Institutional Repository - Research Portal
Dépôt Institutionnel - Portail de la Recherche
researchportal.unamur.be
University of Namur

THESIS / THÈSE

MASTER EN SCIENCES BIOLOGIQUES DES ORGANISMES ET ÉCOLOGIE

Etude des différentes méthodes permettant de retrouver des sites d'atterrissage pour facteur de transcription et élaboration d'une nouvelle méthode de recherche de sites d'atterrissage

Fisse, Jérôme

Award date:
2002

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 25. Apr. 2024

FACULTES UNIVERSITAIRES NOTRE-DAME DE LA PAIX
NAMUR

Faculté des Sciences

Etude des différentes méthodes permettant de retrouver des sites d'atterrissage pour facteur de transcription et élaboration d'une nouvelle méthode de recherche de sites d'atterrissage

Mémoire présenté pour l'obtention du grade de

licencié en Sciences biologiques

Jérôme FISSE

Août 2002

Facultés Universitaires Notre-Dame de la Paix
FACULTE DES SCIENCES
Secrétariat du Département de Biologie
Rue de Bruxelles 61 - 5000 NAMUR
Téléphone: + 32(0)81.72.44.18 - Téléfax: + 32(0)81.72.44.20
E-mail: joelle.jonet@fundp.ac.be - http://www.fundp.ac.be/fundp.html

Etude des différentes méthodes permettant de retrouver des sites d'atterrissage pour facteur de transcription et élaboration d'une nouvelle méthode de recherche de sites d'atterrissage

FISSE Jérôme

Résumé

Lors de l'initiation de la transcription, les facteurs de transcription viennent se fixer sur l'ADN pour réguler la transcription. Ce lieu de fixation s'appelle un site d'atterrissage pour facteur de transcription (S.A.F.T.), ce site a une séquence précise spécifique de son facteur de transcription. Ce travail présente l'analyse de différents *programmes (Oligo-analysis, Dyad-analysis, Motif sampler*) permettant de retrouver des S.A.F.T. ainsi qu'une méthode de comparaison de séquence amont de gènes orthologues. Enfin l'analyse de la distribution du S.A.F.T. de ctrA, nous a amené à mettre au point une nouvelle méthode de recherche de S.A.F.T..

Mémoire de licence en Sciences Biologiques
Août 2002
Promoteur: E. Depiereux
Co-Promoteur: X. De Bolle

4ième primaire :
Monsieur Hadyns emmène toute la classe voir les dinosaures au musée royal des sciences naturelles de Bruxelles. C'est en voyant l'Iguanodon de Bernissart que je choisis d'étudier la Biologie. Merci à l'ingénieur Latinis d'avoir découvert l'Iguanodon en 1878.

4ième rénové :
Madame Bister nous initie aux théories de l'évolution, à la cellule et surtout à l'ADN. Merci madame Bister : votre passion communicative a aiguisé ma curiosité et merci aux autres professeurs qui enseignent avec la même passion.

4ième année d'étude aux facultés :
Voilà j'ai réussi les trois premières années, un grand merci à toutes les filles de ma classe qui m'ont prêté leurs cours pour compléter mes notes (Mélanie, Caroline, Aline, Cindy, Maud, Cathou, Gégé, Sophie, Marie, ...). Merci à tous mes compagnons de guindailles et particulièrement au groupe GRH.

4ième mois de rédaction du mémoire :
Merci au professeur Depiereux de m'avoir accueilli dans son laboratoire, d'avoir pris du temps pour corriger mon mémoire. Et merci d'avoir fait l'effort de retenir mon prénom (je ne m'appelle pas Pascal).
Merci au professeur Vandenhaute pour ses sages conseils durant mon mémoire.
Merci à Xavier De Bolle pour avoir trouvé du temps pour répondre à mes questions. Merci aussi pour ton énergie communicative et pour ta franchise.
Calou et Amélie pour ce surnom débile (Fistule) que vous êtes, heureusement, les seuls à utiliser. Je ne vous remercie pas. Par contre, je remercie Calou et Etienne pour m'avoir humblement expliqué comment faire une bonne blague par E-mail. J'en profite pour dire à Amélie que je n'ai pas le numéro de téléphone des stripteaseurs Canadiens.
Merci à Garçon pour m'avoir aidé moralement du début à la fin (du moins j'espère) de mon mémoire. Et merci au mage Artémis pour sa patience envers Arthuro.

Merci1000 à Benjamin et Christophe pour leurs corrections, sans vous je n'y serais pas arrivé. Merci pour vos soirées sacrifiées aux corrections (merci aussi à Monique). Merci aussi pour toutes ces discussions scientifiques (les guignols de l'info, Georges Bush est vraiment un ..., la fabrication des saucisses zwan, ...).

Merci à mon tuteur pour son extrême discrétion. Tu m'as appris à être autodidacte.
Merci à mon père pour tous ces trajets allez retour facs-maison et aussi pour m'avoir encouragé pendant ces quatre ans. Merci à ma sœur pour cette soirée au resto au milieu de mon blocus de première candi et toutes les intentions de ce genre.
Et je termine par des remerciements en vrac : Nathalie pour son humour et sa franchise ; Nadia pour sa gentillesse, ses conseils sur "comment ranger un bureau" et aussi pour son écoute attentive de mes présentations ; Valérie, tu es tellement altruiste qu'il faudrait créer un

11ième commandement : "TU SERAS AUSSI SYMPA QUE VALERIE" ; la machine à café pour son travail journalier ; Aiko pour ses conseils en stat, pour son bureau et sa sympathie ; à l'équipe du El charro : Michou, Micky Mike, Xanax, Nadine, André, Jacot, Fred, ... et la bouteille de Ricard ; mon petit soleil Guatémaltèque pour m'avoir dégoutté d'Indochine, pour son rire, son sourire et sa joie communicative ; Minou pour sa faculté hors norme à rire d'elle-même, pour ses journées kitch et pour tout ce qui la rend si géniale ; mon petit "King Fish" pour tout ce que tu sais (et que Chouchou ne sais pas).

Abréviations

ORF	*Open reading frame* ou phase ouverte de lecture
CDS	*Coding sequence* ou séquence codante
pCDS	*predicted coding sequence* ou séquence codante prédite
S.A.F.T.	Site d'atterrissage pour facteur de transcription
SD-Box	*Shine-Dalgarno box ou boîte de Shine-Dalgarno*
RBS	*Ribosome biding site* ou site de fixation du ribosome
ARN	Acide ribonucléique
ADN	Acide désoxyribonucléique
RNAP	ARN polymérase
HTH	*helix-turn-helix*
HPK	histidine protéine kinase
RR	régulateur de réponse
RSA-tools	*regulatory sequence analysis-tools ou outils d'analyse de séquences régulatrices*
HMM	Hidden Markov Model ou modèle caché de Markov
GRH	Gestion des ressources humaines
dl	Degré de liberté
F.T.	Facteur de transcription
S.A.F.T.	Site d'atterrissage pour facteur de transcription
pb	paires de base

Introduction

1 Introduction

1.1 Organisation du génome bactérien

1.1.1 Définitions relatives au génome

Beaucoup de définitions relatives au génome sont assez ambiguës. Le rappel des différentes définitions permet de définir clairement les éléments qui composent le génome. Tous les éléments décrit ci-dessous sont visibles sur la figure 1.

Le codon stop représente le signal d'arrêt de la traduction, il existe sous les trois formes suivantes : TGA, TAA, TAG. L'espace compris entre deux codons stop s'appelle une ORF (*Open Reading Frame* ou phase ouverte de lecture).

Le site d'initiation de la traduction, aussi appellé le codon start, est présent sous trois formes : ATG, GTG, TTG. L'espace compris entre un codon start et un codon stop présents dans un même cadre de lecture (voir figure 2) et qui code pour une protéine se nomme CDS (*CoDing Sequence*). Lorsque cette CDS est prédite *in silico*, on parle de pCDS pour *predicted* CDS (séquence codante prédite).

Avant la CDS, on parle de partie amont ou *upstream* et après la CDS, de région aval ou *downstream*. La partie précédant la CDS s'appelle le promoteur. Cette région s'étend sur une longueur variable en fonction de la CDS étudiée. On y trouve différents éléments : (i) des sites d'atterrissage pour facteurs de transcription (S.A.F.T.) décrit au point 1.3 (ii) une *SD-BOX*, qui sera, sur l'ARN, le lieu de fixation du ribosome lors de la traduction (voir figure 3).

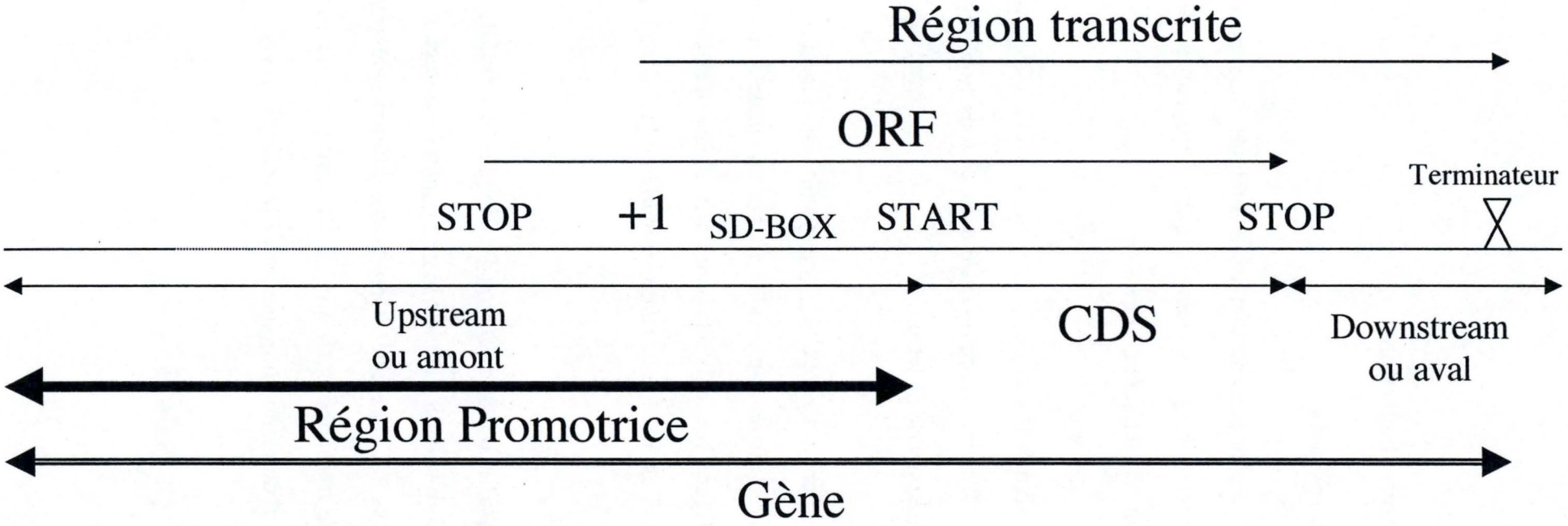

Figure 1 : description du gène procaryote isolé, de l 'ORF, de la CDS, ...La délimitation du codon stop de gauche est arbitraire, à titre d 'exemple.

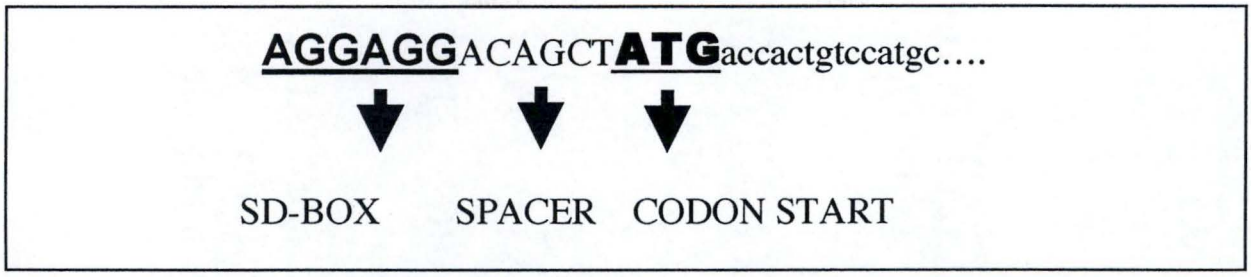

Figure 3 : Ce schéma représente la SD-BOX telle qu'on la trouve chez les bactéries. Elle est séparée du codon start par un *spacer* de taille variable.

(iii) Avant la SD-BOX, on trouve le site +1 qui est le site d'initiation de la transcription. La transcription se termine au terminateur (signal d'arrêt de la transcription présent après la CDS).

De nombreux facteurs de transcription peuvent se positionner dans le promoteur avant le site +1. Cette région précédant le site +1 s'appelle la région promotrice, la longueur de cette région varie d'une CDS à l'autre.

L'espace compris entre le site +1 et le terminateur est la région transcrite. Dans ce travail, nous définissons le gène comme l'espace compris entre le début de la région promotrice et le terminateur. Le site d'initiation de la traduction est le codon start.

Dans le génome, nous pouvons définir deux grands types de régions :

(i) Les régions codantes, qui possèdent une probabilité d'apparition de certains codons bien définie, qui correspond au biais des codons. La probabilité d'apparition des nucléotides est définie par le biais des codons.

(ii) Les régions non codantes où la probabilité d'apparition des nucléotides est pratiquement aléatoire car ces régions peuvent subir des mutations sans altérer la viabilité de l'organisme.

1.1.2 Les groupes de gènes

Les protéines ont une ou plusieurs activités dans la bactérie, appelée collectivement la fonction de la protéine. Un groupe de pCDSs codant pour une fonction qui résulte de l'action de chacune, ont tendance à rester groupées. Lorsque ce groupement se répète chez plusieurs

SANS CADRE DE LECTURE ATCCGTAGCTGATCGCGTAGATCGCGCGATAG

CADRE DE LECTURE 0 ATC CGT AGC TGA TCG CGT AGA TCG CGC GAT AG

ILE ARG SER STOP SER ARG ARG SER ARG ASP /

CADRE DE LECTURE 1 A TCC GTA GCT GAT CGC GTA GAT CGC GCG ATA G

/ SER VAL ALA ASP ARG VAL ASP ARG ALA ILE /

CADRE DE LECTURE 2 AT CCG TAG CTG ATC GCG TAG ATC GCG CGA TAG

/ PRO STOP LEU ILE ALA STOP ILE ALA ARG STOP

Figure 2 : Ce schéma représente le cadre de lecture. La première ligne représente une séquence d 'ADN sans cadre de lecture. Si on divise cette séquence en codon, nous pouvons commencer au premier nucléotide(cadre de lecture 0), au deuxième nucléotide (cadre de lecture 1) ou au troisième nucléotide (cadre de lecture 2). Chacun de ces cadres de lecture défini une séquence d 'acides aminés différente.

espèces proches d'un point de vue évolutif, on parle de cluster de gènes. Si l'ensemble des pCDSs est dans une même unité de transcription, on parle d'opéron.

Certains gènes sont homologues entre eux, c'est-à-dire qu'ils ont divergés à partir d'une même séquence originelle. Souvent, deux séquences homologues possèdent la même fonction. Si deux pCDSs possèdent une fonction semblable dans le même génome, on parle de pCDSs paralogues. Si deux pCDSs possèdent la même fonction dans deux génomes différents, on dit qu'elles sont orthologues.

Lorsque deux CDSs de fonction connue fusionnent en une CDS possédant les deux fonctions, on parle d'événement fusionnel. Une pCDS impliquée dans un événement fusionnel peut être « *as composites* », c'est-à-dire qu'elle fait partie d'un gène fusionné ou « *as components* » ce qui signifie que la pCDS est libre dans le génome.

1.2 La régulation transcriptionnelle chez les procaryotes

1.2.1 Introduction à la régulation transcriptionnelle

La transcription se définit comme étant la synthèse d'une molécule d'ARN complémentaire à un brin d'une molécule d'ADN. Cette molécule d'ARN sera traduite en une séquence d'acides aminés qui, en se repliant sur elle-même, donnera la plupart du temps naissance à une protéine fonctionnelle.
La bactérie peut réguler la production des protéines à différentes étapes, essentiellement lors de la transcription et de la traduction.

La transcription se déroule en trois étapes distinctes : l'initiation, l'élongation et la terminaison. Notre travail se focalisera sur la régulation de la transcription.

L'élément central de ces trois étapes est l'ARN polymérase (RNAP) (voir figure 4). L'ARN polymérase se compose de deux sous-unités α, d'une sous-unité β et d'une sous-unité β' (le core est donc $\alpha_2\beta\beta'$). Chaque sous-unité α possède un domaine responsable de la dimérisation des deux sous-unités α et de l'interaction avec les autres sous-unités de l'ARN polymérase dans sa partie N-terminale. Le domaine C-terminale de la sous unité α peut reconnaître l'ADN au niveau de certains promoteurs et peut interagir avec des régulateurs transcriptionnels, les activateurs par exemple. Les sous-unités β et β' sont responsables de l'activité catalytique de l'ARN polymérase (deHaseth *et al.*, 1998).

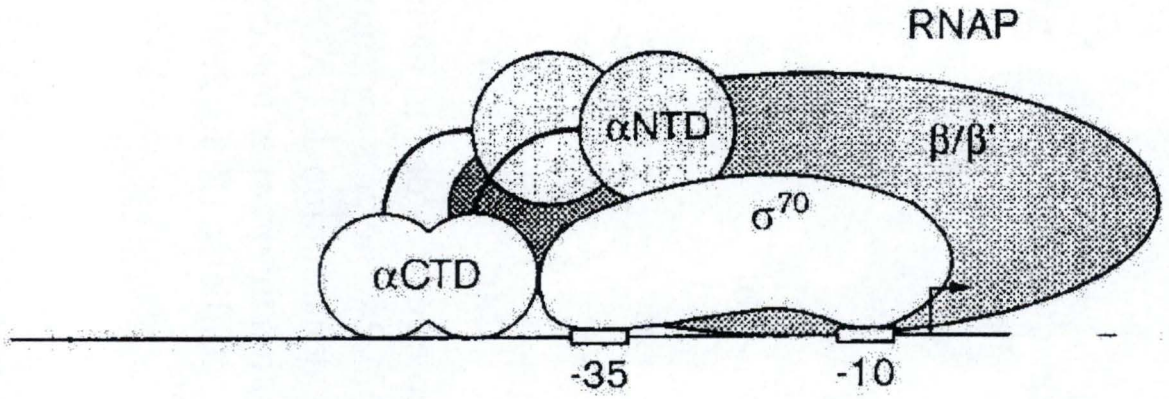

Figure 4 : Représentation de l'holoenzyme, c'est-à-dire la sous-unité σ (σ^{70})et le complexe $\alpha_2\beta\beta'$. Chaque sous-unité α se compose d'une sous-unité C-terminale (αCTD) et C-terminale (αNTD).

Lors de l'initiation, une sous-unité mobile appelée facteur σ (voir figure 4) vient se fixer sur les boîtes –35 et –10 (exception faite pour le facteur σ^{54}, voir chapitre 1.2.4). Lorsque ce complexe ARN polymérase-facteur σ est formé (=holoenzyme), l'enzyme se lie à l'ADN au site d'initiation (site +1) et commence à séparer les deux brins. Entre la fin de l'initiation et le début de l'élongation, le facteur σ est libéré.

La figure 5 montre les différentes étapes de l'initiation. A la première étape, l'holoenzyme se lie faiblement à l'ADN formant ainsi un complexe fermé (RPc1). On parle de complexe fermé car les brins d'ADN ne sont pas encore séparés. Différents changements doivent intervenir avant que l'holoenzyme quitte le promoteur. D'abord, la RNAP se lie plus fortement à

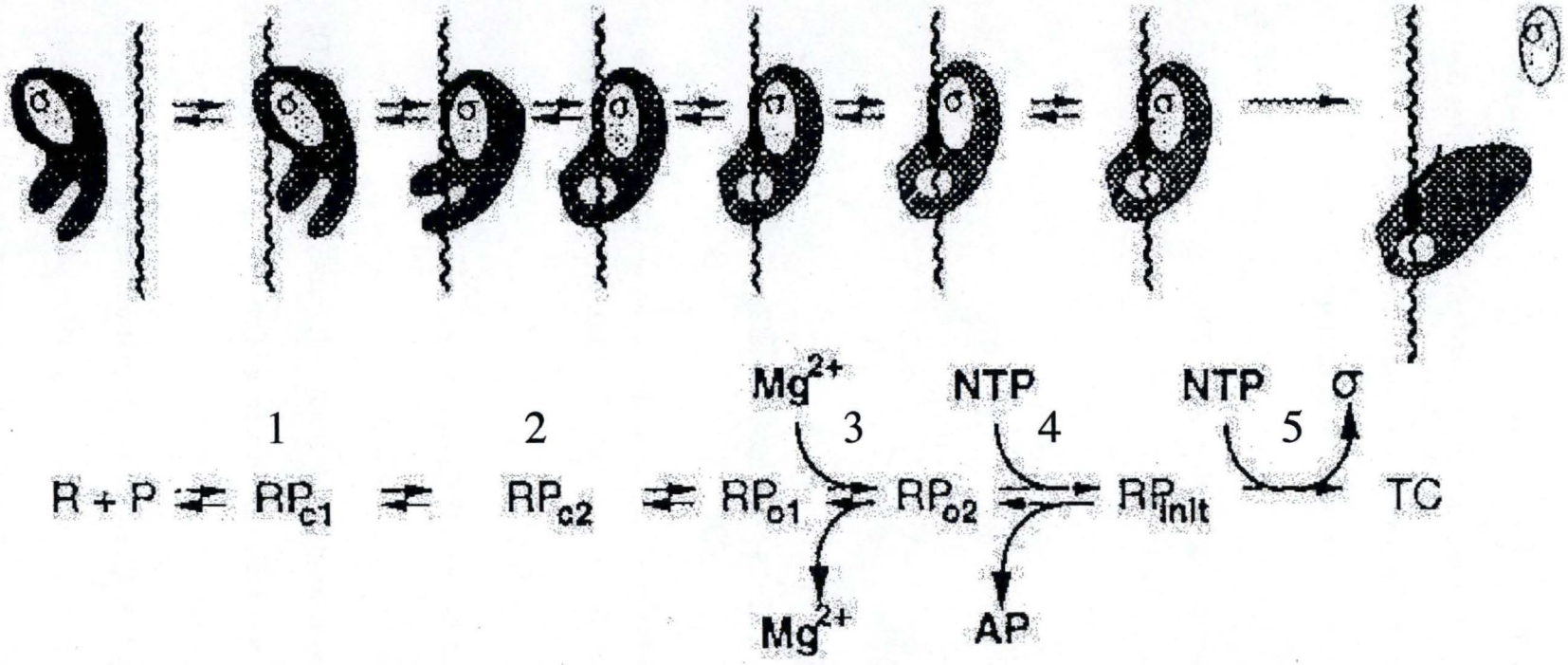

Figure 5 : Représentation des différentes étapes de l'initiation. 1) l'holoenzyme se lie faiblement à l'ADN (RPc1). 2) la RNAP se lie plus fortement à l'ADN (RPc2). 3) le complexe ouvert (Rpo) se forme et engendre la dénaturation de l'ADN. 4) le complexe d'initiation (Rpinit) est formé lors de l'incorporation du premier nucléotides 5) 7 à 12 nucléotides d'ARN sont polymérisées et la sous-unité σ est libérée (Rojo, 2001).

l'ADN, ce qu'on appelle le complexe intermédiaire (RPc2). Ensuite le complexe ouvert (Rpo) se forme et engendre la dénaturation de l'ADN à partir de 10 à 15 pb avant le site +1. La formation du complexe d'initiation (Rpinit) débute lors de l'incorporation du premier nucléotide tri phosphate au site +1. Lorsque 7 à 12 nucléotides d'ARN sont polymérisés, la sous-unité sigma est libérée (Rojo, 2001).

Lors de l'élongation, la RNAP ajoute des nucléotides à l'extrémité 3' de l'ARN en formation tout en avançant sur le brin codant de l'ADN.
Lors de la terminaison, l'enzyme rencontre un site de terminaison sur l'ADN. Il lâche l'ARN néoformé qui est en cours de traduction par les ribosomes.

L'initiation de la transcription est l'étape clé de la régulation chez les bactéries. Il existe deux grandes classes d'agents régulateurs à ce niveau : les facteurs de transcription et les facteurs σ.

1.2.2 Les facteurs de transcription (F.T.)

Les facteurs de transcription sont des protéines mobiles possédant un domaine de liaison à l'ADN. La grande majorité des F.T. chez les procaryotes ont un domaine de liaison de type HTH (helix-turn-helix). Ce domaine se compose généralement d'un segment de 20 acides aminés formant deux hélices α presque perpendiculaires l'une à l'autre, jointes par un plan β (voir figure 6) (Luscombe *et al.*, 2000). Les domaines HTH de ces F.T. reconnaissent souvent des dyades (deux parties conservées séparées d'une distance fixe). Le consensus de chaque domaine HTH a permis de classer les régulateurs transcriptionnels de *Brucella melitensis* en 19 familles. Nous décrivons ici une des 19 familles, le système à deux composants (Stock *et al.*, 2000).

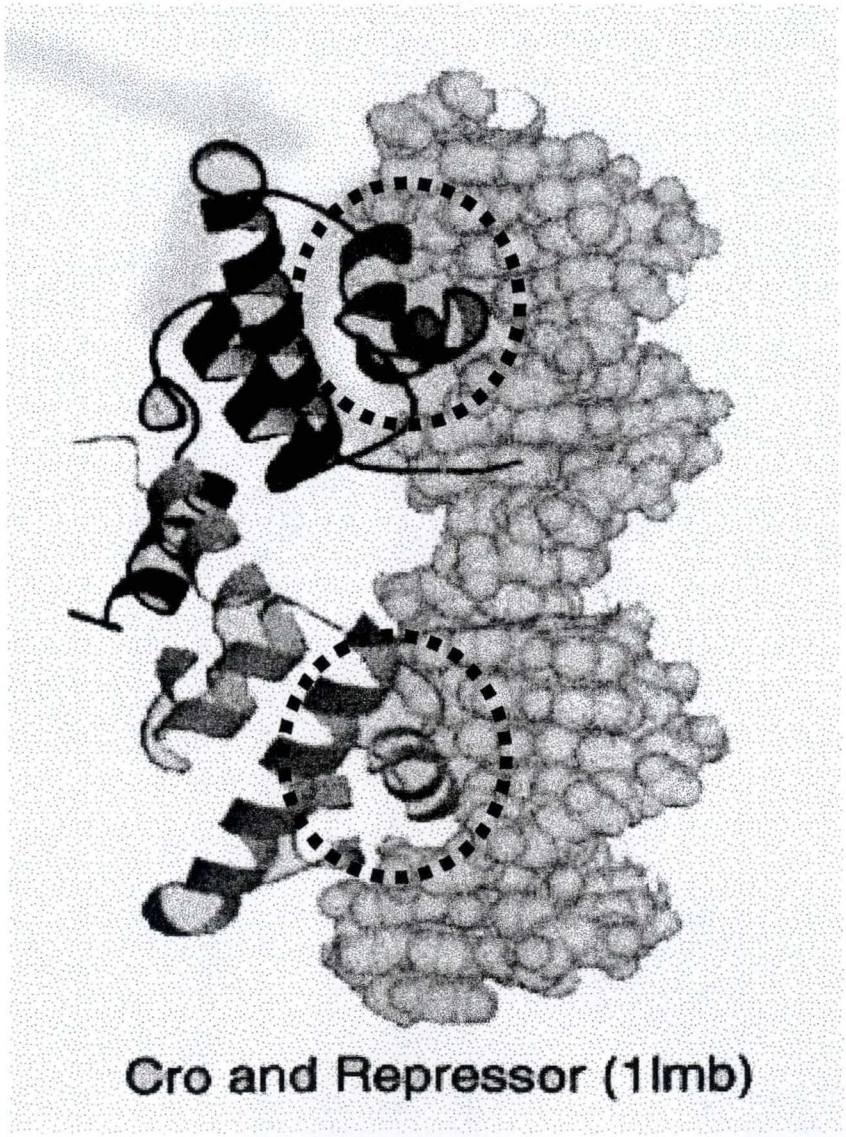

Figure 6 : Les deux domaines entourés par des pointillés sont les domaines de liaisons de type helix-turn-helix (HTH). Ce facteur de transcription se fixe à l 'ADN sur deux motifs séparés d'une distance fixe (Luscombe *et al.*, 2000).

Le système à deux composants, se compose d'une histidine protéine kinase (HPK), ou senseur, et d'un régulateur de réponse (RR). Lorsque le senseur perçoit un stimulus environnemental qui lui est propre, il autophosphoryle un résidu histidine spécifique. Le groupement phosphate est ensuite transféré de l'histidine à l'aspartate du régulateur de réponse, ce qui rend ce dernier actif et apte à se fixer à l'ADN (Stock, et al., 2000).

Les facteurs de transcriptions agissent aux différentes étapes de l'initiation en régulant la vitesse de passage d'un complexe à l'autre. On distingue deux grands mécanismes de régulation de la transcription : la répression et l'activation.

1.2.3 Activation et répression

La répression (Rojo, 2001) peut être effectuée par trois mécanismes distincts (voir figure 7):

a) Par encombrement stérique, en se liant sur le site de fixation de l'ARN polymérase.
b) Par inhibition de la transition du complexe fermé au complexe ouvert. Le répresseur se fixe soit en amont de l'ARN polymérase soit sur l'autre brin d'ADN symétriquement à l'ARN polymérase. Le lien entre le répresseur et l'ARN polymérase empêche la séparation des deux brins d'ADN.
c) Par blocage, le répresseur se place devant l'ARN polymérase et l'empêche d'avancer.

L'activation (Rhodius & Busby, 1998) peut être également effectuée par trois mécanismes distincts (voir figure 8):

a) L'activateur se lie à la partie C-terminale des deux sous-unités α
b) L'activateur se lie directement sur la région 4 du facteur σ (le point de liaison varie en fonction du facteur σ impliqué).

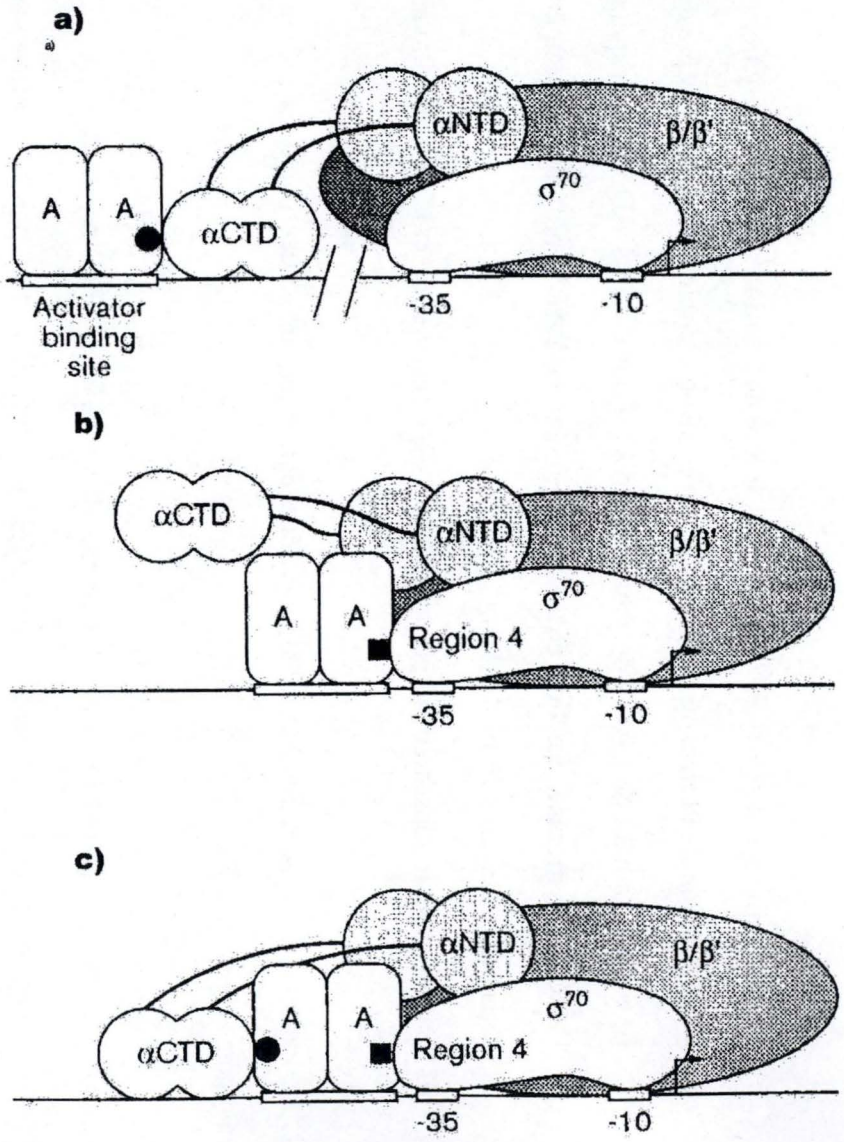

Figure 8 : Représentation des 3 types d'activation.

a) L'activateur se lie à la partie C-terminal des deux sous-unités α.

b) L'activateur se lie directement sur la région 4 du facteur σ (le point de liaison varie en fonction du facteur σ impliqué).

c) L'activateur applique les deux possibilités (Rhodius & Busby,1998).

c) Soit l'activateur applique les deux possibilités.

1.2.4 Les facteurs σ

Comme expliqué précédemment, les facteurs σ sont nécessaires lors de l'initiation de la transcription. Sans eux, l'ARN polymérase ne reconnaît pas les promoteurs. Ils sont donc les premiers agents de régulation à déterminer l'activation ou la non activation de la transcription. On définit deux grandes familles de facteurs sigma : les familles σ^{70} et σ^{54}.

Les facteurs σ de la famille σ^{70} se fixent sur les boîtes –35 et –10 chez *E. coli.* Ces facteurs σ peuvent être regroupés en trois groupes structuralement et fonctionnellement apparentés.

- Groupe 1 : facteurs σ principaux, il s'agit de σ^{70}. Il est responsable de l'initiation de la transcription de la plupart des gènes de croissance et des gènes *housekeeping* exprimés en phase exponentielle de la croissance. Le consensus des boîtes –35 et –10 est respectivement composé de TTGACA et TATAAT
- Groupe 2 : facteurs σ non-essentiels à la croissance exponentielle des cellules. Il s'agit par exemple du facteur σ^{S}, il est responsable de l'initiation de la transcription des gènes spécifiques de la phase stationnaire.
- Groupe 3 : facteurs σ alternatifs. Par exemple le facteur σ^{H} intervient lors de la transcription des gènes de choc thermique.

La famille σ^{54} est différente de la famille σ^{70}. Premièrement, elle ne contient qu'un seul membre : σ^{54}. Ce facteur se fixe aux positions –12 et –24 dans ses promoteurs cibles. Le complexe ARN polymérase-σ^{54} ne peut initier la transcription sans l'aide d'une protéine

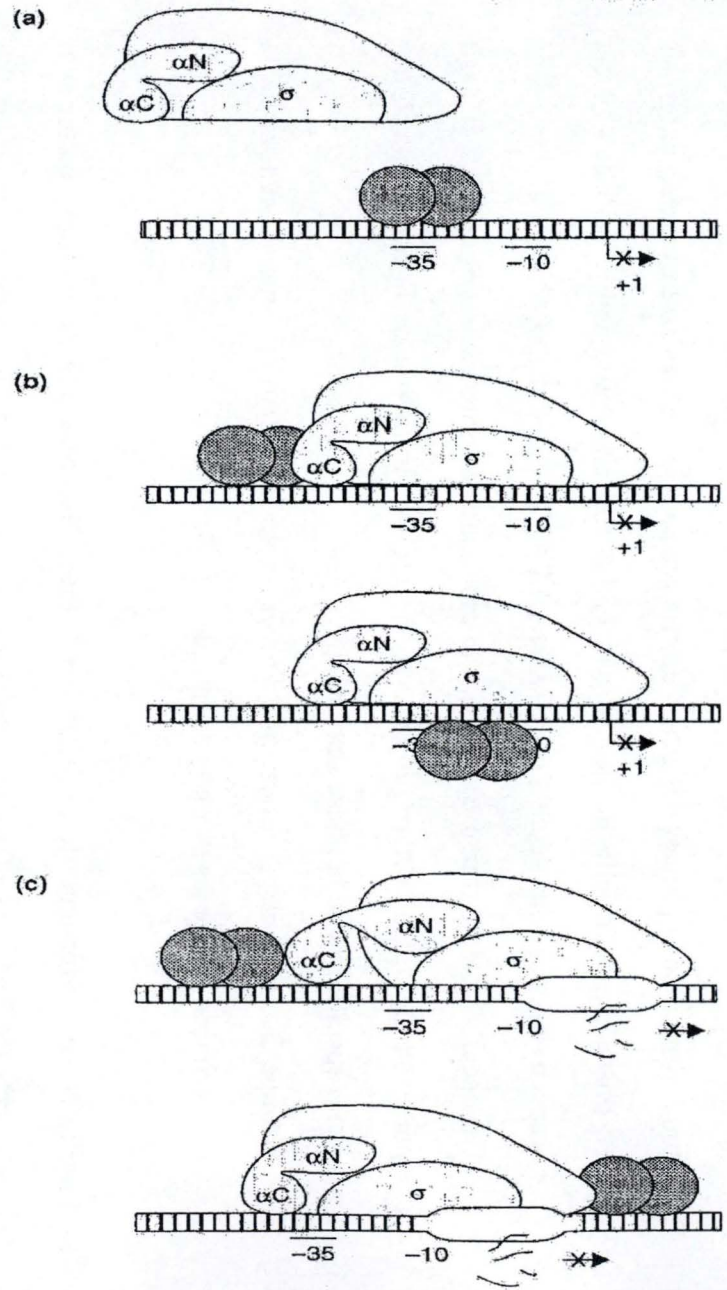

Figure 7 : Représentation des 3 types de répression.

a) Par encombrement stérique.

b) Par inhibition de la transition du complexe fermé au complexe ouvert.

c) Par blocage avant ou après la RNAP (Rojo, 2001).

activatrice. Le facteur σ^{54} est responsable de la transcription des gènes régulés par la disponibilité d'azote et de gènes en réponse au stress (Wosten, 1998).

1.2.5 Hiérarchie de la régulation

Lorsque le facteur de transcription se fixe à l'ADN, il active ou inhibe la transcription d'un seul ou de plusieurs gènes (dans le cas des opérons polycistroniques). Lors de l'activation d'une voie métabolique plus complexe, la régulation de plusieurs opérons différents est nécessaire étant donné que l'ensemble des gènes impliqués dans une voie métabolique est trop grand pour être groupés en un seul opéron. L'ensemble des opérons corégulés s'appelle un régulon et se trouve sous le contrôle d'un régulateur commun. Le niveau de régulation le plus élevé s'appelle le modulon. Ce dernier stade définit la régulation d'un groupe d'opérons intervenant dans une même réaction face à un stimulus extérieur.

1.3 Les sites d'atterrissage pour facteurs de transcription (S.A.F.T.)

1.3.1 Introduction aux S.A.F.T.

Les facteurs de transcription se fixent sur une séquence cis. Cette séquence cis ou S.A.F.T. se situe dans la zone en amont des pCDS. L'analyse de ces séquences cis révèle plusieurs caractéristiques : (i) les S.A.F.T. ont une composition précise, on parle de consensus; (ii) pour chaque facteur de transcription, il existe un S.A.F.T. spécifique; (iii) les S.A.F.T. ne sont pas distribués aléatoirement, leurs positions correspondent à leur action (répresseur proche, activateur lointain).

Il se peut qu'une séquence possède le motif propre à un S.A.F.T. sans pour autant être régulée par le facteur de transcription correspondant. Nous devrions parler de pS.A.F.T. (S.A.F.T. prédit) car pour confirmer que ce pS.A.F.T. est reconnu par un F.T., il faut une évidence

expérimentale. Néanmoins, durant ce travail, nous continuerons à parler de S.A.F.T. tout en gardant à l'esprit qu'il peut ne pas être actif.

1.3.2 Sites conservés

Il existe différents types de motifs conservés composant les S.A.F.T.. (i) Les monades sont composées de 3 à plus de 15 nucléotides dont la séquence est conservée. Le chiffre trois n'est pas dû au hasard car il est généralement admis que 3 paires de bases délimitent l'espace minimal permettant un contact stable entre le facteur de transcription et l'ADN (van Helden *et al.*, 2000(2)). Il n'y a pas de limite supérieure clairement définie, une moyenne observée est de 6 nucléotides c'est-à-dire 2 fois trois paires de bases (van Helden, et al., 2000(2)). (ii) Des S.A.F.T. composés de monades de tailles moyennes mais répétées plusieurs fois sur une courte distance avec des espacements variables. (iii) Le S.A.F.T. est composé de deux monades qui peuvent être différentes, les nucléotides sont conservés et l'espacement est constant, formant ainsi des dyades. La probabilité d'apparition des monades et des dyades dans une séquence aléatoire est comparable. Prenons quelques exemples:

a) Une monade de 6 nucléotides : ATTCAT. La probabilité d'apparition d'un nucléotide est de une chance sur quatre, sans tenir compte des probabilités d'apparition propres aux régions non codantes. Donc, le motif de 6 nucléotides ATTCAT a une chance sur 4096 d'apparaître car

$$\frac{1}{4}\times\frac{1}{4}\times\frac{1}{4}\times\frac{1}{4}\times\frac{1}{4}\times\frac{1}{4}=\frac{1}{4096}$$

b) Une dyade de 2 fois 3 nucléotides avec un espacement de 10 : ATTnnnnnnnnnnCAT. La probabilité d'apparition d'un nucléotide est de une chance sur quatre. Donc les deux motifs de 3 nucléotides espacés de 10 ATTnnnnnnnnnnCAT ont 1 chance sur 4096 d'apparaître car :

$$\frac{1}{4}\times\frac{1}{4}\times\frac{1}{4}\times\frac{1}{4}\times\frac{1}{4}\times\frac{1}{4}=\frac{1}{4096}$$

c) Deux monades de trois nucléotides sans espacement fixe : ATT.....CAT. La probabilité d'apparition d'un nucléotide est d'une chance sur quatre. Donc les deux motifs

de 3 nucléotides d'espacement non fixe ATT.....CAT ont 1 chance sur (4096/le nombre de positions possibles pour le second motif) d'apparaître car

$$\frac{1}{4}\times\frac{1}{4}\times\frac{1}{4}\times(Nombredepositions)\frac{1}{4}\times\frac{1}{4}\times\frac{1}{4}=\frac{Nombredepositions}{4096}$$

La probabilité d'apparition d'une monade et d'une dyade à espacement fixe est donc la même. Le "choix" évolutif entre les deux est spécifiquement dû aux facteurs trans correspondants. On parle souvent de consensus pour définir la composition d'un S.A.F.T. En effet, plusieurs nucléotides sont conservés d'un S.A.F.T. à l'autre pour un même facteur de transcription. Le consensus se compose donc de l'ensemble des nucléotides les plus représentés. En voici un exemple :

Soit une monade de 8 nucléotides retrouvée en amont de gènes co-régulés (représenté dans le tableau 1).

1)	A A T T A C G C
2)	A A T A A C G C
3)	A A T T A C C C
4)	A A T T A C C C
5)	A A T A A C C C
6)	A A T T A C G C
7)	A A T A A C C C
8)	A A T T A C G C
Consensus	A A T W A C S C

Tableau 1 : Les 8 motifs retrouvés et le consensus.

Le consensus est A A T (A/T) A C (C/G) C. Dans plusieurs programmes d'analyses de séquences, A ou T pour une position est symbolisé par W, C ou G par S. On obtient alors

Category	No. of promoters	% of total	% of repressible promoters	% of activatable promoters
Total no. of promoters	132			
Total no. of repressible promoters	91	69	100	
Total no. of activatable promoters	65	49		100
Promoters with repressor and activator sites	23	17	25	35
Promoters regulated by two or more different activators[b]	9	7		14
Promoters regulated by two or more different repressors	4	3	4	
Promoters with repeated homologous activator sites[c]	17	13		26
Promoters with repeated homologous repressor sites[c]	47	36	52	
Promoters with remotely located activator sites[d]	17	13		26
Promoters with remotely located repressor sites[e]	17	13	19	

[a]Percentages are calculated in relation to either the total number of promoters, the total number of repressible promoters, or the number of activatable promoters, as indicated.

[b]Does not include *ompF*, activated by OmpR and helped by IHF.

[c]DnaA sites are not included, as they may be one extended site.

[d]Of these promoters, 13 have a proximal site which binds a different protein.

[e]All of these promoters also contain a proximal site which binds the same protein as the remote one, and thus they are homologous duplications.

Tableau 2 : Analyse de 132 promoteurs, la colonne deux donne le nombre de promoteur concerné, colonne trois donne le pourcentage de promoteur, la colonne quatre et cinq donne respectivement le pourcentage de promoteurs répresseurs et activateurs concernés (Gralla and Collado-Vides, 1996).

comme consensus: **A A T W A C S C**. Toutes les autres combinaisons possibles de A,T,C et G ont une lettre assignée.

1.3.3 Spécificité du complexe cis trans

Dans un génome donné, chaque facteur de transcription possède au moins un élément cis. Dans la plupart des cas, l'élément cis est spécifique de son facteur trans (Gralla & Collado-Vides, 1996). Il existe cependant des sites cis accueillant plus d'un facteur trans et un facteur de transcription peut se fixer sur plusieurs facteurs cis. Mais seuls des régulateurs globaux peuvent se fixer sur plusieurs sites cis. Dans le tableau 2, nous pouvons voir que chez *E. coli*, sur 132 promoteurs analysés (Gralla and Collado-Vides, 1996), 91 sont répresseurs, 65 sont activateurs et 23 promoteurs cumulent les deux fonctions. Le nombre de promoteurs régulés par deux activateurs ou plus est de 9, il est de 4 pour les motifs répresseurs. Le nombre d'activateurs et de répresseurs avec des sites répétés sont respectivement de 17 et de 47. Les promoteurs ayant une position éloignée par rapport au site d'initiation de la transcription sont de 17 pour les activateurs et de 17 pour les répresseurs. Un élément cis peut diverger quelque peu, en fonction de la longueur du motif : plus il est long, moins il est conservé. Si le facteur trans est un régulateur général vital pour la bactérie, le site est souvent conservé de manière stricte.

La conservation du site cis d'une espèce à l'autre dépend de plusieurs facteurs. Premièrement, la proximité évolutive des espèces concernées. Un site cis commun peut être conservé dans deux espèces assez proches. C'est le cas du site de CtrA qui reste très conservé chez *Brucella melitensis, Sinorhizobium melitoti* et *Caulobacter crescentus*. Deuxièmement, CtrA étant un régulateur général essentiel pour la bactérie, la conservation de son site d'atterrissage est due à une forte pression de sélection. De plus, il y a une variation quant aux gènes cibles : CtrA ne régule pas nécessairement les mêmes cibles d'une espèce à l'autre (Bellefontaine *et al.*, 2002) .

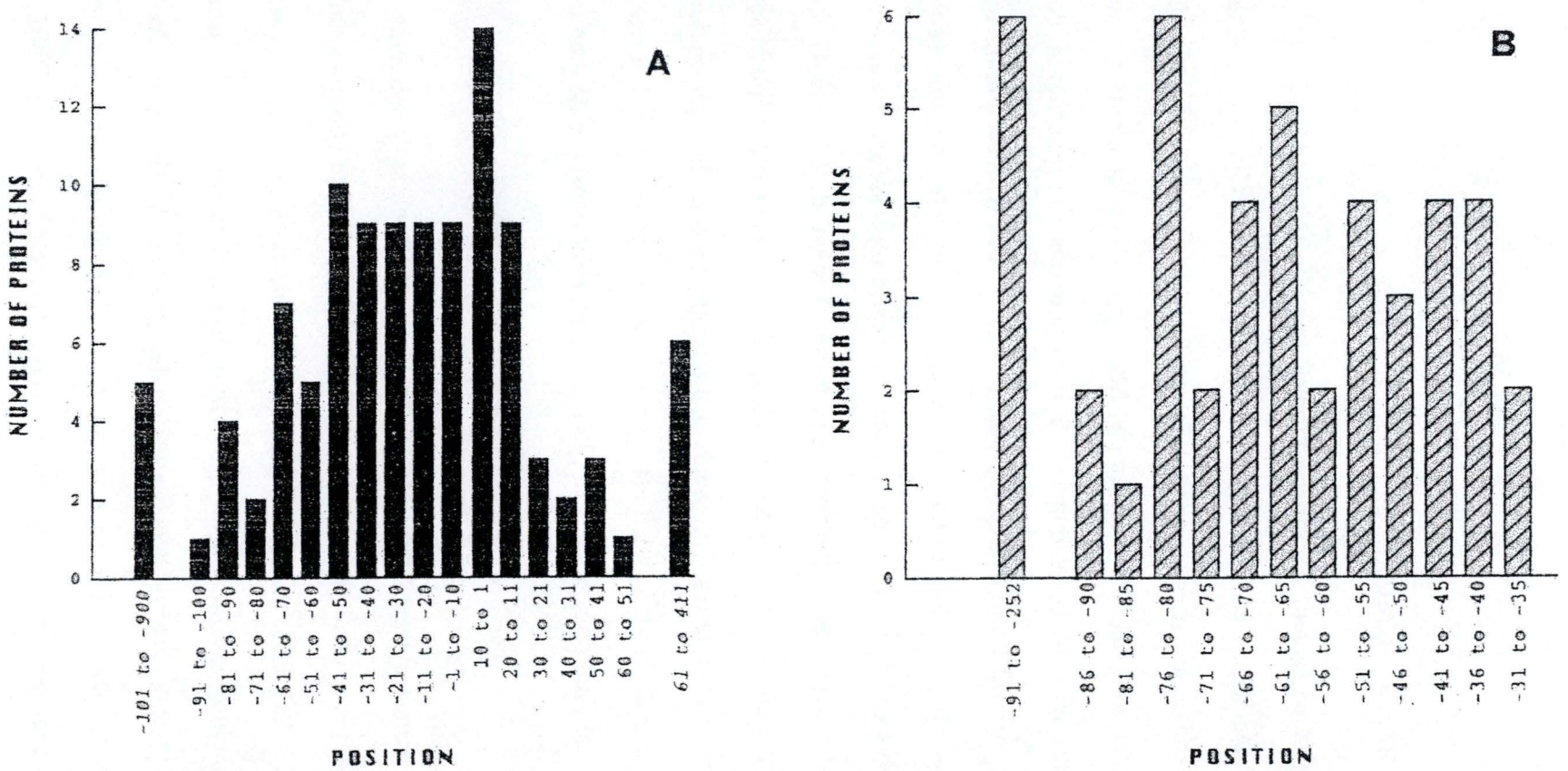

Figure 9 : A) chaque colonne indique le nombre de facteurs de transcription répresseurs qui ont au moins un site centré dans l'intervalle de 10 pb. La colonne du début et de la fin du graphique représentent le nombre de facteurs de transcription répresseurs présents avant -101 et après 61 B) chaque colonne indique le nombre de facteurs de transcription activateurs qui ont au moins un site centré dans l'intervalle de 5 pb. La colonne du début du graphique représente le nombre de facteurs de transcription activateurs présents entre -91 et -252 (Gralla and Collado-Vides, 1996).

1.3.4 Distribution des S.A.F.T. activateurs et répresseurs

Une étude poussée chez *E. coli* sur les F.T. (Gralla and Collado-Vides, 1996), nous permet d'étudier leur distribution. Les F.T. peuvent être activateurs, répresseurs ou les deux à la fois. Dans la figure 9, le graphique 9A représente la distribution sites répresseurs dans les promoteurs tandis que le graphique 9B représente la distribution des sites activateurs dans les promoteurs. Ces deux abscisses sont centrées sur le site d'initiation de la transcription. En analysant les graphiques, on remarque une zone exclusivement répressive, s'étalant de la zone -31 à la zone +60. Les répresseurs y sont très présents car en se fixant au site de fixation de la polymérase, ils empêchent son action. Les motifs activateurs sont seulement présents en amont de la région –31, pour interagir avec les deux sous-unités α ou/et avec le facteur σ.

1.3.5 Recherche de S.A.F.T.

La plupart des méthodes permettant de prédire un S.A.F.T. en amont de pCDSs corégulées partent du même principe : le S.A.F.T. a une composition en nucléotides différente de la région non codante où il se trouve. La différence entre ces programmes réside dans la méthode avec laquelle ils calculent les probabilités d'apparition des nucléotides dans les régions non codantes et par la méthode avec laquelle ils estiment si le motif diffère de l'environnement local.

Ces programmes peuvent rechercher le S.A.F.T. sur base de plusieurs séquences amont de pCDSs corégulées. Il existe cependant beaucoup de facteurs de transcription qui régulent une et une seule pCDS, ceux-là ne pourront jamais être détectés par ces différents programmes.

1.4 Brucella melitensis

1.4.1 Présentation du pathogène

Le genre *Brucella,* agent causal de la brucellose, a été décrit pour la première fois par David Bruce au 19^{ième} siècle. La brucellose, aussi appelée « fièvre de Malte », provoque de fortes

fièvres atypiques chez l'homme. Les principales causes de contamination chez l'homme sont l'ingestion de produits d'animaux contaminés ou le contact répété homme-animal (bergers, fermiers, vétérinaires,...). Il subsiste un problème lors du diagnostic de la maladie car les symptômes sont variés et non spécifiques. Toutefois, les individus susceptibles de contracter la brucellose (fermiers, scientifiques,...) sont prévenus des différents symptômes ce qui permet un diagnostique plus ciblé. Il n'existe malheureusement aucun vaccin efficace pour l'homme.

La brucellose touche les espèces sauvages (bisons, rennes, ours, sangliers, ...) mais elle parasite aussi de nombreux animaux d'élevage (bovins, ovins, caprins, porcins,...) qui contractent la maladie par le contact avec les animaux sauvages.

Chez l'animal, la brucellose touche principalement les organes reproducteurs, ce qui peut provoquer la stérilité chez le mâle et l'avortement chez la femelle gestante. *Brucella* est responsable d'une zoonose la plus commune, et son impact économique est très important. L'argent investi dans le contrôle de la brucellose aux Etats-Unis correspond à 150 millions de dollars par an seulement pour le traitement des bêtes d'élevages. En Amérique latine, les pertes animales sont estimées à 600 millions de dollars par an. Il existe deux vaccins couramment utilisés pour l'animal (Rev 1 et B19). Même si en France, en Belgique et une bonne partie des Etats-Unis, la brucellose est presque complètement éradiquée, elle cause encore beaucoup de pertes financières dues aux contrôles. Pour pouvoir totalement écarter la menace que représente *Brucella,* il faut élucider ses mécanismes de virulence, ... En dehors de l'argument économique, qui nous pousse à mieux connaître *Brucella* afin d'agir plus efficacement il y a un intérêt scientifique fondamental pour l'étude de *Brucella.*

Les bactéries du genre *Brucella* sont des coccobacilles intracellulaires facultatifs, gram négatives non mobiles appartenant à la sous-classe $\alpha 2$ des protéobactéries. Certains pathogènes des plantes (*Sinorhizobium, Mesorhizobium* et *Agrobacterium*) sont phylogénétiquement proches de *Brucella.* Le genre *Brucella* peut être divisé en six espèces en fonction de l'hôte parasité: *B. canis* chez le chien, *B. maris* chez les mammifères marins, *B. melitensis* chez le mouton, *B. abortus* chez la vache, *B. suis* chez le cochon, *B. neotomae* chez

les rongeurs. Parmi ces six espèces, trois sont virulentes chez l'homme, il s'agit de *B. melitensis, B. abortus* et *B. suis.*

Le génome de plusieurs d'espèces bactériennes proches d'un point de vue évolutif est complètement séquencé. Tous ces génomes permettent une comparaison interspécifique efficace car ils sont assez éloignés pour avoir divergés dans les parties aléatoires mais ils restent similaires dans les parties soumises à la pression de sélection. La comparaison de ces génomes montre que les gènes conservent une grande homologie et que les régions non codantes conservent des zones similaires.

La pathogénie de Brucella est actuellement peu comprise. Des recherches intenses ont été récemment menées pour découvrir des facteurs de virulence (Lestrate *et al.*, 2000), c'est-à-dire des acteurs protéiques qui permettent à la bactérie de survivre dans un hôte et de s'y multiplier. L'expression des gènes qui codent pour ces facteurs de virulence est en général strictement contrôlée au niveau transcriptionnel. Ces régulations impliquent donc la liaison de facteurs de transcription (F.T.) (activateurs, répresseurs, facteurs σ) dans le promoteur de ces gènes. Au cours de ce travail, nous avons analysé des méthodes qui permettraient de détecter les sites d'atterrissage pour ces facteurs de transcription (S.A.F.T.).

1.4.2 CtrA

Le seul S.A.F.T. connu et caractérisé chez *Brucella melitensis* est celui de CtrA. Les tests de recherche de sites d'atterrissage seront pratiqués sur celui de CtrA.

La surexpression artificielle de CtrA chez *B. abortus* entraîne une production de bactéries avec une morphologie aberrante. Ce résultat conduit à penser que CtrA intervient dans la régulation du cycle cellulaire. Une analyse *in silico* du génome de *Brucella melitensis* réalisé par A.-F. Bellefontaine (Presses universitaires de Namur, 2001) montre que le S.A.F.T. de CtrA se retrouve entre autres en amont des gènes homologues à *minC, ftsE, ftsK.* Sur base de cette observation, A.-F. Bellefontaine a émis l'hypothèse que ces gènes sont régulés par CtrA chez *Brucella.* Ceci expliquerait le rôle régulateur du cycle cellulaire de CtrA car MinC, FtsE et FtsK seraient impliqués dans la division cellulaire. Le S.A.F.T. de CtrA est une dyade, très

conservée sous la forme de « TTAAN(7)TTAA(C) », le C est entre parenthèses car il n'est pas toujours conservé.

En observant le schéma de régulation de et par CtrA chez *Caulobacter crescentus* et chez *Brucella melitensis*, on remarque déjà que, même si le S.A.F.T. de CtrA ne varie pas d'une espèce à l'autre, les cibles varient (voir figure 10). C'est un élément important à prendre en compte lors de l'analyse interspécifique du S.A.F.T. de CtrA.

*Ctr*A est présent chez *Brucella melitensis*, *Sinorhizobium meliloti et Caulobacter crescentus*, on retrouve la pCDS qui code pour CtrA dans ces différents génomes. Par contre, on ne retrouve par exemple pas la pCDS de *ctr*A chez *Bacillus subtilis et Bacillus halodurans*, qui ne sont pas des α protéobactéries.

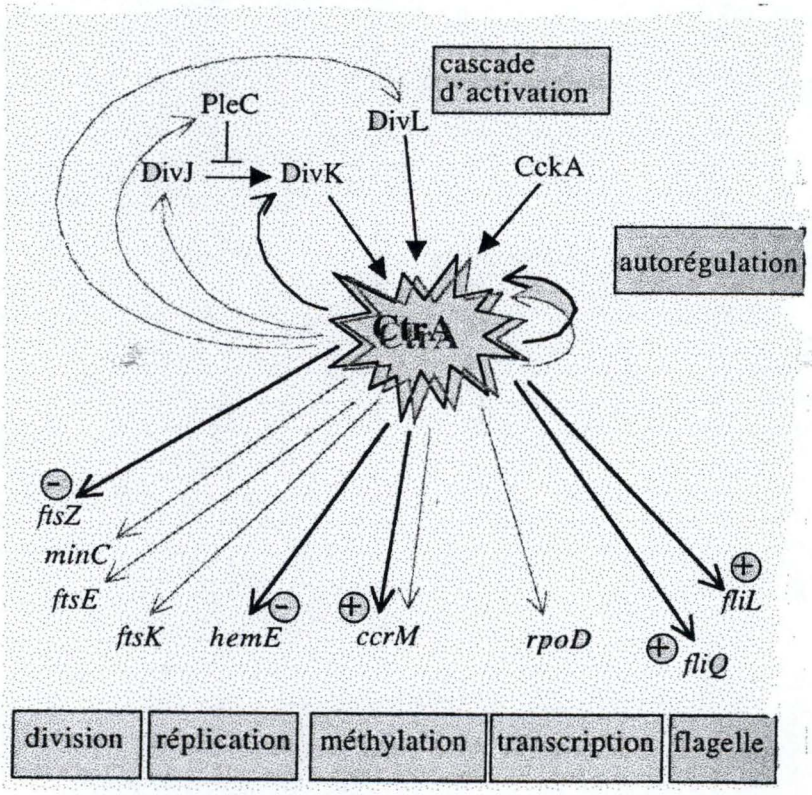

Figure 10 : cette figure représente les régulons dépendant de CtrA chez *C. crescentus* (noir) et *Brucella* (gris). Les flèches à tête épaisse représentent les transferts de phosphate aboutissant à l'activation de CtrA chez *C. crescentus*. Les flèches à tête fine représentent la régulation transcriptionnelle (+ pour activateur et - pour répresseur) médiée par CtrA.

Matériels et méthodes

1.4.3 Le génome de Brucella melitensis

Le génome de *Brucella melitensis* séquencé et annoté (DelVecchio *et al.*, 2002) a été rendu disponible en janvier 2002. Ce génome de 3.29Mb, réparti dans deux chromosomes de tailles différentes (2.117.144BP et 1.177.787BP), compose la matière première de ce travail. Les caractéristiques du génome de *Brucella melitensis* sont définies dans l'annexe 1.

L'annotation officielle des pCDSs de *Brucella melitensis* attribue les noms de BMEI0001 à BMEI2059 aux pCDS du premier chromosome et de BMEII0001 à BMEII1138 pour celle du second. Sur l'ensemble de la séquence génomique, 87% est codante. Le contenu en GC dans l'ensemble du génome est de 57%.

En ce qui concerne les pCDSs, il y en a 3.197 au total dont 2.059 sur le chromosome 1 et 1.138 sur le chromosome 2. Selon DelVecchio *et al* (2002), 78% des ORFs ont une fonction prédite, 15% ont une similarité élevée avec d'autres gènes et les 7% restants n'ont pas de similarité élevée avec les autres gènes. Il y a 66% des pCDSs impliquées dans des clusters orthologues (c'est-à-dire dans un cluster de gène ayant une même fonction chez des espèces différentes) et 26% sont impliquées dans des clusters paralogues (c'est-à-dire ayant une même fonction dans le même organisme). 49% des pCDSs sont présentes dans un cluster de gènes. Ce chiffre est important car chaque cluster de gènes peut se composer de un ou plusieurs opérons. Dans le génome de *Brucella melitensis*, 9% des gènes sont impliqués comme composant d'un gène fusionné (« as composite ») et 30% sont des gènes libres chez *Brucella melitensis* mais sont fusionnés chez une ou plusieurs autres espèces (« as components »).

2 Matériels et méthodes

2.2 RSA Tools

2.2.1 Introduction

Les outils de recherche et d'analyse de séquences RSA-tools (Regulatory Sequence Analysis - tools, voir http://rsat.ulb.ac.be/rsat/) permettent une analyse statistique des différents génomes complètement séquencés et sont développés par Jacques van Helden (van Helden *et al.*, 2000). Chaque outil est expliqué dans la suite du chapitre. Voici les trois situations auxquelles ces différents programmes peuvent répondre :

1) En connaissant des gènes co-régulés et leur S.A.F.T. correspondant, les programmes permettent de définir la position du S.A.F.T. dans le génome d'intérêt.
2) En connaissant des gènes co-régulés mais sans connaître le S.A.F.T. partagé par ces gènes, les programmes permettent de retrouver ce S.A.F.T.
3) On connaît le S.A.F.T. et on veut connaître les gènes co-régulés. On recherche le motif dans le génome, ce qui permet de prédire la position des gènes co-régulés en aval du S.A.F.T. retrouvé.

Outre ces programmes d'analyse du génome, le site web RSA-tools offre plusieurs autres services intéressants. Dans la partie "DATA", plusieurs génomes complètement séquencés et annotés sont disponibles. Pour chaque génome, nous disposons d'informations sur la fonction des protéines encodées par les pCDSs, de l'ensemble des régions non codantes et d'un tableau de fréquences d'apparition des oligonucléotides dans les régions non codantes. La fonction "*Purge sequence*" permet d'enlever toutes les régions répétées connues et communes à chaque séquence telles que la SD-box et les boîtes –35 et -10. L'outil « *Retrieve sequence* » permet de sélectionner des régions de part et d'autre du codon start des pCDSs.

Il faut alors préciser :

1) Le génome d'intérêt
2) Le nom du gène ou de la pCDS
3) Les limites amont et aval (*upstream, downstream*) de la région à sélectionner, par rapport au codon start qui est positionné à zéro.

Exemple : BMEII0348 chez *Brucella melitensis* entre –400 et + 40 nucléotides.

D'après les observations réalisées chez *E. coli* par Jay D.Gralla et Julio Collado-Vides (Gralla and Collado-Vides, 1996), rares sont les S.A.F.T. positionnés plus de 300 nucléotides avant le codon start des pCDSs. Nous avons donc choisi pour un génome procaryotique de sélectionner les régions entre –300 et 0 (codon start). Une recherche sur une plus grande région ne ferait qu'augmenter le taux de faux positifs. De plus, vu la faible distance entre deux gènes chez les bactéries, on se retrouve très rapidement dans la région codante précédente. Comme on peut le voir sur la figure 11, la partie promotrice de la seconde pCDS se superpose à la partie codante de la première pCDS. La région superposée de la partie promotrice de la seconde pCDS est donc aussi conservée que la partie codante de la première pCDS, cette région promotrice superposée a donc le même biais des codons que la séquence codante.

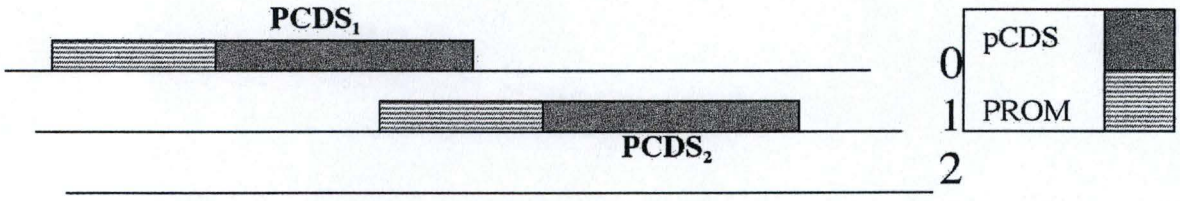

Figure 11 : Dans ce schéma sont représentées les 3 phases de lecture(0,1,2) d'un génome. Ces deux pCDSs sont précédées d'une région promotrice (PROM). La première pCDS se superpose à la région promotrice de la seconde pCDS. Donc, comme la région codante a un biais de codons, la région promotrice de la seconde pCDS a aussi ce même biais de codons.

2.1.2 Oligo-analysis

2.1.2.1 But

Ce programme détecte les monades sur-représentées par rapport à leur fréquence attendue par hasard (van Helden *et al.*, 1998).

2.1.2.2 Explication

Le programme "*Oligo analysis*" (*RSA-tools*) propose plusieurs méthodes pour effectuer la recherche de S.A.F.T. : une méthode utilisant les chaînes de Markov, une méthode se basant sur la fréquence des oligonucléotides dans toutes les régions non codantes et la division lexicale. La méthode utilisée par défaut est « la recherche de fréquence des oligonucléotides dans toutes les régions non codantes ». Toutes ces méthodes différentes estiment la fréquence attendue au hasard (*expected frequency*) du S.A.F.T.. Avec l'estimation basée sur la fréquence des oligonucléotides dans toutes les régions non codantes, la fréquence attendue au hasard pour un motif x est donnée par la formule ci-dessous (van Helden, et al., 1998) :

$$Frequ.attendue = occmot / occallmot$$

Où

1) occmot est l'occurrence du motif dans toutes les régions non codantes du génome
2) occallmot est la somme des occurrences de tous les motifs de même longueur dans toutes les régions non codantes du génome.

Cette fréquence attendue au hasard est une valeur cruciale pour la suite de la recherche de motifs puisque pour chaque motif, on compare sa fréquence d'apparition à sa fréquence attendue. Sur base de cette comparaison, on définit si le motif est anormalement présent ou pas. Il existe plusieurs méthodes pour estimer la fréquence attendue du motif comme décrit plus haut. Deux sont réellement efficaces (van Helden, et al., 1998), « la fréquence des oligonucléotides dans toutes les régions non codantes » et l'estimation de la fréquence

d'apparition par un modèle de Markov d'ordre k (k est au choix de l'utilisateur). Le modèle de Markov sera expliqué plus en détail dans le point 2.2.2.1. Notre estimation de la fréquence attendue s'est axée sur l'option par défaut, car d'après les tests effectués, cette méthode donne les meilleurs résultats (van Helden, et al., 1998).

Lorsqu'on soumet différentes régions promotrices au programme, il nous renvoie les résultats sous la forme d'un tableau comme le tableau 3.

SEQ	identifier	expected_freq	occ	exp_occ	occ_prb	occ_sig	rank
acgtgc	acgtgc|gcacgt	0.0001058140904	10	0.84	2.3e-08	4.33	1
acgtgg	acgtgg|ccacgt	0.0000891309929	8	0.71	8.4e-07	2.76	2
tgccaa	tgccaa|ttggca	0.0002409780744	12	1.92	8.7e-07	2.74	3
ctgcac	ctgcac|gtgcag	0.0001224971878	8	0.97	8.5e-06	1.75	4
cacgtg	cacgtg|cacgtg	0.0001177085210	12	0.94	9.7e-06	1.69	5
cgcacg	cgcacg|cgtgcg	0.0000631795080	6	0.50	1.5e-05	1.52	6
aaacgt	aaacgt|acgttt	0.0003033852167	11	2.41	4.5e-05	1.03	7
cccacg	cccacg|cgtggg	0.0000659600242	5	0.52	0.00021	0.35	8
aacgtg	aacgtg|cacgtt	0.0001513836621	7	1.20	0.00026	0.28	9

Tableau 3 : résultats d'*Oligo-analysis*. *SEQ* et *identifier* = le motif, *expected_freq* = la fréquence attendue, *occ* = l'occurrence du motif, *exp_occ* = l'occurrence attendue du motif, *occ_prob* = la probabilité d'occurrence du motif, *occ_sig* = l'occurrence significative du motif, *rank* = le rang du motif classé selon son occ_sig.

La première colonne (SEQ) présente la composition du motif, la seconde (identifier) donne le motif dans le premier brin et son complémentaire sur l'autre brin.

La troisième colonne, indique l'*expected frequency* ou fréquence attendue.

La quatrième colonne est l'occurrence : elle représente le nombre d'apparitions du motif dans les séquences étudiées. L'occurrence du motif dans les séquences étudiées ne peut pas être comparée à la fréquence attendue, cette fréquence attendue est relative à toutes les séquences non codantes du génome et pas seulement aux séquences étudiées. La fréquence attendue est à mettre en rapport avec le nombre et la taille des séquences étudiées. Ce résultat est en cinquième colonne, c'est-à-dire l'occurrence attendue (*expected occurrence*). Celle-ci se calcule comme suit :

$$occurrenceattendue = p * 2 * \sum_{j=1}^{N} (Lj + 1 - w) = p * T$$

Où

1) p est la probabilité d'apparition du motif estimé par la fréquence d'apparition de ce motif dans toutes les régions non codantes du génome d'intérêt.
2) 2 provient de la recherche sur les deux brins.
3) $\sum_{j=1}^{N}$ (Lj +1-w) = T représente toutes les positions possibles où on peut trouver le motif.
4) N est le nombre de séquences testées, Lj est la longueur de la séquence j (j allant de 1 à N) et w est la longueur du motif.

En sixième colonne, la probabilité d'occurrence (*occurrence probability*) représente la probabilité d'avoir au hasard une occurrence plus grande ou égale à celle observée dans le génome. Celle-ci est calculée par la loi binomiale (T,P). Dans une binomiale, le premier paramètre est le nombre d'épreuves (T) et le second (P) est la chance de succès à l'épreuve. Par exemple, la chance de faire 6 aux dés 3 fois sur une série de 10 lancés se calcule par la binomiale P(X=3), X=BI(10,1/6). Dans notre cas, la binomiale Bi(T,P) est calculée à partir de P, la fréquence attendue du motif et T, le nombre de positions possibles du motif.

La valeur la plus représentative du tableau est l'occurrence significative (colonne 7). Lorsqu'un motif a une occurrence significativement plus grande que 0, il apparaît dans le tableau de résultat. Mais le motif est seulement pris en considération comme un S.A.F.T. potentiel s'il a une occurrence significative supérieure à 2, car les résultats inférieurs à 2 sont trop peu fiables et présentent trop de faux positifs. De plus les vrais S.A.F.T. ont souvent une occurrence plus élevée (van Helden, et al., 2000). Une option du programme permet de fixer une limite d'occurrence significative pour les résultats, le programme sélectionne alors seulement les résultats ayant une occurrence significative supérieure à X où X est fixé par l'utilisateur (van Helden, et al., 1998). Cet indice est calculé par la formule ci-dessous :

$$Occ - sig = -\log_{10}(NDP * p(\geq obs))$$

Où

1) NDP est le nombre de dyades possibles. Soit pour deux trinucléotides qui composent la dyade : $(1/4*1/4*1/4)^2$.
2) p(>=obs) est la probabilité d'observer par hasard une occurrence égale ou plus grande que celle obtenue.

2.1.3 Dyad-analysis

2.1.3.1 But

Ce programme détecte les dyades sur-représentées par rapport à leur fréquence attendue dans le génome (van Helden, et al., 2000).

2.1.3.2 Explication

Pour rappel, une dyade se compose de deux monades et d'un espacement fixe. Le programme permet de retrouver des dyades dans un ensemble de séquences non codantes. Ces dyades retrouvées par le programme doivent obligatoirement être composées de 2 groupes de 3 nucléotides (monades) espacés de 1 à 20 (si c'est 0 on a une dyade de six nucléotides).

Le choix du triplet correspond à une réalité biologique ; en effet, comme nous l'avons déjà dit précédemment, les facteurs de transcription (ayant des S.A.F.T. sous une forme de dyades) reconnaissent généralement l'ADN sur trois bases. La recherche de triplets est aussi un bon compromis statistique car la recherche de dinucléotides génère plus de faux positifs (faible confiance) et, inversement, la recherche de tétrades augmente la confiance en diminuant la puissance. Enfin, le programme est limité par le temps de calcul à la recherche des dyades composées de deux triplets. Pour une dyade de type BBB- (0-20)N -BBB (B=base), il y a $(4)^6$=4096 possibilités pour un espacement donné. Pour une dyade composée de tétrades, il y a

$(4)^8$=65536 possibilités pour un espacement donné, ce qui demande environ seize fois plus de temps calcul (van Helden, et al., 2000).

Un exemple de résultat de *Dyad-analysis* est repris au tableau 4.

dyad_sequence	dyad_identifier	expected_freq	obs_occ	exp_occ	occ_prb	occ_sig	rank
taan{8}taa	taan{8}taalttan{8}tta	0.0002683768868	30	2.40	8.7e-23	17.42	1
gttn{9}tta	gttn{9}ttaltaan{9}aac	0.0002371272854	19	2.12	1.7e-12	7.13	2
ggtn{0}taa	ggtn{0}taalttan{0}acc	0.0002065654575	17	1.84	1.6e-11	6.15	3
ttan{9}taa	ttan{9}taalttan{9}taa	0.0002612204320	28	2.33	3.3e-11	5.84	4
taan{7}tta	taan{7}ttaltaan{7}tta	0.0002852981366	28	2.55	1,00E-10	5.35	5
aacn{6}tta	aacn{6}ttaltaan{6}gtt	0.0003042744978	19	2.72	1.1e-10	5.32	6
gttn{10}taa	gttn{10}taalttan{10}a	0.0002693071531	18	2.40	1.2e-10	5.30	7
aacn{7}taa	aacn{7}taalttan{7}gtt	0.0002389997547	16	2.13	1.2e-09	4.29	8
attn{0}aac	attn{0}aaclgttn{0}aat	0.0001602663032	13	1.43	4.5e-09	3.71	9
gttn{0}aac	gttn{0}aaclgttn{0}aac	0.0001424589362	18	1.27	2.6e-08	2.94	10
ggtn{10}tta	ggtn{10}ttaltaan{10}a	0.0001314627335	10	1.17	4.7e-07	1.69	11
agtn{2}taa	agtn{2}taalttan{2}act	0.0000856616822	8	0.76	1.5e-06	1.19	12
aacn{4}cgt	aacn{4}cgtlacgn{4}gtt	0.0001639998336	10	1.46	3.3e-06	0.84	13
tgtn{0}taa	tgtn{0}taalttan{0}aca	0.0002065654575	11	1.84	3.9e-06	0.77	14
actn{13}cac	actn{13}caclgtgn{13}a	0.0000728479923	7	0.65	5.5e-06	0.62	15
cgtn{0}taa	cgtn{0}taalttan{0}acg	0.0001911324061	10	1.71	1.2e-05	0.27	16
ggtn{1}aac	ggtn{1}aaclgttn{1}acc	0.0001567191855	9	1.40	1.6e-05	0.15	17
aatn{7}taa	aatn{7}taalttan{7}att	0.0003366017488	13	3.01	1.6e-05	0.15	18

Tableau 4 : résultats de *Dyad-analysis*. *dyad_sequence* et *dyad_indentifier* = la dyade, *expected_freq* = la fréquence attendue, *occ* = l'occurrence de la dyade, *exp_occ* = l'occurrence attendue de la dyade, *occ_prob* = la probabilité d'occurrence de la dyade, *occ_sig* = l'occurrence significative de la dyade, *rank* = le rang de la dyade classé selon son occ_sig.

Si la dyade se compose de monades de plus de trois bases, le programme "*Dyad-analysis*" trouvera plusieurs dyades dont les parties conservées se juxtaposeront pour former la dyade complète. Dans l'ensemble des sous programmes du site RSA-tools, le programme "*pattern-assembly*" se charge de rassembler les dyades retrouvées sur-représentées. Nous pouvons reformer une dyade plus grande (*pattern*), en utilisant les résultats repris aux 5 premières lignes du tableau 4, comme nous pouvons le voir dans la figure 12.

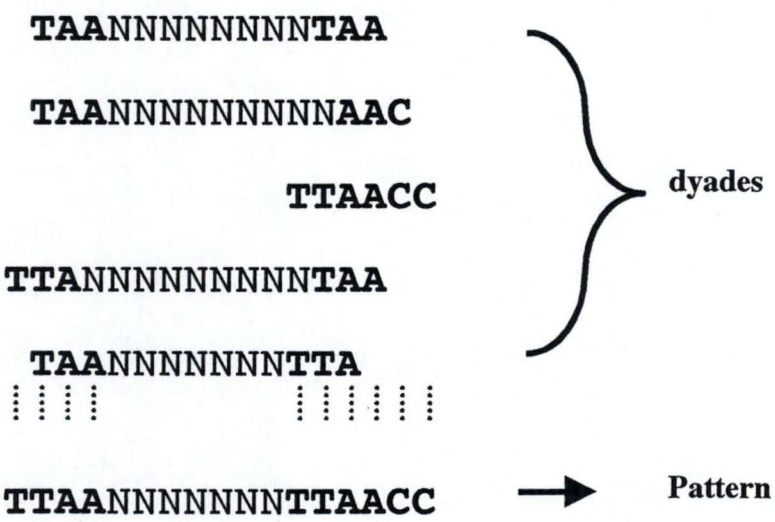

Figure 12 : On retrouve ces dyades dans les séquences sur-représentées par rapport à leur occurrence attendue. En les agençant correctement, on remarque qu'elles se superposent pour former un *pattern*. Ce *pattern* correspond à la vraie dyade.

La seule différence entre *Oligo-* et *Dyad-analysis* réside dans l'estimation de la fréquence attendue. Les autres paramètres se calculent de manière identique.

Il existe deux calibrations différentes pour estimer la fréquence attendue au hasard d'une dyade:

(i) la calibration selon les monades qui composent la dyade. Le calcul s'effectue de la sorte :

Exp-freq. de la dyade = exp-freq. (monade 1) * exp-freq (monade 2)

C'est le produit de la fréquence attendue des deux monades. La fréquence attendue de la monade 1 et 2 se calcule comme dans le programme *Oligo-analysis*.

(ii) La calibration selon la fréquence de la dyade dans toutes les régions non codantes du génome. Cette calibration applique à la dyade le même calcul que le programme "*Oligo analysis*" appliquait aux monades.

Frequ.attendue = occmot / occallmot

OU

1) occmot est l'occurrence de la dyade dans toutes les régions non codantes du génome.
2) occallmot est la somme des occurrences de toutes les dyades possibles, dont l'espacement est de même longueur, dans toutes les régions non codantes du génome.

2.1.4 Genomic scale

2.1.4.1 But

Ce programme renvoie la position d'un motif dans un génome. Il peut également renvoyer le nombre de fois que le motif apparaît devant une certaine pCDS ainsi que la fonction prédite de la protéine encodée par cette pCDS.

2.1.4.2 Explication

On définit un cadre de recherche (par exemple entre –300 bp et +300 bp par rapport au codon start de la pCDS qui est en position zéro) ainsi que le génome dans lequel il faut rechercher le motif. On lui soumet ensuite un ou plusieurs motifs à rechercher dans cette zone. Il peut rechercher le motif exact ou le motif en y autorisant une substitution.

Un exemple de recherche du motif GTTAATCATA entre –300 et +300 par rapport au codon start est repris dans le tableau 5.

PatID	Strand	Pattern	SeqID	Start	End	Matching_word	Score
gttaatcata	D	gttaatcata	BMEI0072	-402	-393	GTTAACCATA	0.9
gttaatcata	R	gttaatcata	BMEI0100	-337	-328	GTTAATGATA	0.9
gttaatcata	D	gttaatcata	BMEI0150	-314	-305	GGTAATCATA	0.9
gttaatcata	R	gttaatcata	BMEI0151	-302	-293	GGTAATCATA	0.9

Tableau 5 : résultats obtenus par *Genomic scale*. PatID et *pattern* = le motif du *pattern*, strand = brin sur lequel le motif est retrouvé, *SeqID* = identification de la pCDS, *Start* et *End* = position du début du motif et de fin du motif, *Matching_word* = motif tel qu'il est retrouvé, *score* = score du motif.

La première et la troisième colonne représentent le motif recherché. Le motif examiné peut être de toutes les formes possibles. Le tableau 6 représente les différentes lettres qui peuvent composer le motif.

A	(Adenine)
C	(Cytosine)
G	(Guanine)
T	(Thymine)
R	= A or G (puRines)
Y	= C or T (pYrimidines)
W	= A or T (Weak hydrogen bonding)
S	= G or C (Strong hydrogen bonding)
M	= A or C (aMino group at common position)
K	= G or T (Keto group at common position)
H	= A, C or T (not G)
B	= G, C or T (not A)
V	= G, A, C (not T)
D	= G, A or T (not C)
N	= G, A, C or T (aNy)

Tableau 6 : les différentes lettres pouvant composer un motif.

Ainsi le motif : G/C T A A/G A/T A A/G G/C se symbolise par les lettres suivantes S T A R W A R S.

La deuxième colonne précise si le motif se trouve sur le brin inverse (reverse= r) ou codant (direct= d). Le numéro de la pCDS se trouve en colonne quatre. Les colonnes 5 et 6 indiquent les positions de début et de fin du motif. Le zéro ne se positionne plus au codon start mais bien à la fin du champ de recherche sélectionné : si la recherche s'effectue de –200 à +200, les résultats s'étendront de 0 à –400. L'avant dernière colonne présente le motif tel qu'il est retrouvé. En effet, lorsqu'on utilise des lettres comme N, R, Y, W ou lorsqu'on permet une substitution, la nature des N ou des W, R, Y, ... n'est pas connue. Le programme nous donne la composition du vrai motif. La dernière colonne indique le score. Le score mesure la similarité entre le motif demandé et le motif trouvé. Le score est généralement de 1.00. Dans l'exemple du tableau 5, nous avons permis la substitution d'une base sur 10, le score du motif est donc de 0.9.

2.2 Motif Sampler

2.2.1 But

Ce programme a pour but de détecter une monade dont la composition diffère par rapport au modèle statistique des séquences environnantes (Thijs *et al.*, 2001) (Thijs *et al.*, 2002). Ce programme est disponible à l'adresse.

www.esat.kuleuven.ac.be/~thijs/Work/MotifSampler.html.

2.2.2 Explication

2.2.2.1 Le modèle caché de Markov

Le programme Motif Sampler utilise une méthode basée sur un modèle caché de Markov. Pour expliquer ce qu'est un modèle caché de Markov, il faut décrire le modèle de Markov. Le modèle de Markov le plus simple est une chaîne de Markov (Krogh *et al.*, 1994). Une chaîne de Markov est une suite d'événements liés par des probabilités. Les événements sont connus et les probabilités sont reprises dans la matrice de transition. Par exemple, la probabilité de passer d'un événement à l'événement suivant sur la figure 13 est de 1. Dans le cadre de ce travail, l'événement est l'apparition d'un nucléotide.

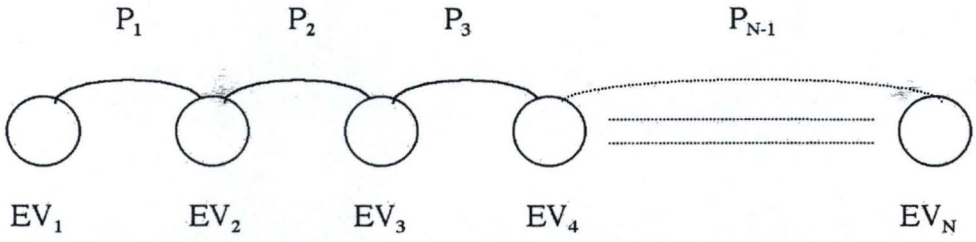

Figure13 : Ce schéma représente une chaîne de Markov. EV = événement et P = probabilité.

Prenons pour exemple une séquence de longueur N : AATC-----T, et établissons une chaîne de Markov sur base de sa composition. La chaîne ressemble à la figure 13 avec pour événement, $EV_1 = A$, $EV_2 = A$, $EV_3 = T$, $EV_4 = C$, ..., $EV_N = T$. Pour cette chaîne de Markov basée sur une seule séquence, la probabilité de passer de l'événement 1 à l'événement 2 ou d'avoir A en deuxième position sachant que l'on a A en première position est de 1.

Si on veut maintenant établir une chaîne de Markov pour modéliser quatre séquences de même longueur N, (voir figure 14) la probabilité de passer d'un événement à l'événement suivant reste toujours de 1 mais la probabilité d'avoir A à la seconde position est de 0,5. Dans ce cas-ci, on a une chaîne cachée de Markov car les événements sont inconnus ou cachés. En effet, on sait qu'on va avoir un nucléotide, mais on ne sait pas lequel, chacun ayant une certaine probabilité d'être émis.

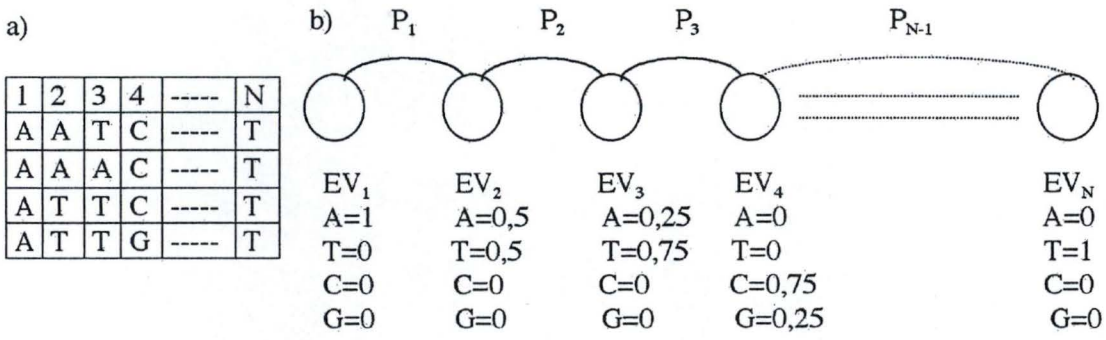

Figure 14 : Chaîne de Markov basée sur quatre séquences. EV = événement, P = probabilité. a) profil des séquences où les nombres 1,2,3,4, ..., N représentent les positions dans le profil. b) chaîne de Markov et matrice d'émission.

Imaginons maintenant que nous voulions établir une chaîne cachée de Markov pour modéliser une série de séquences de la même famille. Certaines sont plus grandes et d'autres plus petites car il peut y avoir des délétions et des insertions.

Pour tenir compte de ces événements de délétions et d'insertions dans la chaîne cachée de Markov, nous devons créer un modèle qui prend en compte ces événements. Ce modèle est visible dans la figure 15, c'est un modèle caché de Markov qui prend en compte les événements de délétions et d'insertions. La probabilité de passer d'un événement à l'événement suivant est toujours égale à 1, mais la probabilité que l'événement suivant soit l'apparition d'un nucléotide n'est plus toujours égale à 1 car on tient compte de la probabilité d'avoir une insertion ou une délétion.

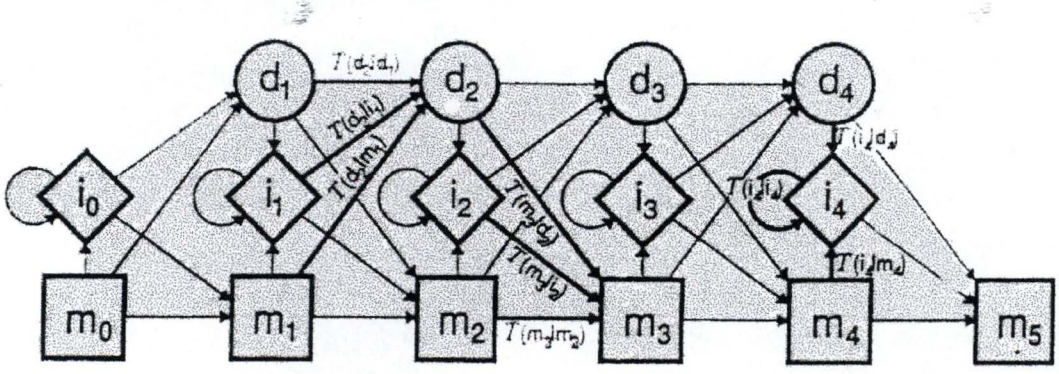

Figure 15 : Anders Krogh *et al* (Krogh, et al., 1994). Ce schéma représente un modèle de Markov avec les délétions et les insertions. Les carrés représentent les événements où on a un acide aminé ou un nucléotide, les losanges représentent une insertion et les ronds une délétion. Toutes les flèches représentent une probabilité de transition d'un état à un autre.

Dans ce modèle, nous ne connaissons aucune des probabilités. Il existe des algorithmes (Krogh, et al., 1994) qui déterminent les probabilités du modèle caché de Markov à partir d'un ensemble de séquences de la même famille.

Pour notre exemple, l'événement N ne dépendait que de l'événement N-1. C'est un modèle de Markov d'ordre 1 ; si l'événement N dépend à la fois de l'événement N-1 et N-2, c'est un modèle de Markov d'ordre 2 (voir tableau 7).

	AA	AT	AC	AG	TT	TA	TC	TG	GG	GC	GT	GA	CC	CA	CT	CG
A	0.20	0.25	0.33	0.21	0.27	0.30	0.10	0.12	0.11	0.15	0.15	0.23	0.12	0.16	0.29	0.28
T	0.28	0.20	0.25	0.33	0.21	0.27	0.30	0.10	0.12	0.11	0.35	0.15	0.23	0.12	0.23	0.15
C	0.15	0.28	0.20	0.25	0.33	0.21	0.27	0.30	0.10	0.12	0.11	0.29	0.29	0.23	0.12	0.16
G	0.37	0.27	0.22	0.21	0.19	0.22	0.33	0.48	0.67	0.62	0.39	0.33	0.36	0.49	0.36	0.41

Nucléotide testé | Nucléotides Précédents | Fréquence d 'apparition de T après GC

Tableau 7 : Probabilité d'une chaîne de Markov d'ordre 2. Donc la probabilité d'apparition du nucléotide testé dépend de ce qui se trouve aux deux positions précédentes. Ici, la probabilité d'avoir T en sachant qu'il y a GC avant est de 0,11.

2.2.2.2 Motif Sampler (Gibbs sampling)

Voici une description de la méthode utilisée par le programme développé par Gert Thijs (Thijs, et al., 2001) (Thijs, et al., 2002).

1. La première étape consiste à sélectionner le génome auquel on applique ce test. Dans le cadre de ce travail, il s'agit du génome de *Brucella melitensis*. Le programme calcule un modèle de Markov d'ordre 3 établi sur base des séquences non codantes du génome testé.

2. Le programme sélectionne au hasard dans chaque séquence testée, un motif de longueur fixe (selon le choix de l'utilisateur) (voir figure 16). Sur base de ces motifs de même longueur, le programme établit une matrice de fréquences pour les motifs. (voir tableau 8)

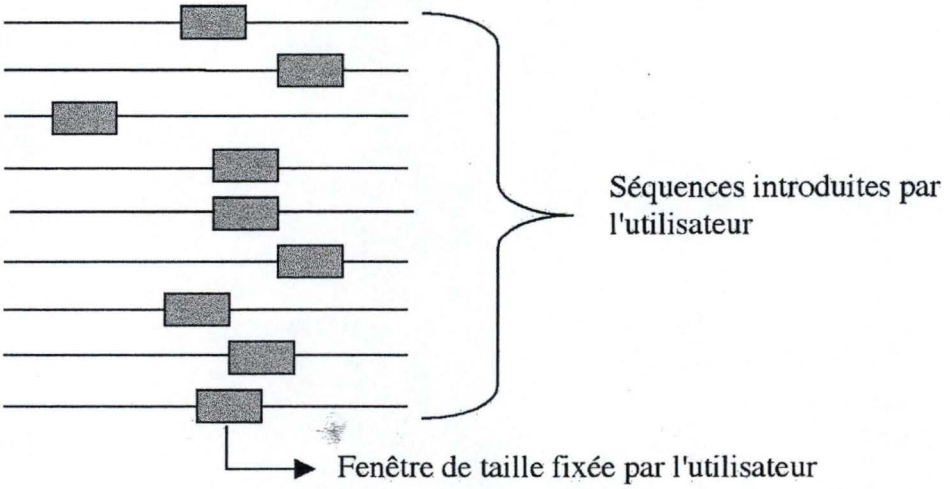

Figure16 : Le programme sélectionne au hasard un motif pour chaque séquence. Ensuite, sur base de ces motifs, il crée une matrice de fréquence.

	position 1	**position 2**	**position 3**	**----------**	**Position N**
A	**0,25**	**1**	**1**	**----------**	**0**
T	**0,25**	**0**	**0**	**----------**	**0**
C	**0,5**	**0**	**0**	**----------**	**0,25**
G	**0**	**0**	**0**	**----------**	**0,75**

Tableau 8 : Tableau de fréquences d'un motif de longueur N.

3. Le programme compare cette matrice de fréquences au modèle de Markov d'ordre 3. Le programme calcule alors la probabilité de retrouver ce motif, avec une certaine fréquence (fréquence observée). Il prend en compte la taille des séquences de départ et la fréquence attendue au hasard de ce motif consensus établit sur base du modèle de Markov d'ordre 3. Le motif consensus le plus sur-représenté par rapport à son occurrence attendue sera gardé comme S.A.F.T. potentiel.

Pour être sûr de trouver ce motif consensus, le programme effectue 300 fois les étapes 1 à 3. A chaque itération, le programme sélectionne un nouveau motif de longueur fixe dans chaque séquence et établit une nouvelle matrice.

Lorsqu'on reçoit les résultats, les indices associés aux motifs sont les suivants :

-"Consensus score" ou score du consensus. Cet indice mesure la conservation du motif. Un motif parfaitement conservé est indiqué par un score de deux, un autre, uniformément distribué, a un score de zéro.

-"Information content" ou contenu informatif. Cet indice calcule la différence entre la fréquence d'apparition observée du motif et la fréquence attendue au hasard pour ce motif sur base du modèle de Markov d'ordre 3. Cette comparaison est mise en rapport avec la taille des séquences testées.

-"log-likelyhood" ou le logarithme de probabilité. Cet indice ne peut être expliqué sans rentrer dans le détail des formules relatives à *Motif Sampler*. Néanmoins, il estime la probabilité que le motif retrouvé ait été engendré par le modèle de Markov d'ordre 3. Plus la valeur est élevée, plus il est improbable que ce motif aie été engendré par le milieu ambiant.

2.3 Artemis

Artemis (Rutherford *et al.*, 2000) est utilisé dans le cadre de cette étude comme un simple visualiseur de fichier. Les génomes sont disponibles en plusieurs formats, nous avons choisi le format .gbk ("Genbank"). Le programme « Artemis » lit ce format et affiche le génome à l'écran en tenant compte des paramètres définis dans le fichier "Genbank" (début et fin des ORFs, des pCDS ainsi que le numéro et la fonction des protéines qu'elles encodent, ...).
Certaines options du programme nous permettent aussi d'annoter le génome et d'indiquer toutes les remarques souhaitées. Nous pouvons aussi rechercher certains éléments d'intérêt

dans le génome ou simplement une pCDS précise. Ce programme est donc d'une grande utilité pour rechercher toutes les régions extra pCDS ou encore pour visualiser les clusters de gènes, les opérons ... afin de se faire une idée claire de la disposition des gènes étudiés. Artemis possède bien d'autres fonctions qui ne seront pas décrites ici. D'autres renseignements sur ce programme de visualisation et d'annotation de séquences, se trouvent sur le site web à l'adresse www.sanger.ac.uk/Software/Artemis/.

2.4 Programme "alicombiJF"

Le but du programme est de générer toutes les dyades ou monades possibles ayant un contenu fixe. Exemple: toutes les monades contenant deux T et deux A : TTAA, AATT, TATA, ATAT, TAAT, ATTA.

Ce programme a été mis au point par Christophe Lambert et est écrit en FORTRAN 77.

2.5 Test de χ^2

2.5.1 But

Le test du χ^2 nous permet d'établir si l'écart de deux fréquences par rapport à la fréquence théorique est significative.

2.5.2 Explication

Dans notre cas, le test de χ^2 va permettre d'établir si la différence de distribution amont/aval d'un motif est significative. Pour ce test, l'indice χ^2 est calculé de la manière suivante.

$$\chi^2 = \frac{(Fobs1 - Fth)^2}{Fth} + \frac{(Fobs2 - Fth)^2}{Fth}$$

OU

1) Fobs1 est la fréquence amont

2) Fobs2 est la fréquence aval

3) Fth est égal à (Fréquence Amont + Fréquence Aval)/2

Nous devons aussi définir un degré de liberté (dl), ce degré de liberté est égal au nombre de variables testées moins 1. Nous comparons deux variables (Fobs1 et Fobs2) ce qui nous donne un degré de liberté de 1.

Lorsque nous connaissons l'indice de χ^2 et le degré de liberté du test, nous pouvons rechercher l'indice théorique dans la table de χ^2. La table de χ^2 est à deux entrées, une pour le degré de liberté et l'autre pour la probabilité que ce soit significatif.

Ex: une probabilité de 0,995 dl=2
Indice de χ^2= 10,6

Nous connaissons le degré de liberté du test ce qui fixe une ligne du tableau. L'ensemble des indices de χ^2 théoriques de cette ligne correspond à des probabilités que la différence soit significative. En comparant l'indice obtenu et les indices théoriques, nous aurons la probabilité de différence significative. Si notre indice de χ^2 est plus grand ou égal à l'indice théorique, la différence observée à une probabilité < P d'être observé (P étant la probabilité prise dans le tableau, de 0,5 à 0,9995).

Résultats

3. Résultats

Le but de ce mémoire est de définir différentes méthodes permettant de retrouver des sites d'atterrissage pour facteurs de transcription dans le génome de *Brucella melitensis*. La première partie des résultats va mettre en évidence la force et la faiblesse des 4 méthodes testées. Ensuite, nous tenterons d'établir une méthode de recherche ne se basant plus sur un ensemble de séquences co-régulées mais sur base d'un caractère intrinsèque aux facteurs de transcription.

3.1 Chapitre 1 : Résultats des différents programmes recherchant le S.A.F.T. de CtrA

Plusieurs programmes permettant de retrouver un S.A.F.T. parmi un ensemble de séquences co-régulées existent déjà. Cette première partie présente les différents résultats obtenus par ces programmes. Lors de notre recherche d'un S.A.F.T. chez *Brucella melitensis,* nous ne recherchons pas un S.A.F.T. au hasard, mais un S.A.F.T. connu et caractérisé en laboratoire chez *Brucella melitensis* : le S.A.F.T. de CtrA (Bellefontaine, et al., 2002). Puisque seul le S.A.F.T. de CtrA est connu chez *Brucella melitensis*, nous n'avons pas pu valider ces méthodes sur d'autres S.A.F.T. Les différents programmes de recherche de S.A.F.T. utilisés pour ces tests ont été décrits dans la partie matériels et méthodes.

3.1.1 "Oligo-analysis"

La première phase du protocole expérimental est identique pour tous les programmes étudiés : il faut tout d'abord sélectionner les séquences d'intérêt. Dans notre cas, les séquences d'intérêt sont toutes les régions en amont du codon start des pCDS de *Brucella melitensis* pouvant posséder le motif CtrA. Pour sélectionner ces séquences, nous utilisons le programme « *Genomic scale* » pour qu'il nous identifie les pCDSs possédant le S.A.F.T. de CtrA. Il est

nécessaire de définir deux paramètres importants pour cette première étape : le motif recherché et la zone de recherche.

Le motif complet du S.A.F.T. de CtrA est la dyade TTAAn(7)TTAAC. En prenant ce motif strict, 15 séquences contenant le S.A.F.T. de CtrA sont sélectionnées (voir tableau 9).

BMEI0531	BMEI0072	BMEI1809
BMEI0532	BMEI0100	BMEI1810
BMEI0738	BMEI0168	BMEI1932
BMEI1242	BMEI0331	BMEII0927
BMEI1279	BMEI0345	BMEII0976

Tableau 9 : Numéro des 15 pCDSs dont la partie amont possède le S.A.F.T. de CtrA.

Le programme « *Genomic scale* » recherche le motif du S.A.F.T. de CtrA dans une zone de recherche précise définie par l'utilisateur, soit dans notre cas la zone située entre –300 et 0 (zéro étant le codon start). Une plus longue zone de recherche diminuerait nos chances de retrouver le S.A.F.T. de CtrA car le bruit autour du S.A.F.T. recherché augmenterait. Une plus courte zone pourrait exclure des résultats, c'est-à-dire des pCDSs possédant le motif du S.A.F.T. de CtrA mais situé en dehors de la zone de recherche. D'après l'étude réalisée chez *E. coli* (Gralla and Collado-Vides, 1996), il y a peu de S.A.F.T. à plus de 300 nucléotides du codon start de la pCDS concernée. Le risque d'obtenir des faux négatifs en choisissant –300 nucléotides comme limite amont est donc réduit. Cette limite semble donc un bon compromis statistique entre la confiance et la puissance du test.

Lorsque l'identité des pCDSs possédant le motif du S.A.F.T. de CtrA est connue, le programme « *Retrieve sequences* » est utilisé pour qu'il nous donne la séquence comprise entre –300 et 0 de toutes les pCDSs possédants CtrA. A la fin de cette première phase, 15 séquences amont des pCDSs possédant le motif de CtrA ont été sélectionnées.

A seq	**identifier**	**Expected_freq**	**occ**	**exp_occ**	**occ_prb**	**Occ_sig**	**rank**
gttaa	gttaalttaac	0.001543697	36	5.75	2.00E-17	14	1
ggtta	ggttaltaacc	0.001371448	17	5.11	2.50E-05	1.89	2
atgaa	atgaalttcat	0.003087559	26	11.49	0.00016	1.09	3
gcgga	gcggaltccgc	0.002577191	22	9.59	0.0004	0.69	4
attaa	attaalttaat	0.001522432	15	5.67	0.0008	0.39	5
aaggt	aaggtlacctt	0.001556456	15	5.79	0.00099	0.29	6
gatga	gatgaltcatc	0.002362412	19	8.8	0.00187	0.02	7
B seq	**identifier**	**Expected_freq**	**occ**	**exp_occ**	**occ_prb**	**Occ_sig**	**rank**
ggttaa	ggttaalttaacc	0.000421774	16	1.56	1.40E-11	7.54	1
attaac	attaaclgttaat	0.000334101	13	1.24	8.10E-10	5.77	2
gttaac	gttaaclgttaac	0.000158755	7	0.59	2.90E-06	2.22	3
tgttaa	tgttaalttaaca	0.000423912	9	1.57	3.90E-05	1.09	4
accttc	accttclgaaggt	0.000393975	8	1.46	0.00014	0.53	5

Tableau 10 : ce tableau présente les résultats obtenu lors d 'une recherche de S.A.F.T de cinq (A) et de six (B) nucléotides sur base de quinze séquences d'intérêts contenant le S.A.F.T. de ctrA. *seq* et *identifier* = le motif, *expected_freq* = la fréquence attendue, *occ* = l'occurrence du motif, *exp_occ* = l'occurrence attendue du motif, *occ_prob* = la probabilité d'occurrence du motif, *occ_sig* = l'occurrence significative du motif, *rank* = le rang du motif classé selon son occ_sig.

Le programme « *Oligo-analysis* » permet de retrouver des motifs entre 1 et 8 nucléotides, mais le S.A.F.T. de CtrA est de 16 nucléotides. Le programme ne peut donc retrouver tout le motif mais seulement une des deux monades de la dyade du S.A.F.T. de CtrA.
Afin de vérifier si ce programme permet de retrouver le S.A.F.T. de CtrA lorsque le cadre de recherche est trop grand, deux motifs de cinq et six nucléotides sont recherchés afin de retrouver le motif TTAAC dans les quinze séquences (sélectionnées entre –300 et 0) contenant le S.A.F.T. de CtrA. La calibration du programme est effectuée sur base de la fréquence attendue pour chaque motif de même longueur dans toutes les régions non codantes de *Brucella melitensis*. (Voir tableau 10).

Lorsque la recherche est effectuée sur un motif de cinq nucléotides, le programme « *Oligo-analysis* » retrouve la demi-dyade du S.A.F.T. de CtrA (TTAAC) 36 fois dans les quinze séquences ce qui définit une valeur d'occurrence significative très élevée (indice de 14). Si nous recherchons un motif de 6 nucléotides, TTAAC est retrouvé à plusieurs reprises mais avec une occurrence significative beaucoup moins élevée. Lors d'un test en aveugle, il nous faut tenir compte de cette observation : si la recherche est effectuée sur un trop grand motif, des faux négatifs peuvent se produire. Il est donc préférable de rechercher des petits motifs. De plus le programme «*Pattern assembly* » permet de rassembler tous les sous-motifs pouvant former un grand motif comme expliqué dans la partie matériels et méthodes.

En conclusion, le programme « oligo analysis » permet de retrouver la demi-dyade du S.A.F.T. de CtrA mais pas la dyade complète. Nous savons qu'en majorité les F.T. des procaryotes sont des HTH et que le site d'atterrissage d'un HTH est une dyade. Cela suggère que le programme « Oligo analysis » n'est pas adapté à la recherche de dyade et donc à la recherche de S.A.F.T. chez *Brucella melitensis*.
En ce qui concerne la taille de recherche, il est préférable de rechercher un petit motif qu'un grand, car un grand cadre de recherche peut passer outre le motif recherché tandis qu'un petit

dyad_sequence	dyad_identifier	expected_freq	obs_occ	exp_occ	occ_prb	occ_s
taan{8}taa	taan{8}taalttan{8}tta	0.00026838	30	2.4	8.70E-23	17.4
gttn{9}tta	gttn{9}ttaltaan{9}aac	0.00023713	19	2.12	1.70E-12	7.13
ggtn{0}taa	ggtn{0}taalttan{0}acc	0.00020657	17	1.84	1.60E-11	6.15
ttan{9}taa	ttan{9}taalttan{9}taa	0.00026122	28	2.33	3.30E-11	5.84
taan{7}tta	taan{7}ttaltaan{7}tta	0.0002853	28	2.55	1.00E-10	5.35
aacn{6}tta	aacn{6}ttaltaan{6}gtt	0.00030427	19	2.72	1.10E-10	5.32
gttn{10}taa	gttn{10}taalttan{10}aac	0.000269307	18	2.4	1.20E-10	5.3
aacn{7}taa	aacn{7}taalttan{7}gtt	0.000239	16	2.13	1.20E-09	4.29
attn{0}aac	attn{0}aaclgttn{0}aat	0.00016027	13	1.43	4.50E-09	3.71
gttn{0}aac	gttn{0}aaclgttn{0}aac	0.00014246	18	1.27	2.60E-08	2.94
ggtn{10}tta	ggtn{10}ttaltaan{10}acc	0.000131462	10	1.17	4.70E-07	1.69
agtn{2}taa	agtn{2}taalttan{2}act	8.5662E-05	8	0.76	1.50E-06	1.19
aacn{4}cgt	aacn{4}cgtlacgn{4}gtt	0.000164	10	1.46	3.30E-06	0.84
tgtn{0}taa	tgtn{0}taalttan{0}aca	0.00020657	11	1.84	3.90E-06	0.77
actn{13}cac	actn{13}caclgtgn{13}agt	0.000072847	7	0.65	5.50E-06	0.62
cgtn{0}taa	cgtn{0}taalttan{0}acg	0.00019113	10	1.71	1.20E-05	0.27
ggtn{1}aac	ggtn{1}aaclgttn{1}acc	0.00015672	9	1.4	1.60E-05	0.15
aatn{7}taa	aatn{7}taalttan{7}att	0.0003366	13	3.01	1.60E-05	0.15

Tableau 11 : ce tableau présente les résultats obtenus par le programme « Dyad analysis » lorsqu'il est calibré selon la fréquence d'apparition des dyades dans toutes les régions non codante et sur base de quinze séquences contenant le S.A.F.T. de ctrA. *Dyad-sequencer* et *dyad-identifier* = le motif, *expected_freq* = la fréquence attendue, *occ* = l'occurrence du motif, *exp_occ* = l'occurrence attendue du motif, *occ_prob* = la probabilité d'occurrence du motif, *occ_sig* = l'occurrence significative du motif, *rank* = le rang du motif classé selon son occ_sig.

cadre de recherche sélectionne au moins une partie du motif recherché. Lors d'une étape suivante, l'agencement de ces petites parties nous permettra de définir sa taille réelle (*Pattern assembly*).

3.1.2 "Dyad-analysis"

Pour tester le programme « Dyad analysis », le processus de sélection des séquences d'intérêt est identique. Nous utilisons d'ailleurs le même ensemble de quinze séquences. Néanmoins, le paramétrage de la recherche est effectué de manière différente. Premièrement, le programme « Dyad analysis » peut être calibré selon deux méthodes, soit la calibration par dyade soit par les monades qui composent la dyade, comme décrit dans le chapitre matériels et méthodes. Deuxièmement, on recherche un motif de forme définie : BBBN(0-20)BBB. B (Base) correspond aux nucléotides composants les deux monades de la dyade et N (aNy) sont les nucléotides aléatoires composants l'espacement (cet espacement peut être constitué de 0 à 20 nucléotides N).

Le programme peut retrouver un motif composé de deux triplets conservés et d'un espacement allant de 0 à 20 dans les quinze séquences introduites et selon la calibration par dyades ou par monades. Ces dyades ont la forme de BBBN(0-20)BBB. Parmi ces dyades, il y en a plusieurs dont l'espacement est de 7 nucléotides ; lorsqu'elles sont assemblées par le programme « Pattern assembly » cela donne le S.A.F.T. de CtrA. Les résultats sont décrits dans le tableau 11.

Au vu de ces résultats nous pouvons conclure que ce programme retrouve bien une dyade commune à plusieurs séquences. Nous avons aussi voulu savoir si : (i) le programme retrouvait toujours ces dyades si toutes les séquences ne la possédaient pas. (ii) savoir laquelle des deux calibrations était la plus efficace. En effet, en fonction de la calibration choisie, les résultats du programme peuvent différer. Ainsi, les deux calibrations ont été étudiées afin de déterminer la méthode donnant les meilleurs résultats.

Afin d'étudier les résultats donnés par les 2 calibrations, les séquences amont des 15 gènes contenant le S.A.F.T. de CtrA sont introduites dans le programme. Celui-ci est calibré avec comme fréquence attendue pour une dyade, la fréquence d'apparition de la dyade dans toutes les régions non codantes du génome de *Brucella melitensis*. Les résultats obtenus sont visibles dans le tableau 11, et il peut être observé que le programme détecte des dyades de la forme suivantes BBBN(0-20)BBB (où B=base), alors qu'une des dyades correspondant au S.A.F.T. de CtrA est TTAn(8)TTA. Cette dyade sera dorénavant appelée CtrA''. CtrA'' apparaît en tête de liste des résultats avec l'indice de signifiance le plus élevé. Lors d'une deuxième étape, nous avons utilisé comme calibration le produit de la fréquence attendue pour les deux monades composant la dyade et nous avons de nouveau retrouvé CtrA'' en haut du tableau de résultats. Néanmoins, ces deux résultats sont différents car la dyade CtrA'' n'obtient pas la même occurrence significative en fonction de la calibration choisie.

Afin de savoir si le programme est capable de retrouver le S.A.F.T. de CtrA dans un groupe de gènes ne contenant pas tous le S.A.F.T. de CtrA, nous avons ajouté des séquences qui ne contiennent pas le S.A.F.T. de CtrA, donc en quelque sorte, dilué le motif dans du bruit. La figure 17 nous montre en abscisse le titre de la dilution, à savoir qu'une dilution 2 représente 15 séquences contenant le S.A.F.T. de CtrA et 15 séquences sans, une dilution 3 représente 15 séquences avec le S.A.F.T. de CtrA et 30 sans et ainsi de suite. L'occurrence significative de la dyade CtrA'' est représentée en ordonnée. Si le programme ne retrouve plus le motif, sa valeur est inférieure à zéro. Par commodité de représentation, toutes les valeurs inférieures à zéro ont été ramenées à zéro. Nous remarquons que la calibration par monade retrouve le motif quel que soit le titre de la dilution. Il augmente même vers les hautes dilutions. Cette augmentation est due à la recherche indirecte du S.A.F.T. de CtrA par CtrA''. En effet, si toutes les séquences permettant de diluer le motif ne contiennent pas le S.A.F.T. de CtrA, elles contiennent peut être CtrA''. C'est cette apparition de CtrA'' dans les séquences de dilution qui augmente l'indice d'occurrence significative. Nous aurions pu enlever toutes les

séquences contenant CtrA'' mais nos résultats auraient été biaisés, car lors de ce test nous devions agir comme si le test avait été réalisé en aveugle. Pour un S.A.F.T. non connu, nous ne pourrions pas enlever ces séquences et donc les résultats seraient aussi affectés par une augmentation de l'occurrence significative.

La calibration par monade repère la dyade parmi un très grand bruit de fond, mais elle finit par repérer tellement de dyades possibles que la dyade du S.A.F.T. de CtrA'' est elle-même noyée dans d'autres résultats. Quant à la calibration par dyade, elle perd très vite la trace de CtrA''. La calibration par monade est décrite comme étant plus puissante mais moins confiante que la calibration par dyade (van Helden, et al., 2000), comme nous pouvons l'observer ici.

Nous avons démontré que ce programme permet de retrouver la dyade CtrA" parmi 1800 séquences dont 15 seulement possèdent le S.A.F.T. de CtrA. Lors du test de ces 1800 séquences, le programme retrouve également beaucoup d'autres dyades donc certaines ont une occurrence significative plus grande que CtrA". Lors d'un test en aveugle, rien ne nous permettrait de definir CtrA" comme étant la dyade composant le S.A.F.T de CtrA. Voilà pourquoi, les tests à grandes échelles ne peuvent être pratiqués, le nombre de S.A.F.T. potentiel serait trop grand pour savoir lequel est réellement un S.A.F.T..

En conclusion, après ce test sur un motif connu et caractérisé, nous déduisons que :

1) Pour un test sur des gènes que l'on sait co-régulés par une dyade, la calibration par dyade est la meilleure car elle est plus confiante (l'occurrence significative est plus élevée).
2) Pour un test sur un ensemble de gènes dont certains seulement sont co-régulés (où le bruit est plus important), la calibration par monade semble la plus apte à retrouver le bon motif dans un bruit de fond élevé, mais la confiance est faible.
3) Pour augmenter la confiance liée au motif, le plus simple est de rechercher ce motif dans le génome et d'observer si sa distribution est corroborée par sa régulation. En

d'autres termes, prendre l'ensemble des pCDSs ayant ce motif en amont et voir si elles correspondent aux pCDSs co-régulés ou aux gènes correspondant à une voie de régulation décrite dans la littérature.

La différence des résultats entre les deux calibrations peut-être mise en évidence en prenant comme exemple une dyade dont l'espacement est de 0.

L'occurrence significative de cette dyade se calcule sur base de sa fréquence attendue. Dans un premier cas, la fréquence attendue est calculée sur base de la fréquence d'apparition de cette dyade dans toutes les régions non codantes du génome (la calibration par les dyades). Dans un second cas, la fréquence attendue est calculée sur base du produit de la fréquence d'apparition des deux triplets composant la dyade (la calibration par monade). Comme on peut le voir sur la figure 18, la fréquence d'apparition de la dyade est moins importante que la fréquence d'apparition des deux triplets. Il en résulte que l'occurrence significative d'une dyade sera plus grande dans le cas de la calibration par dyade car la fréquence attendue sera moindre.

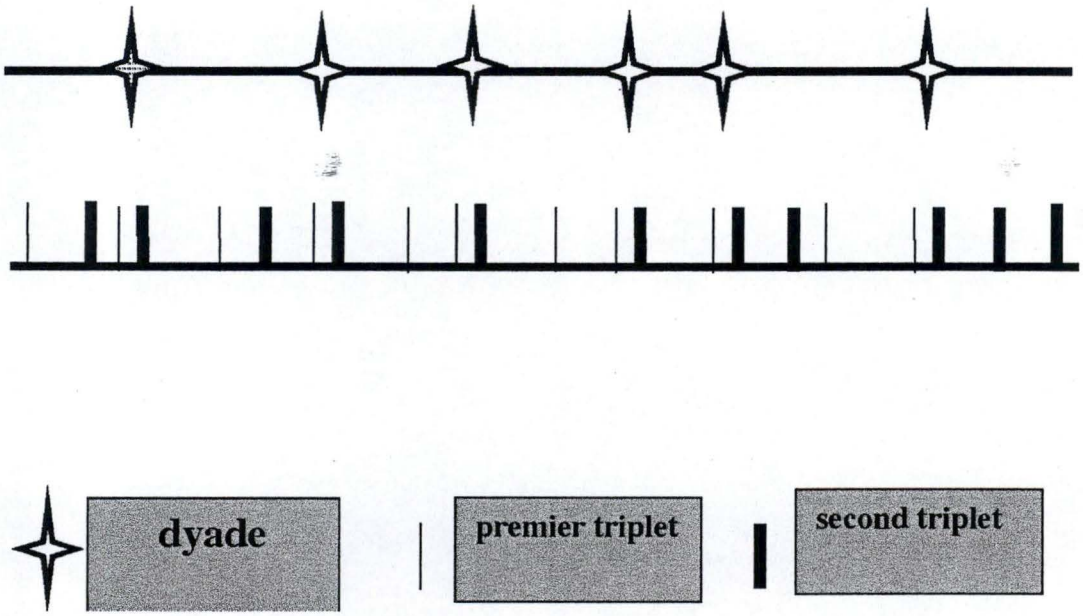

Figure 18 : les deux lignes représentent chacune l'ensemble des régions non codantes du génome. Sur la première ligne, on montre où les dyades apparaissent. Sur la seconde ligne, on montre la distribution des triplets : bruit de fond et dyades réelles

La meilleure méthode reste encore de réaliser le test selon les deux calibrations car les résultats cumulés augmentent notre confiance sur les résultats. Si un motif est trouvé, il peut s'agir seulement d'une partie de la dyade réelle. Comme pour le S.A.F.T. de CtrA, le motif complet n'apparaît pas : seules des parties du motif sont représentées dans les résultats. Comme décrit plus haut, le programme «*Pattern assembly* » permet d'assembler toutes ces parties de dyades en une dyade complète. Plus la dyade assemblée est large, plus nous avons confiance en ces résultats puisqu'un long motif a statistiquement moins de chances d'apparaître au hasard qu'un motif court.

3.1.3 "Motif Sampler"

L'étape de sélection des séquences d'intérêts est identique à celle des deux programmes précédents (*Oligo-analysis* et *Dyad-analysis*). Les tests effectués sur ces trois programmes se basent donc sur les 15 mêmes séquences possédant le motif du S.A.F.T. de CtrA.

Au début de ce travail, un premier test fut effectué avec le programme « Motif Sampler ». A ce moment, le programme ne permettait pas d'utiliser un modèle de Markov propre à *Brucella melitensis*. La calibration se faisait selon un modèle de Markov établi sur base des séquences introduites et les résultats n'étaient pas concluants. Grâce à l'aide de Gert Thijs (Thijs et al., 2001), un nouveau modèle de Markov d'ordre 3 fut établi pour *Brucella melitensis*. sur base de l'ensemble des séquences non codantes de *Brucella melitensis*. Ensuite, sur base des 15 séquences d'intérêts, nous avons recherché des motifs de longueur fixée à 8 (suivant le conseil de Gert Thijs). (voir le tableau 12).

	Motif	consensus	info	log-lik	Fréquence
Motif 1:	nArAwCmG	1.2101	1.2343	450.0665	1
Motif 2:	TTAAnnnT	1.1636	1.2269	731.63	99
Motif 3:	mAwmAAwA	1.141	1.2123	443.0777	2
Motif 4:	AAwsAnrw	1.1413	1.2039	446.0686	4
Motif 5:	AAArrrAA	1.1343	1.1954	496.016	1

Tableau 12 : résultats du programme « *Motif Sampler* » calibré avec un modèle de Markov établi sur base des régions non codantes de *Brucella*. Les colonnes représentent dans l'ordre, le motif, le score du consensus, indice d'information contenue, *log-likelyhood* et finalement le nombre de fois que le motif apparaît sur 300 itérations du programme.

Le tableau 12 représente les cinq meilleurs résultats sur base de l'indice d'information contenue. Le premier motif est sans doute un faux positif puisqu'il n'apparaît qu'une seule fois sur les 300 itérations du programme. La demi-dyade du S.A.F.T. de CtrA apparaît 99 fois, ce qui est assez important. L'indice de vraisemblance (log-likelyhood) est d'ailleurs plus élevé que pour n'importe quel autre motif. Nous pouvons donc considérer que *Motif sampler* a retrouvé le motif de la demi-dyade du S.A.F.T. de CtrA avec succès parmi les 15 cibles présumées de ce régulateur. Toutefois le problème inhérent à cette méthode est le suivant : qu'il existe un motif commun ou pas, le programme trouvera toujours une solution car *Motif Sampler* va converger vers une solution qui ne correspond pas toujours au résultat escompté.

Les deux programmes « *Oligo analysis* » et « *Motif Sampler* » parviennent au même résultat mais leur méthode est différente. Ils ont tous les deux des points fort et des points faibles. « *Oligo analysis* » est moins puissant mais plus confiant que « *Motif Sampler* ». « *Oligo analysis* » retrouve la demi-dyade du S.A.F.T. de CtrA et la classe en première position. « *Motif Sampler* » retrouve beaucoup plus de motifs mais les résultats sont plus incertains puisque TTAA n'est pas en haut de la liste. En conclusion, le meilleur test possible pour rechercher une dyade reste encore une combinaison des deux résultats (*Oligo-analysis* et *Motif Sampler*). La redondance des deux résultats nous permettra d'assurer la présence du ou des motifs avec une plus grande confiance.

3.1.4 Recherche de S.A.F.T. par comparaison de deux régions promotrices de pCDSs orthologues.

Ici, l'approche du problème est différente. Nous comparons les régions promotrices de pCDSs orthologues chez deux espèces évolutivement proches. Nous émettons l'hypothèse que si les pCDSs sont homologues, leurs S.A.F.T. le sont aussi. Nous avons testé cette méthode pour connaître son efficacité.

Nos deux espèces testées sont *Brucella melitensis* et de *Sinorhizobium meliloti*. Ces deux organismes ont été choisis car leur génome ont divergé dans les régions amonts de leurs pCDSs pour ne garder en commun que les S.A.F.T. comme c'est le cas pour le S.A.F.T. de CtrA (Bellefontaine *et al.*, 2002).

Nous utilisons le programme « *blast2* » pour rechercher des parties conservées pouvant contenir des S.A.F.T. Pour retrouver le S.A.F.T. de CtrA, nous comparons la région *upstream* de deux pCDSs orthologues entre *Brucella melitensis* (avec la pCDS BMEI0423 (response regulator CtrA)) et chez *Sinorhizobium meliloti* (avec le pCDS 15075742 (response regulator CtrA)). Nous obtenons les deux blocs conservés repris dans le tableau 13, le premier a un pourcentage d'identité de 78% et le second de 90%.

1)
ggctgcggaaaggggataaaagatgcgcgtcctttgattgaagacgacagtgctatcgc ||| |||||| ||| | | ||||| || || |||| |||||||||| || | || ggcggcggaaaaggggaagactatgcgggttctactgatcgaagacgacagcgcgacggc
acagagcattgagttgatgctcaagtccgagagttt Brucella melitensis ||||||| ||| | |||||||||||| ||||| tcagagcatcgagctcatgctcaagtccgaaagttt Sinorhizobiom meliloti
2)

gaatcatattttgttaaccatttgctggca	Brucella melitensis
|||||| | ||||||||||||| |||||	
gaatcagaatttgttaaccatttggtggca	Sinorhizobiom meliloti

Tableau 13 : Bloc 1 et 2 résultant de l'alignement de la partie amont de 2 pCDS orthologues.

La demi-boîte du S.A.F.T. de CtrA est présente dans le second bloc conservé. Il ne s'agit pas du S.A.F.T. complet de CtrA.

En conclusion, cette méthode de recherche de S.A.F.T. n'est pas efficace car il y a beaucoup trop de contraintes.

Premièrement les deux organismes doivent être assez proches pour garder des parties communes et assez éloignés pour ne pas être trop identique ce qui limite le nombre de comparaisons.

Deuxièmement, on ne trouve les dyades que si elles sont présentes dans une région conservée entre les deux espèces. Si la fenêtre de recherche du programme d'alignement est plus longue que le motif recherché, nous pouvons ne pas voir ce motif.

Troisièmement, d'une espèce à l'autre, les gènes ciblés par un facteur de transcription peuvent différer comme on peut le voir pour les cibles de CtrA (Bellefontaine *et al.*, 2002). Ici nous comparons deux régions promotrices de gènes codant pour un facteur de transcription. Les facteurs de transcription ont tendance à s'autoréguler, ils ont donc plus de chance de retrouver le même motif d'un régulateur dans leurs régions promotrices de ces pCDSs.

3.1.5 Conclusion

Le seul programme efficace pour retrouver le S.A.F.T de CtrA est « *Dyad analysis* » car il prend en compte l'espacement entre les deux parties fixes. Cependant « *Dyad analysis* » souffre d'une faible puissance. « *Motif Sampler* » et « *Oligo analysis* » peuvent s'avérer efficaces si le cadre de recherche est plus petit que les monades composant le S.A.F.T. Il arrive aussi souvent que les deux monades composant la dyade soient identiques, et dans ce cas « *Motif Sampler* » et « *Oligo analysis* » peuvent se montrer adaptés à ce type de

recherche. La méthode par alignement n'est pas très confiante comme nous l'avons montré. Son utilisation nous semble donc peu efficace. L'usage le plus efficace semble «*Dyad analysis*» en premier lieu en utilisant les deux calibrations possibles.

Mais ces différents programmes sont efficaces si et seulement si on leur donne à tester des pCDSs que l'on sait co-régulées. Si on applique le test sur toutes les régions non codantes d'un génome donné, le nombre de résultats est trop grand que pour pouvoir sélectionner un S.A.F.T. ou un autre.

Pour effectuer un *screening* des résultats permettant d'écarter un grand nombre de faux positif, un critère de sélection qui serait propre au S.A.F.T. doit être déterminé.

La seconde partie de notre travail consiste en la recherche d'un tel critère, nous avons étudié dans quelle mesure la distribution des S.A.F.T. en amont et en aval du codon start pourrait être un critère efficace.

3.2 Chapitre 2 : Recherche d'un test spécifique à la détection des S.A.F.T. chez Brucella melitensis

3.2.1 Distribution du S.A.F.T. de CtrA chez Brucella melitensis

Tous les programmes testés au chapitre 1 permettent de retrouver un S.A.F.T. sur base de séquences *upstream* de pCDS co-régulées. Leur efficacité diminue lorsque toutes les séquences introduites ne possèdent pas ce S.A.F.T.. Nous nous sommes demandé s'il était possible de définir une méthode permettant de trouver des S.A.F.T. sans les rechercher parmi plusieurs séquences co-régulés. Pour y parvenir, il faudrait pouvoir définir un comportement particulier des S.A.F.T. Le premier critère important semble être la position des S.A.F.T. par rapport aux sites présumés d'initiations de la traduction (codon start). Comme le S.A.F.T. de CtrA est le seul connu chez *Brucella melitensis*, nous avons observé sa distribution amont-aval.

Grâce au programme « *Genomic scale* », nous pouvons connaître la position exacte de chaque S.A.F.T. de CtrA. Il suffit de définir le motif recherché et la zone de recherche, en gardant à l'esprit qu'une analyse *in silico* du génome génère des faux positifs mais aussi des faux négatifs si la recherche est trop stricte

Nous avons envisagé plusieurs motifs possibles, soit le motif complet (TTAAn(7)TTAAC), ce qui diminue la puissance mais augmente la confiance du test, soit l'inverse en demandant une comparaison à un nucléotide près sur tout le motif, ou en prenant seulement une partie du motif (TTAAn(7)TTAA). Notre choix s'est orienté vers une plus grande confiance au détriment de la puissance. Le motif testé est donc la dyade complète de CtrA, c'est-à-dire TTAAn(7)TTAAC.

Le cadre de recherche se limite à 300 nucléotides avant et après le codon start de chaque CDSs du génome de *Brucella melitensis*. Comme le montre l'analyse de distribution de 173 S.A.F.T. chez *E. coli* (Gralla and Collado-Vides, 1996), les sites éloignés de plus de 300

paires de base du codon start de la CDS sont des événements marginaux. De plus, nous savons aussi que les génomes procaryotiques possèdent en général peu de régions non codantes (seulement 13% du génome dans le cas *Brucella melitensis*) ce qui implique qu'une recherche sur une plus longue distance augmente fortement les chances de se retrouver dans la CDS qui précède ou suit la nôtre. Notre cadre de recherche sera donc de –300 à +300 par rapport au codon start pour une étude sur la distribution d'un motif.

Lorsque le programme *Genomic-scale* recherche le motif CtrA entre –300 et +300 par rapport au codon start de toutes les pCDSs de *Brucella melitensis*, il nous identifie le numéro des pCDSs possédant un ou plusieurs motifs composant le S.A.F.T. de CtrA ainsi que la position du début et de la fin de chaque motif CtrA. Ces S.A.F.T. de CtrA sont classés dans des intervalles de 25 nucléotides. (-300, -275, ..., 275, 300). Le classement s'effectue selon la position du milieu de chaque dyade CtrA, le zéro correspondant au codon start de la pCDS. Les résultats sont ensuite mis en graphique avec en abscisse la position et en ordonnée la fréquence d'apparition des dyades CtrA. (Voir figure 19).

L'analyse de ce graphique révèle que toutes les CDSs possédant le S.A.F.T. de CtrA complet se distribuent entre 0 et –225 à l'exception d'une seule en aval du codon start. Dans l'intervalle allant de 0 à –175, on remarque deux pics consécutifs, un pic à –175 et un pic à –100.

En conclusion de cette étude, nous remarquons que le motif CtrA se distribue préférentiellement en amont du codon start des pCDSs. Pour savoir si cette distribution est le fait du hasard ou une véritable distribution spécifique au motif CtrA, ce motif a été étudié chez d'autres espèces proches d'un point de vue évolutif.

3.2.2 Recherche du S.A.F.T. de CtrA chez d'autres espèces

Pour ce test, nous avons choisi deux espèces proches possédant le motif CtrA, à savoir *Caulobacter crescentus* et *Sinorhizobiom meliloti.* Comme contrôle négatif pour ce test de distribution du S.A.F.T. de CtrA nous avons choisi une espèce plus éloignée *(Bacillus subtilis)* ne possédant pas de régulation par CtrA car la pCDS codant pour CtrA n'a aucune séquence similaire chez *Bacillus subtilis.* Les tests qui suivent s'effectuent de manière identique à celui pratiqué chez *Brucella melitensis.*

3.2.2.1 Caulobacter crescentus

L'analyse de distribution du S.A.F.T. de CtrA chez *Caulobacter crescentus* nous a permis d'élaborer la figure 20. L'analyse de ce graphique révèle une forte distribution de CtrA en amont du codon start. Sur 18 S.A.F.T. de CtrA, 15 sont en amont et 3 sont en aval du codon start. Le ratio (c'est-à-dire le rapport amont/aval des fréquences d'apparition du S.A.F.T. de CtrA) est donc de 5 ce qui est assez élevé pour ne pas être considéré comme un fait du hasard (l'indice de χ^2 est de 8, $P< 0,005$) Nous retrouvons un pic de S.A.F.T. de CtrA à –100 comme chez *Brucella melitensis.*

3.2.2.2 Sinorhizobium melitoti

La distribution du S.A.F.T. de CtrA chez *Sinorhizobium meliloti* est totalement située dans la partie amont (voir figure 21). Le ratio est incalculable puisque la somme des fréquences de la partie aval est égale à zéro ; le rapport est alors infini. Pour pallier à cet inconvénient nous définissons un pseudo-poids pour tous les ratios ; si la somme des fréquences de la partie aval est égale à zéro, nous donnons à celle-ci la valeur 1. La distribution du S.A.F.T. de CtrA chez *Sinorhizobium meliloti* se retrouve exclusivement dans la partie amont, avec une occurrence de 34 (l'indice de χ^2 est de 34, $P< 0,001$).

En conclusion, chez les trois espèces régulées par CtrA, la différence amont/aval est hautement significative. Pour savoir si cette différence significative est toujours présente même chez des espèces non régulées par CtrA, nous l'avons testée chez *Bacillus subtilis.*

3.2.2.3 Bacillus subtilis

Bacillus subtilis n'a pas de pCDS codant pour CtrA. Toute la régulation commune *à C. crescentus, S. meliloti* et *Brucella melitensis* par CtrA n'est pas présente chez *B. subtilis*. La figure 22 nous indique que la distribution du S.A.F.T. de CtrA présente une distribution uniforme. Le ratio est de 1,16 et montre que la distribution n'est pas spécifique à une région ou une autre. Le test de χ^2 donne pour une distribution de 14 en amont et de 12 en aval avec un degré de liberté de 1, une valeur de 0,154 ce qui n'est pas significatif.

3.2.2.4 Conclusion

Nous avons observé que les organismes régulés par CtrA ont une distribution du S.A.F.T. de CtrA fortement concentrée en amont du codon start. Tandis que chez *Bacillus subtilis*, qui n'est pas régulé par CtrA, la distribution est aléatoire. Nous savons que le contenu en nucléotides G et C est plus élevé chez *C .crescentus, S. meliloti* et *Brucella melitensis* que chez *Bacillus subtilis* (voir les fichiers GENBANK relatif à ces génomes : AE008917.gbk et AE008918.gbk pour *Brucella melitensis*). Le motif du S.A.F.T. de CtrA ayant un contenu en AT élevé, nous pouvons penser qu'il y a une plus grande probabilité de retrouver le S.A.F.T. de CtrA par hasard dans les parties non codantes (aléatoire) que dans les parties codantes. Pour vérifier cette hypothèse, il nous suffit de prendre l'ensemble des dyades de même composition en ATGC et d'analyser leur distribution chez *C. crescentus, S. meliloti* et *Brucella melitensis*. Si toutes ces dyades anagrammes ont une distribution excentrée en amont, nous conclurons que le biais des codons justifie cette distribution.

3.2.3 Distribution des anagrammes au S.A.F.T. de CtrA

Pour ce test, il faut créer les anagrammes du S.A.F.T. de CtrA ; c'est à dire toutes les dyades de même espacement et dont le contenu en AGCT est identique à la dyade du S.A.F.T. étudié.

Ici les anagrammes sont établis grâce au programme de génération des anagrammes (voir « alicombiJF » dans matériels et méthodes). Ce programme génère tous les anagrammes possibles contenant un nombre précis de AGCT.

Pour chaque dyade, le programme crée tous les anagrammes possibles pour les deux monades, ensuite il combine les anagrammes de la monade 1 avec les anagrammes de la monade 2 ainsi qu'avec l'espacement fixe de la dyade d'origine.

3.2.3.1 Brucella melitensis

Suite à une analyse de distribution des anagrammes du S.A.F.T. de CtrA, la fréquence d'apparition des différents anagrammes en amont et en aval du codon start est connue. Sur base de ces résultats, nous pouvons calculer le ratio et l'indice de χ^2. Le tableau 14 présente toutes les dyades anagrammes possibles de la boîte CtrA ayant une fréquence attendue égale ou supérieure à cinq pour satisfaire les conditions d'applicabilité du test de χ^2. Le tableau se lit comme suit : la première colonne est la séquence de la dyade, AM est la fréquence en amont, AV la fréquence en aval, "Ratio" est le rapport amont/aval, F.TH est la fréquence théorique ((fréquence amont +fréquence aval)/2) et CHI.2 est l'indice χ^2 de la dyade avec un degré de liberté de 1. Pour rappel, le ratio est un rapport des fréquences amont sur les fréquences aval, si la fréquence aval est égale à zéro, nous l'augmentons à un.

Lors de l'analyse des résultats, nous remarquons que la boîte CtrA est en tête de liste des ratios et des indices de χ^2. Le S.A.F.T. de CtrA est donc, parmi tous les anagrammes possibles, celui qui a le plus grand ratio (c'est à dire 21) et son indice de χ^2 montre que la différence amont aval est hautement significative ($P<0,00005$). La figure 23 représente la distribution des anagrammes en fonction de leur indice de χ^2. L'axe des abscisses représente l'intervalle d'indice χ^2 (allant de 0-0,5 ; 0,5-1 ; ...) et l'axe des ordonnées représente le nombre de dyades anagrammes à la dyade du S.A.F.T. de CtrA comprises dans l'intervalle. On remarque que la distribution des dyades est centrée sur les valeurs basses de χ^2 entre 0 et 6,5. Il semble que seules quelques dyades ont un indice de χ^2 assez élevé pour

Dyade	AM	AV	Ratio	F.TH	Ind.Chi2	Dyade	AM	AV	Ratio	F.TH	Ind.Chi2
TTAAn(7)TTAAC	21	1	21,0	11	18,182	TTAAn(7)TAACT	4	1	4,0	2,5	1,800
ATATn(7)TTAAC	11	0	11,0	5,5	11,000	ATATn(7)TATAC	10	5	2,0	7,5	1,667
TAATn(7)TAATC	11	0	11,0	5,5	11,000	TAATn(7)AATTC	10	5	2,0	7,5	1,667
ATTAn(7)ATATC	12	1	12,0	6,5	9,308	ATATn(7)ATTAC	7	3	2,3	5	1,600
ATATn(7)CTATA	14	2	7,0	8	9,000	ATTAn(7)TTACA	7	3	2,3	5	1,600
TAATn(7)TTCAA	14	2	7,0	8	9,000	ATATn(7)TTCAA	19	12	1,6	15,5	1,581
TTAAn(7)ATATC	10	1	10,0	5,5	7,364	AATTn(7)AACTT	5	2	2,5	3,5	1,286
ATTAn(7)ATTAC	7	0	7,0	3,5	7,000	ATTAn(7)CATAT	5	2	2,5	3,5	1,286
ATTAn(7)TTAAC	9	1	9,0	5	6,400	TATAn(7)AACTT	5	2	2,5	3,5	1,286
TAATn(7)TATCA	9	1	9,0	5	6,400	AATTn(7)AATCT	13	8	1,6	10,5	1,190
TATAn(7)CATAT	11	2	5,5	6,5	6,231	ATATn(7)AACTT	9	5	1,8	7	1,143
ATTAn(7)CTAAT	6	0	6,0	3	6,000	ATATn(7)CATTA	14	9	1,6	11,5	1,087
TAATn(7)ATACT	6	0	6,0	3	6,000	AATTn(7)ACATT	10	6	1,7	8	1,000
AATTn(7)ATTAC	6	0	6,0	3	6,000	ATATn(7)ATTCA	10	6	1,7	8	1,000
TATAn(7)ACTTA	6	0	6,0	3	6,000	ATATn(7)TACAT	6	3	2,0	4,5	1,000
TATAn(7)ATACT	6	0	6,0	3	6,000	ATATn(7)TCATA	6	3	2,0	4,5	1,000
TATAn(7)CATTA	6	0	6,0	3	6,000	AATTn(7)AATTC	11	7	1,6	9	0,889
TATAn(7)TTAAC	6	0	6,0	3	6,000	TATAn(7)ACATT	7	4	1,8	5,5	0,818
ATATn(7)AATTC	14	4	3,5	9	5,556	AATTn(7)CATAT	8	5	1,6	6,5	0,692
AATTn(7)ATTCA	12	3	4,0	7,5	5,400	AATTn(7)TATAC	4	2	2,0	3	0,667
TAATn(7)TTAAC	10	2	5,0	6	5,333	ATATn(7)TTACA	4	2	2,0	3	0,667
AATTn(7)CAATT	15	5	3,0	10	5,000	TAATn(7)AATCT	4	2	2,0	3	0,667
ATATn(7)CAATT	15	5	3,0	10	5,000	TATAn(7)TCATA	4	2	2,0	3	0,667
AATTn(7)TAACT	5	0	5,0	2,5	5,000	TTAAn(7)TATAC	4	2	2,0	3	0,667
ATTAn(7)CTATA	5	0	5,0	2,5	5,000	ATTAn(7)TAATC	5	3	1,7	4	0,500
ATTAn(7)TTCAA	5	0	5,0	2,5	5,000	ATTAn(7)TCAAT	5	3	1,7	4	0,500
TATAn(7)TACAT	5	0	5,0	2,5	5,000	ATATn(7)AATCT	3	5	0,6	4	0,500

s'échapper du groupe formé au début du graphique. Dans le génome de *Brucella melitensis*, le S.A.F.T. de CtrA a le ratio et l'indice de χ^2 le plus élevé de toutes les dyades anagrammes au S.A.F.T. de CtrA. Nous allons maintenant étudier d'autres organismes régulés par CtrA pour déterminer s'ils ont une distribution analogue des anagrammes au S.A.F.T. de CtrA.

3.2.3.2 Caulobacter crescentus

Notre étude de position des anagrammes du S.A.F.T. de CtrA chez *Caulobacter crescentus* est effectuée comme précédemment, mais la fréquence d'apparition des dyades n'est pas assez élevée pour appliquer le test de χ^2 à toutes les dyades. Nous avons donc calculé l'indice χ^2 pour les 5 dyades ayant une fréquence théorique plus élevée ou égale à cinq. Pour les autres dyades anagrammes au S.A.F.T. de CtrA ayant une fréquence inférieure à cinq, la seule valeur permettant de juger leur distribution est le ratio amont/aval.

En observant le tableau 15, on remarque que le motif CtrA est toujours au-dessus du tableau avec un indice de χ^2 extrêmement élevé par rapport aux autres. La figure 24 montre la distribution des dyades en fonction de leur ratio (vu que χ^2 ne peut être calculé). On remarque une disposition analogue à celle observée chez *Brucella melitensis* même si elle est beaucoup plus groupée chez *Caulobacter crescentus*. Le S.A.F.T. de CtrA ferme la marche avec une autre dyade ATATn(7)CATAT, cette dyade est présente chez *Brucella melitensis* avec un indice χ^2 de 2.133 et l'indice est de 2.67 chez *Caulobacter crescentus*. Mais même si leur ratio est égal le test χ^2 montre que la distribution de CtrA est hautement plus significative que celle de ATATn(7)CATAT. Donc Le S.A.F.T. de CtrA a donc bien la dyade la plus significativement en amont.

3.2.3.3 Sinorhizobium melitoti

Lorsqu'on étudie la distribution des anagrammes de CtrA chez *Sinorhizobium meliloti* (voir tableau 16), le même constat s'impose : le motif du S.A.F.T. de CtrA possède l'indice χ^2 et le

TTAAn(7)CTAAT	5	0	5,0	2,5	5,000	AATTn(7)TAATC	6	4	1,5	5	0,400
ATATn(7)TCTAA	7	1	7,0	4	4,500	TATAn(7)AATTC	6	4	1,5	5	0,400
ATTAn(7)ATTCA	7	1	7,0	4	4,500	TATAn(7)ATATC	6	4	1,5	5	0,400
TTAAn(7)CATTA	7	1	7,0	4	4,500	ATATn(7)ATCTA	7	5	1,4	6	0,333
ATATn(7)TATCA	9	2	4,5	5,5	4,455	ATTAn(7)TATCA	7	5	1,4	6	0,333
TAATn(7)TCATA	9	2	4,5	5,5	4,455	ATATn(7)ATCAT	17	14	1,2	15,5	0,290
TTAAn(7)TCAAT	9	2	4,5	5,5	4,455	ATTAn(7)ACTAT	3	2	1,5	2,5	0,200
TAATn(7)CATAT	14	5	2,8	9,5	4,263	ATTAn(7)CTTAA	3	2	1,5	2,5	0,200
AATTn(7)TTCAA	20	9	2,2	14,5	4,172	TATAn(7)TATCA	3	2	1,5	2,5	0,200
ATATn(7)TCAAT	22	11	2,0	16,5	3,667	TTAAn(7)ATTCA	3	2	1,5	2,5	0,200
TAATn(7)ATTCA	6	1	6,0	3,5	3,571	ATTAn(7)AATTC	2	3	0,7	2,5	0,200
TATAn(7)CTAAT	6	1	6,0	3,5	3,571	AATTn(7)ATACT	4	3	1,3	3,5	0,143
TTAAn(7)CAATT	11	4	2,8	7,5	3,267	TTAAn(7)ATCAT	4	3	1,3	3,5	0,143
TATAn(7)CAATT	14	6	2,3	10	3,200	TTAAn(7)CATAT	4	3	1,3	3,5	0,143
TAATn(7)CAATT	9	3	3,0	6	3,000	TATAn(7)AATCT	3	4	0,8	3,5	0,143
ATTAn(7)CATTA	7	2	3,5	4,5	2,778	AATTn(7)ACTAT	5	4	1,3	4,5	0,111
TAATn(7)ATATC	7	2	3,5	4,5	2,778	AATTn(7)ATCTA	5	4	1,3	4,5	0,111
ATATn(7)CTTAA	5	1	5,0	3	2,667	AATTn(7)CATTA	5	4	1,3	4,5	0,111
TTAAn(7)AATCT	5	1	5,0	3	2,667	AATTn(7)TCAAT	5	4	1,3	4,5	0,111
TATAn(7)TTCAA	10	4	2,5	7	2,571	ATATn(7)ACTAT	5	4	1,3	4,5	0,111
TATAn(7)TCAAT	8	3	2,7	5,5	2,273	AATTn(7)ATATC	7	6	1,2	6,5	0,077
ATTAn(7)ATCAT	3	8	0,4	5,5	2,273	ATATn(7)ATATC	8	7	1,1	7,5	0,067
AATTn(7)ATTAC	11	5	2,2	8	2,250	ATATn(7)ACATT	9	8	1,1	8,5	0,059
ATATn(7)CATAT	19	11	1,7	15	2,133	TTAAn(7)TTCAA	10	10	1,0	10	0,000
AATTn(7)ATCAT	15	8	1,9	11,5	2,130	TAATn(7)TCAAT	9	9	1,0	9	0,000
TATAn(7)ATCAT	12	6	2,0	9	2,000	TATAn(7)CTATA	6	6	1,0	6	0,000
AATTn(7)CTTAA	6	2	3,0	4	2,000	AATTn(7)TATCA	5	5	1,0	5	0,000
AATTn(7)TACTA	4	1	4,0	2,5	1,800	AATTn(7)TCATA	4	4	1,0	4	0,000
ATTAn(7)ACATT	4	1	4,0	2,5	1,800	TAATn(7)ATCAT	4	4	1,0	4	0,000
TAATn(7)TACAT	4	1	4,0	2,5	1,800	AATTn(7)CTATA	3	3	1,0	3	0,000
TAATn(7)TTACA	4	1	4,0	2,5	1,800	TAATn(7)ACATT	3	3	1,0	3	0,000

Tableau 14 : Classement de toutes les dyades anagrammes du S.A.F.T. de CtrA chez *Brucella melitensis*. La première colonne représente la dyade après la fréquence amont, la fréquence aval, le ratio, la fréquence théorique, l'indice de χ^2.

ratio le plus élevé. Ce ratio et son indice de χ^2 s'élèvent tout les deux à 34, ce qui est beaucoup plus élevé que chez les autres organismes testés. La figure 25 montre la distribution des dyades en fonction de leur indice χ^2. Sur ce graphique, on remarque que la distribution des dyades anagrammes au S.A.F.T. de CtrA est plus proche du graphique homologue chez *Brucella melitensis* que *Caulobacter crescentus*. La plus grande ressemblance entre ces deux espèces peut aisément s'expliquer par la proximité évolutive entre *Brucella melitensis et S. meliloti*. Les génomes de *C. crescentus* et *Brucella melitensis* sont plus éloignés et donc il y a moins d'homologie. C'est pour cette raison que, durant ce travail, les analyses entre deux génomes se font entre *Brucella melitensis* et *Sinorhizobium meliloti* qui sont plus proches évolutivement (point 3.1.5).

3.2.3.4 Conclusion

En conclusion, hormis le S.A.F.T. de CtrA qui a toujours l'indice de χ^2 le plus élevé, aucune autre dyade ne semble être dans des valeurs élevées d'indice de χ^2. Nous pouvons donc conclure que même si les motifs riches en AT sont légèrement plus distribués en amont qu'en aval, la distribution du S.A.F.T. de CtrA ne peut pas s'expliquer par sa composition en AT.
Mais l'exclusion des régions codantes est peut être due aux codons qui composent TTAAC. Pour répondre à cette question nous avons étudié la distribution des trois sous-motifs composant le S.A.F.T. de CtrA : TTA, TAA, AAC.

3.2.4 Distribution de TTA, TAA, AAC chez Brucella melitensis

Par ce test, nous tentons d'établir si les sous-motifs de TTAAC sont exclus des régions codantes. Il se peut que les trois triplets qui composent TTAAC correspondent à des codons très peu présents dans les régions codantes. Lorsque ces triplets sont dans la phase de lecture, TTA code pour une leucine, AAC pour une aspargine et TAA est un codon stop. TAA ne peut donc pas apparaître dans la phase de lecture de la pCDS. Il a donc plus de chance d'apparaître

Dyade	AM	AV	Ratio	F.TH	CHI.2	Dyade	AM	AV	Ratio	F.TH	CHI.2
TTAAn(7)TTAAC	15	3	5	9	8.00	TATAn(7)TAACT	2	0	2	1	NO
ATATn(7)CATAT	5	1	5	3	2.67	TATAn(7)TATAC	2	0	2	1	NO
ATTAn(7)TCTAA	3	3	1	3	0.00	TATAn(7)TATCA	2	0	2	1	NO
AATTn(7)CAATT	3	2	1.5	2.5	0.20	AATTn(7)AACTT	1	0	1	0.5	NO
TAATn(7)AATCT	4	1	4	2.5	1.80	AATTn(7)ATACT	1	0	1	0.5	NO
AATTn(7)ATCAT	3	1	3	2	NO	AATTn(7)TACTA	1	0	1	0.5	NO
AATTn(7)CATAT	3	1	3	2	NO	AATTn(7)TCTAA	0	1	0	0.5	NO
ATATn(7)ATCTA	4	0	4	2	NO	AATTn(7)TTAAC	1	0	1	0.5	NO
ATATn(7)CATTA	4	0	4	2	NO	AATTn(7)TTACA	1	0	1	0.5	NO
ATATn(7)TACTA	3	1	3	2	NO	ATATn(7)AACTT	1	0	1	0.5	NO
TAATn(7)TCAAT	4	0	4	2	NO	ATATn(7)AATCT	1	0	1	0.5	NO
TATAn(7)AATTC	3	1	3	2	NO	ATATn(7)ACTAT	1	0	1	0.5	NO
TATAn(7)ATTCA	2	2	1	2	NO	ATATn(7)ACTTA	1	0	1	0.5	NO
TTAAn(7)CAATT	4	0	4	2	NO	ATATn(7)CAATT	0	1	0	0.5	NO
TTAAn(7)TCATA	4	0	4	2	NO	ATATn(7)CTAAT	1	0	1	0.5	NO
TTAAn(7)TTACA	4	0	4	2	NO	ATATn(7)CTTAA	1	0	1	0.5	NO
TTAAn(7)TTCAA	4	0	4	2	NO	ATATn(7)TAATC	1	0	1	0.5	NO
AATTn(7)CTATA	2	1	2	1.5	NO	ATATn(7)TATAC	1	0	1	0.5	NO
AATTn(7)TTCAA	3	0	3	1.5	NO	ATATn(7)TATCA	0	1	0	0.5	NO
ATATn(7)ATATC	3	0	3	1.5	NO	ATATn(7)TTAAC	1	0	1	0.5	NO
ATTAn(7)AACTT	2	1	2	1.5	NO	ATTAn(7)AATCT	1	0	1	0.5	NO
ATTAn(7)ATCAT	3	0	3	1.5	NO	ATTAn(7)TATAC	0	1	0	0.5	NO
ATTAn(7)TACTA	2	1	2	1.5	NO	ATTAn(7)TCAAT	1	0	1	0.5	NO
TAATn(7)TAACT	3	0	3	1.5	NO	ATTAn(7)TCATA	1	0	1	0.5	NO
TATAn(7)ATCAT	2	1	2	1.5	NO	ATTAn(7)TTAAC	1	0	1	0.5	NO
TTAAn(7)CATTA	1	2	0.5	1.5	NO	ATTAn(7)TTACA	1	0	1	0.5	NO
TTAAn(7)CTAAT	3	0	3	1.5	NO	TAATn(7)AATTC	0	1	0	0.5	NO
TTAAn(7)TAATC	3	0	3	1.5	NO	TAATn(7)ATCAT	1	0	1	0.5	NO
AATTn(7)AATTC	1	1	1	1	NO	TAATn(7)ATTAC	1	0	1	0.5	NO

dans la région non codante. Pour vérifier si ces éléments interviennent dans la distribution amont du S.A.F.T. de CtrA, nous avons analysé la distribution des triplets qui la composent (voir tableau 17).

La figure 26 nous montre la distribution générale des trois triplets. En abscisse est représentée la fréquence d'apparition et en ordonnée des intervalles de 25 nucléotides par rapport à la position 0 (codon start). Nous remarquons que la distribution de TTA et TAA est identique ce qui est normal puisque TTA sur le brin positif donne TAA dans le sens de lecture du brin négatif. Ensuite, on observe une diminution de la fréquence d'apparition de TAA et TTA lorsqu'on passe le codon start. Pour TTA et TAA le ratio est de 1.65, ce qui signifie qu'il y a trois motifs en amont pour deux en aval. Cela pourrait s'expliquer par l'importance de la phase de lecture. En effet TAA est un codon stop, donc il est exclu une fois sur trois dans la pCDS. AAC a un ratio de 1.11 et il n'est donc probablement pas spécifique de la région amont. Les codons composant la dyade CtrA sont une des raisons de cette distribution en amont. D'ailleurs beaucoup d'anagrammes au S.A.F.T. de CtrA n'ayant pas TTA ou TAA se retrouvent dans les valeurs basses d'indice de χ^2. Par contre, on observe une distribution uniforme du S.A.F.T. de CtrA chez *Bacillus subtilis* où TAA (codon stop) ne semble pas gêner la distribution du motif dans les régions codantes.

En conclusion, le rapport TA/GC ne peut pas expliquer une telle spécificité de la boîte CtrA pour la région amont. Nous aimerions comprendre pourquoi le S.A.F.T. de CtrA a toujours l'indice χ^2 le plus élevé. Pour savoir quelles parties du motif créent cette spécificité, nous avons testé les anagrammes d'un motif légèrement différent, à savoir TTAAN(7)TTAA. Si la distribution est identique avec ces anagrammes nous pourrons définir le nucléotide C comme non-important à la distribution.

AATTn(7)ACATT	1	1	1	1	NO	TAATn(7)TATCA	0	1	0	0.5	NO
AATTn(7)ATATC	2	0	2	1	NO	TAATn(7)TTCAA	1	0	1	0.5	NO
AATTn(7)ATTCA	2	0	2	1	NO	TATAn(7)ATACT	1	0	1	0.5	NO
AATTn(7)CTTAA	2	0	2	1	NO	TATAn(7)CATAT	0	1	0	0.5	NO
AATTn(7)TAACT	2	0	2	1	NO	TATAn(7)CTATA	0	1	0	0.5	NO
ATATn(7)AATTC	1	1	1	1	NO	TATAn(7)TACTA	0	1	0	0.5	NO
ATATn(7)ATTAC	2	0	2	1	NO	TATAn(7)TCATA	0	1	0	0.5	NO
ATATn(7)CTATA	2	0	2	1	NO	TATAn(7)TCTAA	1	0	1	0.5	NO
ATATn(7)TTACA	0	2	0	1	NO	TATAn(7)TTAAC	1	0	1	0.5	NO
ATATn(7)TTCAA	1	1	1	1	NO	TATAn(7)TTACA	0	1	0	0.5	NO
ATTAn(7)AATTC	2	0	2	1	NO	TTAAn(7)AACTT	1	0	1	0.5	NO
ATTAn(7)ATACT	1	1	1	1	NO	TTAAn(7)AATCT	0	1	0	0.5	NO
ATTAn(7)ATTAC	2	0	2	1	NO	TTAAn(7)ACATT	1	0	1	0.5	NO
ATTAn(7)ATTCA	2	0	2	1	NO	TTAAn(7)ATATC	1	0	1	0.5	NO
ATTAn(7)CAATT	2	0	2	1	NO	TTAAn(7)ATCTA	1	0	1	0.5	NO
ATTAn(7)CTTAA	2	0	2	1	NO	TTAAn(7)ATTCA	1	0	1	0.5	NO
ATTAn(7)TTCAA	2	0	2	1	NO	TTAAn(7)CTATA	1	0	1	0.5	NO
TAATn(7)ATACT	2	0	2	1	NO	TTAAn(7)CTTAA	1	0	1	0.5	NO
TAATn(7)CTTAA	2	0	2	1	NO	TTAAn(7)TACAT	1	0	1	0.5	NO
TAATn(7)TAATC	2	0	2	1	NO	TTAAn(7)TATAC	1	0	1	0.5	NO
TATAn(7)ACTAT	2	0	2	1	NO	TTAAn(7)TCAAT	1	0	1	0.5	NO

Tableau 15 : Classement de toutes les dyades anagrammes du S.A.F.T. de CtrA chez *Caulobacter crescentus*. La première colonne représente la dyade après la fréquence amont, la fréquence aval, le ratio, la fréquence théorique, l'indice de χ^2.

3.2.5 Distributions des anagrammes à TTAAn(7)TTAA

Par cette distribution, nous voulons vérifier si le nucléotide C du motif CtrA influence la distribution amont/aval du S.A.F.T. de CtrA. Pour répondre à cette question, nous avons testé tous les anagrammes possibles de TTAAn(7)TTAA. Ces anagrammes sont moins nombreux que pour le S.A.F.T. de CtrA puisqu'il existe seulement 6 possibilités pour la première monade et 6 pour la seconde. Les 36 dyades possibles sont testées grâce au programme « *Genomic scale* » pour connaître leurs positions respectives.

3.2.5.1 Brucella melitensis

Sur le tableau 18, on peut remarquer que TTAAn(7)TTAA a un indice de χ^2 de 23,15, ce qui est plus élevé que TTAAn(7)TTAAC (le motif complet du S.A.F.T. de CtrA). Cette différence minime de distribution amont/aval du S.A.F.T. de CtrA avec ou sans C corrobore les résultats expérimentaux. En effet, plusieurs gènes régulés par CtrA possèdent des sites de liaisons pour CtrA divergeant du site complet ; le nucléotide C ne fait pas toujours partie de l'élément cis de CtrA. Les analyses ultérieures chez *C. crescentus* et chez *S. meliloti* nous apportent plus de renseignements quant à l'importance du nucléotide C.

On remarque que la perte du nucléotide C dans la composition des anagrammes ne donne plus au S.A.F.T. de CtrA la position la plus élevée dans le tableau des résultats car deux dyades ont un ratio plus élevé : TTAAn(7)TAAT et ATTAn(7)TTAA. On remarque aisément que ces deux dyades sont l'image inversée l'une de l'autre, c'est à dire que si on lit la première dans le sens 5'3', l'autre dyade correspond à ce qu'on lirait dans le sens inverse. C'est pour ça que le ratio des deux dyades est égal : il y a 28 sites en amont pour chacune et seulement un en aval. Nous appellerons ces deux dyades des dyades complémentaires.

Lorsqu'on observe la figure 27, on ne remarque plus la distribution groupée similairement aux anagrammes du S.A.F.T. de CtrA chez *Brucella melitensis*. Les dyades semblent avoir des

dyade	AM	AV	Ratio	F.TH	CHI.2	dyade	AM	AV	Ratio	F.TH	CHI
TTAAn(7)TTAAC	34	0	34.00	17	34.00	ATATn(7)ACTAT	4	1	4.00	2.5	1.8
ATATn(7)ATTAC	20	4	5.00	12	10.67	ATTAn(7)AATTC	4	1	4.00	2.5	1.8
TTAAn(7)TCAAT	10	0	10.00	5	10.00	ATTAn(7)ATCTA	4	1	4.00	2.5	1.8
AATTn(7)ATTCA	12	1	12.00	6.5	9.31	ATTAn(7)CATAT	4	1	4.00	2.5	1.8
ATTAn(7)TAATC	9	0	9.00	4.5	9.00	TATAn(7)TCAAT	4	1	4.00	2.5	1.8
ATTAn(7)TTCAA	9	0	9.00	4.5	9.00	TTAAn(7)ATTCA	4	1	4.00	2.5	1.8
AATTn(7)AATTC	15	3	5.00	9	8.00	TTAAn(7)CAATT	4	1	4.00	2.5	1.8
AATTn(7)TTACA	8	0	8.00	4	8.00	ATATn(7)ACATT	7	3	2.33	5	1.6
TAATn(7)ACATT	7	0	7.00	3.5	7.00	TATAn(7)ATTCA	7	3	2.33	5	1.6
TAATn(7)CATAT	7	0	7.00	3.5	7.00	ATATn(7)AATTC	8	4	2.00	6	1.3
AATTn(7)TTCAA	17	5	3.40	11	6.55	ATATn(7)ATCAT	7	12	0.58	9.5	1.3
TATAn(7)ACATT	6	0	6.00	3	6.00	AATTn(7)ACTTA	5	2	2.50	3.5	1.2
TTAAn(7)CTAAT	6	0	6.00	3	6.00	AATTn(7)ATTAC	5	2	2.50	3.5	1.2
AATTn(7)CTAAT	8	1	8.00	4.5	5.44	AATTn(7)ACATT	6	3	2.00	4.5	1.0
TAATn(7)ATTCA	8	1	8.00	4.5	5.44	ATATn(7)TACAT	6	3	2.00	4.5	1.0
TATAn(7)CAATT	8	1	8.00	4.5	5.44	TAATn(7)AATTC	6	3	2.00	4.5	1.0
AATTn(7)CATAT	12	3	4.00	7.5	5.40	AATTn(7)AATCT	8	5	1.60	6.5	0.6
ATATn(7)CAATT	12	3	4.00	7.5	5.40	AATTn(7)ATATC	8	5	1.60	6.5	0.6
AATTn(7)CTATA	5	0	5.00	2.5	5.00	AATTn(7)CTTAA	4	2	2.00	3	0.6
AATTn(7)CAATT	11	3	3.67	7	4.57	ATATn(7)ATCTA	4	2	2.00	3	0.6
AATTn(7)ATCTA	15	6	2.50	10.5	3.86	TAATn(7)ATCAT	4	2	2.00	3	0.6
ATATn(7)TATCA	8	2	4.00	5	3.60	AATTn(7)ATCAT	15	11	1.36	13	0.6
ATATn(7)CTTAA	6	1	6.00	3.5	3.57	AATTn(7)TCATA	5	3	1.67	4	0.5
ATATn(7)TATAC	6	1	6.00	3.5	3.57	TAATn(7)TCATA	5	3	1.67	4	0.5
TTAAn(7)ATCAT	6	1	6.00	3.5	3.57	ATATn(7)AATCT	6	4	1.50	5	0.4
AATTn(7)AACTT	5	1	5.00	3	2.67	TATAn(7)ATATC	6	4	1.50	5	0.4
TTAAn(7)AATTC	5	1	5.00	3	2.67	ATATn(7)CATAT	10	8	1.25	9	0.2
TTAAn(7)CATAT	5	1	5.00	3	2.67	ATATn(7)ATATC	3	2	1.50	2.5	0.2
AATTn(7)TATCA	8	3	2.67	5.5	2.27	ATTAn(7)ATATC	3	2	1.50	2.5	0.2
AATTn(7)TACAT	6	2	3.00	4	2.00	TATAn(7)AACTT	3	2	1.50	2.5	0.2

indices de χ^2 complètement différents et aucune dyade n'est significativement différente par rapport aux autres. Le plus haut pic de fréquence se trouve entre 19 et 19,5. Cela est peut-être dû à la disparition du C qui limitait ce groupe aux basses valeurs d'indices de χ^2.

En conclusion, il n'y a pas une grande différence entre la distribution en amont et en aval du codon start du S.A.F.T. de CtrA avec ou sans le C. La seule différence se voit à la distribution des anagrammes à TTAAN(7)TTAA, En effet, deux dyades ont un ratio et un indice de χ^2 plus élevé que le S.A.F.T. de CtrA sans C. Si ces deux dyades complémentaires ont toujours un indice de χ^2 plus élevé que le S.A.F.T. deCtrA sans C chez *Sinorhizobium meliloti* et *Caulobacter crescentus*, alors nous pourrions émettre l'hypothèse qu'il s'agit aussi de S.A.F.T..

3.2.5.2 Caulobacter crescentus

Ici, la dyade avec l'indice de χ^2 le plus élevé est CtrA (voir tableau 19) sans C. Mais juste derrière on retrouve les dyades complémentaires TTAAn(7)TAAT et ATTAn(7)TTAA. Ces dyades complémentaires n'ont plus l'indice de χ^2 le plus élevé, leurs indices de χ^2 est de 9. Même si l'indice de χ^2 est plus élevé pour le S.A.F.T. de CtrA sans C, le ratio des dyades complémentaires est plus élevé. De plus les deux indices de χ^2 entre TTAAn(7)TTAA et les dyades complémentaires sont proches (9,8 et 9).

La figure 28 montre en ordonnée la fréquence d'apparition des dyades anagrammes à TTAAN(7)TTAA et en abscisse les valeurs de ratio de ces dyades (classées selon les ratio car la fréquence d'apparition est trop basse pour la calculer). Même si pour ce graphique, beaucoup de valeurs sont centrées sur le début, on ne retrouve pas la distribution des dyades anagrammes du S.A.F.T. de CtrA chez *Caulobacter crescentus*.

En conclusion, le nucléotide C ne semble pas avoir d'importance chez *Caulobacter crescentus* car l'indice de χ^2 est proche entre le S.A.F.T. de CtrA avec et sans C. Juste derrière CtrA sans

ATTAn(7)ACATT	6	2	3.00	4	2.00	ATATn(7)TCAAT	4	3	1.33	3.5	0.1
TATAn(7)AATTC	6	2	3.00	4	2.00	ATATn(7)TTCAA	4	3	1.33	3.5	0.1
TATAn(7)ATCAT	2	6	0.33	4	2.00	ATTAn(7)ATCAT	6	6	1.00	6	0.0

Tableau 16 : Classement de toutes les dyades anagrammes du S.A.F.T. de CtrA chez *Sinorhizobiom meliloti*. La première colonne représente la dyade après la fréquence amont, la fréquence aval, le ratio, la fréquence théorique, l'indice de χ^2.

dyade	AM	AV	F.TH	Ratio	CHI2
AAC	17268	15552	16410	1.110339506	89.7213894
TAA	10166	6173	8169.5	1.646849182	975.8277128
TTA	10166	6173	8169.5	1.646849182	975.8277128

Tableau 17 : Distribution des sous motifs de TTAAC. La première colonne représente la dyade après la fréquence amont, la fréquence aval, le ratio, la fréquence théorique, l'indice de χ^2.

C, on retrouve les dyades complémentaires. Les dyades complémentaires ont un ratio plus élevé mais un indice de χ^2 plus faible. Néanmoins, elles restent en haut du tableau avec un indice de χ^2 qui prouve que leur distribution amont est hautement significative (l'indice est de 9) que ce soit chez *Brucella melitensis* ou *Caulobacter crescentus*. L'analyse de la distribution des anagrammes chez *Sinorhizobium meliloti* nous permettra de définir l'importance des dyades complémentaires.

3.2.5.3 Sinorhizobium melitoti

Dans le cas de *Sinorhizobium melitoti*, les dyades complémentaires n'ont plus un indice de χ^2 aussi élevé (voir tableau 20) que pour les études précédentes. Il y a donc une grande différence entre ce que nous observons ici et ce que l'on voit chez les deux autres espèces étudiées en ce qui concerne les dyades complémentaires. La dyade du S.A.F.T. de CtrA sans C reste la plus distribuée en amont car le ratio est de 32 et l'indice de χ^2 est de 29,12.
La dyade CtrA sans C a quasiment le même indice de χ^2 que CtrA. La distribution change très peu entre les deux groupes d'anagrammes (avec et sans C). La figure 29 montre une distribution analogue à *Brucella melitensis* pour les anagrammes de TTAAN(7)TTAA, c'est-à-dire aucun groupe de basses valeurs d'indice de χ^2 et une distribution quasi uniforme des dyades. Le seul point commun entre la figure 29 et 25 c'est l'éloignement de la dyade du S.A.F.T. de CtrA (avec C pour le premier graphique et sans pour le second) par rapport aux autres dyades.

3.2.5.4 Conclusion

En conclusion de ces trois analyses, on peut remarquer que le nucléotide C n'influence pas la distribution du S.A.F.T. de CtrA car ce S.A.F.T. de CtrA avec et sans C a une distribution

dyade	AM	AV	F.TH	ratio	CHI2	dyade	AM	AV	F.TH	ratio	CHI2
ATTAn(7)TTAA	28	1	14.5	28.00	25.14	TAATn(7)TTAA	19	4	11.5	4.75	9.78
TTAAn(7)TAAT	28	1	14.5	28.00	25.14	TTAAn(7)ATTA	19	4	11.5	4.75	9.78
TTAAn(7)TTAA	26	1	13.5	26.00	23.15	AATTn(7)AATT	24	7	15.5	3.43	9.32
AATTn(7)TTAA	26	2	14	13.00	20.57	AATTn(7)TATA	33	13	23	2.54	8.70
TTAAn(7)AATT	26	2	14	13.00	20.57	TATAn(7)AATT	33	13	23	2.54	8.70
AATTn(7)ATAT	54	17	35.5	3.18	19.28	ATATn(7)TATA	39	17	28	2.29	8.64
ATATn(7)AATT	54	17	35.5	3.18	19.28	TATAn(7)ATAT	39	17	28	2.29	8.64
ATATn(7)TTAA	27	3	15	9.00	19.20	TATAn(7)TTAA	20	6	13	3.33	7.54
TTAAn(7)ATAT	27	3	15	9.00	19.20	TTAAn(7)TATA	20	6	13	3.33	7.54
ATTAn(7)ATTA	22	1	11.5	22.00	19.17	TATAn(7)TATA	14	3	8.5	4.67	7.12
TAATn(7)TAAT	22	1	11.5	22.00	19.17	AATTn(7)TAAT	20	12	16	1.67	2.00
ATATn(7)TAAT	33	6	19.5	5.50	18.69	ATTAn(7)AATT	20	12	16	1.67	2.00
ATTAn(7)ATAT	33	6	19.5	5.50	18.69	ATTAn(7)TATA	17	10	13.5	1.70	1.81
AATTn(7)ATTA	38	13	25.5	2.92	12.25	TATAn(7)TAAT	17	10	13.5	1.70	1.81
TAATn(7)AATT	38	13	25.5	2.92	12.25	ATATn(7)ATTA	22	16	19	1.38	0.95
TAATn(7)TATA	21	4	12.5	5.25	11.56	TAATn(7)ATAT	22	16	19	1.38	0.95
TATAn(7)ATTA	21	4	12.5	5.25	11.56	ATATn(7)ATAT	19	14	16.5	1.36	0.76
TAATn(7)ATTA	14	1	7.5	14.00	11.27	ATTAn(7)TAAT	6	6	6	1.00	0.00

Tableau 18 : Classement de toutes les dyades anagrammes à TTAAn(7)TTAA chez *Brucella melitensis*. La première colonne représente la dyade après la fréquence amont, la fréquence aval, la fréquence théorique, le ratio, l'indice de χ^2.

similaire. Par contre, les anagrammes de CtrA sans C ont des indices de χ^2 plus élevés. Les dyades complémentaires ont un indice de χ^2 assez élevé par rapport aux autres anagrammes chez *Brucella melitensis* et *Caulobacter crescentus*. Si ces dyades complémentaires sont des S.A.F.T. réels, alors elles doivent se trouver devant deux ou plusieurs pCDSs co-régulées. Pour vérifier si ces dyades sont des S.A.F.T. potentiels, nous avons demandé au programme « Genomic scale » de donner, pour chacun des trois génomes, la fonction prédite des pCDSs possédant les dyades complémentaires entre –300 et 0 par rapport au codon start. Si parmi les trois génomes (*Brucella melitensis, Sinorhizobiom meliloti* et *Caulobacter crescentus),* on retrouve deux ou plusieurs gènes orthologues, nous pourront émettre l'hypothèse que les dyades complémentaires sont les S.A.F.T. d'un F.T. qui régule ces gènes orthologues. Voici l'ensemble des pCDSs possédant les dyades complémentaires ainsi que les fonctions de ces pCDSs (tableau 21).

BMEI0244	1	TRANSALDOLASE
BMEI0245	1	PRIMOSOMAL PROTEIN N'
BMEI0353	1	PROLINE/BETAINE TRANSPORTER
BMEI0354	1	Hypothetical Membrane Spanning Protein
BMEI0365	1	BIOPOLYMER TRANSPORT EXBB PROTEIN
BMEI0668	1	CALCIUM BINDING PROTEIN
BMEI0677	1	hypothetical protein
BMEI0738	1	hypothetical protein
BMEI0739	1	INTEGRAL MEMBRANE PROTEIN (Rhomboid family)
BMEI0840	1	LEXA REPRESSOR
BMEI0913	1	PENICILLIN-BINDING PROTEIN 6 (D-ALANYL-D-ALANINE CARBOXYPEPTIDASE FRACTION C)
BMEI1079	2	LIPOPROTEIN NLPD
BMEI1305	1	PORIN
BMEI1646	1	ACRIFLAVIN RESISTANCE PROTEIN E
BMEI1866	1	hypothetical protein
BMEI1874	1	Hypothetical Protein
BMEII0158	1	TWO COMPONENT RESPONSE REGULATOR
BMEII0229	1	hypothetical protein
BMEII0421	1	hypothetical protein
BMEII0422	1	FRUCTOSE-1,6-BISPHOSPHATASE
BMEII0651	1	hypothetical protein
BMEII0812	2	POLYPEPTIDE DEFORMYLASE
BMEII0853	1	TWO COMPONENT RESPONSE REGULATOR
BMEII0879	1	PUTATIVE CYTOCHROME P450 YJIB
BMEII0880	1	ACETATE KINASE
BMEII1041	1	CYTOCHROME B561
CC0764	1	chemotaxis protein CheW; identified by match to protein family HMM

dyade	AM	AV	F.TH	Ratio	CHI2	dyade	AM	AV	F.TH	Ratio	CHI2
TTAAn(7)TTAA	17	3	10	5.67	9.8	TATAn(7)TATA	4	0	2	4.00	NO
ATTAn(7)TTAA	9	0	4.5	9.00	9	TTAAn(7)ATTA	2	2	2	1.00	NO
TTAAn(7)TAAT	9	0	4.5	9.00	9	ATATn(7)ATTA	3	0	1.5	3.00	NO
AATTn(7)TAAT	7	1	4	7.00	4.5	ATTAn(7)ATTA	3	0	1.5	3.00	NO
AATTn(7)TATA	7	1	4	7.00	4.5	TAATn(7)ATAT	3	0	1.5	3.00	NO
ATATn(7)ATAT	7	1	4	7.00	4.5	TAATn(7)TAAT	3	0	1.5	3.00	NO
ATTAn(7)AATT	7	1	4	7.00	4.5	AATTn(7)TTAA	1	1	1	1.00	NO
TATAn(7)AATT	7	1	4	7.00	4.5	ATATn(7)TTAA	2	0	1	2.00	NO
AATTn(7)ATTA	3	2	2.5	1.50	0.2	ATTAn(7)TATA	1	1	1	1.00	NO
TAATn(7)AATT	3	2	2.5	1.50	0.2	TATAn(7)TAAT	1	1	1	1.00	NO
AATTn(7)AATT	2	2	2	1.00	NO	TATAn(7)TTAA	2	0	1	2.00	NO
AATTn(7)ATAT	3	1	2	3.00	NO	TTAAn(7)AATT	1	1	1	1.00	NO
ATATn(7)AATT	3	1	2	3.00	NO	TTAAn(7)ATAT	2	0	1	2.00	NO
ATATn(7)TATA	4	0	2	4.00	NO	TTAAn(7)TATA	2	0	1	2.00	NO
TAATn(7)ATTA	3	1	2	3.00	NO	ATATn(7)TAAT	1	0	0.5	1.00	NO
TAATn(7)TTAA	2	2	2	1.00	NO	ATTAn(7)ATAT	1	0	0.5	1.00	NO
TATAn(7)ATAT	4	0	2	4.00	NO	ATTAn(7)TAAT	1	0	0.5	1.00	NO

Tableau 19 : Classement de toutes les dyades anagrammes à TTAAn(7)TTAA chez *Caulobacter crescentus*. La première colonne représente la dyade après la fréquence amont, la fréquence aval, la fréquence théorique, le ratio, l'indice de χ^2.

CC0870	1	hypothetical protein; identified by Glimmer2; putative
CC0903	1	conserved hypothetical protein; identified by Glimmer2; putative
CC1891	1	pentapeptide repeat family protein; identified by match to protein family HMM
CC2287	2	TonB-dependent receptor; similar to GP:9187833; identified by sequence similarity; putative
CC3002	1	oxidoreductase, aldo/keto reductase family; identified by match to protein family HMM
CC3003	1	hypothetical protein; identified by Glimmer2; putative
CC3035	1	cell cycle transcriptional regulator CtrA; identified by match to protein family HMM
SMa2237	1	hypothetical protein; glimmer prediction
SMa2239	1	conserved hypothetical protein; glimmer prediction; similar to hypothetical protein Rv2734,Mycobacterium tuberculosis, E70506
pncA	1	PROBABLE PYRAZINAMIDASE/NICOTINAMIDASE (INCLUDES: PYRAZINAMIDASE, NICOTINAMIDASE) PROTEIN; Product confidence : probable Gene name confidence : probable predicted by Codon_usage predicted byHomology predicted by FrameD
SMc02072	1	HYPOTHETICAL TRANSMEMBRANE PROTEIN; Product confidence : hypothetical Gene name confidence : Hypothetical predicted by Codon_usage predicted by FrameD
fbaB	1	PROBABLE FRUCTOSE-BISPHOSPHATE ALDOLASE CLASS I PROTEIN; Product confidence : probable Gene name confidence : putative predicted by Codon_usage predicted by Homology predicted by FrameD
SMb21079	1	putative transcriptional regulator protein; Product confidence : putative
SMb21080	2	putative response regulator protein; Product confidence : putative
exoY	1	Galactosyltransferase protein

Tableau 21 : ce tableau montre les différentes pCDSs possédant les dyades complémentaires entre –300 et 0 chez *Brucella melitensis, Caulobacter crescentus* et *Sinorhizobiom meliloti.* ainsi que le nombre de fois qu'elles apparaissent et la fonction prédite de la protéine encodée par la pCDS.

Nous avons comparé toutes les CDSs groupées (c'est-à-dire qui sont très proches l'une de l'autre) de Brucella melitensis avec la banque de données « nr » dans l'espoir de retrouver comme pCDSs similaires, les pCDSs présentes dans la liste chez *Caulobacter crescentus* et *Sinorhizobiom meliloti*. La comparaison par BLASTp n'a montré aucune similitude avec les pCDSs de cette liste. Néanmoins, les pCDS BMEI0738 et BMEI0739 sont respectivement similaires aux pCDS SMc1000 et SMc1001 chez *Sinorhizobiom meliloti* ; sans pour autant partager le même motif des dyades complémentaires.

Nous avons aussi testé d'autres pCDS. La paire de pCDSs possédant les dyades complémentaires chez *Sinorhizobiom meliloti* (SMb21079 et SMb21080) nous semblait intéressante car les fonctions prédites semblaient se compléter (« transcriptional regulator protein » et « putative response regulator » ("DATA" site RSA tools). La pCDS SMb21080

Dyade	AM	AV	F.TH	Ratio	CHI2	Dyade	AM	AV	F.TH	Ratio	CHI2
TTAAn(7)TTAA	32	1	16.5	32.00	29.12	AATTn(7)TATA	14	3	8.5	4.67	7.12
AATTn(7)TTAA	21	2	11.5	10.50	15.70	TATAn(7)AATT	14	3	8.5	4.67	7.12
TTAAn(7)AATT	21	2	11.5	10.50	15.70	ATTAn(7)TAAT	7	0	3.5	7.00	7.00
ATATn(7)ATTA	25	5	15	5.00	13.33	ATTAn(7)TTAA	9	1	5	9.00	6.40
TAATn(7)ATAT	25	5	15	5.00	13.33	TTAAn(7)TAAT	9	1	5	9.00	6.40
ATTAn(7)ATTA	11	0	5.5	11.00	11.00	TATAn(7)TTAA	5	0	2.5	5.00	5.00
TAATn(7)TAAT	11	0	5.5	11.00	11.00	TTAAn(7)TATA	5	0	2.5	5.00	5.00
ATATn(7)TTAA	13	1	7	13.00	10.29	TAATn(7)ATTA	8	2	5	4.00	3.60
TTAAn(7)ATAT	13	1	7	13.00	10.29	ATATn(7)ATAT	7	2	4.5	3.50	2.78
AATTn(7)TAAT	14	2	8	7.00	9.00	ATTAn(7)TATA	7	2	4.5	3.50	2.78
ATTAn(7)AATT	14	2	8	7.00	9.00	TATAn(7)TAAT	7	2	4.5	3.50	2.78
AATTn(7)AATT	21	6	13.5	3.50	8.33	ATATn(7)TAAT	13	6	9.5	2.17	2.58
AATTn(7)ATAT	29	11	20	2.64	8.10	ATTAn(7)ATAT	13	6	9.5	2.17	2.58
ATATn(7)AATT	29	11	20	2.64	8.10	TAATn(7)TATA	4	2	3	2.00	0.67
AATTn(7)ATTA	17	4	10.5	4.25	8.05	TATAn(7)ATTA	4	2	3	2.00	0.67
TAATn(7)AATT	17	4	10.5	4.25	8.05	TAATn(7)TTAA	3	0	1.5	3.00	NO
ATATn(7)TATA	27	10	18.5	2.70	7.81	TTAAn(7)ATTA	3	0	1.5	3.00	NO
TATAn(7)ATAT	27	10	18.5	2.70	7.81	TATAn(7)TATA	1	0	0.5	1.00	NO

Tableau 20 : Classement de toutes les dyades anagrammes à TTAAn(7)TTAA chez *Sinorhizobiom meliloti*. La première colonne représente la dyade après la fréquence amont, la fréquence aval, la fréquence théorique, le ratio, l'indice de χ^2.

est similaire à la pCDS BMEII0853 dont la fonction est « two component response regulator ». Ces deux pCDSs possèdent les dyades complémentaires dans leurs régions amont. Cependant nous n'avons pas assez d'éléments pour conclure que les dyades complémentaires sont les S.A.F.T. du F.T. qui régule ces gènes.

Le nucléotide C du motif ne change pas beaucoup cette distribution. Au contraire le S.A.F.T. de CtrA sans C est plus présent en amont que le S.A.F.T. de CtrA avec C. Donc, jusqu'ici, ni la composition de la dyade ni le biais des codons ne peuvent entièrement expliquer cette distribution amont du S.A.F.T. de CtrA. Ainsi, nous émettons l'hypothèse que l'espacement fixe de 7 nucléotides délimite CtrA dans la partie amont des pCDSs. Pour vérifier cette hypothèse nous allons tester tous les anagrammes possibles de TTAAN(8)TTAAC (espacement choisi arbitrairement). Nous saurons alors si c'est l'espacement qui délimite cette distribution amont de CtrA.

3.2.6 Distributions des anagrammes à TTAAn(8)TTAAC

Dans ce test, nous générons toutes les dyades anagrammes au S.A.F.T. de CtrA en changeant l'espacement entre les deux monades qui la compose. Si cet espacement est important, la modification apportée aux anagrammes devrait changer l'ordre de disposition des dyades dans le tableau de ratio et les indices de χ^2 devraient diminuer. Nous pourrons ainsi établir si, oui ou non, l'espacement fixe est important pour la distribution du S.A.F.T. de CtrA.

3.2.6.1 Brucella melitensis

De manière générale dans le tableau 22, on observe directement une diminution des ratios et des indices de χ^2 même si il y a toujours une distribution plus élevée en amont qu'en aval. Cette distribution amont est due au haut contenu en AT du motif et aux sous-motifs composant les différents anagrammes. Mais même si la tendance générale du graphique montre une distribution en amont, aucun motif ne s'échappe du lot comme c'était le cas pour

dyade	AM	AV	Ratio	F.TH	CHI2	Dyade	AM	AV	Ratio	F.TH	CHI2
TTAAn(8)CAATT	13	0	13.00	6.5	13.00	ATTAn(8)ATTAC	6	2	3.00	4	2.00
TATAn(8)ATTAC	12	0	12.00	6	12.00	TATAn(8)CATTA	6	2	3.00	4	2.00
ATATn(8)TCTAA	10	0	10.00	5	10.00	TATAn(8)CTATA	6	2	3.00	4	2.00
ATTAn(8)TTCAA	15	2	7.50	8.5	9.94	TATAn(8)TATAC	6	2	3.00	4	2.00
ATTAn(8)CATTA	12	1	12.00	6.5	9.31	AATTn(8)CTATA	4	1	4.00	2.5	1.80
TATAn(8)TCAAT	11	1	11.00	6	8.33	ATATn(8)TAACT	4	1	4.00	2.5	1.80
AATTn(8)ATTCA	13	2	6.50	7.5	8.07	ATTAn(8)ACTAT	4	1	4.00	2.5	1.80
ATTAn(8)TTAAC	8	0	8.00	4	8.00	ATTAn(8)TCATA	4	1	4.00	2.5	1.80
ATATn(8)TATAC	7	0	7.00	3.5	7.00	TTAAn(8)ACATT	4	1	4.00	2.5	1.80
TATAn(8)CTTAA	7	0	7.00	3.5	7.00	TAATn(8)TCAAT	10	5	2.00	7.5	1.67
TAATn(8)ATTCA	9	1	9.00	5	6.40	AATTn(8)ATATC	14	8	1.75	11	1.64
AATTn(8)TCAAT	15	4	3.75	9.5	6.37	AATTn(8)ACTAT	7	3	2.33	5	1.60
TTAAn(8)AATTC	11	2	5.50	6.5	6.23	ATATn(8)CATAT	8	4	2.00	6	1.33
TTAAn(8)TTCAA	11	2	5.50	6.5	6.23	ATATn(8)TTAAC	8	4	2.00	6	1.33
TAATn(8)TATCA	6	0	6.00	3	6.00	TAATn(8)ACATT	8	4	2.00	6	1.33
TATAn(8)TATCA	6	0	6.00	3	6.00	AATTn(8)CTAAT	5	2	2.50	3.5	1.29
TATAn(8)TTACA	6	0	6.00	3	6.00	ATATn(8)AATCT	5	2	2.50	3.5	1.29
ATATn(8)ATTAC	14	4	3.50	9	5.56	ATATn(8)TTACA	5	2	2.50	3.5	1.29
AATTn(8)TTACA	8	1	8.00	4.5	5.44	TATAn(8)CAATT	5	2	2.50	3.5	1.29
ATATn(8)ACTTA	8	1	8.00	4.5	5.44	TTAAn(8)TAATC	5	2	2.50	3.5	1.29
TATAn(8)TTCAA	12	3	4.00	7.5	5.40	TAATn(8)ATATC	9	5	1.80	7	1.14
ATTAn(8)AATCT	10	2	5.00	6	5.33	ATTAn(8)CAATT	6	3	2.00	4.5	1.00
ATTAn(8)TATCA	7	1	7.00	4	4.50	ATTAn(8)TCAAT	6	3	2.00	4.5	1.00
TAATn(8)ATCAT	7	1	7.00	4	4.50	ATATn(8)ACATT	12	8	1.50	10	0.80
TATAn(8)AACTT	7	1	7.00	4	4.50	AATTn(8)CAATT	8	5	1.60	6.5	0.69
TTAAn(8)TATCA	7	1	7.00	4	4.50	TAATn(8)CAATT	8	5	1.60	6.5	0.69
AATTn(8)TTAAC	9	2	4.50	5.5	4.45	TAATn(8)CATAT	8	5	1.60	6.5	0.69
TATAn(8)AATTC	9	2	4.50	5.5	4.45	AATTn(8)TCTAA	4	2	2.00	3	0.67
TATAn(8)ATATC	9	2	4.50	5.5	4.45	TAATn(8)AATCT	4	2	2.00	3	0.67
ATATn(8)CAATT	21	10	2.10	15.5	3.90	TAATn(8)ACTAT	4	2	2.00	3	0.67
AATTn(8)TCATA	10	3	3.33	6.5	3.77	TTAAn(8)ATCAT	4	2	2.00	3	0.67

le S.A.F.T. de CtrA chez *Brucella melitensis*. Donc l'espacement du S.A.F.T. de CtrA est en partie responsable de cette distribution amont.

Si on observe la figure 30, on remarque une distribution semblable à celle vue précédemment chez *Brucella melitensis*. Pour les anagrammes de TTAAn(7)TTAA, la distribution était différente par rapport au S.A.F.T. de CtrA. Ici, le nucléotide C réapparaît et la distribution ressemble à celle du S.A.F.T. de CtrA dans le même organisme. On peut donc en déduire que le nucléotide C présent dans le motif est responsable de la distribution groupée du début du graphique.

3.2.6.2 Caulobacter crescentus

Chez *Caulobacter crescentus*, les fréquences d'apparitions des dyades sont trop faibles que pour calculer un indice de χ^2. Le tableau 23 présente les dyades anagrammes à TTAANNNNNNNNTTAAC avec un indice de χ^2 lorsqu'il est possible de le calculer. La figure 31 ressemble d'avantage à la distribution des anagrammes de CtrA chez *Caulobacter crescentus*, ce qui montre que le C du motif est responsable de ce groupement de début de graphique.

En conclusion, un espacement de plus change totalement l'ordre des anagrammes d'une espèce à l'autre. Aucun anagramme ne reste à une valeur élevée d'indice χ^2 comparable à CtrA. L'espacement de sept nucléotides du S.A.F.T. de CtrA est une des raisons de sa distribution concentrée en amont.

3.2.6.3 Sinorhizobium melitoti

Les résultats présentés dans le tableau 24 montre le même constat que pour *Caulobacter crescentus*, aucune dyade ne garde une place élevée chez les trois espèces. Néanmoins la dyade TTAAN(8)CAATT a l'indice de χ^2 le plus élevé chez *Brucella melitensis et*

ATATn(8)TATCA	8	2	4.00	5	3.60	ATATn(8)ATCAT	16	12	1.33	14	0.5
TATAn(8)CATAT	8	2	4.00	5	3.60	AATTn(8)TTCAA	10	7	1.43	8.5	0.5
TTAAn(8)ATATC	8	2	4.00	5	3.60	AATTn(8)AATCT	5	3	1.67	4	0.5
AATTn(8)TACTA	6	1	6.00	3.5	3.57	ATTAn(8)TACAT	5	3	1.67	4	0.5
TAATn(8)TAACT	6	1	6.00	3.5	3.57	TATAn(8)TAATC	5	3	1.67	4	0.5
TTAAn(8)ATTAC	6	1	6.00	3.5	3.57	ATTAn(8)TATAC	6	4	1.50	5	0.4
TTAAn(8)TTAAC	6	1	6.00	3.5	3.57	AATTn(8)AATTC	13	10	1.30	11.5	0.3
AATTn(8)CATAT	13	5	2.60	9	3.56	AATTn(8)CATTA	7	5	1.40	6	0.3
ATATn(8)AATTC	16	7	2.29	11.5	3.52	ATATn(8)AACTT	7	5	1.40	6	0.3
AATTn(8)TAATC	9	3	3.00	6	3.00	ATATn(8)CTATA	5	7	0.71	6	0.3
ATATn(8)ATACT	9	3	3.00	6	3.00	ATATn(8)TCAAT	16	13	1.23	14.5	0.3
ATATn(8)CATTA	9	3	3.00	6	3.00	AATTn(8)ATCAT	6	8	0.75	7	0.2
TAATn(8)AATTC	9	3	3.00	6	3.00	AATTn(8)TAACT	3	2	1.50	2.5	0.2
ATATn(8)TTCAA	15	7	2.14	11	2.91	AATTn(8)TACAT	3	2	1.50	2.5	0.2
ATATn(8)ATTCA	12	5	2.40	8.5	2.88	AATTn(8)TATAC	3	2	1.50	2.5	0.2
TATAn(8)ATTCA	7	2	3.50	4.5	2.78	TAATn(8)ATCTA	3	2	1.50	2.5	0.2
AATTn(8)ATTAC	5	1	5.00	3	2.67	TATAn(8)ATCTA	2	3	0.67	2.5	0.2
ATATn(8)TACAT	5	1	5.00	3	2.67	AATTn(8)ACATT	4	3	1.33	3.5	0.1
TAATn(8)TACAT	5	1	5.00	3	2.67	ATATn(8)ACTAT	3	4	0.75	3.5	0.1
TATAn(8)ACATT	5	1	5.00	3	2.67	ATATn(8)ATATC	19	17	1.12	18	0.1
TATAn(8)ACTAT	5	1	5.00	3	2.67	TATAn(8)ATCAT	5	4	1.25	4.5	0.1
TTAAn(8)AATCT	5	1	5.00	3	2.67	AATTn(8)AACTT	5	6	0.83	5.5	0.0
TTAAn(8)TCATA	5	1	5.00	3	2.67	ATTAn(8)ATCAT	6	5	1.20	5.5	0.0
TAATn(8)TAATC	8	3	2.67	5.5	2.27	ATATn(8)TCATA	8	9	0.89	8.5	0.0
TAATn(8)TCATA	8	3	2.67	5.5	2.27	TTAAn(8)TCAAT	11	11	1.00	11	0.0
TAATn(8)TTCAA	8	3	2.67	5.5	2.27	ATTAn(8)ATATC	7	7	1.00	7	0.0
ATATn(8)TAATC	6	2	3.00	4	2.00	AATTn(8)TATCA	4	4	1.00	4	0.0
ATTAn(8)ATACT	6	2	3.00	4	2.00	ATTAn(8)AACTT	4	4	1.00	4	0.0
						ATTAn(8)CATAT	3	3	1.00	3	0.0

Tableau 22 : Classement de toutes les dyades anagrammes à TTAAn(8)TTAAC chez *Brucella melitensis*. La première colonne représente la dyade après la fréquence amont, la fréquence aval, le ratio, la fréquence théorique, l'indice de χ^2.

Sinorhizobiom meliloti. Comme précédemment nous avons voulu savoir dans quelles pCDSs notre dyade était présente chez *Brucella melitensis et Sinorhizobiom meliloti.*

La figure 32 montre une distribution analogue aux deux figures (30 et 31) précédentes. On peut donc en tirer les mêmes conclusions : c'est le nucléotide C qui diminue la concentration en anagrammes à de basses valeurs d'indice de χ^2.

3.2.6.4 Conclusion

En conclusion de cette analyse, nous pouvons suggérer que seul le S.A.F.T. de CtrA garde systématiquement la valeur la plus élevé d'indice de χ^2. Pour savoir si d'autre S.A.F.T. de régulateurs généraux ont la même distribution, nous avons testé le S.A.F.T. de Spo0A chez *Bacillus subtilis* et *Bacillus halodurans.* Nous nous sommes servi de *Brucella melitensis* comme test négatif.

3.2.7 Distribution des anagrammes au S.A.F.T. de Spo0A

Les produits des gènes *spo*0 chez *Bacillus subtilis* sont requis pour l'initiation de la sporulation dans cet organisme. Une mutation de *spo*0 empêche la sporulation. Les effets les plus importants lors d'une mutation de spo0 arrivent lors d'une mutation du locus *spo*0A. *spo*0A est le composant principal de la régulation à partir duquel les signaux environnementaux pour la sporulation agissent. Le S.A.F.T. de Spo0A est TGNCGAA et il se retrouve devant les pCDSs codant pour : Spo0A, Spo0F, SpoIIA (Strauch *et al.*, 1990).

La CDS codant pour Spo0A est présente chez *Bacillus subtilis et Bacillus halodurans* mais pas chez *Brucella melitensis, Sinorhizobium meliloti et Caulobacter crescentus.*

Le motif testé ici est donc celui du site d'atterrissage de Spo0A : TGNCGAA. Nous avons créé toutes les dyades possibles des anagrammes à ce motif en faisant varier les nucléotides de place, en considérant TG comme une première monade et CGAA comme une seconde

dyade	AM	AV	F.TH	Ratio	CHI.2	dyade	AM	AV	F.TH	Ratio	CHI.2
TTAAn(8)CTTAA	9	0	4.5	9	9.00	TAATn(8)AACTT	2	0	1	2	NO
AATTn(8)AATTC	6	2	4	3	2.00	TAATn(8)AATTC	2	0	1	2	NO
AATTn(8)TATAC	4	2	3	2	0.67	TAATn(8)ATACT	2	0	1	2	NO
ATATn(8)ATCAT	4	2	3	2	0.67	TAATn(8)CAATT	2	0	1	2	NO
TAATn(8)TTAAC	5	1	3	5	2.67	TAATn(8)TACAT	2	0	1	2	NO
TATAn(8)ATTAC	5	1	3	5	2.67	TAATn(8)TTCAA	2	0	1	2	NO
AATTn(8)TTAAC	4	1	2.5	4	1.80	TATAn(8)ACATT	2	0	1	2	NO
ATATn(8)TCAAT	3	2	2.5	1.5	0.20	TATAn(8)TCTAA	2	0	1	2	NO
ATATn(8)TTCAA	4	1	2.5	4	1.80	TATAn(8)TTCAA	2	0	1	2	NO
ATTAn(8)TTCAA	4	1	2.5	4	1.80	TTAAn(8)AATCT	2	0	1	2	NO
TTAAn(8)AATTC	4	1	2.5	4	1.80	TTAAn(8)CTATA	2	0	1	2	NO
TTAAn(8)TAACT	5	0	2.5	5	5.00	TTAAn(8)TCTAA	2	0	1	2	NO
AATTn(8)ACATT	3	1	2	3	NO	AATTn(8)ATATC	0	1	0.5	0	NO
ATATn(8)AATTC	2	2	2	1	NO	AATTn(8)ATCAT	1	0	0.5	1	NO
ATATn(8)ATATC	2	2	2	1	NO	AATTn(8)ATCTA	1	0	0.5	1	NO
ATATn(8)ATTCA	3	1	2	3	NO	AATTn(8)ATTAC	0	1	0.5	0	NO
ATTAn(8)TCATA	4	0	2	4	NO	AATTn(8)ATTCA	1	0	0.5	1	NO
TATAn(8)CTAAT	4	0	2	4	NO	AATTn(8)CAATT	1	0	0.5	1	NO
TATAn(8)TACTA	4	0	2	4	NO	AATTn(8)CATAT	1	0	0.5	1	NO
AATTn(8)CTTAA	3	0	1.5	3	NO	AATTn(8)CATTA	1	0	0.5	1	NO
AATTn(8)TACTA	2	1	1.5	2	NO	AATTn(8)TCATA	1	0	0.5	1	NO
ATATn(8)ACTAT	1	2	1.5	0.5	NO	ATATn(8)ATACT	1	0	0.5	1	NO
ATATn(8)ATTAC	3	0	1.5	3	NO	ATATn(8)ATCTA	1	0	0.5	1	NO
ATTAn(8)ACTAT	2	1	1.5	2	NO	ATATn(8)CAATT	1	0	0.5	1	NO
ATTAn(8)ATTAC	2	1	1.5	2	NO	ATATn(8)CATAT	1	0	0.5	1	NO
ATTAn(8)ATTCA	2	1	1.5	2	NO	ATATn(8)CTAAT	1	0	0.5	1	NO
TAATn(8)ATCAT	3	0	1.5	3	NO	ATATn(8)TATAC	0	1	0.5	0	NO
TATAn(8)AATTC	2	1	1.5	2	NO	ATATn(8)TCATA	0	1	0.5	0	NO
TATAn(8)ATTCA	3	0	1.5	3	NO	ATTAn(8)ATACT	1	0	0.5	1	NO
TATAn(8)TATAC	2	1	1.5	2	NO	ATTAn(8)CATAT	1	0	0.5	1	NO
TTAAn(8)ACATT	3	0	1.5	3	NO	ATTAn(8)CTATA	1	0	0.5	1	NO
TTAAn(8)ACTTA	3	0	1.5	3	NO	ATTAn(8)TAATC	1	0	0.5	1	NO

monade. Tous les anagrammes possibles pour les deux monades sont assemblés comme précédemment avec un espacement de un nucléotide. Nous avons ensuite examiné la position de ces dyades grâce à « *Genomic scale* » pour définir leurs positions entre –300 et +300 centré sur 0 (codon start). Vu la taille du motif du S.A.F.T. de Spo0A, nous nous attendions à le retrouver plus souvent qu'un long motif comme le S.A.F.T. de CtrA. Ce qui est le cas, alors que les ratios pour le S.A.F.T. de CtrA ce faisait sur un ordre de fréquence de dizaines d'apparitions du motif, pour le S.A.F.T. de Spo0A, c'est de l'ordre de centaines d'apparitions. Donc les ratios sont établis sur base de données plus importantes.

3.2.7.1 Bacillus subtilis

Pour connaître la distribution du S.A.F.T. de Spo0A et pour savoir si le S.A.F.T. de Spo0A possède le plus grand indice de χ^2 de toutes les dyades anagrammes possibles. Nous avons analysé la position de toutes ces dyades. Les résultats sont présentés dans le tableau 25.
On remarque que le ratio le plus élevé est celui du motif du S.A.F.T. de Spo0A mais il ne possède pas l'indice de χ^2 le plus élevé. L'indice de χ^2 le plus élevé est détenu par une dyade distribuée plus en aval qu'en amont du codon start. On remarque aussi que par rapport à la distribution du motif du S.A.F.T. de CtrA, la distribution est très étroite, c'est-à-dire que l'amplitude (l'écart entre le plus petit et le plus grand ratio) de tous les ratios ne dépasse pas 1. Tous les résultats liés au ratio sont centrés autour de 1.

Contrairement à CtrA chez *Brucella melitensis,* dont l'indice de χ^2 était significativement différent des autres anagrammes, Spo0A chez *Bacillus subtilis* n'a pas d'indice de χ^2 indiquant une telle différence par rapport aux autres anagrammes.

TTAAn(8)TAATC	3	0	1.5	3	NO	TAATn(8)AATCT	1	0	0.5	1	NO
TTAAn(8)TACTA	2	1	1.5	2	NO	TAATn(8)ATCTA	1	0	0.5	1	NO
TTAAn(8)TTACA	3	0	1.5	3	NO	TAATn(8)CATAT	1	0	0.5	1	NO
AATTn(8)AATCT	2	0	1	2	NO	TAATn(8)CTTAA	1	0	0.5	1	NO
AATTn(8)ATACT	1	1	1	1	NO	TAATn(8)TAATC	1	0	0.5	1	NO
AATTn(8)TAATC	2	0	1	2	NO	TAATn(8)TACTA	1	0	0.5	1	NO
AATTn(8)TATCA	2	0	1	2	NO	TATAn(8)ACTAT	0	1	0.5	0	NO
AATTn(8)TCAAT	2	0	1	2	NO	TATAn(8)ACTTA	1	0	0.5	1	NO
AATTn(8)TTCAA	1	1	1	1	NO	TATAn(8)TACAT	1	0	0.5	1	NO
ATATn(8)AACTT	0	2	1	0	NO	TATAn(8)TATCA	0	1	0.5	0	NO
ATATn(8)ACTTA	2	0	1	2	NO	TTAAn(8)AACTT	0	1	0.5	0	NO
ATATn(8)TAACT	2	0	1	2	NO	TTAAn(8)ATACT	1	0	0.5	1	NO
ATATn(8)TATCA	2	0	1	2	NO	TTAAn(8)ATATC	1	0	0.5	1	NO
ATATn(8)TCTAA	2	0	1	2	NO	TTAAn(8)ATCAT	1	0	0.5	1	NO
ATTAn(8)ACATT	2	0	1	2	NO	TTAAn(8)ATTCA	1	0	0.5	1	NO
ATTAn(8)ATCTA	2	0	1	2	NO	TTAAn(8)CAATT	1	0	0.5	1	NO
ATTAn(8)TTAAC	2	0	1	2	NO	TTAAn(8)CATTA	1	0	0.5	1	NO
ATTAn(8)TTACA	2	0	1	2	NO	TTAAn(8)TACAT	1	0	0.5	1	NO

Tableau 23 : Classement de toutes les dyades anagrammes à TTAAn(8)TTAAC chez *Caulobacter crescentus*. La première colonne représente la dyade après la fréquence amont, la fréquence aval, la fréquence théorique, le ratio, l'indice de χ^2.

3.2.7.2 Bacillus halodurans

Le tableau 26 présente les résultats obtenus comme précédemment. Ici le motif du S.A.F.T. de Spo0A n'apparaît plus en premier dans le classement des valeurs de ratio. Il est même très éloigné du haut du tableau. Donc on peut estimer que la première position en fonction du ratio du S.A.F.T. de Spo0A chez *B. subtilis* est un fait du hasard car il ne se répète pas chez *Bacillus halodurans*. Le motif est classé septième dans les scores de ratio et donc nous ne retrouvons pas une distribution du S.A.F.T. de Spo0A analogue à celle du S.A.F.T. de CtrA. Nous allons comparer la distribution du motif du S.A.F.T. de Spo0A entre un organisme régulé par Spo0A et un autre non régulé par Spo0A.

3.2.7.3 Brucella melitensis

Le tableau 27 montre les résultats obtenus chez *Brucella melitensis*. On remarque que tous les motifs sont centrés sur 1 et l'amplitude est très faible. Le motif du S.A.F.T. de Spo0A a un ratio plus petit que 1. Il est donc présent 440 fois en aval et 413 fois en amont. Nous pouvons donc conclure que Spo0A n'a pas de distribution spécifique comme le S.A.F.T. de CtrA lorsque CtrA régule l'espèce étudiée.

3.2.7.3 Conclusion

Cette analyse démontre que le S.A.F.T. de Spo0A n'est pas distribué spécifiquement en amont. Il n'a pas non plus une distribution différente entre une espèce qu'il régule et une autre qu'il ne régule pas. Ces tests démontrent donc que la distribution amont du S.A.F.T de CtrA ne peut pas se généraliser à l'ensemble des sites d'atterrissage pour facteurs de transcription.

Dyade	AM	AV	F.TH	Ratio	CHI2	Dyade	AM	AV	F.TH	Ratio	CHI2
TTAAn(8)CAATT	12	0	6	12.00	12.00	ATATn(8)CATAT	4	1	2.5	4.00	1.80
ATTAn(8)TTAAC	9	0	4.5	9.00	9.00	ATTAn(8)AATTC	4	1	2.5	4.00	1.80
AATTn(8)ACATT	11	1	6	11.00	8.33	TATAn(8)CATTA	4	1	2.5	4.00	1.80
AATTn(8)TCTAA	10	1	5.5	10.00	7.36	TATAn(8)TACTA	4	1	2.5	4.00	1.80
AATTn(8)TTCAA	12	2	7	6.00	7.14	TATAn(8)TATAC	4	1	2.5	4.00	1.80
AATTn(8)CTTAA	7	0	3.5	7.00	7.00	TATAn(8)TCAAT	4	1	2.5	4.00	1.80
ATTAn(8)TCAAT	9	1	5	9.00	6.40	TATAn(8)TCATA	4	1	2.5	4.00	1.80
ATATn(8)TCTAA	6	0	3	6.00	6.00	TTAAn(8)TAATC	4	1	2.5	4.00	1.80
ATTAn(8)AACTT	6	0	3	6.00	6.00	ATATn(8)AACTT	7	3	5	2.33	1.60
TATAn(8)TAATC	6	0	3	6.00	6.00	AATTn(8)TCAAT	8	4	6	2.00	1.33
TAATn(8)TTCAA	8	1	4.5	8.00	5.44	ATATn(8)TACAT	5	2	3.5	2.50	1.29
ATATn(8)CATTA	10	2	6	5.00	5.33	TATAn(8)TATCA	5	2	3.5	2.50	1.29
ATATn(8)AATTC	15	5	10	3.00	5.00	ATATn(8)TCAAT	6	3	4.5	2.00	1.00
AATTn(8)ACTAT	5	0	2.5	5.00	5.00	TAATn(8)CATTA	6	3	4.5	2.00	1.00
AATTn(8)CTAAT	5	0	2.5	5.00	5.00	TTAAn(8)AATTC	7	4	5.5	1.75	0.82
ATATn(8)TTCAA	5	0	2.5	5.00	5.00	AATTn(8)ATCAT	12	8	10	1.50	0.80
ATTAn(8)ATCAT	5	0	2.5	5.00	5.00	AATTn(8)TACAT	4	2	3	2.00	0.67
TAATn(8)ATTCA	5	0	2.5	5.00	5.00	ATTAn(8)ATATC	4	2	3	2.00	0.67
TTAAn(8)ATTCA	5	0	2.5	5.00	5.00	TTAAn(8)ACATT	4	2	3	2.00	0.67
AATTn(8)TAATC	13	4	8.5	3.25	4.76	TTAAn(8)CTTAA	4	2	3	2.00	0.67
TTAAn(8)TACTA	7	1	4	7.00	4.50	AATTn(8)TCATA	7	10	8.5	0.70	0.53
AATTn(8)ATTCA	12	4	8	3.00	4.00	AATTn(8)CATAT	5	3	4	1.67	0.50
AATTn(8)AACTT	6	1	3.5	6.00	3.57	ATATn(8)ACATT	6	4	5	1.50	0.40
TAATn(8)CATAT	6	1	3.5	6.00	3.57	ATATn(8)CAATT	6	4	5	1.50	0.40
TATAn(8)AACTT	6	1	3.5	6.00	3.57	ATATn(8)ATCAT	6	8	7	0.75	0.29
AATTn(8)AATTC	14	6	10	2.33	3.20	ATATn(8)ATATC	9	11	10	0.82	0.20
AATTn(8)ATATC	9	3	6	3.00	3.00	ATATn(8)AATCT	3	2	2.5	1.50	0.20
ATATn(8)ATTCA	7	2	4.5	3.50	2.78	ATATn(8)TAATC	3	2	2.5	1.50	0.20
ATATn(8)TATCA	7	2	4.5	3.50	2.78	ATATn(8)TCATA	3	4	3.5	0.75	0.14
AATTn(8)TAACT	5	1	3	5.00	2.67	TAATn(8)TCATA	3	4	3.5	0.75	0.14
ATATn(8)ATCTA	5	1	3	5.00	2.67	AATTn(8)CAATT	4	3	3.5	1.33	0.14
ATTAn(8)ATTCA	5	1	3	5.00	2.67	ATTAn(8)TTCAA	4	3	3.5	1.33	0.14

4. Discussion, conclusions et perspectives

L'analyse des programmes révèle que seul « Dyad-analysis » est réellement efficace pour retrouver une dyade. Les autres programmes ne permettent pas de retrouver la dyade complète du S.A.F.T. de CtrA. Le défaut de « Dyad-analysis », c'est son manque de puissance et sa force, c'est sa confiance. De plus, les deux calibrations permettent d'avoir soit une confiance plus grande au détriment de la puissance grâce à la calibration par les dyades ou l'inverse par la calibration par les monades.

Les deux autres programmes (*Motif sampler, Oligo-analysis*) sont intéressants pour retrouver des monades chez les procaryotes où la grande majorité des S.A.F.T. sont des dyades (voir paragraphe de l'introduction sur les HTH). Donc ces deux programmes ne sont pas utiles pour les génomes des organismes procaryotes sauf si le programme « Dyad analysis » ne retrouve pas de S.A.F.T. On peut alors utiliser ces deux programmes pour rechercher une des deux monades composant la dyade du S.A.F.T.. On remarque que ces deux programmes y arrivent tous les deux pour le S.A.F.T. de CtrA.

Le test des séquences amont des pCDSs orthologues pratiqué avec « BLAST2 » ne donne pas de bons résultats car les deux organismes doivent être assez proches évolutivement, les dyades doivent se trouver dans une région conservée entre les deux espèces et d'une espèce à l'autre, les gènes ciblés par un facteur de transcription peuvent différer comme on peut le voir pour les cibles de CtrA (Bellefontaine *et al.*, 2002). Dans le cas étudié, nous avons comparé deux régions promotrices de gènes codant pour un facteur de transcription. Les facteurs de transcription ont tendance à s'autoréguler, ils ont donc plus de chance de retrouver le même motif d'un régulateur dans leurs régions promotrices. De plus les parties conservées sont parfois trop importantes pour qu'on puisse y retrouver un motif commun à toutes les séquences.

Mais tous ces programmes sont efficaces si et seulement si on leur donne un petit nombre de séquences co-régulées. Le test pratiqué sur le S.A.F.T. de CtrA par *Dyad-analysis* démontre que pour une faible dilution, on retrouve le S.A.F.T. de CtrA au premier rang des résultats.

TAATn(8)ATATC	5	1	3	5.00	2.67	TAATn(8)ATCAT	4	3	3.5	1.33	0.14
TATAn(8)AATTC	5	1	3	5.00	2.67	TAATn(8)CAATT	4	3	3.5	1.33	0.14
TATAn(8)CAATT	5	1	3	5.00	2.67	TAATn(8)TCAAT	4	3	3.5	1.33	0.14
TTAAn(8)ATCAT	5	1	3	5.00	2.67	AATTn(8)CATTA	4	5	4.5	0.80	0.11
AATTn(8)AATCT	6	2	4	3.00	2.00	TATAn(8)TTCAA	4	5	4.5	0.80	0.11
ATTAn(8)AATCT	6	2	4	3.00	2.00	TATAn(8)ATCAT	5	4	4.5	1.25	0.11
TATAn(8)TACAT	6	2	4	3.00	2.00	AATTn(8)TATCA	4	4	4	1.00	0.00
TTAAn(8)ATATC	6	2	4	3.00	2.00	ATATn(8)ACTTA	3	3	3	1.00	0.00
TATAn(8)CATAT	9	4	6.5	2.25	1.92						

Tableau 24 : Classement de toutes les dyades anagrammes à TTAAn(8)TTAAC chez *Sinorhizobiom meliloti*. La première colonne représente la dyade après la fréquence amont, la fréquence aval, la fréquence théorique, le ratio, l'indice de χ^2.

Par contre, si la dilution est trop importante, le S.A.F.T. de CtrA n'est plus au premier rang des résultats. Lors d'un test en aveugle, rien ne nous permettrait d'identifier CtrA. Une analyse du génome entier génèrerait trop de résultats et nous ne pourrions pas distinguer les S.A.F.T. réels des faux. Pour séparer les bons résultats des mauvais, nous devrions pouvoir déterminer un critère spécifique au S.A.F.T. ; c'est dans ce but que l'étude de sa distribution amont/aval a été envisagée et testée.

Lorsque nous avons observé la distribution du S.A.F.T. de CtrA chez trois espèces proches d'un point de vue évolutif (*Brucella melitensis, Caulobacter crescentus, Sinorhizobium meliloti*), nous avons remarqué que ces trois espèces régulées par CtrA avaient une distribution du S.A.F.T. CtrA significativement plus concentrée en amont du codon start qu'en aval. Cette distribution amont du motif du S.A.F.T. de CtrA pourrait être un critère spécifique aux S.A.F.T. en général lorsqu'ils sont régulateurs chez l'espèce étudiée. Le test de distribution chez *Bacillus subtilis* (espèce non régulée par CtrA), montre une distribution uniforme du motif du S.A.F.T. de CtrA.

Un biais possible serait que cette distribution excentrée en amont chez *Brucella melitensis, Caulobacter crescentus et Sinorhizobium meliloti* soit due à une composition élevée en AT du motif du S.A.F.T. de CtrA. Pour tester si d'autres motifs de même composition pouvaient avoir une distribution analogue, nous avons mis au point le test des anagrammes. Ce test démontre que chez les trois espèces (*Brucella melitensis, Caulobacter crescentus et Sinorhizobium meliloti)* le motif du S.A.F.T. de CtrA a toujours l'indice de χ^2 le plus élevé parmi tous les anagrammes. Nous avons donc testé la distribution des sous-motifs de TTAAC et nous avons remarqué que TTA et TAA étaient moins fréquents en aval qu'en amont. D'ailleurs beaucoup de motifs anagrammes à indice de χ^2 élevé contiennent une ou deux fois ces sous-motifs. Le S.A.F.T. de CtrA les possède quatre fois mais d'autres motifs tel que TTAAn(7)CTTAA possède aussi quatre fois ces deux sous motifs sans pour autant être distribué en amont comme le S.A.F.T. de CtrA. Suite à ce test nous avons observé la distribution des dyades anagrammes au S.A.F.T. de CtrA sans C, TTAAn(7)TTAA garde

dyade	AM	AV	F.TH	Ratio	CHI2
tgwAGCA	424	533	478.5	0.795497186	12.41483804
tgwAAGC	582	509	545.5	1.143418468	4.884509624
tgwAACG	369	431	400	0.856148492	4.805
tgwCGAA	392	339	365.5	1.156342183	3.842681259
gtwACGA	147	180	163.5	0.816666667	3.330275229
tgwAGAC	265	308	286.5	0.86038961	3.226876091
gtwAACG	253	295	274	0.857627119	3.218978102
tgwCAAG	326	282	304	1.156028369	3.184210526
gtwCAGA	215	252	233.5	0.853174603	2.931477516
gtwCAAG	294	328	311	0.896341463	1.8585209
tgwCAGA	447	488	467.5	0.915983607	1.797860963
tgwACGA	282	252	267	1.119047619	1.685393258
gtwACAG	255	227	241	1.123348018	1.626556017
gtwAGCA	164	186	175	0.88172043	1.382857143
tgwACAG	422	455	438.5	0.927472527	1.241733181
tgwGACA	348	330	339	1.054545455	0.477876106
gtwAGAC	113	103	108	1.097087379	0.462962963
tgwGCAA	312	326	319	0.957055215	0.307210031
gtwAAGC	321	308	314.5	1.042207792	0.268680445
gtwGCAA	162	153	157.5	1.058823529	0.257142857
gtwCGAA	165	158	161.5	1.044303797	0.151702786
gtwGACA	195	201	198	0.970149254	0.090909091
gtwGAAC	177	181	179	0.977900552	0.044692737
tgwGAAC	238	241	239.5	0.987551867	0.018789144

Tableau 26 : Classement des dyades anagrammes au S.A.F.T. de Spo0A chez *Bacillus subtilis*. La première colonne représente la dyade après la fréquence amont, la fréquence aval, la fréquence théorique, le ratio, l'indice de χ^2.

l'indice de χ^2 le plus élevé. Néanmoins, d'autres dyades anagrammes sont assez proches, en particulier les dyades complémentaires où les deux sous motifs TAA et TTA n'apparaissent que trois fois. Nous pouvons donc en conclure, premièrement, que le nucléotide C n'influence pas la distribution amont/aval du motif du S.A.F.T. de CtrA et, deuxièmement, que les sous-motifs du S.A.F.T. de CtrA ne peuvent pas expliquer à eux seul la distribution aussi excentrée en amont du S.A.F.T. de CtrA.

Un autre critère a été pris en compte : l'espacement entre les deux monades composant la dyade du S.A.F.T. de CtrA. Nous avons ainsi testé tous les anagrammes possibles à la dyade TTAAn(8)TTAAC et il nous est apparu que l'espacement fixe de la dyade du S.A.F.T. de CtrA semblait avoir de l'importance car un simple changement d'un nucléotide modifiait la distribution des anagrammes et en particulier du S.A.F.T. de CtrA. Nous avons également remarqué que les deux mêmes monades du S.A.F.T. de CtrA séparées de huit nucléotides ne montraient plus une distribution si élevée en amont. Donc la composition et la séquence du motif n'expliquent pas à elles seules cette distribution amont du S.A.F.T. de CtrA.
Donc lorsque CtrA régule certains gènes de l'espèce, le S.A.F.T. de CtrA a une distribution significative en amont (comme chez *Brucella melitensis, Caulobacter crescentus et Sinorhizobium meliloti)* et lorsqu'il n'est pas régulateur pour l'espèce considérée, on observe une distribution uniforme. On en conclut que cette distribution en amont est un caractère intrinsèque du S.A.F.T. de CtrA.

Nous avons également testé le S.A.F.T de Spo0A dont le F.T. est propre à *Bacillus sp.* et non présent chez *Brucella melitensis*. Les tests ne montrent aucune distribution spécifiquement amont du motif du S.A.F.T. de Spo0A ni une distribution différente du S.A.F.T. de Spo0A lorsqu'il est dans un organisme régulé par Spo0A ou non régulé par Spo0A. Donc la distribution amont n'est pas un caractère intrinsèque généralisable à l'ensemble des S.A.F.T..

Dyade	AM	AV	F.TH	Ratio	CHI2
tgwAAGC	513	397	455	1.29	14.79
tgwAGCA	409	507	458	0.81	10.48
gtwCAAG	318	402	360	0.79	9.80
gtwGAAC	265	337	301	0.79	8.61
tgwAACG	484	575	529.5	0.84	7.82
tgwAGAC	200	257	228.5	0.78	7.11
tgwGCAA	480	542	511	0.89	3.76
tgwGAAC	291	247	269	1.18	3.60
gtwAGCA	251	289	270	0.87	2.67
gtwACAG	221	255	238	0.87	2.43
tgwACAG	366	326	346	1.12	2.31
gtwGACA	257	226	241.5	1.14	1.99
gtwCAGA	194	168	181	1.15	1.87
tgwCGAA	479	445	462	1.08	1.25
tgwCAAG	326	300	313	1.09	1.08
gtwAAGC	309	333	321	0.93	0.90
tgwACGA	578	548	563	1.05	0.80
tgwCAGA	267	285	276	0.94	0.59
tgwGACA	321	303	312	1.06	0.52
gtwAACG	381	400	390.5	0.95	0.46
gtwGCAA	365	355	360	1.03	0.14
gtwACGA	323	318	320.5	1.02	0.04
gtwAGAC	125	127	126	0.98	0.02
gtwCGAA	274	272	273	1.01	0.01

Tableau 27 : Classement des dyades anagrammes au S.A.F.T. de Spo0A chez *Bacillus halodurans*. La première colonne représente la dyade après la fréquence amont, la fréquence aval, la fréquence théorique, le ratio, l'indice de χ^2.

De nombreuses nouvelles études pourraient également être menées sur ce sujet et les perspectives sont nombreuses.

Premièrement, il semblerait intéressant d'analyser la distribution amont/aval d'autres S.A.F.T. afin de savoir si cette distribution amont apparaît dans une famille précise de F.T. ou seulement pour des régulateurs généraux touchant beaucoup de cibles comme CtrA.

Deuxièmement, nous pourrions utiliser ce test amont/aval pour filtrer les résultats d'un programme ou comme filtre sur le génome entier. Ce filtre consisterait à garder seulement les S.A.F.T. dont l'indice de χ^2 est plus élevé par rapport à ces anagrammes.

Troisièmement, nous pourrions créer une base de donnés contenant plusieurs fonctions co-régulées et faire une comparaison automatique de cette base de données avec les résultats des S.A.F.T. potentiels. Si un de ces S.A.F.T. se retrouvait en amont de plusieurs pCDSs co-régulées d'après la base de données, nous pourrions définir ce S.A.F.T comme étant un S.A.F.T. réel.

dyade	AM	AV	F.TH	Ratio	CHI2
tgwAACG	382	284	333	1.35	14.42
tgwCAGA	163	109	136	1.50	10.72
gtwACAG	98	58	78	1.69	10.26
gtwGCAA	203	148	175.5	1.37	8.62
gtwGACA	99	63	81	1.57	8.00
tgwAGCA	237	181	209	1.31	7.50
tgwGACA	107	71	89	1.51	7.28
gtwAACG	112	79	95.5	1.42	5.70
gtwCAGA	118	93	105.5	1.27	2.96
gtwAAGC	193	166	179.5	1.16	2.03
gtwACGA	113	132	122.5	0.86	1.47
tgwAAGC	358	329	343.5	1.09	1.22
tgwCAAG	202	181	191.5	1.12	1.15
tgwGCAA	191	172	181.5	1.11	0.99
tgwACAG	168	186	177	0.90	0.92
tgwCGAA	413	440	426.5	0.94	0.85
gtwAGCA	60	52	56	1.15	0.57
tgwAGAC	167	181	174	0.92	0.56
tgwGAAC	159	147	153	1.08	0.47
gtwCGAA	76	70	73	1.09	0.25
gtwCAAG	198	190	194	1.04	0.16
gtwAGAC	31	29	30	1.07	0.07
tgwACGA	177	176	176.5	1.01	0.00
gtwGAAC	143	143	143	1.00	0.00

Tableau 25 : Classement des dyades anagrammes au S.A.F.T. de Spo0A chez *Brucella melitensis*. La première colonne représente la dyade après la fréquence amont, la fréquence aval, la fréquence théorique, le ratio, l'indice de χ^2.

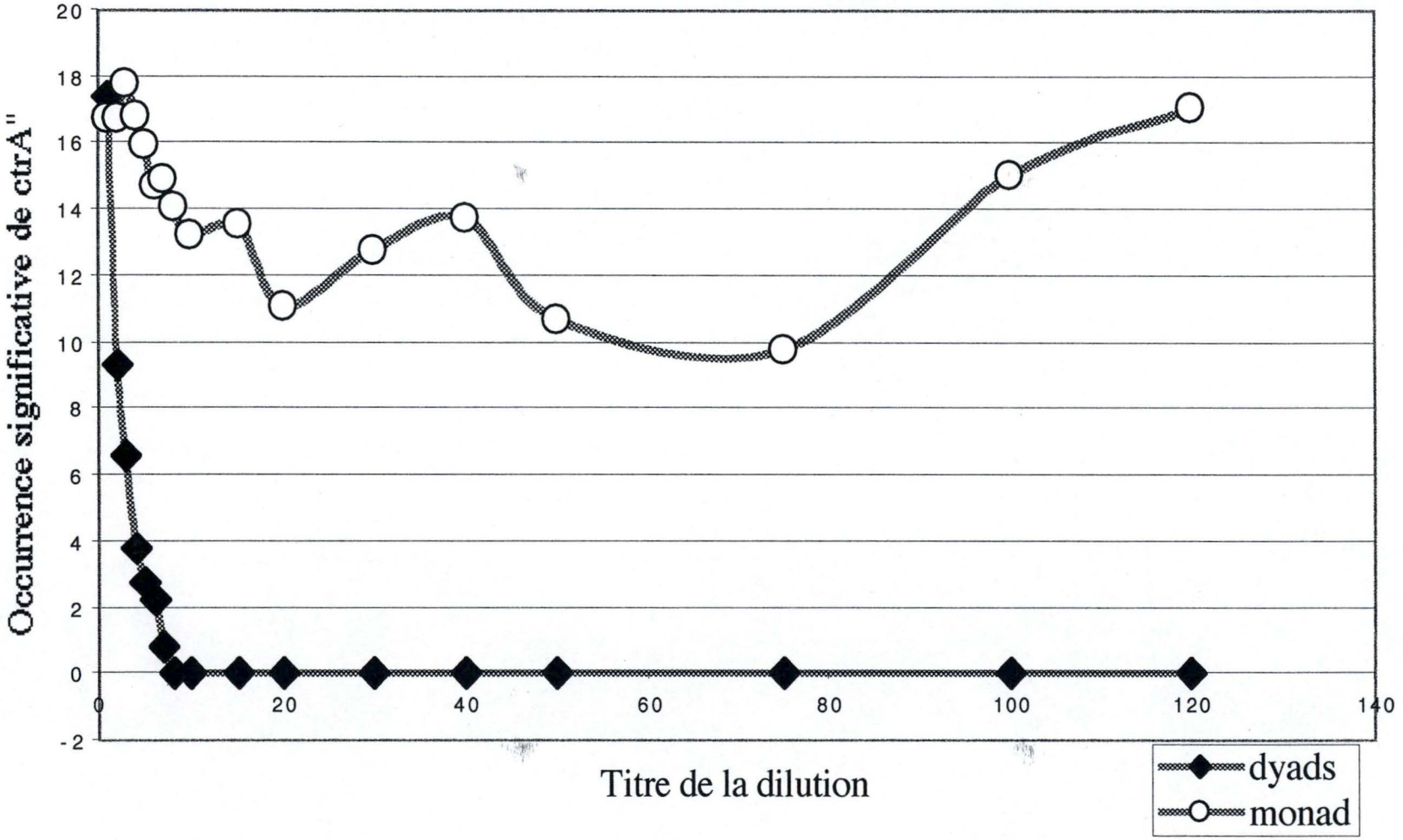

Figure 17 : Dilution du S.A.F.T. de CtrA. Dilution de séquences contenant le S.A.F.T. de CtrA dans des séquences ne contenant pas le S.A.F.T. de CtrA. Les résultats de la calibration par monades est indiqué par des cercles et les résultats de la calibration par dyades avec des losanges.

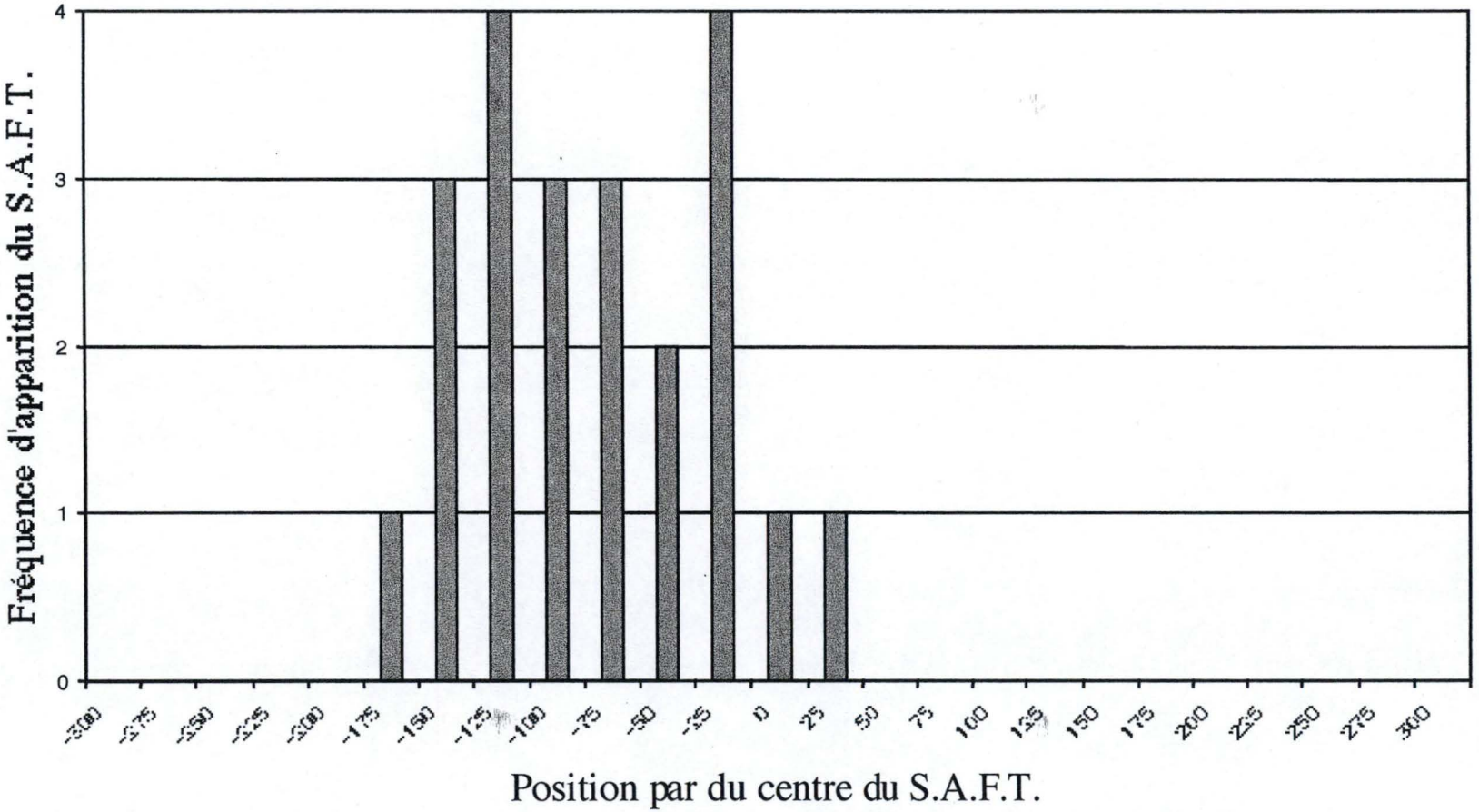

Figure 19 : Distribution du S.A.F.T. de CtrA chez *Brucella melitensis.* Fréquence d 'apparition du S.A.F.T. de CtrA entre -300 et +300 nucléotides par rapport au codon start de toutes les pCDS de *Brucella melitensis.* Le S.A.F.T. est positionné dans un intervalle de 25 nucléotides sur base du centre du S.A.F.T..

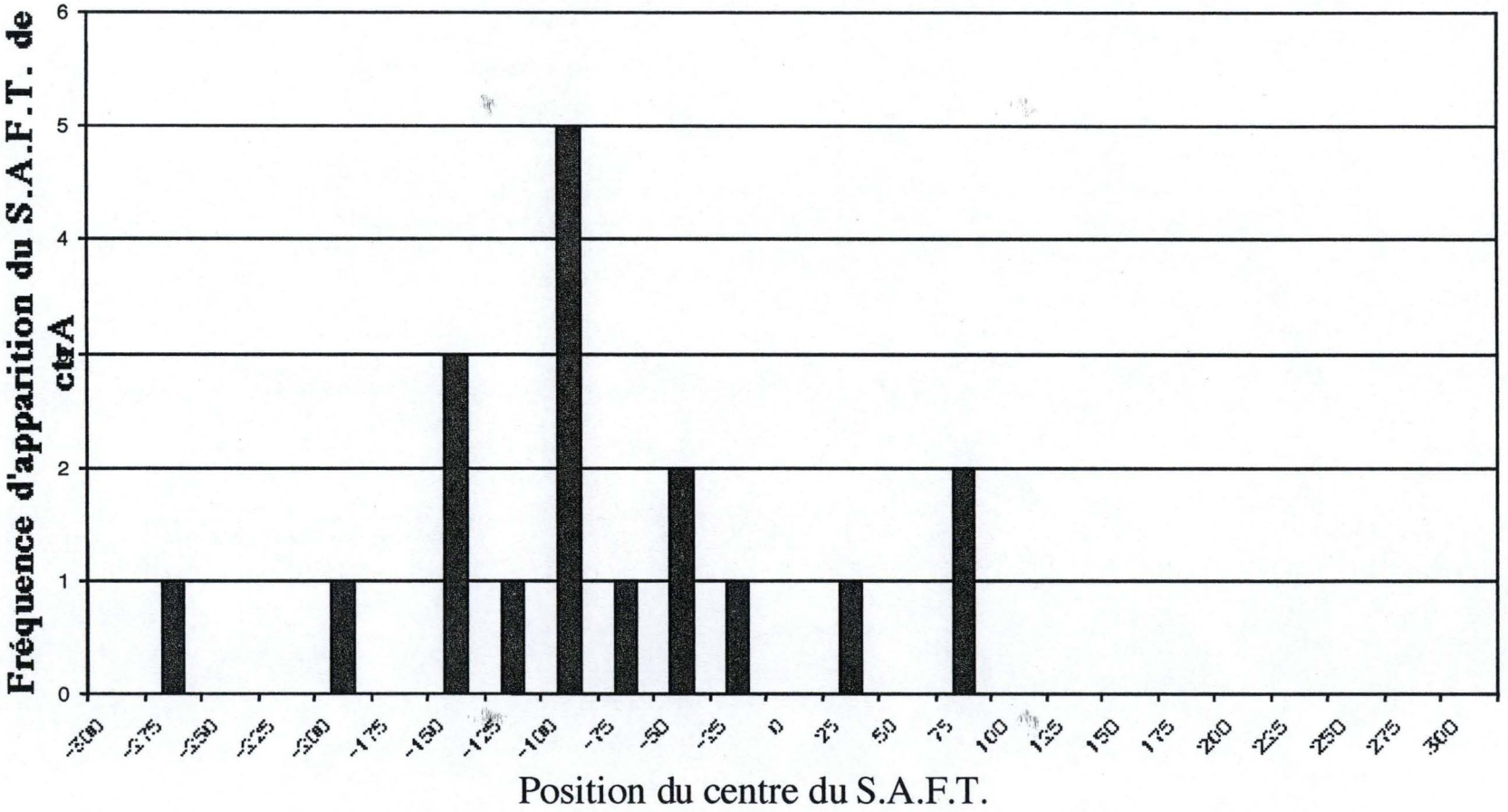

Figure 20 : Distribution du S.A.F.T. de CtrA chez *Caulobacter crescentus*. Fréquence d 'apparition du S.A.F.T. de CtrA entre -300 et +300 nucléotides par rapport au codon start de toutes les pCDS de *Caulobacter crescentus*. Le S.A.F.T. est positionné dans un intervalle de 25 nucléotides sur base du centre du S.A.F.T..

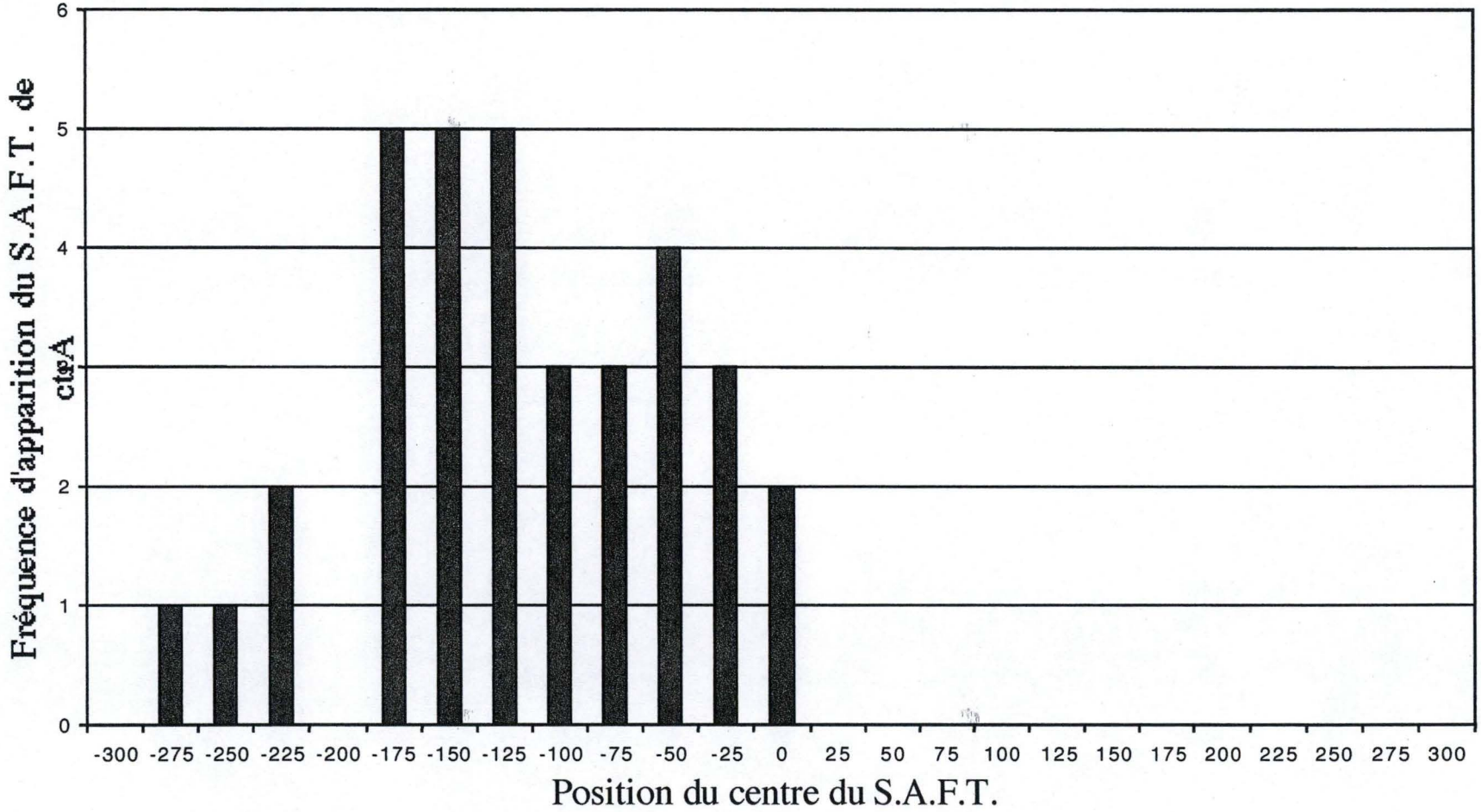

Figure 21 : Distribution du S.A.F.T. de CtrA chez *Sinorhizobium meliloti*. Fréquence d 'apparition du S.A.F.T. de CtrA entre -300 et +300 nucléotides par rapport au codon start de toutes les pCDS de *Sinirhizobiom meliloti*. Le S.A.F.T. est positionné dans un intervalle de 25 nucléotides sur base du centre du S.A.F.T..

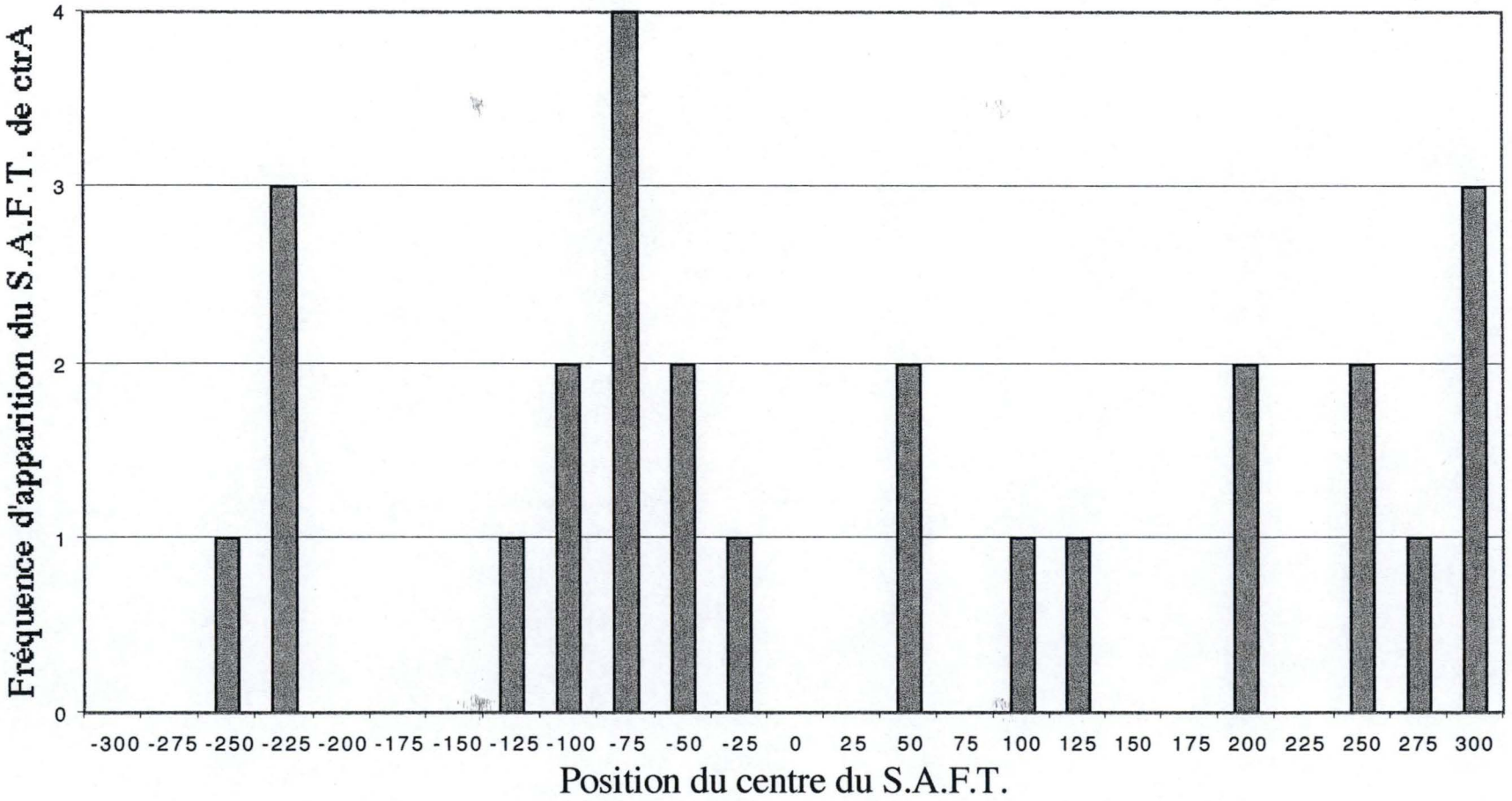

Figure 22 : Distribution du S.A.F.T. de CtrA chez *Bacillus subtilis*. Fréquence d 'apparition du S.A.F.T. de CtrA entre -300 et +300 nucléotides par rapport au codon start de toutes les pCDS de *Bacillus subtilis*. Le S.A.F.T. est positionné dans un intervalle de 25 nucléotides sur base du centre du S.A.F.T..

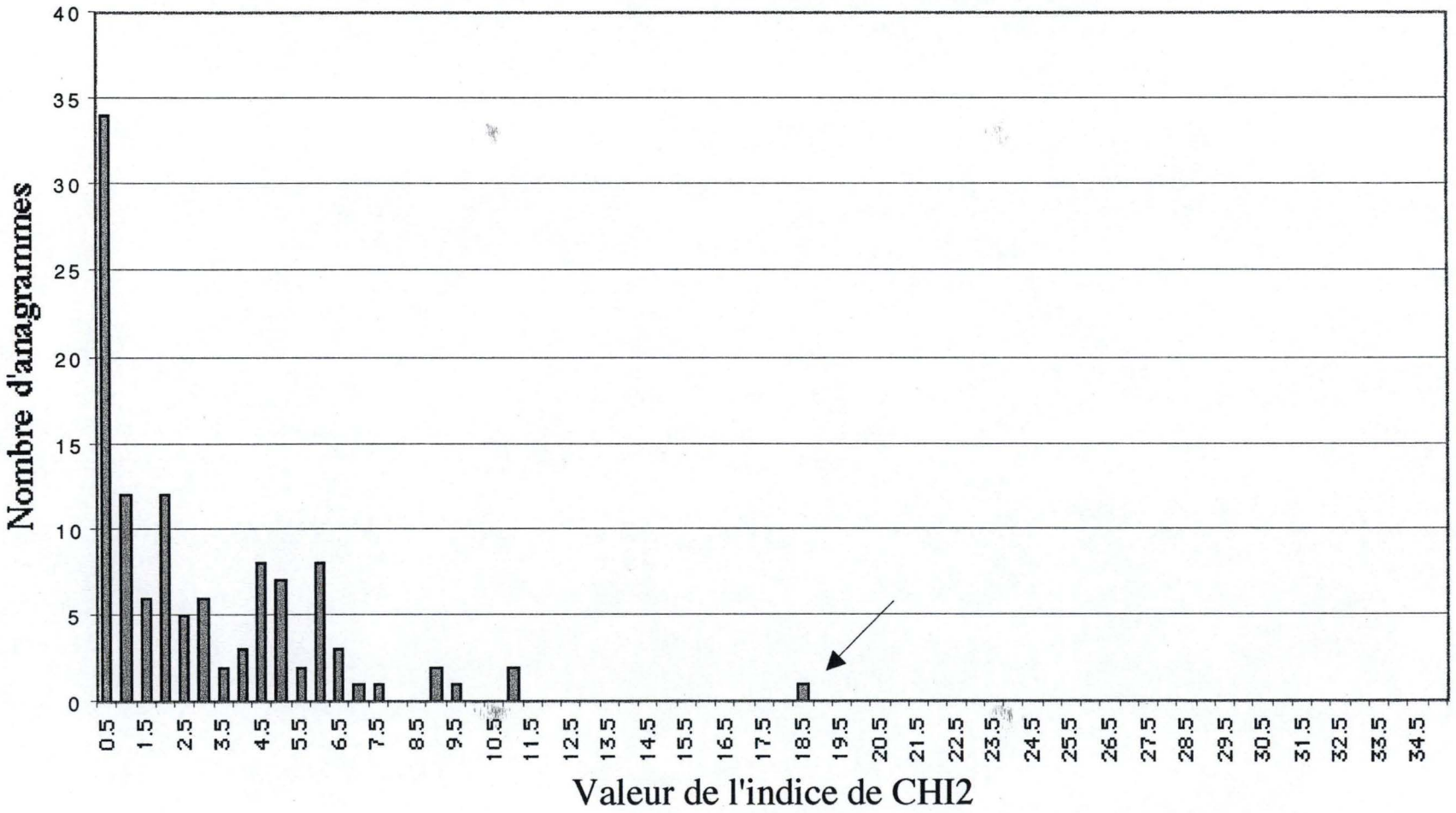

Figure 23 : Distribution des dyades anagrammes au S.A.F.T. de CtrA en fonction de leur indice de CHI2 chez *Brucella melitensis*. Tous les anagrammes sont classés dans des intervalles de 0.5 (indice de CHI2). La flèche indique la dyade du S.A.F.T. de CtrA.

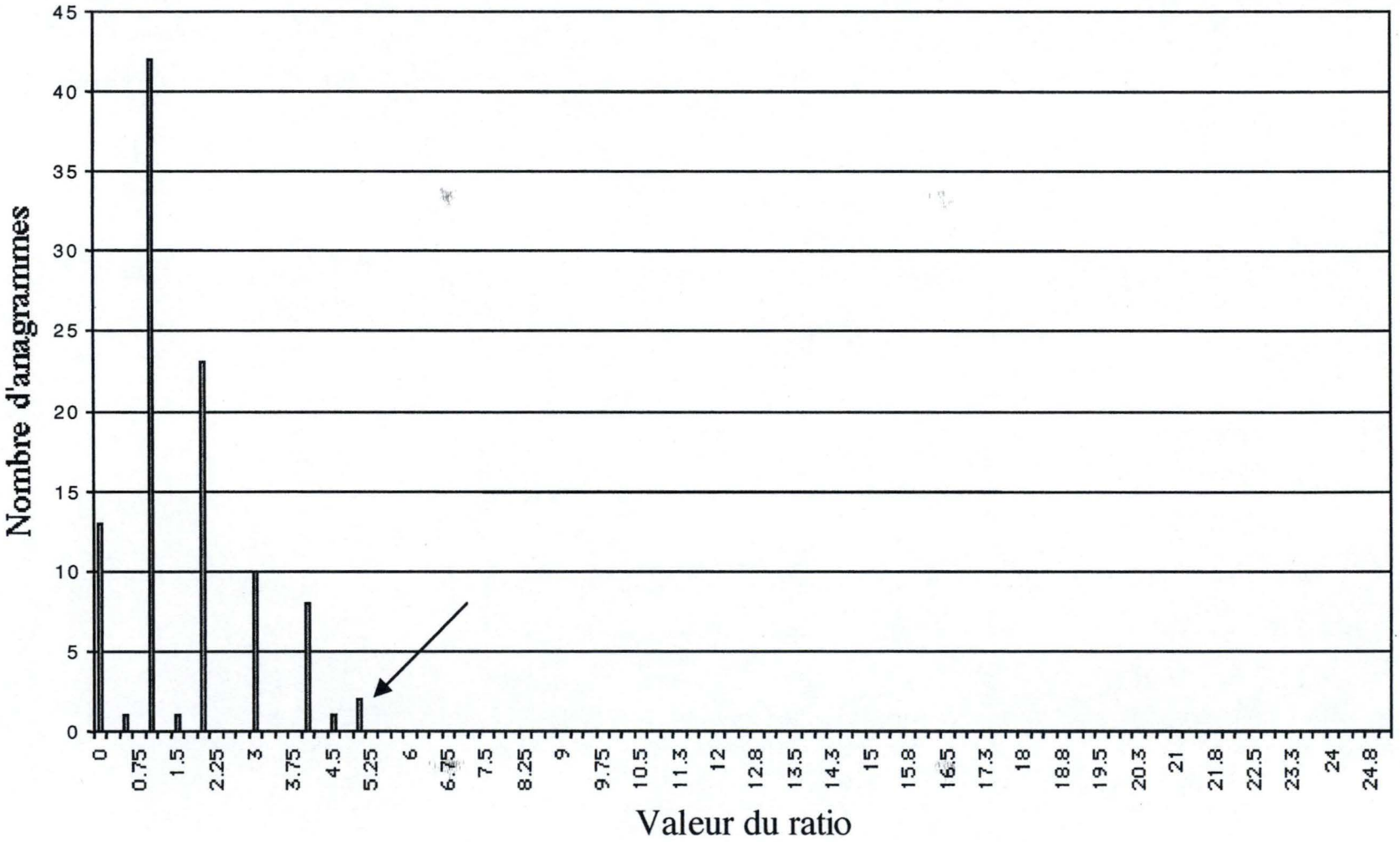

Figure 24 : Distribution des dyades anagrammes au S.A.F.T. de CtrA en fonction de leur ratios chez *Caulobacter crescentus*. Tous les anagrammes sont classés dans des intervalles de 0.25 (ratio). La flèche indique la dyade du S.A.F.T. de CtrA.

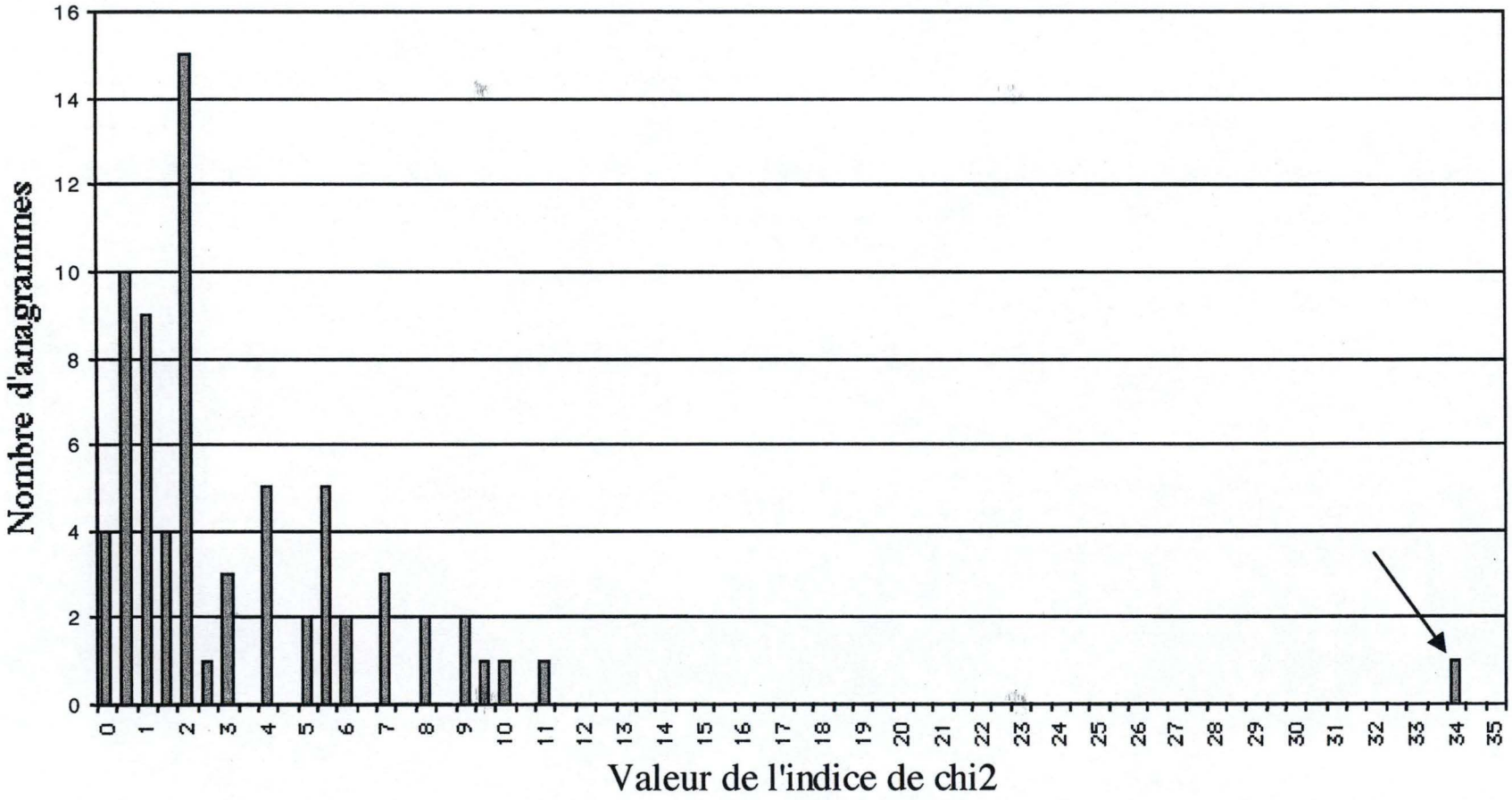

Figure 25 : Distribution des dyades anagrammes au S.A.F.T. de CtrA en fonction de leur indice de CHI2 chez *Sinorhizobium meliloti*. Tous les anagrammes sont classés dans des intervalles de 0.5 (indice de CHI2). La flèche indique la dyade du S.A.F.T. de CtrA.

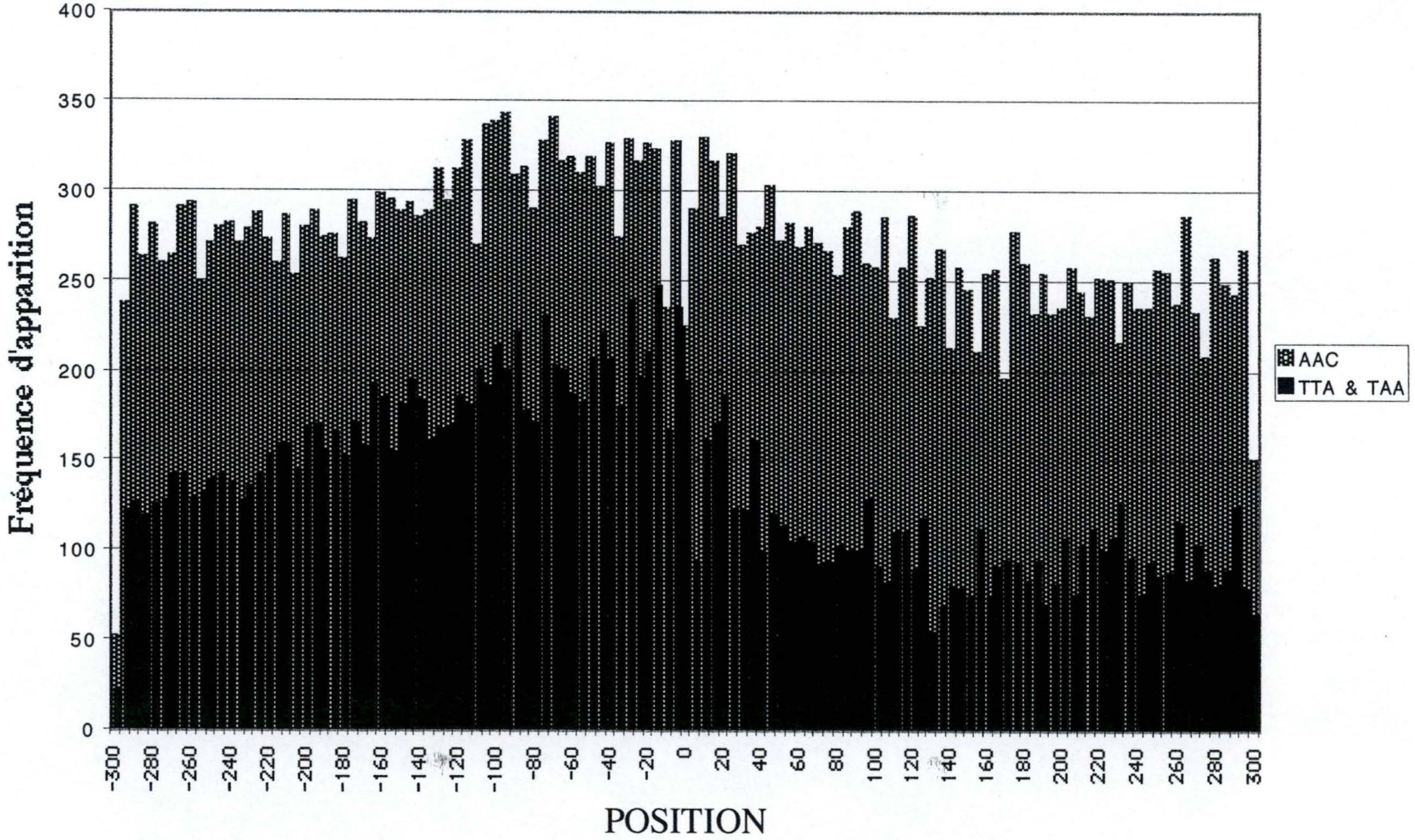

Figure 26 : Distribution de AAC TTA TAA chez *Brucella melitensis.* La partie noire représente la distribution de TTA et TAA. La partie plus claire représente la distribution de AAC. Les deux distributions sont indépendantes.

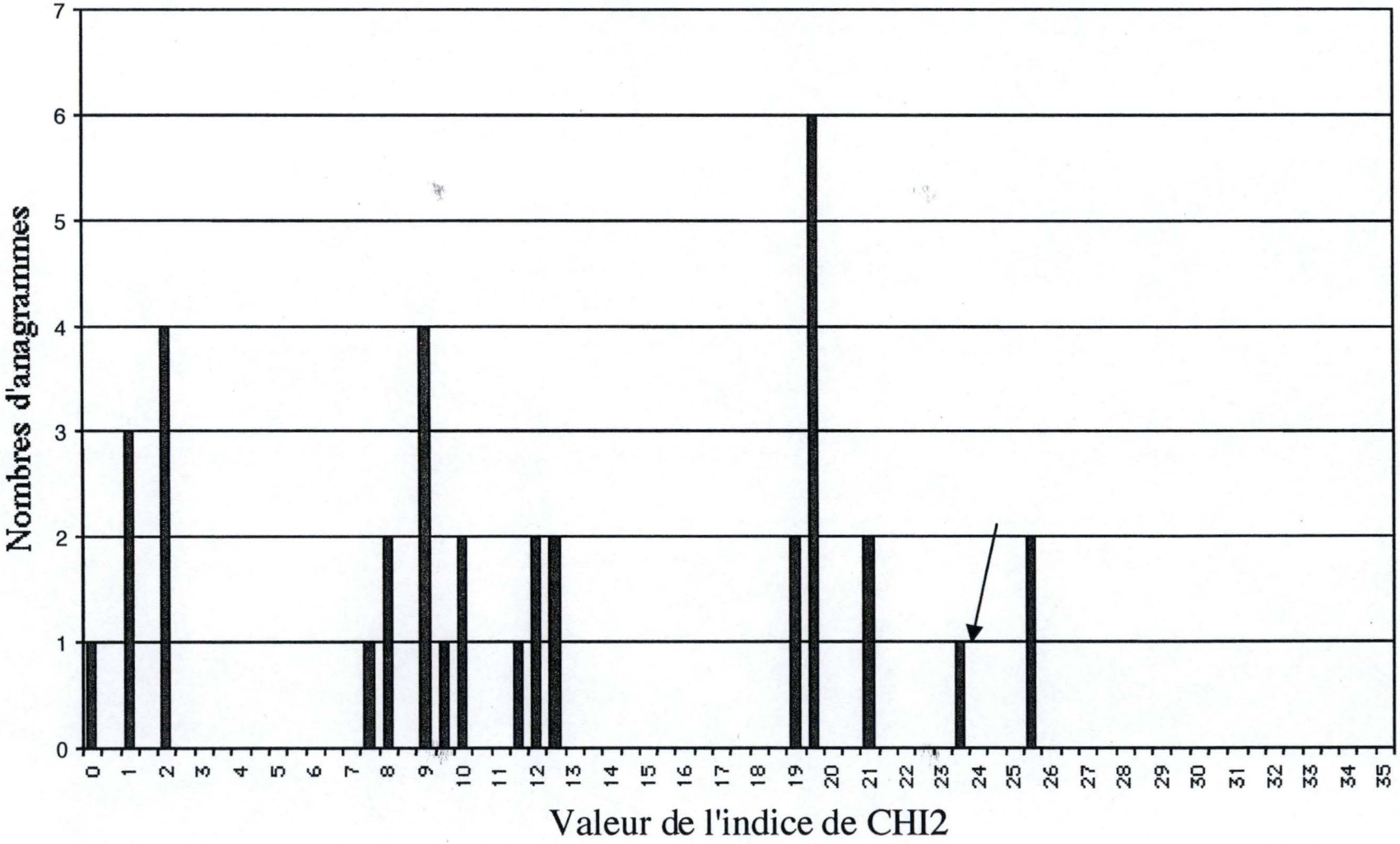

Figure 27 : Distribution des dyades anagrammes à TTAAN(7)TTAA en fonction de leur indice de CHI2 chez *Brucella melitensis*. Tous les anagrammes sont classés dans des intervalles de 0.5 (indice de CHI2). La flèche indique la dyade du S.A.F.T. de CtrA sans C.

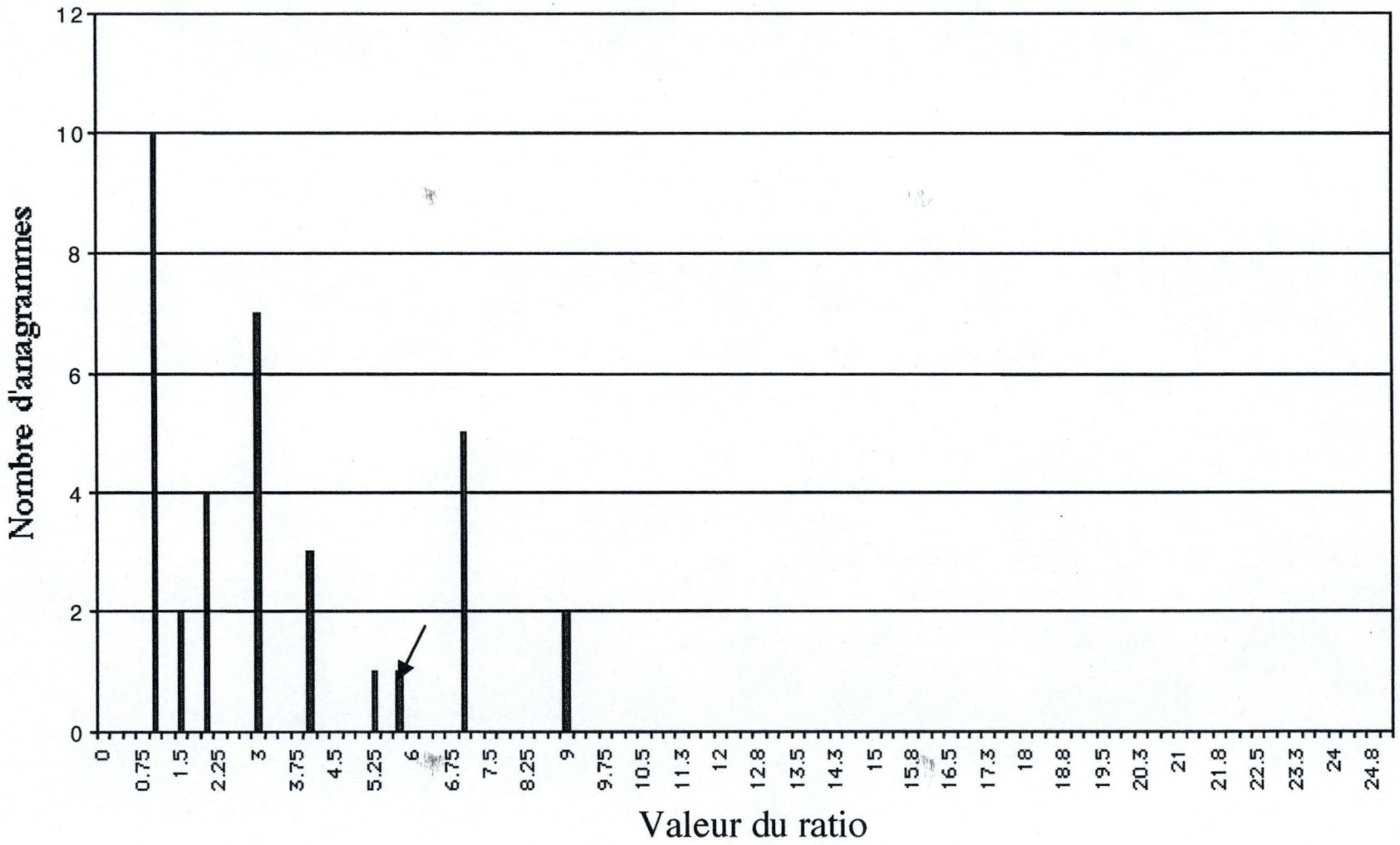

Figure 28 : Distribution des dyades anagrammes à TTAAn(7)TTAA selon leur ratio chez *Caulobacter crescentus*. Tous les anagrammes sont classés dans des intervalles de 0.25 (ratio). La flèche indique la dyade du S.A.F.T. de CtrA sans C.

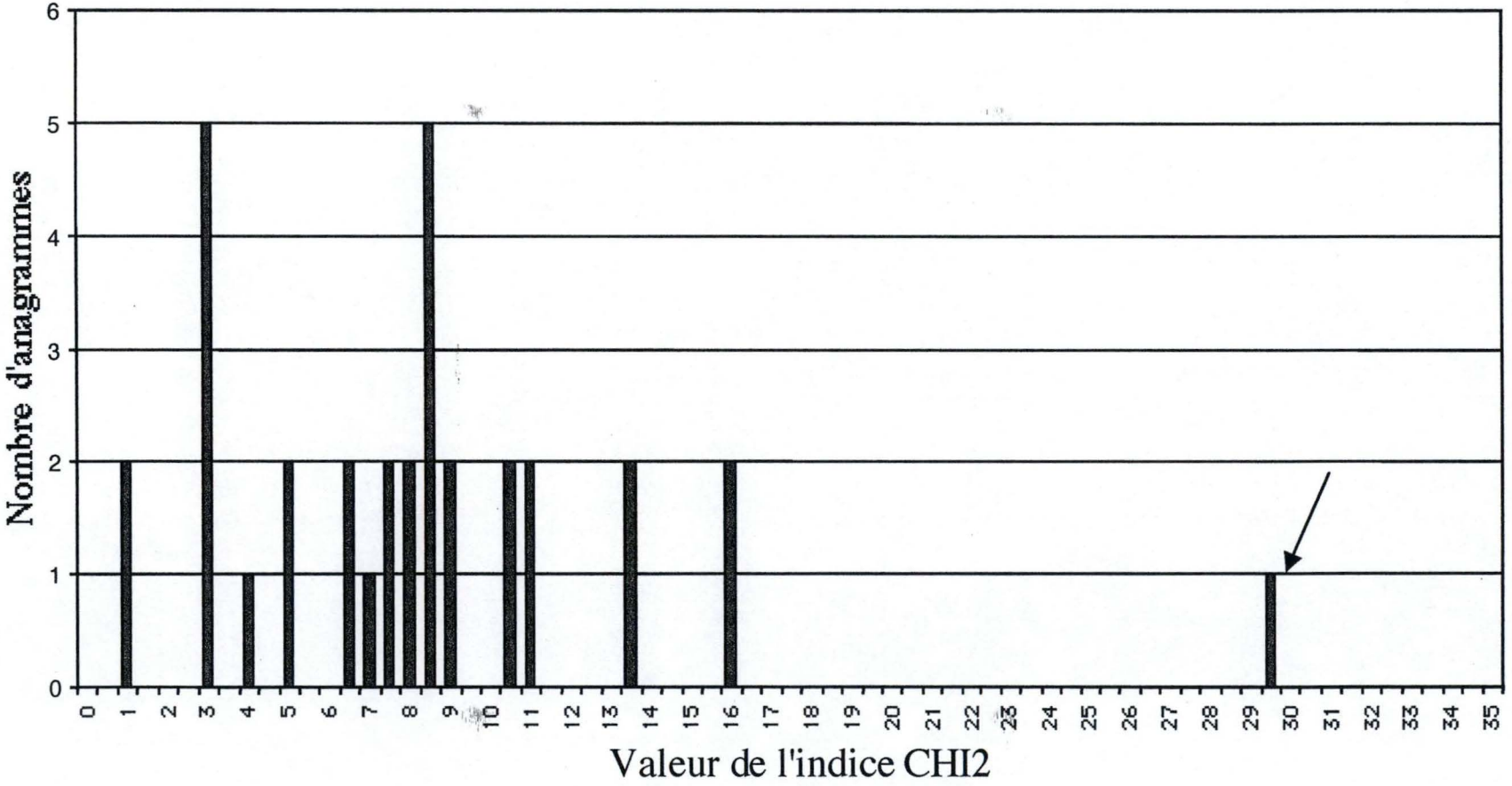

Figure 29 : Distribution des dyades anagrammes à TTAAn(7)TTAA selon leur indice de CHI2 chez *Sinorhizobium meliloti*. Tous les anagrammes sont classés dans des intervalles de 0.5 (indice de CHI2). La flèche indique la dyade du S.A.F.T. de CtrA.

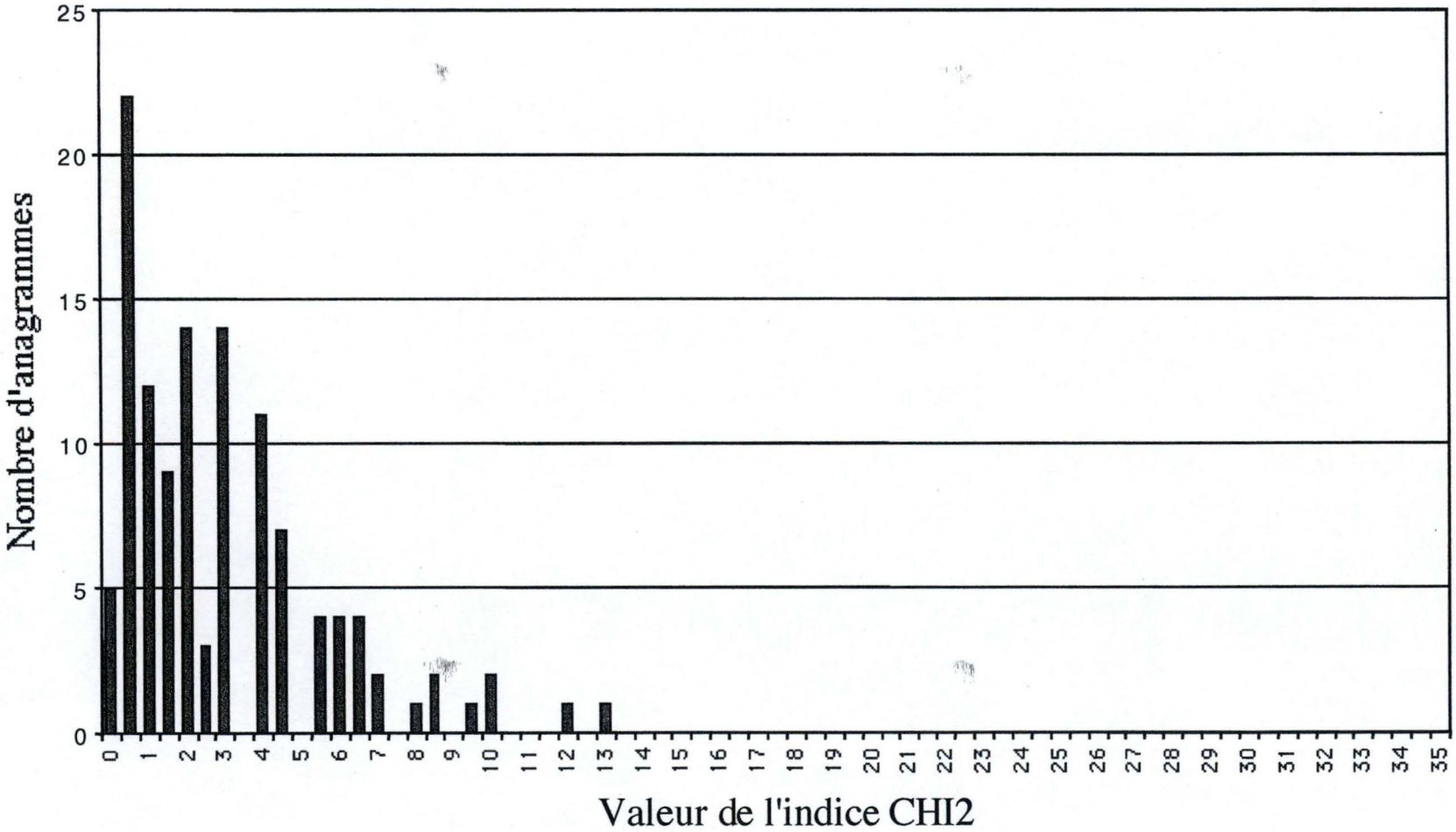

Figure 30 : Tous les anagrammes sont classés dans des intervalles de 0.5 (indice de CHI2). Distribution des dyades anagrammes à TTAAn(8)TTAACselon leur indice de CHI2 chez *Brucella melitensis.*

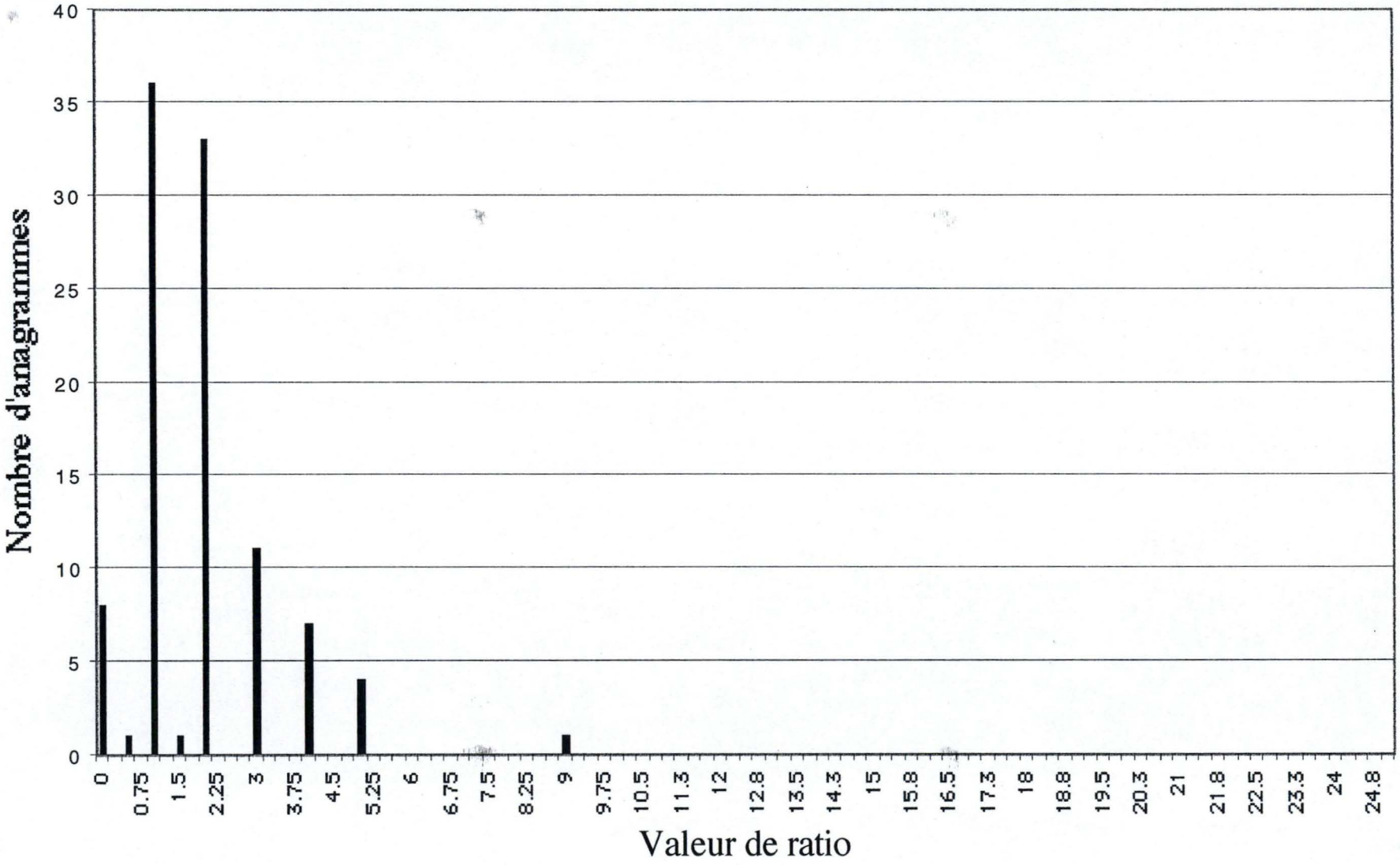

Figure 31 : Distribution des dyades anagrammes à TTAAn(8)TTAAC selon leur ratio chez *Caulobacter crescentus*. Tous les anagrammes sont classés dans des intervalles de 0.25 (ratio).

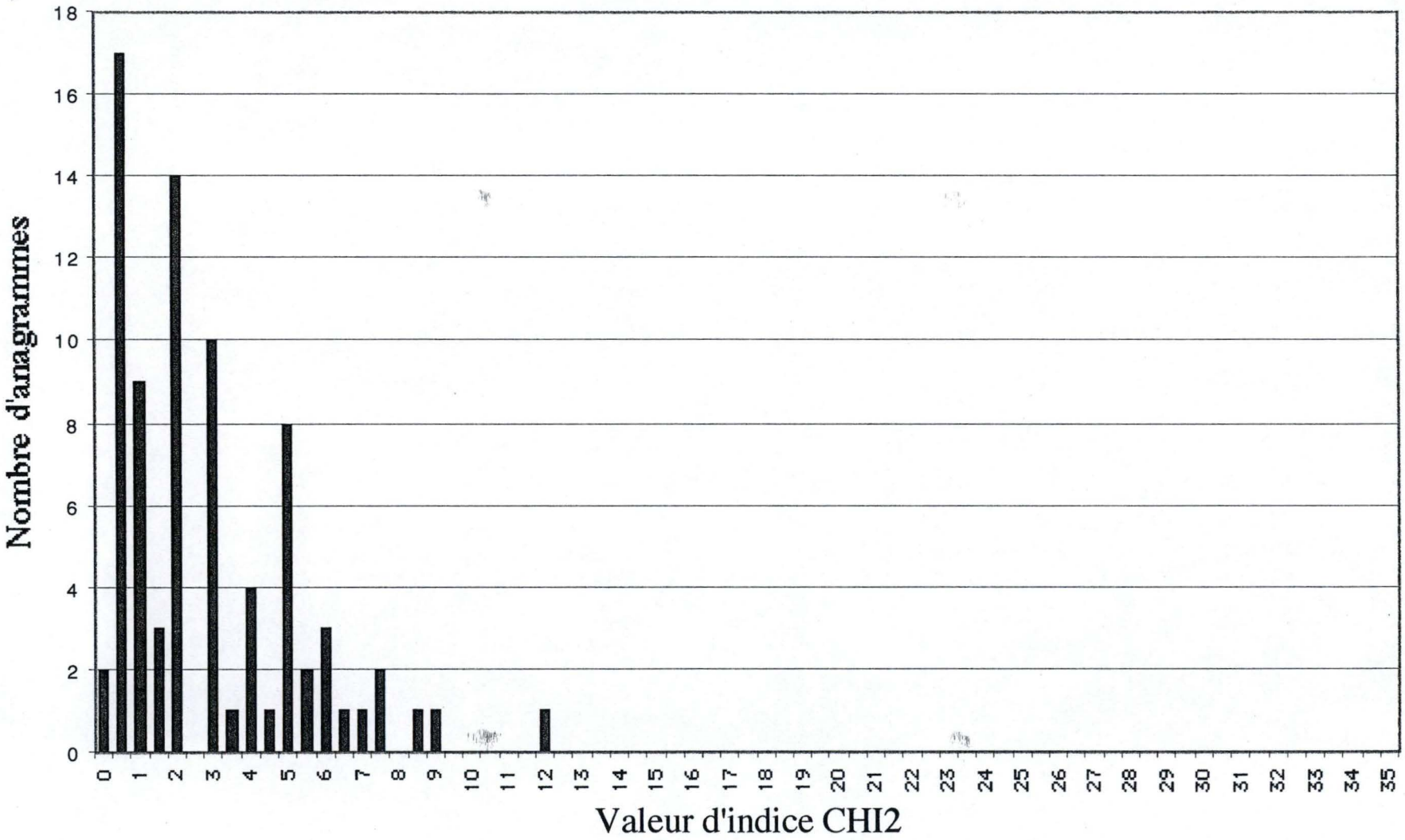

Figure 32 : Distribution des anagrammes à TTAAn(8)TTAAC selon leur indice de CHI2 chez *Sinorhizobium melitoti*. Tous les anagrammes sont classés dans des intervalles de 0.5 (indice de CHI2).

Bibliographie

Bellefontaine, A. F., Pierreux, C. E., Mertens, P., Vandenhaute, J., Letesson, J. J. & Bolle, X. D. (2002). Plasticity of a transcriptional regulation network among alpha-proteobacteria is supported by the identification of CtrA targets in Brucella abortus. *Mol Microbiol* **43**, 945-60.

deHaseth, P. L., Zupancic, M. L. & Record, M. T., Jr. (1998). RNA polymerase-promoter interactions: the comings and goings of RNA polymerase. *J Bacteriol* **180**, 3019-25.

DelVecchio, V. G., Kapatral, V., Redkar, R. J., Patra, G., Mujer, C., Los, T., Ivanova, N., Anderson, I., Bhattacharyya, A., Lykidis, A., Reznik, G., Jablonski, L., Larsen, N., D'Souza, M., Bernal, A., Mazur, M., Goltsman, E., Selkov, E., Elzer, P. H., Hagius, S., O'Callaghan, D., Letesson, J. J., Haselkorn, R., Kyrpides, N. & Overbeek, R. (2002). The genome sequence of the facultative intracellular pathogen Brucella melitensis. *Proc Natl Acad Sci U S A* **99**, 443-8.

Gralla, J. D. & Collado-Vides, J. (1996) Organization and Function of Transcription Regulatory Elements. . Edited by A. press.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* **235**, 1501-31.

Lestrate, P., Delrue, R. M., Danese, I., Didembourg, C., Taminiau, B., Mertens, P., De Bolle, X., Tibor, A., Tang, C. M. & Letesson, J. J. (2000). Identification and characterization of in vivo attenuated mutants of Brucella melitensis. *Mol Microbiol* **38**, 543-51.

Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. (2000). An overview of the structures of protein-DNA complexes. *Genome Biol* **1**.

Rhodius, V. A. & Busby, S. J. (1998). Positive activation of gene expression. *Curr Opin Microbiol* **1**, 152-9.

Rojo, F. (2001). Mechanisms of transcriptional repression. *Curr Opin Microbiol* **4**, 145-51.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944-5.

Stock, A. M., Robinson, V. L. & Goudreau, P. N. (2000). Two-component signal transduction. *Annu Rev Biochem* **69**, 183-215.

Strauch, M., Webb, V., Spiegelman, G. & Hoch, J. A. (1990). The SpoOA protein of Bacillus subtilis is a repressor of the abrB gene. *Proc Natl Acad Sci U S A* **87**, 1801-5.

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**, 1113-22.

Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P. & Moreau, Y. (2002). A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol* **9**, 447-64.

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**, 827-42.

van Helden, J., Andre, B. & Collado-Vides, J. (2000). A web site for the computational analysis of yeast regulatory sequences. *Yeast* **16**, 177-87.

van Helden, J., Rios, A. F. & Collado-Vides, J. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res* **28**, 1808-18.

Wosten, M. M. (1998). Eubacterial sigma-factors. *FEMS Microbiol Rev* **22**, 127-50.

NBRE DE CHROMOSOMES	2
TAILLE DU CHROMO. 1	2,117,144BP
TAILLE DU CHROMO. 2	1,177,787BP
DNA TOTAL SEQUENCE	3,294,931BP
DNA SEQUENCES CODANTES	2,874,027BP(87%)
CONTENU EN GC	57%
PLASMIDES	0
STATISTIQUES GENERALE	
NBRE TOTAL D' ORFs	3,197
NBRE D' ORFs SUR CHROMO 1	2,059
NBRE D' ORFs SUR CHROMO 2	1,138
ORFs AVEC FONCTION CONNUE	2,487-78%
ORFs SANS FONCTION CONNUE	710-22%
ORFs SANS FONCTION CONNUE ET SANS SIMILARITÉ	228-7%
ORFs SANS FONCTION CONNUE MAIS AVEC SIMILARITÉ	488-15%
ORFs DANS CLUSTERS ORTHOLOGUES	2,115-66%
ORFs DANS CLUSTERS PARALOGUES	842-26%
NBRE DE CLUSTERS PARALOGUES	235
ORFs DANS CLUSTERS CHROMOSOMIQUE	1,583-49%
ORFs IMPLIQUÉES DANS ÉLÉMENTS DE FUSION	1,127-35%
ORFs IMPLIQUÉES DANS ÉLÉMENTS DE FUSION AS COMPOSITES	282-9%
ORFs IMPLIQUÉES DANS ÉLÉMENTS DE FUSION AS COMPONENTS	957-30%

Annexe 1 : Caractéristiques du génomes de *Brucella melitensis* selon DelVecchio *et al.* (DelVecchio, 2002).