UNIVERSITE PARIS VIII – XIII

Génopole Evry

APPROCHE "GENETIC PROGRAMMING" DE LA REGULATION DE L'EXPRESSION GENIQUE

ANTHONY PRIEUR

Dr Recherche

François Képès - ATelier de Génomique Cognitive, Génopole d'Evry
Renaud Dumeur - Université Paris VIII

Wolfgang Banzhaf - Memorial University de St-John, Canada

Abstract

The regulation of transcription is one of the basic mechanisms of life, it is essential for cell adaptation to environmental stress, the mechanisms of this regulation are still poorly understood.

This project aims to inform the optimization of the placement of genes along the chromosomes. We have developed a new model of artificial regulatory networks which allowed us to obtain a proof of concept of the validity of chromosomes solenoid model.

Résumé

La régulation de la transcription est un des mécanismes de base de la vie, elle est essentielle à l'adaptation de la cellule aux stress environnementaux, les mécanismes de cette régulation sont encore mal connus.

Ce projet a pour but d'éclairer l'optimisation du placement des gènes le long des chromosomes. Nous avons élaboré un nouveau modèle de réseaux de régulation artificiel qui nous permet d'obtenir une preuve de principe de la validité du modèle d'organisation solénoïde des chromosomes.

Table des matières

Avant Propos

Depuis quelques années la modélisation informatique de processus biologiques est très importante dans les avancées de la recherche fondamentale en biologie.

Ce projet a pour but d'étudier la manière dont s'organise les chromosomes (sous leur forme chromatinienne) dans le noyau de la cellule et d'obtenir une preuve par la simulation de la validité du modèle d'organisation solénoïde des chromosomes.

Nous avons plusieurs exemples [1], [2], où il a été démontré qu'il existe une optimisation de la transcription par organisation spatiale de la chromatine. Les chromosomes possèdent une structuration à grande échelle permettant des rapprochements physiques de différentes parties éloignées. F. Képès propose [3], [4], que ce rapprochement puisse servir à optimiser certains mécanismes biologiques telles les régulations transcriptionnelles, ce qui appuierait le modèle Solénoïde.

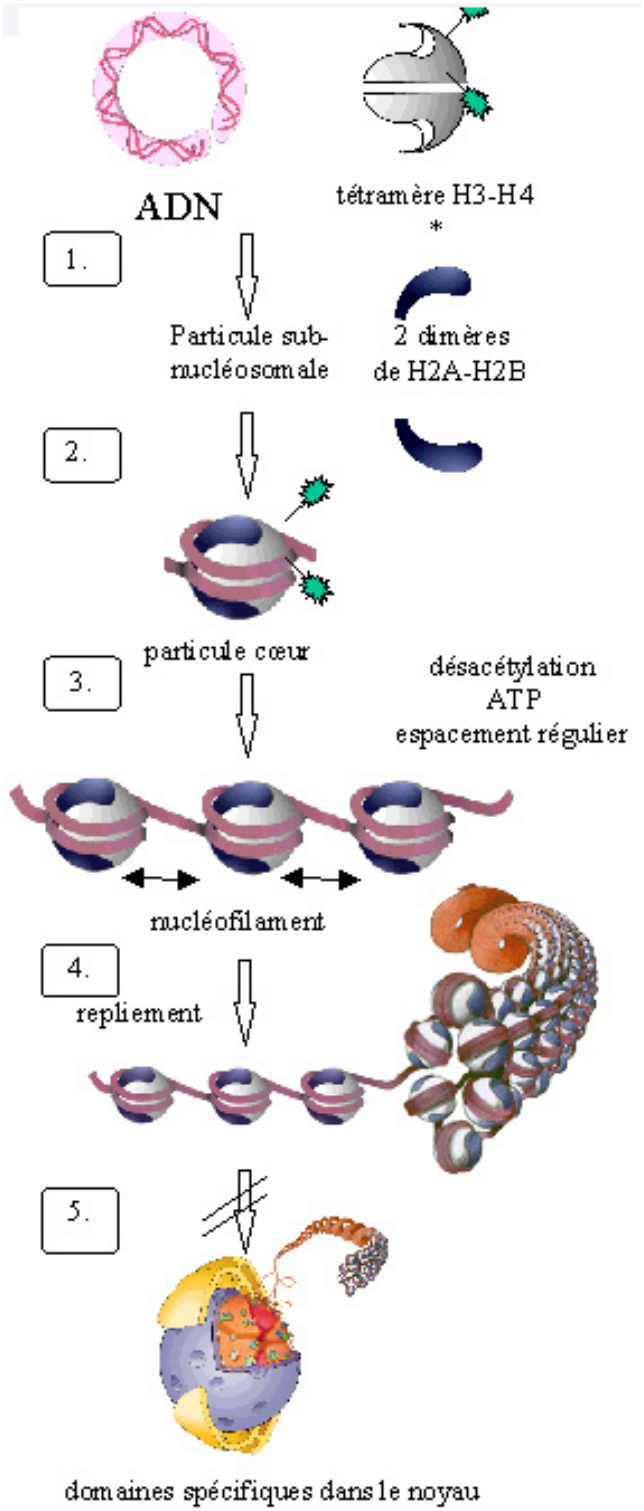
1 Introduction

1.1 Présentation biologique

Les chromosomes des eucaryotes, formés d'un complexe d'ADN, d'ARN et de protéines appelé chromatine, sont des entités dynamiques dont l'état varie énormément en fonction de la phase du cycle cellulaire. Environ deux mètres d'ADN (les 46 chromosomes d'une cellule humaine contiennent de 48 à 240 millions de paires de bases) dans chaque cellule doivent être contenu dans un noyau de quelques nm de diamètre. L'ADN chromosomique a un rapport de compactage supérieur à 8000. Comment l'ADN parvient-il à un tel degré de condensation dans la chromatine ? Des études structurales ont révélé qu'il existe différents niveaux de repliement.

En plus de cet énorme degré de compaction, l'ADN doit être rapidement accessible afin de permettre son interaction avec les machineries protéiques régulant les fonctions de la chromatine : la réplication, la réparation et la recombinaison. Ainsi l'organisation dynamique de la structure chromatinienne influence potentiellement toutes les fonctions du génome.

L'assemblage de l'ADN en chromatine comprend plusieurs étapes qui commencent par la formation de son unité fondamentale, le nucléosome, et finissent par des niveaux d'organisation supérieurs en domaines spécifiques dans le noyau. L'étape suivante est une étape de maturation au cours de laquelle les nucléosomes sont régulièrement espacés et forment le nucléofilament. Ensuite l'incorporation des histones internucléosomales est accompagnée par le repliement du nucléofilament en fibre de 30 nm dont la structure n'est pas élucidée à ce jour. Deux modèles principaux existent : le modèle de type solénoïde et le modèle de type zig-zag. Les différentes étapes de cet assemblage sont décrites dans le schéma ci-dessous.



Les principales étapes de l'assemblage de la chromatine, tiré de :
"Atlas of Genetics and Cytogenetics in Oncology and Haematology" © Infobiogen
(Ici est présenté le modèle de type Solénoïde)

1.2 Transcriptomique

La chromatine évolue dynamiquement, devenant plus ou moins compacte selon l'état de la cellule. En "respirant" ainsi, elle permet aux molécules régulatrices d'accéder aux gènes, ou au contraire en interdit l'accès.

La compaction tri-dimensionnelle de l'ADN dans les chromosomes est très corrélée avec la régulation transcriptionnelle de l'expression des gènes. La transcription individuelle des gènes est modulée par des facteurs de transcription dédiés. Un facteur de transcription est une protéine qui se fixe sur l'ADN en un site spécifique, dans la région régulatoire de son gène cible, dont il va ainsi activer ou réprimer la transcription. Le plus souvent, les facteurs de transcription sont bivalents, c'est-à-dire qu'ils peuvent se fixer sur deux sites similaires de l'ADN. F. Képès a montré que, chez la levure [4] et chez E. coli [3], les gènes cibles qui sont contrôlés par le même facteur de transcription tendent à être soit regroupés, soit régulièrement espacés le long du chromosome. Cette régularité facteur de transcription / cible est plus prononcée que la régularité cible / cible, ce qui suggère que la première engendre la seconde [3]. Le même intervalle est observé pour la plupart des facteurs de transcription sur le chromosome unique et circulaire du nucléoïde bactérien, ou sur un quelconque des 32 bras de chromosomes linéaires dans le noyau de levure.

En principe, on pourrait s'attendre à ce que cette régularité reflète une structure solénoïde dont chaque tour correspondrait à une période. Une première possibilité serait que ce solénoïde résulte de l'action d'un échafaudage physique autour duquel l'ADN s'enroulerait. Cependant, les périodes observées diffèrent entre bras de chromosomes de levure, ou entre souches d' E. coli, alors qu'on s'attendrait à une période constante si l'ADN était enroulé autour d'un échafaudage dédié. Donc il semble plus probable que la structure solénoïde provienne de la dynamique transcriptionnelle.

Spécifiquement, F. Képès a proposé que cette périodicité reflète un niveau supérieur de reploiement de l'ADN chromosomique, qui amène en un même lieu de l'espace géométrique plusieurs gènes co-régulés, autorisant ainsi un accès facile et efficace pour leurs facteurs de transcription associés. Cette

morphogenèse serait causée par la dynamique même qu'elle optimise. En effet la coexistence de plusieurs gènes cibles dans le même lieu favoriserait le recrutement de facteurs de transcription dans cette région. À son tour, une densité plus élevée de facteurs de transcription devrait recruter plus de gènes cibles. Ainsi, le phénomène s'auto amplifie. F. Képès and C. Vaillant suggèrent que cette rétroaction non-linéaire basée sur la cinétique d'association séquence-dépendante des facteurs de transcription bivalents et de leurs sites multivalents serait la force qui ploie l'ADN en une forme homomorphe à un solénoïde, rendant ainsi la transcription et le contrôle génétique plus efficaces [7]. L'explication mécanistique de l'optimisation transcriptionnelle qu'apporte un tel arrangement géométrique existe déjà [8, 9], et il suffit de l'extrapoler ici vers une situation intergénique, à partir d'un cas intragénique où elle a été abondamment établie.

L'idée n'est pas que l'ADN forme un parfait solénoïde géométrique. Il faut et suffit que le reploiement amène les gènes co-régulés dans le même voisinage; aussi s'attend-on à des distorsions locales de la structure, qui néanmoins doit être homomorphe à un solénoïde pour accommoder tous les facteurs de transcription.

2 Modélisation

La modélisation est l'outil privilégié pour approcher le processus évolutionnaire du positionnement génique. Est-ce que l'avantage sélectif qu'amène ce schéma global de transcription peut imposer la périodicité des gènes co-régulés malgré la plasticité des chromosomes à l'échelle temporelle de l'évolution ?

2.1 Etat de l'art

La modélisation du problème nécessite la prise en compte de deux aspects distincts :

i) D'une part les interactions entre gènes (pour la mise en place d'un véritable réseau de régulation).

ii) Et d'autre part la représentation physique du génome (qui sera nécessaire pour déterminer sa configuration spatiale).

Les modèles de réseaux de régulation que l'on trouve dans la littérature ne traitent que : soit le point i), soit le point ii).
Pour le point i) nous avons les modèles déterministes où l'expression des gènes est représentée par leurs concentrations et les interactions par des équations différentielles [11] ; puis les modèles de type René Thomas où les gènes sont représentés par une valeur discrète [10] ; et les réseaux de Petri dans lesquels sont représentés à la fois : gènes, protéines et réactions.
Pour le point ii) nous avons les modèles simulant la structure de l'ADN, pour obtenir une visualisation graphique, pour étudier des réactions chimiques, ou pour étudier les phénomènes d'enroulement et de torsion [12].

Le seul modèle qui tente de faire le pont entre la simulation de la structure physique d'un génome et les interactions géniques, est le modèle développé en 2004 par Sébastien Leclercq [6]. Il s'inspire du modèle de réseaux de régulation génétique évolutif développé par Wolfgang Banzhaf [5] auquel est ajouté une composante de contrainte de structuration spatiale du génome ainsi que des règles d'attractions et de courbures.

2.2 Présentation de notre modèle

Notre approche va dans le sens de celle développée par S.Leclercq, nous allons utiliser une version modifiée de réseaux de régulation génétique évolutif de W. Banzhaf pour modéliser les interactions géniques, auquel nous allons ajouter les règles de structuration spatiale de S. Leclercq, pour rendre compte de la structure physique de l'ADN. Une fois ceci mis en place, nous ajouterons la possibilité, par le biais de mutations changeant l'interconnexion des réseaux, de faire évoluer la carte des interactions et d'étudier informatiquement de possibles repliements alternatifs (on en connaît l'existence – par l'observation – au moins chez E. coli [3]).

Nous utiliserons un algorithme bio-inspiré pour représenter notre modèle, il contient plusieurs phases :

i) *Morphogenèse du génome* : simuler le repliement spatial d'un génome selon les interactions géniques du réseau de régulation.

ii) *Evolution par sélection naturelle* : utilisation d'un algorithme génétique dans le but d'étudier l'influence du réseau de régulation (nombre de gènes, nombre de réseaux).

iii) *Evolution des cartes d'interactions* : étude évolutive de l'interconnexion entre les différents réseaux de régulation (mise en évidence de repliements alternatifs). (Recherche en cours de développement).

Et nous aborderons dans le détail les trois points suivants :

i) *Structuration linéaire du génome* : le génome sera structuré selon un modèle basé sur le modèle de W. Banzhaf [5]. Nous ajouterons de plus la possibilité pour les gènes d'être "déployés", les mutations pouvant ainsi affecter directement leurs structures (ce qui aura pour effet de pouvoir modifier la carte des interactions).

ii) *Morpho-dynamique* : tout comme S. Leclercq, trois règles vont être utilisées pour simuler le repliement spatial du génome :

   a. *Attraction* : les facteurs de transcription et leurs cibles s'attirent mutuellement (bivalence des facteurs de transcription etc.).
   b. *Elongation* : la distance maximale entre deux gènes actifs ne peut excéder leur distance linéaire.
   c. *Courbure* : le génome possède un périmètre minimum de bouclage. Deux gènes situés à une distance linéaire inférieure à ce périmètre ne pourront se trouver à une distance plus petite que la corde d'arc correspondant.

iii) *Evolution génétique* : une fois la morphogenèse mise en place et testée, nous serons à même de faire évoluer à l'aide d'un algorithme génétique toute une population de génomes. Nous pourrons ainsi tester l'influence du nombre de gènes, du nombre de réseaux et des différents paramètres du modèle, sur l'évolution de l'organisation des gènes sur le génome et sur sa configuration spatiale correspondante.

2.3 Observations attendues

i) *Evolution des positions géniques* : sous les différentes contraintes que sont le nombre de gènes et le nombre réseaux de régulation, les génomes étudiés devraient produire des distributions de gènes co-régulés périodiques et des distributions en clusters périodiques. De telles distributions seraient un argument en faveur du modèle solénoïde (cette configuration étant la plus adaptée pour rapprocher spatialement les gènes co-régulés, d'un ou plusieurs réseaux, distribués périodiquement).

ii) *Repliements alternatifs* : La mise en place de mutations pouvant affecter directement la carte d'interaction (interconnexion des réseaux) permettrait de voir apparaître des repliements alternatifs qui seraient préférés au repliement standard sous une certaine pression de l'environnement. (Observation chez E.coli d'un positionnement périodique principal des gènes régulés par le facteur de transcription Crp tous les 92Kb et d'un positionnement périodique alternatif tous les 46Kb). (Recherche en cours de développement).

3 Réseaux de régulation artificiels

3.1 Modèle de W. Banzhaf

Pour rendre compte des interactions géniques dans notre simulation, nous allons utiliser un modèle dérivé du modèle de W. Banzhaf présenté ci-dessous.

Dans ce modèle, les gènes sont constitués d'une suite de bits (256 bits) divisé en quatre régions :

i) Zone Activatrice : 32 bits
ii) Zone Inhibitrice : 32 bits
iii) Zone Promotrice : 32 bits
iv) Zone ORF : 160 bits

Puis un mécanisme de reconnaissance / production est mis en place : la zone ORF du gène est lue et traduite en une séquence de bits (qui sera assimilée aux protéines) de même taille que les zones activatrices et inhibitrices. Si une protéine à le même motif que la zone activatrice ou inhibitrice d'un gène alors on dit que cette protéine régule ce gène qui sera alors appelé gène cible ; et le gène responsable de la production de cette protéine est appelé facteur de transcription. La reconnaissance entre le motif d'une protéine et la zone activatrice ou inhibitrice d'un gène peut être partielle et peut ainsi moduler l'intensité de la régulation. La zone promotrice est un motif spécial qui permet l'identification du commencement et de l'état (actif ou non) d'un gène.

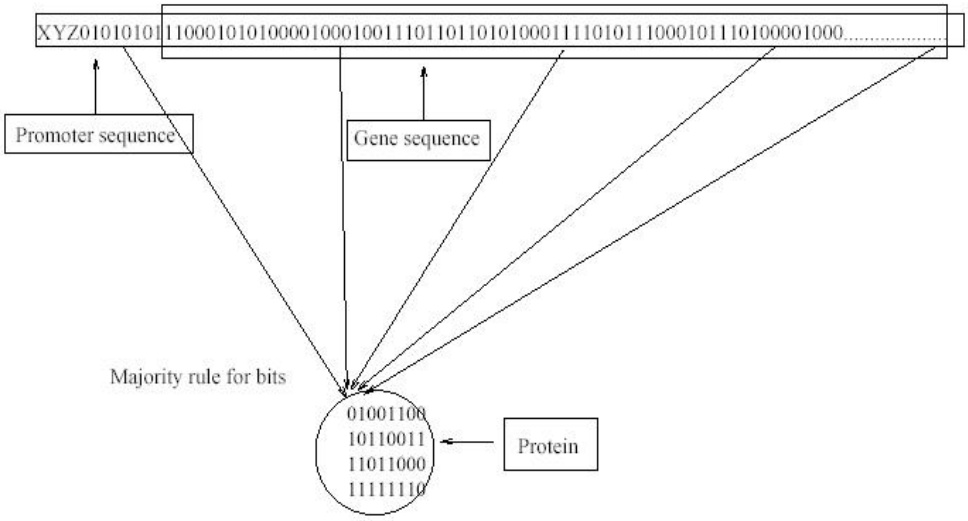


Figure 4.1. The genotype-phenotype mapping. Proteins are produced from genes via the genotype-phenotype mapping function.

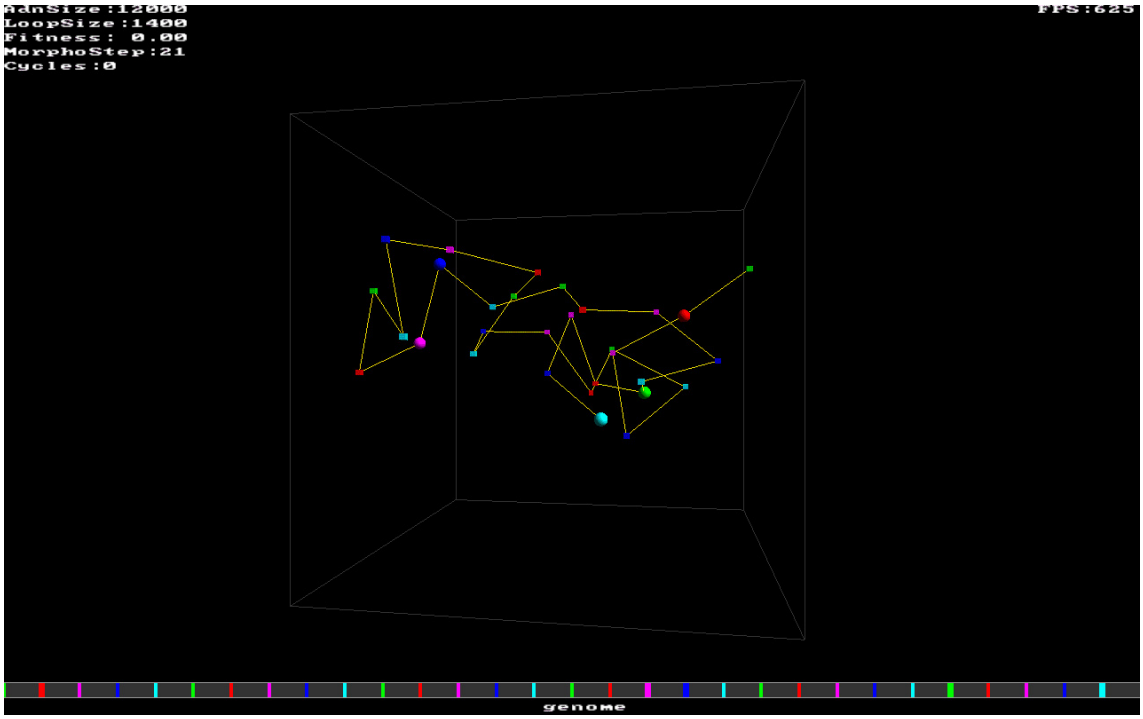
Schéma récapitulatif tiré de l'article de W. Banzhaf [5]

Le génome est alors constitué d'un ensemble de gènes mis bout à bout et contient des gènes actifs et des gènes potentiels (identifiés par leur zone promotrice) qui peuvent, grâce à un système de mutation, apparaître ou disparaître du réseau de régulation.

3.2 Nouveau modèle

Notre modèle reprend l'organisation générale du modèle de W. Banzhaf et traite des interactions entre les gènes selon les mêmes principes. Nous avons ajouté une composante de structuration spatiale du génome pour rendre compte de l'influence des réseaux de régulation sur l'influence de son repliement.

Partant de l'organisation linéaire d'un génome, on attribut à chaque gène actif une position dans un espace à trois dimensions. Nous avons entièrement mis au point un moteur de rendu 3D spécifique qui nous permet de visualiser en temps réel la topologie adoptée par le génome. La librairie graphique utilisée est OpenGL couplée à l'extension AllegroGL ; le programme est totalement portable.



Moteur de rendu 3D permettant de visualiser le repliement du génome

3.2.1 Dynamique et règles physiques

Le repliement du génome est imprimé par la dynamique transcriptionnelle : chaque facteur de transcription attire ses gènes cibles et inversement les gènes cibles attirent le facteur de transcription. Ces forces attractives mises en jeu sont inversement proportionnelles au cube de la distance entre gènes et facteur de transcription.
On ajoute de plus, pour s'approcher le plus possible de la réalité physique, deux règles induites par la physique moléculaire de l'ADN :

i) *Elongation* : la distance maximale entre deux gènes actifs ne peut excéder leur distance linéaire.

ii) *Courbure* : le génome possède un périmètre minimum de bouclage (contraintes physiques entre histones etc...). Deux gènes situés à une distance linéaire inférieure à ce périmètre ne pourront se trouver à une distance plus petite que la corde d'arc correspondant.

3.2.2 Morphogenèse

La règle d'attraction rendant compte de l'effet coopératif d'association des facteurs de transcription et de leurs gènes cibles, et les règles de structure de l'ADN, élongation et courbure, nous permettent de procéder à la simulation de la morphogenèse du génome. Pour avoir une cohérence globale de la structure du génome lors de l'évolution du repliement, on effectue une simulation de type asynchrone :

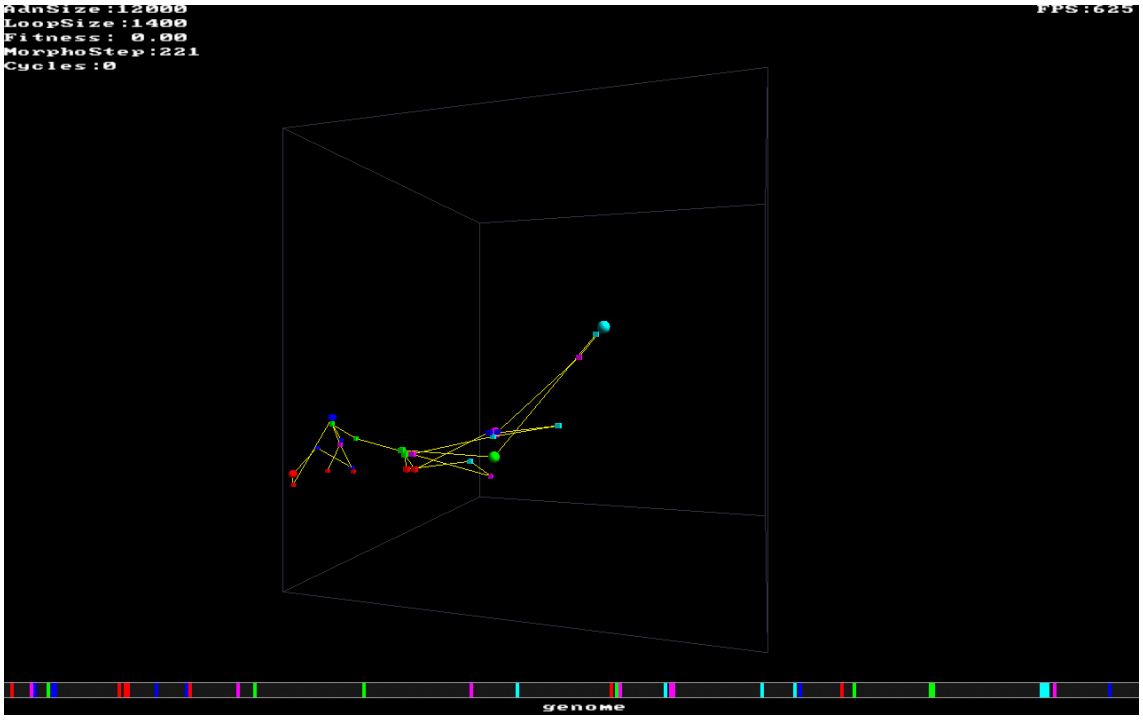
- On déplace un gène selon les différentes forces auquel il est soumis, puis on s'assure qu'il respecte les deux règles d'élongation et de courbure. On corrige éventuellement sa position, puis on procède de même pour tous les gènes actifs.

Notre contrôle sur l'évolution du repliement du génome intervient par le calcul de la valeur de distance moyenne d'interaction (moyenne des distances 3D entre les facteurs de transcription et leurs gènes cibles). Lorsque cette valeur devient stable, cela signifie que le génome à atteint son état de condensation maximal et que la morphogenèse est terminée.

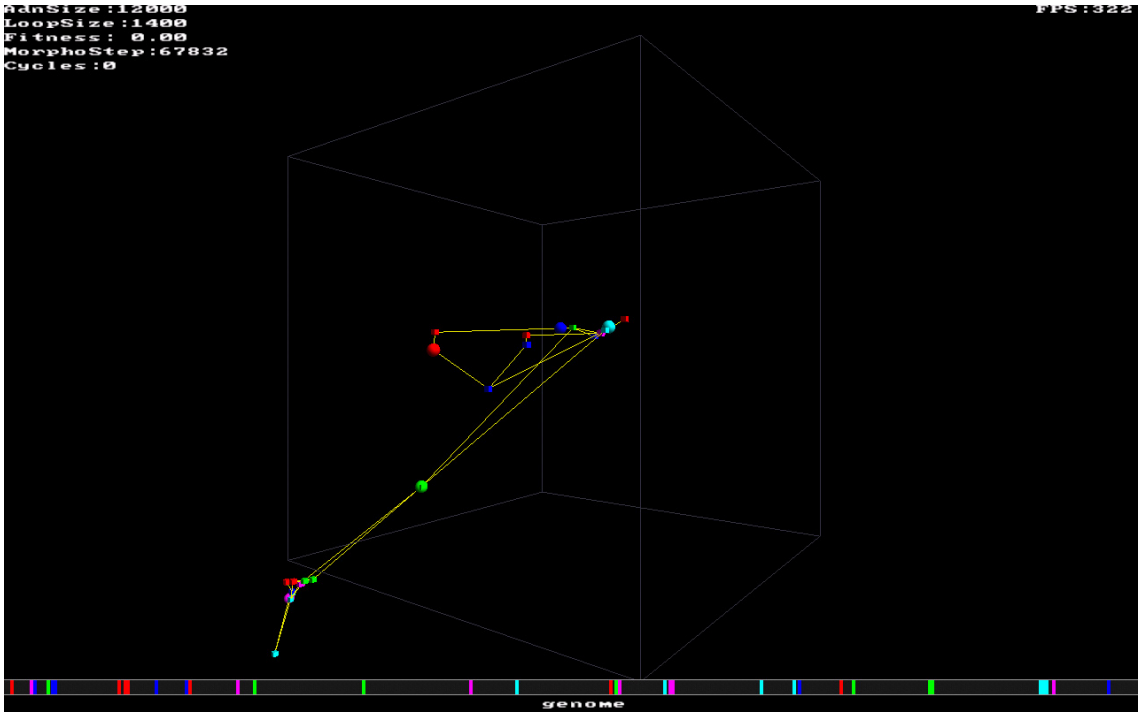
3.2.3 Résultats de la Morphogenèse

1ere famille : La distribution linéaire des gènes est aléatoire :

Nous obtenons une morphogenèse sans structure générale remarquable, les gènes co-régulés ne sont pas proches et le taux de compactage est faible. Voir ci-dessous les captures d'écran retraçant l'évolution d'une morphogenèse de ce type.



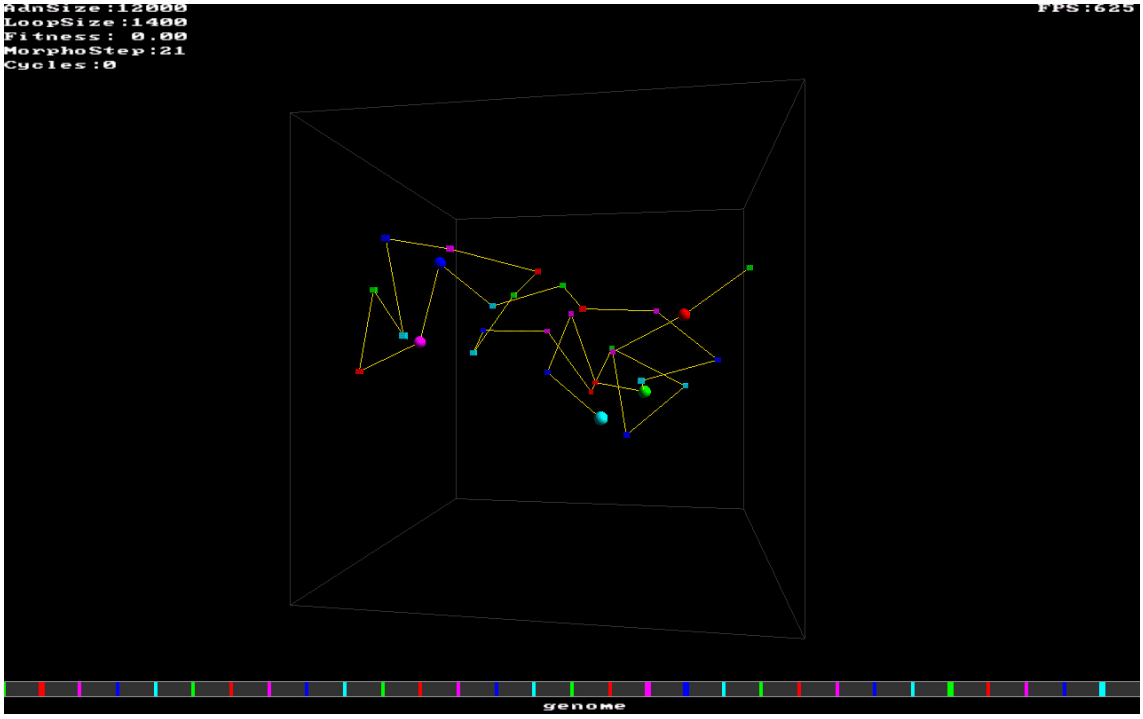
Distribution aléatoire : Début de la simulation



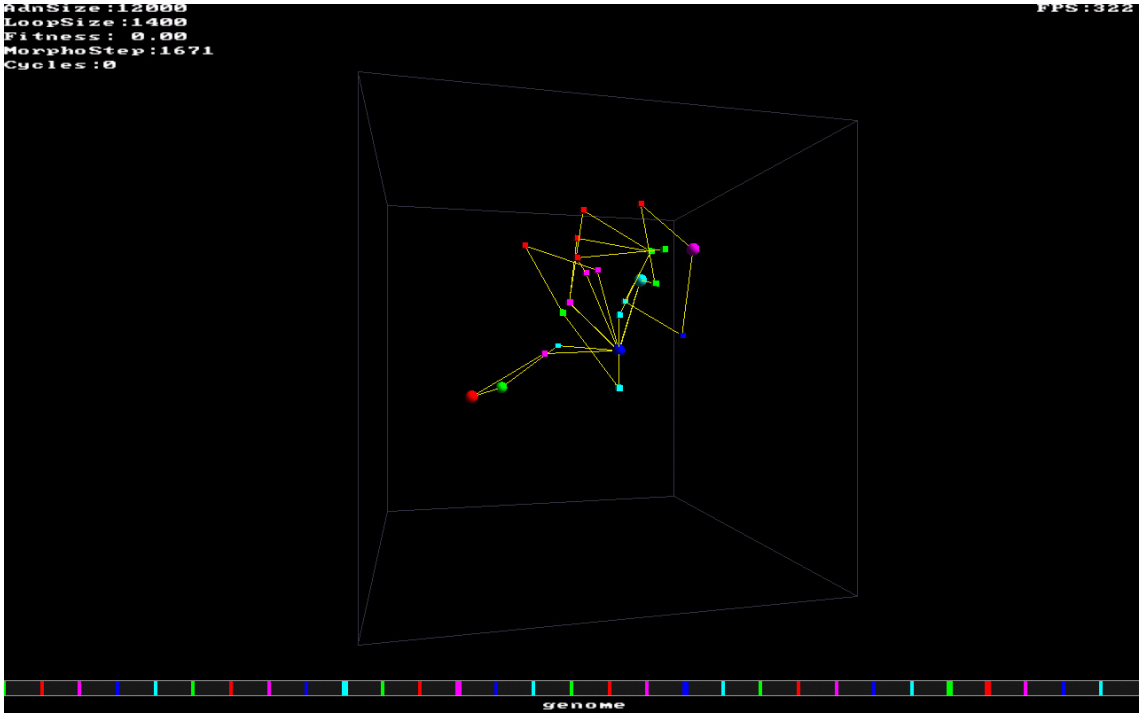
Distribution aléatoire : Mi-simulation

Distribution aléatoire : Morphogenèse stable

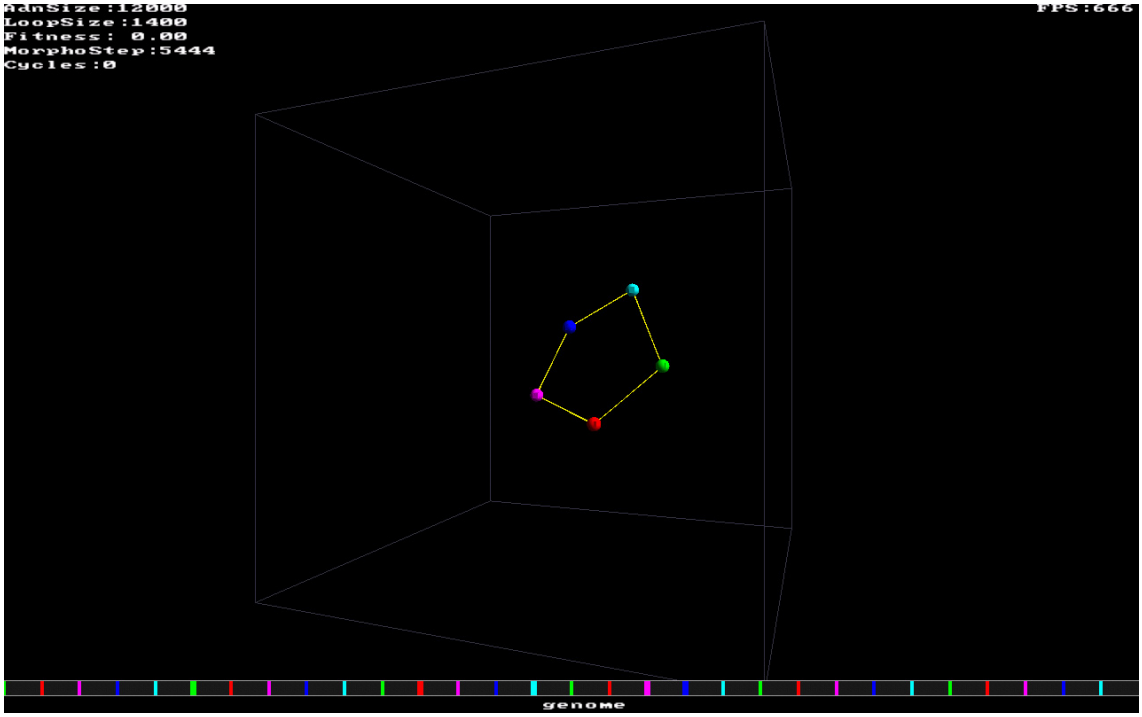
2ème famille : La distribution linéaire des gènes est périodique.



Distribution périodique *: Début de la simulation*



Distribution périodique *: Mi-simulation*



Distribution périodique *: Morphogenèse stable*

Apparition d'une structure homomorphe à un solénoïde

Pour les distributions périodiques, la morphogenèse obtenue fait apparaître que le génome adopte une configuration spatiale homomorphe à un solénoïde ; les gènes co-régulés se sont localisés dans la même zone spatiale et sont maintenant très proches, et le taux de compactage est très élevé. Ce qui confirme, du point de vue strict de la morphogenèse (définie telle que précédemment), qu'une répartition périodique des gènes co-régulés permet un repliement optimal sous la forme d'un solénoïde.

Pour démontrer maintenant que l'avantage sélectif qu'amène ce schéma global de transcription peut imposer la périodicité des gènes co-régulés, il nous faut mettre en place un mécanisme d'évolution, un algorithme évolutionnaire utilisant notre routine de morphogenèse, qui permettrait, en partant d'une distribution aléatoire de gènes, d'arriver, par le biais de mutations, à une

distribution finale globalement périodique qui révélerait la préférence évolutionnaire pour le modèle solénoïde.

3.3 Algorithme Génétique

Pour simuler l'effet de sélection naturelle sous la pression d'un environnement, nous avons développé un algorithme génétique qui va nous permettre de faire évoluer une population de génomes. Les distributions initiales des positions des gènes actifs sur ces génomes sont aléatoires. On cherche à démontrer qu'au cours de l'évolution, les positions géniques vont être de plus en plus régulières au fil des cycles d'évolution.

Notre valeur de Fitness est la valeur de distance moyenne d'interaction, la moyenne des distances 3D entre les facteurs de transcription et leurs gènes cibles. Cette valeur caractérise l'aptitude du génome à rapprocher dans une même région spatiale les différentes familles de gènes co-régulés et donc, lors de la morphogenèse, la capacité de repliement imprimé par la dynamique transcriptionnelle.

Chaque distribution de gènes sur un génome est évaluée avec notre routine de Morphogenèse qui exécute une simulation de repliement sous les différentes forces d'attractions et les différentes règles de structuration définies précédemment ; puis, à stabilité, nous sommes alors à même de déterminer la valeur de Fitness du génome pour la distribution donnée.

Nous sélectionnons ensuite les meilleurs individus de la population que nous clonons et faisons muter ; puis nous les réinsérons dans la population en remplaçant les moins bons individus. La taille de la population reste ainsi constante. Au vu du temps de calcul nécessaire à l'algorithme génétique, nous avons fixé une valeur de 10000 cycles de mutation, et la taille de la population n'excède pas vingt génomes.

Au stade de la rédaction de ce mémoire, les règles de mutations utilisées se résument à la transposition simple – déplacement de position linéaire d'un gène actif–. Sont prévues à venir les mutations affectant directement la structure des gènes, ce qui permettra de faire apparaître ou disparaître un gène d'un réseau (modification de la carte des interactions). Nous pourrons alors étudier la dynamique des variations de concentration de protéines

correspondantes et compléter ainsi le travail entreprit par W. Banzhaf, auquel il manque notre représentation physique du génome et sa configuration spatiale.

3.3.1 Résultats de l'algorithme génétique

Nous allons maintenant chercher à vérifier que l'algorithme génétique appliqué à l'optimisation par la morphogenèse que nous avons mis en place est à même de rendre compte de distributions périodiques – ou en clusters périodiques – de gènes. L'obtention de telles distributions serait un argument fort en faveur du modèle solénoïde. Les génomes étudiés ont une taille de 15000 bits.

Le premier et le second test correspondent à notre programme appliqué à un réseau comportant respectivement 8 gènes pour le test 1 et 30 gènes pour le test 2.

Test 1

1 réseau de 8 gènes : répartition initiale

1 réseau de 8 gènes : répartition finale

Test 2

1 réseau de 30 gènes : répartition initiale

1 réseau de 30 gènes : répartition finale

Afin de pouvoir juger du degré de régularité obtenu, nous allons étudier les résultats des différents tests de manière statistique. Pour comparer le plus objectivement possible la dispersion de distributions, il est nécessaire d'utiliser un indicateur susceptible d'exprimer l'hétérogénéité indépendamment de l'unité de mesure de la variable, du nombre et de l'ordre de grandeur des observations, on utilise généralement le coefficient de variation de la distribution de la variable X, à savoir l'écart-type divisé par la moyenne : $cv(X) = \sigma(X) / \bar{x}$.

Etude statistique du Test 1

Position 1D	Ecart relatif	Ecart moyen	Ecart-type	Coefficient de variation
501				
1877	1376			
3982	2105			
5633	1651			
7708	2075			
10121	2413			
11517	1396			
13786	2269	$\bar{x}$ = 1898	σ(X) = 363	cv(X) = 19%

Etude statistique du Test 2

Position 1D	Ecart relatif	Ecart moyen	Ecart-type	Coefficient de variation
512				
1407	895			
2754	1347			
3850	1096			
4821	971			
6271	1450			
7748	1477			
9193	1445			
10685	1492			
12138	1453			
13416	1278			
14589	1173	$\bar{x}$ = 1280	σ(X) = 179	cv(X) = 14%

Ces deux tests ont pour but de déterminer l'influence de la densité de gènes co-régulés sur un génome dans la répartition finale obtenue après application de l'algorithme génétique.

On observe comme résultats que, pour le génome comportant un réseau de 8 gènes on obtient une répartition périodique assez bien marquée – 19% de dispersion relative – ; et pour le génome comportant un réseau de 30 gènes on constate une distribution périodique plus nette – 14% de dispersion relative –, avec l 'apparition de clusters de deux ou trois gènes co-régulés au niveau des périodes. L'apparition de ces clusters est due au fait que le génome n'a pas la possibilité de placer périodiquement tous les gènes compte tenu de son périmètre de bouclage ; il "trouve" alors cette autre stratégie. Cette expérience reproduite plusieurs fois donne des résultats similaires.

Dans les tests trois tests suivants, nous nous sommes intéressés à des génomes comportant respectivement : 2 réseaux de 10 gènes, 3 réseaux de 10 gènes et 4 réseaux de 8 gènes.

Test 3

2 réseaux de 10 gènes : répartition initiale

2 réseaux de 10 gènes : répartition finale

Test 4

3 réseaux de 10 gènes : répartition initiale

3 réseaux de 10 gènes : répartition finale

Test 5

4 réseaux de 8 gènes : répartition initiale

4 réseaux de 8 gènes : répartition finale

Etude statistique du Test 3

Position 1D	Ecart relatif	Ecart moyen	Ecart-type	Coefficient de variation
VERT 786 2593 4185 5431 7085 8717 10305 11802 13358 14786 ROUGE	VERT 1807 1592 1246 1654 1632 1588 1497 1556 1428 ROUGE	$\bar{x}$ = 1555	σ(X) = 117	cv(X) = 7%

430 2163 3411 4885 6722 8297 9924 11589 13003 14436	 1733 1248 1474 1837 1575 1627 1665 1414 1433	$\bar{x}$ = 1556	σ(X) = 146	cv(X) = 9%

Etude statistique du Test 4

Position 1D	Ecart relatif	Ecart moyen	Ecart-type	Coefficient de variation
VERT 582 2306 3682 5231 6603 8208 10134 11643 13265 14712	VERT 1724 1376 1549 1372 1605 1926 1509 1622 1447	$\bar{x}$ = 1570	σ(X) = 133	cv(X) = 8%
ROUGE 181 1738 3234 4707 6243 7812 9385 11262 12623 14285	ROUGE 1557 1496 1473 1536 1569 1573 1877 1364 1662	$\bar{x}$ = 1567	σ(X) = 91	cv(X) = 6%
BLANC 826	BLANC			

2484 3901 5513 6826 8442 9974 11778 13164 14817	1658 1417 1612 1313 1616 1532 1804 1386 1653	$\bar{x}$ = 1554	σ(X) = 127	cv(X) = 8%

Etude statistique du Test 5

Position 1D	Ecart relatif	Ecart moyen	Ecart-type	Coefficient de variation
VERT 1340 3253 4534 6101 7532 8169 9257 11063	VERT 1913 1281 1567 1431 637 1088 1806	$\bar{x}$ = 1389	σ(X) = 332	cv(X) = 24%
ROUGE 476 2652 4387 6803 8689 10711 12764 14761	ROUGE 2176 1735 2416 1886 2022 2053 1997	$\bar{x}$ = 2041	σ(X) = 149	cv(X) = 7%
BLANC 1702 3416	BLANC 1714			

4737 5754 7220 9295 13101 13596 BLEU 585 2748 4253 5196 7087 9003 10037 10883	1321 1017 1466 2075 3806 495 BLEU 2163 1505 943 1891 1916 1034 846	$\bar{x}$ = 1699 $\bar{x}$ = 1471	σ(X) = 671 σ(X) = 456	cv(X) = 39% cv(X) = 31%

Ces trois tests ont été effectués pour déterminer l'influence du nombre de réseaux de régulation sur le génome dans la répartition finale obtenue après application de l'algorithme génétique.

On observe dans le test 3, où le génome comporte deux réseaux de régulation, des résultats similaires aux deux premières expériences : les deux réseaux se trouvent en alternance et les gènes se positionnent encore de manière périodiques – 7% et 9% de dispersion relative –.

Dans le test 4, où le génome comporte trois réseaux de régulation, on remarque que les gènes de deux des réseaux se sont regroupés – parfois avec une inversion d'ordre – et sont en alternance avec les gènes du troisième réseau. La régularité globale des gènes co-régulés des trois réseaux est toujours remarquable – 8% et 6% de dispersion relative –.

Pour la dernière expérience, nous avons un génome qui possède quatre réseaux de régulation. L'orientation que suggère le test comportant trois réseaux est vérifiée ; on voit que les gènes de trois des quatre réseaux ont tendance à se regrouper (à inversion d'ordre près) tandis que le quatrième

réseaux est plus libre et est en alternance avec ces trois réseaux. Nous obtenons toujours une distribution de gènes ayant une régularité, mais elle est moins marquée que dans les tests précèdent – 31%, 39%, 7% et 24% de dispersion relative –. Un plus grand nombre de cycles de mutation dans l'algorithme génétique amènerait très probablement à une meilleure régularité.

4 Conclusion et Perspectives

Ce projet a nécessité un effort important de mise en œuvre logiciel, avec un programme relativement complexe dont l'optimisation est un des points les plus difficiles, ainsi que des temps de calcul de plusieurs jours pour obtenir les résultats de l'algorithme génétique.

Nos hypothèses de départ était double, d'une part nous supposions que la sélection naturelle était responsable du positionnement périodique des gènes co-régulés le long des chromosomes et d'autre part que cette distribution était imprimée par la dynamique transcriptionnelle et favorisait une organisation de l'ADN en une structure de type solénoïde.

Dans un premier temps nous avons mis en place une routine effectuant la morphogenèse d'un certain génome donné et nous avons constaté qu'une distribution périodique de gènes co-régulés nous donnait bien l'obtention dans notre espace à 3 dimensions d'une structure d'ADN homomorphe à un solénoïde. Nous avons ensuite développé un algorithme génétique dont la fonction de Fitness provient de notre routine de morphogenèse. Nous avons vérifié, d'une part pour un seul réseau de régulation actif, d'autre part, pour plusieurs réseaux actifs, que, sous la pression de la sélection naturelle simulée, nous obtenions des distributions de gènes co-régulés périodiques et des distributions en clusters périodiques. Ce sont des arguments forts en faveur de la validité du modèle d'organisation solénoïde des chromosomes.

La partie concernant l'évolution de la carte d'interaction (interconnexion des réseaux) est toujours en cours de développement et doit se poursuivre l'année prochaine avec un financement de l'ANRS sur 3 ans. Dans cette partie nous devons développer des structures de données à même de pouvoir

supporter le aisément des mutations pouvant affecter le contenu même des gènes, ce qui permettra de pouvoir modifier dynamiquement la carte d'interaction des réseaux. L'étude de ce véritable système dynamique doit nous révéler entre autres, des repliements alternatifs qui seraient, sous certaines conditions environnementales, préférés au repliement standard. Ces repliements alternatifs ont récemment été observé in vitro chez E.coli s'adaptant à un stress oxydatif [3], [4]. L'étude du système composé des parties morphogenèse et algorithme génétique supportant les mutations intragéniques promet d'être très riche.

Glossaire

ADN (acide désoxyribonucléique)

Constituant essentiel des chromosomes, support moléculaire de l'information génétique. Le contenu de cette information est le "code" de synthèse de toutes les protéines de l'organisme. La molécule d'ADN est composée de 2 brins, constitués chacun d'un enchaînement de nucléotides.

ARN (acide ribonucléique) messager

Molécule correspondant à la copie de la séquence codante d'une portion d'ADN, qui assure le passage de l'information génétique dans le cytoplasme de la cellule, où a lieu la synthèse de la protéine.

Chromatine

Substance de base des chromosomes des eucaryotes, correspondant à l'association de l'ADN et des protéines histones.

Chromosome

Molécule d'ADN associée à des protéines qui est présente dans le noyau des cellules. Pour chaque espèce, le nombre de chromosomes par cellule est constant (23 paires pour l'homme). Les cellules eucaryotes (possédant un noyau individualisé) comportent plusieurs chromosomes ; les cellules bactériennes n'en comportent qu'un.

Délétion

Perte d'une partie du matériel génétique pouvant aller d'un seul nucléotide à plusieurs gènes.

Eucaryote

Se dit d'une cellule pourvue d'un noyau figuré (opposé à procaryote).

Facteur de transcription

Protéine qui régule la transcription d'un gène en se fixant sur son promoteur (au niveau des sites de liaison).

Gène

Elément de base du patrimoine génétique qui code pour une protéine ; chaque cellule humaine en contient quelque 30 000 répartis sur 23 paires de chromosomes soit le jeu complet de ses gènes, cependant tous ne s'expriment pas (ne synthétisent pas leur protéine) dans toutes les cellules et tout le temps.

Génome

Ensemble du matériel héréditaire caractéristique d'une espèce ou d'un individu, son support chimique est la molécule d'ADN.

Génomique

Science qui étudie la structure, le fonctionnement et l'évolution des génomes.

Histone

Protéine basique, constituant majeur du nucléosome.

Insertion

Addition d'une séquence d'ADN étranger dans une molécule d'ADN donnée.

Mutation

Modification de la séquence d'ADN par changement d'une (ou plusieurs) base(s) en une autre base.

Nucléosome

Est l'unité fondamentale de la chromatine. Il est composé d'ADN et d'histones. Il constitue le premier niveau de compaction de l'ADN dans le noyau.

Opéron

Unité de transcription constituée par un promoteur, un opérateur et un ou plusieurs gènes de structure.

Phénotype

Manifestation apparente de la constitution du génome.

Génotype

Ensemble des caractères génétiques d'un individu. Son expression conduit au phénotype

Procaryote

Organisme dont la cellule ne possède pas de noyau, contrairement aux eucaryotes. Le génome est constitué d'un unique chromosome circulaire. Il peut aussi coexister des structures extra chromosomiques, sous forme de molécules d'ADN circulaires libres dans le cytoplasme (plasmides), qui se répliquent de façon indépendante.

Promoteur

Séquence d'ADN nécessaire à l'initiation de la transcription et le plus souvent située en amont de la partie transcrite des gènes.

Protéine

Grande molécule constituée par l'enchaînement d'un grand nombre d'acides aminés.

Recombinaison génétique

Ensemble de mécanismes conduisant au réarrangement de séquences d'ADN.

Régulation

Contrôle de l'expression, ou activité, d'un ou de plusieurs gènes.

Traduction

Processus permettant la synthèse d'une chaîne polypeptidique (protéine) à partir d'un brin d'ARN messager. La traduction a lieu au niveau des ribosomes.

Transcription

Processus permettant la copie de l'ADN en ARN. L'ARN synthétiser peut-être de type ARN messager, de transfert ou ribosomique. C'est la première étape du processus qui permet de passer de l'ADN à la protéine, ou plus concrètement du gène à son produit.

Références

[1] Cockell, M, et M & Gasser, S. (1999). Nuclear compartments and gene regulation Curr. Opin. Genet. Dev. 9:199-205

[2] Jackson, D. A., Iborra, F. J., Manders, E. M.M. et Cook, P. (1998). Numbers and organisation of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei Mol. Biol. Cell. 9:1523-1536

[3] Képès, F. Periodic transcriptional organization of the *E. coli* genome. J. Mol. Biol. 340, 957-964 (2004).

[4] Képès, F. Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites. J. Mol. Biol. 329, 859-865 (2003).

[5] Banzhaf, W. (2003). Artificial regulatory networks and genetic programming, Genetic Programming Theory and Practice chapter 4, pages 43-62. Ed. Rick L. Riolo et Bill Worzel.

[6] Leclercq, S. (2004). Influence des interactions transcriptionnelles sur le positionnement des gènes le long d'un génome artificiel. Mémoire de DEA AMIB, Univ. Évry.

[7] Képès, F. & Vaillant, C. Transcription-based solenoidal model of chromosomes. ComPlexUs 1, 171-180 (2003).

[8] Müller-Hill, B. The function of auxiliary operators. Molec. Microbiol. 29, 13-18 (1998).

[9] Dröge, P. & Müller-Hill, B. High local protein concentrations at promoters: strategies in prokaryotic and eukaryotic cells. Bioessays 23, 179-183 (2001).

[10] THOMAS, R., "Regulatory networks seen as asynchronous automata : a logical description.", J. Theor. Biol. 153 ,(1991) 1-23.

[11] Mads Kaern, William J. Blake & James J. Collins. The engineering of gene regulatory networks. (Review article). Annu. Rev. Biomed Eng., 5, 179-206 (2003).

[12] Bouchiat C, Wang MD, Allemand J, Strick T, Block SM, Croquette V. Estimating the persistence length of a worm-like chain molecule from force-extension measurements. Biophys J. 1999 Jan;76(1 Pt 1):409-13.