UNIVERSITÀ DELLA CALABRIA
Dipartimento di ELETTRONICA, INFORMATICA E SISTEMISTICA

UNIVERSITA' DELLA CALABRIA

Dipartimento di Elettronica,
Informatica e Sistemistica

Dottorato di Ricerca in
Ingegneria dei Sistemi e Informatica
XXIII ciclo

Tesi di Dottorato

BIOINFORMATIC METHODS FOR GENE DISCOVERY AND PROTEIN PREDICTION

Ofelia Leone

Coordinatore	Supervisore
Prof. Luigi Palopoli	**Prof. Luigi Palopoli**

DEIS

CONTENTS

1. INTRODUCTION

Nucleic acids and proteins, the most important biomolecules, are linear sequences of nucleotide and aminoacid units respectively, that can be represented by letters. There are two different nucleic acids in cells, the desossiribonucleic acid (DNA), that constitutes the genomes, and the ribonucleic acid, or RNA. Nucleotidic sequences have an alphabet of four characters (A, C, G, T or U), in which every letter represents a nucleotide; protein sequences have an alphabet of 20 characters, each one representing an aminoacid. Thus, from an information point of view, both nucleic acids and proteins are letter strings, on which ad hoc softwares can work. Of course, the character string is only a simplified representation of corresponding nucleic acid or protein, but it represents faithfully the primary structure and it permits to carry out interesting analysis not possible otherwise. For instance, to determine how much two sequences are similar, it is necessary to find the best way to align them.

The development of molecular biotechnologies and, in particular, the improvement of sequencing techniques led to production of enormous amount of biological data. Today we know whole genomic sequences of many species, belonging to all five kingdoms. Computational technology supported the development of modern biology by providing databases and algorithms to get and analyze them. Analysis of biological sequences is the issue of sequence-oriented Bioinformatics, a field of Bioinformatic that focusing on analyzing complete genomes, studying properties of biomolecule strings, rebuilding philogenetic relationships and so on.

Two major goals in sequence analysis are to identify sequences that encode proteins, and to discover sequences that regulate the expression of genes or other cellular processes. In the last years many genomes of different organisms have been completely sequenced, but the simple knowledge of nucleotidic sequences do not imply detailed knowledges on how DNA brings its informations. Thus, in spite of the availability of sequence data, most of informational content of genomes remain still undiscovered. The number of genes contained in sequenced genomes is not completely clear, and many genes are not yet been identified. It is known that DNA

is made not only of genes, but also of sequences involved in regulation of gene expression, and long sequences whose meaning is still unknown. Thus, a relevant challenge is to assign a meaning to millions of nucleotides towards and to understanding the functions of all different parts of genomes.

Several computer programs have been developed to scan genomic sequences in order to find genes, particularly those encoding proteins. These programs usually are based on sequence similarity, thus they are not suitable to be applied to contents where the expected homology between the gene searched for and the known sequences is low. Furthermore, computational methods are very useful in searching for genes with standard structure, but they fail in identification of encrypted genes, that is, the genes hidden in genomes because of the many complications in genic organization.

1.1 Contributions

This thesis has his focus located in the context just mentioned above. From the biological point of view, two are the main problems this thesis deals with: the first, a specific one, is the identification of specific genes; the other, more general, is the discover in plant mitochondrial genomes, of genes encoding proteins not yet identified and difficult to search for using standard search tools and methodologies. In particular the contributions of the thesis can be summarized as follow:

First, as far as methods and tools are concerned, two new approaches to gene search have been proposed to deal with the two problems, that are:

Gene finding by tag search

- ✓ A multi-tool methodology have been developed to discover specific genes;
- ✓ One of the supporting tools is an ad hoc one, specifically designed to implement a particular kind of tag search.

Protein prediction

- ✓ A multi-tool methodology has been defined to predict undiscovered mitochondrial proteins;
- ✓ A system for Open Reading Frame (ORF) extraction from genomes, based on automatic RNA-editing simulation, was designed to support the methodology.

Second, concerning the experimental analysis, we have developed several campaigns, specifically:

Gene finding by tag search

a. The methodology based on tag search was applied to find a specific gene, the *IP6K*, in plants. All sequenced mitochondrial DNA (mtDNA) of plants and nuclear genome of two model plants were analyzed.
b. To show the goodness of the new approach the search of a known gene in plant genome was performed.

Protein prediction

c. In order to predict possible undiscovered protein, the new method for ORF extraction was applied to mtDNA of a model plant.
d. In order to validate our approach the search of known proteins generated by RNA editing was performed.

1.2 Plan of the thesis

This thesis is organized as follow. After this brief introduction, in chapter 2 the state of the art is illustrated, with regard to both biological and technological contexts.

In chapter 3 the problem of specific gene finding is addressed. In particular in section 3.2 the question of *IP6K* identification in plant is illustrated; then the function of the gene we looked for in cellular metabolism is described. In section 3.3 the contribute given by this thesis, to solve this problem, is stated. In particular, in the subsequent section 3.4, the multi-tool methodology based on tag search that we designed is described, as well as other tools supporting the methodology. In section

3.5 the main results of our analysis are illustrated and their biological relevance is discussed.

Chapter 4 concerns the prediction of genes encoding proteins not yet identified with classical searching tools. First, in section 4.2 the peculiar structure of plant mitochondrial DNA is described and the editing phenomenon is these organelles is illustrated. Then, in section 4.3, the contribution given by this thesis in this context is stated. In particular, in section 4.4 a multi-tool methodology for protein prediction is described as well as the system for automatic RNA-editing simulation, that we designed in supporting to the methodology. In section 4.5 the experimental results are illustrated and some further analysis are outlined.

Conclusive remarks are stated in chapter 5.

2. STATE OF THE ART

2.1 Biological context

The importance of data collected from genome sequencings, concerns their biological content. The great interest on nucleotide sequences is because the four letters A, T, C, G, assembled in a suitable way, assume particular meanings. The nucleotide sequences that have a biological meaning, are called *genes*. A gene can be better defined as a nucleotide sequence coding for a functional biological product.

Some genes code for other nucleic acids, the ribosomal RNA (rRNA) and the transfer RNA (tRNA), both implicated in cellular pathway of protein synthesis. However, most of genes codes for proteins, the main macromolecules involved in biochemical and metabolic cellular processes. Actually, only a small percentage of the genomic sequences is known to encode proteins because of the presence of introns within coding regions and other non coding regions in the genome. In superior eukaryotes it is estimated that this percentage is between 1 and 5%, more than 98% being non coding DNA. The role of this large part of genome is not yet completely clarified, and it was termed "junk DNA" because believed without any function. Recently it was found that some of the apparently not meaningful sequences actually contribute to protein synthesis (Labrador M et al, 2001). This discovery opened a new research field and encouraged enormously studies in sequence DNA analyses.

With the publication of the human genome sequence in 2003 (Venter J. Craig et al., 2001), achieved with the international efforts of researchers joined of the Human Genome Project, the "genomic era" was born. Sequencing projects of both prokaryotic and eukaryotic organisms were rapidly completed, and the number of genome sequences available is today in continuous increase. It can be said that now we are entering in the "post-genomic" era, in which the efforts are concentrated on harvesting the fruits hidden in the genomic text.

2.1.1 Generals about gene finding

Once the genome of a species has been sequenced, the first and most important step to understand its meaning, is gene finding. First of all, it is important to know how many genes it contains and their locations, then the nucleotide sequence with the gene structure, and finally the function of the encoded protein. Via the genetic code, the gene determines the exact amino acid sequence of the protein chain. The transcription machinery of the cell reads genes and translates them into the appropriate protein chain. But a gene is much more than a simple coding sequence along the genome; it has a complex internal structure and there are also associated sequence regions that help regulate the transcription.

The first attempts of gene finding were made with biological methods. Starting from known proteins (the gene products), it was possible to create *genetic maps*, establishing the rough location of genes relative to each other on a certain chromosome. There are two different way of mapping: genetic mapping, using classical genetic methods, like pedigree analysis or breeding experiments, and physical mapping, using molecular biology techniques. When the location of a gene is known, it is possible to clone and characterize a gene with recombinant DNA techniques. Another biological method to gene finding is to compare RNA transcripts to genomic DNA sequences by experimental analysis. The genes that can be identified by this method are restricted to ones expressed in the cell at the time of experimental RNA isolation. Indeed, while DNA is static, with the same sequence present in all cells, RNAs are synthesized only when the cell needs them. Generally speaking, genes are expressed according to cell requirement, that depends on the kind of cell or tissue, the distinct steps of cellular differentiation, and on external signals.

Today, with powerful computational resources at the disposal of the research community, gene finding has been largely redefined as a computational problem. Computational programs are available for identifying elements on genomes, a process called *gene prediction*. Then, genome annotation is the following step of attaching biological information to the predicted gene sequence.

2.1.2 Troubles in gene finding

Because of the rapid advancements in genetics and molecular biology knowledge, the gene concept, long regarded as a unit of inheritance, undergoes continuous transformations to accommodate novel structures and modes of action. The advent of DNA sequencing led to the concept of the gene as an open reading frame, and the post-genomic era has challenged the very idea of the gene.

According to the classical concept, each gene corresponds to only one protein, so that the transfer of genetic information results, in a sense, linear. This concept constitutes the *central dogma of biology*, based on univocal correspondence between a gene and the encoded protein, through an intermediate step involving another kind of ribonucleic acid, the messenger RNA or mRNA.

The transfer of information from DNA to mRNA is called *transcription*, while the subsequent step from mRNA to protein is called *translation*. Transcription is the first step leading to gene expression. In this process, a complementary RNA copy of a sequence of DNA is created; the stretch of DNA transcribed into an RNA molecule and translated in a protein is a gene. In eukaryotic cell, a second step is required to get RNAs ready for protein synthesis, that includes some modifications of RNA transcript. The mature mRNA reaches the cellular cytoplasm and are decoded by the ribosome to produce a specific amino acid chain, or polypeptide, that will later fold into an active protein. The correspondence between the nucleic acid alphabet of 4 letters and which of proteins, made of 20 letters, is the *genetic code*. The code is read in triplets, so that each sequence of three nucleotides, called codons, specifies a single amino acid. There are $4^3 = 64$ different codon combinations possible with a triplet codon of three nucleotides; all 64 codons are assigned for either amino acids or stop signals during translation. In particular, the nucleotide triplets *tag*, *tga* and *taa*, are stop codons, not corresponding to any aminoacid in genetic code. The codon *atg*, corresponding to methionine, is considered the start codon, because it is the beginning signal during translation.

However, the effective implementation of these logical steps implies some problems. Confronted with the very small percentage of DNA coding for proteins, the amount of RNA carrying the genetic information, is much larger. Thus, the univocal correspondence “one gene, one protein”, stated by the central dogma of

biology and considered unquestionable until to the year 2000, collapsed after the completion of genome sequencing. Surprisingly, in the human genome, only 25000 genes were found, while in human cells there are at least 90000 different types of proteins.

Some exceptions to the standard gene structure have been discovered that account for part of this disagreement. For instance, it is known that two genes can overlap and that a gene can be codified in part from a strand and in part from the other strand of DNA. However, these mechanisms are not enough to explain the great difference between protein number and DNA sequences that generate them. Therefore, it was considered that the transfer of genetic information is not always linear, but there are some processes that break this linearity and increase the diversity of proteins produced by a certain DNA sequence.

In this respect, it is today clear and accepted that the information content of a single gene can be modified so that the protein diversity of an organism can be increased. These modification processes are prevalently post-transcriptional events, and the most common of them are *trans splicing* and *RNA-editing*.

Trans splicing is a process in which two RNA molecules, produced by different DNA regions (even very distant from one another), are joined into a single RNA molecule able to produce a protein.

Differently, RNA editing is a process in which some bases of an RNA molecule are enzymatically modified, so that its information content is altered. RNA editing occurs in the cell nucleus and cytosol, as well as in mitochondria and plastids. Many molecular editing mechanisms are known, including nucleoside modifications such as cytidine (C) to uridine (U) and adenosine (A) to inosine (I) deaminations, as well as non-templated nucleotide additions and insertions.

The result of editing process is that the amino acid sequence of the encoded protein is effectively altered, so that it differs from that predicted by analyzing the genomic DNA sequence. Unfortunately, but interestingly, the differences between RNAs and their coding sequences can be so large as to hinder both experimental and computational research of genes.

2.1.3 Biological problems of interest

Although there has been a huge progress in developing computational methods for analyzing genomic sequences and finding protein-encoding regions, these methods do not offer complete support to all discovery tasks. Therefore, even if many gene have been identified by computational analysis of genomes, much remains still to be discovered, especially in vegetal organisms.

Very intriguing is the case of genes whose existence is supposed on the basis of specific biological considerations, but they are not yet been discovered. One of these genes is IP6 Kinase (IP6K) in plants, the gene that encodes the enzyme converting inositol hexakisphosphate (IP6 or phitic acid) in diphosphoinositol pentakisphosphate (IP7 or PP-IP5). Although IP6K has not yet been identified in plant chromosomes, there are many clues suggesting its presence in plant cells.

Notably, the classical methods to gene search failed to find IP6K in plant. Thus new approaches are required to identify this interesting and elusive gene.

A more general question is the identification of new proteins in plant mitochondrial genomes. Mitochondrial DNA is very different from nuclear DNA. It is typically made of only one type of circular molecule, occurring in several copies per cell. This molecule contains additional information with respect to nuclear DNA, concerning a certain number of mitochondrial components, such as some tRNAs, rRNAs and proteins. Genes coding for proteins are present on both strands of mtDNA.

There are important structural differences between mitochondrial genomes of animals and plants. While the nucleotide sequences of the former are almost entirely codifing, plant mtDNAs contain a large amount of apparently non-coding sequences. Furthermore, plant mitochondrial genomes have several characteristics that hinders the search of genes. It is considered that the number of proteins present in plant mitochondria is certain higher than Open Reading Frame known in mtDNA to date; identifying all the proteins resident in this organelle represents a major challenge in cell biology.

To discover mitochondrial genes that elude classical gene finding techniques, an approach seems necessary that takes into account the mechanisms of mtDNA gene expression.

2.2 Technological context

The rapid development of biotechnologies introduced the need of computational techniques backing modern biology and there are several biological fields in which computer support is now indispensable. One of the most important function of bioinformatics is to provide systems to collect the enormous amount of biological data and to provide algorithms to analyze them. Many biological databases are available, designed as containers built to store data, so that all users can easily get them. There are primary database (for nucleotide and aminoacid sequences), and specialized databases (including protein motifs and domains, protein structures, genes, transcriptoma, expression profiles, metabolic pathways ecc). The most important primary databases are GenBank (http://www.ncbi.nlm.nih.gov), managed by the National Center for Biotechnology Information (NCBI); the DNA Databank of Japan (DDBJ;http://www.ddbj.nig.ac.jp); and the European Molecular Biology Laboratory (EMBL)/EBI Nucleotide Sequence Database (http://www.embl-heidelberg.de). Collection of genome data is a process in continuous evolution. To have an idea of the growth rate, Genebank had 606 sequences in 1982, and 70 millions in 2007 (Greene E. A, et al., 2000). New sequences are submitted daily to the GenBank, EMBL, and DDBJ databases. The NCBI reviews new entries and updates existing ones. Several biological databases are available in flat-file format, that is sequential files in which every class of information is reproduced on one or more consecutive lines, identified by a code. Most databases provide special tools to process the data, like tools for database screening (BLAST, FASTA), tools for multiple sequence alignments (BLAST, ClustalW, AntiClustAl (Di Pietro et al., 2003), T-Coffee, ProbCons); tools for identification of exons and regulation elements, that is gene and promoter prediction (GeneScan, Promoser). An important development of herewith is the implementation of databases resulting from a generic container with the aim to provide more specific and accurate informations. For instance the EST (Expressed Sequence Tags) database is a collection of partial sequences of expressed genes, useful in finding new genes.

In the post genomic era, the availability of several and whole genome sequences, focused researchers' attention on gene search. The simplest way to find genes in a genome is to scan the nucleotide sequence string searching for *Open Reading*

Frames (or *ORFs).* An ORF is DNA sequences that does not contain any stop codon in a given reading frame; thus, it is made of a stretch of DNA that contains a contiguous set of codons, each of which specifies an amino acid. There are six possible reading frames in every DNA sequence, three starting at positions 1, 2 and 3 of a given strand, and three starting at positions 1,2 and 3 of the complementary sequence, all going in the 5' to 3' direction of the strand[1].

In prokaryotic genomes, DNA sequences that encode proteins are transcribed into mRNA, and the mRNA is usually translated directly into proteins without significant modification. In these organisms an ORF running from the first available start codon on the mRNA to the next stop codon in the same reading frame generally provide a good, even if not assured, prediction of a protein-encoding region. The presence of many in-frame stop codons in a reading frame gives rise to short ORFs that do not encode any protein. In eukaryotic organisms, protein synthesis is a more complex process. Gene transcription begins at specific promoter sequences and it is followed by removal of noncoding sequence (introns) from primary RNA transcript by a splicing mechanism. The mature RNA, arising from post-transcriptional processing, can be translated in the 5' to 3' direction, usually from the first start codon to the first stop codon. As a result of the presence of introns in the genomic DNA sequences of eukaryotes, the ORF corresponding to an encoded gene will be interrupted because of the presence of stop codons in introns.

A considerable percentage of genes identified within genomic sequencing projects encode proteins previously unknown. Hence the need of computational methods to predict new gene structures, that can make it easier to annotates gene and to provide a guide for experimental validation. There are two kind of methods to gene finding: the predictive and the comparative ones. Predictive methods can be content-based, analyzing the global properties of sequences under study, or site-

[1] Each nucleotide in a DNA molecule is made of three components: a nitrogenous base, a phosphate group and a sugar (2-deoxyribose). The backbone of the DNA strand is made from alternating phosphate and sugar residues. The sugars are joined together by phosphate groups that form phosphodiester bonds between the third and fifth carbon atoms of adjacent sugar rings. These asymmetric bonds gives a direction to a strand of DNA. In a double helix the direction of the nucleotides in one strand is opposite to their direction in the other strand: the strands are antiparallel. The asymmetric ends of DNA strands are called the 5′ (five prime) and 3′ (three prime) ends, with the 5' end having a terminal phosphate group and the 3' end a terminal hydroxyl group.

based, that focus on the presence or absence of specific signals (pattern or consensus sequences). Because the single parameter have a very low predictive value, prediction softwares use a combination of content- and site-based approaches. Thus, these methods identify the most likely protein-encoding regions in a DNA sequence on the basis of models of gene structure which incorporate descriptions of the basic transcriptional, translational and splicing signals, as well as length distributions and compositional features of exons, introns and intergenic regions. Several gene prediction softwares are now available, like GeneScan (Burge C et al., 1997), Fgenesh (Salamov A. et al.,2000), HMMgene (Krogh A, 1997), GeneParser (Snyder EE, 1993; Snyder EE et all., 1995). They all have in common the ability to differentiate between characteristic sequences of expressed genes from other non-gene sequences that lack these patterns. Because these gene sequences as well as gene structure (the number and sizes of exons and introns) vary from one organism to another, a program trained on one organism is not generally useful for another organism. Reliability tests of gene prediction programs have shown that the available methods for predicting known gene structure are, in general, anyway, error-prone.

The alternative to this "ab initio" gene discovery is the comparative gene finding, based on sequence similarity.

It consists in performing a database search by translating nucleotidic sequences in all possible reading frames and comparing them to a protein sequence database using the BLASTX or FASTX programs. Homology criteria can allow for the identification of new proteins in the organism under analysis. Alternatively, if a genomic sequence is to be scanned for a gene encoding a particular protein, the protein can be compared to a nucleic acid sequence database that includes genomic sequences and is translated in all six possible reading frames by the TBLASTN or TFASTX/TFASTY systems. For proteins that are highly conserved, these methods can give a very good, albeit approximate, indication of the gene structure. If the proteins are not highly conserved, or if the exon structure of a gene is unusual, these methods may not work.

Available software tools are trained especially for nuclear genes prediction, but they do not take into account phenomena like editing, occurring mostly in mitochondrial genomes. To date much is known about mitochondrial genome of human and several Metazoan, but a big amount of plant mtDNA informational

content is still in hiding. This is because of a major structural complexity of plant mitochondrial genomes, that show an unusual organization and unique gene-processing mechanisms.

Thus, computational methods for gene finding are not exhaustive, because they can work very well for searching some genes, but they fail in solving particular problems in gene finding.

3. GENE FINDING BY TAG SEARCH

3.1 Introduction

Even if several genes have been discovered by computational genome analysis, many challenges still remain open. Indeed, although such computational methods are very helpful in finding canonic genes, there are situations in which they fail in discovering genes encrypted in the genome due to several complications that may possibly arise.

This is the case of *IP6K* gene. This gene belongs to an inositol polyphosphate kinase superfamily, the IPKs (Pfam PF03770), that evolved from a common ancestor. IP6K has been found in all eukaryotes analyzed but not in plant, where it has been searched by common software of gene finding.

Although functionally conserved, *IPK* genes present very low sequence homology in different organisms, with less than 24% identity showed in some pairwise combinations (Ives EB et al., 2000). The sequence identity is limited to a few small regions with high homology. The considerable sequence heterogeneity among the several known *IPKs*, is the cause of failure of common homology search programs in searching this gene in plants, suggesting the need to define new approaches not based on gene sequence homology.

We addressed the problem of identification of *IP6K* gene by proposing a new approach to gene finding based on tag search. In this part of the thesis I will describe the problem and our contribute towards its solution. In particular, this chapter is organized as follow. In the section 3.2 I will describe the metabolic role of the IP6 Kinase and the problem of the lack of this enzyme in plants. In the section 3.3 I shall illustrate our contributions in this context, specifying the methodological approaches (section 3.4) and the experimental results (section 3.5).

3.2 IP6K in Inositol Polyphosphates metabolism

Inositol polyphosphates (IP) are an important class of signaling molecules controlling disparate cellular functions. The first inositol polyphosphate (inositol hexakisphosphate or IP6) was described about 90 years ago in plant seeds (Posternak S., 1919). Inositol polyphosphates derivates from myo-inositol, the most abundant inositol isomer in nature, with the six-carbon ring harboring one axial hydroxyl at the D-2 position and five equatorial hydroxyl groups. Building on this structure a great

diversity of inositol derivatives is achieved with multiple combinations of mono- and pyro-phosphate groups attached to each of the six hydroxyls moieties.

Interest in inositol polyphosphates dramatically increased about thirty years ago when the role of inositol 1,4,5-trisphosphate (Ins(1,4,5) P3) in mobilization of Ca2+ from intracellular stores was discovered (Irvine R., 2003). Today it is known that the cytoplasmic functions of IP include essential structural and signaling roles in vesicular trafficking, actin cytoskeleton rearrangements, and Akt signaling (Strahl and Thorner 2007). Recently some roles in nuclear processes have been discovered for soluble inositol polyphosphates (IPs), like gene expression, (DNA) repair, telomere homeostasis, and kinase-free phosphorylation of proteins within the nucleus (Alcázar-Román AR et al., 2008).

Inositol hexakisphosphate (also known as phytic acid) is the most abundant inositol polyphosphate in eukaryotic cells. It is a major component of plant seeds representing 0,1 – 1% of its dry weight and 60 – 80% of total phosphate content (Shears S.B., 2003). Significantly, IP6 is the precursor of a novel class of more anionic inositol polyphosphates, the inositol pyrophosphates, in which the fully phosphorylated IP6 ring is further phosphorylated to create high-energy pyrophosphate groups (**Fig.3.1**).

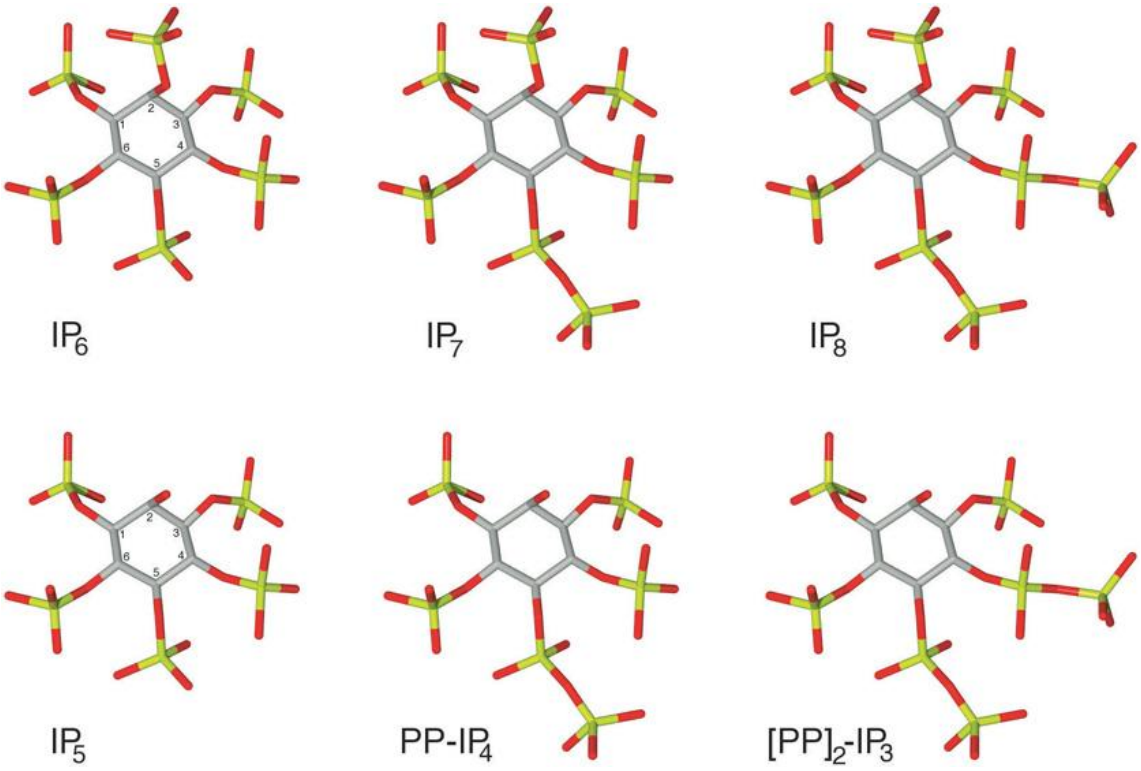


Figure 3.1. Inositol pyrophosphate chemical structure. The figure shows the structure of phytic acid with its pyrophosphate derivate IP7 and IP8, as well as IP5, with the derivate pyrophosphates PP-IP4 and [PP]2-IP3. In the figures, carbon, oxygen and phosphate atoms are coloured grey, red and green, respectively (Bennet et al., 2006).

The best charateryzed inositol pyrophosphates are the diphosphoinositol pentakisphosphate (IP7 or PP-IP5) and the bis-diphosphoinositol tetrakisphosphate (IP8 or [PP]2-IP4), with one and two pyrophosphate group, respectively (Bennett M et al., 2006). Since their discovery in the early 1990s, inositol pyrophosphates have been found in all eukaryotic cells analyzed, from yeast to mammalian neuron.

Inositol pyrophosphates are important cellular messengers that control a wide range of cellular function, including endocytosis (Saiardi A et al., 2002), apoptosis (Gong B et al., 2002), telomere length (Saiardi A., et al., 2005), DNA recombination (Luo H., et al., 2002). The high energy pyrophosphate bond of IP7 can directly donate the beta phosphate to proteins defining a new kind of protein phosphorylation mechanism (Saiardi A., 2004), recently proposed to represent a novel post-

transductional protein modification (serine pyro-phosphorylation) (Bhandari R., et al., 2007).

The first enzyme able to synthesise inositol pyrophosphates was purified to homogeneity from rat brain (Voglmaier S. M., et al., 1996). The new enzyme was called inositol hexakisphosphate kinase (IP6K) and it was show to be able to convert IP6 plus ATP to IP7 and ADP; ATP cannot be substituted by other nucleotides such as CTP or GTP (**Fig.3.2**).

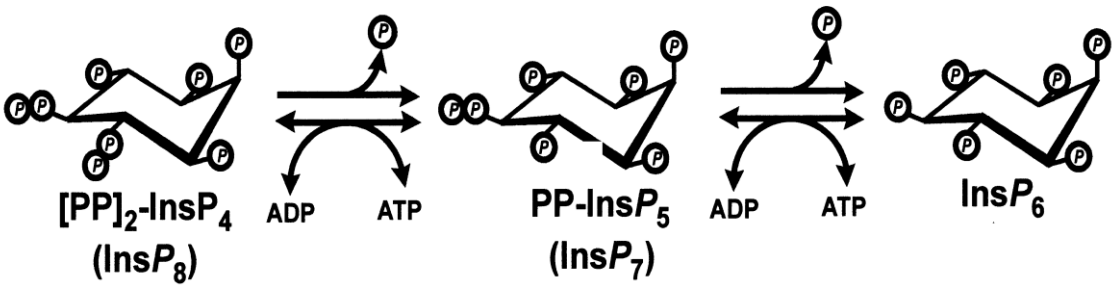


Fig 3.2. Convertion of IP6 plus ATP to IP7; the reaction is catalyzed by IP6K. IP7 plus ATP can be converted in IP8 in a following reaction.

Later the same laboratory characterized one yeast and two mammalian proteins with inositol hexakisphosphate kinase activity, called KCS1 (Kinase C Suppressor 1, also known as yIP6K), IP6K1 and IP6K2, respectively (Saiardi A., et al., 1999). Subsequently, a third mammalian gene, IP6K3, was cloned (Saiardi A. et al., 2001). All the mammalian IP6Ks phosphorylate *in vitro* IP6 to IP7, IP5 to PP-IP4 and [PP]2-IP3 (Saiardi et al., 2001, Saiardi et al., 2000). Later the enzyme was cloned in other mammalians, and its high evolutionary conservation was regularly observed, which facilitated the identification and cloning of IP6K enzymes from distant organisms, including yeast and the amoeba Dictyostelium (Luo H. et al., 2003).

The cloning of IP6Ks also helped to identify an evolutionarily conserved family of inositol polyphosphate kinases known as inositol polyphosphate multi-kinases (IPMKs) (Saiardi A. et al., 1999), that phosphorylate a broad range of substrates to yield a variety of inositol polyposphate reaction products (Stevenson-Paulik J., et al., 2002, , Eskin E. et al., 2002, Larkin M. et al., 2007, Altschul S. et al., 1997),

including some inositol pyrophosphate species (Shears SB, 2004, Luo H. et al., 2003, Loomis W. et al., 1995).

Altogether, the IP6Ks, IMPKs and IP3-3Ks (the enzymes that convert I(1,4,5)P3 to I(1,3,4,5)P4 -Schell M. J. et all., 1999-) belong to an inositol polyphosphate kinase superfamily, the IPKs (PFAM accession number PF03770), that evolved from a common ancestor. Phylogenetic analysis of their sequences predicts that IP6Ks arose first, followed by IPMKs, and lastly IP3-3Ks (**Fig. 3.3**).

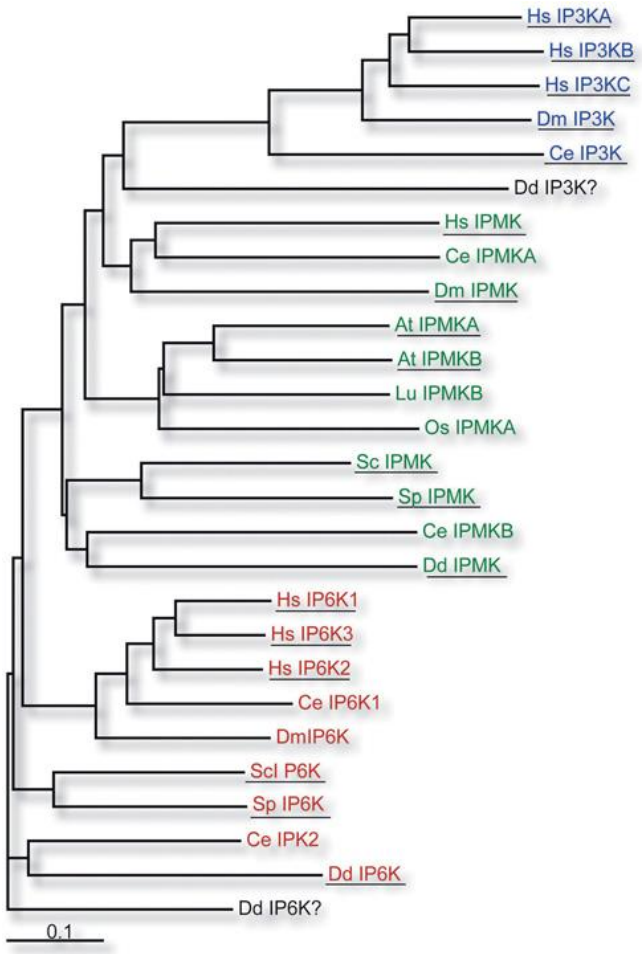


Fig3.3. Phylogenetic tree of IPK proteins family members. The following abbreviations are used: At, *Arabidopsis thaliana*; Ce, *Caenorhabditis elegans*; Dd, *Dictyostelium discoideum*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*; Lu, *Linus utilissimum*; Os, *Oryza sativa*; Sc, *Saccharomyces cerevisiae*; Sp, *Schizosaccharomyces pombe*.

3.2.1 Is there IP6K in plants?

Plants have played a special role in inositol polyphosphate research since in plant seeds was discovered the first IP, the fully phosphorylated inositol ring of phytic acid (IP6). Although plants were instrumental to the early discovery of inositol polyphosphates, and remain experimental systems with large agri-business commercial interests, the majority of studies are now performed in mammalian cells, amoeba or yeast. Nevertheless, studies in plants have demonstrated that IP6 is an endomembrane-acting regulator of calcium-release and controller of K^+ channels in guard cell (Lemtiri-Chlieh F., et al., 2000). IP6 is a major component of plant seeds representing 0,1-1% of its dry weight and 60-80% of total phosphate content (Raboy V., 2003, Shears S.B., 2001). In the late nineties, significant steps were made in the elucidation of the biochemical pathway responsible for IP6 synthesis in crop and model species, enabling subsequent genetic analysis of plant mutants with impaired phytic acid synthesis (Stevenson-Paulik J. et al., 2005, Sweetman Det al., 2006, Xu J. et al. 2005) .

The presence of IP7, the more anionic inositol polyphosphates derived from IP6, has been demonstrated in vegetal organisms, both in monocotyledonous and in dicotyledonous plants (Flores S. 2000, Brearley C. 1996). Furthermore, the conversion of IP6 to IP7 has been detected in Arabidopsis cells and leaf tissue in the presence of ATP, demonstrating IP6- kinase activity in plant extracts (Saiardi A., Azavedo C., unpublished manuscript).

These findings, together with the observed high conservation through the evolution of IP6K, strongly suggest the presence of this enzyme in vegetal cells. Therefore, IP6K enzyme was searched in plant genomes by homology based methods, but all studies have failed to reveal its presence.

Two IPMK proteins (called AtIPK2a and AtIPK2b in Arabidopsis thaliana) have been identified so far (Stevenson-Paulik J. 2002, Xia H, 2003). The substrate ambiguity is a general property of IPKMs, that phosphorylate a broad range of substrates and some IPMK proteins are able to synthesise inositol pyrophosphate species (Nalaskowski M. M., 2002, Saiardi A., 2001, Zhang T.,2001). In rice and barley an IPMK able to phosphorylate all intermediates from inositol bisphosphate to IP6 has been characterized (Josefsen L., 2007). To date IPMKs have been identified in dicotyledonous and in monocotyledonous plants, as well as in algae.

However, in Arabidopsis it has been shown that the two IPMK proteins contribute to inositol 1,3,4,5,6-pentakisphosphate(IP5) production, but do not show any inositol pyrophosphate enzymatic activity (Stevenson-Paulik J. 2002, Xia H, 2003).

Thus, an enzyme must exist in plants converting IP6 in IP7. The observed evolutionary conservation strongly suggests that this enzyme is indeed an IP6 Kinase.

There are many clues connecting IP6K to cell mitochondria. It was shown that human IP6K2 moves from nuclei to mitochondria and provides physiologic regulation of apoptotic process by generating IP7 (Nagata E. et al., 2005). Furthermore, yeast deficient in KCS1 (yeast IP6-Kinase), kcslΔ, do not survive if they are grown in conditions in which survival is dependent from mitochondrial function, thus demonstrating the importance of IP6K for this organelles (Saiardi A., unpublished manuscript).

Some further observations could suggest that the corresponding gene might be found in plant mtDNA, probably encripted and hidden by virtue of editing and/or trans splicing processes. It is known that most of mtDNA information concerns genic products acting inside the mitochondrion itself. Plant mitochondrial genomes have several peculiar characteristics such as the large size (from 200Kb to 2400Kb), the presence of introns and genetic material of chloroplast or nuclear origin (Palmer J et al., 2000). Furthermore, mitochondrial genome is characterized by occurrence of RNA editing and trans splicing mechanism enlarging protein variability (Takenaka M. et al., 2008).

On the basis of the above considerations, we have indeed hypothesized that IP6K gene is present in plants, nested in mitochondrial DNA, probably encrypted and hidden by virtue of editing and/or trans-splicing processes. But before going into more detail on the technique we have developed in order to verify such an hypothesis, let us take a closer look into the RNA editing process, which is of central interest to the methods we propose, which is accounted for next.

3.2.2 RNA editing

RNA editing is a molecular process in which some bases of a RNA molecule are altered by specific enzymes. It may be broadly defined as any process (co- or post-transcriptional) that changes the primary nucleotide sequence of an RNA molecule from that encoded by the corresponding gene. Since the initial discovery in trypanosome mitochondria (Benne R. et al., 1986) various types of RNA editing have been described. The diversity of RNA editing mechanisms includes site-specific insertion or deletion of one or a few nucleotides (insertional/deletional type of editing) as well as specific identity changes of individual nucleotides (conversional type of editing) such as cytidine (C) to uridine (U) and adenosine (A) to inosine (I) deaminations. Most of this systems are localized in mitochondria and act on transcripts of protein-coding genes; however, editing of the transcripts of certain mammalian nuclear genes does occur as well (Hodges P., et al., 1992). The extent of sequence changes introduced by editing can range from massive (e.g. certain trypanosome transcripts, where upwards of 50% of the sequence is contribute by editing process- Stuart K., 1991-) to minimal (e.g. the single nucleotide editing in the mRNA encoding a mammalian glutamate receptor protein –Sommer B. et al., 1991-). The biological consequences of editing differ with the system. Editing in trypanosome mitochondria is required to generate translatable mRNAs from transcripts that otherwise could not produce functional proteins. In other cases, both unedited and edited forms of a mRNA may give rise to biologically active proteins having different physical and/or functional properties. Messenger RNA editing usually involves changes to internal codon positions, but it may also create new initiation and termination codons and alter 5'- and 3'- noncoding regions and intron sequences.

The mechanism and informational basis for the selection of individual nucleotides as editing sites is unknown. No consensus sequence at an editing site has been identified in either mitochondrial or chloroplast RNAs. Nucleotide sequence comparison of mitochondrial editing sites indicated that the edited cytosine usually does not have a purine, especially a guanine, as the 50 nucleotide (Covello P.S., et al., 1990). In addition, some editing sites bear similarities to each other, and this could reflect different 'classes' of editing sites (Gualberto J.M., et al. 1991). Apart from these simple trends, no specific sequence or secondary structural motifs have

been identified from the hundreds of mitochondrial editing sites examined (Gray M.W., 1992; Wissinger B., et al., 1992).

RNA editing is particularly diffuse in plant mitochondria and chloroplasts. Often the genomic information encoding an open reading frame is often incomplete in these organelles, and RNA editing is necessary to yield a functional product.

3.3 Contributions

The hypothesis at the basis of this part of the work is that IP6K gene lies in DNA of plants, in an encrypted form. Therefore, we decided to first search IP6K gene in mtDNA of plants, where the occurrence of mechanisms interrupting the linearity of genetic information is high.

Because of the considerable sequence heterogeneity among the several known IPKs, common homology search programs are not useful to our aim, Furthermore, these softwares cannot detect possible changes in nucleotide sequences due to RNA editing mechanisms.

The main intuition behind here is that some specific gene families, such as all IPK genes, are characterized by the presence of specific tags, short sequences of few amino acids, often corresponding to functional regions. Thus, we decided to use this new approach of looking not for the gene sequence as a whole, but for a specific tag sequence, characterizing IPK gene family. This is possible only when a gene, or a gene family, contains a region (usually a short sequence) that is indispensable and always present in the gene sequence. In fact, alignment studies between IPKs from different organisms allowed to identify several conserved motifs in the amino acids sequence. These motifs comprise the ATP binding site, first characterized in IP3-3K (Communi D. et al., 1993), the C-terminal motif (last 19 amino acids), important for the catalytic activity (Togashi S., 1997), the "SSLL" motif, required for enzymatic activity of IP6K (Saiardi A. et al., 2001) and the P-XXX-D-X-K-X-G tag, a sequence of nine amino acids with four of them very conserved among IPKs (Saiardi A. et al., 1999). Despite the considerable sequence heterogeneity of IPKs, this last motif represents a unique consensus sequence and it can be considered a specific tag of IPK gene family. The consensus sequence P-XXX-D-X-K-X-G (**Fig.3.4**) is a very important functional region, identifying the inositol binding site of the enzyme (Saiardi A. et al., 2000). Here, the functional role explains its strong conservation through evolution.

InsP3 kinase A	243	YLQLQDLLDGFDG**P**CVL**D**C**K**M**G**VRTY	268
InsP3 kinase B	453	YNQMDDLLADFDS**P**CVM**D**C**K**M**G**IRTY	478
InsP6 kinase 1 (KIAA0263)	205	YKFLLLENVVHHFKY**P**CVL**D**L**K**M**G**TRQHGDDA	236
InsP6 kinase 2 (PIUS)	200	YKFILLENLTSRYEV**P**CVL**D**L**K**M**G**TRQHGDDA	231
IPMK (ArgRIII)	112	KQYLVLENLLYGFSK**P**NIL**D**I**K**L**G**KTLYDSKA	143
yInsP6 kinase (KCS1)	758	KFILLEDLTRNMNK**P**CAL**D**L**K**M**G**TRQYGVDA	788
Consensus:		**P**XXX**D**X**K**X**G**	

Fig. 3.4 Alignment of the inositol-phosphate-binding motif of the different inositol polyphosphate kinases. Identical amino acids are shown in bold. The GenBank accession numbers of the different sequences are: rat InsP3 kinase A, GI:124808; rat InsP3 kinase B, GI:1170577; *Saccharomyces cerevisae* yInsP6 kinase, GI:1078508; and IPMK (inositol polyphosphate multikinase), GI:114134. Numbers to the right and left of the sequences indicate their positions in the respective complete amino-acid sequences. The consensus sequence is written in Prosite format, where X represents any amino acid (Saiardi et al., 1999).

Thus, we chose the P-XXX-D-X-K-X-G tag to perform *IP6K* gene search. This tag search, however, had to be supported by a suitable methodology and associated software tools. The following sections indeed account for the multi-step methodology that we developed to discover specific genes and one of the supporting tools specifically designed to implement this particular kind of tag search. Then, the so developed method was indeed applied to searching for IP6K gene. The thus obtained experimental results are also presented below.

3.4 Methodos

3.4.1. Gene finding by tag search: a multi-tool methodology

When a gene is characterized by a tag, the finding of the tag sequence in a genome indicates that possibly this genome contains the gene under study. In fact, the presence of the tag is essential to recognize a gene or a gene family, but a sequence containing the tag is not necessarily the gene in question. Thus, once a

candidate gene sequence is found, further analyses are required to confirm the discovery. To find possible specific, still undiscovered, genes in cell genomes we developed a multi-tool methodology. The new approach is based on tag search, and includes a series of analyses aimed to analyze the candidate gene sequences eventually found.

The tag search is a motif discovery task, since a tag can be viewed as a subsequence whose structure is not completely specified a priori, that is, a special kind of motif. In order to solve the tag search problem, we designed the methodology summarized by the pseudocode illustrated in **Figure 3.5.** In particular, we consider a tag sequence and a mtDNA sequence in input, and we aim at generating all the mtDNA subsequences containing the tag. Then the search can be performed by following the procedure illustrated in Figure 3.5.

Pseudocode

Input:

a string *t* defined on the alphabet □ = (A, C, T, G, X) representing the "tag"

a string *tag* of shape A1-D1-…-D_{n-1}-A_n defined on □ = (Aminoacid, X), representing the "tag"

a string *s* defined on the alphabet □ = (A, C, T, G) representing the mtDNA of a given organism

Output:

a set *S* of strings that are subsequences of *s* containing *tag*

Begin

```
Let S be an empty set
S= 0
while there is a triplet t in s
        if t matches the aminoacid A1
                set j to 1
                while j is smaller than n
                        skip Dj triplets in s
                        get a triplet a in s
                        if a does not match the aminoacid Aj+1
                                break
                        else
                                increment j by one
                        end
                end;
                if j is equal to n
build Si as the portion of s starting from i and matching the tag
                        add to Si 200 triplets after Si
                        add to Si 200 triplets before Si
                        add Si to S
                end
        end
        increment i by one
end

Return S
end
```

Fig.3.5. Tag search pseudocode.

Once a tag is identified, we extracted a nucleotide sequence surrounding the consensus sequence. This sequence was then examined as a candidate gene and submitted to further analyses. To this aim, we implemented the procedure described below by exploiting a set of existing software tools, suitably fixed, as showed in **Figure 3.6**. In particular, for the tag sequence search, we exploited the system L-SME, able to handle different complex kinds of pattern variabilities, proposed by

Fassetti F. et al., 2008. The system is able to take into account both the genetic code degeneration and possible RNA editing events.

The extracted nucleotidic sequences were translated into amino acid sequences by using the Transeq (Rice P. et al., 2000) software.

Then, in order to detect possible homologies, we performed sequence alignments using ClustalW (Larkin M.A., 2007) and BLAST (Altschul S.F., 1997). Finally, using the TBLASTX and TBALSTN algorithms, we screened expressed sequence tag (EST) databases for proteins containing the sequences identified by our tag search. These databases include short fragments of DNA derived from a longer cDNA sequence and representing part of the expressed genome.

The methodology can be, therefore, shortly summarized as follows:

1. (Tag Definition) set a (partially undefined) sequence representing the specific tag to searched for;
2. (Genome Scanning) scan a genome sequence (or a set of genome sequences) to individuate possible instances of the tag;
3. (Post-processing Analysis) analyze the candidate subsequences extracted by the previous step in order to verify the presence of the gene in the considered genomes.

A (somewhat more detailed) summary of the methodology is illustrated in the figure 3.6.

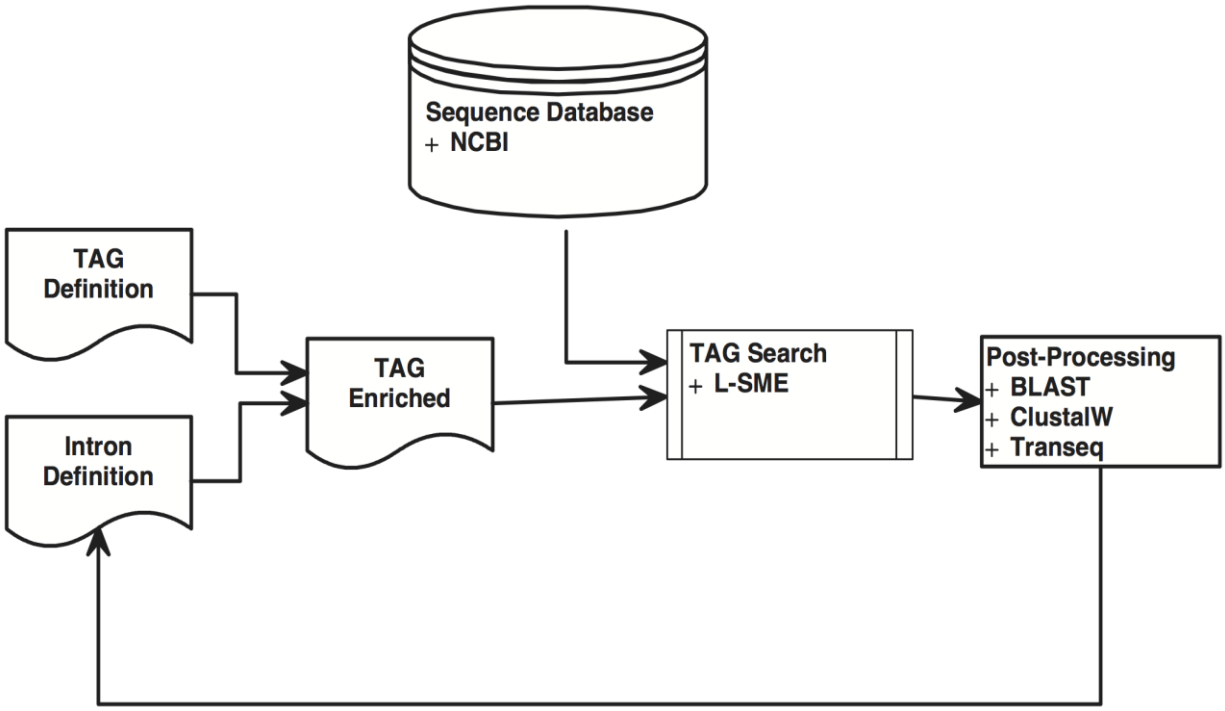


Fig.3.6. Summary of the gene finding methodology based on tag search.

3.4.2. L-SME

L-SME (Fassetti F. et al., 2008) is a system designed to mine general kinds of motifs where several "exceptions" may be tolerated; that is, it is able to handle different complex kinds of pattern variabilities. Roughly speaking, a motif is a pattern composed by two or more substrings (called boxes) separated by a number of symbols (called gaps).

In the following, the main characteristics of L-SME are briefly presented. The basic configuration of L-SME allows the user to specify the minimum and the maximum length of each box composing the pattern and the minimum and the maximum size of each gaps between boxes. But, motivated by biological observation, L-SME supports also approximated matching of the pattern, namely, the occurrences of the pattern may differ in some characters. Also this configuration is fully customizable by the user, indeed it is possible to specify for each box the number of mismatches admitted in the matching measured according to the Hamming or the Levenshtein distances.

Futhermore, L-SME is able to take into account box skips (deletions) and box swaps (box invertions). The former ones consist in mining occurrences where at most a user-defined number of boxes are missing, while the latter ones consist in mining occurrences where at most a user-defined number of swaps between adjacent boxes occur. Finally, L-SME allows the user to specify some anchors for each box, namely, the user can constrain a box to be equal to one of the substrings specified as anchor for that box. Despite the complexity of the addressed pattern variabilities, the system is able to exhibit very good performances.

The flexibility of the method in specifying variable distances between two boxes easily allows possible introns to be taken into account. In order to limit the great variability introduced by considering introns, we adopted an incremental approach consisting in iteratively increasing the number of introns. In particular, we did not take care of any intron at the beginning and, then, we considered the presence of *n* introns by incrementing the distance between *n* pair of boxes of the typically maximum length of an intron (e.g., 100 bases for Arabidopsis thaliana).

3.4.3 BLAST

BLAST (Basic Local Alignment Search Tool) (Altschul S.F., 1997) is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. It is one of the most widely used bioinformatics programs, because it addresses a fundamental biological problem. A BLAST search enables comparison of a query sequence with other sequences or a database of sequences, and identify sequences that resemble the query above a certain threshold. BLAST directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straight-forward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. Input sequences are in FASTA format or Genbank format. BLAST output can be delivered in a variety of formats. These formats include HTML, plain text, and XML formatting.

The following different *types* of BLASTs are available according to the query sequences. Nucleotide Blast: search a nucleotide database using a nucleotide query; Protein Blast: search protein database using a protein query; Blastx: search protein database using a translated nucleotide query; tblastn: search translated nucleotide database using a protein query; tblastx: Search translated nucleotide database using a translated nucleotide query

3.4.4.Clustal

Clustal (Larkin M.A., 2007) is a general purpose multiple sequence alignment tool for DNA or proteins. There are two main variations: ClustalW, command line interface and ClustalX, with a graphical user interface. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen.

The program accepts a wide range on input format, including NBRF/PIR, FASTA, EMBL/Swissprot, Clustal, GCC/MSF, GCG9 RSF, and GDE. The output format can be one or many of the following: Clustal, NBRF/PIR, GCG/MSF, PHYLIP, GDE, or NEXUS. The program includes three main steps: performs a pairwise alignment, creates a phylogenetic tree (or use a user-defined tree), uses the phylogenetic tree to carry out a multiple alignment.

3.5 Experimental results

3.5.1 IP6K gene search

As stated above, several considerations of functional and evolutionary order, led us to suppose that *IP6K* gene is present in plant genomes, most likely in mitochondrial DNA. Thus, we decided to first apply the new method of gene search on mitochondrial DNA of plants, where such a gene could have been nested. Then we looked for the gene in nuclear genome of two model plants, *Arabidopsis thaliana* and *Oryza sativa*.

In the following is specified how the steps listed above have been particularized to achieve our purposes.

Tag Definition

As stated above, the most important tag for *IPK* gene is the P-XXX-DX-K-X-G motif, corresponding to the inositol binding site of the enzyme. Thus, for the identification of *IP6K* gene in plant DNA, we focused on the nucleotide sequence corresponding to this specific IPK tag.

In particular, the sequence associated to this the tag, is made of both symbols in the alphabet Σ = {A,C,G,T}, representing nucleic acids, and a generic symbol X that can be associated to a subset of Σ. This way, step 2 can be carried out by performing an approximate search of the motif represented by the tag sequence.

Genome Scanning

We analyzed *all the published mtDNA sequences* (available at http://www.ncbi.nlm.nih.gov/sites/entrez) and the whole nuclear genome of two plants and performed motif extraction from them.

For the purposes of this research, we looked for the pattern:

[CC{T,C,A,G}] ---------[GA{T,C}] --- [AA{A,G}] --- [GG{T,C,A,G}] ,

where the square brackets delimitate the boxes and the hyphens denote the distances between boxes. For this motif extraction we used L-SME, setting the configuration parameters as reported in **Figure 3.7**. Note that, because of the genetic code degeneration, the fixed aminoacids are specified by more than one nucleotidic triplet. All the possible C->T editing events giving rise to the fixed tag aminoacid are included in the search.

Distance:	Hamming
Number of skips:	0
Number of swaps:	0
Number of boxes:	4

First box length:	3	*Second box length*:	3
Distance from second box:	9	*Distance from third box*:	3
First box anchors:	CCT	*Second box anchors*:	GAT
	CCC		GAC
	CCA		
	CCG		
Third box length:	3	*Fourth box length*:	3
Distance from fourth box:	3	*Fourth box anchors*:	GGT
Third box anchors:	AAA		GGC
	AAG		GGA
			GGG

Fig. 3.7. L-SME configuration parameters.

Post-processing Analysis

For each identified tag, we extracted a sequence of about 1200 nucleotides surrounding the consensus sequence and examined it as a candidate *IP6K* gene. We first translated the extracted nucleotide sequences and then examined the identified amino acid sequences, looking for other IP6Ks conserved domains. Then we performed sequence alignments using both ClustalW and BLAST. Finally we screened expressed sequence tag (EST) databases in order to verify if the candidate gene sequences are actually transcripted.

3.5.1.1. Validation of the method

In order to confirm the goodness of our method, we applied it on nuclear DNA of *Arabidopsis thaliana,* searching a known gene, already identified by biological techniques. Such a gene is the one coding for inositol 1,3,4,5,6-pentakisphosphate2-kinase (InsP5 2-kinase or Ipk1), the enzyme responsible for the production of inositol hexakisphosphate (IP6).

Ipk1s are unique among inositol phosphate kinases in that they phosphorylate the axial 2-position of the inositide ring, whereas other enzymes act on equatorial position of the ring (Gonzales B. et al., 2010) . The family of enzymes responsible for the synthesis of IP6 from IP5 are known as Ipk1. The first Ipk1 gene was

identified in yeast (York J.D. et al., 1999) and in other different fungal species. Although functionally conserved, IPK genes present very low sequence homology in different organisms, with less than 24% identity in pairwise combinations across the fungal proteins. The sequence identity is limited to a few small regions with high homology. This lack of significant homology initially disallowed the discovery of non-fungal Ipk1. After characterization of human Ipk1 (Verbsky J.W. et al., 2002), the gene was cloned in Arabidopsis thaliana using molecular strategy based on the presence of specific tags in the protein (Sweetman D., 2006). As a consequence consequence, we searched Ipk1 in Arabidopsis thaliana genome with the modified L-SME software, exploiting the presence of specific tag in Ipk1 gene family.

As for the validation carried on the *Ipk1* gene, we looked for the pattern composed by the two regions EIKPK and R-XX-MHQ-X-LK characterizing the gene because both present in all Ipk1 genes discovered up to now; they are separated each other from a variable number of amino acids (19 amino acids in human and rat Ipk1, 9 in yeast Ipk1). Then, the corresponding pattern in the genome sequence is:

GA{A,C}AT{T,C,A}AA{A,G}CC{T,C,A,G}AA{A,G}--. . . --

{AGA,AGG,CGT,CGC,CGA,CGG}------{ATG}CA{T,C}CA{A,G}---

{TTA,TTG,CTT,CTC,CTA,CTG}AA{A,G}

where the number of boxes is 11 and the length of each box is 3. The distance between the fifth and the sixth box, corresponding to the distance between the region EIKPK and the region R-XX-MHQ-X-LK, is set to the interval [27 − 57]. As for the distances between the other boxes, they are set according to the above described method.

We found the conserved sequence on chromosome 5 of Arabidopsis thaliana genome. In particular we had only one positive match when we considered the possibility of occurrence of an intron between the third and fourth amino acid of EIKPK motif. The intron length resulted to be 82 nucleotides and distance between two EIKPK and R-XX-MHQ-X-LK motifs 63bp. We extracted a sequence of about 2000 nucleotides around the tags. BLAST allignements showed that the sequence was *Arabidopsis thaliana Ipk1* gene.

This approach led us to easily find the gene, thereby allowing to validate the method, that appears general and very useful when homology search strategies cannot be used.

3.5.1.2. IP6K search results

Due to the numerous suggestions relating IP6K to cell mitochondria, we decided to first perform the IP6K gene search on mitochondrial DNA of plants. To date the full mitochondrial genome sequence is known for 42 different vegetal organisms, belonging to various Phyla, even very distant from one another from the evolutionary point of view. The specific IP6Ks tag (P-XXX-D-XK-X-G) was searched over the overall sequenced mitochondrial genomes available to date and both DNA strands were analyzed. Twentythree genomes out of 42 gave at least one positive match. Interestingly, we noted that some tag sequences (9 amino acids) were identical among different organisms. For each identified tag we extracted a sequence of about 1200 nucleotides surrounding it. To find out possible relevant homologies, we performed alignments among the sequences found in different vegetal organism. All the sequences sharing the same tag showed high similarity in the region surrounding the consensus sequence, while alignment with *IP6K* known genes (Saccharomyces cerevisiae *KCS1* or human *IP6K1*) showed only a weak similarity.

Furthermore, in order to confirm the identity of our putative hit, we looked for other IP6Ks conserved motifs in the identified putative amino acids sequence like the ATP binding site, the C-terminal motif (last 19 amino acids), and the "SSLL" motif. These analyzes led us to focus on the sequence PVGTDRKGG, that was found in mitochondrial genome of Tripsacum dactyloides, Sorghum bicolour, and three different species of Zea genus (Zea mays, Zea perennis and Zea luxurians). Alignment between the 410 aminoacid around the PVGTDRKGG sequence of Tripsacum dactyloides and the human IP6K gene showed an interesting correspondence of the consensus region (see **Figure 3.8**).

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	Tripsacum	410	2	Human	410	6

CLUSTAL 2.0.12 multiple sequence alignment

```
Tripsacum   -PTEILSEY-K--KAISLWWYTSRQFWNFQFQSEHIDPSMDLYV-PLQSCSSFLATSSIF   55
Human       MVVQNSADAGDMRAGVQLEPFLHQVGGHMSVMKYDEHTVCKPLVSREQRFYESLPQAMKR   60
              .:  ::  .   .:.*  :  :   ::.. . . ..  .  *   *   . *. :

Tripsacum   FLGTC--TRNSYVRDSSSENLPVFHSHMR--QESLWATGRHEVIHHVQT-TFRSLGTVYK   110
Human       FTPQYKGTVTVHLWKDSTGHLSLVANPVKESQEPFKVSTESAAVAIWQT-LQQTTGSNGS   119
            *      * . :: ..*: :*.:. . ::  **.: .: .  .:   **   :: *:  .

Tripsacum   S-NSHKWNEKWVHVNDRDLANNNVPSPYGVRDPKAIVE-TVSYSY-LT-AAPLLQGWG-S   165
Human       DCTLAQWPHAQLARSPKESPAKALLRSEPHLNTPAFSLVEDTNGNQVE--RKSFNPWGLQ   177
            . .  :* .  :  . :: . : :  .    :. *:     : .  :      :: ** .

Tripsacum   ASEPYVGRVSASYSSGRRPGKLRRD--GETQLLPVGTDRKGG------GDKLVKKQAYCP   217
Human       CHQAHLTRLCSEYPENKRHRFLLLENVVSQYTHPCVLDLKMGTRQHGDDASEEKKARHMR   237
            . :.:: *:.:.*...:*   *  :   .    *   * * *      . .  **  :

Tripsacum   TPKKQTKKTYAL-QQSAHPLLVASRFHPSP--FRDRRLI-YVQ-SSSDQSARTPDRLCPP   272
Human       KCAQSTSACLGVRICGMQVYQTDKKYFLCKDKYYGRKLSVEGFRQALYQFLHNGSHLRRE   297
            .  :.*.   .:   . :   . .::. .   : .*:*      .:  *  :. .:*

Tripsacum   ILSRTKWNGSLILVTLCPDPSPHVRFYPPATRPTQH-GRPPPHSMLTRAGARFLGSPFPP   331
Human       LLEPILHQLRALLSVIRSQSS--YRFYSSSLLVIYD-GQEPPE--------RAPGSPHPH   346
            :*.    :   :* .: .:.*   ***..:     . *: **.        *  ***.*

Tripsacum   RS-RPGWPACGSGNSPVPW-KKGWLDAGSTPRGAVRT-MISSRPLFAYR-GCLTPLRQLA   387
Human       EAPQAAHGSSPGGLTKVDIRMIDFAHTTYKGYWNEHTTYDGPDPGYIFG---LENLIRIL   403
            .: :..  :. .* : *     .: .:  .     :*   .. * : :    *  * ::

Tripsacum   LPALSCL                                                        394
Human       QDIQEGE                                                        410
              .
```

Fig.3.8 Alignment between the 410 amino acids around the PVGTDRKGG sequence of *Tripsacum dactyloides* and the human IP6K gene (ClustalW2). "*" = residues identical in the two sequences in the alignment; ":" = conserved substitutions; "." = semi-conserved substitutions.

To verify if the Tripsacum dactyloides sequence was an actively transcribed gene, we analyzed the Expressed Sequence Tags (ESTs) databases using the region surrounding the PVGTDRKGG tag of Tripsacum dactyloides. This search failed to find any EST matching indicating that our putative hit is unlikely to be transcribed in mRNA. Finally, we used a region of 50 amino acids of Tripsacum dactyloides mtDNA surrounding the consensus sequence to perform a multiple alignment with corresponding regions of inositol phosphate kinases (IPMK, IP6K, IP3-3K) from different organism using ClustalW2. As shown in **Figure 3.9**, our sequence resulted

to be an outsider. This result indicated that the identified mitochondrial tag does not belong to any subgroup of kinases composing the IPK gene family.

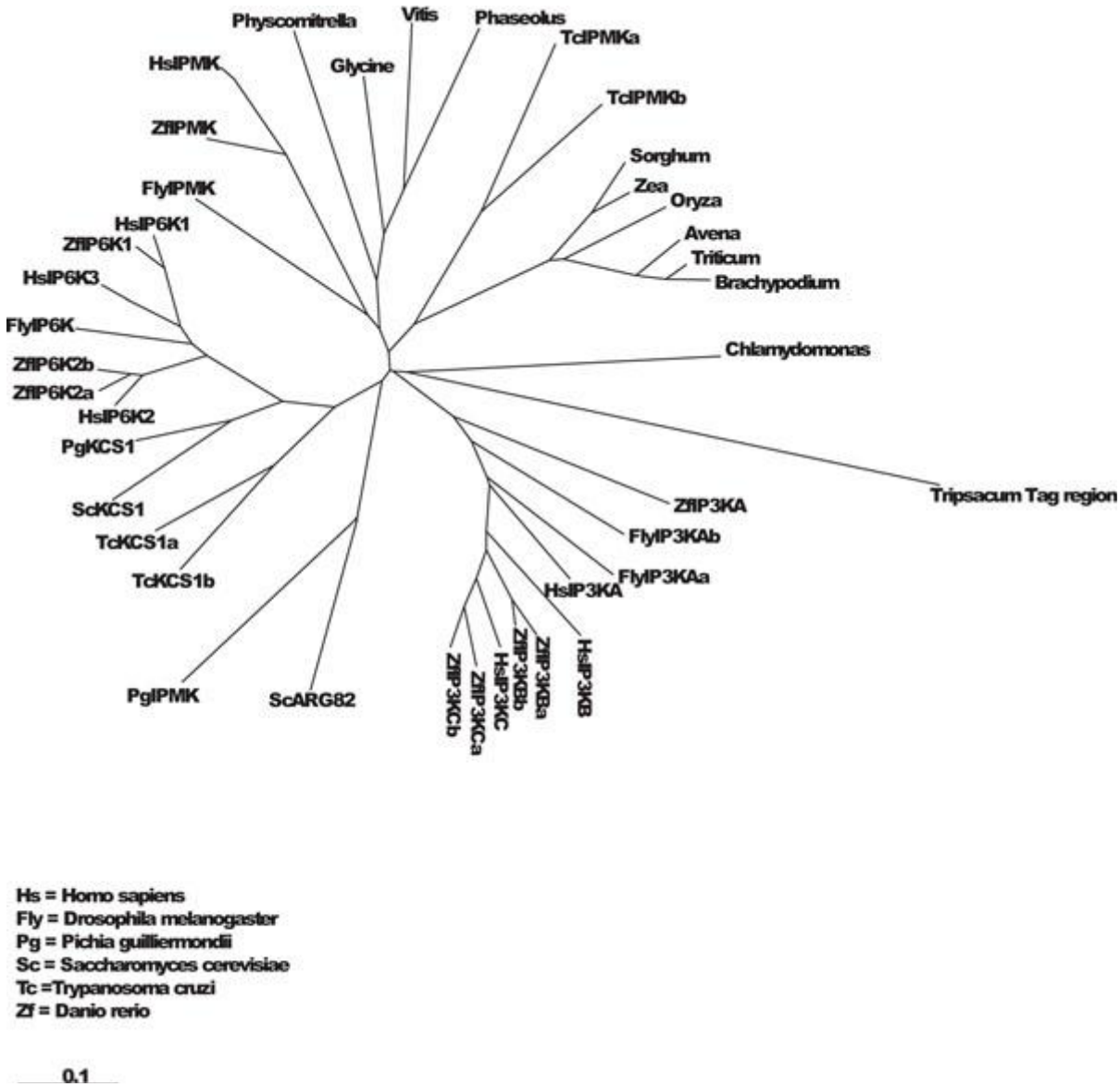


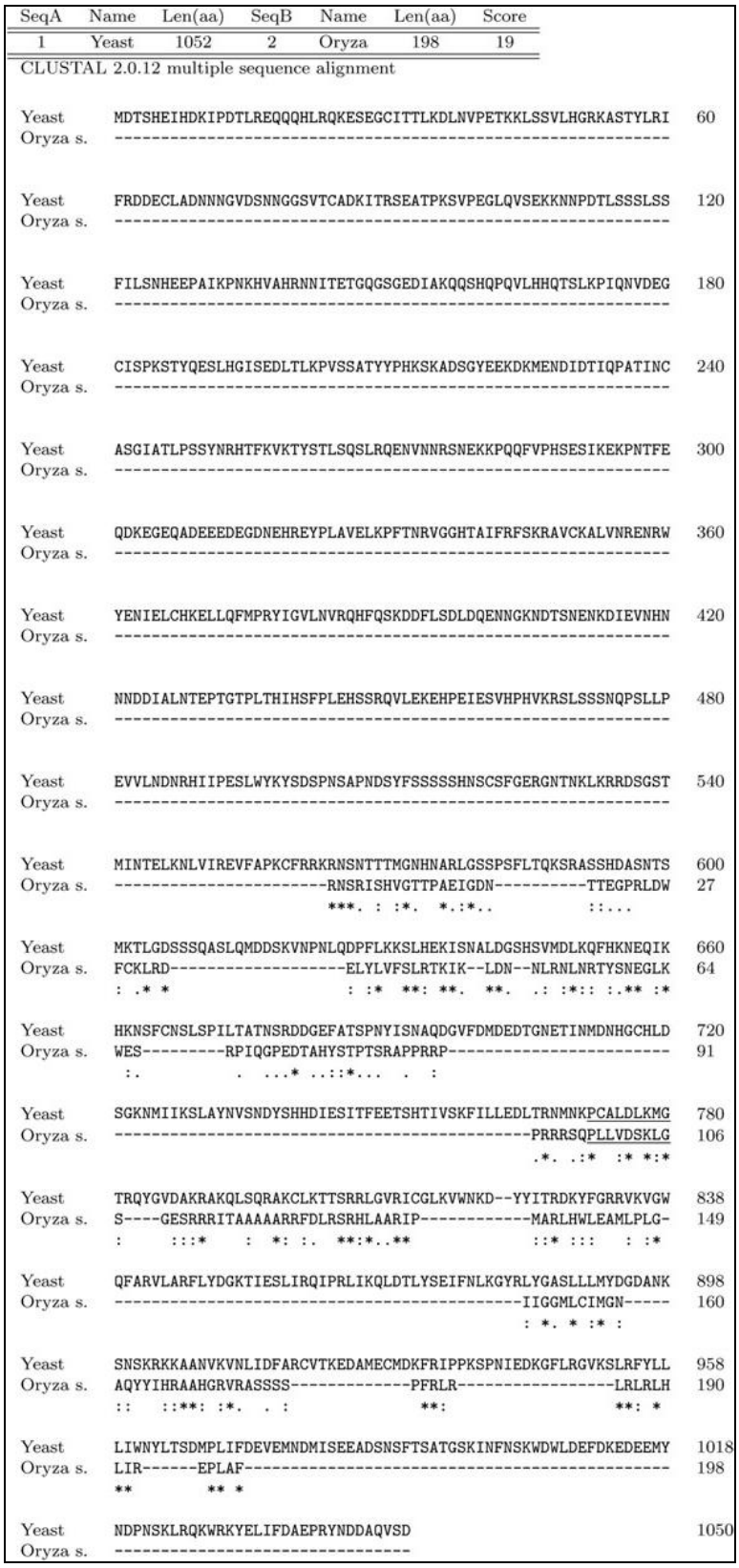
Fig.3.9. Phylogenetic tree from multiple alignment of a 50 amino acid region of *Tripsacum dactyloides* mtDNA surrounding the tag with corresponding regions of inositol phosphate kinase (IPMK, IP6K, IP3-3K) from different organisms (Clustal W2). Branch lengths are proportional to the amount of inferred evolutionary change.

Once excluded the presence of IP6K gene in mtDNA, we decided to look in nuclear genome of plants where, up to now, the search has been performed only by methods based on sequence similarity. We analyzed all chromosomes of *Arabidopsis thaliana* and *Oryza sativa*, a dicotyledonous and a monocotyledon plant respectively.

Arabidopsis thaliana is a small flowering plant, belonging to eudicot, the largest group of flowering plants on the planet. Because of its short generation time and compact size, it is used as a model organism in plant biology and genetics. Its nuclear genome comprises five chromosomes, with a total size of approximately 125 Mb (megabases). It is one of the smallest genome among plants, and it was the first plant genome to be sequenced in 2000 (Initiative T., 2000). *Oryza sativa* (rice) was the second plant genome to be published (Lan S., 2002) the first among monocot. It has the smallest cereal genome consisting of just 430 Mb across 12 chromosomes and it is routinely used as a model organism in cereal genomics.

In each chromosome of both plants, we found dozens of tags, but only few tag sequences per chromosome resulted in good candidates to be specific IP6K tags. In fact, too polar or too big amino acids between the four fixed positions of the tag are not consistent with the tag sequence functionality. In particular we considered as good candidate a tag sequence including amino acids L, V,T,M,I,A,S,G,C between the four fixed positions, and we rejected others. For each identified tag, we extracted a sequence of about 400 amino acids surrounding the tag. Each sequence was examined as a candidate IP6K gene as described above for mtDNA. We did not find any strong homology with known IP6Ks. This result was not surprising, because only a weak similarity is anyhow expected between organisms very distant from an evolutionary point of view. Thus, the selected sequence to be actually interesting was established on the basis of other parameters, like alignment of tag sequences, presence of other conserved amino acids and of sequence in EST database.

A very promising sequence was found on chromosome 5 of *Oryza sativa*, around the tag PLLVDSKLG. The sequence comprises 198 amino acids without any stop codon. As shown in F**igure 3.10**, the ClustalW alignment of this sequence and Saccharomyces cerevisiae *KCS1* gene gave positive score with an alignment in correspondence of the inositol-binding region.



SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	Yeast	1052	2	Oryza	198	19

CLUSTAL 2.0.12 multiple sequence alignment

```
Yeast      MDTSHEIHDKIPDTLREQQQHLRQKESEGCITTLKDLNVPETKKLSSVLHGRKASTYLRI   60
Oryza s.   ------------------------------------------------------------

Yeast      FRDDECLADNNNGVDSNNGGSVTCADKITRSEATPKSVPEGLQVSEKKNNPDTLSSSLSS  120
Oryza s.   ------------------------------------------------------------

Yeast      FILSNHEEPAIKPNKHVAHRNNITETGQGSGEDIAKQQSHQPQVLHHQTSLKPIQNVDEG  180
Oryza s.   ------------------------------------------------------------

Yeast      CISPKSTYQESLHGISEDLTLKPVSSATYYPHKSKADSGYEEKDKMENDIDTIQPATINC  240
Oryza s.   ------------------------------------------------------------

Yeast      ASGIATLPSSYNRHTFKVKTYSTLSQSLRQENVNNRSNEKKPQQFVPHSESIKEKPNTFE  300
Oryza s.   ------------------------------------------------------------

Yeast      QDKEGEQADEEEDEGDNEHREYPLAVELKPFTNRVGGHTAIFRFSKRAVCKALVNRENRW  360
Oryza s.   ------------------------------------------------------------

Yeast      YENIELCHKELLQFMPRYIGVLNVRQHFQSKDDFLSDLDQENNGKNDTSNENKDIEVNHN  420
Oryza s.   ------------------------------------------------------------

Yeast      NNDDIALNTEPTGTPLTHIHSFPLEHSSRQVLEKEHPEIESVHPHVKRSLSSSNQPSLLP  480
Oryza s.   ------------------------------------------------------------

Yeast      EVVLNDNRHIIPESLWYKYSDSPNSAPNDSYFSSSSSHNSCSFGERGNTNKLKRRDSGST  540
Oryza s.   ------------------------------------------------------------

Yeast      MINTELKNLVIREVFAPKCFRRKRNSNTTTMGNHNARLGSSPSFLTQKSRASSHDASNTS  600
Oryza s.   ----------------------RNSRISHVGTTPAEIGDN----------TTEGPRLDW   27
                                ***. : :*.  *.:*..          ::...

Yeast      MKTLGDSSSQASLQMDDSKVNPNLQDPFLKKSLHEKISNALDGSHSVMDLKQFHKNEQIK  660
Oryza s.   FCKLRD-------------------ELYLVFSLRTKIK--LDN--NLRNLNRTYSNEGLK   64
           : .* *                    : :*  **: **.  **.  .: :*:: :.** :*

Yeast      HKNSFCNSLSPILTATNSRDDGEFATSPNYISNAQDGVFDMDEDTGNETINMDNHGCHLD  720
Oryza s.   WES---------RPIQGPEDTAHYSTPTSRAPPRRP------------------------   91
            :.          .  ...* ..::*...  .  :

Yeast      SGKNMIIKSLAYNVSNDYSHHDIESITFEETSHTIVSKFILLEDLTRNMNKPCALDLKMG  780
Oryza s.   ----------------------------------------------PRRRSQPLLVDSKLG  106
                                                         .*. .:*  :* *:*

Yeast      TRQYGVDAKRAKQLSQRAKCLKTTSRRLGVRICGLKVWNKD--YYITRDKYFGRRVKVGW  838
Oryza s.   S----GESRRRITAAAAARRFDLRSRHLAARIP------------MARLHWLEAMLPLG-  149
           :     :::*    :  *: :.  **:*..**              ::* :::   : :*

Yeast      QFARVLARFLYDGKTIESLIRQIPRLIKQLDTLYSEIFNLKGYRLYGASLLLMYDGDANK  898
Oryza s.   ----------------------------------------------IIGGMLCIMGN-----  160
                                                        : *. * :* :

Yeast      SNSKRKKAANVKVNLIDFARCVTKEDAMECMDKFRIPPKSPNIEDKGFLRGVKSLRFYLL  958
Oryza s.   AQYYIHRAAHGRVRASSSS-------------PFRLR-----------------LRLRLH  190
           ::   ::**: :*.  . :              **:                  **: *

Yeast      LIWNYLTSDMPLIFDEVEMNDMISEEADSNSFTSATGSKINFNSKWDWLDEFDKEDEEMY 1018
Oryza s.   LIR------EPLAF-----------------------------------------------  198
           **        ** *

Yeast      NDPNSKLRQKWRKYELIFDAEPRYNDDAQVSD                             1050
Oryza s.   -------------------------------
```

Fig. 3.10 Alignment between the 198 amino acid sequence around the PLLVDSKLG tag of Oryza sativa and the yeast *KCS1* gene (Clustal W2). ”*” = residues identical in the two sequences in the alignment; ”:” = conserved substitutions; ”.” = semi-conserved substitutions. The P-XXX-DX-K-X-G tag is underlined.

Alignment with human IP6K (hIP6K) gave a lower score, but still maintained the correspondence between tags (see **Figure 3.11**).

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	Human	410	2	Oryza	198	9

CLUSTAL 2.0.12 multiple sequence alignment

```
Human     MVVQNSADAGDMRAGVQLEPFLHQVGGHMSVMKYDEHTVCKPLVSREQRFYESLPQAMKR   60
Oryza s.  ------------------------------------------------------------

Human     FTPQYKGTVTVHLWKDSTGHLSLVANPVKESQEPFKVSTESAAVAIWQTLQQTTGSNGSD  120
Oryza s.  -------------------------------------RNSRISHVGTTPAEIGDNTTE   21
                                                ..: ::   *     *.* ::

Human     CTLAQWPHAQLARSPKESPAKALLRSEPHLNTPAFSLVEDTNGNQVERKSFNPWGLQCHQ  180
Oryza s.  GPRLDW----FCKLRDELYLVFSLRTKIKLDN-NLRNLNRTYSNEGLKWESRPIQGPEDT   76
           .  :*    :.:  .*      **:: :*:.  :  :: * .*:  : . .*     .

Human     AHLTRLCSEYPENKRHRFLLLENVVSQYTHPCVLDLKMGTRQHGDDASEEKKARHMRKCA  240
Oryza s.  AHYSTPTSRAPPRRPPR---------RRSQPLLVDSKLGSGESRRRITAAAAARRFDLRS  127
          ** :   *. * .:  *         : ::* ::* *:*: :     :    **::   :

Human     QSTSACLGVRICGMQVYQTDKKYFLCKDKYYGRKLSVEGFRQALYQFLHNGSHLRRELLE  300
Oryza s.  R----HLAARIPMARLHWLEAMLPLG---IIGGMLCIMGNAQYYIHRAAHG-RVRASSSS  179
          :     *..**   :::  :    *      *  *.: *  *   :   :* ::* .  .

Human     PILHQLRALLSVIRSQSSYRFYSSSLLVIYDGQEPPERAPGSPHPHEAPQAAHGSSPGGL  360
Oryza s.  PFRLRLRLRLHLIREPLAF-----------------------------------------  198
          *:  :**  * :**.  ::

Human     TKVDIRMIDFAHTTYKGYWNEHTTYDGPDPGYIFGLENLIRILQDIQEGE            410
Oryza s.  ------------------------------------------------
```

Fig.3.11. Alignment between the 198 amino acid sequence around the PLLVDSKLG tag of *Oryza sativa* and the human IP6K gene (Clustal W2). ”*” = residues identical in the two sequences in the alignment; ”:” = conserved substitutions; ”.” = semi-conserved substitutions. The P-XXX-D-X-K-X-G tag is underlined.

We performed a multiple alignment (ClustalW2) between the region of 50 amino acids of Oryza sativa surrounding the tag and the corresponding regions of inositol

phosphate kinases (IPMK, IP6K, IP3-3K) from different organisms. This analysis revealed (**Figure 3.12**) that the Oryza sativa sequence, although appears dissociated from other IPK family members, shows a certain degree of similarity with *Giardia lamblia IP6K*, that itself appears to be a distant member of the IPK genes family.

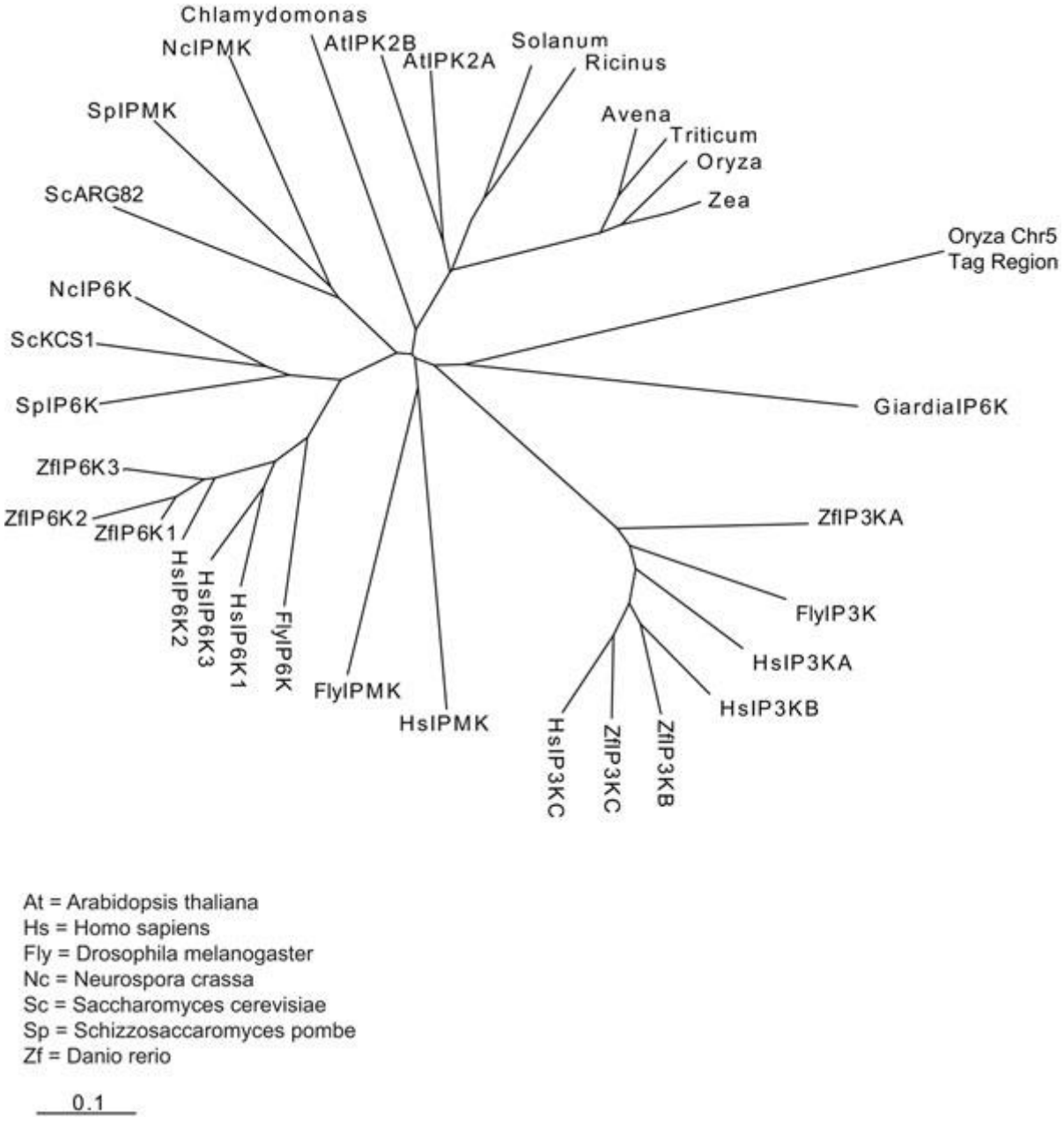


Fig.3.12 A Phylogenetic tree from multiple alignment of a 50 amino acid region of Oryza sativa DNA surrounding the tag with corresponding regions of inositol phosphate kinase (IPMK, IP6K, IP3-3K) from different organisms (ClustalW2). Branch lengths are proportional to the amount of inferred evolutionary change.

Finally, we screened the EST databases using the region surrounding the PLLVDSKLG tag of Oryza sativa. This search showed some matching EST, indicating that the tag sequence is likely to be transcribed in mRNA.

The advantage of our method is that it allows identification of a gene even if many nucleotidic changes have been accumulated during the evolution, so that the homology between homologous loci is now very low. In fact, it is known that *IP6K* is a gene highly conserved through mammalian evolution, but the homology is low when compared with organisms filogenetically very distant, like Yeast. It is possible that in evolutionary stream bringing to plants, many nucleotidic changes occurred, so that plant *IP6K* gene looks quite different both from mammalian and yeast genes. The candidate gene sequence found in nuclear genome of Oryza sativa shows an interesting similarity with yeast *KCS1*, giving a relatively high score when the two sequences are aligned using ClustalW. *KCS1* gene is quite different from mammalian *IP6K* genes. It is bigger, comprising 1052 amino acids against 410 of human IP6K, that lacks the first 305 KCS1 amino acids, and it has some other interruptions as compared to the yeast gene. Very interestingly, the homology region between *Oryza sativa* sequence and KCS1 is indeed clustered in the protein domain corresponding to human IP6K. This result might represent the strong evolution drive of the catalytic IPK domain and the likely conservation of the key feature of this domain in the identified *Oryza sativa* tag.

Furthermore ClustalW alignment shows a correspondence between tags when we compare our sequence with both KCS1 and human or mouse IP6K. This correspondence is still maintained in multiple alignment between our sequence and KCS1, human IP6K and mouse IP6K.

Multiple alignment of a 50 amino acid region of Oryza sativa DNA surrounding the tag with corresponding regions of IPKs from different organisms showed a degree of connection between Oryza sativa tag and Giardia lamblia IP6K sequences. Interestingly, among the different inositol phosphate kinases tested, the best match of Oryza sativa tag region was with a very distant IP6K. This result suggested that the sequence around the identified tag might represent a distant member of the IP6K subfamily of gene as the Giardia IP6K enzyme.

As remarked above, the screening of EST databases showed some matching ESTs. Note that, although EST database represents very powerful tools to study the trascriptome of a specific organism, they are often imperfect. Indeed, their quality is affected by transcript redundancy, low sequence quality and by high transcript truncation rates. Furthermore, these databases only represent the trascriptome of the tissue and developmental stages of the plant from which the mRNA was isolated. Thus, EST databases are not exhaustive, and a negative match does not exclude the expression of (rare) transcripts. This means that the ESTs we found indicate the chromosome region containing the putative plant IP6K is actively transcribed, although such ESTs do not possess the conserved PLLVDSKLG domain. Likely, the identified EST correspond to truncated isoform of the full length mRNA.

In conclusion, we think that this sequence is part of an Oryza sativa gene homologous to mammalian IP6K. In particular we suppose that it is the central part of the gene, comprising the inositol binding site, and it lacks in the N-terminus and the C-terminus sequences, thus indicating the presence of more than one exons in the rice gene. The big evolutive distance between rice and both human and yeast could explain the low similarity observed among these gene.

4. PROTEIN PREDICTION

4.1 Introduction

In plant mitochondria an essential mechanism for gene expression is RNA editing, often influencing the synthesis of functional proteins. RNA editing alters the linearity of genetic information transfer, introducing differences between RNAs and their coding DNA sequences that hind both experimental and computational research of genes.

Complete sequencing of mtDNA of many organisms enabled the identification of canonic genes, but much of the informational content of plant mitochondrial genomes remains still undiscovered, while finding plant mitochondrial proteins and understanding how they integrate into pathways, represent major challenges in cell biology.

We propose a novel strategy useful to intercept candidate coding sequences resulting from some possible editing substitutions on the start and stop codons of a given input organism DNA. Our method is based on the simulation of the RNA editing mechanism, in order to generate candidate Open Reading Frame sequences that could code for some, yet unknown, proteins.

In the following I shall illustrate the problem and our proposal to solve it. In particular, this chapter is organized as follow. Section 4.2 provides a description of mitochondrial DNA in plants and of the structural complications that make the finding of genes difficult. Section 4.3 illustrate our contribution and section 4.4 describes in detail our approach. Section 4.5 eventually presents experimental results.

4.2 The plant mitochondrial DNA

Besides nuclear DNA, cells of aerobic organisms contain another genome, the mitochondrial DNA (mtDNA), located in mitochondria. Mitochondria are dynamic organelles essential for cellular life, death, and differentiation. Although they are best known for ATP production via oxidative phosphorylation, they house myriad other biochemical pathways and are centers for apoptosis and ion homeostasis. Most of the proteins involved in cellular respiration are encoded by mitochondrial genome, while others have nuclear origin. Mitochondrial genome is quite different from nuclear genome. It is typically made by only a type of circular molecule, and it is present in

several copies per cell (from 10 to 10000); mtDNA does not contain histones, and it is maternally inherited, with an high mutational rate (Bonen L., 1991, Gray M.W., 1982). This molecule contains additional informational content with respect to nuclear DNA, concerning some mitochondrial proteins, tRNAs and rRNAs. Mitochondrial have its own ribosomes, responsible for protein synthesis occurring inside the organelle. Messenger RNAs that are synthesized in mitochondria, remain inside the organelle and are translated by mitochondrial ribosomes.

Our knowledge on genic organization of mtDNA arise mostly from sequencing experiments. Improvement of DNA sequencing techniques allowed the accumulation of an enormous amount of sequence data. To date, whole mtDNA sequences are known of several organisms belonging to all five Kingdoms. These data are collected in several databases and web sites, and represent an important resource for genetic studies. The two strands of mtDNA are differentiated by their nucleotide content with the guanine rich strand referred to as the heavy strand, and the cytosine rich strand referred to as the light strand. Genes coding for proteins are located on both DNA strands. Their position has been detected both by biological methods, i.e. aligning mitochondrial mRNA sequences with mtDNA sequences, and by computational methods. In human the heavy strand encodes 28 genes, and the light strand encodes 9 genes for a total of 37 genes (Anderson S. et al., 1981). Of the 37 genes, 13 are for proteins, 22 are for transfer RNA and two are for the small and large subunits of ribosomal RNA. This pattern is also seen among most metazoans, although in some cases one or more of the 37 genes is absent and the mtDNA size range is greater. Mitochondrial DNA dimensions vary among different organisms, but are constant in the same species. Notably mitochondrial chromosomes of superior animals are definitely smaller than fungal or vegetables in general (Ward BL.et al., 1981). Some plant species have enormous mtDNAs (as many as 2,500,000 base pairs per mtDNA molecule) but, surprisingly, even those huge mtDNAs contain almost the same number and kinds of genes as related plants with much smaller mtDNA. An important structural difference is that, while animal mtDNA is almost completely codifying, mitochondrial genome of fungi and plants contains a big amount of DNA not coding for any gene product (Clifton S.W., 2004).

The uniqueness of the plant mitochondrion is not limited to its genome size variability or a few unusual gene products. The mitochondrial genome of plants

distinguishes itself from those of other higher eukaryotes in its unusual organization and gene-processing mechanisms. Much of the variability in genome size in plants is accounted for by the surprisingly complex organizations observed among different plant species. Almost all angiosperms mtDNA contains repeated sequences, whose sizes vary between 2 to more than 12 Kbp. The presence of these recombinationally active repeated sequences leads to a multipartite genome structure generated by both intermolecular and intramolecular recombination events. The presence of repeats in direct and inverted orientation gives rise to subgenomic DNA molecules and inversions within the genome. In particular, direct repeats give rise to deletions, from which small circular molecules can derive, some of which present in very small copy number (*sublimos*). Furthermore in plant mtDNA insertions of chloroplast are present, plasmid and nuclear DNA. Another distinctive characteristic of plant mtDNA, w.r.t. other higher eukaryotes, is that mitochondrial gene can be interrupted by group I or II introns. Certain plant mitochondrial genes are encoded by multiple exons that are interrupted by uncommonly large distances. This unusual gene organization is apparently the result of interruption of intron sequences by recombination events. Proper transcript processing for these genes requires the trans-splicing of the interrupted intron sequences to derive a mature transcript. **Figure 4.1** schematizes the organization of mtDNA of the maize, a monocotyledonous plant.

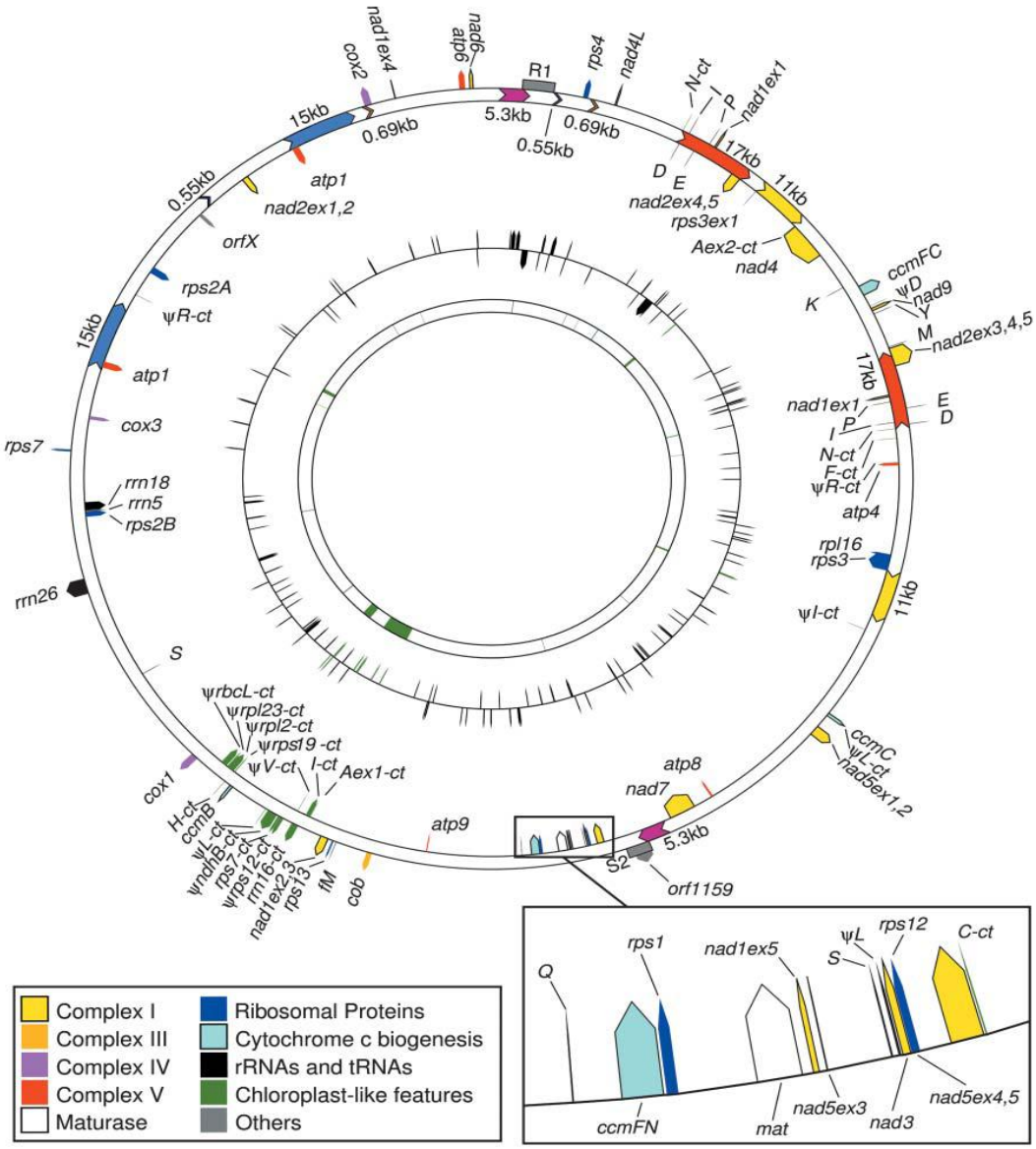


Figure 4.1. Circular map of the maize NB mitochondrial genome generated from sequence data. Known protein-coding, tRNA and rRNA genes, and gene fragments are shown on the outside circle. Colors indicate genes by function: Complex I (nad; yellow),Complex III (cob; orange), Complex IV (cox; purple), Complex V (atp; red), cytochrome assembly (ccm; light blue), ribosomal proteins (dark blue), maturase (white), other ORFs (gray), rRNA and tRNAs (black), and genes transferred from the chloroplast (green). Single-letter designations indicate tRNAs. Large repeats are color coded within the outer ring. Regions homologous to R1 and S2/R2 are indicated by gray blocks. The middle circle indicates positions of ORFs (>99 amino acid predicted sizes). The inner ring shows regions of chloroplast homology with matches of at least 80% identity and lengths of at least 100 bp (green). (Clifton S.W., 2004)

Furthermore, and importantly, plant mitochondria are characterized by very frequent RNA editing events. In these organelle editing mechanism is essential for proper gene expression and occurs not only within an open reading frame to alter amino acid sequence, but especially within start and stop codons that define the open reading frame.

All these peculiar characteristics of plant mitochondrial genomes, make search of coding sequences very complex. Thus, the number and identity of plant mitochondrial protein encoding genes is not yet fully defined, but it is supposed to be larger than that codify by ORFs already identified in mtDNA. The number of known mitochondrial genes varies in different organisms from only 5 genes in *Plasmodium* to nearly 100 genes in jakobid flagellates, with the average across eukaryotes being 40-50 genes (**Fig. 4.2**).

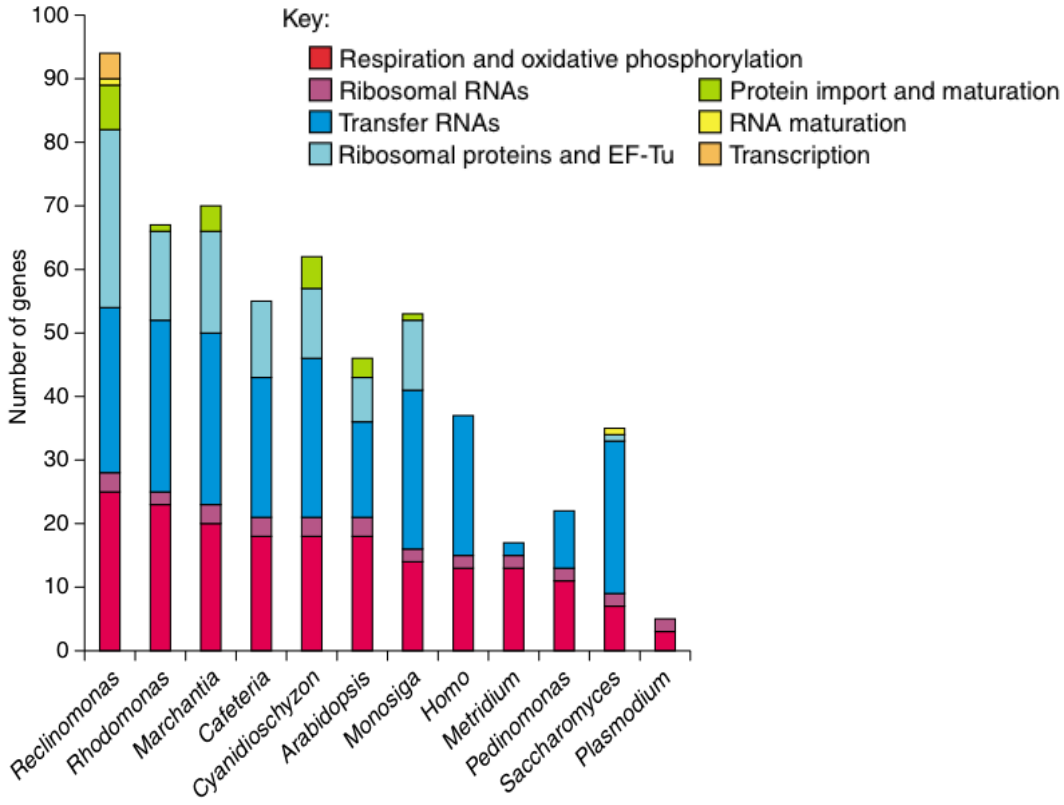


Fig. 4.2. Mitochondrial gene classes and their representation across eukaryotes. Genes included the following functional classes: Respiration and oxidative Phosphorylation:atp1, atp3, atp4, atp6, atp8, atp9; cob; cox1–3; nad1–4, nad4L, nad6–11, sdh2–4. rRNAs: rnl, rns, rrn5. tRNAs: trnA–Y (among others); Ribosomal proteins and EF-Tu: rps1–4, rps7, rps8, rps10–14, rps19; rpl1, rpl5, rpl6, rpl10,

rpl11, rpl14, rpl16, rpl18–20, rpl27, rpl31, rpl32, rpl34, rpl36; tufA. RNA maturation:rnpB; Protein import and maturation: secY, tatC, yejR (ccmF), yejU (ccmC), yejV (ccmB), yejW (ccmA); cox11; Transcription: rpoA-D. (Burger G., et al., 2003).

Despite the difference in number, mitochondrial genes are involved in five basic processes: invariantly in respiration and/or oxidative phosphorylation and translation, and occasionally also in transcription, RNA maturation and protein import. Despite the recent progress, much remains still to discover for the complete definition of mitochondrial proteome. Identification of proteins coded by mtDNA in plant represents an important research goal.

4.2.1 RNA editing in plant mitochondria

In mitochondria and chloroplasts of flowering plants protein variability is increased by mRNA editing. Most RNA editing events are found in the coding regions of mRNAs and usually at first and second position of codon, so that the deriving aminoacid is very often different from that specified by the corresponding unedited codon. Editing can also create new start and stop codons (Hoch B. et al., 1991; Wintz H. et al., 1991) and it can occur in introns (Brennicke A. et al., 1999) and other non translated regions (Schuster W. et al.,1990). The use of editing to generate *aug* start codons might represent another level of regulatory control of gene expression; in fact, the introduction of a translational start codon could make an mRNA rapidly accessible for protein synthesis (Takenaka M. et al., 2008).

The nature, extent, phylogenetic distribution, and general characteristics of RNA editing in angiosperm mitochondria are now well documented (Gray M. et al., 1992). The basic features of editing in plant mitochondria can be sumaryzed as follow:

1) Editing involves almost exclusively C-to-U substitutions in the mitochondrial mRNA, although infrequent reverse (U-to-C) changes have also been reported (Gualberto J.M. et al., 1990).

2) Editing is highly specific, occurring at multiple, usually isolated, C residues, and almost always at the first or second position of an affected codon.

3) In addition to a low proportion of *silent editing* (not changing the corresponding aminoacid) a small amount of *redundant editing* (editing is redundant when there is more than one edited sites at the same codon, and one is sufficient to produce a change) is also observed in plant mitochondrial mRNAs.
4) Editing takes place predominantly within the coding region, usually at internal positions, but with new initiation or termination codons sometimes being created.
5) The number of editing sites varies greatly among mRNAs from different genes.
6) Both complete and incomplete (partial) editing have been observed.

Numerous studies on RNA editing in plant mitochondria clarified that this process is essential for gene expression. In many cases, editing completes the information encoding an open reading frame and the synthesis of functional proteins depends on it (Regina T.M.R. et al., 2002). Given the physiological importance of RNA, identification of sites of RNA editing is essential for molecular, biochemical and phylogenetic studies in plant mitochondria. Experimental analysis, made comparing RNA transcripts and genomic DNA sequences, is the more exhaustive way, but it is also expensive and time consuming. A collection of all sequences post-transcriptionally modified by RNA editing from many organisms, recovered from primary databases and literature, is available on the RNA editing database REDI (http://biologia.unical.it/pyscript/overview.html).

Computational approaches have been used to predict sites of RNA editing, based either on statistical methods (Bundschuh R., 2004) or on evolutionary considerations. The latter ones are based on the observation that often the final effect of editing events is to make mitochondrial proteins more similar in sequence to their homologous in other species (Guadalberto J.M. et al., 1989). For instance, PREPMT (Mower J.P., 2005) and EDIPY (Picardi E. et al., 2005) are both systems exploiting this tendency of RNA editing to "correct" codons that specify unconserved amino acids. Thus, these software can allow the identification of proteins similar to others already described in different organisms, but they do not allow to predict new proteins.

4.3 Contributions

In order to identify new proteins in plant mitochondria, we propose a method for ORF sequences mining from genomes, based on *editing simulation*. Our approach

aims at identifying genes that eluded other gene finding techniques due to RNA editing. Our development is based on the observation that plant mitochondria use editing mechanism on crucial sites, for example to generate start codon *aug* from *acg*. The main idea we pursue is that of simulating such an editing process. Notably, the editing process may generally involve one or more amino acids, whereby a significant number of alternative encrypted sequences are produced. In particular, editing introduces new potential coding sequences, not present on the original DNA sequence, that can be extracted and further analyzed in order to look for unknown proteins, expressed in the organism under analysis.

We developed a system for Open Reading Frame (ORF) extraction from genomes, based on automatic RNA-editing simulation.

In order to extract the ORF sequences from the genome of a given organism, special nucleotide triplets corresponding to the start and stop of an amino acid sequence have to be intercepted on the DNA sequence (*start codons* and *stop codons*). In particular, there exist one start codon, that is *atg*, and three stop codons, that are *tag*, *tga* and *taa*. Although ORF sequences can be easily searched for in a genomic sequence by exploiting one of the existing software tools, such as for example ORF FINDER (NCBI) and STARORF (MIT), taking into account the occurrences of such codons, this is not sufficient to intercept possible proteins for which expression the RNA editing mechanism have intervened. This means that, in plants, several proteins are not found from the ORF sequences returned in output from such existing tools.

The idea exploited in this work is that of *simulating* editing mechanisms possibly causing the presence of proteins that are not imputable to ORF sequences obtained as recalled above. Actually, this is rather meaningful in plants, where mtDNA editing mechanisms can often involve nucleotide triplets leading to start and stop codons.

The sequences extracted from a mtDNA with editing simulator system are then examined to verify if they are new genes. The following analysis provides for a several further steps, performed by common tools. In particular, the sequences extracted by RNA editing simulator are submitted to a two-fold analysis. First, known proteins homologous to the amino acid sequence obtained from the candidate

ORF sequence generated after the editing simulation are searched for in order to predict proteins that are already known to be expressed in other organims, but that have not yet been discovered in the analyzed organism. Second, those candidate ORF sequences that do not correspond to known proteins, of neither the organism under examination, nor other organisms, are analyzed in order to identify potential unknown expressed proteins. This latter kind of predictions are the most interesting ones. Finding of a significant candidate ORF sequence, even one having all the hallmarks of a functional unit, is no guarantee that this is a functional expressed gene. Definitive evidence of gene function is the identification of the corresponding protein, but this criterion has been satisfied in very few cases, the presence of transcripts being considered indicative of gene activity. In order to verify the transcription of our candidate ORFs, we searched the sequences in the EST database (Boguski M.S, 1993), as a further filtering step before returning in output the prediction results.

We applied our method on the mtDNA of *Oryza sativa* (rice), obtaining encouraging results. First of all, our method was able to single out amino acid sequences corresponding to rice proteins for which start codons editing is known to occur, whereby obtaining a first positive assessment of the approach. Second, a number of protein sequences were predicted, some of which are homologous to proteins expressed in other organims, while some others are completely novel ones.

4.4 Methodos

4.4.1 RNA Editing Simulator

We designed a tool for RNA editing simulation, in order to generate novel potential protein sequences, not yet discovered in a given input organism. To this aim, we start from the mtDNA of a specific plant, and suppose that some editing substitutions might have happened causing the generation of some start/stop codons. Among all such possible new codons, only those corresponding to significant potential ORF sequences are taken into account. In particular, only ORF sequences corresponding to amino acid sequences of lenght at least 100 are considered potential proteins. Thus, between a start and a stop at least 300 nucleotides have to occur for interesting ORF sequences to be sigled out. Furthermore, the most frequent

nucleotide substitution caused by editing is *c->u* at the RNA level, that is, *c -> t* if we refer to mtDNA. Thus we consider only this kind of nucleotide substitution in our analysis.

The following example illustrates how new candidate ORF sequences can be generated from the original nucleotide sequence, by simulating possible editing substitutions.

Example. In **Figure 4.3** a portion of the rice mtDNA is shown. In particular, in the considered sequence, there are two stop codons, *taa* and *tag*, highlighted by a wide-hat. Since no start codon occurs between the two stops, no candidate ORF sequences would be extracted without editing simulation. On the contrary, if we consider possible substitutions *c -> t* leading to the generation of new codons, then the start codon *atg* resulting from the triplet *acg* in italic can be indeed intercepted. Since between this start codon and the stop *tag* there are 102 nucleotide triplets, the subsequence highlighted in bold, worth considering as a candidate ORF sequence, can be extracted this way.

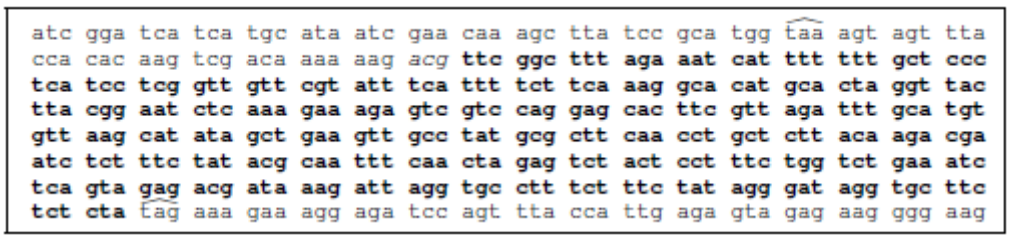


Fig. 4.3. Editing of the start codon *acg -> atg*

Thus, our method starts by considering an input nucleotide sequence (in the case we present in this thesis, this is the mtDNA of a plant). Such a nucleotide sequence *sn* is then scanned in all its three possible reading frames (for both the forward and the reverse cases), by considering all the substitutions *c->t* that can generate new start/stop codons (we call them *edited codons*, while *original codons* are those already occurring in *sn*). Then, the nucleotide subsequences with minimum length equal to 300 between a start and a stop codons are extracted, by taking care that only maximal subsequences are considered. Indeed, if several useful start codons occur before a same stop codon, only the first start codon is considered for the purpose of

extracting the corresponding ORF sequence. All the other start codons are traduced as the corresponding amino acid Methionine (*M*) in the resulting amino acid sequence. This avoids intercepting all the possible subsequences. For what concerns the stop codons, the first one after the chosen start *c*START is considered, if such a *c*STOP is an original codon. If *c*STOP is an edited stop, it is taken into account only if between *c*START and *c*STOP there are at least 300 nucleotides. Otherwise it is discarded, and the next *c*STOP is searched for, by taking care of the same rule. We avoid this way subdividing a potentially significant sequence in several meaningless subsequences, even discarded since not enough long. **Figure 4.4** summarizes the editing simulation method as described above.

Input: A nucleotide sequence *sn*;
Output: A set of amino acid sequences *P*ORF;

1: *P*ORF = Ø;
2: **for each** of the three possible reading frames *fr* of *sn*
3: **repeat**
4: **repeat**
5: read a triplet *t* from *fr* ;
6: **until** *t* is a start codon **or** *editing t* a start codon is achieved;
7: set c_{START} to *t*;
8: **repeat**
9: read a triplet *t* from *fr* ;
10: **until** *t* is a stop codon **or** *editing t* a stop codon is achieved;
11: set c_{STOP} to *t*;
12: let *ni* be the number of nucleotides between c_{START} and c_{STOP};
13: **if** c_{STOP} is an edited stop codon
14: **if** $ni < 300$
15: skip *c*STOP and **goto** step 8;
16: **end if**
17: **end if**
18: **if** $ni \geq 300$
19: extract the nucleotide subsequences *si* between *c*START and *c*STOP;
20: traduce *si* in an amino acid sequence *pi*;
21: $P_{ORF} = P_{ORF} \cup \{pi\}$;
22: **end if**
23: **until** the end of *fr* is reached;
24: **end for**
25: **return** P_{ORF};

Fig. 4.4 The Editing Simulation Module.

The Editing Simulator module is freely available at siloe.deis.unical.it/RNAEditSimulator/ediSim.jar.

4.4.2 Protein prediction: a multi-step methodology

The nucleotide sequences generated by automatic editing simulation are analyzed as candidate new proteins. Thus, the system that we designed to this aim can be view as a multi-tool methodology including a new tool for editing simulation and other similarity tools.

When an amino acid sequence is intercepted for a specific organism, a first criterion to understand its biological interestingness is searching for meaningful homologies with respect to some known proteins belonging to another organism. Among all the candidate ORF sequences generated as explained above, we focus on analyzing only those involving some editing (start and/or stop) codons. Let *S*ORF be the set of such sequences, whose corresponding amino acid sequences are referred to as *candidate protein predictions* in the set *P*ORF.

Candidate protein predictions for the organisms *O* under examination are therefore compared against known proteins by exploiting available alignment algorithms (e.g., Altschul S.F. et al., 1997), in order to single out interesting homologies. Some of the sequences in *P*ORF can be found to be known *O* proteins, in which case we discard them from further analysis. Let b*P*ORF be the resulting amino acid sequences set, that we can divide in two further subsets b*P0*ORF and b*P00*ORF. b*P0*ORF includes amino acid sequences for which significant homologies have been found with respect to some proteins belonging to other organisms, while b*P00*ORF contains the remaining ones. In both cases, a further filtering step is carried out by searching for the presence of possible transcripts by querying the DBEST (Boguski M.S et al.,1993), since this can be considered indicative of gene activity. Eventually, our system returns in output two sets of predicted proteins: *P0* and *P00*, respectively containing proteins in b*P0*ORF and in b*P00*ORF for which trascripts have been found in *O*.

Figure 4.5 graphically illustrates the main steps of our methodology and the associated supporting software tools.

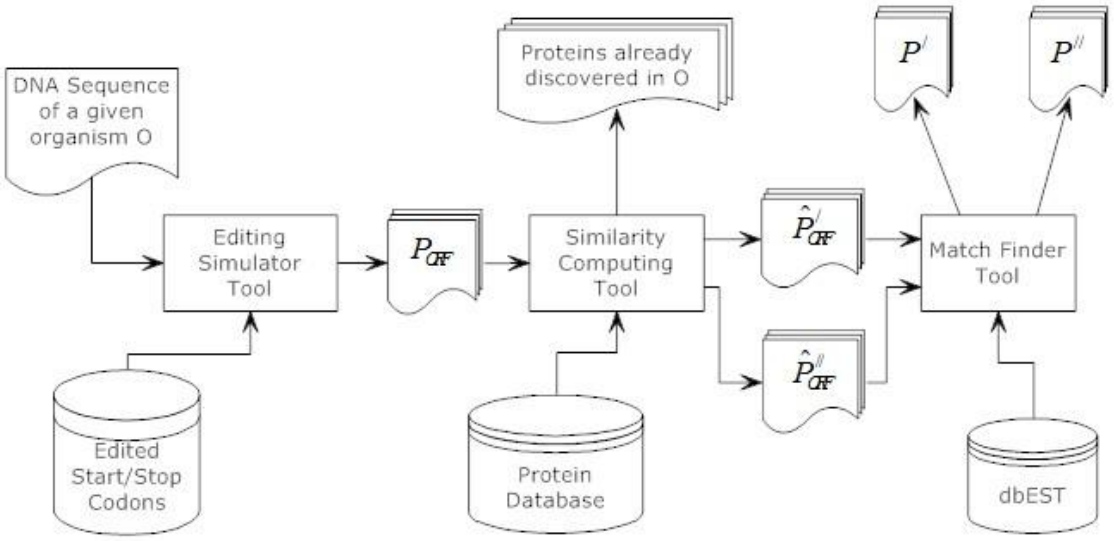


Fig. 4.5. The Protein Prediction methodology based on Editing Simulation.

4.5 Experimental results

4.5.1. Known proteins search

As already explained above, the use of editing to generate *atg* start codons is a mechanism known in flowering plant organelles, and it might represent an important way to increase genetic variability. In order to validate our approach, we at first verified if proteins that are known to be generated by RNA editing in *Oryza sativa* were actually recognized by our system. We found two of such proteins, the NADH dehydrogenase subunit 1 and the NADH dehydrogenase subunit 4.

4.5.2 Mitochondrial proteins prediction

In order to predict possible new mitochondrial proteins, we applied our method on Oryza sativa (rice) mtDNA. The entire mitochondrial genome of rice has been sequenced (Notsu Y. et al., 2002); it was found to be 490*;*520 bp long. To date, 81 genes have been identified, 53 of which coding for proteins.

The automatic simulation of editing on all the potential start and stop codons of rice mtDNA leads to the generation of a total of 176 candidate ORF sequences, among which 138 are those involving edited start and stop codons, consisting of 60 sequences with editing only on the start codon and 78 sequences with editing only on the stop codons. The latter ones seem to be less interesting for our analysis, since

they represent subsequences of ORF sequences that can be generated also by other available ORF finder tools.

We thus focused on the former 60 candidate ORF sequences. Among them, we found 32 sequences corresponding to proteins already described in rice, 7 not known in *Oryza* but homologous to proteins identified in other organisms, and 21 sequences that have been not described before.

The screening of the DBEST database (Boguski M.S. et al., 1993) by TBLASTN (Altschul et al., 1997) gave very interesting results: six candidate ORF sequences from forward DNA strand (Table 1) and seven from reverse strand (Table 2) showed positive matches, indicating their transcription in the organism under study.

SEQUENCE NUMBER	START CODON	STOP CODON	HOMOLOGY WITH OTHER ORGANISMS	HOMOLOGUE PROTEINS	EST
1	54085	354460	None	-	*Oryza sativa*, *Tripsacum dactiloydes*, *Zea mays*, others
2	407800	408127	None	-	*Oryza sativa*, *Bambusa oldhamii*, *Tripsacum dactyloides*, *Zea*, *Triticum aestivum*, *Sorghum bicolor*
3	467635	467935	None	-	*Oryza sativa*, *Sorghum bicolor*, *Zea mays*
4	283844	284180	Some plants, many bacteria	**PG1**	*Oryza sativa*, *Zea mays*, *Cucumus sativus*, *several bacteria*
5	362648	362972	None	-	*Oryza sativa*, *Bambusa oldhamii*, *Zea mays*, *Triticum*, *Sorghum bicolor*, *Vitis vinifera*, others
6	364454	364874	*Zea mays, Trichoplax ashaerens*	*AAR*91184	*Oryza sativa*, *Zea mays*, *Zea mays*, *Bambusa oldhamii*, *Triticum*,*Sorghum bicolor*, *Vitis vinifera*,

Table 4.1. Candidate ORF sequences from forward DNA strand presenting transcription in rice.

Sequence Number	Start Codon	Stop Codon	Homology with other organisms	Homologue Proteins	EST
7	463284	463282	*Persephonella marina* Bacteria	*YP_002730925*	***Oryza sativa***, *Zea mays*, *Sorghum bicolor*, *Vitis vinifera*, *Triticum aestivum*, others
8	362648	231982	None	–	***Oryza sativa***, *Bambusa oldhamii*, *Zea mays*, *Triticum aestivum* *Camellia sinensis*, others
9	449032	449030	None	–	***Oryza sativa***, *Zea mays*, *Carica papaya*, *Tripsacum dactyloides*, *Ricinus comunis*, others
10	31415	314150	*Nicotiana tabacum* *Beta vulgaris* *Arabidopsis thaliana*, other	*YP_173435*	***Oryza sativa***, *Zea mays*,*Bambusa oldhamii* *Tripsacum dactyloides*, *Zea*, *Sorghum bicolor*, others
11	294787	294785	None	–	***Oryza sativa***, *Eichhornia crassipes*, *Bambusa oldhamii*, *Liriodendron tulipifera*, others
12	201127	201125	*Brassica napus*	*YP_717160*	***Oryza sativa***, *Zea mays*, *Tripsacum dactyloides*, *Bambusa oldhamii*, others
13	105517	105515	None	–	***Oryza sativa***, *Triticum aestivum*, *Petunia*, *Tripsacum dactyloides*, *Bambusa oldhami*, others

Table 4.2. Candidate ORF sequences from reverse DNA strand presenting transcription in rice.

Because transcription of an open reading frame indicates gene activity, we directed our further analysis on these 13 transcribed ORFs. The first column in the table contains progressive numbers we exploited to indicate the considered candidate ORF sequences, second and third columns show the position in the nucleotide sequence of the start and the stop codons of each sequence, respectively; fourth and fifth columns highlight possible homologies (proteins are indicated by their name or NCBI accession number). The last column shows organisms where the corresponding transcribed ORF has been found. Among these sequences, five (2 from forward and 3 from reverse strand) were homologous to proteins already known in other organisms, but eight sequences have never been described until now. The evidence of RNA transcription from these sequences, let us suppose that they may indeed represent new genes.

Among the candidate ORF showing homology with proteins already known in other organisms, four are returned by our system as hypothetical proteins. In particular, sequence 6 in Table 1, is homologous to a protein described in *Zea mays* (with NCBI accession number *AAR*91184), a monocotyledon plant, and in

Trichoplax adherens, a Placozoa. Sequence 7 in Table 2 shows homology with a protein described in *Persephonella marina* (*Y P* 002730925) and many bacteria, sequence 10 is homologous to a protein identified in *Nicotiana tabacum* (*Y P* 173435) and other plants, while sequence 12 is homologous to a protein described in *Brassica napus* (*Y P* 717160). DBEST screening showed that all of them are expressed not only in *Oryza sativa*, but in several vegetal organisms. Functional studies can clarify the nature of these proteins. Sequence 4 in Table 1 showed high homology with pG1 protein, a factor involved in transcription regulation, in several plants and many bacteria. The high homology with the same protein in organisms, even very distant from an evolutionary point of view, strongly indicate that our candidate ORF sequence of *Oryza* actually corresponds to the pG1 protein.

4.5.3 Some further analysis

In the previous section we presented the results obtained by applying our approach to predict proteins in rice, thereby showing the effectiveness of our technique. Encouraged by these results, we now discuss further developments we are currently dealing with to increase the number and the quality of the positive predictions (sequences not yet described in the organism under study) returned by our system. Indeed, we observe that many of the predicted ORF sequences in *S*ORF have been discarded, since they did not appear interesting candidates. However, this is true by considering such sequences *as they are* after ORF generation, with and without editing on the start and stop codons. On the other end, RNA editing might occur also on portions *inside* sequences in *S*ORF. In the following, we provide a sketchy illustration about how such a case can be handled.

A first question is to what extent possible RNA editings occurring in each sequence of *S*ORF may influence the prediction process (it is just worth recalling that the only editing we are focusing on here is *c*-> *u* one). Note that if we simulate editing on the sequences in *S*ORF, we should take into account all the possible *c*-> *u* editing configurations that might possibly occur, the number of which is 2^k, where *k* is the number of *c* occurrences in the ORF sequence under consideration. However, for the purposes of our analysis, two or more such configurations are to be considered equivalent as long as they produce the same amino acid. Note, by the way, that since more than one *c* can occur with one single triplet, that triplet can

indeed induce different amino acids via editing – this is the case, for instance, of the amino acid *P*, that corresponds to four triplets including *ccc* and from which, by editing, actually three amino acids, namely *L*, *S* and *F*, can be obtained. Therefore, a quantitative analysis is useful here. Thus, let *a'* be an amino acid containing a *c* such that a substitution *c*-> *u* leads to the generation of an amino acid *a"*≠ *a'*. We say that *a'* is an *editable* amino acid.

Analogously, we call *editable c* each *c* that may cause the generation of a new amino acid after a *c*-> *u* substitution. We exploit the term *editing substitutions* to refer to both *c*-> *u* substitutions and the corresponding *a'*->*a"* substitutions, accordingly to case under analysis (nucleotide sequences or amino acid sequences, respectively). **Figure 4.6** shows the distribution of the number of *c*, editable *c* and editable amino acids for unit of lenght, with respect to all the 176 amino acid sequences generated from rice using our method and discussed in the previous section.

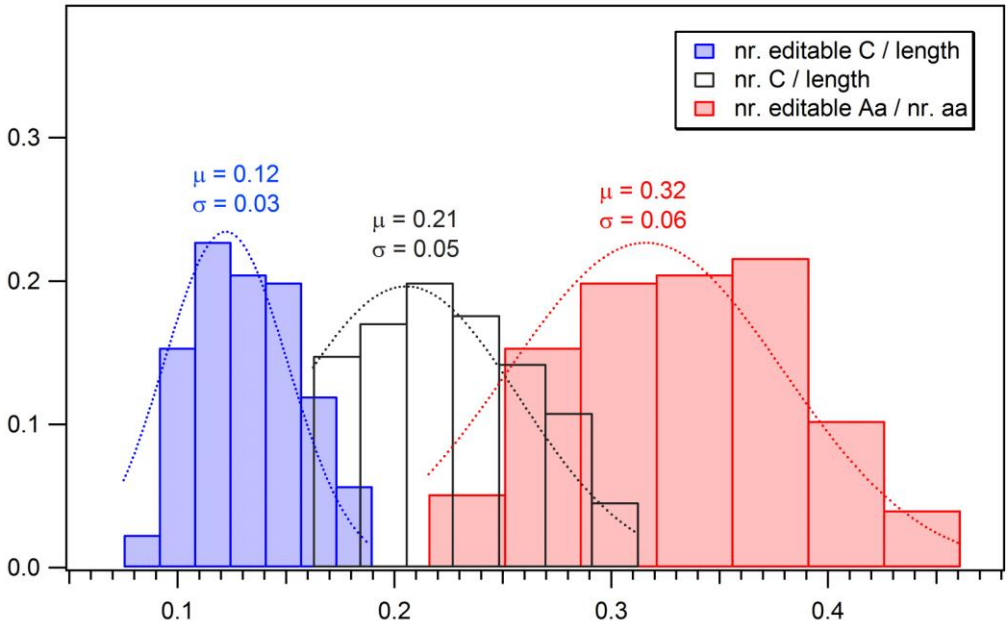


Fig. 4.6. Distribution of editable c and corresponding aminoacid in *S*ORF.

A Gaussian fit has been performed for each distribution: the ascissa corresponding to the peak of each curve fit has been found to agree with the corresponding calculated average value. Moreover, the expected confidence intervals

for normal distributions have been observed: about 64%, 66%, 67% of the set are within one standard deviation for fraction of *c*, editable *c* and editable amino acids, respectively. Two standard deviations from the mean account for about 98%, 97% and 95% of the set for each distribution, respectively.

Looking at Figure 4.4, we observe that the amino acid sequences are more sensible to editing substitutions than the original candidate ORF sequences from which they were obtained. Indeed, the curve fitting editable amino acids results to be translated along the *x*-axis approximatively by a factor of 3 with respect to the curve corresponding to editable *c*. We also observe that, in some cases, editing substitutions involve more than the 40% of an amino acid sequence. This means, first of all, that some of the proteins we discarded in our analysis for they were not transcribed, might be catched again after this further editing-simulation step. Furthermore, also candidate ORF sequences generated without including editing substitutions on the start and stop codons would become worth taking into account (the statistics on such sequences are analogous to those shown in Figure 4.4). This way the number of candidate new proteins would considerable increase. Unfortunately, in order to generate all the different amino acid sequences that can be obtained by this further editing-step, we should tackle the generation of an exponential number of possible configurations, to be then searched for possible homologies and/or transcribed sequences. In order to avoid such an exponential blow-up in the number of candidate ORF sequences to analyze, we plan to follow a strategy that we briefly summarize below.

Let s_i be the amino acid sequence of a candidate protein, obtained according to the procedure illustrated in figure 4.3. We aim at finding already known proteins which are most similar to s_i along all the possible editing substitutions. The idea is to design a novel similarity measure, let us call it *edt_sim*, able to directly take editing into accounts, in order to generate only edited sequences that are homologous to some already known proteins. Roughly speaking, given a candidate protein with amino acid sequence s_i and a known protein with amino acid sequence s_i, the similarity between s_i and s_j is equal to σ if there exists a set of editing substitutions transforming s_i into a sequence $\hat{s}_i$, such that the similarity between $\hat{s}_i$ and s_j is σ. More formally, let φ be an operator transforming an amino acid sequence into another one by applying a finite sequence of editing substitutions. The *edt_sim* between two

sequences s_i and s_j is $edt\_sim(s_i, s_j) = \sigma \Leftrightarrow \exists\, \hat{s}_i = \varphi(s_i) : sim(\hat{s}_i, sj) = \sigma$, where *sim* is one of the classical sequence similarity measures widely employed in literature (i.e., obtained by the BLAST score). It is worth pointing out that we conjecture that such new similarity measure does not require to explicitly compute all the possible editing $\varphi(s_i)$ of a given sequence s_i, but it can be obtained by modifying the substitution matrix (exploited by the algorithm to compute the similarity between two amino-acids) into a new matrix directly encoding also the allowed editing mechanisms.

5. CONCLUSIVE REMARKS

While the pre-genomic era was characterized by the effort to sequence whole genomes, the post-genomic era is addressed on harvesting the fruits hidden in the genomic texts. The major goal of researchers is now interpreting the genomic data that are being uncovered within the diverse genome sequencing projects. This field presents one of the grand challenges of our times.

Genomes appears much more complex than expected, and their informational content is far to be completely understood, although many efforts are making on finding the genes still hidden in genomes. Several software tools have been made in order to find genes inside genomic sequences, but, if computational methods are very useful in searching for genes with standard structure, they fail in identification of encrypted genes.

This thesis aims at giving a contribution to the question of gene discovery. In particular, two main problems are addressed: the first, a specific one, is the identification of specific genes; the other, more general, is the discover in plant mitochondrial genomes, of genes encoding proteins not yet identified and difficult to search for using standard search tools and methodologies. To deal with these problems, two new approaches to gene search have been proposed.

First, we defined a general, semi-automatic methodology to discover the possible presence of specific genes in cells and we applied it to plant genomes in order to search the *IP6K* gene. The intuition behind this work is that some genes or specific gene families, such as all *IPK* genes, are characterized by the presence of specific tags, short sequences of few amino acids, often corresponding to functional regions. The analysis we conducted in plant mitochondria provided the negative, though we argue relevant, result that *IP6K* does not actually occur in vegetable mtDNA. Very interestingly, the tag search in nuclear genomes lead us to identify a promising sequence in chromosome 5 of Oryza sativa. Further analyses are in course to confirm that this sequence actually corresponds to *IP6K* mammalian gene.

The second part of my work was directed to a more general question, that is the identification of new proteins in mitochondrial genomes of plants. It is known that

plant mitochondria use several mechanisms to increase genetic variability, because of which protein products can result very different from DNA generating them. Among these mechanisms RNA editing is the most common. To search new mitochondrial proteins we defined a multi-tool methodology, and we designed ad hoc one of this tools, a system for Open Reading Frame extraction from genomes, providing for editing possibility.

We applied the new approach on mitochondrial DNA of *Oryza sativa*, and we showed that the automatic simulation of editing on the start and stop codons of a nucleotide sequence leads to the generation of relevant candidate ORF sequences. In particular, not only candidate ORF sequences corresponding to proteins already known in the organism under analysis can be generated this way, but also some portions corresponding to protein sequences not yet discovered therein. The latter ones can be divided in *(i)* amino acid sequences presenting high homologies with some protein sequences known in other organisms, and *(ii)* amino acid sequences without any significant homology with other existing proteins. Pertaining to case *(ii)*, those sequences that have been transcribed are the most interesting ones, representing possible novel proteins. Both these and those pertaining to case *(i)* are considered positive predictions. Thirteen out of 60 candidate ORF sequences predicted by simulating editing on start codons, showed positive matches to DbEST. Thus, we directed our further analysis on these 13 transcribed ORFs. Among these sequences, five were homologous to proteins already known in other organisms, but eight sequences have never been described until now.

The evidence of RNA transcription from these sequences let us suppose that they may indeed represent new genes. Further studies can clarify the functional nature of these proteins.

This work might be further extended along several directions. First of all, as far as the *IP6K* gene search is concerned, experiments of molecular cloning and biochemical characterization are required to confirm that the sequence we found is actually the gene in question, and to determine substrate specificity of the enzyme.

Second, with regard to mitochondrial protein prediction, further analysis are required to investigate on the identity of new predicted proteins. Furthermore, as discussed above, it might be very interesting to extend the method we developed in order to increase the number and the quality of the positive predictions it returns.

Finally we observe that, often, proteins with low sequence homology have similar functions and similar secondary/tertiary structures, whereby it appears sensible to comparatively look at such structures for the result assessment purposes.

Acknowledgments

I wish to gratefully acknowledge my Ph.D. supervisor Luigi Palopoli for his teachings, his guidance, his suggestions and his fraternal understanding, especially in more difficult moments. I wish to extend my sincere gratitude to Adolfo Saiardi, for having provided me with ideas, suggestions and continuous support. I want to thank Simona Rombo for her encouragement, for her valuable comments and for the support that she offered me with competence and kindness. I also want to express my gratitude to Fabio Fassetti, for his help, for his availability and his stimulating discussions.

LIST OF FIGURES AND TABLES

LIST OF PUBLICATIONS

Fabio Fassetti, **Ofelia Leone** , Luigi Palopoli, Simona E. Rombo, Adolfo Saiardi. *IP6K gene identifcation in plant genomes by tag searching.* ***BMC Proceedings***, 5(Suppl 2), S1, ISSN 1753-6561, BioMed Central Ltd, London, 2011.

Fabio Fassetti, **Ofelia Leone**, Luigi Palopoli, Simona E. Rombo and Adolfo Saiardi. *IP6K gene identification by tag search.* 6th International Symposium on Bioinformatics Research and Applications (ISBRA 2010). May, 23-26 2010, University of Connecticut, Storrs, Connecticut, USA

Fabio Fassetti, **Ofelia Leone**, Luigi Palopoli, Simona E. Rombo, Adolfo Saiardi. *IP6K gene discovery in plant mtDNA.* Seventh international meeting on computational intelligence methods for bioinformatics and biostatics (CIBB 2010). Palermo, September 16-18 201, *Invited on BMC supplement*

Fabio Fassetti, Claudia Giallombardo, **Ofelia Leone**, Luigi Palopoli, Simona E. Rombo, Adolfo Saiardi. *Predicting Undiscovered Proteins in Plants by the Automatic Simulation of RNA Editing.* Submitted to 16° International Conference on Research in Computational Molecular Biology (RECOMB 2012)

REFERENCES

Alcázar-Román A.R., Wente S.R. (2008). Inositol polyphosphates: a new frontier for regulating gene expression. *Chromosoma*, 117:1–13.

Altschul S. F., Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402.

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG. (1981). Sequence and organization of the human mitochondrial genome. *Nature*. 290:457-65

Benne R., Van Der Burg J., Brakenhoff J.P.J., Sloof P., Van Boom. J.H., Tromp M.C. (1986). Major transcript of the frameshift coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell*, 46: 819-826..

Bennett M., Onnebo S., Azevedo C., Saiardi A. (2006). Inositol pyrophosphates: metabolism and signaling. *Cell Mol Life Sci*, 63:552-564.

Bhandari R., Saiardi A., Ahmadibeni Y., Snowman A.M., Resnick A.C., Kristiansen T.Z., Molina H., Pandey A., Werner J.K., Juluri K.R., Xu Y., Prestwich G.D., Parang K., Snyder S.H. (2007). Protein pyrophosphorylation by inositol pyrophosphates is a posttranslational event. *Proc Natl Acad Sci USA*, 104: 15305-15310.

Boguski M.B., Lowe T.M., and Tolstoshev C.M. (1993) dbEST–database for Expressed Sequence Tags. *Nat .Genet.*, pages 332–333.

Bonen L. (1991) The mitochondrial genome: so simple yet so complex. *Curr. Opin. Genet. Dev.*, 1: 515-522.

Brearley C., Hanke D. (1996). Inositol phosphates in barley (Hordeum vul. L.) aleurone tissue are stereochemically similar to the products of breakdown of InsP6 in vitro by wheat-bran phytase. *Bioch. J.*, 318: 279-286.

Brennicke, A. Marchfelder, and S. Binder.(1999). RNA editing. *FEMS Microbiol. Rev.*, 23:297–316.

Brennicke A., Picardi E., Quagliariello C., and Regina T.M.R. http://biologia.unical.it/pyscript/overview.html RNA Editing database.

Bundschuh R. (2004). Computational prediction of RNA editing sites. *Bioinformatics*, 20: 3214–3220.

Burger C. and Karlin S. (1997). Prediction of complete gene structure in human genomic DNA. *J Mol Biol.*, 268: 78-94

Burger G., Gray M.W., and Lang B.F. (2003). Mitochondrial genomes: anything goes. *TRENDS in Genetics*, 19:709–716.

Communi D., Takazawa K., Erneux C. (1993). Lys-197 and Asp-414 are critical residues for binding of ATP/Mg2+ by rat brain inositol 1,4,5- trisphosphate 3-kinase. *Biochem J* , 291:811-816.

Covello P.S., Gray M.W. (1990). Differences in editing at homologous sites inmessenger RNAs from angiosperm mitochondria. *Nucl Acids Res* 18: 5189–5196.

Clifton S.W., Minx P., Fauron C.M.R., Gibson M., Allen J.O., Sun H., Thompson M., Barbazuk W.B., Kanuganti S., Tayloe C. (2004). Sequence and Comparative Analysis of the Maize NB Mitochondrial Genome1[w]. *Plant Physiology*. 136: 3486–3503.

Di Pietro C., Di Pietro V., Emmanuele G., Ferro A.,Maugeri T., Modica E., et al. (2003). ANTICLUSTAL: Multiple Sequence Alignment by Antipole Clustering and Linear Approximate 1-Median Computation. *Proceeding of the IEEE Computer Society Conference on Bioinformatics.*

Eskin E., Pevzner P.A. (2002). Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, 18: S354-S363.

Fassetti F., Greco G., Terracina G.(2008). Mining loosely structured motifs from biological data. *IEEE Trans. Knowl. Data Eng.*, 20: 1472–1489.

Flores S., Smart C. (2000). Abscisic acid-induced changes in inositol metabolism in Spirodela polyrrhiza. *Planta*, 211: 823-832.

Goff S.A., Ricke D., Lan T.H., Presting G., Wang R., Dunn M., Glazebrook J., Sessions A., Oeller P., Varma H., Hadley D., Hutchison D., et al. (2002). A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica). *Science*, 296: 92-100.

Gonzales B., Banos-Sanz J.I., Villate M., Brearley C.A., Sanz-Aparicio J. (2010). Inositol 1,3,4,5,6-pentakisphosphate 2-kinase is a distant ipk member with a singular inosite binding site for axial 2-oh recognition. *Proc Natl Acad Sci U S* A, 107:9608–9613.

Gray M.W. (1982) Mitochondrial genome diversity and the evolution of mitochondrial DNA. *Can. J. Biochem.,* 60: 157-171.

Gray M.W, Hanic-Joyce P.J, Covello P.S. (1992). AND Transcription, Processing and editing in plant mitochondria. *Annu Rev Plant Physiol Mol Biol* , 43: 145–175.

Greene E. A, Henikoff S. (2000). Getting more from your sequence using the web. *Nature genetics*, 18:209-215.

Gualberto J.M., Lamattina L., Bonnard G., Weil J.H., and Grienenberger J.M. (1989.) RNA editing in wheat mitochondria results in the conservation of protein sequences. *Nature*, 341:660–662.

Gualberto J.M., Weil J. H., Grienenberger J.M. (1990). Editing of the wheat *coxIII* transcript: evidence for twelve C to U and one U to C conversions and for sequence similarities around editing sites. *Nucleic Acids Res*, 18: 3771-3776.

Hoch B., Maier R.M., Appel K., Igloi G.L., Kossel H. (1991). Editing of a chloroplast mRNA. *Nature*. 353:178-80.

Hodges P., Scott J. (1992). Apolipoprotein B mRNA editing: a new tier for the control of gene expression. *Trend Biochem. Scie*., 17: 77-81

Irvine R. (2003). 20 years of Ins(1,4,5)P3, and 40 years before. *Nat Rev Mol Cell Biol*, 4: 586-90.

Ives E.B., Nichols J., Wente S.R., York J.D. (2000). Biochemical and functional characterization of inositol 1,3,4,5,6-pentakiphosphate 2-kinases. *The Journal of Biological Chemistry*, 275: 36575–36583.

Josefsen L., Bohn L., Sorensen M.B., Rasmussen S.K. (2007). Characterization of a multifunctional inositol phosphate kinase from rice and barley belonging to the ATP-grasp superfamily. *Gene*, 397: 114-25.

Krogh A. (1997). Two methods for improving performance of an HMM and their application for gene finding. In Proc. of Fifth Int. Conf. on Intelligent Systems for Molecular Biology, ed. Gaasterland, T. et al., Menlo Park, CA: AAAI Press, 1997, pp. 179-186.

Labrador M., Mongelard F., Plata-Rengifo P., Baxter E.M., Corces V.G., Gerasimova T. (2001). Molecular biology: protein encoding by both DNA strands. *Nature*, 409: 1000-1013.

Larkin M, Blackshields G, Brown N: ClustalW and ClustalX version 2. Bioinf (2007), 23: 2947-2948.

Lemtiri-Chlieh F., MacRobbie E.A. and Brearley C.A. (2000). Inositol hexakisphosphate mobilizes an endomembrane store of calcium in guard cells. *Proc Natl Acad Sci U S A.* 97: 8687-92.

Loomis W., Smith D. (1995). Consensus phylogeny of Dictyostelium. *Experientia*, 51: 1110-1115.

Luo H., Huang Y., Chen J., Saiardi A., Iijima M., Ye K., Huang Y., Nagata E., Devreotes P., Snyder S.H. (2003). Inositol pyrophosphates mediate chemotaxis in Dictyostelium via pleckstrin homology domain PtdIns(3,4,5)P3 interactions. *Cell*, 114: 559-572.

Luo H., Huang Y., Chen J., Saiardi A., Iijima M., Ye K., Huang Y., Nagata E., Devreotes P., Snyder S.H. (2003). Inositol pyrophosphates mediate chemotaxis in Dictyostelium via pleckstrin homology domain PtdIns(3,4,5)P3 interactions. *Cell*, 114: 559-572.

Luo H., Saiardi A., Yu H., Nagata E., Ye Ki, Snyder S.H. (2002). Inositol pyrophosphates are required for DNA hyperrecombination in protein kinase c1 mutant yeast. *Biochemistry*, 41:2509-2515.

M.A. Larkin, G. Blackshields, and N.P. Brown. ClustalW and ClustalX version 2. *Bioinf.*, 23(21):2947–2948, 2007.

MIT. http://web.mit.edu/star/orf/. StarORF, Open Reading Frame Finder Tool.

Morrison B., Bauer J., Hu J., Grane R.W., Ozdemir A.M., Chawla-Sarkar M., Gong B., Almasan A., Kalvakolanu D.V., Lindner D.J.(2002). Inositol hexakisphosphate kinase 2 sensitizes ovarian carcinoma cells to multiple cancer therapeutics. *Oncogene*, 21: 1882-1889.

Mower J.P. (2005). PREP-Mt: predictive RNA editor for plant mitochondrial genes. *BMC Bioinformatics*, 6: 96-107.

Nagata E., Luo H., Saiardi A., Bae B., Suzuki N., Snyder S.H. (2005). Inositol hexakisphosphate kinase-2, a physiologic mediator of cell death. *J Biol Chem*, 280:1634-1640.

Nalaskowski M. M., Deschermeier C., Fanick W. and Mayr G. W. (2002) The human homologue of yeast ArgRIII protein is an inositol phosphate multikinase with predominantly nuclear localization. *Biochem. J.*, 366: 549–556.

NCBI. http://www.ncbi.nlm.nih.gov/projects/gorf/. ORF Finder

Notsu Y., Masood S, Nishikawa T, Kubo N, Akiduki G, Nakazono M, Hirai A. *et al*. The complete sequence of the rice (oryza sativa l.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. *Mol Genet Genomics*, 268(4):434–445, 2002.

Palmer J, Adams K, Cho Y, Parkinson CL, YL YLQ, Song K: Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proc Natl Acad Sci U S A 2000, 97(13):6960-6966.

Picardi E. and Quagliariello C. (2005). EdiPy: a resource to simulate the evolution of plant mitochondrial genes under the RNA editing. *Comput. Biol. Chem.*, 30:77–80.

Posternak S. (1919). Sur la synthese de l'ether hexaphosphorilique de l'inosite avec le principle phosphoorganique de reserve des plantes vertes. *Compt. Rend. Acad. Sci.*, 169: 138-140.

Raboy V. (2003). Myo-Inositol-1,2,3,4,5,6-hexakisphosphate. *Phytochemistry*, 64:1033-43.

Rice P., Longden I., Bleasby A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, 16: 276–277.

Regina T. M. R., Lopez L., Picardi E., Quagliariello C. (2002). Striking differences in RNA editing requirements to express the rps4 gene in magnolia and sunflower mitochondria. *Gene*, 286: 33–41.

Saiardi A., Bhandari R., Resnick A.C., Snowman A.M., Snyder S.H. (2004). Phosphorylation of proteins by inositol pyrophosphates. *Science*, 306: 2101-2105.

Saiardi A, Caffrey J, Snyder S, Shears S: The inositol hexakisphosphate kinase family. Catalytic flexibility and function in yeast vacuole biogenesis. (2000). *J Biol Chem*, 275: 24686-24692..

Saiardi A., Erdjument-Bromage H., Snowman A. M., Tempst P. and Snyder S. H. (1999) Synthesis of diphosphoinositol pentakisphosphate by a newly identified family of higher inositol polyphosphate kinases. *Curr. Biol*., 9: 1323–1326.

Saiardi A., Nagata E., Luo H. R., Sawa A., Luo X., Snowman A. M. et al. (2001) Mammalian inositol polyphosphate multikinase synthesizes inositol 1,4,5-trisphosphate and an inositol pyrophosphate. *Proc. Natl. Acad. Sci. USA*, 98: 2306–2311.

Saiardi A., Nagata E., Luo H. R., Snowman A. M. and Snyder S. H. (2001) Identification and characterization of a novel inositol hexakisphosphate kinase. *J. Biol. Chem.* 276: 39179– 39185.

Saiardi A., Resnick A., Snowman A. (2005). Inositol pyrophosphates regulate cell death and telomere length through phosphoinositide 3-kinase-related protein kinases. *Proc Natl Acad Sci U S A*, 102: 1911-1914.

Saiardi A., Sciambi C., Mc Caffery J. (2002). Inositol pyrophosphates regulate endocytic trafficking. *Proc Natl Acad Sci U S A*, 99: 14206-14211.

Saiardi A: Unpublished manuscript.

Saiardi A., Azavedo C. Unpublished manuscript

Salamov A., Solovyev V. (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome Res*., 10: 516-522.

Schell M. J., Letcher A. J., Brearley C. A., Biber J., Murer H. and Irvine R. F. (1999) PiUS (Pi uptake stimulator) is an inositol hexakisphosphate kinase. *FEBS Lett*, 461: 169–172.

Schuster W., Unseld M., Wissinger B, and Brennicke A.(1990). Ribosomal protein S14 transcripts transcripts are edited in Oenothera mitochondria *Nucleic Acids Res*., 18: 229-33.

Shears S.B. (2001). Assessing the omnipotence of inositol hexakisphosphate. *Cell Signal*, 13:151-158.

Shears S.B. (2004). How versatile are inositol phosphate kinases? *Biochem J* ,377: 265-280.

Snyder E.E., Stormo G.D. (1993). Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Res*, 21: 607-13.

Snyder E.E., Stormo G.D. (1995). Identification of protein coding regions in genomic DNA. *J Mol Biol*, 248: 1-18.

Sommer B., Kohler M., Sprengel R., Seeburg P.H. (1991). RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell*, 67: 11-19

Stevenson-Paulik J., Odom A., York J. (2002). Molecular and biochemical characterization of two plant inositol polyphosphate 6-/3-/5-kinases. *J Biol Chem* 277: 42711-42718.

Stevenson-Paulik J., Bastidas R.J., Chiou S.T., Frye R.A. and York J.D. (2005) Generation of phytate-free seeds in Arabidopsis through disruption of inositol polyphosphate kinases**.** *Proc Natl Acad Sci U S A.* 102: 12612-7

Strahl T., Thorner J. (2007) Synthesis and function of membrane phosphoinositides in budding yeast, Saccharomyces cerevisiae. *Biochem Biophys Acta* 1771: 353–404.

Stuart K. (1991). RNA editing in mitochondrial mRNA from trypanosomatids. *Trend Biochem Sci.*, 16: 68-72.

Sweetman D., Johnson S., Caddick S. E., Hanke D. E., Brearley C. A. (2006). Characteryzation of an arabidopsis inositol 1,3,4,5,6-pentakisphosphate 2-kinase (atipk1). *Biochem J*, 394:95–103, 2006.

Takenaka M., Verbitskiya D., van der Merwea J.A., Zehrmanna A., Brennicke A. (2008). The process of RNA editing in plant mitochondria. *Mitochondrion*, 8: 35-46.

Togashi S, Takazawa K, Endo T, Erneux C, Onaya T: Structural identification of the myo-inositol 1,4,5-trisphosphate-binding domain in rat brain inositol 1,4,5-trisphosphate 3-kinase. Biochem J 1997, 326:221-225.

Venter J.C., Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.* (2001). The Sequence of the Human Genome *Science*, 291**:** 1304-1351.

Verbsky JW, Wilson MP, Kisseleva MV, Majerus PW, and Wenter SR. (2002). The synthesis of inositol hexakiphosphate. characterization of human inositol 1,3,4,5,6-pentakiphosphate 2-kinase. *The Journal of Biological Chemistry*, 277:31857–31862.

Voglmaier S. M., Bembenek M. E., Kaplin A. I., Dorman G., Olszewski J. D., Prestwich G. D. et al. (1996) Purified inositol hexakisphosphate kinase is an ATP synthase: diphosphoinositol pentakisphosphate as a high-energy phosphate donor. *Proc. Natl. Acad. Sci. USA,* 93: 4305–4310.

Ward B.L., Anderson R.S., Bendich A.J. (1981) The mitochondrial genome is large and variable in a family of plants (cucurbitaceae). *Cell*. 25: 793-803

Wintz H., Hanson M.R. (1990). A termination codon is created by RNA editing in the petunia. *Curr Genet*, 268: 252–256.

Wissinger B, Brennicke A, Schuster W. (1992). Regenerating good sense: RNA editing and *trans*-splicing in plant mitochondria. *Trends Genet* , 8: 322–328.

Xia H., Brearley C., Elge S., Kaplan B., Fromm H., Mueller-Roeber B. (2003). Arabidopsis inositol polyphosphate 6-/3-kinase is a nuclear protein that complements a yeast mutant lacking a functional ArgR-Mcm1 transcription complex. *Plant Cell*, 15: 449-463.

Xu J., Brearley C.A., Lin W.H., Wang Y., Ye R., Mueller-Roeber B., Xu Z.H., Xue H.W. (2005). A role of Arabidopsis inositol polyphosphate kinase, AtIPK2alpha, in pollen germination and root growth. *Plant Physiol*., 137: 94-103.

York J.D., Odom A.R., Murphy R., Ives E.B., and Wente S.R. (1999). A phospholipase cdependent inositol polyphosphate kinase pathway required for efficient messanger rna export. *Science*, 285:96–100.

Zhang T., Caffrey J. J. and Shears S. B. (2001) The transcriptional regulator, Arg82, is a hybrid kinase with both monophosphoinositol and diphosphoinositol polyphosphate synthase activity. *FEBS Lett.,* 494: 208–212