UNIVERSITÉ DU QUÉBEC À MONTRÉAL

Utilisation des EST dans la génération d'une nouvelle ressource bioinformatique spécifique à la tolérance du blé au froid

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

MAHDI BELCAID

JUILLET 2006

REMERCIEMENTS

L'accomplissement de ce travail n'aurait jamais pu se faire sans le support sans faille de la Docteure Anne Bergeron qui a su encourager mon esprit scientifique et éveiller mon sens de la curiosité et du savoir. Son expertise et sa rigueur m'ont beaucoup appris tandis que sa gentillesse et sa simplicité ont rendu mon travail sur ce projet agréable et plaisant.

Merci au Docteur Fathey Sarhan pour son support financier et ses précieux conseils pédagogiques et scientifiques sans lesquels je n'aurais pas pu compléter cette recherche.

Merci au Docteur Mario Houde pour son support en biologie, son suivi et son sens critique tout au long de ce projet.

Merci a tous les membres de l'équipe de laboratoire du Docteur Sarhan et du laboratoire de génomique comparée de l'UQAM et plus particulièrement à Yannick Gingras pour ses nombreux conseils techniques, son sens de l'humour et sa gentillesse.

Merci à Nadia pour sa gentillesse, sa générosité et son amitié.

Merci à la Docteure Guylaine Poisson pour tous les conseils tant personnels que professionnels.

Le support de la famille est sans aucun doute un élément majeur dans la réalisation de tout projet. Je tiens donc à remercier mes deux frères Khalil et Si Mohammed pour leurs encouragements et leur appui tout au long de ce projet. Merci a ma sœur Amina pour ses conseils, son support moral et son amitié....

Sans l'appui de mes parents, leur confiance et leur générosité, je n'aurais certainement pas eu le courage de réaliser ce projet jusqu'à la toute fin. Je vous suis très

reconnaissant pour tout ce que vous avez fait pour moi et surtout pour votre amour inconditionnel.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

La bioinformatique constitue un outil de choix dans l'étude des plantes. Lorsqu'utilisées efficacement, la gestion ainsi que l'analyse informatique des données biologiques permettent de réduire considérablement la durée des expériences ainsi que les coûts liés à la recherche tout en élargissant le spectre d'analyses pouvant être effectuées.

L'objectif de cette recherche est de définir un protocole efficace utilisant les outils actuels et mettant en œuvre de nouvelles approches, dans le but de faire le prétraitement, le clustering et l'assemblage d'un ensemble d'EST. Ce mémoire traite des algorithmes et des techniques informatiques utilisés dans l'étude de l'expression différentielle à partir des EST. Les particularités du projet de FGAS sont présentées et une attention particulière est portée sur la manière dont les résultats ont été analysés. À travers la classification en sous-groupes fonctionnels et de l'analyse différentielle digitale, l'identification des gènes potentiellement impliqués dans le phénomène d'acclimatation du blé au froid est effectuée.

Mots clés : Bioinformatique, clustering, assemblage, Expressed Sequence Tags, tolérance au froid.

INTRODUCTION

Durant la dernière décennie, les domaines de la biologie et de la biochimie ont vécu un essor remarquable, dû principalement au progrès scientifique et à l'avancement technologique accomplis dans des domaines connexes, notamment dans le séquençage de génomes et dans l'étude de structures des protéines. Ce progrès a eu pour conséquences la production d'une avalanche de données, dont l'analyse manuelle s'avérait une tâche imposante, voire même impossible. Pour relever ce défi, le domaine de la bioinformatique a été mis en place dans le but de rassembler, gérer, ainsi que pour analyser les données disponibles. En facilitant l'analyse des données et l'extraction d'informations pertinentes, la bioinformatique permet parallèlement de réduire les coûts reliés à la recherche en permettant d'explorer à faible coût, des avenues exigeant autrefois plus de ressources.

Dans le domaine de l'agriculture, par exemple, la bioinformatique s'est avérée d'une grande utilité dans l'étude des plantes, aidant ainsi l'amélioration des récoltes et des conditions de croissance et réduisant la durée des expériences en séparant préalablement les expériences susceptibles de produire des résultats intéressants de celles ayant un moindre impact sur la recherche menée.

Ce mémoire traite des problèmes bioinformatiques liés à la comparaison des différents niveaux d'expression d'EST dans deux types de blé, le premier possédant une résistance au froid et le deuxième requérant des conditions de croissance normales. Les étapes préalables à ce genre d'analyses sont cruciales dans tout projet de génomique comparée, car toute mauvaise préparation ou mauvaise interprétation des données brutes peut biaiser les résultats obtenus. Malgré cela, les étapes préalables sont rarement détaillées dans la littérature scientifique et demeurent souvent un obstacle à la réussite de tout projet relatif aux EST.

Dans le présent document, l'emphase est principalement mise sur les types d'artéfacts susceptibles de corrompre les analyses et la technique utilisée pour les éviter. Pour chacune de ces étapes, les algorithmes les plus utilisés dans le prétraitement des données sont analysés. Les forces et faiblesses de chacune des approches sont discutées et la manière d'interpréter ou de corriger les aberrations obtenues est proposée.

Les étapes abordées dans le document sont les suivantes :

- Le prétraitement des séquences : discute l'impact de la contamination ou de régions de basse complexité sur les étapes conséquentes et propose un protocole efficace, utilisé dans le cadre du projet FGAS, pour détecter et supprimer ces régions.
- Le clustering : introduit les deux approches majeures utilisées dans le clustering de séquences, à savoir, celle utilisant la fréquence de mots et celle se basant sur l'homologie entre séquences. Les résultats de deux outils implémentant chacune de ces approches sont analysés dans le cadre du projet FGAS. L'assemblage et l'annotation des séquences sont brièvement introduits puisque ces deux étapes sont abondamment détaillées dans la littérature.
- L'analyse des résultats : traite de la manière dont les résultats ont été analysés et les outils utilisés pour interpréter les données obtenues.

Les protocoles de contrôle de qualité dans ces types d'expériences se basent généralement sur la comparaison des résultats obtenus avec les résultats connus et publiés. Or, dans le cas présent, peu de choses sont connues sur le blé de manière générale et sur les processus impliqués dans son acclimatation au froid ou à tout autre type de stress. Pour s'assurer de la justesse des résultats obtenus, la comparaison avec les résultats obtenus avec d'autres espèces proches est effectuée pour chaque étape majeure du projet. Les résultats obtenus ont fait l'objet d'un article conjoint avec l'équipe du projet FGAS (*Houde et al. 2006*), et qui est joint en annexe. De plus, une base de données publique a été élaborée pour consolider toute l'information spécifique au projet et sera bientôt disponible en accès libre. Cette base de données constitue un apport majeur et une importante ressource dans l'étude de la tolérance du blé au froid.

Les contributions bioinformatiques issues du projet peuvent être énumérées comme suit :

1. La première contribution consiste en la préparation de séquences d'EST de haute qualité, vides de toute contamination. Ces séquences ont été obtenues suite à un nettoyage itératif, utilisant deux implémentations distinctes dont l'ordre et les paramètres ont soigneusement été choisis.
2. La maîtrise des algorithmes utilisés par ces approches a permis de développer un protocole efficace et unique permettant de résoudre les aberrations de larges clusters. Cette technique, semi-automatique, utilisant les graphes d'adjacences pour détecter les chimères putatives a grandement facilité le processus de correction d'aberrations sans compromettre la qualité ni le temps d'exécution.
3. L'utilisation d'organismes similaires et des méthodes d'annotations complémentaires a permis d'obtenir un pourcentage de séquences annotées hautement supérieur à tout ce qui a été produit jusqu'à présent. La bonne qualité des annotations a permis de détecter, à travers l'analyse d'expression digitale, un grand nombre de gènes associés avec la tolérance du blé au froid et à d'autres conditions de stress.

Le présent document est divisé en quatre chapitres. Le premier introduit de manière brève les bases biologiques nécessaires à la compréhension du problème. Une définition chronologique du processus de génération des EST est présentée et la problématique biologique est définie.

Le chapitre deux discute du processus de prétraitement des séquences. Les étapes de décontamination, de détection de basse complexité et de suppression des répétitions sont toutes définies et les outils utilisés dans chacune de ces étapes sont abordés. L'importance de ces étapes dans les phases subséquentes du projet est développée à travers des exemples sommaires.

Le chapitre trois définit le processus de clustering du point de vue biologique pour ensuite introduire les deux algorithmes utilisés en pratique pour effectuer cette tâche. Une première structure de données, proposée dans l'implémentation du calcul de fréquences de mots et une seconde, proposée dans la jointure de clusters sont présentées et leur complexité brièvement discutée.

Le chapitre quatre est une application, dans le cadre du projet FGAS, des méthodes présentées dans les chapitres précédents. À chacune des étapes, les particularités du projet, le ou les logiciels utilisés ainsi que les résultats obtenus sont présentés. Le schéma de la base de données, créée dans le but de consolider l'information et les résultats finaux, est brièvement décrit.

Pour conclure, nous discuterons des résultats et conclusions obtenus et publiés dans l'article et expliquerons les contributions du projet FGAS, du point de vu bioinformatique, à la recherche sur la résistance du blé au froid.

CHAPITRE I

PRINCIPES FONDAMENTAUX EN BIOLOGIE ET EN BIOINFORMATIQUE

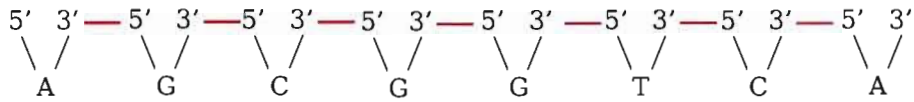
Dans le présent chapitre, nous allons introduire les principes fondamentaux de la biologie et de la bioinformatique, nécessaires à la compréhension des chapitres subséquents. Dans un premier temps, une brève présentation des notions fondamentales en biologie sera effectuée. Ceci permettra d'expliciter les principes d'encodage de l'information dans un génome ainsi que les méthodes par lesquelles les gènes sont transcrits et traduits en protéines. Dans un second temps, les principales notions bioinformatique nécessaires à la compréhension de l'assemblage des séquences, seront brièvement abordées.

1.1 Principes fondamentaux en biologie

1.1.1 Structure de l'ADN

Le matériel génétique est ce qui confère à une espèce les caractéristiques qui lui sont propres et les fonctions de bases nécessaires dans son cycle de vie (*Brenner et al., 01*). Cette substance, appelée *génome*, représente la totalité de l'ADN (Acide desoxyribonucleique) présent dans une cellule à un moment donné. L'ADN est une longue molécule constituée d'une suite de quatre petites sous-unités, appelées *nucléotides* ou *bases*, qui sont : l'Adénine (A), la Thymine (T), la Guanine (G) et la Cytosine (C). Au niveau moléculaire, les paires de bases consécutives sont fixées entre elles par des liaisons chimiques où le carbone $5'$ d'un nucléotide est relié au carbone $3'$ de l'autre pour former

(a)



(b)

```
5′ -AGGCTGATGGATATTG- 3′
3′ -TCCGACTACCTATAAG- 5′
```

Figure 1.1 Les liaisons de carbones entre les nucléotides d'une chaîne d'ADN et sa forme double brin

de longs filaments d'ADN. La molécule est toujours lue, de manière traditionnelle en biologie, du carbone 5' libre vers le carbone 3' libre (Figure 1.1).

L'ADN de la Figure 1.1.a sera raccourcie en supprimant, dans la chaîne d'ADN, les positions de carbone 5' et 3' et ne gardant que les nucléotides le composant. Le brin peut alors être codé par AGCGGTCA, offrant ainsi la possibilité de traduire l'information biologique sous forme de mots et d'en permettre le traitement informatisé.

La structure de l'ADN a été résolue par Watson et Crick comme étant une double hélice où le filament d'ADN est apparié avec son complément inverse. Cette complémentarité, due à l'attraction entre les nucléotides A et T et les nucléotides G et C, confère l'aspect *double-brin* sous lequel l'ADN du génome existe dans la cellule (Figure 1.1.b)

Lorsque apparié, chaque nucléotide est attaché à son complément Watson-Crick par un lien chimique d'hydrogène. Le résultat est appelé *paire de bases*, noté *pb*. Puisque l'appariement est bien défini, seule l'information provenant d'un brin noté de 5' vers 3' est utilisée, l'autre étant facilement déductible à partir de ce dernier.

Dans le cas de la Figure 1.1.b, l'information sur l'ADN double-brin est représentée par le mot ACGTAGTAGCTGCTGA.

L'utilité de l'ADN a été définie, au début des années 40, comme étant le "Livre d'instruction"donnant les directives sur la manière de produire et de réguler les protéines. Certaines des régions de cette longue molécule, appelées *gènes*, possèdent l'information nécessaire pour synthétiser des protéines, tandis que d'autres régions sont impliquées, directement ou indirectement, dans des *processus de régulation* ou, en d'autres termes, le contrôle des niveaux de production individuels.

Alors que certains gènes sont constitués de chaînes continues, d'autres sont constitués de plusieurs mots, appelés *exons*, séparés par des zones tampons appelées *introns* et dont l'utilité n'est pas encore complètement connue (*Brenner et al., 01*).

Les génomes possèdent une taille imposante. Par exemple, le génome de la plante *Arabidopsis thaliana* -une mauvaise herbe - est de 160 millions de paires de bases, réparties en cinq chromosomes, c'est-à-dire en cinq molécules différentes.

1.1.2 Transcription et ARN messagers

Deux étapes majeures sont impliquées dans la génération de l'ARN *messager* (ARNm), considéré comme étant le précurseur de la protéine finale. Ces étapes sont :

1. Le *processus de transcription*, qui permet de copier la région codante pour une protéine à partir de l'ADN. La copie résultante, appelée ARNm possède deux différences significatives avec l'ADN original.
   - Les nucléotides de Thymine (T), sont remplacés par leur équivalent, l'Uracile (U).
   - Un atome d'oxygène est ajouté à toutes les bases, rendant l'ARNm plus actif et lui permettant d'exister sous une forme simple brin.
2. L'ARNm subit des *éditions* à l'intérieur du noyau avant d'être transféré à l'extérieur de ce dernier pour être traduit. Le processus d'édition permet entre autre :

- L'*épissage* de l'ARN, qui consiste en la suppression des introns de l'ARNm et la concaténation des exons. Les combinaisons d'excision d'introns pour des messagers provenant d'un même gène initial peuvent différer produisant ainsi transcrits distincts.
- L'ajout d'une coiffe, composée d'une suite de nucléotides codants pour la localisation cellulaire où la protéine finale sera envoyée.
- L'ajout d'une queue *polyA*, composée d'une suite de nucléotides d'adénosine, servant à identifier la fin de l'ARNm (*Wu et al., 03*).

L'ARNm résultant sera ensuite transporté à l'extérieur du noyau pour être traduit.

1.1.3 Génération de protéines

Dans la Figure 1.2, l'ADN est transcrit en son ARNm complémentaire. Ce dernier sera ensuite édité pour supprimer les introns, représentés par des bases en gras, ainsi que pour ajouter la coiffe et la queue polyA au début et à la fin de l'ARN respectivement. À partir de l'ARNm édité, seule la région codante sera traduite. Chaque acide aminé de la protéine est représenté par sa transcription courte de trois lettres.

La machinerie cellulaire utilise l'ARNm comme modèle pour synthétiser les protéines. Au fur et à mesure que le modèle est lu, chaque suite consécutive de trois nucléotides, appelée *codon*, est transformée en son acide aminé équivalent. La concaténation des acides aminés issus de la traduction de l'ARNm, permet de générer la protéine finale. Il est à noter que la coiffe et la queue polyA, ajoutées durant l'édition de l'ARNm, ne sont pas traduites. De ce fait, un ARNm ayant une région codante de 30 bases génère une protéine composée de 10 acides aminés, *aa* (Figure 1.2). Il existe $4^3 = 64$ codons possibles tandis qu'il n'existe que 20 acide aminés connus. Cette redondance est expliquée par la *dégénérescence* du code génétique qui permet d'utiliser des triplets différents pour coder un seul acide aminé. Par exemple, les triplets GGU, GGC, GGA et GGG codent tous pour l'acide aminé Glycine (*Wu et al., 03*).

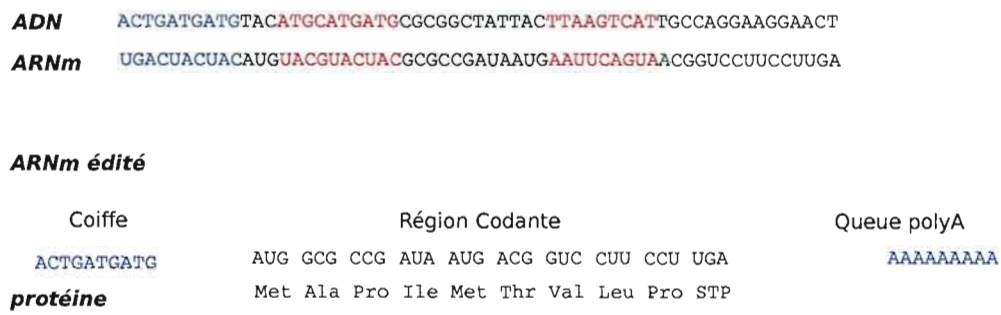


Figure 1.2 Synthèse d'une protéine à partir d'une région codante

Une première théorie proposait que chaque gène code pour une seule protéine. Cependant, la découverte des phénomènes d'épissage alternatifs où certains exons de l'ARNm peuvent aussi être supprimés selon des principes combinatoires et moléculaires complexes, a permis de conclure qu'un gène peut coder pour plusieurs protéines.

Il a été postulé que chaque protéine est définie, en termes de rôle et de structure, par la composition et l'arrangement de ses acides aminés.

1.2 Principes fondamentaux en bioinformatique

Dans le but de mieux de comprendre l'organisation du génome et d'élucider les mécanismes par lesquels les gènes sont régulés et exprimés, la disponibilité de la séquence d'ADN d'un organisme est capitale. Les technologies actuelles ne permettent pas de séquencer plus de 1 000 nucléotides consécutifs. Pour déduire la séquence génomique d'un organisme, souvent composée de plusieurs millions de paires de bases, les laboratoires utilisent une technique nommée *shotgun sequencing*. Cette approche brise l'ADN nucléaire en plusieurs petits fragments qui sont subséquemment séquencés et ensuite assemblés pour reconstituer la séquence initiale du génome. L'assemblage fait ici référence à un processus permettant de générer la chaîne parent initiale, appelée *super-chaîne*, contenant tous les fragments séquencés. Il existe, en théorie, plusieurs super-chaînes possibles et on pourrait penser à celles obtenues par la simple concaténation de tous les fragments dans des ordres différents. Par contre, en pratique il n'existe qu'une seule

Séquence A	A	C	T	T	-	A	G	T	T	-	C	A	G	C
Séquence B	A	-	T	T	T	A	G	C	G	T	C	T	A	G

Figure 1.3 Alignement des deux séquences A et B

super-chaîne représentant l'ADN initial. La meilleure approximation de celle-ci a été définie comme étant la plus courte des toutes les super-chaînes possibles (*Jones et al., 05*). Le problème d'assemblage reviendrait donc à chercher la plus courte chaîne contenant tous les fragments obtenus par cassure de l'ADN initial. Pour mieux comprendre les algorithmes utilisés dans la recherche de la plus courte chaîne commune entre deux fragments, nous allons introduire d'abord la notion d'alignement et de similarité entre deux séquences avant de revenir plus tard sur une définition plus formelle du problème d'assemblage.

1.2.1 Similarité entre séquences

Un *alignement* entre deux séquences A, de longueur n, et B, de longueur m, où m peut être différent de n, consiste en la matrice à deux lignes, tel que la première et deuxième ligne contiennent tous les caractères de A et de B respectivement. Des espaces, représentés par le caractère '-', peuvent être introduit à n'importe quelle position dans l'une ou l'autre des séquences sous la seule condition qu'une colonne doit contenir au moins un nucléotide (Figure 1.3).

On appelle *identité* une colonne ayant le même nucléotide aux deux lignes, tandis qu'une colonne contenant deux caractères différents est appelée *substitution*. Lorsque la colonne contient un espace ou *gap*, elle est nommé *indel*, faisant référence à l'insertion d'un caractère lorsque la ligne du haut contient l'espace où à la délétion ayant eu lieu lorsque l'espace se trouve sur la ligne du bas. L'exemple de la Figure 1.3 contient selon cette définition, six indentités, cinq substitutions et trois indels. Il existe de nombreux alignements possibles, où le nombre d'identités, de substitutions et d'indels peut différer.
La valeur de chacun des ces alignements est déterminée grâce à une fonction qui attribue

Séquence A	A	C	T	T	-	A	G	-	T	T	C	-	A	G	C
Séquence B	A	-	T	T	T	A	G	C	G	T	C	T	A	G	-

Figure 1.4 Alignement altérnatif des deux séquences A et B

à chaque matrice un *score* S déterminé. Par exemple, cette fonction pourrait attribuer un score de $+1$ aux identités, tandis que les colonnes inégalités ou indels recevront un score de 0. Le score global de l'alignement est ensuite calculé en additionnant les scores spécifiques à chaque colonne. Ainsi, le score de la Figure 1.3 est :

$$S = +1+0+1+1+0+1+1+0+0+0+1+0+0+0 = 6$$

Le but ultime de l'alignement est d'évaluer le niveau de similarité entre deux séquences. Étant donné qu'un plus grand score est synonyme d'une plus grande similarité, un bon alignement est représenté par le plus haut score susceptible d'être obtenu par l'algorithme utilisé. Dans le cas de l'exemple 1.3 le meilleur alignement possède un score de neuf et est représenté à la Figure 1.4.

De manière simplifiée, la différence entre les algorithmes d'alignement dépend de deux critères :

1. La fonction de score utilisée : varier celle-ci permet de résoudre une multitude de problèmes d'alignements (*Gusfield, 97*). Par exemple, en donnant un score élevé aux gaps, on peut forcer l'alignement à se produire tout le long de la séquence, produisant ainsi un alignement dit *global*. Par contre, lorsque la similarité entre les séquences est spécifique à une sous-région, l'alignement mettant en évidence cette homologie ne doit pas pénaliser les indels, surtout celles se trouvant aux extrémités des séquences. Ce type d'alignement est dit *local*.
2. Le type de recherche effectuée :
   - Les alignements de type exacts se basent sur des algorithmes utilisant la programmation dynamique. Les résultats obtenus par ces approches sont optimaux au dépend d'une complexité élevée, surtout lorsque de longues séquences sont

alignées. Parmi les algorithmes de programmation dynamique les plus reconnus, on retrouve ceux de Smith-Waterman (*Smith et Waterman, 81*) et Needleman and Wunsch (*Needleman et Wunsch, 70*).

- Les alignements de type inexacts utilisent des heuristiques pour détecter les meilleurs alignements. Cette stratégie sacrifie la garantie d'optimalité pour une amélioration significative du temps d'exécution, estimée entre 10 et 100 fois plus rapide que la recherche utilisant la programmation dynamique. Les heuristiques d'alignement les plus utilisées sont BLAST (*Altchul et al., 97*) et FASTA (*Pearson, 04*).

1.2.2 Assemblage des fragments

Tel que défini précédemment, le problème de l'assemblage revient à trouver la chaîne la plus courte contenant toutes les séquences obtenues par la fragmentation de l'ADN initial. Ce problème, connu aussi sous l'acronyme S.C.S, (Shortest Common Superstring), est connu pour être NP-Complet (*Gallant et al., 80*). Du fait des importantes implications de cette problématique dans les domaines de la compression de données et de l'assemblage génomique, plusieurs heuristiques ont été proposées pour produire une solution approchée. Parmi celles-ci, une approche simpliste vorace consiste en la fusion de deux fragments ayant le plus d'homologie jusqu'à obtention d'une seule séquence finale (Figure 1.5)

L'ADN initial de la Figure 1.5 est brisé en six fragments distincts provenant tous du brin 5'. Dans le cas du fragment *f4*, une erreur s'est produite lors du séquençage à la première base du fragment. Les deux fragments ayant le meilleur score d'alignement seront ensuite fusionnés, générant ainsi une nouvelle chaîne contenant les deux fragments initiaux. Lorsqu'une colonne contient plus qu'un nucléotide, celle-ci sera corrigée par consensus majoritaire sur les autres nucléotides de la colonne lors de l'alignement final. Le fragment, ou *contig*, *f12* obtenu à la dernière étape de l'assemblage constitue une représentation fidèle du brin 5' de l'ADN initial.

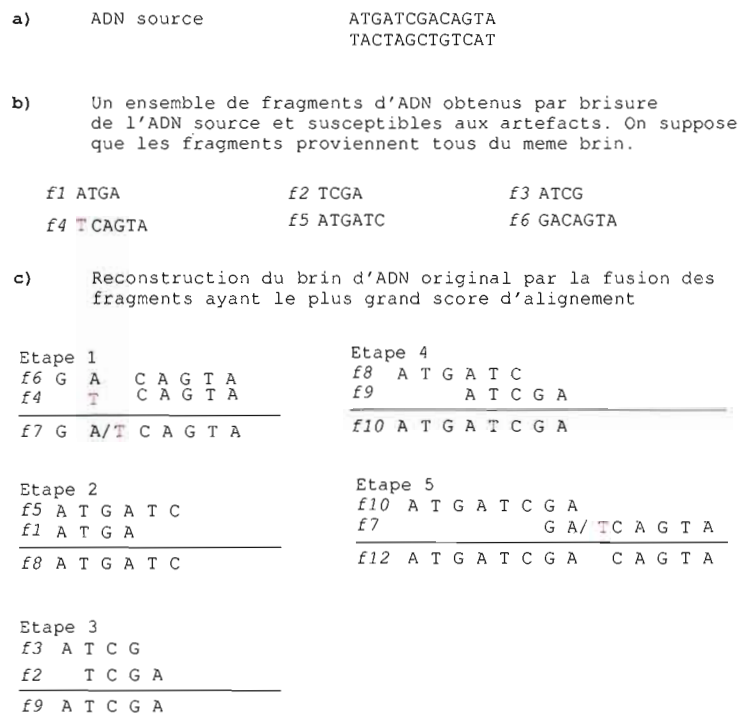


Figure 1.5 Assemblage de fragments d'ADN en utilisant un algorithme vorace

"Bien que l'assemblage de fragments semble être trivial, ce procédé est compliqué par de nombreux facteurs exacerbants" (*Kececioglu et Myers, 95*).

1. Les fragments pourraient ne pas produire la chaîne initiale si les fragments produits ne couvrent pas la totalité de la chaîne.
2. Les séquences produites contiennent des erreurs de séquençage. Ces erreurs peuvent survenir sous plusieurs formes et biaiser la production de la super-chaîne. Les différents types d'erreurs ainsi que leurs effets seront détaillés dans le prochain chapitre.
3. L'assemblage de fragments identiques mais provenant de deux endroits différents dans l'ADN initial est toujours problématique puisqu'on ne possède aucun indice sur la région d'où ces régions parasites sont issues.
4. Puisque les fragments sont obtenus à partir du brin 5' ou 3' de l'ADN, L'orientation du fragment n'est pas préalablement connue (*Kececioglu et Myers, 95*).

CHAPITRE II

GÉNÉRATION DES SÉQUENCES (JEUX DE DONNÉES)

Sur la base des connaissances définies plus tôt, le présent chapitre a pour but de décrire les procédés, d'abord biologiques et ensuite bioinformatiques, utilisés pour extraire et préparer les séquences d'EST. L'emphase sera spécialement mise sur les artéfacts susceptibles de corrompre la qualité des EST et les techniques utilisées pour détecter ces artéfacts et les corriger lorsque cela est possible. Une brève description de chacune des sous-étapes utilisées pour générer et nettoyer les séquences produites dans le cadre du projet FGAS sera exposée.

2.1 Utilité des EST

Le blé représente la céréale la plus importante dans la production alimentaire ainsi qu'une espèce idéale pour étudier la tolérance au froid et à d'autres stress. Malgré l'engouement scientifique et commercial pour cette plante, son génome n'a pas encore été séquencé, dû principalement à sa taille gigantesque, estimée à 16 700 million de bases, ou 127 fois la taille de la plante Arabidopsis (*ornl*).

Une alternative utilisée par plusieurs laboratoires de génomique consiste à obtenir un grand nombre d'EST (Expressed Sequence Tags), courtes séquences provenant d'ARNm, donnant ainsi un accès direct à la partie exprimée du génome. Dans le cas du blé, la partie exprimée est estimée de manière conservatrice à 30 000 gènes ou approximativement 1% de la totalité du génome dont la taille a été approximée à 13,5 milliards de bases. Par

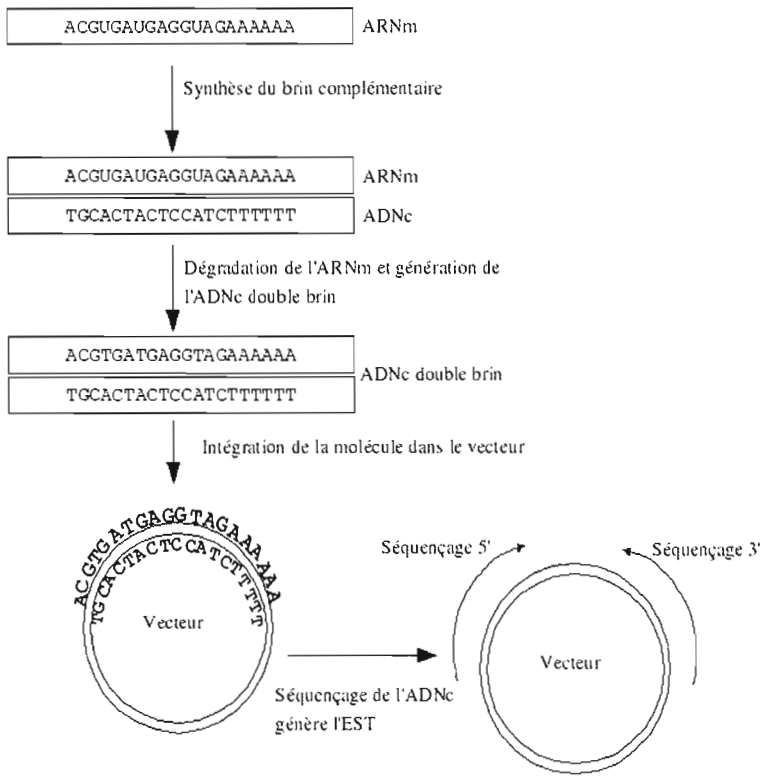


Figure 2.1 Processus de génération des ADNc et des EST

comparaison, les génomes de l'humain et de la souris possèdent respectivement 3 et 2,7 milliards de bases et contiendraient un nombre de gènes similaire à celui contenu dans le blé (*ornl*).

En plus d'être une approche de plus en plus privilégiée lors de l'étude d'espèces dont le génome n'a pas encore été résolu, tel que dans le cas de l'orge (*Close et al., 04*), la tomate (*Fei et al., 04*) et le peuplier (*Sterky, 04*), les EST sont considérés comme étant la commodité la plus produite dans le monde de la génomique des plantes en termes de séquences et du nombre de nucléotides générés (*Rudd, 03*).

2.1.1 Génération des EST

L'ARN utilisé dans la préparation des EST est hautement instable à l'extérieur de la cellule. Pour cette raison, une enzyme est utilisée pour générer le brin complémentaire de l'ARNm, le transformant ainsi en ADNc (Figure 2.1) possédant une plus grande stabilité à l'extérieur de la cellule. L'ADNc est ensuite inséré dans une séquence circulaire artificielle, appelé un *vecteur de clonage*, pour permettre sa sauvegarde et sa réplication à l'intérieur de bactéries. L'ensemble des clones obtenus représente alors une *librairie*. Le séquençage de la librairie consiste en la sélection de clones de manière aléatoire et leur séquençage d'une ou l'autre des extrémités 3' ou 5', générant ainsi des EST.

Puisque les EST proviennent de séquences exprimées dans la cellule sous forme d'ARNm, leur utilisation va au delà du gain de l'information de nature génétique extraite par le séquençage. En effet, lorsque les librairies produites proviennent de conditions expérimentales différentes, la variation dans le taux d'expression d'un même gène d'une librairie à une autre peut être un bon indice sur l'effet de la condition expérimentale sur la cellule.

D'autre part, les EST sont considérés comme une alternative non dispendieuse au séquençage complet. Au delà du coût inhérent au nombre de séquences nécessaires pour reproduire le génome de l'organisme avec fidélité, l'assemblage des EST nécessite moins de temps et certainement moins d'expertise qu'un projet d'assemblage génomique. De plus, leur nature peu répétitive, ainsi que leur courte séquence, facilitent grandement le processus de reproduction des ARNm initiaux et réduisent significativement les possibilités d'erreurs.

Lorsque utilisés comme complément à l'assemblage génomique, l'utilité des EST réside dans la partie spécifique à l'annotation. En plus de servir comme données d'apprentissage pour les outils de prédiction, ils sont souvent utilisés pour la détection et la correction des gènes prédits. Chez *Arabidopsis*, le recours aux EST a permis de détecter 240 gènes qui avaient été "oubliés" par les algorithmes de modélisation de gènes (*Haas et al., 02*). L'approche utilisant les ADNc a aussi largement été utilisée dans l'annotation des

régions non traduites des ARN messagers, ainsi que pour définir ou corriger les limites introns-exons ou dans la détection de phénomènes d'épissage alternatif.

2.2 Survol du projet FGAS

Le projet FGAS (Functional Genomics of Abiotics Stress) est un projet qui regroupe des laboratoires de recherche de plusieurs universités canadiennes, notamment au Québec à travers l'Université du Québec à Montréal. Ce projet a pour but de découvrir les effets de l'exposition du blé aux stress environnementaux, dits *abiotiques*, tels que la chaleur, le froid, la sécheresse ou la salinité, afin de mieux connaître la réponse de cette plante et les étapes de son acclimatation aux conditions hivernales canadiennes.

Le processus d'acclimatation est gouverné par un système génétique induit par exposition graduelle de la plante à de basses températures. Simplement dit, le blé doit être exposé à des températures de plus en plus froides dans le but d'acquérir la résistance qui lui permettra de survivre à l'hiver. Sous des conditions normales, telles qu'une bonne température du sol, la bonne profondeur de la tige, etc., ce processus dure entre 8 et 12 semaines. Durant les quatre premières semaines de cette période, le blé est soumis à une température moyenne de $+9°C$ tandis que la température moyenne durant la deuxième période est généralement aux alentours de $+3°C$. Le blé acclimaté peut ensuite endurer des températures allant jusqu'à $-20°C$ (*umanitoba*).

Les connaissances acquises du projet FGAS ont permis d'améliorer la résistance de certaines plantes face au froid et à d'autres types de stress, augmentant ainsi la productivité et la qualité des récoltes. Le blé représente une espèce de choix pour étudier l'effet du stress abiotique sur les plantes vu son importance économique et la tolérance naturelle de cette plante. Cependant, la taille imposante de son génome et l'existence de six copies de ce dernier dans chaque cellule, rendent son séquençage utilisant les technologies actuelles, laborieux et onéreux.

Pour évaluer l'effet du stress appliqué par l'environnement et la tolérance du blé au froid, le projet *FGAS* a choisi d'utiliser l'approche des EST comme première étape

dans la sélection des gènes régulés à la hausse ou à la baisse dans un blé acclimaté au froid et pouvant survivre à des températures au-dessous du point de congélation, pour, dans une deuxième étape, confirmer leur implication dans le processus d'acclimatation.

Les données relatives au projet FGAS proviennent de 11 librairies ADNc préparées à partir de plantes acclimatées à de basses températures. Le séquençage des ces ADNc a donc permis de générer un total de 110,544 ESTs spécifiques à la tolérance au froid. Le jeu de données généré localement a été enrichi par 280,000 séquences obtenues à partir de différents tissus de plantes de blé, dans des conditions normales dans les projets du NSF et DuPont, deux projets d'EST, utilisant un blé n'ayant subi aucun stress et et dont les données sont publiques.

2.3 Artéfacts de séquençage ou d'origine biologique

Le but de l'assemblage d'EST est de réduire la complexité de la librairie en dérivant un consensus du gène parent par alignement de toutes les séquences transcrites à partir de ce dernier. Pour ce faire, les algorithmes de clustering et d'assemblage se basent sur l'homologie pour regrouper les séquences similaires ensemble. Par conséquent, toute similarité dûe au hasard ou a des artéfacts de séquençage, entre des séquences transcrites à partir de gènes différents peut biaiser les processus de clustering ou d'assemblage.

Les artéfacts d'EST peuvent être subdivisés en deux catégories :

Artéfacts biologiques

- Les queues polyA ou leur complément polyT constituent une caractéristique principale des EST. Ces séquences homo-nucléiques représentent un biais naturel hérité des ARNm dont les séquences ont été dérivées. Bien que la présence des queues polyA/T ne soit pas garantie, leur occurrence est généralement détectée vers le début ou la fin d'un EST.
- Des phénomènes peu connus peuvent causer la fusion de deux séquences durant la préparation d'EST et ainsi produire une *chimère*. Par définition, une chimère est

une séquence provenant de la concaténation d'au moins deux EST indépendants. Les chimères sont généralement d'une longueur supérieure à la taille moyenne d'un EST et peuvent contenir plus d'une queue polyA/T provenant de chacun des séquences ayant servi dans la concaténation.

Artéfacts de séquençage

- Durant le séquençage de clones, la réaction biochimique responsable de la production de l'ordre nucléique ne fournit aucune garantie sur la taille de la séquence générée. De ce fait, la réaction peut se poursuivre même dans le vecteur de clonage et générer ainsi des séquences *contaminées* par le vecteur. La contamination ne se limite pas au vecteur mais peut aussi être de nature bactérienne, humaine ou tout autre type de contamination externe.
- On appelle *basse complexité*, toute région dans une séquence où la composition en terme de nucléotides ou acides aminés semble être non aléatoire ou biaisée. La basse complexité peut être réelle i.e. le gène contient un taux élevé de répétition et de redondance, ou artificiellement générée par le séquenceur. Lorsque artificielle, la basse complexité peut être présente vers le début ou la fin de la séquence, ou se poursuivre tout le long de la séquence, signe que la réaction de séquençage a échoué (*Chou et al., 01*). La basse complexité constitue une part de biais non négligeable dans le clustering des séquences.
- Le produit de la machine à séquencer est appelé *chromatogramme*. Ce dernier est composé de quatre courbes de différentes couleurs (Fig. 2.2), chacune d'entre elles étant assignée à un nucléotide particulier.

  Lors de la prédiction de la séquence, les logiciels d'identification de nucléotides attribuent une qualité à chaque base prédite en se basant sur plusieurs critères, tels l'amplitude des courbes, le bruit de fond, ainsi que sur d'autres paramètres.

  La qualité d'un nucléotide Q est un nombre entier variant entre 0 et 99. Ce paramètre peut être transformé en probabilité d'erreur P_e selon la formule suivante :

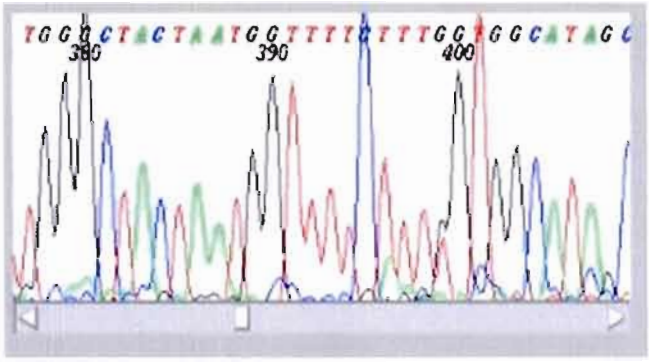


Figure 2.2 Chromatogramme généré par le séquençage d'un EST

$$P_e = \frac{1}{10^{\frac{Q}{10}}} \tag{2.1}$$

Ex. : Selon la formule 2.1, la probabilité d'erreur d'un nucléotide m ayant une qualité Q =20 est P_e =0,01. Ainsi, la probabilité que le nucléotide m ait été mal prédit est seulement 1%. Quoi que le critère de bonne qualité varie d'un projet à un autre, une valeur de Q supérieure à 20 est généralement considérée comme étant acceptable.

Le taux d'erreurs de séquençage se situe entre 2% et 3% de nucléotides erronés dans un projet de séquençage ne contenant pas d'aberrations majeures (*Hiller et al., 96*). Ces erreurs de séquençages se produisent majoritairement au début et vers la fin de la séquence où la réaction est la moins efficace et prennent la forme de mutation, insertions ou délétions (*Ewing et al., 98*).

2.4 Prétraitement des séquences

Plusieurs stratégies de nettoyage et de contrôle de contamination des séquences ont été élaborées. Ces stratégies se basent toutes sur quatre étapes majeures. À savoir : 1-Décontamination, 2-Nettoyage de queues polyA/T, 3-Supression des régions de basse qualité et 4-Détections de chimère et de régions de basse qualité.

Dépendamment de la sensibilité désirée, d'autres tests ou critères de sélection peuvent être rajoutés au processus de contrôle de qualité.

2.4.1 Décontamination de la séquence

Pour effectuer cette étape, les logiciels utilisent des algorithmes d'alignements sensibles aux différences entre séquences, pour détecter les occurrences de contamination dans les EST (*Sanbi_2*). Cross-Match (*phrap*) est un outil développé par Phil Green et utilise une implémentation de l'algorithme de Smith-Waterman pour trouver des identités entre une base de données de vecteurs et les séquences passées en entrées. L'occurrence de ces identités est ensuite remplacée dans la séquence par le caractère "X".

Le logiciel Lucy (Chou et al., 01) utilise une autre approche pour pallier au problème de contamination. Pour détecter la présence de vecteurs, Lucy utilise un fichier contenant *les sites d'épissage* (splice sites) des vecteurs devant être détectés. Les sites d'épissage sont des séquences dont la taille varie entre 100 et 150 nucléotides et qui représentent le préfixe et le suffixe de la séquence du vecteur, ainsi que de son complément inverse. Lucy utilise cette information pour détecter la présence de contamination dans le début et vers la fin de la séquence sans prendre en considération le vecteur dans sa totalité. Ceci a pour résultat de réduire significativement le temps consacré pour repérer les occurrences de vecteur dans les séquences.

Puisque les sites d'épissage sont plus susceptibles de se trouver vers le début et vers la fin de l'EST, où la basse qualité ne permet pas de détecter l'occurrence de manière exacte, Lucy utilise trois fenêtres distinctes où l'identité minimale nécessaire pour inférer la contamination est différente. Cette approche suppose que plus on avance dans la séquence, meilleure est la qualité et donc, meilleure est l'identité entre la séquence et le vecteur (Fig 2.3). Il est à noter que cette approche ne peut être utilisée pour détecter la contamination externe puisque cette dernière peut se produire à n'importe quel endroit de la séquence.

Lucy utilise une première fenêtre de 40 bases et recherche toute occurrence exacte du

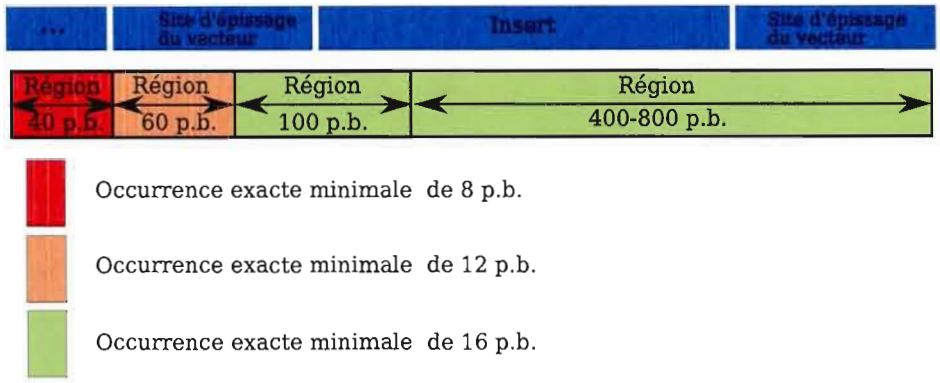


Figure 2.3 Les étapes de recherche de fragments par Lucy

vecteur composée d'au moins huit bases. Dans une deuxième fenêtre de taille 60, Lucy essaye de détecter une occurrence minimale de 12 bases. En cas de succès, Lucy passe alors à la fenêtre suivante de taille 100 où une occurrence minimale de 16 nucléotides est nécessaire pour inférer la présence du vecteur.

A priori, le problème de décontamination semble être trivial et les outils utilisés à cette étape se basent sur des algorithmes peu complexes pour détecter les occurrences du vecteur de clonage. En réalité, la présence d'erreurs provenant du séquençage ou de l'étape d'extraction de bases altère l'occurrence du vecteur dans la séquence. Ainsi les logiciels tels que Lucy et Cross-Match, qui utilisent des règles définies de manière empirique, voient leur tâche grandement compliquée lorsque le problème de contamination est amplifié par la basse qualité des nucléotides. Ceci est le cas du projet FGAS qui souffre particulièrement du problème de basse qualité, indiquant la présence d'un grand nombre de nucléotides erronés.

Une solution simpliste au problème de contamination pourrait consister en la suppression d'un nombre x de bases à partir des extrémités. Ceci pourrait s'avérer un choix intéressant pour tester l'effet hypothétique de la contamination sur les étapes subséquentes dans un projet d'assemblage. Ainsi, si de meilleurs résultats sont obtenus à l'étape de clustering après suppression des extrémités de toutes les séquences, ceci

pourrait être un indice suffisant sur la contamination des séquences. Cette alternative mènerait cependant à la perte d'informations pertinentes lorsque les extrémités supprimées représentent de la séquence réelle et ne constitue donc pas une option sérieuse au problème de décontamination.

Dans le cadre du projet FGAS, les séquences ainsi que leur qualité ont été extraites à partir des chromatogrammes disponibles, grâce au logiciel PHRED utilisé avec les paramètres par défaut. Un examen superficiel des données, s'appuyant sur le calcul du pourcentage de bases ayant une qualité supérieure à 20, a permis de constater que la qualité moyenne des séquences était relativement basse. Pour réduire le biais introduit par la qualité des séquences, un processus de prétraitement rigoureux a été élaboré.

Pour détecter et masquer la contamination externe, ainsi que les occurrences du vecteur dans la séquence, les logiciels Cross-Match et Lucy ont été utilisés pour traiter les séquences de manière indépendante. Pour valider la qualité des résultats, deux échantillons aléatoires, le premier contenant des séquences nettoyées par Lucy et le deuxième constitué de séquences obtenus par nettoyage avec Cross_Match ont été générés. La comparaison des EST de chaque échantillon contre eux-mêmes a permis de déceler une région commune chez un bon nombre de séquences. Cette région, située soit vers la fin ou au début des séquences contaminées et dont la taille moyenne fût estimée à 70 bases à été identifiée comme étant de la contamination par vecteur, non détectée par Lucy ni par Cross-Match. L'alignement de cette région avec le vecteur utilisé (Figure 2.4), a permis de découvrir la présence de mutations, délétions et insertions dans l'occurrence du vecteur chez la majorité des séquences contaminées. Ces modifications sont supposées provenir principalement de l'inefficacité de la réaction de séquençage et cela a pu être confirmé par la basse qualité de la majorité des bases en question.

L'alignement des séquences, numérotées de 1 à 7 avec la séquence du vecteur met en évidence plusieurs mutations, insertions et délétions qui compliquent la tâche de détection du vecteur. Ces différences entre le vecteur et les EST sont principalement dues a des artéfacts de séquençage confirmés par la qualité des bases en question.

```
Vecteur  TTATAATACGACTCACTATAGGGACCACTTTGTACAAGAAAGCTGGGTACGC-GTAAGCTT
1        TTATAATACGGCTCCCTATAGGGCCCCCTTTGTTCAAGAAAACTGGGTCCGC-GTAAACTT
5        TTATAAAACCACTCCCTATAGGGGCCCCTTTGTACAAAAAAACTTGGTACGC-GTAACTTT
4           T-ATACACCTCACAATAGGGCCCACCTTGTACCACAAAATCTGGTACGC-CTAAAATT
6           TAATACGACTCCCCATAGGGCCCACTTTGTAACAAAAACCGGGTAAGCCGTAAGCCT
7               ACGACTCACCATTGGGGCCCCTTTGTACCAGTAAGCCGGGTACCCGTAAA-CCT
3                CGACTCCCAATAGGGACC-CCTTGTACAAGAAA-CTGGGTACGC-TTAAG-TT
2                         ATTGGGCCCACATTTTACAACAAAACCGGGTACGG-GTAAGGTT

Vecteur  GGGCCCCTCGAGGGATACTCTAGAGCGGCCGCC
1        GGGCCCCTCGGGGGATACTTTAGA
5        GGCCCCCTCG-GGGATATTCTAGGGCGGC
4        GGGCCCCTCAAGGGATACTATAGATCGGCCGCC
6        GGGCCCCTCGAAGGATACTTTAGAGCGGCC
7        GGGCCCCTTGAGGGATAATTTTGAGCGGCCGCC
3        GGCCCCCTCGAGGGAAACTCTAGAGCGCCCGCC
2        CGGCCCCTCCAGGGATACT
```

Figure 2.4 Alignement de la séquence du vecteur utilisé avec un ensemble d'EST contaminés ayant un taux d'erreur élevé

Alors que la variation des paramètres de sensibilité ne semblait pas améliorer le nettoyage des séquences en utilisant Lucy, les séquences nettoyées par Cross-Match semblaient bénéficier de paramètres plus restrictifs. Il a cependant été noté que, lorsque l'homologie requise est trop élevée, Cross-Match a tendance à supprimer des régions non contaminées mais pouvant avoir une légère similarité avec le vecteur. Ceci est un effet secondaire signalé dans la documentation (*phrap*) et la seule solution consiste en la variation des paramètres jusqu'à obtention de résultats satisfaisants. Après plusieurs expérimentations, les séquences nettoyées par Cross-Match semblaient s'être défait de la contamination sans ôter les séquences réelles ayant une légère homologie avec le vecteur.

2.4.2 Nettoyage de la queue polyA/T et détections de chimères

Cross-Match n'offre aucun moyen direct de nettoyer les queues polyA/T. Pour effectuer cette tâche, il est possible d'ajouter une séquence de A/T dans la base de données de vecteurs que Cross-Match prend comme étant un contaminant. L'inconvénient de cette approche est que le nettoyage des queues polyA/T ne se limite pas aux régions de début ou de fin de séquence, où ces suites sont le plus susceptibles de se produire,

mais peut aussi bien se produire ailleurs dans la séquence. L'approche utilisée par Lucy se résume dans ce qui suit. Lucy trouve d'abord une première suite minimale de dix nucléotides consécutifs "T" se trouvant dans les 50 premières bases de la séquence décontaminée et ensuite essaie de rallonger ces suites en trouvant toutes les autres occurrences consécutives de 10 nucléotides "T" séparés entre eux par un maximum de trois erreurs. Pour trouver les queues polyA, Lucy effectue la même opération, mais en cherchant à partir de la fin de la séquence et en remplaçant les nucléotides "T" par "A".

La présence de plusieurs polyA/T dans un EST peut être parfaitement naturelle. Cependant dans plusieurs cas, les séquences possédant de tels motifs représentent des chimères formés par la concaténation de deux EST indépendants contenant chacun un polyA/T. La taille d'une séquence peut aussi servir d'indicateur sur la qualité de la séquence. Lorsque la séquence est anormalement longue, il peut s'agir d'une chimère. Pour valider si une séquence est une chimère ou non, un utilisateur peut effectuer une recherche des deux extrémités sur une base de données de séquences et examiner la cohérence des résultats. Encore une fois, ceci n'est pas une garantie que la séquence soit une chimère car il pourrait s'agir de domaines différents appartenant à la séquence.

Dans le cas des séquences FGAS, Lucy s'est avéré très efficace pour trouver les occurrences de queue polyA/T, bien définies, ne contenant pas plus que les deux erreurs allouées par cet outil. L'inspection visuelle d'un échantillon aléatoire de séquences a permis de repérer un bon nombre de polyA/T non détectées par Lucy, principalement pour les mêmes raisons que celles rapportées dans la section précédente, à savoir, des erreurs dûes a la basse qualité.

Pour détecter ce genre d'aberrations, un script se basant sur un algorithme similaire à celui adopté par Lucy mais utilisant une approche moins permissive a été développé. L'outil permet à l'utilisateur de spécifier la longueur minimale d'une occurrence polyA/T recherchée par défaut ainsi que la taille minimale d'une région pouvant séparer deux polyA/T. La poly-queue recherchée est supposée se trouver selon les expériences

empiriques obtenus par Lucy dans les 50 premières bases. Cependant, lorsqu'une polyqueue est localisée après une occurrence de vecteur non détectée, celle-ci peut se trouver au delà des 50 premières bases. Pour cette raison, la recherche réalisée par le script est effectuée dans les 250 bases de chaque extrémité. Lorsqu'une occurrence polyA/T est détectée, seule la plus longue des sous-séquences parmi celle se trouvant en amont de la queue et celle se trouvant en aval, est retenue. L'autre est automatiquement rejetée puisque selon la théorie biologique, elle est considérée comme de la contamination.

En ce qui concerne les chimères contenues dans les données de FGAS, le script permettant de détecter les queues polyA/T a été modifié pour parcourir la séquence complète et non seulement les extrémités afin d'indiquer les séquences possédant plus qu'une occurrence de queue polyA/T. Ces séquences ont été automatiquement comparées avec les protéines de la base de données publique 'nr' et les résultats ont été manuellement analysés, pour confirmer, lorsque cela était possible, quelles séquences représentaient de vraies chimères. Toutes les chimères ne possédant pas plus qu'un site polyA/polyT ne pouvaient être retirées à cette étape.

2.4.3 Nettoyage des régions de basse qualité

Le but de cette étape est de trouver la plus longue région de chaque séquence ayant une qualité moyenne suffisamment élevée, permettant l'utilisation de la séquence avec confiance.

Le début et la fin des séquences sont particulièrement déficients en termes de qualité puisque c'est à ces endroits que la réaction enzymatique est la moins efficace (*Ewing et al., 98*). Lorsqu'on inspecte les chromatogrammes, on remarque que vers les deux extrémités de la séquence générée, les quatre courbes de couleur, ont une basse amplitude et que leurs pics se situent sensiblement aux mêmes niveaux. Typiquement la valeur de qualité obtenue pour ces régions est inférieure à 20. Une approche simpliste pour supprimer ces régions de basse complexité consiste à supprimer n bases de chaque extrémité de la séquence et de s'arrêter lorsque la séquence a atteint une qualité moyenne

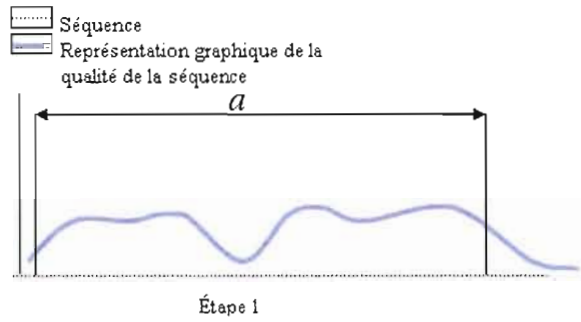


Figure 2.5 Première étape de décontamination de Lucy

supérieure ou égale à Q, où typiquement, $n = 10$ et $Q = 20$.

Lucy utilise une approche plus sophistiquée et plus stricte. La sensibilité implémentée par cette approche se base sur le fait que les nucléotides de basse qualité peuvent se trouver dans une région donnée à une plus haute concentration et biaiser le consensus à cet endroit. Pour détecter ces régions, Lucy emploie les étapes illustrées aux Figures 2.5, 2.6, 2.7.

1. Supprimer la mauvaise qualité se trouvant aux extrémités : à cette étape, il s'agit de trouver deux fenêtres de taille L, au début et vers la fin de la séquence, ayant une qualité minimale Q définie par l'usager. La séquence délimitée par ces deux fenêtres et représentée par la région "a" de la Figure 2.5 est ensuite passée à l'étape suivante.

2. Trouver les zones ayant des taux d'erreurs inacceptables : deux fenêtres de tailles différentes sont utilisées. Une grande fenêtre, 50 nucléotides par défaut, est passée sur la séquence et seules les régions où la qualité moyenne de la fenêtre est suffisamment élevée seront gardées. Ceci a pour effet de générer les régions "b" de la Figure 2.6. Une autre itération est effectuée en utilisant une fenêtre de plus basse taille, idéalement 1/5 de la taille de la fenêtre initiale avec un seuil de qualité moindre. Le but de la fenêtre de grande qualité est d'exclure de larges régions de basse qualité. La deuxième fenêtre est utilisée pour détecter les régions de très

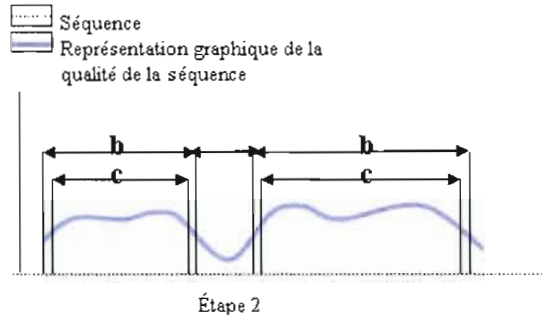


Figure 2.6 Seconde étape de décontamination de Lucy

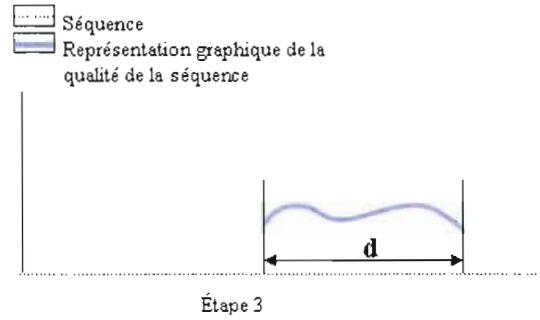


Figure 2.7 Troisième étape de décontamination de Lucy

basse qualité qui pourraient être ignorées par la première fenêtre. La région "*c*" de la Figure 2.6 est obtenue.

3. La dernière étape consiste à trouver la plus longue sous-séquence ayant une qualité moyenne supérieure au minimum spécifié par l'utilisateur. La séquence "*d*" de la Figure 2.7 est obtenue.

Pour extraire la région de plus haute qualité dans chaque séquence du projet FGAS, Lucy a été utilisé avec les paramètres par défaut de manière rcursive. À chaque itération, des séquences ont été rejetées et le processus a été arrêté quand aucun rejet de séquence n'a été obtenu.

2.4.4 Régions de basse complexité

La basse complexité peut être divisée en deux catgories :

1. Artificielle : ce type de basse complexité se caractérise par une basse qualité Q et est souvent supprimée de la séquence (section 2.4.3).
2. Basse qualité due à des répétitions : pour traiter ce genre d'artéfact, l'utilisation d'une base de données de répétitions spécifique à l'organisme est de mise. Dans le cas d'organismes non complètement séquencés, les répétitions de l'organisme le plus proche peuvent être utilisées. RepBase (*girinst*) est une base de données de répétitions continuellement mise à jour offrant la possibilité de scanner le jeu de données contre des répétitions connues dans d'autres espèces et ainsi, supprimer les EST susceptibles de biaiser l'assemblage.

Dans le but de détecter les régions de basse complexité, les séquences du projet ont d'abord été comparées en utilisant l'outil BLAST, avec la base de données de répétitions RepBase, spécifique aux plantes. Là encore, la basse qualité des séquences fût un inconvénient majeur dans la détection des répétitions.
Plusieurs seuils de similarité ont été utilisés pour détecter les répétitions entre un échantillon de séquences choisies manuellement par les biologistes et les répétitions de la base de données RepBase. Les résultats des comparaisons ont été manuellement inspectés pour déterminer la pertinence des résultats et le niveau de sensibilité de chacun des seuils utilisés envers le bruit introduit par la basse qualité des séquences. Les résultats obtenus suggéraient qu'un seuil de $1e^{-18}$ détectait de manière plus efficace que les autres seuils testés les occurrences de répétitions dans les données.

Les répétitions artificielles ont dû être examinées en utilisant une autre approche. Dans le but de formaliser la notion de répétition artificielle dans le cas du projet FGAS, plusieurs dizaines de séquences soupçonnées de contenir de la basse complexité artificielle ont été examinées manuellement. Les séquences ont d'abord été comparées avec les bases de données publiques et toutes celles ne possédant aucune similarité avec les bases de données publiques ont été examinées davantage. Ces séquences semblaient contenir des

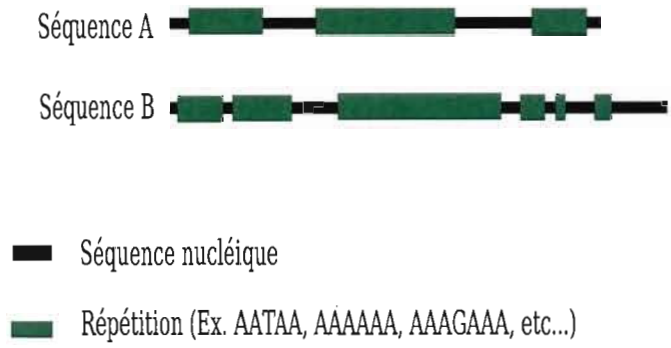


Figure 2.8 Détection de basse complexité dans deux séquences A et B

répétitions de tailles différentes avec une taille de répétition minimale de quatre *pb* et un maximum de une erreur. Une répétition à alors été définie comme étant un mot de taille minimale de quatre bases identiques contenant, au maximum, une erreur .

Toutes les séquences ont été ensuite inspectées et toutes celles contenant plus que 50% de répétitions ont été filtrées(Figure 2.8).

Les régions en vert dans la séquence représentent les basses complexités ou répétitions telles que définies dans le cadre du projet FGAS. Dans le cas de la séquence B, la basse complexité couvre plus que la moitié de la séquence tandis que la séquence A contient moins que 50% de répétitions. Pour cette raison, La séquence A est gardée tandis que la séquence B est rejetée.

À la fin de toutes les étapes de nettoyage, les séquences dont la taille était inférieure à 100 bases ont été rejetées. Seulement 271,226, parmi les 390,000 séquences initiales ont passé l'étape du prétraitement.

CHAPITRE III

CLUSTERING ET ASSEMBLAGE DES EST

Dans ce chapitre, nous allons définir, dans un premier temps, les processus de clustering et d'assemblage considérés comme deux étapes primordiales dans tout projet ayant trait aux EST. Dans la seconde partie de ce chapitre, nous présenterons deux méthodes, l'une se basant sur la fréquence de mots et l'autre sur la similarité par alignement de séquences, pour effectuer le clustering des séquences et nous discuterons des forces et faiblesses de chacune des méthodes. Pour conclure le chapitre, les deux approches les plus utilisées dans l'assemblage des séquences seront brièvement abordées.

3.1 But du clustering et de l'assemblage

Durant les dernières années, les EST ont été utilisés à des tâches aussi diverses que l'étude d'expression à grande échelle ou la construction d'une carte physique de l'humain (*Hudson et al., 95*). Les EST ont permis d'obtenir des résultats intéressants avant même que les efforts de séquençage de certaines espèces soient achevés. Ce succès inattendu a conduit à une montée fulgurante du nombre de séquences disponibles. Les seuls obstacles à une utilisation plus efficace des EST sont les suivants :

1. La quantité et la qualité des séquences disponibles,
2. Le manque de structure permettant de classifier les données de manière à pouvoir en tirer le maximum d'informations.
3. Enfin, la nature fragmentaire des données disponibles (*Haas et al., 02*).

3.1.1 Définition du clustering et de l'assemblage

Pour pallier aux obstacles restreignant l'utilisation efficace des EST, plusieurs projets de classification sont en cours. La classification des données se fait en *clusters*, appelés aussi index, contenant toutes les séquences appartenant à un seul et même gène. Cette étape de clustering permet d'effectuer une fragmentation physique des données et de diviser l'ensemble initial en plusieurs sous-groupes, ou clusters, homogènes.

Pour atténuer des effets de la nature fragmentaire des données, les séquences appartenant à un même cluster sont ensuite assemblées. Cet assemblage a comme résultat la génération de la séquence consensus de l'ARNm parent, ou *contig* et la mise en évidence, lorsque cela s'applique, des autres formes d'épissage, appelées *isoformes*, du gène.

A titre d'exemple, le clustering des séquences présentées à la figure 3.1(b), a permis de générer quatre clusters distincts. Les séquences de chaque cluster sont représentées avec une même couleur pour mettre en évidence le lien d'homologie attendu entre ces séquences. Le cluster 4 représente un cluster singleton, constitué par une seule séquence tandis. Les assemblages des 4 clusters en (c), a permis de générer les consensus des ARNm parents, ou chaque contig est représentés pas les lignes en pointillés dans chaque cluster. Le cluster 2 a permis de générer deux isoformes différents, prouvant ainsi que le gène représenté par le cluster 2, subit un épissage alternatif (Figure 3.2).

L'inspection des deux séquences de l'isoforme 1 et de l'isoforme 2 provenant du cluster 2 permet de constater que la région en gras dans l'isoforme un n'existe pas dans la deuxième séquence. Cette région représente un exon utilisé dans la fabrication de la protéine codée par l'isoforme 2 et omise dans la génération de l'isoforme un. Il s'agit là d'une forme d'épissage alternatif simple appelée omission d'exons 'Exon skipping'.

Les deux étapes de clustering et d'assemblage sont aussi importantes dans la réduction du niveau de redondance des données que dans l'amélioration de la qualité des consensus générés. Les clusters produits permettent aussi d'organiser les données de manière à faciliter l'extraction d'une multitude d'informations, allant des taux d'expression de

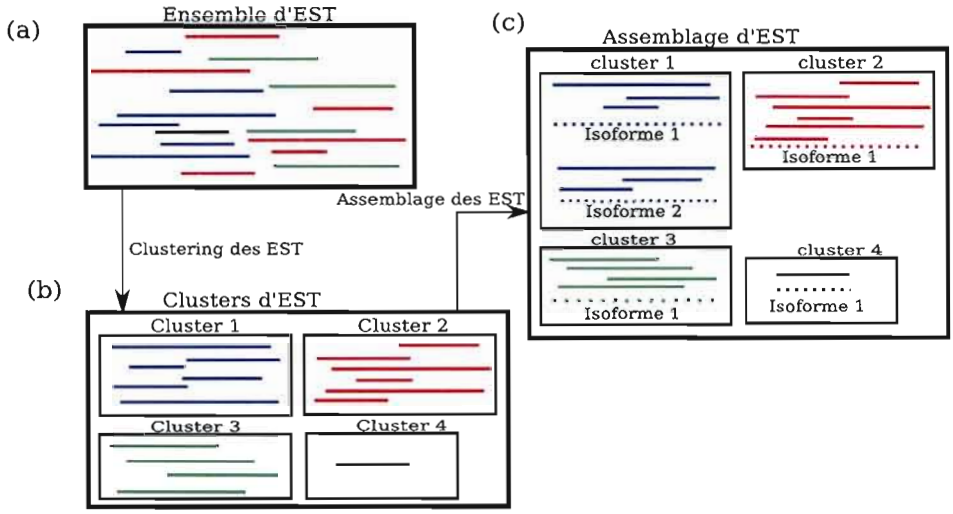


Figure 3.1 Processus de clustering et d'assemblage

```
>Isoforme 1
ACGGATGCTGATCGAATCGATCGATTCGATGCTTATGCTAGCTACG
ATAGTCGCTGAATGACTGCGCATGCATGATGCGATCGATGCATGCA
ATGCATGCATGCAATGCGATCGAGCTGGCTAGGCATGCTGTCGATC
ATGATGAGTCTAACTAGAGTAGAGTCTCGTAGCTAGCTGATGCTAG
CATGCTGACTGATCGACGTAGCTGATCGTAGTCGATGCTGTGTGAC
ATATTATGCCCCTAGGATGCTGCATCGTAGCTGATCGTGCTAGCTG
CTAGCGATCGTAGCTGATCGTAGCTAGCTGTGTAGCTGATCGTAGC
CATGATCGTAGCTTCGTTTTTTTTTAGTCGATGCTAGCTGATGCTG
CTAGCTAGCTGATCTAGCTGATCGTAGCTGACTGATGTGCTGTGAC
CATGCTGAAATCGTAGCTGACTGATCGTAGCTAGCTATATATGCTG
TTTCGATGCTAGTCGATGCTGATGCTGTGTAGCTGATCGTGTGCTA
>Isoforme 2
ACGGATGCTGATCGAATCGATCGATTCGATGCTTATGCTAGCTACG
ATAGTCGCTGAATGACTGCGCATGCATGATGCGATCGATGCATGCA
ATGCATGCATGCAATGCGATCGAGCTGGCTAGGCATGCTGTCGATC
ATGATGAGTCTATGTAGTCGTAGCTAGCTGTGTGTCGTAGCTGAAA
CATGTGATGCTAGCTGATCGTGCTAGCTGATCGTGTGTGTGTATCG
CTAGTATTATATTGCGGTACTAGCTGATGTCGTAGCTAGTCGATTG
CTAGCTAGCTTTTTCATGCATGTAGCTGTATTATTAGCTGATAGCG
ACGTACTGGTGACAAACTAGAGAGTCTCGTAGCTAGCTGATGCTAG
CATGCTGACTGATCGACGTAGCTGATCGTAGTCGATGCTGTGTGAC
ATATTATGCCCCTAGGATGCTGCATCGTAGCTGATCGTGCTAGCTG
CTAGCGATCGTAGCTGATCGTAGCTAGCTGTGTAGCTGATCGTAGC
CATGATCGTAGCTTCGTTTTTTTTTAGTCGATGCTAGCTGATGCTG
CTAGCTAGCTGATCTAGCTGATCGTAGCTGACTGATGTGCTGTGAC
CATGCTGAAATCGTAGCTGACTGATCGTAGCTAGCTATATATGCTG
TTTCGATGCTAGTCGATGCTGATGCTGTGTAGCTGATCGTGTGCTA
```

Figure 3.2 Épissage alternatif d'un gène

certains gènes à la détection de la diversité en termes d'isoformes en passant par la mise en évidence des sites de polymorphisme simples (SNP)(*Stoneking, 01*) et d'autres résultats intéressants. Par abus de langage, il est courant de faire référence par le terme "assemblage" aux deux processus distincts de clustering et d'assemblage des données.

3.2 Techniques de clustering

Tel que mentionné précédemment, le but du clustering est de rassembler les données d'EST disponibles, de manière à regrouper en ensembles tous les EST provenant d'un même gène. Ce regroupement repose sur le principe que la distance, ou similarité, entre les membres d'un même groupe est minimale tandis que la distance entre 2 groupes différents est maximale. Aucun état intermédiaire ne devrait être toléré.

Plusieurs algorithmes et outils peuvent être utilisés pour calculer le taux de similarité entre deux séquences, notamment, pour n'en citer que les plus connus, Smith-Waterman, BLAST, Fasta. Le coût computationnel de ces outils dépend en grande partie de leur sensibilité à détecter des similarités faibles, ainsi que de la taille du jeu de données à traiter. Dans des projets de moindre envergure (quelques centaines à quelques milliers d'EST), l'étape de classification peut grandement bénéficier de l'utilisation d'outils standard de détection d'homologies. Le vrai défi se manifeste quand le nombre de séquences à traiter est grand, rendant le calcul des distances deux à deux trop long. D'autre part, la sensibilité envers les faibles similarités, telle qu'implémentée par les outils standards, n'est pas essentielle dans la classification. Deux séquences descendent généralement d'un même parent si l'homologie entre les deux est flagrante, malgré la tendance à des inconsistances dues aux artéfacts de séquençage et aux régions de basse qualité. Lors du traitement d'un nombre élevé de séquences, les processus d'extension de gaps et d'évaluation des régions ayant peu d'homologie sont moins importants tandis que la vitesse d'exécution devient alors un critère majeur.

Pour classifier des séquences, deux approches majeures sont utilisées. La première utilise la fréquence de mots entre deux séquences pour inférer une distance, tandis que la

deuxième utilise la notion d'alignement pour détecter la similarité entre deux EST.

La première approche abordée dans cette section sera celle se basant sur la fréquence de mots pour établir la distance entre les deux séquences. La deuxième détaillera la technique de clustering par homologie de séquences. Ces méthodes sont celles les plus implémentées dans les outils de clustering actuels, tels TGICL ou $d^2\_cluster$. Bien qu'à priori ces procédés semblent très similaires, ils divergent cependant dans le choix d'algorithmes utilisés pour évaluer la similarité, ainsi que dans la méthode de jointure de séquences semblables.

3.2.1 Clustering par fréquence de mots

Pour calculer la distance entre deux vecteurs de données numériques, une multitude de formules peuvent être utilisées. On peut spontanément penser à la distance euclidienne, la distance de Manhattan ou autres. Les séquences d'ADN sont des suites de nucléotides ayant des propriétés spécifiques telles que leur longueur, le pourcentage de nucléotides du guanine ou cytosine, communément appelé *taux de GC*, le nombre de chaque nucléotide et autres paramètres. Ces paramètres, utiles dans certains domaines, le sont moins pour estimer la distance entre deux séquences. Pour déterminer si les séquences se ressemblent, on peut rechercher la constitution en termes de facteurs de chaque séquence. L'idée est la suivante : si deux séquences sont assez similaires pour être deux produits d'un même gène ou d'un même transcrit, ces deux séquences devraient conserver assez de facteurs de taille w, en commun. Un exemple simple est présenté au Tableau 3.1 .

On constate ainsi au tableau 3.1 que la similarité entre les séquences A et B est plus élevée que celle entre A et C. Cette similarité est détectée par la présence de sept facteurs communs à A et B, contre seulement trois facteurs partagés entre les séquences A et C.

La fréquence pour chaque mot peut alors être vue comme un paramètre du vecteur passé à la formule Euclidienne ou à la formule de Manhattan. Le nombre total de paramètres du vecteur est alors le nombre de mots possibles de taille w, soit 4^w éléments.

Séquence A ACTGATGATTCGCA

Séquence B ACTCATGATTCGTA

Séquence C ACTCGCTCATCGAA

Facteur de longueur trois	Fréquence dans A	Fréqunce dans B	Fréquence dans C
ACT	1	1	1
CTG	1	0	0
CTC	0	1	2
TGA	2	1	0
TCA	0	1	0
GAT	2	1	0
CAT	0	1	1
ATG	1	1	0
ATT	1	1	0
TTC	1	1	0
TCG	1	1	2
CGC	1	0	1
CGT	0	1	0
GCA	1	0	0
GTA	0	1	0

Tableau 3.1 Le nombre de facteurs de longueur trois en commun entre les séquences A, B et C

Par exemple, la valeur de la distance en terme de fréquence de mots, notée *dfm*, entre A et B, notée $dfm(A,B)$ de l'exemple précédent peut être calculée à partir du tableau 3.1 en utilisant la distance *euclidienne*, pour l'ensemble de mots contenus dans ses séquences, comme suit :

$$dfm(A,B) = \sqrt{(freq_i(A) - freq_i(B))^2} \text{ pour tout mot } i \text{ de taille 3}$$
$$dfm(A,B) = \sqrt{0+1+1+1+1+1+1+0+0+0+0+1+1+1+1} = \sqrt{10}$$

3.2.2 Clustering par homologie de séquences

L'homologie fait référence au taux d'identité entre deux séquences A et B, calculé à l'aide d'un alignement et qui a pour but de décrire la similarité entre deux séquences. Lorsque suffisamment élevée, l'homologie permet d'inférer que deux séquences proviennent d'un même gène ou de rejeter l'hypothèse que que les séquences sont reliées lorsque aucune ressemblance n'existe entre elles.

L'alignement de la Figure 3.3(a) met en évidence l'homologie entre une séquence A et une séquence B. Cet alignement est défini par la région L, chevauchante entre les deux séquences et dont la longueur est de 2 bases. Le taux d'identité H de cette région chevauchante est de 18 bases sur 22, soit 81%. L'alignement (b) de la même figure indique l'inclusion complète de C dans A. Ceci signifie que la taille de l'alignement L est égale à la taille de la séquence C dont la longueur est de 22 bases alors que le taux d'identité H est de 17 bases sur 22, soit 77%.

Pour effectuer le clustering d'un ensemble de séquences, le clustering par homologie de séquences effectue une comparaison 2 à 2 des séquences devant être traitées .Les résultats sont ensuite filtrés pour ne garder que les alignements qui respectent une taille L et une homologie H minimales.

Les alignements sont par la suite triés par ordre décroissant de qualité, en commençant

Figure 3.3 Alignements représentant l'homologie entre séquences

par les plus longs alignements ayant les meilleures homologies. Contrairement à la technique de jointure utilisée avec la distance de fréquence de mots, l'approche se basant sur l'homologie de séquences s'effectue en choisissant les premiers alignements de la liste triée pour initier et diriger le clustering. Lors de l'inspection d'un alignement de la liste, impliquant deux séquences S_i et S_j, trois cas sont possibles :

1. Les deux séquences impliquées dans l'alignement n'appartiennent pas à un cluster, dans quel cas un nouveau cluster C_i, où i = nombre de clusters existant + 1, est créé et les séquences S_i et S_j y sont ajoutées.
2. Seulement la séquence S_i appartient au cluster C_i. Dans ce cas, la séquence S_j est insérée dans le cluster C_i.
3. Les deux séquences S_i et S_j appartiennent à deux clusters distincts, C_i et C_j. Une opération UNION(i, j), qui effectue la jointure des deux clusters, est alors effectuée. Les détails de l'opération de jointure seront discutés ultérieurement. Si S_i et S_j appartiennent toutes les deux à C_i, alors l'alignement est évidemment ignoré.

Clustering par inclusion

Le clustering par inclusion est un cas particulier du clustering par homologie de séquences. Cette technique, utilisée surtout dans les bases de données biologiques telles NCBI ou ENSEMBL, consiste à supprimer les séquences qui sont incluses ou approximativement répétées dans d'autres séquences. Ceci a pour effet de réduire le temps de recherche sur la base de données, ainsi que de limiter la redondance dans les résultats obtenus.

Le clustering par inclusion ressemble au clustering standard mais se distingue par la présence, dans chaque cluster, d'une séquence parent qui couvre, ou contient, d'une manière quasi-complète, toutes les autres séquences du cluster. Utilisée dans le cadre des projets d'EST, cette approche a pour but de faciliter la détection d'artéfact en réduisant la taille des données devant être indexées lorsque des clusters de grande taille sont obtenus. Pour savoir si une séquence A est parent, ou contenant, de la séquence B, deux paramètres doivent être spécifiés, à savoir : le pourcentage d'identité minimal PID et le débordement maximal, $OVHANG$.

Le paramètre de débordement indique la taille maximale permise de la partie non chevauchante (Figure 3.3), tandis que le PID spécifie le pourcentage d'homologie minimal requis entre les séquences A et B. L'algorithme ayant pour but d'effectuer le clustering par inclusion est celui même utilisé dans le clustering par homologie de séquences. Cependant, sachant qu'une séquence B ne peut être contenue dans une autre séquence A si $|B| + OVHANG < |A|$, le nombre de comparaisons peut être réduit seulement aux séquences susceptibles d'être incluses l'une dans l'autre.

3.3 La jointure de clusters

Une autre divergence entre les méthodes de clustering repose sur le choix de la technique de jointure utilisée pour former les index finaux. Le clustering par fréquence de mots est souvent apparié à un algorithme utilisant la jointure minimale, appelée aussi jointure simple. Dans la littérature d'analyse de séquences, ce processus est nommé clustering agglomératif.

La jointure utilise les trois opérations fondamentales suivantes pour générer des clusters de gènes :

1. CRÉER_CLUSTER(A) : une opération qui crée un nouveau cluster contenant une séquence passée en paramètre.
2. UNION_CLUSTER(A,B) : effectue l'union de deux clusters, le premier contenant la séquence A et le deuxième contenant la séquence B.

Algorithme 1 (Algorithme de Clustering utilisé avec la distance de fréquence de mots)

```
1: dfm(A,B)
2: for i allant de 1 à n do
3:    CRÉER_CLUSTER_1(S_i)
4: end for
5: for j allant de 1 à n − 1 do
6:    for i allant de j + 1 à n do
7:       if TROUVER_CLUSTER_1 (S_i) != TROUVER_CLUSTER_1 (S_j) then
8:          if dfm(S_i, S_j) <seuil then
9:             UNION_CLUSTER(S_i, S_j)
10:         end if
11:      end if
12:   end for
13: end for
```

3. TROUVER_CLUSTER(A) : retourne l'identificateur, qu'on appellera étiquette, du cluster auquel la séquence A appartient.

L'Algorithme 1 retrace les étapes parcourues pour effectuer la jointure en clusters.

- Toutes les séquences sont placées dans des clusters indépendants.
- On traite en premier la séquence S_1.
- Pour toutes les $n-1$ séquences appartenant aux $n-1$ clusters restants, on effectue une opération de jointure entre *T_cluster[1]* et *T_cluster[i]* si $dfm(S_1, S_i) < seuil$.
- À la deuxième itération de j, les $n-2$ séquences restantes sont parcourues et on joint le cluster auquel S_i appartient avec le cluster contenant S_1 si $dfm(S_1, S_i) < seuil$.
- L'étape précédente est répétée pour les $n-3$ séquences restantes en utilisant la séquence S_2. On ne peut évidemment pas effectuer une opération de jointure sur deux séquences appartenant déjà au même cluster.
- Les étapes précédentes sont répétées pour toutes les séquences S_j avec j allant de 3 à n.

Note : Puisque le calcul de distance est effectué après vérification de l'appartenance des

deux séquences examinées, ceci a pour conséquence d'éviter le calcul de la distance dfm pour toutes les séquences S_i appartenant déjà à un même cluster. Pour le meilleur cas, toutes les séquences sont incluses dans le cluster C_1 et aucun autre calcul n'est ensuite effectué. Au pire cas, n clusters singletons sont générés et n^2 opérations sont effectuées.

Étant donné que le clustering par fréquence de mots utilise la fermeture transitive dans la génération de clusters, deux séquences A et B peuvent se retrouver dans un cluster même si $dfm(A, B) > seuil$. Ceci se produit lorsqu'il existe une séquence C tel que :

$$dfm(A, C) < seuil \text{ et } dfm(B, C) < seuil.$$

La fermeture transitive peut ainsi causer des artéfacts d'assemblage et forcer des transcrits de gènes distincts à se retrouver dans le même index.

3.4 Structures de données pour le calcul de la distance en utilisant la fréquence de mots

Une approche efficace pour le calcul de la distance de fréquence de mots consiste en l'utilisation d'une structure de données, telle un arbre de mots clés, modifiée de manière à permettre la mémorisation de fréquences au fur et à mesure de la construction de l'arbre. On verra que la construction d'une telle structure de données pour deux séquences A et B, sera réduite à une complexité $O(mw + nw)$ où m et n sont les tailles des séquences A et B respectivement et w est la longueur des facteurs.

Un arbre de mots clés est une structure proposée par Aho et Corasik (*Aho et Corasick, 75*), permettant de résoudre plusieurs problèmes de recherche de motifs de manière efficace et élégante. Dans son application standard, un arbre de mots clés représente tous les motifs susceptibles d'être cherchés dans un texte. La Figure 3.4 représente l'arbre des mots clés pour l'ensemble de motifs $\{ACGC, ACT, GAGAC, TTG, CA\}$.

De manière formelle, un arbre de mots clés pour l'ensemble de k motifs $p_1, p_2, \ldots, p_k$

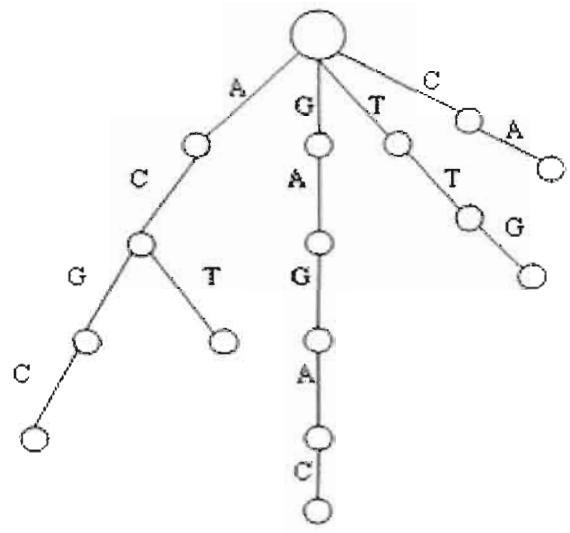


Figure 3.4 Exemple d'arbre de mots clés pour l'ensemble $\{ACGC,\ ACT,\ GAGAC,\ TTG,\ CA\}$

est un arbre enraciné satisfaisant les conditions suivantes. (On suppose par soucis de simplicité qu'aucun des motifs n'est le préfixe d'un autre motif).

- Chaque branche de l'arbre est étiquetée par une lettre de l'alphabet.
- Deux branches provenant d'un même nœud possèdent des étiquettes distinctes.
- Chaque motif $p_i (1 \leq i \leq k)$ parmi l'ensemble des motifs est épelé sur un chemin de la branche allant de la racine à une feuille.

Puisque tous les motifs sont épelés dans l'arbre, le nombre de branches est au pire cas égal à la somme de toutes les lettres composant les motifs. La construction de l'arbre possède donc une complexité $O(N)$, où N est la taille du mot résultant de la concaténation de tous les motifs.

Dans le cas du calcul de fréquence de mots, les motifs recherchés sont les facteurs de taille w existant dans une des deux séquences traitées. Pour optimiser la performance de notre calcul, la fréquence de chaque mot existant dans l'une ou l'autre des séquences sera sauvegardée au fur et à mesure de la construction de l'arbre. Pour ce faire, les feuilles sont utilisées pour garder la fréquence du mot épelé par les branches remontant jusqu'à

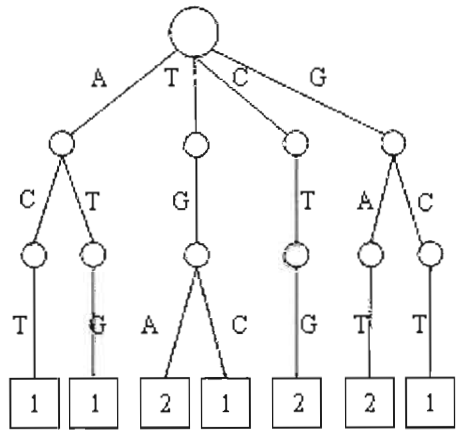


Figure 3.5 Calcul des fréquences dans un arbre de mots clés M_c pour la séquence A

la racine. L'exemple de la Figure 3.5 représente l'arbre de mots clés, M_c, pour l'ensemble de tous les facteurs de taille $w = 3$ de la séquence A ="ACTGATGCTGAT". L'ensemble des facteurs de longueur trois contenus dans cette séquence sont $\{ACT,\ CTG,\ TGA,\ GAT,\ ATG,\ TGC,\ GCT,\ CTG,\ TGA,\ GAT\}$.

Les fréquences des facteurs de A sont stockées dans les feuilles de l'arbre M_c. À partir de la Figure 3.5 on remarque par exemple que les facteurs $TGA,\ CTG,\ GAT$ ont une fréquence de deux et que tous les autres facteurs ont une fréquence de un.

De la même manière et en utilisant la même structure de données, les facteurs de taille w appartenant à la séquence B sont insérés dans le même arbre M_c construit pour la séquence A. La structure des feuilles est modifiée pour stocker les fréquences relatives à la séquence B. Ainsi, en rajoutant la séquence B = "GATGATGAG", où l'ensemble de facteurs de taille w est $\{GAT,\ ATG,\ TGA,\ GAT,\ ATG,\ TGA,\ GAG\}$, on obtient l'arbre de la Figure 3.6. Au niveau des feuilles, la cellule de gauche donne les résultats relatifs à la séquence A, tandis que la cellule de droite représente les résultats obtenus pour la séquence B. Pour permettre l'accès direct aux feuilles de l'arbre sans effectuer un parcours de ce dernier, une file de feuilles est maintenue au fur et à mesure que ces dernières sont créées dans l'arbre M_c. Ceci permet de parcourir les résultats pour

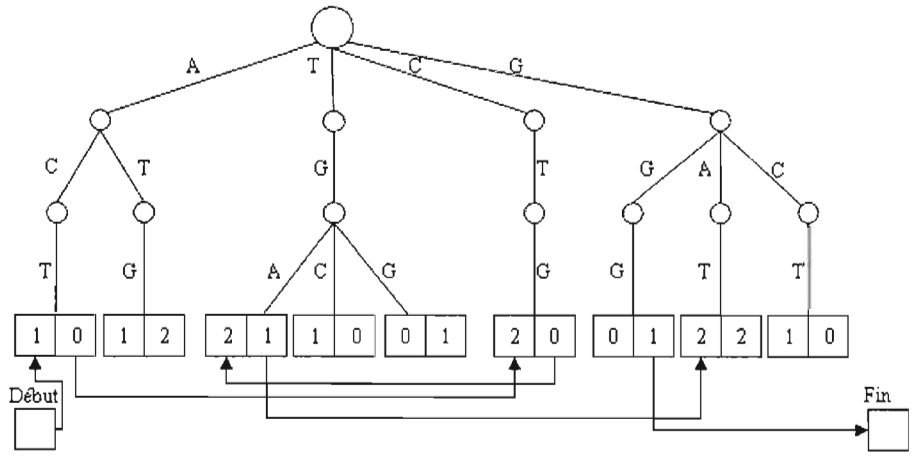


Figure 3.6 Calcul des fréquences dans un arbre de mots clés M_c pour les séquences A et B

calculer la distance de fréquence de mots en un temps $O(m+n)$ au pire cas. Pour alléger la compréhension de la Figure 3.6, seulement quelques références de la file ont été représentées.

Dans le but de calculer la complexité de cette structure de données dans le calcul de la distance de fréquence mots, supposons :
N_1, le nombre de lettres composant les motifs de A,
N_2, la nombre de lettres composant les motifs de B.

On a :
$N_1 \leq (m-w+1)w$
$N_2 \leq (n-w+1)w$

La construction de l'arbre M_c possède au pire cas une complexité $O(N_1+N_2) = O(mw+nw)$.

La somme de toutes les fréquences peut être effectuée en temps $O(m+n)$.
Notons que le choix d'un arbre de suffixes, modifié pour inclure les fréquences de ses

Séquence A ACTGATAGCTGATATCGATAGT

Séquence B ACTGGTAGCTATATCGATTGT

Facteurs de longueur 2	Fréquence dans A	Fréquence dans B	Facteurs de longueur 2	Fréquence dans A	Fréquence dans B
GA	3	1	GT	3	3
AC	1	1	GC	1	1
AG	2	0	TC	1	1
AT	2	3	GT	1	2
CG	1	1	TG	2	1
CT	2	2	TG	0	1

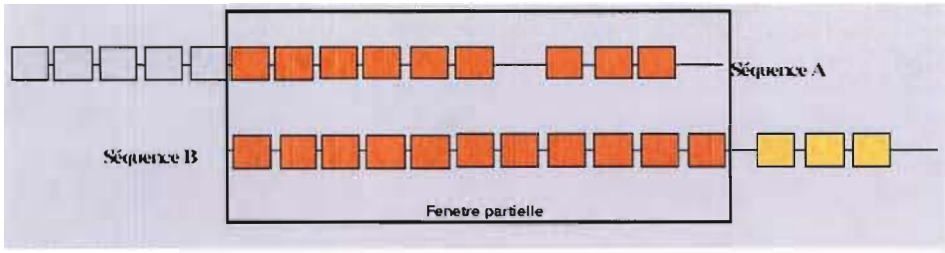
Tableau 3.2 Calcul de la distance dfm pour les séquences A et B avec un facteur $w = 2$

suffixes, implique une complexité quadratique (*Nuallain etl., 04*). Cette implémentation utilise une version modifiée de l'algorithme de Ukkonen [Ukkonen, 95] pour garder dans chaque nœud, ainsi que dans les feuilles, la fréquence de chaque étiquette. La complexité du calcul de la distance de fréquence de mots dans une telle structure de données serait alors $O(m^2 + n^2)$.

De par la nature fragmentaire des EST, l'homologie entre deux séquences A et B peut ne pas être globale. Le chevauchement de deux séquences avec un degré de similarité élevée peut être un indice suffisant sur l'appartenance des deux EST au même transcrit. Les régions non chevauchantes peuvent être dues à une différence de taille entre les deux EST et non pas nécessairement à une divergence entre les séquences. Dans ce cas là, l'inclusion des ces régions non chevauchantes biaiserait le calcul de la distance.

La Figure 3.7 présente un exemple de calcul de la distance dfm sur toute la longueur de séquences chevauchantes et le calcul effectué sur la région du chevauchement. À la Figure 3.7(a) le calcul de la distance est effectué sur la partie chevauchante. Les mots communs partagés par les deux séquences, représentés par les carrés de même couleur, indiquent une forte homologie entre A et B sur la partie chevauchante. Le nombre de

a)



b)

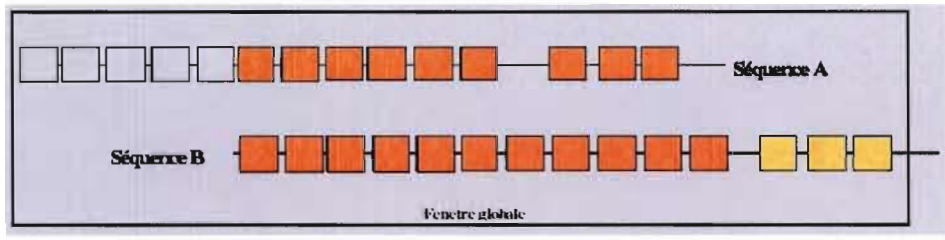


Figure 3.7 Exemple de similarité sur a) la partie chevauchante ainsi que sur b) la globalité des deux séquences A et B

carrés de couleur différente entre les séquences représente les mots différents entre A et B. Selon la figue 3.7.b, aucun mot commun n'existe sur les parties non chevauchantes. Ceci a pour effet de réduire l'homologie, lorsque celle-ci est calculée sur la globalité de la séquence.

L'utilisation de la distance de fréquence de mots pour évaluer la similitude entre deux séquences, transcrites à partir d'un même gène, mais ayant des extrémités non chevauchantes est inefficace. Pour améliorer l'efficacité du calcul de la distance, l'utilisation d'une fenêtre glissante permettant de limiter la région où le calcul est significatif, peut être introduite. La distance de fréquence de mots devient alors le score *dfm* minimal entre toutes les fenêtres de taille k entre deux séquences.

1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	1	1	1	6	6	6	9	9	9	9	9	9

Figure 3.8 Représentation sous forme de tableau de trois clusters d'un ensemble de 14 séquences

3.5 Structures de données utilisées dans la jointure de clusters

Les structures d'ensembles disjoints se prêtent bien aux opérations fondamentales utilisées dans l'algorithme de jointure (Algorithme 1) et représentent une manière efficace de l'implémenter.

Supposons que chaque séquence est étiquetée par sa position dans la liste des séquences. La première séquence aura ainsi l'étiquette 1, la deuxième 2,...et la dernière possèdera l'étiquette n, où n est le nombre total de séquences. On suppose aussi qu'un cluster est identifié par la plus petite étiquette qu'il contient.

Une manière simpliste de représenter la structure de données utilisée par le clustering par fréquence de mots serait d'utiliser un tableau *T_clusters* de taille n, où les indices du tableau représentent les étiquettes des séquences et les cases du tableau contiennent l'étiquette de l'ensemble auquel les séquences appartiennent.

Le tableau de la Figure 3.8 représente un ensemble de 14 séquences, étiquetées de 1 à 14 et regroupées en quatre clusters dont les étiquettes sont les séquences 1, 6, 9.
Note : Le choix d'un tableau est justifié par le nombre, préalablement connu, de séquences à traiter.

Pour faciliter la lecture des algorithmes, la notation suivante sera utilisée.
- S_i représente la séquence ayant l'étiquette i.
- *seuil* est la distance dfm maximale permettant d'inférer si deux séquences sont similaires.

Algorithme 2 (Créer un cluster à partir d'une Séquence A)

1: CRÉER_CLUSTER_1(A)
2: {Initialiser la case représentant la séquence A avec la valeur de son étiquette}
3: i = étiquette de A
4: *T_clusters[i]* = i

Algorithme 3 (Trouver le cluster contenant la Séquence A)

1: TROUVER_CLUSTER_1(A)
2: {Trouver l'étiquette du cluster contant la séquence A}
3: i = étiquette de A
4: retourner *T_clusters[i]*

Ainsi, les opérations CRÉER_CLUSTER_1, UNION_CLUSTER_1, TROUVER_CLUSTER_1 (1) peuvent être définies par les algorithmes 2, 3 et 4 respectivement.

Les opérations CRÉER_CLUSTER_1 et TROUVER_CLUSTER_1 sont des opérations simples qui n'exigent que la consultation d'un seul élément du tableau. Ces opérations peuvent être exécutées en un temps constant. L'opération d'union prend, quant à elle, un temps de l'ordre de n.

En utilisant deux heuristiques et une structure de données alternative, l'opération UNION_CLUSTER_1 (*Aho et Corasick, 75*) peut être implémentée de manière plus efficace. Cette implémentation se base sur les forêts d'arbres enracinés pour représenter les clusters et ne suppose plus qu'un cluster est représenté par la séquence ayant la plus petite étiquette. De manière plus formelle, les forêts possèdent les caractéristiques suivantes :

- Chaque cluster est représenté par un arbre enraciné,
- L'étiquette d'un ensemble est l'élément contenu à la racine,
- Chaque nœud contient un élément de l'ensemble et une référence à son père.
- La racine possède une référence vers elle-même,
- Un tableau de références vers chaque élément est gardé pour ainsi faciliter l'accès aux séquences.

Algorithme 4 (Union de deux clusters contenant les séquences A et B respectivement)

```
1: UNION_CLUSTER_1(A,B)
2: {Modifier les éléments du cluster ayant la plus petite étiquette}
3: i = étiquette de A
4: j = étiquette de B
5: if i <j then
6:    interchanger i et j
7: end if
8: for k allant de 1 à n faire do
9:    if T_clusters[i] = j then
10:      T_clusters[j] = i
11:   end if
12: end for
```

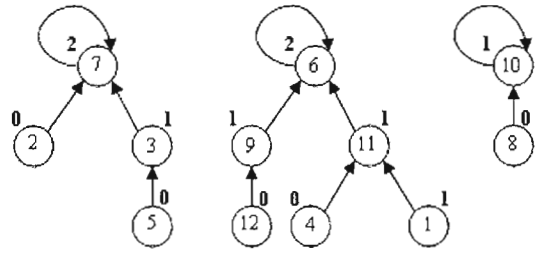


Figure 3.9 Représentation sous forme de forêts de trois clusters

La Figure 3.9 montre les forêts obtenues pour trois clusters ayant comme racines les séquences dont les étiquettes sont 7, 6 et 10.

En utilisant cette structure de données, l'implémentation des opérations de création, de recherche et d'union diffèrent du cas précèdent. Les opérations sont redéfinies comme suit :

1. CRÉER_CLUSTER_2(A) : requiert la création d'un arbre à un seul nœud qui fait référence à lui-même.
2. TROUVER_CLUSTER_2(A) : retourne la racine de l'arbre dans lequel la séquence A est contenue.

3. UNION_CLUSTER_2(A, B) : Crée une référence entre l'arbre contenant une des séquences passées en paramètre et ayant la plus petite racine vers la racine parent de l'autre séquence.

Selon la définition de l'opération UNION_CLUSTER_2, si toutes les séquences appartiennent à un seul cluster, ce dernier sera alors représenté par un arbre dégénéré formé par une liste de séquences. Ceci a pour conséquence d'alourdir l'exécution des opérations TROUVER_CLUSTER_2 et UNION_CLUSTER_2 qui requièrent de remonter jusqu'à la racine. Pour pallier à ce problème, deux heuristiques peuvent être utilisées pour améliorer la performance des opérations de recherche et d'union.

1. L'union par rang : elle a pour but de réduire la hauteur, ou rang, de l'arbre. L'union par rang fait pointer, lors de l'union de deux arbres, la racine de l'arbre de moindre rang vers la racine de l'arbre ayant le rang le plus élevé. Lorsque deux arbres possèdent le même rang, le choix de la racine pointée est alors aléatoire et le rang de l'arbre est ensuite incrémenté.
2. La compression de chemin : l'opération TROUVER_CLUSTER_2 requiert le parcours de tous les nœuds internes en partant de la feuille jusqu'à la racine. Pendant le parcours des nœuds internes, l'heuristique de compression de chemin requiert que chaque nœud parcouru par la recherche fasse directement référence à la racine. Ceci a pour conséquence de réduire la profondeur de chaque branche sur laquelle une opération de recherche est effectuée.

La notation suivante sera utilisée dans les algorithmes des opérations sur les clusters.
- La référence d'un nœud x est notée $p[x]$,
- Le rang d'un nœud x est noté $rang[x]$,
- Le tableau T contient des références vers toutes les séquences S à traiter.

Voici les algorithmes de chaque opération.

La fonction TROUVER_CLUSTER_2 parcourt tous les nœuds intermédiaires, de manière récursive, entre l'élément passé en paramètre et la racine. Tous les éléments consultés durant le processus sont modifiés pour faire référence directement à la racine. La procédure LIER change la référence de la racine de l'arbre ayant le plus petit rang pour pointer

Algorithme 5 (Créer le cluster contenant la Séquence A)

1: CRÉER_CLUSTER_2(A)
2: {Créer un racin faisant référence à elle-même}
3: Créer nœud étiqueté par A
4: T[A] = A
5: P[X] = A
6: Rang[X] = 0

Algorithme 6 (Trouver le cluster contenant la Séquence A)

1: TROUVER_CLUSTER(T[A])
2: {Remonter de manière récursive jusqu'à la racine}
3: A = T[A]
4: If A != p[A]
5: p[A] = TROUVER_CLUSTER_2(p[A])
6: rang(A) = 0
7: Retourner p[A]

vers la racine de l'arbre ayant un rang élevé. Si les deux rangs sont similaires, alors une racine est choisie comme parent et son rang est incrémenté.

Selon un théorème de R. Tarjan (*Tarjan, 75*), une séquence de x opérations CRÉER_CLUSTER_2, TROUVER_CLUSTER_2 et UNION_CLUSTER_2, dont y opérations sont des opérations CRÉER_CLUSTER_2 peut être exécutée sur une forêt d'ensemble disjoints grâce à l'union par rang et à la compression de chemins en un temps $O(x.\alpha(x,y))$ dans le pire cas. $\alpha(x,y)$ fait, dans ce cas, référence à l'inverse de la fonction d'Ackerman où :

$$\alpha(x,y) = min\{i >= 1 : A(i, \left\lfloor \frac{x}{y} \right\rfloor) > log(n)\}$$

On estime $\alpha(x,y) \leq 4$ pour tous les cas pratiques. Par exemple pour tout $y \leq x \leq 10^{80}$

Un exemple mettant en œuvre les opérations d'ensembles disjoints en utilisant les forêts est donné à la Figure 3.10. À l'état initial *a)*, les trois clusters de la Figure 3.10 sont représentés. Le rang est indiqué en gras pour chaque élément de l'arbre. Le tableau de référence est omis dans le but de réduire la complexité de la figure.

Algorithme 7 (Union de deux clusters contenant les séquences A et B respectivement)

```
1: UNION_CLUSTER_2(A,B)
2: {Appelle la fonction LIER sur la racine de l'arbre contenant A
3: et la racine de l'arbre contenant B}
4: LIER(TROUVER_CLUSTER_2(T[A]), TROUVER_CLUSTER_2(T[B]))
```

Algorithme 8 (Lier les clusters auxquels les séquences A et B appartiennent)

```
1: LIER(A,B)
2: if rang[a] >rang[b] then
3:    p[B] = A
4: else
5:    p[A] = B
6: end if
7: if rang[A] >rang[B] then
8:    rang[B] = rang[B] + 1
9: end if
```

L'opération *b)* effectue l'union du cluster contenant la séquence S_4 dont la racine a un rang de un et le cluster contenant la séquence S_8 dont la racine possède un rang deux. La racine ayant le moindre rang est ajoutée comme enfant de la racine ayant le rang deux.

Durant l'opération TROUVER_CLUSTER_2 à l'étape *c)*, la référence des éléments inspectés est modifiée pour pointer sur la racine et leur rang est incrémenté. La dernière opération fusionne deux clusters ayant tous deux un rang de 2. La racine est choisie comme parent et son rang est incrémenté de un.

Cette structure d'ensembles disjoints peut être utilisée avec l'algorithme global de jointure, proposant ainsi une implémentation plus efficace que la recherche exhaustive ou les ensembles disjoints utilisant les tableaux.

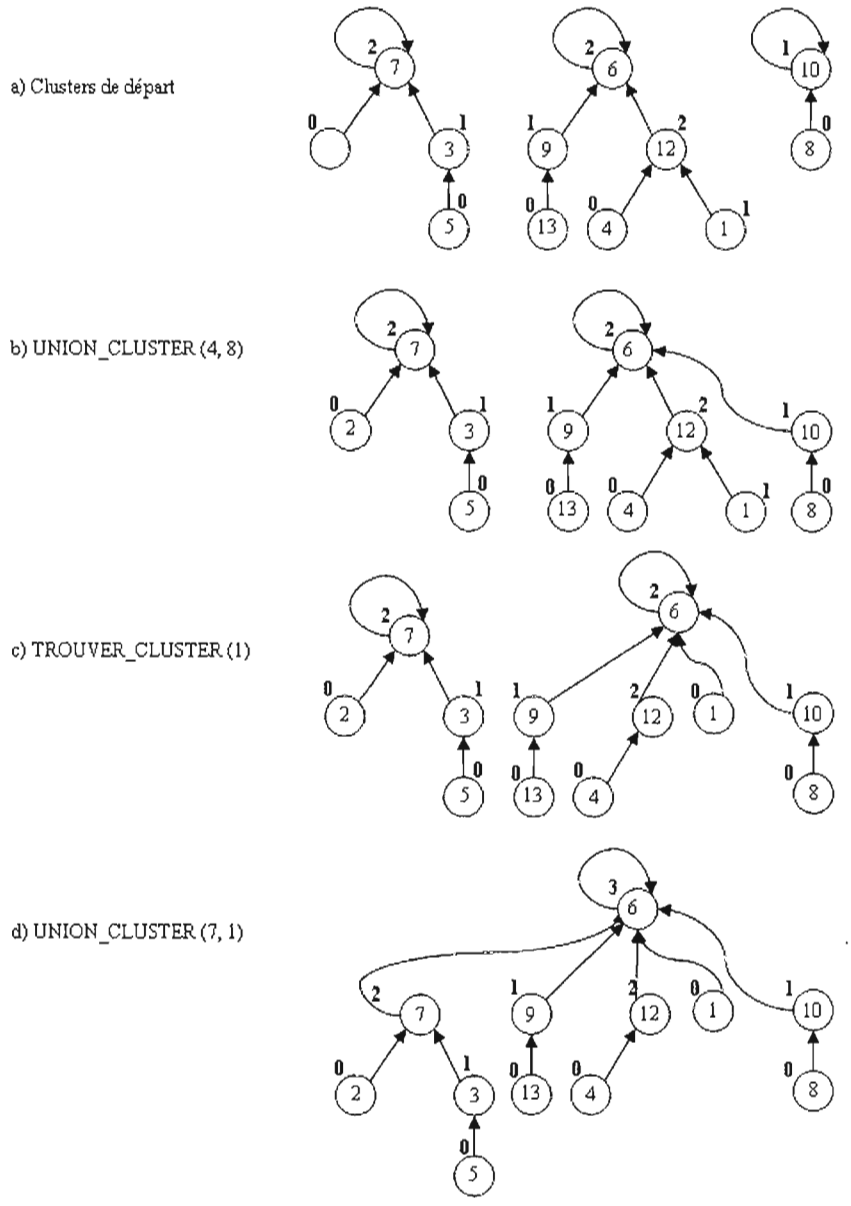


Figure 3.10 Exemple d'opérations sur une structure d'ensemble disjoint implémentée par des forêts

CHAPITRE IV

LE PROJET FGAS

Le présent chapitre est divisé en quatre sections. La première présente les résultats de clustering obtenus dans le cadre du projet FGAS en utilisant les méthodes de clustering définies au chapitre précédent. Dans la deuxième partie, une méthode de résolution des artéfacts de gros clusters est proposée et les résultats de cette dernière sont analysés. L'assemblage de chacun des clusters résultant est ensuite effectué à l'aide des deux méthodes d'assemblage abordées précédemment et les détails en sont donnés dans la troisième partie du chapitre. Pour conclure, les processus d'analyse et d'extraction d'information des contigs et singletons produits par l'assemblage sont abordés et les apports des résultats bioinformatiques au projet FGAS sont discutés.

4.1 Clustering des séquences

Dans le but d'effectuer le clustering des séquences, les deux logiciels, $d^2\_cluster$ (*Burke et al., 99*) et TGICL (*Pertea et al., 03*) ont été employés. Le choix de ces deux outils s'est fait à la lumière de la crédibilité dont ils disposent, puisque ces derniers sont les plus fréquemment utilisés dans les cadres de projets d'assemblage. Le logiciel $d^2\_cluster$ emploie la distance calculée sur la fréquence de mots pour effectuer de manière rapide le clustering d'un ensemble de séquences. Le logiciel TGICL s'appuie en revanche sur une version modifiée de BLAST pour détecter l'homologie entre les séquences devant être traitées. Les résultats sont par la suite utilisés dans le cadre d'un algorithme agglomératif pour effectuer la jointure de clusters.

Bien que plusieurs publications (*Hide et al., 94, Liang et al., 00*) se soient attardées sur la validation des résultats obtenus par d^2_*Cluster* et TGICL, la complexité de la tâche de clustering et le manque de moyens pour confirmer la diversité des résultats restent un obstacle important dans l'évaluation de l'efficacité de chacun de ces outils. Le choix d'utiliser l'une ou l'autre de ces approches reste ambigu et semble être instinctif plus que fondé sur un critère quelconque. Pour cette raison, nous avons décidé dans un premier temps d'effectuer le clusering des séquences en utilisant ces deux outils et de comparer, en un second temps, la taille des clusters obtenus par les deux assemblages avec les tailles de clusters dans Unigene. L'ensemble Unigene désigne les résultats publiés par le NCBI et obtenus en utilisant un protocole spécifique dans l'assemblage de tous les 499,104 EST publiques de blé. Les clusters Unigene sont en constante curation et de nouvelles versions sont publiées sur une base régulière. La version 44 a été utilisée dans la présente comparaison.

La comparaison des résultats obtenus avec des données d'Unigene aura pour but d'appuyer le choix de l'une ou l'autre des méthodes ainsi que de détecter les artéfacts d'assemblage.

Le groupement des séquences en clusters fût d'abord effectué par d^2_*cluster*. Les paramètres ont été modifiés pour que deux séquences soient considérées similaires et ainsi mises dans un même cluster si elles contiennent, au moins une fenêtre de 100 *pb* avec une distance dfm maximale de $\Omega = 0,1$. Pour traiter les 271 226 EST disponibles, 31 heures d'exécution furent nécessaires sur une machine possédant 2 processeurs Pentium 4 2.8 *GHZ* avec 2 *Gig* de mémoire vive. Les données ont aussi été assemblées en utilisant le logiciel TGICL avec les paramètres par défaut. L'exécution du programme a pris 28 minutes en utilisant la même machine utilisée dans de l'assemblage précédent.

4.1.1 Comparaison des résultats de clustering

Le tableau 4.1 représente les fréquences de tailles de clusters, en termes du nombre de séquences contenues, pour chacune des trois méthodes d'assemblage.

Taille du cluster en séquences	Unigene 499,104	d^2_cluster 271,226	TGICL 271,226
4097+	0	1	1
2049-4096	9	1	0
1025-2048	17	3	3
513-1024	47	9	11
257-512	168	38	38
129-256	322	113	133
65-128	682	243	282
33-64	1394	620	667
17-32	2360	1208	280
9-16	3480	2005	2144
5-8	5713	3351	3590
3-4	9758	4906	5194
2	3584	7746	7886
1	7729	33 822	33 635
Nombre total de clusters	35 263	54 066	55 864

Tableau 4.1 Nombre de clusters, de différentes tailles, propres aux méthodes UNIGENE, TGICL et d^2_*Cluster*

Le nombre total de clusters est sensiblement similaire pour TGICL et $d^2\_cluster$ mais il est beaucoup plus élevé que celui obtenu en utilisant l'approche Unigene. Selon Burke et al. (*Burke et al., 99*), l'approche utilisée dans Unigene fait appel à d'autres critères, tels l'annotation des séquences, l'inclusion de séquences complètes d'ARNm, ainsi que l'utilisation d'information 3' et 5' pour faire le clustering. De plus, la curation continuelle des clusters Unigene a pour effet de réduire significativement le nombre d'erreur et d'artéfact de clustering dans chaque nouvelle version. Ceci pourrait se révéler une raison suffisante quand à la disparité de nos résultats avec ceux du NCBI. La première version d'Unigene aurait pu valider une telle hypothèse, cependant cette version n'a pas pu être obtenue.

Deux différences majeures, en termes de résultats entre l'approche du NCBI et les deux autres approches, ont pu être constatées :

1. Le nombre de clusters singletons ou duplets est clairement plus élevé chez TGICL et $d^2\_cluster$.
2. Le nombre de clusters dont la taille est supérieure à deux est plus élevé chez Unigene.

Plusieurs raisons pourraient expliquer la présence d'un nombre si élevé de singletons selon la méthode utilisée. Ces différences sont selon Burke et al. (*Burke et al., 99*) :

- L'échec de la méthode à trouver de l'homologie utilisant ses propres critères de similarité.
- L'introduction de fausses jointures par une séquence (chimères, multidomaines, etc.) dans une méthode et non pas dans l'autre.
- L'utilisation de critères différents pour décider de l'appartenance d'une séquence à un cluster. Par exemple, le fait qu'Unigene utilise les données d'annotation tandis que $d^2\_cluster$ se base seulement sur l'homologie inter-séquences.

On pourrait aussi penser à d'autres raisons triviales, telles que :

- Le nettoyage stringent a supprimé plusieurs séquences pouvant se regrouper avec les singletons ou les clusters de taille deux.
- Le nombre de singletons est effectivement réel et représente le nombre de transcrits

rares dans les conditions de stress abiotique ayant permis de générer les librairies initiales.

- Les séquences d'EST seraient bruitées et auraient ainsi généré un nombre élevé de faux singletons.

4.1.2 Conséquences de la qualité des données et de la sensitivité entre d^2_*cluster* et TGICL sur l'assemblage

Selon Liang et al. (*Liang et al., 00*), le taux de singletons est un bon indicateur sur le nombre de transcrits rares ayant été exprimés. D'après la même source, les logiciels de clustering sont grandement affectés par les erreurs de séquençage et les données de mauvaise qualité. Pour vérifier si la nature des données corrobore, dans notre cas, cette hypothèse, une inspection de 100 séquences, aléatoirement choisies dans des clusters singletons ou duplets, a été effectuée.

Parmi ces séquences, 37% représentaient des artéfacts biologiques putatifs, telles la présence de peu de CG, la présence de longs patrons répétés d'une manière inexacte, etc. Ce pourcentage élevé semble alors soutenir l'hypothèse corrélant la basse qualité des séquences au nombre élevé de singletons et de clusters formés de duplets.

Pour ce qui est du nombre de clusters de taille supérieure à deux. Il a été démontré par Burke et al. (*Burke et al., 99*) que l'algorithme utilisé par d^2_*cluster* engendre un nombre de clusters estimé à 13% moins élevé que celui d'Unigene, principalement dû à la différence en termes de spécificité des algorithmes utilisés par les deux méthodes. Cette différence en termes de spécificité n'est cependant pas un indicateur de la qualité des clusters générés car ce que d^2_*cluster* considère comme étant des isoformes et met dans un même cluster, pourraient être en réalité des paralogues et devraient, comme effectué par Unigene, être séparés. L'inverse pourrait bien entendu aussi être vrai.

En comparant les résultats du tableau 4.1, on constate que le nombre de clusters obtenus par d^2_*cluster* est plus élevé que celui obtenu par TGICL. Une inspection plus approfondie permet de constater que plusieurs des clusters de TGICL ont été séparés

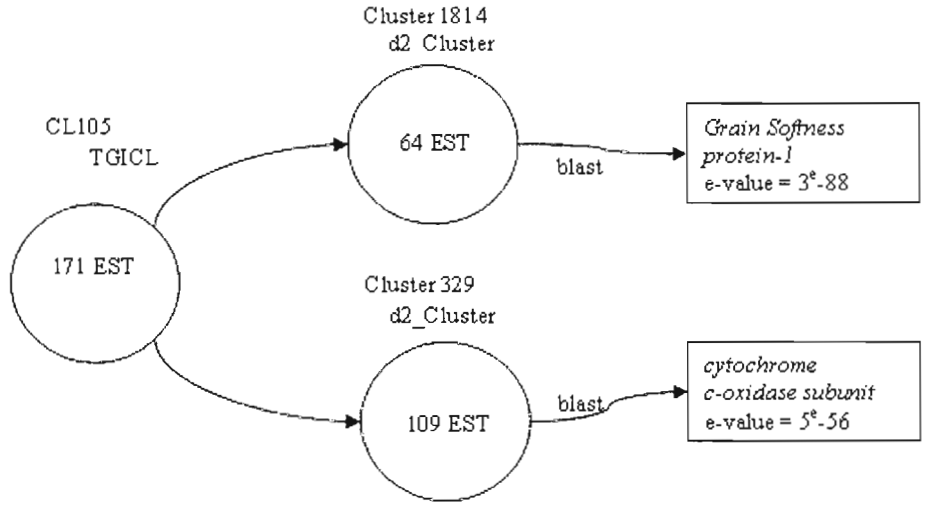


Figure 4.1 Fusion de deux clusters distincts par TGICL

en deux (ou plus) clusters distincts par $d^2\_cluster$. L'exemple de la figure 4.1 illustre ce cas de fusion.

Le CL105[1] de TGICL contient 171 séquences. Ces mêmes séquences ont été séparées en deux clusters disjoints par $d^2\_cluster$, soient *Cluster_329* de 109 séquences et le *Cluster_1814* de 64 séquences.

Dans le but d'évaluer la dissemblance détectée par $d^2\_cluster$, un représentant de chacun des clusters obtenus a été comparé à la base de données de l'ensemble de séquences de protéines publiques, *nr*. La séquence BE422565 provenant du *Cluster_329* démontre une homologie, soutenue par une e-value de $3e^{-88}$, avec la protéine *"grain softness protein-1"*, tandis que la séquence BE425307 du Cluster 1814 possède une similarité dont la e-value est de $5e^{-56}$ avec la protéine *"cytochromec − oxidasesubunit"*.

Dans un deuxième temps, la comparaison de la séquence de la protéine *"grain softness"*,

[1] $d^2\_cluster$ utilise la nomenclature *Cluster_XX* tandis que TGICL utilise la nomenclature *CLXX*, où *XX* fait référence au numéro de cluster.

a été réalisée en utilisant l'algorithme de BLAST, avec tous les EST membres du *Cluster_329* ce qui a permis de constater, en analysant les alignements, que les séquences de ce cluster provenaient toutes du même gène. Utilisant la même méthode, les transcrits du cluster 1814 ont tous été déterminés comme provenant du gène *"cytochrome c-oxidase subunit"*.

L'identification de ces deux clusters comme représentant deux gènes indépendants, illustre l'efficacité supérieure de l'algorithme utilisé dans d^2*_cluster* au détriment d'être beaucoup plus lent que TGICL.

4.1.3 Artéfact du gros cluster

La taille du plus gros cluster produit par Unigene est de 3664, alors que TGICL et d^2*_cluster* ont produit deux clusters dont les tailles sont 39 247 et 25 311 respectivement. Il est évident que ces deux derniers clusters sont les produits d'un artéfact d'assemblage puisque, à moins d'une erreur de manipulation majeure, il est biologiquement impossible d'avoir un pourcentage aussi élevé de transcrits d'un même gène.

Pour analyser, en terme de contenu, les deux gros clusters (*CL1* du TGCIL et *Cluster_18* de d^2*_cluster*) une comparaison des séquences contenues dans chacun des clusters a été effectuée. La figure 4.2 représente les résultats obtenus.

Selon la figure 4.2, 14 966 des séquences du cluster *CL1* ont pu être classifiées par d^2*_cluster* tandis que 1030 séquences du *Cluster_18* ont été placées dans des clusters par TGICL. Il y a cependant 24 281 séquences présentes à la fois dans *Cluster_18* et *CL1*.

La différence entre les tailles des plus gros clusters obtenus par chacune des méthodes peut être due à la sensibilité des deux logiciels et à la tendance de TGICL de fusionner des clusters distincts, comme cela a été observé précédemment. Pour investiguer cette hypothèse, des séquences ont été choisies de manière aléatoire, parmi 14 966 séquences mises dans le *CL1*. Bien que choisie aléatoirement, une de ces séquences, représente une

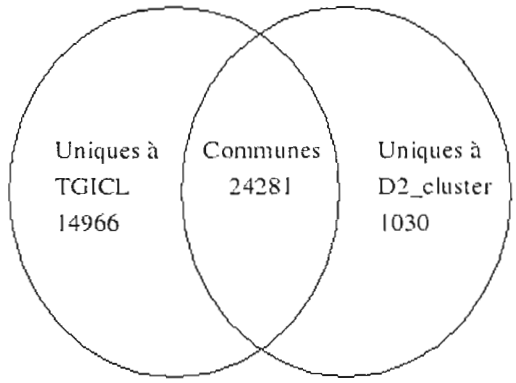


Figure 4.2 Distribution des séquences appartenant aux gros clusters générés par TGICL et par d^2*_Cluster*

protéine très importante de la tolérance au froid *(gi—1657854 Triticum aestivum cold acclimation protein WCOR413 (Wcor413)).*

Cette séquence a été placée par d^2*_Cluster* dans *Cluster_3527* contenant un total de 184 séquences. La comparaison de la séquence de la protéine *Wcor413* contre tous les membres du *Cluster_3527* a permis d'observer que 183, parmi les 184 séquences appartenant à ce cluster, possèdent une forte homologie avec le gène *Wcor*. Seule une séquence ne possédait pas une homologie suffisante pour conclure qu'elle constituait un transcrit de ce gène. TGICL a pu placer seulement 60 des 184 séquences dans un cluster, les autres ont été mises dans CL1.

Certaines séquences de l'échantillon n'ont obtenu aucun hit significatif avec les protéines de la base de données *"nr"* en utilisant l'outil BLASTX. D'autres séquences ont obtenu des hits diversifiés, avec des protéines hautement divergentes, confirmant la différence entre les séquences appartenant à ce cluster.

La même démarche a été effectuée pour les séquences uniques au gros cluster de d^2*_cluster* et celles communes aux gros clusters des deux méthodes. L'analyse des résultats de

BLAST n'a permis de trouver aucune cohérence entre les séquences de chaque échantillon. La conclusion logique est donc que les gros clusters représentent un artéfact de clustering amplifié entre autres par la tendance à fusionner des clusters distincts dans le cas du logiciel TGICL.

La suite des analyses se base uniquement sur les résultats obtenus par la méthode $d^2\_cluster$, étant donnée la meilleure qualité de ces résultats.

4.2 Résolution des artéfacts d'assemblage

Il est possible de représenter les liens de similarité entre les différentes séquences d'un même cluster par un graphe valué. Les sommets du graphe représentent les EST et une arête de poids w relie deux sommets si $w \leq min\_val$. min_val représente la similitude minimale requise entre deux EST provenant d'un même gène.

Le poids w entre deux séquences S_i et S_j, est calculé par la formule suivante :

$$w = \frac{100}{-log(e\text{-}value)}$$

Où *e-value* est le score obtenu par la comparaison de S_i et S_j en utilisant l'outil BLAST et $e\text{-}value \neq 0$.

La figure 4.3 illustre un graphe G, où les sommets en bleu représentent les EST d'un même cluster et les arêtes représentent la similarité entre les sommets qu'elles relient. À défaut d'indiquer le poids w, ce dernier est représenté par la longueur de l'arête. Le graphe est dit connexe, du point de vue topologique, s'il existe un chemin entre chaque paire de sommets (*Brassard et Bratley, 96*).

Dans le but de réduire la taille du *Cluster_18* en regroupant seulement les séquences similaires entre elles, une approche semi-automatique a été mise en place. Cette approche se base sur la recherche de *points d'articulations*, pouvant servir de liens entre différents sous-clusters dans le graphe d'adjacence des séquences. Un point d'articulation est un noeud du graphe G dont la suppression déconnecte le graphe. Il existe trois types de

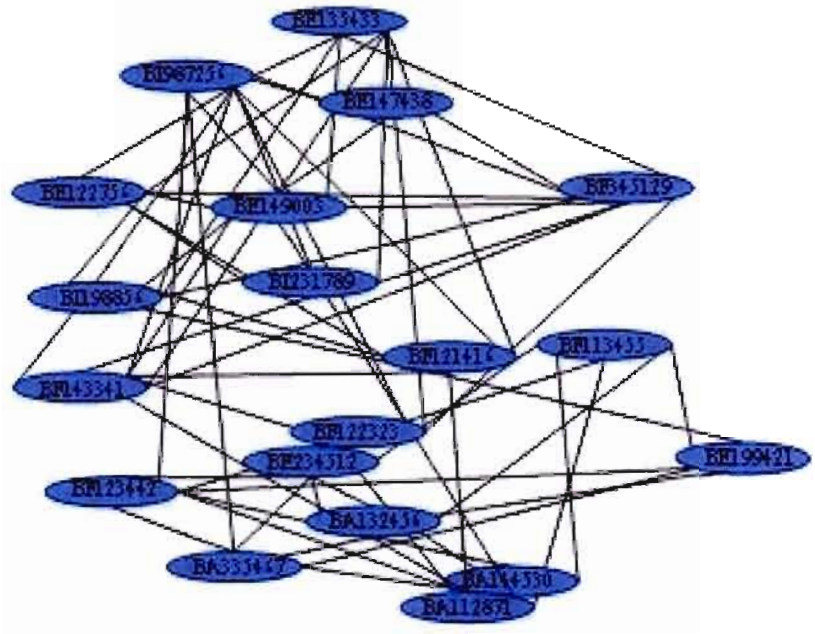


Figure 4.3 Exemple de représentation en forme d'un graphe de proximité d'un cluster

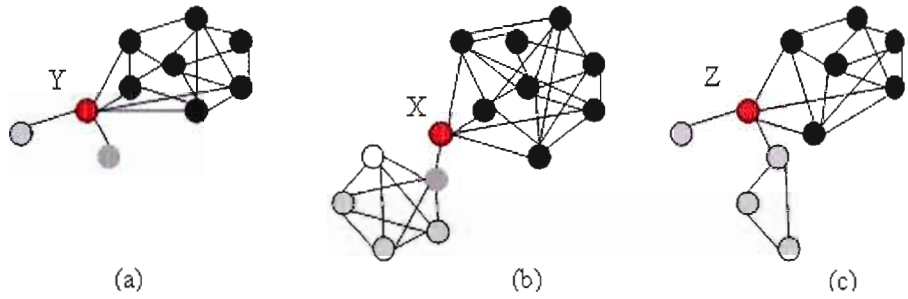


Figure 4.4 Types de points d'articulations dans un graphe G

ponts, appelés aussi points d'articulation, en théorie des graphes.

1. Un pont de type X, représenté par le point rouge du graphe *(b)* à la figure 4.4, relie au moins deux sous-graphes connectés ensemble. La suppression du point d'articulation X génère deux graphes connectés.
2. Un pont de type Y relie un graphe connecté avec un ou plusieurs sommets singletons. Ce type de pont est représenté par le nœud rouge du graphe *(a)* de la Figure 4.4. Sa suppression génère un graphe connexe et des points isolés, en gris.
3. Un point d'articulation de type Z est un sommet du graphe dont la suppression génère au moins deux graphes connexes et au moins un sommet isolé. Ce type de sommets est représenté par le point rouge du graphe *(c)* à la figure 4.4.

De point de vue pratique, un point d'articulation de type X ou Z représente une séquence similaire à deux sous-groupes distincts. Une telle séquence est identifiée dans la plupart des cas, comme étant une chimère ayant causée la jointure de deux clusters différents. Ainsi, la détection et ensuite la suppression de telles séquences ont pour effet de dissoudre ces fausses jointures et réduire la taille éventuelle des clusters (Figure 4.5).

La matrice de distance des éléments appartenant au gros cluster, a été obtenue par BLAST des séquences contre elles même. Les paramètres de coût pour ouvrir et étendre

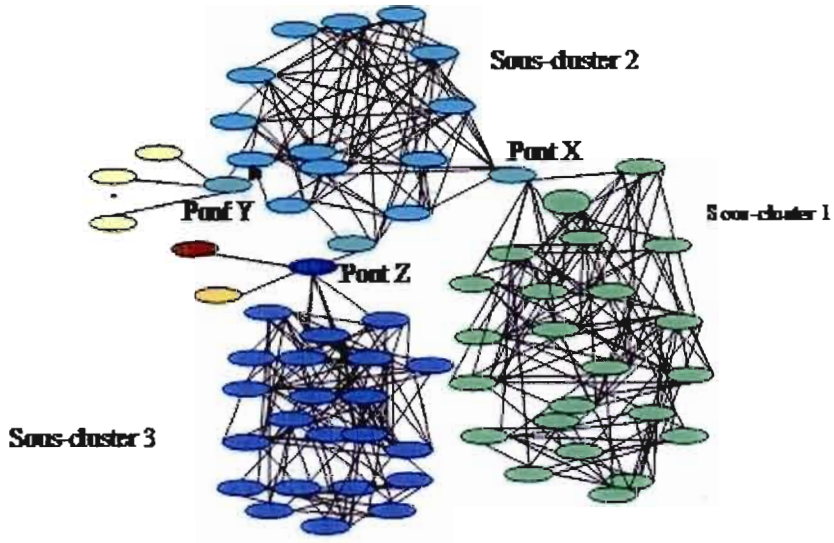


Figure 4.5 Rôle des points d'articulation dans la détection des sous-clusters distincts dans un graphe d'adjacence

un gap, ainsi que la taille d'un mot ont étés choisis comme étant 5, 2 et 7 respectivement. Le filtre de basse complexité a été désactivé dans le but de calculer la distance réelle entre les séquences. La matrice de distance est ensuite passée comme paramètre au programme GRAPH9, qui utilise les informations de distance pour construire le graphe d'adjacence et détecter les points d'articulation de types X, et Z.

Chaque couleur dans le graphe de distance de la figure 4.5 représente des transcrits d'un gène différent. Les trois sous-clusters ainsi que les singletons en jaune sont regroupés par des points d'articulation qui agissent comme chimère et forcent le regroupement de séquences indépendantes ensemble. La suppression de ces points d'articulation aurait pour effet de briser le gros cluster en trois sous-clusters distincts.

L'algorithme utilisé dans GRAPH9 (atgc) n'est pas publié et la complexité de ce logiciel n'est pas décrite. Il est cependant mentionné dans la documentation que l'exécution de ce programme sur des données test d'EST provenant d'*Arabidopsis* a pris plus d'une semaine pour s'achever. le nombre de séquences n'est pas mentionné.

Taille du cluster	Fréquence
257-512	1
129-256	1
65-128	4
33-64	30
17-32	103
9-16	268
5-8	476
3-4	717
2	971
1	0
Total	2517

Tableau 4.2 Tailles moyennes des clusters obtenus avec le clustering par inclusion du *Cluster_18*

L'exécution du programme GRAPH9 avec la matrice de distances des 25 311 séquences appartenant au *Cluster_18* produisait plusieurs erreurs qui interrompaient le calcul. La nature de ces problèmes est inconnue mais l'exécution avec succès du programme sur plusieurs échantillons tests de moindre taille laisse croire que cet échec est dû à la taille du jeu de données.

Pour remédier à ce problème, le clustering par inclusion des données a permis de réduire les 25 311 séquences du *Cluster_18* en 2571 séquences parents en utilisant un pourcentage d'identité PID et une valeur d'OVERHANG de 98% et 20 respectivement. Les tailles moyennes des clusters obtenus sont décrites dans le tableau 4.2.

Le plus gros cluster obtenu est constitué de 297 séquences. La taille de la séquence parent dans ce cluster est de 623 nucléotides et la taille moyenne des séquences qu'elle couvre est de 290 nucléotides.

Taille du sous cluster en EST	Fréquence
257-512	0
129-256	1
65-128	3
33-64	12
17-32	13
9-16	26
5-8	39
3-4	56
2	93
1	380
Total	623

Tableau 4.3 Tailles moyennes des sous-clusters, des séquences parents, obtenus après suppression des points d'articulations X et Z

La comparaison deux à deux des séquences parents a été effectuée en utilisant les mêmes paramètres que préalablement. L'inspection des résultats a permis de produire la matrice de distances passée en entrée à GRAPH9.

Seulement quelques minutes furent nécessaires à l'analyse des données et à l'extraction de points d'articulations potentiels. En total, 104 ponts de type X, et 27 de type Z sont mis en évidence. La suppression de tous les points X et Z a permis de générer un total de 624 sous-graphes et de découvrir des clusters indépendants ayant la distribution de taille présentée au tableau 4.3. Les séquences enfants, ayant été retirées précédemment dans le but d'alléger le processus de détection des points d'articulations, ont été réinsérées dans les clusters contenant leurs parents.

Dans le but de valider les résultats obtenus, une inspection visuelle accompagnée de comparaison de séquences aléatoires d'un échantillon de 50 clusters a été faite. Les

résultats suivants ont été obtenus.

1. Au-delà de 65% des clusters inspectés constituaient des groupes homogènes d'EST provenant d'un même gène parent.
2. Dans 15% des cas, il était impossible de juger l'homogénéité du cluster faute de hits concluants avec la base de données *"nr"*.
3. Les autres clusters inspectés, représentant approximativement 20% de l'échantillon, forment les sous-clusters les plus peuplés et sont constitués majoritairement d'EST provenant de plusieurs gènes abondants dans la cellule, tels Rubisco, histones, protéines ribosomales etc.

Dans le cas du troisième groupe, aucune segmentation manuelle n'a été effectuée étant donnée la forte homologie entre les séquences et le risque de séparer des séquences appartenant à un même cluster (figure 4.6).

Les deux sous-groupes, bleu et vert, de la figure 4.6 représentent deux protéines ribosomales hautement similaires, tandis que le sous-groupe gris représente la protéine Rubisco. Les sous-groupes sont connectés par plus qu'un lien et leur fragmentation manuelle pourrait causer la séparation de séquences appartenant à un même gène parent.

L'utilisation de la technique de détection et, ensuite de suppression de points d'articulation dans un gros cluster s'est avérée d'une grande utilité dans la fragmentation automatique du gros cluster produit par biais d'assemblage. De plus, l'inspection d'un échantillon des clusters obtenus a permis de confirmer une grande cohésion dans les clusters obtenus.

4.3 Assemblage des séquences

Les clusters générés par $d^2\_cluster$ ont été assemblés en utilisant PHRAP et CAP3. Les paramètres de CAP3 utilisés sont ceux qui semblaient être acceptable dans l'assemblage des EST de l'orge (*Close et al., 04*), tandis que les paramètres par défaut ont été utilisés pour PHRAP. (Tableau 4.4).

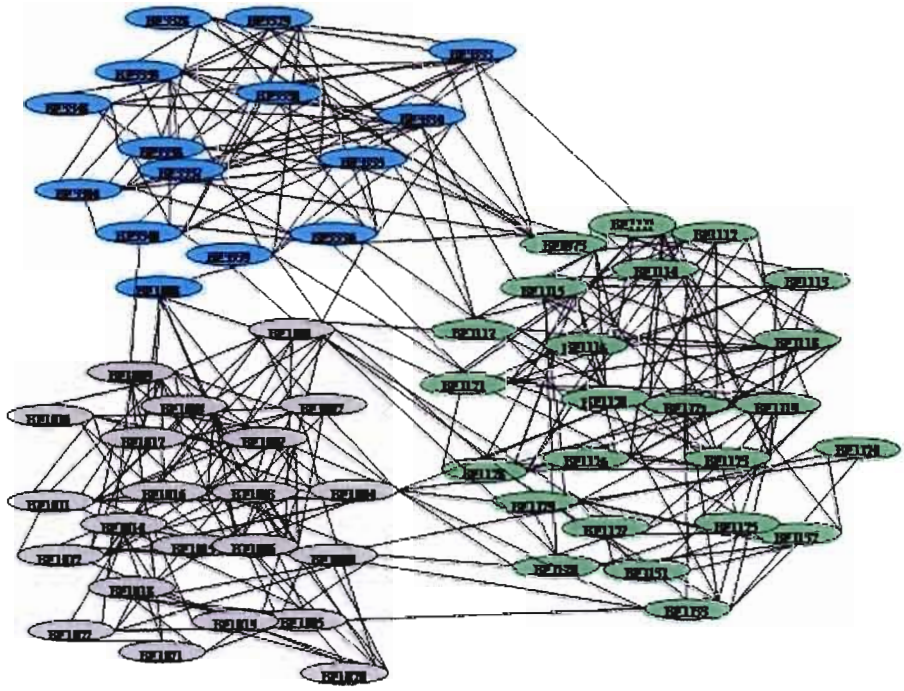


Figure 4.6 Représentation en graphe d'un sous-cluster possédant des séquences abondantes et similaires

Méthode	Contigs	Singlets	% total de singletons + Singlets
Cap3	32,000	7,000	15,5%
PHRAP	50,000	6,000	14%

Tableau 4.4 Résultats de l'assemblage des séquences avec Cap3 et PHRAP

Dans le Tableau 4.4 on constate que le nombre de *singlets*, ou séquences n'appartenant a aucun contig, semble peu varier selon la méthode utilisée tandis que le nombre de contigs obtenus par CAP3, environ 32,000, est significativement inférieur aux 50,000 contigs obtenus à l'aide de PHRAP.

On a procédé à l'évaluation de la présence d'erreurs corrélées et leur impact sur la fidélité des consensus obtenus. En d'autres termes, cette méthode évalue la qualité de l'assemblage vers le début ou la fin des séquences. Ces régions se distinguent, comme il a été mentionné précédemment par la présence d'une taux élevé d'erreurs de séquençage.

4.3.1 Effets des erreurs sur les assemblages

Pour tester l'effet des erreurs de séquençage et des nucléotides de basse qualité sur l'assemblage, les séquences réelles obtenues par le séquençage complet d'ARNm de 10 gènes et obtenues à partir de bases de données publiques ont été comparées avec les séquences consensus dérivées des deux assemblages.

Les divergences entre les séquences et les consensus ont été calculées sur les régions alignées par BLAST pour les contigs de chacune des deux méthodes. Seulement trois parmi les 10 ARNm de départ n'ont pas eu de hits significatifs en utilisant les contigs provenant de l'assemblage avec Cap3 tandis que quatre des dix séquences ne possèdent aucun hit significatif en utilisant les contigs de PHRAP. De plus, les contigs obtenus par Cap3, contiennent peu de divergences sur la longueur des régions alignées, par comparaison aux contigs résultant de l'assemblage par PHRAP. La longueur moyenne des contigs de Cap3 est aussi supérieure a celle des contigs de PHRAP (Tableau 4.5 et Tableau 4.6).

Selon la comparaison des résultats de Cap3 et PHRAP avec les données d'assemblage publiés sur le site du NCBI et en se basant sur le test effectué précédemment, on peut conjecturer que Cap3 produit des résultats plus fiables que ceux obtenus par PHRAP. De plus, il a été démontré par Liang et al. (*Liang et al., 00*), à travers de nombreux tests basés sur des données réelles, ainsi que sur des séquences générées automatiquement,

Gene	Hit	Taille hit	Identités	Gaps
MM2	CL2968Contig1	1181	765/771	0
MC14F1L7-9	CL2849Contig1	841	823/834	0
mc10f1l6-19	CL3571Contig1	1104	935/957	14/957
45G05	CL4758Contig1	776	742/744	0
mc58r1l1-122	CL3064Contig1	1281	990/1022	6/1022
mc18r1l1-143	CL2865Contig2	767	611/625	0
M36	Cl2968Contig1	774	672/694	2/694

Tableau 4.5 Comparaison des contigs de Cap3 avec la séquence complète de 10 ARNm

Gene	Hit	Taille hit	Identités	Gaps
MM2	Contig_5081	776	768/773	0
MC14F1L7-9	Contig_20085	841	821/034	0
mc10f1l6-19	Contig_39127	918	883/897	0
45G05	Contig_48121	722	705/719	0
mc58r1l1-122	Contig_40424	1059	995/1028	6/1028
mc18r1l1-143	Contig_40005	644	629/639	0

Tableau 4.6 Comparaison des contigs de PHRAP avec la séquence complète de 10 ARNm

que Cap3 obtenait, de manière consistante, les assemblages ayant le plus haut degrś de fidélité. Pour ces raisons, les résultats de Cap3 ont étés choisis dans les analyses subséquentes.

4.4 Traduction des EST et annotation des données

L'annotation des séquences est une étape primordiale dans l'étude du modèle selon lequel les gènes sont exprimés. Elle consiste à déterminer avec un haut degré de certitude la fonction putative de la séquence, permettant de comparer des gènes provenant de conditions différentes.

Pour annoter des séquences, deux approches sont possibles. La première se base sur l'apprentissage machine pour pouvoir détecter les patterns, appelés aussi domaines, spécifiques à une fonction particulière, tandis que la deuxième se base sur l'homologie avec des séquences dont la fonction a déjà été établie. L'approche se basant sur l'apprentissage est généralement plus efficace pour détecter les domaines ayant préalablement été répertoriés tandis que la deuxième technique obtient de meilleurs résultats lorsque l'homologie est globale. Bien que les domaines soient de bons indicateurs sur le type d'implication de la séquence dans le métabolisme, le fait qu'une protéine puisse contenir plusieurs domaines augmente la probabilité d'erreur de l'annotation. Une similarité élevée avec une protéine déjà classifiée reste cependant une meilleure approche dans la prédiction de la fonction putative de la séquence chez les EST.

Dans le cas du projet FGAS, l'annotation des contigs et singletons a été effectuée en utilisant l'homologie avec les séquences possédant déjà une annotation. Une base de données contenant toutes les protéines déjà annotées a été téléchargée du site de l'Institut Européen de Bioinformatique, EBI (*ebi*). Les séquences provenant du projet FGAS ont ensuite été comparées à l'aide de l'outil BLASTX avec toutes les protéines préalablement annotées. L'inspection des résultats de BLASTX a ensuite permis d'octroyer à certaines séquences de FGAS l'annotation du polypeptide avec lequel elles partagent une homologie relativement conservative (Une e-value d'au moins $1e^{-25}$).

1-Biological process GO :0008150	2-Transcription GO :0006350
3-Protein metabolism GO :0019538	4-Enzyme regulator activity GO :0030234
5-Nutrient reservoir activity GO :0045735	6-Transcription factor activity GO :0003700
7-Nuclease activity GO :0004518	8-Plasma membrane GO :0005886
9-Secondary metabolism GO :0019748	10-ResponsetoexternalstimulusGO :0009605
11- Carbohydrate binding GO :0030246	12- Response to abiotic stimulus GO :0009628
13- Cell-cell signalling GO :0007267	14- Development GO :0007275
15- Behavior GO :0007610	

Tableau 4.7 Catégories go_slim choisies pour étudier la variation dans les taux d'expression.

Une séquence peut avoir plusieurs annotations si elle possède une homologie dont la e-value est supérieure à $1e^{-25}$ sur la base de données des protéines du EBI. Ceci est biologiquement cohérent car une protéine peut jouer plusieurs rôles et participer à différentes réactions. Cette multifonctionnalité protéique est souvent conférée par la présence de plusieurs domaines lui permettant d'avoir plusieurs types d'interactions.

En utilisant le protocole décrit ci-haut, un total de 29 558 séquences ont pu être annotées. À titre indicatif, 18 310, ou 58% de l'ensemble total, se trouvaient dans des contigs. Tandis que seulement 11 248, ou 30,8% de l'ensemble initial, sont des singletons ou duplets. Dans le but de normaliser la nomenclature établie dans l'annotation des séquences, l'ontologie GO (*web :go*) (Gene Ontology) a été utilisée. Le choix de Gene Ontology s'est fait sur la base de l'acceptation globale, de la part de la communauté. En effet, GO représente l'annotation utilisée dans tous les projets publics d'envergure. Dans le but de faciliter les processus d'analyse des pattrons d'expressions, les annotations ont été regroupées selon des catégories susceptibles de fluctuer dans le cadre du projet FGAS.

L'outil *slim_map* (web :go) distribué avec l'ontologie, ainsi que des scripts Perl développés localement ont été utilisés pour regrouper les annotations dans les catégories choisies par

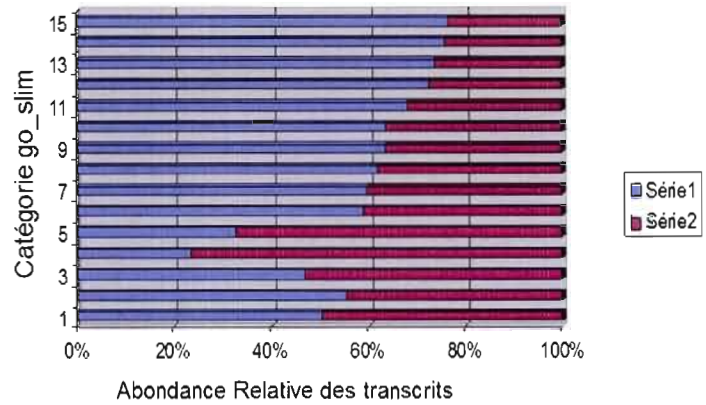


Figure 4.7 Abondance en termes de transcrits annotés de FGAS dans des catégories GO prédéfinies

les biologistes (Tableau 4.7). Le nombre de séquences a été évalué pour chaque groupe et normalisé en calculant le pourcentage pour le nombre total de séquences (Figure 4.7). La figure 4.7 compare l'abondance relative de chaque catégorie GO choisie entre les transcrits de FGAS en bleu et ceux du contrôle en rouge. Le total dans chaque catégorie a été ajusté à 100%.

Le taux d'annotation obtenu avec les données de FGAS est cinq fois supérieur à celui obtenu et publié dans les projets d'assemblage du blé (Version 10,0) par l'organisme TIGR (Institute for Genomic Research). Suite à l'assemblage effectué par TIGR, seulement 6 431 des 44 954, soit 14% de l'ensemble, des contigs du blé ont pu être annotés en utilisant Gene Ontology. La bonne qualité des séquences obtenus suite aux assemblages a permis d'améliorer substantiellement les résultats existant. Ceci représente une contribution significative qui élargit l'ensemble d'annotations spécifiques au blé disponibles dans le cadre des études fonctionnelles.

Selon la figure 4.7 aucune différence majeure entre les catégories 1, 2 et 3 n'est observable. Cependant pour la plupart des autres catégories, biologiquement plus spécifiques au phénomène d'acclimataton, une augmentation du niveau d'expression est identifiée, sauf pour les catégories 4 et 5 ou l'on remarque une baisse remarquable du niveau d'expression. De manière plus générale, on peut conclure que plusieurs gènes ayant différentes fonctions sont sous-représentés ou sur-représentés dans les données de FGAS par rapport à l'échantillon de contrôle composé de (NSF + DuPont).

4.5 Création de la base de données

Dans le but de faciliter l'accès aux données, ainsi que pour consolider et préserver l'intégrité de l'information, une base données mySQL a été développée. La structure des tables et les informations contenues dans la base de données ont été modélisées selon les standards utilisés dans la librairie bioSQL. Aussi, nous nous sommes inspirés de certains logiciels d'annotation (*autofact*) et d'assemblage génomique (*amos, Batzoglou et al., 02*) en adaptant certaines des tables qu'ils utilisent au domaine de l'assemblage d'EST.

La base de données créée peut être séparée en deux parties majeures.

1. La partie un contient l'information pertinente aux séquences, à leur clustering et à l'assemblage et est représentée par les entités en rouge dans le schéma relationnel de la base de données (Figure 4.8).
2. La partie deux représentée par les entités en bleu dans le diagramme(Figure 4.8) et contient l'information relative à l'annotation des données.

L'accès à la base de données est, au moment de la rédaction de ce document, réservé aux membres du projet FGAS et se fait à travers une connexion SSH. La base de données sera accessible au public sous peu à l'aide d'interfaces Web qui permettront d'exécuter divers types de requêtes sur les données.

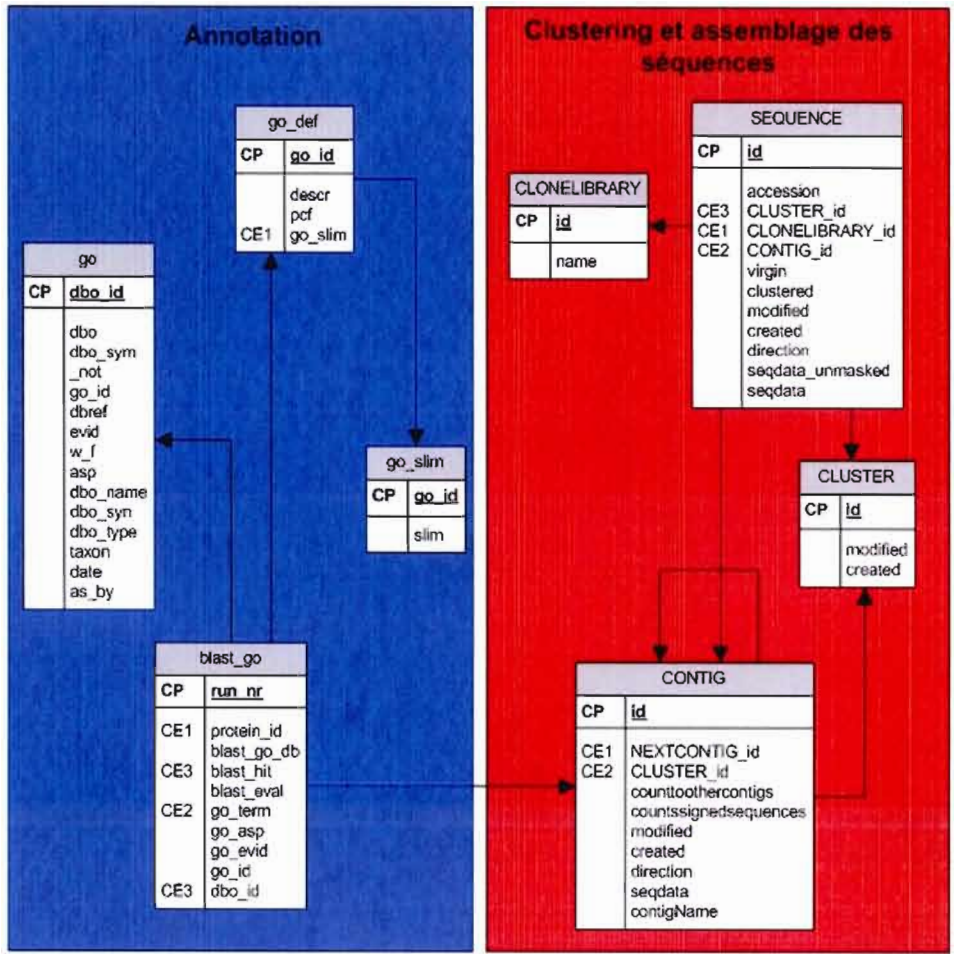


Figure 4.8 Schéma relationnel de la base de données

Les tables relatives à la partie un, en rouge, sont :

Table SEQUENCE	
Contient les informations relatives aux séquences	
Champ	Description
Accession	Numéro d'accession servant comment identificateur unique de la séquence
clean	Champ booléen indiquant si la séquence a été nettoyée ou non
created, modified	Champs indiquant la date de création et celle de modification
seqdata_unmasked	Contien la séquence originale non nettoyée et celle après nettoyage
seqdata	

Table CLONELIBRARY	
Contient l'information relative au clone dont la séquence a été générée	
Champ	Description
id	Numéro de séquence unique du clone
nom	Champ unique de la table contenant le nom du clone. Il est à noter que le nom encode une panoplie d'information sur le clone, telles la direction, la taille ainsi que d'autres informations pertinentes. Nous avons jugé non nécessaire de rajouter des champs additionnels pour ces informations puisqu'elles peuvent être extraites facilement à partir du nom de la séquence

Table CLUSTER	
Contient l'information relative aux clusters générés	
Champ	Description
id	Identificateur unique du cluster
created, modifed	Dates de création et de modification, si il y a lieu, du cluster

Table CONTIG	
Contient l'information relative aux contigs générés	
Champ	Description
id	Identifiant unique du contig
NEXTCONTIG_id	Identifiant du contig suivant. Ce champ est nécessaire puisque les contigs ne se suivent pas de manière séquentielle et cette information permet de connaître le prochain contig
CLUSTER_id	Identifiant unique du cluster auquel le contig appartient
counttoothercontigs	Le nombre d'autres contigs appartenant au même cluster que le contig en question
contigname	Nom composé, encodant l'information sur le cluster auquel le contig appartient ainsi que le numéro de séquence du contig dans le cluster
created, modifed	Dates de création et de modification, si il y a lieu, du contig
seqdata	Séquence nucléique du contig

Les tables relatives à la partie deux, en bleu, sont :

Table go_slim	
Contient les informations sur les sous-catégories générales d'annotation, choisies par les biologistes	
Champ	Description
go_id	Identificateur unique de la sous catégorie
slim	Définition de la catégorie globale

Table go_def	
Contient les définitions des termes de gene ontology utilisés	
Champ	Description
go_id	Code de l'ontologie
descr	Description de l'ontologie
pcf	Code pour définir si l'ontologie décrit un processus biologique, une composant cellulaire, ou une fonction moléculaire
go_slim	Identifiant unique de la sous-catégorie à laquelle appartient l'ontologie

Table blast_go	
Contient les résultats de comparaison par blast des contigs contre la base de données de protéines annotées	
Champ	Description
run_nr	Identifiant unique du hit pour la séquence. Une séquence peut avoir plusieurs hits identifiant les protéines similaires avec une e-value inférieure à $1e^{-25}$
protein_id	Numéro d'accession de la séquence devant être annotée
blast_go_db	La version de la base de données ayant été utilisée dans la comparaison
blast_hit	Identifiant de la séquence similaire au contig
blast_eval	e-value du hit
go_term	L'identifiant de l'ontologie ayant préalablement été attribuée au blast_hit
go_asp	Code pour définir si l'ontologie décrit un processus biologique, une composant cellulaire, ou une fonction moléculaire
go_evid	Code utilisé pour indiquer la manière dont l'annotation a été attribuée
go_id	GO de haut niveau, indiquant la catégorie choisie par les biologistes dans laquelle l'ontologie se situe

Table go	
Cette table provient du logiciel ayant été utilisé pour effectuer l'annotation et contient l'information spécifique à la base de données ayant servi dans l'annotation	
Champ	Description
dbo_id	Identifiant unique de la protéine déjà annotée
db	Nom de la base de données dont la séquence annotée provient
dbo_sym	Symbole de la protéine dans la base de données dont elle provient
go_id	L'identifiant de l'ontologie de la protéine
dbref	Référence de la preuve ayant servi à obtenir l'annotation
evid	Code utilisé pour indiquer la manière dont l'annotation a été attribuée
w_f	Terme non utilisé
asp	Code pour définir si l'ontologie décrit un processus biologique, une composante cellulaire, ou une fonction moléculaire
dbo_name	Nom de la protéine
dbo_syn	Terme non utilisé
dbo_type	Nature de la séquence (protéine, ARNt, etc.)
taxon	Terme non utilisé
date	Date d'ajout de la séquence à la BD
as_by	Organisme ayant effectué l'annotation

4.6 Extraction de données de sur-expression et sous-expression

Dans le but d'avoir une meilleure estimation des gènes régulés à la hausse ou à la baisse, les données de FGAS et NSF-DuPont ont été analysées par DAA (Digital Display Analysis). Pour chaque contig, le nombre d'EST provenant de FGAS, a été divisé par le nombre d'EST de NSF-DuPont. Le rapport a été normalisé de manière à prendre en considération la différence de taille des deux ensembles, soit 54 032 séquences pour FGAS et 196 041 séquences pour NSF-Dupont. Les contigs constitués seulement de séquences FGAS ont été considérés comme étant des séquences uniques à la condition d'acclimatation et le nombre de séquences constituantes a été considéré.

Environ 75% des contigs ont des ratios qui varient par moins de deux fois par rapport aux données de contrôle. Le reste des contigs, représentant 7841 EST, ont un taux de variation supérieur à deux. En choisissant des ratios d'expression plus élevés, 6,6% des séquences varient entre FGAS et les données de contrôle par un taux estimé à 5 et 1,7% lorsqu'un seuil de variation minimal de 10 est choisi. La différence d'expression est due principalement à des gènes qui sont sur-exprimés (Avec un seuil minimal de 5, 1959 gènes sont sur-exprimés et 136 gènes sont sous-exprimés). Une inspection plus détaillée des ces gènes sur-exprimés révèle que ces derniers sont, ou possèdent des homologues, démontrés en laboratoire comme étant des gènes régulés à la hausse sous l'effet d'un stress. Par exemple, pour les 15 premiers contigs les plus abondants, 10 publications ont été trouvées. Ceci confirme l'implication de ces gènes dans des procecssus d'acclimatation et suggère des candidats potentiels pour mieux comprendre la réponse du blé au froid. Ces confirmations in vivo constituent un support important à notre analyse et valident les étapes de la méthode utilisée pour l'obtention des résultats.

L'analyse des 90 plus gros contigs, contenant au moins cinq EST provenant de FGAS et ne contenant aucune séquence du jeu NSF-DuPont, a permis de détecter des gènes abondants et connus pour être régulés par le froid.

Le fait qu'ils soient exprimés seulement dans FGAS et non pas dans les expérience du NSF-DuPont permet de conclure que ces gènes là sont spécifiques au cultivar utilisé et à la condition de stress imposée. En se basant sur les résultats obtenus dans des expériences précédentes ainsi que dans les publications disponibles, il a été démontré que le taux de gènes régulés par le froid est estimé à 10% du nombre total de gènes, évalué de manière très conservative à 30,000 gènes, chez le blé acclimaté. Un nombre similaire de gènes régulés au froid est obtenu dans le cadre du projet FGAS lorsqu'on calcule la somme de :

1- le nombre de contigs ayant un ratio d'expression minimal fixé à cinq entre les données de FGAS et celles du NSF-DuPont

2- le nombre de contigs possédant plus trois EST FGAS et aucun EST provenant de NSF-DuPont.

En effectuant la somme des contigs obtenus dans les deux groupes, 1 et 2, on obtient 2,637 contigs ou (8,3%) des 31 772 contigs générés par l'assemblage. Ce chiffre est similaire à ce qui a été avancé dans le littérature (Close et al., 04, umanitoba) et permet ainsi de valider les résultats d'assemblage et confirmer l'efficacité de la méthode d'annotation choisie.

CONCLUSION

Les résultats obtenus répondent de manière claire aux exigences définies dans la problématique du projet. En plus d'avoir augmenté le pourcentage de séquences annotées et fourni une base de données qui sera sous peu publique, nous avons identifié un nombre important de gènes connus pour participer chez d'autres plantes au processus de tolérance au froid. D'autres gènes peu connus et dont la sur-expression ou sous-expression semblait corroborer une implication dans le processus de tolérance ont aussi été identifiés et représentent des candidats idéals aux études de tolérance en laboratoire. De plus, les expériences bioinformatiques effectuées in-silico ont permis de clarifier le phénomène d'acclimatation en un temps beaucoup plus court que dans le cas d'études au laboratoire, et à coût moins élevé.

Lors de la résolution de l'artéfact de larges clusters, nous avons choisi d'effectuer le clustering par contenance pour ne garder que les séquences parents. Ceci a été décidé dans le but de réduire la taille du jeu de données traitées que le programme d'assemblage ne pouvait gérer. Une alternative à cette approche serait de subdiviser l'ensemble total en plusieurs sous-groupes de taille acceptable et faire l'assemblage de chacun de ses sous-groupes. Le clustering aurait pu alors être effectué sur l'ensemble des contigs et singletons résultants et un autre assemblage une seconde fois effectué pour chaque cluster résultant. Bien que cette méthode aurait pu donner des résultats satisfaisants, la complexité de son implémentation et son manque d'élégance, restent un facteur pour lequel elle n'a pas été choisie.

Ce projet nous a posé d'autres problèmes. Lors d'analyses de sur-expression ou sous-expression expression, nous nous sommes contentés d'utiliser de simples normalisations pour arriver à nos résultats. Lorsque les gènes impliqués dans la condition étudiée sont hautement sur-exprimés ou sous-exprimés, ceci permet tout de même d'arriver à de

résultats fort satisfaisant. Cependant, lorsque la fluctuation dans le niveau d'expression est mineur et donc difficilement détectable, cette méthode n'aurait pas permis de trouver les gènes d'intérêt. Pour cette raison, des méthodes statistiques plus robustes, telles le clustering hiérarchique ou les analyses mutivariées seraient nécessaires. Le développement d'outils prenant en considération ces contraintes dans des projets spécifiques aux EST serait une contribution majeure au domaine.

BIBLIOGRAPHIE

Aho AV, Corasick JM, « Efficient string matching : An aid to bibliographic search »*Communication of the Association for Computing Machinery* 18(6), 1975, 333-340.

Altschul FS, Madden LT, Schaffer AA, Zhang J, Zheng Z, « Gapped BLAST and PSI-BLAST : a new generation of protein database search programs »*Nucleic Acids Res.* 25, 1997, 3389-3402.

Batzoglou S, et al., « ARACHNE : a whole-genome shotgun assembler »*Genome Res* 12(1), 2002, 177-189.

Brenner S, Miller J, « Encyclopedia of genetics »*Academic Press* 2001.

Burke J, Davison D, Hide W, « d2_cluster : A Validated Method for Clustering EST and Full-Length cDNA Sequences »*Genome Res.* 9(11), 1999, 1135-1142.

Chou HH, Holmes MH, « DNA sequence quality trimming and vector removal »*Bioinformatics* 7(12), 2001, 1093-104.

Close TJ, et al., « A new resource for cereal genomics : 22K barley GeneChip comes of age »*Plant Physiol.* 134(3), 2004, 960-968.

Ewing B, Hillier L, Wendl MC, Green P, « Base-calling of automated sequencer traces using phred. I. Accuracy assessment »*Genome Res.* 8, 1998, 175-185.

Ewing B, Green P, « Analysis of expressed sequence tags indicates 35,000 human genes »*Nature Genetics* 25, 1998, 232-234.

Fei ZJ, Tang X, Alba RM, White JA, Ronning CM, Martin GB, Tanksley SD, « Comprehensive EST analysis of tomato and comparative genomics of fruit ripening »*Plant J.* 40, 2004, 47-59.

Gallant KJ, Maier D, Storer AJ, « On Finding Minimal Length Superstrings »*J. Computer System Sci.* 20(2), 1980, 50-58.

Gusfield D, « Algorithms on Strings, Trees, & Sequences »*Cambridge University Press* 1997.

Haas BJ, Volfovsky N, Town CD, Troukhan M, Alexandrov N, Feldmann KA, Flavell RB, White O, Salzberg SL, « Full-length messenger RNA sequences greatly improve genome annotation »*Genome Biol.* 3(6), 2002, electronic publication.

Hide W, Burke J, Davison DB, « Biological evaluation of d2, an algorithm for high-performance sequence comparison »*J. Comput Biol.* 1(3), 1994, 199-215.

Hiller LD, et al., « Generation and analysis of 280,000 human expressed sequence tags »*Genome Res.* 6(9), 1996, 807-828.

Houde M, Belcaid M,Ouellet F, Danyluk J, Monroy FA, Dryanova A, Gulick P, Bergeron A, Laroche A, Links M, MaCarthy L, Crosby WL, Sarhan F, « Wheat EST resources for functional genomics of abiotic stress »*BMC Genomics* 7, 2006, 149.

Hudson TJ, et al., « An STS-based map of the human genome »*Science.* 270(5244), 1995, 1945-1954.

Jones N, Pevzner P, « An Introduction to Bioinformatics Algorithms »*MIT Press* 2000.

Kececioglu JD, Myers EW « Combinatorial Algorithms for DNA sequence assembly »*Algorithmica* 13, 1995, 7-51.

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J, « An optimized protocol for analysis of EST sequences »*Nucleic Acids Res.* 28(18), 2000, 3657-3665.

Needleman SB, Wunsch CD, « A general method applicable to the search for similarities in the amino acid sequence of two proteins »*Journal of Molecular bio.* 148(3), 1970, 443-453.

Nuallain OB, De Rooij S, « Online Suffix Trees with Counts »*Data Compression Conference* 18(6), 2004, 555.

Pearson WR, « Flexible sequence similarity searching with the FASTA3 program package »*Methods Mol. Biol.* 132, 2000, 185-2191.

Rudd S, 200 « Expressed sequence tags : Alternative or complement to whole genome sequences »*Trends Plant Sci.* 8(7), 2003, Epub 2002 May 30.

Smith TF, Waterman SM, « Identification of Common Molecular Subsequences »*Journal of Molecular bio.* 147, 1981, 195-197.

Sterky F, « A Populus EST resource for plant functional genomics »*Proc Natl Acad Sci U S A.* 101(38), 2004, 13951-13956.

Tarjan R, « On the efficiency of a good but not linear set merging algorithm »*Journal of the ACM* 28(3), 577-593

Brassard G, Bratley P, « Fundamentals of algorithmics »*Pearson education* 1996

Pertea G, et al., « TIGR Gene Indices clustering tools (TGICL) : a software system for fast clustering of large EST datasets »*Bioinformatics* 19(5), 2003, 651-652

Torney DC, Davison D, Burkes C, « Computation of d2 : a measure of sequence dissimi-

larity, Computers and DNA »*Computers and DNA, SFI Studies in the Sciences of Complexity* 7, 1990, 109-125.

Ukkonen E, « On-Line Construction of Suffix Trees »*Algorithmica* 14(3), 1995, 249-260.

Wu W, Welsh M, Zhang H, « Gene Biotechnology »*CRC Press* 2, 2003

Stoneking M, « Single nucleotide polymorphisms : From the evolutionary past... »*Nature* 409, 2001, 821-822

Assemblage de séquences
amos http ://amos.sourceforge.net

Points d'articulation dans les graphes
atgc http ://www.atgc.org/BlastParser/Graph9_Program.html

Annotation des EST
autofact http ://www.bch.umontreal.ca/Software/AutoFACT.htm

Portail bioinformatique
ebi http ://www.ebi.ac.uk/

Ontology hierarchique des sequences bilogiques
geneontology http ://www.geneontology.org

Nettoyage des EST
girinst http ://www.girinst.org

Informations relatives au génome humain
ornl http ://www.ornl.gov/sci/techresources /Human_Genome/faq/compgen.shtml#genomesize

Assemblage des séquences
phrap http ://www.phrap.org/phredphrap/general.html

Acclimatation du blé
umanitoba http ://www.umanitoba.ca/afs/agronomists_conf /2001/pdf/struthers.pdf

Assemblage des EST
Sanbi_1 http ://www.sanbi.ac.za/publications/ESTclusteringtutorial.pdf

Algorithmes utilisés dans l'assemblage des EST
Sanbi_2 http ://www.sanbi.ac.za/submissionl.pdf

Calcul de la distance basée sur la fréquence de mots
Sanbi_3 http ://www.sanbi.ac.za/WCD/documentation

BMC Genomics

BioMed Central

Research article

Open Access

Wheat EST resources for functional genomics of abiotic stress

Mario Houde[1], Mahdi Belcaid[2], François Ouellet[1], Jean Danyluk[1], Antonio F Monroy[3], Ani Dryanova[3], Patrick Gulick[3], Anne Bergeron[2], André Laroche[4], Matthew G Links[5], Luke MacCarthy[6], William L Crosby[5] and Fathey Sarhan*[1]

Address: [1]Département des Sciences biologiques, Université du Québec à Montréal, C.P. 8888, Succ. Centre-ville, Montréal QC, H3C 3P8, Canada, [2]Département d'Informatique, Université du Québec à Montréal, C.P. 8888, Succ. Centre-ville, Montréal QC, H3C 3P8, Canada, [3]Biology Department, Concordia University, 7141 Sherbrooke Street West, Montreal QC, H4B 1R6, Canada, [4]Agriculture et Agroalimentaire Canada, Centre de recherches de Lethbridge, 5403, 1st Avenue South, C.P. 3000, Lethbridge AB, T1J 4B1, Canada, [5]Department of Biological Sciences, University of Windsor, 401 Sunset ave, Windsor ON, N9B 3P4, Canada and [6]Department of Computer Science, University of Saskatchewan, 176 Thorvaldson Building, 110 Science Place, Saskatoon SK, S7N 5C9, Canada

Email: Mario Houde - houde.mario@uqam.ca; Mahdi Belcaid - belcaid.mahdi@courrier.uqam.ca; François Ouellet - ouellet.francois@uqam.ca; Jean Danyluk - danyluk.jean@uqam.ca; Antonio F Monroy - amonroy@power2will.com; Ani Dryanova - adryanov@alcor.concordia.ca; Patrick Gulick - pgulick@alcor.concordia.ca; Anne Bergeron - bergeron.anne@uqam.ca; André Laroche - laroche@agr.gc.ca; Matthew G Links - links@uwindsor.ca; Luke MacCarthy - mccarthy@cs.usask.ca; William L Crosby - bcrosby@uwindsor.ca; Fathey Sarhan* - sarhan.fathey@uqam.ca

* Corresponding author

Published: 13 June 2006

BMC Genomics 2006, **7**:149 doi:10.1186/1471-2164-7-149

Received: 08 March 2006
Accepted: 13 June 2006

This article is available from: http://www.biomedcentral.com/1471-2164/7/149

Abstract

Background: Wheat is an excellent species to study freezing tolerance and other abiotic stresses. However, the sequence of the wheat genome has not been completely characterized due to its complexity and large size. To circumvent this obstacle and identify genes involved in cold acclimation and associated stresses, a large scale EST sequencing approach was undertaken by the Functional Genomics of Abiotic Stress (FGAS) project.

Results: We generated 73,521 quality-filtered ESTs from eleven cDNA libraries constructed from wheat plants exposed to various abiotic stresses and at different developmental stages. In addition, 196,041 ESTs for which tracefiles were available from the National Science Foundation wheat EST sequencing program and DuPont were also quality-filtered and used in the analysis. Clustering of the combined ESTs with d2_cluster and TGICL yielded a few large clusters containing several thousand ESTs that were refractory to routine clustering techniques. To resolve this problem, the sequence proximity and "bridges" were identified by an e-value distance graph to manually break clusters into smaller groups. Assembly of the resolved ESTs generated a 75,488 unique sequence set (31,580 contigs and 43,908 singletons/singlets). Digital expression analyses indicated that the FGAS dataset is enriched in stress-regulated genes compared to the other public datasets. Over 43% of the unique sequence set was annotated and classified into functional categories according to Gene Ontology.

Conclusion: We have annotated 29,556 different sequences, an almost 5-fold increase in annotated sequences compared to the available wheat public databases. Digital expression analysis combined with gene annotation helped in the identification of several pathways associated with abiotic stress. The genomic resources and knowledge developed by this project will contribute to a better understanding of the different mechanisms that govern stress tolerance in wheat and other cereals.

Background

Cold acclimation (CA) allows hardy plants to develop the efficient freezing tolerance (FT) mechanisms needed for winter survival. During the period of exposure to low temperature (LT), numerous biochemical, physiological and metabolic functions are altered in plants, and these changes are regulated by LT mostly at the gene expression level. The identification of LT-responsive genes is therefore required to understand the molecular basis of CA. Cold-induced genes and their products have been isolated and characterized in many species. In wheat and other cereals, the expression of several genes during cold acclimation was found to be positively correlated with the capacity of each genotype and tissue to develop FT [1]. Furthermore, abiotic stresses that have a dehydrative component (such as cold, drought and salinity) share some responses. It is therefore expected that, in addition to the genes regulated specifically by each stress, some genes will be regulated by multiple stresses. The availability of wheat genotypes with varying degree of FT makes this species an excellent model to study freezing tolerance and other abiotic stresses. The identification of new genes involved in the cold response will provide invaluable tools to further our understanding of the metabolic pathways of cold acclimation and the acquisition of superior freezing tolerance of hardy genotypes.

Major genomics initiatives have generated valuable data for the elucidation of the expressed portion of the genomes of higher plants. The genome sequencing of *Arabidopsis thaliana* was completed in 2000 [2] while the finished sequence for rice was recently published [3]. The relatively small genome size of these model organisms was a key element in their selection as the first plant genomes to be sequenced with extensive coverage. On the other hand, the allohexaploid wheat genome is one of the largest among crop species with a haploid size of 16.7 billion bp [4], which is 110 and 40 times larger than *Arabidopsis* and rice respectively [5]. The large size, combined with the high percentage (over 80%) of repetitive non-coding DNA, presents a major challenge for comprehensive sequencing of the wheat genome. However, a significant insight into the expressed portion of the wheat genome can be gained through large-scale generation and analysis of ESTs. cDNA libraries prepared from different tissues exposed to various stress conditions and developmental stages are valuable tools to obtain the expressed and stress-regulated portion of the genome. This approach was used in several species such as oat [6], barley [7], tomato [8] and poplar [9]. The sequencing of cDNAs gives direct information on the mature transcripts for the coding portion of the genome that can subsequently be used for gene identification and functional studies. The availability of wheat genomics data in the public datasets has grown rapidly through major initiatives [10,11]. However, additional ESTs are needed to complete the identification of the expressed genes under different growth conditions and from different genotypes. This will contribute to a more complete representation of the genome through identification of new genes and extension of contigs for the majority of genes that have incomplete sequence coverage. Towards this goal, the Functional Genomics of Abiotic Stress (FGAS) program initiated an EST sequencing effort directed toward the study of abiotic stress, with an emphasis on cold acclimation [12]. To increase gene diversity in the EST population and increase the probability of identifying those associated with freezing tolerance, different cDNA libraries were prepared from winter wheat tissues exposed for various times to low temperature, together with select libraries derived from tissues exposed to other stresses or at different developmental stages. In this report, we describe the generation of 73,521 high quality ESTs from wheat stress-associated cDNA libraries. In order to perform the assembly and digital expression analyses, these ESTs were supplemented with wheat ESTs for which sequence quality data was available. These include the NSF [13] and DuPont datasets, which will be referred to as the 'NSF-DuPont' dataset in this report. Digital expression analyses identified a large number of genes that were associated with cold acclimation and other stresses. Expression analyses and functional classification provided important information about the different metabolic and regulatory pathways that are possibly associated with cellular adjustment to environmental stresses. These new EST resources are an important addition to publicly available resources especially in relation to the study of abiotic stresses in cereals.

Results and discussion

The large-scale FGAS wheat EST sequencing project was undertaken to identify new genes associated with abiotic stress and to provide physical resources for functional studies. We have developed a unique wheat EST resource from eleven cDNA libraries prepared from tissues at different developmental stages and exposed to different stress conditions (Table 1). The EST collections from FGAS, NSF and DuPont were analyzed and classified into functional categories.

Assembly and identification of new wheat genes

We have used EST sequences and quality values from the corresponding tracefiles of large datasets (FGAS, NSF and DuPont) to assemble 75,488 different wheat sequences (31,580 contigs, 36,388 singletons and 7,520 singlets). Among these datasets, the FGAS project produced 11,225 unique sequences (2,824 contigs, 6,663 singletons and 1,738 singlets) indicating that the FGAS ESTs encompass a large subset of unique transcripts. These sequences were analyzed using BLASTN on the db_est database and filtered for wheat sequences with two different cut-off e-val-

ues to identify new wheat genes. With an e-25 cut-off value, we found that 2,304 genes had no homologous wheat ESTs (Table 2). After filtering these genes against the wheat protein database with TBLASTX, there were still 2,243 proteins showing no homology to known proteins. With an e-05 cut-off, 1,581 genes had no homologs in wheat. After filtering these against the protein database, 1,470 non-homologous sequences remained. These unique wheat sequences were then BLASTed against *Arabidopsis*, rice, and finally nr db EST (Table 2). In *Arabidopsis*, we found that only 5 of the remaining FGAS wheat sequences had a strong (e-25) similarity using BLASTN while 253 of the remaining sequences had homologs when filtered with the *Arabidopsis* protein database (count down to 1,985). A similar trend was found in *Arabidopsis* using a lower sequence similarity cut-off (e-05). The remaining unique gene count was reduced by several hundred after comparing protein homologs in rice (counts down to 1674 at e-25 and down to 855 at e-05) demonstrating that several genes common between rice and wheat are absent in *Arabidopsis* (Table 2). The remaining unique ESTs were BLASTed against the non redundant database to determine whether homologs were present in other organisms. At an e-05, there were 795 ESTs showing no significant similarity to known domains in genes from other species. It is possible that some of these genes derive from unknown micro-organisms contaminating the plant tissues, and/or from residual genomic DNA in the RNA samples used for cDNA synthesis. However, the majority of these sequences have ORFs encoding proteins larger than 30 amino acids, with an average predicted protein size of over 100 amino acids. This suggests that the unidentified genes do represent novel wheat genes.

The Institute for Genomic research (TIGR) wheat gene index (Release 10.0) shows that only 6,431 of the 44,954 wheat contigs (14%) were successfully allocated a known Molecular Function using Gene Ontology, compared to the classification done for *Arabidopsis* in which 12,558 of the 28,900 contigs (42%) have a known Molecular Function. Therefore, prior to this report, *Arabidopsis* had almost twice as many genes annotated with at least one defined function compared to wheat (12,558 vs 6,431). The classification of the complete dataset (FGAS and NSF-DuPont datasets) allowed the tentative annotation of 43.3% of the genes. As expected, most of the annotated sequences were in contigs (57.6%) while the percentage of annotated singletons/singlets was much lower (30.8%). We have thus been able to functionally annotate 29,556 different sequences, an almost 5-fold increase in annotated sequences compared to TIGR. This is a significant contribution that broadens the available wheat public annotation dataset for downstream functional studies. These results demonstrate that a large number of wheat genes are poorly characterized and stress the fact that major efforts in functional analyses are needed.

Enrichment for stress-regulated genes in the FGAS dataset

Comparative analysis of the FGAS ESTs and NSF-DuPont ESTs based on Gene Ontology (GOslim) showed that several GO classes are more represented in FGAS than in the NSF-DuPont dataset (Figure 1). When general GO classes are compared (GOs 1 to 3; Biological Process, Transcription and Protein Metabolism), no major differences in the number of ESTs were found. Similarly, most GOslim classes showed less than 25% difference between the two datasets. However, GOs 4 and 5 (Enzyme Regulator Activity and Nutrient Reservoir Activity) had a lower representation while GOs 6 to 15 (Transcription Factor Activity, Nuclease Activity, Plasma Membrane, Secondary Metabolism, Response to External Stimulus, Carbohydrate Binding, Response to Abiotic Stimulus, Cell-Cell Signalling, Development and Behavior) were more abundant in the FGAS dataset (Figure 1).

To identify genes that are differentially represented between the two datasets, the relative abundance of ESTs was analyzed and referred to as digital expression analysis. For each contig, the number of ESTs from FGAS (excluding ESTs derived from Suppressive Subtractive Hybridization; SSH) was divided by the number of ESTs from NSF-DuPont and the ratio was normalized to correct for the difference in size between the two datasets (54,032 non SSH EST sequences for the FGAS dataset and 196,041 sequences for the NSF-DuPont dataset). Thus, after normalization, the relative expression level for a contig having 1 EST from each dataset would result in a relative expression of 3.62X in FGAS compared to NSF-DuPont (a ratio of 1 multiplied by 196,041/54,032). Since the SSH technique aims to enrich differentially expressed cDNAs, the ESTs derived from the SSH libraries were analysed separately to avoid a bias in the number of ESTs in a contig, which could invalidate the digital expression analysis approach.

The data indicated that over 75% of the contigs have ratios that vary by less than two-fold, suggesting a similar representation of ESTs between the FGAS (less SSH) and the NSF-DuPont datasets. The remaining 25% of contigs showed more than two-fold difference in abundance (Table 3; see additional file 1: Table1.xls) in the FGAS dataset. When 5- and 10-fold ratios are used as cut-off, 6.6% and 1.7% of the contigs are retained respectively. Most of the differences are due to genes that are over-represented in the FGAS dataset (for the 5-fold cut-off, 1959 genes are over- and 136 genes are under-represented, see Table 3). With a higher cut-off (20-fold differential abundance), only 61 contigs are over expressed and 5 are under-expressed. An analysis of these highly over-repre-

A

GO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
FGAS	26,016	1,378	8,316	262	276	444	230	76	399	971	284	488	31	383	52	39,606
% of Total GOs	65.7	3.5	21	0.66	0.7	1.12	0.58	0.19	1.01	2.45	0.72	1.23	0.08	0.97	0.13	100
NSF-DuPont	86,093	3,704	31,763	2,791	1,878	1,033	522	156	769	1,869	446	615	37	411	53	132,140
% of Total GOs	65.2	2.8	24	2.11	1.42	0.78	0.4	0.12	0.58	1.41	0.34	0.47	0.03	0.31	0.04	100

B

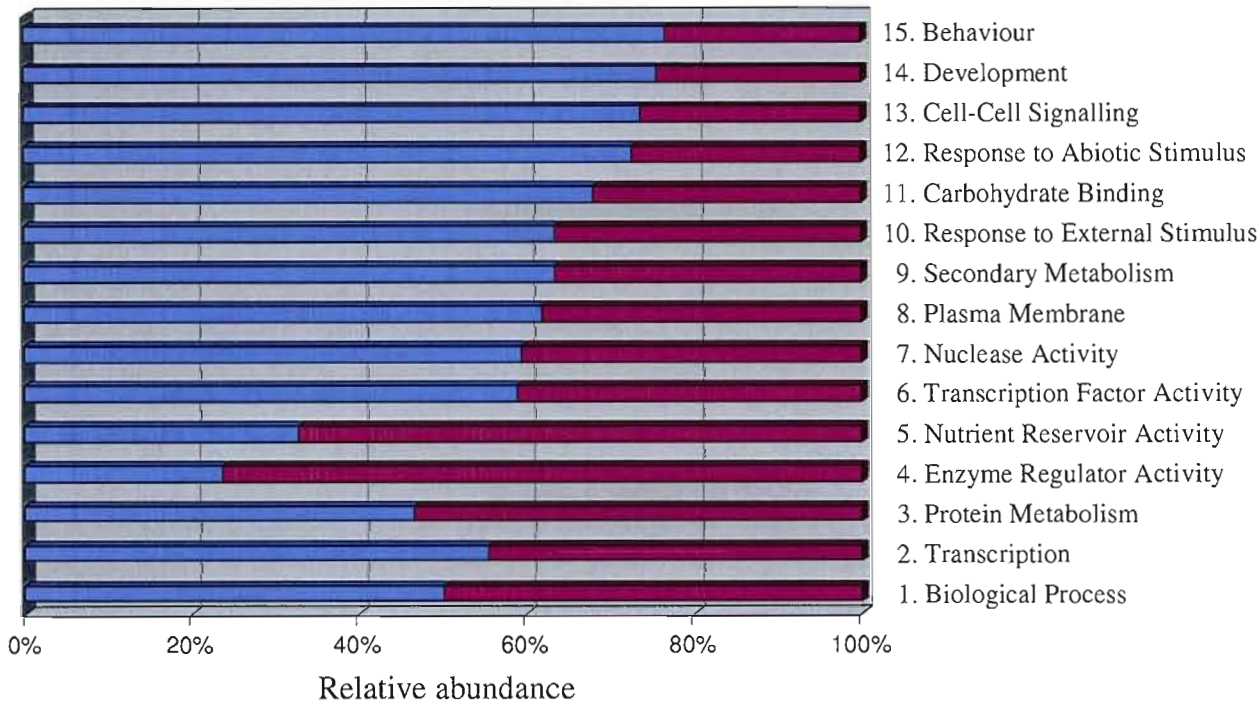


Figure 1

Abundance of annotated ESTs in FGAS contigs relative to NSF-DuPont contigs within select GO classes. A) Number of annotated ESTs. The GO counts were added for each dataset and the percentage of ESTs for each GO was calculated based on this total count. **B)** The relative abundance for each GO is compared between the FGAS (blue) and the NSF-DuPont (red) datasets by comparing the percentage of each GO as determined in A. GO categories: 1. Biological Process GO:0008150; 2. Transcription GO:0006350; 3. Protein Metabolism GO:0019538; 4. Enzyme Regulator Activity GO:0030234; 5. Nutrient Reservoir Activity GO:0045735; 6. Transcription Factor Activity GO:0003700; 7. Nuclease Activity GO:0004518; 8. Plasma Membrane GO:0005886; 9. Secondary Metabolism GO:0019748; 10. Response to External Stimulus GO:0009605; 11. Carbohydrate Binding GO:0030246; 12. Response to Abiotic Stimulus GO:0009628; 13. Cell-Cell Signalling GO:0007267; 14. Development GO:0007275; 15. Behaviour GO:0007610.

sented contigs showed that a good proportion (52%) of these show homology to genes that were previously reported to be over-expressed under stress (see references in Table 4). This high percentage of positive identification suggests that the NSF-DuPont collection was a good reference dataset for digital expression analysis of the FGAS dataset.

Our digital expression analysis relies on the presence of ESTs from both datasets in a same contig (since we cannot

Table 1: Summary of tissues used for the different cDNA libraries generated for the FGAS EST sequencing project.

Library	Growth conditions*	Tissues	High quality EST sequences
Library 2	Control plants; Plants cold acclimated for 1, 23 and 53 days	leaves and crowns	25,240
Library 3	Control plants; Plants cold acclimated for 1, 23 and 53 days; Plants salt stressed for 0.5, 3 and 6 hours	roots	11,382
Library 4	Plants dehydrated on the bench (4 time points) and in a growth chamber (4 time points)	leaves and crowns	2,838
Library 5	Various vernalization and developmental stages through spike formation.	crowns and flowers	6,668
Library 6	Control plants; Plants cold acclimated for short time points (1, 3 and 6 hours) under light or dark conditions	leaves and crowns	7,904
TaLT2	SSH library: Tester: cv. CI14106 cold acclimated for 1 day; Driver: cv Norstar cold acclimated for 21 and 49 days	crowns	2,271
TaLT3	SSH library: Tester: cv. CI14106 cold acclimated for 21 and 49 days; Driver: cv Norstar cold acclimated for 1 day	crowns	1,832
TaLT4	SSH library: Tester: cv. PI178383 cold acclimated for 1 day; Driver: cv Norstar cold acclimated for 21 and 49 days	crowns	2,716
TaLT5	SSH library: Tester: cv. PI178383 cold acclimated for 21 and 49 days; Driver: cv Norstar cold acclimated for 1 day	crowns	2,784
TaLT6	SSH library: Tester: cv. CI14106 cold acclimated for 1 day; Driver: non-acclimated cv. CI14106	crowns	4,961
TaLT7	SSH library: Tester: cv. CI14106 cold acclimated for 21 and 49 days; Driver: non-acclimated cv. CI14106	crowns	4,925

* Libraries 2 to 6 were constructed from wheat cv Norstar.

divide by 0). We have also identified 542 contigs that contained at least 3 ESTs from FGAS but none from the NSF-DuPont dataset (See additional file 2: Table 2.xls). Table 5 lists the 90 genes that contain at least 5 ESTs unique to FGAS, and many of these are similar to genes that have previously been reported to be over-expressed under stress. Although the unique contigs in the FGAS dataset may represent transcripts that are specific to the cultivar used in our study, there is a possibility that they may represent novel genes that are induced by environmental stress.

Table 2: Homology search of FGAS contigs. As a first step, the 11,225 FGAS unique sequences were analyzed using the wheat-filtered db_est (NCBI release 2.2.12, Aug-07-2005). The non-homologous transcripts were then analyzed against the wheat protein database to subtract protein homologs. The remaining transcripts were then analyzed in the same manner against the *Arabidopsis* and rice databases and finally against the nr database. The complete homology search was performed at e-25 and e-05 cut-offs. The numbers indicate the number of genes that do not show any homology at the indicated e-value cut-off.

		e-25	e-05
Wheat	BLASTN db_est	2304	1581
	TBLASTX	2243	1470
Arabidopsis	BLASTN db_est	2238	1470
	TBLASTX	1985	1102
Rice	BLASTN db_est	1845	987
	TBLASTX	1674	855
nr db_est	BLASTN	1623	795

In *Arabidopsis*, microarray experiments have shown that about 10% of the genes are over-or under-expressed by at least two-fold upon exposure to cold acclimation conditions [14]. Based on our previous northern and microarray analyses, we have estimated that the same proportion of wheat genes is cold-regulated (Sarhan *et al.*, unpublished results). If we consider a conservative estimate of 30,000 wheat genes (90,000 if we consider the A, B and D genomes), this means that around 3,000 genes would be cold-regulated. A similar number of genes was identified when we used a 5-fold cut-off differential expression (2,095 differentially expressed contigs, Table 3) and added the 542 contigs having at least 3 ESTs that are unique to the FGAS dataset. Using these criteria, our analyses resulted in a total of 2,637 contigs or 8.4% of the contigs generated in our assembly (31,580 contigs). Considering that 95% of the EST sequences were derived from libraries constructed from cold-acclimated plants, these genes represent candidate genes likely regulated by low temperature and other stresses. However, many of these may be differentially expressed as a consequence of the temperature shift and metabolic adjustment and might not be involved in conferring or regulating increased tol-

Table 3: Contigs containing ESTs that are over or under-represented in the FGAS dataset relative to the NSF-DuPont dataset.

Fold increase/decrease	Over-represented ESTs	Under-represented ESTs	Total	Percent of total contigs (31,772)
20	61	5	66	0.2
10	533	22	555	1.7
5	1959	136	2095	6.6
3	5569	489	6052	19
2	6794	1047	7841	24.7

erance to stress. It would be of interest to analyse these 2,637 genes to identify those relevant to LT tolerance and other stresses in cereals. To verify the conservation of the stress response between wheat and *Arabidopsis*, we first identified the *Arabidopsis* proteins having homology (e-25) to the 2,637 wheat proteins identified in our study, using the TAIR protein database. The homology search resulted in the identification of 1,551 *Arabidopsis* proteins. Most of the genes encoding these proteins are represented on the Affymetrix and MWG microarrays. This allowed us to obtain their expression profiles from the available public data [14,15]. Our analysis indicated that 941 genes are cold-regulated and 890 are drought-regulated (See additional file 1: Table 1.xls and additional file 2: Table 2.xls). There are 678 genes regulated by both stresses, with a total of 1153 different *Arabidopsis* genes that are stress-regulated. Therefore, there are over 44% of the 2,637 putative wheat stress-regulated genes that have a homolog regulated by stress in *Arabidopsis*, suggesting overlapping responses between the two species.

As a complementary approach to identifying new wheat genes that may be differentially expressed, different SSH libraries were produced to identify genes over-expressed after brief (1 day) or long (21–49 days) periods of cold acclimation. Different cultivars that may help to identify other components of freezing tolerance such as pathogen resistance to snow molds were used for these analyses. A total of 3,873 contigs containing 18,610 SSH ESTs were obtained with 2,969 contigs (76.7%) tentatively annotated. Unique contigs from SSH libraries are potentially a good source to mine for new genes associated with cold acclimation. Overall, 225 contigs unique to the SSH libraries (See additional file 3: Table 3.xls) were identified, among which 74 were annotated (Table 6). We found that 11 of the 74 annotated SSH contigs (or 15% of the unique SSH contigs) have corresponding genes (high similarity based on BLASTX e-values) that are over-expressed more than 5-fold in the differentially-expressed FGAS contigs. These results suggest that unique SSH contigs contain candidate genes that could be involved in abiotic stress tolerance.

Metabolic pathways associated with differentially expressed genes

GO slim annotation was used to subdivide the 2,637 stress-regulated genes into function categories to gain insight into their putative role during cold acclimation and abiotic stresses. The results show that a large proportion of these contigs were annotated under a limited number of GO classes (Figure 2). Over 53.7% of the contigs were grouped into 14 GO categories while 27.5% of the contigs were designated "No Gene Ontology" and 4.2% were classified as "Hypothetical Protein", a term used to designate open reading frames predicted from the *Arabidopsis* or rice genomic DNA. The remaining contigs with other GO categories were grouped together in one category (14.6%).

A plethora of physiological and metabolic adjustments occur during cold acclimation and in response to other stresses. The regulation of genes involved in temperature, drought and salt stresses is known to reflect the cross-talk between different signalling pathways [16]. However, few studies have identified multiple genes that are stress-regulated and that belong to a same metabolic pathway. Our analyses enabled us to position several genes in their respective metabolic pathway, suggesting that these pathways are involved in stress responses. Since it is beyond the scope of this report to cover all possible pathways involved, we highlight some of the key elements that likely contribute to the stress response and tolerance. Unless specifically indicated, all enzymes discussed are encoded by transcripts that are over-represented by at least 5-fold in the FGAS dataset.

Amino acid metabolism

Genes encoding proteins involved in primary metabolism pathways have been identified in the contigs with an over-representation of FGAS ESTs and cover several aspects of plant metabolic adjustments. Amino acid metabolism and the TCA cycle are the major pathways that generate precursors for various biological molecules. ESTs encoding several enzymes that are involved in the synthesis of arginine, cysteine, lysine, methionine, serine, phenyla-

lanine, proline and tryptophan are over-represented by more than 5-fold. These amino acids are precursors for the synthesis of several specialized metabolites. Two contigs encode the enzyme delta-1-pyrroline-5-carboxylate synthetase that is involved in proline biosynthesis, a metabolite that was found to increase during cold acclimation and drought stress [17]. Similarly, two contigs encode glutamate decarboxylase (GAD1), which is involved in the synthesis of gamma-aminobutyric acid (GABA), a non protein amino acid known to accumulate during cold acclimation and proposed to function in oxidative stress tolerance [18]. Several contigs encode enzymes involved in the metabolism of cysteine, an important precursor of glutathione involved in the modulation of oxidative stress. These include two different cysteine synthases and a putative O-acetylserine (thiol) synthase (OASTL). Overexpression of different isoforms of OASTL can increase thiol content in different transgenic plants and increase tolerance to abiotic stress such as exposure to elevated levels of cadmium [19].

Lipid metabolism

ESTs encoding different putative lipases and other proteins involved in lipid oxidation (acyl-CoA oxidase, MutT/nudix protein like, dihydrolipoamide acetyltransferase, b-keto acyl reductase, enoyl-ACP reductase, enoyl-CoA-hydratase, 3-hydroxyisobutyryl-coenzyme A hydrolase) are over-represented in the FGAS dataset while the acyl-carrier protein III involved in lipid synthesis is under-represented. These results suggest that lipid degradation occurs concomitantly with a reduction in the synthesis of short chain lipids. On the other hand, ESTs encoding enzymes involved in the synthesis of specialized lipids such as ATP citrate lyase α-subunit and the long chain fatty acid enzyme acetyl-CoA carboxylase are more abundant among FGAS ESTs. ESTs corresponding to several enzymes involved in sterol metabolism are also over-represented, suggesting major lipid modifications in membranes during cold acclimation. ESTs encoding three enzymes involved in the alternate pathway of isopentenyl pyrophosphate and squalene synthesis (1-deoxy-D-xylulose 5-phosphate reductoisomerase, 1-deoxy-D-xylulose-5-phosphate synthase, squalene synthase), three key enzymes of the sterol pathways (cycloartenol synthase, C14-sterol reductase (FACKEL), and 24-methylenelophenol methyltransferase) (Figure 3), and other enzymes such as sterol 4-alpha-methyl-oxidase, which can add to the variety of sterols produced, are also over-represented. The putative over-expression of several enzymes in the sterol pathway supports the previous observation of an increased production of membrane sterols [20]. These authors showed that the concentration of membrane sterols increases during cold acclimation and that this effect is more prominent in tolerant rye cultivars. Interestingly, sitosterol increases while campesterol decreases during acclimation, suggesting that the C24 methyltransferase that is putatively over-expressed in the FGAS dataset may be the SMT-2 transferase that diverts the methylenelophenol into the sitosterol pathway (see Figure 3; [21]). A search through the protein database has shown that the C24 methyltransferase has a much greater homology with SMT2 (7e-143) than with SMT1 (4e-63) supporting that the C24 methyltransferase is SMT2. The over-representation of FGAS ESTs in two contigs encoding stearoyl-acyl-carrier protein desaturase and two contigs encoding CDP-diacylglycerol synthase suggests that other important lipid modifying activities also occur in response to cold acclimation. Stearoyl-acyl-carrier protein desaturase is involved in the desaturation of existing lipids to form double bonds rendering the lipids more fluid at low temperature. This is an important adjustment associated with membrane stability at low temperature [20]. The overexpression of CDP-diacylglycerol synthase was previously shown to favour the synthesis of phosphatidylinositol [22]. In addition, one contig encodes a phosphoethanolamine N-methyltransferase. This enzyme is induced by low temperature and catalyzes the three sequential methylation steps to form phosphocholine, a key precursor of phosphatidylcholine and glycinebetaine in plants – metabolites known to be important in conferring tolerance to osmotic stresses such as low temperature, drought and salinity [23].

Secondary metabolism

Several contigs encode key enzymes involved in the biosynthesis of secondary metabolites such as phenylalanine ammonia lyase, cinnamyl alcohol dehydrogenase, and caffeoyl-CoA O-methyltransferase. Several enzymes are involved in the synthesis of methionine and its derivatives. The digital expression data suggest that the S-adenosylmethionine (SAM) cycle becomes more active during stress since contigs encoding three major enzymes of the cycle (S-adenosylmethionine synthetase, methionine S-methyltransferase, and S-adenosylhomocysteine hydrolase) are over-represented in FGAS. This pathway can provide SAM, the precursor molecule needed for nicotianamine biosynthesis. Four different contigs encoding nicotianamine synthase or nicotianamine aminotransferase are over-represented in FGAS. These enzymes are involved in nicotianamine and phytosiderophores synthesis and were found to be induced under iron deficiency [24,25]. The SAM cycle also provides the one carbon precursor for the methylation steps required for methyltransferase activities. At least 20 different contigs encoding methyltransferases contain ESTs that are over-represented in FGAS.

Transport activity

During cold acclimation, the cell mobilizes several transport systems to adapt to cold conditions. One of the major

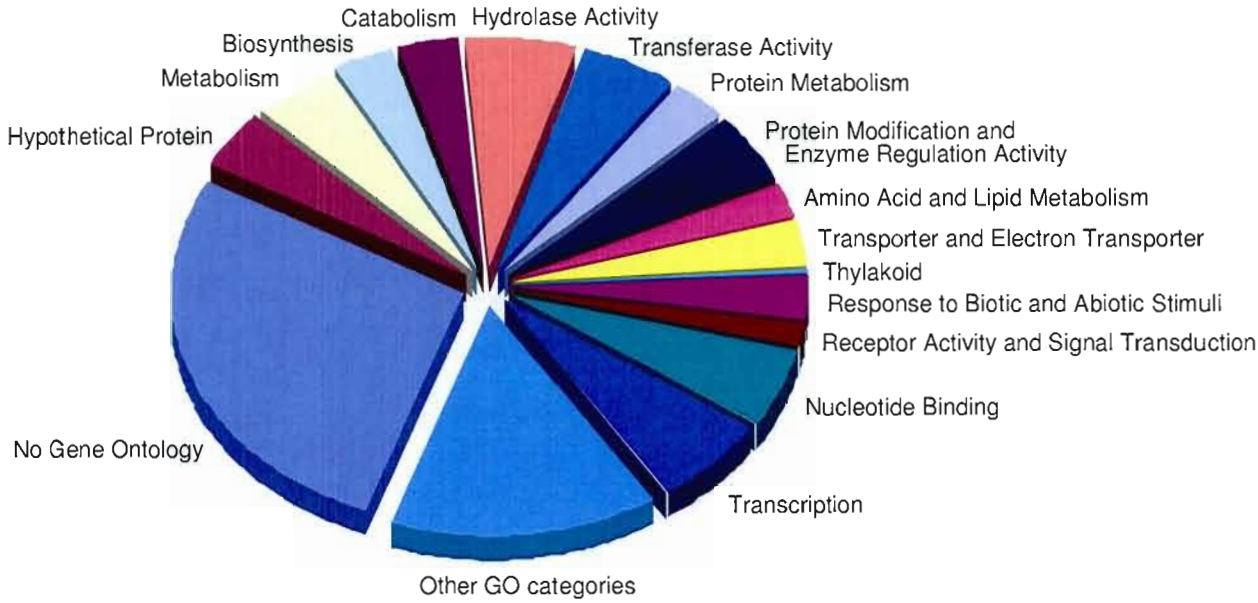


Figure 2

Functional classification of FGAS contigs containing ESTs that are over or under-represented more than 5-fold, or that contain more than 3 unique ESTs. The contigs belonging to the following GO terms were used: GO0008152 Metabolism; GO0009058 Biosynthesis; GO0009056 Catabolism; GO0016787 Hydrolase Activity; GO0016740 Transferase Activity; GO0019538 Protein Metabolism; GO0006464 and GO0030234 Protein Modification and Enzyme Regulator Activity; GO0006519 and GO0006629 Amino Acid and Lipid Metabolism; GO0005215 and GO0005489 Transporter and Electron Transporter Activity; GO0009579 Thylakoid; GO0009607 and GO 0009628 Response to Biotic and Abiotic Stimulus; GO0004872 and GO0007165 Receptor Activity and Signal Transduction; GO000166 Nucleotide Binding; Transcription Factors only from GO0006350 and GO0003677 (other DNA Binding Proteins were transferred to "Other GO categories"); a class was made for the mention "Hypothetical Protein" and for the mention "No Gene Ontology" while the "Other GO Categories" regroups several GO terms with small number of contigs.

effects of extracellular freezing is the reduced apoplastic water pressure and the rapid flow of water from the intracellular compartment to the apoplasm. Some of the consequences include the need for water and ion regulation as well as protection against dehydration. Two different contigs encoding aquaporins are highly abundant in FGAS (a contig with 12 ESTs found only in the FGAS dataset and a contig with ESTs over-represented 18-fold). These proteins likely play an important role in the regulation of the outward water flow. Similarly, several contigs associated with transport of ions or other small solutes are more highly represented, such as anion/sugar transporters, major facilitator superfamily antiporters, MATE efflux family transporters, nitrate transporters, cation exchangers, calcium and zinc transporters, betaine/proline transporters, and amino acid transporters. These different transporters are potential regulators controlling the flow of ions and other solutes that become more concentrated as water is drawn out of the cell during freezing. An interesting transporter activity is the phosphatidylinositol-phosphatidylcholine transfer protein which can contribute to the turnover of these lipids in the membrane. This pathway is involved in the accumulation of the compatible solute betaine that was reported to increase tolerance to drought and freezing [26]. Another mechanism involved in cell protection against higher ionic content include the replacement of water with compatible solutes such glycerol, glucose, sorbitol, proline and betaine. ESTs encoding hydroquinone glucosyltransferase, an interesting enzyme responsible for the synthesis of arbutin, are over-represented over 7-fold in the FGAS dataset. Glycosylated hydroquinone is very abundant in freezing and desiccation tolerant plants. It was suggested to accumulate up to 100 mM in the resurrection plant *Myrothamnus flabellifolia* and to increase membrane stability of artificial liposomes and thylakoids, possibly through the insertion of the phenol moiety in the phospholipid bilayer [27]. These authors showed that the lipid membrane composition is an important element for the cryoprotective effect of arbutin. In support of this observation, several contigs

with an over-representation of FGAS ESTs encoding transporters of compatible solutes and lipid modifying enzymes were identified.

Proteins involved in cryoprotection

One strategy that hardy plants such as wheat use to tolerate subzero temperatures is the accumulation of freezing tolerance associated proteins such as antifreeze proteins (AFPs) and dehydrins [28]. AFPs exhibit two related activities *in vitro*. The first is to increase the difference between the freezing and melting temperatures of aqueous solutions, a property known as thermal hysteresis. The second is ice recrystallization inhibition (IRI), where the growth of large ice crystals is inhibited, thus reducing the possibility of physical damage within frozen tissues [29]. In winter wheat and rye, several AFPs similar to pathogenesis-related proteins such as chitinases, glucanases, thaumatins and ice recrystallization inhibition proteins were identified [30-32]. Many contigs encoding chitinases, β-1,3-glucanases and thaumatin-like proteins contain ESTs that are over-represented in FGAS. Hincha *et al.* [33] reported that different cryoprotective proteins were able to protect thylakoids from freezing injury *in vitro*. Wheat ice recrystallization inhibition proteins are partly homologous to, and were annotated as, phytosulfokine receptors and were present in several contigs containing ESTs over-expressed in FGAS.

The dehydrins are hydrophilic proteins resistant to heat denaturation composed largely of repeated amino acid sequence motifs. They possess regions capable of forming an amphipathic α-helix. These properties may enable them to protect cells against freezing damage by stabilizing proteins and membranes during conditions of dehydration [28]. The most studied dehydrins are the WCS120 family, the WCOR410 and the chloroplastic WCS19 dehydrins. Genes encoding these proteins are highly over-represented in the FGAS dataset (Table 4, Table 5, and see additional file 1: Table 1.xls).

Photosynthesis

During cold acclimation, the chloroplast continues to receive as much light as at normal temperature but its thermal biochemical reactions are reduced. This results in an excess of light energy whereby electrons accumulate mostly in Q_A [34]. The reduced capacity to transfer electrons through PSII requires metabolic adjustments on a short term basis through redox balance, and communication between the chloroplast and the nucleus to modify gene expression for adaptation on a longer term basis. Freezing tolerant plants were previously shown to better cope with photoinhibition than less tolerant cultivars [34]. Although the number of genes classified under the GO "Thylakoids" is only 13, the genes identified indicate that putative changes in expression occur for genes encoding components of both the photosystem I (PSI) and the photosystem II (PSII). Several studies have reported changes in PSII during cold acclimation [34], The D1 and D2 proteins were shown to be sensitive to excess energy and to turn over more rapidly at low temperature and high light [35]. ESTs encoding the D2 protein are over-expressed by 7.2-fold in FGAS suggesting that the PSII adapts to low temperature conditions. On the other hand, the transcript encoding PSII Z is less represented in FGAS. A reduced amount of this protein may lead to a reduction in active antennas and allow a reduction in electron flow towards the PSII. ESTs encoding two other proteins of the PSII complex are over-represented (29.8 kDa and 20 kDa protein). These proteins belong to the same PsbP protein family which has 4 members in *Arabidopsis*. Recent results using RNAi have shown that this lumen protein is both essential and quantitatively related to PSII efficiency and stability. This suggests that their over-expression could improve electron flow through PSII [36,37]. Another limiting factor in the electron flow is the availability of CO_2. Several contigs with over-represented ESTs in the FGAS dataset encode carbonic anhydrase (carbonic anhydrase chloroplast precursor, dioscorin class A and nectarin III). This enzyme is known in C4 plants to concentrate CO_2 at its site of fixation. In the C3 plant wheat, this enzyme was previously shown to be modulated by nitrogen deficiency to maintain optimal CO_2 concentrations [38]. The over-expression of this enzyme could thus help to efficiently use the CO_2 and available light energy at low temperature. Failure to dissipate excess light energy could lead to oxidative stress, which needs to be controlled. A contig encoding a putative serine hydroxymethyltransferase is over-represented in the FGAS dataset. Hydroxymethyltransferases play a critical role in controlling the cell damage caused by abiotic stresses such as high light and salt, supporting the notion that photorespiration forms part of the dissipatory mechanisms of plants to minimize production of reactive oxygen species (ROS) in the chloroplast and to mitigate oxidative damage [39].

Very few studies have documented the modulation of PSI under stress conditions. The excess light or low temperature can decrease stromal NADP/NADPH ratio and it has been proposed that the cytochrome b6f complex can be regulated by the stromal redox potential possibly via a thioredoxin mediated mechanism (see [40]). The PSI components are largely integrated and composed of many subunits making it energetically expensive for the cell to produce. It has been suggested that cells might modulate PSI activity by varying the amount of the small and mobile plastocyanin protein carrying the reducing power [41]. The over-representation of ESTs encoding this protein in FGAS (represented by 27 ESTs within contig CL187Contig5) suggests that this PSI electron relay component becomes more active during cold acclimation and

Table 4: Contigs containing ESTs that are over-represented over 20-fold in the FGAS dataset.

Contig name	Annotation	Fold representation (FGAS/NSF-DuPont)	Reference
CL91Contig4	No Gene Ontology Hit (Wcor413, manual annotation)	163.30	[59]
CL206Contig4	Low molecular mass early light-inducible protein HV90, chloroplast precursor (ELIP)	94.35	[60]
CL386Contig5	Chitinase (EC 3.2.1.14)	68.94	[31]
CL1959Contig1	Legumin-like protein	68.94	[61]
CL117Contig7	No Gene Ontology Hit (Lea/Rab, manual annotation)	61.69	[62,63]
CL10Contig25	Defensin precursor	54.43	[64]
CL347Contig1	COR39 (WCS120 homolog, manual annotation)	52.61	[65]
CL158Contig8	Putative 1-aminocyclopropane-1-carboxylate oxidase	47.17	
CL386Contig1	Chitinase 1	43.54	[31]
CL347Contig2	Cold shock protein CS66 (Wcs120 homolog, manual annotation)	43.54	[65]
CL756Contig2	Hypothetical protein 259116.2b (LEA homolog, manual annotation)	43.54	[66]
CL1620Contig2	No Gene Ontology Hit	32.66	
CL411Contig1	Putative phytosulfokine receptor (Wheat Ice recristallization inhibitor, manual annotation)	32.66	[32]
CL349Contig4	Ferredoxin-NADP(H) oxidoreductase	32.66	[45]
CL1918Contig1	Glycosyltransferase	32.66	[16]
CL2Contig21	Hypothetical protein (Fragment) (Cab binding protein, manual annotation)	32.66	Genbank U73218
CL756Contig3	No Gene Ontology Hit	29.03	
CL2Contig9	Hypothetical protein (Fragment) (Cab binding protein, manual annotation)	29.03	Genbank U73218
CL3270Contig2	No Gene Ontology Hit	29.03	
CL28Contig11	Extracellular invertase (EC 3.2.1.26)	29.03	[67]
CL650Contig2	Cold acclimation protein WCS19	26.30	[68]
CL1442Contig1	Putative major facilitator superfamily antiporter	25.40	
CL1698Contig3	No Gene Ontology Hit	25.40	
CL704Contig4	Legumin-like protein	25.40	[61]
CL4965Contig1	Hypothetical protein P0508B05.10	25.40	
CL4930Contig1	ATP-dependent RNA helicase	25.40	[69]
CL411Contig4	No Gene Ontology Hit (Wheat Ice recristallization inhibitor, manual annotation)	25.40	[32]
CL2910Contig2	CONSTANS-like protein CO6	25.40	
CL117Contig3	No Gene Ontology Hit (Lea/Rab)	25.40	[63]
CL4699Contig1	Cytochrome P450	25.40	
CL4567Contig1	No Gene Ontology Hit	25.40	
CL1631Contig3	Beta-1,3-glucanase	25.40	[33]
CL411Contig3	Putative phytosulfokine receptor (Wheat Ice recristallization inhibitor, manual annotation)	25.40	[32]
CL91Contig8	No Gene Ontology Hit (COR413, manual annotation)	25.40	[59]
CL2020Contig1	No Gene Ontology Hit	23.58	
CL1106Contig2	Putative cytochrome c oxidoreductase	23.58	[70]
CL280Contig5	No Gene Ontology Hit (blt14, manual annotation)	23.58	[71]
CL1911Contig2	Putative cysteine proteinase inhibitor	21.77	[72]
CL3036Contig1	No Gene Ontology Hit hypothetical protein (OSJNBa0062C05.24, manual annotation),	21.77	
CL171Contig6	No Gene Ontology Hit	21.77	
CL202Contig14	No Gene Ontology Hit	21.77	
CL2484Contig2	No Gene Ontology Hit (putative F-Box family, manual annotation)	21.77	
CL3205Contig2	Hypothetical protein At2g43940	21.77	
CL117Contig2	No Gene Ontology Hit (Lea/Rab)	21.77	[63]
CL2663Contig3	Serine carboxypeptidase I precursor (EC 3.4.16.5) (Carboxypeptidase C) (CP-MI)	21.77	
CL4989Contig1	No Gene Ontology Hit	21.77	
CL437Contig6	Putative family II lipase EXL4	21.77	
CL2012Contig3	CIPK-like protein 1 (EC 2.7.1.37) (OsCK1)	21.77	[73]
CL1442Contig3	Putative major facilitator superfamily antiporter (sugar transporter family, manual annotation)	21.77	
CL3511Contig1	Similarity to receptor protein kinase (leucine rich protein similar to TIR1, manual annotation)	21.77	[74]
CL861Contig1	No Gene Ontology Hit	21.77	
CL4814Contig1	Putative cinnamyl alcohol dehydrogenase	21.77	
CL2Contig49	Chlorophyll a/b-binding protein WCAB precursor	21.77	Genbank U73218

Table 4: Contigs containing ESTs that are over-represented over 20-fold in the FGAS dataset. *(Continued)*

CL4798Contig1	No Gene Ontology Hit	21.77	
CL1740Contig2	Hypothetical protein OSJNBa0086E02.13 (Hypothetical protein P0419C04.2) (putative haloacid dehalogenase-like hydrolase, manual annotation)	21.77	
CL4476Contig1	No Gene Ontology Hit (phosphate induced protein, manual annotation)	21.77	
CL4337Contig1	Putative o-methyltransferase	21.77	[75]
CL2623Contig1	No Gene Ontology Hit (lumenal protein subunit of photosystem II, manual annotation)	21.77	
CL3656Contig2	Barwin	21.77	
CL671Contig1	No Gene Ontology Hit	21.77	
CL878Contig3	Putative pollen allergen Jun o 4	21.77	
CL26Contig8	No Gene Ontology Hit	0.050	
CL350Contig1	Photosystem II reaction center Z protein	0.040	
CL185Contig1	Chloroplast 50S ribosomal protein L14	0.037	
CL120Contig2	Lipid transfer protein 1 precursor	0.030	
CL144Contig2	Alpha amylase inhibitor protein	0.026	

may be important in relieving the pressure caused by electrons accumulating in Q_B. The mobile plastocyanin molecule is a limiting factor in the electron transfer from PSII to PSI. The increased expression of plastocyanin may result in an increased activity of PSI under low temperature and may help freezing tolerant plants maintain their energy balance compared to less tolerant plants. We have previously shown that several proteins involved in improving photosynthesis, including plastocyanin, are expressed at low levels under low excitation pressure (20°C/50 μE) but markedly accumulate when transferred to 5°C under the same light regime [42]. A mutation in the PSI-E subunit was also shown to have a great impact on PSII as it becomes easily affected by photoinhibition even under low light [43]. Similarly mutants in the PSI-N subunit, which participates in the docking of PC, are impaired in PSI activity [44]. The over-representation of ESTs encoding the PSI-E and PSI-N subunits in the FGAS dataset could thus provide an integrated response to reduce photoinhibition. In order to maintain a proper NADP/NADPH ratio, the malate valve could be activated to transfer excess reducing power to the cytoplasm [45]. ESTs encoding two PSI components are less abundant in FGAS. One of these is a subunit of the chloroplastic NADH dehydrogenase equivalent to the mitochondrial enzyme. Interestingly, the *FRO1* gene was recently shown to encode the mitochondrial NADH dehydrogenase counterpart which plays a role in controlling ROS and the ability of *Arabidopsis* to respond to low temperature [46]. An excess of ROS in mitochondria was proposed to affect the induction of CBF transcription factors and cold acclimation. The chloroplastic NADH dehydrogenase may also affect the ability to induce CBF if the ROS that accumulate during photoinhibition at low temperature are not detoxified. Tolerant plants may adapt their photosystems to avoid the accumulation of ROS in chloroplasts, thus allowing a strong CBF response and a stable induction of downstream cold-regulated genes. This hypothesis may explain why tolerant plants are able to maintain a strong expression of several freezing tolerance-associated genes while less tolerant plants show transient, reduced expression of these genes at low temperature [1].

Signalling cascades and transcription factors

Among the contigs with an over-representation in FGAS ESTs, we identified several proteins involved in the synthesis or perception of different hormones. These include enzymes of the ethylene, auxin and jasmonic acid metabolism; brassinosteroid LRR receptor, receptor-like kinases CLAVATA2 and PERK1, and phytosulfokine receptor. Contigs encoding several proteins involved in signalling cascades were also found such as calcium binding proteins, diacylglycerol kinase, lipid phosphate phosphatase-2, inositol 1-monophosphatase, GTP-binding proteins, MAP kinases and MAPKK, serine/threonine kinase, CIPK-like protein-1, histidine kinase-2, and protein phosphatases 2A and 2C.

The potentially increased activity of the various signalling pathways is associated with a differential expression of many families of transcription factors (TF; Table 7). The results show that at least 220 contigs contain ESTs encoding TF that are over- or under-represented more than two-fold in the FGAS dataset. Using a more stringent cut-off excludes some TF that may not be strongly regulated, but should also reduce the number of false positives. With a 5-fold cut-off, 151 TF were identified, with 30 of them being contigs unique to FGAS. The most highly represented TF families are the zinc fingers, WRKY, AP2, Myb and NAC. Several members of these families were previously identified as being responsive to various stresses. The most studied members are those of the AP2 family, in particular the CBF/DREB subfamily. CBF members are involved in the cold/drought responses [47]. We have identified 3 different contigs, with a 5-fold over-representation in the FGAS dataset, that contain CBF-like binding factors and 5 unique FGAS contigs containing at least 3 ESTs (annotated as CBF-like, CBF1-like, CBF3-like, C-

Table 5: Contigs containing at least 5 ESTs that are unique to the FGAS dataset.

Contig name	Annotation	Number of ESTs	Reference
CL1638Contig1	No Gene Ontology Hit (no homology)	24	[76]
CL1293Contig2	Wheat cold acclimation protein Wcor80 (Wcs120 homolog, manual annotation)	19	[65]
CL386Contig3	Chitinase I	18	[31]
CL347Contig3	Cold acclimation protein WCS120 (manual annotation)	17	[65]
CL2466Contig1	Putative heat shock protein (E. Coli contaminant, manual annotation)	16	
CL3394Contig1	Nitrogen regulation protein NR(II) (EC 2.7.3.-) (E. coli contaminant, manual annotation)	12	
CL7Contig23	Aquaporin PIP1		[77]
CL40Contig14	Chitinase IV	11	[31]
CL650Contig3	Chloroplast-targeted COR protein (Wcor14c, manual annotation)		[76]
CL1239Contig3	Putative LMW heat shock protein	10	
CL2570Contig1	Hypothetical protein OJ1015F07.4		
CL125Contig7	O-methyltransferase	9	[75]
CL206Contig11	Low molecular mass early light-inducible protein HV90, chloroplast precursor (ELIP)		[60]
CL3635Contig1	No Gene Ontology Hit		
CL4047Contig1	ABA responsive protein mRNA (manual annotation)		[78]
CL52Contig12	No Gene Ontology Hit		
CL52Contig13	No Gene Ontology Hit		
CL619Contig5	WSI76 protein induced by water stress (galactinol synthase, manual annotation)		[79]
CL1228Contig3	Leaf senescence protein-like	8	
CL1293Contig1	Dehydrin (Wcs120 homolog, manual annotation)		[65]
CL2543Contig2	No Gene Ontology Hit		
CL400Contig4	Cysteine protease		[80]
CL4107Contig1	No Gene Ontology Hit		
CL4776Contig1	Probable arylsulfatase activating protein aslB (E. coli contaminant, manual annotation)		
CL1051Contig5	C repeat-binding factor 2	7	[81]
CL2204Contig1	No Gene Ontology Hit (Wheat Ice recristallization inhibitor, manual annotation)		[32]
CL3474Contig1	No Gene Ontology Hit		
CL3792Contig1	No Gene Ontology Hit		
CL4454Contig1	No Gene Ontology Hit		
CL5468Contig1	Ubiquinone/menaquinone biosynthesis methyltransferase ubiE (EC 2.1.1.-) (E. coli contaminant, manual annotation)		
CL833Contig4	Putative EREBP-like protein (putative AP2 domain transcription factor, manual annotation)		
CL1318Contig2	S-like Rnase	6	[82]
CL1368Contig4	Beta-expansin		
CL17Contig3	Type I non-specific lipid transfer protein precursor (Fragment)		[83]
CL20Contig27	No Gene Ontology Hit		
CL2425Contig2	Putative lectin		[84]
CL280Contig2	Low temperature responsive barley gene blt14 (manual annotation)		[62]
CL280Contig4	Cold regulated protein pao29 (similar to blt14 manual annotation)		[62]
CL2910Contig1	CONSTANS-like protein CO6		
CL3212Contig2	No Gene Ontology Hit		
CL3324Contig2	RING zinc finger protein-like		
CL3647Contig2	No Gene Ontology Hit		
CL3778Contig2	Putative phenylalanyl-tRNA synthetase alpha chain		
CL4292Contig1	C2H2 Zinc finger protein (manual annotation)		
CL4895Contig1	No Gene Ontology Hit		
CL5228Contig1	Putative inositol-(1,4,5) trisphosphate 3-kinase		
CL5712Contig1	Putative ABCF-type protein (anthocyanin transport)		
CL5985Contig1	Hypothetical protein P0508B05.10		
CL6056Contig1	Putative calcium binding EF-hand protein (caleosin: lipid body trafficking, manual annotation)		
CL6257Contig1	No Gene Ontology Hit		
CL6493Contig1	No Gene Ontology Hit		
CL861Contig2	No Gene Ontology Hit		
CL1051Contig2	C repeat-binding factor 2	5	[81]
CL1182Contig3	OSJNBa0043A12.18 protein (putative transcription factor)		
CL1279Contig2	Isoflavone reductase homolog (EC 1.3.1.-)		
CL1366Contig3	Putative UDP-glucose: flavonoid 7-O-glucosyltransferase		
CL206Contig6	High molecular mass early light-inducible protein HV58, chloroplast precursor (ELIP)		[60]
CL3647Contig1	No Gene Ontology Hit		

Table 5: Contigs containing at least 5 ESTs that are unique to the FGAS dataset. *(Continued)*

CL4058Contig1	Myb-related protein Hv33	
CL411Contig7	No Gene Ontology Hit (Wheat Ice recristallization inhibitor, manual annotation)	[32]
CL4350Contig2	Similarity to protein kinase	GenBank AY738149
CL4537Contig1	Putative ACT domain-containing protein	
CL4642Contig1	Chitinase 1	[31]
CL4666Contig1	Farnesylated protein 1	[85]
CL4825Contig1	Hypothetical protein P0473D02.6 (Hypothetical protein OJ1368_G08.21)	
CL6137Contig1	No Gene Ontology Hit	
CL6258Contig1	Putative sodium-dicarboxylate cotransporter	
CL6567Contig1	Putative arabinogalactan protein	
CL6634Contig1	No Gene Ontology Hit	
CL6741Contig1	Putative b-keto acyl reductase (fatty acid elongase, waxes biosyntheisis)	
CL6821Contig1	Putative strictosidine synthase (alkaloid biosynthesis)	
CL7090Contig1	No Gene Ontology Hit	
CL721Contig3	No Gene Ontology Hit	
CL7241Contig1	No Gene Ontology Hit	
CL7243Contig1	No Gene Ontology Hit	
CL7272Contig1	Early light-inducible protein	[60]
CL7415Contig1	No Gene Ontology Hit	
CL7455Contig1	ABC1 family protein-like	
CL754Contig3	Chitinase 3	[31]
CL7581Contig1	Aspartate transaminase, mitochondrial	
CL7608Contig1	Putative aspartic proteinase nepenthesin 1	
CL7617Contig1	No Gene Ontology Hit (barley Blt14 homolog, manual annotation)	[62]
CL7686Contig1	No Gene Ontology Hit	
CL7701Contig1	Putative FH protein interacting protein FIP2 (potassium channel tetramerization)	
CL7785Contig1	No Gene Ontology Hit	
CL7794Contig1	No Gene Ontology Hit	
CL807Contig3	Putative diphosphonucleotide phosphatase (calcineurin-like phosphoesterase)	
CL861Contig5	No Gene Ontology Hit	
CL963Contig4	OSJNBb0013O03.11 protein (bHLH transcription factor, manual annotation)	

repeat binding factor 3-like, C-repeat/DRE binding factor 3, CRT/DRE binding factor 2, DRE binding factor-2). Expression profiling using qRT-PCR has confirmed that transcripts corresponding to 7 of the 8 contigs are over-expressed at specific time points during cold acclimation (Sarhan *et al.* unpublished results). Expression of the CBF genes in *Arabidopsis* was shown to be regulated by members of the bHLH family [48]. We have identified 7 contigs encoding bHLH members that are over-represented by two-fold, with two of them being over-represented more than 5-fold (Table 7). However, the genes encoding the bHLH ICE proteins in *Arabidopsis* are not cold-induced. Although the expression pattern with regards to cold inducibility of the ICE genes could be different between wheat and *Arabidopsis*, the isolation of the full length genes, phylogenetic analysis and expression studies are required to determine if any of the over-represented bHLH encode ICE homologs. In addition to the CBFs and bHLH families, several other TF families may be part of other stress components associated with abiotic stress such as drought, salinity, oxidative, etc. Interestingly, several genes that control flowering have also been identified (FLT, Gigantea, MADS, CO, Aintegumenta). These genes are most likely associated with the vernalization response in wheat as was recently shown for *TaVRT1* and *TaVRT2* [49,50].

Conclusion

The large number of ESTs annotated from FGAS and NSF-DuPont datasets represents an important resource for the wheat community. Digital expression analyses of these datasets provide an overview of metabolic changes and specific pathways that are regulated under stress conditions in wheat and other cereals. The information generated will help construct network models of abiotic stress responses that will facilitate computational predictions and direct future experimental work like the development of models such as the "Metabolic pathways of the diseased potato" [51] or MapMan for the analysis of gene expression data in *Arabidopsis* [52]. The results could facilitate the understanding of cellular mechanisms involving groups of gene products that act in coordination in response to environmental stimuli.

Methods

A total of eleven different cDNA libraries were prepared from hexaploid wheat (*Triticum aestivum*) for the FGAS EST sequencing project and are summarized in Table 1.

Cultivar Norstar was used for Libraries 2 to 6 to represent various tissues, developmental stages and stress conditions. Six subtracted cDNA libraries (suppression subtractive hybridization; SSH), named TaLT2 to TaLT7, were also prepared from two different wheat lines (CI14106 and PI178383) and cv Norstar as a complementary approach to isolate differentially expressed transcripts. The "Library 1" and TaLT1 libraries were not used for the large scale EST sequencing FGAS project since the former was not prepared in a Gateway-compatible vector and the latter was generated to optimize the SSH protocol.

Preparation of the cDNA libraries

Growth conditions

For Libraries 2 and 3, the seeds were germinated in water-saturated vermiculite for 7 days at 20°C and 70% relative humidity under an irradiance of 200 μmol m^{-2} sec^{-1} and a 15-hr photoperiod. At the end of this period, the aerial parts (crowns and leaves) and roots of control plants were sampled and individually frozen. Cold acclimation was performed by subjecting germinated seedlings to a temperature of 4°C with a 12-hr photoperiod for 1, 23 and 53 days under an irradiance of 200 μmol m^{-2} sec^{-1}. Seedlings were watered with a nutrient solution (0.5 g/l 20:20:20; N:P:K). Salt stress was induced by watering with the nutrient solution containing 200 mM NaCl for 0.5, 3 and 6-hr. Aerial parts of cold-acclimated plants were sampled for Library 2 and roots of both cold-acclimated and salt-stressed plants were sampled for Library 3.

For Library 4, two different water stress conditions were used. For bench experiments, seeds were germinated for 7 days as described for Library 2. At the end of this period, plants were removed from vermiculite and left at room temperature on the table without water for 1, 2, 3 and 4 days before sampling. For growth chamber experiments, seeds were germinated in a water-saturated potting mix (50% black earth and 50% ProMix) for 7 days under an irradiance of 200 μmol m^{-2} sec^{-1}. The temperature was maintained at 20°C with a 15-hr photoperiod under a relative humidity of 70%. After this period, watering of plants was stopped. Four time points were sampled during a two weeks period; the first after wilting was observed and the last, two weeks later, and consisted of living crown and stem tissues (leaf tissue was yellow and thus not included in the sampled material).

For Library 5, seeds were germinated for 7 days and cold-treated for 49 days (full vernalization) as described for Library 2. Seedlings were then potted in water-saturated potting mix and transferred to flower inducing conditions (20°C and a 15-hr photoperiod). Tissues were sampled as follows: 1 cm crown sections after 30 days of cold treatment; 1 cm vernalized (49-day cold-treated) crown sections that were exposed to flower inducing conditions for 11 days; different developmental stages of spike formation (5 to 50 mm); and different developmental stages of spike and seed formation after the spikes had emerged from the flag leaf (visible).

For Library 6, seeds were germinated for 7 days and cold-treated as described for Library 2, except that cold treatments were performed for short time points (1, 3 and 6 hr) in the light or in the dark. Crown sections (1 cm) and green leaf tissues were harvested individually for each time point and for both exposure conditions.

For SSH libraries TaLT2 to TaLT7, plants were germinated as described for Library 2 except that the light intensity was 275 μm m^{-2} s^{-1} and the cold treatment was performed at 2°C for 1, 21 or 49 days. Crown sections (1 cm) were harvested individually for each time point.

RNA purification and cDNA synthesis

For Libraries 2 and 3, total RNA was isolated using the phenol method [53] except that the heating step at 60°C was omitted, whereas the TRI Reagent method (Sigma) was used for Libraries 4 to 6 and TRIzol (Life Technologies) was used for the TaLT libraries. For Libraries 2 to 6, poly(A)$^+$ RNA was purified from the total RNA samples using two cycles of an oligo(dT)-cellulose affinity batch-enrichment procedure [53] whereas PolyA Pure (Ambion) was used for the TaLT libraries. Total RNAs were subsequently used for cDNA synthesis. For all libraries, cDNA synthesis was initiated with a *Not*I primer-adaptor ($GCGGCCGCCCT_{15}$) using the 'SuperScript™ Plasmid System with Gateway Technology for cDNA Synthesis and Cloning' kit (Invitrogen). For Libraries 3 to 6, methylated dCTP was added to the first strand reaction mix to prevent cleavage by the NotI restriction enzyme used for directional cloning. For Library 6, the 'GeneRacer' kit (Invitrogen) was used prior to first strand synthesis to dephosphorylate truncated and non-mRNAs, remove the 5' cap structure from intact mRNA, and ligate the gene racer RNA oligo 5'-CGACUGGAGCACGAGGACACUGACAUGGACUGAAGGAGUAGAAA-3'. The precipitation steps in the kit were replaced by the RNeasy Mini Protocol for RNA Cleanup (QIAGEN). For this library, the second strand cDNA was synthesized using *Pfx* DNA polymerase (Invitrogen) and the primer 5'-CGACTGGAGCACGAGGACACTGA-3' homologous to the RNA oligo. The 'SuperScript™ Plasmid System with Gateway Technology for cDNA Synthesis and Cloning' kit (Invitrogen) was used for the remaining steps of the construction of Libraries 2 to 6 except that the precipitation steps without yeast carrier tRNA were replaced by the QIAquick PCR purification procedure (QIAGEN). For the TaLT2, 3, 6 and 7 libraries, the Nitro-pyrrole anchored oligo-dT priming technique was used [54]. For TaLT4 and TaLT5 libraries, the SMART cDNA (Clontech) priming kit was used.

Table 6: Annotated contigs that are unique to the TaLT libraries (SSH).

Contig name	Annotation	Number of ESTs	Contigs with similar annotation containing ESTs over-represented in FGAS
CL1246Contig2	Putative high-affinity potassium transporter	29	
CL1122Contig2	Putative phosphoribosylanthranilate transferase	27	7-fold 7e-53 CL10525Contig1
CL1701Contig1	Potential phospholipid-translocating ATPase	23	
CL1961Contig1	Transcriptional factor B3-like	20	
CL1506Contig2	DHHC-type zinc finger domain-containing protein-like	19	
CL2126Contig1	Putative ACT domain-containing protein	19	
CL622Contig3	50S ribosomal protein L22-like	19	
CL2193Contig1	Putative DEAD/DEAH box RNA helicase protein	17	
CL1038Contig2	Pollen-specific calmodulin-binding protein	16	
CL3163Contig1	ATP synthase protein 9, mitochondrial precursor (EC 3.6.3.14) (Lipid-binding protein)	12	
CL3186Contig1	Putative pollen specific protein (Putative ascorbate oxidase)	12	
CL1986Contig1	Putative dCK/dGK-like deoxyribonucleoside kinase	10	
CL3856Contig1	Protein kinase domain	10	
CL2813Contig3	MKIAA0124 protein (Fragment)	9	
CL4654Contig1	Hypothetical protein OSJNBa0088I06.19	8	
CL4703Contig1	40S ribosomal protein S7	8	
CL4937Contig1	Glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12)	8	
CL1038Contig3	Hypothetical protein AT4g28600	7	
CL4812Contig1	Homeobox transcription factor-like	7	
CL4821Contig1	Agglutinin isolectin 3 precursor (WGA3) (Fragment)	7	
CL4846Contig1	Putative aldo/keto reductase family protein	7	
CL10Contig35	Ribosomal protein L10A	6	
CL4Contig25	Phytochrome B (Fragment)	6	
CL5821Contig1	Putative very-long-chain fatty acid condensing enzyme CUT1	6	7-fold 2e-57 CL5480Contig1
CL5833Contig1	Putative UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase-I	6	
CL6515Contig1	NBS-LRR disease resistance protein homologue	6	
CL823Contig3	Putative RNA splicing protein	6	
CL1392Contig2	Heat shock factor-binding protein 1	5	
CL4432Contig2	Putative chromomethylase	5	
CL5300Contig2	Hypothetical protein	5	
CL6924Contig1	Beta-expansin (Fragment)	5	7-fold 8e-48 CL235Contig6
CL6960Contig1	Hypothetical protein OSJNBb0027B08.22 (Hypothetical protein OSJNBa0078D06.5)	5	
CL7305Contig1	Agglutinin (CCA)	5	
CL7698Contig1	Putative resistance gene analog PIC27	5	
CL1101Contig4	Putative amino acid transporter	4	
CL1531Contig2	Putative ZIP-like zinc transporter	4	
CL1739Contig3	Putative ethylene-responsive small GTP-binding protein	4	
CL18Contig7	Putative ribosomal protein L5	4	
CL2037Contig3	Protoporphyrin IX Mg-chelatase subunit precursor	4	
CL2221Contig1	Putative Ribosome recycling factor, chloroplast	4	
CL2305Contig1	Eukaryotic translation initiation factor 3 subunit 12 (eIF-3 p25) (eIF3k)	4	
CL3669Contig2	Putative ascorbate oxidase promoter-binding protein AOBP	4	
CL36Contig7	Adenosylhomocysteinase-like protein	4	
CL3840Contig2	Putative aminopropyl transferase	4	
CL6158Contig2	Cytochrome C6, chloroplast-like protein	4	
CL7225Contig1	P0076O17.10 protein	4	
CL732Contig2	OSJNBa0070C17.10 protein	4	
CL7697Contig1	Heat shock factor protein hsf8-like	4	
CL8407Contig1	Aldo/keto reductase family-like protein	4	11-fold 3e-54 CL3996Contig1
CL9543Contig1	Anthranilate N-benzoyltransferase-like protein (AT5g01210/F7J8_190)	4	
CL10751Contig1	Histone H4-like protein	3	7-fold 6e-46 CL9Contig66
CL10863Contig1	Methionine S-methyltransferase (EC 2.1.1.12) (AdoMet:Met S-methyltransferase)	3	
CL11049Contig1	Transferase family	3	
CL12283Contig1	Putative PPR-repeat containing protein	3	

Table 6: Annotated contigs that are unique to the TaLT libraries (SSH). *(Continued)*

CL12337Contig1	U3 small nucleolar RNA-associated protein 14 (U3 snoRNA-associated protein 14)	3	
CL12711Contig1	Putative lipase/acylhydrolase (Putative anther-specific proline-rich protein)	3	
CL1347Contig2	Omega-3 fatty acid desaturase	3	
CL1402Contig2	Putative VIP2 protein	3	
CL1688Contig3	Putative plastid ribosomal protein L11	3	
CL1Contig342	Protein H2A	3	15-fold 8e-74 CL1Contig113
CL1Contig350	Protein H2A	3	15-fold 2e-47 CL1Contig113
CL1Contig361	60S ribosomal protein L17-1	3	
CL2045Contig1	Cap-binding protein CBP20	3	
CL2470Contig2	Putative inorganic pyrophosphatase	3	7-fold 1e-75 CL2470Contig1
CL2890Contig3	Mak3 protein-like protein	3	7-fold 4e-91 CL2890Contig1
CL3033Contig2	Putative serine/threonine phosphatase	3	
CL3124Contig2	Putative ATP phosphoribosyl transferase	3	
CL4048Contig2	Boron transporter	3	
CL4808Contig2	Putative DNA topoisomerase II	3	
CL617Contig3	Putative calreticulin	3	5-fold 9e-152 CL617Contig1
CL7904Contig1	Hypothetical protein OSJNBb0004M10.19	3	
CL9749Contig1	Putative subtilisin-like proteinase	3	9-fold 3e-20 CL5317Contig1
CL9993Contig1	Hypothetical protein At1g78915	3	
CL4836Contig2	MtN3-like	2	

Suppression Subtractive Hybridization

For the TaLT libraries, SSH was performed on the RNAs isolated from crowns. For the TaLT2 library, RNA from CI14106 cold-acclimated for 1 day was used as tester RNA and subtracted by SSH against the driver RNA from cv Norstar cold-acclimated for 21 and 49 days (equal amounts of cDNAs were pooled together before subtraction). For TaLT3, 21 and 49-day cold-acclimated CI14106 was subtracted against cv Norstar cold-acclimated for 1 day. For TaLT4, 1 day cold-acclimated PI178383 was subtracted against 21 and 49 days cold-acclimated cv Norstar. For TaLT5, 21 and 49 days cold-acclimated PI178383 was subtracted against 1 day cold-acclimated Norstar. For TaLT6, 1 day cold-acclimated CI14106 was subtracted against non-acclimated CI14106. For TaLT7, 21 and 49 days cold-acclimated CI14106 was subtracted against non-acclimated CI14106.

Cloning into vectors

For Libraries 2 to 6, a *Sal*I adaptor (GTCGAC-CCACGCGTCCG) was ligated to the 5' end of the cDNAs synthesized with the *Not*I primer-adaptor to allow for directional cloning. The first two (for Libraries 3 to 5) or five (for Libraries 2 and 6) fractions eluting from size fractionation column chromatography and containing cDNAs larger than 0.5 kb were pooled for ligation with the vector. About 15 ng of *Sal*I-*Not*I-digested cDNAs was ligated with 50 ng of the pCMV.SPORT6 vector, which contains the attB1 and attB2 site-specific recombination sites flanking the multiple cloning sites. Therefore, clones isolated from these libraries can be rapidly transferred into Gateway™ destination vectors using site-specific recombination (Invitrogen). The libraries were then transformed into ElectroMAX™ DH10B cells (Invitrogen) for Library 2 or ElectroTen-Blue™ cells (Stratagene) for Libraries 3 to 6. For TaLT libraries, the PCR-amplified products of SSH were non-directionally cloned into the pGEM-T vector and transformed into DH5α cells.

Assessment of library quality and selection of clones for sequencing

Around 6.0×10^6 primary clones were obtained for Libraries 2 to 6. To determine the average cDNA size, 96 clones were randomly chosen from different libraries and the plasmids digested and characterized on agarose gels. Average insert sizes were estimated at 1300 bp (Library 2: 14% of inserts below 750 bp, 59% between 750 and 1500 bp, and 27% above 1500 bp), 1560 bp (Library 3: 10% below 750 bp, 44% between 750 and 1500 bp, and 46% above 1500 bp), and 1100 bp (Library 6: 17% below 750 bp, 68% between 750 and 1500 bp, and 15% above 1500 bp). Since all libraries contain an average of 6 million different clones, this collection represents an important resource to isolate full length clones for which only truncated cDNAs are available. To reduce the number of ESTs representing highly expressed genes, Libraries 2 to 6 were hybridized to ^{32}P-labelled cDNAs from non-acclimated plants. Colonies showing with the weakest hybridization signals were picked for sequencing.

Bioinformatics

Trimming high quality sequences

Sequence tracefiles were obtained from the FGAS project (110,544 ESTs) and from the NSF (82,332 ESTs; [55]) and DuPont (154,171 ESTs) collections. The latter two collections comprise EST sequences derived from many cDNA libraries prepared from various wheat RNA sources. All sequences were processed as follows. Quality score

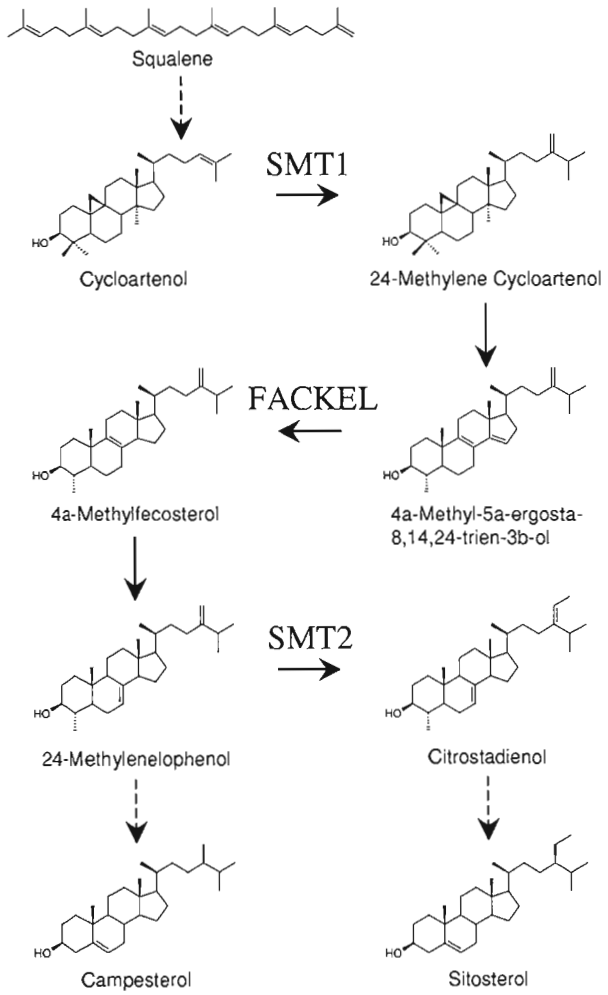


Figure 3

Plant sterols pathway. ESTs encoding several enzymes of the sterol pathways are over-represented in the FGAS dataset. Three enzymes are involved in the production of squalene from which cycloarthenol is obtained. The FACKLE and SMT2 enzymes are involved in the production of sitosterol with a concomitant decrease in campesterol.

sequences were obtained from tracefiles using PHRED [56,57]. Only sequences with mean Q≥20 were retained. Poly(A) or poly(T) regions with length = 14 (± 2 errors) were trimmed and all sequences containing more than one poly(A) and/or poly(T) sequences were flagged as putative chimeras. SeqClean3 with generic Univec DB as well as Lucy4 (using pCMV.SPORT6 and pBlueScript II splice sites) were used with the default settings in an iterative manner. This recursive approach proved more efficient in removing vector and linker sequences, and low quality regions than using either one only once. All resulting high quality sequences were then re-checked for low-complexity and all sequences containing more that 50% repeats were rejected. A repeat was defined as a minimum word size of 4 identical bases with a maximum of 1 error. RepeatMasker2 was used with Repeat DB to mask regions that could eventually bias the assembly. All information pertaining to library details, sequences and data quality scores were stored in a mySQL database. After filtering, 269,562 cleaned ESTs were retained for assembly (73,521 ESTs from FGAS, 68,886 ESTs from NSF and 127,155 ESTs from DuPont).

Clustering, assembly and annotation

Clustering was performed to reduce the redundancy of the dataset and increase the overall quality of the derived consensus sequences. When a small set of sequences (FGAS 73,521 quality-filtered sequences) was used, the clustering performed well through TGICL and d2_cluster. However, when the NSF and DuPont data (196,041 sequences) were added, aberrant large clusters were obtained. This is presumably due to undetected chimeras, multi-domain proteins and the transitive closure technique applied by these applications. These large clusters (38 k sequences for TGICL and 25 k for d2_cluster) contained many unrelated sequences and were difficult to assemble, yielding many incongruent and low quality contigs. To avoid such artifacts, a cluster breaking strategy was used. First, all sequences that could be contained in other ESTs were removed, thereby reducing the dataset to parent sequences. These sequences were then BLASTed against themselves and results were parsed to extract the e-values in order to build an adjacency matrix. The distance (d) between the sequences was calculated based on the level of similarity established using BLAST e-value where d = 100/-log (e-value). Two parent sequences were considered to be part of the same cluster when the BLASTN identity result between them was greater than or equal to 96%. GRAPH9 was used to flag bridges (articulation points where the removal of an EST breaks the link between subclusters) and manually split the large graph into distinct smaller sub-graphs. Other suspicious clusters that were not automatically detected were manually investigated and split when required (Figure 4a). Child ESTs, removed in the first stage were then incorporated into the cluster containing the parent sequence. For example, the largest cluster was broken down using the approach described above and yielded 250 sub-clusters, with the largest being of 6 k sequences (Figure 4b). TGICL and d2_cluster results were compared using randomly chosen clusters that were re-assembled using either clustering tools. It was observed that TGICL had a higher tendency of joining similar genes and falsely splitting sequences from the same gene, thus indicating that d2_cluster was a more reliable clustering tool in our case.

Both CAP3 [58] and PHRAP were tested to assemble the sequences. CAP3 was used on TGICL results using the set-

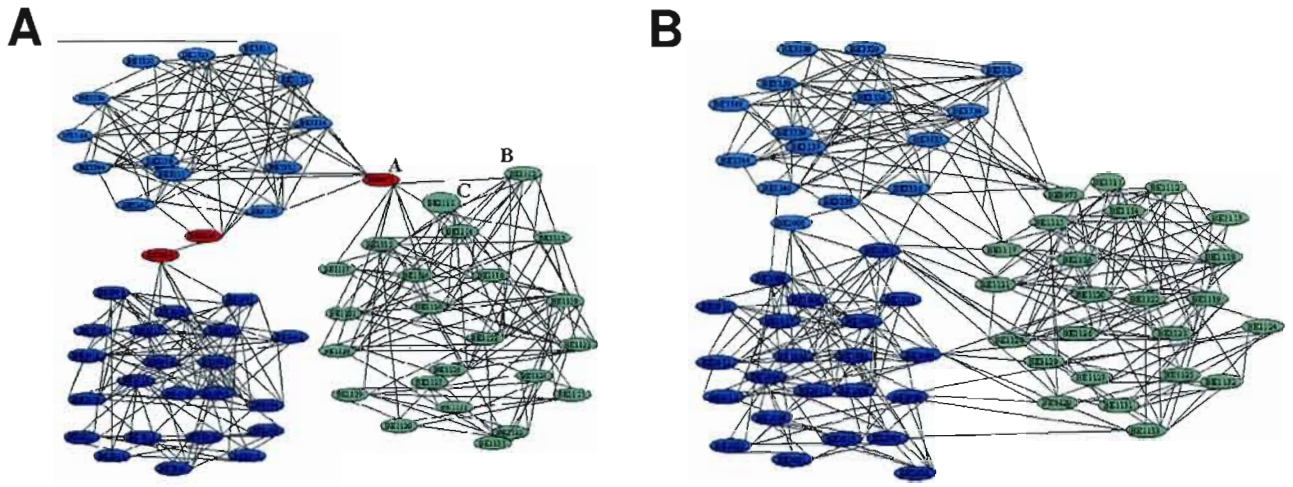


Figure 4
Breaking strategy of large clusters. A breaking strategy was used to reduce the size of large clusters. Each sequence in a cluster was BLASTed against the others and e-values were used to build an adjacency matrix (see Materials and Methods). For example, an e-100 value will result in a distance of 1 cm between two sequences. Only values below e-25 were used for graphical display. GRAPH9 was used to flag bridges (articulation points where an EST links two potential sub-clusters) and manually split a cluster into distinct sub-clusters. **A)** Example of a cluster region where specific ESTs (in red) can be manually transferred to sub-clusters (based on the smallest e-value). **B)** Example of a cluster region that could not be broken into sub-clusters due to the complex interrelations between ESTs.

tings that appeared satisfactory when assembling barley EST sequences [7] while PHRAP was used to assemble d2_cluster results using the default parameters. The first method generated ~32 k contigs while the latter produced over 50 k contigs. The first approach gave results more consistent with the Unigene and TIGR Wheat Gene Index assembly data with respect to contig number, suggesting that PHRAP was less appropriate for assembly of the large dataset used in this study. The total number of singletons and singlets in both cases was similar; 39 k for PHRAP (14% of all ESTs) vs. 42 k for CAP3 (15.5% of all ESTs) and the percentage was close to that found in TIGR (13.3% of all ESTs). Singletons are defined as unique sequences that could not be assembled in a cluster whereas singlets are unique sequences that were assembled in a cluster but could not be assembled in a contig. Based on the TGICL and d2_cluster comparison and on the number of contigs obtained with CAP3 and PHRAP, we chose d2_cluster and CAP3 as the clustering and assembly tools for this project.

We used different annotation tools to increase the number of annotated sequences. The unique assembled sequences produced in our study were annotated after translation using prot4EST and then BLASTed (BLASTX) against a GO-annotated database. All the sequences that did not show sufficient similarity to be functionally classified with this method were investigated with AutoFact where sequences are BLASTed against other complementary databases (ex. PFAM, KEGG, Ribosomal Sequences database) having GO details.

Digital expression analysis

The relative abundance (digital expression) of FGAS ESTs was analysed as follows: 1) among the contigs containing EST sequences present in both the FGAS dataset and NSF-DuPont dataset, abundance was expressed as a ratio of FGAS ESTs (without SSH ESTs) to NSF-DuPont ESTs, after correction for the size (total number of ESTs) in each dataset; 2) contigs that contained only FGAS ESTs were analyzed separately; 3) SSH EST abundance was compared between similar SSH libraries to determine if common ESTs can be identified; and 4) unique SSH contigs were identified as these could represent new genes expressed during cold acclimation.

Identification of homologous genes regulated by stress in Arabidopsis

The 2,637 putative wheat stress-regulated genes identified in our study were BLASTed (TBLASTX) against the *Arabidopsis* proteins TAIR database [12] using a cut-off e-value of e-25. The Protein ID of the homologous *Arabidopsis* proteins were used to identify those that are represented on the Affymetrix ATH1 genome array and the MWG Bio-

Table 7: Transcription factors that are differentially expressed in the FGAS dataset relative to the NSF-DuPont dataset.

Transcription factor family	over-represented 2 to 5-fold	over-represented over 5-fold	Contigs unique to FGAS with at least 3 ESTs	TOTAL
AP2 (ex. CBF1,2,3, Aintegumenta)	4	7	9	20
BHLH (Ex. AtMYC2)	5	2	0	7
BZIP (Ex. FD)	5	3	0	8
CCAAT-box transcription factor	2	1	0	3
DEAD/DEAH box helicase	4	4	0	8
F-box protein family	3	0	0	3
FLOWERING LOCUS T	1	0	1	2
GIGANTEA protein	0	1	1	2
Homeodomain Leucine zipper protein (Ex. ABF3 ABF4, ABA response)	2	2	1	5
MADS box transcription factor (Ex. TaVRT1)	2	0	0	2
MYB (Ex. AtMYB2)	14	7	2	23
NAC-domain containing protein (Ex. RD26 dehydration)	11	8	0	19
PHD finger (Ex. pollen development, chromatin-mediated transcription regulation, a variant of Zn-finger)	2	1	0	3
RING finger containing protein (Ex. HOS1 regulating cold response, A variant of Zn finger)	14	4	3	21
SCARECROW gene regulator-like (Ex. Oxidative stress)	3	1	0	4
WD-repeat containing protein	0	1	0	1
WRKY transcription factor (Ex. Drought, oxidative stress and pathogen induced)	14	7	7	28
Zinc finger protein (Ex. CO, Indeterminate-related)	30	11	6	47
Other Transcription factor-like	113	47	14	174
Other DNA-binding protein	143	46	11	200
Total	372	153	55	580

tech 25 k 50-mer oligonucleotide array. The cold- and drought-regulated genes were then identified from the available published data [14,15].

Authors' contributions

MH, FS, PG, AL and WLC conceived the study and participated in its design and coordination. MH carried out the analyses of the EST datasets and drafted the manuscript. MH, MB and AB carried out the bioinformatics analyses. FO and FS participated in the drafting and editing of the manuscript. JD constructed Libraries 2 to 6. AM, AD and PG prepared the clones from Libraries 2 to 6 for sequencing. AL constructed libraries TaLT2 to TaLT6 and prepared the clones for sequencing. ML, LMcC and WLC carried out the sequencing reactions, the bioinformatics analyses of the FGAS dataset, and submitted the data to Genbank. All authors read and approved the final manuscript.

Additional material

Additional File 1

Contigs containing ESTs that are over- or under-represented at least two-fold in the FGAS dataset compared to the NSF/DuPont dataset. SSH ESTs are not part of this analysis. The contigs containing ESTs over-represented at least 5-fold in FGAS were analyzed by TBLASTX against the Arabidopsis *TAIR database to find homologues (e-25 cut-off). For those that are represented on the Affymetrix and/or MGW microarrays, the expression data with respect to cold or drought regulation was obtained. U, up-regulated; D, down-regulated.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-7-149-S1.xls]

Additional File 2

Contigs containing at least three ESTs that are present only in the FGAS dataset. SSH ESTs are not part of this analysis. The contigs were analyzed by TBLASTX against the Arabidopsis *TAIR database to find homologues (e-25 cut-off). For those that are represented on the Affymetrix and/or MGW microarrays, the expression data with respect to cold or drought regulation was obtained. U, up-regulated; D, down-regulated.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-7-149-S2.xls]

Additional File 3

Contigs containing at least three ESTs that are present only in the TaLT libraries of the FGAS dataset.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-149-S3.xls]

Acknowledgements

This work was funded by Genome Canada (MH, PG, AL, WLC, FS), Genome Prairie (AL, WLC), Génome Québec (MH, PG, FS) and Canarie (MH, FS). We thank the technical staff and students who participated in this study.

References

1. Sarhan F, Ouellet F, Vazquez-Tello A: **The wheat wcs120 gene family. A useful model to understand the molecular genetics of freezing tolerance in cereals.** *Physiol Plant* 1997, **101**:439-445.
2. Initiative TAG: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.** *Nature* 2000, **408**:796-815.
3. Project IRGS: **The map-based sequence of the rice genome.** *Nature* 2005, **436**:793-800.
4. Bennett MD, Leitch IJ: **Nuclear DNA amounts in angiosperms.** *Ann Bot* 1995, **73**:113-176.
5. Sasaki T: **Rice genome analysis: Understanding the genetic secrets of the rice plant.** *Breed Sci* 2003, **53**:281-289.
6. Bräutigam M, Lindlöf A, Zakhrabekova S, Gharti-Chhetri G, Olsson B, Olsson O: **Generation and analysis of 9792 EST sequences from cold acclimated oat, Avena sativa.** *BMC Plant Biol* 2005, **5**:18.
7. Close TJ, Wanamaker SI, Caldo RA, Turner SM, Ashlock DA, Dickerson JA, Wing RA, Muehlbauer GJ, Kleinhofs A, Wise RP: **A new resource for cereal genomics: 22K barley GeneChip comes of age.** *Plant Physiol* 2004, **134**:960-968.
8. Fei Z, Tang X, Alba RM, White JA, Ronning CM, Martin GB, Tanksley SD, Giovannoni JJ: **Comprehensive EST analysis of tomato and comparative genomics of fruit ripening.** *Plant J* 2004, **40**:47-59.
9. Sterky F, Bhalerao RR, Unneberg P, Segerman B, Nilsson P, Brunner AM, Charbonnel-Campaa L, Lindvall JJ, Tandre K, Strauss SH, Sundberg B, Gustafsson P, Uhlen M, Bhalerao RP, Nilsson O, Sandberg G, Karlsson J, Lundeberg J, Jansson S: **A Populus EST resource for plant functional genomics.** *Proc Natl Acad Sci USA* 2004, **101**:13951-13956.
10. Ogihara Y, Mochida K, Kawaura K, Murai K, Seki M, Kamiya A, Shinozaki K, Carninci P, Hayashizaki Y, Shin I, Kohara Y, Yamazaki Y: **Construction of a full-length cDNA library from young spikelets of hexaploid wheat and its characterization by large-scale sequencing of expressed sequence tags.** *Genes Genet Syst* 2004, **79**:227-232.
11. Zhang D, Choi DW, Wanamaker S, Fenton RD, Chin A, Malatrasi M, Turuspekov Y, Walia H, Akhunov ED, Kianian P, Otto C, Simons K, Deal KR, Echenique V, Stamova B, Ross K, Butler GE, Strader L, Verhey SD, Johnson R, Altenbach S, Kothari K, Tanaka C, Shah MM, Laudencia-Chingcuanco D, Han P, Miller RE, Crossman CC, Chao S, Lazo GR, Klueva N, Gustafson JP, Kianian SF, Dubcovsky J, Walker-Simmons MK, Gill KS, Dvorak J, Anderson OD, Sorrells ME, McGuire PE, Qualset CO, Nguyen HT, Close TJ: **Construction and evaluation of cDNA libraries for large-scale expressed sequence tag sequencing in wheat (Triticum aestivum L.).** *Genetics* 2004, **168**:595-608.
12. **Functional Genomics of Abiotic Stress (FGAS).** 2006.
13. Lazo GR, Chao S, Hummel DD, Edwards H, Crossman CC, Lui N, Matthews DE, Carollo VL, Hane DL, You FM, Butler GE, Miller RE, Close TJ, Peng JH, Lapitan NL, Gustafson JP, Qi LL, Echalier B, Gill BS, Dilbirligi M, Randhawa HS, Gill KS, Greene RA, Sorrells ME, Akhunov ED, Dvorak J, Linkiewicz AM, Dubcovsky J, Hossain KG, Kalavacharla V, Kianian SF, Mahmoud AA, Miftahudin, Ma XF, Conley EJ, Anderson JA, Pathan MS, Nguyen HT, McGuire PE, Qualset CO, Anderson OD: **Development of an expressed sequence tag (EST) resource for wheat (Triticum aestivum L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-locus bin-delineated map.** *Genetics* 2004, **168**:585-593.
14. Hannah MA, Heyer AG, Hincha DK: **A global survey of gene regulation during cold acclimation in Arabidopsis thaliana.** *PLoS Genet* 2005, **1**:e26.
15. Rizhsky L, Liang H, Shuman J, Shulaev V, Davletova S, Mittler R: **When defense pathways collide. The response of Arabidopsis to a combination of drought and heat stress.** *Plant Physiol* 2004, **134**:1683-1696.
16. Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y, Kamiya A, Nakajima M, Enju A, Sakurai T, Satou M, Akiyama K, Taji T, Yamaguchi-Shinozaki K, Carninci P, Kawai J, Hayashizaki Y, Shinozaki K: **Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray.** *Plant J* 2002, **31**:279-292.
17. Xin Z, Browse J: **Eskimo1 mutants of Arabidopsis are constitutively freezing-tolerant.** *Proc Natl Acad Sci USA* 1998, **95**:7799-7804.
18. Breitkreuz KE, Allan WL, Van Cauwenberghe OR, Jakobs C, Talibi D, Andre B, Shelp BJ: **A novel gamma-hydroxybutyrate dehydrogenase: identification and expression of an Arabidopsis cDNA and potential role under oxygen deficiency.** *J Biol Chem* 2003, **278**:41552-41556.
19. Sirko A, Blaszczyk A, Liszewska F: **Overproduction of SAT and/or OASTL in transgenic plants: a survey of effects.** *J Exp Bot* 2004, **55**:1881-1888.
20. Uemura M, Steponkus PL: **A contrast of the plasma membrane lipid composition of oat and rye leaves in relation to freezing tolerance.** *Plant Physiol* 1994, **104**:479-496.
21. Holmberg N, Harker M, Gibbard CL, Wallace AD, Clayton JC, Rawlins S, Hellyer A, Safford R: **Sterol C-24 methyltransferase type 1 controls the flux of carbon into sterol biosynthesis in tobacco seed.** *Plant Physiol* 2002, **130**:303-311.
22. Shen H, Dowhan W: **Regulation of phospholipid biosynthetic enzymes by the level of CDP-diacylglycerol synthase activity.** *J Biol Chem* 1997, **272**:11215-11220.
23. Charron JBF, Breton G, Danyluk J, Muzac I, Ibrahim RK, Sarhan F: **Molecular and biochemical characterization of a cold-regulated phosphoethanolamine N-methyltransferase from wheat.** *Plant Physiol* 2002, **129**:363-373.
24. Takahashi M, Yamaguchi H, Nakanishi H, Shioiri T, Nishizawa NK, Mori S: **Cloning two genes for nicotianamine aminotransferase, a critical enzyme in iron acquisition (Strategy II) in graminaceous plants.** *Plant Physiol* 1999, **121**:947-956.
25. Higuchi K, Suzuki K, Nakanishi H, Yamaguchi H, Nishizawa NK, Mori S: **Cloning of nicotianamine synthase genes, novel genes involved in the biosynthesis of phytosiderophores.** *Plant Physiol* 1999, **119**:471-479.
26. Allard F, Houde M, Kröl M, Ivanov A, Huner NPA, Sarhan F: **Betaine improves freezing tolerance in wheat.** *Plant Cell Physiol* 1998, **39**:1194-1202.
27. Hincha DK, Oliver AE, Crowe JH: **Lipid composition determines the effects of arbutin on the stability of membranes.** *Biophys J* 1999, **77**:2024-2034.
28. Breton G, Danyluk J, Ouellet F, Sarhan F: **Biotechnological applications of plant freezing associated proteins.** *Biotechnol Annu Rev* 2000, **6**:59-101.
29. Knight CA, DeVries AL, Oolman LD: **Fish antifreeze protein and the freezing and recrystallization of ice.** *Nature* 1984, **308**:295-296.
30. Gaudet DA, Laroche A, Frick M, Davoren J, Puchalski B, Ergon : **Expression of plant defence-related (PR-protein) transcripts during hardening and dehardening of winter wheat.** *Physiol Mol Plant Pathol* 2000, **57**:15-24.
31. Yeh S, Moffatt BA, Griffith M, Xiong F, Yang DS, Wiseman SB, Sarhan F, Danyluk J, Xue YQ, Hew CL, Doherty-Kirby A, Lajoie G: **Chitinase genes responsive to cold encode antifreeze proteins in winter cereals.** *Plant Physiol* 2000, **124**:1251-1264.
32. Tremblay K, Ouellet F, Fournier J, Danyluk J, Sarhan F: **Molecular characterization and origin of novel bipartite cold-regulated ice recrystallization inhibition proteins from cereals.** *Plant Cell Physiol* 2005, **46**:884-891.
33. Hincha DK, Meins Jr. F, Schmitt JM: **ß-1,3-glucanase is cryoprotective in vitro and is accumulated in leaves during cold acclimation.** *Plant Physiol* 1997, **114**:1077-1083.

34. Öquist G, Huner NP: **Photosynthesis of overwintering evergreen plants.** *Annu Rev Plant Biol* 2003, **54**:329-355.
35. Jansen MA, Mattoo AK, Edelman M: **D1-D2 protein degradation in the chloroplast. Complex light saturation kinetics.** *Eur J Biochem* 1999, **260**:527-532.
36. Ishihara S, Yamamoto Y, Ifuku K, Sato F: **Functional analysis of four members of the PsbP family in photosystem II in Nicotiana tabacum using differential RNA interference.** *Plant Cell Physiol* 2005, **46**:1885-1893.
37. Ifuku K, Yamamoto Y, Ono TA, Ishihara S, Sato F: **PsbP protein, but not PsbQ protein, is essential for the regulation and stabilization of photosystem II in higher plants.** *Plant Physiol* 2005, **139**:1175-1184.
38. Makino A, Sakashita H, Hidema J, Mae T, Ojima K, Osmond B: **Distinctive responses of ribulose-1,5-bisphosphate carboxylase and carbonic anhydrase in wheat leaves to nitrogen nutrition and their possible relationships to CO2 transfer resistance.** *Plant Physiol* 1992, **100**:1737-1743.
39. Moreno JI, Martin R, Castresana C: **Arabidopsis SHMT1, a serine hydroxymethyltransferase that functions in the photorespiratory pathway influences resistance to biotic and abiotic stress.** *Plant J* 2005, **41**:451-463.
40. Scheibe R, Backhausen JE, Emmerlich V, Holtgrefe S: **Strategies to maintain redox homeostasis during photosynthesis under changing conditions.** *J Exp Bot* 2005, **56**:1481-1489.
41. Schöttler MA, Kirchhoff H, Weis E: **The role of plastocyanin in the adjustment of the photosynthetic electron transport to the carbon metabolism in tobacco.** *Plant Physiol* 2004, **136**:4265-4274.
42. N'Dong C, Danyluk J, Huner NP, Sarhan F: **Survey of gene expression in winter rye during changes in growth temperature, irradiance or excitation pressure.** *Plant Mol Biol* 2001, **45**:691-703.
43. Varotto C, Pesaresi P, Meurer J, Oelmuller R, Steiner-Lange S, Salamini F, Leister D: **Disruption of the Arabidopsis photosystem I gene psaE1 affects photosynthesis and impairs growth.** *Plant J* 2000, **22**:115-124.
44. Haldrup A, Naver H, Scheller HV: **The interaction between plastocyanin and photosystem I is inefficient in transgenic Arabidopsis plants lacking the PSI-N subunit of photosystem I.** *Plant J* 1999, **17**:689-698.
45. Scheibe R: **Malate valves to balance cellular energy supply.** *Physiol Plant* 2004, **120**:21-26.
46. Lee BH, Lee H, Xiong L, Zhu JK: **A mitochondrial complex I defect impairs cold-regulated nuclear gene expression.** *Plant Cell* 2002, **14**:1235-1251.
47. Gilmour SJ, Fowler SG, Thomashow MF: **Arabidopsis transcriptional activators CBF1, CBF2, and CBF3 have matching functional activities.** *Plant Mol Biol* 2004, **54**:767-781.
48. Chinnusamy V, Ohta M, Kanrar S, Lee BH, Hong X, Agarwal M, Zhu JK: **ICE1: a regulator of cold-induced transcriptome and freezing tolerance in Arabidopsis.** *Genes Dev* 2003, **17**:1043-1054.
49. Danyluk J, Kane NA, Breton G, Limin AE, Fowler DB, Sarhan F: **TaVRT-1, a putative transcription factor associated with vegetative to reproductive transition in cereals.** *Plant Physiol* 2003, **132**:1849-1860.
50. Kane NA, Danyluk J, Tardif G, Ouellet F, Laliberté JF, Limin AE, Fowler DB, Sarhan F: **TaVRT-2, a member of the StMADS-11 clade of flowering repressors, is regulated by vernalization and photoperiod in wheat.** *Plant Physiol* 2005, **138**:2354-2363.
51. **Metabolic pathways of the diseased potato** 2006 [http://www.scri.sari.ac.uk/TiPP/pps/Chart.pdf].
52. **MapMan** 2006 [http://gabi.rzpd.de/projects/MapMan/].
53. Danyluk J, Sarhan F: **Differential mRNA transcription during the induction of freezing tolerance in spring and winter wheat.** *Plant Cell Physiol* 1990, **31**:609-619.
54. Guo Z, Liu Q, Smith LM: **Enhanced discrimination of single nucleotide polymorphisms by artificial mismatch hybridization.** *Nat Biotechnol* 1997, **15**:331-335.
55. **Index of /NSF/curator/quality** 2006 [http://wheat.pw.usda.gov/nsf/curator/quality].
56. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8**:186-194.
57. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8**:175-185.
58. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9**:868-877.
59. Breton G, Danyluk J, Charron JB, Sarhan F: **Expression profiling and bioinformatic analyses of a novel stress-regulated multispanning transmembrane protein family from cereals and Arabidopsis.** *Plant Physiol* 2003, **132**:64-74.
60. Shimosaka E, Sasanuma T, Handa H: **A wheat cold-regulated cDNA encoding an early light-inducible protein (ELIP): its structure, expression and chromosomal location.** *Plant Cell Physiol* 1999, **40**:319-325.
61. Castillo J, Rodrigo MI, Márquez JA, Zúñiga , Franco L: **A pea nuclear protein that is induced by dehydration belongs to the vicilin superfamily.** *Eur J Biochem* 2000, **267**:2156-2165.
62. Cattivelli L, Bartels D: **Molecular cloning and characterization of cold-regulated genes in barley.** *Plant Physiol* 1990, **93**:1504-1510.
63. Tsuda K, Tsvetanov S, Takumi S, Mori N, Atanassov A, Nakamura C: **New members of a cold-responsive group-3 Lea/Rab-related Cor gene family from common wheat (Triticum aestivum L.).** *Genes Genet Syst* 2000, **75**:179-188.
64. Koike M, Okamoto T, Tsuda S, Imai R: **A novel plant defensin-like gene of winter wheat is specifically induced during cold acclimation.** *Biochem Biophys Res Commun* 2002, **298**:46-53.
65. Houde M, Danyluk J, Laliberté JF, Rassart E, Dhindsa RS, Sarhan F: **Cloning, characterization and expression of a cDNA encoding a 50-kilodalton protein specifically induced by cold acclimation in wheat.** *Plant Physiol* 1992, **99**:1381-1387.
66. Shih MD, Lin SC, Hsieh JS, Tsou CH, Chow TY, Lin TP, Hsing YI: **Gene cloning and characterization of a soybean (Glycine max L.) LEA protein, GmPM16.** *Plant Mol Biol* 2004, **56**:689-703.
67. Livingston III DP, Henson CA: **Apoplastic sugars, fructans, fructan exohydrolase, and invertase in winter oat: responses to second-phase cold hardening.** *Plant Physiol* 1998, **116**:403-408.
68. Chauvin LP, Houde M, Sarhan F: **A leaf-specific gene stimulated by light during wheat acclimation to low temperature.** *Plant Mol Biol* 1993, **23**:255-265.
69. Gong Z, Dong CH, Lee H, Zhu J, Xiong L, Gong D, Stevenson B, Zhu JK: **A DEAD box RNA helicase is essential for mRNA export and important for development and stress responses in Arabidopsis.** *Plant Cell* 2005, **17**:256-267.
70. De Santis A, Landi P, Genchi G: **Changes of mitochondrial properties in maize seedlings associated with selection for germination at low temperature. Fatty acid composition, cytochrome c oxidase, and adenine nucleotide translocase activities.** *Plant Physiol* 1999, **119**:743-754.
71. Phillips JR, Dunn MA, Hughes MA: **mRNA stability and localisation of the low-temperature-responsive barley gene family blt14.** *Plant Mol Biol* 1997, **33**:1013-1023.
72. Massonneau A, Condamine P, Wisniewski JP, Zivy M, Rogowsky PM: **Maize cystatins respond to developmental cues, cold stress and drought.** *Biochim Biophys Acta* 2005, **1729**:186-199.
73. Kim KN, Lee JS, Han H, Choi SA, Go SJ, Yoon IS: **Isolation and characterization of a novel rice Ca2+-regulated protein kinase gene involved in responses to diverse signals including cold, light, cytokinins, sugars and salts.** *Plant Mol Biol* 2003, **52**:1191-1202.
74. Kowalski LR, Kondo K, Inouye M: **Cold-shock induction of a family of TIP1-related proteins associated with the membrane in Saccharomyces cerevisiae.** *Mol Microbiol* 1995, **15**:341-353.
75. N'Dong C, Anzellotti D, Ibrahim RK, Huner NP, Sarhan F: **Daphnetin methylation by a novel O-methyltransferase is associated with cold acclimation and photosystem II excitation pressure in rye.** *J Biol Chem* 2003, **278**:6854-6861.
76. N'Dong C, Danyluk J, Wilson KE, Pocock T, Huner NP, Sarhan F: **Cold-regulated cereal chloroplast late embryogenesis abundant-like proteins. Molecular characterization and functional analyses.** *Plant Physiol* 2002, **129**:1368-1381.
77. Gao YP, Young L, Bonham-Smith P, Gusta LV: **Characterization and expression of plasma and tonoplast membrane aquaporins in primed seed of Brassica napus during germination under stress conditions.** *Plant Mol Biol* 1999, **40**:635-644.
78. Liu JH, Luo M, Cheng KJ, Mohapatra SS, Hill RD: **Identification and characterization of a novel barley gene that is ABA-inducible**

and expressed specifically in embryo and aleurone. *J Exp Bot* 1999, **50:**727-728.
79. Zhao TY, Martin D, Meeley RB, Downie B: **Expression of the maize galactinol synthase gene family: II) Kernel abscission, environmental stress and myo-inositol influences transcript accumulation in developing seeds and callus.** *Physiol Plant* 2004, **121:**647-655.
80. Campalans A, Pages M, Messeguer R: **Identification of differentially expressed genes by the cDNA-AFLP technique during dehydration of almond (Prunus amygdalus).** *Tree Physiol* 2001, **21:**633-643.
81. Kume S, Kobayashi F, Ishibashi M, Ohno R, Nakamura C, Takumi S: **Differential and coordinated expression of Cbf and Cor/Lea genes during long-term cold acclimation in two wheat cultivars showing distinct levels of freezing tolerance.** *Genes Genet Syst* 2005, **80:**185-197.
82. Salekdeh GH, Siopongco J, Wade LJ, Ghareyazie B, Bennett J: **Proteomic analysis of rice leaves during drought stress and recovery.** *Proteomics* 2002, **2:**1131-1145.
83. White AJ, Dunn MA, Brown K, Hughes MA: **Comparative analysis of genomic sequence and expression of a lipid transfer protein gene family in winter barley.** *J Exp Bot* 1994, **45:**1885-1892.
84. Potter E, Beator J, Kloppstech K: **The expression of mRNAs for light-stress proteins in barley: inverse relationship of mRNA levels of individual genes within the leaf gradient.** *Planta* 1996, **199:**314-320.
85. Barth O, Zschiesche W, Siersleben S, Humbeck K: **Isolation of a novel barley cDNA encoding a nuclear protein involved in stress response and leaf senescence.** *Physiol Plant* 2004, **121:**282-293.