UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

BIOINFORMATIC STUDIES ON STRUCTURAL ELEMENTS FOR THE REGULATION OF ALTERNATIVE OXIDASE (AOX) GENE ACTIVITIES

Moisés Geraldes Xavier

Confidential

Mestrado em Engenharia Informática

2007

UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

BIOINFORMATIC STUDIES ON STRUCTURAL ELEMENTS FOR THE REGULATION OF ALTERNATIVE OXIDASE (AOX) GENE ACTIVITIES

Moisés Geraldes Xavier

Projecto orientado pelo Prof. Dr. Francisco José Moreira Couto
e supervisionado pela Prof. Dra. Birgit Arnholdt-Schmitt

Mestrado em Engenharia Informática

2007

Declaração

Moisés Geraldes Xavier, aluno nº 25533 da Faculdade de Ciências da Universidade de Lisboa, declara ceder os seus direitos de cópia sobre o seu Relatório de Projecto em Engenharia Informática, intitulado "Bionformatic studies on structural elements for the regulation of Alternative Oxidase (AOX) gene Activities", realizado no ano lectivo de 2007 à Faculdade de Ciências da Universidade de Lisboa para o efeito de arquivo e consulta nas suas bibliotecas e publicação do mesmo em formato electrónico na Internet.

FCUL, de de 2007

_____

Birgit Arnholdt-Schmitt, supervisor do projecto de Moisés Geraldes Xavier, aluno da Faculdade de Ciências da Universidade de Lisboa, declara concordar com a divulgação do Relatório do Projecto em Engenharia Informática, intitulado "Bionformatic studies on structural elements for the regulation of Alternative Oxidase (AOX) gene Activities".

Évora, de de 2007

_____

Abstract

Alternative Oxidase genes encode a small family of isoenzymes (enzymes with some differences but act in the same chemical reaction). AOX is present in plants, fungi, algae, some yeast, and was also found in some classes of the animal kingdom. The enzymes are responsible for an alternative pathway of respiration that is responsive to stress conditions but also to pathogen attack, as well as growth and stage development.

Scaffold Matrix Attachment Regions (S/MARs) are DNA sequences from 300 to 3000 nucleotides that bound with nuclear proteins serving as anchors for DNA, influencing in this way the DNA organization inside the cell. Several studies have failed to reveal a pattern of organization in the sequences, however some rules have been found that help computer based analysis. Experimental identification of these sequences is hard and time consuming, computer methods could provide a first step selection, and cover larger sequences.

In order to highlight possible links between S/MARs and differential regulation of AOX genes, the first part of this project consists in identifying structurally relevant S/MAR regions in the neighborhood of AOX genes in *Arabidopsis thaliana* and in rice using a selected computer program.

Single Nucleotide Polymorphisms (SNPs) are variations in one nucleotide base among DNA sequences from the same location, from different individuals. These differences could serve as markers to classify a specific set of individuals.

The second part of this project consists in the development of a bioinformatic application that will help in the identification of specific polymorphisms (SNPs) in sequences that are experimentally obtained at the EU Marie Curie Chair in ICAM University of Évora, where this project is performed.

Keywords

Bioinformatic, Scaffold Matrix Attachment Regions, Single Nucleotide Polymorphisms, Alternative Oxidase

Resumo

Os genes da oxidase alternativa (ou AOX) codificam uma pequena família de isoenzimas (enzimas com algumas diferenças mas que actuam nas mesmas reacções químicas), que se encontram nas plantas, fungos, algas, algumas leveduras bem como em algumas classes do reino animal. A AOX é responsável por uma via alternativa de respiração, activada principalmente em condições de stress mas também como reacção a ataques patogénicos, bem como em estádios específicos do desenvolvimento da planta.

As "Scaffold Matrix Attachment Regions" (S/MARs) são sequências de DNA entre 300 e 3000 nucleótidos que se ligam a proteínas do núcleo da célula, servindo como âncoras para o DNA, conferindo-lhe assim uma forma própria no interior da célula. Estudos realizados para determinar uma organização específica destas regiões não produziram muitos resultados, no entanto foram definidas algumas regras que permitem ajudar na detecção computacional destas sequências, uma vez que a detecção experimental é difícil e morosa.

Com vista a estabelecer possíveis relações entre uma regulação diferenciada dos genes da AOX através dos S/MARs, a primeira parte deste projecto consiste em determinar as regiões do DNA com a estrutura de potenciais "S/MARs" na vizinhança dos genes da Oxidase Alternativa na *Arabidopsis thaliana* e no arroz.

"Single Nucleotide Polymorphisms" (SNPs) são diferenças de um nucleótido entre as mesmas regiões de DNA de diferentes indivíduos da mesma espécie. Estas diferenças podem servir para marcar um determinado conjunto de indivíduos.

A segunda parte deste projecto consiste em desenvolver uma aplicação para ajudar na identificação de tipos específicos de polimorfismos, (SNPs) em sequências identificadas na EU Marie Curie Chair, ICAM, Universidade de Évora, onde este projecto foi desenvolvido.

Resumo Alargado

A aplicação da Informática no campo de estudo de outras ciências, revela-se fundamental na concretização de projectos de dimensões consideráveis, que envolvem a gestão e análise de grandes quantidades de dados. A utilização de técnicas informáticas para extrair informação dos dados obtidos, simplifica consideravelmente a dimensão e consequente resolução dos problemas. No contexto dos problemas de Biologia Molecular, a Bioinformática assume um papel preponderante, fornecendo ferramentas e metodologias extremamente úteis.

A Bioinformática estuda informação biológica recorrendo a conceitos de Biologia, Genética, Informática, Estatística, Bioquímica, entre outras, sendo frequentemente dividida em duas áreas, gestão de informação biológica e biologia computacional. A sua aplicação prática estende-se por áreas tais como análise de sequências de DNA, descoberta e anotação de genes, análise evolutiva entre várias espécies, análise das semelhanças e diferenças entre os genomas das várias espécies, regulação de genes, interacções entre proteínas, análise da estrutura das proteínas, etc.

Este projecto integra-se na Cátedra Marie Curie do Instituto de Ciências Agrárias Mediterrânicas (ICAM) da Universidade de Évora inserindo-se no contexto de um projecto mais amplo; "*Stress adaptation in plants – A molecular approach of social-economic interest*", relacionado com a procura de um marcador molecular que espera fornecer novas perspectivas para o melhoramento de plantas.

Neste projecto pretende-se estudar as possíveis implicações de certas sequências na regulação dos genes da Oxidase Alternativa (AOX), bem como desenvolver um aplicação que ajude na identificação de diferenças entre os genes da Oxidase Alternativa de modo a esta ser aceite como um marcador funcional para uma reprogramação eficiente das células em condições de stress [5] [6].

A oxidase alternativa (AOX) existe nas plantas, fungos, algas, algumas leveduras, bem como em algumas classes do reino animal, nomeadamente nos mollusca e nematoda [5]. Esta enzima localiza-se na membrana interna da mitocôndria, sendo responsável por uma via alternativa da cadeia respiratória. Esta via quando activa reduz a quantidade de energia biológica (ATP) e a quantidade de espécies reactivas de oxigénio, ao contrário do que se verifica na via da cadeia respiratória. As espécies reactivas de oxigénio quando presentes em grandes quantidades são tóxicas para as células. A AOX está associada à

produção de calor, à resposta a infecções virais bem como na prevenção da morte celular programada [40] [5] [43].

A AOX é codificada por uma família de genes dividida em duas subfamílias de genes, a *AOX1* e a *AOX2*. Alguns genes da subfamília *AOX1* são expressos preferencialmente sob condições de stress (p.e. ausência de água, excesso de calor, etc.) enquanto que, a expressão dos genes da subfamília da *AOX2* está associada com os estádios específicos de desenvolvimento da planta [27]. Na *Aradidopsis thaliana* (*Thale cress*) foram identificados quatro genes da subfamília *AOX1* (*AOX1a*, *AOX1b*, *AOX1c* e *AOX1d*) e um gene da subfamília *AOX2* (*AOX2*) e no arroz (*Oryza sativa*) apenas quatro genes da subfamília *AOX1* (*AOX1a*, *AOX1b*, *AOX1c* e *AOX1d*) [3] [4].

O DNA de um determinado organismo contém toda a sua informação genética, que determina as suas características fenotípicas. É constituído por uma dupla cadeia de nucleótidos enrolados em dupla hélice. Os nucleótidos são as unidades estruturais do DNA, e são caracterizados por quatro "bases" Adenina (A), Guanina (G), Citosina (C) e Timina (T). Esta dupla cadeia é complementar ou seja sempre que existe uma Adenina numa cadeia existe uma Timina na outra e vice-versa, sendo o mesmo verdade para a Guanina e Citosina.

Para o DNA dar origem a uma proteína, primeiro uma molécula de RNA, complementar ao DNA é sintetizada, sendo depois produzida a proteína a partir da molécula de RNA sintetizada.

Neste projecto pretende-se estudar a localização de certas regiões denominadas *Scaffold Matrix Attachment Regions* (S/MARS) na vizinhança dos genes da AOX na *Arabidopsis thaliana* e no arroz, bem como desenvolver um programa para ajuda na identificação de diferenças, por exemplo *Single Nucleotide Polymorphisms* (SNPs) nos genes da AOX em diferentes espécies.

As *Scaffold Matrix Attachment Regions* (S/MARs) são regiões do DNA que têm entre 300 e 3000 pares de bases aproximadamente, ligando-se a proteínas do núcleo da célula, funcionando como pontos de ancoragem para o DNA, conferindo-lhe uma estrutura própria e compartimentalizando o DNA em regiões funcionais.

A identificação destas regiões experimentalmente é feita através de processos de associação e desassociação dos S/MARs com as proteínas do núcleo recorrendo a procedimentos não triviais e morosos. O recurso a uma identificação destas regiões através

de ferramentas informáticas proporciona um primeiro passo no sentido da redução do número de candidatos, bem como a pesquisa em sequências de grandes dimensões.

Estas regiões não possuem uma estrutura padrão característica, o que dificulta a sua identificação. São no entanto conhecidas pela sua elevada quantidade de Adeninas (A) e Timinas (T), bem como duas subsequências específicas com aproximadamente 200 bases pares de distância. Alguns estudos tentaram caracterizar estas sequências usando modelos probabilísticos, no entanto devido ao pequeno número de sequências existentes nas bases de dados genómicas, muito continua ainda por descobrir em relação aos S/MARs.

A primeira parte deste projecto consiste na identificação destas sequências na vizinhança dos genes da AOX na *Arabidopsis thaliana* e no arroz. Estas duas plantas têm o genoma completamente sequenciado e disponível em bases de dados genómicas. A abordagem adoptada passou em primeiro por seleccionar um programa para a identificação de S/MARs, os vários programas existentes usam diferentes critérios de análise e consequentemente os resultados são diferentes. Seguidamente determinou-se as sequências de DNA para análise na Arabidopsis e no arroz, em primeiro localizaram-se os genes da AOX nas duas espécies, e posteriormente determinou-se o comprimento das sequências para análise em cada uma das espécies. Finalmente as sequências foram analisadas e foi desenvolvido um programa para gerar imagens esquemáticas com os resultados obtidos.

Diferentes bases de dados armazenam e apresentam a informação de diferentes maneiras. Para utilizar esta informação de uma forma estandardizada e útil foi necessário criar algumas pequenas aplicações para converter os dados entre os diferentes formatos. A linguagem de programação usada foi C++, visto possibilitar uma fácil integração com bibliotecas para geração rápida de imagens bem como possuir um variado leque de funções para manipulação de texto e Input/Output, além disso outro aspecto tido em conta foi a portabilidade de código entre Linux e Windows.

No decorrer do projecto foi também considerado importante obter dados sobre a expressão dos genes da AOX na Arabidopsis e no arroz, recorrendo para isso a publicações e a bases de dados acessíveis publicamente, com informações sobre a expressão dos genes (a uma escala genómica), em diversos tecidos da planta, bem como em diversas condições.

Se um gene é transcrito, e posteriormente codificado numa proteína diz-se que o gene é expresso. Com o avanço dos métodos para medição da expressão dos genes, tornou-

se possível efectuar esta medição a uma escala genómica, armazenar os dados numa base de dados para posterior consulta.

A segunda parte deste projecto consiste no desenvolvimento de uma aplicação para a ajuda na identificação de SNPs, em sequências de DNA obtidas no laboratório da Cátedra EU Marie Curie no ICAM.

Single Nucleotide Polymorphisms são variações numa base de DNA nas mesmas regiões genómicas entre vários indivíduos. Estas variações podem influenciar ou não a codificação de proteínas, e consequentemente a sua estrutura. Caso estas variações sejam facilmente detectadas a nível do gene, ou a nível do fenótipo (características visíveis dos genes), podem servir como marcadores para caracterizar um determinado conjunto de indivíduos (o conjunto de indivíduos com essa variação). Neste projecto pretende-se criar uma ferramenta que auxilie a detecção de SNPs em sequências de DNA da Oxidase Alternativa obtidas no laboratório da EU Marie Curie Chair.

Existem várias metodologias e ferramentas para a identificação computacional de SNPs. O primeiro passo consiste em alinhar as sequências de DNA em estudo, procurando as regiões comuns entre elas. A partir deste alinhamento existem várias aproximações para encontrar SNPs. A primeira normalmente consiste em perceber se as diferenças são importantes tendo em conta a qualidade das sequências (diferenças em sequências de fraca qualidade não são indicadores de SNPs), outro factor importante é a quantidade de sequências em estudo, (o mesmo padrão de diferenças em muitas sequências é indicador de SNPs), o número de cada um das bases na sequência, bem como a quantidade de diferenças em cada uma delas podem também ser factores de decisão.

A abordagem efectuada passou em primeiro por identificar os programas que permitissem efectuar uma alinhamento de múltiplas sequências, de seguida seleccionar um algoritmo (já implementado) para a pesquisa de SNPs e criar uma ferramenta que permita ajudar na automatização deste processo, onde através de uma interface gráfica o utilizador selecciona as sequências que pretende analisar, bem como os parâmetros de *input* para cada um dos programas. No fim é gerada uma imagem esquemática comos resultados. Este último ponto encontra-se ainda em desenvolvimento.

Palavras-chave

Bioinformática, *Scaffold Matrix Attachment Regions*, *Single Nucleotide Polymorphisms*, Oxidase Alternativa.

Table of Contents

List of Figures

1 Introduction

In recent years the availability of genomic data has experienced an exponential growth as the price to obtain DNA sequences has decreased. This leaves room for innovation applying informatics as well as mathematical and statistical methodologies, and approaches to solve biological problems, which in turn generates even more data.

Parallel with this growth is the necessity to develop tools and approaches to acquire, store, organize, archive, analyze, or visualize such data. This field of science is commonly known as Bioinformatics, and it could be divided in two major categories: biological information management and computational biology.

Bioinformatics today comprises as major research topics sequence analysis, genome annotation, analysis of gene expression and regulation, computational evolutionary biology, analysis of protein expression and prediction of protein structure among others. In a modern approach Bioinformatics derives knowledge from computer analysis of biological data, not only limited to genetic code but also from population statistical data and scientific literature.

In the latest years several projects took place to obtain the complete genome of some organisms. *Arabidopsis thaliana* (Thale cress) and *Oryza sativa* (Rice) was the first plants to have their genome completely sequenced [1] [2] in 2000 and 2005 respectively, a new set of projects at the genome scale could then take place.

Access to whole genome data provides also information to small scale projects aimed at studying a smaller subset of already known or newly identified genes, since the analysis of this sequences could provide new insights in gene regulation and function across different species.

The Alternative oxidase is an enzyme encoded by a family of genes which is responsible for an alternative pathway of respiration called “non-phosphorylating bypasses”. This enzyme is present in plants, fungi, some yeast, algae, and in animal kingdom, was found so far in phyla (mollusca, nematoda and chordata). Up to date the enzyme has not been identified in humans. This alternative pathway of respiration could be linked to many actions during plants life, which will be depicted in more detail later.

Once the whole genome sequencing projects for Arabidopsis and rice were finished the AOX genes and neighbor regions could be analyzed in these species by bioinformatics tools in search for regions that may have structural and regulatory functions.

One example of these regions is the Scaffold / Matrix Attachment Regions (S/MARs). S/MARs could in a very simple way be defined as DNA regions that are bound to non-histone proteins in the nucleus. When bounded to the nuclear matrix S/MARs serve as anchors, organizing and compartmentalizing the DNA in functional domains.

In order to find possible relations between a differential regulation of AOX genes and the S/MARs regions near these genes, in the first part of this study, a search will be performed, to determine the informatic tool to analyze DNA sequences in the neighborhood of AOX genes, searching for S/MARs. The results obtained will then be summarized in images and / or tables, to better characterize the occurrence that may point to their biological significance.

Another interesting aspect is the differences found in the DNA from different organisms in the same species. The process of understanding these differences involves applying biological computer algorithms to the DNA under study. If computer algorithms present similarities and differences between DNA's a more extensive analysis could be initiated trying to understand their importance and further use.

Single Nucleotide Polymorphisms (SNPs) are variations in one nucleotide between DNA from different organisms of the same specie. SNPs could serve to easily mark differences between different sets of organisms. Integration of all these data is expected to help in the context of a larger project (Stress adaptation in plants – A molecular approach of social-economic interest) at the EU-Marie Curie Chair at the University of Évora, related to understanding the role of AOX genes as functional marker for efficient cell reprogramming and intended to contribute directly to molecular breeding on efficient plant production [5], [6].

The second part of this project aims at creating a program to identify SNPs in DNA sequences (mainly from carrot and olive but possibly from other species) analyzed in the laboratory of the EU-Marie Curie Chair.

This report is organized in six chapters. Chapter 2 introduces basic concepts necessary for understanding the developed work. Chapter 3 describes the objectives of this work, the adopted methodology for each phase of the work, and the adjustments made to

the original planning. Chapter 4 describes the work performed in each phase according with the adopted methodology. Chapter 5 contains a summary about what was done in this project, as well as possible points of improvement. Chapter 6 lists the references used in this work. Appendix 1 contains the results regarding S/MARs identification for *Arabidopsis thaliana* and *Oryza sativa* in the neighborhood of AOX genes and Appendix 2 lists gene expression for AOX genes in *Oryza sativa* found in publicly available databases.

1.1 Objectives

The project is divided in two major parts, first, the identification of S/MARs in the neighborhood of AOX genes in *Arabidopsis thaliana* and rice, and seconds the development of an application to identify SNPs in AOX gene sequences of different species. This first part of the project will introduce several concepts of Molecular Biology and Bioinformatics. Understanding the biology concepts behind each phase is very important since it will facilitate the development of a correct methodology to solve it.

There are several objectives for this phase namely:

- **Understand the concept of S/MARs.**

  What they are, their importance in DNA organization, their structure, influence on gene expression, etc. This will introduce at the same time concepts of Molecular Biology that by themselves need to be understood. More advanced topics will be introduced as the work progresses.

- **Identify the previous work in computational S/MAR discovery.**

  Previous work on, computational S/MAR discovery will show the way how to progress, and will reveal the limitations and possibilities of this approach

- **S/MARs analysis**

  Locating the AOX genes in Arabidopsis and rice, and determine the sequences to be analyzed will introduce the contact with genomic databases, and with literature in order to confirm the correctness of AOX genes location and to have a deeper understanding of AOX gene structure.

- **Integration of results**

  Results must be verified, and presented in an intuitive way in order to facilitate their analysis and presentation.

- **Find evidence of AOX gene expression in Arabidopsis an rice**

  Locate databases and literature with evidence of gene expression for AOX genes in Arabidopsis and rice. This information will be correlated with S/MARs identification results in order to discover possible insights of S/MARs in AOX gene regulation.

The second part of the project will focus on the development of an application to help in SNPs identification for sequences of AOX genes in different species.

- **Understand the concept of SNPs**

  Like S/MARs understanding the concept of SNPs will introduce a new set of topics from Genetics that need to be assimilated, in order to understand better what is intended to do in the next phases.

- **Identify previous work on computational SNP discovery**

  Perform a survey regarding existent SNP scoring algorithms, as well as literature regarding the use of these algorithms.

- **Select a SNP scoring algorithm**

  Select the scoring algorithm to use in this work. This decision must be supported with previous work.

- **Implement an application customized according the necessities of the workgroup**

  After the selection of the algorithm and analysis of previews approaches on SNP computational discovery, the application should be implemented, in order to facilitate the customization and integration of the scoring algorithm to the specific needs in EU Marie Curie Chair, the application should also facilitate the presentation of the results.

- **Test the application and use it to find AOX regulatory sequences in Arabidopsis and rice**

  The application need to be tested, among the workgroup, to ensure accurate results, and usability, users should interact easily with the application.

1.2 Planning

In the beginning of this project, a planning comprising all phases of this project has been elaborated. Objective analysis at the end of each phase and adjustments to the initial planning were adopted and proved useful to establish and tune the methodology. However, this initial planning needed to be adjusted for several reasons justified below.

The first delay is related with the beginning of the project, due to external reasons concerned with the later than expected termination of the graduation. Thus, the project could be started in agreement with the supervisor at the EU Marie Curie Chair only at the beginning of October.

A second unforeseeable bias for the project happened through the fact that the external supervisor of the project, Dr. Stephen Rudd, changed at the start of the project to a position in the industry and could not stick to the agreements. This problem was solved by a rapidly established risk management plan in discussion between the project leader, Birgit Arnholdt-Schmitt and Stephen Rudd. Since then, Dr. Klaus Mayer[1], who was responsibly, involved in the former research activities of Stephen Rudd, agreed to overtake the role as an external advisor for the newly emerging bioinformatics group at the Marie Curie Chair by email contacts instead of Stephen Rudd. During my project, Dr. Klaus Mayer was invited to visit the Chair and to give a talk about his activities and experiences. During that visit, I had about three days of time to discuss my project personally with him. The German scientist is now an official collaborator of the EU Marie Curie Chair.

A third bias concerns discrepancies found when comparing S/MARs identification results, with results already available from other publications [9]. The reason of these discrepancies was caused by differences in the data management for the Arabidopsis genome in different data banks (the 4th release of the Arabidopsis genome available in TIGR, and the 2nd version of the Arabidopsis genome used in [9] See also the discussion at page 37 in the present report). This problem was not promptly identified. After identification, Dr. Klaus Mayer made internal data (concerning new S/MAR annotation based on the most recent data) available that helped to solve the problem rapidly.

[1] http://mips.gsf.de/cgi-bin/webapp/ibi/person/getPerson.pl?Person=MayerKlaus

In the proposal, it is foreseen that the project will contribute to conference presentations and/or a publication. Also, it was foreseen that expression profile data would be used to enable a link between the data of S/MARs occurrence near AOX genes and the activity of these genes. Because of the above mentioned biases, analyses of the expression data, which are most important for any presentation or publication of the project's tasks, has been included before starting with the third and forth task related to SNP analyses. Since analyses of the expression data will be central for the scientific success of the project, I decided to ask for a prolongation for delivering this report after Dr. Francisco Couto (my supervisor at the Department of Informatics) have said it was possible, depending on the justification. Thus, the work related to SNPs started only after the 31th of May. Though successfully treated, the task is more challenging than expected and will still go on even after I will have submitted the project report. Generation of schematic images from SNP analysis text files is in progress, the test phase needs still to be done. Only after these two phases, the programme for SNP analysis in AOX genes currently identified at the Chair can be started to run. Work of the project related to S/MARS and expression data will now be submitted for presentation to a highly competitive conference that takes place in October in the United Kingdom.

2 Basic Concepts

This chapter introduces a set of concepts from molecular biology necessary to understand the problems that will be addressed by informatics tools or approaches in this project. Concepts are presented in a simple way, but as more accurate as possible.

2.1 Basic concepts on DNA

DNA (Deoxyribonucleic acid) contains all the genetic information for the development and functioning of living organisms. In eukaryotes such as animals and plants, DNA is stored inside the cell nucleus (genomic DNA), packed in structures called chromosomes. In prokaryotes, such as bacteria, DNA is in the cell's cytoplasm. DNA could also be found in other cellular organelles such as mitochondria and plastids.

The three-dimensional structure of DNA consists of two long helical strands that are coiled around a common axis forming a double helix. Nucleotides are the structural units of DNA; they are characterized by four bases Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). These nucleotides are complementary, wherever there is an A in one strand there is a T in the other, the same is true for the G and C bases. In RNA the T base is replaced by the U (Uracil) [7].

The genome of an organism comprises its entire DNA (the genotype). In order to synthesize a protein the DNA is first transcribed to RNA (Ribonucleic acid) who migrates to the cytoplasm, the RNA is then translated to a protein in the ribosome (a complex molecular machine composed of both RNA and protein).

Once the DNA is sequenced, the sequence corresponding to one strand of DNA is obtained, this information could be stored as a long string of A's, G's, C's and T's, and in this way be available for further computer analysis. It is only required to know the sequence for one of the strands, since the other strand is complementary (an A in one strand matches a T in the other, the same is true for G and C).

2.2 Gene transcription and gene expression

The discrete functional unit of the DNA that contains the information responsible for the creation of a protein is called gene. For a gene to be transcribed a set of proteins called transcription-factors assemble at a specific DNA site, called promoter that is usually

located (but not necessarily) before the gene. Only one strand of DNA is used for transcription therefore it is called the template strand. An enzyme called RNA polymerase, binds to the transcription-factors thereby initiating transcription.

RNA polymerase melts two strand DNA bonds, separating the two strands in the near region, the direction in which a template DNA strand is transcribed is named downstream. RNA polymerase advances downstream, a complimentary copy of the DNA template strand is produced at the rate of about 1000 nucleotides per minute. We say it is a complementary strand in the sense that Guanine nucleotide bases (G) are replaced for Cytosine (C) ribonucleotide bases and vice-versa, Adenine bases (A) are replaced with Uracil (U). Specific sequences in the template DNA signal the RNA polymerase to terminate transcription. The synthesized RNA molecule is released, and RNA polymerase dissociates from DNA.

In eukaryotic cells, the RNA molecule formed must undergo several transformations before being translated to a protein. There are two types of regions, the coding regions called exons, and non-coding regions called introns. After transcription, introns are removed, thus only the exon regions codify a protein. The RNA with the introns and exons is called a precursor RNA (pre-mRNA), while the RNA only with the exons is called messenger RNA (mRNA).The mRNA traverses the nuclear membrane and is translated to a protein in the ribosomes.

A set of three nucleotides is called a codon. For each codon there is a matching amino acid. Translation begins with the start codon (AUG or Methionine) and ends with any of the stop codons (UAA, UAG and UGA). The RNA regions before the start codons and after the stop codons are called untranslated regions or UTRs. The region before start codon is called 5’ UTR and the region after the stop codon is called 3’ UTR. In this way the direction of transcription is downstream or from 5’ to 3’.

Before introns are removed, a sequence of several Adenine bases is added to the 3' UTR of the pre-mRNA molecule. This sequence is commonly known as the polyA tail and this process is called polyadenylation.

A gene may code for more than one protein, as a result of different mRNA sequences, for a variety of reasons, such as the existence of more than one promoter for each gene, different selection of sites for the polyadenylation or partial transcription of

some introns. This process is known as alternative splicing it may be responsible (among other processes) for the production of different forms of a protein, also called isoforms.

Protein synthesis takes place in the cytoplasm in a complex of RNA and protein called ribosome. As said before a set of 3 mRNA nucleotides is called a codon. For each codon there is a corresponding amino acid, however for each amino acid there is more than one corresponding codon. All amino acids bind to a key element called tRNA (Transfer RNA) which will play a very important role in protein synthesis. Each tRNA has about 73 to 93 nucleotides, and contains a three-nucleotide sequence called anticodon, that matches complementary codon in the mRNA.

Not all genes are transcribed at the same time. For example some genes are expressed in some organs and not in others and vice-versa. If the gene is transcribed to RNA and later translated to a protein it is said that the gene is "expressed" or "not expressed" otherwise.

Gene expression is controlled at transcription by signaling substances that bound to specific receptors in the cell and trigger an intercellular pathway in order to produce a specific protein. However, there are many other chemical and biochemical factors that could influence the process of transcription. The rate of the transcription could also be influenced by means of certain regulatory sequences called enhancers and silencers. Enhancers could be located near or several thousand base pairs away in both upstream and downstream directions, influencing transcription rate. Since they are located at the same strand, they are called cis-acting elements. Other regulatory elements or factors are called trans-acting elements.

A regulatory protein (usually called activator) bind to the enhancer sequence and interacts directly with the transcription complex (RNA polymerase + transcription factors + other set of proteins called coactivators) assembled at the promoter(s) site influencing in this way the transcription rate [7]. Like enhancers there are other sequences called silencers located near or overlapping enhancers, where other type of proteins called repressors bind, inhibiting the binding of activator sequences to enhancers.

2.3 Genome sequencing projects

The technology for sequencing DNA has improved largely since mid seventies when the first methods for DNA sequencing were published by Gilbert W, Maxam A, and Sanger F. Today's methods rely mostly (in a simple description) on fluorescent marking of

nucleotides, and later optical recognition due to different wavelengths associated at each nucleotide.

This evolution has led to a decrease in sequencing prices and allows more ambitious projects to take place, like the human genome project, which has sequenced the whole human genome sequence. Since today's methods can sequence only sequences with 300 to 1000 nucleotides long, sequencing the whole genome of an organism is not a simple task.

The long DNA strands must first be divided into smaller redundant random fragments. Once these smaller fragments are sequenced the longer DNA strands can be assembled by computational means using the overlapped segments of smaller sequences. This is a complex methodology because of the large amounts of information involved and the sequence errors associated with the high repetitive regions of DNA (since a large part of DNA is repetitive, very similar sequences could come from different parts of the larger sequence).

Once DNA has been sequenced, annotation of genes takes place. Possible gene sequences are first predicted using computer programs, a set of rules regarding DNA organization, gene and RNA sequences already available in databases, and through analysis of gene transcription. A locus is assigned to each gene. A locus is an identifier unique for each gene and each organism.

Matching the identified genes to proteins is the next step. This is first performed automatically (by informatics means), translating gene sequence to protein sequence and look for similarities with other characterized proteins. If a match is found a description corresponding to that protein is added to the gene (by using a gene ontology database), otherwise the gene is characterized using general terms, such as: expressed if a RNA transcript matching the gene is found in a database, hypothetical if no RNA transcript is found, putative or probable if the protein exhibit limited sequence similarity, etc.

2.4 Methods for analyzing gene expression

In the latest years the analysis of gene expression has improved very rapidly, is possible today to analyze the expression of several thousand genes simultaneously, searching for genes that are expressed together (co-expressed genes), monitoring gene expression while the cell undergoes some biological process, or search for genes that are expressed differentially in different cell tissues.

Most of these techniques even if refined or reinvented over time rely on the principle of specific probe target hybridization. The attraction of two complementary strands of DNA is so strong that even if separated zip back together at right temperature and salt conditions. This concept commonly known as hybridization is very useful for detecting one strand using the other. A probe is synthesized in order to "hybridize" with the complementary DNA or RNA strand.

The conventional essays for analyzing gene expression are Northern Blotting, and quantitative real-time PCR. Quantitative real-time PCR relies on the detection of a fluorescent signal produced proportionally during the amplification of a PCR product. A PCR or Polymerase Chain Reaction is designed to amplify a specific region of DNA, a mixture of DNA, nucleotide bases, an enzyme to synthesize DNA (Taq polymerase) and probes called primers are heated in order to separate DNA strands and cooled for the probes (primers) to bind at complementary sites. Taq polymerase extends the DNA strands in upstream (3'->5') direction [7]. This process is repeated several times or several cycles. A reaction that cycles 20 times will amplify the specific target sequence 1 million times. In Quantitative Real Time PCR the primers could be fluorescently-labeled, and the signal obtained is proportional to the amount of PCR product.

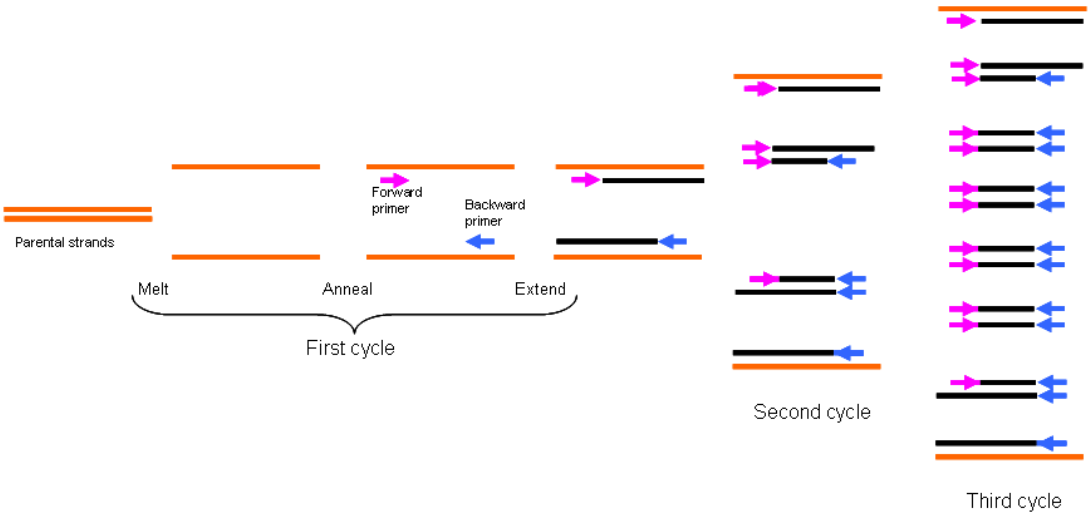


Figure 1 – In this picture are represented the first three cycles of a PCR (Polymerase Chain Reaction). Sequences are heated to be separated, primers anneal to the DNA strands and Taq polymerase enzyme begins to extend the sequences in the direction of the primers. When this procedure is repeated other primers will anneal to the original and extended sequences, (thus the name chain reaction). The number of sequences corresponding to the location amplified at the end of each cycle will be 2^n where n is the number of the cycle.

Newer methods for analyzing gene expression include SAGE (Serial Analysis of Gene Expression), DNA Microarrays and MPSS (Massively Parallel Signature Sequencing).

DNA Microarrays allows measuring the expression of thousand of genes in a single experiment. There are several companies that supply this technology, as well as academic laboratories, with several different methodologies, and different applications, but the basic principles remain true for all of them. Thousands of probe sequences ranging from 25 to 70 or more base pairs called oligonucleotides (chemically synthesized and bounded nucleotides) corresponding to a part of a particular gene or sequence of interest are attached to a surface (such as glass, plastic or silicon) in properly identified spots. Usually to evaluate gene expression, cDNA is used to from two samples that we want to compare. cDNA is the reverse transcription of mRNA, this is, the complement to any mRNA sequence with the U's replaced with T's. Thus cDNA does not have intron sequences.

The cDNAs of the two condition samples are fluorescently marked with different colors (usually Red and Green), then they are mixed and hybridized with the microarray. The spots will then have a specific intensity, corresponding to the cDNA that have hybridized the probes. The ratio of the intensities is measured for each spot. If a condition would stimulate the expression of one gene the spot would show a higher intensity of the color matching that condition. The absence of color in a spot would indicate that the gene is unexpressed in both conditions, and intermediate color intensity would mean that the gene is expressed in both conditions.

Another recent method to measure gene expression is MPSS (Massively Parallel Signature Sequencing), although not so widespread as the DNA Microarrays this is a very powerful technique to measure expression of low expressed genes. In a very simplistic description, MPSS rely on the creation of thousands of small random sequences (17 or 20 base pairs long) called signatures, the next step is matching these signatures to mRNA molecules or transcripts (results of gene transcription). Millions of signatures are created in order to have many copies of each signature. Each set of different signatures are tagged and attached to a 5 micron microbead, signatures are then hybridized with cDNA, (subjected a specific condition of study) and is possible to determine for each microbead how many signature have cDNA attached , with the final goal of quantifying the number of hybridizations for each different set of signatures. These methods and others are depicted in more detail in [18].

With the whole genome sequencing projects releasing a high amount of data in publicly available databases, a new set of projects devoted to understand gene expression in whole genome have begun. Data from microarray and mpss experiments performed under a diverse set of conditions is annotated in relational databases that are publicly available through web interfaces. More complete sets of data are available usually under a subscription fee.

2.5 Chromatin Organization

The DNA of a human cell measures about 2 meters in total length and must be contained in cells with diameters of less than 10μm [7]. It comes then at no surprise that a complex degree of compaction is required.

In eukaryotic cells, DNA is associated with histone proteins, in a highly condensed nucleoprotein called chromatin. DNA is tightly packaged into octamers known as nucleosomes, which consists of four histone proteins (H3, H4, H2A and H2B), around which is wrapped 147 bp (base pairs) of DNA. Individual nucleosomes are connected through 155 to 210 base pairs of DNA. This is the first level of chromatin compaction.

Each nucleosome is associated with one H1 histone protein, and is packed in an irregular spiral arrangement with approximately six nucleosomes per turn, folding into a 30nm chromatin fiber, this is called the second level of chromatin compaction.

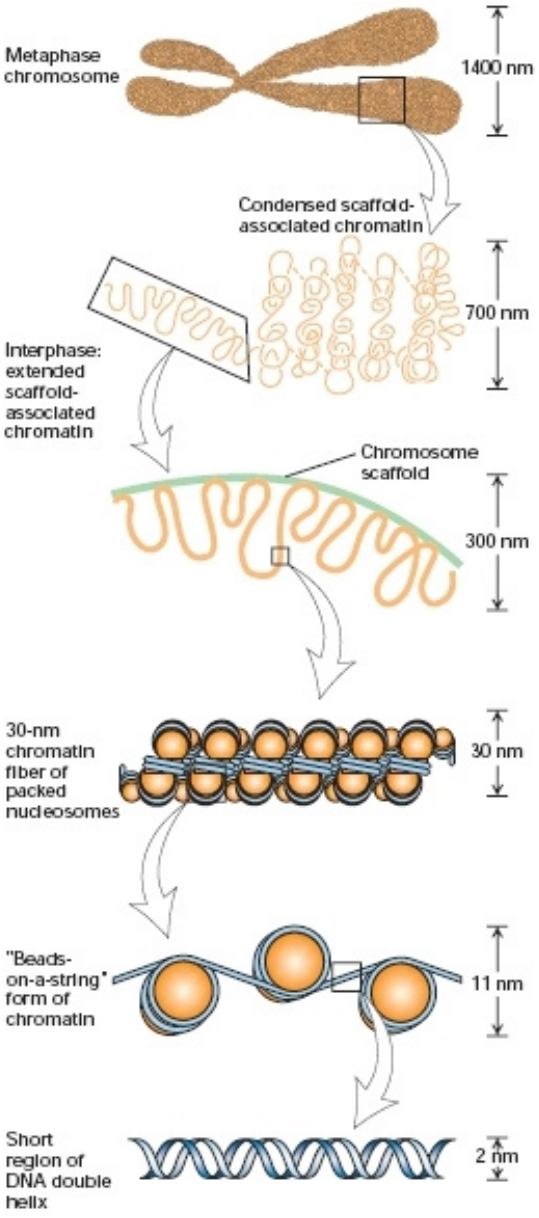


Figure 2- The figure above represents the various levels of chromatin organization, from the DNA level to the chromosome level [7]

The nucleus of the cell is not simply an unorganized container for chromatin, rather, there is a highly organized nuclear substructure comprising a fibrillar network of multiple proteins called the nuclear matrix.

This substructure retains the general shape of the cell nucleus after the removal of major part of DNA and DNA-associated proteins.

30 nm chromatin fiber attaches the nuclear matrix at specific sites called Matrix Attachment Regions or Scaffold Associated Regions (S/MARs), this is called the third level of chromatin compaction.

Besides these three levels of chromatin compaction presented there are more two levels, that are not yet completely understood, as the 30 nm chromatin fiber coils numerous times upon itself, forming fibers of 300nm, 700nm, and 1400nm.

As described before chromatin can assume a coiled from, mostly during the cell division phase called mitosis. When cells exit mitosis the condensed chromosomes uncoil, however, some chromatin regions remain coiled, these regions comprise heterochromatin. The less condensed portions of chromatin are called euchromatin. Despite some exceptions only genes in euchromatin are transcribed, since the high compaction of heterochromatin block the binding of transcription factors.

The organization of DNA in euchromatic and heterochromatic regions can be flexible depending on development, growth and interaction to the environment [6].

2.6 Scaffold Matrix Attachment Regions (S/MARs)

Scaffold Matrix Attachment Regions are 300 to 3000 nucleotides or base pairs long proposed to play an important role in nuclear and chromosome organization. Whenever associated with the nuclear scaffold the S/MARs divide the eukaryotic genomes in individual 30nm chromatin loops, being this way related with the functional compartmentalization of the genome. S/MARs were shown to control gene expression by facilitating interactions between DNA activating complexes and genes, by controlling chromatin accessibility and because of their location near the transcription units and regulatory elements [8].

Rather than only a structural function S/MARs are supposed to play also a regulatory role. S/MARS may additionally be defined as regulatory elements typically found outside the transcribed regions and within introns. They typically augment transcription rates in a

highly context-dependent manner. S/MARs containing genes have been shown to reach overall significantly lower expression levels compared with genes not associated with S/MARs [9].

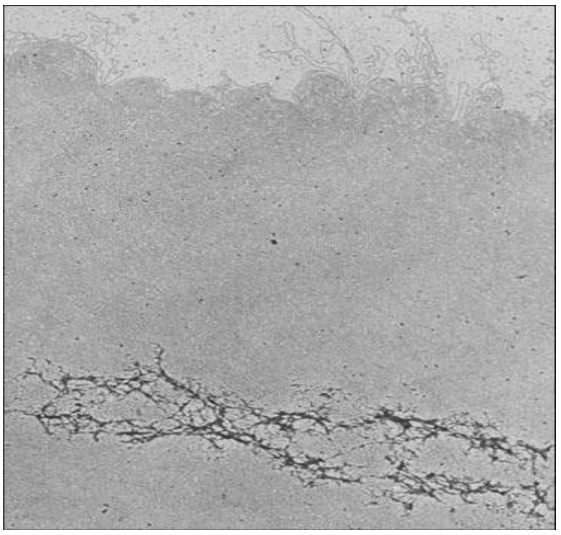


Figure 3 - After removing histone proteins, the scaffold proteins are well visible at electronic microscope, as the darker set of lines, attached to this darker lines are the DNA loops (without histones), the points of attachment between the DNA loops and scaffold proteins are the S/MARs. S/MARs have in this way a very important structural function serving as anchors for chromatin loops [7].

Transcription regulation of gene expression is known to involve formation of dynamic chromatin loops mediated by S/MAR attachment to the nuclear matrix. The attachment of a DNA sequence to the matrix will place the neighboring genes in proximity of the transcription factors. The abundance of S/MARs in the upstream regulatory regions of genes demonstrates their role in transcriptional regulation [10]. So far, this has not been demonstrated in plants, but there are several studies that suggest a correlation between the presence of S/MARs and gene expression [9].

It has been found that actively transcribed genes are associated with the nuclear matrix and preferably in the periphery of the chromosome. Inactive or unexpressed genes are found in tightly folded dense chromatin regions whereas active or expressed genes are found in more unfolded regions characterized by loop bodies [8].

In transgenic plants it has been shown that when a sequence containing a gene flanked by S/MARs is inserted in the plant genome, that gene is expressed. Furthermore genes inserted in plant genome without flanking S/MARs suffer an effect called gene silencing in which the gene is "turned off" by many factors, such as removing this gene from plant genome ("gene elimination"), the position effect (since in transgenic plants a gene may be inserted in different parts of the genome, different genes may suffer from different regulation patterns influenced by neighbor sequences). S/MARs have shown to minimize the gene silencing effects by acting as insulators (the gene would be integrated in an independent loop flanked by S/MARs elements)[12].

The size of S/MARs anchored loops has also been implicated in relative expression efficacy, negative correlation between S/MARs overlapping with genes, or intragenic S/MARs and gene expression was found in [9], and, further, in [13] it is shown that

intragenic S/MARs also have a pronounced specificity for tissues organs and developmental phases.

2.7 *In Silico* S/MARs Annotation

Experimental detection of S/MARs requires substantial effort and is not suitable for large-scale screening of genomic sequences, thus, *in silico* prediction of S/MARs can provide a crucial first step. Experimentally obtained S/MARs provided the work bases for studies in order to determine guideline definitions for S/MARs, who could provide an automated way of annotation.

The most comprehensive data collection of S/MARs publicly available is in S/MARt DB[1], a database dedicated to collect information about S/MARs. The database has up to now information about 559 S/MAR sequences.

An evaluation of the sequences in S/MARt DB was carried in [14] as well as a comparison with other sequences (promoters, exons, transcription factors, and other regulatory sequences) in order to study the nucleotide composition and possible consensus sequences. A consensus sequence is a way of defining a set of sequences where in some positions the nucleotides could change, for example, the sequences ATGG and AAGC could be represented by the consensus sequence AWGS where W stands for a T or A and S stands for a G or C.

The study revealed that S/MARs have a rich content of Adenine and Thymine bases (in some cases higher than 70%), failing to reveal a detectable organization among the nucleotides that could be described by a consensus sequence. Other studies however [15] reveal that most of S/MARs appear to have two separate sequences AATAAYAA and AWWRTAANNWWGNNNC within 200 base pairs, and define a set of statistical rules to help improve S/MARs finding, however care is suggested in the application of these rules as different results may arise from different combination of rules.

There are several informatics programs to identify S/MARs in DNA sequences, such as MAR-Finder (or MAR-Wiz)[2], marscan (part of the EMBOSS bioinformatic package)[3],

[1] http://smartdb.bioinf.med.uni-goettingen.de/

[2] http://futuresoft.org/MAR-Wiz/

[3] http://emboss.sourceforge.net/

ChrClass[1], and SMARTest (as part of the Genomatix GEMS Launcher package)[2]. Different programs rely on different methods to identify S/MARS and produce different types of results [16].

Specificity is the statistical measure of how well, a binary classification test, correctly identifies a condition when it is truly not present. In contrast sensitivity is the statistical measure of how well a binary classification test correctly identifies a condition when it is truly present. Different programs show different tradeoff's regarding sensitivity and specificity.

The low number of S/MARs sequences present in SMARt DB when compared with other genomic sequences, is the major drawback for improvements in both specificity and sensitivity of computer programs to predict S/MARs.

2.8 SNPs

SNPs (Single Nucleotide Polymorphisms) are variations in one nucleotide base among DNA sequences from the same location, from different individuals. These differences may induce changes in the synthesized protein. A different nucleotide in a coding sequence may give rise to a different amino acid, in this case is called a nonsynonymous SNP. However the possible combinations for different codons are $4^3 = 64$ and the number of different amino acids is 20, meaning that are amino acids coded by more than one codon. A SNP that does not change the code from an amino acid is called a synonymous SNP.

However, even some synonymous SNPs may influence a translation to protein as the SNP may occur at an intron/exon splice site, altering in this way the intron and exon structure and consequently the ORF may code for a different protein[19]. An ORF (Open Reading Frame) is usually defined as a stretch of DNA containing at least 100 codons that begins with a start codon and ends with a stop codon. However there are very short genes where this ORF definition is not valid (the number of codons is lower than 100) [7]. Other SNPs may be located in the promoter areas influencing in this way the regulation and expression of the corresponding gene [19].

---

[1] http://www.filesearching.com/cgi-bin/s?t=n&l=en&q=ftp.bionet.nsc.ru/pub/biology/chrclass

[2] http://www.genomatix.de

These changes in the protein structure may or may not be manifested in the phenotype. The phenotype is the visible outcome of a gene's action, such as blue eyes versus brown eyes.

An important feature of SNPs is serving as genetic markers for mapping specific genetic trait characteristics. A genetic marker is a DNA sequence or part of it, easily detectable by an experimental essay (for example if a SNP is linked to a change in the phenotype the detection could be directly observable). A genetic marker marks a specific point in the DNA, therefore it could serve to compare genotypes, explore evolutionary differences in single species, or among several species.

SNPs are therefore important in complex disease mapping, as well as plant breeding. Plant breeding aims at creating specific genotypes or phenotypes, using a wide variety of techniques. When studying human diseases it is important to know if a certain disease regarding a SNP is spread among the population, so the frequency of SNP occurrence is important to define a link between SNP and the disease. In plant breeding it is possible to propagate a specific polymorphism, thus the most important part is the SNP discovery and/or association with a specific phenotype or genotype.

2.9 *In Silico* SNPs identification

Computational SNP discovery, have evolved in the last years. First approaches relayed on visual comparison of several aligned sequences from multiples individuals. A sequence alignment is a way of arranging the DNA or RNA or protein sequences to identify conserved regions, as well as differences in the sequences.

Algorithms to produce the multiple sequence alignments are computational intensive, as the number and length of sequences increases. There are several types of algorithms, programs and frameworks to perform this task. There are several computational methods to perform multiple sequence alignments, some rely on Dynamic Programming and are efficient to perform an optimal alignment, however, are computationally expensive.

Progressive alignment algorithms start with the alignment of the two most related sequences and align each next sequence to those who are already aligned, this is a heuristic algorithm, thus, it does not guarantee an optimal alignment, and the final result is also

dependent from the choice of the two primary sequences. ClustalW is based on progressive alignment algorithms and is one of the most used multiple sequences alignment algorithms.

Iterative alignment algorithms make an initial alignment of groups of sequences, and align the remaining sequences like in the progressive alignments, however the initial sequences are repeatedly realigned. Other than the types referred above algorithms to produce multiple sequence alignments could rely on Hidden Markov Models, genetics algorithms, among other techniques.

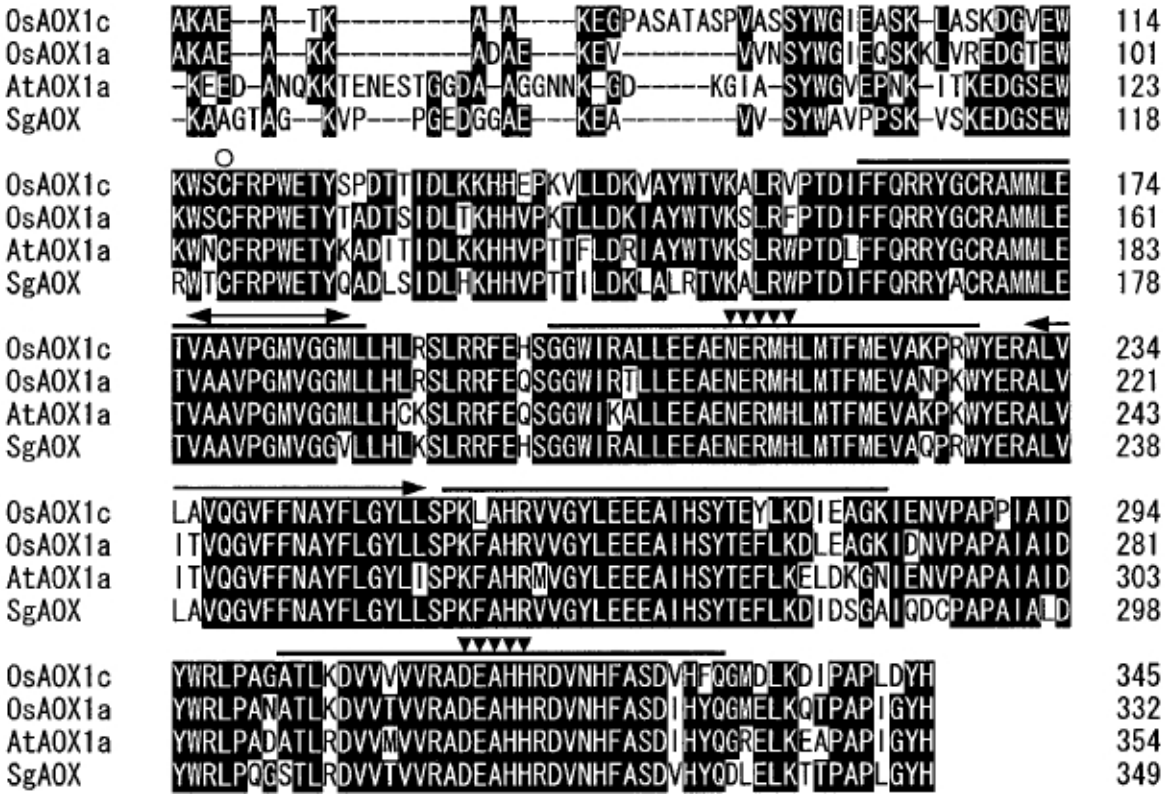


Figure 4 - This image represents a partial protein alignment for AOX genes from different species, the image is adapted from [20] when rice *AOX1c* was identified. Conserved regions are marked black while different regions are marked white.

As previously have been said first SNP identification approaches relied on visual comparison of the aligned sequences, although manual comparison is feasible for a small number of sequences, standard accuracy criteria are hard to establish.

As the number of publicly available EST (Expression Sequence Tags) begun to increase, tools to mine SNPs at a genome scale start to be applied. An EST (Expression Sequence Tag) is a short transcribed sequence of mRNA or cDNA typically ranging from 500 to 800 base pairs. The EST corresponding to a set of previously known regions (as a set of genes) were assembled in a larger sequence called contig. Assemblage and posterior

filtering of SNPs relayed mostly on parameters regarding sequence quality through quality scores [21]. Quality scores are measures assigned to each base present in a sequence, in order to determine if that base has been correctly sequenced to help distinguish between SNPs and sequencing errors is essential [19].

Quality scores are assigned based on wavelengths intensities for each nucleotide after having been sequenced. Sequence errors result from the impossibility to distinguish between the wavelengths corresponding to each type of bases A, T, G and C.

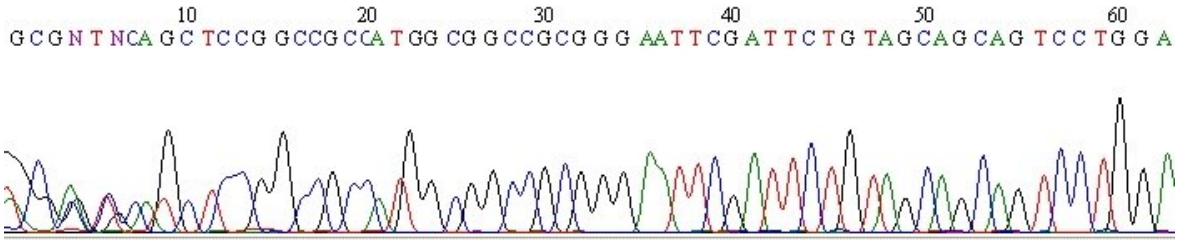


Figure 5 - The image represents a plotted graph with the signal intensities (wavelenghts) from part of a sequence isolated at the EU-Marie Curie Chair. There are several programs capable or reading signal intensities from a file generated when sequencing is performed and piloting the graph. The first bases of the sequence would have a lower quality score, since it is difficult to distinguish between signal intensities while in the middle of the sequence the scores would be higher.

In order to assign a measure of confidence to SNP without regard the source and overall sequence accuracy, a new set of statistical and mathematical models were developed. PolyBayes [22] was designed according these models in order to calculate a Bayesian probability. That mismatches between sequences are true polymorphisms rather than sequencing errors.

2.10 *Arabidopsis thaliana* (Thale Cress) and *Oryza sativa* (Rice)

Arabidopsis is a model plant appreciated by the scientific community, for several reasons, such as having a small genome (~120Mb), a small size and a rapid life cycle (it takes about six weeks from germination to mature seed). Arabidopsis was the first plant to have his whole genome sequenced (in 2000) the resources available today concerning these plants are far extensive and well organized. *Arabidopsis thaliana* has five chromosomes, and belongs to the biological group of dicotyledons.

Figure 6 – *Arabidopsis thaliana*

Oryza sativa is a cereal very important in human diets, being the world's most used crop. This cereal has the smallest genome of all the cereals (~430 Mb). The *Oryza sativa* is a good model for characterizing the genes of other cereals, and associating them with various agronomic traits, cereals number of homologous genes whose order are relatively conserved and numerous resources for a genomic approach, such as excellent genetic maps and efficient techniques for genetic transformation. Furthermore, rice can also be used as a model genome for one of the two main groups of flowering plants, the monocotyledons, in the same way as *Arabidopsis thaliana* is the model for the other group, the dicotyledons. *Oryza sativa* has twelve chromosomes.

Figure 7 – *Oryza sativa* (Rice)

2.11 AOX

Plants, some fungi, algae, eubacteria, protists and more recently in animal kingdom phyla (Mollusca, Nematoda and Chordata) contain an isoenzyme Alternative Oxidase [5] [40]. An isoenzyme usually refers to several proteins that may have rather different structures, but the same biological functions.

The alternative oxidase (AOX) is present in the inner membrane of mitochondria. Mitochondria are organelles of eukaryotic cells, witch are also described as "cellular power", because they generate most of the cell's energy. In the inner membrane of the mitochondria electrons (obtained from photosynthesis and glucose catabolism), pass though several complexes (complex I, II, III, IV and V). In complex IV oxygen is incompletely reduced to water generating reactive oxygen species (ROS). ROS in a higher concentration can cause great damage to cells, including dead.

In complex I, III and IV the proton move from membrane matrix into intermembrane space. This creates a potential energy, referred to as proton-motive force. In complex V, proton-motive force is dissipated by the generation of ATP (biological energy) [41]. This is known as mitochondrial electron transport chain (ECT), cytochrome pathway or oxidative phosphorylation. This pathway is very important for eukaryotic cells because it produces energy (ATP) needed for their metabolism.

An alternative pathway is located between complex II and complex III where the alternative oxidase enzyme is present. The alternative oxidase acts like complex IV reducing oxygen to water, however without generating Reactive Oxygen Species, because oxygen is completely reduced.

The activity of the Alternative Oxidase is regulated by stress conditions, the concentration of Reactive Oxygen Species in cells, pathogen attack, stage development and others [42]. Alternative oxidase can act to decrease formation of ROS [43], is involved in the response to virus infection, heat production, and may prevent programmed cell death. [40] [5] [43].

2.12 AOX Gene Family

Alternative Oxidase is an enzyme encoded by a small family of genes subdivided in two subfamilies, *AOX1* and *AOX2*. In Arabidopsis for instance there are 5 genes (*AOX1a*, *AOX1b*, *AOX1c* and *AOX1d*) and *AOX2*, which have been first reported in [3]. In rice there are 4 genes all belonging to the *AOX1* family (*AOX1a*, *AOX1b*, *AOX1c* and *AOX1d*). *AOX1a* and *AOX1b* were first reported in [4], while *AOX1c* has first been reported in [20]. Once the genomes of Arabidopsis and rice were completely sequenced, *AOX1d* was identified in both Arabidopsis and rice.

In Soybean have been identified 3 genes, one belonging to *AOX1* family and *AOX2* to *AOX2* (*AOX2a* and AOX2b) [25]. AOX genes are distributed unevenly across different species.

2.13 AOX as functional marker

A “functional marker” can be defined as genetic marker, deriving from sequences well known (sequence motifs), from which the function is well characterized. Functional markers are derived from polymorphic sites within genes, and are usually related to phenotype variation.

Genetic markers resulting from functionally defined sequences apply to plant breeding and biodiversity studies, as well as to human genetics. Markers are used in basic plant research and plant breeding to characterize plant germination, assist in gene isolation, etc.

Alternative Oxidase is proposed to be a functional marker for adaptation of plant growth under stress [5]. This idea was been sustained from studies on the metabolic role of alternative oxidase concerning respiration under stress condition, as well as the relation between AOX activity and differential growth. For AOX to be accepted as a functional marker, it is required to discover polymorphic sequences such as SNPs that can be used to genotype and phenotypic characterizations [11].

3 Project Description

3.1 Methodology

3.1.1 Previous work on S/MARs annotations

The first S/MARs annotations were essentially performed by experiments regarding the binding of DNA fragments with nuclear matrices that were experimentally prepared. In [23] is performed a classification of S/MARS isolated from tobacco according to their binding strength, strong, medium and week binders. The most common experimental procedure used to isolate the S/MARs is also described [23].

As referred before under Basic Concepts there were several studies both by laboratory essays and by study of previously identified sequences, aiming to understand the S/MARs functioning at single gene level and at chromatin level. However, the role of S/MARs in maintaining and regulating the cell genetic machinery despite important, remained poorly understood. A summary of S/MARs structure and function could be found in [24].

All S/MAR sequences present in SMARt DB were evaluated by searching for recognizable patterns or "motifs". Sequences from different species were not distinguished, since they have not significant differences, and S/MARs sequences from one specie bind to nuclear matrices of other specie [14].

To compensate the low number of sequences present on the database (245 at that time), a Roulette Wheel selection algorithm was used to generate four additional sets of data. A Roulette Wheel selection algorithm is a process of make a selection (in this case any of the four bases A, T, C, G) using a fitness function who quantifies the optimality of a solution, (in this case the fitness function was the frequency of occurrence of a given base in the S/MARt DB sequences), the probability of selection is proportional to its fitness.

The results have shown that S/MARs have a rich content of Adenines (A) and Thymines (T) especially when analyzing sequences with an AT content > 70%, as 12 % of them contain S/MARs while only 3% contain promoters and 0.3% contain exons. Conversely the portion of sequences with low AT content that contain S/MARs is significantly lower (6.5%) than exons (67%) and promoters (47%).

Searches for consensus sequences describing S/MARs were performed using weight matrices. A weight matrix is a way of describe a pattern or a motif through a matrix. Usually in the rows of the matrix are the four DNA bases or if the matrix is applied to a protein the letters corresponding to the amino acids. And in matrix columns we have the frequency of occurrence for each of the bases or for each of the amino acids.

Weight matrices are usually constructed based on information provided by the multiple alignments of several sequences, so that frequencies for each base or amino acid can be calculated. Weight matrices are more powerful than consensus sequences because they allow a probabilistic treatment.

In the search for consensus sequences describing S/MARS the IUPAC consensus strings have been converted to weight matrices. IUPAC (International Union of Pure and Applied Chemistry) have defined a set of characters to identify variabilities in similar sequences so that they may be expressed by means of a consensus string. Results however were inconclusive.

Another search was conducted especially for sequences of six nucleotides, and for possible transcription factors related sequences but once again the results were inconclusive. Some S/MARs were found in the majority of sequences but not in all. Another interesting study [15] defined a set of probabilistic rules to find S/MARs, however is suggested that these rules may conflict among themselves. A more complete list of results and previous work could be found on [24].

3.1.2 Choosing a S/MARS Annotation program

For the task of annotating S/MARS in the vicinity of AOX genes in *Arabidopsis thaliana* and rice SMARTest were chosen. This decision was based on several considerations discussed below.

For the dataset considered in [16] SMARTest was found to have the best specificity (63%), a lower value that the one specified in [17] (68%). Regarding sensitivity SMARTest values are low (17% in [16]) and (38% in [17]). Analyzing data in [16] also when comparing the values for other programs SMARTest showed lower sensitivity values. However in comparison with the absolute values, regarding for instance ChrClass program for 53 true positives, defined experimentally, ChrClass has identified 105

S/MARs, which shows a great number of false positives, despite lower number of true positives marked by SMARTest, the number of false positives is also significantly lower 6 for SMARTest while 52 for ChrClass.

One reason for the low values of SMARTest sensitivity (identification of true positives) could be, according to the authors, that the current library of weight matrices is derived from a set of sequences where AT rich patterns could be over represented.

According to [16] the other S/MAR finder applications have shown inferior performances finding S/MARs. They show inferior values of specificity when compared with SMARTest and while sensitivity is higher, the number of false positives is also much higher.

While the number of S/MARs identified by SMARTest was low when compared to the other programs, the accuracy prediction was the highest among the applications tested. According to definitions in [16] we could say that SMARTest has a lower efficiency (number of experimental MARs hit by prediction)/(total number of experimental MARs), but a high accuracy (number of true positive prediction)/(total number of predictions).

Other S/MAR studies at whole genome level have been reported in [9], [13] and [17] showing the practical use of SMARTest for S/MAR prediction in long genomic sequences. For the present project, it was concluded that SMARTest should be applied since large part of S/MARs function, importance and identification remains poorly understood, SMARTest is the most accurate tool to identify S/MARs, and, the feasibility of studying S/MAR distribution along large genetic sequences was already demonstrated.

Characteristics of the applications mentioned above are summarized in the table below according with [16]:

Program Name	Strong points	Weak points
SMARTest	Best specificity (or accuracy), (number of true positive prediction)/(total number of predictions); Lowest number of false positives	Lower sensitivity
ChrClass	Best tradeoff between sensitivity and specificity; Highest number of true positives (50,5 % of total predictions)	Highest number of false positives (49,5% of total predictions)
MAR-Finder	Moderate numbers of S/MARs predictions	Lower specificity
marscan	Moderate numbers of S/MARs predictions	Lower specificity

3.1.3 SMARTest

SMARTest analysis is based on a density analysis of S/MARs patterns represented by a weight matrix library generated by comparative sequence analysis of experimentally defined S/MAR sequences. The experimental sequences used to generate the weight matrices library were taken from the SMART DB and EMBL databases, as well as literature.

The sequences selected as the base dataset for S/MARs identification are aligned and the weight matrices are created automatically using the program MatDefine. Another program called MathInspector use the weight matrices created to search if a specific sequence has a certain degree of similarity, expressed in the from of a probability.

SMARTest uses MathInspector to check if a sequence matches the weight matrices defined by MatDefine. The SMARTest input sequence is divided in consecutive blocks of 300 bp (base pairs) called sliding windows, with increments of 5 base pairs. For example if an input sequence with more than 300 base pairs is supplied to SMARTest for analysis, the first 300 base pairs (from base 1 to 300) are processed, then the next 300 base pairs (from 6 to 306) are processed, and so on until the end of the input sequence.

The length of the sliding window is 300 base pairs because this is assumed to be the minimum length of a S/MAR. Using this method of analysis the total length of a possible identified S/MAR may vary, since consecutive sliding windows may have a high degree of similarity with the weight matrices library consequently the length of the S/MAR will be greater than 300.

Results of SMARTest comprise the beginning of S/MAR sequence with the input sequence, the end of the S/MAR sequence, and the length of the S/MAR. SMARTest as well as MatDefine and MathInspector are part of Gems Launcher, a proprietary framework for functional analysis of genomic sequences. In order to use SMARTest, an academic license was obtained from Genomatix Software.

3.1.4 AOX gene sequences for Arabidopsis

Since the complete genome of Arabidopsis was already released in 2000 [1], and because Arabidopsis is a model plant in genetics and plant sciences, the resources available are far extended and easy to find. Thus, it is appropriated to use the sequences of this species for fundamental research.

Examples of databases containing the whole Arabidopsis genome sequence, gene annotations, microarray data, genome browser among others resources are TAIR[1], TIGR[2] , mips[3], ncbi[4], and others.

TAIR (The Arabidopsis Information Resource)[5] comprises a considerable set of tools and data regarding the Arabidopsis genome. A quick search for each one of the AOX genes show a lot of useful information, well grouped, and with several links to more specific areas, such as external links, scientific articles regarding AOX genes, AOX gene sequences, etc.

TAIR is based in TIGR 4th release of Arabidopsis genome, as well as the work performed in this study. In downloads category it is possible to obtain a copy of the Arabidopsis genome for each one of the 5 chromosomes as well as the mitochondrial and plastid genomic sequences for later analysis. Having access to the complete Arabidopsis genome sequence and the details about the genes, the next step is to determine the sequences that need to be analyzed with SMARTest.

The average loop size of Arabidopsis has been estimated to 25kb (25000 base pairs) [26], so the median size between two S/MARs regions attached to the nuclear matrix should be 25kb, since S/MARs are the elements that bind chromatin to the nuclear matrix. It should be considered, however, that this does not mean that only 2 S/MARs should be found in this regions, since other S/MARS sequences may be present but not attached to the nuclear matrix. Moreover, cells in different tissues and organs have a different chromatin organization, so S/MARs that are bounded to the nuclear matrix in some tissues could not be in others and vice-versa.

Since 25kb is only an average measure, a length of 50kb before and after the AOX genes is sufficient for SMARTest to check, if some sequences match SMARTest weight matrices. Since DNA is a helical double strand molecule, binding with the nuclear matrix

---

[1] http://www.arabidopsis.org

[2] http://www.tigr.org/tdb/e2k1/ath1/

[3] http://mips.gsf.de/proj/plant/jsf/athal/index.jsp

[4] http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genomeprj&cmd=ShowDetailView&TermToSearch=13192

[5] http://www.arabidopsis.org

may happen with any of the two strands, therefore both, forward strand and reverse strand, in the neighborhood of AOX genes was analyzed for S/MARs sequences.

3.1.5 AOX gene sequences for rice

The complete rice genome sequence has been published in 2005 [2], by IRGSP (International Rice Genome Sequencing Project). The complete sequences could be downloaded from IRGSP page directly in the from of a text file corresponding to each of the rice chromosomes. However, no information about the genes is available in the IRGSP page.

In the RAP-DB (Rice Annotation Project) home page, it is possible to view these sequences through a genome browser, as well as to perform several types of searches. When searching for Alternative Oxidase genes a lot of inaccurate results are displayed. Comparing against IRGSP build 3.0, 24 gene matches are found for Alternative Oxidase genes. This number of matches is far too high since there are only 4 genes of Alternative Oxidase. Comparing against IRGSP build 4.0, 7 gene matches are found. Once again, genes are overrepresented and incorrectly annotated.

TIGR (The Institute for Genomic Research) has also developed a project for the whole rice genome sequence, the TIGR Rice Genome Annotation[1]. TIGR Rice Genome Annotation integrates a lot of useful tools, and allows several methods for gene search, search for locus, function, homology (similarity), among others. Since AOX genes in rice are annotated under general terms (a functional description is not available), a homology search must be performed with manual selection of the candidates based on literature as described in the next section. The genome sequence of TIGR Rice Database was used for the project, since more powerful search tools are available and gene data of IRGSP regarding Alternative Oxidase is highly inaccurate.

According to mips statistics page for plant genomes[2] rice has an average gene density per 10kb value of 1.66 while Arabidopsis has 2.74. Since no literature was found that estimate the loop size for rice the length of the sequences to analyze was calculated based on average gene density per 10kb. This is likely to be a gross estimation. The number of

[1] http://www.tigr.org/tdb/e2k1/osa1/

[2] http://mips.gsf.de/projects/medicago/statistics

genes in compared sequences was similar or the same, the number of S/MARs was under study. This methodology was adopted since no better term for correlating the number of S/MARs in Arabidopsis and rice was found. Thus, the length of the sequences to be analyzed was calculated in 82530 base pairs before and after the AOX genes.

3.1.6 Data Integration and Analysis

After having S/MARs sequences in the neighborhood of AOX genes identified, Arabidopsis data were compared with supplemental data from [9] to find any discrepancies.

Results from SMARTest annotation program are shown in text format, (the beginning, the end, and the length of the S/MAR) in this way they do not provide a intuitive information regarding how close the S/MARs are from genes, exons and introns, if they flank any gene, and if they could be important for gene regulation.

To understand better the results, a program was created to receive a file with the details about the genes in the analyzed sequence, as well as the details about the S/MARs position predicted by SMARTest. The output is an image representing the genes, and the S/MARs.

Another program was created to draw the position of the analyzed sequence in the chromosome. The images, as well as the output text data and the description of the genes (based on TIGR data) will further be presented together for better understanding in the Appendix 1 of this report. Because of the confidential nature of this work the results of this Appendix should not be publicly available.

The programming language adopted to create the informatic tools described above is C++. This decision has mainly two purposes. The first is to learn the basics of a programming language that was not used during the course in Informatics. The second reason was to facilitate the portability of codes between the Linux and Windows environments without using to the java platform.

GD graphics library had been used to create images. GD is an open source graphics library for dynamic creation of images especially in png, jpeg, and gif file formats, but supports also other types of files. The GD library is currently part of the PHP project, and is available for Linux, Windows, Mac OS X operating systems.

3.1.7 AOX gene Expression in Arabidopsis and rice

In order to investigate, if predicted S/MARs could have some influence on gene regulation, a search for evidences of AOX gene expression in publicly available databases will be performed.

A study of AOX gene expression in Arabidopsis has been reported in [27]. This study used Genevestigator [28] as a tool. Genevestigator is a publicly available database, comprising a repository of microarray data for *Arabidopsis thaliana* and a set of tools to mine the available data. Expression of five AOX genes in Arabidopsis was analyzed using the microarray data of various tissues, development stages and stress conditions [27].

Related to rice, several projects are devoted to gene expression analysis, most of them based on microarray and mpss data. Examples of projects are the Rice Expression Database[1] based on IRGSP data, the NSF Rice Oligonucleotide Array Project[2] , the Yale Virtual Center for Cellular Expression Profiling of Rice[3] and Rice MPSS plus database[4].

Rice Expression Database was the first to be created, it has microarray data, available in the from of photos from experiments performed, however the software to analyze the data (ArrayGauge from Fujifilm company) is not available (only for owners of Fuji microarray technologies), furthermore microarray probes for AOX genes are inconsistent with both IRGSP and TIGR genome sequencing projects.

Rice Oligonucleotide Array Project or Rice Array project has a more accurate and comprehensive set of data based on TIGR, but little evidence of AOX gene expression, only microarray data for one gene (*AOX1a*) and only in two conditions.

Yale Virtual Center for Cellular Expression Profiling of Rice has whole-genome microarray expression data based on TIGR. Evidence of expression for AOX genes is found on different tissues and for all AOX genes.

The rice MPSS plus database includes a comprehensive set of signatures (small sequences of 17 or 20 bp matching a gene) libraries, more than 20 MPSS signature

---

[1] http://red.dna.affrc.go.jp/RED/

[2] http://www.ricearray.org/rice_study.shtml

[3] http://bioinformatics.med.yale.edu/rc/description.jspx

[4] http://bioinformatics.med.yale.edu/rc/description.jspx

libraries derived from diverse tissues, submitted to abiotic stress (cold, drought, and salt treatments) as well as different growth conditions (in light and in the dark) and different developmental stages[29].

Considering the factors mentioned above, a summary of AOX gene expression in rice was created, based on Yale Virtual Center for Cellular expression Profiling, and MPSS plus database.

3.1.8 Previous work on SNP identification

As SNP studies become an important topic in medical research, plant science, agriculture sciences, and other areas, a wide variety of searching techniques has been developed. Most of them are experimental methodologies, but computational methods are also widely applied, especially to perform a first selection that will be validated later by an experimental essay.

As described before, computational methods for SNP identification have undergone a significant evolution, from simple manual comparison of multiple aligned sequences to intuitive commercial applications, freely available online tools, and some approaches that integrate both laboratory experiments and software tools to identify SNPs in a large number of genes.

Techniques to measure gene expression like microarrays are adapted to perform a whole genome mapping of SNPs. DNA Vision[1] is an example of a European company who offers a wide variety of services regarding whole genome SNP genotyping as well as SNP genotyping of previously selected genome areas.

SEQUENCHER from Gene Codes Corporation[2] and HelixTree from Golden Helix, Inc.[3] are examples of quite sophisticated commercial software frameworks for SNP identification.

In parallel with most sophisticated and expensive methods mentioned above a wide variety of SNP scoring algorithms were published in academic papers. SNP scoring

---

[1] http://www.dnavision.be

[2] http://www.genecodes.com/

[3] http://www.goldenhelix.com/helixtree_pbat.html

algorithms assign a confidence score to a specific site of a sequence that have a variation among several individuals.

Bioinformatics Links Directory of Canadian Bioinformatics Workshops[1] provides links for tools and databases recommended from Bioinformatics experts. In the area of sequence polymorphisms are listed 39 resources from which 22 are tools (to use online or offline) to analyze SNPs, and some tools are not listed. Due to this diversity of tools already available the most reasonable approach is to perform a selection of one more and adapt them to the specific requirements of EU Marie Curie Chair laboratory.

PolyBayes [22] program is among the most widely used for computational SNP detection. It was used in several studies to predict SNPs over large amounts of data in Humans [22] [34], and in soybean [35]. It serves also as the SNP detection engine behind the SNP discovery in two large scale projects; the SNP consortium[2] one of three sequencing centers founded by 10 pharmaceutical companies to generate over 300.000 SNP's for public access and clone overlap SNP mining[3] a project in Department of Dermatology, Washington University School of Medicine to help produce a complete reference sequence of the human genome.

3.1.9 PolyBayes

PolyBayes program is mainly divided into three parts each independent of the others, an anchored multiple alignment algorithm, a paralog discrimination algorithm, and the SNP detection algorithm. It has also a set of options to control how the output to the text file is produced.

The purpose of the anchored multiple alignment algorithms is to produce an alignment of the sequences based on that template sequence, or a sequence of reference. Redundant sequences (or the sequences that provide the bases for SNP analysis) are aligned or anchored with the reference sequences forming a cluster. The availability of quality scores is recommended both to produce the alignment of the sequences and later to use the PolyBayes SNP detection algorithm, although they are not required, the same is

---

[1] http://bioinformatics.ca/links_directory/

[2] http://snp.cshl.org/

[3] http://polybayes.wustl.edu/overlapsnp

true for the reference sequence. The anchored alignment algorithm could be bypassed if another alignment is available in the 'ace' file format. That file format has been established by the first two applications to process the sequencing results generating quality base scores (PHRED) and later produce multiple sequence alignments (PHRAP).

The next step in PolyBayes analysis is the paralog discrimination algorithm in the presence of a reference sequence. This algorithm has been designed with the intention to answer the question, if an aligned sequence is likely to have originated from the same genomic location as the reference sequence or otherwise has come from a different but very similar region.

The algorithm attempts to decide whether the numbers of discrepancies observed in the aligned sequence are consistent with the number of sequencing errors expected (based in quality scores) plus the expected differences among the sequences, in case the discrepancies are higher than expected the sequence is considered to come from a different location.

SNP detection algorithm calculates the probability that discrepancies at the analyzed location represent true sequence variation as opposed to sequencing error. The algorithm takes into account several parameters as for instance, nucleotide differences, the quality scores of the nucleotides, the depth of the alignment (number of sequences), the base composition of the sequences, expected polymorphism rates (à priori knowledge), and calculates the probability that the site is polymorphic.

Other approaches have used PolyBayes as a base to implement a machine learning approach in order to improve its prediction accuracy. This approach however was used on large training sets previously classified, and may differ from genomes across different species [36]. The using of a machine learning approach has however increased PolyBayes prediction.

3.1.10 Development of an application to simplify the SNP identification process.

After reviewing several algorithms to identify SNPs, PolyBayes was selected for SNP identifications in the sequences analyzed at the EU Marie Curie Chair. The next step is customizing this tool to the specific needs of the work group at the EU Marie Curie Chair.

A application will be developed in order to accept a set of sequences received directly from the sequencing company (for instance Macrogen), clip parts of the sequence that are not necessary, use PolyBayes program to perform SNP identification and generate an image from PolyBayes text results. Sequences mentioned above are supposed to be from the same genomic location of several plants of the same species. The application should allow the user to adjust PolyBayes input parameters in order to customize SNP identification for AOX genes in several plants, as well as optimize results.

PolyBayes multiple sequence alignment algorithm does not perform the alignments internally, instead it is designed to use another program to perform the alignments (cross_match) who is integrated in another package of programs (PHRED and PHRAP). So, sequenced data is analyzed first with PHRED in order to assign quality scores to each base, then cross_match and PHRAP are used by PolyBayes to perform the alignment.

CAP3 [37] is a program specifically designed to perform multiple sequence alignments and could write the output in the '.ace' file format. Thus PolyBayes could use it directly for SNP identification instead of cross_match and PHRAP, it does not require the existence of quality scores but if available they are used by the alignment algorithm. Sequencing companies usually pre-process sequences in order to supply quality scores to the customer, regarding each sequence, as well as other useful information.

This approach has already been used in the context of other projects [36] [38], it simplifies the analysis process, and gives the user a greater control of the assembly process. Considering this, the application would also allow the user to adjust the input parameters of CAP3 program, providing a default way for sequence assembly.

The developed application must support all the input parameters for CAP3 and PolyBayes, although the user should know the effects of changing input parameters. Both CAP3 and PolyBayes have their specific licenses terms, and thus cannot be freely distributed, but are available free of charge for academic users from their authors.

Sequence files in FASTA format with the sequence text description, and the quality scores are created and supplied to CAP3 with user input parameters; the output is then used by PolyBayes also with user input parameters.

4 Work performed

4.1 Localization of S/MARs within and in the neighborhood of AOX genes

4.1.1 Obtaining genomic data for Arabidopsis and rice

As described under methodology, the sequences used to find S/MARs are based on TIGR 4th release of Arabidopsis genome, and TIGR 5th release of Rice Genome. The first step was to locate the AOX genes in both genomes. For Arabidopsis using TAIR database (The Arabidopsis Information Resource) this was a trivial task, as a simple search by name was enough.

As described, rice AOX genes are not functionally annotated, so another search method was required. A homology search with the complete gene sequences from AOX also did not provide any results. The 3rd exon of Arabidopsis AOX gene is highly conserved among plant species [3], so a homology search with the 3rd exon from *AOX1a* was performed and several matches were found.

Four of these matches showed a high homology in protein with AOX while the others showed only homology with some parts of AOX protein. These four matches were then compared with literature to identify each of the genes.

AOX1a and *AOX1b* genes in Arabidopsis [3] and rice [4] are in a tandem repeat, this is, one immediately after the other in the same strand of DNA. From the four initial matches two match these criteria the locus LOC_OS04g51150 matches *AOX1a* and locus LOC_OS04g51160 matches *AOX1b*, and have the same number of exons as well as the same distance between exons as the ones reported in [4].

AOX1c has been first reported in [30], locus LOC_OS02g47200 has the same distance between exons, and a higher homology at nucleotide and protein levels than locus LOC_OS02g21300, so the former has been classified as *AOX1d*.

The sequence files corresponding to the whole chromosomes where AOX genes are located were downloaded from TIGR ftp server. In Arabidopsis *AOX1a*, *AOX1b* and *AOX1c* are located in the 3rd chromosome, *AOX1d* is located in the 1st chromosome and *AOX2* is located in the 5th chromosome. In rice, *AOX1a* and *AOX1b* are located in the 4th

chromosome while *AOX1c* and *AOX1d* are located in the 2nd chromosome. Since *AOX1a* and *AOX1b* are in a tandem repeat (one after the other in the same DNA strand), the sequence for S/MAR analysis will be practically the same for both genes (1.2kb approximately in Arabidopsis and 1.9kb in Rice), eliminating the necessity of a separate analysis for both genes

4.1.2 Analysis of gene sequences

Chromosome sequence files obtained range from 23 470 804 base pairs from Arabidopsis chromosome 3 to 35 925 388 base pairs from rice chromosome 2. The next step is to extract the sequences vicinity for each gene to perform the analysis. A simple C++ application was created to obtain a specific part of the chromosome file, the input parameters are the position of the first base pair (for the sequence to analyze) in the chromosome file, the position of the last base pair (for the sequence to analyze) in the chromosome file, and the text file where the sequence will be stored.

Since both strands of DNA are to be analyzed another small application was created to obtain the complement sequence of a specific DNA sequence, the input parameters are the filename with sequence input and the filename with the sequence output. Complement strands are treated the same way as "standard" or forward strand regarding gene positioning, however one must consider that gene transcription is processed in the reverse direction.

Sequence analysis with SMARTest was then performed using the Genomatix web interface for GEMS Launcher (the framework in where SMARTest is integrated). Since SMARTest annotation values are relative to the input sequence supplied, the beginning position of the S/MAR in the analyzed sequence is not the same as the beginning position of the S/MAR in the chromosome, to address this problem an application was created to convert the sequence relative coordinates to absolute positions in the chromosome. The input parameters are the file with S/MAR annotations, the beginning of the analyzed sequence in the chromosome and the file, where the absolute S/MAR positioning are to be stored.

4.1.3 Comparing results with previous works

When comparing the achieved results with other studies that have already studies S/MARs positioning at a genome scale for Arabidopsis [10], results were different.

Chromosome data used for S/MAR prediction in [10] was obtained and compared with chromosome data downloaded from TIGR.

There are two techniques for sequencing a whole genome, the map based shotgun sequencing and the whole genome shotgun sequence [44]. Since current sequencing methods can only sequence from 500 to 1000 base pairs, the larger chromosome sequences must first be divided in smaller sequences. The first method constructs a "genome map" of 150 000 bp replicated fragments and latter divide these fragments in smaller overlapping and redundant fragments that can be sequenced, since fragments overlaps it is possible to obtain the larger sequence of 150 000 bp comparing the overlapping sequences. The second method uses a similar approach, but without constructing 150 000 bp fragments, the genome is divided in small redundant overlapping fragments directly ready for sequencing. This approach is more time efficient but the process of assembling the genome sequence is more complex.

One of the major problems comes from highly repetitive genome sequences, (although this is more significant in the whole genome shotgun sequencing method) as these repetitive sequences could be from different places of the genome, and highly repetitive sequences could contain structural sequences in between, these issues may lead to mis-assembled genomes (often with gaps, regions in which none information is known) [45].

When assembling the genome, the coverage is the average number of overlapping sequences in which a nucleotide base is present, for example, a 12X genome coverage means that in average every nucleotide base of the human genome is present in 12 different overlapping sequences. Increasing the genome coverage in high repeat regions is a way to overcome these difficulties. New versions of the genome are released as mis-assembling issues and the gaps in the sequences are removed.

The differences in the results were due to different versions of the Arabidopsis genome sequences. The version used in this work is the most recent version of Arabidopsis genome (the TIGR 4^{th} release of the Arabidopsis genome). The length of the S/MARs was the same but the positioning of the S/MARs was different. Differences in the total length of chromosome files range from hundreds of base pairs (in chromosome 5) to hundreds of thousands (in chromosome 1 and 3).

4.1.4 Generate an image representation from text files

Text files do not provide an intuitive way for visualizing results, since distances between genes and S/MARs, as well as S/MAR positioning and length are not easily visualized.

An image representation is much more adequate, since it allows having a better idea for possible S/MAR influence on gene expression, especially when complementing text data with image data.

The program to generate the images was developed in C++ and has 4 classes as shown in the UML diagram. The input parameters are, the base pair where the sequence analyzed for S/MARs begins in the chromosome, ending base pair, image width, image height, the file with the details about the genes in that sequence, the file with the details about the S/MARs in that sequence.

Information regarding the genes comprises the locus of the gene, their positioning on the chromosome (the position in the genome of the first and the last base pair of the gene), the DNA strand (forward or complement), and a list with the positioning of the exons and introns. Each exon and intron is characterized by an instance of the positioning class, (the description and the first and last base pairs in the chromosome).

Each element of the S/MAR class has a description about the S/MAR represented, the beginning and the end of the S/MAR in the chromosome as well as the DNA strand of the S/MAR (forward or complement). The SequenceDrawer class reads the information concerning the genes and the S/MARs from a text file and creates a list of genes and S/MARs.

Depending on the height and width of the image as well as the size of the sequences to display (supplied by the user), the program calculates the scale, draw the genes and the S/MARs according to that scale in ".png" file format.

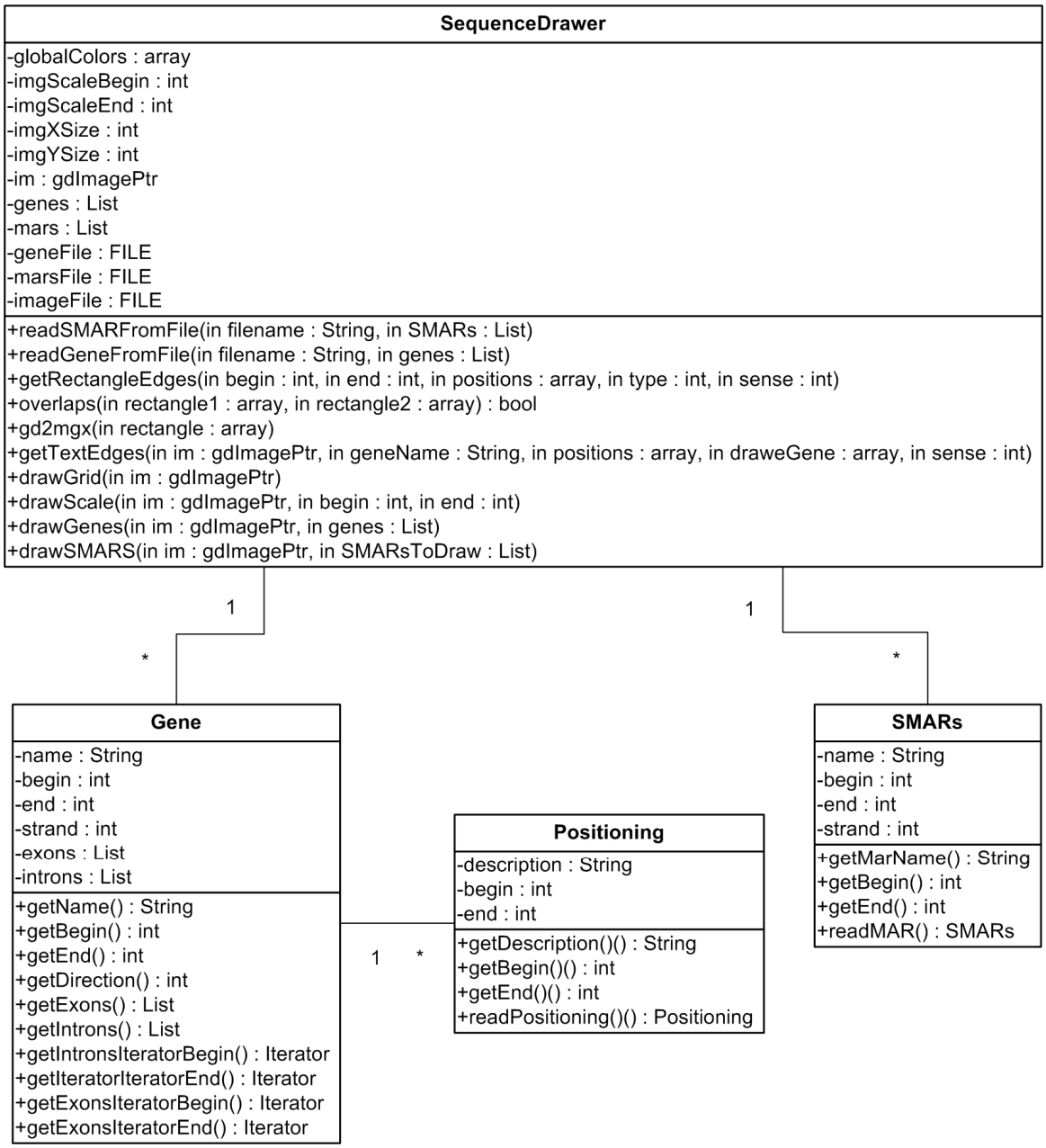


Figure 8 – UML diagram of the application that generates the image schemes based on the details of the genes and the S/MARs supplied in a text file

For Arabidopsis gene details were obtained from TAIR page[1]. Web pages with details for each gene in each one of the analyzed sequences were saved, and an application was created to extract the details from the text web pages and add the details to the text file used to generate the images.

[1] http://www.arabidopsis.org

For rice, gene details were obtained from TIGR Rice Database, more precisely from TIGR Rice Genome Browser, since details about the positioning of exons and introns were not available at individual gene details pages. Genome Browser allows the download of a text file with the details of all the genes in a given interval for any chromosome. Text files were downloaded and an application was created to parse these files and store the details in the text file used to generate the images.

S/MARs text file are created with S/MARs predictions from forward and complement DNA strands joined together, but marking for each S/MAR the strand of which it belongs (forward or complement strand).

To have the notion where each analyzed sequence is placed in the whole chromosome yet another application was created that receives the length of the chromosome as well as the beginning and the end of the analyzed sequence. Images are generated in "PNG" file format.

Images were processed to add more details, text details regarding identification and positioning for each of the S/MARs. Identification, positioning and details for each gene are presented together with the images and are available at the end of this report (Appendix 1).

4.2 Finding evidence of AOX gene expression in rice

Evidence of AOX gene expression for Arabidopsis in available microarray data is well documented in [27], so it was not considered necessary to perform any more searches.

As described before search for AOX expression data in rice is based on two databases, one of microarray data and the other of MPSS data. Both methods have been briefly introduced before, but some more considerations are to be addressed in this section.

The units used for MPSS measure are in TPM (Transcripts Per Million), this is a normalization performed to facilitate comparisons among different signatures scattered among different libraries (different signatures for the same gene may be in different libraries). Normalization is necessary to ensure that comparisons across different libraries accurately reflect biological differences and not only the total number of tags sequenced.

A small expressed gene would have expression values ranging from 1 to 10 TPM, while a much expressed gene may have more than 1000 TPM. The TPM value is obtained

for each signature dividing the RAW expression (number of hybridized signatures) by the total number of signatures and multiplies by 1,000,000.

Microarray and MPSS are different technologies, the units for measuring gene expression are different in kind (color intensity and number of transcripts), and the results from both may not be directly comparable but nevertheless complementary. Microarray have several limitations concerning accurate expression estimation of low expressed genes while the assignment of individual signatures for each gene in MPSS can be difficult in gene families with a high degree of similarity since a signature may match more than a gene[31].

Even when comparing microarray results for different platforms after removing potential sources of variability; it is shown that gene expression results may depend upon the type of microarray used in experiments [32].

Both microarray and mpss experiments are replicated several times in order to provide more data. This redundant data allows the assessment of a 'confidence score' and thus increase the precision of gene expression estimation [33]. In the Yale Rice Project, it is assumed that an expressed gene should be detected in at least 2 of 3 hybridizations/replicates.

Results of gene expression in these databases are presented in Appendix 2 of this report.

4.3 Develop an application to identify regions of Single Nucleotide Polymorphisms in AOX genes

4.3.1 Identification of Application Requirements

An Important part of application development is to understand the client requirements, so that the application could be efficient (address user needs) and usable. This is not an easy task, since most of user requirements are not trivial and not directly perceptible especially when concepts from diverse areas are involved. Therefore the requirements described below are not exhaustive or complete, instead represent the most important points to consider.

- Application should run on Microsoft Windows

Almost all the members of the work group at EU Marie Curie Chair use Windows operating systems on a daily basis, so the software should be developed for the Windows platform.

- Graphic Interface

The programs required for SNP identification have already been described before, both use a command line interface and the results are stored in text files. To improve usability a graphical interface should be developed.

- Full customization of CAP3 and PolyBayes

While usability is an important part, functionality should not be forgotten. Users must have access to all options for CAP3 and PolyBayes programs, so that program customization for different species could be easily performed.

- Text files parsing and image generation

Intuitive visualization of results is also an important point, since in larger sequences, or in sets of sequences, text output may become rather confusing and important results could be lost. Also for publication purposes, images or schemes are easily handled, and more suitable to show results.

- Read and Write access

Users must have read and write access in the directory where the application is installed.

4.3.2 Describing the Application

As described before this application is intended to be used by the researchers at EU Marie Curie Chair to map candidate SNPs in their sequences. They are the only entity that will interact with the application.

4.3.2.1 Sequences

Sequences to use for SNP analysis consist of two types of files, which are received from the sequencing company. One of the text files comprising the nucleotide bases and one text file with the quality scores for each nucleotide in the sequence. Files with quality scores are not mandatory, but recommended. The filenames for text sequences and quality

base scores should have the same filename, only different file extensions, for instance if the sequence text file is “aox1c.txt” the quality scores filename should be “aox1c.qual.1”

Sequences must be contiguous, that is, they must be from the same genomic location and there can be no gaps between them. The minimum number of sequences to perform a SNP analysis is two when one of them is the reference sequence, other studies that do not have a reference sequence consider four the minimum number of sequences for an SNP prediction [39]. However more input sequences should be available for the SNP prediction to be more accurate.

Since sequences come directly from the sequencing company, there is the necessity to clip some parts called “vector extremities”, (in the beginning and possibly in the ending of the sequence), the last 10 base pairs of the vector extremities need to be supplied.

4.3.3 Frameworks used

PolyBayes is a Perl script with some additional modules to perform the mathematical computations. These modules were created by PolyBayes author and are distributed with the program. Perl scripts are designed to be used in an UNIX / LINUX environment, so in order for PolyBayes to be used in Windows a framework with the core binaries of Perl, and the modules to run Perl scripts will be required. ActivePerl[1] is a framework from ActiveState Company for running Perl scripts on Windows, is freely available in under the terms of the ActivePerl Community License.

CAP3 has its own version for Windows.

Once again the programming language chosen was C++, and the IDE to code the interface is Visual Studio C++ 2005, so the .NET Framework Class Library will be used for image generation.

4.3.4 Application description

This is still an undergoing project; the image generation option is not yet implemented. Two Windows Form classes were implemented for user interface. The main Form allows the user to select the sequences, introduce the input data and execute the

[1] http://www.activestate.com/Products/activeperl/

programs, and will latter allow generation of the images with the results, it has also a text box for display information to the user.

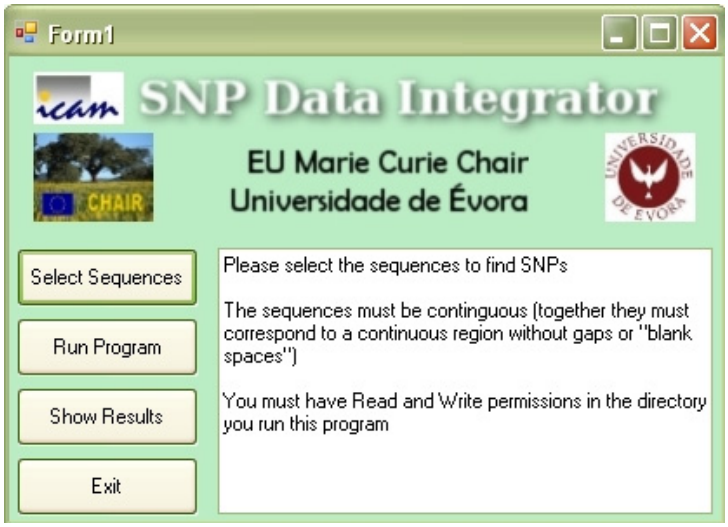


Figure 9 – Main Form, the user has the options to select the sequences, execute the programs to find SNPs and after generate an image based on output text results

Selecting the sequences to analyze will display a selection dialog box that filter the ".txt" and "phd.1" files (the extensions of the nucleotide bases and the quality scores). When user confirms file selection, data in selected files is checked for proper file format (verifications are not exhaustive), if data is in wrong file format a message is shown in a dialog box and the user must repeat the selection process. After file selection, user should hit the run program button, and a new Windows Form is then shown.

The Input Form has three main tabs, the first tab allows the user to introduce the "vector extremities" for clipping, the second controls the input supplied to CAP3 and the third shows the options to be supplied to PolyBayes.

After the DNA is sequenced, the sequence data files received from the sequencing company contain some data that needs to be removed (the cloning vector extremities). Since a sequence may have data regarding the nucleotide bases and the quality scores in separate files, the first step is to match the text sequences to the quality scores, the matching is done by filename, and sequence data identifiers inside the files. For each different filename (discarding extensions) the program asks for the extremities of the cloning vector in the tab "Sequence Clipping" (default values are already in the text input boxes).

The tab "Sequence assembly (cap3)" allows changing default input parameters of CAP3 for sequence assembly. Input parameter description is provided according CAP3

[37] documentation, however users should know the effects of the changes in the assembler parameters.

"SNP Screening (PolyBayes)" tab, allows users to select the input values that will be passed to PolyBayes, they are divided in four sections, options regarding the Anchored Multiple Alignment Algorithm (since CAP3 is used instead cross_match and PHRAP to sequence assembly, there is only one option in this section), the Paralog Filtering Algorithm, SNP screening and the creation of a file report. Options are described according to PolyBayes documentation [22].

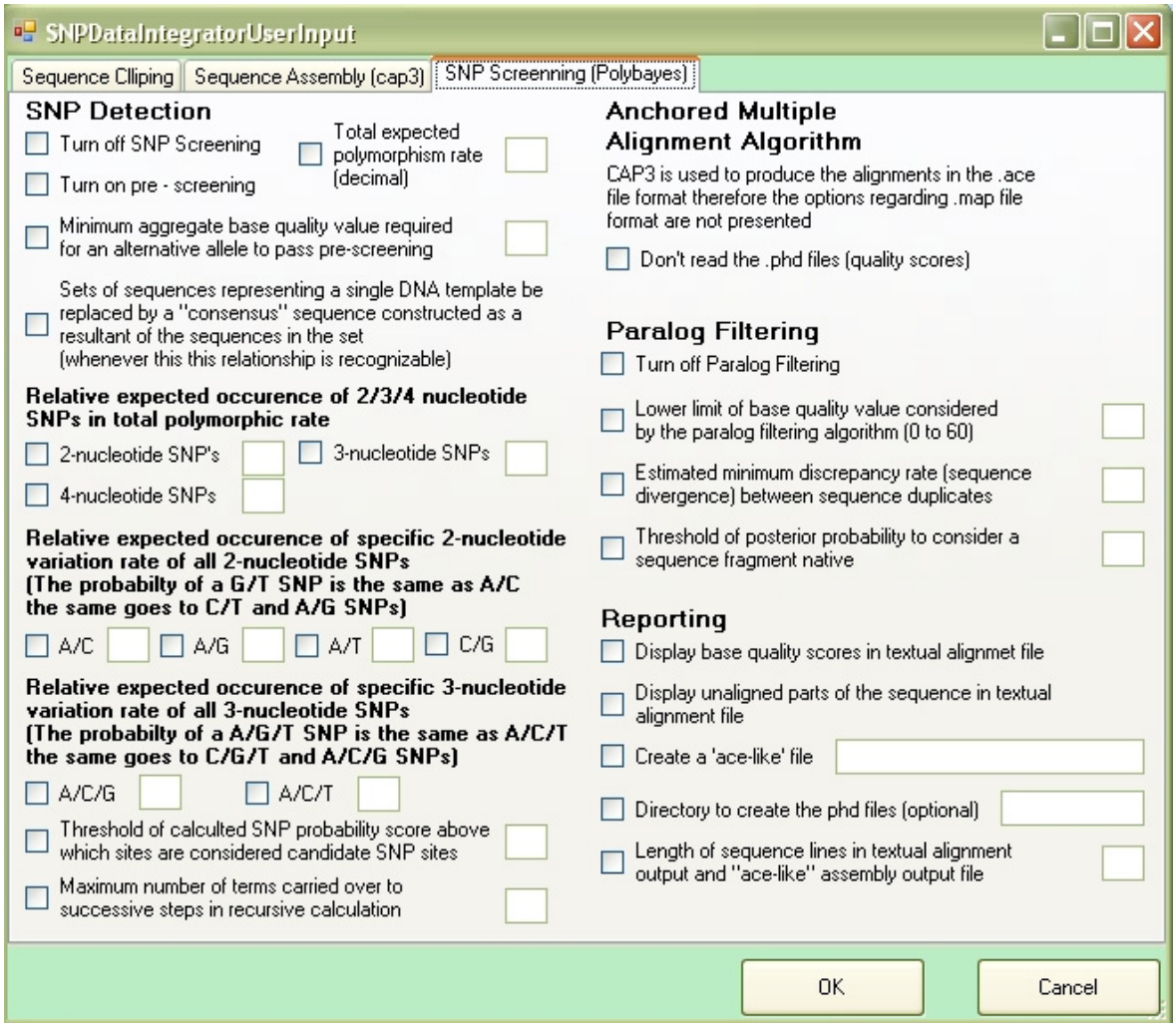


Figure 10 – The Input Form, the application identifies the sequences selected (text files and quality scores), asks for the "vector extremities", the input parameters for CAP3 and PolyBayes, after which runs the CAP3 and PolyBayes with the input parameters introduced by user.

After confirmation from the user, data before and after the extremities is removed, the program creates two files, with all the data of sequence base pairs files and all the data of quality scores, in CAP3 file input format. CAP3 is invoked on those files with the

options that user has selected, generates an ".ace" file with the assembly of the sequences and the quality scores for each sequence, once CAP3 finishes his execution PolyBayes is executed on the ".ace" file generated by CAP3 with the options selected by the user.

Depending on the number of sequences PolyBayes output could be very extensive, this output is indented to be parsed, to generate image schemes. As said before this part of the application is not yet implemented.

Testing phase has also not yet been performed, several studies have been conducted in order to map SNPs in some plants, using computer and experimental approaches [36] [39] [46], results are available on the Internet and will serve as basis to evaluate the accuracy of the application described before.

5 Conclusions

All the programs available that allow the analysis of S/MAR sequences, as well as literature that have already compared these programs were found in order to evaluate the strengths and weaknesses regarding each one of them.

Publicly available databases were used to determine the location of AOX genes in *Arabidopsis thaliana* and rice genomes. For Rice literature was used to locate and confirm the correct location of the genes, since AOX genes are not annotated in rice genome. Once the locus of the genes was found, the size of the S/MARs search area for *Arabidopsis thaliana* and rice was determined.

Several small applications were developed for data conversion between different formats, in order to use the previously selected S/MARTest program to locate S/MAR regions in the vicinity of AOX genes in *Arabidopsis thaliana* and Rice. Schematic images of the results were generated using an application developed for that purpose. Results were then presented using both schematic images and text.

At this point was considered important to have information about expression of AOX genes in both Arabidopsis and rice, to correlate with S/MARs predictions. This information was obtained from publicly available databases. Two databases were used based on data accuracy, and amount of data available.

A survey regarding previous work on computational SNP identification was performed. A long list of SNP scoring algorithms were identified, and the methodology selected was based in the program / algorithm PolyBayes, which was used for computational SNP detection in several studies to predict SNPs over large amounts of data in Humans [22] [34], and in soybean [35] and is the detection engine in two large scale SNP identification projects. PolyBayes assigns a Bayesian probability value to each SNP candidate position providing a more robust model than scoring models based on SNP definition [19].

Since one of the major goals were to create an application customized to use at EU Marie Curie Chair, another program was selected to perform the multiple sequence alignments (CAP3), and a graphic interface was developed to integrate both programs, allowing to use all the options of these two programs. The application is intended to

generate schematic images for the results although this functionality is not yet implemented.

As mentioned before this project suffered biases regarding the original planning. This was due to several reasons. The first is that this program is integrated, in a research project, so all the phases could not be completely compartmentalized in the original planning, which lead to the necessity of adding some phases, particularly the search for AOX gene expression concerning Arabidopsis and rice.

Another reason was the difficulties found when comparing the results for S/MARs identification with previous work, since results did not match and were due to different versions of the Arabidopsis Genome. This problem was not rapidly identified and caused some bias in the original planning.

Furthermore the context of this work requires a biological background, complemented by informatics skills, as means to provide tools for solving the biological problems, thus, a substantial part of the work was devoted to understand the biological concepts, which are not often trivial, and took more time than initially planned. Once the biological problems were depicted and the methodology established, solving the problems using informatics approaches were rather intuitive.

This still and undergoing project, so the further work will be to finish SNP application in order to generate image schemes from PolyBayes text output, testing the functionality of this application with the members of the workgroup. If any problems will be detected, by the test phase a second iteration of the development process will need to be undertaken to solve these problems.

Improvements in computational S/MARs finders are a major step for a better understanding of S/MARs function, and their implications in gene regulation. For this improvement is essential the experimental identification of more S/MARs sequences.

Improvements in SNP prediction programs could pass through integrating a machine learning approach, with already existent scoring algorithms. This approach was already proven to work, but the machine learning algorithm must be trained in large training sets of data, also differences in plant genomes make unfeasible to apply knowledge gained for some plants to another [36].

C++ was the chosen language to implement all the applications, since it allows an easy integration with libraries for rapid image generation, has a many functions for string and Input/Output handling.

6 References

[1] – The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796-815

[2] – International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793-800

[3] – Saisho D, Nambara E, Naito S, Tsutsumi N, Hirai A, Nakazono M (1997) Characterization of the gene family for Alternative Oxidase from *Arabidopsis thaliana*. Plant Molecular Biology 35:585-596

[4] – Ito Y, Saisho D, Nakazono M, Tsutsumi N, Hirai A (1997) Transcript levels of tandem-arranged alternative oxidase genes in rice are increased by low temperature. Gene 203:121-129

[5] – Arnholdt-Schmitt B, Costa JH, Fernandes de Melo D (2006) AOX A functional marker for efficient cell reprogramming under stress? Trends Plant Sci 11:281-287

[6] - Arnholdt-Schmitt B (2006) Efficient cell Reprogramming as a target for functional marker strategies? Towards new prespectives in applied plant nutrition research. J. Plant. Nutr. Soil Sci 168:617-624

[7] – Lodish H, Berk A, Matsudaira P, Kaiser CA, Krieger M, Scott MP, Zipursky L, Darnell J (2003) Molecular Cell Biology. W. H. Freeman 5th Edition.

[8] – Girod PA, Mermond N (2003) Use of Scaffold Matrix Attachment Regions for protein production. Gene Transfer and expression in Mammalian Cells. Chapter 10:359-377 Elsevier Science.

[9] – Rudd S, Frish M, Grote K, Meyers BC, Mayer K, Werner T (2004) Genome Wide In Silico Mapping of Scaffold / Matrix Attachment Regions in Arabidopsis Suggests Correlation of Intagenic Scaffold / Matrix Attachment Regions with gene expression. Plant Physiol 135: 715–722.

[10] – Ganapathi M, Srivastava P, Sutar SKD, Kumar K, Dasgupta D, Singh GP, Brahmachari V, Bramachari SK (2005) Comparative analysis of chromatin landscape in regulatory regions of human housekeeping and tissue specific genes. BMC Bioinformatics 6:126.

[11] – Arnholdt-Schmitt B, (2005) Functional markers and a 'systemic strategy': convergency between plant breeding, plant nutrition and molecular biology. Plant Physiology and Biochemistry 43:817-820

[12] – Nowak W, Gawloska M, Jarmoloswski A, Augustyniak J (2001) Effect of nuclear matrix attachment regions on transgene expression in tobacco plants. Acta Biochimica Polonica vol 48 3:637:646

[13] – Tetko IV, Haberer G, Rudd S, Meyers B, Mewes HW, Mayer K (2006) Spatiotemporal expression control correlates with intragenic scaffold matrix attachment regions (S/MARs) in *Arabidopsis thaliana*. PLoS Comput Biol 2(3): e21.

[14] – Leibich I, Bode J, Reuter I, Wingender E (2002) Evaluation of sequence motifs found in Scaffold / Matrix Attached Regions (S/MARs). Nucleic Acids Research, Vol. 30 15:3433-3442

[15] – Singh GB, Kramer JA, Krawetz SA (1997) Mathematical model to predict regions of chromatin attachment to the nuclear matrix. Nucleic Acids Research Vol 25 7:1419-1425

[16] – Purbowasito W, Suda C, Yokomine T, Zubair M, Sado T, TsuTsui K, Sasaki H (2004) Large-Scale Identification and Mapping of Nuclear Matrix-Attachment Regions in the Distal Imprinted Domain of Mouse Chromosome 7. DNA Research 11:391-407

[17] – Frish M, Frech K, Klingenhoff A, Chartharius K, Leibich I, Werner T (2003) In Silico Prediction of Scaffold/Matrix Attachment Regions in Large Genomic Sequences. Genome Res 12:349-354

[18] – Meyers BC, Galbraith DW, Nelson T, Agrawal V (2004) Methods for Transcriptional Profiling in Plants. Be Fruitful and Replicate. Plant Physiology Vol 135:637-652

[19] – Kwok PY (2003) Single Nucleotide Polymorphisms, Methods and Protocols. Humana Press Inc.

[20] – Saika H, Ohtsu K, Hamanaka S, Nakazono M, Tsutsumi N, Hirai H (2002) *AOX1c*,a novel rice gene for alternative oxidase; comparison with rice *AOX1a* and *AOX1b*. Genes Genet. Syst. 77:31-38

[21] – Garg K, Green P, Nickerson DA (1999) Identification of Candidate Coding Region Single Nucleotide Polymorphisms in 165 Human Genes Using Assembled Expressed Sequence Tags. Genome Res. 9:1087-1092

[22] – Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier LD, Kwok PY, Gish WR (1999) A general approach to single-nucleotide polymorphism discovery. Nature Genetics 23:452-456

[23] – Michalowski SM, Allen GC, Hall Jr. GE, Thompson WF, Spiker S (1999) Characterization of Randomly-Obtained Matrix Attachment Regions (MARs) from Higher Plants. Biochemistry 38:12795-12804

[24] – Chernov IP, Akopov SB, Nikolaev LG (2004) Structure and Functions of Nuclear Matrix Associated Regions (S/MARs). Russian Journal of Bioorganic Chemistry, Vol 30 1:3-14

[25] – Whelan J, Millar AH, Day DA (1996) The alternative oxidase is encoded in a multigene family in soybean. Plant 198:197-201

[26] – Paul AL, Ferl RJ (1998) Higher Order Chromatin Structures in Maize and Arabidopsis. The Plant Cell 10:1349-1359

[27] – Clifton R, Millar AH, Whelan J (2006) Alternative oxidases in Arabidopsis: A comparative analysis of differential expression in the gene family provides new insights into function of non-phosphorylating bypasses. Biochimica et Biophisica Acta 1757:730-741

[28] – Zimmermann P, Hirsch-Hoffman M, Gruissem W (2004) GENEVESTIGATOR. Arabidopsis Microarray Database and Analysis Toolbox. Plant Physiology 136:2621-2632

[29] – Nobuta K, Venu RC, Lu C, Bélo A, Vemaraju K, Kulkarni K, Wang W, Pillay M, Green PJ, Wang G, Meyers BC (2006) An expression atlas of rice mRNAs and small RNAs. Nature Biotechnology 25:473-477

[30] – Saika H, Ohtsu K, Hamanaka S, Nakazono M, Tsutsumi N, Hirai A (2002) *AOX1c*,a novel rice gene for alternative oxidase; comparison with rice *AOX1a* and *AOX1b*. Genes Genet. Syst 77:31-38

[31] - Nakano M, Nobuta K, Vemaraju K, Tej SS, Skogen JW, Meyers BC (2006) Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Research 34, D731-D735

[32] - Tan PK, Downey TJ, Spitznagel Jr. EL, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC (2003) Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Research 31, D5676-D5684

[33] - Causton HC, Quackenbush J, Brazama A (2003) Microarray Gene Expression Data Analysis, A beginner's guide. Blackwell Science Ltd.

[34] - The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928-933

[35] – Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2002) Single Nucleotide Polymorphisms in Soybean

[36] – Matukumalli LK, Grefenstette JJ, Hyten DL, Choi IY, Cregan PB, Tassel CPV (2006) Application of machine learning in SNP discovery. BMC Bioinformatics 7:4

[37] – Huang X, Madan A (1999) CAP3: A DNA Sequence Assembly Program. Genome Res. 9:868-877

[38] – Verjovski-Almeida S, DeMarco R, Martins EAL et al (2003) Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*. Nature Genetics 35:148-157

[39] – Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003) Mining for Single Nucleotide Polymorphisms and Insertions/Deletions in Maize Expressed Sequence Tag Data. Plant Physiology 132:84-91

[40] - Borecky J, Vercesi AE (2005) Plant uncoupling mitochondrial protein and alternative oxidase: energy metabolism and stress. Bioscience Rep. 35: 271–286

[41] - Juszczuk IM, Rychter AM (2003) Alternative Oxidase in higher plants Acta Biochimica Polonica Vol. 50 4:1257-1271

[42] - Vanlerberghe GC (1997) Alternative Oxidase: From Gene to Function. Plant Physiology Plant Molecular Biology 48:703-734

[43] - Umbach AL, Ng VS, Siedow JN (2006) Regulation of plant alternative oxidase activity: A tale of two cysteines. Biochimica et Biophysica Acta 1757:135-142

[44] – Venter JC, Smith HO, Hood L, (1996) A new strategy for genome sequencing. Nature 381, 364 - 366

[45] – Salzberg SL, Yorke JA (2005) Beware of mis-assembled genomes. Bioinformatics 21:4320-4321.

[46] – Schmid KJ, Sörensen TR, Stracke R, Törjék O, Altmann T, Mitchell-Olds T, Weisshaar B (2003) Large-Scale Identification and Analysis of Genome-Wide Single-Nucleotide Polymorphisms for Mapping in *Arabidopsis thaliana*. Genome Res. 13:1250-1257.

Index

Appendix 1(Confidential)

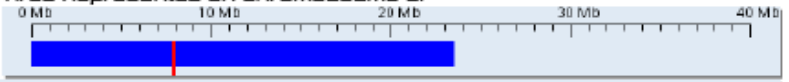
In this Appendix are presented the S/MARs identification results.

The Schematic images and text is presented together, for all AOX genes in *Arabidopsis Thaliana* and *Oryza sativa* (Rice).

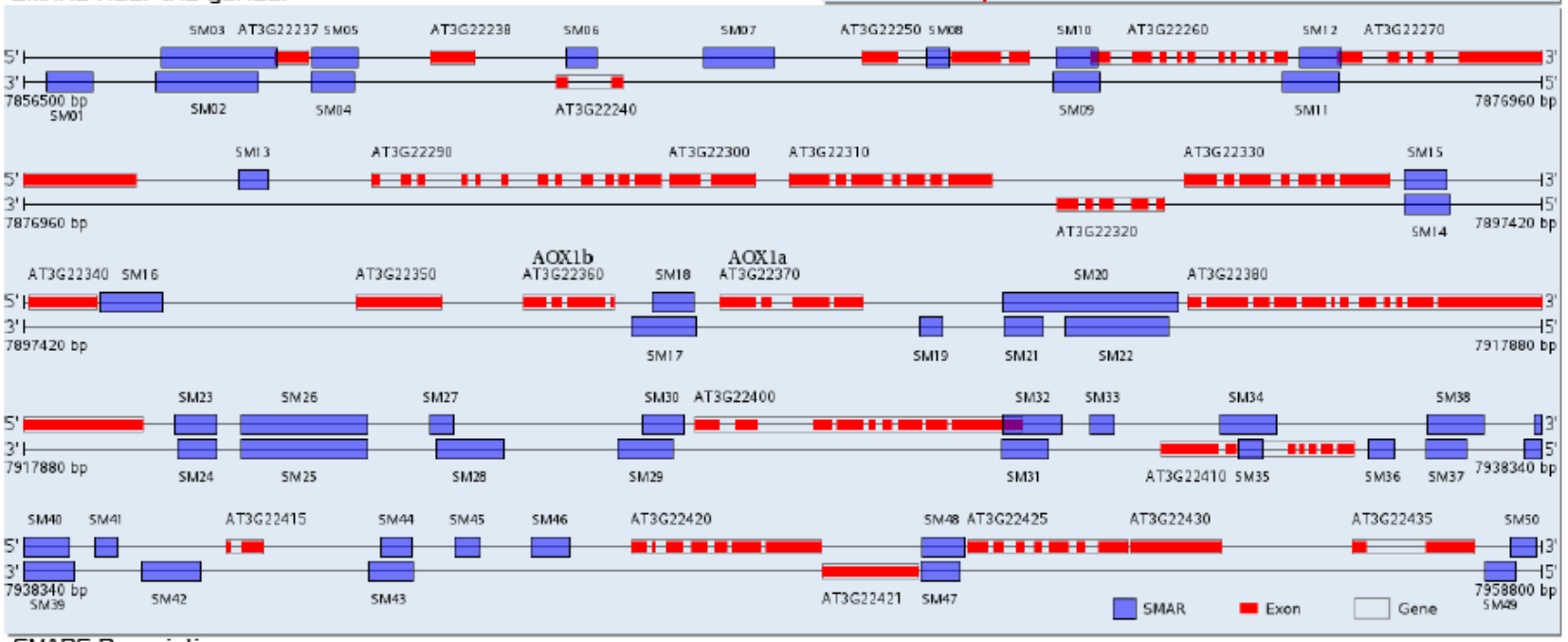
Results comprise a generated image depicting S/MAR sequences in the neighborhood of AOX genes for Arabidopsis and Rice, an image concerning the positioning of the analyzed section in the corresponding chromosome, position and strand of each S/MAR and gene in the chromosome, and the location relative to AOX genes. It is also presented a description of each gene according with database annotations.

AOX1a and AOX1b Genes in *Arabidopsis Thaliana:*

Area Represented on Chromossome 3:



SMARs near the genes:



SMARS Description:

Forward Strand:

Start	End	Length	Id	Position	Start	End	Length	Id	Position
7858361	7859910	1550	SM03	AOX1a-16	7923356	7923680	325	SM27	AOX1a+9
7860396	7861010	615	SM05	AOX1a-14	7926231	7926785	555	SM30	AOX1a+12
7863826	7864225	400	SM06	AOX1a-13	7931081	7931880	800	SM32	AOX1a+14
7865666	7866625	960	SM07	AOX1a-12	7932261	7932590	330	SM33	AOX1a+15
7868686	7868985	300	SM08	AOX1a-11	7934011	7934770	760	SM34	AOX1a+16
7870416	7870975	560	SM10	AOX1a-9	7936806	7937570	765	SM38	AOX1a+20
7873711	7874280	570	SM12	AOX1a-7	7938256	7938965	710	SM40	AOX1a+22
7879866	7880255	390	SM13	AOX1a-6	7939301	7939600	300	SM41	AOX1a+23
7895576	7896155	580	SM15	AOX1a-4	7943156	7943575	420	SM44	AOX1a+26
7898461	7899295	835	SM16	AOX1a-3	7944161	7944500	340	SM45	AOX1a+27
7905896	7906460	565	SM18	AOX1a-1	7945181	7945705	525	SM46	AOX1a+28
7910626	7913000	2375	SM20	AOX1a+2	7950456	7951030	575	SM48	AOX1a+30
7919916	7920485	570	SM23	AOX1a+5	7958391	7958745	355	SM50	AOX1a+32
7920821	7922515	1695	SM26	AOX1a+8					

Reverse Strand:

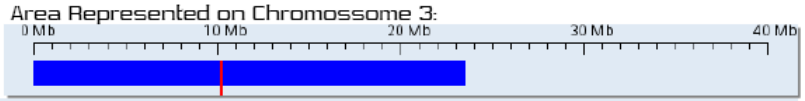
Start	End	Length	Id	Position	Start	End	Length	Id	Position
7856801	7857440	640	SM01	AOX1a-18	7931066	7931700	635	SM31	AOX1a+13
7858281	7859670	1390	SM02	AOX1a-17	7934261	7934595	335	SM35	AOX1a+17
7860381	7860950	570	SM04	AOX1a-15	7936011	7936370	360	SM36	AOX1a+18
7870386	7871020	635	SM09	AOX1a-10	7936796	7937345	550	SM37	AOX1a+19
7873481	7874235	755	SM11	AOX1a-8	7938121	7939025	905	SM39	AOX1a+21
7895571	7896200	630	SM14	AOX1a-5	7939931	7940730	800	SM42	AOX1a+24
7905611	7906485	875	SM17	AOX1a-2	7943001	7943585	585	SM43	AOX1a+25
7909506	7909815	310	SM19	AOX1a+1	7950431	7950965	535	SM47	AOX1a+29
7910641	7911175	535	SM21	AOX1a+3	7958041	7958465	425	SM49	AOX1a+31
7911476	7912865	1390	SM22	AOX1a+4					
7919951	7920480	530	SM24	AOX1a+6					
7920811	7922515	1705	SM25	AOX1a+7					
7923441	7924350	910	SM28	AOX1a+10					
7925886	7926645	760	SM29	AOX1a+11					

AOX1a e AOX1b Genes in *Arabidopsis Thaliana:*

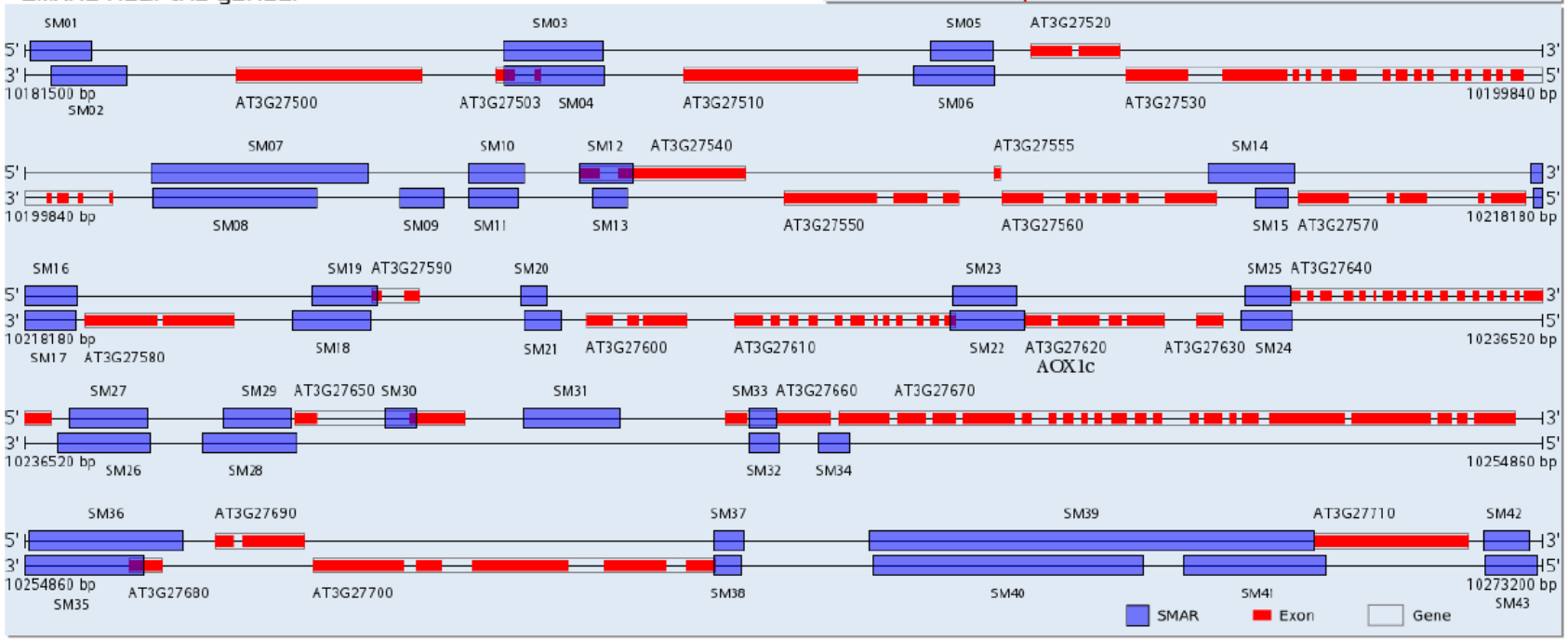
Gene Description

Gene	Begin	End	Strand	Description
AT3G22237	7859884	7860338	Forward	Pseudogene, putative ribulose 1,5-bisphosphate carboxylase, similar to ribulose 1,5-bisphosphate carboxylase
AT3G22238	7862002	7862575	Forward	Pseudogene, putative ribulose 1,5-bisphosphate carboxylase, similar to ribulose 1,5-bisphosphate carboxylase
AT3G22240	7863684	7864586	Reverse	Expressed Protein
AT3G22250	7867813	7870060	Forward	UDP-glucoronosyl/UDP-glucosyl transferase family protein
AT3G22260	7870899	7873551	Forward	OTU-like cysteine protease family protein
AT3G22270	7874228	7878475	Forward	Expressed Protein
AT3G22290	7881672	7885549	Forward	Expressed Protein
AT3G22300	7885675	7886827	Forward	40S ribosomal protein S10, mitochondrial precursor
AT3G22310	7887293	7890026	Forward	DEAD box RNA helicase, putative (RH9)
AT3G22320	7890893	7892330	Reverse	DNA-directed RNA polymerase
AT3G22330	7892623	7895373	Forward	DEAD box RNA helicase, putative
AT3G22340	7897491	7898405	Forward	Copia-like retrotransposon family
AT3G22350	7901912	7903048	Forward	F-box family protein, similar to F-box protein family
AT3G22360	7904163	7905391	Forward	Alternative Oxidase 1b (AOX1b)
AT3G22370	7906800	7908747	Forward	Alternative Oxidase 1a (AOX1a)
AT3G22380	7913126	7919491	Forward	Expressed Protein
AT3G22400	7926941	7931358	Forward	Lipoxygenase, putative
AT3G22410	7933213	7935815	Reverse	Expressed Protein
AT3G22415	7941070	7941571	Forward	Expressed Protein
AT3G22420	7946533	7949103	Forward	Protein kinase family
AT3G22421	7949123	7950403	Reverse	F-box family protein
AT3G22425	7951062	7953230	Forward	Imidazoleglycerol-phosphate dehydratase 1
AT3G22430	7953267	7954490	Forward	Expressed Protein
AT3G22435	7956255	7957891	Forward	XS domain-containing protein

AOX1c Gene in *Arabidopsis Thaliana:*



SMARs near the genes:



SMARS Description:

Forward Strand:

Start	End	Length	Id	Position	Start	End	Length	Id	Position
10181578	10182317	740	SM01	AOX1C-23	10238933	10239747	815	SM29	AOX1C+6
10187288	10188487	1200	SM03	AOX1C-21	10240893	10241257	365	SM30	AOX1C+7
10192443	10193202	760	SM05	AOX1C-19	10242543	10243712	1170	SM31	AOX1C+8
10201378	10203997	2620	SM07	AOX1C-17	10245273	10245617	345	SM33	AOX1C+10
10205213	10205887	675	SM10	AOX1C-14	10254923	10256787	1865	SM36	AOX1C+13
10206553	10207197	645	SM12	AOX1C-12	10263193	10263552	360	SM37	AOX1C+14
10214143	10215197	1055	SM14	AOX1C-10	10265068	10270457	5390	SM39	AOX1C+16
10218043	10218822	780	SM16	AOX1C-8	10272503	10273037	535	SM42	AOX1C+19
10221663	10222437	775	SM19	AOX1C-5					
10224173	10224507	335	SM20	AOX1C-4					
10229398	10230162	765	SM23	AOX1C-1					
10232918	10233477	560	SM25	AOX1C+2					
10237078	10238007	930	SM27	AOX1C+4					

Reverse Strand:

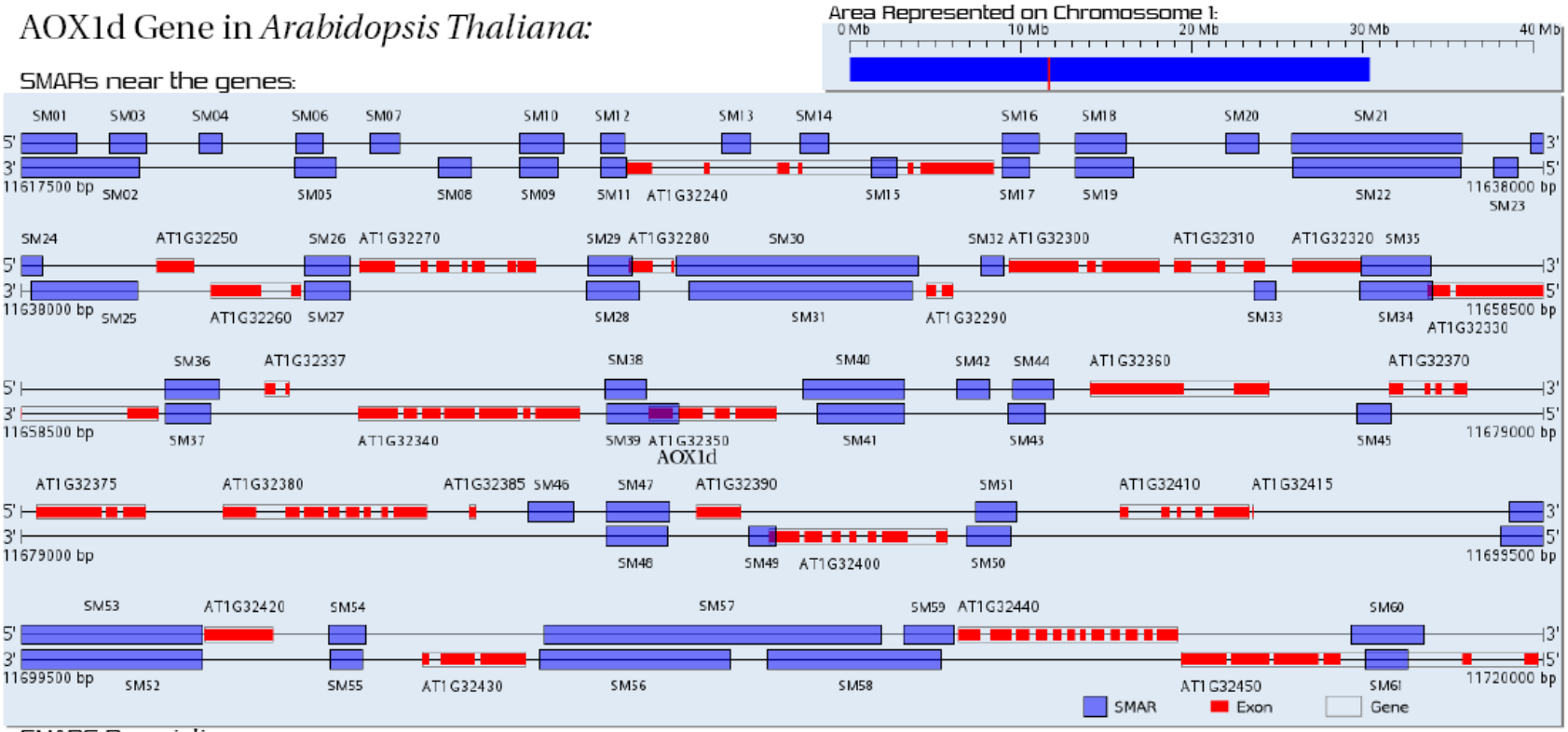
Start	End	Length	Id	Position	Start	End	Length	Id	Position
10181828	10182737	910	SM02	AOX1C-22	10236933	10238042	1110	SM26	AOX1C+3
10187298	10188507	1210	SM04	AOX1C-20	10238678	10239822	1145	SM28	AOX1C+5
10192248	10193217	970	SM06	AOX1C-18	10245268	10245642	375	SM32	AOX1C+9
10201398	10203372	1975	SM08	AOX1C-16	10246118	10246487	370	SM34	AOX1C+11
10204378	10204912	535	SM09	AOX1C-15	10254873	10256312	1440	SM35	AOX1C+12
10205218	10205807	590	SM11	AOX1C-13	10263203	10263527	325	SM38	AOX1C+15
10206713	10207132	420	SM13	AOX1C-11	10265118	10268367	3250	SM40	AOX1C+17
10214713	10215102	390	SM15	AOX1C-9	10268868	10270587	1720	SM41	AOX1C+18
10218063	10218812	750	SM17	AOX1C-7	10272513	10273132	620	SM43	AOX1C+20
10221423	10222367	945	SM18	AOX1C-6					
10224228	10224672	445	SM21	AOX1C-3					
10229373	10230272	900	SM22	AOX1C-2					
10232878	10233497	620	SM24	AOX1C+1					

AOX1c Genes in *Arabidopsis Thaliana:*

Gene Description

Gene	Begin (bp)	End (bp)	Strand	Description
AT3G27500	10184070	10186318	Reverse	Similar to DC1 domain-containing protein
AT3G27503	10187197	10187736	Reverse	Encodes a member of a family of small, secreted, cysteine rich proteins with sequence similarity to SCR
AT3G27510	10189476	10191580	Reverse	Similar to DC1 domain-containing protein
AT3G27520	10193659	10194736	Forward	Unknown protein
AT3G27530	10194806	10200913	Reverse	Vesicle tethering family protein
AT3G27540	10206545	10208543	Forward	Glycosyl transferase family 17 protein
AT3G27550	10209018	10211136	Reverse	Group II intron splicing factor CRS1-related
AT3G27555	10211565	10211636	Forward	tRNA-Asp (anticodon: GTC)
AT3G27560	10211668	10214241	Reverse	Protein kinase (ATN1)
AT3G27570	10215233	10217973	Reverse	Expressed protein
AT3G27580	10218908	10220721	Reverse	A member of a subfamily of Ser/Thr PKs in plants
AT3G27590	10222384	10222952	Forward	Expressed protein
AT3G27600	10224980	10226185	Reverse	Expressed protein
AT3G27610	10226779	10229428	Reverse	Expressed protein
AT3G27620	10230282	10231939	Reverse	Alternative Oxidase 1c (AOX1c)
AT3G27630	10232353	10232655	Reverse	Expressed protein
AT3G27640	10233485	10236858	Forward	Transducin family protein / WD-40 repeat family protein
AT3G27650	10239783	10241848	Forward	LOB domain protein 25 / lateral organ boundaries domain protein 25 (LBD25)
AT3G27660	10244995	10246260	Forward	Encodes oleosin4, a protein found in oil bodies, involved in seed lipid accumulation
AT3G27670	10246371	10254530	Forward	Expressed protein
AT3G27680	10256123	10256527	Reverse	Self-incompatibility protein-related
AT3G27690	10257184	10258248	Forward	Chlorophyll A-B binding protein
AT3G27700	10258354	10263211	Reverse	RNA recognition motif (RRM)-containing protein
AT3G27710	10270443	10272302	Forward	Zinc finger protein-related

AOX1d Gene in *Arabidopsis Thaliana:*



SMARS Description:

Forward Strand:

Start	End	Length	Id	Position	Start	End	Length	Id	Position
11617513	11618262	750	SM01	AOX1d-39	11646833	11650082	3250	SM30	AOX1d-10
11618678	11619187	510	SM03	AOX1d-37	11650938	11651247	310	SM32	AOX1d-8
11619888	11620212	325	SM04	AOX1d-36	11656063	11656997	935	SM35	AOX1d-5
11621203	11621562	360	SM06	AOX1d-34	11660423	11661172	750	SM36	AOX1d-4
11622198	11622592	395	SM07	AOX1d-33	11666363	11666927	565	SM38	AOX1d-2
11624228	11624817	590	SM10	AOX1d-30	11669028	11670407	1380	SM40	AOX1d+1
11625318	11625627	310	SM12	AOX1d-28	11671108	11671547	440	SM42	AOX1d+3
11626938	11627332	395	SM13	AOX1d-27	11671863	11672407	545	SM44	AOX1d+5
11627998	11628372	375	SM14	AOX1d-26	11685833	11686457	625	SM46	AOX1d+7
11630713	11631212	500	SM16	AOX1d-24	11686873	11687742	870	SM47	AOX1d+8
11631703	11632397	695	SM18	AOX1d-22	11691868	11692417	550	SM51	AOX1d+12
11633733	11634177	445	SM20	AOX1d-20	11699068	11701952	2885	SM53	AOX1d+14
11634633	11636922	2290	SM21	AOX1d-19	11703653	11704152	500	SM54	AOX1d+15
11637848	11638287	440	SM24	AOX1d-16	11706553	11711087	4535	SM57	AOX1d+18
11641818	11642437	620	SM26	AOX1d-14	11711388	11712067	680	SM59	AOX1d+20
11645628	11646227	600	SM29	AOX1d-11	11717428	11718407	980	SM60	AOX1d+21

Reverse Strand:

Start	End	Length	Id	Position	Start	End	Length	Id	Position
11617363	11619107	1745	SM02	AOX1d-38	11660443	11661062	620	SM37	AOX1d-3
11621183	11621747	565	SM05	AOX1d-35	11666403	11667362	960	SM39	AOX1d-1
11623128	11623567	440	SM08	AOX1d-32	11669223	11670402	1180	SM41	AOX1d+2
11624218	11624732	515	SM09	AOX1d-31	11671798	11672292	495	SM43	AOX1d+4
11625308	11625667	360	SM11	AOX1d-29	11676503	11676977	475	SM45	AOX1d+6
11628953	11629312	360	SM15	AOX1d-25	11686883	11687702	820	SM48	AOX1d+9
11630718	11631092	375	SM17	AOX1d-23	11688803	11689162	360	SM49	AOX1d+10
11631713	11632492	780	SM19	AOX1d-21	11691738	11692332	595	SM50	AOX1d+11
11634643	11636902	2260	SM22	AOX1d-18	11698943	11701952	3010	SM52	AOX1d+13
11637348	11637672	325	SM23	AOX1d-17	11703663	11704107	445	SM55	AOX1d+16
11638143	11639567	1425	SM25	AOX1d-15	11706483	11709062	2580	SM56	AOX1d+17
11641828	11642432	605	SM27	AOX1d-13	11709558	11711907	2350	SM58	AOX1d+19
11645613	11646327	715	SM28	AOX1d-12	11717623	11718192	570	SM61	AOX1d+22
11647013	11650012	3000	SM31	AOX1d-9					
11654618	11654917	300	SM33	AOX1d-7					
11656033	11657017	985	SM34	AOX1d-6					

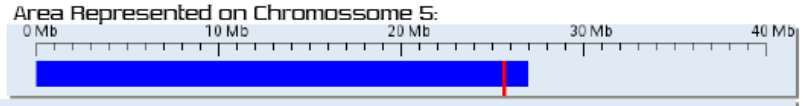
AOX1d Genes in *Arabidopsis Thaliana:*

Gene Description

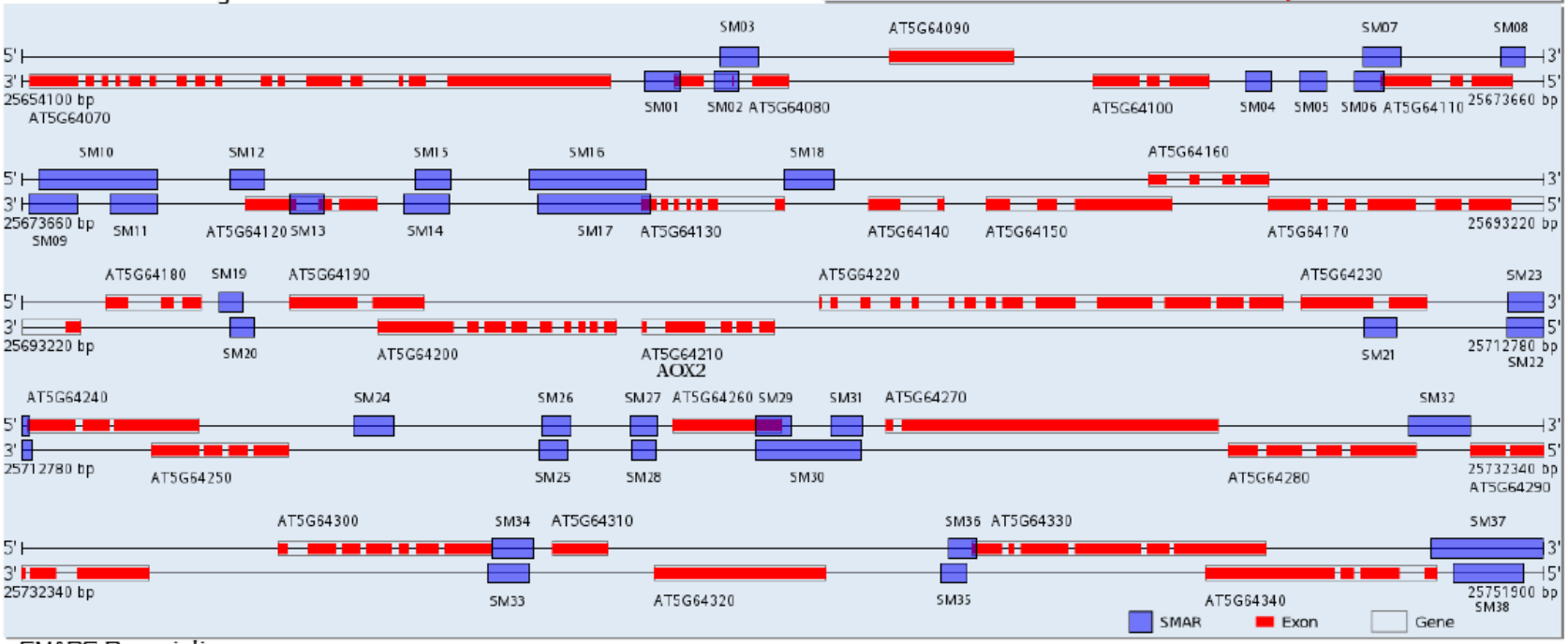
Gene	Begin (bp)	End (bp)	Strand	Description
AT1G32240	11625680	11630606	Reverse	Encodes a member of the KANADI family of putative transcription factors
AT1G32250	11639823	11640323	Forward	Calmodulin, putative
AT1G32260	11640562	11641764	Reverse	Expressed protein
AT1G32270	11642573	11644942	Forward	Syntaxin, putative
AT1G32280	11646200	11646796	Forward	Protease inhibitor / seed storage / lipid transfer protein (LTP) family protein
AT1G32290	11650202	11650547	Reverse	Expressed protein
AT1G32300	11651322	11653345	Forward	FAD-binding domain-containing protein
AT1G32310	11653544	11654771	Forward	Expressed protein
AT1G32320	11655136	11656053	Forward	Mitogen-activated protein kinase kinase (MAPKK), putative (MKK10
AT1G32330	11656966	11660356	Reverse	Heat shock transcription factor family protein
AT1G32337	11661795	11662102	Forward	Expressed protein
AT1G32340	11663050	11666038	Reverse	Zinc finger (C3HC4-type RING finger) family protein
AT1G32350	11666957	11668670	Reverse	Alternative Oxidase 1d (AOX1d)
AT1G32360	11672911	11675319	Forward	Zinc finger (CCCH-type) family protein
AT1G32370	11676945	11677987	Forward	Tobamovirus multiplication protein 2B (TOM2B)
AT1G32375	11679210	11680659	Forward	F-box family protein
AT1G32380	11681732	11684458	Forward	Ribose-phosphate pyrophosphokinase 2 / phosphoribosyl diphosphate synthetase 2 (PRS2)
AT1G32385	11685040	11685124	Forward	snoRNA
AT1G32390	11688109	11688687	Forward	Hypothetical protein
AT1G32400	11689089	11691478	Reverse	Senescence-associated family protein
AT1G32410	11693812	11695546	Forward	Vacuolar protein sorting 55 family protein / VPS55 family proteinacuolar protein sorting 55 family protein / VPS55 family protein
AT1G32415	11695591	11697876	Forward	Pentatricopeptide (PPR) repeat-containing protein
AT1G32420	11701975	11702883	Forward	F-box family protein
AT1G32430	11704912	11706305	Reverse	F-box family protein
AT1G32440	11712142	11715092	Forward	Pyruvate kinase, putative, similar to pyruvate kinase isozyme G
AT1G32450	11715130	11719935	Reverse	Proton-dependent oligopeptide transport (POT) family protein

AOX2 Gene in *Arabidopsis Thaliana:*

Area Represented on Chromossome 5:



SMARs near the genes:



SMARS Description:

Forward Strand:

Start	End	Length	Id	Position	Start	End	Length	Id	Position
25663077	25663571	495	SM03	AOX2-18	25722212	25722676	465	SM29	AOX2+9
25671337	25671831	495	SM07	AOX2-14	25723172	25723591	420	SM31	AOX2+11
25673112	25673426	315	SM08	AOX2-13	25730607	25731396	790	SM32	AOX2+12
25673872	25675406	1535	SM10	AOX2-11	25738392	25738906	515	SM34	AOX2+14
25676337	25676786	450	SM12	AOX2-9	25744262	25744611	350	SM36	AOX2+16
25678712	25679176	465	SM15	AOX2-6	25750447	25751886	1440	SM37	AOX2+17
25680182	25681681	1500	SM16	AOX2-5					
25683462	25684111	650	SM18	AOX2-3					
25695752	25696066	315	SM19	AOX2-2					
25712322	25712881	560	SM23	AOX2+3					
25717037	25717571	535	SM24	AOX2+4					
25719467	25719836	370	SM26	AOX2+6					
25720607	25720951	345	SM27	AOX2+7					

Reverse Strand:

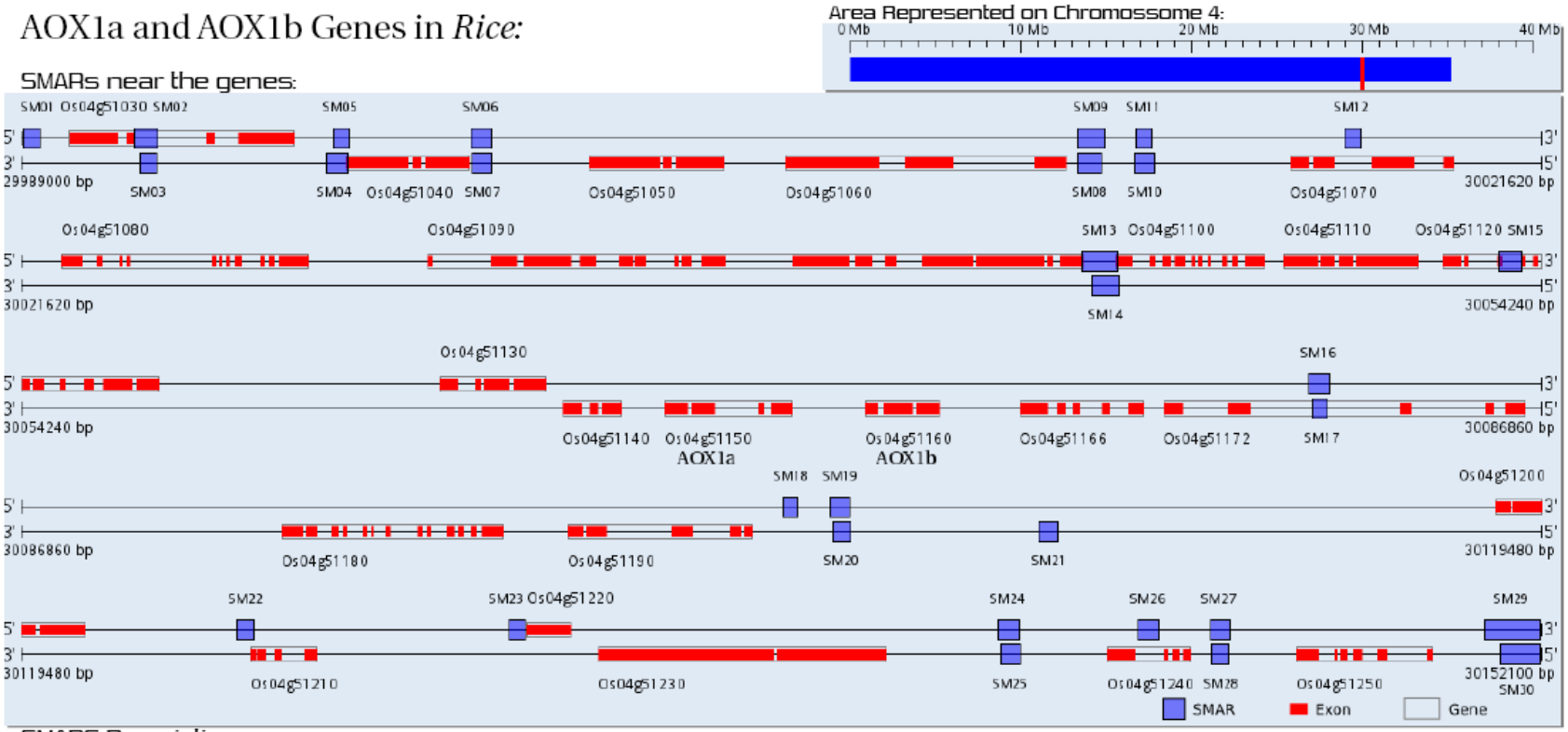
Start	End	Length	Id	Position	Start	End	Length	Id	Position
25662102	25662561	460	SM01	AOX2-20	25719427	25719791	365	SM25	AOX2+5
25662992	25663306	315	SM02	AOX2-19	25720617	25720931	315	SM28	AOX2+8
25669822	25670151	330	SM04	AOX2-17	25722212	25723571	1360	SM30	AOX2+10
25670522	25670866	345	SM05	AOX2-16	25738322	25738866	545	SM33	AOX2+13
25671232	25671606	375	SM06	AOX2-15	25744167	25744486	320	SM35	AOX2+15
25673742	25674381	640	SM09	AOX2-12	25750732	25751646	915	SM38	AOX2+18
25674792	25675406	615	SM11	AOX2-10					
25677092	25677541	450	SM13	AOX2-8					
25678577	25679151	575	SM14	AOX2-7					
25680292	25681736	1445	SM17	AOX2-4					
25695897	25696201	305	SM20	AOX2-1					
25710482	25710896	415	SM21	AOX2+1					
25712307	25712921	615	SM22	AOX2+2					

AOX2 Genes in *Arabidopsis Thaliana:*

Gene Description

Gene	Begin (bp)	End (bp)	Strand	Description
AT5G64070	25654202	25661672	Reverse	Phosphatidylinositol 4-kinase (PI4K)
AT5G64080	25662489	25663954	Reverse	Protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
AT5G64090	25665253	25666846	Forward	Expressed protein
AT5G64100	25667867	25669349	Reverse	Peroxidase, putative
AT5G64110	25671571	25673256	Reverse	Peroxidase, putative
AT5G64120	25676532	25678228	Reverse	Peroxidase, putative
AT5G64130	25681626	25683460	Reverse	Expressed protein
AT5G64140	25684550	25685518	Reverse	40S ribosomal protein S28
AT5G64150	25686057	25688444	Reverse	Methylase family protein
AT5G64160	25688156	25689689	Forward	Expressed protein
AT5G64170	25689692	25693972	Reverse	Dentin sialophosphoprotein-related
AT5G64180	25694306	25695518	Forward	Expressed Protein
AT5G64190	25696658	25698379	Forward	Expressed protein
AT5G64200	25697792	25700858	Reverse	Arginine/serine-rich splicing factor SC35
AT5G64210	25701191	25702890	Reverse	Alternative Oxidase 2 (AOX2)
AT5G64220	25703472	25709441	Forward	Calmodulin-binding protein
AT5G64230	25709679	25711279	Forward	Expressed Protein
AT5G64240	25712846	25715046	Forward	Latex-abundant family protein (AMC3) / caspase family protein
AT5G64250	25714450	25716197	Reverse	2-nitropropane dioxygenase family / NPD family
AT5G64260	25721155	25722542	Forward	Phosphate-responsive protein, putative
AT5G64270	25723885	25728151	Forward	Splicing factor, putative
AT5G64280	25728295	25730699	Reverse	Oxoglutarate/malate translocator, putative
AT5G64290	25731404	25733984	Reverse	Oxoglutarate/malate translocator, putative
AT5G64300	25735630	25738388	Forward	Similar to riboflavin biosynthesis protein, putative
AT5G64310	25739159	25739872	Forward	Arabinogalactan-protein (AGP1)
AT5G64320	25740469	25742677	Reverse	Pentatricopeptide (PPR) repeat-containing protein
AT5G64330	25744563	25748328	Forward	Non-phototropic hypocotyl 3 (NPH3)
AT5G64340	25747561	25750539	Reverse	Encodes a bHLH(basic helix-loop-helix)-type transcription factor SAC51 (suppressor of acaulis 51)

AOX1a and AOX1b Genes in *Rice:*



SMARS Description:

Forward Strand:

Start	End	Length	Id	Position	Start	End	Length	Id	Position
29989053	29989397	345	SM01	AOX1a-15	30124118	30124472	355	SM22	AOX1a+7
29991433	29991947	515	SM02	AOX1a-14	30129963	30130317	355	SM23	AOX1a+8
29995708	29996042	335	SM05	AOX1a-11	30140463	30140907	445	SM24	AOX1a+9
29998673	29999122	450	SM06	AOX1a-10	30143468	30143907	440	SM26	AOX1a+11
30011693	30012267	575	SM09	AOX1a-7	30145028	30145442	415	SM27	AOX1a+12
30012928	30013282	355	SM11	AOX1a-5	30150893	30152142	1250	SM29	AOX1a+14
30017403	30017767	365	SM12	AOX1a-4					
30044398	30045152	755	SM13	AOX1a-3					
30053358	30053832	475	SM15	AOX1a-1					
30081893	30082342	450	SM16	AOX1a+1					
30103218	30103527	310	SM18	AOX1a+3					
30104228	30104652	425	SM19	AOX1a+4					

Reverse Strand:

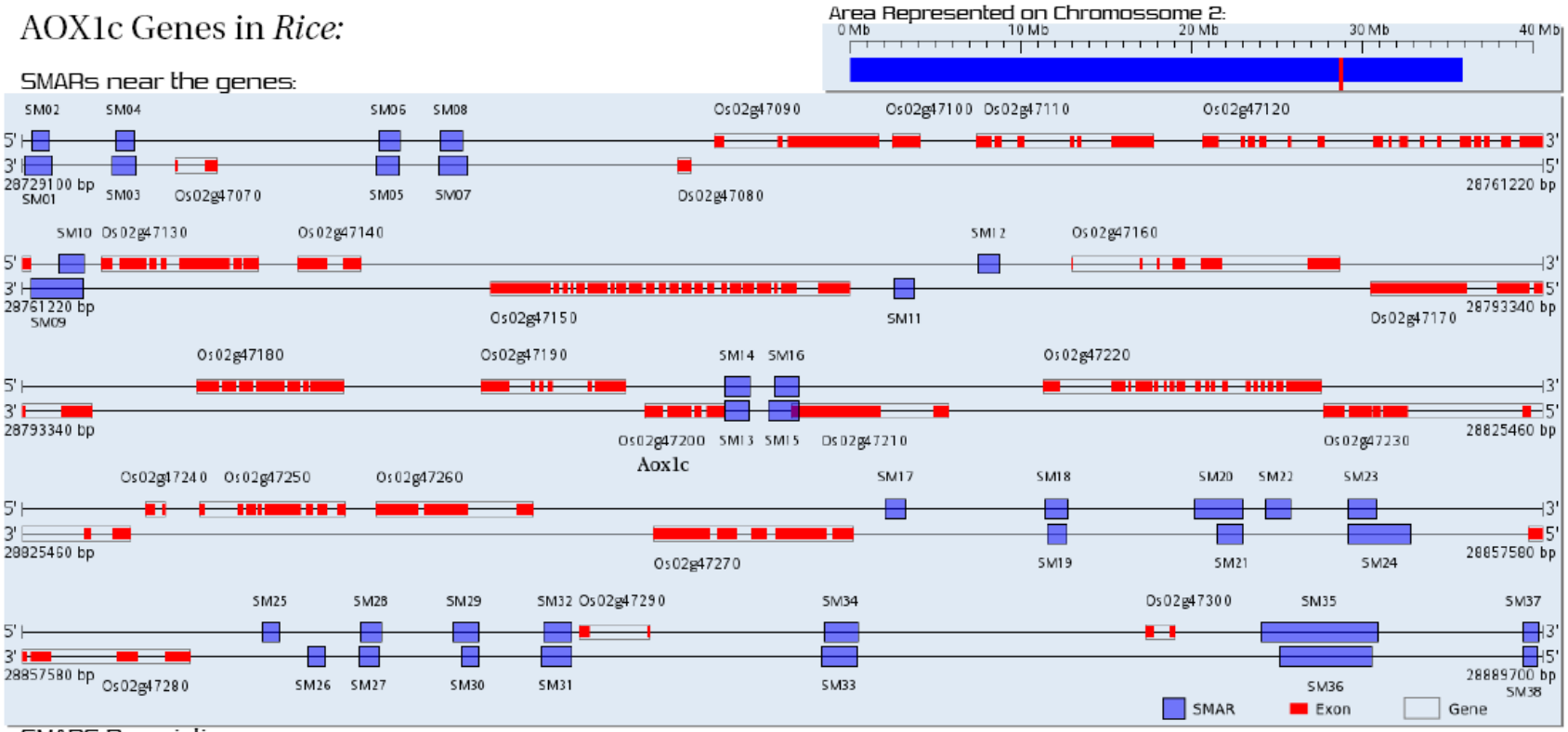
Start	End	Length	Id	Position
29991558	29991917	360	SM03	AOX1a-13
29995553	29996027	475	SM04	AOX1a-12
29998688	29999117	430	SM07	AOX1a-9
30011688	30012202	515	SM08	AOX1a-8
30012913	30013317	405	SM10	AOX1a-6
30044598	30045187	590	SM14	AOX1a-2
30081958	30082287	330	SM17	AOX1a+2
30104303	30104682	380	SM20	AOX1a+5
30108703	30109127	425	SM21	AOX1a+6
30140503	30140927	425	SM25	AOX1a+10
30145048	30145407	360	SM28	AOX1a+13
30151248	30152117	870	SM30	AOX1a+15

AOX1a e AOX1b Genes in *Rice:*

Gene Description:

Gene	Begin(bp)	End(bp)	Strand	Description
Os04g51030	29990047	29994880	Forward	Wall-associated kinase 1, putative, expressed
Os04g51040	29995979	29998608	Reverse	OsWAK50 - OsWAK receptor-like protein kinase, expressed
Os04g51050	30001205	30004086	Reverse	OsWAK53a – OsWAK receptor-like protein kinase, expressed
Os04g51060	30005404	30011424	Reverse	Expressed protein
Os04g51070	30016241	30019772	Reverse	Helix-loop-helix DNA-binding domain containing protein, expressed
Os04g51080	30022506	30027769	Forward	Scramblase family protein, expressed
Os04g51090	30030370	30044424	Forward	Expressed protein
Os04g51100	30045123	30048313	Forward	Nuclear WD protein, putative, expressed
Os04g51110	30048749	30051625	Forward	Cell division cycle protein, putative, expressed
Os04g51120	30052137	30057199	Forward	afl0-protein, putative, expressed
Os04g51130	30063252	30065525	Forward	Calcium-dependent protein kinase substrate protein, putative, expressed
Os04g51140	30065884	30067132	Reverse	Expressed protein
Os04g51150	30068069	30070780	Reverse	Alternative oxidase 1a (AOX1a)
Os04g51160	30072379	30073956	Reverse	Alternative oxidase 1b (AOX1b)
Os04g51166	30075700	30078336	Reverse	Expressed protein
Os04g51172	30078781	30086524	Reverse	Major intrinsic protein, putative, expressed
Os04g51180	30092465	30097221	Reverse	Protein GPR89A, putative, expressed
Os04g51190	30098620	30102567	Reverse	Growth-regulating factor, putative, expressed
Os04g51200	30118520	30120857	Forward	Retrotransposon protein, putative, unclassified
Os04g51210	30124437	30125828	Reverse	Retrotransposon, putative, centromere-specific
Os04g51220	30130348	30131283	Forward	Retrotransposon protein, putative, unclassified
Os04g51230	30131899	30138054	Reverse	Retrotransposon protein, putative, unclassified
Os04g51240	30142799	30144579	Reverse	EF hand family protein, expressed
Os04g51250	30146863	30149773	Reverse	Expressed protein

AOX1c Genes in *Rice:*



SMARS Description:

Forward Strand:

Start	End	Length	Id	Position	Start	End	Length	Id	Position
28729324	28729673	350	SM02	AOX1c-11	28853489	28854083	595	SM23	AOX1c+11
28731084	28731473	390	SM04	AOX1c-9	28862669	28863018	350	SM25	AOX1c+13
28736654	28737088	435	SM06	AOX1c-7	28864724	28865173	450	SM28	AOX1c+16
28737954	28738408	455	SM08	AOX1c-5	28866694	28867233	540	SM29	AOX1c+17
28762024	28762528	505	SM10	AOX1c-3	28868609	28869203	595	SM32	AOX1c+20
28781429	28781883	455	SM12	AOX1c-1	28874539	28875253	715	SM34	AOX1c+22
28808199	28808733	535	SM14	AOX1c+2	28883774	28886233	2460	SM35	AOX1c+23
28809254	28809773	520	SM16	AOX1c+4	28889294	28889648	355	SM37	AOX1c+25
28843689	28844148	460	SM17	AOX1c+5					
28847074	28847573	500	SM18	AOX1c+6					
28850239	28851253	1015	SM20	AOX1c+8					
28851744	28852273	530	SM22	AOX1c+10					

Reverse Strand:

Start	End	Length	Id	Position	Start	End	Length	Id	Position
28729159	28729733	575	SM01	AOX1c-12	28864714	28865123	410	SM27	AOX1c+15
28731004	28731503	500	SM03	AOX1c-10	28866869	28867228	360	SM30	AOX1c+18
28736584	28737063	480	SM05	AOX1c-8	28868544	28869193	650	SM31	AOX1c+19
28737914	28738513	600	SM07	AOX1c-6	28874479	28875228	750	SM33	AOX1c+21
28761419	28762508	1090	SM09	AOX1c-4	28884154	28886108	1955	SM36	AOX1c+24
28779649	28780073	425	SM11	AOX1c-2	28889299	28889603	305	SM38	AOX1c+25
28808194	28808693	500	SM13	AOX1c+1					
28809114	28809763	650	SM15	AOX1c+3					
28847139	28847553	415	SM19	AOX1c+7					
28850719	28851258	540	SM21	AOX1c+9					
28853494	28854813	1320	SM24	AOX1c+12					
28863609	28863983	375	SM26	AOX1c+14					

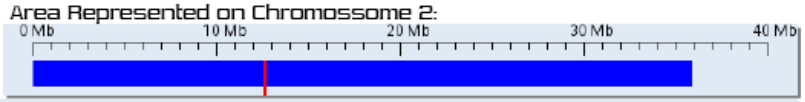
AOX1c Genes in *Rice:*

Gene Description:

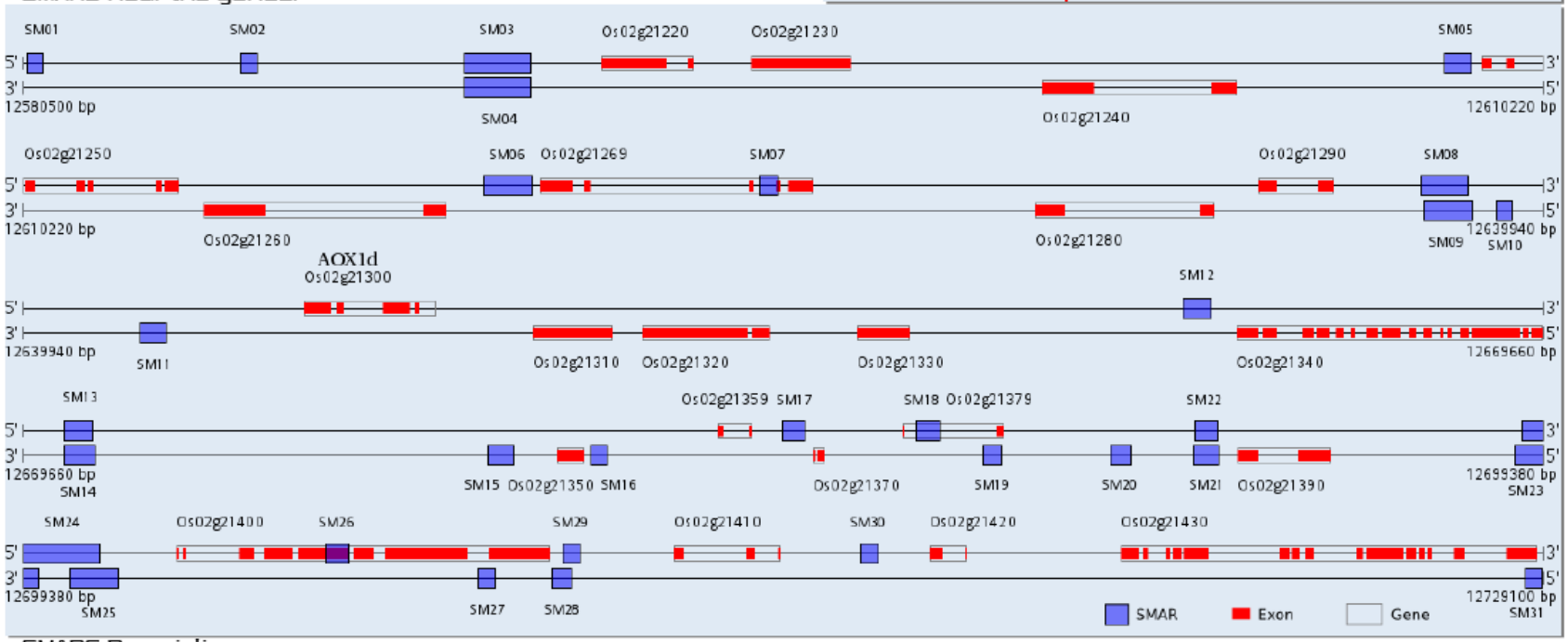
Gene	Begin(bp)	End(bp)	Strand	Description
Os02g47070	28732333	28733209	Reverse	Hypothetical protein
Os02g47080	28742945	28743244	Reverse	Hypothetical protein
Os02g47090	28743728	28747223	Forward	Peptide transporter PTR, putative, expressed
Os02g47100	28747508	28748065	Forward	Hypothetical protein
Os02g47110	28749301	28752996	Forward	ADP-ribosylation factor, putative, expressed
Os02g47120	28754076	28761416	Forward	Region found in RelA / SpoT proteins containing protein, expressed
Os02g47130	28762900	28766231	Forward	Expressed protein
Os02g47140	28767044	28768371	Forward	Ribosomal protein L12, putative, expressed
Os02g47150	28771135	28778718	Reverse	DNA topoisomerase 2, putative, expressed
Os02g47160	28783401	28789056	Forward	OsWAK17 – OsWAK short gene, expressed
Os02g47170	28789714	28794817	Reverse	Expressed protein
Os02g47180	28797042	28800123	Forward	Cell division cycle protein 20, putative, expressed
Os02g47190	28803051	28806106	Forward	Calcium-dependent protein kinase substrate protein, putative, expressed
Os02g47200	28806522	28808209	Reverse	Alternative Oxidase 1c (AOX1c)
Os02g47210	28809597	28812922	Reverse	Cationic amino acid transporter, putative, expressed
Os02g47220	28814926	28820788	Forward	Serine / Fthreonine-protein kinase ICK, putative, expressed
Os02g47230	28820839	28827763	Reverse	Expressed protein
Os02g47240	28828065	28828475	Forward	Hypothetical protein
Os02g47250	28829205	28832269	Forward	Retrotransposon protein, putative, unclassified
Os02g47260	28832951	28836242	Reverse	Retrotransposon protein, putative, unclassified
Os02g47270	28838828	28843028	Reverse	Retrotransposon protein, putative, unclassified
Os02g47280	28857305	28861128	Reverse	Growth-regulating factor, putative, expressed
Os02g47290	28869364	28870848	Forward	Hypothetical protein
Os02g47300	28881319	28881959	Forward	Hypothetical protein

AOX1d Genes in *Rice:*

Area Represented on Chromosome 2:



SMARs near the genes:



SMARS Description:

Forward Strand:

Start	End	Length	Id	Position	Start	End	Length	Id	Position
12580580	12580889	310	SM01	AOX1d-11	12692595	12693019	425	SM22	AOX1d+11
12584775	12585074	300	SM02	AOX1d-10	12698975	12700874	1900	SM24	AOX1d+13
12589140	12590434	1295	SM03	AOX1d-9	12705310	12705754	445	SM26	AOX1d+15
12608320	12608844	525	SM05	AOX1d-7	12709935	12710274	340	SM29	AOX1d+18
12619240	12620179	940	SM06	AOX1d-6	12715780	12716114	335	SM30	AOX1d+19
12624625	12624984	360	SM07	AOX1d-5					
12637580	12638494	915	SM08	AOX1d-4					
12662635	12663189	555	SM12	AOX1d+1					
12670470	12671014	545	SM13	AOX1d+2					
12684520	12684959	440	SM17	AOX1d+6					
12687150	12687594	445	SM18	AOX1d+7					

Reverse Strand:

Start	End	Length	Id	Position	Start	End	Length	Id	Position
12589145	12590429	1285	SM04	AOX1d-8	12700315	12701244	930	SM25	AOX1d+14
12637610	12638589	980	SM09	AOX1d-3	12708280	12708614	335	SM27	AOX1d+16
12639035	12639349	315	SM10	AOX1d-2	12709725	12710114	390	SM28	AOX1d+17
12642230	12642729	500	SM11	AOX1d-1	12728770	12729089	320	SM31	AOX1d+20
12670475	12671069	595	SM14	AOX1d+3					
12678750	12679259	510	SM15	AOX1d+4					
12680770	12681084	315	SM16	AOX1d+5					
12688430	12688804	375	SM19	AOX1d+8					
12690930	12691354	425	SM20	AOX1d+9					
12692570	12693044	475	SM21	AOX1d+10					
12698830	12699679	850	SM23	AOX1d+12					

AOX1d Genes in *Rice:*

Gene Description:

Gene	Begin(bp)	End(bp)	Strand	Description
Os02g21220	12591828	12593598	Foward	Hypothetical protein
Os02g21230	12594757	12596681	Foward	F-box domain containing protein, expressed
Os02g21240	12600448	12604216	Reverse	Ubiquitin-protein ligase, putative, expressed
Os02g21250	12609058	12613255	Forward	Coatomer subunit zeta-1, putative, expressed
Os02g21260	12613764	12618499	Reverse	Ubiquitin-protein ligase, putative, expressed
Os02g21269	12620356	12625647	Forward	Expressed protein
Os02g21280	12630040	12633503	Reverse	Expressed protein
Os02g21290	12634413	12635833	Forward	Hypothetical protein
Os02g21300	12645456	12648008	Forward	Alternative Oxidase 1d (AOX1d)
Os02g21310	12649913	12651466	Reverse	Expressed protein
Os02g21320	12652063	12654527	Reverse	Conserved hypothetical protein
Os02g21330	12656255	12657285	Reverse	Expressed protein
Os02g21340	12663696	12669649	Reverse	ABC-2 type transporter family protein, expressed
Os02g21350	12680130	12680637	Reverse	Expressed protein
Os02g21359	12683271	12683920	Forward	Hypothetical protein
Os02g21370	12685120	12685344	Reverse	Hypothetical protein
Os02g21379	12686872	12688845	Forward	Hypothetical protein
Os02g21390	12693440	12695240	Reverse	Conserved hypothetical protein
Os02g21400	12702399	12709674	Forward	Retrotransposon protein, putative, unclassified
Os02g21410	12712137	12714191	Forward	Hypothetical protein
Os02g21420	12717115	12717819	Forward	Hypothetical protein
Os02g21430	12720879	12728983	Forward	AML1, putative, expressed

Appendix 2 (Confidential)

In this Appendix are presented the results of gene expression evidence found for AOX genes in rice regarding in microarray and MPSS data available in public available databases.

Microarray data in Yale Virtual Center for Cellular Expression Profiling of Rice

Searching this data for AOX genes in Yale Rice Project we can find data for each one of the four AOX genes in rice (*AOX1a*, *AOX1b*, *AOX1c* and *AOX1d*). *AOX1a* has the TIGR locus identifier LOC_OS04g51150, and the Yale microarray identifier: Os054614_01 with the probe: GAGAAGGAGGTGGTGGTCAACAGCTACTGGGGCATCGAGCAGTCGAAGAAGC TGGTGCGGGAGGACGGCA, located approximately in the end of the first exon (bases 333 to 402).

AOX1a microarray expression could be found in:

- Leaf blade (2nd leaf) – Stomata (138,959 in 2 replicates)
- Root elongation zone – Endodermis (142.703 in 4 replicates)
- Root elongation zone – Epidermis (131.204 in 2 replicates)
- Root elongation zone – Cortex (91.224 in 4 replicates)
- Root maturation zone – Endodermis (92.225 in 3 replicates)

And with low intensity values but in almost all replicates

- Coleoptile (24h post imbibition) – (70.19 in 4 replicates)
- Leaf blade (2nd leaf) – Epidermal Long Cell (15.777 in 4 replicates)
- Radicle (24hr post imbibition) – (48.855 in 4 replicates)
- Scutellum (24hr post imbibition) – (35.724 in 4 replicates)

AOX1b has TIGR locus identifier LOC_OS04g51160 and the Yale microarray identifier Os023119_01 with the probe:

TCGTGAGCTACTGGGGCATCCAGCCGCCGAAGCTCGTGAAGGAGGACGGCAC GGAATGGAAGTGGCTCAG located approximately in the end of the first exon.

AOX1b microarray expression could be found in:

- Leaf blade (2nd leaf) – Bulliform (209.616 in 2 replicates)
- Leaf blade (2nd leaf) – Mesophyll (205.383 in 2 replicates)

- Root tip – Vascular bundle (158.51 in 3 replicates)

With low intensity values we can also found *AOX1b* expression in:

- Plumule – Dry Seed (55.105 in 2 replicates)
- Epiblast – 12 hr post imbibition (30.892 in 2 replicates)

AOX1c has TIGR locus identifier LOC_Os02g47200 and the Yale microarray identifier Os053159_01 with the probe:

CAGCTACTGGGGCATCGAGGCGTCGAAGCTCGCGAGCAAGGACGGCGTCGAG TGGAAGTGGTCTTGCTTC located at the end of the first exon.

AOX1c microarray expression is lower than the other AOX genes in rice, since it has low intensity values and is expressed in fewer tissues like:

- Shoot P3 (70.476 in 2 replicates)
- Shoot – Axillary meristem (46.771 in 2 replicates)
- Epiblast – Dry seed (44.838 in 3 replicates)

AOX1d has TIGR locus identifier LOC_Os02g21300 and the Yale microarray identifier Os029770_01 with the probe:

ACACCACAACCCGGCGACGCTCGGTGACAAGGTTGCCAGGTGGACGGTCAAAT CGTTGCGCTGGCCGGTT located in the middle of the second exon

AOX1d microarray expression has one high value in the root tip (vascular bundle) and other small expression values as seen in more detail below:

- Root tip – Vascular bundle (314.416 in 2 replicates)
- Whole leaf blade (fresh) (148.365 in 2 replicates)

Lower intensity values are found in:

- Root tip – Lateral root cap (58.799 in 2 replicates)
- Root tip – Cortex (52.286 in 2 replicates)
- Shoot – Axillary meristem (15.761 in 4 replicates)

MPSS Gene Expression Summary for AOX Genes (According with Rice MPSS database from University of Delaware)

Gene	Signature (20 base pairs)	Expression in TPM (Transcripts per Million)														
		9LA	FLB	9RO	9RR	MC0	MR48	MS3	MS12	MS24	MS96	NCL	NCR	NDR	NLC	NRB
AOX1a	GATCGGATATCACTGAGGAT			37	94	23	64	30		22	538	33		69		71
	GATCCGCACGCTGCTGGAAG										102		39			
AOX1b	GATCATCATTCATCAACGGG			66			22		17		321	14	204		21	
AOX1c	GATCCGGATTCATGTGTGCA	15	16													
	GATCTGTGCTACCGCAGTAG															
	GATCAGTCCTCCTCCGCCAC										19					
AOX1d	GATCCACTACAAGCCCAAGC															
	GATCCACCCGCCGCTGTGCT												11			
	GATCGTTGCAGTTCCTCTCA			33	76											

Gene	Signature (20 base pairs)	Expression in TPM (Transcripts per Million)														
		NSL	NSR	NYL	PLA	PLW	PSN	XC6	XR3	XR12	XR24	XS3	XS6	XS12	XS24	XS48
AOX1a	GATCGGATATCACTGAGGAT	96	35	56	54	23		39	105	46	48	61	60		62	153
	GATCCGCACGCTGCTGGAAG															
AOX1b	GATCATCATTCATCAACGGG															16
AOX1c	GATCCGGATTCATGTGTGCA													214		
	GATCTGTGCTACCGCAGTAG						52									
	GATCAGTCCTCCTCCGCCAC															
AOX1d	GATCCACTACAAGCCCAAGC		50													
	GATCCACCCGCCGCTGTGCT		22													
	GATCGTTGCAGTTCCTCTCA		2019													

(More expressed conditions)

NSR – Young roots stressed in 250 mM NaCl for 24 h (RNA was isolated using TRIzol Reagent)

MS(3-12-24-96) – M. grisea S. *Magnaporthe grisea* or *rice blast fungus* showing susceptible reaction after 96h. *Magnaporthe grisea* is a plant pathogenic fungus (causes an important disease affecting rice)

NCR – 14 Days - Young roots stressed in 4C cold for 24h biological replicate (RNA was isolated using TRIzol Reagent)

XR(3-12-24) - X.oryzae-R (*Xanthomonas oryzae*) after 3, 12 or 24hr - Leaves collected from 2 months old Nipponbare Xa21 plants 3, 12, or 24hr after Xanthamonas Oryzae inoculation showing resistance reaction

XS(3-6-12-24-48) - X.oryzae-S (*Xanthomonas oryzae*) after 3, 6, 12, 24 and 48hr - Leaves collected from 2 months old Nipponbare Xa21 plants 3, 6, 12, 24 and 48hr after Xanthamonas Oryzae inoculation showing susceptible reaction.

Xanthamonas Oryzae are Gram-negative rods and motile with a polar flagellum and is the causal agent of bacterial blight on rice. Bacterial Blight disease is a major rice disease in tropical Asian countries

NSL – Young leaves stressed in 250 mM NaCl for 24h (RNA was isolated using TRIzol reagent)

9RR – Roots (Replicate) - 60 Days Mature Roots - Biological Replicate

(Less expressed conditions by alphabetic order)

9LA – Leaves - 60 Days Mature Leaves

9RO – Roots - 60 Days Mature Roots

FLB – F1 Hybrid 60 days Mature Leaf Replicate B (F1 hybrids is a term used in genetics and selective breeding. F1 stands for Filial 1, the first filial generation seeds/plants or animal offspring resulting from a cross mating of distinctly different parental types)

MC0 – Mock treatment-0hr

MR48 - M. grisea S. *Magnaporthe grisea* or *rice blast fungus* showing resistance reaction after 48h. *Magnaporthe grisea* is a plant pathogenic fungus (causes an important disease affecting rice)

NCL – 14 Days - Young leaves stressed in 4C cold for 24h (RNA was isolated using TRIzol Reagent)

NDR – 14 Days - Young roots stressed in drought for 5 days, biological replicate (RNA was isolated using TRIzol Reagent)

NLC – 60 days - Mature Leaves - Replicate C (RNA was isolated using TRIzol Reagent)

NRB – 60 days - Mature Roots - Replicate B (RNA was isolated using TRIzol Reagent)

NYL – 14 days - Young leaves (RNA was isolated using TRIzol Reagent)

PLA – Rice leaf, beet armyworm damaged, after 24 hr. The beet armyworm is one of the most well-known agricultural pest insects. Feeding on the foliage of plants can completely defoliate small ones. Smaller larvae devour the parenchyma of leaves, so that all that remains is the thin epidermis and veins. Larger larvae tend to burrow holes through thick areas of plants

PLW – Rice leaf, water weevil damaged, after 24 hr. Adult weevils are small (1/8 inch), grayish-black snout beetles which may have a darker brown V-shaped area on their backs, who feed on the upper layer of cells (epidermis of leaves), producing long, narrow window like (skeletonized) slits along the leaf called feeding "scars." Upon reaching host plants, flight muscles of adults degenerate. Larvae chew on roots of developing rice and obtain oxygen from the host plant by means of paired hooks on the upper surface of their second through seventh abdominal segments.

PSN – Rice developing seed, 6 days old, Nipponbare-Grain quality control.

XC(6) - Mock treatment 6hr- Leaves collected from 2 months old Nipponbare Xa21 plants 6hr after wound treatment.