en compte le fait qu'il y a un site sur chacun des deux brins d'ADN) plus deux bases flanquantes (une à chaque extrémité) qu'il est possible de faire varier. Un TF purifié à partir de cellules ou synthétisé *in vitro* est ajouté à la puce, qui est ensuite lavée pour se débarrasser des fixations non spécifiques. La quantité de protéine fixée à un puits donné est déterminée grâce à un anticorps fluorescent contre la protéine. L'enrichissement en protéine est calculé relativement au bruit de fond (anticorps non spécifique par exemple). Il est alors possible d'utiliser ces mesures pour bâtir une PWM du TF (voir par exemple Kinney et al. (2007)).

Une autre méthode utilise aussi des puces à ADN : c'est l'identification de site apparenté (*Cognate Site Identifier* ou CSI) (Warren et al., 2006). Une différence technique avec les PBMs est que l'ADN est d'abord synthétisé en simple brin puis se replie en double brin pour former le site de fixation, évitant ainsi de devoir générer l'ADN double brin à partir de précurseurs.

Par ailleurs, le TF est en compétition avec un marqueur fluorescent qui peut se fixer à l'ADN : il n'est donc pas nécessaire d'utiliser un marquage spécifique sur le TF ou sur un anticorps, ce qui rend la procédure plus généralisable. Finalement, la spécificité du TF est représentée par un « paysage de spécificité » qui encapsule l'information de fluorescence de l'ensemble des variations par rapport à une séquence consensus dans une représentation simple (Carlson et al., 2010).

- **Approche par purification des séquences fixées : SELEX et HT-SELEX**

Mise au point il y a plus de 20 ans, la méthode SELEX (*Systematic Evolution of Ligands by EXponential enrichment*) repose sur la sélection de séquences d'ADN aléatoires par un TF *in vitro* (Oliphant et al., 1989; Tuerk and Gold, 1990; Blackwell and Weintraub, 1990; Wright et al., 1991). Une bibliothèque de sites de fixation potentiels est d'abord générée en synthétisant des séquences d'ADN aléatoires ou en utilisant des séquences génomiques. Les extrémités de ces séquences contiennent des précurseurs permettant l'amplification exponentielle par PCR. Le TF purifié est ajouté aux sites et les séquences fixées sont séparées des séquences non fixées, par exemple par retard sur gel. Après un cycle de sélection, les séquences récupérées sont encore enrichies en séquences de basse affinité pour le TF, car celles-ci sont simplement initialement bien plus abondantes que les séquences de haute affinité. Afin d'augmenter la proportion de séquence de grande affinité, les séquences filtrées sont amplifiées puis filtrées à nouveau, ceci sur plusieurs cycles. À la fin de ce processus, les séquences sélectionnées sont clonées et séquencées, résultant en un nombre typique de moins de $\sim$ 100 séquences indépendantes (Fields et al., 1997). Si les séquences initiales sont issues d'ADN génomique, il est possible d'utiliser l'hybridation des séquences à des puces à ADN. La présence de plusieurs cycles de sélection rend néanmoins la détermination des énergies de fixation moins directe qu'avec les techniques précédentes. Une variante de la technique appelée SELEX-SAGE utilise des multimères de sites à la place de sites uniques et permet de réduire le nombre de cycles de sélection et d'augmenter ainsi le nombre de séquences de fixation obtenues (Roulet et al., 2002), permettant de réaliser des modèles plus précis (Nagaraj et al., 2008).

Depuis la mise au point de la méthode SELEX, des avancées considérables ont été réalisées dans les techniques de séquençage, permettant l'obtention de millions de séquences à la fois : on parle de séquence haut-débit (*high-throughput*) ou encore séquençage massivement parallèle. L'utilisation de ces nouvelles techniques dans l'expérience SELEX a mené à la méthode

HT-SELEX (Nagaraj et al., 2008), aussi appelée Bind-n-Seq (Zykovich et al., 2009). Il est alors possible d'estimer un modèle d'énergie à partir des fréquences d'observation des différentes séquences dès le premier cycle (Nagaraj et al., 2008). Des cycles supplémentaires permettent d'obtenir plus d'information sur les séquences les plus spécifiques, notamment sur la présence de contributions non indépendantes à l'énergie, ou de compenser la faible spécificité d'un TF. L'avantage de cette technique est que la taille des sites de fixation n'est pas limitée. Ainsi, avec une nanomole d'ADN ($\sim 10^{15}$ séquences) on peut couvrir l'ensemble des sites de 25bp possibles. Le séquençage haut-débit permet d'en échantillonner $\sim 10^8$, ce qui est largement suffisant pour contraindre des modèles d'énergie indépendants, même pour des TFs ayant des sites de fixations de taille > 15bp comme c'est souvent le cas chez la bactérie. Cette technique a récemment été poussée encore plus loin (Jolma et al., 2010). En utilisant des protéines marquées, les auteurs ont réalisé un HT-SELEX à partir d'extraits cellulaires, et en ajoutant un code barre aux séquences d'ADN de chaque expérience, ils ont pu analyser les sites de fixation pour plusieurs TFs en parallèle. Ils ont ensuite utilisé cette technique pour obtenir des modèles de spécificité pour 411 TFs humains, la plus grande étude de ce genre réalisée à ce jour (Jolma et al., 2013).

1.4.2 Approche clonale : la technique de simple hybride

Contrairement aux approches précédentes, la technique de simple hybride (*Bacterial one-hybrid* ou B1H) n'est pas purement *in vitro*, au sens où l'interaction protéine-ADN est testée au sein d'une bactérie. Néanmoins, parce que l'interaction n'est pas testée dans son contexte cellulaire d'origine, nous la considérerons comme telle. Cette approche repose sur l'intégration par une bactérie hôte de deux vecteurs d'expression génétique, ou plasmides. Le premier exprime le facteur de transcription d'intérêt fusionné à une sous-unité de l'ARN polymérase (l'appât), c'est la protéine « hybride ». L'autre contient une région de séquence aléatoire représentant un site de fixation potentiel (la proie) en amont d'un promoteur à faible activité. La fixation de cette région par la protéine hybride permet l'activation d'un gène de sélection, généralement *HIS3*, un gène de la levure requis pour la biosynthèse de l'histidine et dont l'homologue bactérien est absent de la souche d'*Escherichia coli* utilisée. La croissance des cellules a lieu dans un milieu ne contenant pas l'histidine. Dans ces conditions, les bactéries n'exprimant pas *HIS3* ne peuvent croître. Ainsi, seules les bactéries au sein desquelles le facteur de transcription se fixe à la proie expriment *HIS3*, croissent et forment des colonies, d'où

la notion de gène de sélection. Par ailleurs, la stringence de la sélection peut être modulée en ajoutant au milieu différentes concentrations de 3-amino-triazole (3-AT), un inhibiteur de *HIS3*. De cette façon l'affinité du site de fixation peut être estimée plus finement.

Dans les études de ce type, les sites de fixation présents au sein des colonies sont séquencés individuellement, ce qui permet d'obtenir environ 50 séquences pour une expérience de sélection donnée. Néanmoins, il semble possible d'utiliser les nouvelles technologies de séquençage pour récupérer l'ensemble des sites de fixation des bactéries présentes sur une plaque (Stormo and Zhao, 2010). À l'instar de la méthode HT-SELEX, on obtient des millions de sites, ceux ayant une plus grande affinité étant présents à plusieurs centaines de milliers d'exemplaires, et ceux ayant une faible affinité étant présent en un seul voire aucun exemplaire.

Notons qu'il est aussi possible d'adopter la démarche inverse, c'est-à-dire de partir de quelques sites de fixation présumés fonctionnels mais pour lesquels on ne connaît pas le TF associé. En utilisant une bibliothèque de plasmides codant pour différents TFs hybrides, il est alors possible de déterminer si l'un d'entre eux possède une affinité importante avec les sites testés.

1.4.3 Approches *in vivo* : ChIP-on-chip, ChIP-seq, DNAse I

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

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

La technique d'immunoprécipitation de la chromatine (ChIP) (fig. 1.11) consiste dans un premier temps à induire la réticulation (*crosslink*) des protéines se liant à l'ADN en traitant les cellules avec de la formaldéhyde. Cette étape permet de transformer les liaisons faibles

FIGURE 1.11 – **Étapes d'une expérience de ChIP-on-chip et ChIP-seq.**
Figure tirée de Visel et al. (2009b). À partir d'extraits cellulaires issus de cultures *in vitro* ou prélevés *in vivo*, plusieurs étapes permettent de récupérer les régions fixées par un TF d'intérêt. D'abord, les protéines, liées de manière non covalente à l'ADN, sont fixées par utilisation d'un agent de réticulation, le formaldéhyde. Puis la chromatine est découpée en fragments de $\sim$ 500bp par utilisation d'ultrasons : on parle de sonication. L'utilisation d'un anticorps spécifique au TF d'intérêt permet de précipiter les fragments d'ADN fixés. Les fragments résultants sont alors soit hybridés à une puce à ADN contenant de nombreuses séquences génomiques (ChIP-on-chip), soit directement séquencés et alignés à un génome de référence (ChIP-seq). La comparaison à un échantillon d'ADN non précipité (« input ») permet alors de définir les régions significativement fixées, avec une résolution plus précise dans le cas du ChIP-seq, où l'ensemble du génome est couvert, par rapport au ChIP-on-chip, pour lequel seulement un sous ensemble du génome est analysé.

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

Dans le cas du ChIP-on-chip, l'échantillon immunoprécipité et l'ADN de départ (*input*) sont marqués avec des colorants fluorescents et hybdridés sur une puce à ADN composée de très nombreux puits contenant des oligonucléotides (courtes séquences d'ADN) correspondant à différentes régions du génome. Dans le meilleur cas, ces oligonucléotides couvrent l'ensemble du génome. Les sites de liaison sont identifiés par l'écart d'intensité entre les signaux de fluorescence des conditions d'immunoprécipitation et d'*input*.

Dans le cas du ChIP-seq, l'échantillon immunoprécipité est analysé par séquençage à haut débit, résultant en une librairie de *reads* d'une longueur typique variant entre 27 et 50bp issus des extrémités des séquences. Ces *reads* sont ensuite alignés sur un génome de référence. À chaque position du génome correspond ainsi un certain nombre de séquences précipitées et d'*input*. En comparant ce nombre au nombre moyen dans le locus et à l'*input*, il est possible d'identifier des pics correspondant à la fixation du facteur (voir par exemple le programme d'appel de pics ChIP-seq MACS (Zhang et al., 2008)).

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

FIGURE 1.12 – **Résolution des expériences ChIP-on-chip et ChIP-seq.** Figure tirée de (Park, 2009), montrant les profils de fixation de la protéine Chromator générés à partir d'expériences de ChIP-on-chip (intensité relative par rapport au contrôle, bleu) et de ChIP-seq (densité de séquences, rouge) dans la lignée cellulaire S2 de *Drosophila melanogaster*. On peut noter la plus grande résolution de l'expérience ChIP-seq pour déterminer les sites de liaison. L'ADN utilisé en *input* de l'expérience de ChIP-seq et servant de contrôle est montré en gris, et les gènes du locus indiqués en noir.

faudrait que tous les oligonucléotides se superposent à une base près, ce qui demanderait un trop grand nombre de puces.

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

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

FIGURE 1.13 – **Expérience d'empreinte à la DNase I chez la levure : vers une résolution au nucléotide près.** Figure tirée de (Hesselberth et al., 2009). (a) Densité de digestion de la DNase I par nucléotide dans une région de 100kb du génome de la levure contenant $\sim$ 50 gènes (boîtes bleues). On voit clairement certaines régions promotrices marquées par la DNase I. (b) Zoom sur deux régions de 2.5kb. (c) Zoom sur des régions de 250bp. Le nombre d'événements de digestion par nucléotide est marqué par des empilements de boîtes noires, révélant des régions protégées : ce sont les empreintes à la DNase I. (d) Les empreintes sont associées à des motifs de régulation de la levure connus. Les pointillés noirs indiquent le nombre moyen de digestion par nucléotide dans le génome ($\sim$ 2 digestions par base).

ouverte ayant des roles de régulation génétique : promoteurs, enhancers...

En combinant la technique de DHS avec le séquençage à haut débit, l'expérience de DNase-seq permet d'identifier tous les types de région de régulation à l'échelle du génome (Thurman et al., 2012). Les régions riches en sites de digestion identifient alors les sites DHS. Par ailleurs, au sein d'un site DHS, il y a de petites régions ($\sim$ 15bp) qui sont protégées de la digestion par DNase I : ce sont les empreintes à la DNase I ou *DNase I footprints* (fig. 1.13). Ces empreintes sont dues à la présence de protéines ou de complexes fixés à l'ADN. Cette technique de détection de sites de liaison par empreinte à la DNase I existe depuis 30 ans mais n'a que récemment été porté à l'échelle génomique. En comparant à des données ChIP-seq ou en utilisant des bases de données de motifs de facteurs de transcription, il est possible d'identifier le facteur correspondant dont les sites de fixation sont alors connus au nucléotide près.

1.5 Les modules de cis-régulation (CRMs)

Nous l'avons vu en section 1.2.2, les séquences d'ADN régulant l'expression génétique – CRMs pour *Cis-Regulatory Modules* – jouent un rôle prépondérant au cours du développement des organismes. Ces CRMs assurent en effet l'orchestration de l'expression de gènes spécifiques aux différentes étapes du développement et aux divers types cellulaires. Ils sont au coeur de l'évolution des réseaux génétiques, car ils dictent les interactions entre gènes. De plus, leur altération peut conduire à de nombreuses pathologies, liées pour la plupart à une expression génétique aberrante. Notamment, la majeure partie des variants génétiques qui sont associés de manière significative à une susceptibilité envers une maladie sont situés hors des régions codant pour des protéines, suggérant qu'un certain nombre affectent non pas la forme de la protéine engendrée mais l'expression du gène la produisant en détruisant une activité CRM. Dans cette partie, nous présentons les différents types de CRMs, leur structure, et leur évolution.

1.5.1 Les différents types de CRMs

Selon leur rôle dans la régulation de l'expression génétique, les CRMs peuvent être distingués en trois catégories.

FIGURE 1.14 – **Les différents types de CRMs et leurs marques épigénétiques.**
Figure tirée de (Hardison and Taylor, 2012). La notion de CRM renvoie à un regroupement de sites de liaison pour un ou plusieurs facteurs de transcription. Les CRMs peuvent être regroupés en plusieurs classes : les promoteurs, les *enhancers/silencers*, et les insulateurs. Les CRMs des différentes classes partagent les marques d'acétylation H3Ac et H4Ac, les promoteurs actifs sont spécifiquement marqués par H3K4me3, et les enhancers et insulateurs par H3K4me1. Les enhancers sont par ailleurs souvent fixés par le co-activateur p300. Enfin, chez les mammifères les insulateurs recrutent CTCF pour bloquer l'activation par les enhancers.

- **Promoteurs**

Les promoteurs permettent la fixation de l'ARN polymérase pour débuter la formation d'un transcrit ARN au site d'initiation de transcription (*Transcription Start Site* ou TSS). Dans les promoteurs fixant l'ARN polymérase II (la majorité des promoteurs eucaryotes), des facteurs de transcription généraux se fixent à un coeur de $\sim$ 100bp autour du TSS afin de faciliter la fixation du complexe de polymérase. Ces coeurs de promoteurs contiennent pour certains des motifs stéréotypés, comme la boîte TATA, et ont un TSS bien déterminé ; néanmoins la plupart des promoteurs des génomes mammifères sont des régions riches en GC et en dinuclétoides CpG (les « îlots CpG ») qui ne possèdent pas de boîte TATA et permettent l'initiation de la transcription dans un interval d'environ 100 bases (Carninci et al., 2006). Au niveau épigénétique, les promoteurs actifs sont caractérisés par une région pauvre en nucléosomes en amont du TSS, flanquée de nucléosomes possédant la marque de méthylation H3K4me3.

- ***Enhancers* et *silencers***

Les *enhancers* et *silencers* sont respectivement définis par leur effet positif ou négatif sur l'expression d'un gène cible. Cet effet peut notamment être observé par transfert d'un plasmide contenant l'élément régulateur en amont d'un gène rapporteur dans un animal transgénique ou dans des cultures cellulaires transfectées (voir 1.6.4). Leur activité ne dépend généralement pas de leur position et de leur orientation sur le plasmide. Selon l'environnement cellulaire, une région régulatrice peut être soit *enhancer* soit *silencer*, en fonction de la nature de co-activateurs ou de co-répresseurs des TFs recrutés. Il y a néanmoins relativement peu de *silencers* caractérisés et l'on utilise le terme d'*enhancers* pour désigner de manière générale ces régions régulatrices.

Les *enhancers* peuvent se situer à des distances variables du gène qu'ils régulent (Maniatis et al., 1987), pouvant parfois aller jusqu'à 1 Mb comme dans le cas de *Shh* chez la souris (Lettice et al., 2003) (voir fig. 1.24). Les enhancers contiennent de multiples sites de fixations de TFs. Cette multiplicité est requise pour l'activité enhancer, comme cela l'a été montré pour le premier enhancer découvert : celui du virus simien 40 (SV40) (Schirm et al., 1987; Ondek et al., 1988). Un gène peut par ailleurs posséder plusieurs enhancers distincts conduisant à des expressions spécifiques dans différents tissus, comme cela l'a été montré dans le cas du gène *eve* chez la *Drosophile* (Wilson and Odom, 2009) ou dans le cas du cluster de gènes de détermination myogénique *Myf5* et *Mrf4* chez les mammifères (Carvajal et al., 2008) (fig. 1.15). Ainsi,

FIGURE 1.15 – **Différents *enhancers* conduisent à différents patterns d'expression.**
(A) Figure tirée de (Wilczynski and Furlong, 2010), montrant le réseau de cis-régulation du gène *even-skipped* chez la *Drosophile*. Des enhancers individuels (boîtes colorées) conduisent à des motifs d'expression distincts (indiqués dans les embryons par la même couleur). Les cercles colorés au sein des régulateurs indiquent les couleurs des CRMs régulés.
(B) Figure tirée de (Carvajal et al., 2008), représentant les différents éléments régulateur du locus *Mrf4/Myf5*. Chaque boîte colorée représente une régulation de la transcription à un stade de développement particulier ou dans une région anatomique spécifique du muscle embryonnaire ou adulte.

les différents enhancers d'un même gène peuvent être vus comme autant de points d'entrée d'un réseau de régulation génétique, représentant diverses fonctions logiques et intégrant différentes information spatio-temporelles pour produire en sortie une expression génétique spatio-temporelle finement contrôlée (Bolouri and Davidson, 2002; Buchler et al., 2003).

Enfin, comme décrit en fig. 1.14, les enhancers sont associés à de hauts niveaux de marque épigénétique H3K4me1 (Heintzman et al., 2009) et sont souvent fixés par le co-activateur p300 (Wang et al., 2005; Heintzman et al., 2009).

- **Insulateurs**

Les insulateurs sont des CRMs qui restreignent l'effet des enhancers sur leur gène cible (Wallace and Felsenfeld, 2007). Ainsi, certains insulateurs possèdent une activité de blocage d'enhancers. Situés entre un enhancer et un promoteur cible, ces insulateurs bloquent l'activité de l'enhancer, conduisant à une réduction de l'expression du gène cible (Chung et al., 1993). Chez les mammifères, la fixation de la protéine CTCF est nécessaire à cette activité de blocage de l'activité enhancer (Bell et al., 1999), alors que chez la *Drosophile* et plusieurs autres insectes il existe au moins quatre protéines additionnelles qui sont suffisantes à la réalisation de cette activité (Schoborg and Labrador, 2010). Par ailleurs, les insulateurs peuvent servir de barrière de protection contre des marques d'hétérochromatine répressives. De tels insulateurs permettent notamment d'éviter les effets de positions – la modification de l'expression d'un gène selon sa position dans le chromosome – lorsqu'ils entourent un gène rapporteur intégré au hasard dans le génome (Recillas-Targa et al., 2002). Cette activité passe notamment par le recrutement de *USF*, protéine qui recrute des enzymes de modification de la chromatine. Un insulateur peut combiner les activités de barrière de protection et de blocage d'enhancer.

De même que les enhancers, les insulateurs peuvent se situer à des distances variables des gènes qu'ils régulent. Il est à noter que la protéine CTCF possède d'autres fonctions que celle d'isolation, et tous les sites de CTCF ne correspondent pas forcément à des insulateurs (Phillips and Corces, 2009).

1.5.2 Grammaire des enhancers : enhanceosome vs billboard

Nous l'avons vu, les CRMs contiennent en général de multiples sites de liaisons (TFBS) pour un ou plusieurs TFs. On parle de *clustering* (regroupement). Lorsque les TFBS correspondent à plusieurs TFs différents, on parle de CRM hétérotypique, et dans le cas où ils correspondent à un même TF, on parle de CRM homotypique. Cette distinction est surtout utile pour décrire les différentes méthodes de prédiction de CRM, car la plupart des CRMs identifiés chez les Métazoaires sont hétérotypiques (Aerts, 2012). L'organisation de ces sites de liaison relève de deux types d'architecture principaux (fig. 1.16).

- **Le modèle "enhanceosome"**

Dans ce modèle, l'architecture des sites de liaison est de première importance. Le paradigme en est l'enhancer du gène humain interferon-β, sur lequel 8 TFs se lient pour former

FIGURE 1.16 – **Deux modèles d'enhancers : enhanceosome et billboard.**
Figure tirée de (Kulkarni and Arnosti, 2003). (A) Dans le modèle enhanceosome, l'enhancer traite l'information des multiples TFs qui le fixent. Un complexe très structuré crée une interface qui recrute la machinerie de transcription basale. L'enhancer peut être vu comme un ordinateur moléculaire qui produit à partir d'entrées multiples un seul signal vers la machinerie de transcription. Le gène cible n'est activé qu'en cas de formation du complexe entier, ce qui fournit un interrupteur binaire on/off seulement activé en cas de stimulus adéquat. La déstabilisation du complexe en changeant par exemple la concentration d'une des protéines permettrait alors d'obtenir une réponse graduelle. (B,C) Modèle d'enhancer « billboard ». Dans ce cas, l'enhancer ne consiste pas en une seule unité de régulation, mais en des sous-unités pouvant contenir différentes informations (répression ou activation par exemple) que la machinerie basale échantillonne soit itérativement (B), soit simultanément (C).

FIGURE 1.17 – **L'enhanceosome de l'interferon-β.**
Figure tirée de (Panne, 2008) représentant le complexe de TFs assemblés sur l'enhanceosome IFN-β, formant une surface de reconnaissance agissant comme une seule unité de régulation.

une surface de reconnaissance continue (Panne, 2008). Les TFBS de cet enhancer se recouvrent les uns les autres, créant au final un complexe de TFs fixés à l'ADN agissant comme une seule unité de régulation (fig. 1.17).

- **Le modèle "billboard"**

La majorité des CRMs adhèrent à ce type d'organisation. L'architecture y est libre : les sites de liaisons n'ont pas de contrainte de nombre, d'ordre, de sens, ou d'espacement (Kulkarni and Arnosti, 2003). De tels CRMs sont propices à une détection informatique basée sur la densité en sites de liaisons pour différents TFs.

1.5.3 Évolution des enhancers

La fonction centrale que jouent les enhancers dans la régulation de l'expression génétique laisse à penser que ceux-ci seraient sous sélection et leur séquence serait donc plus conservée que celle des régions non codantes du génome. De fait, la comparaison de séquences non-codantes entre espèces proches s'avère être un mode de détection puissant des régions de régulation (Prabhakar et al., 2006). Ainsi, l'utilisation de la conservation entre des espèces lointaines comme l'homme et le poisson *Fugu* ou de l'extrême conservation entre des espèces proches comme l'homme, la souris et le rat, permet de détecter des régions ayant une activité enhancer *in vivo* avec un succès proche de 50% (Pennacchio et al., 2006). À l'instar de la régulation de l'interféron-β, de telles séquences très contraintes obéissent à une logique de type « enhanceosome » où la fonction est intimement liée à la séquence.

Contrastant avec cette vision d'enhancers très contraints, plusieurs études pointent vers

FIGURE 1.18 – **Flexibilité du code de cis-régulation au cours de l'évolution chez les *Drosophiles*.**

Figure tirée de (Liberman and Stathopoulos, 2009), montrant l'expression du gène *sog* par ISH chez plusieurs espèces de *Drosophilidæ* (A,D,G,J), ainsi que l'expression d'un rapporteur *LacZ* sous contrôle de la région de régulation de *sog* chez l'espèce considérée (B,E,H,K), dont une représentation schématique est donnée pour chaque cas (C,F,I,L). Les sites de fixation pour différents TFs sont indiqués par un rectangle de couleur.

une plus grande flexibilité des séquences enhancers (Ludwig et al., 2000; Dermitzakis and Clark, 2002; Moses et al., 2006). Supportant l'idée que la plupart des enhancers se comportent selon le modèle « billboard », la grammaire des sites de fixation dans des séquences orthologues apparaît comme étant loin d'être figée (Liberman and Stathopoulos, 2009). Ainsi, l'enhancer régulant le gène *short gastrulation* (*sog*), bien que présentant chez différentes espèces de *Drosophiles* une architecture variable des sites de fixation le composant, conduit à un même motif d'expression (fig. 1.18). Cette idée qu'une panoplie de grammaires conduisent à une même régulation est confortée par les résultats de Zinzen et al. (2009) où des enhancers ayant des « entrées » différentes (i.e étant fixés par des TFs différents pendant des durées variables) produisent des « sorties » similaires, dans ce cas une expression spécifique à un tissu donné.

Supportant l'idée d'une flexibilité de la régulation, plusieurs études ont exhibé l'évolution rapide des sites de liaison de TFs dans le génome (Wilson and Odom, 2009). Une étude de la fixation génomique des facteurs de transcription CEBPα et HNF4α dans les cellules du foie de 5 espèces de vertébrés (l'homme, deux espèces de souris, le chien et le poulet) a notamment montré que les événements de fixation conservés chez les 5 espèces sont très rares ($\sim$ 0.3% des pics humains) et correspondent à des régions ultraconservées proches de gènes importants dans la spécification du foie (Schmidt et al., 2010). Par ailleurs, lors de la perte de fixation dans l'une des espèce, un gain de fixation proche ($\pm$10kb) est observé dans la moitié des cas. Étonnamment, ces changements rapides du câblage du réseau affectent peu l'expression génétique globale (Tirosh et al., 2008; Odom et al., 2007).

Cette évolution est en grande partie due à une évolution de séquence de fixation. Ainsi, une étude récente a utilisé une souris portant le chromosome 21 de l'homme pour comparer la fixation du facteur HNF4α dans un contexte murin par rapport au contexte original (Wilson et al., 2008). Le paysage de fixation sur le chromosome 21 exogène a très précisément récapitulé celui observé chez l'homme (fig. 1.19), montrant que le contexte cellulaire est sensiblement le même chez les deux espèces. Par ailleurs, des modifications épigénétiques ainsi que l'expression des ARNm ont pu être récapitulées.

Reste la question du mécanisme permettant cette évolution rapide. Une étude portant sur 7 facteurs de transcription chez les mammifères a montré qu'une proportion importante ($\sim$ 20%) des régions de fixation de ces TFs se situent au sein de différentes familles de transposons (Bourque et al., 2008) (fig. 1.19). Ces transposons, ou éléments transposables, sont des anciens rétrovirus intégrés dans les génomes mammifères ayant la capacité de se dupliquer pour

FIGURE 1.19 – **Évolution de la fixation de HNF4α chez les mammifères.**
Figure tirée de (Wilson and Odom, 2009), représentant la fixation par ChIP-seq (pics de couleur) du facteur de transcription humain HNF4α (ou son homologue murin HNF4a) chez l'homme, la souris, le macaque, le chien, ainsi qu'une souris transgénique contenant le chromosome 21 humain. Les zones grisées indiquent : (i) une fixation de HNF4α chez l'homme retrouvée sur le chromosome 21 humain de la souris mais pas chez le macaque malgré la proximité de séquence, (ii) une fixation de HNF4a spécifique à la souris, et (iii) une fixation spécifique aux primates qui a lieu sur des éléments transposables.

s'intégrer dans une autre région du génome et jouent un rôle fondamental dans l'évolution des génomes (Cordaux and Batzer, 2009). Leur accumulation dans le génome a vraisemblablement permis d'obtenir un matériau de base permettant de produire par mutations ponctuelles des éléments de régulation *de novo* (Feschotte, 2008). Par ailleurs, les transposons peuvent permettre de diffuser par « copier-coller » des éléments de régulation existant. Ainsi, des vagues d'expansion de transposons spécifiques à différentes espèces de mammifères sont à l'origine de la variabilité des régions de fixation observée dans le cas du facteur CTCF (Schmidt et al., 2012).

1.5.4 Les « shadow enhancers »

L'évolution des éléments de cis-régulation est un mécanisme majeur permettant la diversité animale. Néanmoins, de tels changements pourraient compromettre certaines activités génétiques essentielles. Des expériences de ChIP-on-chip ont suggéré que plusieurs gènes de développement actifs lors du développement précoce de l'embryon de Drosophile possèdent des CRMs secondaires, qui conduisent à des motifs d'expression génétique comparables à ceux produits par des CRMs « primaires » plus proximaux (Zeitlinger et al., 2007). L'expression de « shadow enhancer » a été proposée par Michael Levine en 2008 pour décrire ces CRMs redondants et souvent distaux de plusieurs dizaines de kb du gène régulé (Hong et al., 2008). Il est probable que de tels CRMs soient apparus au cours de l'évolution par duplication du CRM primaire, à l'instar du phénomène de duplication des séquences codant pour des protéines. L'avantage évident que peut conférer la redondance d'un élément de régulation est d'offrir de la robustesse face aux mutations. Par ailleurs, une telle redondance permet de faciliter la divergence et donc la spécialisation des différents CRMs. Ainsi les « shadow enhancers » semblent évoluer plus rapidement que les CRMs primaires auxquels ils sont apparentés (Hong et al., 2008) pour fournir de nouveaux sites de fixation et conduire à de nouvelles activités de régulation sans bloquer la fonction critique de certains gènes de développement.

Un exemple mêlant robustesse et divergence est le cas des multiples CRMs régulant le gène *Svb* chez la Drosophile. Chaque CRM est lié à la production d'un motif distinct de trichomes (excroissances de l'épithélium comparables à des poils) sur la larve : ainsi, plusieurs mutations dans ces différents CRMs sont nécessaires pour observer un changement morphologique conséquent (McGregor et al., 2007). Dans ce même système, il a été montré que deux CRMs supplémentaires, des « shadow enhancers », sont dispensables dans des conditions de

FIGURE 1.20 – **« Shadow enhancer » du gène de segmentation *Hunchback*.**
Figure tirée de (Perry et al., 2011), montrant des hybdridations *in situ* de transgènes contenant le gène rapporteur *LacZ* sous contrôle de l'enhancer proximal (A) ou du « shadow enhancer » plus distal (B) du gène de segmentation *Hb*. (C) Position des enhancers par rapport à *Hb* (bleu : proximal, rouge : distal).

température usuelles, mais requis lorsque les embryons se développent dans des conditions de température extrêmes (Frankel et al., 2010).

Par ailleurs, il a été montré que les gènes de segmentation (ou gènes *gap*) de la Drosophile possèdent tous des « shadow enhancers » (fig. 1.20). Leur rôle semble être d'assurer une plus grande précision spatiale du motif d'expression du gène régulé : la perte de l'un des CRMs, proximal aussi bien que « shadow », conduisant à une expression trop restreinte ou trop répandue spatialement selon le cas (Perry et al., 2011).

1.5.5 Par delà les enhancers : les « super-enhancers »

Récemment, il a été montré que certains groupements d'enhancers peuvent agir comme une même unité de régulation : on parle de *super-enhancers* (Whyte et al., 2013). Ces régions de taille typique $\sim$ 10kb (fig. 1.21), sont fixées par des TFs maîtres et sont associées à des gènes encodant des régulateurs clés de l'identité cellulaire. Identifiés dans les cellules souches embryonnaires (ESCs), ces ensembles d'enhancers sont fixés par le complexe co-activateur Mediator, qui interagit avec la cohésine pour former un anneau permettant de connecter la région de régulation au promoteur (Kagey et al., 2010). Par ailleurs, les gènes associés aux super-enhancers possèdent un niveau particulièrement élevé d'expression et leur knock-down est associée à une perte de l'état souche des cellules.

Ainsi, ce second niveau d'organisation de la régulation pourrait simplifier la modélisation de la régulation du type cellulaire, en passant de millier de traces de fixation pour différents TFs à quelques centaines de super-enhancers contrôlant les gènes clés de l'identité cellulaire.

FIGURE 1.21 – **De l'enhancer au super-enhancer.**
Figure tirée de (Whyte et al., 2013), montrant les profils de ChIP-seq des TFs maîtres Oct4, Sox2, Nanog, Klf4 et Esrrb aux loci de *Gck* et *miR-290-295* dans les cellules souches embryonnaires. Le super-enhancer se distingue du simple enhancer par sa taille (27kb), sa grande concentration en TFs maîtres, notamment Klf4 et Esrrb, et la fixation de la protéine Med1 du complexe Mediator.

1.6 Prédiction et validation des CRMs

1.6.1 Méthodes utilisant la concentration en sites de fixation

Nous l'avons vu, une propriété des CRMs est leur grande concentration en TFBS. Ceci a motivé des approches de prédiction de promoteurs et d'enhancers basées sur leur contenu ou *clustering* en motif (fig. 1.22a). L'avantage de telles approches est qu'elles peuvent être réalisées avec seulement la séquence d'ADN génomique et des modèles de TFs ou motifs (par exemple des PWMs, voir fig. 1.10) représentant les facteurs de transcription impliqués dans le processus étudié. Cependant, les clusters de motifs sont très répandus dans les grands génomes, et sans l'ajout d'informations supplémentaires comme les marques épigénétiques ou l'expression des gènes voisins, ces approches produisent un grand nombre de faux positifs (éléments prédits comme positifs mais étant en réalité négatifs). Par ailleurs, les TFs impliqués ne sont pas toujours connus, et il faut alors apprendre des motifs putatifs à partir de séquences fonctionnelles.

- **Approches utilisant des motifs connus**

L'une des premières investigations basée sur le regroupement de TFBS utilisait 5 motifs connus de la détermination musculaire pour prédire par régression linéaire les CRMs actifs dans le muscle (Wasserman and Fickett, 1998). Le taux de validation était relativement bas, autour de 20%. De même, chez *Drosophila melanogaster*, plusieurs études ont utilisé le clustering de motifs pour prédire des CRMs de différents processus développementaux (par ex Berman et al. (2002)). Ces études ont trouvé de nouveaux enhancers validés expérimentalement (bonne sensibilité) mais avaient des taux de prédiction relativement bas, entre 15 et 30%. L'algorithme *Ahab* (Rajewsky et al., 2002), utilisant un modèle thermodynamique de fixation des TFs sur les CRMs, a quant à lui réussi à prédire un nombre bien plus important de régions fonctionnelles : $\sim$ 80% des modules prédits à proximité de 29 gènes de segmentation chez la drosophile ont effectivement récapitulé le motif d'expression du gène associé (Schroeder et al., 2004). Ce succès semble notamment être dû au fait que ce modèle thermodynamique, basé sur une prise en compte exhaustive de toutes les segmentations possibles des CRMs en motifs et en ADN « background », permet de donner plus de poids au cas où plusieurs sites de faibles affinité pour un TF se trouvent au sein d'un même module, alors que les autres méthodes utilisent généralement un seuil de probabilité relativement élevé (afin d'éviter les faux posi-

FIGURE 1.22 – **Différentes approches pour la prédiction des CRMs.**
Figure tirée de (Hartwell et al., 1999), résumant les différentes approches de prédiction des CRMs. (a) Prédiction de sites sites de fixations (TFBS) pour différents facteurs de transcription. Le regroupement ou *clustering* de TFBS définit un CRM putatif. (b) Le CRM, possédant des contraintes fonctionnelles, est conservé chez plusieurs espèces proches. La séquence d'ADN correspondant au CRM (un enhancer connu de la β-globine) est montrée chez les différentes espèces. Les nucléotides qui diffèrent par rapport à la séquence de référence (la souris) sont indiqués en bleu, et les sites de fixation de TFs connus ou prédits sont indiqués en orange. Ceux-ci correspondent à des séquences complètement conservées : on parle d'empreinte phylogénétique ou *phylogenetic footprint*. (c) Une expérience de ChIP-seq spécifique au TF A produit un pic au niveau du CRM. Il est à noter que des pics peuvent exister même si aucun site de fixation n'est prédit, dû par exemple à une interaction protéine-protéine entre le TF A et un autre TF fixé à l'ADN.

tifs) à partir duquel une séquence est considérée comme fixée par un TF. Par ailleurs, cette étude s'est restreinte à un ensemble de gènes connus pour lesquels les régions à proximité riches en TFBS ont *a priori* plus de chances d'être fonctionnelles. De manière générale, plus le domaine de recherche est étendu (par exemple, le génome entier), plus le nombre de faux positifs augmente.

- **Approches *de novo* où les motifs ne sont pas connus**

Lorsque les motifs (PWMs) ne sont pas connus à l'avance, il faut les générer *de novo* à partir de leur surreprésentation dans des CRMs connus. Par exemple, l'algorithme CisModule permet de générer des motifs et des modules simultanément en utilisant un modèle de mélange hiérarchique (Zhou and Wong, 2004). Lorsqu'il est appliqué aux CRMs musculaires introduits précédemment, il permet de retrouver certains motifs connus et permet de retrouver $\sim 70 - 80\%$ des séquences connues lorsqu'elle sont mélangées avec un nombre similaire de séquences aléatoires. Par ailleurs, l'apprentissage de modèles permettant de discriminer différentes classes de CRMs entre elles plutôt qu'une classe de CRMs par rapport à des séquences aléatoires ou intergéniques peut s'avérer plus fructueux. Ainsi, (Smith et al., 2006) ont utilisé des motifs connus ainsi que des motifs appris *de novo* avec le programme DME (Smith et al., 2005) pour leur capacité à discriminer des séquences appartenant à différents jeux de données de régions promotrices pour bâtir un modèle de régression logistique permettant de prédire l'activité tissu-spécifique dans 45 des 56 tissus humains et murins considérés. Il existe aussi plusieurs méthodes qui n'utilisent pas de motifs du type PWM, mais de purs modèles probabilistes tels que des chaînes de Markov d'ordre 5 ou des regroupements de « mots » de k nucléotides ou k-mers selon des critères de distance de Hamming et surreprésentés dans les séquences d'intérêt, par exemple (Cao et al., 2010a). Ces méthodes sont passées en revue dans (Kantorovitz et al., 2009), et elles peuvent atteindre des sensibilités de $\sim 60\%$ pour la prédiction de CRMs mammifères. L'intérêt est que ces études ne présument pas d'un modèle de fixation des TFs à l'ADN. C'est aussi un désavantage, puisqu'elles sont moins informatives quant au réseau génétique sous-jacent et aux mécanismes de régulation impliqués.

1.6.2 Méthodes utilisant la phylogénie

Les approches utilisant la comparaison des génomes de différentes espèces pour prédire des CRMs sont basées sur l'idée que les séquences de régulation sont plus fortement conser-

vées que l'ADN non fonctionnel les entourant. Nous l'avons vu en 1.5.3, une proportion importante de CRMs ne satisfont pas à cette règle. Cette approche ne permet donc d'étudier que le sous-ensemble de CRMs qui a subi une forte pression de sélection depuis le dernier ancêtre commun aux espèces considérées et ne donne pas accès aux CRMs apparus récemment au sein d'une espèce.

- **Prédictions à partir de la contrainte évolutive seule**

L'alignement de séquences non-codantes orthologues fait apparaître des parties très conservées, avec peu de variations dans les séquences sous-jacentes, entourées de séquences accumulant les variations (fig. 1.22b). De telles séquences conservées sont alors interprétées comme ayant été sous sélection, les substitutions délétères ayant été rejetées au cours de l'évolution (Dermitzakis et al., 2005). Par analogie avec les empreintes à la DNAse I, on parle d'empreinte phylogénétique pour caractériser ces courtes séquences très conservées ($\sim$ 10bp), traces de la fixation putative d'un facteur de transcription. Ces empreintes s'avèrent être un indicateur fiable de fonctionnalité (Kheradpour et al., 2007) et, parce qu'elles ne reposent pas sur des modèles *a priori* de fixation, elles permettent de plus de trouver des motifs de régulation non connus (Xie et al., 2005). Au niveau de séquences plus longues ($\sim$ 100bp), la contrainte évolutive permet de détecter des CRMs entiers. Ainsi, comme nous l'avons vu en 1.5.3, l'utilisation de la conservation extrême permet d'atteindre 50% de taux validation (Pennacchio et al., 2006). Néanmoins, lorsque ces contraintes de conservation extrême (par exemple homme-*Fugu*) sont relâchées, le taux de validation tombe drastiquement, atteignant $\sim$ 5% (Attanasio et al., 2008), montrant la nécessité d'allier le critère de conservation à d'autres données (expression, ChIP...) pour améliorer la prédiction des CRMs.

- **Prédictions utilisant la phylogénie et des motifs connus**

Une approche pour améliorer les prédictions est de combiner les approches précédentes en utilisant à la fois le *clustering* en TFBS et la contrainte évolutive. À l'échelle du génome entier, cette approche permet de filtrer les résultats pour améliorer le signal de détection chez la Drosophile (Sinha et al., 2004). Du côté des mammifères, en utilisant les motifs de la base de données TRANSFAC et la conservation entre l'homme et la souris, Blanchette et al. (2006) ont créé une base de données de modules, PReMods, qui retrouve $\sim$ 17% de CRMs connus et recoupe 40% des fragments occupés par le co-activateur et marqueur de l'activité enhancer p300. D'autres méthodes se sont concentrées sur des types cellulaires bien définis.

Par exemple, la recherche de sites conservés pour des motifs de TF des cellules sanguines connus (Donaldson et al., 2005) a permis de définir des CRMs dont 2 ont été testés et validés.

Certains efforts ont par ailleurs été menés pour sortir du cadre d'une conservation de séquence stricte en modélisant l'évolution d'un CRM fixé par un certain nombre de motifs connus. Par exemple, le modèle MorphMS (Sinha and He, 2007) cherche au sein d'un alignement de deux séquences orthologues des régions prédites par un modèle d'évolution dérivé d'un ensemble de motifs choisis par l'utilisateur. Une extension de cette approche incorpore le gain et la perte de sites de fixation, mais n'a cependant pas encore été appliquée à l'échelle du génome (Majoros and Ohler, 2010).

- **Approches utilisant la phylogénie pour générer des motifs *de novo***

De même que précédemment, tous les motifs ne sont pas connus et il peut être utile d'avoir recours à de l'apprentissage direct à partir de séquences fonctionnelles connues pour aider à la prédiction. Par exemple, l'algorithme ESPER cherche des patterns (TFBS, %GC, etc) surreprésentés dans des alignements multi-espèces de CRMs connus par rapport à des alignements d'ADN *a priori* non fonctionnel (Taylor et al., 2006). Cette méthode n'est pas restreinte à l'analyse de séquences conservées puisqu'elle peut potentiellement capturer des signatures de changements systématiques. La prédiction de régions de haut potentiel de régulation recouvre presque entièrement les prédictions de PReMods, et le test par transfection de ces régions à proximité de gènes exprimés dans les cellules érythroïdes et possédant un site pour un TF spécifique de l'érythroïde mène à un taux de validation de 50%. Une autre méthode consiste à chercher des mots surreprésentés dans un ensemble d'apprentissage de CRMs connus puis à restreindre les prédictions aux régions conservées (Kantorovitz et al., 2009). Les prédictions réalisées ont toutes été validées chez la Drosophile (5/5) comme chez la souris (2/2).

1.6.3 Méthodes utilisant les marques épigénétiques et de ChIP-seq pour des TFs

- **Prédiction des promoteurs**

La méthode la plus fiable de prédiction d'un promoteur utilise le fait qu'il est toujours localisé au niveau d'un TSS, dont la position peut facilement être obtenue en alignant les séquences de l'ARN du gène correspondant sur le génome (Trinklein et al., 2003). Le taux de validation avec cette seule contrainte est très élevé : 91% ont une activité dans au moins un type cellulaire. Par ailleurs, la marque épigénétique H3K4me3 est aussi un indicateur des

promoteurs actifs dans le type cellulaire étudié (Heintzman et al., 2007) (fig. 1.14).

- **Prédiction des enhancers**

La prédiction des enhancers à partir des marques épigénétiques, comme l'acétylation des histones (Roh et al., 2005), la méthylation H3K4me1 (Heintzman et al., 2009), ou encore la présence du co-activateur p300 (Visel et al., 2009a), est très efficace, avec une expression tissue-spécifique dans $\sim$ 80% des cas (Hardison and Taylor, 2012). Par exemple, ces différentes marques, présentes dans différents tissus, peuvent être utilisées comme autant d'entrées d'un modèle de Markov caché pour produire des prédictions fiables de CRMs tissu-spécifiques chez l'homme (Ernst et al., 2011).

En fait, les prédictions d'activité enhancer à partir de ces marques épigénétiques est plus fiable qu'en utilisant la fixation de facteurs de transcription tissu-spécifiques. Par exemple, sur 63 séquences ADN fixées *in vivo* par le facteur spécifique des cellules sanguines GATA1 chez la souris, seulement la moitié conduisent à une activité après transfection dans des cultures cellulaires (Cheng et al., 2008). Ces enhancers fonctionnels sont par ailleurs plus particulièrement associés à un site de fixation *conservé* pour GATA1, montrant à nouveau la nécessité de combiner les approches pour améliorer la détection. Un taux de validation similaire a été observé pour le facteur de différenciation myogénique MyoD, avec 40% de régions fixées ayant une activité après transfection en cellules.

L'utilisation de données de fixation pour plusieurs TFs à la fois semble cependant améliorer le pouvoir de prédiction. Ainsi, Tijssen et al. (2011) ont étudié la co-fixation de GATA1 avec 4 autres TFs hématopoïétiques dans des mégacaryocytes. En s'intéressant aux gènes à proximité de ces régions, ils en ont découvert plusieurs qui n'étaient pas précédemment connus comme étant important dans l'hématopoïèse. Leur fonction a été testée par knock-down, avec dans 8 cas sur les 9 testés une réduction de la production de globules rouges.

1.6.4 Validation expérimentale

Une méthode directe permettant de démontrer qu'un fragment d'ADN régule l'expression génétique consiste en une expérience de gain de fonction dans laquelle un plasmide contenant le CRM prédit à proximité d'un gène rapporteur est introduit par transfection *in vitro* en cellule, permettant un suivi quantitatif de l'activité, ou par transgenèse *in vivo* dans un organisme, auquel cas le suivi est plus qualitatif mais permet d'établir la spécificité spatio-

FIGURE 1.23 – **Méthodes de validation des CRMs par transfection et transgenèse.** Figure tirée de (Hartwell et al., 1999) présentant les méthodes *in vitro* et *in vivo* de validation des CRMs. La région dont on souhaite tester l'activité est insérée dans un plasmide codant pour un gène rapporteur qui est transféré dans une culture cellulaire (transfection, panel a) ou dans un organisme entier (transgenèse, panel b). Dans le cas du test d'un promoteur, le CRM est placé directement en amont du gène rapporteur (on utilise généralement la luciférase *Luc*), alors que dans le cas d'un enhancer, le CRM est placé en amont d'un promoteur minimal de faible activité. L'activité de la luciférase donne une information quantitative sur l'activité de la région testée (boîtes à moustache, panel a). Dans le cas d'une transgenèse, le gène rapporteur généralement utilisé est *lacZ* qui encode la β-galactosidase. La révélation par coloration permet de visualiser en bleu les tissus au sein lesquels l'enhancer est actif.

temporelle (tissu et stade de développement) de l'élément de régulation (fig. 1.23). Ce type d'expérience montre que le CRM prédit est *suffisant* pour reproduire le motif génétique observé. De manière optimale, il faudrait aussi montrer par délétion ciblée de l'élément de régulation au sein du génome que ce dernier est *nécessaire* à l'expression du gène endogène.

FIGURE 1.24 – **Impact physiologique de la délétion et de la mutation d'un enhancer.** Figure tirée de (Visel et al., 2009b). (a) L'enhancer « membre » de *Sonic hedgehog* (*Shh*) est situé à $\sim$ 1 Mb de son gène cible, au sein d'un intron d'un gène voisin. Par transgenèse il conduit l'expression d'un rapporteur dans la partie postérieure du bourgeon de membre chez l'embryon de souris (Lettice et al., 2003). (b) Les souris possédant une délétion ciblée de cet enhancer possèdent des membres tronqués, montrant son importance fonctionnelle au cours du développement (Sagai et al., 2005). (c-e) Différentes mutations ponctuelles dans la séquence enhancer orthologue chez l'homme résultent en une polydactilie pré axiale (présence de doigts ou orteils supplémentaires), montrant l'impact potentiel de variations dans des régions non-codantes (Lettice et al., 2003).

1.6.5 Implication des CRMs dans les maladies humaines

Au cours des dernières décennies, de nombreuses mutations dans les régions codantes des gènes, impliquant des défauts structurels des protéines associées, ont pu être associées à des maladies génétiques. À l'inverse, le rôle des mutations affectant des régions non codantes n'a été que peu exploré, essentiellement du fait de la difficulté d'annoter ces régions correctement afin de définir celles qui pourraient avoir une fonction d'intérêt. Plusieurs études ont cependant pu montrer que des variations affectant des enhancers distaux pouvaient conduire à des pathologies (Visel et al., 2009b).

L'une de ces études concerne l'enhancer spécifique du membre de *Shh* (fig. 1.24). Cet enhancer, initialement décrit chez la souris, se situe à environ 1 Mb de distance de *Shh*, au sein

de l'intron d'un gène voisin. Le séquençage de cet enhancer chez plusieurs individus humains a permis d'associer une douzaine de variations mono-nucléotidiques à la polydactilis pré axiale, c'est-à-dire la présence de doigts ou d'orteils supplémentaires (Lettice et al., 2003). Des études supplémentaires chez la souris ont montré que les variations de séquences observées dans cet enhancer conduisent à une expression ectopique dans la partie antérieure du membre au cours du développement, ce qui est consistent avec la présence de doigts supplémentaires (Masuya et al., 2007). Par ailleurs, la délétion de l'enhancer orthologue de la souris entraîne la troncation des membres (Sagai et al., 2005).

Ainsi, ces résultats montrent l'importance de l'identification des enhancers pour permettre à des études de génétique humaine d'explorer le rôle potentiellement pathologique de mutations dans des régions non codantes fonctionnelles.

1.7 Bases de données

La biologie moderne est caractérisée par l'accumulation de données biologiques qu'il s'agit d'intégrer puis d'interpréter : on parle de biologie integrative. En particulier, depuis le séquençage du génome humain il y a maintenant plus de dix ans (Lander et al., 2001), le nombre de génome séquencés n'a cessé d'augmenter, tandis que dans le même temps le prix du séquençage diminuait drastiquement (fig. 1.25). Afin de permettre la gestion et l'utilisation de ces données, de nombreux outils et bases de données ont été mis à disposition (Wasserman and Sandelin, 2004). Nous évoquons ici ceux qui nous paraissent essentiels du point de vue de la régulation en *cis*.

FIGURE 1.25 – **Évolution du coût de séquençage.**
Figure adaptée de (Sboner et al., 2011), montrant l'évolution du coût du séquençage d'1 Mb d'ADN au cours de la dernière décennie, comparé à une évolution de type « Loi de Moore » où le prix serait diminué de moitié tous les 18 mois.

1.7.1 Obtention de données génomiques

Tout d'abord, les différents génomes séquencés sont à disposition sur des bases de données publiques d'où ils peuvent être téléchargés puis analysés en aval. Parmi les plus généralistes se trouvent la base de donnée de UCSC (UCSC Genome Browser, `http://genome.ucsc.edu`)

et celle de l'EMBL (Ensembl, http://www.ensembl.org) [4].

Sont à disposition les génomes des différentes espèces séquencées pour les différents assemblages réalisés, des alignements des génomes de différentes espèces deux par deux (*pairwise alignments*) ou par groupes d'espèces (*multiple alignments*), ainsi qu'un certain nombre d'annotations essentielles à l'analyse de ces génomes : coordonnées des gènes (TSSs, exons, introns avec potentiellement différents transcrits alternatifs), miRNA ou lincRNA, ontologies associées, coordonnées des séquences répétitives (les *repeats*, en partie liés aux éléments transposables abordés en 1.5.3, et qui sont abondants dans les génomes vertébrés), différentes données ChIP-seq, indices de conservation [5]...

Au final, ces différentes données constituent une base de travail fiable et régulièrement mise à jour. Afin de faciliter leur obtention, il est possible d'utiliser le navigateur de tables de UCSC [6] ou la section BioMart d'Ensembl [7] .

Situé plus en amont, le projet Galaxy (http://galaxyproject.org) permet à l'utilisateur de récupérer des données depuis les différentes banques existantes, puis de leur faire subir divers traitements et analyses par divers outils de bioinformatique. Cet outil, qui peut être utilisé sur internet ou bien localement, a l'avantage de permettre la sauvegarde de plans de travail ou *workflows*, successions de commandes utilisées pour traiter une entrée donnée par différents outils stéréotypés et obtenir directement le résultat final, favorisant une approche conviviale orientée utilisateur.

En guise d'exemple, nous montrons en annexe A des statistiques obtenues aisément à partir d'annotations génétiques présentes sur UCSC et traitées avec Galaxy. Ces statistiques sont les distribution de tailles des régions intergéniques et introniques chez plusieurs espèces : la bactérie *Escherichia coli*, la levure *Saccharomyces cerevisiae*, le ver *Caenorhabditis elegans*, la mouche *Drosophila melanogaster*, la souris, le poulet et l'homme (fig. A.1).

4. Les données sont accessibles sur les pages de téléchargement, respectivement http://hgdownload.cse.ucsc.edu/downloads.html pour UCSC et http://www.ensembl.org/info/data/ftp/index.html pour Ensembl
5. Pour le cas de l'assemblage mm9 de la souris, ces annotations sont accessibles à l'adresse suivante : http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database/
6. http://genome.ucsc.edu/cgi-bin/hgTables
7. http://www.ensembl.org/biomart/martview

1.7.2 Obtention de données sur les TFs

Nous l'avons vu, les données de fixation des TFs (ChIP-seq, ChIP-on-chip) peuvent être obtenues à partir du site UCSC Genome Browser. Ces données sont aussi généralement accessibles sur le site du NCBI (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/) *via* un numéro d'accession donné lors de la publication des données.

De nombreux modèles de TFs ont déjà été bâtis préalablement à l'avènement des données haut-débit de type ChIP-seq, par exemple avec des données SELEX, et il existe des bases de données stockant les PWMs correspondantes : JAPSAR, base de donnée publique[8], et TRANSFAC, qui marche par abonnement[9]. Il est à noter que ces PWMs ayant souvent été construites à partir d'un faible nombre de sites de fixations et de données *in vitro*, elles peuvent être relativement inadaptées à l'analyse de données *in vivo*.

1.7.3 Outils de visualisation

Afin d'avoir une idée plus claire des événements de régulation qui se déroulent à un locus donné, il existe plusieurs outils de visualisation des annotations génomiques et épigénétiques, que ce soit sur le site du NCBI (http://www.ncbi.nlm.nih.gov/gene), sur Ensembl ou sur UCSC Genome Browser. Ce dernier possède notamment l'avantage qu'il est possible d'importer des données personnelles sous un grand nombre de formats, obtenues à partir de la littérature ou à partir de ses propres travaux. Ainsi, nous présentons en figure 1.26 quelques données de ChIP-seq pour des TFs musculaires et pour des marques épigénétiques, ainsi que des prédictions bioinformatiques de sites de fixation conservés pour les homéoprotéines Six réalisée par nos soins. La visualisation sur UCSC Genome Browser permet de rapidement déterminer le mode de régulation putatif du gène *Chrng* : fixation de Six et MyoD au niveau du promoteur et apparition de marques épigénétiques H3K4me1 et H3Ac sur les histones au cours de la différenciation de progéniteurs musculaires.

Par ailleurs, il existe un outil de visualisation complémentaire de ceux cités : le visualisateur de régions conservées au cours de l'évolution ECR Browser (http://ecrbrowser.dcode.org), intégrant de nombreux outils bioinformatiques (Loots and Ovcharenko, 2005). Ce navigateur permet de visualiser la conservation génomique d'un locus donné chez plusieurs espèces plus ou moins lointaines (par exemple souris, homme, vache, grenouille et poisson

8. http://jaspar.cgb.ki.se
9. http://www.gene-regulation.com/pub/databases.html

FIGURE 1.26 – **Visualisation de données ChIP-seq *via* le site UCSC.**
La visualisation de différentes données ChIP-seq et bioinformatiques (bandes noires) permet de mettre en perspective le cas de la régulation de *Chrng* (en bleu au bas de l'image) lors de la différenciation musculaire. Les données ChIP-seq sont issues de la littérature et les données bioinformatiques (MEF3) ont été obtenues au cours de cette thèse. Les rectangles noirs correspondent aux coordonnées des pics obtenus après avoir appliqué un seuil de filtrage du bruit. Les données de fixation de PolII ainsi que les données de méthylation et d'acétylation des histones (H3K4me3 et H3K9Ac, marques de l'activité transcriptionnelle, voir fig. 1.14), tirées de Asp et al. (2011), indiquent que le locus est transcrit lors de la formation de myotubes. Le TF MyoD est fixé au niveau du promoteur de *Chrng* au cours de la prolifération des myoblastes (MB50, 50% de confluence, et MB95, 95% de confluence) et au cours de la différenciation en myotubes MT (Cao et al., 2010b). Le TF Six est co-fixé avec MyoD lors de la différenciation : à T24, 24h après différenciation, et à MT (Liu et al., 2010). De plus, des analyses bioinformatiques montre l'existence dans cette région d'un site de fixation MEF3 pour la protéine Six conservé chez les vertébrés, corroborant une liaison directe de l'ADN par Six. Prises ensemble, la simple visualisation de ces données suggèrent une régulation par Six et MyoD de *Chrng*.

zèbre) afin de cibler l'étude de la régulation sur des régions extrêmement conservées. Il est ensuite possible d'analyser les séquences ultraconservées sélectionnées en utilisant les motifs de la base de donnée TRANSFAC *via* l'outil rVISTA (Loots and Ovcharenko, 2004). Un exemple d'utilisation de cet outil est donné par la découverte de plusieurs régions de régulation fonctionnelles de l'homéoprotéine Six1 possédant une extrême conservation (Sato et al., 2012).

1.7.4 Le projet ENCODE

FIGURE 1.27 – **Les différentes données obtenues par le projet ENCODE.**
Figure tirée du site web du projet ENCODE (`http://genome.ucsc.edu/ENCODE`). Les équipes associées à ce projet utilisent de multiples expériences et méthodes pour identifier les éléments fonctionnels du génome. L'annotation des gènes se fait principalement par le séquençage de l'ARN provenant de différentes sources (tissu, stade de développement), la génomique comparative, des méthodes bioinformatiques ou encore par annotation manuelle. Les éléments de régulation sont quant à eux étudiés *via* des expériences d'hypersensibilité à la DNAseI, l'étude de la méthylation de l'ADN, ou encore les données ChIP-seq de divers facteurs de transcription et histones spécifiquement méthylées ou acétylées.

Le projet ENCODE (pour *Encyclopedia of DNA Elements*) est un consortium de groupes de recherche internationaux financés par le NHGRI (*National Human Genome Research Institute*)

qui a vu le jour afin de systématiser les méthodes permettant l'annotation des génomes et de faciliter l'intégration des nombreuses données obtenues. Son but est de construire une liste exhaustive des éléments fonctionnels du génome humain, qu'ils agissent au niveau de l'ADN, de l'ARN ou des protéines, et des éléments de régulation qui contrôlent l'état cellulaire et l'activité des gènes. Les données sont mises à disposition du public gratuitement sur internet (http://genome.ucsc.edu/ENCODE/). À noter que des projets équivalents existent pour d'autres organismes, comme la souris (http://mouseencode.org), ou encore le ver *Caenorhabditis elegans* et la mouche *Drosophila melanogaster* (http://www.modencode.org).

Totalisant en septembre 2012 plus de 1600 expériences dans plus de 147 types cellulaires, les premières conclusions pointent vers une profusion d'événements de régulation, loin de l'idée d'ADN poubelle (*junk DNA*) : ainsi, 80% du génome est associé à un événement biochimique associé à de la formation d'ARN ou au remodelage de la chromatine, $\sim 400,000$ régions possèdent un état chromatinien caractéristique des enhancers et $\sim 70,000$ des promoteurs (ENCODE Project Consortium et al., 2012). Depuis mai 2013, les données ChIP-seq de 161 TFs couvrant 91 types cellulaires ont été mises à disposition sur UCSC Genome Browser [10].

10. http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeAwgTfbsUniform

Chapitre 2

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

2.1 Observations de corrélations au sein des TFBS **66**

2.2 Modèles existants permettant de décrire la statistique des TFBS **67**

2.2.1 Modèle de référence sans corrélations : la PWM 67

2.2.2 Une PWM généralisée : le modèle GWM 69

2.2.3 Réseaux bayésiens 70

2.2.4 Modèles de mélange 71

2.3 Modèles de maximum d'entropie **72**

2.3.1 Pourquoi maximiser l'entropie ? 72

2.3.2 Maximisation de l'entropie sous contraintes 74

2.3.3 Application aux sites de fixation 75

2.4 Article **77**

2.5 Analyse thermodynamique des modèles **109**

2.5.1 Chaleur spécifique 109

2.5.2 Lien avec les valeurs des champs et des couplages 110

2.6 Conclusion et perspectives **112**

Introduction du chapitre 2

Dans cette partie, nous nous intéressons à la description de l'interaction entre les facteurs de transcription et leurs sites de reconnaissance sur l'ADN. Pendant longtemps, la qualité de cette description a été limitée par la quantité de données disponibles. Ainsi, les expériences de type SELEX (voir 1.4.1), où des expériences de ChIP au cas par cas permettaient de récupérer de l'ordre de quelques dizaines de sites de fixation pour un TF d'intérêt. Or, le modèle PWM, qui est le modèle le plus simple (en terme de nombre de paramètres) que l'on puisse bâtir pour décrire l'interaction possède déjà plusieurs dizaines de paramètres – les fréquences des nucléotides à chaque position –.

Ces données ne permettaient donc pas d'explorer plus en avant des modèles plus complexes de fixation incluant par exemple des termes d'interaction entre nucléotides au sein des sites de fixation. Cependant, les avancées récentes en séquençage à haut débit ont permis l'obtention de données très grande échelle, que ce soit *in vivo* par ChIP-seq ou *in vitro* par HT-SELEX (voir 1.4). Le nombre de sites de fixation obtenus est de l'ordre de quelques milliers, ce qui permet de contraindre des modèles de fixation plus complexe que le modèle PWM.

En utilisant des données ChIP-seq pour un grand nombre de facteurs de transcription de la Drosophile et des vertébrés, nous avons contraint différents modèles de fixation incluant implicitement ou explicitement des interactions entre nucléotides. Nous les avons comparés sur leur capacité à décrire les statistiques de fixation TF-ADN observées *in vivo*. Nous présentons préalablement un survol des observations et modèles existant au sujet des corrélations dans les sites de fixations de facteurs de transcription.

2.1 Observations de corrélations au sein des TFBS

Différents travaux ont mis en exergue l'existence de corrélations entre nucléotides au sein des sites de fixation de TFs. Parce que limitées par la quantité de données alors possible d'obtenir, les premières études de ce genre ont centré leur attention sur quelques corrélations importantes pour des cas particuliers. Ainsi, Man and Stormo (2001) ont observé que la protéine Mnt induit des corrélations entre les positions 16 et 17 de ses sites de reconnaissance *in vitro*. Ils ont mesuré expérimentalement la spécificité aux sites de liaisons contenant tous les variants possibles à ces deux positions. Ils ont ainsi observé que la mutation de la base consensus C en position 17 induisait un changement de préférence en position 16 de la base

A vers la base C. Par ailleurs, Bulyk et al. (2002) ont montré que la protéine EGR1 induisait des corrélations au sein d'un triplet de nucléotides central de leur site de reconnaissance. La prise en compte de ces corrélations dans l'énergie de fixation permettait alors d'améliorer la description des données par rapport au modèle additif PWM.

À une plus grande échelle, Badis et al. (2009) ont utilisé des puces à ADN (technique PBM, cf 1.4.1) pour étudier la fixation *in vitro* de 104 TFs de la souris sur toutes les séquences d'ADN de 10 bp possibles. Pour chaque facteur, plusieurs centaines de séquences de fixation ont ainsi été obtenues. L'étude a révélé l'existence d'une multiplicité de motifs (PWMs) pour la plupart des TFs (seulement 15 étant mieux décrit par un motif unique). Certains motifs reconnaissent notamment des séquences à espacement variable pour lesquelles deux régions spécifiques du site sont séparées par un nombre variable de nucléotides. Enfin, les auteurs ont noté la présence de corrélations fortes dans 19 cas, celles-ci n'étant pas forcément limitées à des dinucléotides mais pouvant impliquer des trinucléotides. Plus récemment, Jolma et al. (2013) ont analysé par HT-SELEX plusieurs centaines de domaines de fixations à l'ADN de TFs humains et de la souris, révélant aussi l'importance d'espacements variables et surtout des corrélations dinucléotidiques entre plus proches voisins.

2.2 Modèles existants permettant de décrire la statistique des TFBS

Différents modèles ont été proposés pour décrire ces corrélations (fig. 2.1). La méthode la plus directe consiste à partir du modèle PWM (fig. 2.1a) et à ajouter des corrélations mutuellement exclusives aux positions les plus corrélées (fig. 2.1b). D'autres méthodes utilisent des structures probabilistes de dépendances sous forme de chaînes de Markov (fig. 2.1c) ou plus généralement de réseau bayésien ou (fig. 2.1d-e). Enfin, une dernière méthode consiste à réaliser des mélanges de modèles afin de capturer des ensembles distincts de corrélations (fig. 2.1f-g).

2.2.1 Modèle de référence sans corrélations : la PWM

Nous l'avons vu, le modèle le plus simple (en termes de nombre de paramètres) décrivant l'interaction entre un TF et son site de reconnaissance sur l'ADN consiste à faire l'hypothèse que les nucléotides contribuent indépendamment à l'énergie de fixation. Cette hypothèse

FIGURE 2.1 – **Différents modèles pour décrire les corrélations entre nucléotides dans les sites de fixation de facteurs de transcription.**
Exemples illustrant différents modèles de fixation sur un site de longueur 5. Pour chaque modèle, la structure du réseau de dépendances sous-jacent est représentée, ainsi que la distribution de probabilité $P(X_1, X_2, X_3, X_4, X_5)$ correspondante, où X_i est une variable aléatoire prenant la valeur (A, C, G ou T) du nucléotide à la position i. Les modèles représentés sont les suivants : (a) PWM (pas de corrélations), (b) GWM (corrélations mutuellement exclusives), (c) chaîne de Markov d'ordre 1 (corrélations entre plus proches voisins), (d) réseau bayésien en arbre (au plus un parent par nœud) ou (e) pas en arbre (le nœud 2 a deux parents), (f) mélange de PWMs et (e) mélange d'arbres à dépendances fixées.

conduit au modèle PWM (section 1.3.2 et fig.2.1a), qui s'écrit [11] :

$$P(X_1, \cdots, X_k) = \prod_{i=1}^{K} P(X_i) \tag{2.1}$$

où $P(X_i)$ est la probabilité marginale d'observer le nucléotide $X \in \{A, C, G, T\}$ à la position i. Un tel modèle possède $3K$ paramètres – 3 paramètres $P(X_i)$ par position, la normalisation des probabilités permettant de fixer le paramètre restant –. Pour une longueur de site typique $K = 10$, le modèle PWM contient 30 paramètres à contraindre, sachant qu'un « modèle »

11. Comme nous l'avons signalé en 1.3.2, le terme PWM (*Position Weight Matrix*) réfère en fait à la matrice des poids $\log(P(X_i)/\pi_{X_i})$ où π_{X_i} est une distribution neutre indépendante de la position (dite distribution *background*), par exemple calculée sur des régions intergéniques.

complet paramétrant la distribution jointe sans faire d'hypothèse comporterait $4^{10} - 1 \sim 10^6$ paramètres.

2.2.2 Une PWM généralisée : le modèle GWM

Une première méthode permettant de complexifier le modèle PWM consiste à intégrer explicitement des groupes mutuellement exclusif [12] de nucléotides corrélés au sein du modèle (fig. 2.1b). Une telle méthode fut d'abord employée par Benos et al. (2002) pour prendre en compte des corrélations préalablement définies entre nucléotides plus proches voisins. De manière plus générale, Zhou and Liu (2004) ont développé un modèle de matrice de poids généralisée (GWM pour *Generalized Weight Matrix*) qui prend en compte de manière systématique les corrélations permettant d'améliorer le modèle indépendant. Pour ce faire, les auteurs utilisent une méthode de Monte-Carlo par chaîne de Markov (MCMC) : des corrélations sont ajoutées ou enlevées au hasard au modèle et acceptées selon la règle de Metropolis-Hastings (Krauth, 2006). Cette acceptation est proportionnelle au facteur de Bayes, une quantité qui permet de comparer des modèles possédant des nombres de paramètres différent [13] . Ce facteur est défini par le rapport entre la probabilité de générer les données D (les séquences de fixation) avec un modèle M_1 de paramètres θ_1 plutôt qu'avec un autre modèle M_2 de paramètres θ_2 :

$$BF = \frac{P(D|M_1)}{P(D|M_2)} = \frac{\int P(D|\theta_1, M_1)P(\theta_1|M_1)d\theta_1}{\int P(D|\theta_2, M_2)P(\theta_2|M_2)d\theta_2} \tag{2.2}$$

Le modèle final consiste en un ensemble de paramètres décrivant des positions indépendantes et des positions corrélées mutuellement exclusives . En analysant les données TRANSFAC, les auteurs ont noté que dans 25% des cas (22/95) le modèle GWM était significativement meilleur que le modèle PWM (facteur de Bayes supérieur à 6).

Cette méthode a par la suite été utilisée sur des données ChIP-seq pour 4 TFs mammifères – NRSF, STAT1, CTCF et ER – (Hu et al., 2010). En utilisant les 10% des pics les plus importants comme ensemble d'apprentissage et en se restreignant aux régions de 200bp centrées autour du sommet du pic ChIP, les auteurs ont réalisé un échantillonnage de Gibbs (Casella and George, 1992) pour obtenir les sites de fixation suivant les hypothèses que (1) chaque pic contient au plus un seul site de fixation (modèle ZOOPS pour *Zero or One Occurrences Per*

12. Les corrélations entre des couples de positions (i,j) et (j,k) ne peuvent être admises au sein du même modèle.
13. Sous certaines approximation, ce facteur peut se rapporter à une différence de valeurs du BIC, introduit dans l'article en 2.4.

Sequence), (2) la probabilité *a priori* d'avoir un site à une certaine position sur la séquence est plus forte autour du sommet du pic, et (3) les sites sont décrits par un modèle GWM. L'étude a révélé l'existence de corrélations fortes limitées aux nucléotides plus proches voisins dans les quatre cas étudiés. Les nucléotides participant aux corrélations se situaient à des positions ayant un faible contenu en information dans le modèle PWM. Enfin, les auteurs ont noté la présence de plusieurs triplets de nucléotides voisins corrélés.

2.2.3 Réseaux bayésiens

Une généralisation du modèle GWM consiste à supprimer la condition d'exclusion mutuelle des paires de nucléotides corrélés en décrivant de manière plus générale le réseau de dépendance entre positions. Une telle description est possible en utilisant le langage des réseaux bayésiens. Les dépendances y sont représentées par un graphe orienté acyclique[14] G, dont les nœuds sont les variables X_i et les liens représentent les conditionnements d'une variable avec des variables parentes (fig. 2.1e). La probabilité jointe s'écrit :

$$P(X_1, \cdots, X_k) = \prod_{i=1}^{K} P(X_i | P_i^G) \tag{2.3}$$

où P_i^G est l'ensemble (pouvant être vide) des parents de X_i dans G. Le nombre de paramètres peut rapidement devenir grand : si l'on note N_i le nombre de parents de X_i, alors le nombre de paramètres du modèle est $3 \sum_{i=1}^{K} 4^{N_i}$.

Lorsque les différents nœuds possèdent au plus un parent, on parle d'arbre bayésien (fig. 2.1d). Ce type de réseau bayésien généralise notamment le cas des chaînes de Markov d'ordre 1, où chaque nœud dépend du nœud précédent (fig. 2.1c). Le nombre de paramètres est alors restreint, puisqu'il est au plus de $3 \cdot 4K$.

L'avantage des arbres bayésiens est qu'il existe des algorithmes efficaces permettant de trouver la meilleure structure d'arbre (Friedman et al., 1997). De tels modèles d'arbres ont été utilisés pour décrire les données de 95 TFs de Transfac (Barash et al., 2003). Dans $\sim 25\%$ des cas (22/95), le modèle d'arbre bayésien s'avère significativement meilleur qu'un modèle PWM, ce qui est du même ordre de grandeur que pour le modèle GWM vu en 2.2.2.

14. Un graphe orienté acyclique est un réseau dont les liens sont orientés et au sein duquel il n'est pas possible de revenir à son point de départ en suivant les flèches

2.2.4 Modèles de mélange

Dans les cas précédents, nous avons présenté des modèles capturant des dépendances « locales » entre paires de nucléotides. Néanmoins, il peut exister des dépendances plus largement réparties entre les positions, comme cela a déjà été observé empiriquement (Badis et al., 2009; Jolma et al., 2013). De telles corrélations à plus grande échelle peuvent être modélisées en supposant que le facteur de transcription possède plusieurs « modes » de fixation. Ceux-ci peuvent par exemple correspondre à différentes conformations de la protéine sur son site de fixation, chaque configuration possédant ses propres préférences de fixation. Ces modes sont décrits par une variable aléatoire T (le *type* de fixation) de probabilité $P(T)$. Il est ensuite possible de décrire la fixation au sein de chaque mode par l'un des modèles précédents.

- **Mélange de PWMs**

Le cas le plus naturel consiste à utiliser comme modèle de fixation le modèle PWM, c'est-à-dire que dans chaque mode il y a indépendance entre les positions. La probabilité d'observer un site est alors donnée par la somme sur les différents modes de fixation de la probabilité de fixer un site, conditionnée par la probabilité d'être dans ce mode :

$$P(X_1, \cdots, X_K) = \sum_{T=1}^{N} P(T) \prod_{i=1}^{K} P(X_i|T) \tag{2.4}$$

où N est le nombre de modes de fixation. Ce modèle a plusieurs avantages. D'abord, le nombre de paramètres reste linéaire en K : pour décrire $P(T)$ et les N PWMs il faut $N - 1 + 3KN$ paramètres. Ce nombre reste dont raisonnablement faible devant le nombre de paramètres requis pour complètement décrire les interactions à deux nucléotides, qui croît comme K^2. Ensuite, le modèle a une interprétation claire qui peut permettre de mettre en exergue un mécanisme biologique sous-jacent.

Ce type de modèle permet de dépasser le modèle PWM dans un nombre substantiel de cas. Ainsi, Barash et al. (2003) ont montré que $\sim$ 40% des TFs de Transfac (36/95) sont significativement mieux représentés par un mélange de 2 PWMs que par une seule PWM. En utilisant des données *in vitro* plus précises pour 104 TFs de la souris, Badis et al. (2009) ont montré que $\sim$ 85% (89/104) étaient mieux représenté par une combinaison de PWMs que par une PWM seule, plaidant pour un portée générale de l'existence de « motifs secondaires ».

- **Mélange d'abres**

De la même manière que les PWMs, il est possible d'étendre les modèles d'arbres en réalisant un mélange d'arbres. Intuitivement, ceci permet de capturer des dépendances additionnelles en gardant un nombre de paramètres linéaire en fonction de la taille du motif. Un tel modèle semble posséder des performances comparables au mélange de PWM, et améliore la description des TFs de Transfac dans $\sim$ 40% des cas (35/95) (Barash et al., 2003).

2.3 Modèles de maximum d'entropie

FIGURE 2.2 – **Illustration d'un système dont on veut maximiser l'entropie.**
Le système explore K états $\{s_1, \cdots, s_K\}$ au cours du temps. Au bout de N observations, un état s_i a été observé n_i fois. À chaque état correspond un certain nombre de quantités $\mathcal{O}_\alpha(s_i)$ dont seule la valeur moyenne empirique $\langle \mathcal{O}_\alpha(s) \rangle_r$ est connue de l'observateur (on parle d'« observable »).

2.3.1 Pourquoi maximiser l'entropie ?

Le concept d'entropie remonte aux prémisses de la physique statistique (Jaynes, 1978). Dans l'essence, il peut être compris de la manière suivante. Supposons qu'un système com-

porte K états distincts $\{s_1, \cdots, s_K\}$. Au cours du temps, le système explore les différents états (fig. 2.2). Au bout de N observations, chaque état a été observé un nombre n_i de fois. La question sous-jacente au calcul de l'entropie est la suivante : sans connaissance *a priori* sur le système, que puis-je dire de ces n_i ? Prenons l'exemple de la figure 2.2. On a certaines valeurs pour les n_i ($n_1 = 3$, $n_2 = 1$, etc.), et on aimerait savoir de combien de manières il est possible de réaliser un tel ensemble de valeurs. Notons ce nombre $\mathcal{N}(n_1, \cdots, n_K)$. Il est donné par la formule suivante :

$$\begin{aligned}\mathcal{N}(n_1, \cdots, n_K) &= \binom{N}{n_1}\binom{N-n_1}{n_2}\cdots\binom{N-\sum_{i=1}^{K-1} n_i}{n_K} \\ &= \frac{N!}{(N-n_1)!n_1!} \times \frac{(N-n_1)!}{(N-n_1-n_2)!n_2!} \times \cdots \times \frac{(N-\sum_{i=1}^{K-1} n_i)!}{0!n_K!}\end{aligned} \tag{2.5}$$

soit

$$\mathcal{N}(n_1, \cdots, n_K) = \frac{N!}{n_1!n_2!\cdots n_K!} \tag{2.6}$$

Il convient alors de s'intéresser au logarithme de cette quantité. En effet, dans le cas où les nombres d'observation sont grands $n_i \gg 1$, ceux-ci s'expriment simplement grâce à la formule de Stirling :

$$\log(n!) \xrightarrow[n\to\infty]{} n\log(n) - n \tag{2.7}$$

On peut alors écrire

$$\begin{aligned}\log\mathcal{N}(n_1, \cdots, n_K) &= N\log(N) - N - \sum_{i=1}^{K}(n_i\log(n_i) - n_i) \\ &= \sum_{i=1}^{K} n_i \log(\frac{N}{n_i}) \\ &= -N\sum_{i=1}^{K}\frac{n_i}{N}\log(\frac{n_i}{N})\end{aligned} \tag{2.8}$$

On note l'apparition des probabilités empiriques $f(s_i) = n_i/N$ d'observer l'état s_i, qui tendent asymptotiquement (dans la limite « thermodynamique » $N \to \infty$) vers les « vraies » probabilités $P(s_i)$. L'entropie est définie dans cette limite comme étant égale à $1/N \log\mathcal{N}(n_1, \cdots, n_K)$, soit

$$\boxed{S[P] = -\sum_{\{s\}} P(s) \log P(s)} \tag{2.9}$$

où $\{s\} = \{s_1, \cdots, s_K\}$ dénote l'ensemble des états accessibles. L'idée est alors la suivante : nous souhaitons savoir quels états le système a le plus probablement visité au cours des N transitions. Sans connaissance *a priori* sur le système, il est plus probable que les nombres $(n_1, \cdots, n_K)$ obtenus soient ceux qui sont réalisés le plus souvent, c'est-à-dire ceux qui maximisent la quantité $\mathcal{N}(n_1, \cdots, n_K)$ et donc au final l'entropie. Par ailleurs, les fluctuations relatives des quantités n_i sont de l'ordre de $1/\sqrt{n_i}$ (Sethna, 2006). Ainsi, la solution de maximum d'entropie domine largement les autres solutions possibles dans la limite thermodynamique.

2.3.2 Maximisation de l'entropie sous contraintes

Notons $\mathcal{O}_\alpha(s)$ une quantité attachée à s (fig. 2.2). En thermodynamique, une telle quantité correspond par exemple à l'énergie d'un état. L'observateur n'a lui accès qu'aux valeurs moyennes de telles quantités sous-jacentes. À l'état d'équilibre thermodynamique, l'échantillonnage des états est réalisé au sein de la distribution de probabilité de maximum d'entropie $P(s)$, et les valeurs moyennes calculées avec les fréquences empiriques $f(s)$ doivent donc être compatibles avec les valeurs moyennes calculées avec la distribution de probabilité $P(s)$:

$$\sum_{\{s\}} P(s)\mathcal{O}_\alpha(s) = \sum_{\{s\}} f(s)\mathcal{O}_\alpha(s) \tag{2.10}$$

Nous souhaiterions maintenant connaître la distribution $P(s)$ la moins biaisée (i.e de maximum d'entropie) qui satisfait les contraintes de l'éq. 2.10 imposées par l'observation des données (l'information que possède l'observateur). Ce problème revient à maximiser le Lagrangien suivant :

$$\mathcal{L} = -\sum_{\{s\}} P(s) \log P(s) + \lambda \left(\sum_{\{s\}} P(s) - 1 \right) + \sum_\alpha \beta_\alpha \sum_{\{s\}} \left(P(s) - f(s) \right) \mathcal{O}_\alpha(s) \tag{2.11}$$

où les paramètres λ et β_α sont les multiplicateurs de Lagrange correspondant respectivement à la contrainte de normalisation de la distribution de probabilité et aux informations qu'a l'observateur sur certaines valeurs moyennes du système (éq. 2.10). La maximisation de ce Lagrangien est obtenue en annulant la dérivée fonctionnelle par rapport à la distribution de probabilité $P(s)$:

$$\frac{\delta \mathcal{L}}{\delta P(s)} = 0 = -\ln P(s) - 1 + \lambda + \sum_{\alpha} \beta_\alpha \mathcal{O}_\alpha(s) \tag{2.12}$$

En utilisant la normalisation des probabilités, il est possible de trouver λ, et la solution se met finalement sous la forme

$$\boxed{P(s) = \frac{1}{\mathcal{Z}} e^{-\mathcal{H}(s)}} \tag{2.13}$$

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

$$\mathcal{H} = \sum_{\alpha} \beta_\alpha \mathcal{O}_\alpha(s) \tag{2.14}$$

et $\mathcal{Z}$ est la fonction de partition permettant la normalisation de la distribution $P(s)$:

$$\mathcal{Z} = \sum_{\{s\}} \exp[-\mathcal{H}(s)]. \tag{2.15}$$

Remarque

Il est possible de montrer que la maximisation de l'entropie, partant des contraintes de l'éq. 2.10 sur les valeurs moyennes pour en arriver à une forme exponentielle de la distribution de probabilité, est le contrepoint d'une maximisation de la vraisemblance partant d'une forme exponentielle pour en arriver aux mêmes contraintes sur les valeurs moyennes (Grendar Jr and Grendar, 2001; Jaynes, 1982).

2.3.3 Application aux sites de fixation

- **Corrélations à un point : le modèle PWM**

Dans le cas qui nous intéresse, un état s correspond à une séquence d'ADN appartenant à l'ensemble $\{s\}$ des sites de fixation d'un facteur de transcription. Considérons maintenant l'observable quantifiant la présence du nucléotide a à la position i d'un site :

$$\mathcal{O}_{i,a}(s) = \delta(s_i, a) \tag{2.16}$$

où δ est la fonction de Kronecker qui vaut 1 lorsque le nucléotide à la position i du site s_i vaut a et 0 sinon. De cette définition il suit que la valeur moyenne sur les fréquences empiriques

$$\sum_{\{s\}} f(s)\mathcal{O}_{i,a}(s) = f_{i,a} \tag{2.17}$$

se réduit à la fréquence du nucléotide a à la position i. Notons $h_i(a)$ le multiplicateur de Lagrange correspondant et $\mathcal{A} = \{A, C, G, T\}$. On trouve alors

$$\begin{aligned}\mathcal{H}(s) &= \sum_{i=1}^{L} \sum_{a\in\mathcal{A}} h_i(a)\delta(s_i, a) \\ &= \sum_{i=1}^{L} h_i(s_i)\end{aligned} \tag{2.18}$$

Les différentes positions étant indépendantes, la fonction de partition $\mathcal{Z}$ peut par ailleurs se scinder en différentes fonctions de partition par position : $\mathcal{Z} = \prod_{i=1}^{L} \mathcal{Z}_i$. On obtient au final

$$P(s) = \frac{1}{\mathcal{Z}} e^{-\sum_{i=1}^{L} h_i(s_i)} = \prod_{i=1}^{L} \frac{e^{-h_i(s_i)}}{\mathcal{Z}_i} \tag{2.19}$$

On retrouve le modèle PWM introduit dans l'éq. 2.13.

- **Corrélations à deux points : le modèle de Potts**

Il est maintenant relativement direct de complexifier le modèle en ajoutant l'observation des couples d'interaction au sein des sites de fixation :

$$\mathcal{O}_{i,a,j,b}(s) = \delta(s_i, a)\delta(s_j, b) \tag{2.20}$$

La corrélation à deux points entre le nucléotide a en position i et b en position j s'écrit donc

$$\sum_{\{s\}} f(s)\mathcal{O}_{i,a,j,b}(s) = f_{i,a,j,b} \tag{2.21}$$

où $f_{i,a,j,b}$ est la fréquence empirique d'observation de la paire de nucléotide (a, b) aux positions (i, j). Notons $J_{i,j}(a, b)$ le multiplicateur de Lagrange correspondant. L'Hamiltonien sous les contraintes imposées par les équations 2.17 et 2.21 s'écrit :

$$\begin{aligned}\mathcal{H}(s) &= \sum_{i=1}^{L} \sum_{a\in\mathcal{A}} h_i(a)\delta(s_i, a) + \sum_{i=1}^{L-1} \sum_{j>i} \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{A}} J_{i,j}(a, b)\delta(s_i, a)\delta(s_j, b) \\ &= \sum_{i=1}^{L} h_i(s_i) + \sum_{i=1}^{L-1} \sum_{j>i} J_{i,j}(s_i, s_j)\end{aligned} \tag{2.22}$$

Le modèle de maximum d'entropie est finalement

$$\boxed{P(s) = \frac{1}{\mathcal{Z}} e^{-\sum_{i=1}^{L} h_i(s_i) - \sum_{i=1}^{L-1} \sum_{j>i} J_{i,j}(s_i, s_j)}} \tag{2.23}$$

On reconnaît le modèle de Potts inhomogène de champs magnétiques locaux h_i et de termes d'interaction $J_{i,j}$ couramment utilisé dans la description des verres de spins (Baxter, 2007).

2.4 Article

L'article qui suit décrit l'analyse de données de fixation *in vivo* à grande échelle pour plusieurs TFs drosophiles et mammifères. Différents modèles sont comparés, incluant ou non des dépendances : un modèle PWM, un modèle de mélange de PWMs, et un modèle de Potts.

Beyond position weight matrices: nucleotide correlations in transcription factor binding sites and their description

Marc Santolini, Thierry Mora, and Vincent Hakim
Laboratoire de Physique Statistique, CNRS, Université P. et M. Curie, Université D. Diderot, École Normale Supérieure, Paris, France.

The identification of transcription factor binding sites (TFBSs) on genomic DNA is of crucial importance for understanding and predicting regulatory elements in gene networks. TFBS motifs are commonly described by Position Weight Matrices (PWMs), in which each DNA base pair independently contributes to the transcription factor (TF) binding, despite mounting evidence of interdependence between base pairs positions. The recent availability of genome-wide data on TF-bound DNA regions offers the possibility to revisit this question in detail for TF binding *in vivo*. Here, we use available fly and mouse ChIPseq data, and show that the independent model generally does not reproduce the observed statistics of TFBS, generalizing previous observations. We further show that TFBS description and predictability can be systematically improved by taking into account pairwise correlations in the TFBS via the principle of maximum entropy. The resulting pairwise interaction model is formally equivalent to the disordered Potts models of statistical mechanics and it generalizes previous approaches to interdependent positions. Its structure allows for co-variation of two or more base pairs, as well as secondary motifs. Although models consisting of mixtures of PWMs also have this last feature, we show that pairwise interaction models outperform them. The significant pairwise interactions are found to be sparse and found dominantly between consecutive base pairs. Finally, the use of a pairwise interaction model for the identification of TFBSs is shown to give significantly different predictions than a model based on independent positions.

Author Summary

Transcription factors are proteins that bind on DNA to regulate several processes such as gene transcription or epigenetic modifications. Being able to predict the Transcription Factor Binding Sites (TFBSs) with accuracy on a genome-wide scale is one of the challenges of modern biology, as it allows for the bottom-up reconstruction of the gene regulatory networks. The description of the TFBSs has been to date mostly limited to a simple model, where the affinity of the protein for DNA, or binding energy, is the sum of independent contributions from uncorrelated amino-acids bound on base pairs. However, structural aspects are of prime importance in proteins and could imply appreciable correlations throughout the observed binding sequences. Using a statistical physics inspired description and high-throughput ChIPseq data for a variety of Drosophilae and mammals TFs, we show that such correlations exist and that accounting for their contribution greatly improves the predictability of genomic TFBSs.

Introduction

Gene regulatory networks are at the basis of our understanding of a cell state and of the dynamics of its response to environmental cues. Central effectors of this regulation are Transcription Factors (TF) that bind on short DNA regulatory sequences and interact with the transcription apparatus or with histone-modifying proteins to alter target gene expressions [1]. The determination of Transcription Factor Binding Sites (TFBSs) on a genome-wide scale is thus of importance and is the focus of many current experiments [2]. An important feature of TF in eukaryotes is that their binding specificity is moderate and that a given TF is found to bind a variety of different sequences *in vivo* [3]. The collection of binding sequences for a TF-DNA is widely described by a Position Weight Matrix (PWM) which simply gives the probability that a particular base pair stands at a given position in the TFBS. The PWM provides a full statistical description of the TFBS collection when there are no correlations between nucleotides at different positions. Provided that the TF concentration is far from saturation, the PWM description applies exactly at thermodynamic equilibrium in the simple case where the different nucleotides in the TFBS contribute independently to the TF-DNA interaction, such that the total binding energy is the mere sum of the individual contributions [4, 5].

Previous works have reported several cases of correlations between nucleotides at different positions in TFBSs [6–9]. A systematic *in vitro* study of 104 TFs using DNA microarrays revealed a rich picture of binding patterns [10], including the existence of multiple motifs, strong nucleotide position interdependence, and variable spacer motifs, where two small determining regions of the binding site are separated by a variable number of base pairs. Recently, the specificity of several hundred human and mouse DNA-binding domains was investigated using high-throughput SELEX. Correlations between nucleotides were found to be widespread among TFBSs and predominantly located between adjacent flanking bases in the TFBS [9]. The relevance of nucleotide correlations remains however debated [11].

On the modeling side, probabilistic models have been proposed to describe these correlations, either by explicitly identifying mutually exclusive groups of co-varying

nucleotide positions [7, 12, 13], or by assuming a specific and tractable probabilistic structure such as Bayesian networks or Markov chains [9, 14, 15]. However, the extent of nucleotide correlations in TFBSs *in vivo* remains to be assessed, and a systematic and general framework that accounts for the the rich landscape of observed TF binding behaviours is yet to be applied in this context. The recent breakthrough in the experimental acquisition of precise, genome-wide TF-bound DNA regions with the ChIPseq technology offers the opportunity to address these two important issues. Using a variety of ChIPseq experiments coming both from fly and mouse, we first show that the independent model generally does not reproduce well the observed TFBS statistics for a majority of TF. This calls for a refinement of the PWM description that accounts for interdependence between nucleotide positions.

The general problem of devising interaction parameters from observed state frequencies has been recently studied in different contexts where large amounts of data have become available. These include describing the probability of coinciding spikes [16, 17] or activation sequences [18, 19] in neural data, the statistics of protein sequences [20, 21], and even the flight directions of birds in large flocks [22]. Maximum-entropy models accounting for pairwise correlations in the least constrained way have been found to provide significant improvement over independent models. The PWM description of TF binding is equivalent to the maximum entropy solely constrained by nucleotide frequencies at each position. Thus, we propose, in the present paper, to refine this model by further constraining pairwise correlations between nucleotide positions. This corresponds to including effective pairwise interactions between nucleotides in an equilibrium thermodynamic model of TF-DNA interaction, as already proposed [23]. When enough data are available, the TFBS statistics and predictability are found to be significantly improved in this refined model. We consider, for comparison, a model that describes the statistics of TFBSs as a statistical mixture of PWMs [14] and generalizes previous proposals [24, 25]. This alternative model can directly capture some higher-order correlations between nucleotides but is found to be outperformed for all considered TF by the pairwise interaction model.

We further show that the pairwise interaction model accounts for the different PWMs appearing in the mixture model by studying its energy landscape: each basin of attraction of a metastable energy minimum in the pairwise interaction model is generally dominantly described by one PWM in the mixture model. Significant pairwise interactions between nucleotides are sparse and found dominantly between consecutive nucleotides, in general qualitative agreement with *in vitro* binding results [9]. The proposed model with pairwise interactions only requires a modest computational effort. When enough data are available, it should thus generally prove worth using the refined description of TFBS that it affords.

Results

The PWM model does not reproduce the TFBS statistics

We first tested how well the usual PWM model reproduced the observed TFBS statistics, *i.e.* how well the frequencies of different TFBSs were retrieved by using only single nucleotide frequencies. For this purpose, we used a collection of ChIPseq data available from the literature [26–28], both from *D. Melanogaster* and from mouse embryonic stem cells (ESC) and a myogenic cell line (C2C12). The TFBSs are short L-mers (we take here $L = 12$), which are determined in each few hundred nucleotides long ChIP-bound region with the help of a model of TF binding. One important consequence and specific features of these data, is that the TFBS collection is not independent of the model used to describe it. Thus, in order to self-consistently determine the collection of binding sites for a given TF from a collection of ChIPseq sequences, we iteratively refined the PWM together with the collection of TFBSs in the ChIPseq data (see Figure and *Methods*). This process ensured that the frequency of different nucleotides at a given position in the considered ensemble of binding sites was exactly accounted by the PWM. We then enquired whether the probability of the different binding sequences in the collection agreed with that predicted by the PWM, as would be the case if the probabilities of observing nucleotides at different positions were independent. Figure 2 displays the results for three different TFs, one from each of the three considered categories: Twi (Drosophila), Esrrb (mammals, ESC), and MyoD (mammals, C2C12). For each factor, the ten most frequent sequences in the TFBS collection are shown. For comparison, Figure 2 also displays the probabilities for these sequences as predicted by the PWM built from the TFBS collection. The independent PWM model strongly underestimates the probabilities of the most frequent sequences. Moreover, the PWM model does not correctly predict the frequency order of the sequences and attributes comparable probabilities to these different sequences, in contrast to their observed frequencies.

The relative entropy or Kullback-Leibler divergence (DKL) is a general way to measure the difference between two probability distributions [29]. In order to better quantify the differences between the observed binding sequence frequencies and the PWM frequencies, we computed the DKL between these distributions for all the considered TF, as shown in Figure 2D. For each transcription factor T, part of the differences comes from the finite number N(T) of its observed binding sites. The results are thus compared for each factor T to DKLs between the PWM probabilities and frequencies obtained for artificial sequence samples of size N(T) generated with the same PWM probabilities. For most TFs (22 out of 28), the difference between the observed binding sequence frequencies and the PWM frequencies is signifi-

FIG. 1: **Workflow.** An initial Position Weight Matrix (PWM) is used to find a set of binding sites on ChIPseq data. Models are then learned using single-point frequencies (independent), two-point correlations (pairwise) or a mixture of independent models learned on sites clustered by K-Means (mixture) with increasing complexity, *i.e.* increasing number of features in the model. Finally the models with best Bayesian Information Criteria (BIC) are used to predict new sites until convergence to a stable set of sites.

cantly larger than expected from finite size sampling. In the following we focus on these 22 factors for which the PWM description of the TFBSs needs to be be refined. It can be noted that the 6 factors for which the PWM description appears satisfactory are predominantly those for which the smallest number of ChIP sequences is available (see Table 1 and Figure S1).

Pairwise interactions in the binding energy improve the TFBS description

The discrepancy between the observed statistics of TFBSs and the statistics predicted by the PWM model calls for a re-evaluation of the PWM main hypothesis, namely the independence of bound nucleotides. As recalled above, the inverse problem of devising interaction parameters from observed frequencies of "words" has been recently studied in different contexts. It has been proposed to include systematically pairwise correlations between the "letters" comprising the words to refine the independent letter description. In the case of a two-letter alphabet, the obtained model is equivalent to the classical Ising model of statistical mechanics[30]. In the present case, the 4-nucleotide alphabet (A,C,G,T) leads to a model equivalent to the so-called inhomogeneous Potts model [30] (hereafter called pairwise interaction model), a generalization of the Ising model to the case where spins assume q values and their fields and interaction parameters depend on the sites considered. In this analogy, nucleotides are spins with $q = 4$ colors.

In practice, the probability of observing a given word $(s_1...s_L)$ in the dataset is expressed as $P[s_1...s_L] = (1/\mathcal{Z})\exp(-\mathcal{H}[s_1...s_L])$, where $\mathcal{Z}$ is a normalization constant. $\mathcal{H}$ is formally equivalent to a Hamiltonian in the language of statistical mechanics, and reads:

$$\mathcal{H}[s_1...s_L] = -\sum_{i=1}^{L} h_i(s_i) - \sum_{i=1}^{L}\sum_{j<i} J_{i,j}(s_i,s_j), \quad s_i \in \{A,C,G,T\} \tag{1}$$

The "magnetic fields" h_i at each site i, along with the interaction parameters J_{ij} between nucleotides at positions i and j, are computed so as to reproduce the frequency of nucleotide usage at each position in the TFBS as well as the pairwise correlations between nucleotides at different positions (see *Methods*). In principle, the number of parameters in the model is sufficient to reproduce the observed values of all pairwise correlations between nucleotides. This however would result in over-fitting the finite-size data with an irrealistically large number of parameters. Therefore, to obtain the model parameters we instead maximized the likelihood that the data was generated by the model with a penalty proportional to the numbers of parameters involved, as provided by the Bayesian Information Criterion (BIC) [31]. Similarly to the procedure followed for the PWM, the pairwise interaction model and the collection of TFBSs for a given factor were iteratively refined together, as schematized in Figure .

Figure 3 shows the improvement in the description of TFBS statistics when using the final pairwise interaction model, for the three factors chosen for illustrative

FIG. 2: **Observed TFBS frequencies are poorly predicted by a PWM model.** The observed frequencies of the most represented binding site sequences for the TF Twist (A), Esrrb (B) and MyoD (C) are shown (gray bars) as well as the probabilities of these sequences as predicted by the PWM model (blue bars). (D) Kullback-Leibler Divergence (DKL) between the observed probability distribution and the independent model distribution (blue). As a control we show the mean (cyan bars) along with two standard deviations of the DKL between the independent model and a finite sample drawn from it (see Methods). A discrepancy between the observed and predicted sequence probabilities is reported for 22 out of 28 factors.

purposes. Where the independent model failed at reproducing the strong amplitude and non-linear decrease in the frequencies of the most over-represented TFBSs, the pairwise interaction model provides a substantial improvement in reproducing the observed statistics. The improvement is most apparent when comparing the frequencies of the ten most observed TFBSs between the model and the ChIPseq data (Figure 3 A, C, E), and is further shown by the statistics of the full collection of TFBSs (Figure 3 B, D, F).

The pairwise model ranks binding sites differently from the PWM

Precise predictions of TFBSs are one important output of ChIPseq data. Moreover, they condition further validation experiments such as gel mobility shift assays or mutageneses. We therefore found it worth assessing the difference in TFBS predictions between pairwise and independent models.

First, we compared the set of ChIP sequences retrieved by the independent and pairwise models model at the cutoff of 50% TPR (True Positive Rate) used in the learning scheme, as shown in Figure 4A. The non overlapping set of ChIPseq sequences (*i.e.* sequences that were picked by one model but not by the other) was found to range

FIG. 3: **Models with correlations improve TFBS statistics prediction.** The observed frequencies (gray bars) of the most represented TFBSs for Twist (A), Esrrb (B) and MyoD (C) TFs, are shown together with the probabilities of these sequences predicted by the independent energy model (blue bars), the pairwise model taking into account interactions between nucleotides (red bars), and the K-means mixture model (green bars). (B,D,F) show the comparison between frequencies for all binding sequences and predicted sequence probabilities for the three models (same color code). The probability predictions of the pairwise model and to a lesser extent of the mixture model are in much better agreement with the observed frequencies than those of the PWM model.

from a few percent for TF like Esrrb, up to about 15 % for Twist. Thus, even when stemming from the same ChIPseq data, the two models can be learnt from significantly distinct set of sites.

Second, using the set of ChIPseq peaks on which the pairwise model was learned, we looked for the best predicted sites on each ChIPseq bound fragment using both the pairwise and PWM models (Figure 4B).

The overlap was found to be about 80% on average. The overlap between the sets comprising the two best TFBSs of each ChIPseq was also computed. This resulted in an overlap increase or decrease between the

FIG. 4: **Overlap between predicted sites.** (A) Venn diagrams showing the overlap between the ChIP predicted by the independent (blue) and pairwise (red) models. (B) Difference (one minus the proportion of shared sites) between the best sites predicted by pairwise and PWM models on ChIPseq peaks (light red), and the same quantity when including the next best predicted sites on each peak (dark red). In several cases (*e.g.* Fosl1, Max, n-Myc, Srf, Stat3, Usf1), the difference between predicted sites is much smaller when the two best sites are considered, indicating that the pairwise model and the PWM model rank differently the two best sites in ChIP peaks with multiple bound sites.

prediction of the two models depending on the average of number of binding sites per retrieved ChIPseq fragment. In a few cases (*e.g* CTCF, Esrrb), the inclusion of the second best TFBS increased the difference between the two models. This generally happened when the ChIPseq fragments were retrieved with typically a single TFBS above threshold (*e.g.* for Essrb the TFBS specificity was fixed to retrieve 50% of 18453 ChIPseq and about 11000 fragments where found by the two models—see Table I). In these cases, the low specificity TFBSs tended to differ more between the two models than the very specific ones. In several other cases (*e.g* for Fosl1, Max, n-Myc, USF1), the inclusion of the second best predicted binding sites (Figure 4B) greatly increased the overlap between the two model predictions. This corresponded to cases for which the retrieved fragments contained on average two of more TFBSs about the specificity threshold (Table I). This showed that for these cases the prediction difference between the two models arose predominantly from a different ranking of the best TFBSs.

In conclusion, the TFBS predictions made by the two models can differ significantly both in the rank of ChIPseq fragments and in the rank of binding sites on these fragments.

Comparison with a PWM-mixture model

When described by a PWM, the binding energies of a TF for different nucleotides sequences form a simple energy well with a single minimum at a preferred consensus sequence. Some authors have instead analyzed the binding specificity of transcription factors by introducing multiple preferred sequences [24, 25]. A model of this type that naturally generalizes the PWM description consists of using multiple PWMs [14]. We found it interesting to investigate this approach based on a mixture of PWMs and compare it with the pairwise interaction model to get some insights into potentially important high-order correlations that would not be captured by the pairwise model. As precisely described in *Methods*, an initial mixture of K PWMs was generated by grouping into K clusters the TFBS data for a given TF. Similarly to the pairwise interactions, the number of clusters K was constrained, to avoid over-fitting, by penalizing the corresponding model score using the BIC. For a given TF, the PWM mixture and the collection of TFBSs in the ChIPSeq data were refined iteratively until convergence, usually reached after 10 iterations. The results are shown in Figure 5A for the three representative factors,

FIG. 5: **Model selection.** (A) Minimisation of the Bayesian information criterion (BIC, see *Methods*) is used to select the optimal number of model parameters and avoid over-fitting the training set. The evolution of the BIC is shown for the pairwise model (crosses) and the PWM-mixture model (lines). Colors from dark blue to red indicate the number of interations (see Fig.).
(B) Kullback-Leibler divergences (DKL) between the independent, K-means and pairwise distributions and the observed distribution for the different TFs, for the BIC optimal parameters. We also show the DKL of the pairwise model with a finite-size sample of sequences drawn from it (pink, see *Methods*). Error bars represent two standard deviations.

Twi, Esrrb and MyoD.

The best description of Twi ChIPSeq data is, for instance, provided by a mixture of 5 PWMs, which corresponds to 184 independent parameters. The mixture model yields a significant improvement when compared to the single-PWM model for Twi, and milder ones for Essrb and MyoD. In the three cases however, it proves inferior to the pairwise model.

More generally, Figure 5B shows the performances of the different models for all studied TFs using the Kullback-Leibler Divergence or DKL between the data distribution $P(s)$ and the models distributions $P_m(s)$. On the one hand, the mixture model improves the description of the binding data for 12 out of 27 TFs as compared to the single PWM model. The mixture model gives in particular strong improvements in the cases for which the binding sites have a palindromic structure (eg Twi, MyoD, Myog, Max, USF1). This feature often stems from the fact that the TF binds DNA as a dimer, which could give some concreteness to the mixture model: the recruitment of different partners by bHLH factors like MyoD or Myog could indeed lead to a mixture of TFs binding the same sites. On the other hand, the pairwise model clearly outperforms the other models in all cases studied.

As in the PWM case, the finite size of the datasets

leads us to expect fluctuations in the estimation of the DKL. In order to assess the magnitude of these finite-size fluctuations, we computed the average DKL between the best-fitting (pairwise) model and a finite-size artificial sample drawn from its own distribution, as shown in Figure 5B. Values of this DKL that are larger than the one obtained with the real dataset are indicative of overfitting, while the opposite case would suggest that the model is incomplete. In all cases, however, the DKL obtained with this control procedure was within error bars of the value computed with respect to the observed sample, with the exception of NRSF, MyoD, and Myog, as seen in Figure 5B. Thus, the pairwise model is generally the best possible model, insofar as the available dataset allows us to probe.

The metastable states of the pairwise interaction model

In order to more directly relate the pairwise interaction and the mixture models, it is useful to consider the energy landscape of the pairwise interaction model in the space of all possible TFBSs. By contrast with the simple, single-minimum energy well of the PWM model, the pairwise interaction model has multiple metastable energy minima. The energy landscape of the pairwise interaction model can thus be seen as a collection of energy wells, each centered on its metastable energy minimum. The span of the different energy wells in sequence space can be precisely defined as the basins of attraction of the different metastable minima in an energy minimizing procedure (see *Methods*). This allows one to associate each observed TFBS to a particular energy minimum. This defines basins of attraction that are used to build representative PWMs for each metastable minimum together with a weight—the number of sequences in the basin of attraction—for this energy minimum. We compared each metastable minimum to the PWMs of the mixture model, by calculating the DKL between the PWM computed from the sequences in its basin of attraction and the PWMs of the mixture model. This gave an effective distance which allowed us to associate each metastable state to the nearest PWM of the mixture model.

Using this procedure, we computed the set of PWMs and weights corresponding to the 27 considered TF pairwise interaction models. Examples are shown in Figure 6. In the case of Twi, the PWMs of the pairwise model ("metastable PWMs") can be clearly associated to the $K = 5$ PWMs of the mixture model. For MyoD, three of the 5 "metastable PWMs" can be clearly assigned to PWMs of the mixture model. The other two have a more spread out representation. The case of Esrrb is similar with one "metastable PWM" in clear correspondence with one PWM of the mixture model, and the other one less clearly so. The correspondence between the two models is shown in Figure S2 for the other TFs for which the mixture model uses more than a single PWM. This representation allows one to identify some features captured by the pairwise model. For example, in the case of Twist, most of the correlations are coming from the two nucleotides at the center of the motif, which take mainly 3 values among the 16 possible: CA,TG and TA. In the case of MyoD, the representation makes apparent the interdependencies between the two nucleotides following the core E-Box motif, and the restriction to the three main cases of CT, TC and TT.

Properties of the pairwise interactions

The computation of the interaction parameters allows an analysis of some of their properties. In particular, it is interesting to quantify their strengths and measure the typical distance between interacting nucleotides. We address these two questions in turn.

The concept of Direct Information was previously introduced to predict contacts between residues from large-scale correlation data in protein families [33]. We used it here to measure the strength of the pairwise interaction between two nucleotides. Using the previously generated interaction parameters from the pairwise model, we built the Normalized Direct Information (NDI), a quantity which varies from 0 for non-existing interactions, to 1 when interactions are so strong that knowing the amino acid identity at one position entirely determines the amino acid identity at the other position (see *Methods*). Heatmaps displaying the results for the representative Twist, Esrrb and MyoD factors are shown in Figure 7 and in Figures S3 for the other factors. An important observation is that the direct information between different nucleotides is rather weak—usually smaller than 10%—but substantially larger than the direct interaction between nucleotides in the surrounding background (1-3%, see Figure S4). It is interesting to note that such weak pairwise interactions give rise to a substantial improvement in the description of TFBS statistics, similarly to what was previously found in a different context [16]. The pairwise interactions are furthermore observed in Figure 7 to be concentrated on a small subset of all possible interactions. This can be made quantitative by computing the Participation Ratio of the interaction weights, an indicator of the fraction of pairwise interactions that accounts for the observed Direct Information (see Methods). Here, typical values of $10 - 20\%$ were found (Figure 7 and Table I), showing that the interactions tend to be concentrated on a few nucleotide pairs.

The interaction weights can also be used to measure the typical distance between interacting nucleotides. To that purpose, we computed the relative weight of the Direct Information as a function of the distance between nucleotides (see Methods). Figure 8 A shows box plots that summarize the results for the considered Drosophilae and mammalian TFs. Both plots show a clear bias towards nearest-neighbor interactions with a strong peak at $d = 1$, and a rapid decrease for $d \geq 2$. Finally, the

FIG. 6: **Metastable states.** The DNA sequence variety described by each model is illustrated using weblogos [32]. Shown are PWMs built from all sites, from the PWM-mixture model, and from the basins of attraction of the pairwise interaction model for Twist (A), Esrrb (B), and MyoD (C). The metastable PWMs are grouped under the mixture PWMs with smallest distance (measured by DKL, in bits). Heatmaps showing the DKLs between metastable PWMs and mixture PWMs are displayed on the right for each factor (minimal DKLs are in black). The proportions of sites used for each logo are also indicated and serve to denote the corresponding PWM.

dominant pair interactions are on average located in the flanking regions of the BS in clear anti-correlation with the most informative nucleotides which are on average in the central region (Figure 8 B). These observations for TF binding *in vivo* agree with similar ones made from a large recent analysis of TF binding *in vitro* [9]. The fact that for pair correlations to be important, nucleotide variation at a given location is required, may be one way to rationalize them.

FIG. 7: **Nucleotide pair interactions.** Heat maps showing the values of the Normalized Direct Information between pairs of nucleotides. The matrix is symmetric by definition. PWMs are shown on the side for better visualization of the interacting nucleotides. The participation ratio R is indicated below each heat map.

TABLE I: **Participation Ratios**

Name	Part. Ratio
Bin	0.11
Mef2	0.19
Twi	0.28
E2f1	0.13
Esrrb	0.11
Klf4	0.16
Nanog	0.10
n-Myc	0.09
Oct4	0.24
Sox2	0.12
Tcfcp2l1	0.12
Zfx	0.10
CEBPB	0.05
CTCF	0.23
E2f4	0.14
Fosl1	0.09
Max	0.18
MyoD	0.09
Myog	0.09
NRSF	0.27
TCF3	0.19
USF1	0.07

Alternative representation of interactions by Hopfield patterns

Using a simple binary description of neurons, JJ Hopfield suggested, in a classic piece of work [34], that neural memories could be attractors corresponding to patterns arising from pair interactions between neurons. These interaction patterns can be computed in the present case. They offer an alternative way to analyze the patterns of correlation from the pair-interactions between positions, as already proposed in a mean-field context in [35]. Because the matrix of interactions J_{ij} is symmetric, it can be diagonalized in an orthonormal basis of eigenvectors ξ^k, the Hopfield patterns in the present case, with corresponding real eigenvalues λ_k. These orthonormal eigenvectors correspond to the Hopfield patterns in the present case. The Potts energy (Eq. (1)) for a binding sequence $s_1 \cdots s_L$ can be rewritten in terms of the Hopfield patterns as (see Methods):

$$\mathcal{H} = -\sum_i h_i(s_i) - \frac{1}{2}\sum_{k=1}^{4L} \lambda_k \left(\sum_{i=1}^{L} \xi_i^k(s_i) \right)^2 . \qquad (2)$$

Although here the presence of the diagonal h term prevents the patterns to be metastable energy states, they can still be useful to analyze the interaction matrix. This spectral decomposition of the interaction matrix is also similar in spirit to a principal component analysis, and even equivalent in the case of Gaussian variable. One can thus wonder how many patterns are needed to well approximate the full matrix of interactions J. To address this question, one can rank the eigenvalues λ_k in order of decreasing moduli and note J_p the restriction of the interaction matrix generated by the first p eigenvalues and their associated patterns. The full interaction matrix naturally corresponds to J_{48}. Approximate interaction matrices obtained by keeping different numbers of dominant patterns are shown in Figure 9 for the three considered representative factors. Pairs of successive patterns appear to provide the main interaction domains in this representation, as is particularly clear in the case of MyoD. One can see in Figure 9 that J_6 already closely approximates the full interaction matrix, a reflection in the present representation, that the important interactions are concentrated on a few links between pairs of nucleotides.

FIG. 8: **Properties of the pair interactions.** (A) Distances between interacting nucleotides. The box plots show the relative importance of the Normalized Direct Information as a function of the distance between interacting nucleotides. Red dots denote average values. (B) Sum of normalized direct informations in the TFBSs at a given position, averaged over all considered factors (blue line). The average site information content relative to background as a function of position is also shown (red line). In both quantities, the average over the two TFBS orientations has been taken.

Discussion

The availability of ChIPseq data for many TFs has led us to revisit the question of nucleotide correlations in TFBSs. In order to perform a fully consistent analysis of this type of data, we have developed a workflow in which the TFBS collection and the model describing them are simultaneously obtained and refined together. We have found that when a sufficiently large number of TFBSs is available, the PWM description does not account well for their statistics. The general presence of correlations that follows from this finding, agrees with previous reports for particular transcription factors [6, 8, 24] and with the conclusions of large scale *in vitro* TF binding studies [9, 10].

In order to refine the PWM description, we have analyzed a model with pairwise interactions [23], and a PWM mixture model [14]. Data overfitting is a concern for multi-parameter models and has been addressed by putting a penalty on the parameter number using the BIC. While the mixture-model improved in some cases the PWM description, especially for palindromic binding sites, a much more significant and general improvement was found with the pairwise interaction model. The success of the pairwise interaction model agrees with the results of its recent application (however, without the BIC) to high-throughput *in vitro* binding data [23]. It moreover shows that, at least in the case we considered, pairwise interactions are sufficient to account for higher-order correlations, and that an explicit description like the one provided by the PWM-mixture model is not necessary. For example, for Essrb, metastable states arising from nearest-neighbor interactions reproduce a triplet of flanking nucleotides with a variable spacer from the core motif (Figure S5).

Our detailed analysis of the obtained interaction mod-

FIG. 9: **Representation of interactions by Hopfield patterns.** The full interaction matrix J is approximated by a matrix J_p built from the p Hopfield patterns with highest eigenvalue moduli. We show J_2, J_4, J_6 and the full matrix J in the basis (i, b) with $i = \{1, ..., 12\}$ and $b =$\{A,C,G,T\}. Color bars are shown on the first row for each factor. For MyoD, the correspondence between successive pairs of patterns and distinct interaction domains is seen particularly clearly. In all cases the full interaction matrix is already well approximated by J_6.

els for different TFs shows that the weights of pairwise interactions are generally weak. The most important are only about 10 % of the PWM weights, but significantly above the interaction weights in the surrounding background DNA (of the order 1-3% by the same measure). Nonetheless, collectively these interactions significantly improve the model description of the TF binding data as found in other examples [16].

We have here obtained the pairwise interaction models based on the principle of maximum entropy, constrained to account for the pair-correlations measured in the data. This approach has already been followed in a variety of biological contexts, from populations of spiking neurons [16, 17] to protein sequences [20] to bird flocks [22]. An interesting feature of these interaction models is their non-convexity, which allows for the existence of many lo-

cal maxima in the probability distribution of sequences, or local minima of energy. This was noted for repertoires of antibodies in a single individual [21], where many of these local states were observed and suggested as possible signatures of past infections. In a very different context, local probability maxima in the probability distribution of retinal spiking patterns was reported and linked to error-correcting properties of the visual system [36]. In the present case of TFBSs, these local mimima reflect the multiplicity of binding solutions and resemble the individual PWMs of the mixture model. Pairwise interaction models thus somehow incorporate models of multiple PWMs while outperforming them.

The previously considered case of protein sequences shares many similarities to the statistics of TFBSs, since correlations in protein sequences as in TFBSs reflect both structural and functional contraints. In proteins families, correlations are usually interpreted as resulting from the co-evolution of residues interacting with each other in the protein structure. These effects are hard to distinguish from phylogenic correlations or other observational biases. Nonetheless, the inference of interaction models from data was successfully used to predict physical contacts between amino-acids in the tertiary structure [37], and to aid molecular dynamics simulations in predicting protein structure [38–40]. In the case of TFBSs, comparison between *in vitro*[9, 10] and *in vivo* binding data may help to disentangle the different possible origins of the found correlations and seems worth pursuing. It appears similarly interesting to study how much of the found pair correlations can be explained on the basis of structural data. Finally, the role of nucleotide interaction in TFBS evolution [41] should be considered and could improve the reconstruction of TFBSs from multi-species comparison [42–44].

Independently of these future prospects, we have found that the TFBSs predicted from ChIP-seq data significantly depended on the model used to extract them. Since the pairwise interaction model and the developed workflow significantly improve TFBS description and require a modest computational effort, they should prove worthy tools in future data analyses.

Materials and Methods

Genome-wide data retrieval

We use both ChIP-on-chip data from *Drosophila Melanogaster* and ChIPseq data from *Mus Musculus*. Data was retrieved from the litterature [26, 27] and from ENCODE data accessible through the UCSC website `http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeCaltechTfbs/`, for a total of 27 TFs. Among them, there are 5 developmental Drosophilae TFs: Bap, Bin, Mef2, Tin and Twi, 11 mammalian stem cells TFs: c-Myc, E2f1, Esrrb, Klf4, Nanog, n-Myc, Oct4, Sox2, Stat3, Tcfcp2l1, Zfx, and 11 factors involved in mammalian myogenesis: Cebpb, E2f4, Fosl1, Max, MyoD, Myog, Nrsf, Smad1, Srf, Tcf3, Usf1. Overall, there are between 678 and 38292 ChIP peaks, with average size 280bp. DNA sequences were masked for repeats using RepeatMasker [45].

Background models

It is important to discriminate the statistics of the motifs proper from that of the background DNA on which motifs are found. Besides particular nucleotides frequencies, the background DNA can exhibit significant nucleotide correlations, for instance arising from CpG depletion in mammalian genomes (Figure S4). For each ChIPseq data, we used, as background, all sites from both strands of the sequences. This serves to learn independent and pairwise background models which were used as reference models to score the corresponding TFBS models. The position information content in all plotted PWM logos is measured with respect to the nucleotide background frequencies (*i.e.* the independent background model)

Initial PWM refinement

Along with the ChIPseq data for the different factors, we also retrieved corresponding PWMs from the literature [26] or from TRANSFAC database [46]. These initial PWMs were refined according to the following protocol.

Given ChIPseq data (bound regions) for a given TF and an initial PWM of length L ($L = 12$ was taken for all computations in the present paper), we scanned both strands of each bound region and attributed to all observed L-mers a score defined as the ratio between the PWM and background models probabilities. A cutoff was set such that half of the bound regions had at least one predicted TFBS with a score above the cutoff, setting a True Positive Rate (TPR) of 50%. This heuristic criterium overcame the problem of False Positives among the ChIPseq peaks that might have polluted the data. This defined a training set of N L-mers with probability higher than the cutoff. Bound sites were again predicted using the same cutoff. This procedure was repeated until stabilization of the predicted sites to a fixed subset. This resulted in a refined PWM with its associated set of bound sites.

Independent model evaluation

The independent model consist of a matrix of single nucleotide probabilities of size $4 \times L$, where L is the width of the binding site. In a first approximation, the parameters appearing in the matrix can be estimated from a set of binding sites by computing the observed frequency $f_{b,i}$ of

TABLE II: **Information about the TFs**

Name	N_{chip}	$\Delta_{\text{inde-mixture}}$	$\Delta_{\text{inde-pairwise}}$	$\Delta_{\text{mixture-pairwise}}$	N_{inde}	N_{mixture}	N_{pairwise}
Bap	678	0	12	12	2205	2208	2117
Bin	1857	2	80	81	1300	1298	1228
Mef2	4545	0	161	161	3681	3681	3665
Tin	1791	0	40	40	1333	1333	1310
Twi	3211	182	141	128	3810	3862	3722
c-Myc	3038	0	95	95	2996	2996	2920
E2f1	17367	0	877	877	16625	16625	14915
Esrrb	18453	172	160	167	11243	11333	11275
Klf4	9404	0	97	97	5912	5912	5913
Nanog	8022	0	111	111	6196	6196	6224
n-Myc	6367	0	54	54	6981	6981	6954
Oct4	3147	0	74	74	3187	3187	3079
Smad1	907	0	24	24	690	690	667
Sox2	3523	0	95	95	2306	2306	2293
STAT3	2099	54	58	62	2308	2264	2231
Tcfcp2l1	22406	0	418	418	16691	16691	16649
Zfx	9152	0	203	203	6473	6473	6473
CEBPB	14500	399	337	334	8275	8322	8267
CTCF	32958	360	492	579	17087	17098	17060
E2f4	4132	248	590	517	4643	5146	3879
Fosl1	5981	0	90	90	5088	5088	5039
Max	8751	24	70	81	12531	12495	12386
MyoD	33969	717	679	665	25416	25430	25344
Myog	38292	1116	584	835	29520	29334	29647
NRSF	13756	639	672	488	13183	14363	13440
SRF	2370	1	34	35	2929	2928	2948
TCF3	9453	185	277	257	8528	8690	8775
USF1	8956	11	14	12	8628	8619	8625

For each TF, we show the number N_{chip} of ChIP sequences retrieved, the numbers $\Delta_{\text{inde-pairwise}}$, $\Delta_{\text{inde-mixture}}$, and $\Delta_{\text{pairwise-mixture}}$ of different ChIP sequences used for training between either two models, and the numbers N_{inde}, N_{mixture}, and N_{pairwise} of TFBSs used to learn each model.

nucleotide b at position i. However, this frequency fluctuates around the "true" probability due to finite sample size, and for example unobserved nucleotides could actually have a low probability of being observed provided that the number of observations be high enough. It is usual to correct for this effect by using the Bayesian pseudo-count approach stemming from Laplace's rule of succession [3]. The probability to observe nucleotide b at position i is given by:

$$p_{i,b} = \frac{n_{i,b} + \alpha_b}{N + \sum_b \alpha_b} \tag{3}$$

where $n_{i,b}$ is the number of observed b at position i, N is the total number of sites, and α_b's are the pseudo-counts, or prior probabilities to observe nucleotide b at position i. The pseudo-counts were all set to 1, however no significant effect was noted when changing this value, as expected from the large number of observations.

Kullback-Leibler divergence

The Kullback-Leibler divergence is a measure of distance between two probability distributions p and q of a variable s, and is defined as:

$$\text{DKL}(p\|q) = \sum_s p(s) \log \frac{p(s)}{q(s)}. \tag{4}$$

Throughout this paper, when a DKL is calculated between a finite sample and a model distribution, p corresponds to the sites frequencies in the sample, and q to the model distribution. When the DKL is calculated between a PWM of a basin of attraction of a metastable

state and a PWM from the mixture model, p is used for the former, and q for the latter.

Estimation of the fluctuations due to finite sampling: DKL *vs* self

To estimate whether the description of the data by a model (*e.g.* independent or pairwise) could be improved or was consistent with the finite number N of observed sequences, we computed the 'self' DKL between the distribution of a set of N sequences drawn from the model distribution and the model distribution itself. This procedure was repeated 100 times. TFs for which the independent model DKL was smaller than or within two standard deviations of the self DKL were discarded for later analysis.

Derivation of the pairwise interaction model

Information theory offers a principled way to determine the probabilities of a set of states given some measurable constraints. It consists in maximizing a functional known as the entropy[47, 48] over the set of possible probability distributions given the imposed constraints. Here, we wish to determine the probability $P(s)$ of a DNA sequence s of length L, in the set of TFBSs for a transcription factor, given the constraints that the probability distribution P retrieves the one- and two-point correlations observed in a set of bound DNA sequences. We denote by $\mathcal{A}$ the alphabet of possible nucleotides, $\mathcal{A} = \{A, C, G, T\}$ and by s_i the nucleotide at position i in the sequence s so that $s = s_1 \cdots s_L$. With these notations, the entropy with the considered constraints translates into the the following functional:

$$\begin{aligned}\mathcal{L} = -\sum_{\{s\}} P(s)\ln P(s) + \lambda\left(\sum_{\{s\}} P(s) - 1\right) + \sum_{i=1}^{L}\sum_{a\in\mathcal{A}} h_i(a)\left(\sum_{\{s\}} P(s)\delta(s_i,a) - P_i(a)\right) \\ + \sum_{i=1}^{L-1}\sum_{j>i}\sum_{a\in\mathcal{A}}\sum_{a'\in\mathcal{A}} J_{i,j}(a,a')\left(\sum_{\{\sigma\}} P(a)\delta(s_i,a)\delta(s_j,s') - P_{i,j}(a,a')\right),\end{aligned} \tag{5}$$

where $P_i(a)$ (resp. $P_{i,j}(a,a')$) is the probability of having nucleotide a at position i (resp. nucleotides a and a' at position i and j) in the TFBS data set. The function δ denotes the Kronecker $\delta-$function defined by $\delta(a,a') = 1$ if $a = a'$,and 0 otherwise. The first term in Eq. (5) is the entropy of the probability distribution to be found and the other terms are the given constraints along with their Lagrangian multipliers. Maximization of the functional $\mathcal{L}$ is performed in a usual way by setting the functional derivative with respect to the probability distribution P to zero:

$$\frac{\delta\mathcal{L}}{\delta P(s)} = 0 = -\ln P(s) - 1 + \lambda + \sum_{i=1}^{L} h_i(s_i) + \sum_{i=1}^{L-1}\sum_{j>i} J_{i,j}(s_i,s_j). \tag{6}$$

Finally, using the constraint $\sum_{\{s\}} P(s) = 1$, one finds the probability distribution that maximizes entropy given the constraints that it reproduces the observed one- and two-point correlations:

$$P[s] = \exp[-\mathcal{H}(s)]/\mathcal{Z}, \tag{7}$$

where $\mathcal{H}(s)$ is the inhomogeneous Potts model Hamiltonian,

$$\begin{aligned}\mathcal{H}[s_1...s_L] = -\sum_{i=1}^{L} h_i(s_i) - \sum_{i=1}^{L}\sum_{j<i} J_{i,j}(s_i,s_j), \\ s_i \in \{A, C, G, T\}.\end{aligned} \tag{8}$$

The normalization constant $\mathcal{Z}$ is the partition function,

$$\mathcal{Z} = \sum_{\{s\}} \exp[-\mathcal{H}(s)]. \tag{9}$$

Gauge fixing

The probability distribution of sequences, as given by Eqs. (7, 8), is invariant under shifts of the local fields $h_i(a)$ and under transformations between the interaction terms $J_{i,j}(a,a')$ and the local fields. In order to uniquely determine $\mathcal{H}$, this arbitrariness needs to be taken care of by adding further conditions that uniquely fix the in-

teraction parameters, a process known as gauge fixation [20] that we detail here.

a. Local fields. Since it amounts to changing the reference energy and is cancelled by the normalization, the probability is invariant with respect to the following global shift of the $h_i(a)$

$$h_i(s_i) \to \tilde{h}_i(s_i) = h_i(s_i) + \varepsilon_i. \tag{10}$$

We choose to fix this invariance by minimizing the square norm $S_i = \sum_{a\in\mathcal{A}} \tilde{h}_i(a)^2$ of local field terms with respect to the gauge degree of freedom. The corresponding gauge-fixing condition reads

$$\sum_{a\in\mathcal{A}} \tilde{h}_i(a) = 0. \tag{11}$$

This condition can be imposed on any set of fields h_i by using the symmetry (10) and redefining the fields as follows,

$$h_i(s_i) \to h_i(s_i) - \frac{1}{4}\sum_{a\in\mathcal{A}} h_i(a). \tag{12}$$

b. Interaction terms. Another invariance stems from the fact that contributions can be shifted between local fields and interaction energies. Namely, the following change of variables does not affect the probability:

$$J_{ij}(s_i,s_j) \to \tilde{J}_{ij}(s_i,s_j) = J_{ij}(s_i,s_j)+\psi_i(s_i)+\phi_j(s_j)+C_{i,j}, \tag{13}$$

since the local fields ψ_i and ϕ_j can be redistributed in h and the constant $C_{i,j}$ gives an energy reference for the interacting nucleotides that is cancelled by the normalization process. Again, a gauge condition is obtained by minimizing the square norm $S_{i,j} = \sum_{a,a'\in\mathcal{A}}[\tilde{J}_{ij}(a,a')]^2$ of interaction terms with respect to the gauge degrees of freedom. This yields the conditions:

$$\sum_{a\in\mathcal{A}} \tilde{J}_{i,j}(a,a') = \sum_{a'\in\mathcal{A}} \tilde{J}_{i,j}(a,a') = 0. \tag{14}$$

These can be imposed on a set a of J_{ij} parameters by redefining them as follows:

$$\begin{aligned} J_{ij}(s_i,s_j) \to & J_{ij}(s_i,s_j) + \frac{1}{16}\sum_{a,a'\in\mathcal{A}} J_{i,j}(a,a') \\ & - \frac{1}{4}\sum_{a\in\mathcal{A}} J_{i,j}(a,s_j) - \frac{1}{4}\sum_{a\in\mathcal{A}} J_{i,j}(s_i,a). \end{aligned} \tag{15}$$

Determination of the pairwise interaction model from the data

The parameters of the inhomogeneous Potts model in Eq. (8), giving the energy of an observed sequence of length L, must be computed from the data. The parameters h and J represent the energy contributions respectively coming from individual nucleotides and from their interactions. The PWM model is the particular case where all the interaction parameters vanish: $J_{i,j}(a,a') = 0$.

To build the model, we start from the PWM description, characterized by the set of initial $h_i(a) = \log p_{i,a}$ and the interaction parameters J's set to zero. We add one interaction parameter $J_{i,j}(a,a')$ at a time, corresponding to the pair of nucleotides whose pairwise distribution predicted by the model differs most from data, as estimated by a binomial p-value. We then fit the augmented model to data, use this model to select a new set binding sites from the reads, and repeat the whole procedure. In each of these steps, fitting is performed by a gradient descent algorithm:

$$J \;\to\; J + \epsilon\left[c_2^{\text{data}} - c_2^{\text{model}}\right], \tag{16}$$

$$h \;\to\; h + \epsilon\left[c_1^{\text{data}} - c_1^{\text{model}}\right], \tag{17}$$

where c_1 and c_2 are matrices of size $4\times L$ and $4L\times 4L$ respectively corresponding to the single- and two-point frequencies, and superscripts denote whether the matrices are computed from the data or from the model distribution. This algorithm converges to the set of parameters $(\{\tilde{h}_i\},\tilde{J}_{i,j})$ that match all single marginals and the pairwise marginals of interest. The number of interaction parameters that are being added is controlled by the Bayesian Information Criterion, or BIC (Figure 5). The BIC computes the opposite log-likelihood and adds a penalty proportional to the number of parameters involved. This adverts the over-fitting of a finite dataset with an extravagant number of parameters. The procedure is iterated until minimization of the BIC, yielding the best pairwise model with the full set of parameters $(\{h_i(a)\},\{J_{i,j}(a,a')\})$. As in the case of the PWM model, we score each sequence using the ratio between the TF and background pairwise models and impose a score cutoff so as to select a set of bound sites yielding 50% TPR, on which a new pairwise model is learned. This process is iterated until convergence to a stable set of bound sites.

BIC computation

Consider a sample $X = (X_1,\ldots,X_N)$ of N TFBSs drawn from an unknown distribution function f we wish to estimate. To this extent, several models $\{M_1,\ldots,M_m\}$ are proposed, each model M_i having a density g_{M_i} with parameter θ_i of dimension K_i. It is straightforward to see that, as K_i increases, the fit to the observed sample as measured by the likelihood function $g_{M_i}(X|\theta_i)$ increases as well, the limiting case being when f is estimated as the sample distribution. However, such an estimator is inappropriate to account for new, yet unobserved TFBSs, *i.e.* it is not predictive. Such a case where the number of parameters used to estimate a distribution becomes of the order of the size of the sample is known as overfitting. The BIC allows to overcome overfitting by penalizing high dimension parameters. Using

Bayes Rule, and a uniform a priori distribution on the models, we have

$$P(M_i|X) \propto P(X|M_i). \tag{18}$$

That is, the probability of the model given the data can be inferred from the probability that the data is generated by the model. The latter is obtained by marginalizing the joint distribution of the data and the parameters over the space of parameters Θ:

$$P(X|M_i) = \int_\Theta P(X,\theta|M_i)d\theta = \int_\Theta g_{M_i}(X|\theta)P(\theta|M_i)d\theta. \tag{19}$$

For a unidimensional parameter θ, the likelihood $g_{M_i}(X|\theta)$ is maximized at some particular $\hat{\theta}_i$ with an uncertainty (or width) proportional to $1/\sqrt{N}$ in the limit of large N. Assuming a broad prior, then for large N the integral is dominated by the likelihood which is concentrated around its maximum. One can then approximate the integral by the area of the region of height the maximum likelihood and of width $1/\sqrt{N}$, that is $g_{M_i}(X,\hat{\theta}_i)/\sqrt{N}$. This result can be retrieved analytically using the method of steepest descent. For a number K_i of parameters, one gets a total volume $g_{M_i}(X,\hat{\theta}_i)/N^{K_i/2}$ [31]. Taking the logarithm yields the BIC condition:

$$BIC_i = -2\log(P(X|M_i)) \simeq -2\log(g_{M_i}(X,\hat{\theta}_i)) + K_i\log(N). \tag{20}$$

In the present case, the sample X is the set of observed TFBSs and the model M_i determines the probability $P_{M_i}(s)$ of belonging to X,

$$\log(g_{M_i}(X,\hat{\theta}_i)) = \sum_{s\in X}\log[P_{M_i(\hat{\theta}_i)}((s)]. \tag{21}$$

The interpretation of Eq. (20) is clear: adding new parameters improves the fit, but also adds new sources of uncertainty about these parameters due to the finite size of the data. This uncertainty disappears as $N \to \infty$, since the log-likelihood scales with N while the correction scales with $\log(N)$.

Finally, Eq. (20) is a functional over models, the chosen model M_{BIC} is the one that minimizes it,

$$M_{BIC} = \underset{M_i}{\operatorname{argmin}}\ BIC_i. \tag{22}$$

PWM mixture model

We investigated an approach based on a mixture of PWMs. For that purpose, we used a comparable setup as for the pairwise model. However, instead of adding correlations to a given PWM, new PWMs were added to a mixture model. More precisely, a mixture of K PWMs, with $1 \leq K \leq 10$, was generated by using a K-means algorithm with a Hamming distance metrics on the initial set of bound sites. This resulted in K clusters, each comprising n_k sites among the initial N sites. A PWM was generated on each of these clusters, with probability distribution $\mathcal{P}_k$. The mixture model of order K was then defined as [31]:

$$\mathcal{P}[s] = \sum_{k=1}^{K} p_k\mathcal{P}_k[s], \tag{23}$$

where $p_k = n_k/N$ is the cluster weight. Because a PWM has $3\times L$ degrees of freedom (L of them being constrained by the summation of nucleotide probabilities to one) and there are $K-1$ free weight parameters, the number of parameters corresponding to a mixture of order K is $3LK + (K-1)$. As previously, the model showing minimal BIC score was used for sites detection, a new set of PWMs and weights p_k was generated by clustering the set of detected sites and the procedure was iterated until convergence to a stable set of sites.

Metastable minima of the pairwise interaction model and their basins of attractions

We defined the basins of attraction of a pairwise interaction model energy landscape, in the following fashion. Let s be a site with energy $\mathcal{H}(s)$. We looked for the nucleotides that could be changed to minimize $\mathcal{H}(s)$. If such nucleotides existed, one of them was chosen at random, and its value was updated. One local minimum of the energy landscape, or metastable state, was reached when no such nucleotide could be found. The basin of attraction of a metastable state was then defined as the ensemble of sites that fell to this metastable state when their energy was minimized following the above procedure. We computed metastable states and their basins of attraction for the final set of bound sites obtained with the best pairwise model. A PWM was learned on each basin of attraction, leading to a set of representative PWMs, with different weights representing different proportions of bound sites in their basins.

Computation of the Direct Information

We wanted to build a quantity based solely on direct interactions $J_{i,j}$ between nucleotides, discarding indirect interactions. To this end, we used the interaction parameters obtained from the pairwise model to build the direct dinucleotide probability function:

$$P^d_{i,j}(a,a') = e^{\tilde{h}_i(a)+\tilde{h}_j(a')+J_{i,j}(a,a')}/\mathcal{Z}_{i,j}, \tag{24}$$

where

$$\mathcal{Z}_{i,j} = \sum_{a,a'} e^{\tilde{h}_i(a)+\tilde{h}_j(a')+J_{i,j}(a,a')}.$$

The 8 effective fields $\tilde{h}_i$ and $\tilde{h}_j$ were fully determined by the constraints that the direct probability function matches the observed one-point frequencies:

$$\begin{aligned} \sum_{a'} P^d_{i,j}(a,a') &= P_i(a), \qquad a' \in \{A,C,G,T\}, \\ \sum_{a} P^d_{i,j}(a,a') &= P_j(a'), \qquad a \in \{A,C,G,T\}. \end{aligned} \tag{25}$$

The normalization of the probabilities $\sum_a P_i(a) = 1$, served to reduce this system to 6 equations. The fields $\tilde{h}_i(a)$, which are determined up to a constant, were fixed by the gauge condition that they vanished for the nucleotide A, $\tilde{h}(A) = 0$. The system was solved using the Levenberg-Marquadt algorithm with $\lambda = 0.005$.

The Direct Information [37] was then defined as:

$$DI_{i,j} = \sum_{a,a'} P^d_{i,j}(a,a') \log_2 \left(\frac{P^d_{i,j}(a,a')}{P_i(a)P_j(a')} \right). \tag{26}$$

As there is no upper bound for this direct information, we built a normalized version of the direct information:

$$NDI_{i,j} = \frac{DI_{i,j}}{\sqrt{S_i S_j}}, \tag{27}$$

where S_i denotes the entropy at position i. Note that $S_i = DI_{i,i}$ so that $NDI_{i,i} = 1$ for this maximally correlated case. On the contrary, independent nucleotides give $NDI_{i,j} = DI_{i,j} = 0$.

Participation Ratio

For each TF, an interaction weight was defined for each pair of nucleotides as

$$w_{i,j} = NDI_{i,j} / \sum_{i \neq j} NDI_{i,j}. \tag{28}$$

Self-interactions have no meaning here and were attributed $w_{i,i} = 0$. Let us note $N = L(L-1)$ the number of possible interactions. Using our weight, one writes the Participation Ratio as:

$$R = \frac{1}{N \sum_{i \neq j} w^2_{i,j}}. \tag{29}$$

The interpretation is simple: if all weights are equal, $w_{i,j} = 1/N$ and $R = 1$, that is all possible interactions are represented. Conversely, if only one interaction accounts in the distribution budget, then $R = 1/N$, meaning that only one of all possible interactions is represented.

Distance between interactions

The previously defined interaction weights were averaged over all possible pairs of nucleotides at a given distance d of one another, yielding the distance distribution:

$$P(|i-j| = d) = \mathcal{Z}^{-1} \frac{1}{N-d} \sum_{|i-j|=d} w_{i,j}, \tag{30}$$

where

$$\mathcal{Z} = \sum_{d=1}^{N-1} \frac{1}{N-d} \sum_{|i-j|=d} w_{i,j} \tag{31}$$

is a normalization factor. Note that we introduced a correction $1/(N-d)$ to account for finite-size effects, namely the fact that randomly distributed interactions will lead to an overrepresentation of nearest neighbours interactions just because these are more numerous.

Interaction matrix and Hopfield patterns

In the Hamiltonian shown in (1), only $16L(L-1)/2$ terms appear in the interaction budget: indeed, we forbid self-interations (already accounted for by the local field h) and do not count the interactions twice. However, we can straightforwardly extend the interaction matrix to a full symmetric matrix $\hat{J}_{(i,a),(j,b)}$ of size $(4L)^2$, with $4L$-valued indices $(i,a), i \in \{1, \cdots, L\}, a \in \mathcal{A}$. The matrix $\hat{J}$ is such that for $i > j$, $\hat{J}_{(i,a)(j,b)} = J_{i,j}(a,b)$ with furthermore $\hat{J}_{(i,a),(i,b)} = 0$ and $\hat{J}_{(i,a),(j,b)} = \hat{J}_{(j,b),(i,a)}$. The energy of a sequence s can then be written with these notations

$$\sum_{1 \leq i < j \leq L} J_{i,j}(s_i, s_j) = \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \hat{J}_{(i,s_i),(j,s_j)} = v(s)^\dagger \hat{J} v(s), \tag{32}$$

where in the last equality the † sign denotes vector transposition and we have introduced the $4L$ vector $v(s)$ associated to sequence s, $v(s)_{i,a} = 1$ if $a = s_i$ and $v(s)_{i,a} = 0$ otherwise.

Since the matrix $\hat{J}$ is symmetric, it can be diagonalized in an orthonormal basis of eigenvectors ξ^k, $k = 1, \cdots, L$ with real eigenvalues λ_k,

$$\hat{J} = \sum_k \lambda_k \xi^k \, \xi^{k\dagger}. \tag{33}$$

Denoting by $\xi^k_{(i,a)}$ the coordinates of the k-th eigenvector then, one can rewrite Eq. (32) as

$$\sum_{1 \leq i < j \leq L} J_{i,j}(s_i, s_j) = \frac{1}{2} \sum_{k=1}^{4L} \lambda_k \left(\sum_{i=1}^{L} \xi^k_{(i,s_i)} \right)^2. \tag{34}$$

Finally, the full Hamiltonian is given by:

$$\mathcal{H} = -\sum_i h_i(s_i) - \frac{1}{2} \sum_{k=1}^{4L} \lambda_k \left(\sum_{i=1}^{L} \xi^k_{(i,s_i)} \right)^2. \tag{35}$$

Acknowledgments

We wish to thank PY Bourguignon and I Grosse for stimulating discussions at a preliminary stage of this work.

[1] F. Spitz and E. E. Furlong, Nat. Rev. Genet. **13**, 613 (2012).
[2] J. A. Stamatoyannopoulos, Genome Res. **22**, 1602 (2012).
[3] W. W. Wasserman and A. Sandelin, Nat. Rev. Genet. **5**, 276 (2004).
[4] O. G. Berg and P. H. von Hippel, J Mol Biol **193**, 723 (1987).
[5] G. D. Stormo and D. S. Fields, Trends Biochem Sci **23**, 109 (1998).
[6] T. K. Man and G. D. Stormo, Nucleic Acids Res **29**, 2471 (2001).
[7] P. V. Benos, M. L. Bulyk, and G. D. Stormo, Nucleic Acids Res **30**, 4442 (2002).
[8] M. L. Bulyk, P. L. F. Johnson, and G. M. Church, Nucleic Acids Res **30**, 1255 (2002).
[9] A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, et al., Cell **152**, 327 (2013).
[10] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, et al., Science **324**, 1720 (2009), URL `http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19443739&retmode=ref&cmd=prlinks`.
[11] Y. Zhao and G. D. Stormo, Nat. Biotechnol. **29**, 480 (2011).
[12] Q. Zhou and J. S. Liu, Bioinformatics **20**, 909 (2004).
[13] M. Hu, J. Yu, J. M. G. Taylor, A. M. Chinnaiyan, and Z. S. Qin, Nucleic Acids Res **38**, 2154 (2010).
[14] Y. Barash, G. Elidan, N. Friedman, and T. Kaplan, in *Proceedings of the seventh annual international conference on Research in computational molecular biology* (ACM, 2003), pp. 28–37.
[15] E. Sharon, S. Lubliner, and E. Segal, PLoS Comput Biol **4**, e1000154 (2008).
[16] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature **440**, 1007 (2006), URL `http://www.nature.com/nature/journal/v440/n7087/full/nature04701.html`.
[17] J. Shlens, G. D. Field, J. L. Gauthier, M. I. Grivich, D. Petrusca, A. Sher, A. M. Litke, and E. J. Chichilnisky, J Neurosci **26**, 8254 (2006), URL `http://www.jneurosci.org/cgi/content/full/26/32/8254`.
[18] Y. Ikegaya, G. Aaron, R. Cossart, D. Aronov, I. Lampl, D. Ferster, and R. Yuste, Science **304**, 559 (2004).
[19] A. Roxin, V. Hakim, and N. Brunel, J. Neurosci. **28**, 10734 (2008).
[20] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc Natl Acad Sci USA **106**, 67 (2009), URL `http://www.pnas.org/content/106/1/67.long`.
[21] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Proc Natl Acad Sci USA **107**, 5405 (2010), URL `http://www.pnas.org/content/107/12/5405`.
[22] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, Proc Natl Acad Sci USA **109**, 4786 (2012).
[23] Y. Zhao, S. Ruan, M. Pandey, and G. D. Stormo, Genetics **191**, 781 (2012).
[24] Y. Cao, Z. Yao, D. Sarkar, M. Lawrence, G. J. Sanchez, M. H. Parker, K. L. MacQuarrie, J. Davison, M. T. Morgan, W. L. Ruzzo, et al., Dev Cell **18**, 662 (2010).
[25] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass, Mol Cell **38**, 576 (2010).
[26] R. Zinzen, C. Girardot, J. Gagneur, M. Braun, and E. Furlong, Nature **462**, 65 (2009).
[27] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, et al., Cell **133**, 1106 (2008).
[28] I. Dunham and et al., Nature **489**, 57 (2012).
[29] T. Cover and J. Thomas, *Elements of information theory* (Wiley-interscience, 2006).
[30] R. Baxter, *Exactly solved models in statistical mechanics* (Dover Publications, 2008).
[31] C. Bishop et al., *Pattern recognition and machine learning* (Springer New York, 2006).
[32] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, Genome Res **14**, 1188 (2004).
[33] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc Natl Acad Sci USA **106**, 67 (2009).
[34] J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).
[35] S. Cocco, R. Monasson, and V. Sessak, Phys Rev E Stat Nonlin Soft Matter Phys **83**, 051123 (2011).
[36] G. Tkacik, E. Schneidman, M. J. B. II, and W. Bialek, arXiv **q-bio.NC** (2006), 4 pages, 3 figures, q-bio/0611072v1, URL `http://arxiv.org/abs/q-bio/0611072v1`.
[37] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc Natl Acad Sci USA **108**, E1293 (2011).
[38] J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic, Proc Natl Acad Sci USA **109**, 10340 (2012).
[39] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks, Cell **149**, 1607 (2012).
[40] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, PLoS ONE **6**, e28766 (2011).
[41] M. Lassig, BMC Bioinformatics **8 Suppl 6**, S7 (2007).
[42] A. Moses, D. Chiang, D. Pollard, V. Iyer, and M. Eisen, Genome biology **5**, R98 (2004).
[43] R. Siddharthan, E. Siggia, and E. van Nimwegen, PLoS Comput Biol **1**, e67 (2005).
[44] H. Rouault, K. Mazouni, L. Couturier, V. Hakim, and F. Schweisguth, Proc Natl Acad Sci U S A **107**, 14615 (2010).
[45] Z. Bao and S. R. Eddy, Genome Res **12**, 1269 (2002).
[46] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Prüss, I. Reuter, and

F. Schacherer, Nucleic Acids Res **28**, 316 (2000).
[47] E. Jaynes, Physical review **108**, 171 (1957).
[48] C. Shannon, Bell Syst Tech J **27**, 623 (1948).

Supporting Figures

FIG. S1: **Dependence of the fit on the number of ChIP sequences**. For each TF, the number of available ChIP sequences is plotted *vs.* the improvement in the description of its TFBS statistics, provided by the he pairwise model as compared to the PWM independent model. The latter is quantified by the ratio of DKL between the respective model probability distributions and the experimental ones provided by the ChIP data, $\mathrm{DKL_{pw}}/\mathrm{DKL_{inde}}$. The improvement afforded by the pairwise model is clearly seen to be correlated to the number of ChIP sequences available.

CEBPB

Independent model

Mixture model

55% 45%

Pairwise model

26% 37%

25% 11%

CTCF

Independent model

Mixture model

24% 20% 20% 20% 17%

Pairwise model

18% 82%

E2f4

Independent model
Mixture model
63%
37%
Pairwise model
29%
24%
8%
<1%
4%
<1%
3%
<1%
1%
<1%
26%
<1%
2%
<1%
1%
1%
1%
<1%

Mixture PWMs
0.63
0.37
Metastable PWMs
0.29
0.26
0.24
0.08
0.04
0.03
0.02
0.01
0.01
0.01
0.01
<0.01
<0.01
<0.01
<0.01
<0.01
<0.01
16
14
12
10
8
6
4
2

Max

Independent model
Mixture model
35%
27%
21%
17%
Pairwise model
32%
2%
2%
32%
20%
12%
Mixture PWMs
0.35
0.27
0.21
0.17
0.32
0.32
0.20
0.12
0.02
0.02
Metastable PWMs
13
12
11
10
9
8
7
6
5
4

Myog

Independent model
Mixture model
25%
25%
21%
18%
11%
Pairwise model
14%
30%
10%
23%
9%
9%
5%
Mixture PWMs
0.25
0.25
0.21
0.18
0.11
0.30
0.23
0.14
0.10
0.09
0.09
0.05
Metastable PWMs
26
24
22
20
18
16
14
12
10
8

NRSF

Independent model

Mixture model

Pairwise model

Tcf3

Independent model
Mixture model
56%
44%
Pairwise model
31%
12%
9%
6%
8%
5%
6%
4%
4%
4%
2%
2%
2%
2%
2%
2%
1%
1%

Mixture PWMs
0.56
0.44
0.31
0.12
0.09
0.08
0.06
0.06
0.05
0.04
0.04
0.04
0.02
0.02
0.02
0.02
0.02
0.02
0.01
0.01
Metastable PWMs
14
12
10
8
6
4

FIG. S2: Same as Figure 6 of the main text for all considered factors described by a mixture model with two or more PWMs.

Bin
0.12
0.1
0.08
0.06
0.04
0.02
0
CEBPB
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
CTCF
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
e2f1
0.1
0.08
0.06
0.04
0.02
0
E2f4
0.6
0.5
0.4
0.3
0.2
0.1
0
Fosl1
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Klf4
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Max
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Mef2
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Myog
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Nanog
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
n-Myc
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0

FIG. S3: Same as Figure 7 of the main text for the other considered factors.

FIG. S4: **Background correlations** (A,B,C) Heat maps showing the correlations between nucleotides in the ChIP data of the 3 factors from the main text. Because of translation invariance, we only show the correlations between a nucleotide (rows) and the next nearest (first four columns) to farthest (last four columns) nucleotides, using the binding site length of $L = 12$. We see in the Drosophila data the appreciable presence of repeated sequences (of type AA, TT, CC, and GG). In the mammalian data sets, we observe the known CpG depletion. (A',B',C') Heat maps showing the values of the Normalized Direct Information between pairs of nucleotides.

FIG. S5: **Variable spacer length** We learned a pairwise model for Esrrb including the 4 flanking nucleotides on the left of the main motif. (A) The metastable states of this model show a feature not captured in the main text where binding sites are defined symmetrically around the center of mass of the information content: namely a 'CAG' trinucleotide with variable spacer length from the main motif. This feature is apparent in the first 3 logos shown here. (B) The contribution of this trinucleotidic interaction to the Direct Information is captured through strong direct links between the 4 flanking nucleotides, showing that the pairwise model is implicitly able to capture higher order correlations. Logos from the PWM model are surrounding the heatmap for clarity.

2.5 Analyse thermodynamique des modèles

2.5.1 Chaleur spécifique

En plus des résultats présentés dans l'article, nous nous sommes intéressés à une quantité classique de la thermodynamique : la chaleur spécifique ou capacité calorifique. Considérons un modèle décrit par la statistique de Boltzmann à la température inverse $\beta = 1/T$ (on omet la constante de Boltzmann k en l'intégrant à l'énergie) :

$$P(s) = \frac{1}{\mathcal{Z}} e^{-\beta E(s)} \tag{2.24}$$

Le cas de l'équation 2.23 correspond au cas particulier $\beta = 1$. Nous voulons voir comment l'amplification ou la diminution globale de l'écart entre les énergies affecte la possibilité du système d'explorer les différents états possibles. À température nulle ($T \to 0$ ou $\beta \to \infty$), le système reste dans le niveau fondamental de minimum d'énergie et de probabilité 1, alors qu'à des températures non nulles le système à l'énergie E_0 transite vers un état d'énergie supérieure E_1 avec une probabilité $\propto \exp\left(-\beta(E_1 - E_0)\right)$. Lorsqu'un paysage énergétique est composé de plusieurs puits d'énergie séparés par des barrières énergétiques importantes, on s'attend à avoir une (ou plusieurs) températures critiques à partir desquelles de fortes différences d'énergie deviennent franchissables. L'énergie moyenne peut alors être significativement affectée, sautant soudainement à une nouvelle valeur du fait du poids des nouveaux états explorés.

La chaleur spécifique permet de caractériser ces sauts soudains d'énergie moyenne lors de la variation de la température, caractéristiques des transitions de phase. Elle mesure simplement la variation de l'énergie moyenne lors d'une variation de température :

$$C(T) = \frac{d\langle E \rangle}{dT} \tag{2.25}$$

où

$$\langle E \rangle = \sum_{\{s\}} E(s) \frac{e^{-\beta E(s)}}{\mathcal{Z}} \tag{2.26}$$

Cette chaleur spécifique peut par ailleurs s'écrire sous une forme plus utile :

$$\begin{aligned}\frac{d\langle E\rangle}{dT} &= -\beta^2\frac{d\langle E\rangle}{d\beta}\\ &= -\beta^2\left[\sum_{\{s\}}E(s)\left(-E(s)e^{-\beta E(s)}\right)\frac{1}{\mathcal{Z}}+\sum_{\{s\}}E(s)e^{-\beta(E(s))}\left(-\frac{d\mathcal{Z}}{d\beta}\frac{1}{\mathcal{Z}^2}\right)\right]\\ &= \beta^2\left[\langle E^2\rangle-\langle E\rangle^2\right]\end{aligned} \tag{2.27}$$

Ainsi, la chaleur spécifique $C(T)$ est directement accessible en regardant les corrélations de l'énergie sur l'ensemble des états du système, ce qui peut se calculer simplement à partir des modèles de fixation. Nous avons calculé la variation de $C(T)$ en fonction de la température pour les modèles indépendant et avec dépendances obtenus en 2.4 pour les différents TFs étudiés. Une température fictive est introduite dans les modèles en multipliant les énergies par β, afin de se placer dans le cadre de l'équation 2.24. Les résultats sont montrés en figure 2.3 (modèle indépendant en bleu, modèle de Potts en rouge). On observe pour la plupart des facteurs l'existence de deux pics de chaleur spécifique pour des températures de l'ordre de $T \sim 10^{-1}$ et $T \sim 5$ (par exemple, $T = 0.4$ et $T = 2.8$ dans le cas du modèle indépendant de Twist). Il y a de légères variations entre les deux modèles : notamment, le premier pic semble renforcé par le modèle de Potts dans plusieurs cas (par exemple, E2f4, NRSF, TCF3 ou Twist). Néanmoins, le nombre de pics (ou de transitions de phases) reste le même.

2.5.2 Lien avec les valeurs des champs et des couplages

Afin de comprendre l'existence des pics de chaleur spécifique et les énergies (températures) associées, il faut revenir aux modèles d'énergie. Lorsque l'on regarde l'histogramme des valeurs absolues de h_i obtenues dans les modèles indépendant des différents TFs étudiés, on trouve plusieurs valeurs typiques autour de 10^{-4}, 1 et 10 (fig. 2.4A). Celles-ci peuvent s'expliquer de la manière suivante. Dans le modèle indépendant, les champs sont simplement le logarithme naturel de la probabilité d'observer un nucléotide a à une position i donnée $h_i(a) = -\log P_i(a)$ (la jauge est choisie telle que $\mathcal{Z}_\rangle = 1$). En valeur absolue, les champs h_i proches de 0 ($h_i \sim 10^{-4} - 10^{-3}$) correspondent aux nucléotides très conservés (toujours observés), les valeurs autour de 1 correspondent à des nucléotides dégénérés (i.e également observés : $|\log(1/4)| \sim 1.4$) et les valeurs autour de 10 correspondent aux nucléotides qui ne sont jamais observés, au pseudocount près (pour un pseudocount de 1 et 10^4 séquences, $|\log(10^{-4})| \sim 9.2$). On peut maintenant mieux comprendre les pics de chaleur spécifique. À

FIGURE 2.3 – **Chaleur spécifique pour différents TFs.**
La chaleur spécifique (l'équivalent de la capacité calorifique en thermodynamique) $C(T) = d\langle E\rangle/dT$ est tracée en fonction de la température kT (échelle logarithmique) pour les différents TFs considérés. Le modèle indépendant (bleu) et le modèle de Potts avec interactions (rouge) sont comparables dans la plupart des cas.

température nulle, seuls les sites consensus sont accessibles. Lorsque la température se rapproche de 1, les nucléotides dégénérés d'énergie $h_i \sim 1$ deviennent accessibles, augmentant significativement la valeur de l'énergie moyenne (premier pic). Puis, lorsque la température se rapproche de 10, les nucléotides non observés d'énergie $h_i \sim 10$ deviennent à leur tour accessibles, augmentant à nouveau l'énergie moyenne (deuxième pic).

Dans le cas du modèle de Potts (fig. 2.4B), les champs h_i prennent des valeurs proches de celles obtenues avec le modèle indépendant. Par ailleurs, les interactions $J_{i,j}$ sont réparties autour d'un mode centré autour de $J_{i,j} \sim 0.5$, ce qui correspond l'échelle d'énergie du premier pic. Ainsi, le renforcement du premier pic de chaleur spécifique par rapport au cas indépendant observé pour plusieurs TFs de la figure 2.3 peut s'expliquer par l'effet des termes d'interaction $J_{i,j}$.

2.6 Conclusion et perspectives

Nous avons analysé les dépendances au sein des sites de fixation liés *in vivo* pour différents facteurs de transcription Drosophiles et mammifères. Nous avons comparé les performances d'un modèle PWM, d'un modèle de mélange de PWMs, et d'un modèle de Potts, en utilisant un critère bayésien (BIC) pénalisant les modèles à grand nombre de paramètres. Nous avons exhibé l'existence de corrélations faibles dont la prise en compte permet de significativement améliorer la description des données, le modèle de Potts étant significativement supérieur aux deux autres modèles dans la plupart des cas (22/28). Les interactions ont été étudiées systématiquement, montrant notamment une prépondérance des interactions entre plus proches voisins. Nous avons établi une correspondance entre les PWMs du modèle de mélange et les PWMs décrivant les états métastables du paysage énergétique généré par le modèle de Potts. Enfin, nous avons montré que les corrélations pouvaient être groupées en patterns de Hopfield ou « mémoires », et qu'un petit nombre était suffisant à reconstruire le paysage d'interactions.

Une perspective intéressante de ce travail serait de conduire la même analyse sur des données grande échelle obtenues *in vitro* par la méthode HT-SELEX (Jolma et al., 2013). Notamment, certains des facteurs que nous avons étudiés *in vivo* sont représentés dans ces données, et il serait intéressant de voir les différences entre les modèles obtenus. Notamment, retrouve-t-on les mêmes corrélations ? Peut-on exhiber des spécificités de la fixation *in vivo*,

FIGURE 2.4 – **Histogrammes des valeurs des champs h et couplages J.**
Histogrammes réalisés à partir des valeurs obtenues pour l'ensemble des TFs. Les champs et les couplages sont montrés en valeur absolue sur une échelle logarithmique d'espacement 0.05, et les valeurs nulles ne sont pas représentées. (A) Champs h_{inde} dans le modèle indépendant. (B) Champs h_{pw} et couplages J dans le modèle de Potts.

où l'on s'attend à avoir des effets provenant de diverses sources (fixation de nucléosomes, superposition de sites de fixations, ...) ? Ces questions feront certainement l'objet d'un prochain travail.

Chapitre 3

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

3.1 Les approches existantes pour la recherche de motifs et de modules de régulation . . . 116

3.1.1 MEME : une approche de type Espérance-Maximisation . . . 116

3.1.2 STUBB : une méthode utilisant des motifs connus, les corrélations entre sites de fixations, et la phylogénie . . . 119

3.1.3 MONKEY : vers des modèles phylogénétiques plus complexes . . . 122

3.1.4 Méthodes utilisant des collections d'oligonucléotides . . . 125

3.1.5 Approches sans motifs ou *motif-blind* . . . 125

3.2 Article . . . 127

3.3 Calcul de la moyenne de la postérieure par une méthode MCMC . . . 156

3.3.1 Principe de l'algorithme de Metropolis-Hastings . . . 156

3.3.2 Application au calcul de la postérieure . . . 158

3.3.3 Illustration sur un exemple . . . 159

Introduction du chapitre 3

Dans le chapitre 2, nous avons vu comment décrire l'interaction TF-ADN lorsque des sites de fixation sont connus. Dans ce chapitre, nous adoptons une démarche plus générale. Nous connaissons l'activité de régulation d'un certain nombre de CRMs, et nous souhaitons savoir quels TFs s'y fixent (recherche de motifs), et si le génome contient d'autres CRMs avec la même activité (recherche de modules). Un algorithme permettant précisément de réaliser ces étapes a été développé précédemment par Hervé Rouault et appliqué au cas de la différenciation des organes sensoriels de la Drosophile (Rouault et al., 2010). Nous présentons ici Imogene, l'extension de cet algorithme au cas des mammifères, ainsi que son utilisation comme outil de classification de CRMs associés à différentes régulations.

Avant de rentrer dans le détail d'Imogene, nous présentons les méthodes existantes de recherche de motifs dans des CRMs. Le problème général est le suivant : étant données des CRMs conduisant à une même régulation (l'ensemble d'apprentissage), peut-on construire des modèles de sites de fixation qui « expliquent » cette co-régulation, c'est-à-dire qui prédisent l'existence de sites sur les CRMs mais pas sur des séquences ne participant pas à la corégulation ?

3.1 Les approches existantes pour la recherche de motifs et de modules de régulation

Nous avons déjà introduit différentes méthodes de prédiction de motifs et modules en introduction (section 1.6). Ici nous décrivons plus en avant certaines approches. Pour une revue plus exhaustive, le lecteur intéressé pourra se référer à Aerts (2012).

3.1.1 MEME : une approche de type Espérance-Maximisation

L'une des premières approches pour la prédiction *de novo* de motifs à partir de séquences a été celle de MEME (Bailey and Elkan, 1994), un algorithme basé sur la méthode d'espérance-maximisation ou EM (*expectation maximization*) utilisée précédemment dans ce cadre par Lawrence and Reilly (1990). Cet algorithme se base sur une approche générative pour décrire les processus probabilistes qui ont permis la génération des séquences CRMs (fig. 3.1).

FIGURE 3.1 – **Illustration de l'approche espérance-maximisation.**
Figure tirée de (D'haeseleer, 2006) décrivant l'approche EM. Un premier modèle de motif est construit à partir d'un site initial. Ce modèle permet de pondérer l'ensemble de sites sur les séquences (étape E). En rouge sont montrés les meilleurs sites pour chaque séquence. En utilisant ces poids, il est possible de construire un modèle de vraisemblance maximale (étape M). La méthode originale de Lawrence and Reilly (1990) fait l'hypothèse qu'il y a exactement un site de fixation par séquence, condition qui est relâchée par MEME.

- **Vraisemblance d'une séquence**

Notons $S = \{S_1, \cdots, S_L\}$ une séquence de taille L.[15] Supposons qu'il y a exactement un site de régulation par CRM. C'est l'approche de Lawrence and Reilly (1990), et cette condition est relâchée par MEME, qui autorise l'utilisateur à préciser un nombre moyen de sites par séquence. La probabilité que la séquence possède un site de taille K à la position i étant donné le modèle $\mathcal{M}$ s'écrit

$$P(S|i, \mathcal{M}) = P_0(S_{1,i-1}) \times P(S_{i,i+K-1}|\mathcal{M}) \times P_0(S_{K,L}) \tag{3.1}$$

où $S_{i,j}$ dénote la séquence entre les positions i et j incluses, $P(S_{i,i+K-1}|\mathcal{M})$ est la probabilité

15. On concatène les deux brins d'ADN dans cette séquence. On suppose en effet qu'ils participent équiprobablement à la fixation. Ainsi la séquence génomique double brins est de longueur $L/2$.

de générer la séquence de taille K débutant à la position i avec le modèle $\mathcal{M}$ (voir section 2.2), et $P_0(s)$ est la probabilité neutre ou *background* de générer la séquence s. Cette dernière est généralement prise comme étant une chaîne de Markov $\mathcal{P}_k$ d'ordre k petit (0 ou 1) :

$$P_0(S_{i,j}) = \prod_{l=i}^{j} \mathcal{P}_k(S_l | S_{l-k,l-1}) \tag{3.2}$$

Ainsi, l'équation 3.1 décrit la probabilité de générer la séquence S avec le modèle *background*, sauf à la position i où un site est généré avec le modèle de fixation $\mathcal{M}$. La probabilité de générer la séquence s'obtient finalement en sommant sur les positions pondérées par la probabilité *a priori* $P(i)$ que le site soit à la position i :

$$P(S|\mathcal{M}) = \sum_{i=1}^{L-K+1} P(i) P(S|i, \mathcal{M}) \tag{3.3}$$

Cette probabilité est généralement prise uniforme, mais on peut y incorporer certaines informations *a priori*, comme le nombre de *reads* d'une expérience de ChIP-seq par exemple.

- **Apprentissage du modèle**

Maintenant que nous savons exprimer la vraisemblance d'une séquence régulée par le motif $\mathcal{M}$, nous pouvons apprendre le meilleur modèle possible l'ayant générée : c'est la maximisation de la vraisemblance. Soit un ensemble de séquences $\mathcal{S}$ constitué de M séquences co-régulées $S[1], \cdots, S[M]$. Ces séquences étant supposées indépendantes, la vraisemblance que ces données soient générées par un modèle $\mathcal{M}$ est le produit sur les séquences de la quantité $P(S[m]|\mathcal{M})$. Il est plus utile dans ce cas de regarder la log-vraisemblance, s'écrivant alors comme une somme :

$$l(\mathcal{S}|\mathcal{M}) = \sum_{m=1}^{M} \log P(S[m]|\mathcal{M}) \tag{3.4}$$

Nous désirons obtenir le modèle $\mathcal{M}$ maximisant cette quantité [16]. Cependant, nous ne connaissons pas les positions exactes des sites, qui sont des « variables cachées » et il n'existe pas de méthode d'estimation simple permettant de résoudre ce problème. C'est à ce stade qu'intervient la méthode Espérance-Maximisation (EM) (Dempster et al., 1977). L'algorithme EM est une méthode itérative qui part d'un modèle initial $\mathcal{M}^0$, permettant de calculer les poids des positions dans les séquences (étape E d'espérance), puis estime le meilleur modèle

16. La distribution *background* est supposée fixée (par exemple la chaîne de Markov peut être apprise sur un grand nombre de séquences intergéniques non codantes).

$\mathcal{M}^1$ étant données ces poids (étape M de maximisation). L'itération a lieu jusqu'à convergence vers un maximum local.

Notons $\mathcal{M}^t$ le modèle à l'itération t. La probabilité qu'un site à la position i dans la séquence $S[i]$ soit un site de fixation s'écrit $P(i|S[m], \mathcal{M}^t)$. On définit la log-vraisemblance moyenne d'un modèle $\mathcal{M}$ à l'itération t par :

$$Q(\mathcal{M}|\mathcal{M}_t, \mathcal{S}) = \sum_m \sum_i P(i|S[m], \mathcal{M}^t) \log P(\mathcal{S}, i|\mathcal{M}) \tag{3.5}$$

Le modèle suivant $\mathcal{M}^{t+1}$ est celui qui maximise cette quantité :

$$\mathcal{M}^{t+1} = \underset{\mathcal{M}}{\operatorname{argmax}} Q(\mathcal{M}|\mathcal{M}_t, \mathcal{S}) \tag{3.6}$$

L'équation 3.5 se scinde en une partie dépend de $\mathcal{M}$ et une partie *background* qui n'en dépend pas (eq. 3.1), que l'on peut dont ignorer pour ce qui est de la maximisation. Ainsi, le modèle $\mathcal{M}$ maximise la quantité suivante :

$$Q(\mathcal{M}|\mathcal{M}_t, \mathcal{S}) = \sum_m \sum_i P(i|S[m], \mathcal{M}^t) \log P(S_{i,i+K-1}|\mathcal{M}) \tag{3.7}$$

Chaque K-mer des séquences de $\mathcal{S}$ est donc pris en compte dans l'apprentissage en proportion de la croyance courante $P(i|S[m], \mathcal{M}^t)$ que c'est un site de fixation.

Pour résumer on a deux étapes :

- étape E : utiliser $\mathcal{M}^t$ pour attribuer un poids à chaque K-mer des séquences
- étape M : apprendre $\mathcal{M}^{t+1}$ qui a la plus grande vraisemblance de générer les données pondérées par $\mathcal{M}^t$.

Reste le problème de choisir un modèle initial adéquat pour ne pas converger vers un maximum de vraisemblance local. MEME adopte pour cela une approche semi-exhaustive. Les différents K-mers possibles des séquences d'apprentissage sont successivement utilisés pour générer un modèle initial. L'algorithme EM est itéré une fois, et le modèle de plus grande probabilité est finalement gardé come motif initial pour une itération complète.

3.1.2 STUBB : une méthode utilisant des motifs connus, les corrélations entre sites de fixations, et la phylogénie

L'algorithme STUBB (Sinha et al., 2003) décrit les séquence par un processus génératif de type modèle de Markov caché (HMM pour *Hidden Markov Model*). Il est basé sur l'algo-

rithme Ahab (Rajewsky et al., 2002) – lui-même basé sur l'algorithme MobyDick (Bussemaker et al., 2000) –, qui peut être vu comme une extension de MEME au cas où les séquences contiennent plusieurs sites de fixations pour différents motifs. Notamment, ces méthodes ont l'intérêt tout comme MEME de ne pas avoir de seuil arbitraire pour définir un site car elles moyennent sur toutes les segmentations possibles de la séquence. On parle notamment de modèle thermodynamique (voir 1.6.1). La différence entre STUBB et Ahab est qu'il introduit deux informations supplémentaires : les corrélations entre motifs et la phylogénie. Enfin, l'algorithme utilise comme condition initiale des PWMs de motifs connus pour être impliqués dans la co-régulation étudiée.

- **Description du modèle HMM**

Nous décrivons d'abord l'algorithme Ahab (Rajewsky et al., 2002). Notons W l'ensemble des motifs initiaux. Le modèle HMM à l'ordre 0 (HMM0) utilisé décrit la génération d'une séquence S de la manière suivante : à chaque pas, le processus choisit un motif $w_i \in W$ avec une probabilité p_i ou le motif *background* w_b avec une probabilité $1 - \sum_i p_i$. Une fois le motif w choisi, une séquence est échantillonnée à partir de la PWM de w et est ajoutée à la séquence S. Le processus est itéré jusqu'à ce que la séquence générée atteigne une taille L. La séquence de motifs choisis au cours de la procédure définit une segmentation T. La probabilité que la séquence observée soit générée par ce processus de paramètres $\theta = \{w_i, p_i\}$ est

$$P(S|\theta) = \sum_T P(S, T|\theta) \tag{3.8}$$

et peut être calculée par programmation dynamique. Le score d'une séquence est en comparant cette probabilité et la probabilité $P(S|\theta_b)$ que la séquence soit générée uniquement par le modèle *background* :

$$F(S) = \underset{\theta}{\operatorname{argmax}} \log \left(\frac{P(S|\theta)}{P(S, \theta_b)} \right) \tag{3.9}$$

Le paramètre θ (c'est-à-dire les p_i) qui maximise le membre de droite est obtenu grâce à un algorithme de type EM (Sinha et al., 2003).

- **Ajout des corrélations entre motifs**

Des informations sur les corrélations entre motifs sont introduites dans θ sous la forme de probabilités de transition p_{ij} que le motif choisi soit j lorsque le premier motif précédent

non-*background* est i. Parce que le nombre de paramètres devient grand, seules les corrélations importantes sont ajoutées.

- **Incorporation de l'information phylogénétique**

Enfin, STUBB utilise des informations provenant de la comparaison phylogénétique avec des séquences provenant d'autres espèces. Les séquences des différentes espèces sont d'abord alignées, puis la probabilité de générer l'alignement est calculée à l'aide d'un modèle phylogénétique. Ce modèle permet de prendre en compte le fait que les séquences homologues sont corrélées du fait qu'elles dérivent d'un ancêtre commun. Dans le cas de Stubb, les espèces sont supposées liées par un topologie étoilée, c'est-à-dire que les espèces partagent un seul ancêtre commun. Le modèle d'évolution suppose que les différentes bases de la séquence évoluent indépendamment, à la même fréquence, et que la probabilité de fixation d'une mutation $b \to b'$ à la position i est proportionnelle au poids de la PWM du nucléotide b à cette position. Ce modèle est identique au modèle que nous introduisons dans l'article (voir section 3.2), et qui est inspiré du modèle de Felsenstein (Felsenstein, 1981). La probabilité $P(\sigma|w)$ de générer l'alignement σ de séquences s avec le motif w de taille L_w s'écrit alors :

$$P(\sigma|w) = \prod_{i=1}^{L_w} \left[\sum_b w_{i,b} \prod_{s \in \sigma} (q_s \delta_{b,s_i} + (1 - q_s) w_{i,s_i}) \right] \tag{3.10}$$

où w_{i,s_i} est la probabilité de générer le nucléotide s_i à la position i pour le motif w, $\delta_{x,y} = 1$ si $x = y$ et 0 sinon, et $q_s = e^{-\lambda t_s}$ est la probabilité de conserver un nucléotide au cours de l'évolution, qui est une fonction du taux de mutation neutre λ et du temps d'évolution t_s entre l'ancêtre commun et l'espèce s. En résumé, pour chaque position i, un nucléotide b est généré chez l'ancêtre commun avec une probabilité $w_{i,b}$, puis ce nucléotide est soit conservé chez l'espèce s avec une probabilité q_s ou bien il mute avec une probabilité $1 - q_s$, et une nouvelle base est sélectionnée selon les poids définis par w_i. Pour des espèces proches, $q \sim 1$ et le fait d'observer des bases différentes à des positions homologues diminue fortement $P(\sigma|w)$, même si leur fréquence *a priori* donnée par w est identique : l'information phylogénétique permet dans ce cas de rejeter un « mauvais » modèle qui ne le serait pas si les séquences orthologues étaient considérées comme indépendantes. Pour des espèces lointaines, $q \sim 0$ et tout se passe comme si les séquences de σ étaient indépendantes.

3.1.3 MONKEY : vers des modèles phylogénétiques plus complexes

Dans la section précédente, nous avons vu qu'utiliser l'information phylogénétique permettait de mieux attribuer un site à un motif (cf cas où les espèces sont proches). Néanmoins, le modèle phylogénétique sous-jacent était relativement simple. L'algorithme MONKEY (Moses et al., 2004) propose d'utiliser des modèles d'évolution plus complexes pour détecter les sites conservés à partir de motifs connus.

Les motifs w sont décrits par le modèle PWM et le *background* par les fréquences des nucléotides π_b. Le score d'un site est défini par la fraction des probabilités de générer la séquence s par l'un ou l'autre des modèles (*log-likelihood ratio* ou LLR) :

$$\mathcal{S}(s) = \log \frac{P(s|w)}{P(s|\pi)} = \sum_i \log \frac{w_{i,b(i)}}{\pi_b} \tag{3.11}$$

Le but est de généraliser ce score au cas de l'alignement σ de séquences s, en réalisant cette fois le calcul sur l'ancêtre commun A des séquences. L'ancêtre commun, ainsi que tous les ancêtre communs intermédiaires pour une topologie d'arbre T quelconque, ne sont pas observés et il faut donc sommer sur tous leurs états (séquences) possibles étant donnés l'alignement observé σ, l'arbre T et le modèle d'évolution. La solution générale de ce problème a été donnée par Felsenstein (1981). Notamment, nous pouvons nous concentrer sur le cas à deux espèces, le cas général procédant par récurrence à partir de ce cas simple. Nous pouvons aussi nous concentrer sur une position i donnée, puisque les positions sont indépendantes.

Considérons un alignement de deux nucléotides s_1 et s_2. On définit un nouveau score comme étant le LLR comparant l'hypothèse que s_1 et s_2 représentent un site conservé pour un motif w et l'hypothèse que les bases ont été tirées dans le *background* :

$$\mathcal{S}_{\text{cons}}(s_1, s_2) = \log \frac{P(s_1, s_2|w, T, R_w)}{P(s_1, s_2|w, T, R_{\text{back}})} \tag{3.12}$$

où R_w et R_{back} sont des matrices de taux de transition décrivant le processus de substitution au cours de l'évolution pour le cas d'un site de fixation du motif w et pour le *background*. Ainsi la transition de la base b à la base b' s'écrit

$$p_{b \to b'} = \left(e^{Rd}\right)_{b,b'} \tag{3.13}$$

où R est la matrice de taux de transition et d le temps évolutif. Dans le cas simple à deux espèces, l'arbre T est en étoile, avec des distances d_1 et d_2 de l'ancêtre commun pour les espèces contenant les bases s_1 et s_2. Par ailleurs, les espèces évoluent indépendamment depuis leur séparation. Notant b la base sur l'ancêtre commun et M le modèle (motif ou *background*) considéré, on a finalement :

$$\begin{aligned} P(s_1, s_2|M, T, R_M) &= \sum_b P(s_1|b, d_1, R_M) P(s_2|b, d_2, R_M) P(b|M) \\ &= \sum_b \left(e^{R_M d_1}\right)_{b,s_1} \left(e^{R_M d_2}\right)_{b,s_2} P(b|M) \end{aligned} \tag{3.14}$$

Le calcul sur un nombre quelconque d'espèces se fait finalement récursivement (Felsenstein, 1981), jusqu'à la racine de l'arbre où $P(b|M)$ vaut $w_{i,b}$ pour le motif et π_b pour le *background*.

Plusieurs modèles peuvent être choisis pour les matrices de transition. Pour ce qui est du *background*, il peut être décrit par des modèles neutres. Par exemple le modèle de Felsenstein (Felsenstein, 1981) est le modèle le plus simple dont la distribution d'équilibre redonne les fréquences du *background*. Le modèle HKY (Hasegawa et al., 1985) est une variante qui inclut le fait que les mutations entre bases de même nature chimique (purine ou pyrimidine), appelées transitions, sont 2 fois plus fréquentes que les autres mutations, appelées transversions. Ce modèle est utilisé dans MONKEY pour décrire le *background*. Pour ce qui est du motif, les taux de transition dépendent de la position au sein du site : par exemple les bases dégénérées mutent plus vite que les bases très conservées (Moses et al., 2003). Pour prendre cette variation en compte, une possibilité est de modifier un modèle neutre pour utilisant à la place des fréquences *background* les fréquences données par la PWM : c'est ce qui est fait dans Stubb avec le modèle Felsenstein dans l'éq.3.10. Dans MONKEY, les auteurs utilisent un modèle plus complexe, appelé modèle HB, préalablement introduit par Halpern et Bruno pour étudier l'évolution des régions codantes (Halpern and Bruno, 1998). Dans ce modèle, le taux de substitution $r^i_{a,b}$ de la base a vers la base b en position i est à une constante près le produit de la probabilité de mutation neutre (indépendante de la position) de a vers b par une probabilité de fixation $f^i_{a,b}$:

$$r^i_{a,b} \propto Q_{a,b} f^i_{a,b} \tag{3.15}$$

où $Q = R_{\text{back}}$ est la matrice de transition du modèle *background*. La probabilité de fixation peut être obtenue en utilisant un résultat classique de génétique des populations (Kimura, 1962) :

$$f^i_{a,b} \simeq \frac{2s}{1 - e^{-2Ns}} \tag{3.16}$$

$$f^i_{b,a} \simeq \frac{-2s}{1 - e^{2Ns}} \tag{3.17}$$

où N est la taille effective de la population (le facteur 2 vient du fait que la population est diploïde), s la valeur adaptative relative de la base b par rapport à la base a en position i, et l'évolution est quasi-neutre ($2Ns \ll 1$). On a donc :

$$\frac{f^i_{a,b}}{f^i_{b,a}} \simeq e^{2Ns} \tag{3.18}$$

Par ailleurs, en supposant que les substitutions vérifient le bilan détaillé à l'équilibre, on peut écrire

$$\frac{w_{i,b} Q_{b,a}}{w_{i,a} Q_{a,b}} = \frac{f^i_{a,b}}{f^i_{b,a}} \tag{3.19}$$

Finalement, en combinant les équations 3.16, 3.18 et 3.19, on obtient la matrice de transition du modèle HB en position i sous la forme :

$$\boxed{R(i)_{a,b} = Q_{a,b} \frac{x \log x}{x - 1}} \tag{3.20}$$

où

$$x = \frac{w_{i,b} Q_{b,a}}{w_{i,a} Q_{a,b}} \simeq e^{2Ns} \tag{3.21}$$

traduit l'effet de la sélection. En développant autour de $x = 1$ et en restant au premier ordre, on a

$$\frac{x \log x}{x - 1} \simeq \frac{1}{2}(1 + x) \tag{3.22}$$

Ainsi, dans le cas neutre où $x = 1$, la matrice de transition se réduit à la matrice *background* : $R(i)_{a,b} = Q_{a,b}$. Cependant, lorsque la base b est très conservée par rapport à la base a ($x > 1$), les substitutions de a vers b sont plus fréquentes que sous le modèle neutre : $R(i)_{a,b} > Q_{a,b}$.

3.1.4 Méthodes utilisant des collections d'oligonucléotides

Alors que les méthodes basées sur l'enrichissement en k-mers présentées en 3.1.5 s'intéressent au contenu général d'un CRM en k-mers, d'autres méthodes tentent de regrouper les k-mers en groupes associés à un régulateur putatif. Par exemple, Cao et al. (2010a) ont introduit un algorithme de recherche de motifs destiné à l'étude de données ChIP-seq. Ici, un motif est simplement défini comme une collection de k-mers. Le but est de trouver les motifs qui discriminent le mieux un ensemble de séquences positives (des pics de ChIP-seq) d'un ensemble de séquences *background*. L'algorithme énumère d'abord tous les k-mers, mesure leurs fréquences, et ajuste pour chacun un modèle de régression logistique mesurant la capacité à classifier les séquences. Le k-mer le plus important est choisi comme graine. Puis toutes les variations à distance de Hamming de 1 et 2 de cette graine sont énumérées, et sont ajoutées au motif si elles permettent d'améliorer la régression. Lorsqu'un motif final est obtenu, toutes ses occurrences sont masquées et un nouveau motif est appris. Un algorithme similaire, HOMER, a été développé par Heinz et al. (2010). La différence majeure est que HOMER utilise la collection de k-mers obtenue pour générer une PWM qui est ensuite raffinée sur les séquences.

3.1.5 Approches sans motifs ou *motif-blind*

L'algorithme MEME utilise en son cœur un modèle de motif $\mathcal{M}$ permettant d'attribuer une probabilité $P(S|\mathcal{M})$ à une séquence donnée. Ce modèle peut être un modèle PWM ou un modèle plus complexe avec interactions, cf 2.2. Néanmoins, il existe certaines méthodes cherchant à décrire plus généralement la statistique des mots au sein des CRMs sans chercher à associer ces statistiques à des motifs ayant une caractérisation biochimique précise. De telles approches sont dites sans motifs (*motif-blind*). Nous en recensons ici quelques-unes (voir Kantorovitz et al. (2009) pour plus de détails).

- **Modèles basés sur des chaînes de Markov**

Plusieurs modèles basés sur des chaînes de Markov ont été proposés. Ainsi, l'algorithme PFRSampler de Grad et al. (2004) consiste en un apprentissage de modèles de Markov d'ordre 5 sur des séquences d'intérêt et sur des séquences *background*, ces séquences étant préalablement filtrées par la conservation phylogénétique. Il est ensuite possible de calculer la vraisemblance qu'une séquence donnée soit générée par l'un ou l'autre des modèles, de manière similaire à l'éq.3.3. Le score d'une séquence S de taille L est alors défini comme étant la dif-

férence des log-vraisemblances qu'elle soit générée par le modèle d'intérêt $\mathcal{M}_{\text{train}}$ et par le modèle *background* $\mathcal{M}_{\text{back}}$:

$$\text{Score}(S) = \log \frac{P(S|\mathcal{M}_{\text{train}})}{P(S|\mathcal{M}_{\text{back}})} = \sum_{i=1}^{L} \log \frac{T_{\text{train}}(S_i|S_{i-k,i-1})}{T_{\text{back}}(S_i|S_{i-k,i-1})} \tag{3.23}$$

où T_{train} et T_{back} sont les probabilités de transition associées au deux modèles, $S_{i,j}$ est la séquence entre les positions i et j incluses, et k est l'ordre de la chaîne de Markov (ici $k = 5$). Cette méthode détecte donc la *signature* globale d'un CRM plutôt que la présence de sites de fixation pour des TFs particuliers. Cette méthode a aussi été implémentée par Ivan et al. (2008) sous le nom de *Markov Chain Discrimination* (MCD), avec la différence notable que les auteurs n'utilisent pas la phylogénie. Une généralisation de cette approche a été proposée par (Kazemian et al., 2011) sous le nom d'*Interpolated Markov Model*. Au lieu d'utiliser une chaîne de Markov à un ordre donné, les auteurs réalisent une interpolation entre des chaînes de Markov d'ordres 0 à 5, en ne gardant pour chaque ordre que les transitions importantes. Ceci leur permet de capturer les signatures présentes à différentes résolutions.

- **Modèles basés sur des enrichissements en k-mers**

D'autres modèles sont basés sur la statistique des mots de k nucléotides (k-mers) dans les séquences d'apprentissage. Par exemple, Kantorovitz et al. (2007) ont introduit une mesure de similarité entre séquences basée sur le nombre de k-mers qu'elles ont en commun. Les auteurs définissent le score D_2 par

$$D_2(S_1, S_2) = \sum_{\{w\}} N_1(w) N_2(w) \tag{3.24}$$

où S_i est la séquence i, $\{w\}$ est l'ensemble des k-mers, et $N_i(w)$ est le nombre de k-mers w dans la séquence i. Ce score est grand si les séquences partagent de nombreux k-mers, ie si elles ont une régulation commune. Ce score est normalisé pour produire le z-score de mesure de similarité $D2z$:

$$D2z(S_1, S_2) = \frac{D_2(S_1, S_2) - E(D_2)}{\sigma(D_2)} \tag{3.25}$$

où $E(D_2)$ et $\sigma(D_2)$ sont l'espérance et l'écart-type de la distribution de $D_2(S_1, S_2)$ calculées théoriquement sous l'hypothèse que les séquences S_1 et S_2 sont indépendantes et sont générées par un modèle *background* de type chaîne de Markov.

D'autres méthodes pour attribuer un score à une séquence par similarité à des séquences d'apprentissage ont été introduites par Kantorovitz et al. (2009). Étant données des séquences d'apprentissage, les 200 k-mers ($k = 6$) les plus représentés par rapport à un modèle *background* sont sélectionnés selon leur z score, c'est-à-dire le nombre d'écarts-type séparant le nombre $n(w)$ de fois que le mots apparaît dans le training set du nombre de fois moyen $\lambda(W)$ qu'il devrait apparaître sous un modèle *background* (Sinha and Tompa, 2000). Étant donnés ces mots sur-représentés, il est possible de définir un score basé sur la statistique de Poisson (modèle PAC pour *Poisson Additive Conditional*) :

$$PAC(S) = \frac{1}{200} \sum_{w} F(\lambda(w), n(w) - 1) \tag{3.26}$$

où $F(\lambda, x)$ est la distribution de Poisson cumulative de paramètre λ, donnant une valeur faible (proche de 0) si $n(w) \simeq \lambda(w)$ et maximale (proche de 1) si $n(w) \gg \lambda(w)$. D'autres scores sont définis par une approche de classification linéaire pondérant les comptes de k-mers (WSC pour *Weighted Sum of Counts*)

$$WSC(S) = \sum_{w} \beta(w) n(w) \tag{3.27}$$

où $\beta(w)$ est un poids reflétant l'association avec l'ensemble d'apprentissage. Ce poids peut être le rapport de la fréquence du mot dans l'ensemble d'apprentissage et de sa fréquence dans le *background* (modèle HexDiff, Chan and Kibler (2005)), le logarithme de cette quantité (Rouault et al., 2010), ou encore le z score introduit précédemment mesurant la sur-représentation du k-mer dans l'ensemble d'apprentissage (méthode HexYMF, Kantorovitz et al. (2009)).

3.2 Article

Imogene: identification of motifs and cis regulatory modules underlying gene co-regulation

Hervé Rouault[1,2,+], Marc Santolini[3,+] , François Schweisguth[1,2] and Vincent Hakim[3]

[1] Institut Pasteur, Developmental Biology Department,75015 Paris, France
[2] CNRS, URA2578, F-75015 Paris, France
[3] Laboratoire de Physique Statistique, CNRS, Université P. et M. Curie, Université Paris-Diderot, École Normale Supérieure, Paris, France.
[+] Have contributed equally

Email: Hervé Rouault - herve.rouault@pasteur.fr; Marc Santolini - santolin@lps.ens.fr; François Schweisguth*- francois.schweiguth@pasteur.fr; Vincent Hakim*- hakim@lps.ens.fr;

*Corresponding author

April 17, 2013

Abstract

Cis-regulatory modules (CRMs) and motifs play a central role in tissue and condition-specific gene expression. Their identification could be facilitated by the development of suitable bio-informatic tools. Here we present and test *Imogene* an algorithm that we have implemented in a publicly available software (http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::imogene). Starting from a small training set of mammalian or fly CRMs that drive similar gene expression profiles, *Imogene* determines *de novo* cis-regulatory motifs that underlie this co-expression as well as to predict on a genome wide scale other CRMs with a regulatory potential similar to the training set. The algorithm makes central use of information provided by other sequenced genomes, based on previously developed statistical tools and explicit evolutionary models. We further show, using two mouse neural and limb specific CRM sets as well as CRMs active during fly mesoderm development, that CRMs related to different developmental programs can be distinguished on the basis of *Imogene de novo* generated motifs. We thus expect *Imogene* to be a useful tool to decipher transcriptional gene regulation in higher eukaryotes.

Background

The identification and functional characterization of the non-coding sequences that direct the spatio-temporal specificity of gene expression in eukaryotes is of fundamental importance in developmental biology [1] and can find crucial applications in medicine [2]. These regulatory sequences are generally located distally from gene promoters and termed enhancers or more generically cis-regulatory modules (CRMs) since they can either enhance or repress gene expression [3]. They usually are of the order of 500 nucleotides (nts) long and can be located as far as several mega base-pairs away from the transcription start sites (TSSs) of the genes that they regulate. CRMs are composed of transcription factor binding sites (TFBSs) which bring spatio-temporal

specificity to the expression of their target promoters [4]. Detailed studies in both flies and vertebrates [5] have shown that CRMs contains multiple binding sites for transcription factors that can be either identical (homotypic clustering) or different (heterotypic clustering). Homotopic clustering can provide cooperatively and sharp on-off gene expression whereas heterotypic clustering allows for combinatorial gene regulation. The extent to which the order and relative positioning of the different TFBSs in CRMs matter, remains however debated [6, 7].

With the advent of ChIP-seq techniques, genome-wide studies are providing large amount of data on the binding loci of tissue-specific transcription factors [8], as well as on other factors that regulate transcription e. g. by modifying chromatin structure (p300, CTCF, histone marks, etc) [9, 10]. This protein binding data has helped the identification of numerous CRMs specific of well-defined developmental processes and it has brought important information on CRM structure. However, genome wide studies suffer from limitations. A full characterization of regulatory mechanisms would require ChIP-seq analysis to be performed for every potential regulatory factor, on every tissue, at multiple developmental stages. The results would also have to be obtained for the often heterogeneous cells that constitute the tissue of interest instead of being averaged over them as it usually needs to be the case. Finally, and very importantly, binding cannot be equated to functional regulation.

Therefore, *in silico* identification of CRMs form a useful complement to genome-wide binding studies. By classic case-by-case studies or through larger scale analyses [11], as previously described, several CRMs have been identified as active players in the co-regulation of a subset of genes, in specific biological systems or in the formation of different organs at various stages of developement. Identifying the important binding sites on these known sequences would help to bypass some of the limitations of large scale studies by providing information on the factor involved, both known and new, as well as on the existence of a regulatory grammar. It should also help one to determine other CRMs providing specific expression patterns, a difficult task at present given the absence of close association [12] between CRMs and their target genes. These labor-intensive experimental tasks could be eased by bioinformatics. To this end, we have previously developed [13] statistical tools to determine cis-regulatory elements *de novo*, in a set of input DNA sequences encoding a common transcriptional regulation. They allow the determination of regulatory elements from input DNA sequences without any prior information on the transcription factors acting in cis or on their binding sites. They make by central use of the phylogenetic information contained in the aligned DNA sequences of related species. The method was applied to the *D. melanogaster* gene expression program in sensory organ precursor cell (SOPs), a specific type of neural progenitor cells [13]. Predicted motifs included already characterized TFBS as well as new motifs and were successfully tested by mutational analysis. These motifs were used to rank intergenic DNA fragments genome wide for their regulatory potential in SOPs. Of the top 29 predicted CRMs, 38% were found by transgenic assays to direct transcription in SOP. A larger fraction (65%) drove more generally transcription in neural precursors.

This successful application to a *Drosophila* transcriptional program led us to wonder how the tools developed in ref. [13] would perform in the case of mammalian CRMs. The tool performance certainly needed to be assessed since the task of determining cis-regulatory elements is considerably more challenging for mammalian genomes, which are are an order of magnitude richer in intergenic sequences than *Drosophila* ones. To tackle this challenge, we have developed *Imogene*, a computer algorithm and software that we present here and characterize. *Imogene* aims at:

1. predicting cis-regulatory sequences (of about 10 nt long) responsible for specific gene co-regulation within these CRMs, as well as to build a set of Probability Weight Matrices (PWM) or motifs [14, 15] characterizing the DNA-binding specificity of the associated putative factors.

2. predicting novel CRMs at the genomic scale with the same expression pattern as the starting set of CRMs, based on the set of build PWMs.

Imogene is based on the statistical tools introduced in ref. [13] and described in detail there. It extends this previous version by:
i) allowing the use of a refined evolutionary model,
ii) including the eutherian genomes and correspond-

ing alignment data, as described below,
iii) being accessible through the publicly available interface that we have developed and that we here describe.

In the following, the general methodoly of *Imogene* is first presented. Then *Imogene* performance is tested on two sets of mammalian CRMs pertaining to neural tube and limb developmental programs during embryogenesis. We then consider the distinct but related task of discriminating CRMs with different specificities, rather than discriminating a set of specific CRM from backgroung intergenic sequences. The discrimination of the two sets of mammalian neural tube and limb CRMs is first addressed. To further assess the performance of *Imogene*, it is applied to the discrimination of five sets of mesodermal fly CRMs, a task previously considered in ref. [16]. Finally, the developed publicly available *Imogene* interface is presented.

Results and Discussion

Description of Imogene

Imogene has two modes that can be used in succession, as sketched on Figure 1 and summarized here (see *Methods* for details of their implementation).

The first mode, *genmot*, aims at extracting statistically meaningful PWMs from a "training set" of functionally related CRMs on a reference genome (the mouse *M. musculus* genome for mammals; the *D. melanogaster* genome for flies). The size of the training set could in principle be unlimited, but in practice computer execution time requires it to stay below 100 Kbp. It should also be above a few Kbp to provide a sufficient amount of information (a training set of about 20 Kbp appears as a good compromise). Starting from a chosen training set, *Genmot* performs its task in two steps (I and II in Figure 1): I. *Genmot* first enlarges the training set with orthologous sequences in other related sequenced genomes, shown in Figure 2 (for the mouse, the 11 other aligned eutherian sequenced genomes with high coverage presently available on the Ensembl project [43]) the 11 other *Drosophilae* sequenced genomes [44] for the fly). This comparative genomics step results in the creation of the "enlarged training set" (step I in Figure 1).

II. In this second central step, *Genmot* build PWMs of given length ℓ (10 nt is the default value) by scanning the training set, in an iterative manner (step II in Figure 1). Each sequence of ℓ nucleotides in the training set is used in turn to create an initial PWM using a Bayesian prior. This PWM is then refined by scanning the training set to find all the PWM binding sites in the training set, i.e. all ℓ nucleotide long sequences in the training set that have a binding score above a generation threshold score S_g, chosen at the procedure onset ($S_g = 13$ bits is the default value). These binding sites are filtered using conservation, that is only sites that have orthologues in distant species are kept. A shift in alignment between a binding site on the reference species and its orthologues in other species to correct for eventual alignment errors (20nt is the shift default value). The ensemble of conserved binding sites and their orthologues serve, using an evolutionary model, to build a refined PWM. The procedure is then iterated by finding the binding sites of the refined PWM and using them to build a further refined PWM until convergence to a stable set of binding sites.

The need of an evolutionary model to properly assemble binding sites [25,26,45] is simply explained. A binding site in the reference genome and its orthologues are all related through descent from their last common ancestor, and cannot therefore be considered as independent observations. In order to correctly quantify the amount of information provided by the observation of orthologous sites, one has to estimate their potential of change through mutation since their last common ancestor. To account for this, *Imogene* can, in its present implementation, make use of either one of two evolutionary models of TFBS evolution at the user choice. The first option, *"Felsenstein"*, is a simple and computationally fast model proposed in [45]. Mutations are generated at the same rate in a PWM binding site than in the background intergenic sequences. However, the mutated nucleotide in a binding site is drawn according to its frequency in the PWM at the mutated position. This is analogous to the simplest model of DNA evolution [46] but with nucleotides neutral relative abundances replaced by PWM nucleotide frequencies. This *Felsenstein* model is the simplest model that provides at evolutionary equilibrium, nucleotide frequencies that agree with those prescribed by the PWM at the different positions in the binding site. The second option, *"Halpern-Bruno"* [47] uses an evolutionary model that is more complex than the *Felsenstein* model and that has previously been used for TFBS evolution in [25]. It allows for the inclusion of different mutational prob-

abilities between different bases in the neutral background intergenic mutation model. Additionally, it includes a fitness-dependent fixation probability for a mutation in a TFBS, based on classical population genetics estimates for the fixation of a mutant allele appearing in an homogeneous population [48]. The relative fitnesses of different nucleotides are determined by the requirement that binding site convergence to evolutionary equilibrium leads to the PWM nucleotide frequencies (see Methods for details).

The described procedure produces a PWM for every ℓ nucleotide long sequence in the training set. In a series of final steps, this long list is pruned and ranked based on comparing the PWM bindings sites on the training set to a "background" set of intergenic sequences in the reference genome (20 Mb of *M. Musculus* or *D. melanogaster* genomic DNA). The PWM corresponding to repeated sequences are first eliminated on the basis of the non-poissonian distribution on their binding sites in the background set. Then for each remaining PWM, the distribution of its conserved binding sequences on the training set is compared to the distribution of the PWM conserved binding sequences on a set of background intergenic sequences. The larger the statistical deviation between the two distributions, the larger its score and the more meaningful the PWM is deemed. In a final step, PWMs in the ranked list are compared and, among similar ones, only the highest scoring one is kept. Although the identity of the transcription factors corresponding to the different PWMs of interest is not directly assessable by the algorithm, the comparison between the produced PWMs and existing databases can provide relevant information on their identity, as will be shown in the following sections.

In its second mode, *scangen*, *Imogene* determines intergenic sequences in the reference genome that are considered as putative CRMs with the same functional specificity as the training set. This second mode (step III in Figure 1) is based on the inferred PWMs in the *genmot* mode. The algorithm scan the entire non-coding repeat-masked reference genome and find all the conserved binding sites above the scanning binding score S_s for the N first PWMs in the ranked list. The intergenic sequences of a given length (the default value is 1000 nt) are then scored according to their similarity to the training set in their content of PWM binding sites. The closest the similarity in its motif content with the training set, the most likely an intergenic sequences is deemed to be functionally related to the training set.

Application to eutherian developmental programs

In order to assess *Imogene* performance on mammalian transcriptional regulation, we applied it to two sets of mammalian specific CRMs, that have previously been identified starting from p300 Chip-seq data and functionally tested in a transient transgenic assay for activity in stage 10 mouse embryo [11, 49]. We chose CRMs active in neural tube and limb, as characterized in the VISTA website (http://enhancer.lbl.gov). For each developmental program, a subset of CRMs was visually selected for specificity and strength of expression in the tissue of interest, from the provided expression pattern. Among these selected sets, 2 limb CRMs and 4 neural tube CRMs contained no sequence that could possibly be used to learn motifs by *Imogene*, due to its conservation requirements, either because of repeat masking or because of low conservation (see *Methods*). Elimination of these uninformative sequences produced curated training sets of 29 neural and 39 limb CRMs.

A cross-validation scheme was then used to measure *Imogene* predictability power (see *Methods* for details). In brief, for each developmental program, the CRMs were divided into a training set composed of 15 CRMs chosen at random, and a test set composed of the other CRMs used as True Positives. These test CRMs were ranked against a 'background test set', a set of ~ 60 regions of 1Kb taken from the flanking sequences of the initial set of CRMs (see *Methods*).

The training set was used for motifs generation using *Imogene genmot* mode. This procedure was conducted for both evolutionary models using different values of the generation parameter S_g and scanning threshold S_s to obtain the optimal values of these parameters for each model and each training set(see Figure 3 and Figure S1). For different parameter sets, the test CRMs as well as the intergenic sequences of the background set were scored. The proportion of retrieved test set CRMs above a given score (True Positive Rate or TPR) was plotted against the proportion of appearing test background regions above the same score (False Positive Rate or FPR) as this score decreased, to produce a so-called ROC curve [**?**]. The ROC curves corresponding to different parameters values were then compared using the Area Under ROC Curve (AUC), a quantity

that is maximal at best prediction.
Figure S1 shows the AUC as a function of the number of motifs N for different values of the scanning threshold S_s. One can see that the AUC increases quickly with the 5 first motifs generated, and has nearly converged to its maximum value when 10 motifs are kept. Therefore we restricted ourselves to $N = 10$ motifs, and constrained the other parameters using AUC maximization. Figure 3 shows the ROC curves obtained for the optimal parameters which are seen to be similar for both models and both training sets. Figure S1 shows how changing N impacts these ROC curves, making it clear that $N = 10$ is already nearly optimal.
For the limb CRMs, 60% of the test set CRMs are retrieved at 5% FPR whereas an even larger proportion of 70% is obtained for the neural tube CRMs. The *Halpern-Bruno* and the *Felsenstein* models are seen in Figure 3 to yield very similar results in both cases, with a slight superiority for the *Halpern-Bruno* one. It should be noted that the chosen measure really provides a lower estimate of *Imogene* success rate since we considered as 'False Positives'. all flanking CRMs sequences, whereas, in reality, some could be *bona fide* positive CRMs.

In the cross-validation procedure, different ranked lists of motifs were created for each randomly drawn test set. In order to provide a list of motifs generated by the algorithm, we ran *Imogene* on the full set of CRMs for each class. The corresponding 10 best motifs are shown in Figure S2. The closest TRANSFAC PWM assigned to each motif by *Imogene* PWM distance is also shown in Figure S2 . Previously characterized motifs belonging to the considered developmental programs appear in each class (e.g. Oct and NeuroD motif in the neural CRMs). The motif content of each CRM is also provided in Figures S3, S4. It is seen that the 10 best motifs appear on most CRMs of the training set.

Discrimination of tissue-specific CRMs in the mouse

Given *Imogene* ability to distinguish specific CRMs from background sequences, we found it interesting to apply it to the related but distinct task of distinguishing different classes of CRMs. The question was previously considered for *D. melanogaster* CRMs based on ChIP-seq data at different developmental time points [16], as detailed in the next section. It consists in learning features that distinguishes the CRMs of a given class from the CRMs of other classes, in order to be able to predict the class of a newly observed CRM. The task differs from distinguishing CRMs from background intergenic sequences since learning motifs shared among different classes, for instance characterizing the binding of generic CRM factors, is of no use for discrimination purposes. As a test case, we considered the neural tube and limb sets of mammalian CRMs used in the previous section. Given the nature of the task, we selected in each set the CRMs with an expression that appeared mostly restricted to neural tube and limb. This yielded 12 neural and 15 limb CRMs.

As in ref. [16], we used a leave-one-out cross validation (LOOCV) scheme in which the learning set constituted all but one of the elements of a class, the remaining one being used as a test sequence. The process can be summarized as follow. We call the class of interest the positive class and the classes against which we wish to learn the negative classes. The LOOCV process begins with the exclusion of a (positive or negative) CRM which serves as an unobserved test CRM. Then, a set of N motifs is learnt on the remaining CRMs of each class, yielding positive and negative motifs. These motifs are used to build a weighted score giving positive (resp. negative) contributions to positive (resp. negative) motifs (see *Methods*). Finally, the rank of the test CRM among all CRMs is kept as an indicator of the classification. We expect positive CRMs to be on top of the list and have low ranks while negative CRMs should be attributed high ranks. After processing for all CRMs, we have a list of ranks for the positive and negative CRMs that can be represented by a ROC curve indicating the True Positives Rate and False Positive Rate for increasing rank. We optimized parameters (the threshold for motifs generation S_g, the threshold for sequences scanning S_s, and the number of motifs N used to score sequences) by maximizing the Area Under the ROC Curve for a FPR ≤ 0.2.

The results are shown in Figure 4. We focus on the results obtained with the *Halpern-Bruno* evolutionary model, which does slightly better than the *Felsenstein* model, as for the previous CRMs ranking task. Results (motifs, thresholds) are nonetheless very comparable in the two cases. Motifs are shown on the right of the ROC plots and were generated on the positive classes with optimal parameters. The two classes were optimally discriminated using only 2 motifs in each class, with specifities $S_g = 11$, $S_s = 8$, comparable to that found in the learning

task of the previous section.The best ranking motif of the neural CRMs was found to be unequivocally associated to the Transfac Oct1/Pou2f3 Transcription Factor, known to be involved in the neural tube formation [51].

Discrimination of Drosophila tissue-specific CRMs

In order to further test the discriminating power of *Imogene de novo* generated motifs, we applied it to the CRM classification task reported in ref. [16]. In this work, previously characterized *D. melanogaster* CRM were divided in 5 classes corresponding to the different tissue types in which they were active: mesoderm (Meso), somatic muscle (SM), visceral muscle (VM), mesoderm and somatic muscle (Meso & SM) and visceral and somatic muscle (VM & SM). Ref. [16] made use of a collection of Chip-seq binding data for different factors and at different developmental time points to attribute to each CRM a total of 15 peak height values. It was then tested whether classical machine learning techniques could be used to discriminate the different CRM classes, on the basis of these extensive data. This was indeed found possible with a high success rate in a standard cross-validation scheme: CRMs predicted with probability higher than 95% to belong to a given class were indeed found to belong to that class with a high success rate of 80%.

This led us to wonder whether *Imogene* would succeed in classifying these different CRMs, without using any binding data, but rather on the basis of combinations of *de novo* motifs that it would itself generate. We used the set of well-characterized CRMs belonging to 5 different classes assembled in ref. [16]. We then proceeded similarly to the previous case with eutherian CRMs.

Imogene results are shown together with the machine learning results of ref [16] in Figure 5. For clarity, we here show results obtained with the *Felsenstein* model. Results obtained with the *Halpern-Bruno* model are comparable. Strikingly, without any binding data *Imogene* prediction rates are comparable to the machine learning ones, in the high specificity range (FPR $\leq 5\%$) used for CRM prediction in [16]. Its performance is even better for the Meso and SM classes at high score. The latter case is of particular interest. The machine learning algorithm essentially used Mef2 ChIPseq peak heights to predict SM CRMs, resulting in an incorrect classification at high scores since this TF is required for the differentiation of all muscle types. However, the use of the specific Mef2 motif obtained *de novo* from the SM training set allows one to restore a correct classification at high score (Figure 5C).

On the side of each ROC plot, the *de novo* motifs generated on the whole training set are displayed. The number of motifs shown is the optimal number used for CRM scoring in the leave-one-out cross-validation. Among the generated motifs, one can recognize 4/5 TFs for which ChIPseq data was used in [16], namely Twist (motif 2, Meso & SM), Mef2 (motif 1, SM), Bin and Tin (motifs 1 and 2, VM). The Bap motif was not found by the algorithm, and correspondly it was not shown to be of importance in ref. [16].

in summary, our analysis indicates that *Imogen* can not only identify *de novo* functionnally relevant binding sites within a set of CRMs but can also be used to identify the more subtle differences in binding sites that underlie functional differences between related sets of CRMs.

Web interface

The described algorithm is available through a user-friendly web interface that provides motif and CRM predictions for the community. This interface is powered by the Pasteur Institute Internet server through the mobyle framework [52]. The input web page and an example output web page are shown in Figure 6 and 7.

The input form (see Figure 6) is divided into several sections. One of the two available algorithm modes should be chosen at start:

- genmot: given a list of coordinates of typically 15 enhancers of 1 kb (training set), generates *de novo* motifs ranked by p-value.
- scangen: given the previously generated motifs, produces a list of genome-wide predicted CRMs with conserved binding sites. The rank of a CRM is based on a poissonian score that takes into account the motif content, as explained in [13].

The group of species considered should also be specified. The algorithm can be used on Drosophilae (with reference species *D. melanogaster*) or mammals (with reference species *Mus musculus*). The different algorithm parameters such as the seeked motif width, threshold specificity for binding sites or

allowed position shifts between different species (see *Methods* for a detailed description) are set by default to values that have been found to provide reasonable results. They can be modified by the user to optimize the results for the considered training set.

In mode *genmot*, the user should enter the training set CRM coordinates. The chosen evolutionary model for the TFBS should also be specified. The *Felsenstein* mode is computationnally faster than the *Halpern-Bruno* one. The results of the two modes have been found to be comparable, with a slight superiority in the the performance of the *Halpern-Bruno* model (see Figure 3 and 4).

In mode *scangen*, the algorithm scores and ranks intergenic sequences in the reference species, using a list of motifs, as described in the first *Results* section and in *Methods*. The list of *de novo genmot* motifs is used as input. The user can set the length of the ranked sequences (1 Kb is the default value) and the number of scoring motifs (5 is the default value for computational speed but this can be changed to improve results, see Figure S1)

An example of *Imogene* output is displayed in Figure 7. The *genmot* mode creates from the provided training set a list of ranked motifs together with their significance and over-representations (see *Methods*). The positions of these motifs on the CRM of the training set and on their homologous sequences in other species are also provided, as illustrated in Figure 7A for 2 motifs. Figure 7B shows the output of the *scangen* mode for these two motifs. The ordered list of best-ranking intergenic sequences is given together with information on the closest TSSs.

Conclusions

We have presented *Imogene*, a computer algorithm able to predict *de novo* relevant motifs in functionally related sets of CRMs and able to infer novel CRMs with a low false positive rate in both drosophilae and mammalian genomes. *Imogene* mode of inference internally makes use of quantitative models for binding site evolution. This allows it to systematically exploits the information available in multiple sequenced-genomes, and to work efficiently from a CRM set of modest size.

Numerous algorithms have already been developed to try and map cis-regulatory information underlying transcriptional regulation (see e.g. [3,14,17–19] for recent reviews). *Imogene* differs from previous methods in several respects. It has been specially designed to decode cis-regulatory regulation in a small set of CRMs while other algorithms are aimed at the analysis of large datasets such as whole ChIP-seq peak regions [20]. It works *de novo* while many algorithms make use of well-characterized binding motifs [21–29]. Several that work *de novo* try and distinguish regulatory sequences by their entire content in short nucleotide sequences [30–35]. Phylogenetic conservation between multiple sequenced genomes has been shown to provide useful information on cis-regulatory motifs [36–38], but cannot *per se* address the question of specific spatio-temporal expression. In order to analyze a small set of CRMs with well-characterized expression, *Imogene* makes full use of several sequenced genomes instead of focusing a single genome [29] or simply comparing it with another one [39–41]. Moreover, it does not simply add orthologous sequences [42] but uses an evolutionary model to properly weigh this additional information.

The algorithm which lies at *Imogene* core was previously applied to gene co-regulation in *Drosophila* [13]. Motifs predicted to be important for Sensory-Organ-Precursors development were confirmed by site-directed mutagenesis. A significant fraction of top predicted new CRMs were also shown to direct expression in SOP or more generally in the peripheral nervous system. The ability of the algorithm to provide meaningful information on cis-regulatory elements in *Drosophila* was further confirmed in a subsequent application to epidermal morphogenesis and trichome development [53]. The algorithm provided an informative PWM for the master regulator Ovo/Shavenbaby and predicted as well a functionally important novel motif.

In spite of its successful application to gene co-regulation in *Drosophila*, it was not clear that *Imogene* would be able to decipher cis-regulatory information in the notoriously more difficult case of mammalian gene expression. We have here provided bio-informatic evidence that *Imogene* provide meaningful results in this case also. *Imogene* was shown to successfully recognize CRMs belonging to neural and limb development programs solely based on motifs that it has constructed *de novo* on other CRMs. Furthermore, the created PWMs appear to comprise both known and new motifs, in strong analogy with the previous studied cases in the fly.

There is currently numerous cases for which a

small number of CRMs belonging to the same program of gene expression has been characterized. Therefore, the possibility to use *Imogene* should provide helpful service to the community. We have further shown that *Imogene* can discriminate between classes of CRMs.In this task, it should play a complementary role to ChIP-seq data that are currently obtained for many developmental programs. Whereas, ChIP-seq provides information on the binding of already known factors, *Imogene* is able to propose new PWMs and help to identify new involved DNA-binding cofactors and their binding sites. We thus believe that *Imogene* is a useful complement to existing softwares [19]. A user-friendly version has been made it publicly available on the Pasteur Institute web platform http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::imogene. The full computer code is also freely available at http://github.com/hrouault/Imogene.

Acknowledgments

We wish to thank I Leroux, S Meilhac and B Robert who helped us to characterize the patterns of expression of the mammalian CRMs used in the present work. We acknowledge the Centre d'Informatique pour la Biologie at the Pasteur institute for its help in the design of a mobyle front-end to *Imogene*. This work was supported by core funding from Centre National de la Recherche Scientifique, Ecole Normale Supérieure and Institut Pasteur and by a specific grant from the Agence Nationale pour la Recherche (ANR-08-BLAN-0235).

Methods

Genome alignments

The alignments were downloaded from ftp://ftp.ensembl.org/pub/release-63/emf/ensembl-compara/epo_12_eutherian for eutherians and from http://www.biostat.wisc.edu/~cdewey/fly_CAF1/data for Drosophilae. For the latter case, we have used the alignments engineered by A. Caspi with the help of the Mercator and MAVID programs. In both cases, the alignments were processed through a customized script to produce alignments in fasta format, mask for coding sequences (CDS) and simple repeats (see below).

Annotations

The CDS coordinates were downloaded from ftp://ftp.ensembl.org/pub/release-64/gtf/mus_musculus for eutherians (mm9 coordinates) and from ftp://ftp.flybase.net/releases/FB2011_06/dmel_r5.38/gff for Drosophilae (release 5 coordinates). In the case of eutherians, the TSS coordinates were obtained separately from http://hgdownload.cse.ucsc.edu/goldenPath/mm9/database. Eutherian alignments repeat sequences were already masked for repeat sequences, appearing as small capital nucleotides in the emf files. Drosophilae alignments were masked using the coordinates indicated in the gff file.

Background sequences

Imogene computes the statistical overrepresentation of the predicted motifs by comparing them to 20 Mb of background intergenic region. The script that generates the random coordinates is included in the distribution of *Imogene* as well as the actual coordinates of the produced intergenic regions.

Training sets

The two mammalian training sets (limb, neural tube) were obtained from http://enhancer.lbl.gov, based on the work of [11, 49]. They were manually curated to produce a high-quality data set, with respectively 41 CRMs for the limb, and 33 for the neural tube. We further pruned out uninformative CRMs for which no motifs could be generated, either because of repeat masking or because of lack of conservation. More precisely, the reference species sequence was scanned using a window size corresponding the motif size. If a sequence did not contain any masked nucleotide, we looked in the other species for any unmasked sequence in the surrounding neighborhood of ± 20nt, our flexibility criterium when defining a conserved instance. If putative orthologous sequences were found in enough species to satisfy our conservation requirements (see below), the site was declared as a putative conserved site for a regulatory motif. This filtering step resulted in final sets of 39 CRMs for the limb and 29 for the neural tube. The Drosophilae training sets were obtained from [16]. Coordinate files are given as Supplemental Material.

Mammalian predictions

Training and background test sets For each class, the CRMs were divided into a learning set composed of 15 CRMs chosen at random, the other CRMs (~ 20) defining the test set of 'True positives'. In addition, a set of background test regions was built using the 1Kb flanking sequences of the full list of CRMs Such an 'adapted' background test set was used to provide a more stringent and informative test of the algorithm by preventing discrimination on the training set from the background test set, based on other features than the seeked high-information-content motifs , such as a local composition bias. Furthermore, in order to avoid biaising the results towards the true positives, uninformative sequences for *Imogene* (i.e sequences where no binding site could possibly be found given *Imogene* conservation requirements) were also removed from this background test set. These regions were also filtered for uninformative elements, This yielded background test sets of 72 CRMs for the limb and 57 for the neural tube.

Cross-validation protocol The learning set was used to learn the motifs content. The 10 best motifs were then used to score test set CRMs and background regions. Because the length of the training set CRMs could vary, we decided to keep for each test sequence the best scoring 1kb fragment. This process was repeated 40 times, and both generation and scanning threshold were varied. The retrieval rate of test set CRMs (True Positives) among background elements (False Positives) as a function of the score was used to build a ROC curve. The Area Under ROC Curve or AUC, a quantity that varies between 0 for absolute misclassification, 0.5 for random classification, to 1 for perfect classification, was used to evaluate the quality of prediction. The parameter set yielding the highest AUC was chosen as the best set.

Leave-one-out cross-validation for the CRM discrination task.

Let us note $\mathcal{C}_i$ the tissue class of interest. There are M_i corresponding CRMs. Let N_c denote the total number of classes. Our goal is to find the particular motif signature that distinguishes these M_i CRMs from the $N_c - 1$ other classes of CRMs. This signature corresponds in our case to a number N of top ranked motifs with generation and scanning thresholds S_g and S_s. These are the three parameters we wish to constrain with a leave-one-out cross-validation (LOOCV) procedure.

Let us detail this procedure in the case where we distinguish class $\mathcal{C}_i$ from the other classes $\mathcal{C}_j$. The M_i CRMs of $\mathcal{C}_i$ are termed 'positive' CRMs and the M_j CRMs of each of the other classes are termed 'negative' CRMs. Let us note $M = \sum_i M_i$ the total number of CRMs. The LOOCV consists in withdrawing one 'test' CRM from these M CRMs, learn the motifs on the $M - 1$ resulting CRMs, and use them to score the let alone test CRM. For the learning step, motifs are generated with threshold S_g on each class (one class being deprived of one CRM), yielding N_c sets of motifs: one set of positive motifs from class $\mathcal{C}_i$ and $N_c - 1$ sets of negative motifs from the other classes. The N top ranked motifs from each set are then used to scan the M CRMs for conserved instances with scanning threshold S_s. Each CRM E is scored with respect to these N_c sets of motifs by:

$$S(E) = \sum_{j=1}^{N_c} (2\delta_{j,i} - 1) S_N^{\mathcal{C}_j}(E) \qquad (1)$$

where $S_N^{\mathcal{C}_j}(E)$ is the CRM score for the N top motifs of class $\mathcal{C}_j$ as defined later in the *Methods* section, and $\delta_{j,i} = 1$ if $j = i$, and 0 otherwise. This score simply gives positive contributions if positive motifs are found on the CRM, and negative contributions if negative motifs are found. This scoring procedure allows to rank the test CRM among the other $M - 1$ CRMs. Ties are resolved by using the mean rank among equally scored CRMs. Here, the rank is used rather than the raw score of the test CRM to avoid any artifact stemming from normalization problems. Indeed, the raw score is dependent on the generated motifs, which differ at each step of the LOOCV. This procedure is repeated over all M CRMs, yielding a corresponding list of M ranks. This list is finally used to build a ROC curve discriminating True Positives (CRMs from class $\mathcal{C}_i$) from False Positives (the other CRMs). The discrimination is quantified by the area under the ROC curve for a False Positive Rate FPR $\leq 20\%$, which we note AUC20 and that we want to maximize.

In our case, we used a $2D$ parameters grid with S_g varying between 7 and 13 bits by steps of 1, and S_s varying between $S_g - 5$ and S_g by steps of 1. Both *Felsenstein* and *Halpern-Bruno* models were used for motif generation. For each parameter set, the number of motifs used for scanning was increased from 1 to a maximum number of 10 (actually never at-

tained) until the addition of a new motif decreased the AUC20, yielding an optimal number of motifs N. Finally, for each class, the parameter set $\{S_g, S_s, N\}$ yielding the highest AUC20 was selected as the best parameter set.

Motifs identification

In order to identify the known TFs that might correspond to the *de novo* generated motifs, we used Transfac database [54]. In order to avoid uninformative matches, we kept Transfac motifs that had an information content greater than 8 bits, a threshold approximately corresponding to 4 conserved nucleotides. This gets rid of 170 vertebrate motifs and 32 insect motifs, yielding a total of respectively 765 and 37 motifs.

Each *de novo* motif was compared to all Transfac motifs from the corresponding clade (vertebrates or insects) using the PWM distance introduced in [13]. During the comparison, motifs are shifted to find the best match, with a minimal match of 5 nts. The shift is simply introduced by adding flanking nucleotides with background frequency on either side. the closest candidate was kept for identification.

Main program

The main program is written in C++ and adapted from the program used in a previous study [13]. It is distributed under the GNU GPL licence and available as a git repository at http://github.com/hrouault/Imogene.

Binding site and CRM scores

Binding sites as well as CRMs are scored in the same manner as in [13].

For a given PWM with the weight $w_{i,b}$ for the base b at position i, the score of a sequence s_i is defined as:

$$S = \sum_i w_{i,s_i} \tag{2}$$

A sequence is considered as a binding site when $S > S_{th}$, where S_{th} is the score threshold defined by the user of *Imogene*.

An CRM E is scored with respect to a set of motifs m_i by:

$$S(E) = \sum_i n(E, m_i) \log(\lambda_i^t / \lambda_i^b) \tag{3}$$

where $n(E, m_i)$ is the number of binding sites for the motif m_i on E and λ_i^t, λ_i^b and the average number of binding sites per base on the training set and background respectively. It is important to note that the previously found motif binding sites are masked when scanning with successive motifs. Thus motifs with lower ranks that resemble high-ranking motifs, but could not be fused properly, do not increase artificially the CRM weight by predicting the same binding sequences twice.

Evolutionary models

Imogene can use two different evolutionary models, which vary in complexity and computational time. To simplify the computation, we suppose in both models that the bases within a site evolve independently from each other.

Felsenstein The simplest models of nucleotides evolution are copied on model of neutral selection. This procedure has been proposed by Sinha *et al* [26, 45] with the Felsenstein model of neutral evolution [46]. In this TFBS evolution model, the transition probability from nucleotide b to b' at position i in two sites at evolutionary distance d writes:

$$p^i_{b \to b'} = q\ \delta_{b,b'} + (1-q)\ w_{i,b'} \tag{4}$$

where $\delta_{b,b'}$ is the Kronecker symbol, $w_{i,b'}$ is the mean frequency of base b at position i of the site (as given by the PWM model), and q is the probability of conservation for an evolutionary distance d under neutral selection (see below).

When two species are close to one another, $q \sim 1$ and the probability that the observed bases are identical is high. On the contrary, when the two considered species are distant ($q \sim 0$), the observed bases are uncorrelated and reflect the PWM probabilities $w_{i,b}$.

The probability of conservation q can then be computed within this model by setting the PWM probabilities $w_{i,b}$ to the mean genomic frequencies π_b:

$$q = \exp\left(-\frac{d}{1/2 + 4\pi_{\mathrm{A,T}}\pi_{\mathrm{C,G}}}\right) \tag{5}$$

with $\pi_{\mathrm{A,T}}$ (resp. $\pi_{\mathrm{C,G}}$) the common genomic frequency of A and T (resp. G and C).

Halpern-Bruno The Halpern-Bruno model (HB) differs in two ways from the simplest *Felsenstein* model. It uses the more complex Hasegawa, Kishino and Yano model (HKY) [55] for the neutral evolution of nucleotides and adds a fixation probability based on fitness differences for the evolution of nucleotides within the TFBS.

The HKY model improves on the Felsenstein model by taking into account the observed dependance of the mutation rate on the chemical nature of the bases. Mutations between bases of the same chemical nature (purine or pyrimidine), also called transitions, are generally more frequent than the other type of mutations, called transversions. This is encapsulated in the HKY model by the parameter κ which is the ratio of transition rate to the transversion rate. It is measured to be $\kappa = 2$ in flies and $\kappa = 3.7$ in mammals.

Within a TFBS, the HB model extends the HKY model to take into account an additional purifying selection on the nucleotide identities. It is formulated by the following transition probabilities:

$$p_{b\to b'} = \exp(t\mathbf{H})_{b,b'} \tag{6}$$

where $\mathbf{H}$ is the rate matrix defined by:

$$H_{b,b'} = \begin{cases} \pi_b \; h_{b'\to b} & \text{if } b \neq b' \\ -\sum_{b'\neq b} H_{b,b'} & \text{if } b = b' \end{cases} \tag{7}$$

The evolutionary time t is expressed in term of the evolutionary distance by:

$$t = \frac{d}{1/2 + 4\kappa\, \pi_{\mathrm{A,T}}\pi_{\mathrm{C,G}}} \tag{8}$$

Finally, the transition rates are defined by:

$$h_{b\to b'} = \frac{w_{\mathrm{b'}}}{\pi_{\mathrm{B'}}} \; \frac{\log\left(\frac{\pi_{\mathrm{B}} w_{\mathrm{b'}}}{\pi_{\mathrm{B'}} w_{\mathrm{b}}}\right)}{w_{\mathrm{b'}}/\pi_{\mathrm{B'}} - w_{\mathrm{b}}/\pi_{\mathrm{B}}} \alpha_{b\to b'} \tag{9}$$

with $\alpha_{b\to b'} = \kappa$ for a transition and $\alpha_{b\to b'} = 1$ for a tranversion.

Inference

The algorithm performs Bayesian inference in order to infer the frequencies $w_{i,b}$ based on observations of binding sites, as previously described in [13]. In the Bayesian framework, one can write the Posterior distribution of w_i given the observation of a set of aligned nucleotides $\{\mathcal{A}\}$ as

$$\mathcal{P}(w_i|\{\mathcal{A}\}) \;\propto \prod_{a\in\{\mathcal{A}\}} \mathcal{P}(a|w_{i,b}) \prod_{b\in\{A,T,C,G\}} w_{i,b}^{\alpha_b - 1} \tag{10}$$

where we omit the normalization factor. The first product is the likelihood function and the second one is the prior, taken to be a Dirichlet distribution with pseudo-counts α_β computed as in [13]. In the idealistic case where the aligned nucleotides would represent independent observations (infinitely distant species), the likelihood reduces to a multinomial distribution and the posterior writes:

$$\mathcal{P}(w_i|\{\mathcal{A}\}) \;\propto \prod_{b\in\{A,T,C,G\}} w_{i,b}^{N_b + \alpha_b - 1} \tag{11}$$

where N_b is the number of times the base b is observed in $\{\mathcal{A}\}$. This formula allows simple analytic formulations for the estimator of mean and maximum posterior probability. The estimator of the mean posterior distribution is expressed as:

$$\tilde{w}_{i,b} = \frac{N_b + \alpha_b}{\sum_b N_b + \alpha_b} \tag{12}$$

The estimator of maximum probability has the same value if one uses the "transformed" posterior where $\alpha_b \to \alpha_b + 1$.

In our case, we take into account the evolutionary history that correlates the orthologous sites to the one on the reference species. In that case, $\mathcal{P}(a|w_{i,b})$ is a polynomial function of the $w_{i,b}$'s and generally lacks a simple analytical formulation.

Mean Posterior Estimation

The transformed posterior distribution is maximized by using a simplex algorithm implemented by the GNU GSL to fit the mean estimator. The initial value for the estimation is taken to be the mean estimator in the independent species regime given in Eq. (11). This allows to start close to the quadratic region and ensures fast convergence. The estimation was consistently retrieved when using an independent MCMC approach to compute the mean estimator.

A simple example of nucleotide inference using the two evolutionary models

To illustrate the inference of ancestral nucleotides and the main features of the two models, we con-

sider a dinucleotidic genome with bases X and Y and the simple alignment shown in Figure S5 with an ancestral species at equal evolutionary distance from the reference species and a daughter species. We suppose that the observed nucleotide at position i of an observed binding site is X both in the reference and the orthologous species.

Our goal is to infer the frequencies w_Y and $w_X = 1 - w_Y$. First, there are two simple cases. For $d = 0$, the observations of the same nucleotide in the two evolutionary branches really constitute only one observation of X. On the contrary, for very long evolutionary branches $d \to \infty$, the two instances of nucleotide X form two independent observations. Using the previous result (Eq. (12)) with $\alpha_X = \alpha_Y = \alpha$, the estimator of the maximum transformed posterior distribution for N_X and N_y independent instances of X and Y is:

$$w_Y = \frac{N_Y + \alpha}{N_Y + N_X + 2\alpha} \tag{13}$$

Thus, for $d = 0$, the inferred frequency is:

$$w_Y = \frac{\alpha}{1 + 2\alpha} \tag{14}$$

while for $d \to \infty$, it tends toward:

$$w_Y = \frac{\alpha}{2 + 2\alpha} \tag{15}$$

Between these two extreme cases, an evolutionary model has to be used to estimate w_Y, for finite evolutionary branches of length d.

For the Felsenstein model, the likelihood function writes:

$$\begin{aligned} \mathcal{P}(\mathcal{A}|w) &= w_X \left[q + (1-q)w_X\right]^2 + w_Y (1-q)^2 w_X^2 \\ &= q^2 w_X + (1-q^2) w_X^2 \end{aligned} \tag{16}$$

where $\mathcal{A}$ stands for the simple alignment drawn on fig. S5 and we used $w_X = 1 - w_Y$. From this expression it can clearly be seen that the evolutionary model simply interpolates between the independent species case ($d \to \infty$, $q = 0$) where there are two observations of base X: $\mathcal{P}(w|\mathcal{A}) = w_X^2$, and the fully correlated case ($d = 0$, $q = 1$) where the two species merge and we have only one observation: $\mathcal{P}(w|\mathcal{A}) = w_X$. The corresponding mean, $w_{Y,me}$ and maximum likelihood, $w_{Y,ma}$ analytic estimates for finite d read

$$\begin{aligned} w_{Y,me} &= \frac{\alpha}{2} \frac{1+q^2}{\alpha + 1 + \alpha q^2} \\ w_{Y,ma} &= \frac{1}{4(\alpha+1)(1-q^2)} \Big[3\alpha + 2 - (\alpha+1)q^2 \\ &- \sqrt{[\alpha + 2 - 3(\alpha+1)q^2]^2 + 8q^2(1-q^2)(\alpha+1)^2}\Big] \end{aligned}$$

Note that for the maximum likelihood estimate, $w_{Y,ma}$, the prior exponent $\alpha + 1$ has been used instead of α as explained above. So, the two estimates coincides at $q = 0$ and $q = 1$. Both estimates are plotted as of function of the evolutionary distance d in Figure S5 ($\alpha = 0.1$).

For the Halpern-Bruno model, the analogous results have been computed numerically and are also shown for comparison in Figure S5. The Halpern-Bruno model results are seen to be closer to the large distance limit than the Felsenstein model ones. Moreover, the difference between the nature of the estimates is seen to dominate the difference between the evolutionary models.

Phylogenetic trees

The phylogenetic trees used within *Imogene* are displayed in Figure 2. For drosophilae, the distances are taken from Heger and Pontig [56]. For eutherian, they are extracted from the Ensembl website (www.ensembl.org).

Conservation requirements

Imogene builds PWM form binding sites that have conserved instances in different species. The conservation requirements is that orthologous instances are found in at least 3 distant species, including the reference species. For mammals, the 5 following groups of related species are composed of: *Mus musculus* and *Rattus norvegicus*; *Callithrix jacchus*, *Macaca mulatta*, *Pongo abelii*, *Gorilla gorilla*, *Homo sapiens* and *Pan troglodytes*; *Bos taurus*; *Sus scrofa*; *Canis familiaris*; *Equus caballus*. Similarly for flies, there are 5 groups composed of: *Drosophilae melanogaster*, *sechellia*, *simulans*, *yakuba* and *erecta*; *Drosophila ananassae*; *Drosophilae pseudoobscura* and *persimilis*; *Drosophila willistoni*; *Drosophilae grimshawi*, *mojavensis* and *virilis*.

A site instance must be found in at least 3 of these 5 groups (with an allowed shift of up to 20

nt with the reference species)to be considered conserved by *Imogene*.

Selection of CRMs

A number N of motifs are used to scan the genome for conserved instances above a given threshold. Instances are ranked according to their genomic position and grouped in successive CRMs of size L such as to maximize clustering. The position E_i of the center of the enhancer i is chosen to be the center of the motifs cluster:

$$E_i = \frac{X_1 + X_N + w - 1}{2} \tag{17}$$

where X_1 and X_N are the starting positions of the first and last TFBSs in the cluster and w is the width of the motif.

Statistical analysis

All statistical analyses were performed using R [57].

Authors contributions

H. R., M. S., F. S., V. H. designed research. H. R., M. S. performed research and wrote the software. H. R., M. S., F. S., V. H. wrote the paper.

References

1. Davidson EH: *The regulatory genome: gene regulatory networks in development and evolution.* Burlington, MA: Academic 2006, [http://www.loc.gov/catdir/enhancements/fy0668/2006445256-d.html].
2. Dorer DE, Nettelbeck DM: **Targeting cancer by transcriptional control in cancer gene therapy and viral oncolysis**. *Adv Drug Deliv Rev* 2009, **61**(7-8):554–71.
3. Hardison RC, Taylor J: **Genomic approaches towards finding cis-regulatory modules in animals**. *Nat. Rev. Genet.* 2012, **13**(7):469–483.
4. Lelli KM, Slattery M, Mann RS: **Disentangling the many layers of eukaryotic transcriptional regulation**. *Annu. Rev. Genet.* 2012, **46**:43–68.
5. Levine M: **Transcriptional enhancers in animal development and evolution**. *Curr. Biol.* 2010, **20**(17):R754–763.
6. Arnosti DN, Kulkarni MM: **Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?** *J Cell Biochem* 2005, **94**(5):890–8.
7. Swanson CI, Evans NC, Barolo S: **Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer**. *Dev Cell* 2010, **18**(3):359–70.
8. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions**. *Science* 2007, **316**(5830):1497–502.
9. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome**. *Cell* 2007, **129**(4):823–37.
10. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells**. *Nature* 2007, **448**(7153):553–60.
11. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA: **ChIP-seq accurately predicts tissue-specific activity of enhancers**. *Nature* 2009, **457**(7231):854–8.
12. Amano T, Sagai T, Tanabe H, Mizushina Y, Nakazawa H, Shiroishi T: **Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription**. *Dev. Cell* 2009, **16**:47–57.
13. Rouault H, Mazouni K, Couturier L, Hakim V, Schweisguth F: **Genome-wide identification of cis-regulatory motifs and modules underlying gene coregulation using statistics and phylogeny**. *Proc Natl Acad Sci U S A* 2010, **107**(33):14615–20.
14. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements**. *Nat. Rev. Genet.* 2004, **5**(4):276–287.
15. Stormo G, Fields D: **Specificity, free energy and information content in protein–DNA interactions**. *Trends in biochemical sciences* 1998, **23**(3):109–113.
16. Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EE: **Combinatorial binding predicts spatio-temporal cis-regulatory activity**. *Nature* 2009, **462**:65–70.
17. Su J, Teichmann SA, Down TA: **Assessing computational methods of cis-regulatory module prediction**. *PLoS Comput. Biol.* 2010, **6**(12):e1001020.
18. Elnitski L, Jin VX, Farnham PJ, Jones SJ: **Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques**. *Genome Res.* 2006, **16**(12):1455–1464.
19. Aerts S: **Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets**. *Curr. Top. Dev. Biol.* 2012, **98**:121–145.
20. Machanick P, Bailey TL: **MEME-ChIP: motif analysis of large DNA datasets**. *Bioinformatics* 2011, **27**(12):1696–1697.

21. Berman B, Nibu Y, Pfeiffer B, Tomancak P, Celniker S, Levine M, Rubin G, Eisen M: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome**. *Proceedings of the National Academy of Sciences* 2002, **99**(2):757.

22. Halfon M, Grad Y, Church G, Michelson A: **Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model**. *Genome research* 2002, **12**(7):1019.

23. Rebeiz M, Reeves N, Posakony J: **SCORE: A computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data**. *Proceedings of the National Academy of Sciences* 2002, **99**(15):9888.

24. Schroeder M, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia E, Gaul U: **Transcriptional control in the segmentation gene network of Drosophila**. *PLoS biology* 2004, **2**:1396–1410.

25. Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB: **MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model**. *Genome Biol* 2004, **5**(12):R98.

26. Siddharthan R, Siggia E, van Nimwegen E: **PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny**. *PLoS Comput Biol* 2005, **1**(7):e67.

27. Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J: **Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity**. *Cell* 2006, **124**:47–59.

28. Pierstorff N, Bergman C, Wiehe T: **Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA**. *Bioinformatics* 2006, **22**(23):2858.

29. Herrmann C, Van de Sande B, Potier D, Aerts S: **i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules**. *Nucleic Acids Res* 2012.

30. Nazina A, Papatsenko D: **Statistical extraction of Drosophila cis-regulatory modules using exhaustive assessment of local word frequency**. *BMC bioinformatics* 2003, **4**:65.

31. Abnizova I, te Boekhorst R, Walter K, Gilks W: **Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the Drosophila genome: the fluffy-tail test**. *BMC bioinformatics* 2005, **6**:109.

32. Chan B, Kibler D: **Using hexamers to predict cis-regulatory motifs in Drosophila**. *BMC bioinformatics* 2005, **6**:262.

33. Leung G, Eisen M, Provart N: **Identifying Cis-Regulatory Sequences by Word Profile Similarity**. *PLoS ONE* 2009, **4**(9):e6901.

34. Kantorovitz M, Kazemian M, Kinston S, Miranda-Saavedra D, Zhu Q, Robinson G, G"ottgens B, Halfon M, Sinha S: **Motif-Blind, Genome-Wide Discovery of cis-Regulatory Modules in Drosophila and Mouse**. *Developmental Cell* 2009, **17**(4):568–579.

35. Brody T, Yavatkar AS, Kuzin A, Kundu M, Tyson LJ, Ross J, Lin TY, Lee CH, Awasaki T, Lee T, Odenwald WF: **Use of a Drosophila genome-wide conserved sequence database to identify functionally related cis-regulatory enhancers**. *Dev Dyn* 2012, **241**:169–89.

36. Xie X, Lu J, Kulbokas E, Golub T, Mootha V, Lindblad-Toh K, Lander E, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals**. *Nature* 2005, **434**(7031):338–345.

37. Ettwiller L, Paten B, Souren M, Loosli F, Wittbrodt J, Birney E: **The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates**. *Genome Biology* 2005, **6**(12):R104.

38. Stark A, Lin M, Kheradpour P, Pedersen J, Parts L, Carlson J, Crosby M, Rasmussen M, Roy S, Deoras A, et al.: **Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures.** *Nature* 2007, **450**(7167):219.

39. Wang T, Stormo G: **Combining phylogenetic data with co-regulated genes to identify regulatory motifs**. *Bioinformatics* 2003, **19**(18):2369.

40. Grad Y, Roth F, Halfon M, Church G: **Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D. pseudoobscura**. *Bioinformatics* 2004, **20**(16):2738.

41. Zhao G, Schriefer L, Stormo G: **Identification of muscle-specific regulatory modules in Caenorhabditis elegans**. *Genome research* 2007, **17**(3):348.

42. Busser BW, Taher L, Kim Y, Tansey T, Bloom MJ, Ovcharenko I, Michelson AM: **A machine learning approach for identifying novel cell type-specific transcriptional regulators of myogenesis**. *PLoS Genet* 2012, **8**(3):e1002531.

43. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Harrow J, Herrero J, Hubbard TJ, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SM: **Ensembl 2012**. *Nucleic Acids Res.* 2012, **40**(Database issue):84–90.

44. Clark A, Eisen M, Smith D, Bergman C, Oliver B, Markow T, Kaufman T, Kellis M, Gelbart W, Iyer V, et al.: **Evolution of genes and genomes on the Drosophila phylogeny**. *Nature* 2007, **450**(7167):203–218.

45. Sinha S, van Nimwegen E, Siggia ED: **A probabilistic method to detect regulatory modules**. *Bioinformatics* 2003, **19 Suppl 1**:i292–301.

46. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach**. *J Mol Evol* 1981, **17**(6):368–76.

47. Halpern AL, Bruno WJ: **Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies**. *Mol Biol Evol* 1998, **15**(7):910–7.

48. Kimura M: **On the probability of fixation of mutant genes in a population**. *Genetics* 1962, **47**:713–9.

49. May D, Blow MJ, Kaplan T, McCulley DJ, Jensen BC, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Afzal V, Simpson PC, Rubin EM, Black BL, Bristow J, Pennacchio LA, Visel A: **Large-scale discovery of enhancers from human heart tissue**. *Nat. Genet.* 2011, **44**:89–93.

50. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B: **A map of the cis-regulatory sequences in the mouse genome**. *Nature* 2012.

51. Kiyota T, Kato A, Altmann CR, Kato Y: **The POU homeobox protein Oct-1 regulates radial glia formation downstream of Notch signaling**. *Dev Biol* 2008, **315**(2):579–92.

52. Neron B, Menager H, Maufrais C, Joly N, Maupetit J, Letort S, Carrere S, Tuffery P, Letondal C: **Mobyle: a new full web bioinformatics framework**. *Bioinformatics* 2009, **25**(22):3005–3011.

53. Menoret D, Santolini M, $\cdots$, Payre F, Plaza S: **Decoding the transcriptional program of epidermal cell morphogenesis**. *(Submitted)* 2012.

54. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes**. *Nucleic Acids Res* 2006, **34**(Database issue):D108–10.

55. Hasegawa M, Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA**. *J Mol Evol* 1985, **22**(2):160–74.

56. Heger A, Ponting CP: **Variable strength of translational selection among 12 Drosophila species**. *Genetics* 2007, **177**:1337–1348.

57. R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2011, [http://www.R-project.org/]. [ISBN 3-900051-07-0].

Figures

Figure 1: ***Imogene* workflow.** The algorithm takes as input a list of functionally related CRMs. Homologous sequences from closely related species are automatically retrieved (I) and scanned in order to generate a list of putative transcription factor motifs (II). These motifs fuel the last step consisting in the inference of related novel CRMs (III). These predicted CRMs can finally be compared to a set of test CRMs to evaluate the predictability power (IV).

Figure 2: **Phylogenetic trees and phylogenetic distances used by *Imogene*.** The branch lengths represent the evolutionary distances d used by the evolutionary models at the motif construction stage.

Figure 3: **Analysis of well characterized developmental processes.** We tested the algorithm on mammal CRMs expressed at E11.5 in neural tube (A) and limb (B). For each class, CRMs were divided into a training set and a test set. Motifs were learned on the training set and used to score CRMs from the test set along with background regions consisting of the CRMs 1kb flanking sequences. Finally, True Positive Rates or TPR (resp. False Positive Rates or FPR) were defined as the proportion of test set CRMs (resp. background sequences) recovered above a given score. ROC plots summarize the results averaged over 40 trials. Insets emphasize the FPR $\leq 5\%$ region. Evolutionary models, along with thresholds S_g and S_s used for motifs generation and sequences scanning are indicated. F and HB models respectively stand for Felsenstein and Halpern-Bruno models. Black dashed lines show random discrimination.

Figure 4: **Pattern recognition (mammals).** ROC plots showing the discrimination between limb and neural CRMs. Neural and limb classes are compared to each other. Thick lines correspond to a leave-one-out cross-validation (LOOCV) scheme with a score function based on the *de novo* generated motifs from *Imogene*, while colored dashed lines represent the discrimination based on the whole training set, systematically showing overfitting compare to LOOCV. Two evolutionary models are used: Felsenstein (solid dark red line, $S_g = 11$, $S_s = 9$) and Halpern-Bruno (solid light red line, $S_g = 11$, $S_s = 8$). Black dashed line show random discrimination.

Figure 5: **Pattern recognition (Drosophilae).** Recognition of classes of CRMs expressed in 5 tissue types: mesoderm (meso), somatic muscle (sm), visceral muscle (vm) , mesoderm and somatic muscle (meso & sm) and visceral and somatic muscle (vm & sm). ROC plots are obtained using a leave-one-out cross-validation scheme. Two classifiers are compared: a Support Vector Machine using 15 ChIPseq peak heights (grey), and *Imogene* using the *de novo* generated motifs with Felsenstein evolutionary model (red). The following thresholds were used: meso ($S_g = 12$, $S_s = 12$), meso & sm ($S_g = 10$, $S_s = 10$), sm ($S_g = 9$, $S_s = 4$), vm ($S_g = 10$, $S_s = 10$), vm & sm ($S_g = 11$, $S_s = 8$).

Figure 6: **Web based interface : input web page.** A copy input web page for *Imogene* powered by the mobyle bioinformatics framework is shown.

A

Motifs

Color	Rank	Logo	P-value (log10)	Over-representation
	1		-79.4772	55.122
	2		-76.2578	38.1757

Motifs instances in the training set

>MusMus MRPS9(intragenic)_1_42945168_42946091 1 42945168 42946091

```
CAACTTGTTA  CACGGATGGG  TTGCACGCAG  CGAAGCTGTG  GAAAATCTGT  GCCTTTTAAC
TTTTCTACTT  AATCACGGTT  GTAGCATTGC  CTTTAGACTG  TATGCTACAT  TAATTCTCTT
CCTGCCTTCT  GCCTTCATCC  CAAGTTTCAC  GGGAAAAAGT  AAAGTGTGCA  GGTCTTACAG
AGGAGCCTTA  TCAAACAGCT  GTCATCTGAC  AAGCCATTTG  CATTTGTTTT  GGCTGAAATG
GAGCAACCCA  AGGGCAAGAT  CTTTTGTTGC  ATTCCATCAT  AATGAAGAAA  TTACACATTG
TGTAAGAGGC  CTGGCTTTAT  TTTTAGTTTG  CTTGTGTGCT  TTAAAAGGTA  TTGCTCCAGA
AACTGATGGG  ATAGAATTTT  ACCG
```

Motifs presence in alignments

MRPS9(intragenic)_1_42945168_42946091

B

Score	Coordinate	Closest TSS	Relative distance to closest TSS (bp)	5 surrounding TSSs
48.1146	chr15:81014639-81015638	Mkl1	7048	Sgsm3;Mkl1;Mkl1;4930483J18Rik;Mchr1;
34.2492	chr3:143836754-143837753	Lmo4	29042	A830019L24Rik;Gm6260;Lmo4;Lmo4;Lmo4;
34.2492	chr12:51291776-51292775	Prkd1	458934	Foxg1;3110039M20Rik;Prkd1;G2e3;Scfd1;
33.8818	chr14:23564465-23565464	Gm10248	349828	Zfp503;1700112E06Rik;Gm10248;Kcnma1;Dlg5;
30.9743	chr2:63807707-63808706	Fign	128862	Gca;Kcnh7;Fign;Grb14;Cobll1;

Figure 7: **Web based interface : output web page.** Example of an output web page for *Imogene* powered by the mobyle bioinformatics framework. A. Result page for the genmot mode. Two motifs were generated from the neural tube full training set (default is 5), using the same parameters as in Figure 3. Results are shown for the training set sequence MRPS9(intragenic). For display purposes, the beginning of the sequence, which contains no instances for the motifs, was cut in the middle panel. In the alignments, thick lines correspond to sequences and thin lines to gaps. B. Result page for the scangen mode. The two generated motifs were used to score putative regulatory sequences of 1kb in the mouse genome at optimal threshold $S_s = 10$. The 5 best ranking sequences are shown (default is 200).

Supplementary Figures

Figure S1: **Dependence of the predictions on the number of scoring motifs** ROC plots obtained at optimal scanning threshold using the Halpern-Bruno evolutionary model are shown for the neural tube (A) and limb (B) cases. Different curves are shown corresponding to sequences scored with different number of motifs: 1, 5 and 40 (light-color dashed lines), 10 (thick line). The ROC curves obtained for 10 motifs correspond to the ones shown in Fig. 3. To assess the degree of convergence, we computed the Area Under ROC Curve as a function of the number of motifs used (A',B',C'). We show the curves corresponding to the choice of different scanning thresholds S_s. In all cases, 10 motifs were sufficient for the AUC to converge. The optimal S_s was chosen as the one maximizing the AUC for 10 motifs.

Figure S2: **Motifs learnt on the full training sets.** The 10 best ranking motifs generated on the three CRMs training sets are shown together with the closest Transfac motifs (see *Methods* for a description of motif distance computation)

	Mot1	Mot2	Mot3	Mot4	Mot5	Mot6	Mot7	Mot8	Mot9	Mot10
ZIC4-ZIC1_9_91261697_91263041	2	0	1	0	0	0	2	1	1	0
TCF4(intragenic)_18_69658816_69660452	0	0	0	1	1	2	0	0	0	0
CEI-IRX1_13_72435297_72436784	3	4	2	2	0	3	0	1	3	2
NBEA(intragenic)_3_55768657_55770664	0	1	1	2	1	0	0	0	3	1
AKT3(intragenic)_1_179080168_179081586	2	1	1	3	2	1	2	0	2	0
FOXG1B-PRKD1_12_51291542_51292872	4	2	1	0	2	0	0	3	1	0
DACH1(intragenic)_14_98553917_98556433	5	0	2	4	0	2	3	1	3	2
FAM44A-CPEB2_5_42914188_42915270	1	0	1	3	1	1	2	0	1	0
IRX4-IRX2_13_73170587_73173631	0	0	0	2	0	0	0	0	4	0
EBF1(intragenic)_11_44469978_44471372	2	3	1	1	0	3	4	1	0	0
ATG4C-FOXD3_4_99240573_99241457	0	0	0	0	0	0	0	0	0	0
CYLD-SALL1_8_91462919_91464123	0	0	1	1	1	0	0	0	1	0
POU2F1(intragenic)_1_167864366_167866439	5	0	4	1	0	0	0	3	0	3
APG4C-FOXD3_4_99040833_99042291	0	0	0	0	0	0	0	0	0	0
MGC14798-HH114_2_115363420_115365044	2	2	2	0	1	0	0	2	0	0
MGST1-LMO3_6_138199417_138201368	5	1	1	1	1	3	0	3	1	1
APG4C-FOXD3_4_98961102_98962673	2	2	2	2	3	1	4	0	0	0
FLJ46321-RASEF_4_73149468_73150526	0	0	2	4	0	1	1	1	0	1
TCF12(intragenic)_9_71823775_71824538	1	0	0	1	2	0	1	0	0	1
BMPER(intragenic)_9_23182371_23184296	2	1	1	1	0	2	1	0	1	0
SOX21-ABCC4_14_118834760_118836087	1	6	2	1	3	0	1	3	3	2
FANCL-BCL11A_11_25256346_25257683	0	2	1	0	3	0	2	0	0	0
DERA(intragenic)_6_137772070_137773298	1	5	0	1	1	1	2	0	0	0
MRPS9(intragenic)_1_42945168_42946091	1	1	0	2	1	1	1	0	0	0
YTHDF3-BHLHB5_3_16776170_16778776	2	2	0	1	0	0	0	0	1	0
STXBP6-NOVA1_12_47121350_47122759	1	4	2	3	0	2	2	0	0	0
IDH3B-CPXM1_2_130177541_130178125	0	0	0	0	0	0	0	0	0	0
LOC347487-SOX3_X_57972482_57973750	3	0	1	2	1	1	2	2	1	3

Figure S3: **Neural CRMs and motifs.** List of the neural CRMs used in this study. The number of motifs of different types on each CRM is given for the 10 best-ranking neural motifs shown in Figure S2
.

	Mot1	Mot2	Mot3	Mot4	Mot5	Mot6	Mot7	Mot8	Mot9	Mot10
hs1435_7_106105018_106107143	1	1	2	2	0	3	3	0	3	0
hs126_14_97485454_97486724	5	1	2	0	0	2	1	3	1	0
hs1477_2_59400401_59401189	2	0	1	1	2	1	1	1	1	0
hs521_1_91610325_91611486	0	1	2	4	0	1	0	0	8	0
mm422_2_4477190_4478921	0	0	1	1	0	0	0	0	0	0
hs1432_13_91326599_91329775	0	0	0	1	0	0	0	0	0	0
hs1433_3_30003454_30008202	8	4	8	5	5	5	4	5	6	1
hs208_9_100171947_100173392	2	2	3	3	5	1	1	1	4	2
hs1507_1_75765578_75770167	1	0	5	4	3	0	1	0	7	1
hs774_3_5329674_5330756	4	2	1	0	0	2	0	2	0	0
hs919_15_50496379_50498196	3	1	1	0	2	1	1	1	2	3
hs326_19_45568075_45569359	1	0	4	1	2	3	3	0	0	1
hs72_8_91978407_91979282	1	1	2	3	2	2	0	0	2	1
hs1484_4_97888231_97891318	0	1	0	0	2	0	0	1	1	0
mm423_2_4508631_4509808	0	0	1	0	0	0	0	0	0	0
mm428_5_38308981_38309833	0	2	1	0	0	0	1	0	4	0
hs741_3_66874217_66875516	4	2	1	0	1	2	0	2	1	0
hs1148_12_119941220_119942766	0	0	1	0	0	0	0	0	0	0
hs1109_13_79503055_79504129	2	1	1	1	0	1	1	0	1	0
hs2041_9_96280544_96283360	2	0	0	0	0	0	1	0	0	0
hs1473_13_56260379_56262548	1	1	7	1	8	0	0	0	1	0
hs1434_14_23833434_23842485	1	1	7	5	3	0	4	2	4	3
hs1465_6_51144711_51148222	0	2	6	3	1	1	1	0	3	0
mm94_6_122342623_122346341	0	0	2	1	1	0	0	0	3	0
hs1452_10_45612931_45614502	0	0	0	0	0	0	2	0	1	2
hs1468_10_125358093_125366026	0	0	1	0	0	1	0	0	0	0
hs1586_13_15640807_15642666	0	1	1	1	1	2	0	0	3	0
hs1273_12_9344323_9346407	2	2	2	1	3	4	1	4	3	1
hs1278_2_137073444_137074711	1	1	5	3	0	0	0	1	0	1
hs1500_14_22281464_22282917	0	0	4	1	2	0	0	0	2	0
mm458_15_63025492_63026343	2	0	1	0	0	1	0	0	4	0
hs388_12_26576441_26577229	4	4	2	2	1	0	0	0	0	1
hs1491_14_25804749_25806653	1	0	6	0	6	0	0	0	3	3
hs1428_3_99469238_99471067	0	2	4	2	2	0	1	0	3	0
hs1430_6_52917020_52919645	5	1	4	1	2	1	1	0	2	0
hs1475_16_72685882_72688547	0	0	1	0	1	0	1	4	0	1
hs1448_2_171555881_171562133	1	3	5	1	0	0	1	0	1	0
hs644_12_34884495_34885741	0	5	4	1	0	1	1	0	2	0

Figure S4: **Limb CRMs and motifs.** List of the limb CRMs used in this study. The number of motifs of different types on each CRM is given for the 10 best-ranking limb motifs shown in Figure S2
.

Figure S5: **Simple example of motif inference with Felsenstein and Halpern-Bruno evolutionary models** The inference of an ancestral base is compared in the simple case of two species at a phylogenic distance d from their common ancestor, for a two nucleotide alphabet,X and Y. The mean and maximum likelihood estimate of observing Y in the common ancestor given that the two species share an X is shown as a function of evolutionary distance d, for the Felsenstein or Halpern-Bruno evolutionary models. The likelihood is always smaller with the Halpern-Bruno model, reflecting the model greater evolutionary rate.

3.3 Calcul de la moyenne de la postérieure par une méthode MCMC

Nous avons voulu savoir si l'approximation réalisée par Imogene lors du calcul du maximum de l'estimateur de la postérieure modifiée donnait un résultat effectivement proche du calcul de la moyenne de l'estimateur de la postérieure non modifiée $\mathcal{P}(w_i|\{\mathcal{A}\})$, où w_i est le vecteur de poids de la PWM à la position i et $\{\mathcal{A}\}$ est l'ensemble des alignements de nucléotides observés à cette position dans les sites de fixation. Néanmoins, l'estimation de la moyenne de cette distribution est difficile, puisque nous n'avons pas de moyen simple de l'échantillonner. Afin de contourner ce problème, nous avons eu recours à une méthode de Monte-Carlo par chaînes de Markov ou MCMC basée sur l'algorithme de Metropolis-Hastings (Krauth, 2006).

3.3.1 Principe de l'algorithme de Metropolis-Hastings

L'algorithme de Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) permet d'échantillonner la distribution de la distribution postérieure en utilisant le parcours d'une chaîne de Markov ayant cette distribution pour loi stationnaire. Une tel processus de Markov est défini par des probabilités de transition $P(w \to w')$ entre deux états w et w'. Il converge vers une distribution stationnaire $\pi(w)$ unique sous deux conditions : (1) les transitions sont réversibles et le processus satisfait le bilan détaillé $\pi(w)P(w \to w') = \pi(w')P(w' \to w)$, (2) le processus est ergodique, c'est-à-dire que tout état est accessible et qu'il n'y a pas de cycles. L'algorithme de Metropolis-Hastings repose sur la construction d'une chaîne de Markov ayant ces propriétés et dont la distribution d'équilibre $\pi(w)$ est la probabilité que l'on cherche à échantillonner $P(w)$. Pour cela, on part de l'équation du bilan détaillé, que l'on peut écrire

$$\frac{P(w \to w')}{P(w' \to w)} = \frac{P(w')}{P(w)} \tag{3.28}$$

La transition $P(w \to w')$ est ensuite décomposée en deux sous-étapes, la proposition (*proposal*) et l'acceptation (*acceptance*) :

$$P(w \to w') = \underbrace{g(w \to w')}_{\text{proposition}} \cdot \underbrace{A(w \to w')}_{\text{acceptation}} \tag{3.29}$$

En insérant dans l'éq. 3.28 on obtient

$$\frac{A(w \to w')}{A(w' \to w)} = \frac{P(w')}{g(w \to w')} \frac{g(w' \to w)}{P(w)} \tag{3.30}$$

Plusieurs choix de la fonction d'acceptation sont possibles pour satisfaire cette équation (Hastings, 1970). Un choix courant, dit choix de Metropolis, est :

$$A(w \to w') = \min\left(1, \frac{P(w')}{g(w \to w')} \frac{g(w' \to w)}{P(w)}\right) \tag{3.31}$$

Un aspect très pratique de cette quantité est qu'elle est invariante sous multiplication de la distribution $P(w)$ par un facteur non nul. Autrement dit, la distribution n'a pas besoin d'être normalisée. Dans un cadre bayésien, cela veut dire que l'on peut remplacer la postérieure par le produit de la vraisemblance et du *prior*.

La méthode de Metropolis-Hastings se résume donc ainsi :

1. Initialiser w à une valeur prise au hasard.
2. Choisir un nouvel état w' tiré selon $g(w \to w')$
3. Accepter l'état avec une probabilité donnée par $A(w \to w')$. Si le nouvel état n'est pas accepté, alors $w' = w$.
4. Itérer jusqu'à convergence

Au final, w étant tiré selon la distribution $P(w)$, sa moyenne est estimée en sommant les poids $w(t)$ obtenus au cours des N itérations réalisées :

$$\langle w \rangle \simeq \hat{w}_N = \frac{1}{N} \sum_{t=1}^{N} w(t) \tag{3.32}$$

Quant au critère de convergence, une possibilité est d'utiliser le Théorème Centrale Limite (TCL). Celui-ci stipule que la moyenne de n variables aléatoires indépendantes et identiquement distribuées selon une loi de moyenne μ et d'écart-type σ de valeurs finies suit, pour n grand, une loi normale de moyenne μ et d'écart-type $\sigma/\sqrt{n}$. Dans l'approche MCMC, les échantillons successifs w_i ne sont pas indépendants à cause du fait qu'on les tire selon la loi $g(w \to w')$. Il faut donc calculer le temps de décorrélation T pour lequel $\langle w(t) w_i(t+T) \rangle \simeq 0$, puis utiliser les échantillons $w(t)$ obtenus toutes les T itérations comme variables indépendantes. L'application du TCL permet alors d'arrêter les itérations lorsqu'une certaine précision désirée est atteinte, par exemple lorsque $\sigma/\sqrt{n} < 0.05 \cdot \mu$, c'est-à-dire lorsque les variations

de l'estimateur sont de l'ordre de 5%. L'écart-type σ étant lui même estimé à partir des échantillons de l'algorithme, il faut aussi s'assurer que lui-même a convergé vers une valeur stable pour appliquer le TLC.

3.3.2 Application au calcul de la postérieure

Dans notre cas, nous souhaitons utiliser l'algorithme de Metropolis-Hastings pour calculer la valeur moyenne du vecteur de poids w_i en position i de la PWM selon la distribution postérieure $\mathcal{P}(w_i|\{\mathcal{A}\})$:

$$\langle w_i \rangle = \int w_i \mathcal{P}(w_i|\{\mathcal{A}\}) \mathrm{d}w \tag{3.33}$$

Il nous faut pour cela définir une loi de proposition $g(w_i \to w_i')$ pertinente. Dans notre cas, les poids w_i doivent rester dans le simplexe de dimension 3 défini par $w_A, w_C, w_G > 0$ et $w_A + w_C + w_G < 1$, le poids w_T étant entièrement déterminé par la normalisation des probabilités $w_T = 1 - w_A - w_C - w_G$. La distribution naturelle possédant cette propriété est la loi de Dirichlet $\mathrm{Dir}(\alpha)$, de paramètres $\alpha = \{\alpha_A, \alpha_C, \alpha_G, \alpha_T\}$ et de densité de probabilité

$$f(w) = \frac{1}{B(\alpha)} \prod_{b \in \{A,C,G,T\}} w_{i,b}^{\alpha_b - 1} \tag{3.34}$$

où $B(\alpha)$ est la fonction bêta multinomiale permettant la normalisation. Cette distribution est la même que celle obtenue dans le cas d'observations indépendantes (cf article). Afin d'accélérer l'échantillonnage MCMC, nous avons cherché à régler les paramètres α de manière à être au plus proche de la distribution $\mathcal{P}(w_i|\{\mathcal{A}\})$. Dans le cas de N sites indépendants, celle-ci suit une loi de Dirichlet de paramètres $\alpha_p + N_i$, où α_p est le vecteur de pseudo-counts et N_i le vecteur donnant les nombres d'observations des nucléotides à la position i au sein des différentes séquences. Dans le cas d'un arbre phylogénétique fini, le nombre effectif d'observation est moins grand que le nombre total de séquences du fait des corrélations entre sites. Nous avons donc défini les paramètres de notre *proposition* comme étant

$$\alpha_b = \alpha_p + N_{\mathrm{eff}} \cdot w_i \tag{3.35}$$

où $N_{\mathrm{eff}} = N_{\mathrm{sites}} \cdot N_{\mathrm{spe}}/2$ avec N_{sites} le nombre d'alignements observés et N_{spe} le nombre d'espèces dans l'alignement (12 dans les deux cas, Drosophile et mammifères). Grossièrement, cela revient à dire que le modèle d'évolution réduit d'un facteur 2 le nombre de séquences

indépendantes. Nous avons calculé que le taux d'acceptation (proportion de mouvements proposés qui sont acceptés) pour ce paramètre était de l'ordre de 50%, une valeur généralement considérée comme raisonnable (Krauth, 2006). Nous obtenons finalement l'expression pour la *proposition* :

$$g(w \to w') = \frac{1}{B(\alpha)} \prod_{b \in \{A,C,G,T\}} (w'_{i,b})^{\alpha_{p,b} + N_{\text{eff}} \cdot w_{i,b} - 1} \tag{3.36}$$

Le vecteur w_i est initialisé à la valeur qu'il prendrait si toutes les séquences orthologues étaient des observations indépendantes (cf article) :

$$w_{i,b}(0) = w_{i,b}^{\text{inde}} = \frac{N_{i,b} + \alpha_b}{N_{\text{tot}} + \sum_b \alpha_b} \tag{3.37}$$

Le poids $w_i(1)$ suivant est ensuite tiré selon la probabilité de transition $g(w_i(0) \to w_i(1))$. Les différentes quantités de l'équation 3.31 sont ensuite calculées et la transition est acceptée avec probabilité $A\,(w_i(0) \to w_i(1))$.

3.3.3 Illustration sur un exemple

Étudions maintenant un exemple concret. Nous présentons la méthode sur le cas présenté en fig. 3.2A et nous utilisons le modèle d'évolution *Felsenstein*. Le vecteur de poids w_i est initialisé au cas indépendant w_i^{inde}. La proposition $g(w_i^{\text{inde}} \to w)$ (éq. 3.36) est montrée en fig. 3.2B. On voit notamment comment la distribution de Dirichlet permet de rester dans le simplexe dans les cas w_A et W_T proches de 0. La valeur finale de l'estimation de la moyenne de la postérieure obtenue après convergence de la chaîne (voir ci-dessous) est aussi montrée : elle est relativement proche de la moyenne de la distribution, indiquant que le choix de la valeur initiale est effectivement judicieux.

La chaîne MCMC est ensuite lancée. Le taux d'acceptation est calculé comme valant 62%. Les 500 premiers échantillons de w_i sont montrés en figure 3.3A. On note que certains points semblent corrélés : diminutions ou augmentations successives de la valeur courante $w_i(t)$ sur plusieurs itérations. Pour quantifier cet effet, nous avons mesuré la corrélation temporelle des échantillons. Celle-ci est donnée par

FIGURE 3.2 – **Conditions initiales.**

(A) Sites utilisés pour le MCMC. On s'intéresse ici à la 2ème position, contenant au total 10 G, 8 C et 1 A.

(B) Distribution de la distribution de Dirichlet correspondant à la proposition $g(w_i^{\text{inde}} \to w)$. L'échantillonnage consiste en 50,000 tirages aléatoires, et le bin est de 0.02. Les lignes verticales en pointillés indiquent l'estimation de la moyenne obtenue après convergence (voir fig. 3.4).

$$C_b(\tau) = \frac{1}{N}\sum_{t=1}^{N-\tau} w_{i,b}(t)w_{i,b}(t+\tau) - \left(\frac{1}{N}\sum_{t=1}^{N} w_{i,b}(t)\right)^2 \tag{3.38}$$

avec dans ce cas $N = 50,000$. Le logarithme de cette quantité est montrée en figure 3.3B. L'intérêt du logarithme est de mettre en exergue le caractère exponentiel de la décorrélation :

$$C_b(\tau) \propto e^{-\tau/T_b} \tag{3.39}$$

Les temps de décorrelation $T_b \sim 10$ sont estimés en ajustant une droite sur les 20 premiers points de la courbe. Maintenant que l'on connait le temps de corrélation entre deux échantillons, il est possible d'obtenir des échantillons indépendants en les choisissant à des

FIGURE 3.3 – **Corrélations entre les échantillons.**

(A) Extrait de l'échantillonnage de w_i. Les 500 premiers échantillons w_i de la chaîne MCMC sont montrés.

(B) Fonction d'auto-corrélation de la chaîne MCMC. La corrélation $C_b(t)$ est réalisée entre des échantillons séparés de $t-1$ itérations. Une droite est ajustée sur les 20 premièrs points, permettant d'obtenir la pente $-T_b$ donnant le temps de décorrélation.

intervalles grands devant T_b. Dans notre cas, nous avons choisis un intervalle de 30 itérations.

Nous souhaitons maintenant étudier la convergence de la chaîne MCMC. Pour cela, nous utilisons le Théorème Central Limite (TCL, cf 3.3.1). Celui-ci stipule que pour un nombre n suffisamment grand d'échantillons indépendants, l'écart-type de la distribution de l'estimateur empirique de la moyenne $\hat{w}_n$ se comporte comme $\sigma_b/\sqrt{n}$, où σ_b est l'écart-type de la distribution $\mathcal{P}(w_i|\mathcal{A})$. Ce dernier doit lui-même être estimé à partir de la chaîne MCMC. Nous présentons en figure 3.4A la valeur de l'estimation $\sigma_b(n)$ obtenue pour n itérations indépendantes. On voit que cette valeur atteint rapidement en ~ 100 itérations une valeur stable, et que dans tous les cas les fluctuations sont faibles, de l'ordre de 10% de $\hat{w}_n$. Il paraît donc raisonnable d'utiliser cette valeur de σ_b pour le TCL. Nous traçons en figure 3.4B la quantité $\sigma_b(n)/\sqrt{n}$. La chaîne est considérée comme convergée lorsque cette valeur est inférieure ou égale à 5% de la valeur moyenne estimée $\hat{w}_n$ (ligne grise) dans les 4 cas. Ce seuil est bien entendu arbitraire et dépend de la précision voulue par l'utilisateur sur l'estimation. Néanmoins, une plus grande précision implique un plus long temps de calcul.

Nous comparons finalement en figure 3.5 la valeur de l'estimation du Maximum A Posteriori (MAP) de la postérieure modifiée (voir article) obtenue avec la méthode de descente de gradient et celle de la moyenne de la postérieure obtenue avec l'approche MCMC, en fonction du nombre d'itérations (toutes les itérations, pas seulement les échantillons indépendants). L'approche MCMC, bien que convergeant vers un état proche de celui donné par la descente de gradient, met beaucoup plus de temps à converger : plus de $10,000$ itérations, alors que la descente de gradient n'en requiert que 100. Au vu de la faible différence entre les deux résultats, nous utilisons dans Imogene l'approche de maximisation de la postérieure modifiée, ce qui permet un gain de temps considérable pour l'algorithme (au minimum un facteur 10).

FIGURE 3.4 – **Estimation de la convergence.**
(A) Écart-type σ_b de la distribution $w_{i,b}$ estimé à partir de n échantillons indépendants. Ces échantillons sont pris toutes les 30 itérations au sein de la chaîne MCMC, soit environ 3 fois le temps de décorrélation. On observe qu'au bout de 100 itérations l'écart-type est stabilisé. Par ailleurs les fluctuations sont relativement faibles, au plus de l'ordre de $\sim$ 10% de la moyenne.
(B) La convergence est estimée grâce au Théorème Centrale Limite. L'écart-type de l'estimateur de la moyenne se comporte comme $\sigma_b/\sqrt{n}$. La précision demandée correspond à un écart-type $\leq$ 5% de la moyenne pour les 4 bases, ce qui correspond à l'intersection la plus tardive entre les courbes noires et grises (dans notre cas le cadran du bas à droite).

FIGURE 3.5 – **Comparaison des approches par MCMC et par descente de gradient.** L'approche de maximisation de (l'opposé de) la postérieure modifiée par descente de gradient (rouge, cf article) est comparée à l'approche de calcul de la moyenne de la postérieure par MCMC (noir). Les deux méthodes sont initialisées au même w_i^{inde}. Alors que la descente de gradient converge rapidement ($\sim$ 100 itérations), l'approche MCMC converge plus lentement, dans ce cas plus de 10,000 itérations. Au final les deux approches convergent vers des quantités proches (lignes pointillées rouges et noires).

Annexe A

Statistiques génomique

En guise d'exemple, nous montrons en figure A.1 des statistiques obtenues aisément à partir d'annotations génétiques présentes sur UCSC et traitées avec Galaxy. Ces statistiques sont les distribution de tailles des régions intergéniques et introniques chez plusieurs espèces : la bactérie *Escherichia coli*, la levure *Saccharomyces cerevisiae*, le ver *Caenorhabditis elegans*, la mouche *Drosophila melanogaster*, la souris, le poulet et l'homme.

FIGURE A.1 – **Distribution des tailles intergéniques et introniques chez différentes espèces.** Distributions log-log de la taille des régions intergéniques (A) et introniques (B) chez différentes espèces. Les histogrammes sont réalisés avec un intervalle de 0.05 puis lissés avec l'estimateur local LOESS de paramètre $span = 0.3$ (logiciel R). (A) Les régions intergéniques sont définies comme les régions complémentaires aux régions transcrites (données UCSC), celles-ci étant préalablement fusionnées pour éviter les redondances liées aux multiples transcrits d'un même gène. De la bactérie à l'homme, on observe une inflation de la quantité de génome non codant. (B) Les régions introniques sont définies par le fait qu'elles sont entourées par deux exons d'un même gène. Pour pouvoir être épissés lors de la maturation des preARNm, les introns doivent posséder des sites d'épissage, imposant une borne inférieure à leur taille pour que l'ARNm final soit fonctionnel.

Bibliographie

Aerts, S. (2012). Chapter 5 - Computational Strategies for the Genome-Wide Identification of cis-Regulatory Elements and Transcriptional Targets. *Current Topics in Developmental Biology : Transcriptional Switches During Development*, 98 :121–145. (Pages 41 et 116.)

Alon, U. (2007a). An Introduction to Systems Biology : Design Principles of Biological Circuits (Mathematical and Computational Biology Series vol 10). (Page 17.)

Alon, U. (2007b). Network motifs : theory and experimental approaches. *Nat Rev Genet*, 8(6) :450–461. (Page 17.)

Asp, P., Blum, R., Vethantham, V., Parisi, F., Micsinai, M., Cheng, J., Bowman, C., Kluger, Y., and Dynlacht, B. D. (2011). PNAS Plus : Genome-wide remodeling of the epigenetic landscape during myogenic differentiation. *Proceedings of the National Academy of Sciences*, pages 1–11. (Page 62.)

Attanasio, C., Reymond, A., Humbert, R., Lyle, R., Kuehn, M. S., Neph, S., Sabo, P. J., Goldy, J., Weaver, M., Haydock, A., Lee, K., Dorschner, M., Dermitzakis, E. T., Antonarakis, S. E., and Stamatoyannopoulos, J. A. (2008). Assaying the regulatory potential of mammalian conserved non-coding sequences in human cells. *Genome Biol*, 9(12) :R168. (Page 53.)

Aurell, E., d'Hérouël, A., Malmnäs, C., and Vergassola, M. (2007). Transcription factor concentrations versus binding site affinities in the yeast S. cerevisiae. *Physical biology*, 4 :134. (Page 27.)

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C.-F., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R., and Bulyk, M. L. (2009). Diversity and Complexity in DNA Recognition by Transcription Factors. *Science*, 324(5935) :1720–1723. (Pages 67 et 71.)

Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2 :28–36. (Page 116.)

Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. (2003). Modeling dependencies in protein-DNA binding sites. In *Proceedings of the seventh annual international conference on Research in*

computational molecular biology, RECOMB '03, pages 28–37. ACM, New York, NY, USA. ISBN 1-58113-635-8. (Pages 70, 71 et 72.)

Bartel, D. P. (2009). MicroRNAs : target recognition and regulatory functions. *Cell*, 136(2) :215–33. (Page 14.)

Baxter, R. (2007). *Exactly Solved Models in Statistical Mechanics*. Dover Books on Physics Series. DOVER PUBN Incorporated. ISBN 9780486462714. (Page 77.)

Baylies, M. K., Bate, M., and Ruiz Gomez, M. (1998). Myogenesis : a view from Drosophila. *Cell*, 93(6) :921–7. (Page 19.)

Bell, A. C., West, A. G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, 98(3) :387–96. (Page 41.)

Benos, P. V., Bulyk, M. L., and Stormo, G. D. (2002). Additivity in protein-DNA interactions : how good an approximation is it ? *Nucleic Acids Res*, 30(20) :4442–51. (Page 69.)

Berg, O. and von Hippel, P. (1987). Selection of DNA binding sites by regulatory proteins : Statistical-mechanical theory and application to operators and promoters. *Journal of molecular biology*, 193(4) :723–743. (Page 21.)

Berg, O. G., Winter, R. B., and von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, 20(24) :6929–48. (Page 21.)

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., 3rd, and Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24(11) :1429–35. (Page 29.)

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M., and Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci U S A*, 99(2) :757–62. (Page 50.)

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev*, 16(1) :6–21. (Page 13.)

Blackwell, T. K. and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science*, 250(4984) :1104–10. (Page 30.)

Blanchette, M., Bataille, A. R., Chen, X., Poitras, C., Laganière, J., Lefèbvre, C., Deblois, G., Giguère, V., Ferretti, V., Bergeron, D., Coulombe, B., and Robert, F. (2006). Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res*, 16(5) :656–68. (Page 53.)

Blau, H. M., Pavlath, G. K., Hardeman, E. C., Chiu, C. P., Silberstein, L., Webster, S. G., Miller, S. C., and Webster, C. (1985). Plasticity of the differentiated state. *Science*, 230(4727) :758–66. (Page 8.)

Bolouri, H. and Davidson, E. H. (2002). Modeling DNA sequence-based cis-regulatory gene networks. *Dev Biol*, 246(1) :2–13. (Page 40.)

Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H. H., and Liu, E. T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res*, 18(11) :1752–1762. (Page 45.)

Brazma, A., Parkinson, H., Schlitt, T., and Shojatalab, M. (2001). A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays. http://www.ebi.ac.uk/microarray/biology_intro.html. (Page 5.)

Buchler, N. E., Gerland, U., and Hwa, T. (2003). On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A*, 100(9) :5136–41. (Page 40.)

Bulyk, M. L., Johnson, P. L. F., and Church, G. M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30(5) :1255–61. (Page 67.)

Bussemaker, H. J., Li, H., and Siggia, E. D. (2000). Building a dictionary for genomes : identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci U S A*, 97(18) :10096–100. (Page 120.)

Campbell, C. T. and Kim, G. (2007). SPR microscopy and its applications to high-throughput analyses of biomolecular binding events and their kinetics. *Biomaterials*, 28(15) :2380–92. (Page 29.)

Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G. J., Parker, M. H., Macquarrie, K. L., Davison, J., Morgan, M. T., Ruzzo, W. L., Gentleman, R. C., and Tapscott, S. J. (2010a). Genome-wide MyoD binding in skeletal muscle cells : a potential for broad cellular reprogramming. *Developmental Cell*, 18(4) :662–74. (Pages 52 et 125.)

Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G. J., Parker, M. H., Macquarrie, K. L., Davison, J., Morgan, M. T., Ruzzo, W. L., Gentleman, R. C., and Tapscott, S. J. (2010b). Genome-wide MyoD binding in skeletal muscle cells : a potential for broad cellular reprogramming. *Developmental Cell*, 18(4) :662–74. (Page 62.)

Carlson, C. D., Warren, C. L., Hauschild, K. E., Ozers, M. S., Qadir, N., Bhimsaria, D., Lee, Y., Cerrina, F., and Ansari, A. Z. (2010). Specificity landscapes of DNA binding molecules elucidate biological function. *Proc Natl Acad Sci U S A*, 107(10) :4544–9. (Page 30.)

Carninci, P., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6) :626–35. (Page 39.)

Carvajal, J. J., Keith, A., and Rigby, P. W. J. (2008). Global transcriptional regulation of the locus encoding the skeletal muscle determination genes Mrf4 and Myf5. *Genes & development*, 22(2) :265–76. (Pages 39 et 40.)

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3) :167–174. (Page 69.)

Chan, B. Y. and Kibler, D. (2005). Using hexamers to predict cis-regulatory motifs in Drosophila. *BMC Bioinformatics*, 6 :262. (Page 127.)

Cheng, Y., King, D. C., Dore, L. C., Zhang, X., Zhou, Y., Zhang, Y., Dorman, C., Abebe, D., Kumar, S. A., Chiaromonte, F., Miller, W., Green, R. D., Weiss, M. J., and Hardison, R. C. (2008). Transcriptional enhancement by GATA1-occupied DNA segments is strongly associated with evolutionary constraint on the binding site motif. *Genome Res*, 18(12) :1896–905. (Page 55.)

Chung, J. H., Whiteley, M., and Felsenfeld, G. (1993). A 5′ element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell*, 74(3) :505–14. (Page 41.)

Cordaux, R. and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10) :691–703. (Page 47.)

Davis, R. L., Weintraub, H., and Lassar, A. B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell*, 51(6) :987–1000. (Page 9.)

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38. (Page 118.)

Dermitzakis, E. T. and Clark, A. G. (2002). Evolution of transcription factor binding sites in Mammalian gene regulatory regions : conservation and turnover. *Mol Biol Evol*, 19(7) :1114–21. (Page 45.)

Dermitzakis, E. T., Reymond, A., and Antonarakis, S. E. (2005). Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat Rev Genet*, 6(2) :151–7. (Page 53.)

D'haeseleer, P. (2006). How does DNA sequence motif discovery work ? *Nat Biotechnol*, 24(8) :959–61. (Page 117.)

Djordjevic, M., Sengupta, A. M., and Shraiman, B. I. (2003). A biophysical approach to transcription factor binding site discovery. *Genome Res*, 13(11) :2381–90. (Page 24.)

Donaldson, I. J., Chapman, M., Kinston, S., Landry, J. R., Knezevic, K., Piltz, S., Buckley, N., Green, A. R., and Göttgens, B. (2005). Genome-wide identification of cis-regulatory sequences controlling blood and endothelial development. *Hum Mol Genet*, 14(5) :595–601. (Page 54.)

ENCODE Project Consortium, et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414) :57–74. (Page 64.)

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345) :43–9. (Page 55.)

Felsenstein, J. (1981). Evolutionary trees from DNA sequences : a maximum likelihood approach. *J Mol Evol*, 17(6) :368–76. (Pages 121, 122 et 123.)

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat Rev Genet*, 9(5) :397–405. (Page 47.)

Fields, D. S., He, Y., Al-Uzri, A. Y., and Stormo, G. D. (1997). Quantitative specificity of the Mnt repressor. *J Mol Biol*, 271(2) :178–94. (Page 30.)

Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F., and Stern, D. L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, pages 1–5. (Page 48.)

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, 29(2-3) :131–163. (Page 70.)

Furusawa, C. and Kaneko, K. (2012). A Dynamical-Systems View of Stem Cell Biology. *Science*, 338(6104) :215–217. (Page 6.)

Gerland, U., Moroz, J., and Hwa, T. (2002). Physical constraints and functional characteristics of transcription factor–DNA interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 99(19) :12015. (Pages 21, 24, 25 et 26.)

Giocomo, L. M., Moser, M.-B., and Moser, E. I. (2011). Computational models of grid cells. *Neuron*, 71(4) :589–603. (Page 23.)

Grad, Y. H., Roth, F. P., Halfon, M. S., and Church, G. M. (2004). Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D.pseudoobscura. *Bioinformatics*, 20(16) :2738–50. (Page 125.)

Graf, T. and Enver, T. (2009). Forcing cells to change lineages. *Nature*, 462(7273) :587–94. (Page 9.)

Greer, E. L. and Shi, Y. (2012). Histone methylation : a dynamic mark in health, disease and inheritance. *Nat Rev Genet*, 13(5) :343–57. (Page 14.)

Grendar Jr, M. and Grendar, M. (2001). MiniMax Entropy and Maximum Likelihood : complementarity of tasks, identity of solutions. In *AIP Conference Proceedings*, volume 568, page 49. (Page 75.)

Gurdon, J. B. and Melton, D. A. (2008). Nuclear reprogramming in cells. *Science*, 322(5909) :1811–5. (Page 9.)

Halpern, A. and Bruno, W. (1998). Evolutionary distances for protein-coding sequences : modeling site-specific residue frequencies. *Molecular biology and evolution*, 15(7) :910. (Page 123.)

Hammond, S. M., Caudy, A. A., and Hannon, G. J. (2001). Post-transcriptional gene silencing by double-stranded RNA. *Nat Rev Genet*, 2(2) :110–9. (Page 14.)

Hannon, G. J. (2002). RNA interference. *Nature*, 418(6894) :244–51. (Page 14.)

Hardison, R. C. and Taylor, J. (2012). Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews Genetics*, 13(7) :469–483. (Pages 38 et 55.)

Hartwell, L., Hopfield, J., Leibler, S., and Murray, A. (1999). From molecular to modular cell biology. *Nature*, 402(6761) :47. (Pages 51 et 56.)

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2) :160–74. (Page 123.)

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1) :97–109. (Pages 156 et 157.)

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., and Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, 39(3) :311–8. (Page 55.)

Heintzman, N. D., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243) :108–12. (Pages 40 et 55.)

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, 38(4) :576–89. (Page 125.)

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., Fields, S., and Stamatoyannopoulos, J. A. (2009).

Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods*, 6(4) :283–9. (Page 36.)

Hong, J.-W., Hendrix, D. A., and Levine, M. S. (2008). Shadow enhancers as a source of evolutionary novelty. *Science*, 321(5894) :1314. (Page 47.)

Hu, M., Yu, J., Taylor, J. M., Chinnaiyan, A. M., and Qin, Z. S. (2010). On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic acids research*, 38(7) :2154–2167. (Page 69.)

Ivan, A., Halfon, M. S., and Sinha, S. (2008). Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs. *Genome Biol*, 9(1) :R22. (Page 126.)

Jaynes, E. T. (1978). Where do we stand on maximum entropy ? pages 1–105. (Page 72.)

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9) :939–952. (Page 75.)

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., Bonke, M., Palin, K., Talukder, S., Hughes, T. R., Luscombe, N. M., Ukkonen, E., and Taipale, J. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*, 20(6) :861–73. (Page 31.)

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., and Taipale, J. (2013). DNA-Binding Specificities of Human Transcription Factors. *Cell*, 152(1-2) :327–39. (Pages 31, 67, 71 et 112.)

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., Taatjes, D. J., Dekker, J., and Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314) :430–435. (Page 48.)

Kantorovitz, M. R., Kazemian, M., Kinston, S., Miranda-Saavedra, D., Zhu, Q., Robinson, G. E., Göttgens, B., Halfon, M. S., and Sinha, S. (2009). Motif-blind, genome-wide discovery of cis-regulatory modules in Drosophila and mouse. *Developmental Cell*, 17(4) :568–79. (Pages 52, 54, 125 et 127.)

Kantorovitz, M. R., Robinson, G. E., and Sinha, S. (2007). A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, 23(13) :i249–55. (Page 126.)

Kaufmann, S. (1993). The origins of order. (Page 7.)

Kazemian, M., Zhu, Q., Halfon, M. S., and Sinha, S. (2011). Improved accuracy of supervised CRM discovery with interpolated Markov models and cross-species comparison. *Nucleic Acids Res*, 39(22) :9463–72. (Page 126.)

Keim, C. N., Martins, J. L., Abreu, F., Rosado, A. S., de Barros, H. L., Borojevic, R., Lins, U., and Farina, M. (2004). Multicellular life cycle of magnetotactic prokaryotes. *FEMS Microbiol Lett*, 240(2) :203–8. (Page 5.)

Kheradpour, P., Stark, A., Roy, S., and Kellis, M. (2007). Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res*, 17(12) :1919–31. (Page 53.)

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics*, 47(6) :713. (Page 124.)

Kinney, J. B., Tkacik, G., and Callan, C. G. (2007). Precise physical models of protein-DNA interaction from high-throughput data. *Proc Natl Acad Sci USA*, 104(2) :501–6. (Page 29.)

Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F., and Kageyama, Y. (2010). Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. *Science*, 329(5989) :336–9. (Page 14.)

Krauth, W. (2006). *Statistical mechanics : algorithms and computations*, volume 13. Oxford University Press. (Pages 69, 156 et 159.)

Kulessa, H., Frampton, J., and Graf, T. (1995). GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes Dev*, 9(10) :1250–62. (Page 9.)

Kulkarni, M. M. and Arnosti, D. N. (2003). Information display by transcriptional enhancers. *Development*, 130(26) :6569–75. (Pages 42 et 43.)

Lander, E. S., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860–921. (Page 59.)

Lässig, M. (2007). From biophysics to evolutionary genetics : statistical aspects of gene regulation. *BMC Bioinformatics*, 8(Suppl 6) :S7. (Pages 20 et 25.)

Lawrence, C. E. and Reilly, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7(1) :41–51. (Pages 116 et 117.)

Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thompson, C., and Simon, I. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594) :799. (Pages 16 et 17.)

Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet*, 12(14) :1725–35. (Pages 39, 57 et 58.)

Liberman, L. M. and Stathopoulos, A. (2009). Design flexibility in cis-regulatory control of gene expression : Synthetic and comparative evidence. *Developmental Biology*, 327(2) :578–589. (Pages 44 et 45.)

Liu, Y., Chu, A., Chakroun, I., Islam, U., and Blais, A. (2010). Cooperation between myogenic regulatory factors and SIX family transcription factors is important for myoblast differentiation (SI). *Nucleic acids research*. (Page 62.)

Liu, Y.-H., Jakobsen, J. S., Valentin, G., Amarantos, I., Gilmour, D. T., and Furlong, E. E. M. (2009). A systematic analysis of Tinman function reveals Eya and JAK-STAT signaling as essential regulators of muscle development. *Developmental Cell*, 16(2) :280–91. (Page 18.)

Loots, G. G. and Ovcharenko, I. (2004). rVISTA 2.0 : evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*, 32(Web Server issue) :W217–21. (Page 63.)

Loots, G. G. and Ovcharenko, I. (2005). Dcode.org anthology of comparative genomic tools. *Nucleic Acids Res*, 33(Web Server issue) :W56–64. (Page 61.)

Ludwig, M. Z., Bergman, C., Patel, N. H., and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, 403(6769) :564–7. (Page 45.)

Maerkl, S. and Quake, S. (2007). A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809) :233. (Page 28.)

Majoros, W. H. and Ohler, U. (2010). Modeling the evolution of regulatory elements by simultaneous detection and alignment with phylogenetic pair HMMs. *PLoS Comput Biol*, 6(12) :e1001037. (Page 54.)

Man, T. K. and Stormo, G. D. (2001). Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res*, 29(12) :2471–8. (Page 66.)

Maniatis, T., Goodbourn, S., and Fischer, J. A. (1987). Regulation of inducible and tissue-specific gene expression. *Science*, 236(4806) :1237–45. (Page 39.)

Masuya, H., Sezutsu, H., Sakuraba, Y., Sagai, T., Hosoya, M., Kaneda, H., Miura, I., Kobayashi, K., Sumiyama, K., Shimizu, A., Nagano, J., Yokoyama, H., Kaneko, S., Sakurai, N., Okagaki, Y., Noda, T., Wakana, S., Gondo, Y., and Shiroishi, T. (2007). A series of ENU-induced single-base substitutions in a long-range cis-element altering Sonic hedgehog expression in the developing mouse limb bud. *Genomics*, 89(2) :207–14. (Page 58.)

McGregor, A., Orgogozo, V., Delon, I., Zanet, J., Srinivasan, D., Payre, F., and Stern, D. (2007). Morphological evolution through multiple cis-regulatory mutations at a single gene. *Nature*, 448(7153) :587–590. (Page 47.)

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21 :1087. (Page 156.)

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs : simple building blocks of complex networks. *Science*, 298(5594) :824–7. (Page 17.)

Moses, A., Chiang, D., Pollard, D., Iyer, V., and Eisen, M. (2004). MONKEY : identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol*, 5(12) :R98. (Page 122.)

Moses, A. M., Chiang, D. Y., Kellis, M., Lander, E. S., and Eisen, M. B. (2003). Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol*, 3 :19. (Page 123.)

Moses, A. M., Pollard, D. A., Nix, D. A., Iyer, V. N., Li, X.-Y., Biggin, M. D., and Eisen, M. B. (2006). Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol*, 2(10) :e130. (Page 45.)

Nagaraj, V. H., O'flanagan, R. A., and Sengupta, A. M. (2008). Better estimation of protein-DNA interaction parameters improve prediction of functional sites. *BMC Biotechnol*, 8(1) :94. (Pages 30 et 31.)

Nurse, P. and Hayles, J. (2011). The Cell in an Era of Systems Biology. *Cell*. (Page 10.)

Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., MacIsaac, K. D., Rolfe, P. A., Conboy, C. M., Gifford, D. K., and Fraenkel, E. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet*, 39(6) :730–2. (Page 45.)

Odom, D. T., Zizlsperger, N., Gordon, D. B., Bell, G. W., Rinaldi, N. J., Murray, H. L., Volkert, T. L., Schreiber, J., Rolfe, P. A., Gifford, D. K., Fraenkel, E., Bell, G. I., and Young, R. A. (2004). Control of pancreas and liver gene expression by HNF transcription factors. *Science*, 303(5662) :1378–81. (Pages 16 et 17.)

Oliphant, A. R., Brandl, C. J., and Struhl, K. (1989). Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides : analysis of yeast GCN4 protein. *Mol Cell Biol*, 9(7) :2944–9. (Page 30.)

Ondek, B., Gloss, L., and Herr, W. (1988). The SV40 enhancer contains two distinct levels of organization. *Nature*, 333(6168) :40–5. (Page 39.)

Panne, D. (2008). The enhanceosome. *Curr Opin Struct Biol*, 18(2) :236–42. (Page 43.)

Park, P. J. (2009). ChIP-seq : advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10) :669–80. (Page 35.)

Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B. L., Couronne, O., Eisen, M. B., Visel, A., and Rubin, E. M. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118) :499–502. (Pages 43 et 53.)

Perry, M. W., Boettiger, A. N., and Levine, M. (2011). Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *Proc Natl Acad Sci U S A*, 108(33) :13570–5. (Page 48.)

Phillips, J. E. and Corces, V. G. (2009). CTCF : master weaver of the genome. *Cell*, 137(7) :1194–211. (Page 41.)

Prabhakar, S., Poulin, F., Shoukry, M., Afzal, V., Rubin, E. M., Couronne, O., and Pennacchio, L. A. (2006). Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res*, 16(7) :855–63. (Page 43.)

Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E. D. (2002). Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics*, 3 :30. (Pages 50 et 120.)

Recillas-Targa, F., Pikaart, M. J., Burgess-Beusse, B., Bell, A. C., Litt, M. D., West, A. G., Gaszner, M., and Felsenfeld, G. (2002). Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *Proc Natl Acad Sci U S A*, 99(10) :6883–8. (Page 41.)

Roh, T.-Y., Cuddapah, S., and Zhao, K. (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev*, 19(5) :542–52. (Page 55.)

Rouault, H., Mazouni, K., Couturier, L., Hakim, V., and Schweisguth, F. (2010). Genome-wide identification of cis-regulatory motifs and modules underlying gene coregulation using statistics and phylogeny. *Proceedings of the National Academy of Sciences*, 107(33) :14615. (Pages 116 et 127.)

Roulet, E., Busso, S., Camargo, A. A., Simpson, A. J. G., Mermod, N., and Bucher, P. (2002). High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol*, 20(8) :831–5. (Page 30.)

Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development*, 132(4) :797–803. (Pages 57 et 58.)

Sato, S., Ikeda, K., Shioi, G., Nakao, K., Yajima, H., and Kawakami, K. (2012). Regulation of Six1 expression by evolutionarily conserved enhancers in tetrapods. *Dev Biol*, 368(1) :95–108. (Page 63.)

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., and Gerstein, M. B. (2011). The real cost of sequencing : higher than you think ! *Genome Biol*, 12(8) :125. (Page 59.)

Schirm, S., Jiricny, J., and Schaffner, W. (1987). The SV40 enhancer can be dissected into multiple segments, each with a different cell type specificity. *Genes Dev*, 1(1) :65–74. (Page 39.)

Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, Â., Kutter, C., Brown, G. D., Marshall, A., Flicek, P., and Odom, D. T. (2012). Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. *Cell*, 148(1-2) :335–348. (Page 47.)

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., Talianidis, I., Flicek, P., and Odom, D. T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding SI. *Science*, 328(5981) :1036–40. (Page 45.)

Schoborg, T. A. and Labrador, M. (2010). The phylogenetic distribution of non-CTCF insulator proteins is limited to insects and reveals that BEAF-32 is Drosophila lineage specific. *J Mol Evol*, 70(1) :74–84. (Page 41.)

Schones, D. E. and Zhao, K. (2008). Genome-wide approaches to studying chromatin modifications. *Nat Rev Genet*, 9(3) :179–91. (Page 15.)

Schroeder, M. D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E. D., and Gaul, U. (2004). Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol*, 2(9) :E271. (Page 50.)

Sethna, J. (2006). *Statistical Mechanics : Entropy, Order Parameters and Complexity*. Oxford Master Series in Physics. OUP Oxford. ISBN 9780198566779. (Page 74.)

Shen-Orr, S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics*, 31(1) :64–68. (Page 17.)

Shumaker-Parry, J. S., Aebersold, R., and Campbell, C. T. (2004). Parallel, quantitative measurement of protein binding to a 120-element double-stranded DNA array in real time using surface plasmon resonance microscopy. *Anal Chem*, 76(7) :2071–82. (Page 29.)

Sinha, S. and He, X. (2007). MORPH : probabilistic alignment combined with hidden Markov models of cis-regulatory modules. *PLoS Comput Biol*, 3(11) :e216. (Page 54.)

Sinha, S., Nimwegen, E. V., and Siggia, E. (2003). A probabilistic method to detect regulatory modules. *Bioinformatics*, 19(Suppl 1) :i292. (Pages 119 et 120.)

Sinha, S., Schroeder, M. D., Unnerstall, U., Gaul, U., and Siggia, E. D. (2004). Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC Bioinformatics*, 5 :129. (Page 53.)

Sinha, S. and Tompa, M. (2000). A statistical method for finding transcription factor binding sites. *Proc Int Conf Intell Syst Mol Biol*, 8 :344–54. (Page 127.)

Slutsky, M. and Mirny, L. A. (2004). Kinetics of protein-DNA interaction : facilitated target location in sequence-dependent potential. *Biophys J*, 87(6) :4021–35. (Page 21.)

Smith, A. D., Sumazin, P., Xuan, Z., and Zhang, M. Q. (2006). DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci U S A*, 103(16) :6275–80. (Page 52.)

Smith, A. D., Sumazin, P., and Zhang, M. Q. (2005). Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci U S A*, 102(5) :1560–5. (Page 52.)

Stormo, G. and Fields, D. (1998). Specificity, free energy and information content in protein-DNA interactions. *Trends in biochemical sciences*, 23(3) :109–113. (Page 22.)

Stormo, G. D. and Zhao, Y. (2007). Putting numbers on the network connections. *Bioessays*, 29(8) :717–21. (Page 28.)

Stormo, G. D. and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. *Nature Reviews Genetics*, 11(11) :751–60. (Pages 27 et 32.)

Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4) :663–76. (Page 10.)

Taylor, J., Tyekucheva, S., King, D. C., Hardison, R. C., Miller, W., and Chiaromonte, F. (2006). ESPERR : learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res*, 16(12) :1596–604. (Page 54.)

Thurman, R. E., et al. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414) :75–82. (Page 37.)

Tijssen, M. R., Cvejic, A., Joshi, A., Hannah, R. L., Ferreira, R., Forrai, A., Bellissimo, D. C., Oram, S. H., Smethurst, P. A., Wilson, N. K., Wang, X., Ottersbach, K., Stemple, D. L., Green, A. R., Ouwehand, W. H., and Göttgens, B. (2011). Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev Cell*, 20(5) :597–609. (Page 55.)

Tirosh, I., Weinberger, A., Bezalel, D., Kaganovich, M., and Barkai, N. (2008). On the relation between promoter divergence and gene expression evolution. *Mol Syst Biol*, 4 :159. (Page 45.)

Trinklein, N. D., Aldred, S. J. F., Saldanha, A. J., and Myers, R. M. (2003). Identification and functional analysis of human transcriptional promoters. *Genome Res*, 13(2) :308–12. (Page 54.)

Tuerk, C. and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment : RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968) :505–10. (Page 30.)

U.S. Department of Energy (2001). Genomes to life : accelerating biological discovery (Office of Biological and Environmental Research and Office of Advanced Scientific Computing Research of the U.S. Department of Energy). http://genomicscience.energy.gov/roadmap/GTLcomplete_web.pdf. (Pages 11 et 12.)

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors : function, expression and evolution. *Nat Rev Genet*, 10(4) :252–63. (Page 13.)

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., and Pennacchio, L. A. (2009a). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231) :854–8. (Page 55.)

Visel, A., Rubin, E. M., and Pennacchio, L. A. (2009b). Genomic views of distant-acting enhancers. *Nature*, 461(7261) :199–205. (Pages 33 et 57.)

Waddington, C. H. et al. (1957). 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. (Pages 6 et 7.)

Wallace, J. A. and Felsenfeld, G. (2007). We gather together : insulators and genome organization. *Curr Opin Genet Dev*, 17(5) :400–7. (Page 41.)

Wang, Q., Carroll, J. S., and Brown, M. (2005). Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Mol Cell*, 19(5) :631–42. (Page 40.)

Warren, C. L., Kratochvil, N. C. S., Hauschild, K. E., Foister, S., Brezinski, M. L., Dervan, P. B., Phillips, G. N., Jr, and Ansari, A. Z. (2006). Defining the sequence-recognition profile of DNA-binding molecules. *Proc Natl Acad Sci U S A*, 103(4) :867–72. (Page 29.)

Wasserman, W. and Fickett, J. (1998). Identification of regulatory regions which confer muscle-specific gene expression1. *Journal of molecular biology*, 278(1) :167–181. (Page 50.)

Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4) :276–87. (Pages 23 et 59.)

Weintraub, H., Tapscott, S. J., Davis, R. L., Thayer, M. J., Adam, M. A., Lassar, A. B., and Miller, A. D. (1989). Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proc Natl Acad Sci U S A*, 86(14) :5434–8. (Page 18.)

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2) :307–19. (Pages 48 et 49.)

Wilczynski, B. and Furlong, E. E. M. (2010). Challenges for modeling global gene regulatory networks during development : Insights from Drosophila. *Developmental Biology*, 340(2) :161–169. (Page 40.)

Wilson, M. D., Barbosa-Morais, N. L., Schmidt, D., Conboy, C. M., Vanes, L., Tybulewicz, V. L. J., Fisher, E. M. C., Tavaré, S., and Odom, D. T. (2008). Species-specific transcription in mice carrying human chromosome 21. *Science*, 322(5900) :434–8. (Page 45.)

Wilson, M. D. and Odom, D. T. (2009). Evolution of transcriptional control in mammals. *Curr Opin Genet Dev*, 19(6) :579–85. (Pages 39, 45 et 46.)

Winter, R. B., Berg, O. G., and von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor–operator interaction : kinetic measurements and conclusions. *Biochemistry*, 20(24) :6961–77. (Page 21.)

Winter, R. B. and von Hippel, P. H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. The Escherichia coli repressor–operator interaction : equilibrium measurements. *Biochemistry*, 20(24) :6948–60. (Page 21.)

Wright, W. E., Binder, M., and Funk, W. (1991). Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. *Mol Cell Biol*, 11(8) :4104–10. (Page 30.)

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, 434(7031) :338–45. (Page 53.)

Zeitlinger, J., Zinzen, R. P., Stark, A., Kellis, M., Zhang, H., Young, R. A., and Levine, M. (2007). Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev*, 21(4) :385–90. (Page 47.)

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9) :R137. (Page 34.)

Zhao, Y., Granas, D., and Stormo, G. D. (2009). Inferring binding energies from selected binding sites. *PLoS Comput Biol*, 5(12) :e1000590. (Page 24.)

Zhou, Q. and Liu, J. S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6) :909–916. (Page 69.)

Zhou, Q. and Wong, W. H. (2004). CisModule : de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A*, 101(33) :12114–9. (Page 52.)

Zinzen, R., Girardot, C., Gagneur, J., Braun, M., and Furlong, E. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269) :65–70. (Page 45.)

Zykovich, A., Korf, I., and Segal, D. J. (2009). Bind-n-Seq : high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. *Nucleic Acids Res*, 37(22) :e151. (Page 31.)