Relations entre l'organisation des sites de fixation des facteurs de transcription, la fonction des gènes et l'expression des gènes chez Arabidopsis thaliana: vers une annotation des sites de fixation.

Virginie Bernard

Université d'Evry-Val d'Essonne

Ecole Doctorale des Génomes Aux Organismes

Thèse

Présentée pour obtenir le grade de Docteur en sciences
de l'université d'Evry-Val d'Essonne

Spécialité Bioinformatique

par

Virginie Bernard

Relations entre l'organisation des sites de fixation des facteurs de transcription, la fonction des gènes et l'expression des gènes : vers une annotation des sites de fixation chez *Arabidopsis thaliana*

Soutenue le 11 décembre 2009, devant le jury composé de :

Dr. Jacques van Helden, Directeur du laboratoire de Bioinformatique des Génomes et des Réseaux, Bruxelles	Rapporteur
Pr. Alain Denise, Professeur à l'Université Paris-Sud 11	Rapporteur
Pr. Bernard Prum, Professeur à l'Université d'Evry-Val d'Essonne	Examinateur
Dr. Thierry Lagrange, Directeur de Recherche CNRS, Perpignan	Examinateur
Dr. Alain Lecharny, Directeur de Recherche CNRS, Evry et Orsay	Directeur de thèse

Université d'Evry-Val d'Essonne

Ecole Doctorale des Génomes Aux Organismes

Thèse

Présentée pour obtenir le grade de Docteur en sciences
de l'université d'Evry-Val d'Essonne

Spécialité Bioinformatique

par

Virginie Bernard

Relations entre l'organisation des sites de fixation des facteurs de transcription, la fonction des gènes et l'expression des gènes : vers une annotation des sites de fixation chez *Arabidopsis thaliana*

Soutenue le 11 décembre 2009, devant le jury composé de :

Dr. Jacques van Helden, Directeur du laboratoire de Bioinformatique des Génomes et des Réseaux, Bruxelles	Rapporteur
Pr. Alain Denise, Professeur à l'Université Paris-Sud 11	Rapporteur
Pr. Bernard Prum, Professeur à l'Université d'Evry-Val d'Essonne	Examinateur
Dr. Thierry Lagrange, Directeur de Recherche CNRS, Perpignan	Examinateur
Dr. Alain Lecharny, Directeur de Recherche CNRS, Evry et Orsay	Directeur de thèse

Remerciements

La page des remerciements, la page la plus lue d'une thèse… Je n'ai pas eu la fameuse «angoisse de la page blanche» en écrivant cette thèse. Mais c'est en écrivant mes remerciements, que j'ai le plus hésité. Cette partie doit en quelques ~~lignes~~ pages présenter toute ma reconnaissance à mon entourage. Un challenge qui n'est pas si simple, tout thésard et tout docteur sera d'accord je pense. Pendant la rédaction de cette thèse, je me suis constituée une liste des personnes à remercier. Des personnes qui m'ont aidée à faire avancer le projet de ma thèse bien sûr, mais aussi toutes les personnes qui m'ont permis de décompresser, de prendre un peu de recul par rapport à mon travail! Les personnes qui m'ont soutenue, qui m'ont encouragée… qui étaient là en cas de besoin. J'espère n'avoir oublié personne dans ces pages… Je présente mes excuses si c'est le cas.

Tout d'abord, je remercie Michel Caboche, ancien directeur de l'Unité de Recherche en Génomique Végétale ainsi que Héribert Hirt, directeur, et Anne-Françoise Adam-Blondon, vice-directrice de l'unité depuis le milieu de ma thèse qui m'ont tous les trois accueillie depuis mon Master 2. Je remercie également l'Ecole Doctorale «des Génomes Aux Organismes» de l'université d'Evry, qui a financé ces trois années de thèse.

Je remercie Alain Lecharny, mon directeur de thèse, alias mon «grand chef» de m'avoir accueillie au sein de l'équipe de Bioinformatique de l'URGV pour ces quelques années. Merci Alain d'avoir cru en moi pour commencer une thèse après mon Master professionnel. Merci aussi de m'avoir laissée me «dissiper» à mes activités associatives, à mes recherches de postdoc tout en me faisant confiance pour la progression du travail de thèse. Merci enfin d'avoir guidé mes premiers pas de chercheuse tout en me laissant des libertés qui m'ont permis d'explorer mes propres pistes et d'apprendre aussi par moi-même. Un grand merci aussi pour le temps passé à me lire et me relire en particulier ces derniers temps… pour les articles, pour la thèse, pour les résumés divers que j'ai rédigés… Et merci pour ta patience quand ta thésarde spammeuse te demandait beaucoup ☺. Merci Alain !!

Je remercie Véronique Brunaud, qui a co-encadré ce travail et a été ma «petite chef» pendant mes années de thèse, après avoir été ma responsable de stage pendant mon Master2. L'adjectif «petite» devant chef n'est pas du tout proportionnel à la reconnaissance que j'ai pour toi Véro. Tu as été là dès mes débuts à l'URGV. Tu m'as toujours fait confiance pour cette thèse. Un grand grand merci pour ton écoute et tes encouragements pendant certains moments plus difficiles. Tu as aussi été disponible pour participer avec moi à une formation encadrant/doctorant, pour témoigner à la journée satellite de JOBIM que j'ai co-organisée, pour assister à ma présentation du Nouveau Chapitre de la Thèse. Merci pour tout ça, ce sont des choses qui ont compté et je suis heureuse que tu aies été là ! Je n'aurais jamais été jusqu'au bout sans toi, j'en suis convaincue, donc un grand merci, encore ! Un grand merci aussi pour les lectures et les corrections de la thèse et des articles ! Toi aussi je t'ai mené la vie dure … Merci d'avoir comme Alain toujours répondu à mes demandes fréquentes ces derniers temps. J'espère très sincèrement te revoir après cette thèse. A Vancouver, février 2011 c'est ça ☺ ?!? Je réserve les pistes de ski ☺ !

Je remercie Eric Ruelland et Philippe Bessières d'avoir été membres de mon comité de thèse. Vous avez contribué à l'amélioration de cette thèse en donnant votre avis sur les publications et en m'aidant à corriger mes présentations orales, en soulevant que je faisais quelques simplifications de langage mal choisies. J'espère ne plus (trop) les faire, j'y travaille !

Je remercie les membres de mon jury, mes rapporteurs, Alain Denise et Jacques van Helden, ainsi que mes examinateurs, Bernard Prum et Thierry Lagrange, d'avoir accepté d'évaluer mon travail. Je vous remercie tous de m'avoir fait l'honneur d'assister à ma soutenance, je vous remercie aussi sincèrement pour le travail que vous avez fait préparer notre discussion.

Je tiens également à remercier mes collègues de bureau, qui ont été nombreux au cours de ces trois années de thèse ! C'est vous qui m'avez le plus supportée, moi et mes «désagréments au quotidien». Je pense notamment à ma musique, la variété française, qui n'est pas du goût de tout le monde, à ma grande frilosité, chauffage en hiver et pas trop de climatisation en été, et aussi à mes sursauts au moindre bruit (certains diront mes cris de terreur, il ne faut pas trop les écouter). Je remercie mes collègues qui ont souffert de tout cela (enfin pas trop j'espère !): Cécile, David, Magali, Yang et Romain pour mon bureau de début de thèse, mais aussi Séverine, Marie Laure et Alain, collègues de bureau à mi-temps pour mon bureau de fin de thèse.

Je remercie sincèrement certains anciens collègues partis il y a un an et qui m'ont manqué. Je pense au Dr. David Armisén ☺, parti en postdoc en Irlande après sa soutenance il y a un an. Moi aussi j'ai beaucoup apprécié nos «chat» comme tu disais. Aujourd'hui, un an après toi, c'est à moi d'écrire mes remerciements en début du manuscrit de thèse. Tu avais raison, ça fait plaisir d'en être à ce stade. Je remercie également du fond du cœur Séverine pour sa spontanéité, son respect, sa très grande gentillesse et sa disponibilité. Merci de m'avoir encouragée dans cette aventure de la thèse lors de mes moments de doutes ! Je te remercie aussi pour ta patience lorsqu'une sonnerie de portable t'a un peu (beaucoup) dérangé pendant une nuit du brainstorming et pour ton courage lorsque ta voisine de bureau a failli te faire avoir une attaque cardiaque. Enfin un grand merci pour tes corrections des fautes de cette thèse, même entre deux sorties escalades. Tu as été une collègue, voisine et tu es devenue une amie. L'an prochain, à JOBIM, je mangerai peut-être encore avec ton chef, mais aussi avec toi j'espère ☺.

Je remercie l'ensemble de mes collègues de l'équipe bioinformatique. Cécile, petite dédicace… je vais faire les remerciements dans le même ordre que celui des bureaux ☺.

Je remercie Sébastien pour sa disponibilité pour discuter de questions scientifiques (ou non), et pour son volontariat pour relire des textes… Chacun de tes commentaires a apporté un plus à ma thèse. Merci pour ta minutie ☺. Merci aussi pour la discussion «cafétéria» ☺. Elle m'a été bien utile. Enfin, merci pour les soirées «bioinfo-ludiques» que j'ai beaucoup appréciées ! Vivement janvier !

Je remercie Jean-Philippe Tamby, pour ses excellents conseils esthétiques afin d'améliorer mon site Web, mais aussi pour ses commentaires pertinents sur n'importe quel texte… Tu prends toujours ça très au sérieux et avec ta lecture minutieuse tu déniches l'erreur que personne n'avait vue… Ce document a été nettement amélioré après ta relecture. Un grand merci. Et une question… mais quel est ton secret?? Et un merci spécial pour les «Dodo» du pays ☺ et pour les «Poulet Tandoori»…

Je remercie Franck Samson pour son soutien en particulier lors de mes démarches de recherche de postdoc. Ton intérêt m'a fait très plaisir, et tu m'as toujours encouragée à poursuivre. Ça m'a motivée à continuer l'aventure ! Merci aussi pour les pauses café bien agréables que j'ai souvent passées avec toi, à parler de tout et de rien ☺. Merci pour tes encouragements pour la rédaction et pour tes programmes écrit plus rapidement que ton ombre quand j'en avais besoin. Et bonne continuation à Jouy-en-Josas ☺ !! Contente pour toi, très sincèrement ! Et le champagne alors ???

Je remercie Philippe Grevet, dit «Phiphi», et Jean-Luc Collomb grâce à qui le système informatique de l'URGV fonctionne au mieux. Phiphi, tu m'as bien dépannée à différentes reprises, suite à un «rm *» (dans mon répertoire contenant tous mes programmes… tu te souviens ?) mais aussi suite à divers problèmes informatiques. Tu es toujours disponible pour les autres ☺. Tu vas me manquer chez les Grizzli comme tu dis ! Jean-Luc, tu m'as permis d'avoir de la couleur pour imprimer ma thèse… et ça… c'était un gros soulagement !. Merci à tous les 2 !

Je remercie Marie-Laure Martin-Magniette, dite Martin-Martignette pour ses conseils en statistiques à différents moments de mon Master 2 mais aussi et surtout pendant ma thèse. Tu as toujours été disponible pour discuter. Merci également pour tes conseils pour mener à bien cette thèse et pour tes recommandations pour certaines formations que j'ai particulièrement appréciées.

Je remercie Cécile Guichard, pas pour le café comme d'autres l'ont fait (dédicace), mais pour son écoute et sa compréhension. Tu as toujours été disponible. Merci pour ton soutien Cécile, pour tes encouragements et tes mails qui me faisaient bien plaisir quand tu étais en congé. Je te remercie aussi pour ton intérêt et tes encouragements pour différentes démarches, notamment celles de recherche de postdoc ☺. Un énorme merci également pour le temps passé à traquer les fautes dans ce document … c'est fou les bêtises qu'on peut laisser hein ! Et merci aussi pour ta patience pour tenter de comprendre mes explications parfois peu claires sur une fameuse courbe… mais promis… tu comprendras avant mon départ ☺. Je m'y engage ! Si si !!! Un grand merci Cécile pour tout…!

Je remercie enfin Odile, Sandra, Marta et Audrey pour les quelques repas, pauses 4h ou café partagés. Bonne poursuite les filles ☺. Et on encourage plus souvent les thésards… mais aujourd'hui, je voudrais aussi encourager Odile… dont le cher et tendre doit finir de rédiger sa thèse ☺.

Je remercie deux équipes avec qui j'ai crée des liens au cours de la thèse. L'équipe de Vincent Colot, partie à l'ENS à Paris. Je remercie plus particulièrement Agnès B., Alexis, Felipe, Martine, qui m'ont toujours accueillie avec beaucoup de gentillesse et d'amitié lorsque je suis passée leur faire un coucou. Agnès merci de m'avoir sortie pendant ma période «rédaction en autarcie». Je remercie aussi l'équipe de Jean Pierre Renou et de Claire Lurin avec qui j'ai créé des liens qui resteront je l'espère.

Je remercie aussi les personnes de l'URGV qui contribuent à maintenir une ambiance détendue et des liens entre les membres de l'unité. Tout d'abord, le COES, ou Comité d'Organisation des Evènements Sympa. Continuez ces petites fêtes les filles, ce sont des moments de détente bien agréables ! Et puis, merci aussi à Sandrine Balzergue d'avoir été officiellement «responsable anniversaire» avec moi pendant quasiment mes cinq années à l'URGV. J'espère que même en baisse l'équipe des anniversaires «été» assurera l'an prochain ☺. Enfin, je remercie le groupe du volley du mercredi… ça n'a pas toujours été facile avec moi et mes deux mains gauches, mais ça a été des moments bien sympa. Aujourd'hui, mission accomplie, je sais quelles sont les limites du terrain!

En quelques mois, 5 thésards de l'unité vont soutenir: Samuel, Mathieu, Imen, Aloïs et moi. Une pensée pour chacun... Et bon courage s'il vous reste encore à boucler article ou manuscrit ! Merci tout particulièrement à Samuel pour sa franchise, ses bons conseils et son humour. J'ai beaucoup apprécié nos conversations et j'espère sincèrement qu'on restera en contact après cette thèse. Une belle occasion de faire du tourisme: se rendre visite !!! Bon courage à tous les thésards qui soutiendront plus tard: Cléa, Ronan, Aymeric, Souha, Amélie, Clément.

Je remercie également mes anciens professeurs du Master de Bioinformatique de Rouen, qui restera pour les anciens dont je fais partie le «Master EGOISt» ! Je pense en particulier à Hélène Dauchel, qui m'a accompagnée pendant mon apprentissage, et qui m'a fait confiance pendant mes années de Master, mais aussi après. Merci Hélène de m'avoir confié les enseignements «annotation des promoteurs» aux M2.2, cette expérience m'a énormément plu ! Merci aussi de m'avoir fait confiance, avec Sandrine Rousseau, pour organiser les 10 ans de la formation. Comme dirait Sandrine, «EGOISt un jour, EGOISt toujours». J'en profite pour te dire merci Sandrine pour tes encouragements et tes mails en période «rédaction de thèse». J'espère bien qu'on se rendra visite au Canada !!! Côte Ouest puis Est.

Je remercie également l'équipe des réflexives et l'équipe du Nouveau Chapitre de la Thèse, en particulier Barbara Filler qui m'a permis de bien réfléchir à ma thèse et mon orientation.

Je remercie tous mes collègues qui m'ont aidée et conseillée pour mes démarches postdoc, ceux de l'URGV, de l'ENS, de Rouen, de Paris 7, des associations. Merci aussi à Alain Lecharny, Hélène Dauchel, Aurélien Mazurie et Alexandre de Brevern pour leurs lettres de recommandations. Et enfin un GRAND merci à Jennifer pour sa disponibilité pour les corrections d'anglais en tout genre, CV, site web, diapositives… Si tu as besoin de quoi que ce soit, n'hésite pas à me demander !

Je remercie le conseil d'administration des deux associations dans lesquelles je travaille.

Tout d'abord Alexandre de Brevern, Jean Tristan Brandenbourg, Guilhem Faure, Peter Schmidtke, Sébastien Lementec, Benjamin Broyer et Cyprien Guerin, de l'association des anciens élèves de la Licence et du Master de Paris 7 (AMBI pour les intimes). Je vous ai un peu moins accompagnés ces derniers mois à cause de ma thèse. Désolée, et promis, je me rattrape pour la journée rencontre en novembre. Dans tous les cas, les 2 années passées avec vous à gérer cette association ont été réellement très bénéfiques pour moi et chaque réunion me donnait plus de motivation pour retourner au travail le lendemain. Un merci tout particulier à Guilhem, pour ses mails détente et à Alex, toujours là (à toute heure) pour de très bons conseils, tous sujets confondus. J'espère pouvoir poursuivre l'aventure avec vous-même ! Dans tous les cas, je pense bien à vous JT, Guilhem, Peter pour votre fin de thèse. Continuez même si il y a des moments difficiles ! Et j'ajouterai longue vie à la gazette, longue vie à Mr tortue ☺ et longue vie à l'AMBI !

Je remercie Magali Michaut, David Thybert, Cédric Saule, Anne-Laure Gaillard, Loic Paulevé, Florent Boussardé, de l'association qui a fait parler d'elle… l'association des «jeunes bioinformaticiens de France» (alias JeBiF). J'ai participé avec vous à l'aventure journée satellite à JOBIM, une belle expérience. Un merci tout particulier à Magali, qui même à 7h du matin, à peine réveillée, est prête à discuter avec une thésarde en pleine recherche de postdoc. David, si tu as gagné ton pari «clope» (ce que je souhaite !), chaque année, tu gagneras 2 jours et demi pour profiter de ton fils ! Si ça ne te motive pas je ne comprend pas ☺ ! Bon courage à toi à l'EBI, et aussi bon courage à toi, Cédric, Anne-Laure, Loïc pour vos fins de thèses. Longue vie à JeBif !

Je tiens aussi à remercier sincèrement mes amis… je pense en particulier au groupe IMA, au groupe ZION, au groupe P7, aux copains d'Angers, Estelle, Jo, Sophie & co. Ces derniers mois, je vous ai peu ou pas vus. Merci à vous tous d'avoir compris mon implication dans mon travail, de m'avoir encouragée pour la rédaction, et surtout merci de m'avoir apporté des bouffées d'énergies à différentes occasions lors de soirées jeux ! Allez on se programme une murder ? J'organise ! Enfin, je remercie ma petite Cyanne qui m'a permis de faire des coupures après les journées de travail. J'espère que tu viendras nous voir vite !

Je remercie particulièrement des amis qui sont d'anciens thésards aujourd'hui docteurs pour leurs encouragements en phase «rédaction». Je pense déjà à Anne-Laure, Mériam et Fabrice… merci pour vos mails et coups de téléphone. On se voit bientôt, promis ! Merci tout particulièrement à Greg, mon coach de recherche de postdoc, mon QG lors de mes déplacements en Amérique du Nord … que dis-je… mon MENTOR ☺. Bref… qu'est ce que j'aurais fait sans toi Greg ?!!?!!? ☺ Merci, un grand merci ! Et bientôt, c'est à toi de venir me rendre visite ☺.

Je remercie aussi ma famille et ma belle-famille… en premier lieu ma maman, qui m'a soutenue pendant toutes ces années, qui m'a encouragée dans chacune de mes démarches et qui m'a permis d'être aujourd'hui là où je suis. Merci aussi à ma belle-famille qui a compris mes absences j'espère. Un pardon serait plus approprié qu'un merci pour famille et belle-famille peut-être... Promis, je me rattrape dès que possible ! Pardon en particulier à toi Martine d'emmener ton fils si loin, mais on espère que tu viendras nous voir ☺. On compte sur toi !

Enfin, the last but not the least, Will je te remercie pour ta patience à m'écouter parler de mon projet (que tu peux en partie présenter aujourd'hui j'en suis sûre !), pour tes conseils en statistiques, pour ta compréhension lorsque j'étais moins (pas du tout ?) disponible, pour ta persévérance pour m'encourager en cas de baisse de confiance. Il manque de la place pour tout te dire, mais merci, un grand merci du fond du coeur… cette thèse est aussi pour toi, parce que je sais que tu en as tout de même bavé avec ta thésarde lorsque je ne faisais que rédiger pendant plusieurs semaines... sans ne plus rien gérer. Aujourd'hui, je dois aussi te dire merci d'accepter de me suivre à Vancouver. Merci d'avoir été toujours si motivé pour ce projet. Il va bientôt se concrétiser. Je suis très heureuse de pouvoir me lancer dans cette aventure avec toi ☺.

Sommaire

Liste des figures

Introduction

Résultats et discussions

Conclusions générales

Liste des tables

Introduction

Résultats et discussions

Conclusions générales

Abréviations

Abréviations officielles

ADN	Acide Désoxiribo-Nucléique
ADNc	ADN complémentaire
ARN	Acide Ribo-Nucléique
ARNm	ARN messager
ARNt	ARN de transfert
ARNpm	ARN pré-messager
ARN pol II	ARN polymérase II
bp	Paire de bases
BRE	Elément reconnu par TFIIB
ChIP	Immunoprécipitation de la chromatine
chip-ChIP	Puce à ADN réalisée à la suite d'une ChIP
CRM	Module d'éléments régulateurs
DCE	Elément central en aval
DPE	Elément en aval
EST	Marqueurs de séquences transcrites
GO	Ontologie des gènes
GTF	Facteur de transcription général
GST	Séquence spécifique d'un gène
Inr	Elément initiateur
kb	Kilo base
MAP	Protéine activée par des mitogènes
MTE	Elément de 10 bases
TAF	Facteur de transcription associé à la TBP
TBP	Facteur de transcription reconnaissant la boîte TATA
TIC	Complexe d'initiation de la transcription
TF	Facteur de transcription
TFBS	Site de fixation de facteur de transcription

TSS	Site d’initiation de la transcription
UTR	Région non transcrite

Abréviations non officielles

FF	Fenêtre fonctionnelle
HE	Forte intensité d’expression
HE+	Très forte intensité d’expression
LE	Faible intensité d’expression
LE+	Très faible intensité d’expression
NS	Non significatif
PP	Position préférentielle
PLM	Motif ayant une position préférentielle
SMS	Score de l’écart maximal à la moyenne
SR	Hybridation spécifique
WR	Hybridation constitutive

Préambule

Au cours de ma thèse, j'ai présenté mon travail à des biologistes, à des bioinformaticiens et à des statisticiens. Chaque expérience m'a appris à adapter mon discours en fonction du public concerné. En écrivant ce document, j'ai donc essayé de le rendre accessible à ces trois catégories de personnes avec lesquelles je serai amenée à travailler par la suite. J'ai donc commencé ce manuscrit par une introduction relativement généraliste de biologie avant d'entrer plus en détail dans le projet de cette thèse.

* * *

Le processus permettant la synthèse d'une protéine à partir d'un gène est aujourd'hui connu mais son initiation l'est moins. De multiples partenaires protéiques sont impliqués dans cette initiation mais leur coopération fonctionnelle n'est pas toujours définie et de nouveaux éléments sont chaque année identifiés comme contribuant à cette première étape. Finalement, la synthèse d'une protéine est très contrôlée à chacune de ses étapes: c'est la régulation de l'expression des gènes. Elle permet la survie et le développement de l'organisme en régulant la synthèse de protéines en fonction des besoins. Aujourd'hui cette régulation est mieux connue et l'idée autrefois admise qu'un gène conduisait à une seule protéine n'est plus considérée. Un gène peut s'exprimer et être la matrice servant à synthétiser diverses protéines, tout comme il peut être transcrit sans être traduit ou encore ne pas s'exprimer et donc ne pas conduire à la synthèse d'une protéine. De plus, certains gènes ne s'expriment qu'en réponse à un stress, un manque de nutriment ou en fonction des conditions environnementales.

Néanmoins, deux questions majeures restent à élucider: «Quels sont les éléments permettant l'initiation de la synthèse des protéines ?» ainsi que «Quels processus sont mis en place pour contrôler la régulation de l'expression des gènes ?». Il existe un grand nombre de processus contribuant chacun à cette régulation. Lorsque tous seront compris et que toutes les régulations pourront être comprises conjointement, il sera alors possible de répondre à cette question : «Quels gènes s'expriment dans quels tissus, dans quels organes et dans quelles conditions, à quel moment ?».

Aujourd'hui, pour comprendre les processus impliqués dans la régulation de l'expression des gènes, chacun est étudié indépendamment des autres. La régulation de l'initiation de la transcription est un des processus majeur car il est la première étape de la synthèse d'une protéine. Les autres processus dépendent donc de lui. Différentes approches peuvent être utilisées pour mieux comprendre ce mécanisme majeur: *in vivo*, *in vitro* ou *in silico*. Les approches *in vivo* ou *in vitro* ont l'avantage de proposer des résultats expérimentalement validés. Néanmoins elles sont limitées par différentes contraintes et ne permettent pas une analyse exhaustive. Les analyses *in silico* permettent de prédire des données qui devront par la suite être validées (ou non) par une analyse *in vivo* ou *in vitro*.

Une étape intermédiaire est indispensable aux analyses *in silico*: la connaissance de la séquence du génome ou tout au moins des gènes considérés par l'étude. De plus, pour augmenter la pertinence, des annotations fonctionnelles et structurales peuvent être utilisées.

Une stratégie largement exploitée pour explorer un processus biologique est l'étude d'un organisme modèle, c'est-à-dire un organisme qui est plus simple à analyser. Par exemple, la plante *Arabidopsis thaliana* est un organisme modèle du règne des plantes qui est très étudié. Elle présente les avantages d'être de petite taille, de croissance rapide et son génome est un des plus petits du monde végétal aujourd'hui connus. Les prédictions proposées chez un organisme modèle pourront être transposées à d'autres organismes d'intérêt. Par exemple, les gènes mis en évidence chez *A. thaliana* faciliteront l'identification des gènes dans des génomes de plus grandes tailles et d'organisation plus complexe.

* * *

La thèse présentée dans ce document se place dans le cadre de l'analyse de la régulation de l'initiation de la transcription des gènes chez les eucaryotes. Les régions en amont des gènes d'*A. thaliana*, appelées les promoteurs, ont été étudiées afin d'identifier des éléments régulateurs c'est-à-dire de courtes séquences d'ADN susceptibles d'être impliquées dans la régulation de l'expression des gènes. Le travail réalisé a pris en considération l'ensemble des données disponibles, allant de la séquence du génome au transcriptome.

L'approche développée au cours de cette thèse exploite des compétences en bioinformatiques et en biostatistiques afin d'identifier automatiquement des éléments régulateurs putatifs puis d'analyser les caractéristiques fonctionnelles des gènes qui les contiennent.

Les résultats obtenus en utilisant cette approche sont de deux natures. Premièrement, une analyse globale des promoteurs propose une cartographie des promoteurs d'*A. thaliana*, une ressource comprenant une liste d'éléments régulateurs et leurs caractéristiques, ainsi qu'une liste des gènes les contenant. Cette première partie du travail laisse place à de nombreuses perspectives d'analyses qui seront en partie présentées dans ce document. Deuxièmement, des analyses spécifiques de groupes de gènes ont été réalisées. Ces sous-groupes peuvent être constitués par rapport à une architecture précise de leurs promoteurs, c'est-à-dire la présence ou l'absence d'éléments régulateurs. Les résultats obtenus proposent l'existence de nouveaux éléments régulateurs. Ces prédictions qui n'ont pas été proposées lors de l'analyse globale soulignent le fait que des études de sous-groupes de gènes sont complémentaires d'une analyse globale et peuvent ouvrir de nouvelles pistes d'investigation.

Introduction

Dans le cadre de cette thèse, je me suis intéressée à la régulation de l'expression des gènes chez un organisme eucaryote : la plante modèle *Arabidopsis thaliana*. L'introduction de ce travail est exclusivement tournée vers une présentation des connaissances relatives à la régulation de la transcription chez les eucaryotes. Dans une première partie sont introduites des notions de biologie moléculaire relatives à l'expression et la régulation des gènes codant des protéines. Les gènes à ARN n'ont pas été analysés dans ce travail. En deuxième partie, différentes approches possibles pour identifier les sites de fixation des facteurs de transcription ainsi que des bases de données les indexant seront présentées. Puis une troisième partie sera consacrée aux analyses bioinformatiques ayant pour objectif d'identifier et de caractériser des éléments régulateurs du promoteur central. Enfin, le projet de thèse et son contexte seront présentés.

1 Expression et régulation des gènes codant des protéines chez les eucaryotes

La régulation de l'expression des gènes est un ensemble de processus qui permet de moduler la production de molécules essentielles au développement de l'organisme. Différentes catégories de gènes existent, certains codant des protéines, d'autres étant des gènes à ARN par exemple. Dans la première partie de cette introduction, quelques notions sont introduites, à savoir ce qu'est l'expression d'un gène et comment cette expression est régulée.

1.1 Expression des gènes

Le patrimoine génétique est contenu dans le noyau des cellules eucaryotes, dans des structures appelées les chromosomes. L'ensemble des chromosomes d'une cellule constitue le génome d'un organisme. Chaque chromosome est formé d'une longue molécule d'Acide Désoxyribo-Nucléique (ADN). Les gènes des parties d'ADN répartis le long des chromosomes (Figure 1-1).

Un organisme eucaryote est constitué de une à plusieurs cellules ayant le même patrimoine génétique, et donc les mêmes gènes. Chaque gène codant une protéine va contribuer en fonction des besoins de l'organisme à la production des molécules qui seront impliquées dans le développement et la survie de l'organisme.

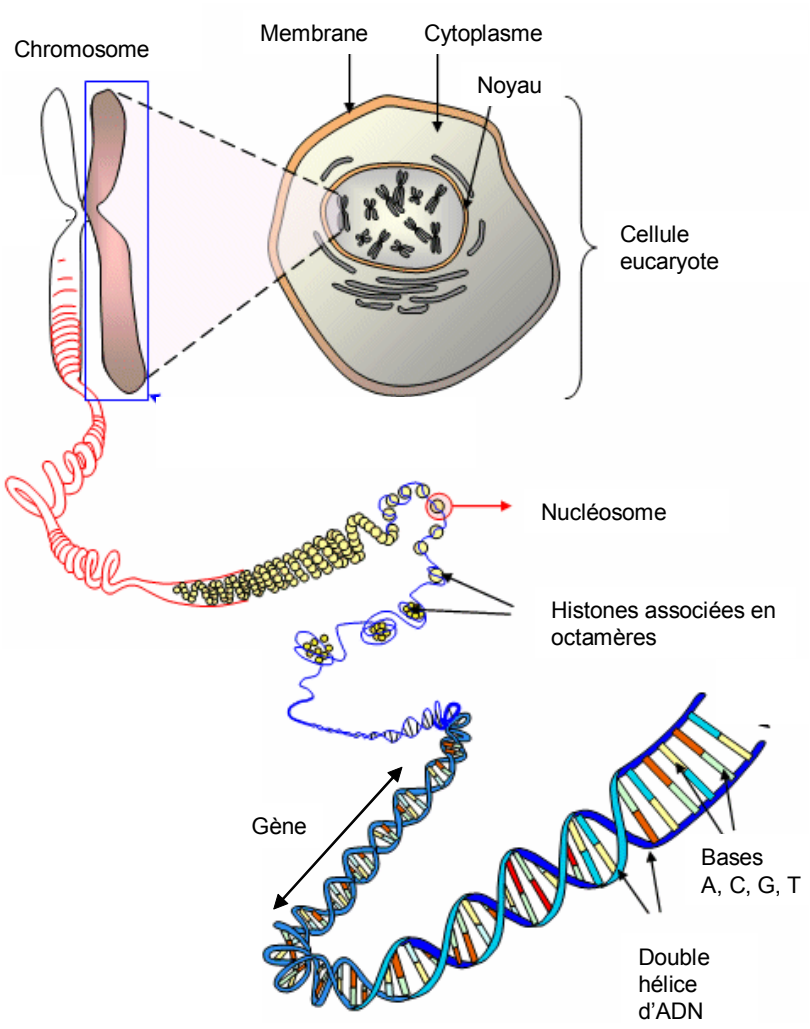


Figure 1-1 : De la cellule à l'ADN chez les eucaryotes.

Figure adaptée à partir d'une image du National Human Genome Research Institut.

1.1.1 Du gène à la protéine

La synthèse d'une protéine repose sur deux processus qui se succèdent et sont dépendants l'un de l'autre. Dans un premier temps une molécule d'Acide Ribo-Nucléique (ARN) est synthétisée à partir du gène. C'est la transcription. Dans un deuxième temps, une protéine est synthétisée à partir de l'ARN.

a) Transcription

Chez les eucaryotes, la transcription est un processus biologique qui se déroule dans le noyau des cellules. A partir de la séquence d'ADN d'un gène codant, des copies d'ARN sont synthétisées. Dans le cadre de la transcription de gènes codant des protéines, l'ARN est un ARN messager (ARNm). La Figure 1-2 illustre les différentes étapes de la transcription.

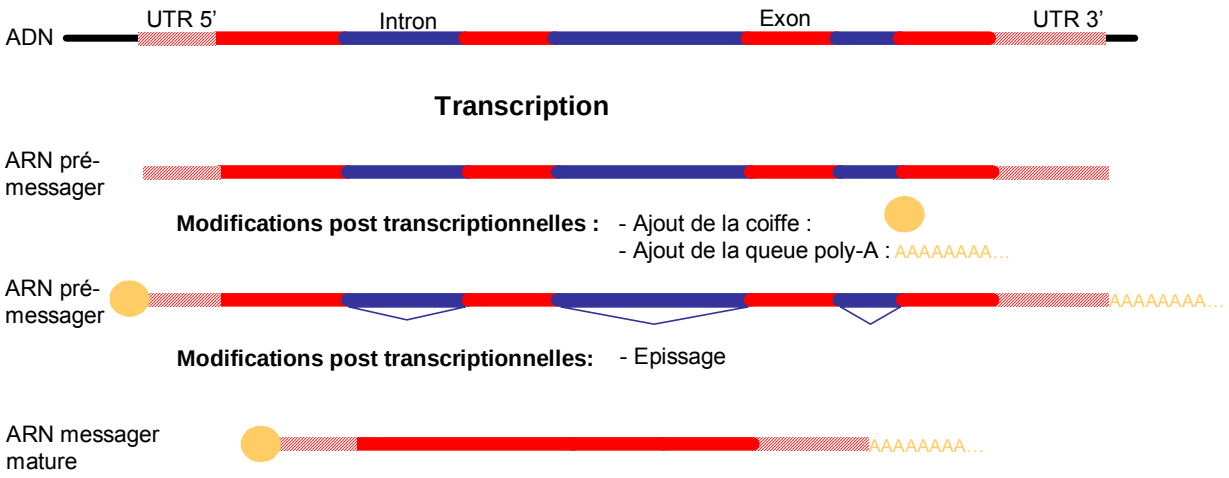


Figure 1-2 : Transcription et maturations post-transcriptionnelles.

i) Formation de l'ARN pré-messager

L'ARN polymérase est l'enzyme qui permet avec un ensemble d'autres protéines de catalyser ce processus biologique. L'ensemble de ces molécules forme le Complexe d'Initiation de la Transcription ou TIC («Transcription Initiation Complex»). Il existe trois catégories d'ARN polymérases dépendant des types d'ARN qu'elles synthétisent et donc des gènes qu'elles transcrivent. Dans le cadre de la transcription de gènes codant des protéines, une enzyme, l'ARN polymérase II (ARN pol II), est impliquée. Trois étapes successives sont nécessaires pour la transcription d'ADN en ARN pré-messager (ARNpm).

Premièrement, l'ARN pol II est recrutée par un complexe protéique et se fixe sur une région en amont de la séquence codante, le promoteur (Figure 1-3). C'est l'initiation de la transcription. Deuxièmement, l'enzyme progresse le long du gène en synthétisant au fur et à mesure une molécule d'ARNpm à partir du brin sens d'ADN, qui sert de matrice. L'ARNpm synthétisé est complémentaire de l'ADN. Cette étape est l'élongation. Troisièmement, la transcription se termine : l'ARNpm et l'ARN pol II sont libérés lorsque l'enzyme reconnaît un signal de fin de transcription sur la séquence d'ADN parcourue. L'ARNpm est clivé au niveau de ce site. C'est la terminaison.

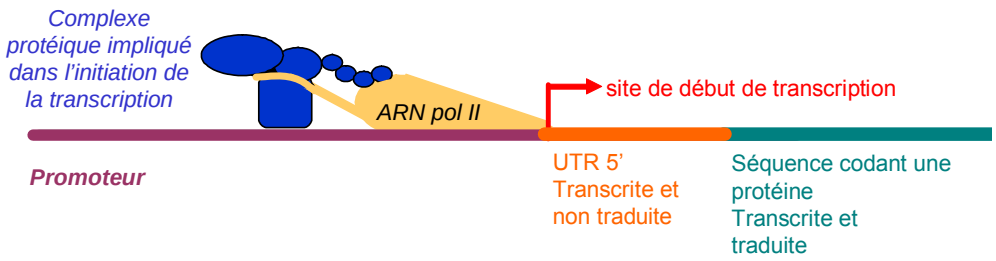


Figure 1-3 : Représentation schématique d'une ARN pol II et de son complexe protéique.

Chez les eucaryotes, l'ARNpm nécessite différentes maturations avant d'être fonctionnel et de pouvoir être traduit en protéines dans le cytoplasme.

ii) Maturation de l'ARNpm en ARN messager

Différentes catégories de maturations vont rendre le transcrit obtenu, l'ARNpm, fonctionnel ou vont modifier ses fonctions. Elles peuvent avoir lieu au cours de la transcription, comme les premières maturations, ou bien après, comme les deux dernières :

- Ajout de la coiffe méthyl-guanosine en extrémité 5' de la molécule d'ARNpm qui intervient dès le début de la transcription et permettra à l'ARNpm d'être reconnu par les ribosomes lors de la traduction ;
- Edition des ARN, c'est-à-dire l'ajout, la suppression ou la conversion de nucléotides transcrits à partir du brin d'ADN. Les conséquences sont de possibles changements d'acides aminés, avec l'apparition d'un codon stop ou d'un codon initiateur ou encore une modification des sites d'épissage. La désamination d'une cytosine en uracile de l'ARNpm de APOB provoque par exemple la synthèse de deux protéines différentes chez *Homo sapiens* (Powell *et al.*, 1987) : un des transcrit est non édité l'autre l'est, ce qui induit une terminaison plus précoce ;
- Ajout d'une queue poly-A en extrémité 3' de la molécule d'ARNpm lors de la terminaison de la transcription. Lorsque l'ARN a été clivé au niveau du site de polyadénylation, une polymérase intervient, la polyA Polymérase. Chez les eucaryotes, cette enzyme ajoute environ 200 adénines à l'extrémité de l'ARN synthétisé. Cet ajout d'une queue poly-A permet à l'ARNpm d'avoir un temps de demi-vie augmenté (Dreyfus and Regnier, 2002) ;
- Epissage. Les gènes codants des protéines sont constitués d'une succession d'introns, séquences non codantes, et d'exons, séquences codantes. Les ARNpm sont synthétisés à l'identique de l'ADN ; ils comportent donc ces séquences codantes et non codantes. L'épissage est une excision des introns, puis une jonction des exons restants. Néanmoins, l'épissage peut être alternatif. Lors de certains processus d'épissage, certains introns sont conservés, tandis que des exons peuvent être excisés. Cet épissage alternatif permet d'obtenir une grande variété d'ARN messagers. Ainsi, à partir d'un gène, différentes protéines peuvent être synthétisées. Chez les plantes, 20% des gènes sont soumis à l'épissage alternatif (Barbazuk *et al.*, 2008).

Suite à cette dernière étape de maturation, les ARNm obtenus migrent vers le cytoplasme par les pores nucléaires. Ils sont alors des ARNm matures et fonctionnels. Ils peuvent être traduits en protéines.

b) Traduction

La traduction est un processus biologique qui se déroule dans le cytoplasme des cellules (Sonenberg and Hinnebusch, 2009). Il permet à partir d'une molécule d'ARNm de synthétiser une protéine qui sera fonctionnelle après plusieurs modifications post-

traductionnelles. L'ensemble de ces modifications va principalement permettre de réguler l'activité de la protéine, de la marquer pour qu'elle soit reconnue par d'autres molécules, de l'adresser à son compartiment cellulaire. La Figure 1-4 illustre l'ensemble du processus, transcription et traduction, permettant la synthèse de multiples protéines. Ce processus ne sera pas plus détaillé dans le cadre de ce document qui présente un projet de thèse concernant l'initiation de la transcription.

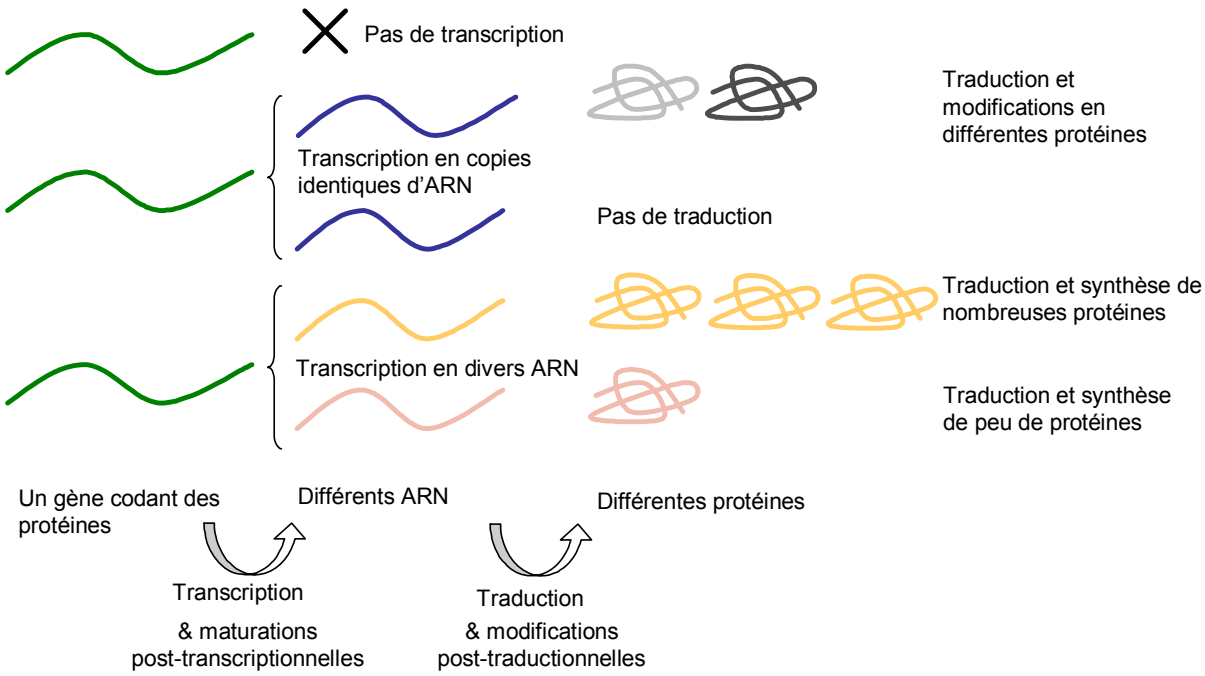


Figure 1-4 : Du gène à plusieurs protéines.

Chaque étape de la transcription et de la traduction subit des régulations. L'ensemble de ces régulations permet, à partir d'un gène, d'obtenir différents ARN qui chacun synthétisera diverses protéines.

1.1.2 Expression des gènes et le transcriptome

La différenciation cellulaire implique que chaque cellule d'un organisme multicellulaire acquiert une fonction spécifique. En conséquence, les gènes vont s'exprimer par rapport aux besoins et aux fonctions de la cellule. L'expression d'un gène dépend de l'emplacement de la cellule qui le contient dans l'organisme (organe, tissu), du stade de développement de l'organisme, de facteurs environnementaux comme un manque de nutriment par exemple… Notons que la présence d'ARNm dans une cellule n'est pas la preuve qu'une protéine correspondante a été synthétisée. Le transcriptome désigne l'ensemble des ARN synthétisés par un génome et présents à un instant donné.

a) Quelques techniques pour analyser le transcriptome

L'analyse du transcriptome a pour premier objectif d'identifier les gènes qui s'expriment dans un organisme, un organe ou un tissu, à un moment donné, dans une condition expérimentale donnée. Elle a également comme deuxième objectif, pour les gènes qui s'expriment, d'identifier leur intensité d'expression, c'est-à-dire l'équilibre entre la synthèse et

la dégradation d'ARN. L'intensité d'expression est proportionnelle à la quantité de transcrits dans une cellule à un instant t. Différentes techniques existent pour mettre en évidence la présence de transcrits dans les cellules, comme celles décrites ci-dessous présentées par ordre chronologique d'apparition.

Les étiquettes de transcrits des gènes ou EST («Expressed Sequence Tag») sont séquencées et indexées dans des banques. Elles représentent une extrémité d'environ 300 à 500 nucléotides du transcrit. Elles ont permis d'identifier plusieurs milliers de gènes chez *A. thaliana* (Cooke *et al.*, 1996; Asamizu *et al.*, 2000; White *et al.*, 2000) et sont également exploitées afin d'estimer l'expression relative des gènes qui est proportionnelle au nombre d'EST associées à un gène donné. Ne représentant que des portions d'ADN complémentaires (ADNc), leur quantité peut ne pas toujours être suffisante pour identifier des unités de transcription complètes : la qualité des analyses des EST dépend du nombre d'EST séquencées (Audic and Claverie, 1997). Actuellement, plus de 1.5 millions d'EST d'*A. thaliana* ont été séquencées et sont indéxées dans dbEST (Boguski *et al.*, 1993).

L'analyse en série de l'expression des gènes ou technique SAGE («Serial Analysis of Gene Expression») permet d'identifier et de quantifier des ARNm d'un échantillon (Velculescu *et al.*, 1995; Velculescu *et al.*, 1997). De petits fragments d'étiquettes de transcrits d'une dizaine de bases sont obtenus grâce à une digestion enzymatique des ARNm puis concaténés et séquencés sous forme d'un ADN synthétique qui contient un ensemble de fragments. Le nombre d'étiquettes séquencées reflète le niveau d'expression d'un gène. La technique SAGE a permis d'améliorer l'annotation du génome d'*A. thaliana* en identifiant de nouveaux transcrits et de nouveaux gènes (Robinson *et al.*, 2004; Robinson and Parkin, 2008).

De même, la technique du séquençage massif des signatures en parallèle ou MPSS («Massively Parallel Signature Sequencing») permet d'identifier et de quantifier des ARNm d'un échantillon. A la différence de la technique SAGE, elle est basée sur le séquençage en parallèle de milliers d'étiquettes d'ADNc générées à partir d'ARNm qui sont immobilisées sur des micro-billes (Brenner *et al.*, 2000). Cette technique a, elle aussi, contribué à l'identification de nouveaux transcrits chez *A. thaliana* (Nobuta *et al.*, 2007).

Les ADN complémentaires ou ADNc dits «pleine longueur» permettent l'obtention d'un ADNc théoriquement complet. Différentes technologies ont été développées (Clepet *et al.*, 2004; Seki *et al.*, 2004) et ont toutes contribué à une meilleure annotation des génomes et en particulier des régions promotrices. Ceci a été d'un grand intérêt pour l'analyse des séquences régulatrices dans les promoteurs (Castelli *et al.*, 2004; Molina and Grotewold, 2005; Alexandrov *et al.*, 2006). Comme pour chacune des techniques présentées, plus les conditions expérimentales d'obtention de transcrits sont diverses, plus elles permettent d'obtenir une image complète des transcrits d'un génome. Les ADNc pleine longueur aujourd'hui disponibles peuvent ne pas être entiers. Ceci est dû aux rétrotranscriptions qui font partie intégrante des technologies et qui sont souvent incomplètes. Néanmoins, par abus de langage, le terme d'ADNc pleine longueur est utilisé dans ce document.

Les développements récents du séquençage à haut débit donne une nouvelle dimension à ces approches de séquençage d'ARN en amplifiant la capacité à décrire l'inventaire des gènes transcrits et l'intensité de cette transcription (Morozova *et al.*, 2009).

b) Puces à ADN

La technologie des puces à ADN, développée en 1995 (Schena *et al.*, 1995), permet d'analyser un transcriptome à un instant précis en mesurant simultanément l'expression de l'ensemble des gènes dans un échantillon biologique donné. Cette technologie est une adaptation de la technique du Northern blot basée sur l'hybridation moléculaire entre des séquences d'ADN prédéfinies fixées sur une lame de verre et des séquences d'ADNc extraites d'un échantillon biologique.

Concrètement, une puce à ADN est un support solide, une lame de verre, sur lequel un ensemble de fragments d'ADN sont ancrés. Un fragment d'ADN est appelé ici une «sonde» et il représente un fragment de génome ou de gène. Un «spot» regroupe plusieurs sondes identiques à une position définie sur la puce. Plusieurs milliers de spots sont présents sur une puce à ADN (Figure 1-5).

Au cours du temps, la technologie des puces à ADN s'est déclinée sous différentes formes (micro-array, puces à oligonucléotides) et différentes applications. Le principe de l'ensemble des puces à ADN reste le même. Il repose sur la propriété de l'ADN simple brin à s'hybrider avec sa séquence complémentaire : deux brins d'ADN complémentaires formeront spontanément un duplex. Une puce à ADN est donc une surface constituée de sondes représentant chacune une région du génome, un gène par exemple, afin qu'un ADNc marqué déposé sur une puce puisse s'y hybrider et mettre en évidence l'expression du gène qu'il représente. Les étapes principales lors d'expériences de puces à ADN sont schématisées Figure 1-5 et listées ici :

1. Extraction des ARN des cellules ;
2. Amplification et transformation des ARN en ADNc ;
3. Marquage des ADNc (fluorescence / radioactivité) ;
4. Dépôt des ADNc marqués sur la puce et hybridation ;
5. Lecture de l'ensemble des spots ;
6. Analyse des données.

Lorsqu'une molécule d'ADNc marquée est hybridée à sa sonde, elle est ensuite mise en évidence par des procédés optiques ou radioactifs en fonction de la technique de marquage de l'ADNc. Les analyses d'images de puces à ADN ainsi que des analyses statistiques permettent par la suite d'identifier les sondes qui ont été hybridées et donc les transcrits qui sont présents dans l'échantillon biologique analysé. Ainsi, les sondes propres à un gène peuvent révéler si le gène est exprimé ou non. L'intensité des signaux mesurés, proportionnelle au taux d'ADNc hybridés aux sondes, permet de déterminer la quantité relative de transcrits correspondant à un gène exprimé.

Notons que certaines puces peuvent supporter deux échantillons simultanément (Schena *et al.*, 1995; DeRisi *et al.*, 1997; Crowe *et al.*, 2003). Elles permettent de marquer deux échantillons avec deux fluorochromes différents et donc d'étudier l'expression différentielle de ces deux échantillons. Ce sont des puces dites bicolores.

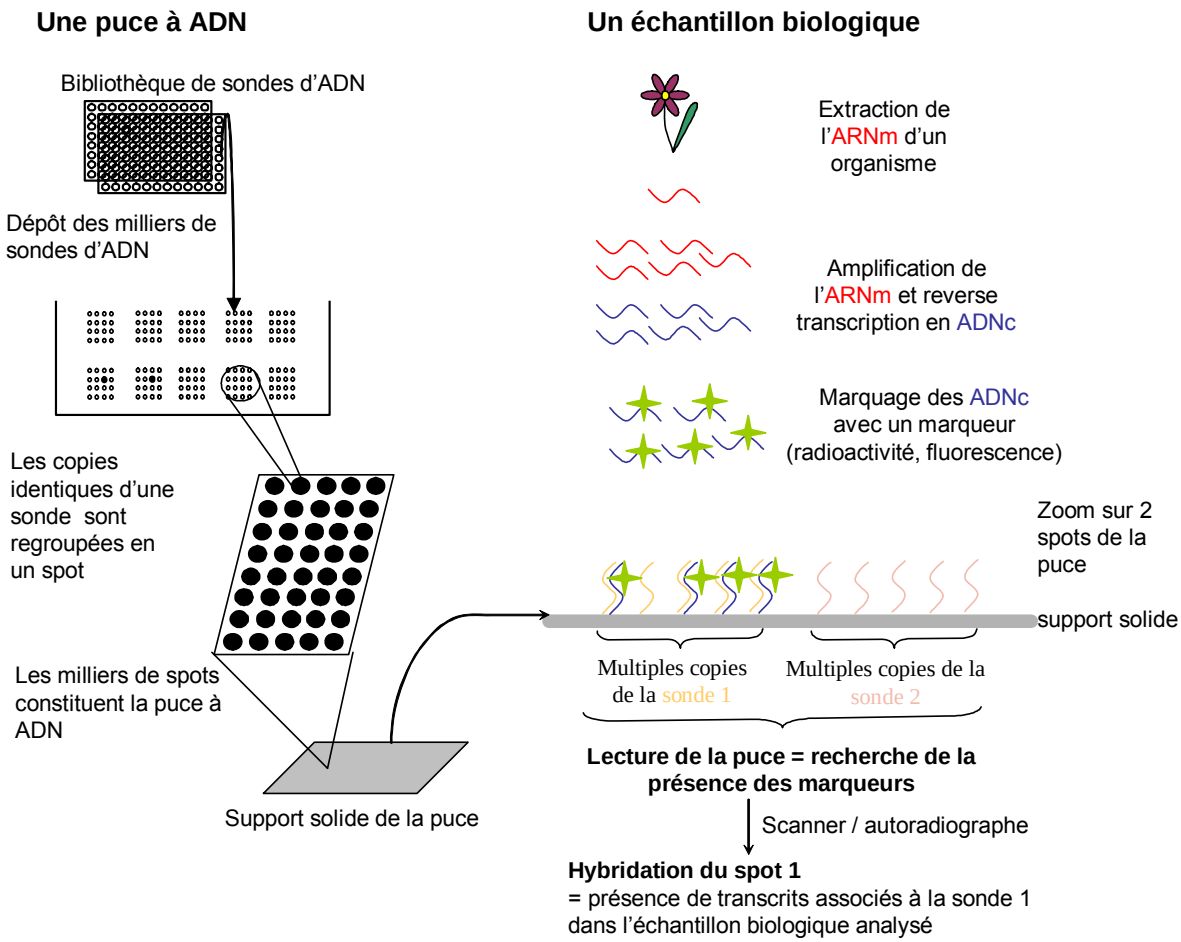


Figure 1-5 : Les puces à ADN : principe.

Certaines utilisations de la technologie des puces à ADN nécessitent de définir et donc de connaître les gènes d'un organisme pour constituer les sondes. Ainsi, les puces à ADN qui sont dédiées à l'analyse globale de l'expression des gènes ne permettent pas une étude exhaustive de l'ensemble des gènes pour les génomes qui ont une annotation partielle. Différentes puces sont donc seulement constituées d'une partie des gènes d'un génome. De plus si un gène a une séquence qui est très similaire à celle d'un autre gène de l'organisme étudié, il pourra être difficile voire impossible de construire deux sondes suffisamment différentes pour les distinguer efficacement.

Les données générées par SAGE, MPSS et puces à ADN ne sont pas directement comparables et sont complémentaires les unes des autres. Chacune est susceptible de contribuer à l'annotation du génome (Coughlan *et al.*, 2004).

Ainsi, tandis que le génome d'un organisme ne peut varier que peu au cours des stades de développement d'un organisme, le transcriptome et le protéome, c'est-à-dire l'ensemble des protéines synthétisées dans une cellule, varient à chaque instant. En effet, l'expression des gènes est un processus dynamique finement régulé par des mécanismes cellulaires qui

peuvent intervenir à chaque étape du processus. Tous les gènes sont régulés. Un gène peut être ou non accessible à l'ARN pol II ; puis toutes les étapes de la transcription, de la maturation des ARNpm, de la traduction et des modifications des protéines sont sujettes à des régulations. Le projet présenté dans ce document est centré sur une étape primordiale de la régulation de l'expression des gènes : l'initiation de la transcription.

1.2 Régulation de l'initiation de la transcription

L'initiation de la transcription est le premier stade de la synthèse d'une protéine. Les étapes suivantes dépendent donc d'elle. Elle est régulée à l'échelle de certaines modifications épigénétiques, puis à celle des régulations induites par les promoteurs et les partenaires fonctionnels qu'ils contiennent.

1.2.1 Régulations épigénétiques

Les régulations épigénétiques correspondent aux modifications héréditaires de l'expression des gènes qui n'impliquent aucun changement de la séquence d'ADN elle-même. Une régulation épigénétique peut rendre des gènes inactifs dès la transcription.

L'initiation de la transcription d'un gène n'est possible que lorsque le gène est accessible, c'est-à-dire que l'ARN pol II peut être recrutée dans la région en amont de la séquence codante. L'accessibilité de l'ADN est donc capitale pour permettre la synthèse d'une protéine. La chromatine est la forme sous laquelle la molécule d'ADN est pelotonnée dans le noyau. Elle peut être sous forme condensée dite hétérochromatine ou sous forme décondensée dite euchromatine. Ces deux formes résultent de l'enroulement de l'ADN autour de protéines, les histones, qui sont organisées en octamères et forment un complexe appelé le nucléosome. Les nucléosomes maintiennent les molécules d'ADN dans le noyau sous forme compactée. De plus, ils jouent un rôle dans l'accessibilité de l'ADN. Ce n'est que sous forme d'euchromatine que l'ARN pol II peut s'arrimer en amont des séquences codantes. Néanmoins, ces structures peuvent être déplacées le long de l'ADN sous l'action de complexes de remodelage (Segal *et al.*, 2006).

Dans le noyau des cellules, la chromatine est majoritairement sous forme d'hétérochromatine. Des modifications de la condensation de la chromatine sont requises pour permettre à l'ARN pol II d'accéder à l'ADN. Ce degré de la chromatine est réglé par des modifications des extrémités N-terminales des histones, comme des acétylations ou des méthylations. L'hétérochromatine est en effet hypo-acétylée et hyper-méthylée comparée à l'euchromatine (Peng and Karpen, 2008). La méthylation d'une histone aura un effet condensateur et l'acétylation un effet décondensateur sur la structure de la chromatine.

En conclusion, différents processus épigénétiques contribuent à la régulation de l'expression des gènes. Son importance quantitative est difficile à estimer car cette régulation est aujourd'hui moins connue que les autres régulations plus directement liées à la séquence génomique.

1.2.2 Promoteurs

Le promoteur est la région en amont des séquences codantes qui contribue à l'initiation de la transcription et à sa régulation. Il interagit avec de multiples partenaires, comme l'ARN pol II entre autres, pour permettre au gène qu'il régule d'être transcrit. Dans cette partie, les différentes régions des promoteurs ainsi que les protéines et facteurs impliqués dans la régulation de l'expression des gènes sont introduits.

a) Facteurs de transcription et leurs sites de fixation

Les facteurs de transcription et leurs sites de fixation sont appelés «Transcription Factor» TF et «Transcription Factor Binding Site» TFBS. Ils sont conjointement impliqués dans l'initiation de l'expression des gènes et / ou dans sa régulation. Les TFBS sont de courtes séquences d'ADN sur lesquelles peuvent se fixer des TF. Un TF est une protéine constituée d'un domaine de fixation à l'ADN ou à un autre TF et d'un domaine de régulation de la transcription. Il existe deux catégories de TF : les facteurs de transcription généraux, ou GTF («General Transcription Factor»), et les facteurs spécifiques de transcription. Ils interviennent à deux moments différents de la transcription.

i) Facteurs de transcription

Lorsqu'un gène va être transcrit, différents GTF se fixent à leur site de fixation situé dans le promoteur en amont de la séquence codante, et forment le Complexe d'Initiation de la Transcription ou TIC. Ce complexe recrute l'ARN pol II qui se fixe dans la région en amont du site de début de la transcription ou TSS («Transcription Start Site»). Ces GTF sont nécessaires à l'initiation de la transcription. Lorsque le complexe d'initiation de la transcription est formé et que l'ARN pol II est arrimée, des facteurs de transcription spécifiques vont participer à la modulation de la transcription du gène. Fixés à leur TFBS spécifique, ils pourront inhiber ou activer la transcription du gène qu'ils régulent.

Environ 45% des TF connus sont spécifiquement observés chez les plantes (Riechmann *et al.*, 2000) d'où une distinction entre les TF de plantes et ceux de mammifères ou d'autres organismes. Différentes banques de données sont dédiées à l'indexation des TF d'*A. thaliana* comme DATF pour «Database of *Arabidopsis* Transcription Factors» (Guo *et al.*, 2005) ou AtTFDB pour «*Arabidopsis* transcription factor database» (Davuluri *et al.*, 2003). De plus des banques de données sont aussi dédiées au règne végétal comme PlantTFDB pour «Plant Transcription Factor Databases» (Guo *et al.*, 2008). Au total, ces ressources indexent aujourd'hui 2182 gènes codant des facteurs de transcription chez *A. thaliana*. La base de données FLAGdb^{++} (Samson *et al.*, 2004) donne un accès aisé à ces gènes classés en 75 familles (travail réalisé par Magalie Leveugle, URGV).

ii) Sites de fixation des facteurs de transcription

Les sites de fixation de ces TF, les TFBS, sont des séquences d'ADN d'environ 5 à 15 bases de long qui sont reconnues par les TF. Ils sont observés dans les séquences promotrices, mais aussi dans les introns (Vyas *et al.*, 1992; Sieburth and Meyerowitz, 1997; Deyholos and Sieburth, 2000) et dans les UTR (Larkin *et al.*, 1993; van Helden *et al.*, 2000; Graber *et al.*, 2002; Hulzink *et al.*, 2003). Les séquences des TFBS étant courtes, elles peuvent être observées à de nombreux emplacements du génome. Néanmoins, elles ne sont fonctionnelles, c'est-à-dire reconnues par leur facteur de transcription spécifique, que

dans des conditions précises, à un emplacement particulier par exemple. Ces séquences sont alors appelées éléments régulateurs.

Les TFBS sont des séquences souvent dégénérées, c'est-à-dire admettant à certaines positions une variabilité de 2 à 4 bases parmi A, C, G et T (Annexe I). Cette variabilité au sein des TFBS peut être expliquée par deux principales raisons. (i) La fixation chimique du domaine de fixation à l'ADN d'un TF sur son TFBS spécifique ne nécessite pas la reconnaissance de l'ensemble des bases du TFBS, certaines étant peu ou pas impliquées. La reconnaissance entre les TF et l'ADN ne suit pas des règles simples comme c'est le cas de la reconnaissance ADN / ADN par complémentarité (Bareket-Samish *et al.*, 2000; Svozil *et al.*, 2008). (ii) De plus, certains sites de fixation peuvent avoir différentes fonctions régulatrices et être reconnus par différents TF. L'affinité des facteurs de transcription avec leur site de fixation n'est donc pas toujours la même et une variabilité de la fixation existe (Muller, 2001).

Les sites de fixation, pour être fonctionnels, doivent être accessibles pour que les facteurs de transcription puissent s'y fixer. Comme nous avons vu lors de la présentation de la régulation épigénétique, l'ADN est souvent condensé et donc inaccessible (Whitehouse and Tsukiyama, 2009). Le rôle des nucléosomes dans l'architecture du promoteur a fait l'objet de nombreux travaux chez *S. cerevisiae*, chez *D. melanogaster* ainsi que d'autres métazoaires (Leach *et al.*, 2001; Boyle *et al.*, 2008; Hartley and Madhani, 2009). Diverses estimations de l'importance des propriétés et de la séquence de l'ADN dans l'occupation des régions promotrices par les nucléosomes ont été proposées (Kiyama and Trifonov, 2002; Thastrom *et al.*, 2004; Segal *et al.*, 2006; Henikoff, 2008; Miele *et al.*, 2008; Whitehouse and Tsukiyama, 2009). Une analyse à l'échelle génomique a montré que le positionnement précis de certains TFBS au sein des promoteurs leur permet d'être accessibles à la surface des nucléosomes et d'être reconnus par les facteurs de transcription (Ioshikhes *et al.*, 1999).

Des contraintes topologiques peuvent imposer à des TFBS d'être (i) à une distance précise d'autres TFBS, pour former un CRM («Cis Regulatory Module») c'est-à-dire un complexe de TFBS (Yuh *et al.*, 2001; Blanchette *et al.*, 2006) ou (ii) à une distance précise du TSS pour permettre d'initier la transcription (Grace *et al.*, 2004). Par exemple, la boîte TATA est observée à de multiples endroits du génome, mais elle n'est fonctionnelle qu'à une position précise : environ 30 bases en amont du TSS chez les métazoaires (Patikoglou *et al.*, 1999). De plus, une distance préférentielle entre la boîte TATA et d'autres TFBS a été identifiée biologiquement (Dion and Coulombe, 2003; Tabach *et al.*, 2007). A ce jour, aucune ressource ne propose de cartographie des TFBS : la distinction entre les occurrences fonctionnelles et le bruit de fond n'est pas suffisante actuellement pour permettre de les représenter (Blanchette and Sinha, 2001; Loganantharaj, 2006).

Il est à noter que tout comme les TF peuvent être généraux ou spécifiques, il existe des TFBS qui peuvent être qualifiés de généraux et d'autres de spécifiques. Certains sont en effet impliqués dans la formation du TIC (Burke and Kadonaga, 1997; Lim *et al.*, 2004; Juven-Gershon *et al.*, 2008) et d'autres dans des régulations spécifiques, suite à un stress biotique ou abiotique ou dans un tissu particulier par exemple (Rushton *et al.*, 1996; Laloi *et al.*, 2004).

Ainsi, conjointement, les sites de fixation et leurs TF contrôlent l'expression des gènes. Ils vont contrôler l'initiation de la transcription, et activer ou inhiber l'expression des gènes

qu'ils régulent dans un système complexe pouvant comprendre plusieurs cascades de régulation et qui est régulé à différents niveaux hiérarchiques (van Driel *et al.*, 2003).

b) Promoteur central, proximal et distal

Différentes régions fonctionnelles des promoteurs ont été décrites en fonction des TFBS qu'elles contiennent. Elles se distinguent également les unes des autres en fonction de leur distance par rapport au TSS (Figure 1-6). Il est à noter que ces régions ne sont pas fonctionnellement indépendantes. Par exemple, le projet ENCODE (Consortium, 2004) a montré qu'en moyenne les séquences entre -300 et -50 par rapport au TSS contribuent positivement à l'activité du promoteur central (Cooper *et al.*, 2006).

Le promoteur central est la région la plus proche du TSS et débute environ 50 bases en amont de ce site (Novina and Roy, 1996; Smale, 2001). C'est dans cette région que le TIC va se former, *via* l'intervention des GTF et de leurs TFBS spécifiques. Le promoteur central est décrit comme étant la région minimale indispensable à l'initiation de la transcription.

Le promoteur proximal est une région qui s'étend sur quelques centaines de bases en amont du TSS. Cette région contient principalement des TFBS qui sont reconnus par des facteurs spécifiques de transcription. Elle est impliquée dans la régulation spécifique de l'expression de gènes précis.

Le promoteur distal est celui le plus éloigné du TSS. Il peut s'étendre sur plusieurs milliers de bases en amont de la séquence codante (Barton *et al.*, 1997). Dans ces régions, les TFBS sont reconnus par des facteurs spécifiques de transcription. Dans le promoteur distal, des repliements de l'ADN permettent de rapprocher deux régions distantes et donc deux TFBS qui peuvent interagir fonctionnellement.

Au regard des connaissances actuelles, les TFBS situés dans le promoteur proximal et dans le promoteur central sont les plus nombreux (Higo *et al.*, 1999; Davuluri *et al.*, 2003; Bryne *et al.*, 2008). Néanmoins, comme précisé précédemment, les TFBS peuvent être observés dans les introns, dans les UTR ou à de grandes distances d'un gène, parfois plusieurs gènes en amont ou en aval d'un gène qu'ils régulent.

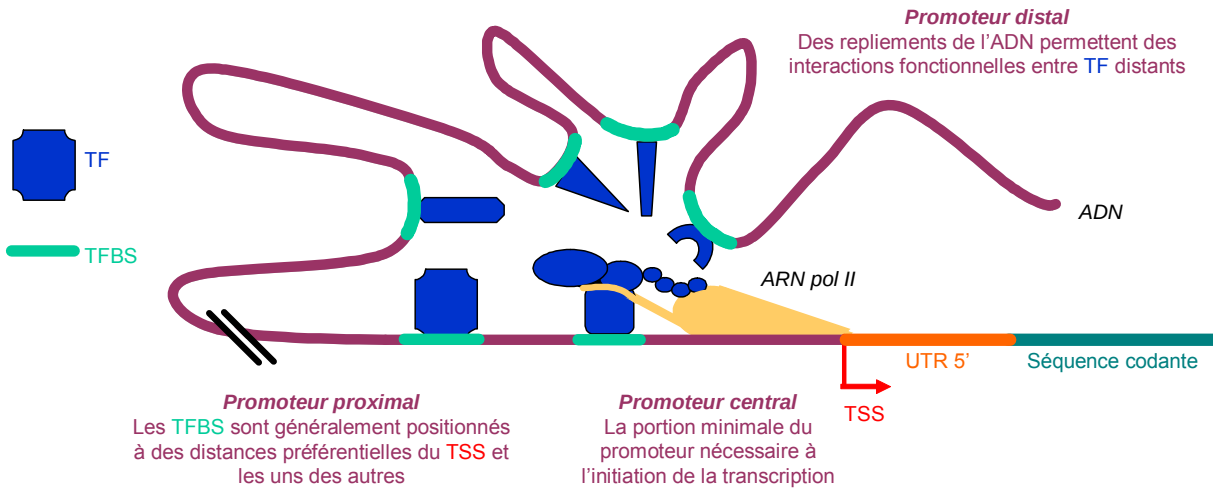


Figure 1-6 : Promoteur d'un gène codant des protéines chez les eucaryotes.

c) Modularité des sites de fixation des facteurs de transcription

L'interaction entre les facteurs de transcription et leurs sites de fixation cibles permet de réguler l'expression des gènes. Néanmoins, cette régulation est plus complexe, car elle concerne généralement plusieurs éléments régulateurs qui interagissent fonctionnellement (Chattopadhyay *et al.*, 1998; Blazquez and Weigel, 2000) comme schématisé Figure 1-6. Les TFBS des eucaryotes supérieurs agissent en modules ou CRM. Au sein d'un CRM il est fréquent que plusieurs facteurs de transcription se fixent sur différentes occurrences d'un même élément régulateur (Berman *et al.*, 2002). Une base de données comme TRANScompel souligne le rôle clef des interactions entre différents TFBS et TF (Kel-Margoulis *et al.*, 2002). La distance entre deux TFBS est un paramètre conservé entre *Mus musculus* et *H. sapiens* (Lu *et al.*, 2008).

Comparé à des prédiction de modules de motifs, il a été suggéré que la recherche de motifs seuls conduit à moins de TFBS fonctionnels chez les mammifères (Cawley *et al.*, 2004; Wasserman and Sandelin, 2004). Par exemple, chez *H. sapiens*, des études expérimentales récentes réalisées à l'échelle génomique ont démontré que le facteur de transcription E2F1 est recruté par deux sites de fixation (Rabinovich *et al.*, 2008). Chez les eucaryotes supérieurs, les approches *in silico* pour prédire l'existence de TFBS prennent le critère de modularité des TFBS en considération pour identifier non plus un site de fixation unique, mais des ensembles de TFBS présents sur un même promoteur et susceptibles de réguler conjointement l'expression d'un gène (Blanchette *et al.*, 2006; Ferretti *et al.*, 2007).

d) Définition de la position du site d'initiation de la transcription

La fin de la région promotrice est délimitée par le TSS (Figure 1-6). L'obtention de la position de ce site est capitale au regard d'une approche prenant en compte l'information positionnelle car la distance préférentielle des TFBS n'a été observée que vis à vis du TSS. Cependant la définition de la position du TSS est difficile.

Différents critères aident à la prédiction *in silico* de régions susceptibles de contenir un à plusieurs TSS. Certaines études proposent d'exploiter les biais compositionnels en bases des promoteurs à l'approche du TSS (Shi and Zhou, 2006; Yang *et al.*, 2007) ou encore la présence de TFBS (Rombauts *et al.*, 2003; Yamamoto *et al.*, 2007) pour positionner le TSS. De plus, des caractéristiques structurales comme les courbures intrinsèques de l'ADN et la préférence dans le positionnement des nucléosomes peuvent permettre de prédire une région promotrice (Barrera and Ren, 2006; Heintzman *et al.*, 2007; Akan and Deloukas, 2008; Boyle *et al.*, 2008; Henikoff, 2008). Néanmoins, une prédiction *in silico* de la position des TSS des gènes n'est pas actuellement possible (Rombauts *et al.*, 2003) même si des approches sont de plus en plus performantes (Abeel *et al.*, 2008; Abeel *et al.*, 2008) et permettent de définir des régions transcriptionnellement actives.

Aujourd'hui, une méthode à haut débit, pour définir avec la plus grande exactitude la position du TSS, est le séquençage des transcrits d'un organisme et leur alignement avec la séquence du génome de l'organisme considéré. L'obtention de transcrits pleine longueur est possible (Clepet *et al.*, 2004; Seki *et al.*, 2004) et a permis récemment, lorsque les transcrits ont été suffisamment nombreux, d'avoir une vision des TSS à l'échelle génomique chez *H. sapiens* et *A. thaliana* (Carninci, 2006; Tanaka *et al.*, 2009; Yamamoto *et al.*, 2009).

Pour conclure cette partie, la régulation de l'expression des gènes est un processus complexe qui fait intervenir un très grand nombre de partenaires qui interagissent fonctionnellement. Ses grands principes sont connus, mais des découvertes restent à faire afin de mieux comprendre ce processus et son impact sur l'expression des gènes. Les analyses réalisées dans le cadre de cette thèse avaient pour objectif de contribuer à cette connaissance.

2 Identification et indexation des éléments régulateurs de l'expression des gènes

Dans ce chapitre, les approches *in vitro* et *in silico* qui ont été développées pour identifier les TFBS sont introduites, suivies de différentes ressources d'informations les concernant. Enfin, quelques-unes des banques de données qui indexent les TFBS sont présentées, en axant sur les ressources dédiées aux plantes, et plus particulièrement à *A. thaliana*, l'organisme modèle étudié dans le cadre de cette thèse.

2.1 Approches pour identifier les éléments régulateurs

Deux grandes catégories d'approches pour identifier les éléments régulateurs sont aujourd'hui exploitées : les approches *in vitro* ou *in vivo*, qui permettent d'identifier et de valider expérimentalement les TFBS propres à un promoteur, et les approches *in silico*, qui permettent de prédire l'existence d'éléments régulateurs à petite ou à grande échelle, en considérant un gène ou le génome.

2.1.1 Approches de biologie expérimentale

Seules les expériences en biologie expérimentale, surtout celles *in vivo*, peuvent réellement permettre de caractériser un élément régulateur fonctionnel. Les approches *in silico* permettent de prédire des candidats qui pourront ou non être validés *in vitro* ou *in vivo*. Sans être exhaustives, différentes expériences en biologie expérimentale pour identifier les TFBS peuvent être mentionnées.

a) Retard sur gel

Le retard sur gel est une technique mise au point en 1989 et qui permet de mettre en évidence les interactions protéine / acide nucléique. Cette technique est plus connue sous le nom d'EMSA, pour «*electrophoretic mobility shift assay*» (Garner and Revzin, 1981).

Le principe de cette approche est qu'un complexe protéine / acide nucléique est plus lent à migrer sur un gel d'électrophorèse que l'acide nucléique seul. Lors d'une analyse de retard sur gel, l'échantillon témoin est l'acide nucléique qui migre seul. L'autre est un échantillon d'acide nucléique test avec une protéine susceptible de s'y fixer. Si c'est le cas, le complexe sera plus lent à migrer sur le gel d'électrophorèse que l'acide nucléique seul et sera dit «retardé».

Dans le cadre de l'étude des éléments régulateurs, cette technique permet de mettre en évidence la fixation d'une protéine, un facteur de transcription, sur un acide nucléique, son site de fixation spécifique. Très sensible, cette technique permet de comparer des vitesses de migration, et d'étudier l'affinité d'un complexe formé par un TF et son TFBS.

b) Empreinte à la DNase I

La technique de l'empreinte à la DNAse I mise au point en 1978 permet de mettre en évidence l'emplacement d'une séquence d'ADN sur laquelle une protéine d'intérêt s'est fixée. Cette technique est également connue sous le nom de footprinting à la DNAse (Galas and Schmitz, 1978).

Le principe de cette méthode est qu'une séquence d'ADN sur laquelle est fixée une protéine ne peut pas être digérée par une enzyme. Cette technique exploite la propriété des DNases de type I qui catalysent la dégradation de l'ADN en nucléotides. Un échantillon contrôle contient de l'ADN et de la DNase I : il est digéré intégralement. L'échantillon test contient l'ADN mis en contact avec une protéine puis de la DNase I. S'il y a eu fixation, la protéine protège de la dégradation et laisse son «empreinte». La comparaison de la digestion des échantillons permet de connaître les sites rendus inaccessibles par la fixation de la protéine.

Dans le cadre de la recherche de TFBS, cette technique peut être utilisée lorsqu'un TF est identifié mais que sa cible n'est pas connue. Cette technique présente l'avantage de mettre en évidence des fixations quelle que soit l'affinité protéine / ADN.

c) Immunoprécipitation de la chromatine

L'immunoprécipitation de la chromatine ou ChIP pour «Chromatine ImmunoPrecipitation» est une technique mise au point en 2000 pour mettre en évidence *in vivo* les sites de fixation de protéines d'intérêt comme les TF par exemple (Orlando, 2000).

Cette technique permet d'isoler des fragments d'ADN qui ont interagit avec une protéine. Elle se déroule en quatre étapes : formation *in vivo* d'une liaison covalente TF / TFBS par action du formaldéhyde ; extraction et découpage de l'ADN en petits brins par sonication ; immuno-précipitation de la protéine d'intérêt par un anti-corps spécifique ; séparation du complexe protéine / ADN par la protéinase K pour ne conserver que l'ADN. Des fragments d'ADN qui ont interagi *in vivo* avec la protéine d'intérêt sont identifiés, un TF dont la cible n'est pas connue par exemple. Ces fragments d'ADN peuvent être séquencés.

Cette technique présente l'intérêt de mettre en évidence des interactions *in vivo*, tandis que les deux méthodes précédentes mettent en évidence des interactions *in vitro*. Cependant l'immunoprécipitation de la chromatine ne peut être appliquée que lorsqu'un anticorps hautement spécifique dirigé contre un TF d'intérêt est disponible.

d) Immunoprécipitation de la chromatine sur puce

L'Immunoprécipitation de la chromatine sur puce, plus connue sous la dénomination de ChIP on chip ou ChIP-chip combine la technologie des puces à ADN (page 22) à la technique ChIP décrite ci-dessus. Développée en 2000, elle permet de localiser à l'échelle génomique l'ensemble des séquences reconnues par un TF d'intérêt (Ren *et al.*, 2000).

Le ChIP-chip est une approche qui peut être résumée en trois étapes. Dans un premier temps, une ChIP est réalisée afin d'extraire les fragments d'ADN qui ont interagi avec le TF étudié. Puis, les fragments sont marqués et déposés sur une puce à ADN ou «chip» afin qu'ils s'hybrident avec leurs séquences complémentaires (Figure 1-5). Enfin, l'analyse des résultats de la puce permet de positionner sur le génome les fragments d'ADN qui ont interagi avec la protéine d'intérêt. Une expérience de ChIP-chip ne peut pas être appliquée sans la disponibilité d'un anticorps spécifiquement dirigé contre un TF d'intérêt.

L'analyse systématique des promoteurs par puce ou projet SAP («Systematic Analysis of Promoters») coordonné par Pierre Hilson (VIB, Ghent) a permis de produire une puce dédiée à l'analyse des promoteurs d'*A. thaliana*. Les puces SAP peuvent être exploitées en

ChIP-chip. La version SAPv2 représente 22260 promoteurs. Les fragments d'ADNc d'une taille allant jusqu'à 2kb sont déposés sur la puce et doivent répondre à différents critères : (i) la spécificité, c'est-à-dire être une région unique du génome de la plante modèle et (ii) la longueur qui dit être suffisante pour contenir les TFBS présents au niveau de la région régulatrice et permettre une analyse efficace. Cette puce a été utilisée pour l'analyse des interactions entre l'histone acétyltransférase GCN5 et l'ADN (Benhamed *et al.*, 2008). Elle a permis de préciser le rôle majeur que pourrait jouer GCN5 dans l'initiation de la transcription des gènes.

A ce jour, une application majeure de la technique ChIP-chip est la technologie des puces à ADN de type tiling-array qui couvrent l'ensemble d'un génome, c'est-à-dire à la fois régions codantes et non codantes. L'utilisation de ces puces à ADN permet d'étudier les phénomènes de transcription en ayant la possibilité de s'affranchir de l'annotation structurale qui peut être biaisée. Par exemple, le génome d'*A. thaliana*, séquencé il y a 9 ans (AGI, 2000), voit son annotation fonctionnelle évoluer constamment. La dernière mise à jour du génome de la plante modèle disponible sur le site Web «The *Arabidopsis* Information Resource» (TAIR) le 19 juin dernier fait état de 27379 gènes codant des protéines (Swarbreck *et al.*, 2008). Ainsi, depuis la mise à jour précédente datant d'avril 2008, 739 nouveaux gènes ont été identifiés, et 1254 structures de gènes ont été actualisées (TAIR9 Genome Release). L'annotation des génomes évolue constamment, chaque nouvelle mise à jour apportant des corrections. A l'inverse des tiling-arrays, les puces à ADN classiques et les puces SAP dépendent de ces évolutions. Les tiling-arrays présentent l'intérêt de pouvoir analyser de manière exhaustive les transcrits d'un génome. L'étude du transcriptome d'*A. thaliana* par tilling-array en 2003 a proposé l'existence de 2397 nouveaux gènes dans le génome en plus des gènes prédits par les outils traditionnels de prédiction (Yamada *et al.*, 2003). Ces prédictions concernent de nouveaux transcrits fonctionnels, mais aussi du bruit de fond de transcription attendu dans les génomes eucaryotes (Struhl, 2007).

2.1.2 Approches *in silico*

Depuis le séquençage du premier génome complet, un phage, (Sanger *et al.*, 1977), plus de cent génomes d'eucaryotes ont été séquencés. Les informations structurales et fonctionnelles en découlant ont été rapidement complétées par de grandes quantités de transcrits avec, des banques d'EST dans un premier temps, puis des technologies à haut débit. Ces données ont été souvent exploitées pour la recherche des TFBS *in silico* qui demeure un challenge de longue date pour la bioinformatique (Tompa *et al.*, 2005). Les recherches sont basées sur différentes hypothèses et exploitent la séquence, les transcrits et / ou les prédictions de la position du TSS afin de prédire l'existence de TFBS, ce qui a conduit au développement de différents algorithmes. Sans être exhaustive, cette partie présente les hypothèses de travail exploitées dans chaque grande catégorie d'approches *de novo*. D'autres approches peuvent exploiter les séquences consensus et les caractéristiques connues de TFBS validés biologiquement, comme leurs contraintes topologiques.

a) Recherche de motifs sur-représentés

Lors de la réplication de l'ADN ou en conséquences d'agents mutagènes, des erreurs ou mutations peuvent être induites dans les génomes. Ces modifications de séquence ne sont pas observées dans n'importe quelle région. Les mutations dans les régions fonctionnelles

comme les gènes sont non favorables à la survie et à la reproduction d'une espèce. Elles ne sont pas conservées et apparaissent donc contre sélectionnées. Ainsi, les structures fonctionnelles d'un génome comme les TFBS peuvent être mises en évidence par leur nature exceptionnelle (Schbath, 1995; van Helden *et al.*, 1998). Une variation de la séquence de ces éléments régulateurs pouvant entraîner la non reconnaissance des facteurs de transcription et donc un défaut de régulation de l'expression des gènes, la sélection limite leur mutation.

La recherche de motifs sur-représentés au sein des promoteurs exploite cette caractéristique des TFBS pour identifier des motifs qui sont de potentiels éléments régulateurs de l'expression des gènes. Des études ont montré que des motifs fonctionnels sont des motifs statistiquement exceptionnels (Schbath, 1995).

Les méthodologies recherchant des sur-représentations de motifs ont été analysées afin de juger leur fiabilité sur des génomes eucaryotes de différentes complexités (Xue *et al.*, 2004). L'approche utilisée par les auteurs compare le comptage observé dans les régions intergéniques au comptage attendu dans l'ensemble du génome. Les comptages sont réalisés pour un ensemble de TFBS validés dans quatre organismes : *S. cerevisiae*, *Saccharomyces pombe,* le virus d'Epstein-Barr et *D. melanogaster*. Les résultats mettent en évidence que les TFBS sont sur-représentés dans les régions intergéniques des génomes eucaryotes inférieurs exclusivement. Les sur-représentations des TFBS chez *D. melanogaster* sont identifiées uniquement en adaptant la démarche c'est-à-dire en comparant le comptage observé au comptage attendu dans les exons. Cette adaptation permet de s'affranchir des comptages dans les introns, constituant 20% du génome de *D. melanogaster*. La difficulté à identifier les TFBS par une recherche de motifs sur-représentés est plus grande chez les eucaryotes supérieurs qui ont des génomes plus complexes (Xue *et al.*, 2004). De plus, les résultats montrent la nécessité de développer des approches adaptées à chaque organisme, une démarche pouvant être pertinente chez une espèce et ne pas l'être chez une autre. La recherche de motifs sur-représentés est à l'origine de différents outils constamment améliorés.

i) MotifSampler

MotifSampler utilise un algorithme de Gibbs sampling ou échantillonnage de Gibbs (Lawrence *et al.*, 1993; Neuwald *et al.*, 1995) pour identifier la présence de motifs consensus sur-représentés (Thijs *et al.*, 2001). Pour résumer, à partir d'un jeu de promoteurs, l'algorithme recherche un motif de taille n (proposée entre 5 et 15 bases) présent avec le plus de similitudes possibles dans les séquences du jeu de promoteurs considéré. La recherche débute avec un motif pris au hasard. Par modifications successives de sa séquence, l'outil s'approche d'un consensus sur-représenté jusqu'à ce qu'il soit stable et donc qu'il représente au mieux le jeu de séquences. Chaque utilisation d'un tel algorithme génère donc des listes de motifs sur-représentés différentes, dépendant du motif pris au hasard en début de recherche. Depuis les premiers algorithmes, des améliorations sont toujours proposées (Defrance & Helden, 2009; Schultheiss *et al.*, 2009).

ii) R'MES : Recherche de Motifs Exceptionnels dans les séquences

R'MES est un logiciel développé par Sophie Schbath *et al.* au laboratoire Mathématique et Informatique du Génome du centre INRA de Jouy en Josas (Hoebeke & Schbath, 2006).

La recherche de motifs exceptionnels, c'est-à-dire sur ou sous-représentés, est réalisée en comparant le comptage observé dans les séquences analysées au comptage attendu. Ce dernier est obtenu en utilisant les chaînes de Markov. Le choix de l'ordre du modèle de Markov permet de définir les probabilités d'observer une base A, C, G et T dans la séquence (ordre 0) ou les probabilités d'observer chacune des 4 bases après les mono, di, tri-nucléotides (ordre 1 à 3 respectivement) etc… Pour un motif de longueur l, l'ordre maximal qui s'ajuste au mieux est l'ordre l-1. En sortie de R'MES, un score est associé à chaque motif étudié. Une différence significative, c'est-à-dire un score supérieur ou inférieur au seuil de 3, indique que le motif est respectivement sur- ou sous-représenté.

iii) Regulatory Sequence Analysis Tools : RSAT

Des outils de recherche de motifs sur-représentés ont été mis au point avec le génome de *S. cerevisiae* (van Helden *et al.*, 1998) et sont proposés sur le portail de RSAT. Ce premier travail proposait de mettre en évidence des sur-représentations d'oligonucléotides en comparant leur comptage dans des régions intergéniques au comptage dans le génome de l'organisme étudié. L'outil initialement développé a été progressivement amélioré et adapté pour pouvoir être appliqué à des génomes d'eucaryotes supérieurs (van Helden *et al.*, 2000; Hulzink *et al.*, 2003; van Helden, 2003; Defrance *et al.*, 2008; Thomas-Chollier *et al.*, 2008; Turatsinze *et al.*, 2008; Defrance & Helden, 2009). RSAT met à la disposition des utilisateurs les outils d'analyses développés par van Helden *et al.* par une interface qui permet d'exploiter différentes approches pour un même jeu de promoteurs. A ce jour, RSAT inclut une page permettant d'obtenir les séquences de 1475 organismes puis propose l'analyse du lot de séquences téléchargées par différentes grandes catégories d'approches présentées par la suite. Ce site très complet propose la recherche de motifs sans *a priori* aussi bien que l'identification de TFBS connus.

b) Génomique comparative

Les régions fonctionnelles peuvent être conservées entre espèces ayant divergé (Sumiyama *et al.*, 2001; Woolfe *et al.*, 2005). Des études de génomique comparative ou empreintes phylogénétiques ont été menées pour identifier des structures communes à différents génomes comme les primates (Wasserman *et al.*, 2000; Boffelli *et al.*, 2003), les levures (Tompa, 2001; Kellis *et al.*, 2003), les plantes (Colinas *et al.*, 2002). Lors de l'étude des régions non codantes, une hypothèse de travail propose que les mécanismes de régulation de l'expression des gènes, nécessaires à la survie de l'organisme, soient également soumis à une pression de sélection. Les TFBS seraient alors conservés entre espèces. En effet les séquences non codantes conservées sont enrichies en TFBS (Levy *et al.*, 2001). Le projet ENCODE, The «ENcyclopedia Of DNA Elements» (Consortium, 2004) a montré l'intérêt des approches par génomique comparative pour des recherches d'éléments régulateurs dans les génomes de *H. sapiens* et de *M. musculus* (King *et al.*, 2007). De plus, l'organisation de TFBS impliqués dans la régulation de gènes co-exprimés dans un même processus est conservée entre gènes orthologues (Brown *et al.*, 2007).

Notons que les démarches par comparaison de génomes nécessitent différents choix. Le premier porte sur le nombre de génomes à considérer, deux ou plus. Plus ce nombre est important, plus les prédictions seront sensibles, mais moins elles seront spécifiques. Le deuxième choix porte sur la distance évolutive entre les organismes. L'ancêtre commun de

deux espèces doit être suffisamment éloigné pour qu'un nombre important de mutations se soit accumulé dans les régions non fonctionnelles et qu'il n'y ait pas de sur-estimation des séquences conservées, donc des TFBS. Cependant, l'ancêtre commun de deux espèces doit également être suffisamment proche pour que les régions fonctionnelles n'aient pas trop divergé et qu'il n'y ait pas de sous-estimation des TFBS candidats. En effet les séquences promotrices évoluent rapidement. Seule une faible conservation même entre gènes orthologues des céréales a été observée (Guo & Moose, 2003). Même dans une situation jugée très favorable de gènes sans paralogue mais avec un orthologue unique (Armisén *et al.*, 2008), aucune conservation entre les paires d'orthologues d'*A. thaliana*, d'*Oryza sativa*, de *vitis vinifera* et de *Populus trichocarpa* n'a été observée (Armisen, 2008). Des recherches de motifs conservés ont révélé une plus grande conservation entre *A. thaliana* et *Raphanus sativus* ou encore entre *O. sativa* et *P*. *trichocarpa*. Cependant cette conservation pourrait être due à une plus grande conservation générale des promoteurs et non à la conservation spécifique d'éléments régulateurs potentiels (Armisen, 2008).

Une évolution adaptative rapide des TFBS par mutations ponctuelles existe (Berg *et al.*, 2004), notamment après duplication des gènes de *S. cerevisiae* (Papp *et al.*, 2003). Il existe également un turnover des TFBS chez les mammifères (Dermitzakis & Clark, 2002) et un turnover des TSS produisant des TSS alternatifs pourrait être un cas extrême susceptible de diminuer l'efficacité des analyses de génomique comparative (Frith *et al.*, 2006). Une autre des difficultés d'une approche par génomique comparative pour la prédiction de TFBS vient du fait que la conservation de différentes régions du génome n'est pas la même. Certaines régions conservées n'ont pas de fonction évidente tandis que des régions non conservées peuvent être fonctionnelles (Nobrega & Pennacchio, 2004; King *et al.*, 2007).

Une autre limite de la génomique comparative pour une identification exhaustive de TFBS est l'importance de la présence des TFBS spécifiques d'une espèce (Dermitzakis & Clark, 2002). Les auteurs n'ont pas retrouvé 32 à 40% des TFBS fonctionnels identifiés chez *H. sapiens* dans les promoteurs des gènes orthologues de *M. musculus*. Les approches par empreintes phylogénétiques sont donc susceptibles de générer des faux négatifs, certains TFBS étant spécifiques d'une espèce ou d'autres ayant divergé. Pour une mise en évidence la plus pertinente possible de TFBS conservés entre espèces, le paramètre de conservation d'ordre des TFBS dans les séquences, et celui d'orientation devraient être considérés (Blanchette & Sinha, 2001). En plus des faux-négatifs, le taux de faux-positifs est souvent élevé lors de la recherche de TFBS (Blanchette & Sinha, 2001; Tompa, 2001). Par une approche utilisant les empreintes phylogénétiques, le taux de faux positifs a été réduit d'environ 85% par comparaison à une méthodologie n'exploitant qu'une seule séquence (Sandelin *et al.*, 2004).

Le site Internet de prédiction des TFBS ConSite propose une méthodologie de génomique comparative appelée «Phylogenetic footprinting» ou empreinte phylogénétique pour rechercher des séquences conservées entre gènes orthologues au sein de différentes espèces (Sandelin *et al.*, 2004). A partir des séquences promotrices d'un couple de gènes orthologues, un alignement global est réalisé. Individuellement, pour chacune des séquences, une recherche de TFBS potentiels est réalisée en exploitant les données de JASPAR. Dans cette base de données, les sites sont représentés sous forme de matrices et sont issus d'approches validées expérimentalement (Bryne *et al.*, 2008). Enfin, les matrices de TFBS obtenues pour les deux séquences sont comparées et seules celles conservées à

une même position de l'alignement dans les deux séquences sont considérées comme de potentielles cibles des TF.

c) Recherche de motifs conservés à une position préférentielle

Les TFBS dans le promoteur central et dans le promoteur proximal peuvent coopérer fonctionnellement comme discuté page 29. Plusieurs TFBS étant présents dans une même région, cette caractéristique leur impose des contraintes de position, afin de ne pas perturber la reconnaissance de complexes TF / TFBS et afin de pouvoir coopérer avec leurs partenaires. C'est pour cette raison que les régions promotrices proximales et centrales, à proximité du TSS, contiennent des TFBS qui sont sensibles à la position. Par exemple la boîte TATA a été observée à des distances fixes de différents sites reconnus par des facteurs de la transcription spécifique (Dion & Coulombe, 2003; Tabach *et al.*, 2007).

En complément de la fréquence des motifs, la recherche de biais positionnels a été présentée dès la fin des années 1980 (Mengeritsky & Smith, 1987). Par la suite, les premiers, Kielbasa *et al.* (2001) puis Fitzgerald *et al.* (2004) ont suggéré d'ajouter la position préférentielle des motifs par rapport au TSS comme critère de sélection pour identifier des TFBS potentiels. Les approches exploitant cette recherche de motifs conservés à une position préférentielle du TSS sont aujourd'hui de plus en plus exploitées pour identifier de nouveaux TFBS, notamment chez les métazoaires (Berendzen *et al.*, 2006; Bernard *et al.*, 2006; Bellora *et al.*, 2007; Bellora *et al.*, 2007; Tabach *et al.*, 2007; Yamamoto *et al.*, 2007; Casimiro *et al.*, 2008). La démarche proposée par Yamamoto *et al.* a été mise au point pour l'étude du génome d'*A. thaliana*. Elle est très similaire à celle qui a été entreprise lors de cette thèse avant la publication de Yamamoto *et al.* (2007b).

d) Vers une meilleure prédiction de l'identification des TFBS

Chacune des démarches *in silico* proposées aujourd'hui présente des points faibles. La détection de TFBS au sein des promoteurs est encore un challenge complexe qui n'est pas résolu (Tompa *et al.*, 2005). Il est donc apparu nécessaire de combiner aux approches déjà existantes des informations supplémentaires et / ou de combiner les différentes hypothèses de travail.

Deux catégories d'analyses peuvent être réalisées pour étudier les promoteurs d'un génome complet : une analyse génomique de tous les promoteurs en même temps ou une analyse de groupes de promoteurs. Pour cette deuxième option, des TFBS communs enrichis et attendus dans un jeu de gènes caractérisés peuvent conduire à des prédictions contenant moins de faux positifs (Mitasiunaite *et al.*, 2009).

Le regroupement est exploité de plus en plus avec l'avènement des technologies de puces à ADN qui permettent de constituer des groupes de gènes ayant des profils d'expression similaires . D'autres regroupement sont réalisés en fonction de l'annotation des gènes (Cora *et al.*, 2004). L'hypothèse de travail est alors qu'un tel regroupement permet d'obtenir un enrichissement en TFBS communs qui sont impliqués dans la régulation spécifique des gènes analysés. Par exemple une analyse d'empreinte phylogénétique sur des groupes de gènes co-exprimés peut être réalisée (Vandepoele *et al.*, 2006). De même des groupes de gènes peuvent être constitués en fonction de leur annotation fonctionnelle, disponible *via* la ressource Gene Ontology (Ashburner *et al.*, 2000). Chez *H. sapiens*, de

telles analyses ont révélé que pour chaque groupe de gènes ayant une même fonction biologique, un à plusieurs éléments régulateurs sont plus particulièrement présents (Long *et al.*, 2004). L'efficacité du regroupement de gènes a été démontrée *via* des validations expérimentales des prédictions *in silico* (GuhaThakurta, 2006).

Lors de comparaisons de 13 outils d'analyse des promoteurs *in silico*, Tompa et al (2005) ont révélé les faiblesses de chacun à prédire efficacement des TFBS ainsi que leur propension à générer des faux positifs. Ce travail les a conduit à proposer de combiner différentes hypothèses de travail pour améliorer l'efficacité des prédictions (Tompa *et al.*, 2005). Différentes combinaisons ont été élaborées depuis, comme le couplage d'un Gibbs sampling avec une recherche de position préférentielle des motifs par rapport au TSS (Molina & Grotewold, 2005) ou encore une empreinte phylogénétique couplée à une recherche de position préférentielle des motifs par rapport au TSS (Bellora *et al.*, 2007).

e) Il n'y a pas d'approche idéale

Finalement, le choix de l'approche à utiliser dépend de multiples critères. Tout d'abord, la question biologique : est-ce que la recherche de TFBS a pour objectif de découvrir de nouveaux TFBS ou bien d'identifier la présence de TFBS connus, indexés dans des banques de données ? Une deuxième question à considérer concerne la taille du jeu de données à étudier. Analyser un jeu de quelques promoteurs ou un jeu couvrant un génome complet n'a pas les mêmes implications statistiques et informatiques. Une approche efficace sur des petits jeux pourra ne plus l'être sur des grands et inversement. Il restera alors à savoir le type de données à analyser, c'est-à-dire s'il s'agit de promoteurs pour lesquels le TSS a été positionné ou bien de séquences non annotées pour lesquelles n bases en amont du codon initiateur ont été extraites. De plus, les outils disponibles pour la recherche de TFBS sont basés sur des algorithmes différents qui peuvent être efficaces pour l'analyse des séquences de certains organismes tandis qu'ils ne le seront pas pour d'autres (Tompa *et al.*, 2005). En effet, la complexité des régions intergéniques n'est pas la même chez tous les eucaryotes, certains génomes ayant une plus grande présence en éléments transposables que d'autres par exemple (Buisine *et al.*, 2008; Han *et al.*, 2009). L'utilisateur d'un outil doit donc se renseigner avant de l'exploiter pour vérifier s'il est adapté à sa problématique. Notons également que chaque approche propose des paramètres par défaut que l'utilisateur peut modifier. Comme pour chaque outil d'analyse développé, l'utilisateur doit prendre connaissance de ces paramètres afin de les adapter à ses besoins, mais aussi afin de pouvoir interpréter les résultats avec pertinence.

En conclusion les approches *in silico* permettent des analyses de génomes complets, mais restent des prédictions. Combiner les deux approches, *in silico* puis *in vitro* / *in vivo* est une des meilleures stratégies à adopter pour obtenir des résultats en filtrant des candidats d'intérêt *via* la bioinformatique et en les validant (ou non) en biologie expérimentale (Rombauts *et al.*, 2003). Des technologies combinant biologie expérimentale et bioinformatique ouvrent des perspectives d'étude intéressantes (King *et al.*, 2007), comme par exemple le ChIP-séquençage qui avec l'avènement du séquençage à haut débit, promet d'être la technologie phare dans les années à venir (Park, 2009).

2.2 Indexation des sites de fixation

La revue Nucleic Acids Research présente annuellement une édition dédiée aux bases de données en biologie. Dans l'édition de 2009, la catégorie «Transcriptional regulator sites and transcription factors» répertoriait 64 bases de données consacrées aux TF et à leurs sites de fixations. Notons que ces bases de données concernent tous les organismes, procaryotes comme eucaryotes. Il existe parmi elles des bases de données dédiées au règne végétal et plus spécialement à *A. thaliana*.

Le site Web *AGRIS* est une ressource dédiée à la plante modèle *A. thaliana* (Davuluri *et al.*, 2003). Ce site met à la disposition du public deux bases de données : AtcisDB pour «*A. thaliana* cis-regulatory database» qui contient des TFBS identifiés et AtTFDB pour «*A. thaliana* transcription factor database» qui contient les TF. AtcisDB est constituée de 99 TFBS identifiés par des méthodes de biologie expérimentale. Pour chaque TFBS, son nom, son motif consensus et une ou plusieurs références bibliographiques sont disponibles.

PLACE est une base de données dédiée aux TFBS identifiés chez les plantes (Higo *et al.*, 1999). Depuis deux ans, PLACE n'est plus mise à jour par son administrateur Kenichi Higo, mais présente néanmoins une ressource pertinente pour les recherches concernant les TFBS présents chez les plantes. Cette base est construite à partir de la lecture d'articles relatant l'identification de TFBS. Chaque entrée est associée à des publications détaillant la découverte de l'élément régulateur. La dernière mise à jour, qui date de 2007, présente 469 TFBS, dont 99 sont des TFBS observés chez *A. thaliana*.

D'autres bases de données comme la version public n° 7.0 de Transfac (Wingender *et al.*, 1996; Wingender, 2008) ou PlantCARE (Lescot *et al.*, 2002) proposent respectivement 58 et 41 TFBS propres à *A. thaliana*, qui sont comprises dans les bases PLACE ou AGRIS. Sur le même principe que PLACE et AGRIS, Transfac et PlantCARE proposent une à plusieurs références bibliographiques associées à chaque TFBS.

Ainsi, les connaissances et preuves expérimentales concernant les éléments régulateurs et les facteurs de transcription sont encore faibles. Le pourcentage de TF validés par rapport aux TFBS est très disproportionné : chez *A. thaliana*, près de 2200 gènes codant des protéines existent (Samson *et al.*, 2004), tandis que le cumul des bases de données de TFBS ne permet de disposer que de 140 séquences validées. Les TFBS sont des séquences difficiles à identifier et dont la validation fonctionnelle est longue. C'est pourquoi aujourd'hui une étude impliquant des approches *in silico* et des approches *in vitro* / *in vivo* semble être la stratégie la plus efficace susceptible de prédire des résultats pertinents. La bioinformatique est une science qui peut permettre de restreindre le champ d'investigation de la biologie expérimentale.

3 Promoteur central des organismes eucaryotes : apport des études *in silico*

Les connaissances actuelles de l'architecture des promoteurs centraux chez les eucaryotes reposent pour l'essentiel sur des analyses du règne des animaux. Des études *in silico* ont contribué à une meilleure connaissance des TFBS qui les constituent. Dans une première partie, les études ayant permis l'identification des TFBS généraux sont présentées avant de préciser les connaissances propres à différents organismes eucaryotes.

3.1 Identification des sites de fixation du promoteur central

Un minimum de cinq éléments régulateurs généraux peuvent être impliqués dans la formation du complexe d'initiation de la transcription ou TIC (Bajic *et al.*, 2006; Juven-Gershon *et al.*, 2008). Tous ont été identifiés chez des animaux, mais n'ont pas forcément été observés dans d'autres règnes. Ils sont représentés schématiquement dans la Figure 3-1.

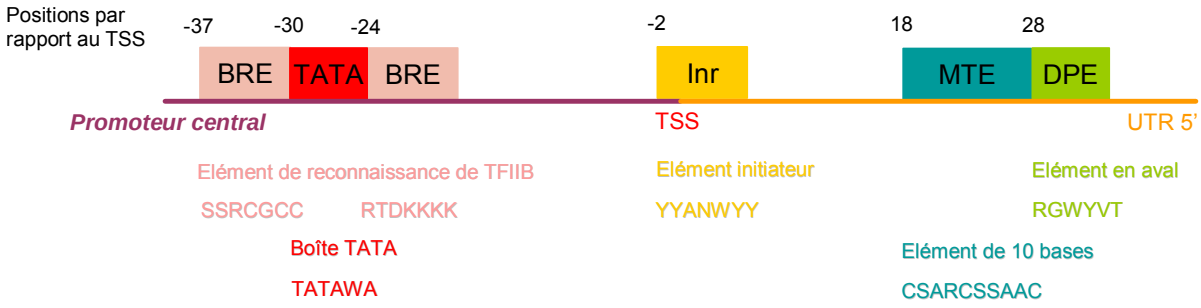


Figure 3-1 : Représentation des éléments observés dans le promoteur central des gènes à ARN pol II chez les mammifères.

Les positions sont indiquées par rapport au TSS et correspondent à la position du début des éléments régulateurs. Les noms et les séquences consensus chez les mammifères sont indiqués sous la représentation schématique du promoteur. Boîte TATA (Mathis & Chambon, 1981) ; Inr (Javahery *et al.*, 1994) ; DPE (Burke & Kadonaga, 1997) ; MTE (Lim *et al.*, 2004) ; BRE (Lagrange *et al.*, 1998; Deng & Roberts, 2005). Code IUB définissant les bases dégénérées disponible en Annexe I (NC-IUB, 1985).

3.1.1 Elément initiateur

L'élément initiateur est une séquence observée à un emplacement très strict proche du TSS (Smale & Kadonaga, 2003). Son motif consensus est YYANWYY chez les mammifères où il est présent dans 46% des gènes 2 bases en amont du TSS (Javahery *et al.*, 1994). Le A est la première base transcrite par l'ARN pol II. Cet élément régulateur est le plus observé au sein des promoteurs des protozoaires aux métazoaires, mais son motif consensus varie selon les organismes et selon l'architecture des promoteurs (Carninci, 2006; Yamamoto *et al.*, 2007).

Un facteur de transcription TAF («TBP-Associated Factor») du TFIID reconnaît l'Inr. Des études ont démontré que cet élément régulateur permet une meilleure fixation du TIC et ainsi une activité transcriptionnelle plus importante (Smale & Kadonaga, 2003).

3.1.2 Boîte TATA

La boîte TATA est un des premiers éléments régulateurs de la transcription des gènes identifiés chez les eucaryotes (Mathis & Chambon, 1981). Elle est impliquée dans la formation du TIC en étant reconnu par un facteur de transcription spécifique : la TBP «TATA-Binding protein». La boîte est riche en bases A et T et est conservée chez les métazoaires environ 30 bases en amont du TSS (Patikoglou *et al.*, 1999).

Il a été longtemps considéré que la boîte TATA était essentielle à la transcription des gènes. Aujourd'hui, cette hypothèse n'est plus valable. En remplaçant la séquence de la boîte TATA par des séquences aléatoires, l'initiation de la transcription est toujours assurée et une transcription supérieure à celle des gènes sauvages est même parfois induite (Singer *et al.*, 1990). Finalement, pour initier la transcription, même en l'absence de boîte TATA, les promoteurs peuvent être reconnus par des protéines reconnaissant la boîte TATA (Tsai & Sigler, 2000).

3.1.3 Elément en aval ou DPE

Le DPE («Downstream Promoter Element») a été identifié chez *D. melanogaster,* avec pour motif consensus RGWYVT (Burke & Kadonaga, 1996). Cet élément a été mis en évidence lors d'une étude de promoteurs sans boîte TATA qui était menée afin d'identifier les motifs pouvant être impliqués dans la formation du TIC. Le DPE est conservé chez l'humain (Burke & Kadonaga, 1997). Il n'a pas été signalé chez les plantes. Il est observé dans l'UTR 5' 28 bases en aval du TSS et est reconnu par un TAF du TFIID. Associé à l'Inr, cet élément régulateur permet le recrutement de la TBP *via* l'intervention de facteurs qui lui sont associés. La distance entre l'Inr et le DPE est très conservée et une modification même minime entraîne le non recrutement de la TBP (Kutach & Kadonaga, 2000). La boîte TATA et le DPE peuvent être présents sur un même promoteur. Dans ce cas, le DPE n'intervient pas pour recruter la TBP si la boîte TATA le fait. Lorsque la boîte TATA est rendue non fonctionnelle, le DPE contribue à la transcription du gène (Burke & Kadonaga, 1996).

3.1.4 Elément de 10 bases ou MTE

Le MTE («Motif Ten Element») est localisé dans l'UTR 5' et a pour motif consensus CSARCSSAAC (Kutach & Kadonaga, 2000). Suite au constat que 31% des gènes ne contiennent ni boîte TATA ni DPE, le MTE a été découvert chez *D. melanogaster*. Il est fonctionnel lorsqu'il est 18 à 27 bases en aval du TSS. Le MTE peut être présent avec la boîte TATA sur un même promoteur, et agir en module avec elle (Lim *et al.*, 2004) tout comme il coopère fonctionnellement avec l'Inr (Ohler *et al.*, 2002). Cet élément est conservé chez *H. sapiens* (Lim *et al.*, 2004). Il n'a pas été signalé chez les plantes.

3.1.5 Elément reconnu par TFIIB ou BRE

Le BRE («TFIIB Recognition Element») a été identifié chez les mammifères et encadre la boîte TATA. Il est constitué de deux sous unités : SSRCGCC observée en -37 par rapport au TSS (Lagrange *et al.*, 1998) et RTDKKKK en -24 (Deng & Roberts, 2005). Le BRE est

reconnu par TFIIB qui comme TFIID est impliqué dans la formation du TIC. A ce jour, aucune séquence BRE n'a été identifiée chez *S. cerevisiae* ni chez *A. thaliana* (Lagrange *et al.*, 1998; Shahmuradov *et al.*, 2003). Il est observé chez *H. sapiens* mais sa fonction reste mal connue. Il pourrait intervenir dans la définition de la direction de la transcription.

3.1.6 Autres TFBS du promoteur central

Différentes études dédiées au promoteur central de petits groupes de gènes ont mis en évidence de nouveaux éléments. Par exemple l'élément central en aval ou DCE («downstream core element») est composé de trois sous-éléments observés sur un même coté de l'hélice d'ADN, dans l'UTR 5'. Le DCE est observé dans les promoteurs de gènes de la β globuline de *H. sapiens* (Lewis *et al.*, 2000; Lewis *et al.*, 2005). De tels motifs sembleraient être impliqués dans des fonctions plus spécifiques. Récemment, des éléments ayant un fonctionnement n'impliquant pas nécessairement les TAF ont été identifiés, ce sont les éléments XCPE1 et XCPE2 (Anish *et al.*, 2009).

3.1.7 Eléments régulateurs connus non indispensables

La Table 3-1 présente un bilan de la présence des cinq motifs dans trois catégories d'eucaryotes : les mammifères, les plantes et les protozoaires. L'Inr, la boîte TATA, le DPE, le MTE et le BRE doivent être considérés comme les éléments contribuant mais n'étant pas indispensables à l'initiation de la transcription (Muller *et al.*, 2007). Aucun n'est présent dans l'ensemble des gènes codant d'un génome. L'image qui émerge pour les métazoaires est qu'au lieu du dogme ancien proposant une liaison entre TBP et boîte TATA pour initier la transcription, des jeux distincts de TF reconnaîtraient une grande variété de TFBS et de promoteurs centraux (Muller *et al.*, 2007). Le concept du promoteur basal générique (Buratowski, 1997) était une simplification qui ne peut plus être soutenue.

Nom du motif	BRE	Boîte TATA	Inr	MTE	DPE
Mammifères	Oui	Oui	Oui	Oui	Oui
Plantes	Non	Oui	Oui	*?*	*?*
Protozoaires	Non	Oui	Oui	*?*	Non

Table 3-1 : Eléments régulateurs dans le promoteur central de différentes familles d'organismes.

Résumé basé sur différents travaux. Concernant le BRE : Lagrange *et al.* (1998), Yang *et al.* (2007). Concernant la boîte TATA : Mathis & Chambon (1981), Shahmuradov *et al.* (2003), Basehoar *et al.* (2004). Concernant l'Inr : Javahery *et al.* (1994), Carninci *et al.* (2006), Yamamoto *et al.* (2006). Concernant le MTE : Lim *et al.* (2004). Concernant le DPE : Burke & Kadonaga (1997), Yang *et al.* (2007).

Ce bilan permet de mettre en évidence la complexité du processus d'initiation de la transcription des gènes à ARN pol II. D'autres éléments spécifiques, comme le DCE (Lewis *et al.*, 2000; Lewis *et al.*, 2005) ou XCPE1, XCPE2 (Anish *et al.*, 2009) peuvent intervenir. De plus, il n'est pas exclu que d'autres éléments régulateurs du promoteur central ne soient pas encore découverts et qu'ils puissent être impliqués dans l'initiation de la transcription. Cette présentation des TFBS du promoteur central ne décrit certainement qu'une partie des éléments régulateurs, en particulier chez les protozoaires ou les plantes dont le promoteur central a été moins étudié. Il reste probablement des éléments à identifier. Les approches *in*

silico contribuent à une meilleure connaissance de ces éléments du promoteur central des eucaryotes et de leur organisation.

3.2 Promoteur central chez *Saccharomyces cerevisiae*

S. cerevisiae est le premier organisme eucaryote dont la séquence a été publiée (Goffeau, 1996). En tant que génome de référence, de nombreuses études *in silico* ont d'abord été développées spécifiquement avec *S. cerevisiae* avant d'être généralisées à d'autres organismes : c'est un organisme modèle dans le développement d'outils de prédiction de TFBS. Il a aussi contribué à la connaissance fonctionnelle d'éléments régulateurs du promoteur central.

3.2.1 Boîte TATA chez les protozoaires

a) Observation sur une large région

Une des difficultés de l'analyse de la boîte TATA ou des gènes qui la contiennent provient de la connaissance non stricte de sa séquence consensus. En effet, des variants de ce motif, observés au même emplacement, rendent les boîtes TATA difficiles à distinguer et induisent en erreur les études conduites pour identifier un motif consensus fiable (Loganantharaj, 2006). Une richesse en A et T présente dans les promoteurs eucaryotes à l'emplacement de la boîte TATA ne remplit pas *de facto* la fonction de la boîte TATA (Singer *et al.*, 1990).

Une étude des promoteurs de *S. cerevisiae* a permis d'établir un motif consensus de l'élément régulateur chez les protozoaires (Basehoar *et al.*, 2004). Elle impose que les séquences pour être identifiées comme boîtes TATA fonctionnelles :

- coïncident avec une séquence TAWWNNNN ;
- soient conservées dans les orthologues de quatre espèces de Saccharomyces ;
- soient présentes dans la région de la boîte, [-150, -20].

Ces critères ont permis de définir la séquence TATAWAWR comme motif consensus, cette séquence a été validée biologiquement (Basehoar *et al.*, 2004). Chez *S. cerevisiae*, 19% des promoteurs contiennent cet élément régulateur dans sa région fonctionnelle prédite : [-150, -20]. La distribution de la séquence TATAWAWR chez *S. cerevisiae* et chez *H. sapiens* montre des différences de contraintes topologiques du motif entre les deux organismes (Yang *et al.*, 2007). En effet, la boîte TATA chez les protozoaires possède des contraintes moins strictes que chez des métazoaires. Elle est préférentiellement positionnée dans la région [-150, -20] par rapport au TSS chez les protozoaires.

b) Biais fonctionnels et évolutifs des gènes contenant une boîte TATA

Basehoar *et al.* (2004) ont comparé des gènes avec et sans boîte TATA. Les gènes contenant le motif consensus TATAWAWR ont tendance à :

- évoluer plus rapidement, ce biais pouvant être dû à la plus forte présence des gènes ayant une boîte TATA dans les régions subtélomériques qui évoluent plus rapidement chez *S. cerevisiae* (Kellis *et al.*, 2003) ;

- s'exprimer à un niveau particulièrement faible ou fort, ce qui témoignerait d'une capacité de modulation de leur expression en fonction des besoins spécifiques de la cellule ;
- être impliqués dans des processus biologique spécifiques, comme par exemple être activés en réponse à un stress ;
- être, plus que les autres gènes, soumis à des régulations liées aux nucléosomes et au remodelage de la chromatine, et moins dépendants des facteurs associés à la TBP (Basehoar *et al.*, 2004).

Deux modes de régulation de l'expression des gènes pourraient exister (Basehoar *et al.*, 2004). Celui dit «simple», permettrait aux gènes comme les gènes de ménage de s'exprimer à un taux constant, de manière constitutive *via* peu (ou pas) de régulations. Ce mode ne nécessiterait qu'un complexe ARN pol II, TFIID et les facteurs associés à la TBP. Le deuxième mode pourrait permettre aux gènes nécessitant d'être exprimés dans des conditions spécifiques, en réponse à un stress par exemple, d'être réprimés ou exprimés à un niveau de base en condition «normale» et d'être exprimés au niveau requis en conditions spécifiques. Ce mode ferait intervenir une plus grande diversité de régulations : conformation de l'ADN, activateurs ou inhibiteurs de l'expression des gènes.

c) Motifs consensus incompatibles entre protozoaires et métazoaires

Pour rechercher la présence des variants fonctionnels de la boîte TATA, Yang *et al.* (2007) ont défini une séquence appelée TATA-532 car elle comprend 532 différents motifs variants. Son motif consensus inclut HWHWWWWR et exclut CAYTTTWR, MAMAAAAR et CTYAAAAR pour coïncider au mieux aux critères définissant une boîte TATA (Yang *et al.*, 2007). Ce motif consensus est pertinent chez *H. sapiens* où il est sur-représenté dans une région stricte, mais pas chez *S. cerevisiae* où TATA-532 est observée dans 96% des promoteurs si elle est recherchée dans la région de la boîte TATA (Yang *et al.*, 2007).

Ainsi, la boîte TATA comme ses variants fonctionnels, bien que conservés au sein des eucaryotes, présentent des spécificités chez les métazoaires (Yang *et al.*, 2007).

3.2.2 Conservation de l'élément initiateur

L'Inr identifié chez *D. melanogaster* est caractérisé chez *S. cerevisiae* comme chez les métazoaires par une position préférentielle très stricte à l'emplacement du TSS (Yang *et al.*, 2007). Le motif consensus YYANWYY est conservé chez les métazoaires.

La boîte TATA ne serait donc pas l'élément du promoteur central le plus conservé chez les eucaryotes : l'Inr l'est plus. Il est caractérisé par des contraintes topologiques similaires entre eucaryotes, et en complément, il est l'élément du promoteur central le plus largement observé. Chez *S. cerevisiae*, 40% des gènes possèdent un Inr (Yang *et al.*, 2007).

3.2.3 BRE et le DPE : des éléments régulateurs non conservés

Les promoteurs des gènes de *H. sapiens* et de *S. cerevisiae* ont été analysés en comparant la présence de deux éléments régulateurs du promoteur central (Table 3-1 page 43) : le DPE et le BRE (Yang *et al.*, 2007). Aucun n'a été observé chez *S. cerevisiae* alors

qu'ils sont présents chez *H. sapiens*. Ils ne semblent pas être conservés au sein de l'ensemble des eucaryotes.

En conclusion, les analyses de promoteurs de *S. cerevisiae* ont permis de souligner certaines spécificités des éléments régulateurs du promoteur central propres aux organismes protozoaires. Ils sont en effet les seuls eucaryotes à avoir une boîte TATA qui n'est pas soumise à des contraintes positionnelles strictes environ 30 bases en amont du TSS. De plus ils ne contiennent pas les éléments DPE et BRE observés chez les mammifères. Malgré ces spécificités, les approches mises au point pour l'analyse des promoteurs chez des protozoaires ainsi que les hypothèses sous-jacentes sont souvent exploitées pour étudier des organismes ayant un génome plus complexe. Les biais fonctionnels et évolutifs observés chez les protozoaires peuvent en effet être conservés chez les métazoaires s'ils sont réellement propres à un élément régulateur.

3.3 Promoteur central chez *Homo sapiens*

La séquence du génome de *H. sapiens* a été publiée en 2001 (Lander *et al.*, 2001). Une des spécificités de son génome et plus généralement des promoteurs de mammifères est la présence d'îlots CpG.

3.3.1 Ilots CpG

Les îlots CpG sont un constituant majeur des promoteurs des gènes chez les mammifères. Il s'agit de régions de 0.3 à 3 kb contenant une forte proportion de dinucléotides CpG, c'est-à-dire d'une base C suivie d'une base G liées par une liaison phosphodiester. Ces sites constituent 1% du génome de *H. sapiens* ; ce faible pourcentage est corrélé au faible pourcentage de régions codantes du génome : moins de 5% (Lander *et al.*, 2001). Les îlots CpG sont observés dans le promoteur proximal. Ils sont naturellement contre-sélectionnés dans les génomes, les bases C suivies d'une base G étant méthylées et sujettes à des désaminations. Néanmoins, plus de la moitié des promoteurs des gènes humains contiennent des îlots CpG (Antequera & Bird, 1993; Ioshikhes & Zhang, 2000). Ces régions sembleraient être impliquées dans une régulation épigénétique en permettant ou non l'accès de la machinerie de transcription au promoteur. Les cytosines des îlots CpG sont en effet non méthylées lorsque le gène s'exprime et des défauts de méthylation peuvent avoir pour conséquences des maladies liées à une sur expression des gènes par exemple (Lujambio & Esteller, 2007). Néanmoins, leur rôle précis reste aujourd'hui non défini.

Les recherches d'îlots CpG chez d'autres eucaryotes que les mammifères n'ont jamais abouti. Chez les plantes, *A. thaliana* comme *O. sativa*, des recherches à l'échelle génomique n'ont pas permis d'identifier d'îlots CpG (Yamamoto *et al.*, 2007), même lorsque les différences de composition en bases entre *A. thaliana* et les mammifères ont été prises en compte (Rombauts *et al.*, 2003).

3.3.2 Sites de fixation des facteurs de transcription généraux

a) Organisation en modules

Entre gènes orthologues *H. sapiens* / *M. musculus*, la conservation des bases est plus grande dans la région du promoteur central que dans des régions plus distantes du TSS (Jin

et al., 2006). Les éléments BRE, TATA, Inr, MTE et DPE (Figure 3-1) qui sont conservés entre gènes orthologues *H. sapiens* / *M. musculus* ont été prédits comme fonctionnels (Table 3-2). L'absence de ces éléments dans une partie des promoteurs montre qu'aucun n'est essentiel au promoteur central chez *H. sapiens*, la même conclusion étant valable chez *M. musculus* (Jin *et al.*, 2006). L'Inr, l'élément le plus observé, se retrouve dans moins de 63% des promoteurs.

	BRE	Boîte TATA	Inr	MTE	DPE
Pourcentage de promoteurs contenant l'élément	21.7%	16.5%	62.7%	56.9%	12.0%

Table 3-2 : TFBS observés dans le promoteur central chez *Homo sapiens* (Jin *et al.*, 2006).

Les cinq éléments régulateurs de la Table 3-2 ne sont que très rarement présents au sein d'un même promoteur (Jin *et al.*, 2006). Deux à trois éléments sont plus souvent observés dans un même promoteur et pourraient agir en modules. Dans ces situations, l'écart entre les éléments est particulièrement conservé afin de conférer au promoteur une architecture optimale. Comme démontré *in vitro* (Lim *et al.*, 2004), la modularité entre la boîte TATA, l'Inr et le MTE, a été retrouvée par les auteurs lors de cette étude. En complément, ils proposent une modularité entre BRE, Inr et MTE (Jin *et al.*, 2006).

Chez *M. musculus*, les boîtes TATA fonctionnelles sont dans la région [-32, -29] par rapport au TSS (Ponjavic *et al.*, 2006). Le dinucléotide initiateur YR est observé au sein des gènes contenant une telle boîte TATA. Pour les occurrences de boîtes TATA non fonctionnelles en position -33 ou en amont, le dinucléotide YR s'étend en amont de -1 et la boîte TATA s'étend en aval de -33 : les deux éléments sont ainsi distants de 28 à 31 bases, comme le sont les boîtes TATA fonctionnelles et le dinucléotide initiateur YR. Ce biais représenterait une compensation de l'organisation du promoteur central pour obtenir la conformation «idéale» de l'écart boîte TATA - YR (Ponjavic *et al.*, 2006). Les occurrences de la boîte TATA non fonctionnelles en position -29 ou en aval n'ont pas le dinucléotide initiateur YR. L'utilisation de TSS alternatifs en aval du premier TSS rencontré pourrait compenser cette absence (Ponjavic *et al.*, 2006). Ces hypothèses sont en accord avec des travaux récents montrant chez *H. sapiens* (Carninci, 2006) comme chez les plantes (Yamamoto *et al.*, 2009) que l'organisation et l'utilisation des TSS alternatifs dépend de la présence d'une boîte TATA dans le promoteur.

b) Organisation du promoteur central, influence sur la transcription

Chez *H. sapiens* et *M. musculus*, quatre groupes de promoteurs établis selon la richesse en A+T par rapport à la richesse en C+G en amont et en aval des TSS sont caractérisés par des propriétés différentes en terme de présence en éléments régulateurs mais aussi en terme de «dinucléotide initiateur» (Bajic *et al.*, 2006). Les dinucléotides initiateurs statistiquement associés à un des quatre groupes de promoteurs (Table 3-3) ne sont pas tous en accord avec la séquence YYANWYY, identifiée chez *D. melanogaster* et conservée avec les mêmes contraintes topologiques chez *H. sapiens* (Yang *et al.*, 2007). De plus ils ont chacun un environnement en bases spécifique et les TFBS qui sont associés à ces quatre groupes montrent des biais significatifs. D'autres séquences Inr aujourd'hui non caractérisées pourraient donc exister (Bajic *et al.*, 2006).

		Richesse en bases en aval du TSS	
		A et T	C et G
Richesse en bases en amont du TSS	A et T	GG	TA
	C et G	TA	CG

Table 3-3 : Dinucléotides initiateurs et composition en bases des promoteurs.

Les quatre groupes de promoteurs ont été constitués en fonction de leur richesse en bases en amont et aval du TSS (Bajic *et al.*, 2006). A chacun est associé un dinucléotide initiateur statistiquement plus présent chez *H. sapiens* et *M. musculus*.

Chaque groupe de gènes correspondant à une annotation fonctionnelle de la Gene Ontology (Ashburner *et al.*, 2000) a plus tendance à avoir une des quatre organisations de promoteurs. Les différentes catégories de promoteurs pourraient donc être reliées à des expressions spécifiques *via* différentes organisations des promoteurs contenant différents TFBS et ayant une composition en bases spécifique (Bajic *et al.*, 2006). Cette hypothèse rejoint celle de Basehoar *et al.* (2006) proposant l'existence de plusieurs modes de régulation de l'expression des gènes.

c) Contraintes topologiques pour définir un motif consensus de la boîte TATA

Au cours des dernières années, les études consacrées à la boîte TATA chez *H. sapiens* ont présenté des résultats très différents, allant pour les extrêmes de moins de 3% des gènes contenant l'élément régulateur (FitzGerald *et al.*, 2004) à 64% (Trinklein *et al.*, 2003). Ces différences sont majoritairement dues aux contraintes imposées par les différents auteurs. Ne prenant pas en considération la contrainte positionnelle du motif, Trinklein *et al.* (2003) ont utilisé le motif consensus TAWWWW soit 16 séquences de longueur 6, et l'ont recherché dans la région [-550, +50]. Plus stricts, Fitzgerald *et al.* (2004) ont utilisé le motif consensus TATAAAD, soit 3 séquences de longueur 7, et l'ont recherché dans une fenêtre fonctionnelle de 20 bases. Le rôle des contraintes topologiques apparaît comme capital pour définir une liste de gènes contenant une boîte TATA (Ponjavic *et al.*, 2006). La boîte TATA n'est fonctionnelle chez les mammifères que dans la région stricte [-32, -29] par rapport au TSS (Ponjavic *et al.*, 2006). Les gènes ayant une boîte TATA dans cette région ont plus fréquemment une expression spécifique en fonction des tissus.

Le motif consensus de la boîte TATA pourrait ne pas être exclusivement constitué des bases A et T. Chez *H. sapiens*, parmi les différentes séquences conformes au motif consensus TATAWAWN (Juo *et al.*, 1996) deux sont apparues comme ayant des contraintes topologiques plus marquées : les motifs TATA(A/T)AAG (Shi & Zhou, 2006). Les extensions de ces motifs avec une base C ou G, en amont comme en aval, conservent les mêmes contraintes topologiques que la boîte TATA. Les contraintes topologiques de ces extensions sont conservées dans les promoteurs d'autres eucaryotes, notamment *M. musculus*, *Danio rerio* et *D. melanogaster*. Ces séquences pourraient donc constituer un motif consensus plus approprié pour identifier la présence d'une boîte TATA reconnue comme site de fixation de la TBP (Shi & Zhou, 2006).

Se basant sur les contraintes topologiques et une étude des couples de gènes orthologues dont une partie des résultats est résumée Table 3-2, Jin *et al.* (2006) proposent

que moins de 20% des gènes de mammifères possèdent la boîte TATA. Ce pourcentage est en accord avec les prédictions établies chez les métazoaires (Basehoar *et al.*, 2004).

3.3.3 Caractéristiques fonctionnelles communes pour annoter les éléments régulateurs

Une analyse de l'annotation de la Gene Ontology, réalisée avec les gènes contenant exclusivement une boîte TATA, un Inr, ces deux motifs ou aucun de ces deux motifs, a mis en évidence des biais fonctionnels spécifiques dans chacune de ces catégories (Yang *et al.*, 2007). La Table 3-4 résume les principaux biais révélés par cette étude. Chacun des groupes de gènes caractérisés par la présence ou non des éléments régulateurs du promoteur central est plus particulièrement impliqué dans une catégorie fonctionnelle précise.

Contenu des gènes	Fonctions biologiques les plus fréquentes
Boîte TATA	- Organogenèse - Réponse à des stimulus biotiques
Boîte TATA et Inr	- Assemblage des nucléosomes - Adhésion cellulaire
Inr	- Processus biologiques basiques (synthèse des protéines, processus des ARNm)
Aucun des 2 éléments	- Autres processus biologiques basiques (croissance cellulaire, transport intracellulaire, métabolisme…)

Table 3-4 : Biais fonctionnels des gènes de *H. sapiens* contenant une boîte TATA et / ou un Inr (Yang *et al.*, 2007).

L'annotation fonctionnelle des gènes contenant une boîte TATA est la plus éloignée de celle des gènes ne contenant aucun élément du promoteur central. La présence de cette boîte associée ou non à l'Inr induirait chez les gènes une tendance à être plus impliqués dans la régulation de l'expression de fonctions spécifiques. Ce biais est en accord avec une étude réalisée chez *S. cerevisiae* (Basehoar *et al.*, 2004) et avec des observations suggérant que les gènes exprimés dans des conditions spécifiques contiennent plus fréquemment une boîte TATA (Smale & Kadonaga, 2003).

Cette analyse *in silico* de groupes de gènes en fonction de l'architecture de leurs promoteurs démontre l'intérêt de la recherche de biais pour aider à l'annotation fonctionnelle d'un élément régulateur. La tendance des gènes contenant un élément régulateur à être plus ou moins impliqués dans une fonction est un premier pas vers la connaissance du rôle de l'élément régulateur.

En conclusion, les analyses de promoteurs de mammifères font partie des analyses les plus complètes. La connaissance de l'organisation du promoteur central de ces eucaryotes est probablement la plus aboutie. Les éléments régulateurs présentés dans la Table 3-1 (page 43) sont observés chez les mammifères. Cependant, malgré ces connaissances, la région de la boîte TATA reste aussi méconnue que chez les protozoaires. Au début de cette thèse, aucune liste des variants fonctionnels de la boîte TATA n'était établie.

3.4 Promoteur central chez *Arabidopsis thaliana*

Le génome d'*A. thaliana* a été le premier génome végétal dont la séquence a été publiée en 2000 (AGI, 2000). Son annotation est aujourd'hui une des plus accomplie. Comparé à d'autres génomes végétaux, le génome d'*A. thaliana* a une organisation plus compacte. Les cinq chromosomes qui le constituent ont une longueur approximative de 135 méga-bases et la dernière mise à jour du TAIR («The *Arabidopsis* Information Ressource»), ressource dédiée à cette plante modèle, fait état de 27379 gènes codant des protéines (Swarbreck *et al.*, 2008). Comparés aux autres eucaryotes, les promoteurs de plantes présentent des spécificités.

3.4.1 Caractéristiques spécifiquement observées chez les plantes

Deux caractéristiques des régions proches du TSS distinguent les génomes de plantes des autres génomes : le biais compositionnel en C et G ou GC-skew à l'emplacement du TSS et la présence de microsatellites tout particulièrement dans les UTR 5'.

a) GC-skew

Dans le génome de la plante *A. thaliana*, un biais a été identifié concernant la composition en G par rapport à celle en C dans la région du promoteur central (Tatarinova *et al.*, 2003). Ce biais appelé GC-skew est identifié en représentant la distribution (C-G) / (C+G) sur le brin transcrit. A taux de C et G égaux, cette représentation est linéaire et égale à 0. Chez *A. thaliana* et plus généralement les plantes, à l'approche du TSS, elle devient positive et révèle une plus grande présence en base C sur le brin transcrit (Figure 3-2). Dans l'UTR 5' le rapport (C-G) / (C+G) redescend progressivement (Tatarinova *et al.*, 2003; Fujimori *et al.*, 2005).

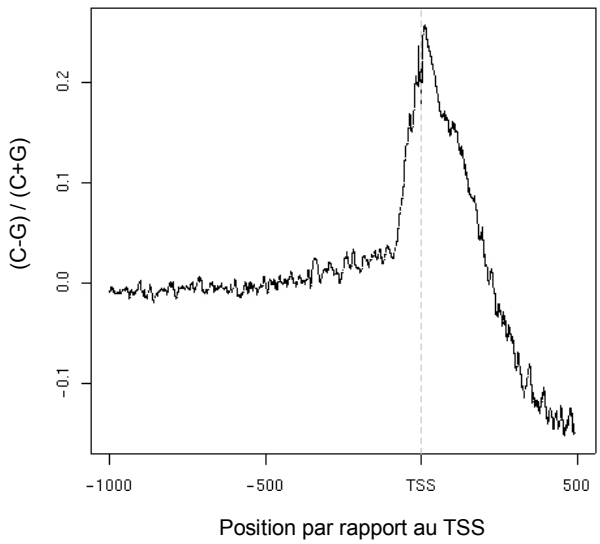


Figure 3-2 : GC-skew chez *A. thaliana*.

Représentation du GC-skew au sein du jeu de 14927 promoteurs, c'est-à-dire du rapport (C-G)/(C+G). Une fenêtre glissante de taille 10 a été utilisée. La composition en bases C et G a été étudiée sur le brin sens.

Le GC-skew a été mis en évidence chez *A. thaliana* et observé chez *O. sativa*, mais ni chez *H. sapiens* ou *D*. *melanogaster* ni chez *S. cerevisiae* (Fujimori *et al.*, 2005). Malgré une composition en bases très différente entre *A. thaliana* et *O. sativa*, la région [-50, 50] étant constituée de 37% de C+G chez *A. thaliana* et de 53% chez *O. sativa*, la valeur du GC-skew chez ces deux espèces est très similaire à l'emplacement du TSS. C'est la position la plus biaisée dans chacune des analyses (Fujimori *et al.*, 2005). Environ 70% des gènes d'*O. sativa* et d'*A. thaliana* sont caractérisés par une valeur de GC-skew supérieur à 0.33, soit un biais compositionnel important (Fujimori *et al.*, 2005).

Trois hypothèses ont été avancées pour tenter d'expliquer l'existence du GC-skew à cet emplacement :

- Lors de la transcription, des substitutions des bases C en T sur le brin antisens, soit G en A sur le brin sens, dues à la désamination des cytosines, pourraient créer un déséquilibre entre les bases C et G (Tatarinova *et al.*, 2003; Alexandrov *et al.*, 2006) ;
- La transversion du G en C pourrait être plus importante dans la région du TSS sur le brin sens (Fujimori *et al.*, 2005) ;
- Les TFBS orientés sur un brin préférentiel dans le promoteur central pourraient avoir une composition en bases plus riche en C qu'en G (Fujimori *et al.*, 2005).

Les promoteurs des gènes les plus fortement exprimés ont un GC-skew plus marqué (Fujimori *et al.*, 2005; Alexandrov *et al.*, 2006). Ce biais est en accord avec les deux premières hypothèses permettant d'expliquer l'existence du GC-skew : les gènes qui s'expriment plus fortement ont leurs promoteurs accessible à la machinerie de base de la transcription plus longtemps, et donc sont susceptibles de subir plus de substitutions donc plus de modifications (Alexandrov *et al.*, 2006). Aucun AT-skew n'est observé avec le même profil que le GC-skew, ce qui aurait dû être le cas si la première explication jouait un rôle majeur pour créer le biais (Fujimori *et al.*, 2005; Civan & Svec, 2009). Si le GC-skew était induit par les TFBS et leur orientation préférentielle, les gènes caractérisés par une expression plus forte devraient contenir plus d'éléments régulateurs orientés que les autres gènes. Ce biais n'a pas été démontré et reste une hypothèse à ce jour. C'est pourquoi, les hypothèses pourraient n'être valables que conjointement (Fujimori *et al.*, 2005; Civan & Svec, 2009).

Finalement, l'origine du GC-skew et son lien avec le niveau d'expression des gènes ne sont pas encore élucidés.

b) Microsatellites

Un microsatellite est une portion d'ADN de quelques bases qui est répétée. Le nombre de répétitions et donc la longueur du microsatellite peuvent être très variables entre organismes, entre individus, mais aussi entre allèles (Oliveira *et al.*, 2006). Différentes natures de microsatellites existent, dépendant de la séquence répétée, de la localisation dans le génome et de l'organisme qui la contient. Les microsatellites sont observés dans différents génomes eucaryotes (Toth *et al.*, 2000; Morgante *et al.*, 2002), avec certaines spécificités propres aux plantes.

Sans *a priori*, la recherche de répétitions parfaites et imparfaites de mono-, di- jusqu'à pentanucléotides a été réalisée dans différents génomes de plantes : *A. thaliana*, *O. sativa*, *Glycine max*, *Zea mays*, *Triticum aestivum* (Morgante *et al.*, 2002). Des différences quantitatives et qualitatives distinguent la présence de microsatellites dans les régions transcrites et non transcrites. Les premières sont plus riches en microsatellites riches en AG et CT, tandis qu'une plus forte présence en bases T et A est observée dans les régions non transcrites (Morgante *et al.*, 2002; Fujimori *et al.*, 2003).

Les plantes dicotylédones (*A. thaliana* et *G. max*) n'ont pas la même composition en microsatellites que les monocotylédones. *O. sativa* est l'organisme contenant le plus de répétitions de CGG et CCG ; *Z. mays* et *T. aestivum* en contiennent également, mais en moins grande quantité (Morgante *et al.*, 2002). Cette spécificité des monocotylédones pourrait s'expliquer par leur richesse en C+G différente des dicotylédones.

Néanmoins, contrairement aux mammifères, les plantes partagent des répétitions de AAG/CTT et de AG/CT dans les UTR 5' exclusivement (Fujimori *et al.*, 2003; Li *et al.*, 2004). Les microsatellites riches en bases C et T sont les plus observés. Les différentes répétitions possibles de séquences microsatellites sont en partie conservées entre orthologues. Ceci permet de supposer que leur présence au sein des génomes de plantes est ancienne (Zhang *et al.*, 2006).

De par leur nature répétée, les microsatellites sont des séquences qui sont particulièrement soumises à des mutations. Il n'est donc pas rare d'observer une purine au milieu d'une séquence de pyrimidines répétées, ou encore l'insertion d'une base T ou C au milieu d'une séquence TC ou TTC répétée parfaitement sur plusieurs dizaines de bases par exemple. C'est pourquoi quantifier les microsatellites dans un génome est une problématique complexe. Les analyses proposant de quantifier les microsatellites doivent prendre en compte l'existence de répétitions parfaites mais aussi non parfaites (Morgante *et al.*, 2002). Chez *A. thaliana*, en comptabilisant l'ensemble des microsatellites présents dans les différentes régions du génome, 0.85% des chromosomes sont constitués de ces séquences répétées (Morgante *et al.*, 2002) avec des différences majeures entre régions.

Le rôle des microsatellites particulièrement présents dans les UTR 5' de plantes n'est aujourd'hui pas réellement connu. Différentes hypothèses ont été proposées. A l'échelle de quelques gènes et lors de l'étude de la répétition de CTT et de GAA, une corrélation positive entre la longueur des répétitions des microsatellites et le niveau d'expression des gènes a été identifiée chez *A. thaliana* (Zhang *et al.*, 2006). La nature et longueur des microsatellites influenceraient l'expression des gènes. Une autre hypothèse propose que la présence de polypurines ou de polypyrimidines joue un rôle sur la conformation de l'ADN et donc sur la transcription (Fujimori *et al.*, 2003). Des répétitions de TC, TTC, GA ou GAA, sont en effet susceptibles d'induire des courbures ou une tendance à la rigidité (Bolshoy *et al.*, 1991). Une modification de conformation pourrait avoir des conséquences sur le positionnement des nucléosomes (Ioshikhes *et al.*, 1999; Kiyama & Trifonov, 2002), l'accès à l'ADN et donc la transcription des gènes.

3.4.2 Eléments régulateurs du promoteur central

a) Elément initiateur

Le motif consensus de l'Inr identifié chez *D. melanogaster* est YYANWYY (Javahery *et al.*, 1994). Il est fonctionnel chez les plantes (Nakamura *et al.*, 2002) mais peu observé chez *A. thaliana* (Yamamoto *et al.*, 2007). L'hypothèse de l'existence d'un motif consensus Inr spécifique aux plantes a donc été considérée par différentes études dont certaines ont été publiées au cours de cette thèse.

Une étude des dinucléotides dans la région du TSS a montré une forte présence de CA mais aussi de TA, TG et CG positionnés une base en amont du TSS (Yamamoto *et al.*, 2007). Les dinucléotides TG et CG sont nouveaux par rapport au motif consensus identifié chez les mammifères (Javahery *et al.*, 1994). Ils sont néanmoins en accord avec le dinucléotide initiateur pyrimidine-purine proposé chez les mammifères (Carninci, 2006), à la différence près que CG est minoritaire chez *A. thaliana*. La séquence consensus observée chez les plantes est YR. Elle est présente dans 77% des promoteurs chez *A. thaliana* et dans près de 60% des promoteurs chez *O. sativa* (Yamamoto *et al.*, 2007).

Cette séquence YR est très courte et a été appelée le «dinucléotide initiateur» par d'autres auteurs (Bajic *et al.*, 2006). Néanmoins, par abus de langage, dans l'ensemble du document il sera nommé «Inr-YR».

Des études très récentes ont conduit à définir une séquence d'Inr plus longue. La boîte TATA se retrouve significativement associée aux gènes contenant le motif YTCAY (Yamamoto *et al.*, 2009).

b) Vers une séquence consensus de la boîte TATA

En se basant sur les critères définissant une boîte TATA proposés sur le site de l'EPD («The Eukaryotic Promoter Database»), une base de données de promoteurs d'eucaryotes, le motif consensus TATA(A/T)A(T/A)A a été proposé chez les plantes en construisant une matrice résumée dans la Table 3-5 A (Shahmuradov *et al.*, 2003). Parmi les 305 promoteurs de plantes analysés lors de cette étude, plus de 56% contenaient une boîte TATA.

Plus récemment, une analyse du promoteur central des gènes d'*A. thaliana* a permis de proposer une matrice plus robuste (Table 3-5 B) établie avec près de 13000 séquences ayant un TSS dont la position est déterminée en biologie expérimentale (Molina & Grotewold, 2005). Cette matrice considère un plus large environnement de la boîte et propose une extension du motif consensus pour obtenir des prédictions plus spécifiques des gènes contenant la boîte TATA. En étudiant l'emplacement des motifs, la position préférentielle proposée par cette étude est 32 bases en amont du TSS, position en accord avec la position observée chez les eucaryotes (Ponjavic *et al.*, 2006). Le pourcentage alors prédit de promoteurs contenant une boîte TATA est de 29%.

Ces deux exemples d'analyses montrent l'apport d'une analyse globale pour prédire avec plus de pertinence un motif consensus robuste et une liste de gènes contenant un motif d'intérêt. La meilleure connaissance de la position du TSS chez *A. thaliana* a fortement contribué à améliorer ces analyses (Carninci *et al.*, 1996; Castelli *et al.*, 2004; Clepet *et al.*, 2004).

A

Position	-2	-1	1	2	3	4	5	6	7	8	9	10
%A	28	16	3	95	0	100	62	97	38	73	13	30
%C	27	63	1	0	4	0	0	0	1	8	42	42
%G	17	5	0	0	0	0	0	2	0	10	28	16
%T	28	16	96	5	96	0	38	1	61	9	18	11
consensus		c	T	A	T	A	A/T	A	T/A	A		

B

Position	-4	-3	-2	-1	1	2	3	4	5	6	7	8	9	10	11	12
%A	23	26	24	25	0	100	0	99	41	99	36	91	24	44	30	39
%C	24	26	23	40	0	0	0	0	0	0	0	0	29	23	27	20
%G	13	18	11	15	0	0	0	0	0	0	0	9	19	15	16	16
%T	40	30	41	20	99	0	100	0	59	0	64	0	27	18	27	24
consensus	t	n	t	c	T	A	T	A	T/A	A	T/A	A	n	a	n	a

Table 3-5 : Matrices de fréquences de nucléotides de la boîte TATA.

Pour l'établissement de la matrice **A**, parmi 305 promoteurs de plantes, ceux contenant une boîte TATA ont été étudiés (Shahmuradov *et al.*, 2003). Pour la matrice **B**, parmi 12749 promoteurs d'*A. thaliana*, ceux contenant une boîte TATA ont été étudiés. Pour les 2 matrices, les pourcentages en rouge soulignent les bases servant à former un motif consensus.

En conclusion, les connaissances de l'organisation du promoteur central chez les plantes sont encore limitées et ne permettent pas comme chez les mammifères d'avoir une représentation biologiquement validée de l'architecture des éléments régulateurs. Des recherches d'éléments régulateurs dans des jeux de gènes ne contenant pas la boîte TATA ont permis d'identifier de nouveaux TFBS chez les mammifères (Burke & Kadonaga, 1996; Ohler *et al.*, 2002) tandis que de telles analyses n'ont pas conduit à de nouvelles prédictions chez les plantes (Molina & Grotewold, 2005). Deux éléments sont conservés et observés chez les plantes comme chez les mammifères : la boîte TATA et l'Inr. Les séquences consensus de ces éléments peuvent néanmoins être divergentes entre ces deux classes taxonomiques. De plus, des travaux récents ont présenté des éléments, décrits par d'autres comme des microsatellites, en tant qu'élément régulateur du promoteur central : le Y-patch (Yamamoto *et al.*, 2007; Civan & Svec, 2009).

Tandis qu'au début de cette thèse, les connaissances du promoteur central chez *A. thaliana* étaient réduites à l'existence de la boîte TATA, différents travaux ont contribué à l'obtention d'informations plus complètes (Molina & Grotewold, 2005; Yamamoto *et al.*, 2007; Yamamoto *et al.*, 2007; Civan & Svec, 2009; Yamamoto *et al.*, 2009). Le projet de cette thèse a été mené dans un climat très concurrentiel et des travaux présentant une méthodologie et des résultats similaires aux nôtres ont été publiés au cours de ces trois années.

4 Projet de thèse

4.1.1 Contexte

L'analyse des promoteurs d'*A. thaliana* a été réalisée dans l'équipe de Bioinformatique de l'Unité de Recherche en Génomique Végétale, l'URGV. Cette équipe ainsi que les équipes de biologie expérimentale de l'URGV ont permis l'obtention de données essentielles au projet qui ont complété des données publiques. Par exemple, l'équipe Génomique Fonctionnelle des Plantes Cultivées a contribué à générer des séquences d'ADNc pleine longueur chez *A. thaliana* qui ont été utilisées pour définir la position du TSS (Clepet *et al.*, 2004). De plus, l'équipe Génomique Fonctionnelle chez *A. thaliana* gère une plateforme transcriptome avec notamment des puces CATMA («Complete *Arabidopsis* Transcriptome Micro-Array») dont les résultats ont été analysés lors des études fonctionnelles. L'objectif du projet CATMA était de concevoir une puce reposant sur des sondes de haute qualité *via* l'exploitation de séquences spécifiques des gènes, les GST (Gene-Specific Tag) couvrant la majorité des gènes d'*A. thaliana* (Sclep *et al.*, 2007). L'URGV a utilisé des GST de 150 à 500 paires de bases (bp) conçues par le logiciel SPADS («Specific Primers & Amplicons Design Software») (Thareau *et al.*, 2003) pour produire ces puces. Les gènes considérés lors de la recherche de GST proviennent du TIGR («The Institute for Genomic Research») et des prédictions du logiciel EuGene (Schiex *et al.*, 2001). La version v3 de CATMA contient 30343 sondes représentant les 26751 gènes codants des protéines prédits chez *A. thaliana*. Les puces CATMA sont aussi efficaces que d'autres puces pour réaliser des analyses de transcriptomes (Allemeersch *et al.*, 2005). Enfin, des collaborations entre les équipes de biologie expérimentale et l'équipe de Bioinformatique sont fréquentes à l'URGV et permettent de disposer de jeux de gènes basés sur les expérimentations qui sont pertinents à analyser *in silico*. Dans le cadre de ce projet, une collaboration avec les membres de l'équipe MAP-kinases («Mitogen Activated Proteins») a notamment été possible.

L'équipe de Bioinformatique de l'URGV est particulièrement impliquée dans des projets concernant *A. thaliana* (Aubourg *et al.*, 2000; Brunaud *et al.*, 2002; Lurin *et al.*, 2004; Aubourg *et al.*, 2005; Rivals *et al.*, 2006; Aubourg *et al.*, 2007; Armisén *et al.*, 2008; Deveaux *et al.*, 2008). Elle a développé deux bases de données exploitées intensément au cours de ce projet :

- FLAGdb^{++} (Samson *et al.*, 2004) propose un accès aux annotations structurales de génomes des plantes modèles *A. thaliana*, *O. sativa*, *P. trichocarpa* et *V. vinifera*. Les annotations du génome d'*A. thaliana* proviennent de différentes sources dont le TAIR (Swarbreck *et al.*, 2008), le TIGR, et de l'utilisation de différents outils publics comme EuGene (Schiex *et al.*, 2001) ou internes à l'équipe de Bioinformatique. Par exemple, tous les transcrits disponibles pour un organisme donné ont été utilisés pour définir au mieux les extrémités des ARN et positionner les TSS des gènes.
- CATdb (Gagnot *et al.*, 2008) propose un accès à une large collection de données relatives au transcriptome d'*A. thaliana* produites par la plateforme CATMA de l'URGV, puis analysées et stockées par l'équipe de Bioinformatique. Aujourd'hui, 39 publications exploitent les données disponibles dans cette base. CATdb contient plus de 6900 hybridations distribuées dans près de 300 expériences qui reflètent une grande diversité de conditions expérimentales.

Le projet de thèse qui était proposé devait générer des données pour compléter l'annotation des régions promotrices débutée dans FLAGdb^{++} et permettre de savoir s'il est possible d'exploiter l'ensemble des données structurales et fonctionnelles afin de contribuer à l'annotation des éléments régulateurs.

4.1.2 Objectifs et hypothèses

Le projet de thèse présenté dans ce document concernait (i) l'identification exhaustive des éléments régulateurs du promoteur central et proximal de la plante modèle *A. thaliana* en exploitant leurs contraintes topologiques et (ii) la caractérisation structurale et fonctionnelle des gènes qui les contiennent. L'objectif était de contribuer à l'annotation de la fonction d'éléments régulateurs candidats *via* l'identification de caractéristiques fonctionnelles communes des gènes.

Nous proposions de développer une stratégie de recherche pour identifier des motifs étant de potentiels éléments régulateurs dans les promoteurs de la plante modèle *A. thaliana*. Notre idée était de combiner une approche globale à une approche par sous-groupes de gènes pour prédire l'existence d'éléments régulateurs généraux comme d'éléments régulateurs spécifiques.

La première hypothèse de travail justifiant cette méthodologie était que des motifs sur-représentés dans une région précise des promoteurs seraient des éléments régulateurs de l'expression des gènes, reconnus par des facteurs de transcription ou impliqués dans la conformation de l'ADN. La deuxième hypothèse était que ces éléments régulateurs potentiels qui sont caractérisés par des contraintes positionnelles auraient plus de chance d'être fonctionnels dans la région précise où ils sont sur-représentés.

Nous avons souhaité développer cette approche afin de définir une liste de gènes contenant chaque élément régulateur candidat, l'ensemble des résultats pouvant conduire à une cartographie des promoteurs d'*A. thaliana*.

Dans un deuxième temps, nous avons souhaité analyser des groupes de gènes contenant une même architecture de leurs promoteurs. L'hypothèse était que l'identification d'annotations fonctionnelles communes de ces groupes de gènes pourrait contribuer à l'annotation fonctionnelle des éléments régulateurs.

En complément, au regard des faibles connaissances concernant l'organisation du promoteur central chez les plantes, nous avons décidé de nous orienter vers une analyse approfondie de cette région. En particulier un objectif était d'identifier s'il existait d'autres éléments régulateurs à l'emplacement de la boîte TATA et de les caractériser. L'hypothèse de ce travail était que des éléments impliqués dans l'initiation de la transcription des gènes chez les plantes pourraient être découverts en exploitant l'approche par l'analyse de sous-groupes de promoteurs.

Résultats et discussions

5 Identification de courtes séquences d'ADN caractérisées par des contraintes topologiques

L'approche mise au point au cours de cette thèse permet d'identifier les courtes séquences d'ADN, appelées par la suite des «motifs», qui sont sur-représentées dans une région précise au sein des promoteurs. Dans les parties suivantes, l'approche développée au cours de cette thèse est présentée afin de comprendre comment de tels motifs avec des contraintes de position sont identifiés. Ces motifs sont appelés par la suite les PLM («Preferentially Located Motifs»), et l'approche qui permet de les identifier, l'approche PLM.

5.1 Jeu de promoteurs

Dans le cadre de ce travail, un promoteur comprend l'ensemble de l'UTR 5' et 1kb en amont du TSS de la séquence codante localisée en aval. Pour identifier les motifs caractérisés par une position préférentielle au sein des promoteurs, tous sont alignés selon le TSS de leur gène, cette position étant essentielle pour utiliser l'approche PLM.

5.1.1 Construction du jeu de promoteurs d'*A. thaliana*

Nous avons exploité les transcrits disponibles dans la base de données FLAGdb^{++} (Samson *et al.*, 2004) pour déterminer la position du TSS.

a) Identification du début de transcription

Les annotations structurales d'*A. thaliana* disponibles dans FLAGdb^{++} et issues du TAIR R.6 (Swarbreck *et al.*, 2008) permettent de définir les positions des gènes et des ARNm. Ces informations ont été complétées grâce au travail de l'équipe de Bioinformatique qui a permis d'étendre l'unité de transcription en utilisant les transcrits disponibles, c'est-à-dire les EST et les ADNc pleine longueur s'ils sont disponibles comme illustré *via* un exemple Figure 5-1.

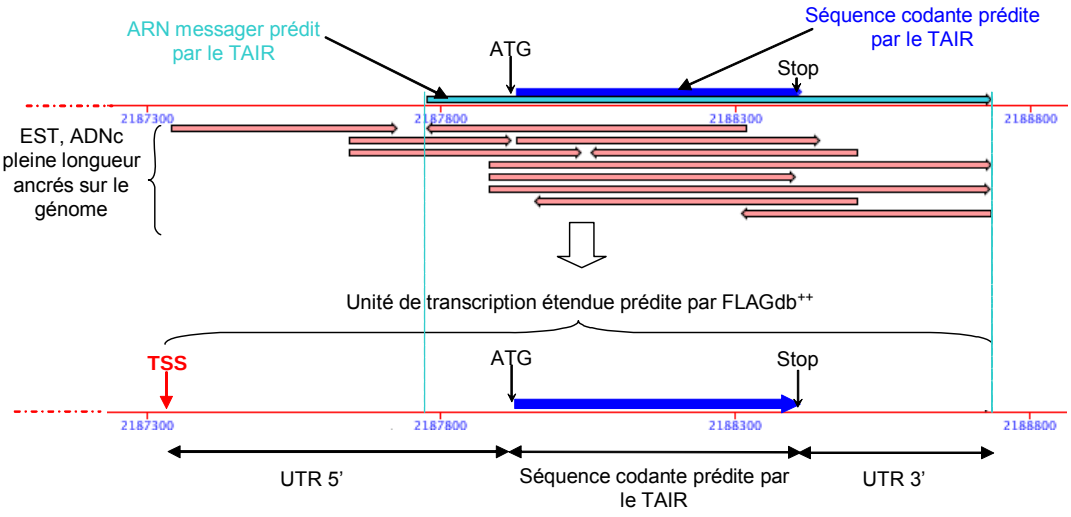


Figure 5-1 : Extension de l'unité de transcription du gène AT3G06890.

Image issue de FLAGdb^{++}. Les transcrits associés à un gène sont exploités pour définir l'unité de transcription. Les transcrits sont associés à un gène lorsqu'ils le chevauchent ou lorsqu'ils chevauchent un transcrit lui-même chevauchant le gène. Cette extension de l'unité de transcription permet de définir plus précisément la position du TSS le plus en amont de la séquence codante.

Ce travail permet dans 33% des cas d'obtenir une unité de transcription en 5' plus grande que celle prédite par l'ARNm du TAIR (Figure 5-2). Ainsi, 8798 gènes ont un TSS prédit qui est plus en amont que le TSS prédit par le TAIR. Les TSS prédits par le TAIR sont susceptibles de montrer soit la présence de TSS alternatifs soit l'existence de mauvaises définitions de la position du TSS, dues à des rétrotranscriptions non complètes par exemple.

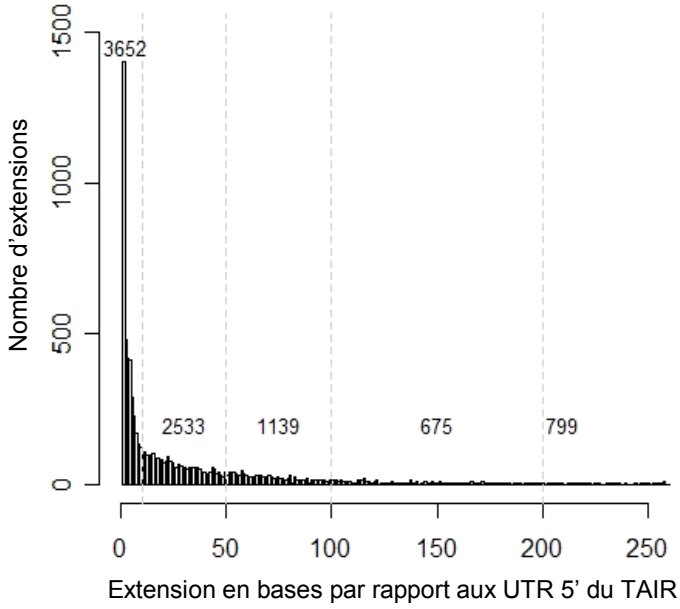


Figure 5-2 : Extension des unités de transcription en 5' par rapport aux annotations du TAIR.

Histogramme de la différence en nombre de bases entre les deux prédictions et nombre d'extensions correspondant aux différentes classes délimitées par les traits en pointillés gris. Par rapport aux prédictions des ARNm du TAIR, 8798 unités de transcriptions ont été étendues dans FLAGdb^{++}.

L'histogramme de la taille des UTR 5' des gènes codant des protéines met en évidence deux phases : entre les longueurs de 1 à 50 bases, une augmentation progressive de la fréquence de promoteurs puis après les longueurs de 50 bases une diminution rapide (Figure 5-3). C'est pourquoi le seuil de 50 bases a été conservé comme valeur minimale pour identifier un UTR 5' non tronqué. Ce critère strict a été choisi afin de rendre le jeu de promoteurs étudié le plus fiable possible. Parmi les 26751 gènes codant des protéines prédits chez *A. thaliana*, 18451 ont un UTR de longueur supérieure ou égal à 1 base et 14927 ont un UTR considéré comme vraisemblable, de taille supérieure à 50.

Les TSS alternatifs ne sont pas considérés par l'approche PLM, bien qu'ils aient été très récemment décrits chez *A. thaliana* (Tanaka *et al.*, 2009). Utiliser l'ensemble des TSS alternatifs pour extraire le jeu de promoteurs reviendrait à lisser une potentielle sur-représentation locale d'un motif. Il a donc été choisi de ne considérer qu'un seul TSS par gène pour cette approche qui recherche des biais de position préférentielle. Des études récentes ont confirmé ce choix. Chez les plantes, l'analyse de Tanaka *et al.* (2009) a permis de distinguer les deux types d'organisation des TSS ayant des régions promotrices différentes. La composition en bases dans les promoteurs est atypique lorsque les TSS les plus proches de la région codante sont considérées. De plus, les TFBS connus sont

présents tout particulièrement dans les promoteurs des TSS les plus en amont. Ces TSS sont conservés aux mêmes emplacements entre gènes orthologues d'*A. thaliana* et d'*O. sativa*. Les autres TSS ne sont pas conservés. Pour ces raisons, le TSS considéré dans ce travail est le TSS le plus en amont du codon initiateur, c'est-à-dire le plus éloigné de la séquence codante.

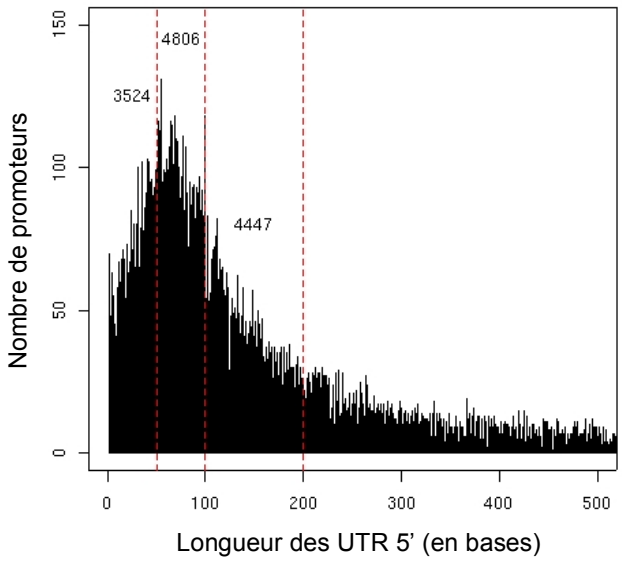


Figure 5-3 : Histogramme des longueurs des UTR 5'.

Nombre de promoteurs dans les différentes classes de longueur d'UTR 5' délimitées par les traits en pointillés rouges. Au total, l'UTR 5' de 18451 gènes a été défini.

b) Extraction du jeu de promoteurs

Les UTR sont des régions où des éléments régulateurs ont été identifiés (Kutach & Kadonaga, 2000; Graber, 2003; Lim *et al.*, 2004). Nous avons donc conservé ces régions lors de l'extraction du jeu de promoteurs. De plus, les éléments régulateurs ayant des contraintes positionnelles sont localisés dans quelques centaines de bases en amont du TSS (Higo *et al.*, 1999). Ce n'est que dans les 500 bases en amont du TSS que les éléments régulateurs sont strictement conservés aux mêmes emplacements entre *A. thaliana* et *Raphanus sativus* (Armisen, 2008). C'est pourquoi pour les 14927 gènes dont la position du TSS est vraisemblablement correcte, 1000 bases en amont du TSS et l'ensemble de l'UTR 5' sont extraits (Figure 5-4). Ces séquences sont considérées par la suite comme les promoteurs. Il est à noter que tous n'ont donc pas la même longueur, les plus courts étant de 1050 bases, les plus longs de plus de 1500 bases. Pour les études présentées dans ce document, ces 14927 promoteurs ont été analysés.

Au cours de ma thèse, différentes améliorations de l'annotation du génome d'*A. thaliana* ont entraîné des mises à jour de FLAGdb^{++} et donc de mon jeu de promoteurs. Les résultats présentés dans la partie concernant l'analyse globale du génome ont été générés avec un jeu plus ancien que les résultats présentés dans la partie concernant l'étude de sous-groupes de promoteurs. Pour cette raison, certaines caractéristiques décrivant les PLM peuvent être légèrement différentes.

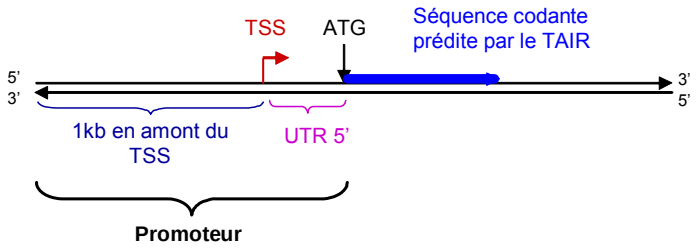


Figure 5-4 : Extraction des promoteurs d'*A. thaliana*.

Pour chaque gène dont le TSS est prédit, 1000 bases en amont du TSS ainsi que l'ensemble de l'UTR 5' sont extraits et constituent un promoteur. Le seuil de 50 bases est utilisé pour sélectionner les UTR 5' considérés comme fiables.

5.1.2 Validation des positions du TSS dans le jeu de 14927 promoteurs

Depuis son séquençage en 2000 (AGI, 2000), des caractéristiques spécifiques autour des gènes ont été identifiées dans la séquence du génome d'*A. thaliana*. La conservation de ces biais de séquences a été recherchée dans le jeu complet de promoteurs en première validation de sa fiabilité.

a) Etude du GC-skew

Le GC-skew est un biais compositionnel en bases C et G observé chez les plantes. Il a été retrouvé en étudiant les 14927 promoteurs de notre jeu d'étude (Figure 5-5). Sans lissage de la représentation, la position à laquelle le biais est maximal est précisément observée une base en amont du TSS, la position du TSS ayant un GC-skew négatif.

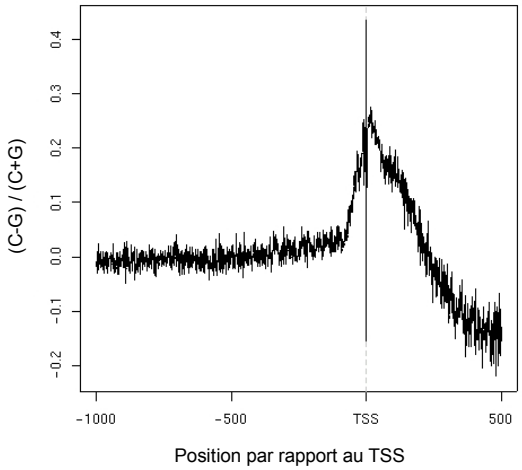


Figure 5-5 : Etude du biais compositionnel en bases C et G dans la région du TSS chez *A. thaliana*.

Au sein du jeu de 14927 promoteurs d'*A. thaliana*, sur le brin sens, représentation du GC-skew, c'est-à-dire du rapport (C-G)/(C+G). Une fenêtre glissante de une base de longueur a été utilisée.

b) Composition en nucléotides dans la région du TSS

Tandis que la richesse en bases A, C, G et T est relativement constante dans les promoteurs, dans les quelques bases en amont et en aval du TSS, des fluctuations ont été observées (Alexandrov *et al.*, 2006). En -1 et à l'emplacement du TSS, la richesse respective en bases C puis A révèle la présence de l'Inr (Yamamoto *et al.*, 2007). L'étude des 14927 promoteurs révèle les mêmes biais qualitatifs de composition en bases en -1 et au TSS (Figure 5-6).

Néanmoins, des différences quantitatives distinguent les deux observations. Elles peuvent s'expliquer par une différence entre les jeux de promoteurs due aux approches utilisées pour positionner le TSS. Les TSS exploités par Alexandrov *et al.* (2006) proviennent de jeux d'ADNc pleine longueur du Ceres et du RIKEN. La méthodologie exploitée identifie les extrémités des ADNc pleine longueur en les marquant par l'ajout d'un groupe biotine qui sera reconnu par des billes magnétiques recouvertes de streptavidine (Carninci *et al.*, 1996). Les fragments d'ADNc sont clonés dans le vecteur Lambda Zap II après qu'une queue d'oligo (dG/dC) ait été ajoutée à leur séquence. Cette méthode intitulée «*biotinylated CAP trapper*» exploite cette queue poly C pour définir le début de l'ADNc pleine longueur. La première base n'étant pas un C sera considérée comme le début de la séquence. Si une ou plusieurs cytosines sont les premières bases de l'ADNc pleine longueur, la prédiction proposée par l'approche de Carcini *et al.* (1996) entraînera un décalage d'autant de bases qu'il y a de cytosines consécutives. Ces décalages auront pour conséquences une position du TSS erronée.

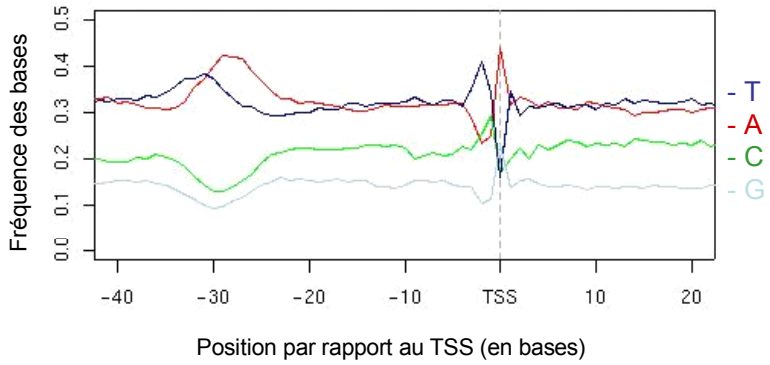


Figure 5-6 : Distribution des bases A, C, G et T dans la région du TSS chez *A. thaliana*.
Représentation au sein de la région [-40, +20] chez *A. thaliana* lors de l'étude des 14927 promoteurs.

La construction des 14927 promoteurs analysés pendant cette thèse n'exploite pas exclusivement les ADNc pleine longueur provenant de cette méthode «*biotinylated* CAP-trapper». Pour définir la position du TSS, nous utilisons également des EST et des ADNc pleine longueur obtenus en exploitant la fixation d'un oligonucléotide à l'extrémité 5' d'un ARN grâce à la T4 DNA ligase (Clepet *et al.*, 2004). Finalement, environ la moitié des 14927 promoteurs ont un TSS positionné par une approche ADNc pleine longueur, les autres étant complétés par les autres transcrits disponibles. L'utilisation conjointe de différentes sources de transcrit permet d'obtenir une position plus fiable du TSS.

En conclusion, les biais de séquences attendus dans les promoteurs d'*A. thaliana* sont retrouvés dans le jeu de 14927 promoteurs et permettent de valider notre jeu de promoteurs.

5.2 Identification des PLM

L'approche PLM globale pour identifier des PLM est schématisée Figure 5-7 et détaillée par la suite.

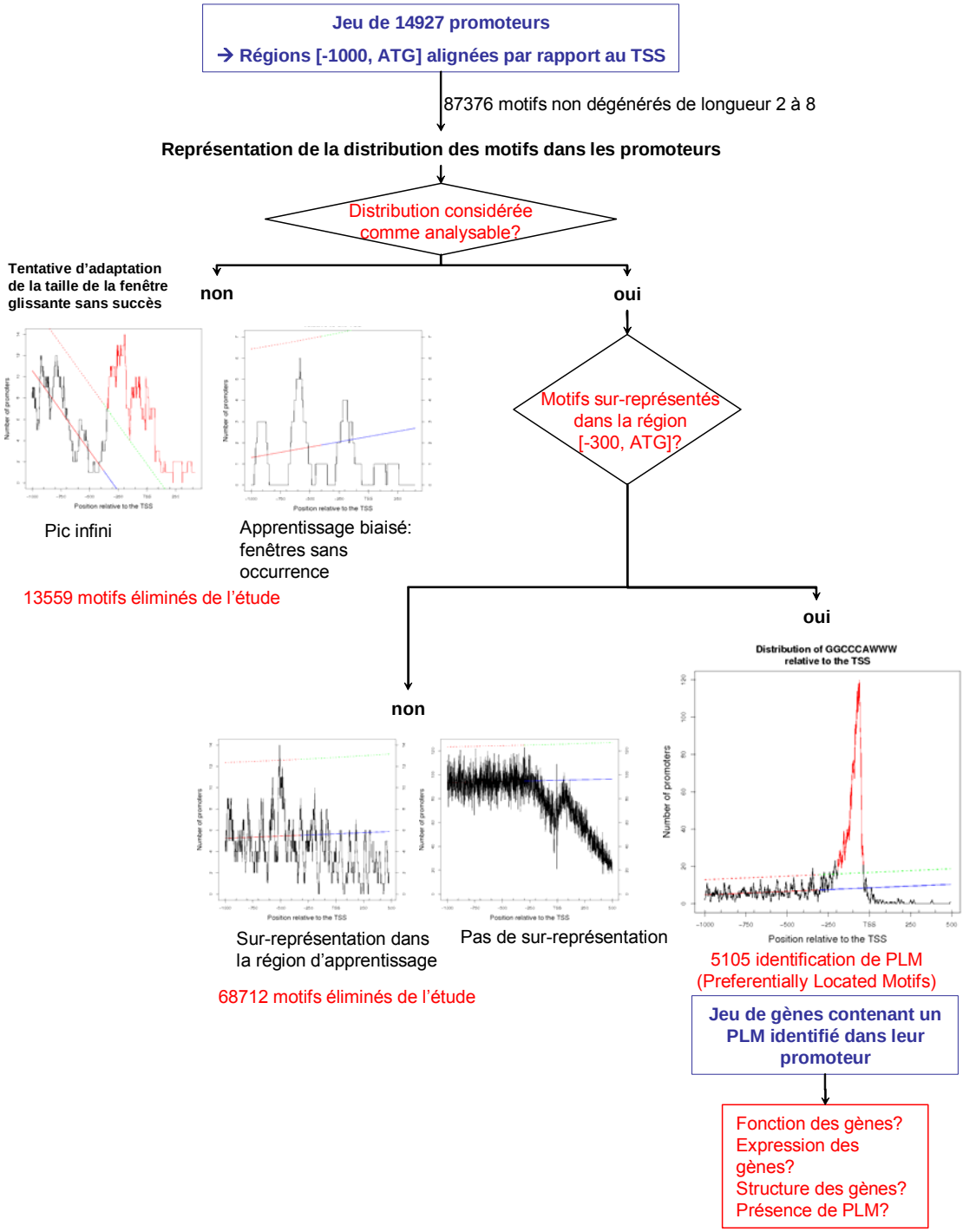


Figure 5-7 : Organigramme de l'approche utilisée pour identifier les PLM automatiquement.

5.2.1 Construction de la distribution des motifs

Afin d'identifier des motifs caractérisés par une position préférentielle au sein des promoteurs, ceux-ci sont alignés par rapport au TSS. Pour chaque motif considéré, toutes ses occurrences sont recherchées dans les 14927 promoteurs. La position d'un motif est définie par la première position de sa séquence. Puis les promoteurs alignés sur leur TSS sont parcourus par une fenêtre glissante qui se déplace de la position -1000 jusqu'au codon initiateur. Cette fenêtre a initialement une taille de 1 base puis cette taille est adaptée en fonction des contraintes de chaque distribution comme discuté page 69. Le nombre de séquences promotrices possédant le motif étudié est compté dans chaque fenêtre glissante, et non pas le nombre de motifs pour ne pas surestimer les motifs répétés. La représentation graphique de la distribution d'un motif candidat par rapport au TSS est générée comme illustré Figure 5-8.

Il faut noter dans l'ensemble de ce manuscrit de thèse, les distributions sont représentées en effectuant la recherche d'un mot exclusivement sur le brin sens.

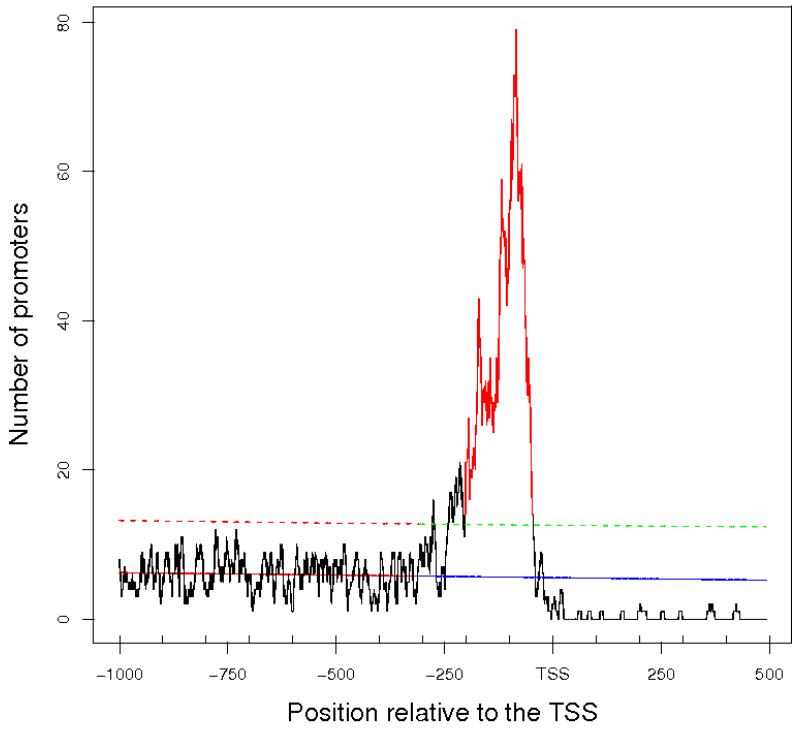


Figure 5-8 : Distribution de ATGGGCC dans les 14927 promoteurs alignés par rapport au TSS.

Pour chaque distribution d'un motif, l'axe des abscisses représente les positions -1000 à 500 par rapport au TSS. L'axe des ordonnées représente le nombre de promoteurs contenant le motif étudié à une position donnée. Le modèle de distribution est appris dans la région d'apprentissage [-1000, -300] où la droite moyenne (droite continue) et la borne supérieure de l'intervalle de confiance (droite hachurée) sont estimées (en rouge) et étendues à la région [-299, 500] (bleu et vert).

5.2.2 Identification automatique des motifs ayant des contraintes positionnelles

a) Apprentissage du modèle de distribution

Pour identifier de manière automatique la distribution non linéaire des motifs positionnés à une distance préférentielle du TSS, un apprentissage du modèle de distribution est réalisé pour chaque motif étudié. La région [-1000, -300] est utilisée en tant que région de référence pour réaliser cet apprentissage car (i) les éléments régulateurs soumis à des contraintes topologiques sont majoritairement attendus dans le promoteur proximal et (ii) parmi les 87376 motifs de longueur 2 à 8, seuls 69 présentent des distributions non linéaires dans la région d'apprentissage [-1000, -300] (Cf. chapitre «Identification des distributions non unifomes» ci-dessous). Ainsi, dans cette région d'apprentissage, la droite moyenne de forme $y = bx + a$[1] reflétant le bruit de fond de présence d'un motif est estimée (Figure 5-8, trait plein), avec le nombre de séquences «y», la position par rapport au TSS «x», le coefficient de régression «a» et la constante «b». Puis un intervalle de confiance à 99% est construit (Figure 5-8, trait discontinu). La droite et l'intervalle de confiance obtenus sont identifiés par une régression linéaire *via* la fonction «predict.lm» du logiciel R (Dessau & Pipper, 2008). Cette fonction permet de prendre en compte l'éloignement de la région d'apprentissage. En effet, l'extension d'un modèle de distribution appris dans une région est plus susceptible d'être efficace proche de la région d'apprentissage qu'à 1kb de cette région. C'est pourquoi cette fonction permet d'être de plus en plus exigeant en s'éloignant de la région [-1000, -300] pour identifier les biais et donc les PLM.

b) Identification des distributions mettant en évidence un motif sur-représenté localement

Pour une distribution donnée, lorsque la borne supérieure de l'intervalle de confiance est franchie, un motif ayant une contrainte positionnelle est identifié. Un score est attribué à chaque distribution et donc à chaque motif. Un classement peut donc distinguer les PLM en fonction de leurs contraintes topologiques plus ou moins fortes. Un score appelé SMS («Score of Maximal Square relative to the base line»), a été défini comme illustré Figure 5-9 A. Il correspond au rapport :

$$\frac{\text{(hauteur du pic - hauteur de la droite moyenne)}}{\text{(hauteur de la borne supérieure de l'intervalle de confiance - hauteur de la droite moyenne)}}$$

L'ensemble des SMS supérieurs à 1 concerne les distributions non linéaires et donc des PLM. Néanmoins, pour avoir un jeu de motifs ayant des contraintes topologiques plus strictes, seuls les SMS supérieurs à 3 sont considérés pour l'analyse globale des motifs présentée par la suite (Figure 5-9 B et Figure 5-10).

c) Délimitation de la fenêtre fonctionnelle d'un motif

[1] Annotation anglophone de $y = ax + b$

Lorsqu'un PLM est obtenu, les bornes du pic doivent être définies. Pour cela, la position préférentielle du motif est exploitée puis la première position en amont (et en aval) du motif qui est en dessous de l'intervalle de confiance est extraite. Ainsi, une fenêtre fonctionnelle est identifiée. Elle correspond aux positions où le PLM est sur-représenté et donc ses bornes sont celles où ce PLM est probablement sous pression de sélection et donc potentiellement fonctionnel.

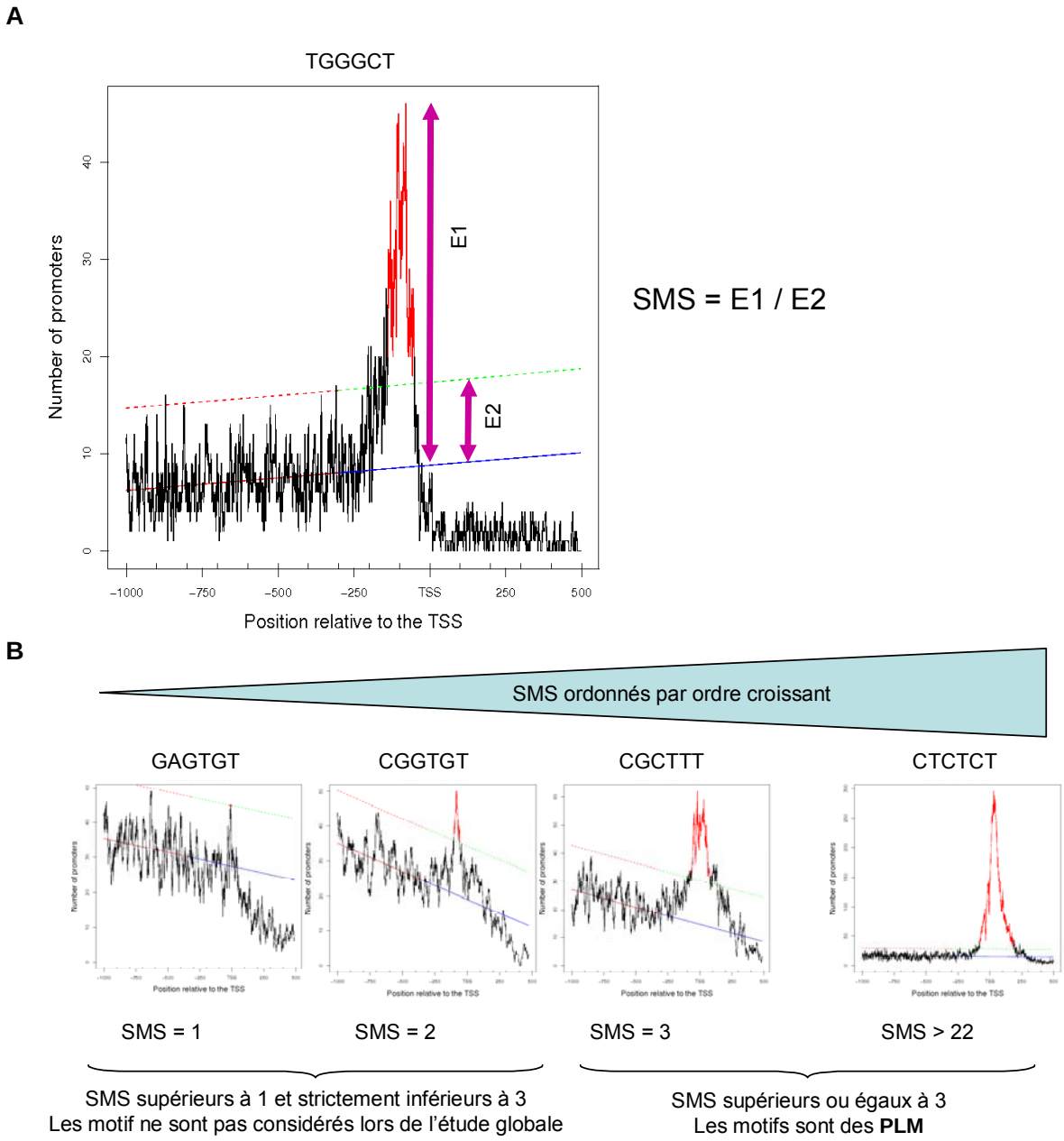


Figure 5-9 : Attribution du score à un motif et sélection des PLM.

(A) Le SMS ou Score de l'écart maximal à la moyenne (Score of Maximal Square relative to the base line) correspond au rapport E1 / E2. La fenêtre fonctionnelle d'un PLM est la région du promoteur comprise entre les bornes du pic rouge. **(B)** Le SMS permet de trier les PLM en fonction de leurs contraintes topologiques.

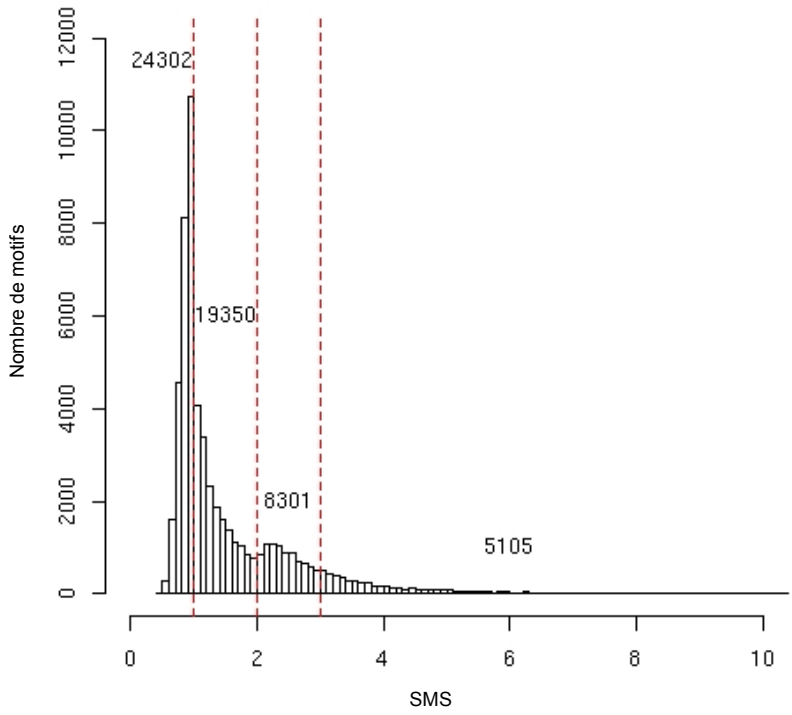


Figure 5-10 : Histogramme des SMS obtenus lors de l'analyse des motifs de longueur 2 à 8.
Nombre de motifs dans les différentes classes de SMS délimitées par les traits en pointillés rouges.

En complément, l'orientation des motifs a été étudiée afin de caractériser les PLM. Les sites de fixation peuvent être sensibles à l'orientation, et être significativement plus observés sur un des deux brins d'ADN ou être non sensibles et observés équitablement sur les deux brins. Afin de quantifier les motifs sensibles ou non à l'orientation, la présence de couples «motif + complément inverse» est recherchée. Par définition, un motif palindromique est insensible à l'orientation. Pour l'ensemble des motifs non palindromiques, différentes situations ont été considérées. Le complément inverse (CI) d'un motif M peut ne pas être un PLM. Le motif M est alors sensible à l'orientation. Si CI est un PLM, il peut être caractérisé par une fenêtre fonctionnelle différente de celle de M. Dans ce cas M et son CI sont tous les deux sensibles à l'orientation. Enfin, si M et son CI partagent une même fenêtre fonctionnelle chevauchante, le nombre d'occurrences des motifs sur les deux brins est comparé. Une p-value significative (Test exact de Fisher - p-valeur inférieure à 5e-2) indique que les motifs sont sensibles à l'orientation.

Pour chaque motif correspondant à un PLM, les caractéristiques qui le décrivent sont :

- La représentation graphique de sa distribution par rapport au TSS ;
- Le score ou SMS ;
- La position préférentielle, c'est-à-dire la position du sommet du pic (PP) ;
- Les bornes préférentielles délimitant la fenêtre fonctionnelle du motif (FF) ;
- L'orientation préférentielle ou non des PLM comme définie ci-dessus ;

- La liste des gènes qui possèdent dans leurs promoteurs le PLM dans sa fenêtre fonctionnelle ainsi que leur pourcentage par rapport aux gènes étudiés.

d) Adaptation de la taille de la fenêtre glissante

La position du TSS est définie expérimentalement pour de plus en plus de gènes. Les approches exploitant une recherche de motifs ayant une position préférentielle par rapport à ce site sont donc de plus en plus développées (Kielbasa *et al.*, 2001 ; Berendzen *et al.*, 2006 ; Bernard *et al.*, 2006 ; Bellora *et al.*, 2007 ; Yamamoto *et al.*, 2007). Chacune étudie la distribution de motifs *via* une fenêtre glissante parcourant les séquences génomiques. La taille de cette fenêtre est fixée en fonction de la taille du motif étudié. L'hypothèse sous-jacente peut se résumer ainsi : un motif de petite taille est plus fréquent dans les séquences et donc la taille de la fenêtre glissante ne nécessite pas d'être grande pour observer un nombre d'occurrences suffisant (et inversement).

Néanmoins, il est connu que la richesse en bases dans les régions n'est pas de 25% pour chacune. Les bases A et T sont par exemple particulièrement majoritaires dans les promoteurs d'*O. sativa* (Fujimori *et al.*, 2005). Ainsi, un motif de longueur n constitué exclusivement de A et de T ne sera pas toujours moins présent qu'un motif plus court, mais constitué de bases C et G. Cette remarque ouvre sur la question des conséquences du choix de cette taille pour l'identification et la caractérisation des motifs d'intérêt, conséquences qui n'ont pas été étudiées. Diverses questions en découlent. Quel est l'impact de ce choix de taille de fenêtre glissante sur l'identification des PLM ? Quel est l'impact sur la liste des promoteurs contenant un motif d'intérêt ? L'article ci-dessous soumis en octobre propose de comparer deux approches : l'approche PLM avec la taille de la fenêtre glissante qui est adaptée à chaque distribution et la même approche sans cette adaptation.

Dans l'approche PLM, une régression linéaire du premier degré est réalisée pour apprendre le modèle de distribution plutôt que d'utiliser une droite moyenne y = a. En effet, la visualisation de plusieurs distributions a révélé que tous les modèles de distributions n'ont pas une orientation horizontale. Néanmoins, aucune étude n'a soulevé ce biais lors de la recherche d'éléments régulateurs dans les promoteurs. Différentes questions ont été posées avant de réaliser le travail présenté ci-dessous. Quelle est la part des distributions de PLM pour laquelle les occurrences en motifs à l'approche de la région fonctionnelle augmentent ou diminuent ? Une augmentation est-elle plus fréquemment observée qu'une diminution ? Dans cet article, une analyse de la pente de chacune des distributions de PLM a permis de répondre à ces questions et d'ouvrir sur des hypothèses permettant de justifier les biais observés.

Hormis des modifications liées à la mises en pages, le manuscrit de cet article est tel qu'il était lors de la soumission.

Improved detection of motifs with preferential location in promoters

Virginie Bernard[1], Alain Lecharny[1, 2] and Véronique Brunaud[1, §]

[1] Unité de Recherche en Génomique Végétale (URGV), UMR INRA 1165 - CNRS 8114 - UEVE, 2 Rue Gaston Crémieux, 91057 Evry Cedex, France

[2] Université Paris-Sud, Institut de Biotechnologie des Plantes (IBP), UMR CNRS 8618 -UPS, Bâtiment 630, 91405 Orsay Cedex, France

[§]Corresponding author

Abstract

Many transcription factor binding sites (TFBSs) involved in gene expression regulation are preferentially located relative to the transcription start site. *In silico* approaches use this property. The study of local over-representation of motifs using a sliding window to scan the promoters is an accurate approach. Nevertheless, consequences of the size of sliding window have never been analysed. We proposed an automatic adaptation of this size relative to each motif distribution profile. This approach allows a better characterization of motif topological constraints and of the lists of genes containing a motif. Moreover, our approach highlighted the non-constant number of spurious motif occurrences that could be counter-selected close to their functional area. Thus, we propose to adapt the sliding window size to each motif distribution and to consider the non-constant presence of motifs in promoters for a better accuracy of *in silico* prediction of TFBSs and a more sensible cartography of promoters.

Résumé

Les sites de fixation des facteurs de transcription (TFBS), impliqués dans la régulation de l'expression des gènes, sont nombreux à être observés à une distance préférentielle des sites de débuts de transcription. Cette propriété est exploitée dans des approches prédictives *in silico*. Avec une fenêtre glissante parcourant les promoteurs, la recherche de sur-représentations locales de motifs a démontré son efficacité. Néanmoins, les conséquences du choix de la taille de cette fenêtre n'ont jamais été étudiées. Nous proposons une approche qui adapte automatiquement cette taille de fenêtre aux distributions. Cette approche permet de caractériser avec plus de précision les contraintes topologiques des motifs et des listes de gènes les contenant. De plus, l'utilisation de notre approche a mis en évidence que le nombre d'occurrences non fonctionnelles des motifs n'est pas constant et que les motifs pourraient

être contre sélectionnés à l'approche de leur région fonctionnelle. Ainsi, pour accroître l'efficacité des approches *in silico* recherchant des TFBS, nous proposons d'utiliser une taille de fenêtre glissante qui s'adapte à chaque distribution de motif et de considérer la fréquence non constante des éléments dans les promoteurs. Cela conduit à une cartographie des promoteurs étudiés statistiquement plus sensible.

Introduction

In eukaryotes, many biochemical processes are highly regulated at different levels of gene transcription including initiation, elongation and termination. The transcription initiation for protein-coding genes requires the formation of the pre-initiation complex, which implies Transcription Factors (TFs), co-factors and the RNA polymerase II (Lagrange *et al.*, 1998; Butler and Kadonaga, 2002; Borukhov and Nudler, 2008). The formation of this complex begins with TFs binding to short DNA sequences named Transcription Factor Binding Sites (TFBSs). These short sequences are mainly observed in non-coding sequences, in an area upstream of the genes, called promoter (Coulombe and Burton, 1999; Smale and Kadonaga, 2003). Three promoter regions can be distinguished: (i) the core promoter, 50 bases upstream of the Transcription Start Site (TSS) (Novina and Roy, 1996; Smale, 2001), which is the minimal DNA region that recruits the basal transcription machinery to direct efficient and accurate transcription initiation; (ii) the proximal promoter in few hundred base pairs upstream of the core promoter; (iii) the distal promoter that in some organisms may extend up to thousands base pairs upstream of the regulated gene (Barton *et al.*, 1997). TFs are involved through a complex cooperation in transcription initiation regulation. Nevertheless, experimentally identified binding sites are known for only a few of them in spite of the accumulation of data as in plants for instance (Palaniswamy *et al.*, 2006). Indeed, two specialized databases AGRIS and PLACE indexed only 140 consensus experimentally identified in *Arabidopsis thaliana* (Higo et al., 1999; Davuluri et al., 2003)

Ab initio approaches to predict TFBSs are generally based on the following assumptions: regulatory elements (i) are over-represented relative to their genomic environment or (ii) are highly observed in a sub-set of promoters. Promoter clustering should be done before the *ab initio* prediction in order to group gene promoters that are expected to

be under the control of the same regulatory elements. This clustering may lie on biological hypotheses like either (i) the preferential conservation of functional regulatory elements during evolution (Sumiyama *et al.*, 2001; Blanchette and Tompa, 2002; Van Hellemont *et al.*, 2005; Woolfe *et al.*, 2005; Monsieurs *et al.*, 2006) or (ii) the presence of the same TFBSs in genes either co-expressed or involved in the same metabolic pathway (Caselle *et al.*, 2002; Aerts *et al.*, 2003; Vandepoele *et al.*, 2006; Wu *et al.*, 2007). Moreover, the preferential location of motifs in promoters has been suggested as a supplementary criterion to identify biologically relevant TFBSs (Kielbasa *et al.*, 2001; FitzGerald *et al.*, 2004). Indeed, several regulatory elements, display a preferential location relative to the TSS in eukaryotic genomes (Bucher and Trifonov, 1988; Burke *et al.*, 1998; Patikoglou *et al.*, 1999; Lim *et al.*, 2004). The location of a regulatory element theoretically may depend on two complementary mechanisms of selection: (i) a positive selection based on the role of the regulatory element in gene transcription regulation and (ii) a negative selection against inaccurate regulatory elements, *i.e.* elements appearing by mutation at a non-functional location (Hahn *et al.*, 2003). Thus, it is generally assumed that motifs preferentially observed at a specific location of promoters are good candidates for elements involved in the regulation of gene expression (Abnizova *et al.*, 2005; Berendzen *et al.*, 2006; Yamamoto *et al.*, 2007; Casimiro *et al.*, 2008). For instance, taking into account the preferential location of the TATA-box consensus sequence allowed a more accurate prediction of the functional occurrences of this element (Molina and Grotewold, 2005; Ponjavic *et al.*, 2006; Yang *et al.*, 2007). Approaches combining the positional information relative to the TSS and the frequency of occurrences of a putative regulatory element have been used at the genomic level to characterize promoters (Kielbasa *et al.*, 2001; FitzGerald *et al.*, 2004; Berendzen *et al.*, 2006; Maston *et al.*, 2006; Bellora *et al.*, 2007; Yamamoto *et al.*, 2007). In all these studies, the sliding window used to scan the genomic sequence and to count the element occurrences had a fixed size for a given

length of motif. Nevertheless, genome-wide analyses have never characterized the consequences of using a fixed size of sliding windows. In this work, we examined the interest to adapt the sliding window size for an optimized definition of topological constraints on promoter regulatory elements. We searched for motifs exhibiting a preferential location in *A. thaliana* promoters, called hereafter the Preferentially Located Motifs (PLMs) and we compared the results obtained with either a fixed or an adaptive size of the sliding window used to build the motif distribution. Our results showed a moderate effect of the sliding window size on the number of PLMs identified but a strong effect on the number of promoters containing a given PLM. Furthermore, our approach allowed the characterization of the slope of the distribution based on the linear regression applied to a wide learning region. Our results showed a wide proportion of model distributions decreasing in the 700 bp of the learning area used for PLM identification. This has consequences on the accuracy of PLM identification and suggests a frequent purifying selection on spurious motifs in the proximal and core promoter regions.

Materials and methods

Promoter set building

We used FLAGdb^{++} (Samson *et al.*, 2004), an integrative database around plant genome structures, to define the transcriptional units of *A. thaliana* by aligning all the available transcripts to gene models excluding pseudogenes. More than 450000 Expressed Sequence Tags and full-length cDNA available for *A. thaliana* genes from TAIR R.6 have been used (Swarbreck *et al.*, 2008). We excluded promoters with a 5' UTR smaller than 50 bases that could be doubtful TSSs. Finally, the promoter set was made of 14927 sequences

extending 1000 bases upstream of the defined TSS and containing the whole predicted 5' UTR.

Preferentially Located Motif identification

For each motif, we extracted all its occurrences in the promoter set. The motif location corresponded to the position of the first base of its sequence. We built motif distributions by scanning the promoter using a 1-base-long sliding window with a 1-base-shift. Then the sliding window was increased step by step up to the size providing motif representation fulfilling to criteria as explained in the results section. The promoter sequences were divided into two regions. First, the [-1000, -300] region of the distal promoter was used to learn the distribution model using a simple linear regression and to build a 99% confidence intervals. Second, in the [-300, ATG] region, we searched for not evenly distributed motifs, *i.e.* motifs showing a distribution exhibiting a peak of motif over-representation at discrete positions.

Different features were associated to each motif with a preferential location in promoters: its distribution, the size of the sliding window used to scan the promoter, the preferential location, *i.e.* the position of the top of the peak in the motif distribution, the functional window between the peak bounds, the list of genes containing in their promoter the motif and a score weighting the topological constraint on the given motif. For each functional window, the number of promoters containing the motif is given rather than the number of occurrences to avoid favouring repeated motifs. The Score of Maximal Square relative to the base line or SMS was the ratio (peak height - base line) / (upper bound - base line). We only considered motifs characterized by a SMS higher than 3 as a Preferentially Located Motifs (PLMs) in order to obtain a confident list of accurate motifs.

Statistical analyses

Statistical analyses were performed with the R statistical software (Dessau and Pipper, 2008). We used R for the PLM identification. We performed two-sided t-statistic test and-

values with a Bonferroni correction in order to distinguish constant from non-constant distribution model in the distal promoter. We performed one-sided Fisher exact tests using the Bonferroni correction for comparisons of percentages between two independent samples: the percentages of promoter containing a given PLM identified by two different approaches.

Results

We analysed the distribution of the 87376 motifs from 2- to 8-base-long to identify PLMs, *i.e.* motifs characterized by a statistically significant preferential location relative to the TSS in the [-300, ATG] area. We defined for each of these PLMs the preferential location relative to the TSS, *i.e.* the top of the distribution peak and the functional window derived from the bounds of the region in which the PLM is locally over-represented. We identified 5105 PLMs, exhibiting topological constraints (Figure 1) consistent with the four promoter regions previously described (Yamamoto *et al.*, 2007). At the top of the graph in Figure 1, the two regions contain PLMs with large peak widths, above 50 bases and up to 350. PLMs located upstream of -50 are putative regulatory elements involved in the specific regulation of small gene sets as indicated by the presence in this promoter region of PLMs corresponding to the G-box CACGTG (Menkens and Cashmore, 1994) or the SORLPI2 element, GGGCC (Hudson and Quail, 2003), both indexed in the databases AGRIS (Davuluri *et al.*, 2003) and PLACE (Higo *et al.*, 1999). PLMs between -50 and the ATG codon mainly reflect the microsatellite presence in plant 5' UTRs and core promoter (Morgante *et al.*, 2002; Fujimori *et al.*, 2003; Molina and Grotewold, 2005). At the bottom of Figure 1, two promoter regions contain PLMs with peak widths less than 50 bases. Many of the PLMs in the group preferentially located between -42 and -24 contain the TATA-box or sequences related to this regulatory element (Mathis and Chambon, 1981; Joshi, 1987; Singer *et al.*, 1990). The fourth

group of PLMs is preferentially located in the [-6, +7] area and thus it overlaps the TSS, *i.e.* the region of the Initiator element or Inr (Doelling and Pikaard, 1995). In this group many PLMs contain the YR initiating dinucleotide identified as Inr in *A. thaliana* (Yamamoto *et al.*, 2007).

The optimal size of sliding windows is highly variable for a given motif length

In published methodologies, window sizes were fixed relative to the motif length (Abnizova *et al.*, 2005; Berendzen *et al.*, 2006; Yamamoto *et al.*, 2007; Casimiro *et al.*, 2008; Civan and Svec, 2009). The rationale was that there is a global relationship between the length of a motif and the number of its occurrences. This assumption would have been adequate whether considering the non-informative regions of a genome and regions where all bases had would be observed with the same relative frequencies. Nevertheless, it is not fully appropriate due to the divergence between base frequencies in genomes (AGI, 2000; Lander *et al.*, 2001; Swarbreck *et al.*, 2008) and in the framework of a positional constraint onto regulatory elements. Therefore, we expected that an arbitrary choice of the sliding window size might have negative consequences on PLM identification and characterization. Indeed, changing the window size has two influences on the results. First, a too large window size smoothes the distribution (Figure 2.A compared to 2.B) and as consequence, PLMs may be missed (Figure 2.C compared to Figure 2.D). Second, decreasing the window size may generate multiple peaks (Figure 2.G compared to 2.H) and leads to abnormally sharp functional windows when motif occurrences are not sufficient for an accurate learning of the distribution upstream –300 relative to the TSS. All these examples highlight the interest in optimizing the sliding window size to each distribution profile. In an approach called hereafter Wadapt, the sliding window is first 1-base-long and it is extended up to 100-base-long while (i) the distribution learning is not relevant due to the presence of at least one

sliding window without motif occurrence in the learning area (Figures 2.E compared to 2.F) or (ii) the PLM functional window may be enlarged to fuse adjacent peaks within one peak (Figures 2.G). Among the 5105 PLMs identified herein, we analysed the optimal size of the sliding windows (Figure 3). The representation shows a large variation in the optimal size of the sliding window no matter what is the motif length. Thus the sliding window size has to be optimized to each distribution profile whatever the motif length. Indeed the optimal size is linked to the length of the motif due to the by chance number of occurrences. Nevertheless, results in Figure 3 clearly indicate that the optimal size is also highly dependent on the topological constraints that differ from one PLM to the other ones.

Adaptation of the sliding window size decreases false negative Preferentially Located Motif identification

We compared the PLM lists returned by the Wadapt approach and by a variant approach called hereafter W15. W15 used a fixed sliding window 15-base-long. This length was one of the most observed sizes obtained for the 6-base-long PLMs in Wadapt analyses (Figure 3) and the size used in previous hexamer analyses (Yamamoto *et al.*, 2007). In this comparison, with both approaches, we only considered the 6-base-long motif size to avoid nested PLM identification.

We identified 525 PLMs shared by W15 and Wadapt (Supplementary data 1) and 207 PLMs specific to only one of the two approaches of which 174 were specific to Wadapt. Nevertheless, we were aware that the SMS threshold of 3 used for PLM identification was stringent. Thus, we computed the same analyses with a SMS threshold of 1.5, 2 and 2.5 and obtained similar results (Table I): whatever the SMS threshold, W15 identified less PLMs than Wadapt. The SMS analysis of the 207 PLMs with a score higher than 3 in only one of both approaches highlighted that the 33 SMS specific of W15 had a score lower but close to 3 with Wadapt (Figure 4.A). On the contrary, the 174 PLMs specifically found by Wadapt did

not have all such a high SMS in W15 (Figure 4.B) and 7 W15 SMS were lower than 1, *i.e.* their distribution was always under the upper bond of the confidence interval. These 7 PLMs have a small sliding window size in Wadapt and contain a CA sequence nested in a A and T rich sequence, characterizing the Inr expected at a very strict position around the TSS (Yamamoto *et al.*, 2007). In these 7 cases, a too wide sliding window in W15 drastically smoothed the peaks and the PLMs were missed (Figure 2.C compared to 2.D). Other motifs whose SMS is higher than 3 in Wadapt and between 1 and 2 in W15, had wider sliding windows in Wadapt than in W15. A high variability in the distribution led to a lower SMS in W15 than in Wadapt (Figure 2.E compared to 2.F). Clearly, these results showed that to adapt the sliding window size allows a better definition of the topological constraints and strongly suggested a more accurate identification of PLMs with Wadapt.

Adaptation of the sliding window size delivers more specific sets of genes containing a Preferentially Located Motif

In complement, we questioned if using fixed or adapted sliding window sizes might have consequences on the number of promoters containing a given PLM. Preferential locations were poorly affected by the sliding window size (Figure 5.A), while the functional window sizes were highly dependent on the approach (Figure 5.B). Thus, the remaining question was how significant a difference in the functional window sizes may or not lead to a difference in the number and nature of the genes containing a given PLM in their promoter. To answer, for each of the 525 PLMs identified in both W15 and Wadapt, we compared the number of genes containing in their promoters a given PLM within the two predicted functional windows. This number was higher for 398 PLMs from W15 and higher for 89 PLMs from Wadapt (Figure 6). Among the 525 shared-PLMs, 243 (46%) were characterized by a significant difference in the number of promoters containing them in both approaches. The Table II illustrates the 25 most biased PLMs and clearly shows a direct correlation

between the number of genes containing a given PLM and the functional window size. Wider functional windows and so higher number of promoters with a given motif were observed (i) for 188 PLMs in W15 that is less able to define strict functional windows (Figures 2.A compared to 2.B) and (ii) for 55 PLMs in Wadapt that added companion peaks to the main peak identified by W15 (Figure 2.G compared to 2.H). To understand which prediction is the most accurate for these 243 PLMs, we examined the distributions from both approaches. Some PLMs from the group of 188 cases contained the canonical TATA-box TATA(A/T)A or a related sequences (Mathis and Chambon, 1981; Singer *et al.*, 1990). A visual examination of the TATAAA distribution showed that a 1-base-long sliding window was the most efficient to characterize the functional window bounds with precision (Figures 2.A compared to 2.B and Figure 6 red dots). Same conclusions suited for PLMs containing the YR-Inr (Table II PLMs with stars, Figure 2.C compared to 2.D and Figure 6 green dots).

As a conclusion, the adaptation of the sliding window size led either, in some cases, to a wider functional window size in Wadapt than W15 or, in most of the cases, to a sharper functional window size in Wadapt. Clearly, the Wadapt approach predicted functional window bounds often fitting better with the distributions than the fixed size approach and therefore defined more accurate lists of genes containing a given motif in their promoters.

Upstream their functional region, many candidate regulatory elements are counter-selected

From each model distribution, in the [-1000, -300] learning area, we defined a linear equation of the form $y = bx + a$. In this equation, the b sign or slope indicates if the model distribution increases or decreases in the direction of the TSS. The analysis of the 699 slopes of the 6-base-long PLMs identified (i) 294 model distributions with a slope not different from zero, *i.e.* the distributions were constant in the learning area, (ii) 327 model distributions with a significantly negative slope and (iii) 78 with a significantly positive one (Figure 7). High

absolute model distribution slopes were observed since 43 distributions were characterized by a slope lower than -0.04 and 21 others by a slope higher than 0.04 (Table III and Figure 8). Thus, many PLM model distributions are decreasing. This effect is accentuated in most of the PLM distributions where an under-representation is noticed close to the motif functional window (Figures 2.B, 2.D, 2.H and 8.B for instance). All together, our observations suggest that many PLMs might be counter-selected close to their functional window. It might be in order to avoid binding competition particularly within the accessible DNA area that is the region between the nucleosomes closest of the TSS, upstream and downstream the site (Yuan et al., 2005; Hartley and Madhani, 2009; Venters and Pugh, 2009)

In complement, we may point out onto the fact that 29 PLMs would not have been found whether the model distribution slope had not been considered, *i.e.* whether an a priori distribution model in the learning area would be in the form y=b equation as in previous studies (Yamamoto et al., 2007). It may also be stressed that only the precise slope characterization together with the sliding window size adaptation allowed the observations sportily an interesting evolutionary hypothesis concerning the counter-selection of spurious regulatory element in the core promoter.

Conclusions and discussion

Our work provides the first evaluation of the role on the sliding window size on the quality of prediction of motifs characterized by a preferential location in promoters. We developed an approach based on the automatic adaptation of the sliding window size to the distribution of each motif in promoters. We identified a wide variation among the optimal window sizes. The comparison between our approach and the same approach but with a fixed window size featured significant differences. An adapted window size lowers the risk of false negative PLM identification. Furthermore, the number of genes containing a given PLM in

their promoters is significantly different for half of the comparisons between Wadapt and W15. Using a fixed sliding window size may result in missing PLMs or in losing a part of the promoters containing a given PLM.

In Wadapt the topological constraints on a motif are improved by the optimal definition of the model distribution depending on the sliding window size adaptation and the definition of the background with the linear regression of first order.

For several PLMs, the motif distribution increases toward the functional window. Most of these PLMs are characterized by a base A or T richness and a wide functional window, *i.e.* a low topological constraint. These characteristics suggest that the related PLMs might be often involved in the chromatin conformation around the TSS. More PLMs exhibit a decrease of their motif density along the promoter toward their functional window. Our hypothesis is that they are good candidates for Transcription Binding Site and that when approaching the functional region, a counter-selection might be implicated in the decrease of spurious motifs in the more accessible DNA around the TSS.

Acknowledgements

We thank Sébastien Aubourg and Jean-Philippe Tamby for their advice and their helpful comments on the manuscript.

References

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

Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B (2003) Toucan: deciphering the cis-regulatory logic of coregulated genes. Nucleic Acids Res **31:** 1753-1764

AGI (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408:** 796-815

Barton MC, Madani N, Emerson BM (1997) Distal enhancer regulation by promoter derepression in topologically constrained DNA in vitro. Proc Natl Acad Sci U S A **94:** 7257-7262

Bellora N, Farre D, Alba MM (2007) Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters. BMC Genomics **8:** 459
Berendzen KW, Stuber K, Harter K, Wanke D (2006) Cis-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves. BMC Bioinformatics **7:** 522
Blanchette M, Tompa M (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res **12:** 739-748
Borukhov S, Nudler E (2008) RNA polymerase: the vehicle of transcription. Trends Microbiol **16:** 126-134
Bucher P, Trifonov EN (1988) CCAAT box revisited: bidirectionality, location and context. J Biomol Struct Dyn **5:** 1231-1236
Burke TW, Willy PJ, Kutach AK, Butler JE, Kadonaga JT (1998) The DPE, a conserved downstream core promoter element that is functionally analogous to the TATA box. Cold Spring Harb Symp Quant Biol **63:** 75-82
Butler JE, Kadonaga JT (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev **16:** 2583-2592
Caselle M, Di Cunto F, Provero P (2002) Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes. BMC Bioinformatics **3:** 7
Casimiro AC, Vinga S, Freitas AT, Oliveira AL (2008) An analysis of the positional distribution of DNA motifs in promoter regions and its biological relevance. BMC Bioinformatics **9:** 89
Civan P, Svec M (2009) Genome-wide analysis of rice (Oryza sativa L. subsp. japonica) TATA box and Y Patch promoter elements. Genome **52:** 294-297
Coulombe B, Burton ZF (1999) DNA bending and wrapping around RNA polymerase: a "revolutionary" model describing transcriptional mechanisms. Microbiol Mol Biol Rev **63:** 457-478
Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E (2003) AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC Bioinformatics **4:** 25
Dessau RB, Pipper CB (2008) ["R"--project for statistical computing]. Ugeskr Laeger **170:** 328-330
Doelling JH, Pikaard CS (1995) The minimal ribosomal RNA gene promoter of Arabidopsis thaliana includes a critical element at the transcription initiation site. Plant J **8:** 683-692
FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C (2004) Clustering of DNA sequences in human promoters. Genome Res **14:** 1562-1574
Fujimori S, Washio T, Higo K, Ohtomo Y, Murakami K, Matsubara K, Kawai J, Carninci P, Hayashizaki Y, Kikuchi S, Tomita M (2003) A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription. FEBS Lett **554:** 17-22
Hahn MW, Stajich JE, Wray GA (2003) The effects of selection against spurious transcription factor binding sites. Mol Biol Evol **20:** 901-906
Hartley PD, Madhani HD (2009) Mechanisms that specify promoter nucleosome location and identity. Cell **137:** 445-458
Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res **27:** 297-300

Hudson ME, Quail PH (2003) Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data. Plant Physiol **133:** 1605-1616

Joshi CP (1987) An inspection of the domain between putative TATA box and translation start site in 79 plant genes. Nucleic Acids Res **15:** 6643-6653

Kielbasa SM, Korbel JO, Beule D, Schuchhardt J, Herzel H (2001) Combining frequency and positional information to predict transcription factor binding sites. Bioinformatics **17:** 1019-1026

Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH (1998) New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. Genes Dev **12:** 34-44

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. Nature **409:** 860-921

Lim CY, Santoso B, Boulay T, Dong E, Ohler U, Kadonaga JT (2004) The MTE, a new core promoter element for transcription by RNA polymerase II. Genes Dev **18:** 1606-1617
Maston GA, Evans SK, Green MR (2006) Transcriptional Regulatory Elements in the Human Genome. Annu Rev Genomics Hum Genet **7:** 29-59
Mathis DJ, Chambon P (1981) The SV40 early region TATA box is required for accurate in vitro initiation of transcription. Nature **290:** 310-315
Menkens AE, Cashmore AR (1994) Isolation and characterization of a fourth Arabidopsis thaliana G-box-binding factor, which has similarities to Fos oncoprotein. Proc Natl Acad Sci U S A **91:** 2522-2526
Molina C, Grotewold E (2005) Genome wide analysis of Arabidopsis core promoters. BMC Genomics **6:** 25
Monsieurs P, Thijs G, Fadda AA, De Keersmaecker SC, Vanderleyden J, De Moor B, Marchal K (2006) More robust detection of motifs in coexpressed genes by using phylogenetic information. BMC Bioinformatics **7:** 160
Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet **30:** 194-200
Novina CD, Roy AL (1996) Core promoters and transcriptional control. Trends Genet **12:** 351-355
Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grotewold E (2006) AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. Plant Physiol **140:** 818-829
Patikoglou GA, Kim JL, Sun L, Yang SH, Kodadek T, Burley SK (1999) TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. Genes Dev **13:** 3217-3230
Ponjavic J, Lenhard B, Kai C, Kawai J, Carninci P, Hayashizaki Y, Sandelin A (2006) Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. Genome Biol **7:** R78
Samson F, Brunaud V, Duchene S, De Oliveira Y, Caboche M, Lecharny A, Aubourg S (2004) FLAGdb++: a database for the functional analysis of the Arabidopsis genome. Nucleic Acids Res **32:** D347-350
Singer VL, Wobbe CR, Struhl K (1990) A wide variety of DNA sequences can functionally replace a yeast TATA element for transcriptional activation. Genes Dev **4:** 636-645
Smale ST (2001) Core promoters: active contributors to combinatorial gene regulation. Genes Dev **15:** 2503-2508
Smale ST, Kadonaga JT (2003) The RNA polymerase II core promoter. Annu Rev Biochem **72:** 449-479
Sumiyama K, Kim CB, Ruddle FH (2001) An efficient cis-element discovery method using multiple sequence comparisons based on evolutionary relationships. Genomics **71:** 260-262
Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res **36:** D1009-1014
Van Hellemont R, Monsieurs P, Thijs G, de Moor B, Van de Peer Y, Marchal K (2005) A novel approach to identifying regulatory motifs in distantly related genomes. Genome Biol **6:** R113
Vandepoele K, Casneuf T, Van de Peer Y (2006) Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. Genome Biol **7:** R103

Venters BJ, Pugh BF (2009) A canonical promoter organization of the transcription machinery and its regulators in the Saccharomyces genome. Genome Res **19:** 360-371

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE, Elgar G (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol **3:** e7

Wu RZ, Chaivorapol C, Zheng J, Li H, Liang S (2007) fREDUCE: Detection of degenerate regulatory elements using correlation with expression. BMC Bioinformatics **8:** 399

Yamamoto YY, Ichida H, Matsui M, Obokata J, Sakurai T, Satou M, Seki M, Shinozaki K, Abe T (2007) Identification of plant promoter constituents by analysis of local distribution of short sequences. BMC Genomics **8:** 67

Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E (2007) Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene **389:** 52-65

Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. Science **309:** 626-630

Tables

SMS threshold	Shared PLMs	W15 specific PLMs	Wadapt specific PLMs
3	525	33	174
2.5	760	20	351
2	1124	22	474
1.5	1641	75	282

Table I: With different SMS thresholds, comparison of the number of PLMs identified in the W15 and Wadapt approaches

With SMS threshold from 1.5 to 3, the second column indicates the number of motifs identified as PLMs by both W15 and Wadapt, the third column the number of motifs being PLM only in W15 and the fourth column those found only by Wadapt.

PLM sequence	W15		Wadapt		
	Number of promoters	Sliding window size	Number of promoters	Functional window size	Functional window size
AAAAAG	5826	28	**6878**	323	251
AAAAGC	**2141**	2	179	9	205
AAAAGG	**1976**	2	164	7	180
AAACCC	**2664**	4	624	34	242
AAAGCT	1078	25	**1499**	169	110
AAGGGC	198	31	**397**	258	102
ACCCCA	142	19	**320**	100	41
ACGACG	320	21	**760**	254	97
ACGCCG	78	28	**258**	168	49
CACAAA *	**1507**	1	91	3	103
CATCTT *	**1729**	2	66	3	175
CCAAAC	831	24	**1660**	216	100
CCAACC	226	28	**598**	181	59
CCACAC	251	23	**571**	148	61
CCGTCA	150	26	**562**	308	64
CCTTTT	886	19	**1694**	157	78
CGCCAC	159	28	**436**	252	77
CGTCAC	309	18	**620**	233	110
CTCAAA *	**1493**	1	43	1	124
CTTCAA *	**1876**	1	43	1	180
CTTCAC *	**1682**	2	120	7	214
CTTCAT *	**1730**	2	116	4	171
CTTTTC	1573	26	**2064**	176	123
GCCTTT	287	17	**521**	124	67
GCGAAA	207	26	**450**	148	58

Table II: List of the 25 PLMs predicted by both W15 and Wadapt approaches but with a significant difference in the number of promoters containing the PLMs

For these 25 PLMs, the difference is significant with a p-values lower than 1e-16. The stars highlight the PLMs including a YR initiating dinucleotide and characterized by a distribution with a sharp peak close to the TSS.

Rank	PLM	b-value	Preferential position	Functional window width
1	AACAGA	-6,69e-2	0	322
2	AAGGAG	-6,46e-2	32	127
3	CAGACA	-5,58e-2	1	311
4	GGGCAA	-4,79e-2	-75	274
5	GAAAAG	-4,36e-2	-118	307
6	TAGCTC	-4,30e-2	48	233
7	AAAGCT	-4,00e-2	42	169
8	AGAACC	-3,86e-2	35	165
9	CTCCAG	-3,60e-2	55	273
10	GTCACT	-3,35e-2	6	240
11	CTGATC	-3,31e-2	61	207
12	TCATCC	-3,27e-2	10	163
13	CACTGT	-3,18e-2	24	167
14	AAGCCT	-3,17e-2	-48	206
15	CTCAGC	-3,13e-2	14	199
...				
685	AATTCC	2,80e-2	51	213
686	CGATTT	3,26e-2	19	120
687	TTTTCC	3,73e-2	24	204
688	TTTCTT	5,49e-2	17	146
689	TTCTTT	6,84e-2	29	133
690	CCAATT	7,95e-2	53	291
691	AAACCA	1,09e-1	65	302
692	TTTTCT	1,12e-1	21	91
693	AACCAA	1,21e-1	66	297
694	AAAAAG	2,15e-1	28	323
695	CCAAAA	2,16e-1	48	352
696	AAAAAC	3,13e-1	36	296
697	AATTAA	3,38e-1	32	157
698	TAATAA	3,66e-1	45	225
699	ACAAAA	4,14e-1	45	184

Table III: List of the 30 PLMs showing the lowest and the highest regression-coefficients for their distributions in the distal promoters

The PLMs are ranked according to the regression-coefficients. The 15 first rows and the 15 last rows list the PLMs with respectively the most decreasing and the most increasing distributions. All the distribution slopes (b-values) are significantly different from 0 (p-values lower than 1e-30). The two last columns indicate the topological constraints of the PLMs.

Figures

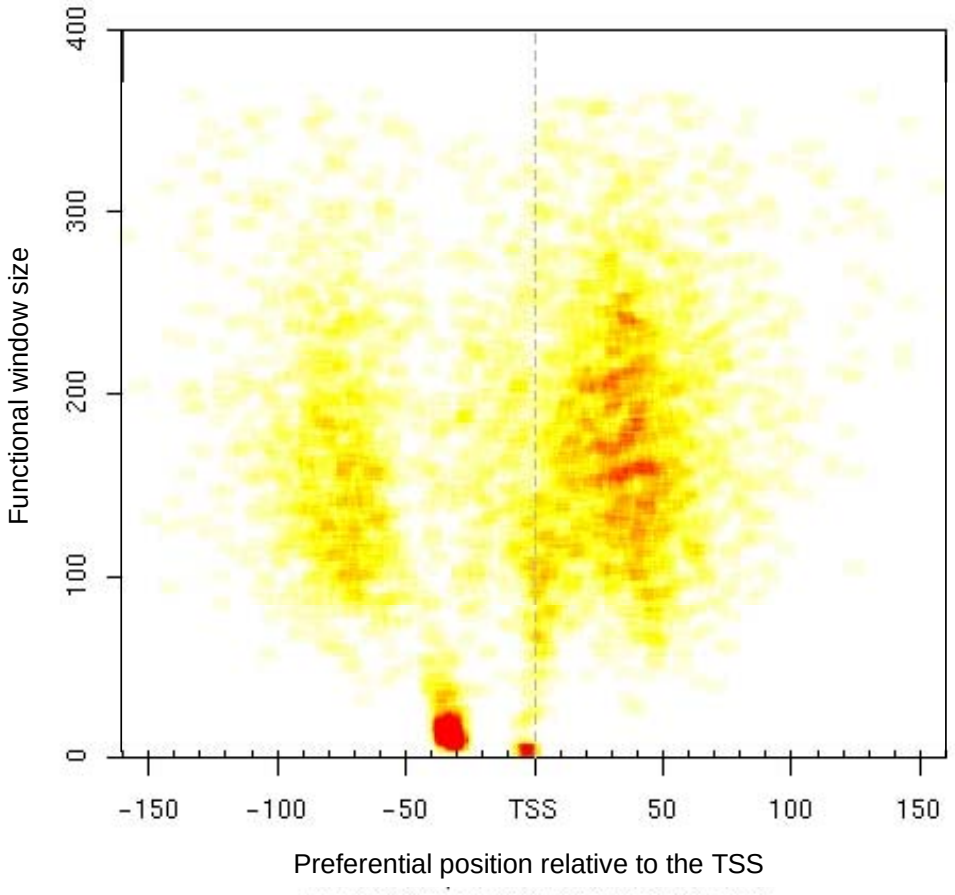


Figure 1: Overview of the *Arabidopsis thaliana* promoter organization according to the topological constraints on the PLMs.

The densities of PLMs with the same preferential position and functional window width are represented by a gradient of colors from yellow (lower density) to red (higher density). Only PLMs whose constraints are shared by at least 30 PLMs are considered. The vertical grey dashed line indicates the TSS position.

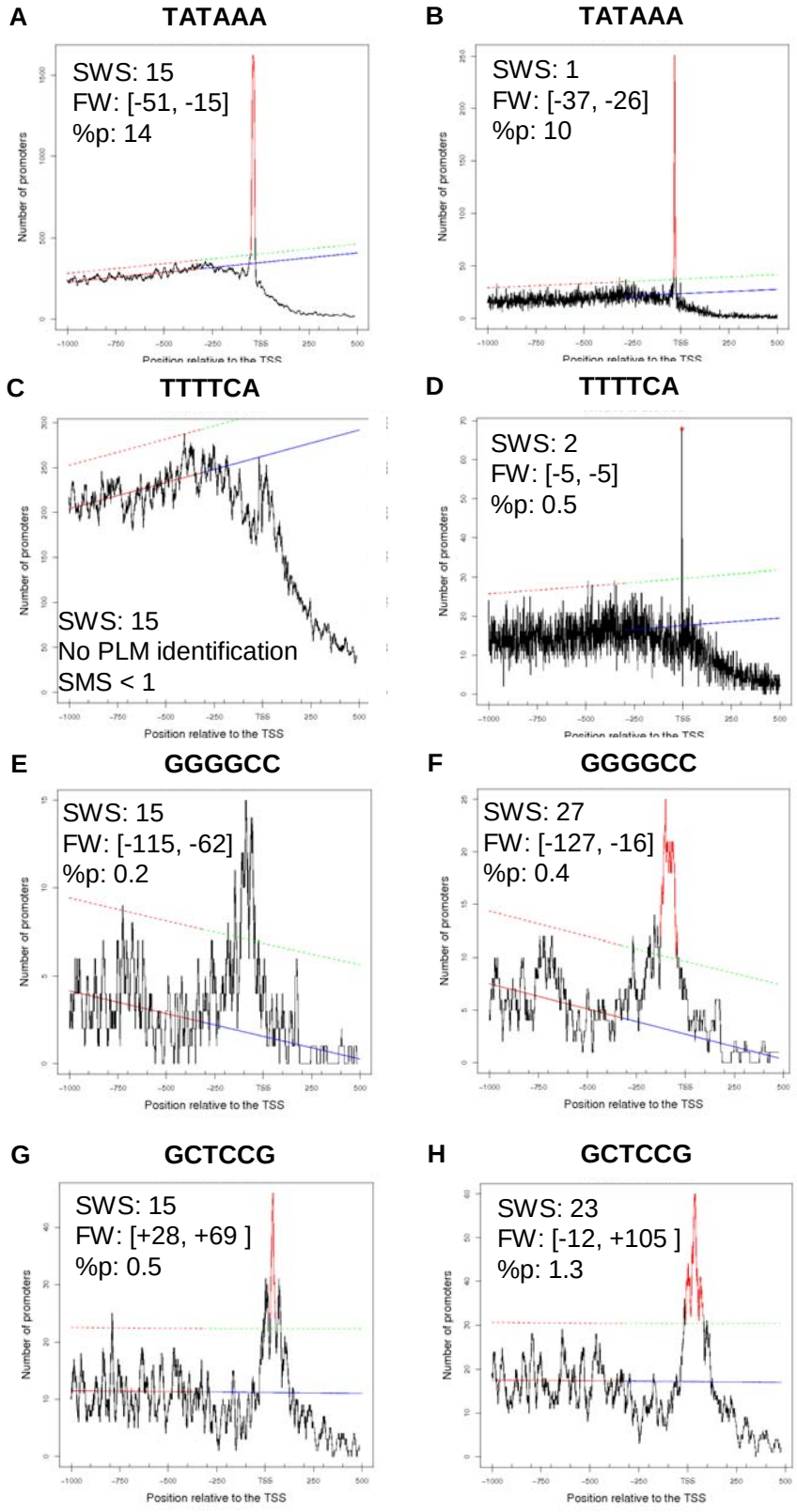


Figure 2: Consequences of the choice of the size of the sliding window on the motif distributions and features

Each distribution was learned in the [-1000, -300] area where the base line (continuous red line) and the upper bound of the confidence interval (dashed red line) are estimated. Red parts of distributions indicate the PLM functional window. Each row corresponds to distributions of the same motif, but with two different sliding window sizes. The sliding window size in the first column is 15-base-long while it is the adapted size designed by the Wadapt method in the second column. **A** and **B**, distributions of TATAAA. **C** and **D**, distributions of TTTTCA. **E** and **F**, distributions of GGGGCC. **G** and **H**, distributions of GCTCCG. Feature legends: SWS: Sliding window size. FW: Functional window. %p: percentage of promoters containing the given sequence motif in the bounds of the PLM functional window.

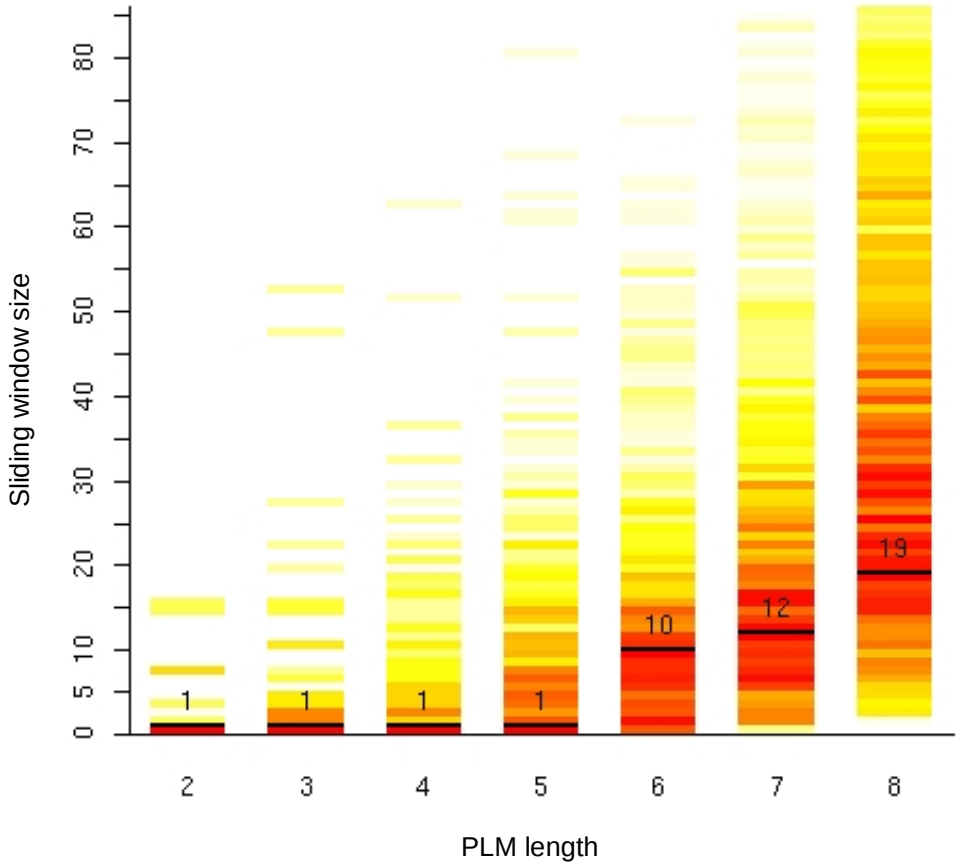


Figure 3: The sliding window size relative to the PLM length

Each column shows the adapted sliding window size in function of the PLM length ranked from 2 to 8-base-long. For each PLM length, we computed the percentage of PLMs with a given sliding window size. The percentages are represented by a gradient of colour from yellow (lower percentages) to red (higher percentages).

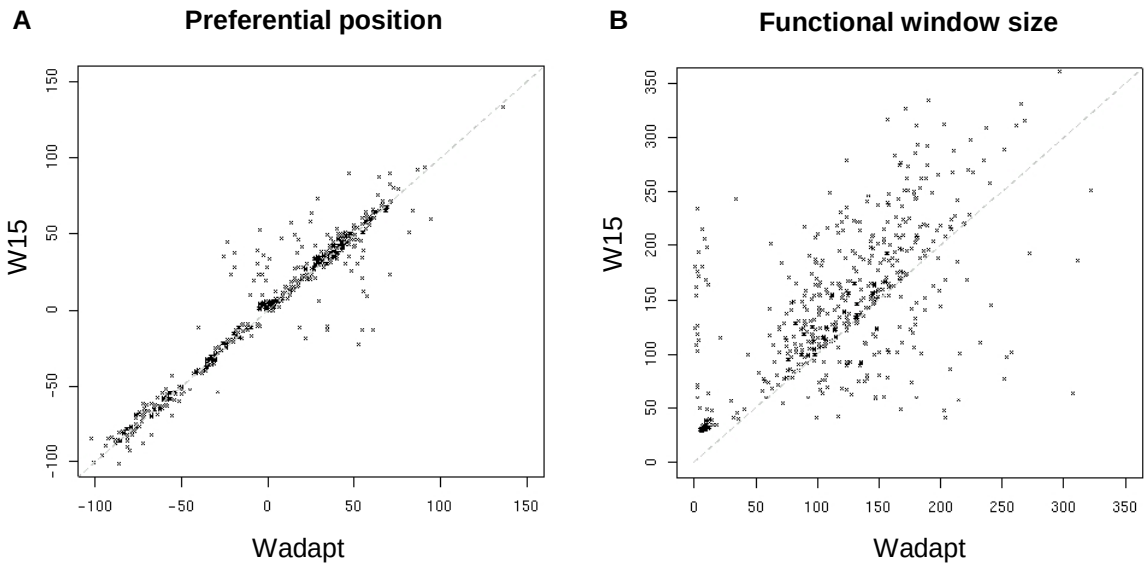


Figure 4: SMS of motifs identified as PLMs either by W15 or Wadapt

(**A**) SMS calculated with Wadapt of the 33 motifs being PLMs specific to W15. (**B**) SMS calculated with W15 of the 177 motifs being PLMs specific to Wadapt. The motifs with SMS lower than 1 are grouped together at the SMS value lower than 1.

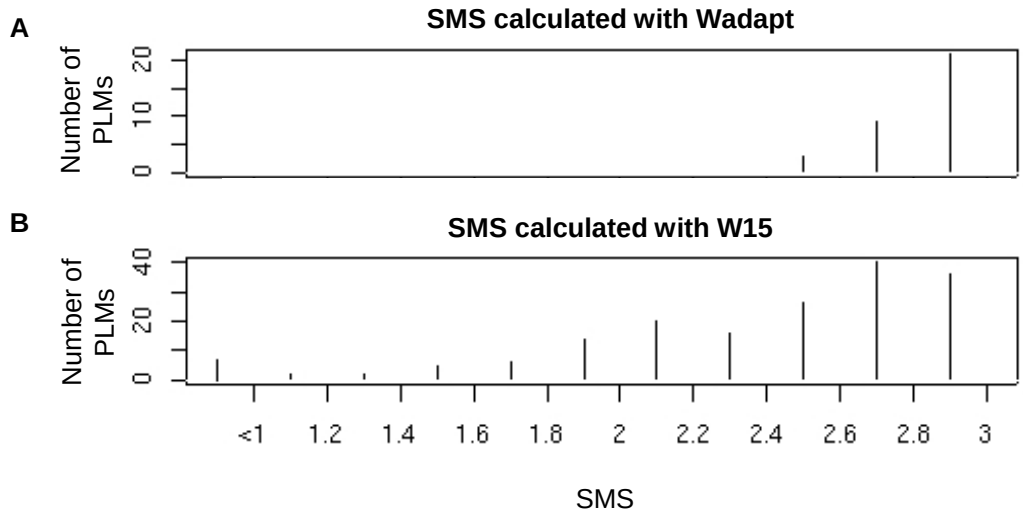


Figure 5: Topological constraints of PLMs given by both W15 and Wadapt

W15 and Wadapt approaches shared the identification of 525 PLMs with SMS higher than 3. (**A**) Dot plot of the preferential positions of the 525 PLMs (**B**) Dot plot of the functional window size.

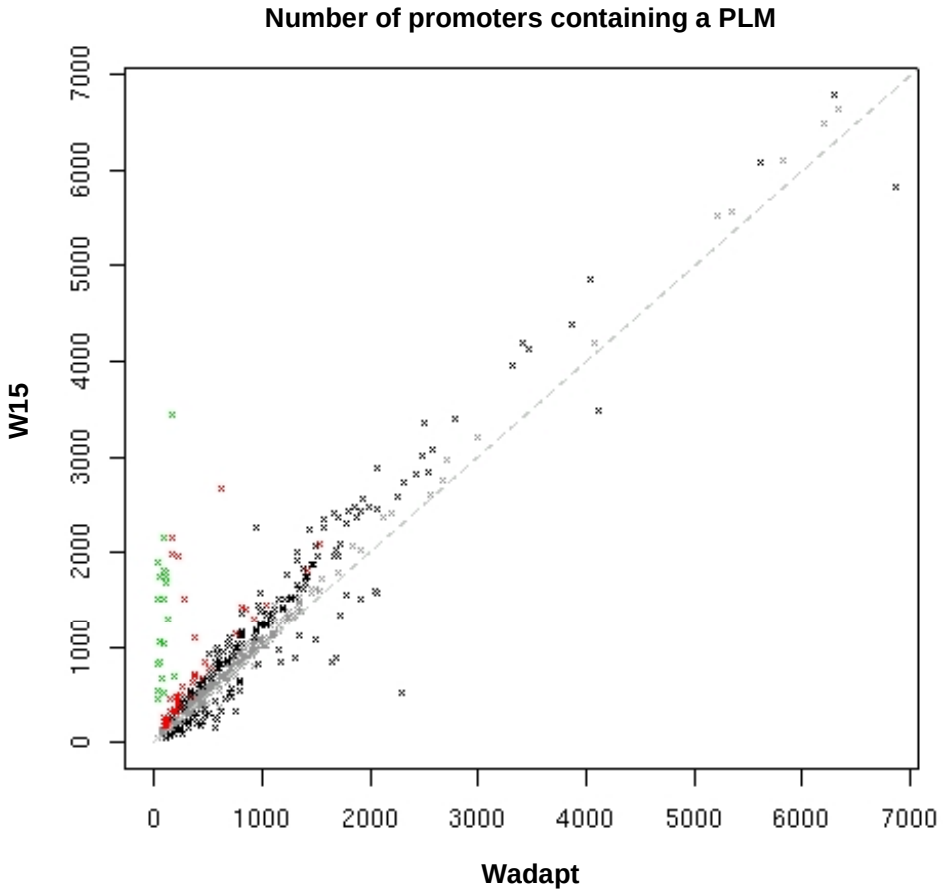


Figure 6: Number of promoters containing a PLM from either W15 or Wadapt

For each of the 525 PLMs identified by both methods, number of promoters containing a given PLM with W15 is plotted against the number with Wadapt. Grey dots are for PLMs with no statistical difference between both approaches. Other coloured dots are for PLMs significantly more observed in promoters with one of both approaches. Green dots are for PLMs sharing the topological constraints of the canonical TATA-box, *i.e.* a sharp peak in [-39, -26]. Red dots are for PLMs sharing the topological constraints of the plant core-Inr YR: a sharp peak close to the TSS.

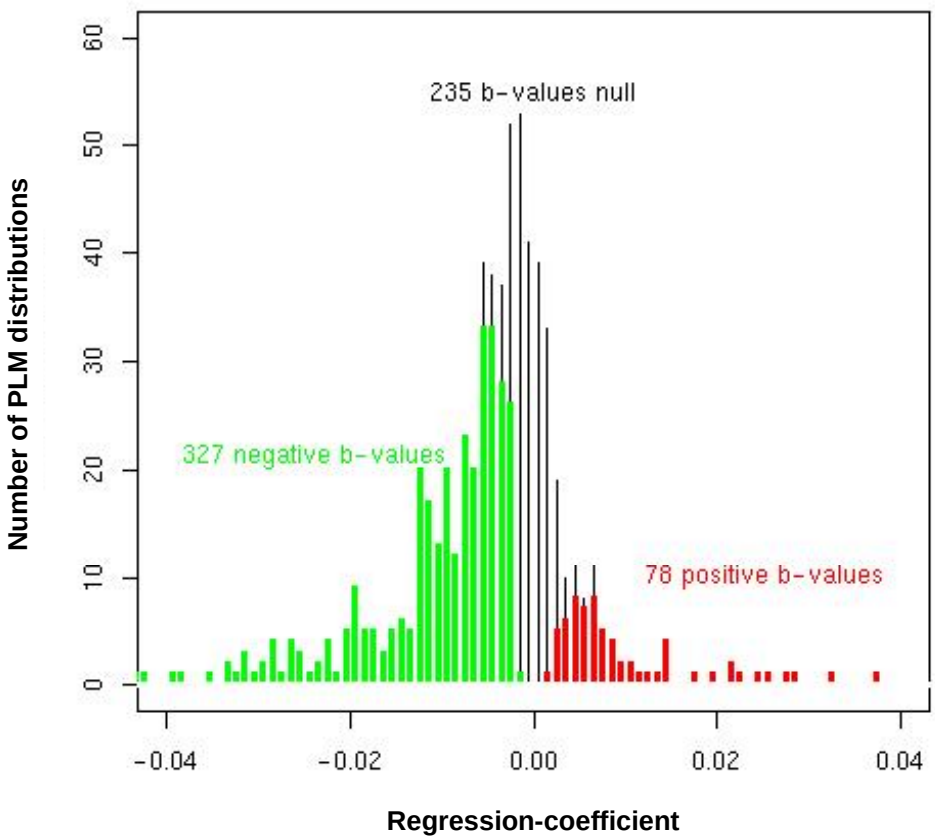


Figure 7: Slopes of the PLM distributions in the distal promoters of *A. thaliana*

All motif distributions are analysed with a linear regression in the [-1000, -300] distal promoter. The histogram represents the b-values, *i.e.* the regression-coefficient of the 699 distributions of 6-base-long PLMs. In black are the regression-coefficients not significantly different from zero. In green and in red, the regression-coefficients are respectively significantly lower and higher than zero. For these PLMs the density of motifs decreases or increases respectively from -1000 to -300.

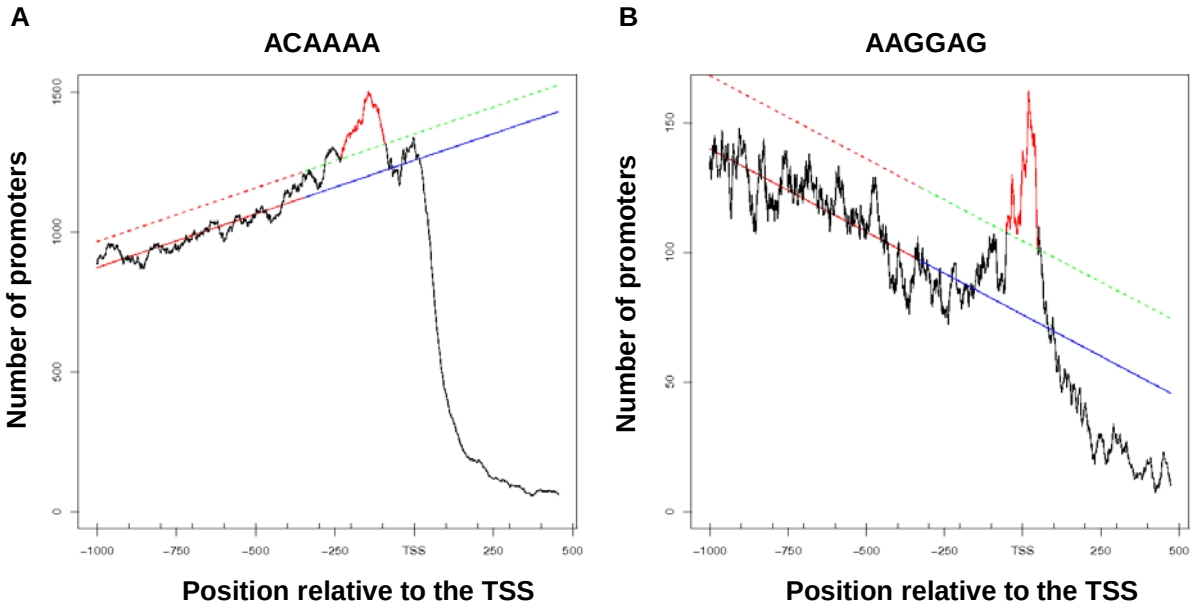


Figure 8: Distributions of two PLMs exhibiting high distribution slopes in the distal promoters

In **A**, the distribution of the ACAAAA PLM exhibits a linear regression-coefficient higher than +0.4. This distribution is one of the most increasing in the learning area. In **B**, the distribution of AAGGAG exhibits a linear regression-coefficient less than -0.06. This distribution is one of the most decreasing in the learning area.

Supplementary data

Supplementary data 1

List of the 525 PLMs identified with both Wadapt and W15 and ranked by their SMS in Wadapt. SWS: Sliding window size. PP: Preferential Position. FW: Functional Window. NbP: Number of promoters containing the given PLM in its functional window. FWS: Functional Window Size.

Cf. Annexe VI.A

En conclusion, ce travail a permis de montrer l'importance du choix de la taille de la fenêtre glissante sur l'identification des motifs ayant des contraintes topologiques.

Des conséquences modérées mais néanmoins existantes ont été observées sur l'identification de faux négatifs si la taille de fenêtre glissante n'est pas adaptée. Plus impactée, la taille de la fenêtre fonctionnelle souffre d'une taille de fenêtre glissante non adaptée qui peut :

- soit lisser la représentation si la taille de la fenêtre glissante est trop grande. Cela augmente donc la prédiction de la taille de la fenêtre fonctionnelle ;
- soit induire beaucoup de fluctuations si la taille de la fenêtre glissante est trop petite. Cela génère des représentations plus difficilement exploitables et pour lesquelles une fenêtre fonctionnelle peut être à tort prédite étroite.

Ce résultat a des conséquences sur la construction d'une liste de gènes ayant le motif étudié dans leurs promoteurs. Ainsi, dans l'avenir, nous proposons qu'une telle adaptation de la taille de la fenêtre glissante soit appliquée aux méthodologies dont l'objectif est de caractériser des motifs ayant des contraintes topologiques.

L'étude des pentes au sein de la région d'apprentissage a souligné la présence majoritaire de distributions non constantes (58%). Parmi ces distributions, celles à pentes décroissantes sont les plus observées (47%) et pourraient être associées à des PLM étant des TFBS et dont les séquences - qui sont susceptibles d'être fonctionnelles dans une région précise exclusivement - pourraient subir une contre sélection hors de leur fenêtre fonctionnelle.

5.2.3 Validation de l'approche PLM

a) Distribution des TFBS disponibles dans AGRIS et PLACE

En premier lieu, l'approche PLM a été étudiée pour vérifier son efficacité pour la recherche de motifs d'intérêt.

Dans la première partie de la thèse, deux bases de données PLACE (Higo *et al.*, 1999) et AGRIS (Davuluri *et al.*, 2003) ont été présentées. Elles indexent les sites de fixation des facteurs de transcription identifiés par des approches biologiques. Combinées, les deux bases contiennent 140 TFBS différents. Notons néanmoins que les TFBS sont pour certains inclus les uns dans les autres.

Les distributions de ces 140 motifs ont été étudiées dans le jeu de 14927 promoteurs d'*A. thaliana* et parmi eux, 64 sont caractérisés par une position préférentielle dans le promoteur proximal. Notons que deux tiers des TFBS ne sont pas des PLM. Cela est expliqué par plusieurs raisons.

- Premièrement, l'ensemble des promoteurs de gènes n'est pas étudié. Des TFBS peuvent être spécifiquement présents dans des promoteurs non étudiés. Certains des TFBS sont *de facto* absents dans les 14927 promoteurs.
- Deuxièmement, parmi les 140 TFBS dans les bases de données PLACE et AGRIS, certains sont très longs, et peuvent aller jusqu'à une longueur de 33 bases (Figure 5-11). De tels motifs ne peuvent pas être étudiés avec l'approche

PLM : ils sont trop rares et l'apprentissage du modèle de distribution n'est pas réalisé correctement.

- Troisièmement, les TFBS ne sont évidemment pas tous caractérisés par une contrainte positionnelle dans la région du promoteur central [-299, ATG].

Néanmoins, les 64 TFBS identifiés par l'approche PLM en tant que motifs soumis à des contraintes topologiques permettent une première validation de l'approche PLM. De plus, l'analyse de la présence de ces 64 TFBS dans différents groupes de promoteurs est un atout considérable lors de l'étude de petits groupes de gènes. Des caractéristiques fonctionnelles communes peuvent être mises en évidence dans un sous-groupe de promoteurs comme présenté dans la suite des résultats de cette thèse.

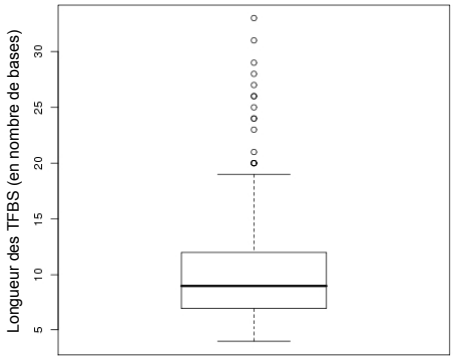


Figure 5-11 : Les longueurs des 140 TFBS connus chez *A. thaliana*.

Longueur des 140 TFBS indexés dans les bases de données PLACE et AGRIS.

b) Etude d'un jeu de promoteurs dont les TFBS sont connus

i) Le jeu de promoteurs étudié

Un jeu de 55 gènes appartenant à la famille des ARN hélicases à boîte DEAD a été exploité pour valider l'emplacement de TFBS connus (Mingam *et al.*, 2004). Une annotation structurale validée par des expertises biologiques a permis de mettre en évidence la présence de deux boîtes régulatrices : la boîte Télo de motif consensus AAACCCTA (Manevski *et al.*, 1999; Manevski *et al.*, 2000) et le motif II de motif consensus GGCCCA (Hudson & Quail, 2003). Parmi les 55 gènes de la famille étudiée, 43 peuvent être étudiés par l'approche PLM car ils appartiennent au jeu de 14927 promoteurs.

ii) Les PLM identifiés

Les distributions des motifs de 5 à 8 bases ont été étudiées afin d'identifier les PLM dans ce jeu de promoteurs connu. Cette analyse ayant été réalisée en début de thèse, alors que la taille de la fenêtre glissante ne s'adaptait pas encore automatiquement, les résultats présentés sont obtenus avec une taille de fenêtre glissante fixe de 50 bases.

Lors de cette étude, 6 PLM ont été identifiés. Des chevauchements de motifs caractérisés par des contraintes topologiques similaires permettent de les regrouper (Table

5-1). Finalement, deux motifs peuvent résumer ces 6 PLM : AAACCCTA et GGCCCA correspondant aux résultats attendus. L'emplacement de ces deux éléments serait capital pour réguler efficacement l'expression des gènes qui les possèdent (Mingam *et al.*, 2004)

Séquence du PLM	Nom du TFBS (séquence)	SMS (classement)	Fenêtre fonctionnelle	Nombre de promoteurs Parmi les 43 contenant le PLM
AACCCT	Boîte Télo (AAACCCTA)	5.78 (3)	[-101,91]	20
AACCC		4.99 (1)	[-102,111]	24
GGCCCA	Motif II (GGCCCA)	6.26 (4)	[-158, 2]	9
GGCCC		9.33 (5)	[-211, 5]	16
GCCCA		10.82 (6)	[-219, 5]	12
GCCCAT		5.67 (2)	[-153,-13]	13

Table 5-1 : PLM mis en évidence lors de l'étude des 43 gènes.

Liste des 6 PLM identifiés avec en rouge les bases qui coïncident avec le motif consensus de la boîte Télo (Manevski *et al.*, 1999 ; Manevski *et al.* 2000) et en bleu celles qui coïncident avec le motif consensus du Motif II (Tremousaygue *et al.*, 2003).

Du fait de l'analyse d'un jeu réduit, les contraintes topologiques des deux TFBS établies au sein des 43 promoteurs (Figure 5-12 A et B et Table 5-1) sont moins précises que celles observées dans le jeu global (Figure 5-12 C et D). Elles sont néanmoins évidentes. Dans le jeu de 14927 promoteurs, 6.4% des gènes possèdent une boîte Télo de motif consensus AAACCCTA dans sa région fonctionnelle [-53, 136] et 10.1% possèdent un Motif II de motif consensus GGCCCA dans sa région fonctionnelle [-192, -33] (Annexe II.B rang 247 et Annexe II.A rang 11). Pour les 43 promoteurs de l'étude PLM, les pourcentages sont de 37.2% et 30.2% respectivement pour la boîte Télo et le Motif II en utilisant les mêmes régions fonctionnelles que celles de l'analyse globale. Ces deux pourcentages sont significativement supérieurs aux pourcentages dans le jeu global (p-values respectives de 2e-16 et 1e-5 - test exact de Fisher).

Durant cette étude, un petit jeu de promoteurs dont les éléments régulateurs sont connus a été analysé par l'approche PLM qui a retrouvé les résultats attendus. Ainsi, un nombre restreint de séquences peut être analysé par l'approche PLM à la recherche de contraintes topologiques.

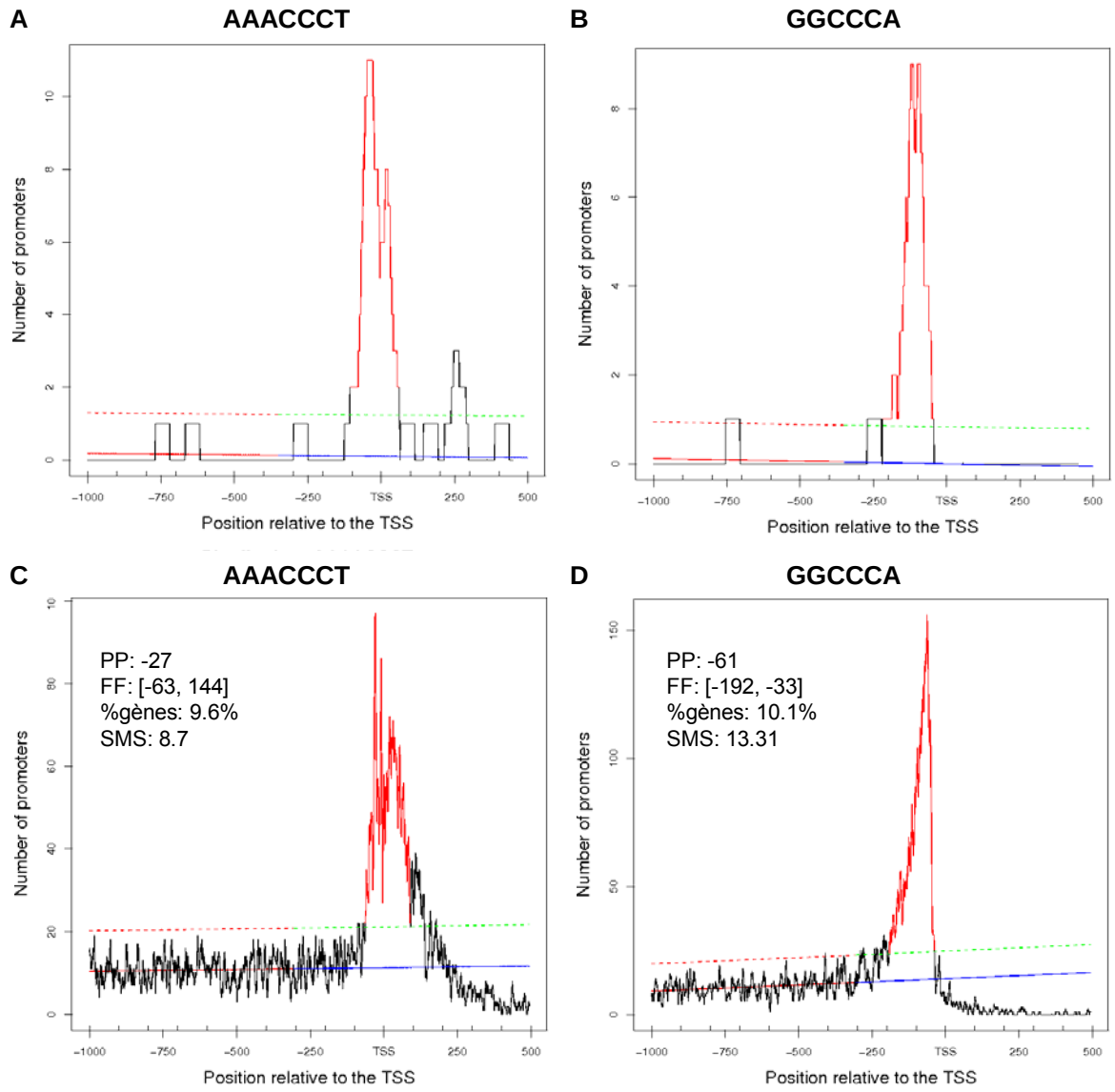


Figure 5-12 : Distributions des PLM AAACCCT et GGCCCA.

Représentation de la boîte Télo de séquence consensus AAACCCT (**A, C**) et du Motif II de séquence consensus GGCCCA (**B, D**) dans le jeu des promoteurs des gènes appartenant à la famille des hélicases à boîte DEAD (**A, B**) et dans l'ensemble des 14927 promoteurs (**C, D**). Une taille de fenêtre glissante fixe de 50 bases est utilisée pour les distributions A et B. Pour chaque PLM, sont renseignés sa position préférentielle (PP), sa fenêtre fonctionnelle (FF), le pourcentage de gènes contenant le PLM dans ses promoteurs (%gènes) et son score le SMS.

En conclusion, les validations du jeu de données tout comme de l'approche PLM ont permis de poursuivre l'étude avec ce jeu de 14927 promoteurs pour rechercher des motifs caractérisés par des contraintes topologiques.

c) Contrôle négatif

Afin d'estimer la propension de l'approche PLM à identifier des faux positifs, nous l'avons appliquée à deux jeux de séquences de références. Ces deux jeux sont constitués de 14927 séquences pour avoir la même taille que celle du jeu de promoteurs analysé au cours de cette thèse. Le premier jeu de référence est constitué de séquences d'*A. thaliana* sélectionnées aléatoirement sur les 5 chromosomes nucléaires et qui ne correspondent pas aux 14927 séquences promotrices. Le deuxième jeu de référence est constitué de séquences générées par l'outil «random sequence» du portail RSAT (Thomas-Chollier *et al.*, 2008). L'outil a permis de générer des séquences aléatoires basées sur les modèles de Markov calibrés sur l'ensemble des régions non-codantes d'*A. thaliana*.

La recherche de PLM dans ces deux jeux de référence a été réalisée pour les 4096 motifs de longueur 6. Parmi ces motifs, moins de 8% ne sont pas analysables car lors de la construction de leurs distributions, l'adaptation de la taille de la fenêtre glissante n'aboutit pas. Pour plus de 59% des motifs, la sur-représentation la plus importante est localisée dans la région d'apprentissage du modèle de distribution (Table 5-2). La Figure 5-7 page 65 illustre en partie ces résultats. Le caractère aléatoire de ces distributions est montré par (i) leurs faibles scores : les valeurs de SMS de ces motifs sont inférieures à 3 et moins de 2% sont comprises entre 2 et 3 ; et par (ii) leur largeur de fenêtre de sur-représentations, très étroite, souvent sur une seule position. En comparaison avec les pics de sur-représentations retenus et présentés dans ce document, c'est-à-dire les pics des PLM, les sur-représentations dans la région d'apprentissage ne semblent pas avoir de caractéristiques pouvant leur conférer une réalité biologique.

	Séquences génomiques aléatoires	Séquences aléatoires (RSAT)
Distribution non analysable : adaptation de la fenêtre glissante sans succès	8%	3%
Sur-représentation dans la région d'apprentissage [-1000, -300]	59%	63%
Sur-représentation dans la région [-299, +500] et SMS inférieur ou égale à 1	15%	14%
Sur-représentation dans la région [-299, +500] et SMS supérieur à 1	18%	19%

Table 5-2 : Recherche de PLM dans deux jeux de contrôles négatifs. Les quatre catégories de distributions obtenues.

Parmi les distributions retenues par l'approche PLM car la sur-représentation est localisée dans la région [-299, +500], un score a été attribué aux motifs (Table 5-2). Pour les deux jeux de référence considérés, les SMS significatifs, c'est-à-dire supérieurs à 1, sont majoritairement compris entre 1 et 2 (Figure 5-13 et Table 5-2) et aucun motif n'a un SMS supérieur à 3. En conclusion, ce contrôle négatif, *a posteriori*, valide la pertinence du choix de ne considérer que les motifs ayant un SMS supérieur à 3 lors de l'étude des 14927 promoteurs.

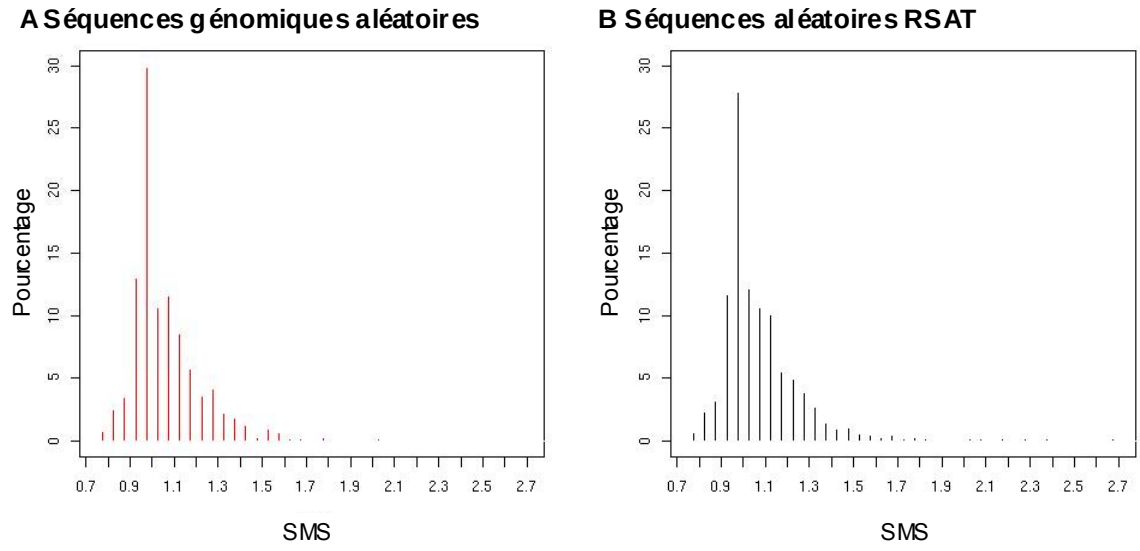


Figure 5-13 : Histogrammes des SMS obtenus lors de l'analyse des jeux de contrôles négatifs.

SMS obtenus lors de l'analyse des 14927 séquences génomiques aléatoires (A) et des 14927 séquences aléatoires RSAT (B).

5.3 Motifs sur-représentés dans l'ensemble des séquences ou localement

L'approche PLM a été comparée avec une autre approche permettant l'identification de motifs sur-représentés sans prise en compte de la position.

5.3.1 Logiciel R'MES

R'MES pour «*Recherche de Mots Exceptionnels dans une Séquence*» est un logiciel développé au laboratoire Mathématique et Informatique du Génome de l'INRA de Jouy en Josas (Schbath, 1995; Hoebeke & Schbath, 2006). Il détecte les motifs de taille t prédéfinie par l'utilisateur dont la fréquence d'apparition observée est significativement différente de celle attendue dans un lot de séquences nucléiques modélisées (chaînes de Markov).

5.3.2 Recherche de motifs d'intérêt de 6 nucléotides

Sans tenir compte des motifs dégénérescents, il existe 4096 motifs de longueur 6 constitués des 4 bases A, C, G et T. L'ensemble de ces motifs a été étudié dans le jeu de 14927 promoteurs afin d'identifier les motifs d'intérêt. Un motif d'intérêt est un motif sur-représenté localement pour l'approche PLM et c'est un motif sur- ou sous-représenté dans l'ensemble des séquences pour R'MES. Le modèle utilisé pour l'étude avec R'MES est le modèle maximal, c'est-à-dire 5. Le diagramme de Venn de la Figure 5-14 révèle les motifs d'intérêt partagés et ceux propres à chaque approche. Ces derniers sont disponibles en Annexe III. Les PLM ne sont pas tous identifiés comme étant des motifs exceptionnels par R'MES et des motifs sous-représentés peuvent être des PLM.

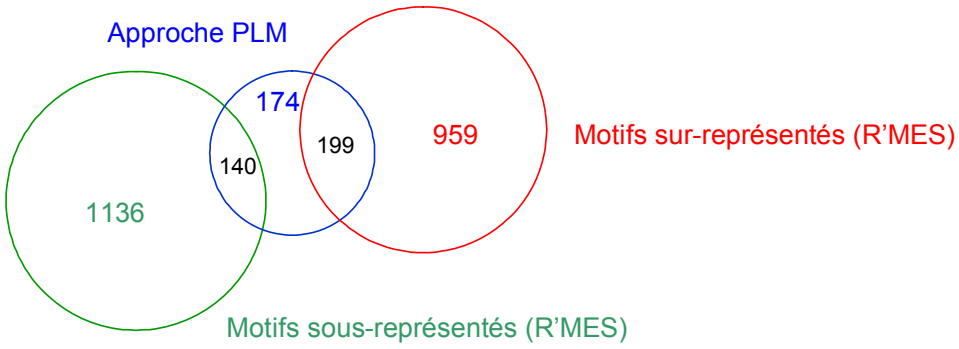


Figure 5-14 : Motifs d'intérêt identifiés par l'approche PLM et par R'MES.

Représentation des motifs d'intérêt de 6 bases qui sont sur- et sous-représentés d'après R'MES respectivement en rouge et vert. Les motifs d'intérêt d'après l'approche PLM sont en bleu. R'MES a été lancé en utilisant l'ordre maximal.

En complément, les scores des motifs analysés par les deux approches ont été comparés. La représentation des scores des motifs définis par R'MES par rapport aux scores des PLM (Figure 5-15) montre que les PLM sont aussi bien des motifs sous-représentés, sur-représentés ou non exceptionnels, y compris ceux ayant des contraintes topologiques élevées. Les motifs d'intérêt communs sont relativement peu nombreux : combiner les deux approches reviendrait donc à éliminer des motifs d'intérêt.

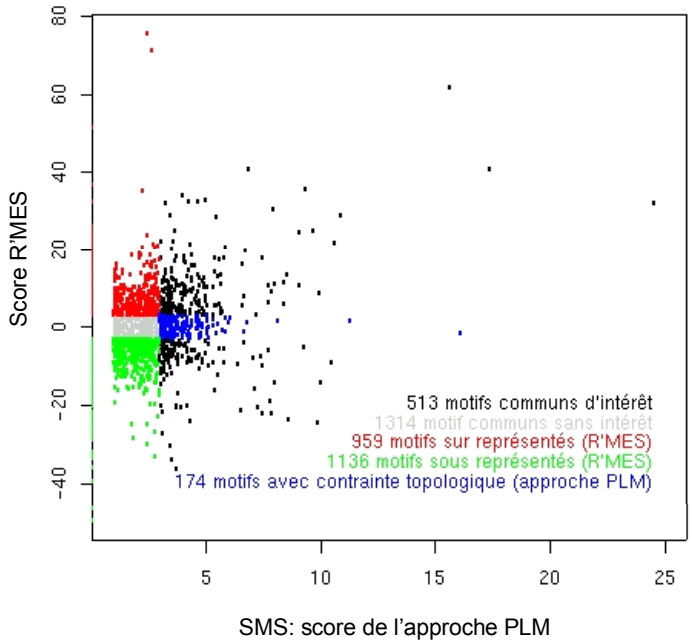


Figure 5-15 : Comparaison des scores de motifs par l'approche PLM et par R'MES.

Représentation des scores R'MES en fonction des SMS attribué par l'approche PLM pour l'ensemble des motifs de longueur 6. Chaque point représente un des 4096 motifs.

En conclusion de cette analyse, une approche de recherche de motifs sur- ou sous-représentés dans des promoteurs présente majoritairement un intérêt lors d'études du

promoteur distal. Les contraintes de position des éléments régulateurs ne sont pas ou peu présentes dans de telles régions car comme discuté précédemment, des repliements de l'ADN permettent de rapprocher les régions qui «coopèrent» fonctionnellement. Le motif ATTACA est par exemple caractérisé par une forte sur-représentation d'après le logiciel R'MES, tandis qu'il ne présente aucune contrainte de position préférentielle avec l'approche PLM (Figure 5-16 A). Le logiciel R'MES peut également être utilisé lorsque les TSS ne sont pas définis.

L'approche PLM présente tout son intérêt dans la région du promoteur central et proximal. Dans ces régions, les contraintes de position concernant les éléments régulateurs sont très importantes. Le motif ATAAAA, par exemple, est soumis à de fortes contraintes topologiques sans néanmoins être identifié comme étant un motif exceptionnel par R'MES (Figure 5-16 B).

Ainsi, les deux approches répondent à la même question : quels sont les motifs susceptibles d'être des éléments régulateurs dans mon jeu de séquences, mais n'utilisent pas la même hypothèse de travail. C'est pourquoi elles ne doivent pas être utilisées dans les mêmes circonstances pour être les plus efficaces possible. L'approche PLM pourra être utilisée pour une étude du promoteur central et proximal si le TSS est défini. R'MES pourra être utilisé pour une étude du promoteur distal, ou du promoteur central et proximal, si le TSS n'est pas défini.

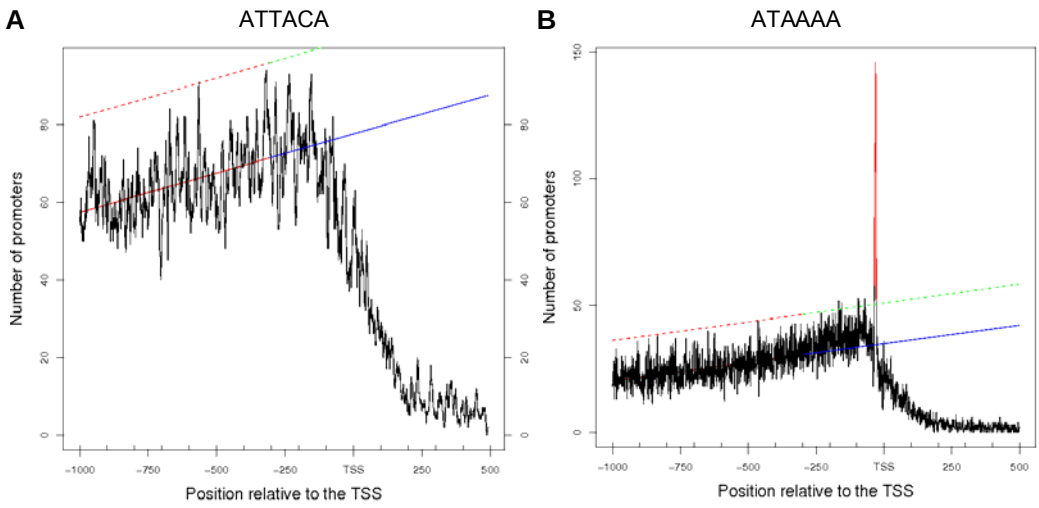


Figure 5-16 : Distributions des motifs ATTACA et ATAAAA dans le jeu de 14927 promoteurs d'*A. thaliana*.

Distributions **(A)** du motif ATTACA qui est sur-représentés lors de l'utilisation de R'MES (score supérieur à 11) et **(B)** du motif ATAAAA qui n'est pas exceptionnel lors de l'utilisation de R'MES, mais est un PLM caractérisé par un SMS de 6,7 avec l'approche PLM.

6 Cartographie des promoteurs chez *A. thaliana*

6.1 Recherche des PLM

6.1.1 Longueur des motifs étudiés

Les TFBS identifiés biologiquement peuvent être de petite taille, comme illustré Figure 5-11, où le plus court TFBS a une taille de 4 bases. Plus court encore, le dinucléotide YR, avec Y pour une pyrimidine et R pour une purine est le motif consensus proposé de l'Inr chez *A. thaliana* (Yamamoto *et al.*, 2007). Enfin, des répétitions de di- et de trinucléotides observées dans le promoteur central des plantes pourraient être des éléments impliqués dans la régulation de l'expression des gènes (Zhang *et al.*, 2006). C'est pourquoi de petits motifs de longueur 2 ont été étudiés afin de réaliser la cartographie des promoteurs d'*A. thaliana*.

Les TFBS peuvent être très longs comme les expériences le montrent Figure 5-11. Néanmoins, plus de la moitié des TFBS mis en évidence chez *A. thaliana* a une séquence de moins de 8 bases de long. L'approche PLM est efficace lorsque l'apprentissage est réalisé avec un nombre suffisant d'occurrences dans la région d'apprentissage. C'est pourquoi des motifs jusqu'à une taille de 8 bases ont été analysés. Au total, ainsi, les distributions de 87376 motifs de longueur 2 à 8 bases ont été analysées.

6.1.2 PLM identifiés - comparaison aux analyses précédentes

Parmi les 87376 motifs de longueur 2 à 8 étudiés, 5105 (5.8%) sont des PLM ayant un SMS supérieur à 3.

Différentes études à l'échelle du génome d'*A. thaliana* ont été publiées avant et au cours de ma thèse (Molina & Grotewold, 2005; Vandepoele *et al.*, 2006; Yamamoto *et al.*, 2007; Yamamoto *et al.*, 2007). L'approche utilisée par Yamamoto *et al.* (2007), c'est-à-dire l'approche LDSS (pour «Local Distribution of Short Sequences») est relativement proche de l'approche PLM. Les deux méthodologies recherchent des motifs caractérisés par des biais de positions préférentielles au sein du promoteur. Néanmoins, des différences méthodologiques notables sont listées dans la Table 6-1 et ont été discutées précédemment dans le manuscrit sur la méthodologie PLM présenté dans cette thèse à la page 70.

Les motifs d'intérêt des deux approches ont été comparés. En considérant exclusivement les motifs de longueur 6, l'approche LDSS identifie 247 motifs d'intérêt. Avec l'approche PLM, 658 motifs sont des PLM.

	L'approche PLM	Approche LDSS
Région d'apprentissage	[-1000, -300]	[-1000, -500]
Région de recherche des motifs d'intérêt	[-299, ATG]	[-500, TSS]
Longueur des motifs étudiés	2 à 8	6 et 8
Apprentissage	Régression linéaire : y = bx+a	Droite moyenne : y = a
Fenêtre glissante	Adaptée à chaque distribution	De longueur 15 pour les hexamères, 21 pour les octamères

Table 6-1 : Différences de paramètres et méthodologiques entre l'approche PLM et l'approche LDSS (Yamamoto *et al.*, 2007b).

Parmi les 247 motifs d'intérêt de l'approche LDSS, 16 ne sont pas des PLM car leurs SMS sont inférieurs à 3. Néanmoins, tous leurs SMS sont compris entre 2,3 et 2,9. Ces 16 motifs sont identifiés comme ayant des contraintes topologiques significatives par l'approche PLM mais ne sont pas étudiés du fait du seuil de SMS utilisé pour analyser les motifs les plus pertinents.

Au total, 388 motifs identifiés comme PLM avec l'approche PLM ne sont pas des motifs d'intérêt pour Yamamoto *et al.* Parmi ces motifs, 240 soit plus de 60% sont caractérisés par une position préférentielle dans l'UTR 5'. Ce premier élément de différence entre les deux approches est capital pour expliquer les différences de motif d'intérêt observées. Les motifs restants sont principalement affectés par deux facteurs introduits précédemment : (i) la taille de la fenêtre glissante qui est fixée à une longueur de 15 dans l'approche LDSS tandis qu'elle est adaptée à chaque distribution dans l'approche PLM et (ii) l'inclinaison de la pente n'est pas prise en considération dans l'approche LDSS. L'importance de ces deux paramètres a été discutée page 69. Enfin, deux derniers paramètres sont susceptibles de jouer un rôle dans la différence d'identification de motifs d'intérêt entre les deux approches : la région d'apprentissage, plus courte chez Yamamoto *et al.* (2007) peut être moins efficace et enfin, les jeux de promoteurs utilisés ne sont pas exactement identiques.

L'approche LDSS et l'approche PLM sont très similaires. Néanmoins, les résultats de cette comparaison dévoilent des différences notables. Cette comparaison des résultats des deux approches est en accord avec les résultats du manuscrit de l'article présentés page 69 : le choix des paramètres peut avoir des conséquences importantes sur l'identification ou non de motifs d'intérêt.

6.1.3 Contraintes topologiques des PLM

Chaque PLM identifié est caractérisé, entre autre, par deux contraintes topologiques qui permettent de les classer : la position préférentielle et la largeur de la fenêtre fonctionnelle. La distribution des 5105 PLM en fonctions de ces deux contraintes (Figure 6-1) met en évidence quatre régions.

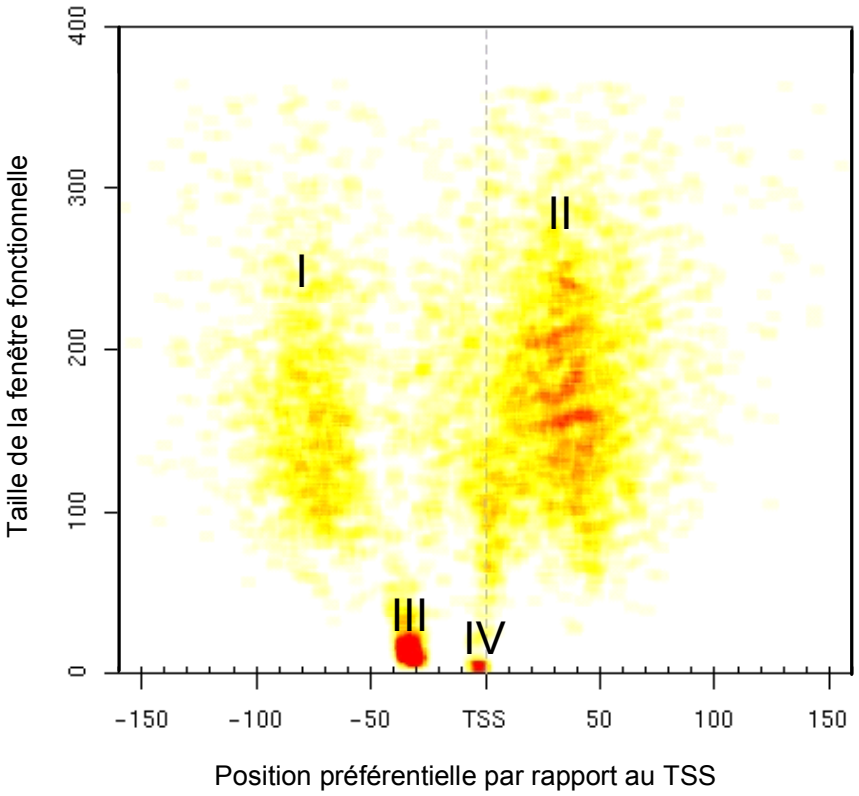


Figure 6-1 : Contraintes topologiques des 5105 PLM chez *A. thaliana*.

La densité en PLM caractérisés par les mêmes contraintes topologiques est représentée par un gradient de couleur allant du jaune (faible densité) au rouge (forte densité). Seuls les PLM dont les contraintes topologiques sont partagées avec 30 autres PLM sont représentés sur ce graphique. Les régions I à IV sont caractérisées par différentes contraintes topologiques détaillées dans le texte.

Une première caractéristique distingue les PLM ayant de larges fenêtres fonctionnelles : les régions I et II de la Figure 6-1 sont caractérisées par des fenêtres fonctionnelles de plus de 50 bases pouvant même atteindre plus de 350 bases de large. Les positions préférentielles des PLM de la région I couvrent la région [-158, -51], celles de la région II la région [-50, ATG]. Les régions III et IV concernent des PLM ayant des fenêtres fonctionnelles étroites de moins de 50 bases. La région III contient des PLM ayant des positions préférentielles dans l'intervalle [-42, -24] ; la région IV contient des PLM ayant des positions préférentielles dans l'intervalle [-6, 7]. Quatre distributions de motifs appartenant à une de ces quatre régions sont données en exemple Figure 6-2. L'ensemble des critères de ces quatre régions détaillé dans ce chapitre est résumé dans la Table 6-2.

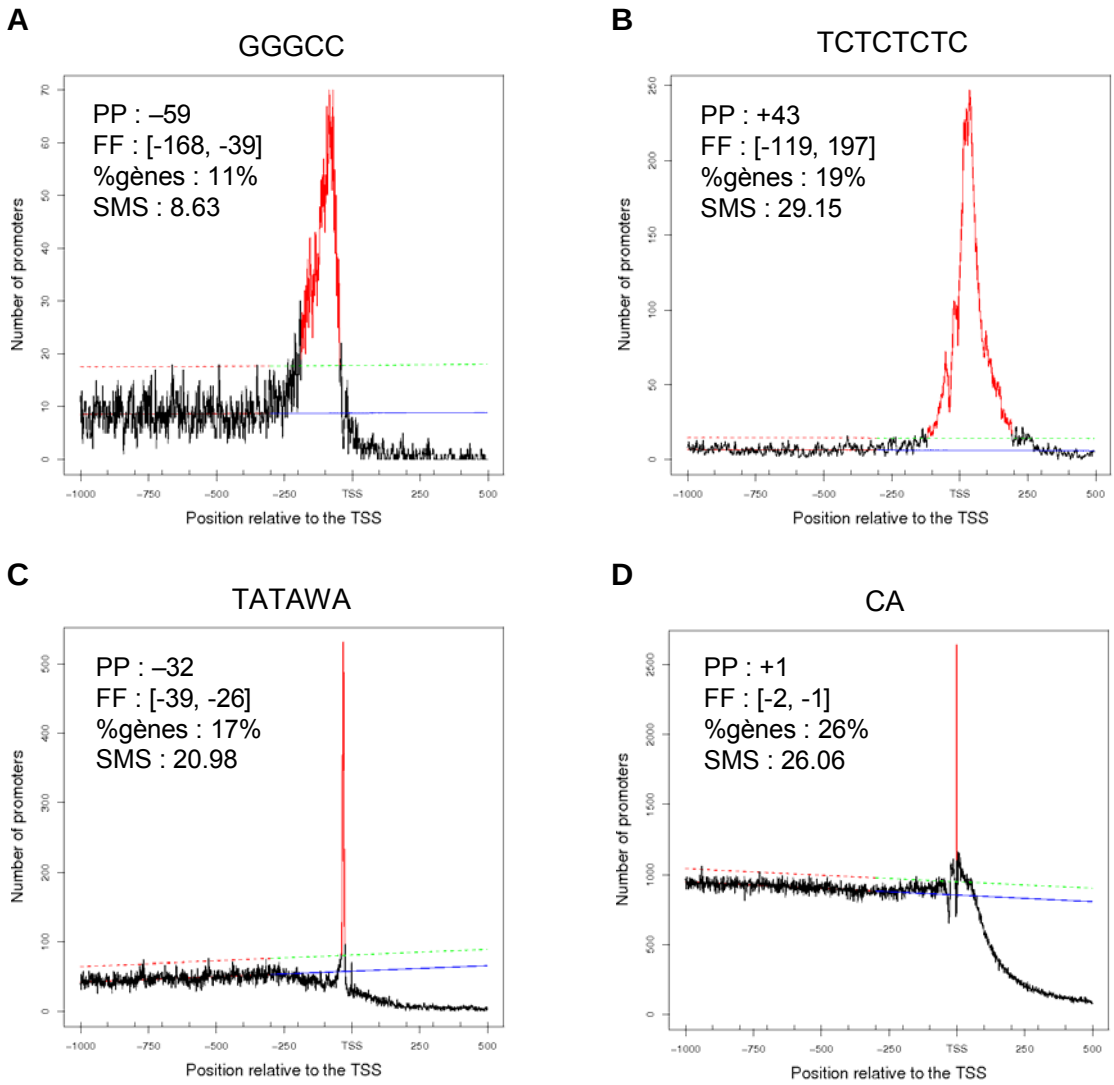


Figure 6-2 : Exemple de distributions et des caractéristiques de PLM des 4 régions déterminées Figure 6-1.

Distributions **(A)** de GGCCC, motif non orienté, PLM de la région I ; **(B)** de TCTCTCTC, motif orienté, PLM de la région II ; **(C)** de TATAWA, motif consensus de la boîte TATA, motif orienté, PLM de la région III et (**D)** de CA, motif orienté, PLM de la région IV. Pour chaque PLM, sont renseignés sa position préférentielle (PP), sa fenêtre fonctionnelle (FF), le pourcentage de gènes contenant le PLM dans leurs promoteurs (%gènes) et son score le SMS.

Région	Contraintes topologiques des PLM	Nombre de PLM (nombre ; % de leader)	Nombre de PLM orientés (%)	Contenu en base de la région	Contenu en base des PLM	p-value	Catégories de PLM		Promoteurs contenant les PLM
I	Position préférentielle dans [-158, -51]	1217 (80 ; 6.6%)	741 (60.9%)	A 35.5% C 17.1% G 15.8% T 31.6%	A 23.7% C 30.6% G 29.8% T 15.9%	NS <1e-16 <1e-16 NS	PLM de séquences différentes qui sont globalement riches en C et G		de petits groupes: 3/4 des PLM dans moins de 2.2% des promoteurs
II	Position préférentielle dans [-50, ATG] Fenêtre fonctionnelle > 50 bases	3421 (91 ; 2.7%)	3244 (94.8%)	A 29.0% C 20.1% G 16.1% T 34.8%	A 16.0% C 40.1% G 12.8% T 31.1%	NS <1e-16 NS NS	PLM ayant une séquence répétée: TFBS connus:	* Riche en C et T: Microsatellites TC et TTC * Riche en A et G: Microsatellites GA et GAA * Telo-box AAACCCTA dans la région [-53, 135] * GCC-box, GCCGCC dans la région [10, 119] * Inr CA une base en amont du TSS	98% 74% 6.4% 3.3% 17%
III	Position préférentielle dans [-42, -24] Fenêtre fonctionnelle < 50 bases	405 (4 ; 1%)	337 (83.2%)	A 35.1% C 17.9% G 13.2% T 33.8%	A 45.7% C 12.3% G 4.6% T 37.3%	<1e-16 NS NS 3e-5	PLM illustrant la présence:	* De la boîte TATA (séquence canonique TATAWA) observés dans la région [-39, -26] * Des variants de la boîte TATA partageant les mêmes contraintes topologiques que TATAWA	17% 28%
IV	Position préférentielle dans [-6, +7] Fenêtre fonctionnelle < 50 bases	67 (2 ; 3%)	65 (97%)	A 31.5% C 22.3% G 14.5% T 31.8%	A 30.9% C 35.8% G 1.8% T 31.5%	NS 2e-9 NS NS	Contenant CA ou TG illustrant:	* L'inr CA une base en amont du TSS * L'Inr TG une base en amont du TSS	17.8% 9.5%

Table 6-2 : Principales caractéristiques des PLM issus des 4 régions identifiées chez *A. thaliana*.

Dans chaque ensemble de PLM des régions I à IV, l'hypothèse d'un enrichissement en base dans les PLM (colonne 6) par rapport à la richesse en bases dans la région (colonne 5) a été testée *via* un test de comparaison des pourcentages (colonne 7, p-values du test unilatéral exact de Fisher). Les données en rouge représentent un pourcentage en base dans les PLM supérieur au pourcentage dans la région considérée. NS : pas de différence significative (p-values >1e-2).

6.2 PLM des régions I à IV et leurs spécificités

6.2.1 Région I, les PLM préférentiellement positionnés en amont de -50 : divers éléments régulateurs

a) Contraintes topologiques des PLM

Une position préférentielle en amont de la position -50 par rapport au TSS caractérise 1217 PLM (Annexe II A). La moitié des PLM ont une fenêtre fonctionnelle d'au moins 159 bases (Figure 6-3 A) et ils sont pour la majorité préférentiellement positionnés dans la région [-88, -67] (Figure 6-3 B). Les 25 PLM caractérisés par les meilleurs SMS sont disponibles dans la Table 6-3. La composition en nucléotide de l'ensemble des PLM de cette région est riche en G (29,8%) et C (30,6%) en comparaison avec la richesse attendue dans cette

région (Table 6-2). Malgré les fenêtres fonctionnelles larges, les PLM sont présents dans peu de promoteurs (Figure 6-3 C). Plus des trois-quarts des PLM de la région sont contenus dans moins de 2.2% des promoteurs analysés. Néanmoins, 93% des gènes contiennent au moins un de ces PLM présents dans un petit nombre de promoteurs. L'ensemble des PLM de la région I est donc réparti parmi tous les promoteurs d'*A. thaliana*.

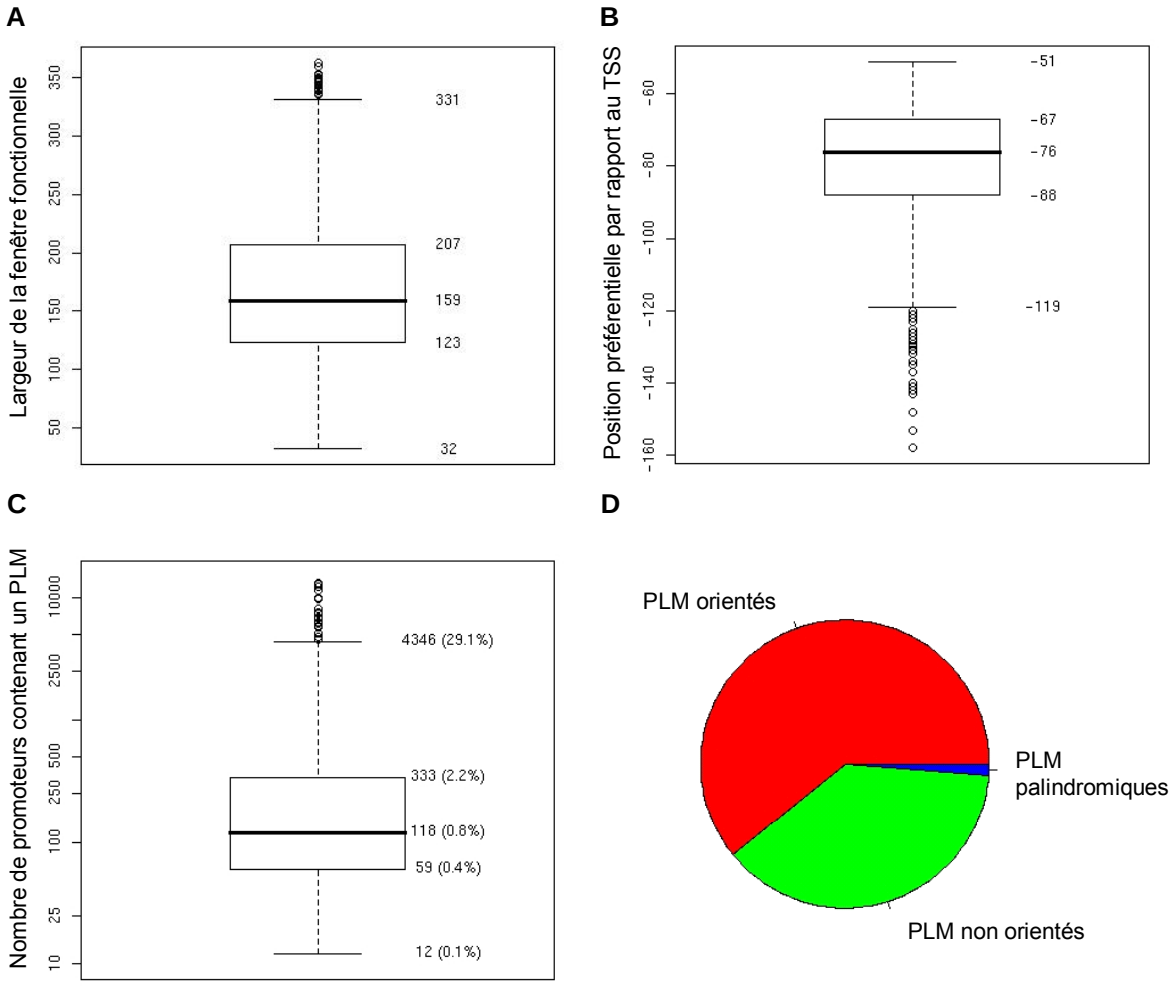


Figure 6-3 : Caractéristiques des PLM de la région I.

Rang	Séquence du PLM	SMS	Position préférentielle	Fenêtre fonctionnelle	Largeur de cette fenêtre	Nombre de gènes contenant le PLM
1	AGGCCCAT	20.44	-73	[-247, -18]	230	474
2	AAGGCCCA	14.98	-65	[-218, -12]	207	466
3	AGGCCCA	14.95	-68	[-189, -32]	158	892
4	ATAGGCCC	14.91	-80	[-321, +7]	328	168
5	GGCCCATA	14.79	-74	[-242, -4]	239	271
6	AAAGGCCC	14.47	-72	[-266, +19]	285	258
7	GCCCATTA	14.29	-86	[-171, -22]	150	371
8	GGCCCAAT	14.00	-64	[-232, -16]	217	374
9	GCCCAATA	13.75	-64	[-220, -18]	203	377
10	GGCCCAT	13.40	-62	[-163, -33]	131	665
11	GGCCCA	13.31	-61	[-192, -33]	160	1508
12	ATGGGCCT	13.19	-88	[-249, -31]	219	419
13	TAGGCCCA	13.01	-70	[-198, -30]	169	326
14	TAGGCCC	12.83	-70	[-198, -31]	168	353
15	AATGGGCC	12.54	-79	[-210, -32]	179	398
16	AAGGCCC	12.24	-70	[-211, -31]	181	491
17	ACGTGGCA	12.15	-72	[-231, +10]	241	248
18	GGCCCATT	11.76	-58	[-189, -27]	163	380
19	ACCCGACC	11.75	-61	[-208, +113]	321	113
20	CCCGACCC	11.73	-66	[-172, +110]	282	89
21	AGGCCCAA	11.68	-65	[-185, -28]	158	402
22	GGCCCAA	11.63	-67	[-170, -31]	140	672
23	AGGCCC	11.57	-62	[-190, -35]	156	978
24	GCCCATAA	11.32	-74	[-237, +5]	242	214
25	ACGGCCCA	11.24	-76	[-186, -10]	177	90

Table 6-3 : Les 25 PLM caractérisés par les meilleurs scores de la région I.

b) Orientation des PLM

L'étude de l'orientation des 1217 PLM de la région I met en évidence plus d'un tiers de motifs non orientés, c'est-à-dire non sensibles à l'orientation sur les brins d'ADN (Figure 6-3 D). Cette proportion est très élevée par rapport aux résultats des autres régions (Table 6-2).

c) Etude de la diversité des PLM

Comme nous pouvons le constater dans la Table 6-3 et aussi en Annexe II A, les PLM peuvent être chevauchants ou inclus dans d'autres PLM. L'approche PLM qui étudie les motifs de longueur 2 à 8 va par définition générer des PLM inclus dans d'autres PLM partageant les mêmes contraintes topologiques. Un «noyau» commun à plusieurs PLM peut être identifié. Une addition de bases en amont ou en aval d'une séquence noyau peut avoir pour conséquence une diminution du score associé aux motifs.

Afin d'estimer la réelle diversité en PLM dans la région I, une approche pour identifier les noyaux des PLM appelés les leaders a été développée. Un leader ne doit pas être considéré comme définissant un motif fonctionnel, mais plutôt comme étant le noyau d'une famille de motifs ayant des contraintes topologiques communes et différentes séquences flanquantes susceptibles d'être impliquées dans des fonctions spécifiques. Les TFBS présents dans les bases de données AGRIS et PLACE (Higo *et al.*, 1999; Davuluri *et al.*, 2003) représentent ces familles de motifs. Par exemple, 2 TFBS contiennent la sous séquence GGGCC.

En Annexe IV, une identification de leader est présentée comme exemple.

Au sein de la région I, 80 leaders sont présents (Annexe II A). Ainsi, 6.6% des 1217 PLM de la région I permettent de représenter la variabilité des séquences des PLM de la région. Ce pourcentage est le plus élevé des quatre régions et met en évidence une plus grande diversité des PLM au sein de la région I (Table 6-2).

d) De nouveaux éléments régulateurs?

Comme présenté en introduction, les éléments régulateurs plus éloignés du TSS ont moins de contraintes positionnelles strictes que ceux du promoteur central, dans la région [-50, +50]. De plus, ils sont bien souvent présents sur les deux brins indifféremment : ils sont moins sensibles à l'orientation que les TFBS du promoteur central (Table 6-2). Ces critères caractérisent bien les PLM de la région I. Ils sont observés dans de larges fenêtres fonctionnelles, et sont non orientés. De plus, ils sont observés dans des petits groupes de gènes. L'ensemble de ces résultats suggère que la région I contient des PLM qui sont de potentiels TFBS.

Ces PLM ont été comparés aux TFBS connus. Parmi les 1217 PLM de la région I, 31 (2.5%) ont une séquence appariée parfaitement avec un des 140 motifs consensus de TFBS identifiés biologiquement chez *A. thaliana* (Higo *et al.*, 1999; Davuluri *et al.*, 2003). Il faut noter que parmi ces 140 motifs consensus, seulement 47 sont des motifs non dégénérés de longueur 2 à 8 bases, c'est-à-dire partageant les caractéristiques de nos PLM. Parmi les séquences appariées partiellement, le degré d'appariement a été distingué en fonction de l'alignement avec la séquence du TFBS. Un appariement avec des bases non dégénérées A, C, G ou T n'a pas le même poids qu'un appariement avec une base N (Figure 6-4). Finalement, 80 PLM du groupe (6.6%) sont appariés avec des TFBS très dégénérés et 335 (27.5%) sont mésappariés avec les 140 TFBS. Ces derniers sont de potentiels nouveaux éléments régulateurs spécifiques de petits groupes de gènes.

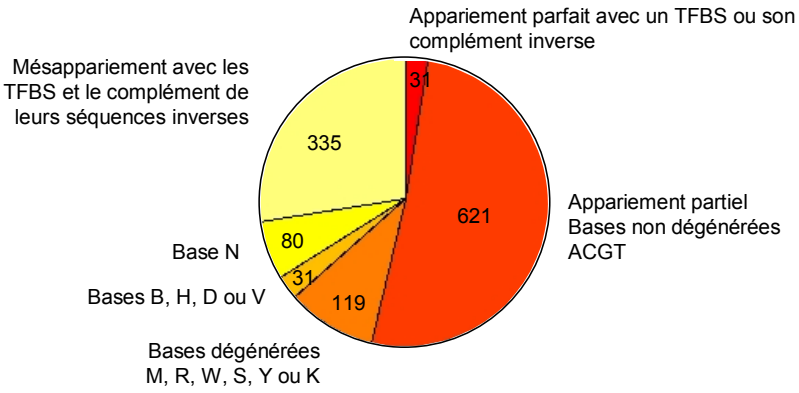


Figure 6-4 : Les appariements des PLM avec les 140 TFBS connus.

6.2.2 Région II, les PLM préférentiellement positionnés dans la région en aval de -50 : motifs répétés

La région s'étendant de la position -50 au codon ATG contient 3421 PLM (Annexe II B) ayant une large fenêtre fonctionnelle supérieure à 170 bases pour plus de la moitié des motifs (Figure 6-5 A). Ces PLM sont caractérisés par de fréquentes répétitions de dinucléotides ou de trinucléotides comme illustré dans la Table 6-4 qui contient les PLM de la région caractérisée par les SMS les plus élevés.

La contrainte d'orientation est plus importante pour les TFBS de cette région qui chevauche le TSS que pour les TFBS de la région I, faisant partie du promoteur proximal (Figure 6-5 D et Figure 6-3 D). Les TFBS de la région II sont en effet plus proches du TSS et sont positionnés dans le promoteur central où l'orientation préférentielle des motifs est plus souvent observée.

Les 3421 PLM sont riches en C (40,1%) par rapport au pourcentage attendu dans la région (Table 6-2 page 84). De plus, 19.5% des PLM ne sont constitués que de deux bases : (i) 11,6% sont ont une séquence Y_n avec Y pour C ou T, (ii) 4,4% ont une séquence M_n avec M pour A ou C, (iii) 2,2% ont une séquence R_n avec R pour A ou G et (iv) 1,3% ont une séquence S_n avec S pour C ou G. Aucune séquence W_n avec W pour A ou T, ni K_n avec K pour G ou T ne caractérisent les PLM de cette région. Malgré le caractère répété des PLM, 91 leaders sont nécessaires pour représenter la variabilité des PLM de la région II (Annexe II B). Ils représentent 2.7% des 3421 PLM de la région II. Les PLM de cette région sont donc moins diversifiés que ceux de la région I qui nécessite une plus forte proportion de leaders.

Rang	Séquence du PLM	SMS	Position préférentielle	Fenêtre fonctionnelle	Largeur de cette fenêtre	Nombre de gènes contenant le PLM
1	TCTCTCTC	29.15	+43	[-119, +197]	316	2848
2	CTCTCTCT	29.11	+44	[-123, +202]	325	2625
3	CTCTCTC	26.79	+43	[-96, +187]	283	3603
4	TCTCTCT	22.67	+37	[-81, +196]	277	4536
5	CTCTCT	21.37	+29	[-88, +189]	277	5873
6	TCTCTC	17.75	+29	[-32, +174]	206	5708
7	CTTCTCTC	16.87	+25	[-75, +209]	284	1198
8	CTCTCTCG	16.68	+50	[-56, +196]	252	530
9	CTCTCTTC	16.51	+20	[-79, +215]	294	1270
10	CTCTCTCC	16.50	+35	[-80, +170]	250	760
11	CCTCTCTC	16.46	+21	[-100, +166]	266	710
12	CTCTC	16.33	+30	[-31, +175]	206	7343
13	CTCTCTCA	15.43	+40	[-96, +200]	296	959
14	TTCTCTCT	15.23	+38	[-74, +206]	280	2209
15	TCTTCTTC	15.13	+26	[-68, +219]	287	3125
16	TCTCT	15.05	+40	[-71, +190]	261	9889
17	TCTC	14.80	+40	[-67, +170]	237	12592
18	TCT	14.49	+29	[-58, +162]	220	14648
19	TC	14.44	+29	[-27, +129]	156	14902
20	CTCTATAT	14.37	-39	[-79, -8]	72	186
21	CTTCTTCT	14.28	+28	[-52, +211]	263	2481
22	TCTCTCTT	14.24	+38	[-78, +206]	284	2167
23	CT	14.07	+30	[-58, +97]	155	14916
24	TCTCTCCT	13.89	+39	[-68, +147]	215	575
25	TTCTCTC	13.73	+27	[-45, +170]	215	2967

Table 6-4 : Les 25 PLM caractérisés par les meilleurs scores de la région II.

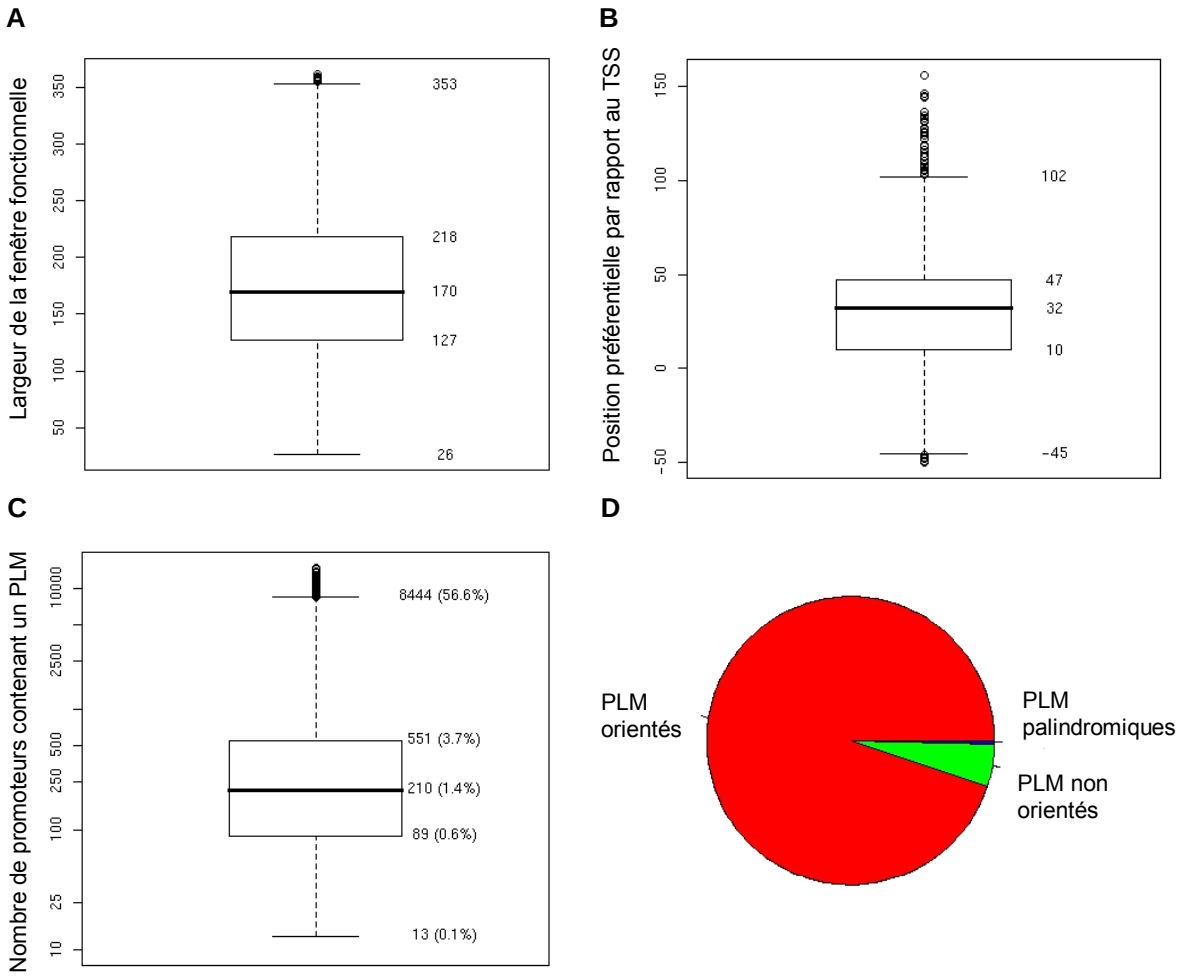


Figure 6-5 : Caractéristiques des PLM de la région II.

a) Microsatellites TC, TTC, GA et GAA dans les UTR 5'

Les 397 PLM ayant une séquence Y_n sont caractérisés par des SMS parmi les plus élevés de l'analyse globale. Par exemple, le SMS de TC_4 est proche de 30 (Table 6-4 rang 1 et Figure 6-5 B). La taille médiane des plus longues répétitions de motifs Y_n est de 11 bases et jusqu'à 112 Y consécutifs sont observés (Figure 6-6 Y). Les PLM R_n sont des compléments inverses des Y_n. Les PLM R_n et Y_n sont tous observés sur le brin sens aux mêmes emplacements de l'UTR 5'. Plus de 93% des motifs Y_n sont en effet orientés préférentiellement sur le brin sens. Les répétitions riches en bases T et C sont plus fréquemment observées. De plus, les caractéristiques des PLM R_n se distinguent des PLM Y_n par une moins grande longueur de répétitions (Figure 6-6 R et Y). Les répétitions R_n ont une taille médiane de 8 bases, tandis que les Y_n ont une taille médiane de 11 bases.

Parmi les 14927 promoteurs, 98% contiennent au moins un des 343 PLM Y_n avec n supérieur à 5. Ce résultat distingue ces PLM comme étant parmi les plus observés de la région II (Figure 6-5 C). En comparaison, les PLM R_n sont présents dans 74% des gènes (Table 6-2 page 84).

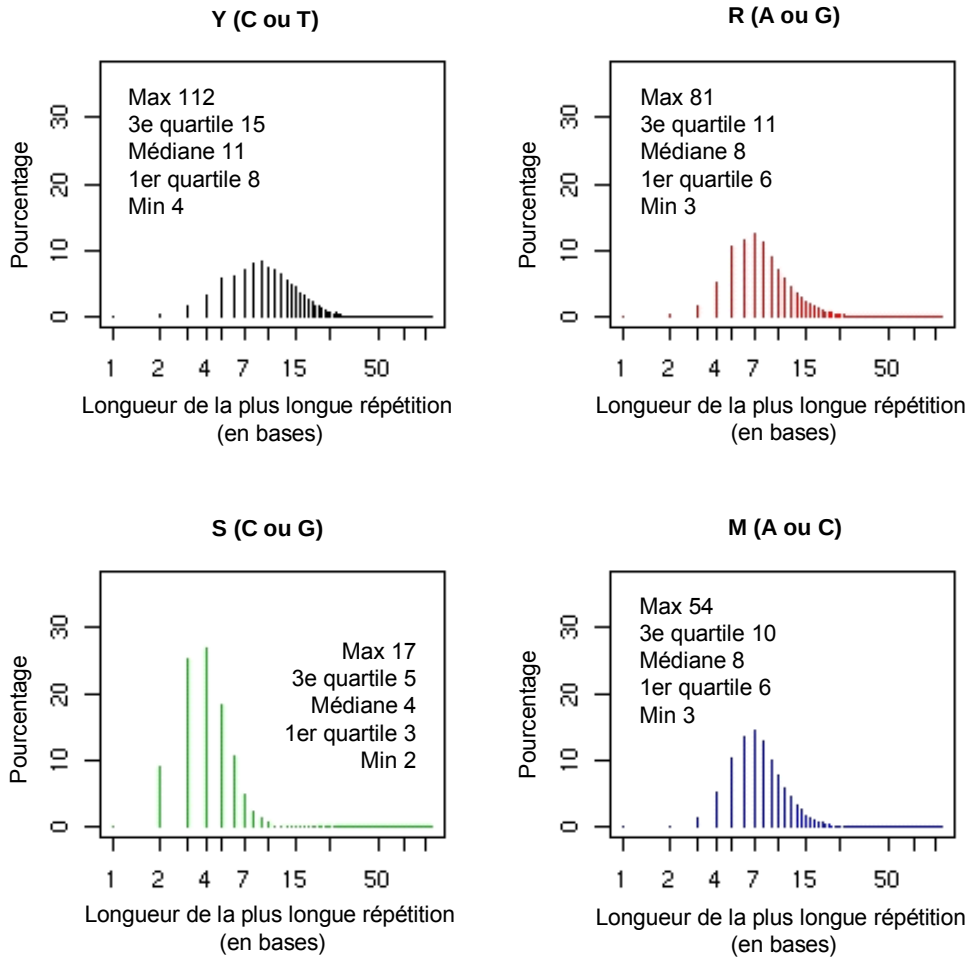


Figure 6-6 : Longueurs des répétitions de Y, R, S et M dans les UTR 5'.

Recherche de la plus longue séquence exclusivement constituée de 2 nucléotides dans les UTR 5' des 14927 gènes. Représentation de l'histogramme des longueurs maximales observées. Les 1er et 3^{e} quartiles mettent en évidence les longueurs séparant 25% de données les plus petites et 75% les plus grandes. Les 1er, 3^{e} quartiles et la médiane (2^{e} quartile) distinguent quatre groupes de données ayant le même nombre de gènes.

i) Répétitions parfaites et imparfaites

Les répétitions TC et TTC sont les plus fréquentes. Parmi les 14927 promoteurs, 38% contiennent le motif $(TC)_3$ et 42% le motif $(TTC)_2$. Les deux séquences sont répétées parfaitement jusqu'à 41 fois pour TC et jusqu'à 36 fois pour TTC. En complément des répétitions parfaites, des répétitions imparfaites sont fréquemment observées : l'insertion d'une base au sein d'un motif répété est illustrée dans la Table 6-4 rang 7 et 9 avec l'insertion d'une pyrimidine ou rang 8 et 13 avec l'insertion d'une purine. Parmi les 3421 PLM de la région, 2277 sont constitués de plus de 60% de pyrimidines, 1087 d'entre eux contenant une seule purine. Ils représentent les PLM riches en C et T qui peuvent être parfaitement répétés ou non.

Ces répétitions parfaites ou imparfaites de pyrimidines principalement conservées dans l'UTR 5' des gènes représentent des microsatellites. Chez les plantes, ces séquences sont présentes dans toutes les régions du génome, mais elles sont tout particulièrement observées dans les UTR 5' (Morgante *et al.*, 2002; Fujimori *et al.*, 2003). Au sein des 14927 promoteurs d'*A. thaliana*, 93% contiennent ces séquences. C'est la classe de motif la plus observée au sein des promoteurs.

ii) Etude des données d'expression en fonction de la longueur des répétitions des microsatellites

Une corrélation entre la longueur des microsatellites et l'expression des gènes a été observée chez *H. sapiens*, *Neisseira gonorrhoeae* et *Gallus gallus* (Xu & Goodridge, 1998; Li *et al.*, 2004). A l'échelle de quelques gènes chez *A. thaliana*, cette même corrélation a été observée (Zhang *et al.*, 2006). Néanmoins, aucune étude à l'échelle génomique n'a aujourd'hui établi de corrélation entre l'intensité d'expression des gènes et la longueur des microsatellites

Parmi les 14927 gènes dont les bornes du promoteur sont supportées par des données expérimentales, 11161 ont des données d'expression disponibles dans la base de données CATdb et ont été analysées sur les puces CATMA (Crowe *et al.*, 2003; Gagnot *et al.*, 2008). La Table 6-5 liste les groupes de gènes analysés, contenant différentes longueurs de GA, GAA, TC et TTC parfaitement répétés. Quels que soient les motifs étudiés GA, GAA, TC ou TTC, l'intensité d'expression des groupes de gènes n'est pas corrélée avec la longueur de la répétition (Figure 6-7). La comparaison des groupes extrêmes de taille 3 et supérieure à 10 ne présente aucun biais non plus.

Longueurs des répétitions parfaites de	3	4	5	6	7	8	9	10	11	12	13	14	15
GA	3742	3943	2841	1161	918	374	282	115	125	297	-		
GAA	2025	3999	3825	1850	1111	659	338	193	161	375	-		
TC	1854	2710	3262	1943	1725	853	590	315	279	163	131	436	-
TTC	1308	2553	3036	2440	1727	1195	716	450	320	199	144	135	371

Table 6-5 : Jeux de gènes contenant différentes tailles de répétitions de GA, GAA, TC et TTC.

Pour l'ensemble des promoteurs, la longueur maximale de répétition parfaites des 4 séquences microsatellites GA, GAA, TC et TTC a été recherchée. La dernière colonne renseignée d'une ligne correspond au nombre de gènes contenant au minimum la longueur de microsatellite indiquée.

Dans une seconde étape, l'impact de la présence de ces répétitions sur le pourcentage d'hybridation, c'est-à-dire le pourcentage d'expériences où un gène s'exprime a été étudié, c'est-à-dire le pourcentage d'expériences où une expression a été détectée. Pour les différents jeux de gènes étudiés, la longueur des répétitions parfaites de microsatellites n'est pas non plus corrélée avec le pourcentage d'hybridation.

En conclusion, les microsatellites retrouvés dans la région des UTR 5' sont très présents au sein des gènes d'*A. thaliana*. Néanmoins, aucune corrélation n'a pu être mise en évidence entre leur nature, leur longueur et l'expression des gènes. Ces séquences sont présentes sur de très larges fenêtres fonctionnelles allant de régions en amont du TSS jusqu'à l'UTR 5'. Leur présence pourrait avoir un rôle conformationnel et modifier la stabilité du double brin d'ADN de la région des UTR 5'.

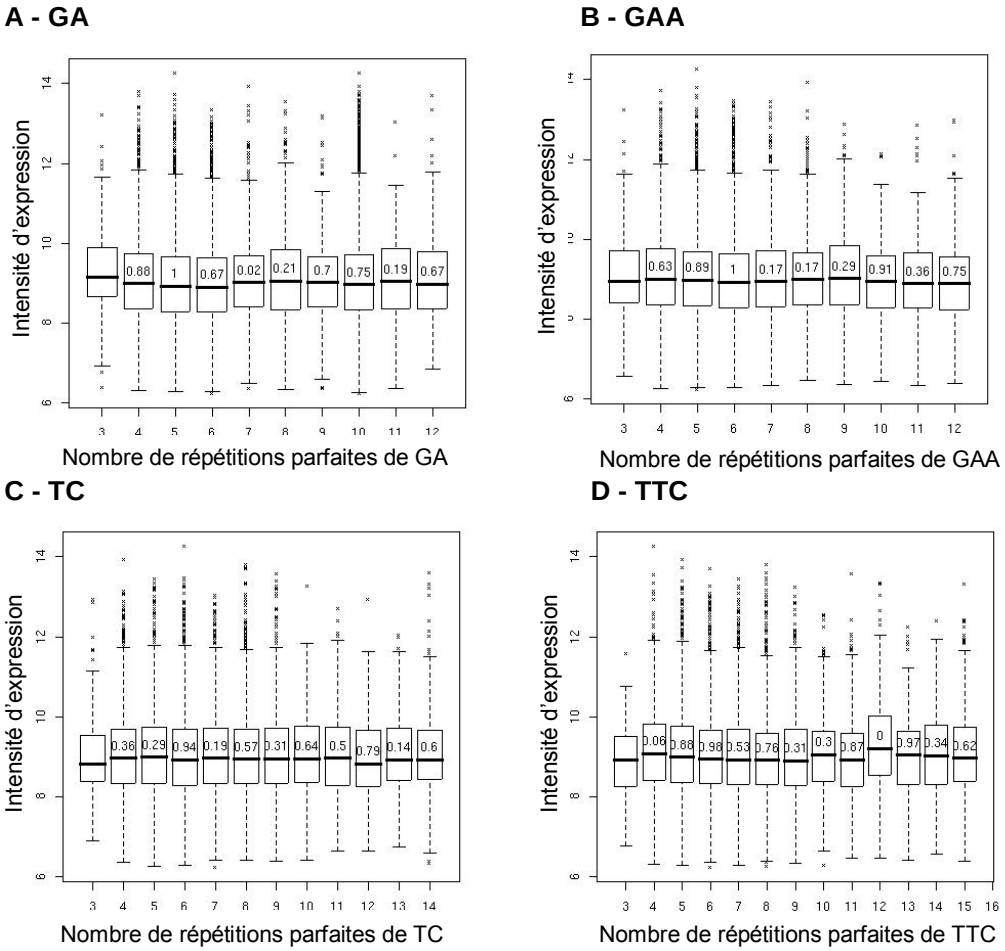


Figure 6-7 : Expression des gènes en fonction du nombre de répétitions de GA, GAA, TC et TTC.

Représentation des données d'intensité d'expression des gènes contenant différentes tailles de répétitions de séquences GA **(A)**, GAA **(B)**, TC **(C)** et TTC **(D)** (données de la Table 6-5). Chaque distribution des données d'un jeu de gènes contenant n répétitions (appelé jeu 1) est comparée à la distribution du jeu de gènes contenant n+1 répétitions (Jeu 2) par un test de Wilcoxon unilatéral. Les p-values obtenues sont notées au dessus de la valeur médiane des représentations du jeu 2.

b) TFBS connus dans les UTR 5'

En complément des microsatellites, dans la région II, deux catégories de PLM exclusivement constitués de bases C et G ou A et C ont été mises en évidence. Les mêmes analyses que celles effectuées avec les PLM Y_n et R_n ont été réalisées pour les PLM M_n et S_n. La médiane des longueurs maximales de répétitions de M et de S est respectivement de 4 bases et de 8 bases (Figure 6-6). Ces PLM révèlent la présence d'éléments régulateurs identifiés biologiquement chez *A. thaliana* et qui ne sont pas répétés comme le sont les microsatellites. Parmi les TFBS qui sont des PLM, seulement ceux présentés ci-dessous sont préférentiellement positionnés dans l'UTR 5'.

i) BoîteTélo

Les 148 PLM M_n ont une médiane de largeur de pic de 142 bases et leur position préférentielle est à l'emplacement du TSS. Néanmoins, deux catégories de PLM M_n ont été identifiées :

Certains PLM contiennent plusieurs séquences CA comme par exemple CACCAAAC ou CCACACAC. Le dinucléotide est répété trois fois consécutives dans 3% des promoteurs, mais très rarement plus. L'existence de ces PLM contenant CA à l'emplacement du TSS est en accord avec la forte présence de l'élément initiateur CA ou Inr-CA. L'importance de cet élément initiateur sera discutée ci-dessous, dans la partie 6.2.4 de ce travail.

Les autres PLM M_n sont constitués d'une suite de A puis d'une suite de C, comme AAACCC par exemple. Ces motifs s'apparient avec la boîte Télo de motif consensus AAACCCTA (Audic & Claverie, 1998; Manevski *et al.*, 1999; Manevski *et al.*, 2000). Le motif AAACCCTA est lui-même un PLM de la région II. Cet élément régulateur est impliqué dans la régulation de l'expression des gènes codant des protéines ribosomales. La distribution du motif AAACCCTA prédit une fenêtre fonctionnelle dans l'intervalle [-53, 136] et une position préférentielle 27 bases en amont du TSS (Figure 6-5 B). Ce PLM est présent dans 6.4% des gènes. Les motifs s'appariant à la séquence AAACCC sont orientés préférentiellement sur le brin sens. Le complément inverse de AAACCCTA, par exemple, n'est pas identifié comme PLM.

ii) Boîte GCC

Les 43 PLM S_n ont une largeur de pic de 171 bases et sont préférentiellement observés 48 bases en aval du TSS. Ces PLM s'apparient au motif GCCGCC ou le chevauchent. GCCGCC est le motif consensus de la boîte GCC (Shinozaki & Yamaguchi-Shinozaki, 2000). Cet élément régulateur est observé dans les gènes de réponse à l'éthylène. L'orientation des PLM associés à cet élément régulateur est sensible au brin. La boîte GCC est préférentiellement observée sur le brin sens. Le PLM GCCGCC est préférentiellement observé 49 bases en aval du TSS (Figure 6-8). Ce motif est présent dans 2% des promoteurs dans sa fenêtre fonctionnelle [10, 119].

En résumé, deux catégories de PLM ont été identifiées lors de l'étude de la région II. Les PLM riches en Y ou en R représentent des microsatellites dans les UTR 5' des gènes et correspondent aux PLM fortement représentés au sein des promoteurs dans la Figure 6-5 C et à ceux orientés sur un brin préférentiel Figure 6-5 D. Les autres PLM montrent la présence de TFBS dans les UTR 5'. Ils sont présents dans peu de promoteurs.

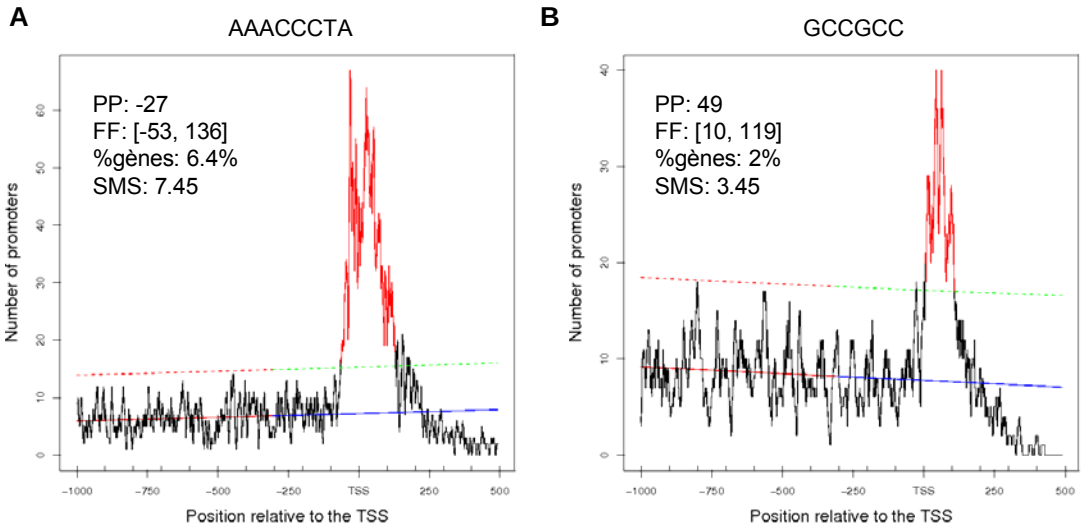


Figure 6-8 : Distribution de 2 TFBS connus de la région II.

Distribution de deux TFBS identifiés comme étant des PLM dans la région II : la boîte Télo, de séquence consensus AAACCCTA **(A)** et la boîte GCC, de séquence consensus GCCGCC **(B)**. Pour chaque PLM, sont renseignés sa position préférentielle (PP), sa fenêtre fonctionnelle (FF), le pourcentage de gènes contenant le PLM dans ses promoteurs (%gènes) et son score le SMS.

6.2.3 Région III, les PLM à fortes contraintes topologiques en amont du TSS

a) Contraintes topologiques des PLM

Dans la région [-42, -24], 405 PLM ont été identifiés (Annexe II C). Les fenêtres fonctionnelles de ces PLM sont de faibles envergures, puisque 47% d'entre elles sont inférieures à 13 bases (Figure 6-9 A). De plus ces 405 PLM sont localisés dans une région très étroite : 62% d'entre eux sont observés préférentiellement dans l'intervalle [-31, -35] (Figure 6-9 B). Les 25 PLM ayant les meilleurs SMS de cette région sont disponibles dans la Table 6-6.

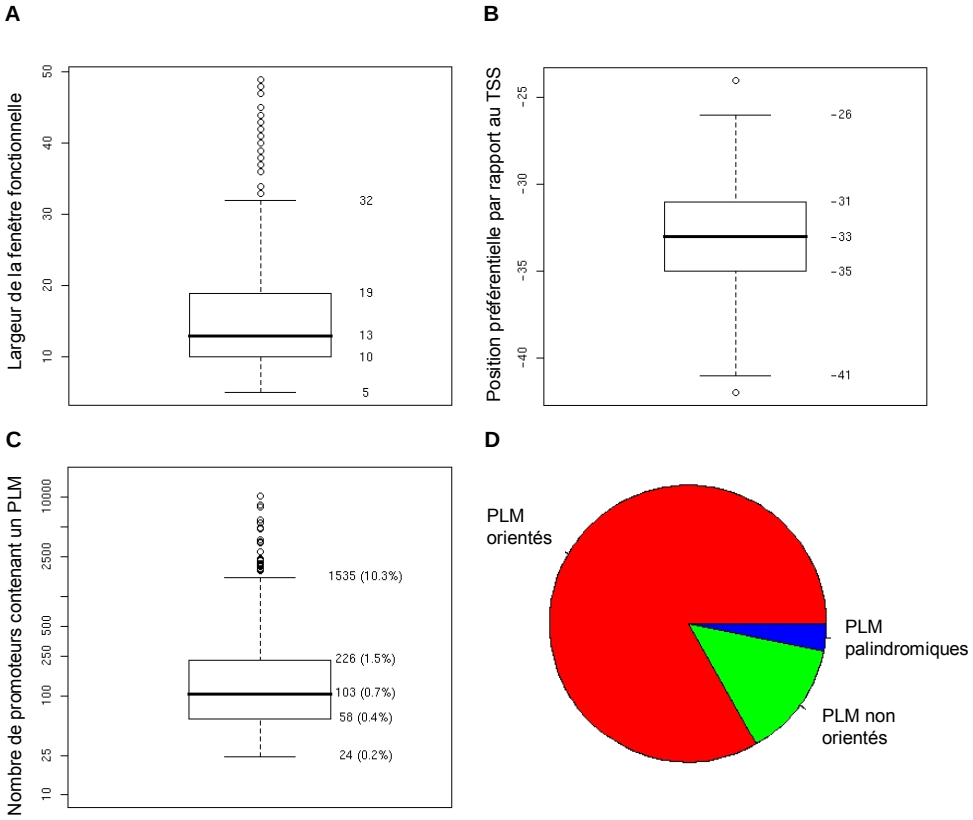


Figure 6-9 : Caractéristiques des PLM de la région III.

Rang	Séquence du PLM	SMS	PP	PP_de TATAWA	FF	Largeur de cette fenêtre	Promoteurs contenant le PLM
1	CTATAAAT	32.08	-33	-32	[-43, -20]	24	272
2	TATAAATA	28.39	-32	-32	[-57, -22]	36	569
3	CTATATAA	26.62	-34	-33	[-53, -21]	33	278
4	CTATAAA	23.57	-33	-32	[-38, -26]	13	425
5	TCTATATA	22.69	-34	-32	[-56, -25]	32	319
6	ATAAATAC	22.42	-31		[-48, -20]	29	219
7	CTATATA	21.41	-33	-32	[-41, -27]	15	511
8	TATATAAA	20.58	-32	-32	[-41, -25]	17	472
9	TCTATAAA	20.48	-34	-32	[-44, -23]	22	190
10	ATATAAAC	19.03	-31	-30	[-41, -21]	21	219
11	CTATATAT	18.42	-33	-32	[-55, -25]	31	307
12	TATAAAT	18.31	-32	-32	[-37, -26]	12	645
13	CTCTATA	17.77	-35		[-51, -25]	27	241
14	CTATAAAA	15.44	-33	-32	[-43, -22]	22	152
15	ATATAAA	15.37	-32	-31	[-39, -26]	14	718
16	TCTATAT	15.24	-35		[-41, -28]	14	296
17	TATAAA	15.21	-32	-32	[-37, -26]	12	1535
18	TAAATAC	14.88	-31		[-36, -24]	13	219
19	TATATA	14.60	-33	-33	[-37, -26]	12	1401
20	TATA	14.52	-31		[-36, -26]	11	3682
21	ATAAATA	14.42	-32		[-40, -26]	15	663
22	CTATAA	14.13	-34		[-37, -27]	11	499
23	TATATAA	13.75	-33	-33	[-36, -27]	10	690
24	CCTATA	12.96	-35		[-41, -28]	14	247
25	CCTATAA	12.54	-35		[-44, -24]	21	139

Table 6-6 : Les 25 PLM caractérisés par les meilleurs scores de la région III.

Les séquences TATAWA inclues dans des PLM sont soulignées dans la 2e colonne. Pour chaque sous séquence, la position du premier T de TATAWA est renseignée dans la 5e colonne.

b) Boîte TATA : un élément régulateur observé dans moins de 18% des gènes

Seuls 4 leaders, soit 1% des 405 PLM de la région III, représentent la variabilité des PLM de la région III. Il s'agit des PLM ATA, TAT, AA et TTTA (Annexe II C). Ces leaders exclusivement constitués de bases A et T sont en accord avec la richesse en A (45.7%) et T(37.3%) au sein des 405 PLM par rapport à la région III (Table 6-2 page 84). En effet, la boîte TATA est attendue dans cette région (Patikoglou *et al.*, 1999). Cet élément régulateur est caractérisé par le motif consensus TATAWA avec W pour A ou T (Figure 6-2 C page 83). Une position préférentielle 32 bases en amont du TSS et une fenêtre fonctionnelle dans la région stricte [-39, -26] caractérisent ce PLM qui est présent dans 2606 promoteurs. Cette présence le positionne parmi les motifs les plus représentés de la région, puisque 51% des PLM sont observés dans moins de 103 promoteurs, 103 étant la médiane observée sur le jeu des 405 PLM (Figure 6-9 C).

Le motif TATAWA comprend le motif palindromique TATATA. La présence de tels motifs est plus fréquemment observée dans la région III que dans les autres régions. Treize PLM palindromiques constitués des bases A et T sont identifiés dans cette région (Figure 6-9 D). Dans la fenêtre fonctionnelle [-39, -26], 1537 promoteurs contiennent TATAAA et 567 autres son complément inverse TTTATA. Ces résultats indiquent une forte préférence de la boîte TATA pour le brin sens (p-value < 1e-16).

i) Localisation dans la région stricte [-33, -31]

Parmi les 405 PLM de la région III, 87 incluent le motif TATAWA. Ces motifs sont caractérisés par des contraintes topologiques parmi les plus strictes (Table 6-6, rang 1 à 5 par exemple). Pour ces 87 PLM, une position préférentielle «corrigée» a été déterminée. Elle correspond à la position préférentielle du premier T de la sous séquence TATAWA si elle existe (Table 6-6, colonne 5). Le T1 préférentiel est quasiment exclusivement localisé dans la région [-33, -31]. 1354 promoteurs contiennent le motif TATAWA dans cette fenêtre [-33, -31], soit plus de la moitié des 2606 promoteurs contenant une boîte TATA dans sa fenêtre fonctionnelle [-39, -26].

ii) Comparaison aux autres approches

Avec l'approche PLM, 17.5% des gènes d'*A. thaliana* contiennent une boîte TATA définie par la présence de TATAWA en [-39, -26]. Des analyses utilisant une matrice de fréquence ont prédit la présence de la boîte régulatrice dans 28.8% des promoteurs d'*A. thaliana* (Molina & Grotewold, 2005). La différence entre les deux pourcentages peut être expliquée par la dissimilitude majeure entre les deux approches plus que par une non adéquation entre les jeux de promoteurs. En effet, avec les deux approches, 15% des promoteurs contiennent la séquence TATAWAWA dans la régio [-50, -1]. L'utilisation d'une matrice entraîne une tolérance de motifs dégénérés. Nous pouvons supposer que des variants de la boîte TATA doivent être assignés comme boîtes TATA lors de l'approche par une matrice de comptage des nucléotides. De tels PLM ont été assignés comme variants de la boîte TATA lors de l'approche PLM. Comme discuté lors de l'identification et de la caractérisation de ces variants à la page 103, les gènes contenant différents variants ne partagent pas toutes les caractéristiques fonctionnelles ni d'expression des gènes contenant le PLM TATAWA.

L'hypothèse posée lors de cette thèse est qu'un motif doit partager plus qu'une similarité de séquence pour être considéré comme une boîte TATA fonctionnelle. Un motif ressemblant à une boîte TATA mais dont la présence n'induit pas les caractéristiques fonctionnelles partagées par les gènes contenant une boîte TATA (Basehoar *et al.*, 2004; Moshonov *et al.*, 2008) est donc plus susceptible d'être un variant de séquence mais pas un variant fonctionnel de la boîte TATA.

Ainsi, l'estimation du nombre de gènes contenant une boîte TATA par une approche considérant les positions semble être plus réaliste que l'estimation par matrice.

iii) Boîtes TATA simples, étendues et chevauchantes

Les 87 PLM contenant une séquence TATAWA ont des fenêtres fonctionnelles qui couvrent globalement l'intervalle [-62, -13]. Dans cet intervalle, certains promoteurs contiennent plusieurs occurrences de TATAWA chevauchantes ou non. Les cas non chevauchant sont minoritaires par rapport à l'ensemble des séquences. Ils ne caractérisent que 257 promoteurs (1.7% de 14927), ils n'ont pas été considérés.

Ont été identifiés 1119 promoteurs (7.5% de 14927) qui contiennent une unique occurrence de TATAWA et 1040 promoteurs (7.0%) qui contiennent une séquence étendue TATATAWA, sans autre WA suivant le dernier A. De plus, 190 promoteurs (1.3%) contiennent trois chevauchements de TATAWA ou plus. La plus longue séquence $(TA)_nWA$ est obtenue pour le gène AT1G13300. Elle est constituée de 21 occurrences chevauchantes et est longue de 46 bases. L'emplacement de ces boîtes TATA chevauchantes Figure 6-10 illustre la large région sur laquelle ces motifs peuvent s'étendre. Elles vont de la position -62 à la position -8 par rapport au TSS.

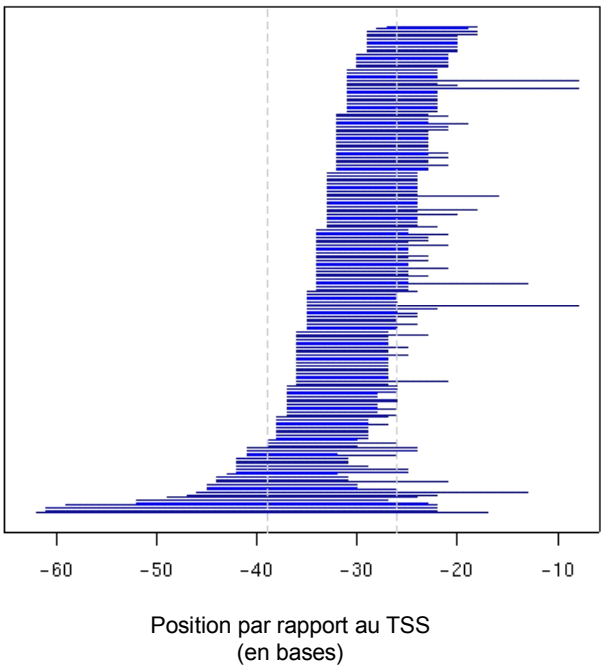


Figure 6-10 : Extension des boîtes TATA chevauchantes.

Représentation de l'emplacement des boîtes TATA se chevauchant sur une longueur minimale de 10 bases pour former une séquence (TA)nWA n étant au moins égal à 5. Les traits en pointillé gris indiquent les bornes de la fenêtre fonctionnelle de la boite TATA [-39, -26].

Ces résultats nous indiquent que certaines boîtes TATA pourraient avoir une histoire évolutive similaire à celle des microsatellites, et s'étendre *via* la répétition du dinucléotide TA (Zhang *et al.*, 2006). Néanmoins, leur extension est moindre par rapport aux extensions de TC ou de TTC dans les UTR 5' (Cf. page 89) : une sélection négative pourrait s'opposer à leur extension.

iv) Environnement des boîtes TATA

Plusieurs PLM de la région III sont caractérisés par un appariement partiel avec la séquence TATAWA, comme par exemple le PLM CTCTATA, rang 13 de la Table 6-6. Au total, 92 PLM ont une séquence TATAWA tronquée de une base en 3' ou en 5'. Il est intéressant de noter que 70% des promoteurs contenant ces motifs tronqués contiennent effectivement une séquence TATAWA complète. Ces PLM peuvent décrire l'environnement de la boîte TATA.

La sur-représentation de la base C en amont des boîtes TATA (Table 6-7) concorde avec des analyses précédentes (Shahmuradov *et al.*, 2003; Molina & Grotewold, 2005) et justifie la présence du PLM CTATAAAT au premier rang des SMS de la région (Table 6-6). Cette base C est d'autant plus observée en amont des boîtes TATA étendues qui sont 42.5% à être précédées d'une base C. Aucune autre sur-représentation n'a été identifiée en amont. En aval des boîtes TATA étendues et chevauchantes, les bases C et G sont sur-représentées. En aval des boîtes simples, le A est sur-représenté (39.9%) et le C est sous-représenté (13.5%). En complément, notons que cette dernière base est néanmoins sur-représentée en aval des TATAWAA : 27% de ces séquences sont suivies d'un C. Ce pourcentage étant significativement supérieur au pourcentage observé dans la région. Ainsi, les boîtes TATA sont souvent entourées de bases C. Ces résultats suggèrent un rôle fonctionnel de cette base qui pourrait être impliquée dans la reconnaissance de l'élément régulateur par un facteur de transcription.

A - les boîtes TATA simples

	-2		-1			+1		+2	
A	27.6%	5e-8	22.3%	1e-20	Boîtes TATA simples TATAWA	39.9%	4e-4	11.3%	<1e-30
C	22.5%	4e-5	32.6%	<1e-30		13.5%	5e-5	24.9%	2e-9
G	13.1%	NS	11.9%	NS		14.0%	NS	24.8%	<1e-30
T	36.8%	NS	33.1%	NS		32.6%	NS	39.0%	1e-4

B - les boîtes TATA étendues

	-2		-1			+1		+2	
A	29.0%	2e-7	17.5%	<1e-30	Boîtes TATA Etendues TATATAWA	19.8%	2e-27	34.5%	NS
C	24.5%	5e-8	42.5%	<1e-30		37.3%	<1e-30	27.3%	4e-14
G	31.2%	NS	15.3%	NS		22.6%	1e-16	19.5%	7e-9
T	33.3%	NS	24.7%	2e-10		20.3%	1e-21	18.7%	4e-27

C - les boîtesTATA chevauchantes

	-2		-1			+1		+2	
A	22.1%	6e-5	21.6%	3e-5	Boîtes TATA chevauchantes $(TA)_nWA$	20.0%	4e-6	34.2%	NS
C	23.2%	NS	37.4%	7e-11		29.5%	3e-5	24.7%	NS
G	14.2%	NS	20.0%	NS		21.6%	5e-4	17.9%	NS
T	40.5%	NS	21.1%	9e-5		28.9%	NS	23.2%	9e-4

Table 6-7 : Environnement des boîtes TATA.

Etude des bases observées en amont et en aval des TATA simples (**A**), étendues (**B**) et chevauchantes (**C**). Les pourcentages de bases observées ont été comparés aux pourcentages de bases attendues dans la région II (Table 6-2 page 84). Les tests exacts de Fisher avec correction de Bonferroni ont mis en évidence les bases observées qui sont plus ou moins fréquemment présentes, respectivement en rouge et vert. NS pour les p-values non significatives supérieures à 5e-2.

c) Variants de la boîte TATA

Des variants de la boîte TATA ont été identifiés comme étant des sites de fixation permettant de former le complexe d'initiation de la transcription (TIC pour Transcription Initiation Complex) (Joshi, 1987; Singer *et al.*, 1990). Ces séquences peuvent avoir différents impacts sur l'expression des gènes (Moshonov *et al.*, 2008). Parmi les PLM de la région III, 268 de longueur supérieure à 5 ont des séquences qui ne s'apparient pas avec la séquence de la boîte TATA. Ces PLM peuvent contribuer à une meilleure connaissance de l'environnement de la boîte TATA, comme discuté précédemment, mais ils peuvent également représenter des PLM de séquences très similaires à la boîte TATA, et ayant les mêmes contraintes topologiques : des variants de la boîte TATA. L'identification et la caractérisation de ces séquences sont détaillées page 103.

6.2.4 Région IV. Les PLM à fortes contraintes topologiques chevauchant le TSS

a) Contraintes topologiques des PLM

Au sein de l'intervalle [-6, 7], 62 PLM sont caractérisés par un pic fin et ont une fenêtre fonctionnelle de moins de 50 bases (Annexe II.D). Ces PLM sont majoritairement caractérisés par des fenêtres fonctionnelles bien plus étroites que 50 bases. La médiane des tailles de fenêtres fonctionnelles est de 3 bases comme l'illustre la Figure 6-11 A. Ces 62 PLM sont présents dans des petits groupes de promoteurs (Figure 6-11 B). De plus, ils sont orientés sur un seul brin d'ADN pour 60 (97%) d'entre eux (Figure 6-11 D). Les 25 PLM ayant les meilleurs SMS sont détaillés dans la Table 6-8. Une forte présence de motifs préférentiellement positionnés dans la fenêtre [-1, -4] par rapport au TSS est observée (Figure 6-11 C). En comparaison avec la richesse en bases attendue dans la région (Table 6-2 page 84), les 62 PLM sont riches en base C (35.8%).

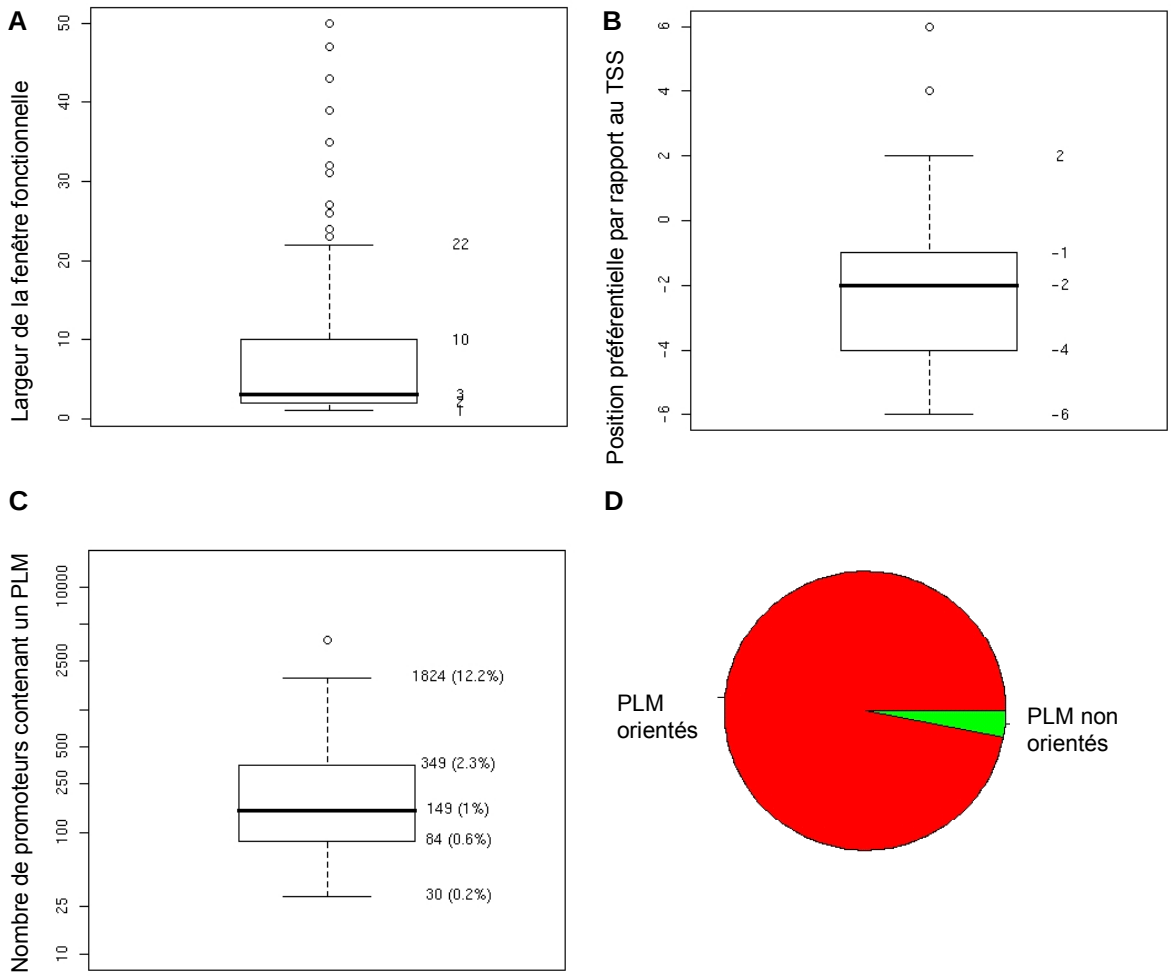


Figure 6-11 : Caractéristiques des PLM de la région IV.

b) Deux motifs leaders : CA et TG, préférentiellement positionnés en -1

Le dinucléotide CA est le PLM caractérisé par le meilleur score de la région (Figure 6-2 page 83). Il est inclus dans 57 des 62 PLM (Table 6-8). Deux leaders (3% des 67 PLM), les dinucléotides CA et TG, décrivent la variabilité en PLM de la région IV. Ces deux motifs, rang 1 et 24 dans la Table 6-8, sont des séquences complément inverse l'une de l'autre. CA et TG sont tous les deux observés préférentiellement au même emplacement, une base en amont du TSS. Le dinucléotide CA est sur-représenté sur la fenêtre fonctionnelle [-2, -1]. Au total, 1102 promoteurs (7,4%) contiennent ce motif en -2 mais il est tout particulièrement observé en -1 où 2639 promoteurs (17.7%) permettent de former le pic fin de la Figure 6-2 D page 83. Le dinucléotide TG n'est exclusivement sur-représenté qu'une base en amont du TSS, position à laquelle 1416 promoteurs le contiennent. La prédominance de CA une base en amont du TSS est en accord avec les résultats d'une analyse de promoteurs validés biologiquement de 217 plantes dicotylédones (Shahmuradov *et al.*, 2003) et d'une étude de 12242 promoteurs d'*A. thaliana* (Alexandrov *et al.*, 2006).

Rang	Séquence du PLM	SMS	Position préférentielle	Fenêtre fonctionnelle	Largeur de cette fenêtre	Nombre de gènes contenant le PLM
1	CA	18.59	-1	[-2, -1]	2	3741
2	TCA	18.24	-2	[-3, -2]	2	1670
3	CTCA	14.35	-3	[-4, -3]	2	473
4	TCAC	12.87	-2	[-6, -2]	5	649
5	CTTCA	12.37	-4	[-5, -3]	3	274
6	TTCA	11.85	-3	[-4, -3]	2	642
7	TCTTCA	11.70	-5	[-6, -4]	3	169
8	TCATC	11.33	-2	[-3, -2]	2	232
9	TCAT	11.24	-2	[-3, -2]	2	584
10	CAT	10.82	-1	[-2, -1]	2	1138
11	CAA	9.95	-1	[-1, -1]	1	931
12	TTCAC	9.14	-3	[-6, -3]	4	243
13	CTCAT	8.32	-3	[-4, -3]	2	159
14	CTTCAT	7.93	-5	[-6, -3]	4	116
15	CATCA	7.78	-1	[-2, -1]	2	166
16	TCTCAT	7.64	-4	[-5, -4]	2	88
17	TCATCA	7.61	-2	[-3, -2]	2	107
18	TCACA	7.34	-2	[-2, -2]	1	124
19	TCTTCAT	7.30	-5	[-25, -3]	23	194
20	TTCAT	7.23	-3	[-4, -3]	2	228
21	CTCATC	7.06	-4	[-11, -2]	10	136
22	TCATT	6.74	-2	[-3, -2]	2	208
23	CATT	6.44	-1	[-1, -1]	1	287
24	TG	6.28	-1	[-1, -1]	1	1416
25	TTCATC	6.27	-4	[-5, -2]	4	102

Table 6-8 : Les 25 PLM caractérisés par les meilleurs scores de la région IV.

Les séquences CA inclues dans des PLM sont soulignées dans la 2e colonne.

c) Inr-YR présent dans la moitié des gènes

L'élément initiateur ou Inr est observé à un emplacement strict proche du TSS (Corden *et al.*, 1980; Lo & Smale, 1996). Chez *A. thaliana*, le dinucléotide YR a été proposé comme Inr (Yamamoto *et al.*, 2007). Avec l'approche PLM, les 4 motifs YR, c'est-à-dire des motifs CA, TG, TA et CG ont été étudiés. Chacun des motifs est un PLM, mais seuls CA et TG sont assignés à la région IV. Le PLM TA est un motif de la région III, région où une richesse en A et en T est attendue. Le PLM CG est assigné à la région II du fait de sa large fenêtre fonctionnelle. Néanmoins, dans la région stricte du TSS, les 4 dinucléotides sont caractérisés par un pic une base en amont du TSS (Figure 6-12).

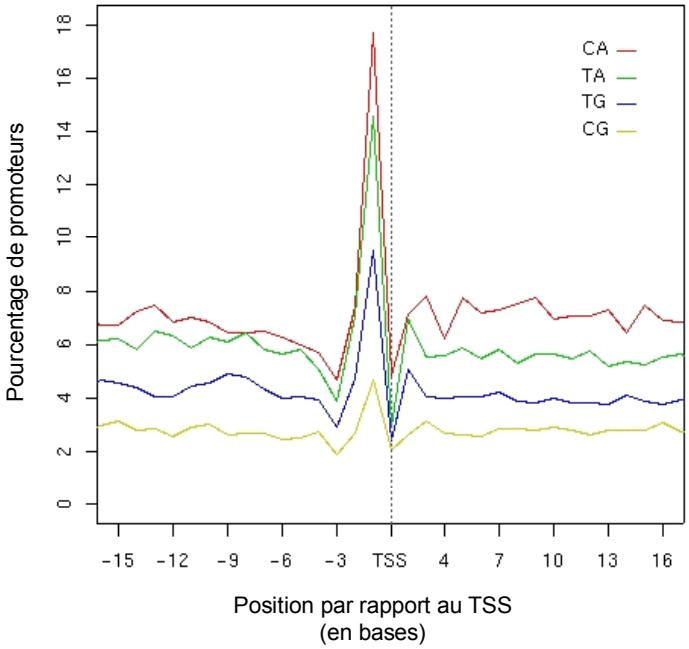


Figure 6-12 : Distribution des 4 dinucléotides YR dans la région du TSS.

Les 14927 promoteurs d'*A. thaliana* sont analysés et une fenêtre glissante de une base de large est utilisée pour représenter ces distributions.

Les 12 autres dinucléotides possibles ne sont pas caractérisés par un tel pic une base en amont du TSS. Ce résultat met en évidence une présence majoritaire du CA présent dans près de 18% des promoteurs en -1. Les autres dinucléotides sont moins observés (Table 6-9).

	CA	TA	TG	CG
Pourcentage de promoteurs contenant dinucléotides en -1	17.8%	14.6%	9.5%	4.7%

Table 6-9 : Présence des dinucléotides YR une base en amont du TSS.

L'étude des trinucléotides n'a pas permis d'étendre le motif consensus de l'Inr, en accord avec la recherche des leaders qui sont des dinucléotides. Seuls CA et TG sont leaders. En complément, l'extension du motif YR a été considérée mais n'a pas présenté non plus d'amélioration du SMS.

Les PLM de la région IV sont donc des motifs qui signalent la présence de l'Inr-YR. Les résultats qualitatifs sont en accord avec les analyses précédemment menées sur le génome d'*A. thaliana* (Alexandrov *et al.* 2006,.Yamamoto *et al.* 2007b) Néanmoins, comme lors de l'analyse des biais attendus dans la région du TSS (partie 5.1.2, page 64), les résultats quantitatifs diffèrent. En effet, la présence de l'Inr est observée dans 88% des promoteurs d'après Yamamoto *et al.* (2007b) tandis qu'il est observé dans moins de la moitié des 14927 promoteurs lors de ce travail de thèse. L'étude menée par Yamamoto *et al.* (2007a) a été réalisée exclusivement à partir de TSS dont la position a été défini par des ADNc pleine longueur du RIKEN. C'est pourquoi l'hypothèse d'un décalage de la position du TSS discutée page 64 peut expliquer cette différence.

7 Etude approfondie de la région du promoteur central

Dans sa fenêtre fonctionnelle [-39, -26], la boîte TATA est présente dans 17.5% des promoteurs d'*A. thaliana*. Néanmoins, les gènes ne contenant pas de boîte TATA sont tout de même reconnus (i) par les protéines se fixant sur les boîtes TATA (TBP pour «TATA-Binding Proteins») et (ii) par divers facteurs de transcription associés à ces TBP (Tsai & Sigler, 2000). Certains variants sont également impliqués dans la formation du TIC (Nakamura *et al.*, 2002), mais peu d'études ont été réalisées pour les identifier et les caractériser fonctionnellement. C'est pourquoi une partie des travaux de cette thèse a été consacrée à l'identification et à la caractérisation de variants de la boîte TATA puis à leur évolution putative par rapport à la boîte TATA. Dans cette première partie, l'approche mise au point pour étudier ces variants est présentée.

Ces résultats ont été présentés dans un poster dans le cadre de la conférence internationale ISMB/ECCB («international conference on Intelligent Systems for Molecular Biology / European Conference on Computational Biology») à Stockholm (Annexe V).

7.1 Approche pour identifier les variants de la boîte TATA

L'identification de variants de la boîte TATA peut être réalisée en considérant exclusivement la séquence génomique (Moshonov *et al.*, 2008). Dans le cadre de cette thèse, plusieurs contraintes ont été ajoutées afin de définir une liste de variants les plus probablement fonctionnels qui seront par la suite appelés par abus de langage des variants fonctionnels. Ils doivent avoir les mêmes contraintes topologiques que la boîte TATA, être observés dans un jeu de promoteurs sans boîte TATA et être conservés au sein de génomes divergents.

7.1.1 Etude des promoteurs sans boîte TATA

Au sein de la région III, la présence de PLM qui ont les mêmes contraintes topologiques que la boîte TATA mais qui ne s'apparient pas à la séquence TATAWA ont été observés. Ces PLM peuvent être des variants fonctionnels de la boîte TATA, mais aussi de simples motifs chevauchant une boîte TATA. En effet, de courts motifs (TAT par exemple) comme des motifs de 6 bases de long (NTATAT par exemple) peuvent être des PLM uniquement parce qu'ils chevauchent une séquence TATAWA. La distribution du motif TATATG en est un exemple. Ce PLM est caractérisé par les mêmes contraintes topologiques que la boîte TATA lors de l'étude du jeu complet de promoteurs (Figure 7-1 A). Néanmoins, le pic est perdu lors de l'étude des promoteurs ne contenant pas de boîte TATA (Figure 7-1 B). Afin d'identifier exclusivement les variants fonctionnels et d'éviter ces PLM dont les contraintes topologiques sont dues à un chevauchement avec la boîte TATA, le jeu de promoteurs étudié ne contenait pas de séquence TATAWA dans sa fenêtre fonctionnelle. C'est dans ce contexte que les variants caractérisés par une à trois bases différentes de TATAWA ont été recherchés.

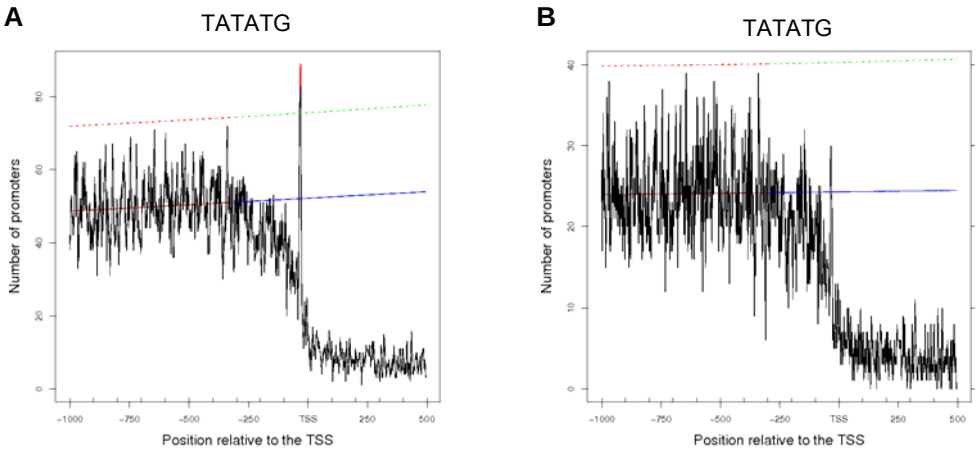


Figure 7-1 : Conséquences possible du chevauchement d'un motif avec une séquence TATAWA.

Distribution du motif TATATG dans les 14927 promoteurs d'*A. thaliana* **(A)** et dans les12321 promoteurs ne contenant pas de boîte TATA **(B)**.

7.1.2 Conservation des variants dans deux génomes divergents

Une des hypothèses de travail était que les variants fonctionnels de la boîte TATA sont comme la boîte TATA probablement conservés au sein de différents organismes. Afin d'identifier une liste de variants, deux génomes ont été étudiés en parallèle : le génome d'*A. thaliana* et le génome d'*O. sativa*, deux plantes qui ont divergé il y a environ 150 millions d'années (Wolfe *et al.*, 1989). Retrouver des motifs communs aux deux organismes étudiés et partageant les mêmes contraintes topologiques que la boîte TATA est un argument supplémentaire permettant de proposer l'hypothèse que ces motifs sont des variants fonctionnels de la boîte TATA.

La même démarche que lors de la construction du jeu de promoteurs d'*A. thaliana* a été appliquée pour construire le jeu de promoteurs d'*O. sativa*. Les données mises à disposition par le TIGR R.3, The Institute for Genomic Research (Ouyang *et al.*, 2007), et intégrées dans FLAGdb^{++} ont été exploitées. Ainsi, 18012 promoteurs d'*O. sativa* ont été analysés.

La distribution du motif TATAWA présente les mêmes contraintes topologiques chez *A. thaliana* et chez *O. sativa*. Lors de l'étude des promoteurs des deux organismes, la séquence canonique de la boîte TATA est préférentiellement positionnée 32 bases en amont du TSS, et sa fenêtre fonctionnelle est la région [-39, -26]. Chez *A. thaliana*, 2606 promoteurs (17.5%) contiennent une boîte TATA dans cette fenêtre fonctionnelle. Chez *O. sativa*, dans la même région, 2601 promoteurs (14.4%) contiennent l'élément régulateur.

7.2 Identification de 15 variants de la boîte TATA

Les résultats obtenus et les jeux de données exploités aux différentes étapes de la recherche de variants sont illustrés Figure 7-2.

Arabidopsis thaliana

Oryza sativa

①

Recherche des séquences TATAWA

14927 promoteurs

18012 promoteurs

TATATA, TATAAA

②

Recherche parmi 32 motifs ayant une substitution par rapport à TATAWA de variantsΔ1 potentiels

12321 promoteurs sans boîte TATA

15411 promoteurs sans boîte TATA

TATATT

AATAAA , CATAAA , GATAAA
TTTAAA , TTTATA
TAAATA, TACATA
TATTTA , TATTAA
TATATC, TATAAG

= 11 TATAΔ1

③

Recherche parmi 211 motifs ayant deux substitutions par rapport à TATAWA de variantsΔ2 potentiels

9512 promoteurs sans boîte TATA ni variant TATAΔ1

13494 promoteurs sans boîte TATA ni variant TATAΔ1

GATTAA, GTTAAA, TTTAAG, TTTATC

CTTAAA, GATAAG, TACTTA, TCTTAA

CAAATA, TACAAG, TATACC, TATCTC, TATTTC, TGCAAA

= 4 TATAΔ2

④

Recherche parmi les motifs ayant trois substitutions par rapport à TATAWA des variantsΔ3 potentiels

9113 promoteurs sans boîte TATA ni variant TATAΔ1 ou TATAΔ2

13208 promoteurs sans boîte TATA ni variant TATAΔ1 ou TATAΔ2

TCTTAT, CTAAAA

ACAAAA, CAAAAG

= 0 TATAΔ3

Figure 7-2 : Schéma global de l'identification des variants de la boîte TATA.

La recherche des variants de la boîte TATA s'arrête étape 4 où aucun nouveau variant n'est identifié. La colonne de gauche représente les jeux de promoteurs utilisés chez *A. thaliana*, la colonne de droite chez *O. sativa*. Les cercles bleus et rouges montrent respectivement les variants potentiels identifiés chez *A. thaliana* et *O. sativa*. Les motifs en gras entourés des deux cercles sont les variants fonctionnels, identifiés par les deux jeux de promoteurs.

7.2.1 Variants ayant une substitution par rapport à la séquence TATAWA

Toutes les séquences proches de TATAWA ne sont pas fonctionnelles, et *in vitro*, les variants ayant une seule substitution sont plus observés (Kiran *et al.*, 2006). C'est pourquoi, dans une première étape, les variants de la boîte TATA ayant une séquence avec une seule substitution par rapport à la séquence TATAWA sont recherchés. Ils seront appelés par la suite les variants TATAΔ1.

La séquence TATAWA correspond à deux motifs, TATATA et TATAAA. Toutes les substitutions de chacun des deux motifs ont donc été considérées indépendamment. En effet, tandis qu'une substitution pourra mener à l'identification d'un variant si le W de TATAWA est un A par exemple, cela ne sera pas toujours le cas si le W est un T (ou inversement). GATATA par exemple n'est pas un PLM, tandis que GATAAA en est un (Figure 7-3). Ce résultat met en évidence l'intérêt de considérer des motifs non dégénérés.

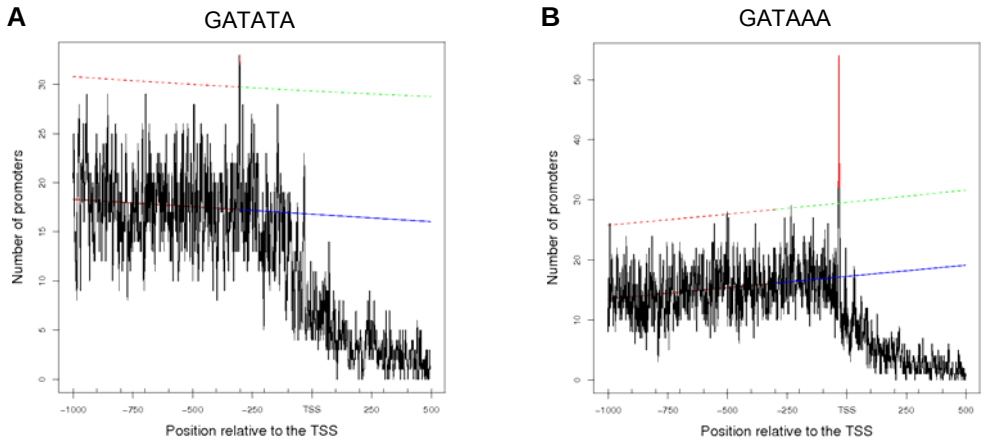


Figure 7-3 : Distribution de GATATA et GATAAA dans les promoteurs d'*A. thaliana* sans boîte TATA.

Dans les jeux de 12321 promoteurs d'*A. thaliana* et de 15411 promoteurs d'*O. sativa* ne contenant pas de boîte TATA, les distributions des 32 motifs ayant une substitution par rapport aux deux séquences TATAWA ont été analysées (Figure 7-2). Parmi ces 32 motifs, 11 sont des TATAΔ1 c'est-à-dire qu'ils partagent les contraintes topologiques strictes de la boîte TATA chez les deux organismes étudiés (Table 7-1).

Par la suite, les variants seront distingués en fonction de la position de la substitution qui a permis leur obtention. Par exemple, le TATAΔ1 GATAAA a été obtenu en substituant le 1er T de TATAAA en G. Il sera donc appelé TATAΔ1 T1.

Les 11 séquences des TATAΔ1 et leurs occurrences sont listées dans la Table 7-1. En premier lieu, nous pouvons constater que les substitutions de T en A ou de A en T sont les plus fréquentes. Ces TATAΔ1 représentent 6 des TATAΔ1 et sont bien plus présents dans les promoteurs que les 5 autres TATAΔ1. Ce résultat est en accord avec les observations qui ont pu être réalisées lors d'une étude chez *Nicotiana tabacum* (Kiran *et al.*, 2006). Des

substitutions en C ou en G sont présentes, mais elles restent minoritaires et sont principalement en bordure de la séquence canonique de la boîte TATA. La 5e base de la séquence, la base W, est très conservée, et aucune substitution n'est observée à cette position. Les bases C et G sont contre sélectionnées aux positions centrales.

	T1	A2	T3	A4	W5	A6
A	689	-	849	-	-	-
	AATAAA		TAAATA			
C	244		179			292
	CATAAA		TACATA			TATATC
G	217					187
	GATAAA					TATAAG
T	-	611+514	-	498+274	-	
		TTTATA+TTTAAA		TATTTA+TATTAA		

Table 7-1 : Les TATAΔ1 et leurs occurrences.

Les 11 TATAΔ1 sont ordonnés en fonction de la base qui a été substituée entre leur séquence et la séquence TATAWA. Notons que pour la colonne des substitutions du A2 par exemple, deux variants avec une substitution en T sont observés. Ceci est du à la considération de la base W comme A puis comme T. Pour chacun des 11 TATAΔ1 de la boîte TATA, le nombre au dessus de leur séquence correspond au nombre de promoteurs d'*A. thaliana* le contenant dans la fenêtre fonctionnelle [-39, -26].

Sur l'ensemble des 14927 gènes étudiés chez *A. thaliana*, 3733 contiennent au moins un des 11 TATAΔ1, et parmi eux, 2809 ne contiennent pas de boîte TATA canonique. Ces chiffres montrent l'existence possible dans les promoteurs d'une boîte TATA et de un ou plusieurs variants. Chaque promoteur peut en effet contenir plusieurs motifs ayant des contraintes topologiques dans la région [-39, -26].

7.2.2 Variants ayant 2 substitutions ou plus par rapport à TATAWA

Dans une deuxième étape, les TATAΔ2 ont été recherchés. Les 211 motifs ayant 2 substitutions par rapport à la séquence TATAWA ont été considérés. Pour les mêmes raisons que celles présentées lors de l'identification des TATAΔ1, la distribution des motifs a été étudiée dans les jeux de promoteurs ne contenant ni la boîte TATA ni les 11 TATAΔ1 dans la région [-39, -26]. Ainsi, 9512 promoteurs d'*A. thaliana* (63.7%) et 13494 promoteurs d'*O. sativa* (74.9%) ont été analysés. Quatre motifs partagent les mêmes contraintes topologiques que la boîte TATA et sont conservés dans les deux organismes (Figure 7-2 étape 3). Ce sont des TATAΔ2. Ils sont présents dans 471 promoteurs d'*A. thaliana*, dont 399 qui ne contiennent ni boîte TATA ni TATAΔ1 dans la fenêtre fonctionnelle [-39, -26].

La recherche de TATAΔ3 avec la même approche n'a pas permis d'identifier de nouveaux variants (Figure 7-2 étape 4).

Grâce à cette étude complète, une liste de 15 variants TATAΔ1 et TATAΔ2 a donc été constituée. Aucun autre motif de longueur 6 n'a été identifié comme variant sur le jeu de promoteurs sans boîte TATA, ni variant TATAΔ1 ou TATAΔ2.

7.2.3 Evolution des boîtes TATA et des variants

Les substitutions ayant permis d'obtenir les TATAΔ1 à partir des 2 séquences de boîte TATA créent un possible lien évolutif entre des motifs (Figure 7-4). Ceci est également

valable pour tous les TATAΔ2 qui proviennent de un ou deux TATAΔ1 avec lesquels ils ont une base de substituée. Le nombre de promoteurs contenant un des 11 TATAΔ1 ou un des 4 TATAΔ2 est représenté dans la Figure 7-4. Les promoteurs contenant un TATAΔ2 sont moins nombreux que ceux contenant un TATAΔ1, tout comme les promoteurs contenant les TATAΔ1 sont moins nombreux que les promoteurs contenant une boîte TATA.

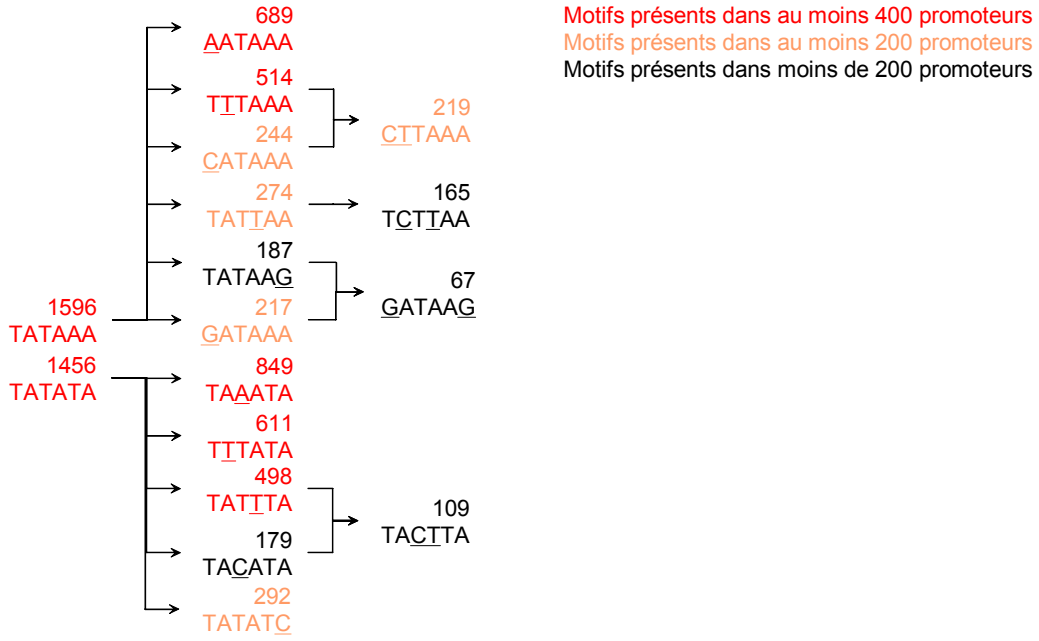


Figure 7-4 : Evolution de la séquence de la boîte TATA en variants.

Les motifs de la colonne 1 sont des boîtes TATA canoniques, ceux de la colonne 2 les TATAΔ1 et ceux de la colonne 3 les TATAΔ2. La valeur en indice représente le nombre de promoteurs d'*A. thaliana* contenant le motif dans la région [-39, -26]. Les bases soulignées sont les bases substituées par rapport à la séquence TATAWA. Chaque flèche montre qu'une substitution permet de passer du motif de gauche au motif de droite.

7.2.4 Etude fonctionnelle *in silico* des variants de la boîte TATA

Afin de révéler les caractéristiques fonctionnelles potentielles induites par la présence des variants fonctionnels, des jeux de gènes dont les promoteurs ne contiennent qu'un seul motif dans la région [-39, -26] ont été constitués (Table 7-2). Seuls les groupes de gènes constitués d'au moins 100 identifiants ont été conservés afin de pouvoir réaliser des tests pertinents.

	Boîte TATA	AATAAA	CATAAA	GATAAA	TTTAAA	TTTATA	TATTAA	TATTTA	TATAAG
Nombre de gènes	1496	424	147	143	269	134	206	224	119

Table 7-2 : Jeux de gènes contenant soit une unique boîte TATA soit un unique variant.

Les jeux contenant un minimum de 100 promoteurs ont été sélectionnés pour pouvoir réaliser des analyses statistiques. Les bases en rouge correspondent aux bases substituées par rapport à la séquence TATAWA.

Chacun des groupes de gènes de la Table 7-2 a été étudié pour caractériser fonctionnellement les différents TATAΔ1. Les comparaisons réalisées par la suite consistent à confronter les résultats d'un groupe de gènes aux résultats de l'ensemble des autres gènes pour lesquels un promoteur a été défini.

a) Fonction des gènes

L'annotation des gènes est un indicateur du rôle des TFBS qui régulent leur expression (Walther *et al.*, 2007; Boden & Bailey, 2008). Le projet GO (Gene Ontology) met à la disposition une ressource pour représenter des annotations concernant les gènes d'une manière standardisée (Ashburner *et al.*, 2000). La GO contient trois catégories d'informations : la fonction moléculaire des gènes, les processus biologiques dans lesquels ils sont impliqués et les composants cellulaires vers lesquels les protéines qu'ils codent sont dirigées. Les annotations de la GO propres aux gènes d'*A. thaliana* sont disponibles au TAIR (Swarbreck *et al.*, 2008) et permettent de disposer d'annotations classifiées en 16 localisations subcellulaires, 15 fonctions moléculaires et 14 processus biologiques. Ces données ont été exploitées afin de caractériser les groupes de gènes de la Table 7-2.

i) Processus biologiques des gènes

La Table 7-3 A détaille les processus biologiques dans lesquels sont impliqués les groupes de gènes contenant exclusivement un variant dans la fenêtre fonctionnelle [-39, -26] ou exclusivement une boîte TATA. Pour un groupe de gènes donné, le pourcentage de gènes annotés dans un processus biologique est comparé au pourcentage caractérisant tous les autres gènes (Table 7-3 première colonne). Les tests unilatéraux exacts de Fisher ont mis en évidence une implication plus grande des gènes contenant une boîte TATA dans des processus de réponses à des stress ou des stimulus. Ce résultat est en accord avec un résultat récemment publié (Walther *et al.*, 2007). Les gènes contenant une boîte TATA sont plus fréquemment impliqués dans des processus biologiques spécifiques que d'autres gènes, comme cela avait été observé chez *S. cerevisiae* et *H. sapiens* (Yang *et al.*, 2007;

Moshonov *et al.*, 2008). En parallèle, ils sont moins impliqués dans des processus biologiques basiques, comme le métabolisme des protéines, de l'ADN et de l'ARN ou la transduction de signaux.

Les gènes contenant un variant de la boîte TATA sont moins nombreux que les gènes contenant une boîte TATA (Table 7-2). Néanmoins des biais sont observés pour le variant AATAAA, un TATAΔ1 T1. Aucun des autres variants ne présentent de biais fonctionnels. Pour les catégories d'annotations fonctionnelles qui sont biaisées, les gènes contenant AATAAA ont des implications opposées à celles des gènes contenant une boîte TATA. Ils sont moins fréquemment associés à des réponses à un stress ou à un stimulus. (Table 7-3 A).

A - Processus biologique

	14927 gènes	TATAWA	AATAAA	CATAAA	GATAAA	TTTAAA	TTTATA	TATTAA	TATTTA	TATAAG
Réponses à des stimulus biotiques ou abiotiques	**8.8**	11.5 7e-4	5.4 6e-3	7.5 NS	4.8 3e-2	7.3 NS	10.3 NS	8.8 NS	8.1 NS	12.0 NS
Réponses à des stress	**8.8**	11.5 1e-3	4.7 7e-4	6.8 NS	8.0 NS	8.9 NS	10.3 NS	8.8 NS	10.2 NS	11.1 NS
Métabolisme de l'ADN ou de l'ARN	**1.4**	0.7 5e-3	1.0 NS	4.5 1e-2	1.6 NS	1.6 NS	0.8 NS	0.6 NS	1.0 NS	1.9 NS
Métabolisme des protéines	**14.7**	12.2 6e-3	17.3 NS	15.8 NS	17.6 NS	14.1 NS	15.9 NS	15.8 NS	15.7 NS	17.6 NS
Signal de transduction	**4.5**	3.3 1e-2	4.7 NS	4.5 NS	3.2 NS	3.2 NS	3.2 NS	5.3 NS	6.1 NS	7.4 NS

B - Localisation subcellulaire

	14927 gènes	TATAWA	AATAAA	CATAAA	GATAAA	TTTAAA	TTTATA	TATTAA	TATTTA	TATAAG
Paroi cellulaire	**3.1**	5.8 1e-6	2.2 NS	0.8 NS	0.8 NS	2.7 NS	1.8 NS	3.6 NS	4.2 NS	2.9 NS
Cytosol	**2.8**	5.0 4e-5	3.4 NS	2.5 NS	1.6 NS	2.3 NS	2.7 NS	3.0 NS	5.3 NS	5.9 NS
Ribosome	**2.6**	4.2 1e-3	3.9 NS	2.5 NS	1.6 NS	2.7 NS	4.5 NS	3.0 NS	5.8 NS	6.9 NS
Extracellulaire	**2.2**	3.1 3e-2	0.8 NS	0.8 NS	2.4 NS	1.4 NS	2.7 NS	1.8 NS	3.2 NS	1.0 NS
Chloroplaste	**15.4**	10.9 5e-6	16.0 NS	18.2 NS	22.0 NS	13.6 NS	11.7 NS	15.5 NS	19.0 NS	10.8 NS
Mitochondrie	**5.6**	3.2 3e-5	5.3 NS	4.1 NS	11.4 NS	5.4 NS	3.6 NS	7.1 NS	6.8 NS	4.9 NS
Plastides	**6.1**	3.6 6e-5	6.7 NS	4.1 NS	8.1 NS	7.7 NS	1.8 NS	7.1 NS	7.9 NS	4.9 NS

Table 7-3 : Fonction des gènes contenant un PLM dans la région [-39, -26].

Les pourcentages de la colonne «14927gènes» servent de références. Pour chaque annotation fonctionnelle, le pourcentage pour un groupe de gènes contenant un motif est comparé au pourcentage pour tous les autres gènes (tests unilatéraux de Fisher avec correction de Bonferroni). Sont mis en évidence les enrichissements (rouge) et les appauvrissements (vert) en gènes pour une annotation donnée. Seuls les catégories de GO biaisées dans une catégorie de gènes sont renseignées. Les résultats soulignés correspondent à des biais opposés entre les gènes contenant une boîte TATA et les gènes contenant un variant. NS : pas de différence significative (p-value > 5e-2).

Ces résultats montrent que les gènes contenant AATAAA semblent avoir acquis des fonctions spécifiques qui les distinguent des gènes contenant une boîte TATA, et ce malgré la séquence très proche de ce TATA∆1 T1 et de TATAWA. Ces observations supportent l'hypothèse que de petites variations de la boîte TATA sont liées à des modifications de l'expression des gènes (Mingam *et al.*, 2004). Néanmoins, toutes les variations ne semblent pas avoir les mêmes effets, les trois TATA∆1 T1 n'ayant pas tous de biais significatifs.

ii) Localisation subcellulaire des produits des gènes

L'étude de ces groupes de gènes a été poursuivie en considérant la localisation subcellulaire des produits des gènes. Comme pour les processus biologiques, la localisation subcellulaire des produits des gènes contenant une boîte TATA est la plus différente des autres gènes. Les produits de ces gènes sont plus souvent dirigés vers les membranes cellulaires, le cytosol, les ribosomes et le compartiment extracellulaire. Ils sont moins dirigés vers les chloroplastes, mitochondries et plastides (Table 7-3 B). L'observation de ces biais indiquent que les produits des gènes possédant une boîte TATA sont plus particulièrement associés à des interfaces inter et intra cellulaires.

Les gènes contenant les variants n'ont pas de biais significatifs révélés lors de ces analyses (Table 7-3 B). Ce résultat n'indique pas que ces gènes n'ont pas de localisation subcellulaire communes, mais peut laisser supposer que les données analysées ne conduisent pas à un résultat significatif.

b) Structure des gènes

La structure des gènes contenant les 9 motifs de la Table 7-2 a été considérée. Chez *H. sapiens*, il a été montré récemment que les gènes contenant une boîte TATA sont globalement plus courts que d'autres gènes (Moshonov *et al.*, 2008). Les auteurs ont notamment observé des ARNm mais surtout des introns plus courts au sein des gènes contenant une boîte TATA. Au cours de cette thèse, les mêmes biais ont été observés pour les gènes d'*A. thaliana* contenant une boîte TATA. Lors de cette étude, en exploitant les informations disponibles dans FLAGdb^{++} (Samson *et al.*, 2004), différentes structures des unités de transcription ont été exploitées : les longueurs des UTR 5', UTR 3', CDS et introns. Les études précédentes avaient considéré les ARNm dans leur globalité pour conclure que les 2 UTR des gènes contenant une boîte TATA étaient plus courts. La distinction entre la longueur des UTR 5' et des UTR 3' permet de mettre en évidence le rôle majeur de la longueur de l'UTR 5', tandis que celle des UTR 3' n'est pas biaisée. L'UTR 5' est la structure qui joue le rôle majeur pour expliquer la plus petite taille des unités de transcription (Table 7-4). Ce résultat est en accord avec le biais observé en comparant la longueur des UTR 5' des promoteurs d'*A. thaliana* contenant ou non une boîte TATA (Molina & Grotewold, 2005). L'hypothèse proposée par ces auteurs pour expliquer cette différence est que la longueur de l'UTR pourrait influencer l'assemblage du complexe d'initiation de la transcription.

Comme observé précédemment, parmi les gènes possédant un variant, ceux possédant un TATA∆1 T1 AATAAA sont les seuls possédant un biais. Le biais identifié est à nouveau opposé aux caractéristiques des gènes contenant une boîte TATA. Ces gènes contenant AATAAA ont des unités de transcription plus longues que les autres gènes (Table 7-4).

	14927 gènes	TATAWA	AATAAA	CATAAA	GATAAA	TTTAAA	TTTATA	TATTAA	TATTTA	TATAAG
Unité de transcription	**2250**	1936 <1e-30	2369 4e-3	2299 NS	2286 NS	2316 NS	2030 NS	2216 NS	2359 NS	2105 NS
UTR 5'	**158**	111 <1e-30	180 NS	165 NS	158 NS	171 NS	144 NS	152 NS	172 NS	150 NS
CDS	**1086**	966 2e-14	1139 NS	1074 NS	1071 NS	1056 NS	963 NS	1035 NS	1100 NS	909 NS
Introns	**588**	521 7e-5	665 NS	561 NS	700 NS	591 NS	484 NS	472 NS	578 NS	635 NS

Table 7-4 : Structure des gènes contenant un PLM dans la région [-39, -26].

Les médianes des longueurs de la colonne «14927 gènes» servent de références. Pour chaque structure, la distribution des longueurs pour un groupe de gènes contenant un motif est comparée à celle pour tous les autres gènes (tests unilatéraux de Wilcoxon avec correction de Bonferroni). Sont mis en évidence les enrichissements en structures longues (rouge) ou courtes (vert) dans un groupe de gènes. Seuls les structures biaisées sont renseignées. Les résultats soulignés correspondent à des biais opposés entre les gènes contenant une boîte TATA et les gènes contenant un variant. NS : pas de différence significative (p-value > 5e-2).

En conclusion de cette étude structurale, les gènes contenant le variant AATAAA se distinguent des gènes contenant d'autres variants, car ils présentent des biais significatifs, toujours opposés aux biais des gènes contenant la boîte TATA.

c) Expression des gènes

La propension des gènes contenant une boîte TATA à s'exprimer avec une plus forte intensité a été révélée chez *H. sapiens* (Moshonov *et al.*, 2008) mais aussi chez *S. cerevisiae* (Basehoar *et al.*, 2004). Chez *A. thaliana*, ce biais a été observé en étudiant le nombre d'EST associé aux gènes (Molina & Grotewold, 2005). Les données transcriptome présentées dans le contexte du projet de la thèse permettaient une analyse *in silico* plus précise de l'impact de la présence de la boîte TATA et des variants sur l'expression des gènes. Parmi les 14927 gènes, 11161 ont des données d'expression qui ont été mises en évidence via une analyse des puces CATMA (Crowe *et al.*, 2003). Chacun des 11161 gènes peut être caractérisé par (i) son intensité d'expression médiane et (ii) le pourcentage d'expériences dans lesquelles il s'est exprimé. Ces deux catégories d'informations sont annotées respectivement «intensité d'expression» et «pourcentage d'hybridation».

Lors de l'étude globale des données d'expression CATMA, chaque gène est représenté Figure 7-5 en fonction de son intensité d'expression médiane et de son pourcentage d'hybridation. Au total, 4 groupes ont été constitués en fonction de cette Figure. Pour séparer les gènes en fonction de leur intensité d'expression élevée ou non, une régression polynomiale est utilisée. Elle permet de regrouper 4371 gènes ayant une intensité d'expression élevée (HE pour «High Expression») et 6790 gènes ayant une intensité d'expression faible (LE pour «Low Expression») par rapport à un pourcentage d'hybridation donné. La courbe permettant de distinguer les gènes HE des gènes LE est caractérisée par deux points de courbure. Les pourcentages de 15% et de 85% ont été sélectionnés pour différencier les hybridations dites spécifiques (SR pour «Small Range of hybridization») de celles dites constitutives (WR pour «Wide Range of hybridization»). La catégorie SR est constituée de 4510 gènes qui s'expriment dans moins de 15% des expériences de puces considérées. La catégorie WR comprend 1241 gènes qui s'expriment dans plus de 85% des

expériences. L'ensemble des données de ces 4 groupes de gènes est disponible dans la Table ci-dessous.

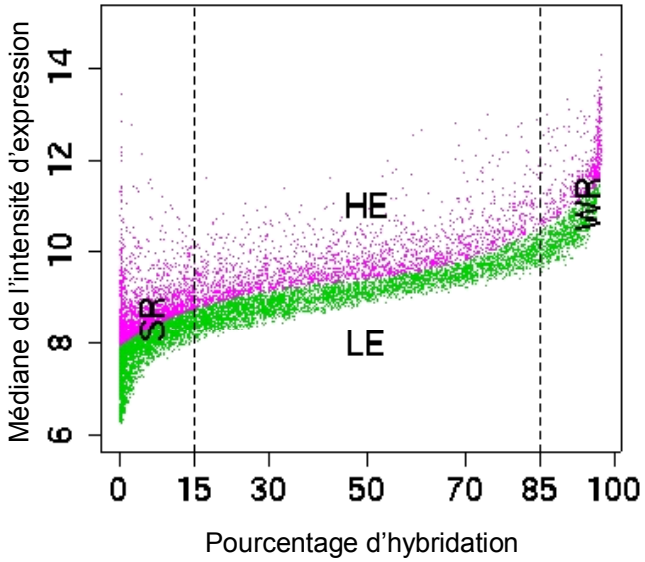


Figure 7-5 : Analyse des données d'expression des gènes d'*A. thaliana*.

Chaque point représente l'expression d'un gène. Une régression polynomiale distingue les intensités d'expression fortes et faibles (respectivement en magenta et vert) notés HE et LE. Cette courbe met en évidence deux points de courbure dans la courbe qui permettent de séparer les gènes en fonction de leur pourcentage d'hybridation. Les hybridations spécifiques et celles constitutives sont respectivement en amont de 15% d'hybridation et en aval de 85% d'hybridation (traits en pointillés). Elles sont notées respectivement SR et WR.

Intitulé du groupe	**Caractéristiques des gènes**	**Critère d'association d'un gène à ce groupe**	**Nombre**
HE	Intensité d'expression élevée	Intensité d'expression au dessus de la droite de régression	4371
LE	Intensité d'expression faible	Intensité d'expression au dessous de la droite de régression	6790
SR	Expression dans des conditions spécifiques	Pourcentage d'hybridation inférieur à 15%	4510
WR	Expression constitutive	Pourcentage d'hybridation supérieur à 85%	1241
SR-HE+	Intensité d'expression élevée dans des conditions spécifiques	Pourcentage d'hybridation inférieur à 15% et intensité d'expression au dessus de la borne supérieure de l'intervalle de confiance	618
SR-LE+	Intensité d'expression faible dans des conditions spécifiques	Pourcentage d'hybridation inférieur à 15% et intensité d'expression au dessous de la borne inférieure de l'intervalle de confiance	822
WR-HE+	Intensité d'expression élevée et constitutive	Pourcentage d'hybridation supérieur à 85% et intensité d'expression au dessus de la borne supérieure de l'intervalle de confiance	244
WR-LE+	Intensité d'expression faible et constitutive	Pourcentage d'hybridation supérieur à 85% et intensité d'expression au dessous de la borne inférieure de l'intervalle de confiance	385

Table 7-5 : Groupes de gènes en fonction de leur expression.

Par la suite, une analyse des données extrêmes a été réalisée afin de considérer les gènes des catégories SR et WR ayant des intensités d'expression particulièrement élevées ou faibles (Table 7-5). Les gènes SR qui ne s'expriment qu'avec une intensité extrême sont

des gènes répondant à des régulations précises et qui ne s'expriment qu'à un très faible niveau (SR-LE+) ou qu'à un très fort niveau (SR-HE+) pour répondre aux besoins spécifiques de l'organisme. Les gènes WR qui ne s'expriment qu'avec une intensité extrême peuvent être des gènes de ménage s'exprimant toujours à un taux élevé (WR-HE+) ou relativement faible (WR-LE+). Pour constituer ces groupes un intervalle de confiance à 60% a été construit (Figure 7-6) afin d'avoir un nombre de gènes dans chaque groupe considéré comme suffisant, c'est-à-dire plus de 100 gènes dans chacune des 4 catégories.

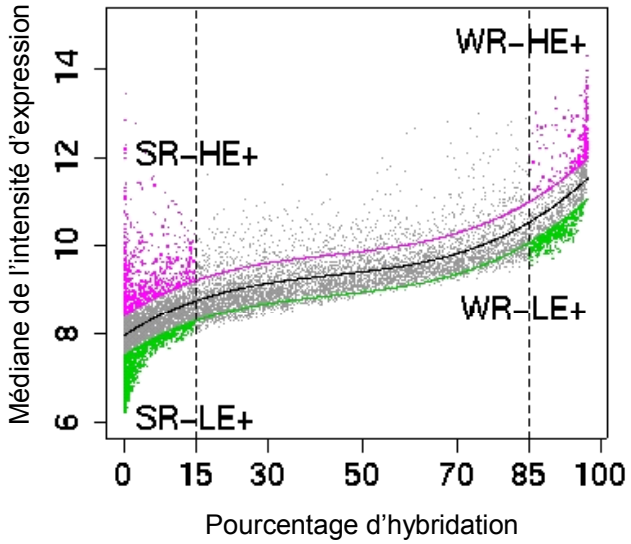


Figure 7-6 : Quatre classes des données d'expression extrêmes.

Un intervalle de confiance à 60% est construit pour extraire les gènes ayant des intensités extrêmes très élevées (HE+) ou très faibles (LE+) au sein des catégories des gènes SR et WR. La Table 7-5 détaille les caractéristiques de chacune des classes.

Huit groupes de gènes ont été constitués en fonction des données d'expression. La présence de variants ou de la boîte TATA dans ces 8 classes de gènes peut alors être considérée.

Dans la première colonne de la Table 7-6, le pourcentage de l'ensemble des 11161 gènes dans chacun des groupes d'expression de la Table 7-5 est indiqué en référence. Les pourcentages correspondant aux gènes contenant un variant ou une boîte TATA sont ensuite comparés aux pourcentages de l'ensemble des autres gènes comme réalisé précédemment. Les gènes qui possèdent une boîte TATA sont plus fréquemment des gènes ayant une intensité d'expression forte (Table 7-6 A). Aucun biais n'a été identifié concernant le pourcentage d'hybridation, c'est à dire le pourcentage d'expériences au cours desquelles les gènes s'expriment. Dans ces 4 premières catégories, aucun variant ne présente de biais.

A

	11161 gènes	TATAWA		AATAAA		CATAAA		GATAAA		TTTAAA		TTTATA		TATTAA		TATTTA		TATAAG	
LE	**60.8**	46	2e-24	65.4	NS	62.7	NS	66.1	NS	64.8	NS	63	NS	61.7	NS	51.8	NS	53	NS
HE	**39.2**	54	0	34.6	NS	37.3	NS	33.9	NS	35.2	NS	37	NS	38.3	NS	48.2	NS	47	NS
SR	**40.4**	41	NS	35.8	NS	36.4	NS	37.5	NS	41.2	NS	48	NS	30.5	NS	37.3	NS	41	NS
WR	**11.1**	10	NS	13.1	NS	15.5	NS	16.1	NS	12.5	NS	13	NS	15.6	NS	16.9	NS	10	NS

B

	11161 gènes	TATAWA		AATAAA		CATAAA		GATAAA		TTTAAA		TTTATA		TATTAA		TATTTA		TATAAG	
SR-LE+	**7.4**	5.0	8e-4	5.7	NS	8.2	NS	7.1	NS	7.9	NS	7.0	NS	4.5	NS	9.6	NS	9.0	NS
SR-HE+	**5.5**	8.5	1e-5	3.9	NS	6.4	NS	3.6	NS	4.2	NS	11	NS	5.2	NS	6.0	NS	7.0	NS
WR-LE+	**3.4**	2.5	4e-2	6.0	6e-3	7.3	NS	4.5	NS	4.6	NS	2.0	NS	3.9	NS	3.0	NS	1.0	NS
WR-HE+	**2.2**	2.4	NS	2.1	NS	2.7	NS	0.9	NS	2.8	NS	2.0	NS	2.6	NS	4.2	NS	3.0	NS

Table 7-6 : Expression et pourcentage d'hybridation des gènes contenant un PLM dans la région [-39, -26].

Les pourcentages de la colonne «11161 gènes» servent de références. Pour chaque catégorie d'expression, le pourcentage correspondant à un groupe de gènes contenant un motif donné est comparé au pourcentage obtenu dans tous les autres gènes (tests unilatéraux exacts de Fisher avec correction de Bonferroni). Sont mis en évidence les enrichissements (rouge) ou appauvrissements (vert) des gènes dans une catégorie d'expression. Les résultats soulignés correspondent à des biais opposés entre les gènes avec une boîte TATA et ceux avec un variant. NS : pas de différence significative (p-value > 5e-2). Les abréviations sont détaillées Table 7-5.

L'étude des expressions extrêmes (Table 7-6 B) révèle une plus forte fréquence des gènes contenant une boîte TATA dans la catégorie des expressions spécifiques à très forte intensité SR-HE+, et une plus faible fréquence dans les deux groupes à intensité d'expression très faible : SR-LE+ et WR-LE+. Le groupe de gènes contenant les variants TATAΔ1 T1 AATAAA a des biais opposés et sont enrichis en gènes WR-LE+. Comme précédemment décrit lors des études de la fonction et de la structure, seuls les gènes contenant ce variant T1 ont des caractéristiques opposées aux gènes contenant une boîte TATA.

Les résultats de ce travail sont en accord avec des travaux réalisés chez *S. cerevisiae* qui ont démontré que la région de la boîte TATA est une région soumise à une évolution fonctionnelle rapide (Basehoar *et al.*, 2004). Des modifications d'une seule base de la séquence TATAWA ont en effet des conséquences fonctionnelles importantes. De plus, les gènes contenant une boîte TATA forment un groupe très biaisé comparé aux autres gènes. Ils ont tendance à avoir une unité transcriptionnelle plus courte, en particulier concernant la longueur de l'UTR 5' et du CDS, comme observé chez *H. sapiens* (Moshonov *et al.*, 2008) ; ils sont plus souvent impliqués dans des réponses spécifiques, une réponse à un stress ou un stimulus par exemple comme observé chez *S. cerevisiae* et chez *H. sapiens* (Yang *et al.*, 2007) et ils s'expriment plus fortement (Basehoar *et al.*, 2004 ; Yang *et al.*, 2007 ; Moshonov *et al.* 2008). Ainsi, les caractéristiques des gènes contenant une boîte TATA semblent être conservées entre les mammifères, les plantes et les protozoaires. Ce résultat montre l'intérêt d'une étude systématique des biais structuraux et d'expression de gènes partageant une même organisation de leurs promoteurs. La présence de la boîte TATA au sein de gènes

induit des biais structuraux et fonctionnels. Ces biais peuvent aider à l'annotation d'un motif ou d'un module de motifs.

7.2.5 Variant AATAAA : les plus divergents

Les gènes contenant un des variants de la boîte TATA ne présentent pas tous les mêmes biais. Deux catégories peuvent être mises en évidence.

Les gènes contenant les variants GATAAA, CATAAA, TTTATA, TTTAAA, TATTAA, TATTTA et TATAAG ne présentent pas de biais fonctionnels ni structuraux par rapport aux autres gènes. Ce résultat ne permet donc pas de confirmer que ces séquences puissent être des variants fonctionnels. Néanmoins, les données qui ont été analysées ne sont pas exhaustives et les gènes contenant ces motifs pourraient avoir d'autres caractéristiques.

Le variant TATAΔ1 T1 AATAAA induit des biais le distinguant des autres variants et de la boîte TATA (Table 7-7). La première base de la séquence TATAWA est donc susceptible d'être substituée en 3 bases différentes, mais les conséquences fonctionnelles et structurales résultantes peuvent être importantes, en particulier lors d'une substitution en A. Ces analyses laissent supposer que ce motif AATAAA pourrait être un élément régulateur impliqué dans un mécanisme de régulation différent de celui des boîtes TATA ou de ses autres variants. Ce motif pourrait s'être différencié fonctionnellement de la boîte TATA canonique et des variants.

		Tous les gènes	TATAWA	AATAAA
Fonction	Réponse à des stimulus biotiques ou abiotiques	**8.8**	11.5	5.4
	Réponse à des stress	**8.8**	11.5	4.7
Structure	Unité de transcription	**2250**	1936	2369
Expression	WR-LE+	**3.4**	2.5	6.0

Table 7-7 : Bilan des biais mis en évidence au sein des gènes contenant AATAAA par rapport aux autres gènes.

Les résultats de la Table 7-3 (Fonction), Table 7-4 (Structure) et Table 7-6 (Expression) sont résumés dans ce tableau où sont considérées exclusivement les données biaisées dans au moins un des groupes de gènes contenant AATAAA. Les cases jaunes correspondent aux biais opposés à ceux des gènes contenant une boîte TATA.

7.2.6 Intérêt de considérer les motifs indépendamment

Chez *H. sapiens*, en constituant des jeux de variants de la boîte TATA basés exclusivement sur des comparaisons de séquences, des jeux de variants ont été analysés (Moshonov *et al.*, 2008). La seule ressemblance de séquences n'est pas corrélée avec la fixation de la TBP (Kiran *et al.*, 2006). C'est pourquoi la prise en compte de l'ensemble des motifs ayant un ou deux mésappariement(s) avec la séquence TATAWA pour identifier les variants fonctionnels de la boîte TATA pourrait être considérée comme non pertinente. Malgré cela, des biais ont été mis en évidence. Plus les gènes possèdent une séquence proche de TATAWA, c'est-à-dire un variant avec un mésappariement, plus ils ont une intensité d'expression élevée et plus leur unité transcriptionnelle est compacte (Moshonov *et al.*, 2008). En regroupant l'ensemble des gènes contenant un variant, ces mêmes biais sont retrouvés chez *A. thaliana* (Bernard *et al.* soumis - page 117). Néanmoins, une telle

approche revient à considérer que l'ensemble des séquences proches de la boîte TATA canonique ont les mêmes fonctions, or nous venons de proposer différentes catégories de variants fonctionnels ou ayant divergé fonctionnellement. Considérer les motifs indépendamment permet de les caractériser les uns par rapport aux autres et non dans leur globalité.

En conclusion, une liste de 15 séquences dérivées de la boîte TATA a été prédite. Les caractéristiques des gènes contenant ces séquences permettent de distinguer un variant TATAΔ1 T1 qui induit des biais fonctionnels et structuraux différents des autres gènes : le PLM AATAAA. Le rôle fonctionnel de la base T1 semblerait être capital. Des variations de la base pourraient jouer un rôle sur la reconnaissance et / ou la fixation de la TBP sur son site.

7.3 De nouveaux motifs à l'emplacement de la boîte TATA : les motifs-TC

7.3.1 Motifs-TC : une nouvelle classe d'éléments régulateurs chez les plantes

Les résultats obtenus lors de l'analyse des variants ont permis de confirmer la présence attendue d'une grande richesse en PLM riches en A et en T dans les promoteurs d'*A. thaliana* et d'*O. sativa* dans la région de la boîte TATA. En complément, différentes observations ont permis de proposer l'hypothèse de l'existence de motifs riches en bases T et C à l'emplacement de la boîte régulatrice.

Des PLM riches en C et T et répétés représentant la présence de microsatellites ont été observés dans de larges fenêtres fonctionnelles (partie 6.2.2 page 88). Leur présence a été observée dans les UTR 5' (Fujimori *et al.*, 2003) et leur fenêtre fonctionnelle s'étend en amont du TSS sur une centaine de bases. Cette présence dans le promoteur central a été observée lors d'autres études et pourrait être due à une extension des microsatellites (Molina & Grotewold, 2005) ou bien à un nouvel élément régulateur du promoteur central : le «Y-patch» (Yamamoto *et al.*, 2007). Le Y-patch est observé sur une large région dans le promoteur central et est abondamment présent dans les promoteurs de plantes et non des mammifères (Yamamoto *et al.*, 2007).

Cette richesse en C et T qui s'étend sur de larges régions est néanmoins susceptible de dissimuler la présence de PLM riches en C et T mais ayant des contraintes topologiques plus strictes que celles associées à des microsatellites. Nous avons proposé cette hypothèse, appuyée sur le constat que la représentation de certains PLM riches en C et T présente un pic secondaire à l'emplacement de la boîte TATA (Figure 6-2 page 83). De plus, les séquences des TATAΔ2 laissent apparaître une plus grande richesse en C et T que les autres variants. C'est pourquoi l'analyse de la région de la boîte TATA a été poursuivie, pour les promoteurs qui ne contiennent pas de richesse en A et T dans la région en explorant l'éventualité d'identifier un nouveau motif. Ce travail est décrit dans un article soumis et présenté dans les pages suivantes.

Hormis des modifications liées à la mises en pages, le manuscrit de cet article est tel qu'il était lors de la soumission.

TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation

Virginie Bernard[1], Véronique Brunaud[1] and Alain Lecharny[1, 2§]

[1]Unité de Recherche en Génomique Végétale (URGV), UMR INRA 1165 -CNRS 8114 - UEVE, 2 Rue Gaston Crémieux, 91057 Evry Cedex, France

[2]Université Paris-Sud, Institut de Biotechnologie des Plantes (IBP), UMR CNRS 8618 -UPS, Bâtiment 630, 91405 Orsay Cedex, France

[§]Corresponding author

Email addresses:

VBe: bernard@evry.inra.fr

VBr: brunaud@evry.inra.fr

AL: lecharny@evry.inra.fr

Abstract

Background

The TATA-box and TATA-variants are regulatory elements involved in the formation of a transcription initiation complex. Both have been conserved throughout evolution in a restricted region close to the Transcription Start Site (TSS). However, less than half of the genes in model organisms studied so far have been found to contain either one of these elements. Indeed different core-promoter elements are involved in the recruitment of the TATA-box-binding protein. Here we assessed the possibility of identifying novel functional motifs in plant genes, sharing the TATA-box topological constraints.

Results

We developed an *ab-initio* approach considering the preferential location of motifs relative to the TSS. We identified motifs observed at the TATA-box expected location and conserved in both *Arabidopsis thaliana* and *Oryza sativa* promoters. We identified TC-elements within non-TA-rich promoters 30 bases upstream of the TSS. As with the TATA-box and TATA-variant sequences, it was possible to construct a unique distance graph with the TC-element sequences. The structural and functional features of TC-element-containing genes were distinct from those of TATA-box- or TATA-variant-containing genes. *Arabidopsis thaliana* transcriptome analysis revealed that TATA-box-containing genes were generally those showing relatively high levels of expression and that TC-element-containing genes were generally those expressed in specific conditions.

Conclusions

Our observations suggest that the TC-elements might constitute a class of novel regulatory elements participating towards the complex modulation of gene expression in plants.

Background

Over the past genomics has markedly changed our view on core-promoter organization [1]. For instance, the TATA-box can no longer be regarded as a general feature of polymerase II core-promoters [2]. Indeed, only a small fraction of eukaryotic genes actually harbour a TATA-box: less than 20% of genes in both human [3] and yeast [4]. While TATA-box-containing promoters support the direct binding of TATA-box-Binding Proteins (TBPs) and thereby also the formation of the pre-initiation complex, TATA-less promoters are recognized by multiple TBP-related proteins and other TBP-associated Transcription Factors (TFs) involved in the recruitment of TBP [5]. Indeed some TATA-variants and other alternative elements allow the initiation of transcription and participate towards defining distinct patterns of expression [4, 6-9].

Several core-promoter elements or general Transcription Factor Binding Sites (TFBSs) have been previously identified in eukaryotes. They are characterized by a strong positional preference relative to the Transcriptional Start Site (TSS) as for instance the TATA-box in the [-30, -25] area [8], the Initiator element, (Inr), around the TSS [10], the downstream promoter element in the [+28, +33] area [11], or the IIB recognition element immediately upstream of certain TATA-boxes [12]. The position of binding sites of proteins belonging to the transcription complex is important for the functioning of promoters since it determines both the TSS location [13] and the transcription direction [5]. Thus, a strong positional conservation of a novel regulatory element would strongly indicate a functional role. This concept has led to a generation of tools that in contrast to previous TFBS predictors [14, 15] are based on the positional densities of oligonucleotides rather than on their frequency of occurrence. These tools have been used to characterize core-promoter elements in several model genomes including plants [16-22].

All together, the core-promoter elements listed above seem unable to account for the transcription of all the RNA-polymerase-II transcribed genes. Less conserved core-promoter elements present in small gene sets have been described in previous studies at the gene level. For instance, in the human cytosolic phospholipase A2-alpha gene, an AAGGAG motif in the [-35, -30] area binds TBP and is critical for basal transcriptional activity [23]. In other studies, a TBP has been shown to bind to a TAAGAGA element in the [-23, -17] region of the hepatitis B virus S gene [24]. These experimental observations suggest that core-promoter elements specific to small sets of genes remain to be disclosed. A study of large-scale structural properties of DNA in promoters indicated that the instability of DNA around -30 relative to the TSS necessary for transcription may be due to as yet unidentified motifs other than the TATA-box [25].

It is therefore clear from and despite this growing amount of data that the code embedded within core-promoter sequences has not yet been fully deciphered. In this work, we used an *in silico* hypothesis-driven approach to predict novel elements potentially recognized by the transcriptional complex. We explored the bioinformatics-based evidence that sequences other than the TATA-box and TATA-variants but located in the same region relative to the TSS may be functional core-promoter elements. We therefore searched for short sequences exhibiting similar positional constraints to those of the TATA-box and identified pyrimidine-rich elements distinct from the pyrimidine tract [21, 26] as candidate elements for about 18% of the plant genes. To determine their potential functional role, we investigated any association between such identified TC-elements and specific features of the genes containing them, as has been previously shown for the TATA-box.

Results

Less than 39% of *A. thaliana* promoters contain a TATA-box or a TATA-variant

Our approach was based on three steps. First, we searched for all 6 base long motifs with a statistically significant preferential position within the 300 nucleotides upstream of the TSS and called these motifs the Preferentially Located Motifs (PLMs). This method was first described by FitzGerald *et al.* [27] for the analysis of human promoters and was then applied by others to plant promoters [19, 21]. Our results correlated well with previous studies in terms of the spatial representation of the PLMs within the plant promoters. Second, for each PLM, we precisely defined: (i) the preferential position relative to the TSS, *i.e.* the top of the peak, (ii) the functional window, *i.e.* the peak width, derived from the peak boundaries, (iii) the Score of Maximal Square relative to the base line (SMS, see Materials and Methods section) ranking the PLMs in terms of their topological constraint and (iv) the list of genes containing the studied PLM within the functional window. Third, to increase the chance of identifying functional PLMs, we searched for their conservation in 14927 *A. thaliana* genes and 18012 *O. sativa* genes with experimentally supported TSSs. A well-annotated genomic sequence is available for both of these species that diverged approximately 150 million years ago [28, 29]. As previously described within the 50 nucleotides upstream of the TSS of both *A. thaliana* [19, 21] and *O. sativa* [26] we found PLMs made up exclusively or almost exclusively of: (i) T and A nucleotides and exhibiting strong topological constraints, *i.e.* a sharp peak, and (ii) T and C nucleotides and exhibiting low topological constraints, *i.e.* a wide peak. Among the T and A rich motifs, we found the canonical TATA-box defined by the TATAWA consensus (W for A or T) and TATA-box variants [8]. We wondered whether by analysing the whole promoter set, other motifs with the same strong topological constraints could be missed. It is reasonable to predict that the presence of a PLM characterized by a wide functional window overlapping the TATA-box expected area might hide the strong

topological constraints of a PLM specific to a small promoter set. To address this issue, we built up promoter sets by successively subtracting from the whole set of promoters, those characterized by different classes of PLMs, as described below. We firstly considered the conservation of PLMs at the genome level and then their conservation at the orthologues level.

TATAWA is a particularly well-conserved PLM since it is found in the same promoter region in both plants and animals [30]. Confirming previous results, we found this PLM in both *A. thaliana* and *O. sativa* genomes in a preferential position 32 bases upstream of the TSS and strictly located within the [-39, -26] region (Figure 1). A total of 2606 (17.5%) and 2601 (14.4%) promoters in *A. thaliana* and *O. sativa* respectively contained a TATAWA within the [-39, -26] functional window. In contrast to the motif counting method highly prone to false positives, our method may slightly underestimate the number of functional motifs. Indeed, as our aim was to characterize and compare different sets of genes each containing defined PLMs (TATA-box or other) within their promoters, we chose only "clean" sets ,*i.e.* those containing only one regulatory element to the detriment of completeness.

Several T and A rich motifs partially matching the TATAWA sequence have been shown at the TATA-box expected position [31, 32]. We observed PLMs with the same positional constraints as TATAWA, *i.e.* showing a sharp peak in the [-39, -26] region. These PLMs may be (i) functional motifs hereafter named as TATAΔ-PLMs where Δ stands for variant; or (ii) motifs being PLMs due to their overlap with the TATAWA sequence. This overlap represents an inherent drawback of the method since motifs shorter than 6 nucleotides (TAT for instance) as well as 6 base long motifs (NNTATA for instance) may appear as PLMs. TATA-variant and TATA-box are distinct regulatory elements that need differentiating [33]. For this reason, we searched for TATAΔ-PLMs in sets of promoters not containing the canonical TATA-box in the [-39, -26] region. In this way, we were able to ascertain that the PLMs found within the promoter sets were *bona fide* elements under topological constraint. In *A.*

thaliana and *O. sativa* respectively 12321 (82.5%) and 15411 (85.6%) promoters were found without a TATAWA in the [-39, -26] region. Relative to the TATAWA sequence, 32 possible motifs diverged at one position. Out of these 32 motifs, only 11 were found to represent PLMs (TATAΔ1-PLMs) conserved in both species (Figure 2, column 2). Consistent with *in vivo* experimental results [34], T to A or A to T substitutions were the main differences observed among the TATAΔ1-PLMs. Motifs with a base C or G substitution at the central positions were not considered as PLMs and were counter-selected. Motifs with a C or G substitution at the T1 or A6 positions were, however, considered as TATAΔ1-PLMs. There remained 9512 (63.7%) and 13494 (74.9%) of *A. thaliana* and *O. sativa* genes respectively containing neither a TATAWA nor a TATAΔ1-PLM. We identified only 4 conserved TATAΔ2-PLMs amongst the 211 possible motifs (Figure 2, column 3) and no conserved TATAΔ3-PLM. Finally, at the TATA-box expected position, 4151 *A. thaliana* (27.8%) and 3045 *O. sativa* promoters (16.9%) contained a TATAΔ-PLM. Hereafter, TATAΔ- and TATAWA-PLMs are collectively referred to as TA-PLMs.

The number of genes containing the TA-PLMs decreased from the TATA-box to the TATAΔ2-PLMs (Figure 2). Interestingly, even though most of the TATAΔ-PLMs were found to be T and A rich as expected, 3 out of the 4 TATAΔ2-PLMs (CTTAAA, TACTTA and TCTTAA) contained a base C. This suggested that some T and C rich PLMs, hereafter referred to as TC-PLMs, could have been missed in the analysis of the whole promoter set. Indeed, it has been shown that, in plants, the region overlapping the TSS is characteristically rich in T and C nucleotides described as CpT microsatellites or TC-microsatellites [35, 36]. We observed these microsatellites during the analyses of the whole promoter set within a number of different 6-base-long PLMs. They are characterized by a wide functional window centred downstream of the TSS often extending several hundred bases. We sometimes observed a small secondary peak located around 30 bases upstream of the TSS which disturbed the broad distribution of some of the TC-microsatellites (Figure 3.A). These TC-

microsatellites may hide $TC_{[-39, -26]}$-PLMs, *i.e.* TC-PLMs with high topological constraints similar to those of TA-PLMs present within specific sub-sets of genes.

Conserved $TC_{[-39, -26]}$-elements are observed in almost 18% of *A. thaliana* promoters

We hypothesised that $TC_{[-39, -26]}$-PLMs with narrow functional windows could be functional alternatives to TA-PLMs. This being the case we expected to be able to detect these PLMs more easily in promoter sets within which they predominate or within promoters not containing TA-PLMs. We therefore selected a TA-less promoter set by excluding the TA-PLM-containing promoters from the whole promoter dataset. Out of the 16 possible dinucleotides, only 3 were found to be conserved PLMs and all shared the TATA-box topological constraints: TA, AT and AA. A total of 1745 *A. thaliana* promoters (11.7%) and 5089 *O. sativa* promoters (28.2%) contained no TATA-box or conserved TATAΔ-PLM or the 3 conserved dinucleotide-PLMs.

We examined the possibility of these TA-less promoter sets representing sets of poorly annotated promoters due to an incorrect prediction of the TSS position. In both TA-less promoter sets, three arguments supported the validity of the TSS position prediction. First, the observed $TC_{[-39, -26]}$-PLMs were conserved between *A. thaliana* and *O. sativa* and in both plants they exhibited the same strict topological constraints (see below). Second, the distributions of several motifs known as TFBSs supported by experimental analyses [37, 38], such as the Inr or the GGCCC element, shared the same topological constraints in both the whole and the TA-less promoter sets (data not shown). Third, as expected, the GC-compositional strand bias or GC-skew expected in plant promoters was observed at the TSS location in both promoter sets (data not shown).

We therefore searched the TA-less promoter sets for the presence of 6-base-long conserved PLMs. Out of the 4096 possible motifs, we found 29 conserved PLMs exhibiting a strict functional window in the [-39, -26] area relative to the TSS. These PLMs were present in

2645 *A. thaliana* (17.7%) and in 2331 *O. sativa* (12.9%) promoters. In agreement with our hypothesis, the 29 PLMs were $TC_{[-39, -26]}$-PLMs, *i.e.* were comprised of T and C bases only. Figure 3 illustrates how the $TC_{[-39, -26]}$-PLMs can be clearly distinguished in the TA-less promoter set while they were missed in the whole promoter set. Indeed, using the whole promoter set, it is only possible to find the TTCTTC-PLM, which is characterized by a wide functional window and by a preferential position 29 bases downstream of the TSS (Figure 3.A). On the contrary, using the TA-less promoter set, we observed both the wide functional window PLM and a distinct TTCTTC-PLM characterized by a sharp peak centred 33 bases upstream of the TSS, *i.e.* sharing the TA-motif topological constraints (Figure 3.B). We propose that TA-PLMs may be recognized and thus be functional in a T and C rich environment whereas the same might not be true for $TC_{[-39, -26]}$-PLM. As a consequence, large pyrimidine-rich regions could have been preferentially maintained during evolution in promoters with a TA-PLM whereas they may have been counter-selected, at least upstream of the TSS, in promoters with a functional $TC_{[-39, -26]}$-PLM. It is important to note that only $TC_{[-39, -26]}$-PLMs exhibited the TATA-box topological constraints in the TA-less promoter set. Altogether, our results (i) confirmed the importance of TC-microsatellites ($TC_{\{n\}}$ or $TTC_{\{n\}}$ for instance - Figure 3.A) in plant promoters and (ii) predicted the existence of a novel class of functional elements, the $TC_{[-39, -26]}$-PLMs characterized by a sharp peak in the [-39, -26] region and observed in almost 18% of *A. thaliana* promoters.

We investigated the putative presence of $TC_{[-39, -26]}$-PLMs in other eukaryotic genomes. We analyzed 15802 *Homo sapiens* promoters and 15833 *Mus musculus* promoters with an experimentally supported TSS [39]. We observed neither $TC_{[-39, -26]}$-PLMs nor any other PLMs at the TATA-box expected region. Thus, both the TC-microsatellites and the $TC_{[-39, -26]}$-PLMs observed in plants are absent in both human and mouse. These observations suggest an evolutionary link between the presence of the $TC_{[-39, -26]}$-PLMs and the TC-microsatellites.

$TC_{[-39, -26]}$-elements derived from three seed motifs

As for all identified TATAΔ-PLMs, any 6-base-long $TC_{[-39, -26]}$-PLM may be a functional PLM, be a part of a larger functional PLM or contain a smaller functional PLM. Any of the $TC_{[-39, -26]}$-PLMs can potentially overlap by 5 consecutive bases with at least one other $TC_{[-39, -26]}$-PLM. Among the promoters studied, we found only 3 overlapping $TC_{[-39, -26]}$-PLMs: CTTCTT, TTCTTC and TCTTCT. The trinucleotide repetition made up of 2 T bases and one C base is characteristic of self-overlapping motifs involved in DNA recognition by transcription factors [40]. Furthermore, we analyzed the effect of extending each of the 29 $TC_{[-39, -26]}$-PLMs on SMS, *i.e.* the distribution score. Extended PLMs with higher SMS scores than those of the initial 6-base-long PLM were considered to be functional candidates. CTTCTT and TTCTTC could be extended up to the 9-base-long TCTTCTTCT PLM that exhibited the highest SMS (Additional file 1). Together, these results provide evidence in favour of the TCTTCTTCT motif being a functional $TC_{[-39, -26]}$-PLM. Extension of most of the other $TC_{[-39, -26]}$-PLMs did not generate PLMs with higher SMS than the SMS of the initial 6-base-long PLM. Note that the TC-microsatellites observed in the gene 5' UnTranslated Regions (UTR) could be extended up to a 13-base-long PLM (data not shown). The shortest core-PLMs were CTC, TCT and CTT, found within 27 $TC_{[-39, -26]}$-PLMs, suggesting that these trinucleotides could be the core of the functional $TC_{[-39, -26]}$-PLMs. Both the trinucleotide repetitions and the different T and C environments might therefore have a role in $TC_{[-39, -26]}$-PLM function.

Three observations are in favour of a putative role of most $TC_{[-39, -26]}$-PLMs independently of the presence of other PLMs in the same region. First, $TC_{[-39, -26]}$-PLMs were often observed without any of the other PLMs sharing the same preferential position in [-39, -26]. Second, $TC_{[-39, -26]}$-PLMs are not extensions of either a TATA-box or its variants. Thus, among the $_{[-39, -26]}$-PLM-containing promoters, 80% contained PLM-motifs belonging to only one PLM class,

either a TATA-box , a TATAΔ- or a $TC_{[-39, -26]}$-PLM. Third, it was possible to construct oriented graphs displaying the motif divergence for both TA- and TC-PLMs (Figure 2 and 5). Concerning the TA-PLMs, the graph root or seed is the TATAWA-PLM. For a given TA-PLM, the number of promoters containing it depends on both the distance from the seed, *i.e.* the number of substitutions between the given PLM and the seed-PLM and the existence or not of a more divergent sequence. We applied the same approach to the conserved $TC_{[-39, -26]}$-PLMs and found that they could be organized into a unique, closed and oriented graph (see Material and Methods). Three seeds were suggested: TCTTCT, TTTCTT and TTCTTC (Figure 5). Similar to that observed for TA-PLMs, these three $TC_{[-39, -26]}$-PLMs were those most frequently observed in promoters with the number of promoters containing the other $TC_{[-39, -26]}$-PLMs depending on both the distance from these seeds and the existence or not of a more divergent sequence. In conclusion, all $_{[-39, -26]}$-PLMs can be detected independently in different promoters and may be organized into different groups with apparent evolutionary links between the given PLM and the related seed.

$TC_{[-39, -26]}$-element-containing genes are preferentially involved in protein metabolism

While most promoters contain only one class of $_{[-39, -26]}$-PLM, some contain more than one (Figure 4). To investigate the functional significance of the different $_{[-39, -26]}$-PLM classes we decided to construct four promoter sets each characterized by the exclusive presence of one of the classes of PLM in the [-39, -26] region. We distinguished (i) the 1496 promoters containing only a TATA-box, (ii) the 1919 containing only a $TC_{[-39, -26]}$-PLM, (iii) the 2773 containing only a TATAΔ-PLM and (iv) as a negative reference the 7194 promoters without any $_{[-39, -26]}$-PLM, called hereafter the $_{[-39, -26]}$-PLM-less set.

Gene Ontology (GO) annotations [41] have been used previously in the prediction of the functional role of TFBS [42]. Here we analysed the GO annotations of the four classes of genes defined above, according to the four promoter sets. As far as biological processes are

concerned, the most conspicuous observation was a preferred association between genes containing a TATA-box and responses to different stimuli (Table 1, Biological Process). This result was expected since in human and in yeast the TATA-box-containing genes are more frequently involved in specific biological processes [4, 9]. Interestingly, we also found that $TC_{[-39,\ -26]}$-PLM-containing genes are more frequently involved in protein metabolism than any other gene class (Table 1, Biological Process), *i.e.* a basic biological process. Neither the TATAΔ-containing genes nor the $_{[-39,\ -26]}$-PLM-less genes showed any significant bias for any biological process categories. We also analysed the partitioning of genes between the different GO cellular components. Again, the TATA-box-containing genes presented the most biased partitioning. The products of these genes are more often constituents of the cell wall, the cytosol and the ribosomes and less frequently constituents of the chloroplasts, the plastids, the Golgi apparatus and the mitochondria compared to the products of all the other genes (Table 1, Cellular Components). Products of the $TC_{[-39,\ -26]}$-PLM-containing genes exhibited little biased partitioning between the GO cellular component categories. Indeed, only the cell wall component was significantly less represented. It seems biologically sound to have found a positive correlation between the response to different stimuli and the cell wall GO categories in the TATA-box-containing genes and a negative correlation between the protein metabolism and the cell wall categories for $TC_{[-39,\ -26]}$-PLM-containing genes. These observations suggest that $TC_{[-39,\ -26]}$-PLMs might have a functional role in transcriptional control that could frequently differ from and perhaps even oppose the TATA-box functional role. It is of particular interest that our observations support the hypothesis that small variations in the TATA-box sequence are linked to large changes in gene expression [6, 43]. Unfortunately, due to the relatively small number of promoters containing a unique $TC_{[-39,\ -26]}$-PLM, it was deemed impossible to perform the same comparisons between the $TC_{[-39,\ -26]}$-PLM seeds alone and their predicted variants as that performed for the TA-motifs. It should be noted that by analysing all the genes containing a $TC_{[-39,\ -26]}$-PLM alone we decreased the power of our

statistical tests. We were however able to demonstrate a significant difference between the TATA-box and the $TC_{[-39, -26]}$-PLM sets concerning protein metabolism and cell wall categories (p. values of $3e^{-5}$ and $1e^{-9}$ respectively).

Only the canonical TATA-box is spatially linked to the Initiator element

Recent progress in mammalian genomics has defined a more functional image of core-promoters. Two functional categories of core-promoters have been described [44]. The single peak promoters have a tightly defined TSS position within a few base pairs while in broad peak promoters several TSSs may be found within small clusters spaning tens of base pairs. A TATA-box is more likely to be found within single peak promoters than in broad peak promoters [44]. In genome-wide studies, only the most upstream TSSs are often considered and are defined by the available transcript sequences. There are three main categories of core-promoters described in human [9], *i.e.* TATA-box-containing promoters with an Inr, TATA-box-containing promoters without Inr and TATA-less promoters. We observed the same relative representation of these three categories in *Arabidopsis*. Consistent with mammalian studies [44], recent evidence suggests that *A. thaliana* genes containing a TATA-box tend to have a sharper dominant peak of TSS compared to that of other genes [45].

The distance between TATA-box and Inr is important for accurate transcription initiation [46]. We considered the distance between the Inr [21], called here YR-TSS, and the conserved PLMs characterized using our approach. As expected, the YR-TSS is a conserved PLM whose functional window is one base upstream of the TSS in both plants (data not shown). We first analysed the 1496 *A. thaliana* promoters containing a TATA-box and we represented the distances between each TATAWA in the [-39, -26] region and each dinucleotide CA, TA, TG and CG up to 70 bases downstream of the TATAWA (Figure 6.A). The dinucleotide CA, but not TA, TG or CG showed a strong preferred distance of 30 to 33 bases from TATAWA, a distance consistent with the preferential position of both PLMs and with mammalian results

[47]. Furthermore, none of the YR motifs showed a preferential distance either from TATAΔ- or from $TC_{[-39, -26]}$-PLMs (Figure 6.B and 6.C). Finally, our results provide evidence of a link between the canonical TATA-box and the CA-TSS: 30 to 33 bases preferentially separate the CA from the first T of TATAWA.

Genes containing the $TC_{[-39, -26]}$-element are generally long

In human, the presence of a TATA-box in promoters is often associated with a compact gene structure [6]. In *A. thaliana*, we observed the same bias towards a more compact structure of TATA-box-containing genes (Table 2). In contrast, the $TC_{[-39, -26]}$-PLM-containing genes and the $_{[-39, -26]}$-PLM-less genes were overall relatively longer while the gene structure of TATAΔ-PLM-containing genes showed no bias. When present, the bias in gene size was mainly explained by the 5'UTR and by the length of the coding sequence. However, to a lesser extent, the enrichment in compact genes amongst those containing a TATA-box may be explained by the shorter cumulative length of introns, and by the higher percentage of intronless genes (Table 2).

Based on the counting of Expressed Sequence Tags (ESTs) [4] or on microarray data [6], gene size and gene expression levels have been shown previously to be inversely correlated in both human and yeast. This being the case, our observations could be due to the fact that, in general, TATA-box genes have higher expression levels than other *A. thaliana* genes [19, 21]. The differences in gene sizes we observed are consistent with a lower mean expression of TATAΔ-PLM- than of TATA-box-containing genes. Both $TC_{[-39, -26]}$-PLM-containing and $_{[-39, -26]}$-PLM-less genes could be expressed at lower levels. TATA-box-containing genes have been shown to generally display relatively high specificity of expression compared to housekeeping genes that frequently show high expression levels (Table 1). Therefore, the question remains as to understand whether the presence of regulatory elements in the [-39, -26] region of promoters is linked directly to gene function *per se* or to the transcription level

that in turn is associated with gene function. To resolve this we analysed a large set of transcriptome data, distinguishing the two different components of transcription: specificity and level.

Specific role played by each class of regulatory element found in the [-39, -26] region on different components of transcription

In eukaryotic genomes, various TFBSs have been linked to specific gene expression patterns [4, 6, 9, 48]. We searched for such a link with PLMs in the [-39, -26] region using *A. thaliana* transcriptome data obtained with the Complete *Arabidopsis* Transcriptome MicroArray (CATMA) [49] and available through the Complete *Arabidopsis* Transcriptome database: CATdb [50]. The dataset included 1044 hybridizations based on 522 different samples covering numerous developmental stages, biotic and abiotic stresses and mutants. We used normalized data on which positive hybridizations have previously been determined (see Methods in Aubourg *et al.*, [51]). For any one gene, the relative number of positive hybridizations was considered as a measure of the range of expression, *i.e.* of the specificity. To measure the global level of expression we computed the median of the signal intensities. The relationship between the median signal intensity and the percentage of hybridizations was clearly not linear and suggested the existence of different classes of genes with respect to transcriptional regulation. We thus clustered the transcriptome data into four classes (Figure 7 and Table 3, first rows). First, genes were separated in two classes depending on their expression level above or below the distribution model line, respectively HE for those with High levels of Expression and LE for those with Low levels of Expression. Second, two classes were defined relating to the specificity of hybridizations and delimited by the two inflections in the cloud of points. The first class clustered genes positively hybridized in less than 15% of microarray hybridizations (SR for Small Range) and the second class clustered those with more than 85% of hybridizations (WR for Wide Range). Interestingly, each of

these four transcriptional clusters was predominantly made up of genes containing one of a specific class of $_{[-39, -26]}$-PLMs. Thus, (i) genes within the transcriptional SR group, *i.e.* those that hybridized in specific conditions predominantly contained $TC_{[-39, -26]}$-PLM; (ii) genes within the transcriptional WR group, *i.e.* housekeeping genes predominantly contained a TATAΔ-PLM; (iii) the transcriptional HE group was enriched with the TATA-box-containing genes and (iv) the transcriptional LE group was enriched with the $_{[-39, -26]}$-PLM-less genes. These results clearly indicate a link between the different motifs within the core-promoter and transcriptional features of the gene and are in favour of a functionally independent role of the TATA-box, the TATAΔ- and the $TC_{[-39, -26]}$-PLMs. Our observations provide further evidence for the involvement of the TATA-box in the regulation of transcription level. In addition, our results suggest that both TATAΔ- and $TC_{[-39, -26]}$-PLMs might be involved in defining the specificity of transcription. The role of the TATAΔ-PLMs might be to promote a relatively wide range of expression while on the other hand $TC_{[-39, -26]}$-PLMs may impose narrower limits. The adverse effect of TATAWA and $_{[-39, -26]}$-PLM-less within genes on the extreme level of transcription (see Material and Methods) was particularly evident in the SR group (Table 3, last two rows). The most Highly Expressed genes (HE+ genes) were predominantly those containing TATAWA whereas the most Lowly Expressed genes (LE+ genes) were mainly $_{[-39, -26]}$-PLM-less genes. We propose that in the presence of a TATA-box, the recognition of the TSS depends directly on a small number of TBPs and / or of TBP-related protein(s). The relaxed recognition by TBPs in TATAΔ-containing promoters might be responsible for a bias towards the large range of expression. In the absence of TATA-box or TATAΔ, the recruitment of the TBP might be mediated by different TBP-associated TFs depending on the gene. Interaction of the different TBP-associated TFs with other specific factors might therefore explain the relatively higher specificity of transcription observed for the $TC_{[-39, -26]}$-PLM gene set.

The class of regulatory element in the [-39, -26] region is not conserved in higher plant orthologous gene pairs

With the increasing availability of transcriptome data, the search for conserved TFBSs within promoters of genes exhibiting similar transcriptional patterns has gathered much attention. Surprisingly, the conservation of core-promoter elements, widely considered as being those most conserved, has not received such large interest. One recent study addressing this question did cast doubt on this widely accepted notion [52]. Concordantly, in several studies on orthologous relationships between large gene families, we observed no significant conservation of the presence of a TATA-box between plant orthologues (data not shown). Even single copy genes present as orthologues in both *A. thaliana* and *O. sativa* are no more conserved than random pairs of genes [53]. The co-occurrence of different elements in the [-39, -26] promoter region could, at least in part, explain an apparent non-conservation at the gene level despite the observed conservation at the genomic one. This prompted us to re-examine the question comparing core-promoter elements in the [-39, -26] region in *A. thaliana* and *O. sativa*. We analysed the conservation of the three classes of $_{[-39, -26]}$-PLMs in 5805 pairs of orthologous genes in *A. thaliana* and *O. sativa* characterized by an experimentally supported TSS. While the conservation of TATAWA appeared relatively low (17%), it was significantly higher than that expected by chance (6%) (Table 4). This was not the case for TATAΔ- or $TC_{[-39, -26]}$-PLMs, which exhibited no conservation between orthologous genes (Table 4). Our results showing that the TATA-box is more involved at the transcriptional level and that both TATAΔ- and $TC_{[-39, -26]}$-PLMs are more involved in specificity are in line with finding. All together our observations are in accordance with the fact that gene transcription levels are positively correlated between orthologous genes in *A. thaliana* and *O. sativa* [53] and that the correlation between transcriptional range and intensity is rather weak [51]. Nevertheless, TATAΔ- and $TC_{[-39, -26]}$-PLMs are conserved, both at the sequence level *per se* and at the occurrence level within the whole genome.

Discussion and conclusions

In *A. thaliana* and *O. sativa*, we identified a novel class of putative TFBS involving TC-elements that are preferentially located within the same core-promoter region as TATA-boxes.

We showed that these TC-elements are structurally distinct from the previously described TC-microsatellites observed in the 5'UTR of plant genes. Nevertheless, the presence of both TC-elements and TC-microsatellites in higher plants but not in vertebrates suggests an evolutionary link between these two promoter elements.

Previous reports have described the presence of a pyrimidine rich element, named the Y-patch, in plant core-promoters [21, 26], and its tendency to be associated with both the TATA-box and the Inr motif [45]. Y-patch motifs and TC-microsatellites are two names used to globally describe T and C rich sequences widely surrounding the TSS. In promoters, these two elements are characterized by two non identical but overlapping groups of motifs. Our results show that the TC[-39,-26]-PLM exhibits specific characteristics distinguishing it from the other Y-patch motifs. Y-patch motifs show frequent occurrence in plant promoters, are present in a wide area around the TSS (see Fig.1) and may be extended from a 6 to a 10-base-long element without decreasing the score associated to their local overrepresentation (SMS). Indeed the TC[-39,-26]-PLMs were only observed in a sub-set of 18% of *A. thaliana* promoters at the TATA-box expected position, were associated with a sharp functional window, were 6-base-long and could not be extended without decreasing their SMS.

Specific promoter features were frequently observed when genes were classified into groups, *i.e.* the TC-element-containing gene group, the TATA-box-containing gene group or the TATAΔ-PLMs-containing gene group. First, our data indicate that the TATA-box preferentially associates with a CA-TSS. Neither the TATAΔ- nor the TC-elements exhibited this apparent functional association. In addition, no association between the TATA-box and

any of the three other YR-TSS was observed. Second, TC-element-containing genes were predominately large and involved in protein metabolism while TATA-box containing genes were predominantly compact and involved in response to stress and stimulus. Indeed, gene function, specificity and level of expression are linked features. Using an original approach we have been able to distinguish the possible involvement of different $_{[-39, -26]}$-elements in the control of either the level or the specificity of expression. A global analysis of CATMA-transcriptome data indicated preferential links between the presence of: i) a TATA-box and high gene expression; ii) TC-elements and high specificity of gene expression; and iii) TATAΔ-PLMs and broad expression patterns, as in housekeeping genes. All together our observations suggest that TC-elements might be considered as a novel class of plant promoter elements linked directly or indirectly to the regulation of gene expression.

The presence of all three elements, *i.e.* TATA-boxes, TC- and TATAΔ-elements, have been observed in two plants that diverged about 150 millions of year ago [29]. We observed a global conservation, *i.e.* conservation in the relative number of genes containing one of the three motifs in the two genomes. Nevertheless, only the TATA-box showed a low level of conservation between orthologous gene pairs while conservation of either TC- or TATAΔ-elements between orthologous gene pairs was no higher than that found between any gene pair. On the one hand, the low level of conservation at the orthologous gene level is in line with the rapid evolution of core regulatory motifs after gene duplication in *A. thaliana* [43] and with the notion that *A. thaliana* and *O. sativa* may have independently evolved novel TSSs [54]. On the other hand, the conservation of the relative number of genes with either a TATA-box, a TC- or a TATAΔ-element suggests a global conservation of the interspecies variability, noise level and evolvability of gene expression; three processes involving TATA sequences [55-57]. The three classes of core-promoter elements observed in the [-39, -26] region are each present in about 20% of the promoters. In this study, we used “clean” classes of promoters, *i.e.* those containing only one of the three classes of elements, however in

several other promoters, more than one element from a given class may be present. Therefore many functional combinations are expected. Experimental data have shown that the TATA-box is involved in the direct binding of TBP [5] and at least some TC- and TATAΔ-elements might be involved in the recognition by other TFs recruiting TBP to form the transcription initiation complex at the right position. Consistent with this notion, both TC- and TATAΔ-elements are linked to the specificity of expression that could be mediated through different TFs under the control of various signals. An alternative would be that TATA-boxes, TATAΔ-elements and TC-elements are all responsible for DNA instability around -30 but at different levels [25] promoting a continuous range of control on gene transcription.

Either way, we believe that our observations call for further biological analysis and that our predictions will be useful for future assessment of the relationships between promoter architecture and gene expression. In this respect, future studies in model plants should be based on a better description of the different organization of core-promoters and include a more precise definition of the location of alternative TSSs based on new sequencing technologies.

Materials and Methods

Promoter sets

Arabidopsis thaliana and *Oryza sativa* promoters were built from transcripts, Expressed Sequence Tags and full-length cDNA (The *Arabidopsis* Information Resource - TAIR R.6 [41] and The Institute for Genomic Research - TIGR R.3 [58] respectively). We used FLAGdb^{++} [59], an integrative database of plant model genomes, to define the transcriptional units by aligning all the available transcripts to gene models, excluding pseudogenes. We excluded promoters with a 5'UTR smaller than 50 bases in order to avoid truncated UTR. Our promoter sets included 14927 *A. thaliana* and 18012 *O. sativa* sequences extending 1kb upstream of the predicted TSS and containing the whole 5'UTR. The 15802 *Homo sapiens*

and the 15833 *Mus musculus* promoters were retrieved from the DBTSS [39]. To maintain a consistent approach to that used to construct the plant promoter set, we selected only the TSSs located the most upstream of genes.

Identification of Preferentially Located Motifs

For each motif analysed, we extracted all occurrences within a promoter set. The motif position corresponds to the position of the first base of the sequence. Motif distributions were determined using a one-base-long sliding-window with a one-base shift. For each window, we counted the number of promoters containing the motif rather than the number of occurrences to avoid favouring repeated motifs. The promoter sequences were then divided into two regions. First, the [-1000, -300] region was used to learn the distribution model using a simple linear regression and determine a 99% confidence interval. Second, within the [-300, 500] region, we searched for non-evenly distributed motifs, *i.e.* those exhibiting a peak above the confidence interval. We increased the sliding-window size from 1 to 100 bases step-by-step when a second peak was detected or when the learning area contained at least one window within which no motif was detected, in order to avoid non-accurate learning of the distribution model.

For each PLM we recorded: (i) the motif distribution, (ii) the window size, (iii) the preferential position, *i.e.* the position of the peak top in the distribution, (iv) the peak boundaries allowing the functional window to be located, (v) the list of promoters containing the motif and (vi) the Score of Maximal Square relative to the base line or SMS ,*i.e.* the ratio (peak top minus base line)/(upper bound of the confidence interval minus base line). We only considered PLMs characterized by an SMS greater than 1.

Sequence distance graph and seed motifs

Out of all $_{[-39,\ -26]}$-PLMs, we searched for those differing by a single substitution and represented the links by means of a graph. For both TA-PLM and TC-PLM graphs we

identified the seed(s), *i.e.* PLM(s) only connected in the graph to PLMs less present in promoters than itself. The graph is thus oriented relative to the seed(s). The TA-PLM graph comprises the TATA-box and all the 15 TATAΔ-PLMs. The $TC_{[-39, -26]}$-PLM graph was first constructured with the PLMs observed in more than 200 promoters. Then, remaining PLMs were added considering only links with the PLMs making up the graph and not those with the remaining PLMs.

Extended motif analyses

PLM relevance is ranked by the SMS value. For each PLM, we analyzed the SMS of all motifs extended by one base upstream or downstream of the initial PLM. We referred to these motifs as extended-motifs. Selected extended-motifs were only those that (i) were a PLM, (ii) shared the initial PLM constraints, *i.e.* a sharp peak in the [-39, -26] region and (iii) were characterized by an SMS 1.1-fold higher than that of the initial PLM, *i.e.* exhibited a stronger topological constraint.

Gene structure and gene function

All information about gene structure was obtained from $FLAGdb^{++}$ [59] including: (i) the median of the lengths of frames, coding sequences (CDSs), 3'UTRs and 5'UTRs, first introns and cumulated introns; and (ii) the percentage of intron-less genes.

Information about gene function was obtained from the TAIR Gene Ontology (GO) categories [60] [41]. We used the GO Biological Function and GO Cellular Component categories. For a given gene set, we calculated the percentage of genes within that category relative to the remaining *A. thaliana* genes with an experimentally supported TSS. Finally, we only considered previously annotated genes, *i.e.* we excluded the "unknown" and the "other" categories.

Gene expression

The transcriptome data used in this work were obtained using the CATMA v2 microarray [61]. They include 522 hybridized samples extracted from 40 different projects covering 12 organ types. Expression data were downloaded from CATdb [50] containing all the transcriptome data generated by the CATMA-URGV platform. They had also been deposited in either the NCBI Gene Expression Omnibus [62] or the European Bioinformatics Institute ArrayExpress [63] repositories. Out of the 14927 *A. thaliana* genes with an experimentally supported TSS, we analyzed the expression data of 11161 genes specifically relating to one CATMA probe. Each gene was spotted according to (i) the percentage of samples in which it had been detected, called the hybridization percentage and (ii) its median expression level. We performed a third order linear regression allowing us to cluster 4371 genes with a High level of Expression (HE) and 6790 with a Low level of Expression level (LE). Different hybridization ranges were also defined according to two inflections in the linear regression curve. The Small Range (SR) class included 4510 genes characterized by a hybridisation percentage of less than 15% and the Wide Range (WR) class included 1241 genes characterized by a hybridisation percentage of over 85%. Second, we built a 60% confidence interval allowing the identification, within each class, of the highest and the lowest expression levels (respectively HE+ and LE+).

Orthologous genes

We selected those *A. thaliana* and *O. sativa* orthologous gene pairs being Bidirectional Best Hits [64] with BLASTp [65] resulting in 5805 orthologous pairs of genes having an experimentally supported TSS. The presence of a PLM in both genomes led to the distinction between the presence of this PLM within (i) *A. thaliana* promoters but not in their respective *O. sativa* orthologue; (b) *O. sativa* promoters but not in their respective *A. thaliana* orthologue; or (c) both orthologous genes. For each PLM-class, we performed comparisons

between expected and observed conservation within orthologous gene pairs. The c/(a+c) ratio indicates the level of observed conservation in *O. sativa* with respect to *A. thaliana*. The expected conservation value is given by the ratio b/(5805-a-c), *i.e.* the presence of a given PLM-class in *O. sativa* orthologous genes from the PLM-class-less *A. thaliana* genes.

Statistical analysis

Statistical analyses were performed with the R statistical software [66]. We used R for (i) the regression analysis leading to PLM identification (Figure 1) and the characterization of expression categories (Figure 7), and for (ii) motif distributions (Figures 3 and 6). We performed Fisher exact one-sided tests using the Bonferroni correction to compare percentages between two independent samples. We searched for statistical decreases or increases in the percentages of one set of genes (i) in GO annotation categories (Table 1), (ii) intron-less (Table 2, last row), (iii) in expression categories (Table 3) and (iv) with a given motif (Table 4). Structural gene features, even after log modification, cannot be assumed to be normally distributed. We therefore performed Wilcoxon-Mann-Whitney one-sided tests using the Bonferroni correction to compare two independent structural gene feature distributions. We considered the hypothesis that structural data might result in higher or lower values for a given gene set compared to all other genes, *i.e.* the 14927 genes minus the considered gene set (Table 2, first rows). Each p value less than 5% with the Bonferroni correction was considered significant.

Abbreviations

CATMA: Complete *Arabidopsis* Transcriptome MicroArray

TFBS: Transcription Factor Binding Site

GO: Gene Ontology

HE: High Expression

LE: Low Expression

PLM: Preferentially Located Motif

SMS: Score of Maximal Square relative to the base line

SR: Small Range

TBP: TATA-binding protein

TSS: Transcriptional Start Site

WR: Wide Range

Author contributions

VBe performed the analyses and wrote the manuscript. VBr coordinated the analyses and helped to write the manuscript. AL managed the study and helped to write the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Marie-Laure Martin-Magniette for her statistical advices, Jérémy Just and Franck Samson for their advices for orthologous gene analysis, Sébastien Aubourg for his helpful comments on the manuscript and Jean-Philippe Tamby, Fabrice Touzain and Joan Sobota for their careful reading of the manuscript.

References

1. Muller F, Demeny MA, Tora L: **New problems in RNA polymerase II transcription initiation: matching the diversity of core promoters with a variety of promoter recognition factors**. *J Biol Chem* 2007, **282**(20):14685-14689.
2. Gross P, Oelgeschlager T: **Core promoter-selective RNA polymerase II transcription**. *Biochem Soc Symp* 2006(73):225-236.
3. Shi W, Zhou W: **Frequency distribution of TATA Box and extension sequences on human promoters**. *BMC Bioinformatics* 2006, **7 Suppl 4**:S2.
4. Basehoar AD, Zanton SJ, Pugh BF: **Identification and distinct regulation of yeast TATA box-containing genes**. *Cell* 2004, **116**(5):699-709.
5. Tsai FT, Sigler PB: **Structural basis of preinitiation complex assembly on human pol II promoters**. *Embo J* 2000, **19**(1):25-36.
6. Moshonov S, Elfakess R, Golan-Mashiach M, Sinvani H, Dikstein R: **Links between core promoter and basic gene features influence gene expression**. *BMC Genomics* 2008, **9**(1):92.
7. Nakamura M, Tsunoda T, Obokata J: **Photosynthesis nuclear genes generally lack TATA-boxes: a tobacco photosystem I gene responds to light through an initiator**. *Plant J* 2002, **29**(1):1-10.
8. Patikoglou GA, Kim JL, Sun L, Yang SH, Kodadek T, Burley SK: **TATA element recognition by the TATA box-binding protein has been conserved throughout evolution**. *Genes Dev* 1999, **13**(24):3217-3230.
9. Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E: **Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters**. *Gene* 2007, **389**(1):52-65.

10. Javahery R, Khachi A, Lo K, Zenzie-Gregory B, Smale ST: **DNA sequence requirements for transcriptional initiator activity in mammalian cells**. *Mol Cell Biol* 1994, **14**(1):116-127.

11. Burke TW, Kadonaga JT: **The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila**. *Genes Dev* 1997, **11**(22):3020-3031.

12. Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH: **New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB**. *Genes Dev* 1998, **12**(1):34-44.

13. Gershenzon NI, Trifonov EN, Ioshikhes IP: **The features of Drosophila core promoters revealed by statistical analysis**. *BMC Genomics* 2006, **7**:161.

14. Thompson W, Rouchka EC, Lawrence CE: **Gibbs Recursive Sampler: finding transcription factor binding sites**. *Nucleic Acids Res* 2003, **31**(13):3580-3585.

15. van Helden J: **Regulatory sequence analysis tools**. *Nucleic Acids Res* 2003, **31**(13):3593-3596.

16. Bellora N, Farre D, Alba MM: **Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters**. *BMC Genomics* 2007, **8**:459.

17. Berendzen KW, Stuber K, Harter K, Wanke D: **Cis-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves**. *BMC Bioinformatics* 2006, **7**:522.

18. Kielbasa SM, Korbel JO, Beule D, Schuchhardt J, Herzel H: **Combining frequency and positional information to predict transcription factor binding sites**. *Bioinformatics* 2001, **17**(11):1019-1026.

19. Molina C, Grotewold E: **Genome wide analysis of Arabidopsis core promoters**. *BMC Genomics* 2005, **6**(1):25.

20. Narang V, Sung WK, Mittal A: **Computational modeling of oligonucleotide positional densities for human promoter prediction**. *Artif Intell Med* 2005, **35**(1-2):107-119.

21. Yamamoto YY, Ichida H, Matsui M, Obokata J, Sakurai T, Satou M, Seki M, Shinozaki K, Abe T: **Identification of plant promoter constituents by analysis of local distribution of short sequences**. *BMC Genomics* 2007, **8**:67.

22. Defrance M, Touzet H: **Predicting transcription factor binding sites using local over-representation and comparative genomics**. *BMC Bioinformatics* 2006, **7**:396.

23. Cowan MJ, Yao XL, Pawliczak R, Huang X, Logun C, Madara P, Alsaaty S, Wu T, Shelhamer JH: **The role of TFIID, the initiator element and a novel 5' TFIID binding site in the transcriptional control of the TATA-less human cytosolic phospholipase A2-alpha promoter**. *Biochim Biophys Acta* 2004, **1680**(3):145-157.

24. Bogomolski-Yahalom V, Klein A, Greenblat I, Haviv Y, Tur-Kaspa R: **The TATA-less promoter of hepatitis B virus S gene contains a TBP binding site and an active initiator**. *Virus Res* 1997, **49**(1):1-7.

25. Abeel T, Saeys Y, Bonnet E, Rouze P, Van de Peer Y: **Generic eukaryotic core promoter prediction using structural features of DNA**. *Genome Res* 2008, **18**(2):310-323.

26. Civan P, Svec M: **Genome-wide analysis of rice (Oryza sativa L. subsp. japonica) TATA box and Y Patch promoter elements**. *Genome* 2009, **52**(3):294-297.

27. FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C: **Clustering of DNA sequences in human promoters**. *Genome Res* 2004, **14**(8):1562-1574.

28. Wolfe KH, Gouy M, Yang YW, Sharp PM, Li WH: **Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data**. *Proc Natl Acad Sci U S A* 1989, **86**(16):6201-6205.

29. Chaw SM, Chang CC, Chen HL, Li WH: **Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes**. *J Mol Evol* 2004, **58**(4):424-441.

30. Smale ST, Kadonaga JT: **The RNA polymerase II core promoter**. *Annu Rev Biochem* 2003, **72**:449-479.

31. Joshi CP: **An inspection of the domain between putative TATA box and translation start site in 79 plant genes**. *Nucleic Acids Res* 1987, **15**(16):6643-6653.

32. Singer VL, Wobbe CR, Struhl K: **A wide variety of DNA sequences can functionally replace a yeast TATA element for transcriptional activation**. *Genes Dev* 1990, **4**(4):636-645.

33. Loganantharaj R: **Discriminating TATA box from putative TATA boxes in plant genome**. *Int J Bioinform Res Appl* 2006, **2**(1):36-51.

34. Kiran K, Ansari SA, Srivastava R, Lodhi N, Chaturvedi CP, Sawant SV, Tuli R: **The TATA-box sequence in the basal promoter contributes to determining light-dependent gene expression in plants**. *Plant Physiol* 2006, **142**(1):364-376.

35. Fujimori S, Washio T, Higo K, Ohtomo Y, Murakami K, Matsubara K, Kawai J, Carninci P, Hayashizaki Y, Kikuchi S *et al*: **A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription**. *FEBS Lett* 2003, **554**(1-2):17-22.

36. Morgante M, Hanafey M, Powell W: **Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes**. *Nat Genet* 2002, **30**(2):194-200.

37. Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E: **AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors**. *BMC Bioinformatics* 2003, **4**:25.

38. Higo K, Ugawa Y, Iwamoto M, Korenaga T: **Plant cis-acting regulatory DNA elements (PLACE) database: 1999**. *Nucleic Acids Res* 1999, **27**(1):297-300.

39. Wakaguri H, Yamashita R, Suzuki Y, Sugano S, Nakai K: **DBTSS: database of transcription start sites, progress report 2008**. *Nucleic Acids Res* 2008, **36**(Database issue):D97-101.

40. Drawid A, Gupta N, Nagaraj VH, Gelinas C, Sengupta AM: **OHMM: a Hidden Markov Model accurately predicting the occupancy of a transcription factor with a self-overlapping binding motif**. *BMC Bioinformatics* 2009, **10**(1):208.

41. Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M *et al*: **The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community**. *Nucleic Acids Res* 2003, **31**(1):224-228.

42. Boden M, Bailey TL: **Associating transcription factor-binding site motifs with target GO terms and target genes**. *Nucleic Acids Res* 2008, **36**(12):4108-4117.

43. Mingam A, Toffano-Nioche C, Brunaud V, Boudet N, Kreis M, Lecharny A: **DEAD-box RNA helicases in Arabidopsis thaliana: establishing a link between quantitative expression, gene structure and evolution of a family of genes**. *Plant Biotechnol J* 2004, **2**(5):401-415.

44. Carninci P: **Tagging mammalian transcription complexity**. *Trends Genet* 2006, **22**(9):501-510.

45. Yamamoto YY, Yoshitsugu T, Sakurai T, Seki M, Shinozaki K, Obokata J: **Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis**. *Plant J* 2009.

46. Zhu Q, Dabi T, Lamb C: **TATA box and initiator functions in the accurate transcription of a plant minimal promoter in vitro**. *Plant Cell* 1995, **7**(10):1681-1689.

47. Ponjavic J, Lenhard B, Kai C, Kawai J, Carninci P, Hayashizaki Y, Sandelin A: **Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters**. *Genome Biol* 2006, **7**(8):R78.

48. Bajic VB, Tan SL, Christoffels A, Schonbach C, Lipovich L, Yang L, Hofmann O, Kruger A, Hide W, Kai C *et al*: **Mice and men: their promoter properties**. *PLoS Genet* 2006, **2**(4):e54.

49. Sclep G, Allemeersch J, Liechti R, De Meyer B, Beynon J, Bhalerao R, Moreau Y, Nietfeld W, Renou JP, Reymond P *et al*: **CATMA, a comprehensive genome-scale resource for silencing and transcript profiling of Arabidopsis genes**. *BMC Bioinformatics* 2007, **8**:400.

50. Gagnot S, Tamby JP, Martin-Magniette ML, Bitton F, Taconnat L, Balzergue S, Aubourg S, Renou JP, Lecharny A, Brunaud V: **CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform**. *Nucleic Acids Res* 2008, **36**(Database issue):D986-990.

51. Aubourg S, Martin-Magniette ML, Brunaud V, Taconnat L, Bitton F, Balzergue S, Jullien PE, Ingouff M, Thareau V, Schiex T *et al*: **Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome**. *BMC Genomics* 2007, **8**:401.

52. Ma X, Zhang K, Li X: **Evolution of Drosophila ribosomal protein gene core promoters**. *Gene* 2009, **432**(1-2):54-59.

53. Armisen D, Lecharny A, Aubourg S: **Unique genes in plants: specificities and conserved features throughout evolution**. *BMC Evol Biol* 2008, **8**:280.

54. Tanaka T, Koyanagi KO, Itoh T: **Highly diversified molecular evolution of downstream transcription start sites in rice and Arabidopsis**. *Plant Physiol* 2009, **149**(3):1316-1324.

55. Tirosh I, Weinberger A, Carmi M, Barkai N: **A genetic signature of interspecies variations in gene expression**. *Nat Genet* 2006, **38**(7):830-834.

56. Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL: **Genetic properties influencing the evolvability of gene expression**. *Science* 2007, **317**(5834):118-121.

57. Raser JM, O'Shea EK: **Control of stochasticity in eukaryotic gene expression**. *Science* 2004, **304**(5678):1811-1814.

58. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L *et al*: **The TIGR Rice Genome Annotation Resource: improvements and new features**. *Nucleic Acids Res* 2007, **35**(Database issue):D883-887.

59. Samson F, Brunaud V, Duchene S, De Oliveira Y, Caboche M, Lecharny A, Aubourg S: **FLAGdb++: a database for the functional analysis of the Arabidopsis genome**. *Nucleic Acids Res* 2004, **32**(Database issue):D347-350.

60. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

61. Hilson P, Allemeersch J, Altmann T, Aubourg S, Avon A, Beynon J, Bhalerao RP, Bitton F, Caboche M, Cannoot B *et al*: **Versatile gene-specific sequence tags for Arabidopsis functional genomics: transcript profiling and reverse genetics applications**. *Genome Res* 2004, **14**(10B):2176-2189.

62. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles--database and tools update**. *Nucleic Acids Res* 2007, **35**(Database issue):D760-765.

63. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG *et al*: **ArrayExpress--a public repository for microarray gene expression data at the EBI**. *Nucleic Acids Res* 2003, **31**(1):68-71.

64. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling**. *Proc Natl Acad Sci U S A* 1999, **96**(6):2896-2901.

65. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

66. **The Comprehensive R Archive Network** [http://cran.r-project.org].

Figure legends

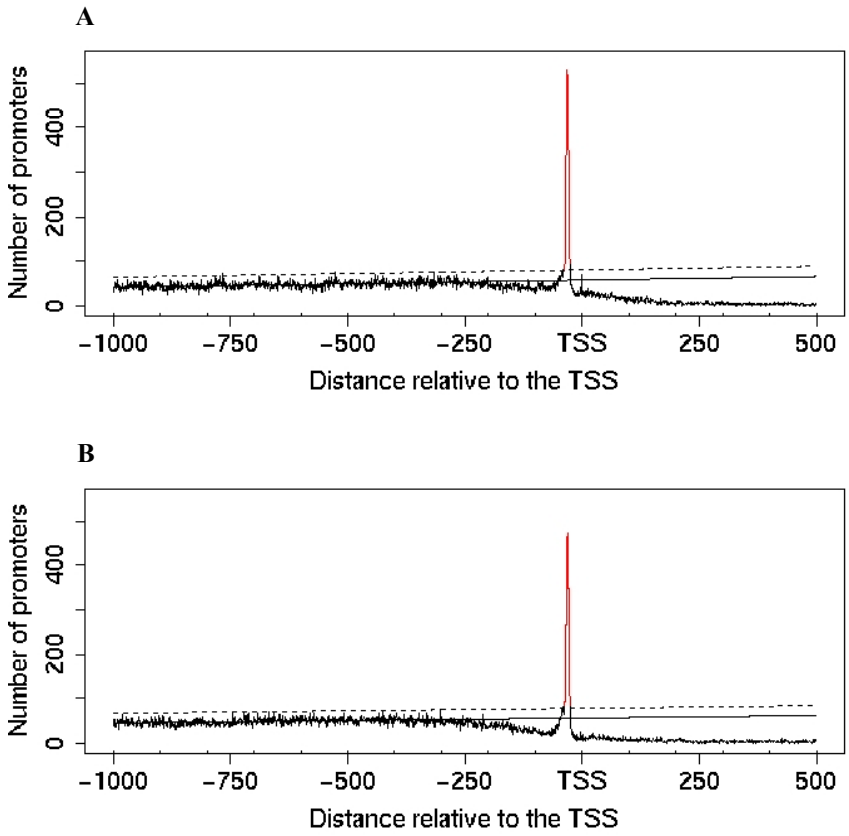


Figure 1: Distribution of TATAWA in plant promoters aligned relative to the TSS.

The canonical TATA-box defines the TATAWA-PLM. The distribution model is learned in the [-1000, -300] region where the base line (continuous line) and the upper bound of the confidence interval (dashed line) are estimated and applied to the [-300, 500] region. Distributions are shown for the *Arabidopsis thaliana* set containing 14927 promoters (A) and for the *Oryza sativa* set containing 18012 promoters (B). The TATAWA-PLM is preferentially positioned 32 bases upstream of the TSS and a [-39, -26] functional window in both genomes. *A. thaliana* and *O. sativa* are characterized by an SMS of 21 and 19 respectively.

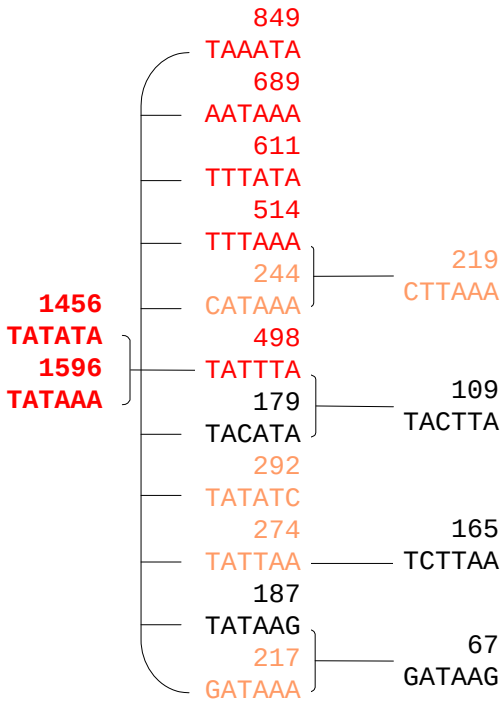


Figure 2: Sequence distance graph of TA-PLMs present in *A. thaliana* promoters.

The TATA-box PLM sequence (first column), TATAΔ1-PLM sequences (second column) and TATAΔ2-PLM sequences (third column) are organized in an oriented graph. In red the motifs observed in more than 350 promoters; in orange the motifs observed in more than 200 promoters and in black the motifs observed in up to 200 sequences. Each edge between two PLMs reflects the presence of one substitution leading from one PLM to another one. Numbers of *A. thaliana* promoters containing a given motif are indicated above the sequences. The TATAWA motifs are seeds (in bold).

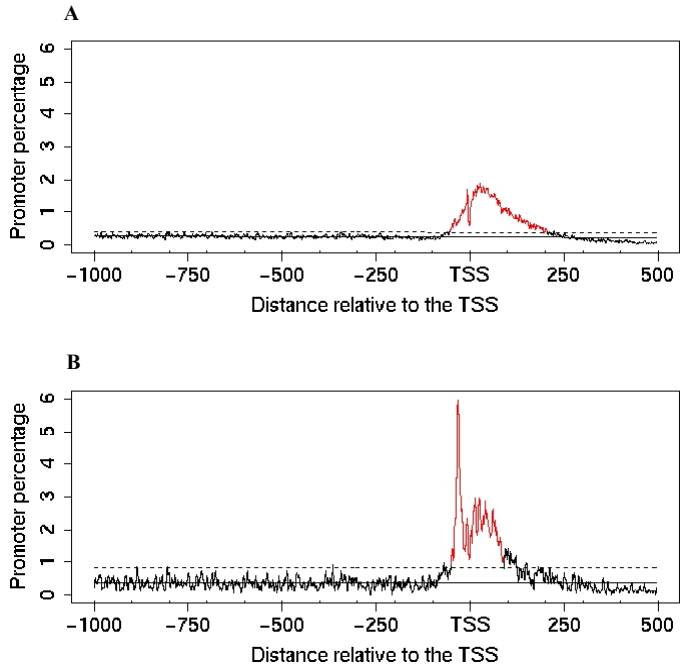


Figure 3: Distributions of TTCTTC in *A. thaliana* promoters aligned relative to the TSS.

TTCTTC is one of the 29 conserved $TC_{[-39, -26]}$-PLMs. Distributions are shown for the whole promoter set containing 14927 promoters (A) and for the 1745 TA-less promoters (B). The y-axis percentage depends on the promoter number in each data set. TTCTTC is preferentially positioned at +29 (SMS 10) and -33 (SMS 8) in the whole promoter set and in the TA-less promoter set respectively. In both cases, the size of the window used to scan the promoters was 2 bases.

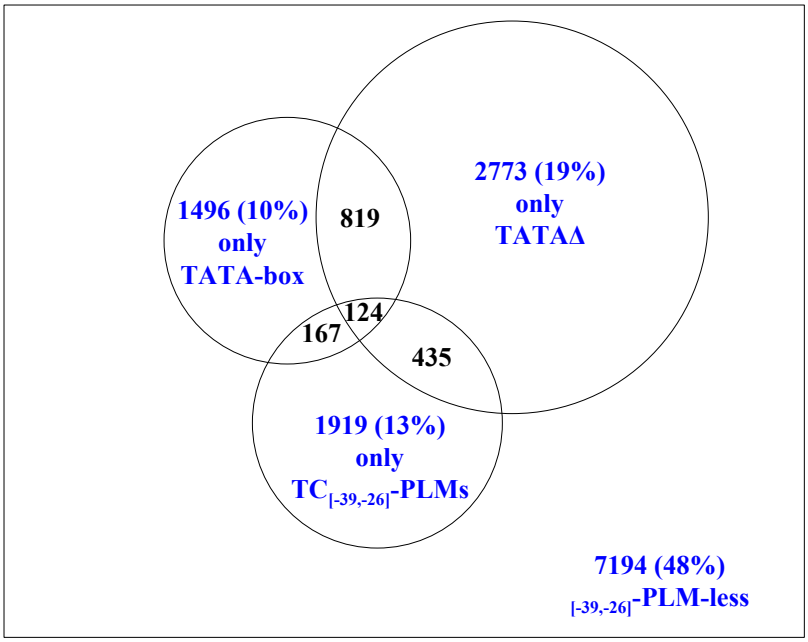


Figure 4: Presence in *A. thaliana* promoters of the three PLM classes characterized by a sharp functional window and preferentially located in the [-39, -26] region.

The square represents the 14927 *A. thaliana* promoters (the whole promoter set) with experimentally known TSS. The central Venn diagram illustrates the overlaps between the three PLM classes in promoters. In the [-39, -26] region, 2606 promoters contained a canonical TATA-box, 4151 a TATAΔ-PLM and 2645 a $TC_{[-39, -26]}$-PLM. In black the overlap between the three classes. In blue, the promoter sets analysed in the study, *i.e.* the promoter sets with PLMs from only one of the three classes and the promoter set that does not contain PLMs of the three classes.

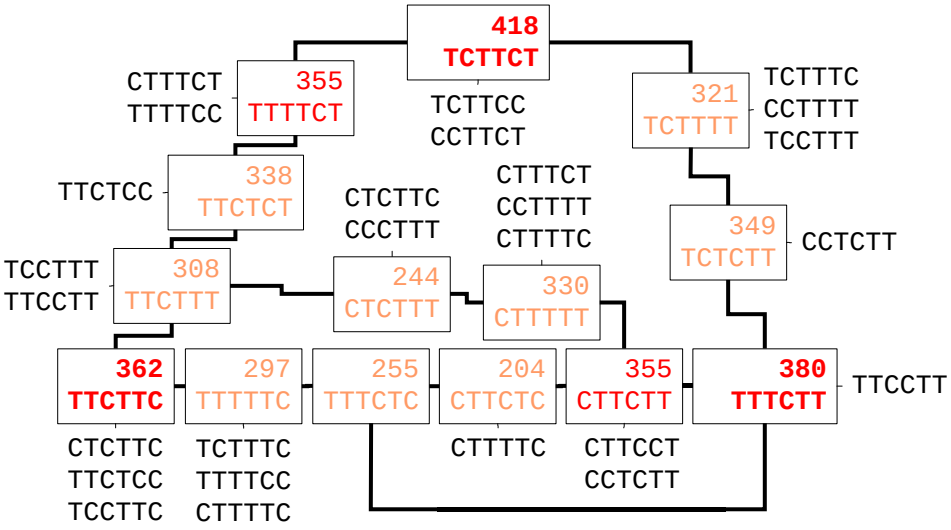


Figure 5: Sequence distance graph of the $TC_{[-39,\ -26]}$-PLMs present in *A. thaliana* promoters.

For each $TC_{[-39,\ -26]}$-PLM observed in more than 200 *A. thaliana* promoters, the number of promoters containing the PLM is indicated above the sequence (surrounded data). In red PLMs observed in more than 350 promoters; in orange, PLMs in more than 200 promoters and in black PLMs observed in up to 200 sequences. $TC_{[-39,\ -26]}$-PLMs observed in more than 200 promoters are organized in an oriented and closed graph. Each edge between two PLMs reflects the presence of one substitution leading from one PLM to another. Edges between the less observed PLMs (in black) are not considered. The three seeds of this graph (in bold) are PLMs of which all the directly connected motifs are less frequently observed than the seed-PLMs themselves. Only one graph is possible.

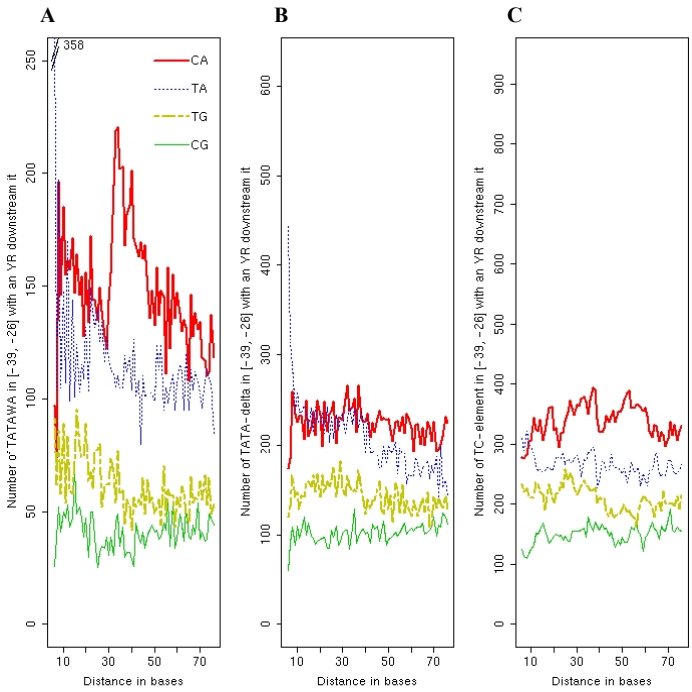


Figure 6: Distance between the four YR-TSSs relative to the three different classes of $_{[-39, -26]}$-PLMs in *A. thaliana* promoters.

We computed the distance observed between each PLM of a given class of $_{[-39, -26]}$-PLMs and all dinucleotides CA (red line), TG (green line), TA (blue line) and CG (yellow line) present downstream of the PLMs. The distance on the x-axis corresponds to the number of bases between the first base of a $_{[-39, -26]}$-PLM and the first base of an YR-TSS. For instance, a distance of 8 bases between a TATAWA and a CA corresponds to the TATAWANNCA sequence in a promoter. We analysed the three gene sets containing a unique class of $_{[-39, -26]}$-PLMs and represented the distance between (A) the TATAWA-, (B) the TATAΔ- and (C) the TC$_{[-39, -26]}$-PLMs and the four Inr. Analyses done for each PLM led to the same representation as the global analysis done for the PLM class it belongs to: for instance, the analysis for the TTCTTC-PLM led to a (C) representation.

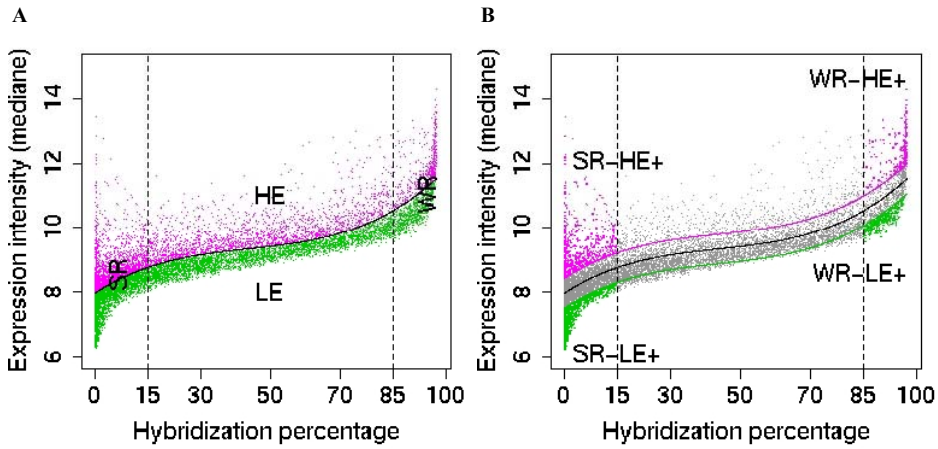


Figure 7: Subsets of *A. thaliana* genes characterized by different expression intensities and specificities

Each gene is plotted according to the percentage of plant samples by which the gene specific probe has been hybridized (x-label) and to the median of expression intensities observed in the different samples (y-label). (A) First, we estimated the base line for all data (in black) and we characterized two gene groups: the highly expressed genes (HE, in magenta, 4371 genes) above the base line and the lowly expressed genes (LE, in green, 6790 genes) below the base line. Second, we identified two inflections in the base line corresponding to 15% and 85% of hybridizations (vertical dashed black lines). These limits characterized two extreme sets of genes, the genes exhibiting the smallest range of hybridizations (SR, 4510 genes) and the gene exhibiting the widest range of hybridizations (WR, 1241 genes). (B) From the SR and WR gene sets, we defined 4 sets of genes with extreme hybridization intensities by estimating an upper (magenta line) and lower (green line) bound at the 60% confidence interval. Thus, there are 618 SR-HE+, 822 SR-LE+, 244 WR-HE+ and 385 WR-LE+ genes.

Tables

Gene sets		All genes	Only TATA-box		Only $TC_{[-39, -26]}$-PLMs		Only TATAΔ-PLMs		$_{[-39, -26]}$-PLM-less	
GO Biological process	Response to abiotic or biotic stimulus	8.8	**12**	**($7e^{-4}$)**	7.7	(NS)	7.7	(NS)	8.1	(NS)
	Response to stress	8.8	**11**	**($1e^{-3}$)**	8.6	(NS)	8.2	(NS)	7.7	(NS)
	Protein metabolism	15	12	(NS)	<u>18</u>	<u>($1e^{-3}$)</u>	16	(NS)	15	(NS)
GO Cellular component	Cell wall	3.1	**5.8**	**($1e^{-6}$)**	<u>1.8</u>	<u>($6e^{-4}$)</u>	3.1	(NS)	2.7	(NS)
	Ribosome	2.6	**4.2**	**($1e^{-3}$)**	1.8	(NS)	3.1	(NS)	2.4	(NS)
	Cytosol	2.8	**5**	**($4e^{-5}$)**	1.9	(NS)	3.2	(NS)	2.4	(NS)
	Golgi apparatus	1.4	<u>0.3</u>	<u>($3e^{-5}$)</u>	1.3	(NS)	1.8	(NS)	1.6	(NS)
	Mitochondria	5.6	<u>3.2</u>	<u>($3e^{-5}$)</u>	4.8	(NS)	6	(NS)	6.6	(NS)
	Plastid	6.1	<u>3.6</u>	<u>($6e^{-5}$)</u>	5.2	(NS)	7	(NS)	7	(NS)
	Chloroplast	15	<u>11</u>	<u>($5e^{-6}$)</u>	15	(NS)	16	(NS)	17	(NS)

Table 1: Functional category profiles of *A. thaliana* genes for the four promoter classes.

Gene products have been retrieved from the GO categories [60] provided by TAIR [41]. In the table, each number not in parenthesis is the percentage of genes annotated in a given GO category. We performed the two one-sided Fisher exact tests allowing the identification of enrichment (bold) or impoverishment (underlined) of GO categories in a given set of genes when compared with all the other genes, *i.e.* genes within the whole gene set minus genes within the considered gene set. NS indicates a non-significant difference between percentages. P-values in parenthesis are less than 5% with the Bonferroni correction. Results are only shown for GO categories exhibiting at least one bias.

Gene sets		All genes	Only TATA-box	Only $TC_{[-39, -26]}$-PLMs	Only TATAΔ-PLMs	$_{[-39, -26]}$-PLM-less
Length median	Gene	2293	1936 ($5e^{-23}$)	**2385 ($8e^{-8}$)**	2293 (NS)	**2334 ($2e^{-7}$)**
	5'UTR	158	112 ($7e^{-38}$)	**175 ($3e^{-5}$)**	161 (NS)	**173 ($9e^{-10}$)**
	CDS	1086	966 ($5e^{-14}$)	**1185 ($3e^{-10}$)**	1071 (NS)	**1119 ($1e^{-4}$)**
	All introns	588	521 ($1e^{-4}$)	605 (NS)	588 (NS)	614 (NS)
Percentage	Intron-less	18.8	**24.5 ($4e^{-9}$)**	16.7 (NS)	17.5 (NS)	17.1 (NS)

Table 2: Structural gene features of *A. thaliana* genes for the four promoter classes.

Structural gene features have been assigned by querying the $FLAGdb^{++}$ database [59]. For median length data, we performed two one-sided Wilcoxon tests allowing the identification of enrichment in wide (bold) or in large structures (underlined) in a set of genes compared with all the other genes, *i.e.* genes within the whole gene set minus genes within the considered gene set . For intron-less gene percentages, we performed two one-sided Fisher exact tests allowing the identification of higher (bold) or lower (underlined) percentages in a gene set in comparison with all the other genes. NS indicates a non-significant difference. P-values in parenthesis are less than 5% with the Bonferroni correction. Both the first intron and 3'UTR lengths are never biased (data not shown).

Gene sets	All genes	Only TATA-box	Only $TC_{[-39, -26]}$-PLMs	Only TATAΔ-PLMs	$_{[-39, -26]}$-PLM-less
HE	39.2%	**53.8% (<1e^{-30})**	36.7% (NS)	38.1% (NS)	34.5% (1e^{-22})
LE	60.8%	46.2% (2e^{-24})	63.3% (NS)	61.9% (NS)	**65.5% (<1e^{-30})**
SR	40.4%	41% (NS)	**44.1% (<1e^{-30})**	37.4% (<1e^{-30})	40.2% (NS)
WR	11.1%	10.3% (NS)	9.3% (NS)	**13.2% (<1e^{-30})**	11.2% (NS)
WR-HE+	2.2%	2.4% (NS)	1.6% (NS)	2.2% (NS)	2.3% (NS)
WR-LE+	3.4%	2.5% (NS)	3.1% (NS)	4.0% (NS)	3.8% (NS)
SR-HE+	5.5%	**8.5% (1e^{-5})**	5.3% (NS)	5.3% (NS)	4.6% (8e^{-6})
SR-LE+	7.4%	5% (<1e^{-30})	8.6% (NS)	6.8% (NS)	**8.1% (<1e^{-30})**

Table 3: Expression of *A. thaliana* genes for the four promoter classes.

The 8 expression categories are described in the Figure 7. HE, High Expression; LE, Low Expression; SR Small Range of expression; WR, Wide Range of expression; HE^+ Highest Expressions; LE^+, Lowest Expressions. The whole gene set contains the 11161 genes for which we have both a known TSS and CATMA-expression data. We performed two one-sided Fisher exact tests allowing the identification of higher (bold) or lower (underlined) percentages in a gene set in comparison with all other genes, *i.e.* genes within the whole gene set minus genes within the considered gene set. NS indicates non-significant difference. P-values in parenthesis are less than 5% with the Bonferroni correction.

	Number of ortholog pairs with PLM in			Percentage of PLM conservation	
	A. thaliana (a)	*O. sativa* (b)	Both species (c)	observed	expected by chance:
Only TATAWA	393	343	82	**17% ($7e^{-15}$)**	6%
Only TATAΔ-PLMs	997	531	138	12% (NS)	11%
Only $TC_{[-39, -26]}$-PLMs	655	602	98	13% (NS)	12%

Table 4: Conservation of the three $_{[-39, -26]}$-PLM classes in orthologous gene pairs between *A. thaliana* and *O. sativa.*

In the 5805 pairs of orthologous genes between *A. thaliana* and *O. sativa*, we searched for the number of genes containing a given $_{[-39, -26]}$-PLM (a) in *A. thaliana* but not in their respective *O. sativa* ortholog, (b) in *O. sativa* but not in their respective *A. thaliana* ortholog and (c) in both orthologous genes. The c/(a+c) ratio indicates the level of observed conservation for each PLM class in *O. sativa* in regard to *A. thaliana*. Then we compared the ratio to the by chance value, *i.e.* the ratio b/(5805-a-c): the presence of a given PLM in *O. sativa* orthologous genes from the PLM-less *A. thaliana* genes. We performed two one-sided Fisher exact tests allowing the identification of significantly different percentages. The underlined data indicate that the observed percentage is higher than expected. NS indicates non-significant difference. P-values in parenthesis are less than 5% with the Bonferroni correction.

Additional files

Additional file 1 - Effect on the score of an extension by pyrimidines of $TC_{[-39, -26]}$-PLMs.

First column, motifs that do not exhibit an increase in their SMS (in parentheses) when extended. Second column, motifs that, when extended exhibit an increase in their original SMS due to the generation of an other 6-base-long motif (underlined sequence) characterized by an higher SMS (underlined SMS). Note that the SMS of the 7- or 8-base-long motifs are all lower than the SMS of the underlined 6-base-long motifs. Third column, motifs that can be extended in 7- or 9-base-long motifs. Note that the two long extensions are made of three repeats of TCT.

Cf. Annexe VI.B

Cet article présente une recherche exhaustive de motifs de 6 bases de long au sein des promoteurs ne contenant pas de richesse en AT, ni TA, ni AA dans la région de la boîte TATA. Des PLM caractérisés par les mêmes contraintes topologiques que celles de la boîte TATA et conservés entre *A. thaliana* et *O. sativa* ont été mis en évidence. Parmi eux, les 29 PLM dont les séquences sont constituées exclusivement des bases C et T sont les motifs-TC. Ces motifs ne sont pas présents chez les mammifères considérés lors de cette étude : *H. sapiens* et *M. musculus*.

Les caractéristiques fonctionnelles et structurales des motifs-TC les distinguent de la boîte TATA et de ses variants. En particulier, les gènes contenant ces motifs ont plus tendance à :

- être impliqués dans des fonctions liées au métabolisme des protéines ;
- avoir des unités de transcription étroites, avec en particulier des CDS et des UTR 5' courts ;
- s'exprimer dans un faible pourcentage des expériences analysées.

En conclusion, cette étude a permis de confirmer la présence suspectée d'une nouvelle classe de motifs : les motifs-TC. Ils sont potentiellement impliqués dans la régulation de l'expression des gènes. Ces motifs, sont présents chez *A. thaliana*, *O. sativa*, mais aussi chez le peuplier, mais n'ont pas été identifiés dans des génomes en dehors du règne végétal.

a) Motifs-TC et les microsatellites riches en bases C et T: deux catégories d'éléments à distinguer

Les motifs-TC ressemblent beaucoup aux séquences riches en C et T associés à des microsatellites. Ces deux catégories d'éléments ont été analysées et leurs caractéristiques ont été comparées afin de pouvoir définir leurs divergences.

i) Les séquences des motifs-TC sont plus riches en T

La liste des 29 motifs-TC a été comparée à la liste des 46 motifs de 6 bases de long constitués exclusivement de bases T ou C mis en évidence dans la région II lors de l'analyse globale et qui représentent les microsatellites dans les UTR 5' d'*A. thaliana*. Leurs contraintes topologiques sont résumées dans la Table 7-8.

	Motifs-TC	Microsatellites riches en C et T
Position préférentielle	-32	Dans l'UTR 5'
Largeur de la fenêtre fonctionnelle	Etroite - moins de 20 bases	Large - du promoteur central à l'ATG
Prévalence : motifs observés dans	17% des gènes	près de l'ensemble des promoteurs
Composition en bases des séquences	Plus riches en T	Plus riches en C
Extension des motifs	Faible voir inexistante	Large

Table 7-8 : Caractéristiques des motifs-TC et des microsatellites riches en bases C et T.

La Figure 7-7 présente ces deux listes de motifs ainsi que leurs chevauchements. Certains motifs sont spécifiquement observés dans une seule des deux catégories. Les séquences les plus riches en T sont plus spécifiquement associées aux motifs-TC tandis que celles les plus riches en C sont plus associées aux microsatellites.

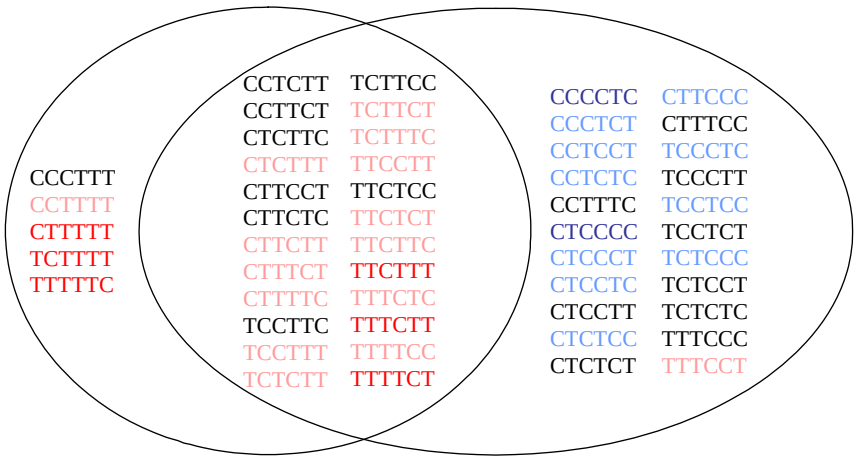


Figure 7-7 : Les séquences des motifs-TC et des microsatellites riches en bases C et T.

La couleur associée à chaque motif dépend de sa richesse en bases C et T. Les motifs en noir sont constitués d'autant de C que de T. Les motifs en bleu sont plus riches en C, ceux en rouge sont plus riches en T. Les dégradés de couleur illustrent une richesse en bases C ou T qui peut être de 4 ou de 5 bases, et donc d'une couleur pâle ou vive. Les motifs dans le cercle de gauche sont les 29 motifs-TC. Les motifs dans le cercle de droite sont les 46 motifs représentant les microsatellites. Les motifs partagés sont entourés de deux cercles.

Deux catégories de séquences correspondant aux motifs-TC existent. Cinq sont seulement des motifs-TC (Figure 7-7 à gauche) et les 24 autres sont des séquences de motifs-TC et de microsatellites riches en bases C et T (Figure 7-7 au centre). La distribution de ces deux catégories de séquences montre des contraintes topologiques à l'emplacement de la boîte TATA pour chacune (Figure 7-8), 24 d'entre elles ayant de plus une sur-représentation sur 100 à 200 bases dans les UTR 5' (Figure 7-8 B).

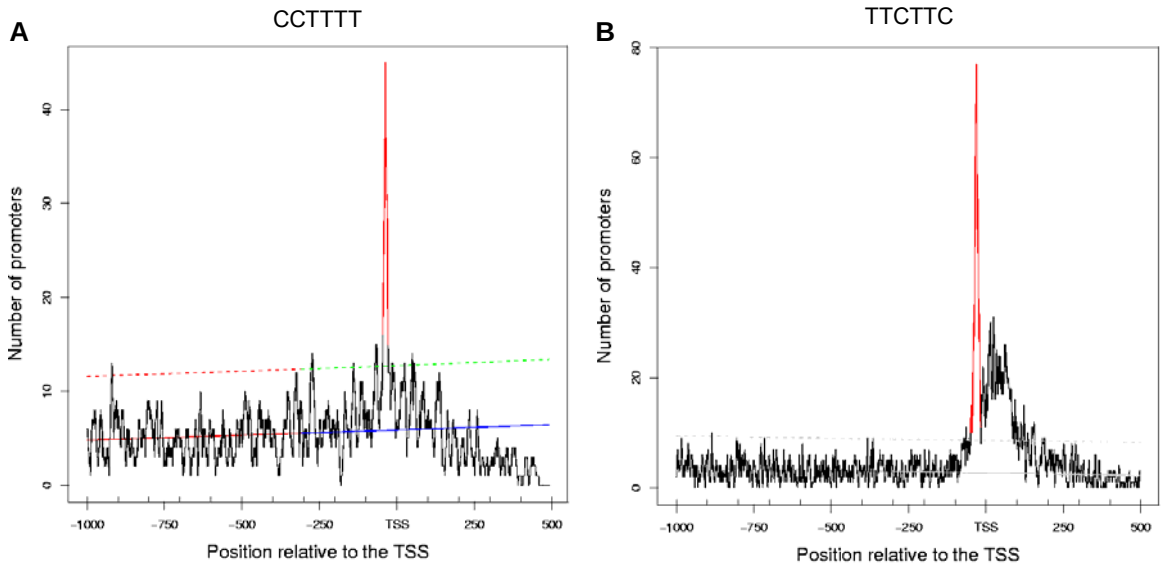


Figure 7-8 : Exemples de distributions de motifs-TC.

Les deux distributions de motifs sont réalisées au sein des 1745 promoteurs ne contenant pas les dinucléotides TA, AT ni AA à l'emplacement de la boîte TATA.

ii) Les motifs-TC ne s'étendent pas autant que les microsatellites

L'approche présentée page 86 pour identifier les leaders au sein des PLM a été adaptée ici pour identifier les motifs pouvant ou non être étendus. Si M est un motif de longueur n, un motif étendu de M est un motif incluant M et qui est long de n+1 bases. Chaque motif possède donc 8 motifs étendus en ajoutant une des 4 bases A, C, G ou T en amont ou en aval de M. Un motif étendu est dit positif lorsque son SMS est supérieur au SMS du motif M, et que les contraintes topologiques des deux motifs sont les mêmes.

Les extensions des 29 motifs-TC et les microsatellites TC et TTC répétés parfaitement ont été analysées. Tandis que les motifs-TC sont peu ou pas étendus (Table 7-9), les microsatellites peuvent être étendus jusqu'à la longueur de 18 bases (Table 7-10).

Motif-TC (SMS)	Motif étendu positif (SMS)	Longueur du motif étendu
CTCTTT (8.86)		6
CCCTTT (3.87)		6
CCTCTT (4.92)		6
CCTTCT (5.66)		6
CCTTTT (4.5)		6
CTTTCT (8.79)		6
CTTTTC (4.59)		6
TCTTTC (7.7)		6
TTCCTT (4.44)		6
TCCTTT (4.42)		6
TTTTTC (6.13)		6
TTTTCC (3.52)		6
TTTCTC (10.28)		6
TCTTCT (12.15)		6
TCTCTT (11.27)		6
TTTCTT (6.29)		6
TTTTCT (6.22)		6
TCCTTC (5.46)		6
TCTTTT (4.35)	TCTCTTTT (5.89) ; 8.86	8
CTTTTT (4.68)	CTTTTTCT (5.79) ; 6.13	8
CTTCCT (8)	TTCTTCCT (9.38) ; 12.19	8
TCTTCC (8.28)	TTCTTCCT (9.38) ; 12.19	8
TTCTTT (4.83)	TTCTTTCT (5.5) ; 8.79	8
CTTCTC (10.31)	TCTTCTC (12. 06) ; 12.15	7
CTCTTC (11.26)	CTCTTCT (13. 01)	8
TTCTCT (9.24)	TTCTCTC (10.84)	7
TTCTCC (6.87)	TTCTCCT (7.66)	7
TTCTTC (12.19)	TCTTCTTCT (18.94)	9
CTTCTT (12.49)	TCTTCTTCT (18.94)	9

Table 7-9 : Extension des motifs-TC.

Si un motif-TC peut être étendu, le motif étendu résultant est renseigné dans la deuxième colonne. Les bases en noir sont les bases permettant l'extension, les bases en rouge correspondant au motif-TC. Si le motif étendu conduit à la génération d'un autre motif-TC, ce dernier et son SMS sont soulignés.

Microsatellites riches en C et T	Motif étendu positif (SMS)	Longueur du motif étendu
TC répété	TC(7)T (46.44)	15
CT répété	CT(7)C (43.07)	15
TTC répété	TTC(6) (31.75)	18
CTT répété	CTT(4)C (32.65)	13
TCT répété	TCT(5) (31.81)	15

Table 7-10 : Extension des microsatellites riches en bases C et T.

Ainsi, les caractéristiques des motifs-TC et des microsatellites riches en bases C et T permettent de les distinguer, comme résumé dans la Table 7-8. Ce sont deux éléments différents.

b) Y-patch, des microsatellites riches en bases C et T ?

Les motifs-TC ont également des similarités de séquences avec le Y-patch qui a été présenté comme étant un nouvel élément régulateur du promoteur central pendant cette thèse (Yamamoto *et al.*, 2007). Le Y-patch est observé abondamment dans les promoteurs d'*A. thaliana* comme d'*O. sativa* mais pas dans ceux des mammifères, sur une large région chevauchant le TSS (Yamamoto *et al.*, 2007). Son motif consensus identifié chez *O. sativa* est CYTCYYCCYC, qui est un motif plus riche en bases C que T (Civan & Svec, 2009). Ces caractéristiques permettent de distinguer le Y-Patch des motifs-TC, mais pas des microsatellites riches en bases C et T (Table 7-8). En complément, les séquences des Y-patch sont plus appariées avec les séquences spécifiques de microsatellites (19/22) qu'avec les séquences spécifiques des motifs-TC (1/5) comme illustré Figure 7-9.

Ainsi, le Y-patch proposé par Yamamoto *et al.* (2007b) est différent des motifs-TC. Notre hypothèse est que le Y-patch pourrait être une représentation de la présence des microsatellites riches en bases C et T en amont du TSS, comme proposé par Molina *et al.* (2006) et non un nouvel élément régulateur du promoteur central.

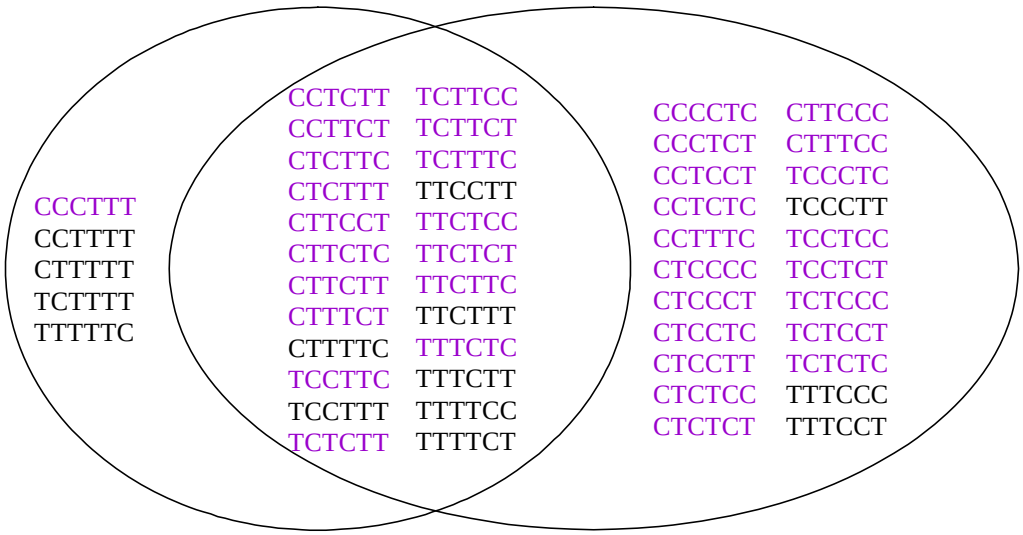


Figure 7-9 : Les séquences Y-patch par rapport aux motifs-TC et aux microsatellites riches en C et T.

Le diagramme de Venn de la Figure 7-7 est reproduit dans cette figure, avec les séquences des motifs-TC à gauche et les séquences des microsatellites à droite. Les motifs en violet correspondent à des séquences de motifs Y-patch (Yamamoto *et al.*, 2007).

c) Pourquoi les motifs-TC n'ont pas été identifiés dans l'analyse globale ?

Les motifs-TC n'ont pas été identifiés lors de l'étude globale en raison du grand nombre de microsatellites riches en bases C et T qui sont observés dans presque l'ensemble des UTR 5'. Les motifs-TC sont alors moins observés. En effet, les motifs-TC et les

microsatellites partagent des séquences communes (Figure 7-7). Lors de l'analyse globale, le pic formé par les microsatellites dissimule le pic formé par les motifs-TC (Table 7-11). Leur présence n'est donc pas clairement visible au sein de la distribution du motif TTCTTC dans le jeu de 14927 promoteurs. Néanmoins, étudier des promoteurs ne contenant pas de richesse en TA, AT ni AA à l'emplacement de la boîte TATA a pour conséquence d'inverser les tendances (Table 7-11). Les motifs-TC ne sont donc visibles que s'ils sont accumulés dans un jeu de promoteurs.

	32 bases en amont du TSS	29 bases en aval du TSS
14927 promoteurs	31	99
1745 promoteurs sans TA, AT ni AA en [-39, -26]	26	9

Table 7-11 : Nombre d'occurrences de TTCTTC.

Le motif TTCTTC est un motif qui montre la présence en microsatellite et en motifs-TC. Il est préférentiellement observé 29 bases en aval du TSS en tant que microsatellite et 32 bases en amont du TSS en tant que motif-TC. Le tableau présente le bilan du nombre de promoteurs contenant ce motif à ces deux emplacements dans les 14927 promoteurs et dans les 1745 promoteurs sans TA, AT ni AA en [-39, -26].

En conclusion, les motifs-TC sont une nouvelle classe d'éléments régulateurs spécifiquement observée chez les plantes, conservée entre *A. thaliana* et *O. sativa* et contribuant à la régulation de l'expression des gènes. Ils partagent les mêmes contraintes topologiques que la boîte TATA : ils sont observés dans la région [-39, -26]. Environ 18% des gènes d'*A. thaliana* contiennent un motif-TC, tandis que 17% contiennent une boîte TATA et 28% un variant fonctionnel. Ainsi, 46% des gènes de cette plante modèle sont caractérisés par la présence d'un élément régulateur ayant des contraintes topologiques strictes environ 30 bases en amont du TSS.

7.4 Caractéristiques de l'Inr-YR

7.4.1 Motif consensus incomplet ?

Les PLM de la région IV représentent la présence de l'élément initiateur chevauchant le TSS de motif consensus YR chez les plantes (Yamamoto *et al.*, 2007). La position de l'Inr est très conservée, une base en amont du TSS. A ce jour aucune validation expérimentale n'a confirmé ces prédictions. Au cours de ce travail de thèse, l'étude des bases autour de l'Inr n'a pas révélé de biais permettant d'étendre le motif. La plus grande richesse en A et T observée est en accord avec la richesse en bases dans la région du TSS. Néanmoins, ce résultat ne permet pas de conclure que le motif YR est le motif consensus complet de l'Inr. En effet, il peut être envisagé différentes hypothèses pour expliquer cela. Il est possible que :

- le motif consensus soit entouré de bases quelconques (N) et que à quelques bases en amont ou en aval du dinucléotide YR, une à plusieurs bases d'intérêt soient impliquées dans la reconnaissance de la séquence Inr ;
- le motif consensus soit très dégénéré et qu'une approche considérant les bases une à une ne permette pas d'identifier de biais significatif ;

- différents motifs consensus existent et qu'une approche globale ne soit pas efficace pour établir un motif consensus. Une approche par sous-groupe possédant un même Inr serait plus adaptée.

Les TFBS sont en effet des éléments souvent très dégénérés et qui contiennent des bases neutres N mais aussi des bases laissant la possibilité d'observer en une position 2 à 3 bases différentes. Il est donc concevable que l'Inr-YR ne soit que le centre d'un Inr plus grand.

7.4.2 Prolongation de la richesse en dinucléotide initiateur dans les UTR 5'

Quatre dinucléotides constituent le motif Inr chez *A. thaliana* : CA, TA, TG et CG, ordonnés du plus au moins fréquent. La distribution du CA met en évidence un épaulement en aval de la position préférentielle qui laisse supposer que la présence s'étend en aval du TSS (Figure 7-10). De plus, dans la Table 6-8 page 101, la séquence de différents PLM montre que les CA sont regroupés très proches les uns des autres. Les occurrences de chacun des 4 dinucléotides CA, TA, TG et CG ont été analysées dans des fenêtres de 5 bases de largeur en aval du TSS dans le jeu de 14927 promoteurs (référence) et dans les jeux de promoteurs contenant un des 4 Inr. Dans les 5 bases en aval du TSS, le dinucléotide CA est sur-représenté exclusivement dans le jeu de promoteurs contenant l'Inr-CA (Table 7-12 A). Le même biais est observé en aval de chacun des Inr, reflétant une extension de la présence de chaque Inr. Ces biais ne sont pas observés au-delà de 25 bases en aval du TSS (Table 7-12 B à F). Ainsi, chaque Inr-YR voit sa présence se prolonger sur quelques bases en aval du TSS. Aucun biais de la sorte n'a été mis en évidence en amont de la position -1.

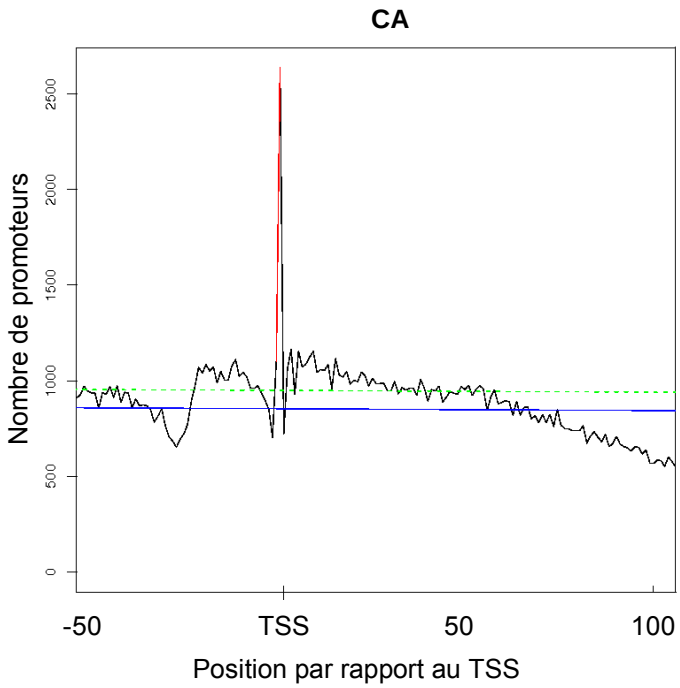


Figure 7-10 : Distribution du dinucléotide CA [-50, 100] chez *A. thaliana*.

	Intervalle	Jeu de promoteurs	CA		TA		TG		CG	
A.		14927 promoteurs	30,2 %		24,3 %		18,3 %		12,4 %	
		Inr-CA	33,4 %	4e-5	21,2 %	NS	12,3 %	NS	8,9 %	NS
	[TSS, 4]	Inr-TA	26,4 %	NS	27,5 %	9e-5	14,6 %	NS	9,5 %	NS
		Inr-TG	21,2 %	NS	22,3 %	NS	22,0 %	1e-4	11,1 %	NS
		Inr-CG	21,7 %	NS	19,8 %	NS	18,1 %	NS	19,7 %	8e-9
B.		14927 promoteurs	32,4 %		25,1 %		18,6 %		13,1 %	
		Inr-CA	39,6 %	<1e-16	24,3 %	NS	15,9 %	NS	12,3 %	NS
	[5, 9]	Inr-TA	32,0 %	NS	27,7 %	NS	18,8 %	NS	12,5 %	NS
		Inr-TG	30,2 %	NS	25,1 %	NS	22,5 %	4e-5	13,8 %	NS
		Inr-CG	28,2 %	NS	25,5 %	NS	17,0 %	NS	19,0 %	3e-6
C.		14927 promoteurs	31,4 %		24,3 %		18,1 %		13,0 %	
		Inr-CA	34,7 %	2e-5	24,3 %	NS	15,2 %	NS	11,4 %	NS
	[10, 14]	Inr-TA	31,9 %	NS	25,5 %	NS	17,6 %	NS	12,3 %	NS
		Inr-TG	28,7 %	NS	25,5 %	NS	22,2 %	1e-5	12,8 %	NS
		Inr-CG	27,7 %	NS	20,4 %	NS	23,3 %	2e-4	19,0 %	2e-6
D.		14927 promoteurs	30,5 %		25,0 %		17,9 %		13,4 %	
		Inr-CA	33,1 %	5e-4	24,6 %	NS	16,1 %	NS	11,3 %	NS
	[15, 19]	Inr-TA	31,4 %	NS	26,6 %	NS	17,0 %	NS	11,8 %	NS
		Inr-TG	28,2 %	NS	25,7 %	NS	21,4 %	2e-4	14,1 %	NS
		Inr-CG	27,8 %	NS	21,1 %	NS	17,8 %	NS	18,8 %	1e-5
E.		14927 promoteurs	30,4 %		24,5 %		17,9 %		13,8 %	
		Inr-CA	33,8 %	2e-5	23,5 %	NS	15,1 %	NS	11,3 %	NS
	[20, 24]	Inr-TA	30,2 %	NS	25,0 %	NS	18,0 %	NS	12,3 %	NS
		Inr-TG	29,0 %	NS	25,8 %	NS	17,7 %	NS	16,0 %	NS
		Inr-CG	27,7 %	NS	20,4 %	NS	18,8 %	NS	19,8 %	2e-6
F.		14927 promoteurs	29,2 %		24,0 %		18,0 %		13,2 %	
		Inr-CA	30,7 %	NS	24,2 %	NS	15,7 %	NS	11,7 %	NS
	[25, 29]	Inr-TA	30,2 %	NS	25,5 %	NS	18,5 %	NS	12,0 %	NS
		Inr-TG	27,7 %	NS	25,4 %	NS	18,9 %	NS	14,7 %	NS
		Inr-CG	29,0 %	NS	21,0 %	NS	19,0 %	NS	15,3 %	NS

Table 7-12 : Etude de l'extension des dinucléotides CA, TA, TG et CG en aval des Inr-YR.

Pourcentage de gènes contenant le dinucléotide CA, TA, TG et CG dans des fenêtres de 5 bases de large au sein des régions [TSS, 4] (**A**) à [25, 29] (**F**) dans l'ensemble des 14927 promoteurs (1ère ligne, référence) et dans les jeux de promoteurs élément -32. Les sur-représentations de dinucléotides dans un sous-groupe de gènes par rapport aux autres gènes (rouge) ont été mis en évidence par des tests exacts de Fisher unilatéraux avec correction de Bonferroni. NS : pas de différence significative (p-value > 1e-2).

Lors de l'étude des promoteurs d'*A. thaliana*, Yamamoto *et al.* (2009) ont relevé la présence du dinucléotide CA dans l'UTR 5'. Cette présence pourrait être le reflet de l'extension dans la région des UTR 5' proche du TSS de l'Inr-CA qui a été mise en évidence lors de notre étude. Chez *H. sapiens*, une extension de l'Inr a été mise en évidence et

pourrait être associée à une organisation du promoteur central conduisant à la conformation «idéale» avec chacun des TSS alternatifs (Ponjavic *et al.*, 2006). Cette hypothèse peut être appliquée chez *A. thaliana*. Cette prolongation de la présence des dinucléotides Inr est également susceptible de pointer sur l'existence de TSS alternatifs proches au sein de ces promoteurs. Chez *H. sapiens* et très récemment chez *A. thaliana*, il a été démontré que les gènes contenant une boîte TATA sont plus souvent associés à un regroupement de TSS présents dans une région peu étendue (Carninci, 2006; Yamamoto *et al.*, 2009). De plus, nous avons montré que chez *A. thaliana*, les gènes contenant une boîte TATA contiennent plus souvent que les autres gènes un Inr-CA ou -TA (Cf. paragraphes suivants et Table 7-13). L'hypothèse selon laquelle au sein des gènes contenant un Inr il existerait des TSS alternatifs proches peut donc être posée. Des analyses supplémentaires pourraient permettre de tester cette hypothèse en exploitant les données récemment publiées mettant à la disposition des jeux de TSS alternatifs chez *A. thaliana* (Yamamoto *et al.*, 2009).

Enfin, l'étude présentée ici précise que l'extension d'un Inr dépend de sa séquence. Seuls les dinucléotides CA sont étendus en aval des Inr-CA, seuls les dinucléotides TG sont étendus en aval des Inr-TG etc. Ce résultat permet de supposer que des TSS alternatifs proches pourraient être gouvernés par une même catégorie d'Inr.

7.5 Boîte TATA et l'Inr-CA : deux éléments en module

Parmi les 14927 promoteurs, 46% contiennent une boîte TATA, un variant ou un motif-TC dans la région de la boîte TATA. Par la suite, ces trois catégories d'éléments sont appelés les «éléments -32», appellation relative à leur position préférentielle. De même, 46% (6933) contiennent un Inr : CA, TA, TG ou CG. Les liens architecturaux entre ces deux régions ont été considérés afin de définir les éléments qui pourraient interagir fonctionnellement.

7.5.1 Des éléments en module

Les TFBS chez les eucaryotes agissent en modules (Blanchette *et al.*, 2006), comme dans le promoteur central par exemple avec le DPE qui interagit fonctionnellement avec l'Inr pour initier la transcription. Chez *A. thaliana*, la boîte TATA et l'Inr-CA sont plus souvent présents au sein d'un même promoteur (Table 7-13). Près de 28% des promoteurs contenant une boîte TATA contiennent un Inr-CA, ce pourcentage étant moins important dans les promoteurs contenant un variant fonctionnel (21.2%) et dans les promoteurs contenant un motif-TC (17.5%). Pour ces deux dernières catégories d'éléments -32, des Ā□□□□□□□□□□ÌĤ□□□□□□□□□Ŧ□□□□Ĥ□par motif n'ont pas mis en évidence de biais. Enfin, l'Inr-TA et la boîte TATA sont également plus souvent observés conjointement dans les gènes.

	Inr-CA Pourcentage (P-value)	Inr-TA Pourcentage (P-value)	Inr-TG Pourcentage (P-value)	Inr-CG Pourcentage (P-value)
14927 promoteurs	17.8%	14.6%	9.5%	4.7%
Boîte TATA (2606 promoteurs - 17%)	27.6% <1e-16	17.5% 4e-6	8.6% NS	3.9% NS
Variant (4151 promoteurs - 28%)	21.2% NS	16.1% NS	9.2% NS	4.6% NS
Motif-TC (2645 promoteurs - 18%)	17.5% NS	13.9% NS	9.6% NS	4.5% NS

Table 7-13 : Etude de la présence simultanée des éléments régulateurs dans le promoteur central.

Pourcentage en Inr au sein des 14927 promoteurs (référence) et au sein des promoteurs contenant une boîte TATA, un variant ou un motif-TC. Les sur-représentations de la présence en Inr dans un des groupes sont mises en évidence par des tests exacts de Fisher unilatéraux avec correction de Bonferroni. NS : pas de différence significative (p-value > 1e-2).

Cette association boîte TATA / Inr-CA a été observée par d'autres auteurs récemment (Yamamoto *et al.*, 2009) qui ont proposé que ces éléments interagiraient fonctionnellement pour initier la transcription des gènes. Nos résultats confirment ces hypothèses, et proposent en complément une association boîte TATA / Inr-TA.

En complément de ces associations entre la boîte TATA et les Inr-CA et -TA, la recherche d'associations entre des éléments -32 et des TFBS validés expérimentalement a été réalisée. Des regroupements au sein de mêmes promoteurs des motifs-TC et de TFBS ayant des séquences riches en bases C et G ont été mis en évidence. Au regard de ces résultats, une étude de la composition en C+G dans les promoteurs a été réalisée pour les trois groupes de gènes contenant les éléments -32. Les gènes contenant les motifs-TC ont des promoteurs plus riches en C+G que les autres gènes (Figure 7-11).

Chez *H. sapiens*, les gènes ne contenant pas de boîte TATA contiennent plus que d'autres gènes des îlots CpG (Carninci *et al.*, 2006) et sont globalement plus riches en C+G (Yang *et al.*, 2007). Les mêmes biais sont donc conservés des mammifères aux plantes. Ainsi, tandis que les îlots CpG n'ont pas été identifiés chez *A. thaliana* (Rombauts *et al.*, 2003 ; Yamamoto *et al.*, 2007a), des TFBS riches en C et G sont observés dans le promoteur proximal des gènes.

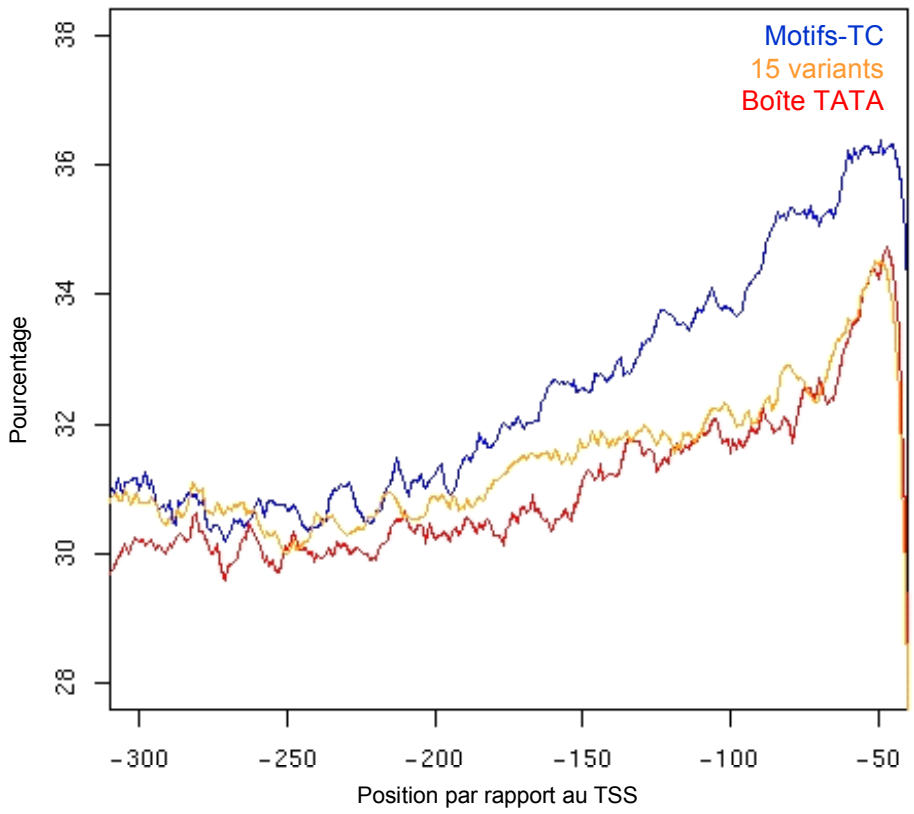


Figure 7-11 : Composition en C+G dans les promoteurs.

Représentation du taux de bases C+G dans les promoteurs des gènes contenant un motif-TC, une boîte TATA ou un variant fonctionnel. Fenêtre glissante de 10 bases utilisée pour représenter ces distributions.

7.5.2 Distance préférentielle

Chez les mammifères, deux éléments du promoteur central le DPE et l'Inr (Kutach & Kadonaga, 2000) ont une distance préférentielle leur permettant d'être fonctionnels. Des distances préférentielles entre la boîte TATA et l'Inr ont été mises en évidence chez *H. sapiens* (Ponjavic *et al.*, 2006). De telles distances préférentielles n'ont pas été mises en évidence chez *A. thaliana* entre l'Inr-CA et la boîte TATA. Les éléments -32 et les Inr sont tous préférentiellement positionnés dans une région très stricte par rapport au TSS. C'est pourquoi l'étude de la distance entre ces éléments a été analysée afin de tester l'hypothèse qu'ils pourraient interagir fonctionnellement en complexes ou CRM (Yuh *et al.*, 2001; Blanchette *et al.*, 2006).

La distribution de la distance entre les 4 dinucléotides CA, TA, TG et CG en aval des éléments -32 a été construite. Par exemple, chaque occurrence de CA a été recherchée en aval de chaque occurrence de boîte TATA dans sa fenêtre fonctionnelle.

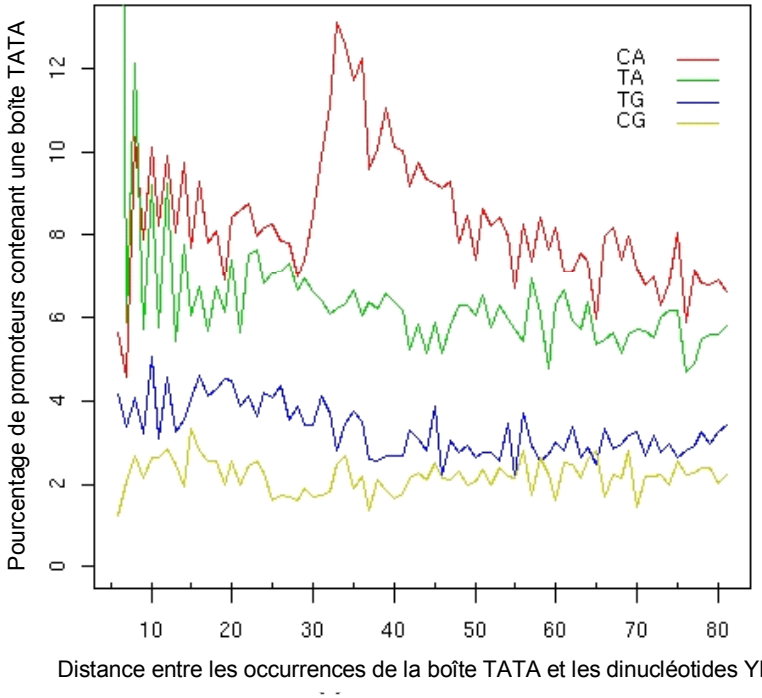


Figure 7-12 : Distance préférentielle entre les Inr et la boîte TATA ?

Représentation de la distance entre chaque dinucléotide CA (rouge), TA (vert), TG (bleu) et CG (jaune) observé en aval d'une occurrence de boîte TATA dans la région [-39, -26]

La Figure 7-12 met en évidence une périodicité dans les premières bases après une occurrence de TATAWA pour les motifs CA et TA. Ces dinucléotides sont observés à des distances impaires de la boîte TATA. La périodicité du A au sein de la boîte TATA est donc étendue en aval du motif sur quelques bases. Ce biais pourrait être du à l'extension des boîtes TATA décrite page 97.

Parmi les quatre dinucléotides, seul CA est observé à une distance préférentielle des occurrences de TATAWA dans [-39, -26]. La Figure 7-12 montre une plus forte présence des CA 30 à 33 bases en aval des boîtes TATA. Ces écarts sont en accord avec les positions préférentielles des deux motifs : 32 bases en amont du TSS pour la boîte TATA et -1 pour l'Inr-CA. La quasi-totalité des promoteurs contenant une occurrence de CA 30 à 33 bases en aval d'une boîte TATA, contiennent un Inr-CA. Les autres dinucléotides ne présentent pas de tels biais, malgré leurs contraintes topologiques similaires à celles de CA.

Les mêmes analyses réalisées pour chacun des variants fonctionnels et pour chacun des motifs-TC n'ont pas relevé de biais. L'ensemble des motifs regroupés non plus.

En conclusion, ces résultats permettent pour la première fois de préciser qu'en plus d'être présents dans un même promoteur, la boîte TATA et l'Inr sont observés à une distance préférentielle l'un de l'autre : 30 à 33 bases séparent préférentiellement le CA du TATAWA. Ces motifs constituent une dyade TATAWA-d{30-33}-CA. La distance entre TFBS est une caractéristique conservée entre *M. musculus* et *H. sapiens* (Lu *et al.*, 2008). D'après ces résultats, la distance préférentielle entre la boîte TATA et l'Inr est une caractéristique conservée des mammifères (Ponjavic *et al.*, 2006) aux plantes.

8 Approche PLM pour l'identification d'éléments régulateurs spécifiques

L'approche PLM a été appliquée à des jeux d'intérêts expérimentaux *via* une collaboration avec l'équipe MAP-kinase de l'URGV.

Les MAP-kinases ou MAPKs sont des protéines qui catalysent la phosphorylation des protéines activées par différents facteurs externes, des hormones, ou des stress abiotiques ou biotiques. Les MAPKs sont actives en cascade : une MAPK est activée par une MAPKK qui elle-même est activée par une MAPKKK qui peut transférer des signaux des récepteurs divers. Chez les plantes, des voies de signalisations intracellulaires sensibles aux stress biotiques et abiotiques permettent d'activer ces cascades de MAPK et ainsi de transmettre les signaux reçus de la membrane vers les effecteurs intracellulaires (Colcombet & Hirt, 2008). L'équipe de Heribert Hirt (URGV) développe des projets de recherche centrés autour de la signalisation des réponses aux stress chez les plantes, dont un des objectifs est d'identifier les rôles des MAP-kinases intervenant dans des réseaux de régulation. Afin d'étudier une voie de signalisation, des expériences ont été réalisées avec les puces CATMA avec des mutants de la MAPKKK MEKK1, des MAPKKs MKK1 et MKK2, et la MAPK MPK4. Toutes ces MAPK constituent le module de cette voie de signalisation. Dix expériences sur 40 puces CATMA ont été menées. Ainsi, deux listes de gènes ont été constituées : une regroupant 185 gènes sur-exprimés dans la voie de signalisation des MAK-kinases, une autre regroupant 56 gènes sous-exprimés dans la même voie. Afin de mieux caractériser ces gènes et notamment d'identifier les éléments susceptibles d'être impliqués dans leur régulation, une étude de ces deux jeux de gènes a été réalisée. Dans la suite de la présentation de cette étude, ces jeux seront appelés les jeux «MAPK+» et «MAPK-» et représenteront respectivement les gènes sur-exprimés et sous-exprimés dans les expériences décrites. Parmi les 185 gènes MAPK+ et les 56 gènes MAPK-, respectivement 180 et 50 appartiennent au jeu des 14927 gènes ayant un TSS caractérisé et ont été étudiés par l'approche PLM. Les analyses réalisées sur ces deux jeux de promoteurs sont schématisées Figure 8-1.

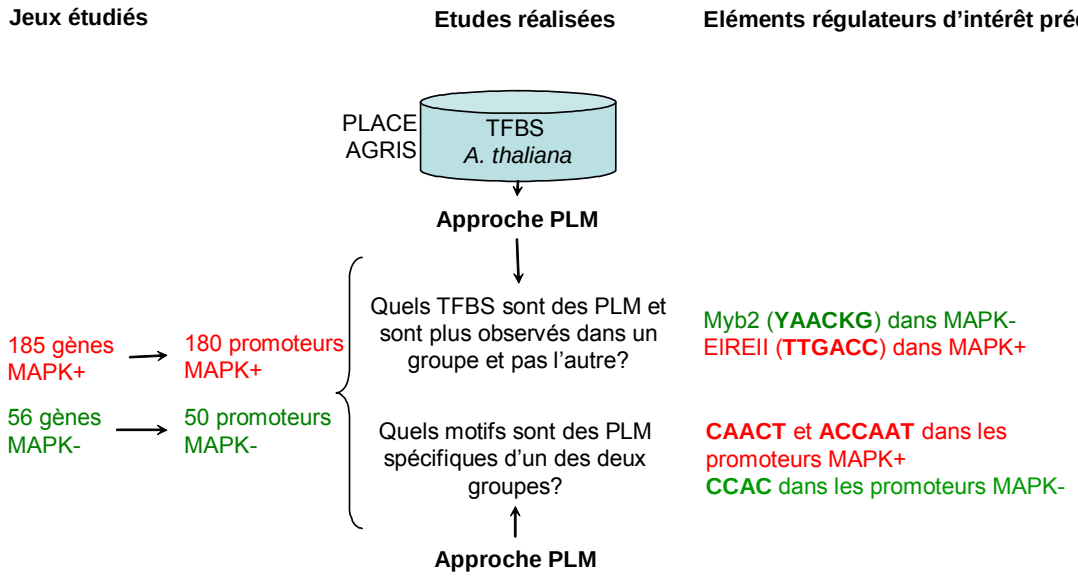


Figure 8-1 : Résumé de l'étude des promoteurs MAPK+ et MAPK-.

8.1 Quels TFBS connus sont associés aux promoteurs ?

La présence de chacun des 66 TFBS de PLACE et de AGRIS (Higo *et al.*, 1999; Davuluri *et al.*, 2003) ayant des contraintes topologiques a été étudiée au sein des promoteurs MAPK+, MAPK- et du jeu complet. Des biais opposés entre les gènes MAPK+ et MAPK- ont été mis en évidence (Table 8-1). L'élément régulateur Myb2, de séquence consensus YAACKG, a tendance à être présent dans les promoteurs des gènes MAPK- tandis que l'élément EIRE, de motif consensus TTGACC, est plus observé dans le jeu MAPK+. Ces deux éléments sont impliqués dans des réponses à des stress. Myb2 intervient dans des réponses à une déshydratation (Biedenkapp *et al.*, 1988). EIREII intervient lors d'un stress causé par la présence de pathogènes. Il est reconnu par des facteurs de transcription possédant le domaine WRKY, domaine dont le nom provient de la conservation en acides aminés au sein des protéines le contenant (Rushton *et al.*, 1996).

TFBS			14927 promoteurs		Promoteurs MAPK+		Promoteurs MAPK-	
Séquence	Nom	Référence	Fenêtre fonctionnelle	Pourcentage	Pourcentage (nombre)	p-value	Pourcentage (nombre)	p-value
YAACKG	Myb2	(Biedenkapp *et al.*, 1988)	[-155, -48]	6.8%	4.3% (6)	2e-2	16% (8)	4e-3
TTGACC	EIRE	(Rushton *et al.*, 1996)	[-234, -18]	6%	11% (15)	1e-2	Pas d'occurrence	

Table 8-1 : Présence des TFBS PLM dans les 14927 promoteurs et dans les promoteurs MAPK+ et MAPK-.

Les sur-représentations (rouge) et les sous-représentations (vert) ont été mises en évidence par des tests exacts de Fisher bilatéraux. NS : pas de différence significative (p-value > 1e-2). EIRE pour «Elicitor Responsive Element».

8.2 Recherche de nouveaux PLM dans les groupes MAPK+ et MAPK-

En complément, l'approche PLM a été exploitée pour la prédiction d'éléments régulateurs de longueur 2 à 6 dans chacun des jeux de gènes de cette étude.

Lors de l'étude du jeu MAPK+, 2 PLM ont été mis en évidence : CAACT et ACCAAT et lors de l'étude du jeu MAPK-, le PLM CCAC a été identifié (Table 8-2). Ces 3 PLM ont été comparés aux TFBS indexés dans les bases de données PLACE et AGRIS. La recherche d'appariement a été réalisée en considérant les motifs consensus et leur séquence complément inverse (Table 8-3). Les deux PLM identifiés dans le jeu de promoteurs MAPK+, CAACT et ACCAAT, sont appariés à des TFBS qui ne sont pas des PLM lors de l'étude de l'ensemble des 14927 promoteurs. Ce sont donc deux nouveaux éléments régulateurs potentiels. Le PLM identifié dans le jeu de promoteurs MAPK- est apparié avec un TFBS qui est un PLM : SORLIP1 de motif consensus GCCAC (Hudson & Quail, 2003). Le SMS de CCAC, inférieur à celui de GCCAC, laisse supposer que les contraintes positionnelles de CCAC sont dues à son inclusion dans la séquence de SORLIP1, et non à des contraintes qui lui sont propres. Le PLM CCAC n'est donc pas un nouvel élément régulateur mais reste un PLM spécifique des MAPK-.

	Fenêtre glissante	SMS	Position préférentielle	Fenêtre fonctionnelle	Nombre de promoteurs contenant le PLM	Pourcentage	Largeur de la fenêtre fonctionnelle
CAACT	78	3.12	59	[-159, 153]	56	40.3%	312
ACCAAT	81	3.10	-158	[-253, -55]	21	15.11%	199
CCAC	75	3.15	51	[-174, 130]	35	70%	304

Table 8-2 : Caractéristiques des PLM identifiés dans le jeu MAPK+ et MAPK-.

PLM MAPK+ ou MAPK-			TFBS de PLACE et AGRIS			
PLM	MAPK+	MAPK-	Nom	Séquence consensus (ou sa séquence complément inverse ci)	Référence	Est un PLM?
CAACT	PLM	Non PLM	CAREs	CAACTC	(Sutoh & Yamauchi, 2003)	Non
ACCAAT	PLM	Non PLM	ERSEII-like	CGTGGACCAAT (ci)	(Yamamoto *et al.*, 2004)	Non
CCAC	Non PLM	PLM	SORLIP1	GCCAC	(Hudson & Quail, 2003)	Oui

Table 8-3 : Appariements des PLM MAPK+ et MAPK- avec les TFBS connus.

Dans la colonne des séquences consensus ou de leur séquence complément inverse, les bases soulignées sont appariées avec la séquence du PLM MAPK+ ou MAPK-. CARE pour «CAACTC Regulatory Elements». ERSE pour «Endoplasmic Reticulum Stress response Element». SORLIP pour «Sequences Over-Represented in Light-Induced Promoters».

La comparaison de la présence des 2 PLM CAACT et ACCACT dans le jeu de 14927 promoteurs et dans le jeu MAPK+ montre leur plus forte association avec le sous-groupe de promoteurs MAPK+ (Table 8-4).

PLM		14927 gènes	Promoteurs MAPK+	
Séquence	Fenêtre fonctionnelle	Pourcentage	Pourcentage	p-value
CAACT	[-159, TSS]	15%	21%	2e-5
ACCAAT	[-253, -55]	10%	15%	1e-2

Table 8-4 : CAACT et ACCAAT dans les promoteurs MAPK+ et dans l'ensemble des promoteurs.

Deux tests unilatéraux exacts de Fisher ont permis de mettre en évidence une plus forte présence (rouge) ou une plus faible présence (vert) d'un motif donné.

En conclusion de ces analyses, l'utilisation de la cartographie des promoteurs qui a pu être réalisée *via* l'approche PLM et les TFBS indexés dans les bases de données a permis de proposer la présence de deux éléments régulateurs connus : Myb2 est plus particulièrement présent dans les gènes MAPK- et EIREII dans les gènes MAPK+. L'opposition des biais en présence de ces TFBS entre le jeu MAPK+ et MAPK- pourrait montrer que la présence de l'EIRE induirait une sur-expression des gènes MAPK+ tandis que la présence du Myb core entraînerait une sous-expression des gènes MAPK-. De plus, les deux PLM CAACT et ACCAAT identifiés dans le jeu MAPK+ sont susceptibles d'être de nouveaux éléments régulateurs. Afin de valider ces prédictions, des études expérimentales sont actuellement en cours.

Conclusions générales

9 Discussion générale

9.1 Limites de l'approche PLM

9.1.1 Etude restreinte aux promoteurs dont la position du TSS est connue

Au total, le TSS de 14927 gènes a été défini avec des contraintes strictes, soit 55% des gènes codant une protéine chez *A. thaliana* (Swarbreck *et al.*, 2008). L'ensemble des promoteurs n'a pas pu être étudié faute de transcrits qui auraient permis de définir la position du TSS. L'annotation fonctionnelle des 14927 gènes a été comparée à celle des autres gènes. Les gènes non considérés lors de cette étude sont essentiellement des gènes dont le rôle fonctionnel n'est pas connu (Table 9-1).

	14927 gènes dont les promoteurs sont analysés	11824 autres gènes
Localisation subcellulaire inconnue	21.5%	53.8%
Fonction moléculaire inconnue	30.6%	44.2%
Processus biologique inconnu	36.0%	52.4%

Table 9-1 : Annotation fonctionnelle des gènes étudiés par l'approche PLM et des autres gènes.

L'annotation fonctionnelle des gènes est issue du TAIR (Swarbreck *et al.*, 2008). Seules les annotations significativement sur-représentées dans le groupe des 11824 autres gènes sont indiquées ici (test unilatéral exact de Fisher). Toutes les p-values correspondantes sont inférieures à 1e-16.

Les gènes qui s'expriment dans des conditions spécifiques sont ceux dont l'annotation fonctionnelle et la position du TSS sont les plus difficiles à définir. Les expériences relatives au transcriptome actuellement disponibles ne couvrent pas toujours leurs conditions d'expression spécifiques. Ces gènes sont les plus susceptibles d'être écartés lors de notre recherche de PLM. Néanmoins, les gènes analysés par l'approche PLM sont ceux dont l'annotation fonctionnelle est la mieux définie, ce qui est un critère important pour permettre des recherches de caractéristiques fonctionnelles communes au sein des gènes ayant une même architecture de leurs promoteurs.

9.1.2 Considération exclusive des TSS les plus en amont des séquences codantes

L'existence de TSS alternatifs a récemment été identifiée dans le génome d'*A. thaliana* (Tanaka *et al.*, 2009). Chez *A. thaliana*, parmi les gènes contenant plusieurs TSS, l'analyse d'un grand nombre d'ADNc pleine longueur a mis en évidence la présence de 6,5 TSS en moyenne par gène (Yamamoto *et al.*, 2009). Ce chiffre est probablement sous-estimé, et pourrait être plus justement évalué avec une collection d'ADNc plus complète (Tanaka *et al.*, 2009). En pratique, l'approche PLM peut considérer les TSS alternatifs. Cependant, nous n'avons pris en compte qu'un TSS par gène pour constituer le jeu des 14927 promoteurs. Ceci est une conséquence de l'extension de l'unité de transcription utilisée pour définir la position du TSS le plus en amont de la séquence codante (Figure 5-1 page 60). De plus, l'importance des TSS alternatifs chez *A. thaliana* n'a été identifiée que récemment. Néanmoins, les conséquences de l'exploitation exclusive des TSS les plus en amont étaient

modérées, ces TSS permettant de définir des contraintes topologiques plus fortes. Si plusieurs TSS par gène sont considérés, une occurrence d'un motif pourrait être fonctionnelle par rapport à un TSS mais pas par rapport à d'autres. Ceci influerait sur les distributions de motifs en lissant les sur-représentations locales de motifs qui permettent d'identifier les PLM.

En complément, des travaux récents ont révélé que les TSS les plus en amont des séquences codantes sont les TSS forts qui conduisent le plus souvent à la transcription des gènes (Tanaka *et al.*, 2009). Les TFBS observés à un emplacement précis des promoteurs sont positionnés par rapport à ces TSS et non par rapport aux autres TSS s'ils existent (Tanaka *et al.*, 2009). Ainsi, le jeu de promoteurs utilisé dans notre étude exploite le TSS ayant un rôle fonctionnel dans l'initiation et la régulation de la transcription.

9.1.3 Nécessité de posséder un nombre de promoteurs suffisant

L'approche PLM utilise un apprentissage du modèle de distribution de chaque motif. Plus le jeu de promoteurs étudié est petit, plus la variance est grande et donc plus l'intervalle de confiance est grand. C'est pourquoi moins de PLM sont susceptibles d'être identifiés dans des sous-groupes de promoteurs et certains jeux constitués de trop peu de promoteurs ne pourront pas être considérés. Néanmoins, les différentes études réalisées au cours de cette thèse ont toujours permis d'identifier des motifs caractérisés par des contraintes topologiques. Par exemple, lors de l'étude des jeux de MAP kinases, 50 gènes ont été étudiés, ce qui a conduit à de nouveaux éléments régulateurs potentiels. De plus, les prédictions de PLM lors d'analyses de peu de promoteurs sont plus robustes car les PLM sont caractérisés par des contraintes particulièrement fortes.

9.1.4 Recherche de contraintes topologiques dans la région [-299, ATG] uniquement

La région utilisée pour l'apprentissage du modèle de distribution s'étend sur 700 bases, dans l'intervalle [-1000, -300]. Les motifs que nous identifions comme PLM sont donc nécessairement dans la région [-299, ATG]. Ce choix a été fait suite au constat qu'aucun motif n'était caractérisé par une sur-représentation locale en amont de -300. Néanmoins, des TFBS peuvent être situés dans d'autres régions comme dans les promoteurs distaux, les UTR 3' ou les introns (Vyas *et al.*, 1992 ; Larlin *et al.*, 1993 ; Sieburth & Meyerowitz, 1997 ; Graber *et al.*, 2002 ; Hulzink *et al.*, 2003). Bien que les régions considérées au cours de cette thèse soient uniquement en amont des gènes, notre approche pourrait être appliquée à l'étude des UTR 3'.

De plus, des TFBS peuvent ne pas être soumis à une contrainte de position préférentielle ; en particulier les TFBS dans le promoteur distal peuvent coopérer fonctionnellement avec des TFBS du promoteur proximal grâce à des repliements de l'ADN. L'approche PLM ne permet pas d'identifier ces TFBS. Néanmoins, notons que les TFBS observés dans le promoteur central et proximal, les régions considérées par cette étude, sont souvent soumis à des contraintes positionnelles. Cependant, seuls neuf TFBS validés expérimentalement chez *A. thaliana* ont une contrainte positionnelle en amont de -300 (Higo *et al.*, 1999).

9.1.5 Motifs de 8 bases maximum, non dégénérés et qui ne sont pas en modules

Les PLM identifiés pour réaliser la cartographie des promoteurs d'*A. thaliana* ont une longueur limitée à 8 bases. Cette longueur est inférieure à la taille maximale attendue pour les TFBS : jusqu'à 15 bases de long (Bulyk, 2003). Ce choix s'est imposé à la suite de plusieurs constats. Premièrement, à titre d'exemple, étudier l'ensemble des motifs de 2 à 10 nucléotides de long reviendrait à analyser près de 1.4 millions de motifs ce qui n'était pas envisageable compte tenu des ressources informatiques que nous pouvions mobiliser. Deuxièmement, les motifs plus longs sont statistiquement moins observés et donc potentiellement trop peu présents pour permettre un apprentissage pertinent du modèle de distribution. Néanmoins, pour identifier des PLM plus longs, il est possible de réaliser une étude de l'extension des motifs comme celle réalisée pour les motifs-TC ou les microsatellites.

Même si l'approche PLM est capable de représenter des distributions de motifs dégénérés comme les TFBS le sont souvent, les motifs étudiés pour réaliser la cartographie des promoteurs d'*A. thaliana* sont des motifs non dégénérés. Une étude exhaustive des motifs dégénérés de 6 bases de long reviendrait à étudier plus de 7.5 millions de motifs, ce qui n'est pas envisageable du point de vue des ressources informatiques. De plus, les résultats seraient très difficiles à interpréter : les PLM chevauchant ou inclus dans d'autres PLM seraient alors encore plus nombreux. Ce problème abordé par Blanchette *et al.* (2001) se retrouve pour un grand nombre d'outils de prédiction *in silico*, outils qui génèrent souvent un grand nombre de faux positifs. C'est pourquoi le choix a été fait de rechercher des PLM non dégénérés. Ils peuvent par la suite être regroupés s'ils sont caractérisés par les mêmes contraintes topologiques et des séquences appariées *via* la recherche de leaders. Il sera alors possible de définir si les contraintes topologiques mises en évidence sont plus importantes pour les séquences non dégénérées ou pour la séquence dégénérée résultante. Cela pourra conduire à un consensus dégénéré.

L'approche PLM propose une liste d'éléments régulateurs potentiels, mais sans considérer leur présence simultanée au sein d'un même promoteur. Cependant les modules de TFBS sont nombreux (Chattopadhyay *et al.*, 1998 ; Blazquez & Weigel, 2000) et plus pertinents à mettre en évidence *in silico* que les motifs seuls (Cawley *et al.*, 2004 ; Wasserman & Sandelin, 2004). Néanmoins, grâce à la cartographie des promoteurs proposée par l'approche PLM, il est possible de rechercher les PLM présents simultanément dans des promoteurs. De telles études menées au cours de la thèse ont permis de mettre en évidence l'association de la boîte TATA et de l'Inr-CA. Elles peuvent être généralisées à l'ensemble des PLM obtenus.

Finalement, certains gènes ne sont pas pris en considération dans la cartographie des promoteurs et tous les éléments régulateurs ne sont pas identifiés. Comme proposé par Tompa *et al.* (2005), il sera nécessaire de combiner différentes approches pour contourner ces limites. Les autres limitations discutées, relatives à la taille des motifs étudiés ou à la non considération des motifs dégénérés, peuvent être contournées par des études complémentaires réalisées à la suite de l'approche PLM.

9.2 Exploitation de la cartographie d'*A. thaliana* et des contraintes topologiques des motifs

9.2.1 Détermination du taux d'erreurs dans l'assignation de la position des TSS

Pour un gène donné, un nombre insuffisant de transcrits conduit à définir un UTR 5' tronqué et donc une mauvaise assignation de la position du TSS. Néanmoins, l'analyse de PLM caractérisés par des contraintes strictes nous permet de penser que notre jeu de 14927 promoteurs est correct.

Le taux d'erreurs dans la détermination de la position du TSS peut être estimé en exploitant la distance préférentielle entre plusieurs PLM comme la boîte TATA et l'Inr-CA par exemple. Au total, 288 gènes contiennent dans leurs promoteurs la dyade TATAWA-d{30-33}-CA dans sa région fonctionnelle. Nous avons considéré que la position du TSS des 288 gènes était correcte et que les dyades TATAWA-d{30-33}-CA dont le T est hors de la région [-39, -26] ne sont pas fonctionnelles. Au total, 3.8% des 288 promoteurs contiennent une dyade dans les 100 bases en amont de -39. Ce pourcentage montre le bruit de fond de la présence de dyades non fonctionnelles. En extrapolant ces résultats, parmi les 14927 promoteurs, 570 TATAWA-d{30-33}-CA sont attendues dans les 100 bases en amont de -39. Finalement, 718 sont observées, ce qui permet de supposer que moins de 1% des promoteurs (148) pourraient avoir un TSS dont la position n'est pas fiable.

Dans le jeu de 14927 promoteurs d'*A. thaliana* étudiés, les distributions de 88 motifs présentent un pic parfaitement symétrique par rapport à la position préférentielle. C'est le cas lorsque la position préférentielle est au centre de la fenêtre fonctionnelle. Par exemple la séquence canonique de la boîte TATA ou l'Inr-YR ont un pic symétrique. Ces distributions suggèrent que la position de la majorité des TSS des gènes contenant ces 88 PLM est correctement assignée et / ou qu'il y a peu de TSS alternatifs au sein des gènes qui les contiennent. Récemment, il a été montré chez *H. sapiens* (Carninci, 2006) et *A. thaliana* (Yamamoto *et al.*, 2009) que les gènes possédant une boîte TATA ou un Inr ont plus souvent des TSS alternatifs mais qu'ils sont regroupés dans un intervalle restreint. L'hypothèse d'une bonne définition de la position du TSS parmi les gènes contenant ces PLM est donc la plus envisageable. Une distribution du consensus de la boîte TATA canonique ou de l'Inr-YR et l'analyse de la symétrie du pic sont de bons indicateurs pour juger de la fiabilité du jeu de promoteurs. Ce critère ne doit cependant pas être le seul pris en compte.

9.2.2 Explication de biais compositionnels observés dans les promoteurs de plantes

Le GC-skew est un biais compositionnel en bases mis en évidence chez les plantes qui a été présenté page 63 (Fujimori *et al.*, 2005). Il est susceptible d'être provoqué par la présence des microsatellites riches en bases C et T spécifiquement dans les UTR 5' de plantes (Fujimori *et al.*, 2005). Néanmoins, les microsatellites sont majoritairement observés dans l'UTR 5', ce qui ne permet de justifier ni la présence du biais le plus fort à l'emplacement du TSS, ni les fluctuations mises en évidence entre la position -1 et le TSS décrites au cours de notre étude. Au regard des résultats de cette thèse, l'Inr-YR et plus particulièrement l'Inr-CA ainsi que son extension sur quelques bases en aval du TSS pourraient être impliqués dans la formation du GC-skew. Leurs présences pourraient permettre d'expliquer (i) la valeur maximale et la valeur minimale du GC-skew observées

respectivement en -1 et à l'emplacement du TSS et (ii) le maintien d'un biais élevé sur quelques bases en aval du TSS. La présence des motifs-TC et la richesse en C et T dans le promoteur central (qui peut être due à une extension des microsatellites) pourraient contribuer à l'augmentation progressive du GC-skew en amont du TSS. Jusqu'à aujourd'hui, les microsatellites riches en bases C et T, dans le promoteur central comme dans les UTR 5', n'ont été mis en évidence que chez les plantes, ce qui est le cas également pour les motifs-TC. Ces biais pourraient expliquer pourquoi le GC-skew est un biais compositionnel spécifique aux plantes.

9.3 Distinction entre TFBS et éléments impliqués dans la conformation de l'ADN

Les PLM peuvent être des TFBS ou des éléments impliqués dans la conformation de l'ADN. Nous proposons différentes hypothèses permettant de distinguer ces deux catégories d'éléments. Les résultats exploités pour atteindre cet objectif sont les contraintes topologiques des PLM et l'évolution du nombre d'occurrences des motifs hors de leur région fonctionnelle présentée dans le manuscrit de l'article qui décrit la méthodologie (page 70).

Trois classes de PLM peuvent être définies. La Table 9-2 résume nos hypothèses de classement des PLM suivant leur propriétés. (i) Près de 47% des PLM voient leur nombre d'occurrences diminuer entre le promoteur distal et leur région fonctionnelle comme la boîte GCC de séquence consensus GCCGCC (Figure 6-8 B page 94) ou l'Inr-CA (Figure 6-2 D page 83). Ces motifs pourraient être contre-sélectionnés à l'approche de leur région fonctionnelle. (ii) Plus de 11% des PLM sont caractérisés par un nombre d'occurrences qui augmente à l'approche de la région fonctionnelle. Une partie d'entre eux, après cette augmentation, voient leur nombre d'occurrences diminuer en amont de la région fonctionnelle. C'est le cas des PLM représentant la séquence canonique de la boîte TATA ou ses variants par exemple (Figure 6-2 C page 83). De tels motifs pourraient être contre-sélectionnés quelques bases en amont de leur région fonctionnelle. (iii) Parmi les 11% de PLM caractérisés par une augmentation du nombre d'occurrences des motifs, 15 ont des caractéristiques qui les distinguent des autres PLM. Ils ont une fenêtre fonctionnelle large, de plus de 200 bases, et sont fréquemment constitués de séquences contenant une très faible diversité en bases, c'est-à-dire contenant une à deux bases seulement. C'est le cas des PLM AAAAAG ou AAAAAC ou encore des microsatellites riches en bases C et T présents principalement dans les UTR 5' (Figure 6-2 B page 83). De tels PLM pourraient être des éléments plus susceptibles d'être impliqués dans la conformation de l'ADN. Ainsi, les hypothèses proposées ici et résumées Table 9-2 permettent de classer une partie des PLM soit en tant que sites de fixation potentiels, *i.e.* ceux étant contre-sélectionnés en amont de leur fenêtre fonctionnelle, soit en tant qu'éléments impliqués dans la régulation de la conformation de l'ADN.

	TFBS	Eléments impliqués dans la conformation de l'ADN
Evolution du nombre d'occurrences	**Diminution** entre le promoteur proximal et la fenêtre fonctionnelle ou Augmentation dans le promoteur distal et **diminution** en amont de la fenêtre fonctionnelle	**Augmentation** entre le promoteur proximal et la fenêtre fonctionnelle
Bases constituant les PLM	-	Très faible diversité en bases
Largeur de la fenêtre fonctionnelle	-	Fenêtres larges de plus de 200 bases

Table 9-2 : Caractéristiques proposées pour distinguer les TFBS des éléments impliqués dans la conformation de l'ADN.

9.4 Organisation des TSS alternatifs en fonction de l'architecture des promoteurs

Certains motifs ont une distribution asymétrique par rapport à la position préférentielle du motif étudié et ont, dans la fenêtre fonctionnelle du motif, une augmentation lente puis une diminution rapide. Sur l'ensemble des PLM, 412 (8%) ont une augmentation observée sur une longueur en bases deux fois plus grande que la diminution, comme GGGCC par exemple (Figure 6-2 A page 83). Les séquences de ces PLM partagent une richesse en bases C et G qui représentent près de la moitié de leur composition. Parmi les 412 PLM, plus d'un sur cinq contient la sous-séquence GGCC par exemple. Ces distributions asymétriques peuvent mettre en évidence la présence de TSS alternatifs et / ou une mauvaise assignation de la position du TSS. L'estimation du taux d'erreurs dans le jeu de 14927 promoteurs nous a permis de supposer que ces PLM sont présents dans des gènes ayant des TSS alternatifs.

Une étude des TSS alternatifs chez les mammifères a montré la présence de deux catégories d'organisation parmi les gènes en possédant. Certains gènes ont des TSS observés dans une région restreinte avec un TSS dominant ; d'autres ont des TSS dans une région large sans TSS dominant (Carninci, 2006). Les promoteurs contenant des îlots CpG ont plus fréquemment des TSS alternatifs organisés dans une large région. Aucun îlot CpG n'a été identifié chez *A. thaliana* (Rombauts *et al.*, 2003; Yamamoto *et al.*, 2007). Néanmoins, les PLM riches en C et G dont la distribution est asymétrique pourraient, comme les îlots CpG, être associés à une présence de TSS alternatifs sur une large région. Sans être comparable du point de vue de la largeur de la région qu'ils occupent, les îlots CpG comme ces PLM apportent une richesse en C et G aux promoteurs. Chez les mammifères, l'impact de la richesse en A+T ou en C+G dans les régions promotrices sur l'organisation de la région du TSS a été démontrée (Bajic *et al.*, 2006). C'est pourquoi nous proposons que la richesse en C et G dans les promoteurs puisse chez *A. thaliana* comme chez *H. sapiens* être associée à une organisation similaire des TSS alternatifs.

A ce jour, aucune étude n'a été réalisée pour étudier l'organisation des TSS alternatifs des gènes contenant un motif-TC. Plusieurs critères permettent néanmoins de proposer

l'hypothèse que ces gènes aient des TSS alternatifs présents sur une région large sans TSS dominant. Premièrement, les caractéristiques fonctionnelles et structurales des gènes contenant une boîte TATA ou un motif-TC sont opposées. Il peut être envisagé que les gènes contenant un motif-TC aient une organisation des TSS opposée à celle des gènes contenant une boîte TATA. Ces derniers ont des TSS alternatifs sur une région restreinte et ont un TSS dominant. Deuxièmement, nous avons mis en évidence que les gènes possédant les motifs-TC ont des promoteurs plus riches en bases C et G. Le même biais a été observé chez *H. sapiens* au sein des gènes dont les promoteurs ne contiennent pas de boîte TATA (Carninci *et al.*, 2006 ; Yang *et al.*, 2007). De plus, ces gènes d'*H. sapiens* sont fréquemment associés à la présence de TSS alternatifs présents dans une région large sans TSS dominant (Carninci *et al.*, 2006). La même organisation de TSS alternatifs peut être retrouvée chez les gènes d'*A. thaliana* partageant une même richesse en bases C et G dans leurs promoteurs.

9.5 Pertinence de la recherche de caractéristiques communes aux gènes contenant un élément régulateur

Différentes études, dont cette thèse, ont montré l'existence de biais structuraux et fonctionnels au sein des gènes contenant un élément régulateur donné. Par exemple, les gènes contenant une boîte TATA ont tendance à être plus courts, à être impliqués dans des processus biologiques spécifiques et à s'exprimer à une intensité plus forte que les autres gènes. La conservation de ces biais des protozoaires aux métazoaires permet de supposer que ces biais sont dus à la présence de cette boîte. Néanmoins, il est à noter que des études ne prenant pas en considération la présence des éléments régulateurs ont démontré que des gènes d'*A. thaliana* ayant un UTR 5' court sont plus souvent exprimés avec une forte intensité (Chung *et al.*, 2006). Ainsi il faut se demander si les biais caractérisant les gènes contenant une boîte TATA sont réellement dus à sa présence ou à d'autres caractéristiques. Dans ce cas précis, pour le vérifier, il serait intéressant d'analyser les biais de quatre groupes de gènes dont les UTR 5' sont courts ou longs et dont les promoteurs contiennent ou non une boîte TATA. La robustesse des tests statistiques serait diminuée mais l'étude permettrait de conclure quant à la caractérisation des biais induits par la présence d'éléments régulateurs.

Les éléments régulateurs agissent en modules et la considération d'un seul élément régulateur peut être limitante : l'expression d'un gène est régulée par un ensemble d'éléments. Plusieurs associations de TFBS sont possibles et peuvent conduire à différentes régulations : un TFBS au sein de deux modules différents peut induire des régulations différentes. Enfin, la conformation de l'ADN, la méthylation de l'ADN et la présence des nucléosomes sont impliquées dans la régulation de l'expression des gènes et peuvent activer ou inhiber la transcription. Ainsi, même si des biais sont mis en évidence au sein d'un groupe de gènes ayant une même architecture des promoteurs, ils peuvent être dus à la présences des PLM, mais également à d'autres paramètres qui ne sont pas pris en compte par l'approche PLM. Ces biais sont donc des pistes pour déterminer le rôle fonctionnel des TFBS, mais ne permettent pas de conclure définitivement à leur rôle.

9.6 Conservation et évolution des TFBS

Les gènes orthologues sont susceptibles de conserver une même organisation de leurs TFBS (Brown *et al.*, 2007). Ce critère est utilisé pour identifier des TFBS (Blanchette & Tompa, 2002; Vandepoele *et al.*, 2006), mais il n'est pas pertinent si les espèces ont trop ou trop peu divergé. Les gènes orthologues d'*H. sapiens* et de *M. musculus* ont conservé une même organisation des TFBS généraux dans leurs promoteurs. Ceci a permis de prédire que 17% des gènes de mammifères contiennent une boîte TATA (Jin *et al.*, 2006). Chez *A. thaliana* et *O. sativa*, sans considérer les gènes orthologues, une même organisation des promoteurs est observée (Yamamoto *et al.* 2007b). Cependant, entre gènes orthologues, nous avons montré que la conservation de la boîte TATA est très faible. Parmi les 5805 paires d'orthologues obtenues par une recherche de Bi-Directional Best Hit (Overbeek *et al.*, 1999), moins de 4% possèdent une boîte TATA chez les deux organismes. Malgré ce faible pourcentage, ce TFBS est le plus conservé des éléments observés dans la région -30 : les motifs-TC et les variants fonctionnels de la boîte TATA, ne sont pas conservés entre gènes orthologues. Globalement, nous avons donc montré que les trois catégories d'éléments régulateurs présents en -30 par rapport au TSS ne sont pas conservés entre gènes orthologues. Ces résultats pourraient laisser supposer qu'*A. thaliana* et *O. sativa* ne sont pas de bons candidats pour des études en génomique comparative pour la recherche de TFBS conservés. Néanmoins, il est intéressant de rechercher une conservation globale des éléments régulateurs potentiels entre ces deux organismes. Observer les mêmes motifs-TC chez *A. thaliana* et *O. sativa* est un argument pour juger de la robustesse de la prédiction de ces éléments.

Un turnover de TFBS spécifiques a été observé chez *H. sapiens* et *M. musculus* (Dermitzakis & Clark, 2002). Les résultats de différentes études concernant les éléments régulateurs observés environ 30 bases en amont du TSS pourraient laisser penser qu'un turnover pourrait également contrôler l'évolution de ces éléments. Les gènes contenant une boîte TATA évoluent plus rapidement (Basehoar *et al.*, 2004). De plus, une évolution est possible entre la séquence canonique des boîtes TATA et les séquences des variants fonctionnels (Figure 7-4 page 108), mais aussi entre les motifs-TC (manuscrit de l'article page 117). Ceci pourrait enfin expliquer l'absence de conservation des éléments entre gènes orthologues.

Les gènes dont le promoteur contient une boîte TATA ont conservé les mêmes caractéristiques structurales et fonctionnelles des protozoaires aux métazoaires (Basehoar *et al.*, 2004; Molina & Grotewold, 2005; Yang *et al.*, 2007; Moshonov *et al.*, 2008), et la même organisation de leurs TSS alternatifs des plantes aux mammifères (Carninci, 2006; Tanaka *et al.*, 2009; Yamamoto *et al.*, 2009). C'est donc probablement la présence de la boîte TATA qui induit ces caractéristiques. Qu'elle ne soit pas conservée entre orthologues mais qu'elle soit présente dans les mêmes proportions au sein des génomes eucaryotes permet de supposer qu'il existe une pression de sélection à l'échelle génomique qui imposerait un taux de présence de la boîte TATA.

Les microsatellites riches en bases C et T comme les motifs-TC sont deux catégories d'éléments dont la présence a été observée uniquement chez les plantes. Ces deux éléments sont conservés quantitativement entre *A. thaliana* et *O. sativa*. Les microsatellites s'étendent en amont du TSS (Molina & Grotewold, 2005), mais ne chevauchent pas les

motifs-TC. Une partie des séquences riches en C et T conduit à une sur-représentation locale dans les promoteurs qui est à la fois caractéristique des motifs-TC et des microsatellites. Ces critères pourraient témoigner d'un lien évolutif entre ces deux éléments. Après une évolution potentielle, ils ont pu se différencier pour avoir aujourd'hui des caractéristiques structurales et fonctionnelles qui les distinguent. Tandis que les microsatellites sont observés sur une large région dans environ 93% des UTR 5' de gènes chez *A. thaliana*, seulement 18% contiennent un motif-TC caractérisé par des contraintes topologiques très strictes. Ainsi, les microsatellites sont probablement impliqués dans la conformation de l'ADN tandis que nous supposons que les motifs-TC sont des sites de fixation des facteurs de transcription.

10 Conclusions et perspectives

Ce document présente une cartographie des promoteurs proximaux d'*A. thaliana* obtenue par l'utilisation de l'approche PLM qui recherche les motifs présents à une distance préférentielle du TSS. Il propose l'existence d'une nouvelle classe d'éléments régulateurs : les motifs-TC observés à l'emplacement de la boîte TATA, des motifs conservés au sein des plantes mais absents des génomes de mammifères étudiés.

10.1 Atouts de l'approche PLM

L'approche PLM permet d'obtenir automatiquement une liste d'éléments régulateurs potentiels, avec leurs contraintes topologiques, c'est-à-dire leurs positions préférentielles et leurs fenêtres fonctionnelles, et des listes de gènes qui les contiennent. Les motifs à étudier, le jeu de promoteurs à analyser et le seuil du SMS (3 par défaut) restent des paramètres que l'utilisateur peut définir. L'approche PLM a démontré son efficacité à la fois lors de l'analyse globale des promoteurs d'*A. thaliana*, et lors d'analyses de sous-groupes de gènes. Elle a permis de prédire l'existence de nouveaux éléments régulateurs observés à l'échelle génomique ou à l'échelle de sous-groupes de gènes. Ces résultats montrent l'intérêt d'utiliser les deux approches conjointement pour obtenir une cartographie des promoteurs plus complète. De plus, l'approche PLM a permis de caractériser les contraintes topologiques de certains TFBS connus indexés dans les bases de données AGRIS et PLACE (Higo *et al.*, 1999; Davuluri *et al.*, 2003).

Le choix de la taille de la fenêtre glissante est un paramètre important pour identifier les PLM. Il est géré automatiquement dans l'approche PLM. Lors d'études visant à rechercher des caractéristiques fonctionnelles communes à un groupe de gènes ayant la même architecture de promoteurs, il est important d'éviter la présence de gènes faux positifs qui aurait pour effet d'affaiblir les analyses statistiques et / ou d'induire des résultats biaisés.

Enfin, cette approche peut être exploitée pour étudier les promoteurs d'autres eucaryotes dont les transcrits permettent de définir la position du TSS comme *O. sativa*, *H. sapiens* ou *M. musculus*, tous trois analysés au cours de cette thèse.

10.2 Cartographie des promoteurs d'*A. thaliana*

Au début de cette thèse, quelques études avaient mis en évidence des biais compositionnels en bases dans la région du TSS, premières indications de la présence de l'Inr (Shahmuradov *et al.*, 2003; Fujimori *et al.*, 2005; Alexandrov *et al.*, 2006). Une seule analyse réalisée à l'échelle de l'ensemble des promoteurs chez *A. thaliana* avait été proposée par Molina *et al.* (2005) et concernait exclusivement le promoteur central. Aucune étude n'avait proposé de cartographie des promoteurs proximaux d'*A. thaliana*. Cela a été publié pendant le déroulement de cette thèse par Yamamoto *et al.* (2007b). Les différences méthodologiques entre leur approche et l'approche PLM permettent de confirmer leurs analyses mais aussi de compléter certains résultats. L'approche PLM utilisée pour l'analyse globale des 14927 promoteurs d'*A. thaliana* a permis d'identifier 5105 PLM. Le bilan des PLM mis en évidences au cours de ce travail est schématisé Figure 10-1. Il est à noter que cette représentation ne doit pas être considérée comme un promoteur type. Aucun promoteurs ne contient l'ensemble des éléments décrits.

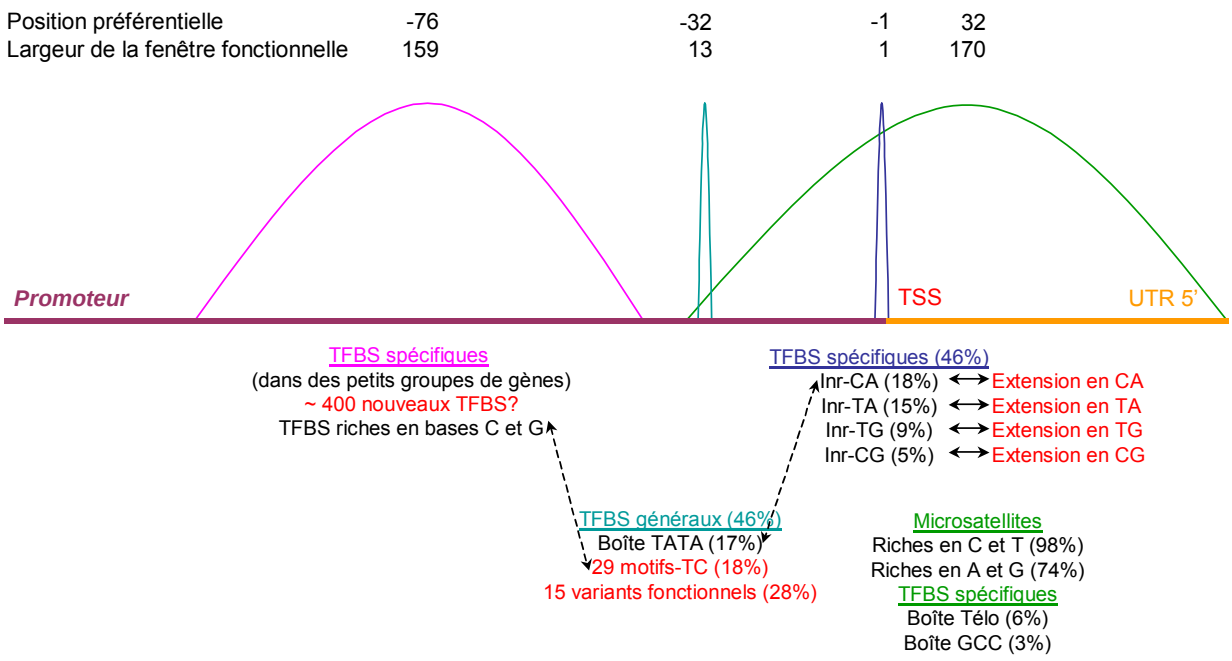


Figure 10-1 : Les éléments régulateurs potentiels et leurs associations.

Bilan des PLM identifiés au cours du projet, avec en rouge, les éléments nouveaux. Pour les quatre grandes catégories d'éléments régulateurs, représentées de 4 couleurs, la médiane des positions préférentielles et celle des largeurs de fenêtres fonctionnelles sont renseignées. Les flèches entre deux éléments mettent en évidence qu'ils peuvent être associés au sein de mêmes promoteurs.

10.2.1 Eléments régulateurs spécifiques

Les 1217 PLM qui ont une position préférentielle en amont de la position -50 sont près des trois quarts à avoir une séquence qui est appariée avec des TFBS connus chez *A. thaliana*. Les PLM restants sont observés dans des petits groupes de gènes et sont plus fréquemment orientés sur un brin préférentiel. Ces caractéristiques en font de potentiels nouveaux éléments régulateurs de l'expression des gènes.

Les TFBS agissent souvent en modules avec d'autres éléments régulateurs qui sont leurs partenaires fonctionnels (Berman *et al.*, 2002; Blanchette *et al.*, 2006). Il serait intéressant d'analyser les associations de PLM préférentiellement positionnés dans le promoteur proximal. La recherche de motifs présents simultanément au sein des promoteurs et la recherche de caractéristiques fonctionnelles communes des gènes les contenant seraient donc une contribution à la caractérisation fonctionnelle de ces motifs, comme cela a été réalisé pour les éléments observés à l'emplacement de la boîte TATA. Il serait également intéressant d'analyser des modules entre ces PLM et ceux du promoteur central, caractérisés pas des contraintes topologiques strictes. En effet, chez les mammifères des associations entre les gènes ne possédant pas de boîte TATA et la présence d'îlots CpG ont été observées (Carninci, 2006).

10.2.2 Microsatellites dans les UTR

La majorité des 3421 PLM qui ont une fenêtre fonctionnelle large et une position préférentielle en aval de -50 par rapport au TSS montrent une richesse en microsatellites

TC, TTC, GA et GAA dans l'UTR 5' des plantes (Morgante *et al.*, 2002; Fujimori *et al.*, 2003). Nous avons mis en évidence que 98% des gènes contiennent ces séquences au sein de leur UTR 5'. A l'échelle génomique, les analyses réalisées au cours de cette thèse n'ont pas mis en évidence de corrélation entre la longueur des répétitions et l'expression des gènes, ce qui avait été proposé à l'échelle de quelques gènes (Zhang *et al.*, 2006). Ces séquences répétées pourraient être impliquées dans la conformation de l'ADN et avoir une influence sur la fixation des nucléosomes.

10.2.3 Boîte TATA, variants fonctionnels et motifs-TC

Les 405 PLM à fenêtre fonctionnelle de taille inférieure à 50 et positionnés préférentiellement dans la région [-42, -24] sont riches en A et T et montrent la présence de la boîte TATA, de ses variants fonctionnels et de ses variants de séquence.

Chez *A. thaliana*, les connaissances relatives aux éléments régulateurs observés à l'emplacement de la boîte TATA étaient faibles en 2006. Cette thèse a permis de proposer une liste stricte de 2606 promoteurs (17%) contenant la boîte TATA canonique, c'est-à-dire une occurrence de TATAWA dans la région fonctionnelle [-39, -26]. La recherche de caractéristiques fonctionnelles partagées par les gènes contenant une boîte TATA chez *A. thaliana* a conduit aux mêmes résultats que ceux observés chez d'autres organismes. Chez les eucaryotes, la présence de la boîte TATA dans des promoteurs est associée à des structures de gènes plus courtes, une intensité d'expression plus élevée et est impliquée dans des fonctions spécifiques, liées au stress.

De plus, une liste de 15 variants supposés fonctionnels car correspondant à des PLM aux contraintes fortes et conservés entre *A. thaliana* et *O. sativa* a été définie. Ils sont observés dans 28% des gènes chez *A. thaliana*. L'étude des gènes contenant les variants a permis de distinguer plusieurs catégories d'éléments. Le variant AATAAA pourrait avoir évolué et s'être différencié fonctionnellement de la boîte TATA et des autres variants de séquence. Les autres variants ne présentent pas de biais significatifs lors des études menées au cours de cette thèse.

Enfin, nous avons mis en évidence une nouvelle classe d'éléments présents au même emplacement que la boîte TATA. Ces éléments sont caractérisés par les mêmes contraintes topologiques strictes et sont nommés les motifs-TC. Ils sont conservés au sein des promoteurs d'*A. thaliana* et d'*O. sativa*. Ces éléments sont comme les microsatellites absents du règne des mammifères et pourraient avoir un lien évolutif avec eux. Les 18% de gènes contenant les motifs-TC ont des caractéristiques fonctionnelles et structurales qui les distinguent des autres gènes, ce qui permet de supposer que les motifs-TC auraient un rôle spécifique différent de celui des autres éléments régulateurs observés à l'emplacement de la boîte TATA.

Ainsi, près de la moitié des promoteurs d'*A. thaliana* sont caractérisés par un élément régulateur présent dans la région [-39, -26]. Les différents rôles de ces éléments peuvent conduire à des régulations différentes de la transcription des gènes. Les variants fonctionnels et les motifs-TC ont été prédits *in silico*. Pour être caractérisés comme étant réellement des TFBS, une des perspectives de ce travail serait de valider expérimentalement le rôle de ces motifs dans la régulation de l'expression des gènes.

10.2.4 Elément initiateur

Les PLM à fenêtre fonctionnelle de taille inférieure à 50 et chevauchant le TSS dans l'intervalle [-6, 7] contiennent très souvent le dinucléotide CA. Ces motifs mettent en évidence la présence attendue de l'Inr-YR. Le travail réalisé au cours de cette thèse a permis d'apporter des connaissances nouvelles concernant l'Inr, identifié par Yamamoto *et al.* (2007b) chez *A. thaliana*. Tout d'abord, l'Inr-CA est le plus observé parmi les quatre Inr-YR. Il est présent dans 18% des promoteurs une base en amont du TSS, l'Inr-YR étant dans 46% des promoteurs chez *A. thaliana*. Cette présence majoritaire n'est pas le seul critère qui distingue l'Inr-CA des autres Inr. Cet élément est plus souvent observé avec la boîte TATA au sein des promoteurs et est distant de 30 à 33 bases en aval de cet élément. Ils pourraient agir en module et collaborer fonctionnellement pour réguler la transcription des gènes. Enfin, une extension de la présence de chacun des dinucléotides YR en aval de l'Inr a été mise en évidence. Elle pourrait indiquer la présence de TSS alternatifs dans une région stricte qui seraient caractérisés par un même dinucléotide initiateur.

10.3 Approche PLM pour aider à l'annotation fonctionnelle des gènes

L'approche PLM va être exploitée dans un projet de l'équipe de Bioinformatique de l'URGV qui a pour objectif de développer une méthodologie permettant de prédire le rôle de gènes non annotés chez *A. thaliana*. Ce projet va s'appuyer notamment sur une analyse de ressources transcriptomiques et sur un ensemble d'analyses et de prédictions bioinformatiques incluant la recherche de PLM. Après avoir regroupé des gènes co-exprimés (Maugis *et al.*, 2009),l'approche PLM pourra permettre de prédire l'existence de potentiels éléments régulateurs spécifiques d'un groupe de gènes.

Une partie de ce projet exploitera (i) les TFBS validés expérimentalement, qui sont indexés dans les bases de données PLACE et AGRIS (Davuluri *et al.*, 2003 ; Higo *et al.*, 1999) et qui sont des PLM, et (ii) la cartographie des promoteurs d'*A. thaliana* générée lors de cette thèse. La recherche de sur-représentation de ces motifs dans différents groupes de gènes co-exprimés pourra permettre d'associer des TFBS à des groupes de gènes. Cela servira notamment à désigner des groupes de gènes d'intérêt susceptibles d'être régulés par un même ensemble de TFBS.

Ainsi, cette thèse propose une approche pour l'identification automatique d'éléments régulateurs potentiels. Mise au point chez *A. thaliana*, elle peut être utilisée chez d'autres organismes eucaryotes sous réserve de la connaissance de leur génome et d'une quantité de transcrits suffisante pour définir les unités de transcription des gènes. Les analyses réalisées ont conduit à une cartographie des promoteurs de cette plante modèle qui ouvre des perspectives sur de multiples analyses. Cette ressource peut être exploitée pour analyser des groupes de gènes ayant une même architecture de leurs promoteurs. Des études ayant pour objectif d'identifier les caractères communs de tels groupes de gènes ont permis de mieux caractériser les éléments régulateurs qu'ils contiennent. Ces résultats permettent de contribuer à la connaissance de leur rôle et donc de mieux connaître les éléments impliqués dans la régulation de la transcription des gènes.

Références

Abeel T, Saeys Y, Bonnet E, Rouze P, Van de Peer Y (2008) Generic eukaryotic core promoter prediction using structural features of DNA. Genome Res **18:** 310-323

Abeel T, Saeys Y, Rouze P, Van de Peer Y (2008) ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles. Bioinformatics **24:** i24-31

AGI (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408:** 796-815

Akan P, Deloukas P (2008) DNA sequence and structural properties as predictors of human and mouse promoters. Gene **410:** 165-176

Alexandrov NN, Troukhan ME, Brover VV, Tatarinova T, Flavell RB, Feldmann KA (2006) Features of Arabidopsis genes and genome discovered using full-length cDNAs. Plant Mol Biol **60:** 69-85

Allemeersch J, Durinck S, Vanderhaeghen R, Alard P, Maes R, Seeuws K, Bogaert T, Coddens K, Deschouwer K, Van Hummelen P, Vuylsteke M, Moreau Y, Kwekkeboom J, Wijfjes AH, May S, Beynon J, Hilson P, Kuiper MT (2005) Benchmarking the CATMA microarray. A novel tool for Arabidopsis transcriptome analysis. Plant Physiol **137:** 588-601

Anish R, Hossain MB, Jacobson RH, Takada S (2009) Characterization of transcription from TATA-less promoters: identification of a new core promoter element XCPE2 and analysis of factor requirements. PLoS One **4:** e5103

Antequera F, Bird A (1993) Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A **90:** 11995-11999

Armisen D (2008) Les gènes uniques chez les plantes: caractéristiques, évolution et promoteurs. Université d'Evry, Evry

Armisén D, Lecharny A, Aubourg S (2008) Unique genes in plants: specificities and conserved features throughout evolution. BMC Evol Biol **8:** 280

Asamizu E, Nakamura Y, Sato S, Tabata S (2000) A large scale analysis of cDNA in Arabidopsis thaliana: generation of 12,028 non-redundant expressed sequence tags from normalized and size-selected cDNA libraries. DNA Res **7:** 175-180

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25:** 25-29

Aubourg S, Boudet N, Kreis M, Lecharny A (2000) In Arabidopsis thaliana, 1% of the genome codes for a novel protein family unique to plants. Plant Mol Biol **42:** 603-613

Aubourg S, Brunaud V, Bruyere C, Cock M, Cooke R, Cottet A, Couloux A, Dehais P, Deleage G, Duclert A, Echeverria M, Eschbach A, Falconet D, Filippi G, Gaspin C, Geourjon C, Grienenberger JM, Houlne G, Jamet E, Lechauve F, Leleu O, Leroy P, Mache R, Meyer C, Nedjari H, Negrutiu I, Orsini V, Peyretaillade E, Pommier C, Raes J, Risler JL, Riviere S, Rombauts S, Rouze P, Schneider M, Schwob P, Small I, Soumayet-Kampetenga G, Stankovski D, Toffano C, Tognolli M, Caboche M, Lecharny A (2005) GeneFarm, structural and functional annotation of Arabidopsis gene and protein families by a network of experts. Nucleic Acids Res **33:** D641-646

Aubourg S, Martin-Magniette ML, Brunaud V, Taconnat L, Bitton F, Balzergue S, Jullien PE, Ingouff M, Thareau V, Schiex T, Lecharny A, Renou JP (2007) Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome. BMC Genomics **8:** 401

Audic S, Claverie JM (1997) The significance of digital gene expression profiles. Genome Res **7:** 986-995

Audic S, Claverie JM (1998) Visualizing the competitive recognition of TATA-boxes in vertebrate promoters. Trends Genet **14:** 10-11

Bajic VB, Tan SL, Christoffels A, Schonbach C, Lipovich L, Yang L, Hofmann O, Kruger A, Hide W, Kai C, Kawai J, Hume DA, Carninci P, Hayashizaki Y (2006) Mice and men: their promoter properties. PLoS Genet **2:** e54

Barbazuk WB, Fu Y, McGinnis KM (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges. Genome Res **18:** 1381-1392

Bareket-Samish A, Cohen I, Haran TE (2000) Signals for TBP/TATA box recognition. J Mol Biol **299:** 965-977

Barrera LO, Ren B (2006) The transcriptional regulatory code of eukaryotic cells--insights from genome-wide analysis of chromatin organization and transcription factor binding. Curr Opin Cell Biol **18:** 291-298

Barton MC, Madani N, Emerson BM (1997) Distal enhancer regulation by promoter derepression in topologically constrained DNA in vitro. Proc Natl Acad Sci U S A **94:** 7257-7262

Basehoar AD, Zanton SJ, Pugh BF (2004) Identification and distinct regulation of yeast TATA box-containing genes. Cell **116:** 699-709

Bellora N, Farre D, Alba MM (2007) Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters. BMC Genomics **8:** 459

Bellora N, Farre D, Mar Alba M (2007) PEAKS: identification of regulatory motifs by their position in DNA sequences. Bioinformatics **23:** 243-244

Benhamed M, Martin-Magniette ML, Taconnat L, Bitton F, Servet C, De Clercq R, De Meyer B, Buysschaert C, Rombauts S, Villarroel R, Aubourg S, Beynon J, Bhalerao RP, Coupland G, Gruissem W, Menke FL, Weisshaar B, Renou JP, Zhou DX, Hilson P (2008) Genome-scale Arabidopsis promoter array identifies targets of the histone acetyltransferase GCN5. Plant J **56:** 493-504

Berendzen KW, Stuber K, Harter K, Wanke D (2006) Cis-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves. BMC Bioinformatics **7:** 522

Berg J, Willmann S, Lassig M (2004) Adaptive evolution of transcription factor binding sites. BMC Evol Biol **4:** 42

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB (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:** 757-762

Bernard V, Brunaud V, Serizet C, Martin-Magniette ML, Caboche M, Aubourg S, Lecharny A (2006) Sélection de motifs candidats pour la régulation des gènes chez *Arabidopsis thaliana* sur des critères topologiques. *In* Journée Ouvertes de la Bioinformatique et des Mathématiques, Bordeaux, pp 17-28

Biedenkapp H, Borgmeyer U, Sippel AE, Klempnauer KH (1988) Viral myb oncogene encodes a sequence-specific DNA-binding activity. Nature **335:** 835-837

Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, Coulombe B, Robert F (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res **16:** 656-668

Blanchette M, Sinha S (2001) Separating real motifs from their artifacts. Bioinformatics **17 Suppl 1:** S30-38

Blanchette M, Tompa M (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res **12:** 739-748

Blazquez MA, Weigel D (2000) Integration of floral inductive signals in Arabidopsis. Nature **404:** 889-892

Boden M, Bailey TL (2008) Associating transcription factor-binding site motifs with target GO terms and target genes. Nucleic Acids Res **36:** 4108-4117

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science **299:** 1391-1394

Boguski MS, Lowe TM, Tolstoshev CM (1993) dbEST--database for "expressed sequence tags". Nat Genet **4:** 332-333

Bolshoy A, McNamara P, Harrington RE, Trifonov EN (1991) Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. Proc Natl Acad Sci U S A **88:** 2312-2316

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell **132:** 311-322

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol **18:** 630-634

Brown CD, Johnson DS, Sidow A (2007) Functional architecture and evolution of transcriptional elements that drive gene coexpression. Science **317:** 1557-1560

Brunaud V, Balzergue S, Dubreucq B, Aubourg S, Samson F, Chauvin S, Bechtold N, Cruaud C, DeRose R, Pelletier G, Lepiniec L, Caboche M, Lecharny A (2002) T-DNA integration into the Arabidopsis genome depends on sequences of pre-insertion sites. EMBO Rep **3:** 1152-1157

Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res **36:** D102-106

Buisine N, Quesneville H, Colot V (2008) Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. Genomics **91:** 467-475

Bulyk ML (2003) Computational prediction of transcription-factor binding site locations. Genome Biol **5:** 201

Buratowski S (1997) Multiple TATA-binding factors come back into style. Cell **91:** 13-15

Burke TW, Kadonaga JT (1996) Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. Genes Dev **10:** 711-724

Burke TW, Kadonaga JT (1997) The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. Genes Dev **11:** 3020-3031

Burke TW, Willy PJ, Kutach AK, Butler JE, Kadonaga JT (1998) The DPE, a conserved downstream core promoter element that is functionally analogous to the TATA box. Cold Spring Harb Symp Quant Biol **63:** 75-82

Carninci P (2006) Tagging mammalian transcription complexity. Trends Genet **22:** 501-510

Carninci P, Kvam C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Izawa M, Muramatsu M, Hayashizaki Y, Schneider C (1996) High-efficiency full-length cDNA cloning by biotinylated CAP trapper. Genomics **37:** 327-336

Casimiro AC, Vinga S, Freitas AT, Oliveira AL (2008) An analysis of the positional distribution of DNA motifs in promoter regions and its biological relevance. BMC Bioinformatics **9:** 89

Castelli V, Aury JM, Jaillon O, Wincker P, Clepet C, Menard M, Cruaud C, Quetier F, Scarpelli C, Schachter V, Temple G, Caboche M, Weissenbach J, Salanoubat M (2004) Whole genome sequence comparisons and "full-length" cDNA sequences: a combined approach to evaluate and improve Arabidopsis genome annotation. Genome Res **14:** 406-413

Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell **116:** 499-509

Chattopadhyay S, Ang LH, Puente P, Deng XW, Wei N (1998) Arabidopsis bZIP protein HY5 directly interacts with light-responsive promoters in mediating light control of gene expression. Plant Cell **10:** 673-683

Chung BY, Simons C, Firth AE, Brown CM, Hellens RP (2006) Effect of 5'UTR introns on gene expression in Arabidopsis thaliana. BMC Genomics **7:** 120

Civan P, Svec M (2009) Genome-wide analysis of rice (Oryza sativa L. subsp. japonica) TATA box and Y Patch promoter elements. Genome **52:** 294-297

Clepet C, Le Clainche I, Caboche M (2004) Improved full-length cDNA production based on RNA tagging by T4 DNA ligase. Nucleic Acids Res **32:** e6

Colcombet J, Hirt H (2008) Arabidopsis MAPKs: a complex signalling network involved in multiple biological processes. Biochem J **413:** 217-226

Colinas J, Birnbaum K, Benfey PN (2002) Using cauliflower to find conserved non-coding regions in Arabidopsis. Plant Physiol **129:** 451-454

Consortium EP (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science **306:** 636-640

Cooke R, Raynal M, Laudie M, Grellet F, Delseny M, Morris PC, Guerrier D, Giraudat J, Quigley F, Clabault G, Li YF, Mache R, Krivitzky M, Gy IJ, Kreis M, Lecharny A, Parmentier Y, Marbach J, Fleck J, Clement B, Philipps G, Herve C, Bardet C, Tremousaygue D, Hofte H, et al. (1996) Further progress towards a catalogue of all Arabidopsis genes: analysis of a set of 5000 non-redundant ESTs. Plant J **9:** 101-124

Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM (2006) Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome Res **16:** 1-10

Cora D, Di Cunto F, Provero P, Silengo L, Caselle M (2004) Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs. BMC Bioinformatics **5:** 57

Corden J, Wasylyk B, Buchwalder A, Sassone-Corsi P, Kedinger C, Chambon P (1980) Promoter sequences of eukaryotic protein-coding genes. Science **209:** 1406-1414

Coughlan SJ, Agrawal V, Meyers B (2004) A Comparison of Global Gene Expression Measurement Technologies in Arabidopsis thaliana. Comp Funct Genomics **5:** 245-252

Crowe ML, Serizet C, Thareau V, Aubourg S, Rouze P, Hilson P, Beynon J, Weisbeek P, van Hummelen P, Reymond P, Paz-Ares J, Nietfeld W, Trick M (2003) CATMA: a complete Arabidopsis GST database. Nucleic Acids Res **31:** 156-158

Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E (2003) AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC Bioinformatics **4:** 25

Defrance M, Helden JV (2009) Info-gibbs: a motif discovery algorithm that directly optimizes information content during sampling. Bioinformatics

Defrance M, Janky R, Sand O, van Helden J (2008) Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. Nat Protoc **3:** 1589-1603

Deng W, Roberts SG (2005) A core promoter element downstream of the TATA box that is recognized by TFIIB. Genes Dev **19:** 2418-2423

DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **278:** 680-686

Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol **19:** 1114-1121

Dessau RB, Pipper CB (2008) ["R"--project for statistical computing]. Ugeskr Laeger **170:** 328-330

Deveaux Y, Toffano-Nioche C, Claisse G, Thareau V, Morin H, Laufs P, Moreau H, Kreis M, Lecharny A (2008) Genes of the most conserved WOX clade in plants affect root and flower development in Arabidopsis. BMC Evol Biol **8:** 291

Deyholos MK, Sieburth LE (2000) Separable whorl-specific expression and negative regulation by enhancer elements within the AGAMOUS second intron. Plant Cell **12:** 1799-1810

Dion V, Coulombe B (2003) Interactions of a DNA-bound transcriptional activator with the TBP-TFIIA-TFIIB-promoter quaternary complex. J Biol Chem **278:** 11495-11501

Dreyfus M, Regnier P (2002) The poly(A) tail of mRNAs: bodyguard in eukaryotes, scavenger in bacteria. Cell **111:** 611-613

Ferretti V, Poitras C, Bergeron D, Coulombe B, Robert F, Blanchette M (2007) PReMod: a database of genome-wide mammalian cis-regulatory module predictions. Nucleic Acids Res **35:** D122-126

FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C (2004) Clustering of DNA sequences in human promoters. Genome Res **14:** 1562-1574

Frith MC, Ponjavic J, Fredman D, Kai C, Kawai J, Carninci P, Hayashizaki Y, Sandelin A (2006) Evolutionary turnover of mammalian transcription start sites. Genome Res **16:** 713-722

Fujimori S, Washio T, Higo K, Ohtomo Y, Murakami K, Matsubara K, Kawai J, Carninci P, Hayashizaki Y, Kikuchi S, Tomita M (2003) A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription. FEBS Lett **554:** 17-22

Fujimori S, Washio T, Tomita M (2005) GC-compositional strand bias around transcription start sites in plants and fungi. BMC Genomics **6:** 26

Gagnot S, Tamby JP, Martin-Magniette ML, Bitton F, Taconnat L, Balzergue S, Aubourg S, Renou JP, Lecharny A, Brunaud V (2008) CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. Nucleic Acids Res **36:** D986-990

Galas DJ, Schmitz A (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res **5:** 3157-3170

Garner MM, Revzin A (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. Nucleic Acids Res **9:** 3047-3060

Goffeau A (1996) 1996: a vintage year for yeast and Yeast. Yeast **12:** 1603-1605

Graber JH (2003) Variations in yeast 3'-processing cis-elements correlate with transcript stability. Trends Genet **19:** 473-476

Graber JH, McAllister GD, Smith TF (2002) Probabilistic prediction of Saccharomyces cerevisiae mRNA 3'-processing sites. Nucleic Acids Res **30:** 1851-1858

Grace ML, Chandrasekharan MB, Hall TC, Crowe AJ (2004) Sequence and spacing of TATA box elements are critical for accurate initiation from the beta-phaseolin promoter. J Biol Chem **279:** 8102-8110

GuhaThakurta D (2006) Computational identification of transcriptional regulatory elements in DNA sequence. Nucleic Acids Res **34:** 3585-3598

Guo A, He K, Liu D, Bai S, Gu X, Wei L, Luo J (2005) DATF: a database of Arabidopsis transcription factors. Bioinformatics **21:** 2568-2569

Guo AY, Chen X, Gao G, Zhang H, Zhu QH, Liu XC, Zhong YF, Gu X, He K, Luo J (2008) PlantTFDB: a comprehensive plant transcription factor database. Nucleic Acids Res **36:** D966-969

Guo H, Moose SP (2003) Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. Plant Cell **15:** 1143-1158

Han Y, Burnette JM, 3rd, Wessler SR (2009) TARGeT: a web-based pipeline for retrieving and characterizing gene and transposable element families from genomic sequences. Nucleic Acids Res **37:** e78

Hartley PD, Madhani HD (2009) Mechanisms that specify promoter nucleosome location and identity. Cell **137:** 445-458

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet **39:** 311-318

Henikoff S (2008) Nucleosome destabilization in the epigenetic regulation of gene expression. Nat Rev Genet **9:** 15-26

Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res **27:** 297-300

Hoebeke M, Schbath S (2006) R'MES: Finding Exceptional Motifs, version 3. User guide. *In*,

Hudson ME, Quail PH (2003) Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data. Plant Physiol **133:** 1605-1616

Hulzink RJ, Weerdesteyn H, Croes AF, Gerats T, van Herpen MM, van Helden J (2003) In silico identification of putative regulatory sequence elements in the 5'-untranslated region of genes that are expressed during male gametogenesis. Plant Physiol **132:** 75-83

Ioshikhes I, Trifonov EN, Zhang MQ (1999) Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. Proc Natl Acad Sci U S A **96:** 2891-2895

Ioshikhes IP, Zhang MQ (2000) Large-scale human promoter mapping using CpG islands. Nat Genet **26:** 61-63

Javahery R, Khachi A, Lo K, Zenzie-Gregory B, Smale ST (1994) DNA sequence requirements for transcriptional initiator activity in mammalian cells. Mol Cell Biol **14:** 116-127

Jin VX, Singer GA, Agosto-Perez FJ, Liyanarachchi S, Davuluri RV (2006) Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. BMC Bioinformatics **7:** 114

Joshi CP (1987) An inspection of the domain between putative TATA box and translation start site in 79 plant genes. Nucleic Acids Res **15:** 6643-6653

Juo ZS, Chiu TK, Leiberman PM, Baikalov I, Berk AJ, Dickerson RE (1996) How proteins recognize the TATA box. J Mol Biol **261:** 239-254

Juven-Gershon T, Hsu JY, Theisen JW, Kadonaga JT (2008) The RNA polymerase II core promoter - the gateway to transcription. Curr Opin Cell Biol **20:** 253-259

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature **423:** 241-254

Kel-Margoulis OV, Kel AE, Reuter I, Deineko IV, Wingender E (2002) TRANSCompel: a database on composite regulatory elements in eukaryotic genes. Nucleic Acids Res **30:** 332-334

King DC, Taylor J, Zhang Y, Cheng Y, Lawson HA, Martin J, Chiaromonte F, Miller W, Hardison RC (2007) Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data. Genome Res **17:** 775-786

Kiran K, Ansari SA, Srivastava R, Lodhi N, Chaturvedi CP, Sawant SV, Tuli R (2006) The TATA-box sequence in the basal promoter contributes to determining light-dependent gene expression in plants. Plant Physiol **142:** 364-376

Kiyama R, Trifonov EN (2002) What positions nucleosomes?--A model. FEBS Lett **523:** 7-11

Kutach AK, Kadonaga JT (2000) The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. Mol Cell Biol **20:** 4754-4764

Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH (1998) New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. Genes Dev **12:** 34-44

Laloi C, Mestres-Ortega D, Marco Y, Meyer Y, Reichheld JP (2004) The Arabidopsis cytosolic thioredoxin h5 gene induction by oxidative stress and its W-box-mediated response to pathogen elicitor. Plant Physiol **134:** 1006-1016

Lander ES *et al.* (2001) Initial sequencing and analysis of the human genome. Nature **409:** 860-921

Larkin JC, Oppenheimer DG, Pollock S, Marks MD (1993) Arabidopsis GLABROUS1 Gene Requires Downstream Sequences for Function. Plant Cell **5:** 1739-1748

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science **262:** 208-214

Leach KM, Nightingale K, Igarashi K, Levings PP, Engel JD, Becker PB, Bungert J (2001) Reconstitution of human beta-globin locus control region hypersensitive sites in the absence of chromatin assembly. Mol Cell Biol **21:** 2629-2640

Lescot M, Dehais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouze P, Rombauts S (2002) PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res **30:** 325-327

Levy S, Hannenhalli S, Workman C (2001) Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics **17:** 871-877

Lewis BA, Kim TK, Orkin SH (2000) A downstream element in the human beta-globin promoter: evidence of extended sequence-specific transcription factor IID contacts. Proc Natl Acad Sci U S A **97:** 7172-7177

Lewis BA, Sims RJ, 3rd, Lane WS, Reinberg D (2005) Functional characterization of core promoter elements: DPE-specific transcription requires the protein kinase CK2 and the PC4 coactivator. Mol Cell **18:** 471-481

Li B, Xia Q, Lu C, Zhou Z, Xiang Z (2004) Analysis on frequency and density of microsatellites in coding sequences of several eukaryotic genomes. Genomics Proteomics Bioinformatics **2:** 24-31

Lim CY, Santoso B, Boulay T, Dong E, Ohler U, Kadonaga JT (2004) The MTE, a new core promoter element for transcription by RNA polymerase II. Genes Dev **18:** 1606-1617

Lo K, Smale ST (1996) Generality of a functional initiator consensus sequence. Gene **182:** 13-22

Loganantharaj R (2006) Discriminating TATA box from putative TATA boxes in plant genome. Int J Bioinform Res Appl **2:** 36-51

Long F, Liu H, Hahn C, Sumazin P, Zhang MQ, Zilberstein A (2004) Genome-wide prediction and analysis of function-specific transcription factor binding sites. In Silico Biol **4:** 395-410

Lu J, Luo L, Zhang Y (2008) Distance conservation of transcription regulatory motifs in human promoters. Comput Biol Chem **32:** 433-437

Lujambio A, Esteller M (2007) CpG island hypermethylation of tumor suppressor microRNAs in human cancer. Cell Cycle **6:** 1455-1459

Lurin C, Andres C, Aubourg S, Bellaoui M, Bitton F, Bruyere C, Caboche M, Debast C, Gualberto J, Hoffmann B, Lecharny A, Le Ret M, Martin-Magniette ML, Mireau H, Peeters N, Renou JP, Szurek B, Taconnat L, Small I (2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell **16:** 2089-2103

Manevski A, Bardet C, Tremousaygue D, Lescure B (1999) Characterization and properties of heteromeric plant protein complexes that interact with tef cis-acting elements in both RNA polymerase II-dependent promoters and rDNA spacer sequences. Mol Gen Genet **261:** 892-900

Manevski A, Bertoni G, Bardet C, Tremousaygue D, Lescure B (2000) In synergy with various cis-acting elements, plant insterstitial telomere motifs regulate gene expression in Arabidopsis root meristems. FEBS Lett **483:** 43-46

Mathis DJ, Chambon P (1981) The SV40 early region TATA box is required for accurate in vitro initiation of transcription. Nature **290:** 310-315

Maugis C, Celeux G, Martin-Magniette ML (2009) Variable selection for clustering with Gaussian mixture models. Biometrics **65:** 701-709

Mengeritsky G, Smith TF (1987) Recognition of characteristic patterns in sets of functionally equivalent DNA sequences. Comput Appl Biosci **3:** 223-227

Miele V, Vaillant C, d'Aubenton-Carafa Y, Thermes C, Grange T (2008) DNA physical properties determine nucleosome occupancy from yeast to fly. Nucleic Acids Res **36:** 3746-3756

Mingam A, Toffano-Nioche C, Brunaud V, Boudet N, Kreis M, Lecharny A (2004) DEAD-box RNA helicases in Arabidopsis thaliana: establishing a link between quantitative expression, gene structure and evolution of a family of genes. Plant Biotechnol J **2:** 401-415

Mitasiunaite I, Rigotti C, Schicklin S, Meyniel L, Boulicaut JF, Gandrillon O (2009) Extracting signature motifs from promoter sets of differentially expressed genes. In Silico Biol **9:** S17-39

Molina C, Grotewold E (2005) Genome wide analysis of Arabidopsis core promoters. BMC Genomics **6:** 25

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet **30:** 194-200

Morozova O, Hirst M, Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet **10:** 135-151

Moshonov S, Elfakess R, Golan-Mashiach M, Sinvani H, Dikstein R (2008) Links between core promoter and basic gene features influence gene expression. BMC Genomics **9:** 92

Muller CW (2001) Transcription factors: global and detailed views. Curr Opin Struct Biol **11:** 26-32

Muller F, Demeny MA, Tora L (2007) New problems in RNA polymerase II transcription initiation: matching the diversity of core promoters with a variety of promoter recognition factors. J Biol Chem **282:** 14685-14689

Nakamura M, Tsunoda T, Obokata J (2002) Photosynthesis nuclear genes generally lack TATA-boxes: a tobacco photosystem I gene responds to light through an initiator. Plant J **29:** 1-10

NC-IUB (1985) Nomenclature Committee of the International Union of Biochemistry (NC-IUB). Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Biochem J **229:** 281-286

Neuwald AF, Liu JS, Lawrence CE (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Sci **4:** 1618-1632

Nobrega MA, Pennacchio LA (2004) Comparative genomic analysis as a tool for biological discovery. J Physiol **554:** 31-39

Nobuta K, Vemaraju K, Meyers BC (2007) Methods for analysis of gene expression in plants using MPSS. Methods Mol Biol **406:** 387-408

Novina CD, Roy AL (1996) Core promoters and transcriptional control. Trends Genet **12:** 351-355

Ohler U, Liao GC, Niemann H, Rubin GM (2002) Computational analysis of core promoters in the Drosophila genome. Genome Biol **3:** RESEARCH0087

Oliveira EJ, Gomes Pádua J, Imaculada Zucchi M, Vencovsky R, Carneiro Vieira ML (2006) Origin, evolution and genome distribution of microsatellites. Genetics and Molecular Biology **29:** 294-307

Orlando V (2000) Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. Trends Biochem Sci **25:** 99-104

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis J, Haas B, Wortman J, Buell CR (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res **35:** D883-887

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A **96:** 2896-2901

Papp B, Pal C, Hurst LD (2003) Evolution of cis-regulatory elements in duplicated genes of yeast. Trends Genet **19:** 417-422

Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet **10:** 669-680

Patikoglou GA, Kim JL, Sun L, Yang SH, Kodadek T, Burley SK (1999) TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. Genes Dev **13:** 3217-3230

Peng JC, Karpen GH (2008) Epigenetic regulation of heterochromatic DNA stability. Curr Opin Genet Dev **18:** 204-211

Ponjavic J, Lenhard B, Kai C, Kawai J, Carninci P, Hayashizaki Y, Sandelin A (2006) Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. Genome Biol **7:** R78

Powell LM, Wallis SC, Pease RJ, Edwards YH, Knott TJ, Scott J (1987) A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell **50:** 831-840

Rabinovich A, Jin VX, Rabinovich R, Xu X, Farnham PJ (2008) E2F in vivo binding specificity: comparison of consensus versus nonconsensus binding sites. Genome Res **18:** 1763-1777

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA (2000) Genome-wide location and function of DNA binding proteins. Science **290:** 2306-2309

Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, Yu G (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science **290:** 2105-2110

Rivals E, Bruyere C, Toffano-Nioche C, Lecharny A (2006) Formation of the Arabidopsis pentatricopeptide repeat family. Plant Physiol **141:** 825-839

Robinson SJ, Cram DJ, Lewis CT, Parkin IA (2004) Maximizing the efficacy of SAGE analysis identifies novel transcripts in Arabidopsis. Plant Physiol **136:** 3223-3233

Robinson SJ, Parkin IA (2008) Differential SAGE analysis in Arabidopsis uncovers increased transcriptome complexity in response to low temperature. BMC Genomics **9:** 434

Rombauts S, Florquin K, Lescot M, Marchal K, Rouze P, van de Peer Y (2003) Computational approaches to identify promoters and cis-regulatory elements in plant genomes. Plant Physiol **132:** 1162-1176

Rushton PJ, Torres JT, Parniske M, Wernert P, Hahlbrock K, Somssich IE (1996) Interaction of elicitor-induced DNA-binding proteins with elicitor response elements in the promoters of parsley PR1 genes. Embo J **15:** 5690-5700

Samson F, Brunaud V, Duchene S, De Oliveira Y, Caboche M, Lecharny A, Aubourg S (2004) FLAGdb++: a database for the functional analysis of the Arabidopsis genome. Nucleic Acids Res **32:** D347-350

Sandelin A, Wasserman WW, Lenhard B (2004) ConSite: web-based prediction of regulatory elements using cross-species comparison. Nucleic Acids Res **32:** W249-252

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (1977) Nucleotide sequence of bacteriophage phi X174 DNA. Nature **265:** 687-695

Schbath S (1995) Étude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquence exceptionnelle dans les séquences d'ADN. *In*, Paris V, Université René Descartes

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:** 467-470

Schiex T, Moisan A, Rouzé P (2001) EuGène: A eucaryotic gene finder that combines several sources of evidence. Lect Notes Comput Sci **2066:** 111-125

Schultheiss SJ, Busch W, Lohmann JU, Kohlbacher O, Ratsch G (2009) KIRMES: kernel-based identification of regulatory modules in euchromatic sequences. Bioinformatics **25:** 2126-2133

Sclep G, Allemeersch J, Liechti R, De Meyer B, Beynon J, Bhalerao R, Moreau Y, Nietfeld W, Renou JP, Reymond P, Kuiper MT, Hilson P (2007) CATMA, a comprehensive genome-scale resource for silencing and transcript profiling of Arabidopsis genes. BMC Bioinformatics **8:** 400

Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J (2006) A genomic code for nucleosome positioning. Nature **442:** 772-778

Seki M, Satou M, Sakurai T, Akiyama K, Iida K, Ishida J, Nakajima M, Enju A, Narusaka M, Fujita M, Oono Y, Kamei A, Yamaguchi-Shinozaki K, Shinozaki K (2004) RIKEN Arabidopsis full-length (RAFL) cDNA and its applications for expression profiling under abiotic stress conditions. J Exp Bot **55:** 213-223

Shahmuradov IA, Gammerman AJ, Hancock JM, Bramley PM, Solovyev VV (2003) PlantProm: a database of plant promoter sequences. Nucleic Acids Res **31:** 114-117

Shi W, Zhou W (2006) Frequency distribution of TATA Box and extension sequences on human promoters. BMC Bioinformatics **7 Suppl 4:** S2

Shinozaki K, Yamaguchi-Shinozaki K (2000) Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. Curr Opin Plant Biol **3:** 217-223

Sieburth LE, Meyerowitz EM (1997) Molecular dissection of the AGAMOUS control region shows that cis elements for spatial regulation are located intragenically. Plant Cell **9:** 355-365

Singer VL, Wobbe CR, Struhl K (1990) A wide variety of DNA sequences can functionally replace a yeast TATA element for transcriptional activation. Genes Dev **4:** 636-645

Smale ST (2001) Core promoters: active contributors to combinatorial gene regulation. Genes Dev **15:** 2503-2508

Smale ST, Kadonaga JT (2003) The RNA polymerase II core promoter. Annu Rev Biochem **72:** 449-479

Sonenberg N, Hinnebusch AG (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell **136:** 731-745

Struhl K (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol **14:** 103-105

Sumiyama K, Kim CB, Ruddle FH (2001) An efficient cis-element discovery method using multiple sequence comparisons based on evolutionary relationships. Genomics **71:** 260-262

Sutoh K, Yamauchi D (2003) Two cis-acting elements necessary and sufficient for gibberellin-upregulated proteinase expression in rice seeds. Plant J **34:** 635-645

Svozil D, Kalina J, Omelka M, Schneider B (2008) DNA conformations and their sequence preferences. Nucleic Acids Res **36:** 3690-3706

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res **36:** D1009-1014

Tabach Y, Brosh R, Buganim Y, Reiner A, Zuk O, Yitzhaky A, Koudritsky M, Rotter V, Domany E (2007) Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site. PLoS ONE **2:** e807

Tanaka T, Koyanagi KO, Itoh T (2009) Highly diversified molecular evolution of downstream transcription start sites in rice and Arabidopsis. Plant Physiol **149:** 1316-1324

Tatarinova T, Brover V, Troukhan M, Alexandrov N (2003) Skew in CG content near the transcription start site in Arabidopsis thaliana. Bioinformatics **19 Suppl 1:** i313-314

Thareau V, Dehais P, Serizet C, Hilson P, Rouze P, Aubourg S (2003) Automatic design of gene-specific sequence tags for genome-wide functional studies. Bioinformatics **19:** 2191-2198

Thastrom A, Bingham LM, Widom J (2004) Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. J Mol Biol **338:** 695-709

Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics **17:** 1113-1122

Thomas-Chollier M, Sand O, Turatsinze JV, Janky R, Defrance M, Vervisch E, Brohee S, van Helden J (2008) RSAT: regulatory sequence analysis tools. Nucleic Acids Res **36:** W119-127

Tompa M (2001) Identifying functional elements by comparative DNA sequence analysis. Genome Res **11:** 1143-1144

Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol **23:** 137-144

Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res **10:** 967-981

Trinklein ND, Aldred SJ, Saldanha AJ, Myers RM (2003) Identification and functional analysis of human transcriptional promoters. Genome Res **13:** 308-312

Tsai FT, Sigler PB (2000) Structural basis of preinitiation complex assembly on human pol II promoters. Embo J **19:** 25-36

Turatsinze JV, Thomas-Chollier M, Defrance M, van Helden J (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nat Protoc **3:** 1578-1588

van Driel R, Fransz PF, Verschure PJ (2003) The eukaryotic genome: a system regulated at different hierarchical levels. J Cell Sci **116:** 4067-4075

van Helden J (2003) Regulatory sequence analysis tools. Nucleic Acids Res **31:** 3593-3596

van Helden J, Andre B, Collado-Vides J (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol **281:** 827-842

van Helden J, del Olmo M, Perez-Ortin JE (2000) Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. Nucleic Acids Res **28:** 1000-1010

Vandepoele K, Casneuf T, Van de Peer Y (2006) Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. Genome Biol **7:** R103

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science **270:** 484-487

Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, Jr., Hieter P, Vogelstein B, Kinzler KW (1997) Characterization of the yeast transcriptome. Cell **88:** 243-251

Vyas P, Vickers MA, Simmons DL, Ayyub H, Craddock CF, Higgs DR (1992) Cis-acting sequences regulating expression of the human alpha-globin cluster lie within constitutively open chromatin. Cell **69:** 781-793

Walther D, Brunnemann R, Selbig J (2007) The regulatory code for transcriptional response diversity and its relation to genome structural properties in *A. thaliana*. PLoS Genet **3:** e11

Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE (2000) Human-mouse genome comparisons to locate regulatory sites. Nat Genet **26:** 225-228

Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet **5:** 276-287

White JA, Todd J, Newman T, Focks N, Girke T, de Ilarduya OM, Jaworski JG, Ohlrogge JB, Benning C (2000) A new set of Arabidopsis expressed sequence tags from developing seeds. The metabolic pathway from carbohydrates to seed oil. Plant Physiol **124:** 1582-1594

Whitehouse I, Tsukiyama T (2009) Opening windows to the genome. Cell **137:** 400-402

Wingender E (2008) The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Brief Bioinform **9:** 326-332

Wingender E, Dietze P, Karas H, Knuppel R (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res **24:** 238-241

Wolfe KH, Gouy M, Yang YW, Sharp PM, Li WH (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci U S A **86:** 6201-6205

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE, Elgar G (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol **3:** e7

Xu G, Goodridge AG (1998) A CT repeat in the promoter of the chicken malic enzyme gene is essential for function at an alternative transcription start site. Arch Biochem Biophys **358:** 83-91

Xue W, Wang J, Shen Z, Zhu H (2004) Enrichment of transcriptional regulatory sites in non-coding genomic region. Bioinformatics **20:** 569-575

Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, Pham P, Cheuk R, Karlin-Newmann G, Liu SX, Lam B, Sakano H, Wu T, Yu G, Miranda M, Quach HL, Tripp M, Chang CH, Lee JM, Toriumi M, Chan MM, Tang CC, Onodera CS, Deng JM, Akiyama K, Ansari Y, Arakawa T, Banh J, Banno F, Bowser L, Brooks S, Carninci P, Chao Q, Choy N, Enju A, Goldsmith AD, Gurjal M, Hansen NF, Hayashizaki Y, Johnson-Hopson C, Hsuan VW, Iida K, Karnes M, Khan S, Koesema E, Ishida J, Jiang PX, Jones T, Kawai J, Kamiya A, Meyers C, Nakajima M, Narusaka M, Seki M, Sakurai T, Satou M, Tamse R, Vaysberg M, Wallender EK, Wong C, Yamamura Y, Yuan S, Shinozaki K, Davis RW, Theologis A, Ecker JR (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. Science **302:** 842-846

Yamamoto K, Yoshida H, Kokame K, Kaufman RJ, Mori K (2004) Differential contributions of ATF6 and XBP1 to the activation of endoplasmic reticulum stress-responsive cis-acting elements ERSE, UPRE and ERSE-II. J Biochem **136:** 343-350

Yamamoto YY, Ichida H, Abe T, Suzuki Y, Sugano S, Obokata J (2007) Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. Nucleic Acids Res **35:** 6219-6226

Yamamoto YY, Ichida H, Matsui M, Obokata J, Sakurai T, Satou M, Seki M, Shinozaki K, Abe T (2007) Identification of plant promoter constituents by analysis of local distribution of short sequences. BMC Genomics **8:** 67

Yamamoto YY, Yoshitsugu T, Sakurai T, Seki M, Shinozaki K, Obokata J (2009) Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis. Plant J

Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E (2007) Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene **389:** 52-65

Yuh CH, Bolouri H, Davidson EH (2001) Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. Development **128:** 617-629

Zhang L, Zuo K, Zhang F, Cao Y, Wang J, Zhang Y, Sun X, Tang K (2006) Conservation of noncoding microsatellites in plants: implication for gene regulation. BMC Genomics **7:** 323

Annexes

I. Code IU-PAC

Symbole	Base	Origine de la désignation
A	A	Adénine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
M	A ou C	aMine
R	A ou G	puRine
W	A ou T	Faibles interactions (Weak)
S	C ou G	Fortes interactions (Strong)
Y	C ou T	pYrimidine
K	G ou T	Cétone (Keto)
B	C, G ou T	Tout sauf A - B suit A dans l'alphabet
D	A, G ou T	Tous sauf C - D suit C dans l'alphabet
H	A, C ou T	Tous sauf G – H suit G dans l'alphabet
V	A, C ou G	Tous sauf T - V suit U (l'uracile) qui suit T
N	A, C, G ou T	N'importe quelle base (aNy)

(NC-IUB, 1985)

II. Listes complètes des PLMs contenus dans les différentes régions

Légende pour chaque tableau A à D :

Colonne 1: Rang du PLM en fonction de son score

Colonne 2: Séquence du PLM

Colonne 3: Taille de la fenêtre glissante utilisée pour représenter la distribution

Colonne 4: Score du PLM: le SMS

Colonne 5: Position préférentielle

Colonne 6: Bornes de la fenêtre fonctionnelle

Colonne 7: Largeur du pic c'est-à-dire de la fenêtre fonctionnelle

Colonne 8: Nombre et pourcentage de promoteurs contenant le PLM dans la région fonctionnelle

Colonne 9 : Caractérisation du PLM qui peut ou non être un leader

Les quatre fichiers correspondant sont disponibles sur le CD-ROM.

A. PLMs contenus dans la région I

ftp://urgv.evry.inra.fr/Publications/AnnexeIIA.pdf

B. PLMs contenus dans la région II

ftp://urgv.evry.inra.fr/Publications/AnnexeIIB.pdf

C. PLMs contenus dans la région III

ftp://urgv.evry.inra.fr/Publications/AnnexeIIC.pdf

D. PLMs contenus dans la région IV

ftp://urgv.evry.inra.fr/Publications/AnnexeIID.pdf

III. Séquences et scores des PLMs non identifiés par R’MES

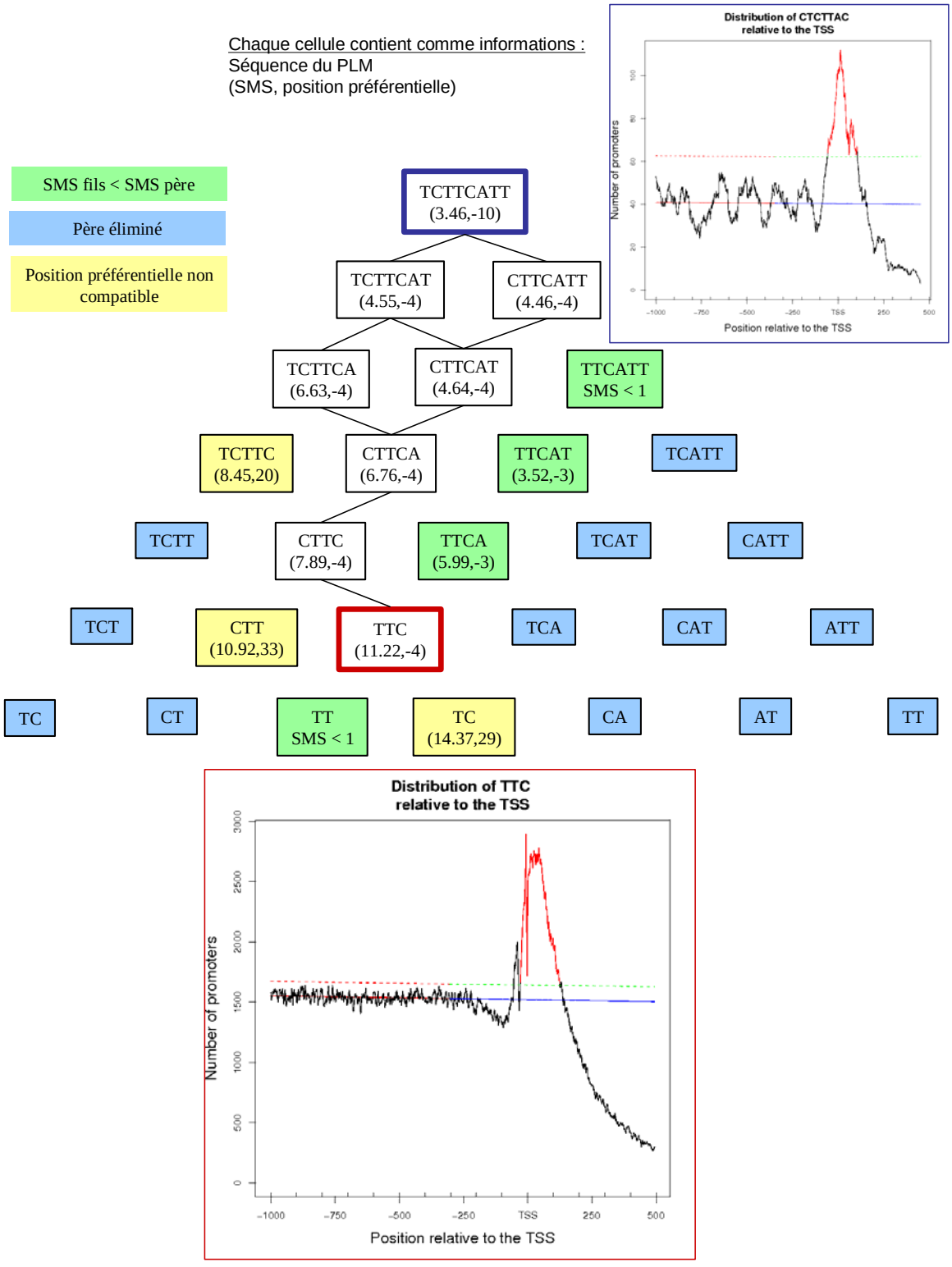
Liste des motifs identifiés comme ayant une position préférentielle par l’approche PLM et non identifiés comme exceptionnels par R’MES.

Les motifs sont listés par ordre alphabétique.

Séquence des motifs	Score approche R'MES	Score approche PLM	Séquence des motifs	Score approche R'MES	Score approche PLM
AAAACG	0,5905	4,66	CCAAAA	0,0478	4,86
AAAGCT	-2,9147	4,72	CCACTC	-1,3861	4,03
AAAGGC	-1,4221	3,51	CCACTT	0,8122	3,11
AAAGTC	2,5965	3,04	CCATCT	-1,6142	4,82
AAATAA	2,9929	3,01	CCATTA	-1,2892	3,49
AAATAG	-1,8635	3,69	CCCAAA	-1,9483	4,73
AACACA	1,7175	3,74	CCCACA	0,6238	3,79
AACCTC	0,763	4,01	CCCACC	-0,2654	5,04
AACGCC	2,1556	3,19	CCCACG	-2,9615	3,1
AACGGC	0,5273	4,06	CCCATC	-1,8362	3,35
AACGGT	0,2625	3,17	CCCCCA	1,562	3,6
AACTCC	-0,5849	3,42	CCCCGA	-0,7857	3,13
AACTTC	-2,331	3,18	CCCCTC	2,2977	4,19
AAGCCT	-2,4186	3,49	CCCGAC	0,2566	4,76
AAGCTC	-2,0804	4,29	CCCGCC	1,5591	3,51
AAGGCC	2,9834	5,06	CCCGGT	-2,8541	3,97
AAGGGC	1,0486	4,22	CCCGTC	-1,3346	3,44
AAGGGG	2,3607	3,15	CCCTCC	-1,217	5,89
AATACC	0,9753	3,32	CCCTCT	2,2259	5,85
AATCTC	1,7327	4,7	CCCTTA	0,2525	3,34
ACACAA	-0,3228	3,44	CCGAAA	-2,9583	3,14
ACACGC	-1,7454	3,37	CCGATT	-2,046	3,76
ACACGT	2,8716	4,63	CCGCCG	-0,6128	3,77
ACACTC	-1,1927	3,46	CCGCGT	-2,2775	3,24
ACCCAA	2,3276	3,29	CCGCTC	-0,8159	3,05
ACCGAC	-1,4535	3,1	CCGGTC	0,2055	3,07
ACCGGT	-2,7253	4,23	CCGTTC	2,8896	3,1
ACCTCT	-2,7572	3,51	CCTATA	1,3954	11,26
ACGTCG	-2,7307	4,28	CCTCAA	2,6549	3,59
ACTCTG	-1,3045	3,17	CCTCAC	0,1646	4,31
ACTTCC	1,7887	3,19	CCTCCC	-0,9356	3,68
AGCCCT	-2,4114	3,15	CCTCCG	-1,7985	3,21
AGGGGC	1,4868	3,56	CCTCGC	-0,8598	3,22
AGGGGT	0,0219	3,42	CCTCTA	-0,0045	4,11
AGGGTA	1,0885	4,03	CCTTCG	-0,3643	3,49
AGTCTC	-0,5984	3,83	CGACGC	0,3752	4,18
ATAAAA	-0,7674	5,73	CGATCT	-2,4097	3,62
ATACAC	0,5775	3,2	CGATTT	-0,9889	3,43
ATCGCC	0,8936	3,2	CGCCGC	2,8273	4,56
ATTCCC	-2,1772	3,12	CGCGCG	1,19	3,54
CAAACA	2,3937	3,3	CGCGTA	-2,9309	3,61
CAACGC	0,3773	3,4	CGCGTC	-2,5901	3,52
CAAGCC	1,2662	4,02	CGCTCC	0,9503	3,53
CAATCT	-1,8026	3,08	CGGTTA	2,0562	3,57
CACCAG	-1,6846	3,05	CGTCGC	1,5221	5,07
CACTCA	-2,0035	3,96	CGTCTT	2,8309	5,08
CACTCG	-2,0192	4,22	CGTGGA	1,82	3,4
CAGGCC	2,3369	3,29	CGTGTA	1,454	3,31

Séquence des motifs	Score approche R'MES	Score approche PLM	Séquence des motifs	Score approche R'MES	Score approche PLM
CGTTCC	1,6753	3,01	GTCTAT	-0,2236	3,6
CGTTCT	-0,6182	3,67	TAAAGC	2,0634	3,66
CTATTT	0,9139	3,03	TAACCG	2,3245	3,07
CTCAAC	-0,2164	3,04	TAAGGC	2,9368	3,23
CTCAGG	2,5919	3,19	TAATGG	1,8132	3,2
CTCATC	2,2789	6,07	TACCCT	0,6389	4,04
CTCATT	-1,7873	4,25	TATAAA	-1,7392	16,14
CTCCAT	0,8129	3,06	TCACTT	0,4776	3,14
CTCCCG	-1,7336	3,17	TCCACC	-0,0878	3,14
CTCCGT	-1,6004	5,05	TCCACG	0,2581	3,02
CTCCTT	-1,7555	5,47	TCCCCA	-0,2715	3,76
CTCGCC	-2,7611	5,15	TCCCCC	0,5004	3,03
CTCGTC	-0,2607	4,11	TCCGCC	0,7332	3,35
CTGCTC	2,0642	3,51	TCCGTC	1,7217	4,58
CTTATA	-2,3699	4,79	TCCTCT	1,4618	8,13
CTTCAA	-0,0965	3,04	TCCTTT	-2,9172	4,3
CTTCCC	0,3774	5,35	TCGCTC	0,8659	5,51
CTTGCT	-1,8044	3,07	TCGCTT	1,2581	4,74
CTTTCC	-0,1694	5,1	TCGTCG	0,583	4,15
GACGTG	1,1852	3,78	TCGTTC	2,7921	3,98
GATCTC	-0,4962	4,63	TCTCAC	1,2195	6,78
GCACGC	0,8449	3,33	TCTCAT	1,2085	6
GCCGTC	2,1463	5,31	TCTCCT	-1,0967	6,68
GCCTCT	-0,8718	3,9	TCTCGA	-1,3974	4,27
GCGTCT	2,4267	3,14	TCTCGT	-1,4663	3,33
GCTCGC	-1,7364	3,06	TCTGCA	-0,846	3,11
GGACCC	-1,8383	3,17	TCTGCC	-0,6111	3,07
GGCAAA	2,2774	3,48	TGCTTC	1,5633	3,55
GGCCGG	1,6609	3,4	TGTCTC	-1,0968	3,58
GGCCTT	2,6018	3,76	TTCACA	-0,2638	3,96
GGCGAT	-0,2224	3,95	TTCACC	0,8759	4,83
GGGCAA	0,3028	3,5	TTCACT	0,5724	4,42
GGGGGT	-2,9293	3,42	TTCCCT	-0,5537	4,62
GGGGTA	0,3178	4,19	TTCCGC	-0,9708	3,15
GGGGTT	0,2068	3,71	TTCGCC	0,2176	4,42
GGGTAA	-1,7046	3,06	TTTATA	2,2819	3,73
GGGTCC	-0,5069	3,19	TTTCCG	0,9676	3,09
GTCGCC	1,6167	4,38	TTTCGC	2,1296	3,65
GTCGTT	0,0838	3,8			

IV. Illustration d'une identification de leader



A partir de la racine, c'est-à-dire le PLM dont la séquence est la plus longue, on descend dans la pyramide tant que les PLM inclus sont caractérisés par une position préférentielle similaire et que leur SMS est supérieur à celui du PLM dans lequel ils sont inclus. Dans cet exemple, TTC est le leader de 7 autres PLM.

V. Poster présenté à une conférence internationale

Poster présenté à l'ECCB ISMB lors du Student Council et lors de la session poster de la conférence générale du 27 juin au 2 juillet 2009 à Stochholm.

TATAvariant identification, characterization and functional classification in plant genomes

Virginie Bernard[1], Véronique Brunaud[1] and Alain Lecharny[1,2]

[1] Unité de Recherche en Génomique Végétale (URGV), UMR INRA 1165 -CNRS 8114 -UEVE, 2 Rue Gaston Crémieux, 91057 Evry Cedex, France

[2] Université Paris-Sud, Institut de Biotechnologie des Plantes (IBP), UMR CNRS 8618 -UPS, Bâtiment 630, 91405 Orsay Cedex, France

http://www.versailles.inra.fr/urgv/bioinformatics.htm bernard@evry.inra.fr ~ http://v.bernard.bioinfo.free.fr

Abstract

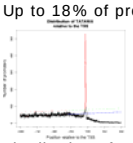
Taking advantage of the TATAbox topological constraints we identified TATAvariants sharing the same constraints and being conserved in *Arabidopsis thaliana* and *Oryza sativa*. This work led to TATAvariant characterization, distinguishing some motifs relative to the specific function, structure and expression of their related genes.

Background

The **TATAbox** and **TATAvariants** are regulatory elements involved in the Transcription Initiation Complex formation [1, 2]. Both have been conserved throughout evolution. They are observed **in a restricted region about 30 bases upstream of the Transcription Start Site** (TSS). The TATAbox is observed in less than 20% of eukaryotic genes.
Less is known about the TATAvariants and their functions. In this work, first, we **identified functional TATAvariants**. Second we **compared structural and functional features** of different sub-sets of TATAvariant-containing genes and TATAbox-containing genes.

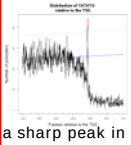
TATAvariant identification

We assumed that the **preferential distance relative to the TSS** and the **conservation throughout evolution** are both key features to **discriminate between functional and random motifs**. Thus we searched for motifs (i) observed in a strict location relative to the TSS [3] and (ii) being conserved in both *A. thaliana* and *O. sativa* core promoters with experimentally supported TSSs.
Distribution of the canonical TATAbox TATAWA is characterized by a sharp peak in the [-39, -26] area in both organims.
Up to 18% of promoters in both organisms contained the canonical TATAbox in this area.



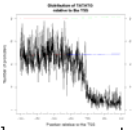
Are motifs sharing the TATAWA topological constraints
* functional motifs (TATAvariants)? or
* false-positive motifs due to overlap with a TATAbox sequence?
We distinguish them:

Distribution of TATATG in *A. thaliana* whole set:



a sharp peak in [-39, -26]
Functional motif?

and in the promoter set without TATAWA:

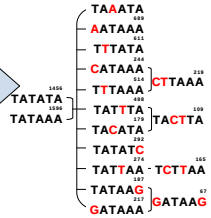


no more peak
→ **False-positive motif**

Conclusions:
In the whole set, the sharp peak is due to TATATA and TATATG overlap.

That is why **we searched for TATAvariants in plant promoter sets without TATAbox.**

We identified **11 conserved TATAvariants with one mismatch** and **4 with two mismatches**.



The number of promoters containing a TATAvariant depends on the number of substitutions with TATAWA.

Substitutions T→A or A→T are the main ones observed

C and G are avoided at central positions relative to the TATAWA sequence

The first T in TATAWA may be substituted by any other base

More than **28% of *A. thaliana* promoters contain a TATAvariant** at the TATAbox expected position

TATAvariant functional analyses

We built up 8 *A. thaliana* promoter sets characterized by the exclusive presence of one class of motif: only a TATAvariant or only a TATAbox. Each TATAvariant analysed has one mismatch:

Gene function [4]
Product of genes more or less directed to different cell components?

Gene more or less involved in different biological functions?

Gene structure [5]
Short or wide gene length?

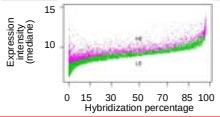
Gene, CDS, 5'UTR, introns length

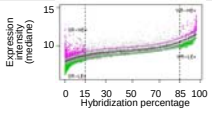
Gene expression [6]
High or Low Expression intensity? HE, LE
Wide or Short Range of hybridization? WR, SR

Extrem categories?

	All genes	TATAbox	AATAAA	CATAAA	GATAAA	TTTAAA	TTTATA	TATTAA	TATTTA	TATAAG
Number of genes	14927	1496	424	147	143	269	134	206	224	119
Cell wall	3.1%	5.8%	2.2%	0.8%	0.8%	2.7%	1.8%	3.6%	4.2%	2.9%
Chloroplast	15.4%	10.9%	16.0%	18.2%	22.0%	13.6%	11.7%	15.5%	19.0%	10.8%
Extracellular	2.2%	3.1%	0.8%	0.8%	2.4%	1.4%	2.7%	1.8%	3.2%	1.0%
DNA, RNA metabolism	1.4%	0.7%	1.0%	4.5%	1.6%	1.6%	0.8%	0.6%	1.0%	1.9%
Responses to stimulus	8.8%	11.5%	5.4%	7.5%	4.8%	7.3%	10.3%	8.8%	8.1%	12.0%
Responses to stress	8.8%	11.5%	4.7%	6.8%	8.0%	8.9%	10.3%	8.8%	10.2%	11.1%
Gene length	4000-	2250	1936	2368	2299	2286	2316	2029	2215	2359	2105
	2000-										
High Expression HE	39.2%	53.8%	34.6%	37.3%	33.9%	35.2%	37.0%	38.3%	48.2%	47.0%
Low Expression LE	60.8%	46.2%	65.4%	62.7%	66.1%	64.8%	63.0%	61.7%	51.8%	53.0%
SR-HE+	5.5%	8.5%	3.9%	6.4%	3.6%	4.2%	11.0%	5.2%	6.0%	7.0%
WR–LE+	3.4%	2.5%	6.0%	7.3%	4.5%	4.6%	2.0%	1.0%	3.9%	1.0%

Genes containing TATAvariant with a substitution downstream of the first base are genes poorly biased relative to all other genes or genes sharing biases with TATAbox-containing genes





Genes containing a TATAbox or a TATAvariant NATAWA have opposite biases on gene function, structure and expression

Conclusions & perspectives

With an *ab initio* approach selecting motifs with a strong preferential position, we identified **38% of *A. thaliana* genes containing either a TATAbox** or **a TATAvariant**. Substitutions leading from TATAbox to TATAvariants mainly concerned the bases on the sides.
We distinguish TATAvariant relative to the substitution position. One substitution in TATAWA may lead to wide functional, structural and expression variations. The most conspicuous observation was **opposite biases between TATAWA and TATAvariant NATAWA.** Genes containing **TATAvariant with a G at the last position share short structure and high expression level** with TATAbox-containing genes and are likewise involved in **specific biological fonctions**. Our results might be indicative of the **adaptation of the TATAbox expected region, which could promote changes in expression of functionally diversified genes with specific motif in the [-39, -26] area.**

We plan to combine this analysis to a **global proximal promoter analysis**. Our extensive description of *A. thaliana* promoters provides the necessary material to further analyse the question of the functional link between **regulatory element organisation** and **gene function**

[1] Patikoglou GA *et al.* (1999) Genes Dev 13(24): 3217
[2] Muller F *et al.* (2007) J Biol Chem 282(20): 14685
[3] Bernard V *et al.* (2006) JOBIM: 17
[4] Rhee SY *et al.* (2003) Nucleic Acids Res 31(1): 224
[5] Samson F *et al.* (2005) Nucleic Acids Res 33(Database issue): D347
[6] Gagnot S *et al.* (2008) Nucleic Acids Res 36(Database issue): D986

VI. Annexes des manuscrits des articles présentés dans ce travail de thèse

A. Manuscrit introduit page 70

ftp://urgv.evry.inra.fr/Publications/AnnexeVIIA.xls

B. Manuscrit introduit page 117

ftp://urgv.evry.inra.fr/Publications/AnnexeVIIB.xls

VII. Ressources Internet

AGRIS *Serveur dédié à l'analyse de la régulation des gènes chez A. thaliana*
http://*Arabidopsis*.med.ohio-state.edu/AtcisDB/bindingsites.html

CATdb Base de données transcriptome *A. thaliana*
http://urgv.evry.inra.fr/CATdb

CATMA Puces micro-array couvrant le génome d'A. thaliana
http://www.catma.org

ConSite Site dédié à la recherche de TFBS
http://asp.ii.uib.no:8090/cgi-bin/CONSITE/consite

DATF Base de données de TF d'*A. thaliana*
http://datf.cbi.pku.edu.cn

EPD Base de données de promoteurs eucaryotes
http://www.epd.isb-sib.ch

FLAGdb^{++} Base de données génomique dédiée à 4 plantes modèles
http://urgv.evry.inra.fr/FLAGdb

MotifSampler Site dédié à la recherche de TFBS
http://homes.esat.kuleuven.be/~thijs/BioDemo/MotifSampler.html

NAR - page des banques de données Nucleic Acids Research
http://217.169.56.209/nar/database/subcat/1/4

PLACE Base de données de TFBS de plantes
http://www.dna.affrc.go.jp/PLACE

PLANTCARE Base de données de TFBS de plantes
http://bioinformatics.psb.ugent.be/webtools/plantcare/html

PlantTFDB Base de données de TF de plantes
http://planttfdb.cbi.pku.edu.cn

R'MES Recherche de Mots Exceptionnels dans une Séquence
http://migale.jouy.inra.fr/outils/mig/rmes

RSAT Outil d'analyse des séquences régulatrices
http://rsat.ulb.ac.be/rsat

TAIR Ressource d'information concernant *A. thaliana*
http://www.*Arabidopsis*.org/news/news.jsp

TransFac Base de données de TF et leurs TFBS
http://www.gene-regulation.com/pub/databases.html

URGV Unité de Recherche en Génomique Végétale
http://www.versailles.inra.fr/urgv

Résumé

Les protéines sont synthétisées via la transcription de l'ADN en ARN, puis la traduction de l'ARN en protéine. Les principaux mécanismes impliqués lors de la transcription et de la traduction sont aujourd'hui assez bien connus. En revanche, leurs régulations sont encore mal identifiées.

Les sites de fixation des facteurs de transcription ou éléments régulateurs présents dans les promoteurs des gènes sont impliqués dans ce processus. Leur présence conduit à la régulation de l'expression des gènes dans des conditions particulières, en réponse à un stimulus ou dans un tissu spécifique. Une meilleure connaissance de l'architecture des promoteurs et donc de la régulation de l'expression des gènes est aujourd'hui accessible à l'échelle génomique, en exploitant l'annotation globale des génomes, les collections de transcrits disponibles et les données transcriptomiques. Néanmoins, les analyses actuelles ne considèrent souvent que l'identification des éléments régulateurs sans prendre en compte leur organisation dans les promoteurs, ni la fonction des gènes associés.

Certains éléments régulateurs peuvent être conservés à une position préférentielle au cours de l'évolution. L'hypothèse de ce travail est que des éléments régulateurs peuvent être prédits en recherchant des séquences sur-représentées à une position précise du promoteur. De plus, des éléments régulateurs impliqués dans une même voie de régulation de l'expression sont attendus simultanément dans les promoteurs. L'identification d'annotations fonctionnelles communes au sein de gènes ayant une même organisation des éléments régulateurs dans leurs promoteurs pourrait contribuer à l'annotation fonctionnelle de ces éléments régulateurs.

Chez *A. thaliana*, génome modèle chez les plantes, nous avons mis au point une approche pour caractériser des éléments régulateurs observés dans une région préférentielle. Ce travail a permis de proposer une cartographie des promoteurs d'*A. thaliana* en identifiant 5105 motifs caractérisés par une sur-représentation locale dans les promoteurs centraux et proximaux. De plus, l'étude du promoteur central, où est attendue la boîte TATA, élément régulateur observé dans tous les règnes, a été approfondie. Une liste de 15 variants fonctionnels de la boîte TATA a été identifiée, ainsi qu'une nouvelle classe d'éléments régulateurs qui sont caractérisés par des mêmes contraintes topologiques que la boîte TATA : les motifs-TC. Ils sont conservés chez *A. thaliana* et *O. sativa*, mais absents dans les promoteurs des gènes de mammifères. Les 18% de gènes *d'A. thaliana* contenant un motif-TC ont une propension à être exprimés dans des conditions expérimentales spécifiques. L'ensemble des résultats a permis de supposer que ce nouvel élément participe à la régulation de l'expression des gènes chez les plantes supérieures. L'étude de la région du TSS où l'élément initiateur YR est attendu chez *A. thaliana* a mis en évidence une extension de chaque dinucléotide dans l'UTR 5', qui peut être due à la présence de TSS alternatifs dans une région restreinte. Enfin, des associations entre éléments régulateurs ont été mises en évidence, notamment la boîte TATA et l'Inr-CA, mais aussi les motifs-TC et d'autres éléments régulateurs particulièrement riches en bases C et G.

Ainsi, la recherche de caractéristiques fonctionnelles communes aux gènes possédant une même organisation d'éléments régulateurs pourra permettre de contribuer à l'annotation fonctionnelle des éléments régulateurs.