

CHALMERS | GÖTEBORG UNIVERSITY

A project report submitted for the award of
MSc in Bionformatics

**Comparative Genomic Study of upstream
Open Reading Frames**

Marija Cvijovic

Supervisor: Per Sunnerhagen

Examiner: Olle Nerman

Table of Contents

<i>Preface</i>	5
<i>Acknowledgements</i>	6
<i>Abstract</i>	7
1 INTRODUCTION	8
1.1 MOTIVATION	8
1.2 OUTLINE.....	8
2 BACKGROUND	9
2.1 CELL	9
2.2 DNA	10
2.2.1 <i>Genes</i>	11
2.2.2 <i>DNA replication</i>	12
2.3 FROM DNA TO PROTEIN	13
2.3.1 <i>From DNA to RNA (transcription)</i>	13
2.3.2 <i>From RNA to protein (translation)</i>	14
2.4 GENE EXPRESSION	15
2.4.1 <i>Posttranscriptional control</i>	15
2.5 UNTRANSLATED REGIONS OF MRNA	18
2.5.1 <i>Structural characteristics of untranslated regions</i>	18
2.5.2 <i>Upstream Open Reading Frames</i>	19
2.5.3 <i>Different modes of action of uORFs</i>	19
3 DATA ACQUISITION AND CATALOGUE	20
3.1 SACCHAROMYCES CEREVISIE AS MODEL ORGANISM	20
3.2 SACCHAROMYCES GENOME DATABASE (SGD).....	21
3.3 INFORMAX	21
3.4 MAKING THE CATALOGUE	22
4 STUDY	24
4.1 DISTRIBUTION OF UORFs	25
4.2 MECHANISM OF UORFs	25
4.2.1 <i>GCN4</i>	25
4.2.2 <i>YAP1 and YAP2</i>	26
4.3 GENOME STRUCTURE AND EVOLUTION.....	27
4.4 GENOME ALIGNMENT AND SYNTENY.....	28
4.4.1 <i>Synteny</i>	28
4.4.2 <i>Alignment</i>	28
4.5 EXTENDED ANALYSIS	30
5 RESULTS	32
6 CONCLUSIONS	34
6.1 FUTURE WORK	34
REFERENCES	35
APPENDIX A	37

List of Figures

FIGURE 2-1 – DNA STRUCTURE.....	10
FIGURE 2-2 – GENE STRUCTURE	11
FIGURE 2-3 - MRNA STRUCTURE	18
FIGURE 3-1 - MAIN PAGE OF THE CATALOGUE	22
FIGURE 3-2 - CPA1 GENE PAGE.....	23
FIGURE 4-1 - DIFFERENT TYPES OF UORF DISTRIBUTION (NOT DRAWN TO SCALE)	25
FIGURE 4-2 - EVOLUTIONARY TREE.....	27
FIGURE 4-3 - SYNTENY VIEWER FOR CLN3 GENE (NEIGHBOURING GENES ARE ALSO REPRESENTED).....	28
FIGURE 4-4 – INFORMAX SCREENSHOT	29
FIGURE 4-5 - SCHEMATIC REPRESENTATION OF GCN4 UORFs IN DIFFERENT ORGANISMS	31

List of Tables

TABLE 4-1 - LIST OF STUDIED GENES.....	21
TABLE 5-1 – RESULTS.....	30

Preface

This report is Master Thesis in Bioinformatics and it is a part of International Masters Program at Chalmers and Göteborg University.

The research project has been carried out in Lundberg Laboratory, Göteborg within the department of Cell and Molecular Biology.

Examiner:

Olle Nerman, Department of Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden

Supervisor:

Per Sunnerhagen, Department of Cell and Molecular Biology, Göteborg University, Göteborg, Sweden

Acknowledgements

		1		12
1	(1)	TACKSAMYCK	E	T
human	(1)	-----P	E	R
Consensus	(1)		E	

I would like to thank to my supervisor Per Sunnerhagen for invaluable help and patience for a mathematician that came into biology playground.

Also, many thanks to my examiner Olle Nerman for giving me all those ideas and suggestions.

Special thanks goes to all people gathered around Bioinformatics programme at Chalmers and Göteborg University for enjoyable and unforgettable time I had in Sweden.

Marija Cvijovic

December 2004, Göteborg

Abstract

The untranslated regions of mRNA molecules are involved in several post-transcriptional regulatory pathways. The 5'UTR is the sequence between the 5' terminal cap structure and the initiation codon for protein synthesis. The 5' end (the leader) can accurately regulate the amount of protein synthesised from a specific mRNA. In a fraction of mRNAs, upstream open reading frames (uORFs) are present. A general method for detecting uORFs with a regulatory role is of great importance. The ribosome can recognise the AUG of a uORF as an initiation codon, translate the downstream sequence into protein and terminate before the main ORF is translated.

Although the complete genome of the yeast *Saccharomyces cerevisiae* has been sequenced, the number of mRNAs containing uORFs is not known. It is predicted that in *S.cerevisiae* 200 genes (3 %) have uORFs. The facts that mRNA start sites are not known, and that some genes have more than one promoter, constitute major problems in the identification of real uORFs.

In this study, a collection of 18 genes with known uORFs was examined. It has been observed that some genes have a very long leader (*CLN3* – 864nt), while some are extremely short (*DCD1* – 33nt). Also, a longer UTR doesn't necessarily correlate to a longer uORF or higher uORF frequency (*CLN3* – UTR 864nt, one uORF – 4 codons long).

A comparative analysis has been done for the yeast *Saccharomyces cerevisiae* and six related species (*S. paradoxus*, *S. mikatae*, *S. bayanus*, *S. castellii*, *S. kluyveri*, *S. kudriavzevii*). We aligned sequences (1000 nt upstream of the initiation codon) using a pairwise sequence alignment method, and then labelled all identified uORFs. It has been observed that uORFs in some genes, where a functional role for the uORFs had been established by experiment, show a striking conservation. Upstream ORFs from other genes, where no indication of such regulation was available, displayed a lower degree of conservation.

1 Introduction

1.1 Motivation

Biology is one of the most rapidly expanding and diverse areas in the sciences. The problems encountered in biology are frequently complex and often not totally understood. Mathematical models provide means to better understand the processes and unravel some of the complexities.

Identification of elements responsible for posttranscriptional control represents one of the biggest problems in modern biology.

Real challenge is to find those conserved motifs and characterize them in order to make a computational tool, which will be able to scan entire genome and distinguish elements that have crucial role in translational level.

1.2 Outline

Chapter 2 gives overview of some basic concepts of molecular biology that are important to understand the main idea of this project.

Chapter 3 describes type of data we used in this study and how catalogue that contains all, so far, identified genes with uORFs was made.

Chapter 4 brings description of methods used in this study and explains function of uORFs.

Chapter 5 is dedicated to results we achieved and how they can be used in order to fully solve addressed problems.

Chapter 6 is discussion of work that has been done and potential continuation of this project.

2 Background

2.1 Cell

Cell is the structural and functional unit of all living organisms. Each cell stores its own set of instructions for carrying out all specialized functions.

The six kingdoms of living things are divided into two major groups, Prokaryotes and Eukaryotes. There are two prokaryote kingdoms and four eukaryote kingdoms.

Kingdom	Evolved	Structure
Prokaryotes:		
Bacteria	3 to 4 billion years ago	Unicellular
Archaea	3 to 4 billion years ago	Unicellular
Eukaryotes:		
Protista	1.5 billion years ago	Unicellular Sometimes
Fungi	1 billion years ago	Unicellular or Multicellular
Animalia	700 million years ago	Multicellular
Plantae	500 million years ago	Multicellular

Prokaryotes

The simplest of the cells and the first type of cells that evolve were prokaryotic cells. They are organisms without nucleus, mitochondria or any other membrane bound organelles. Bacteria are the best known and most studied form of prokaryotic organisms. Prokaryotes are unicellular organisms that do not develop or differentiate into multicellular forms. They are capable of inhabiting almost every place on Earth and every surface of our body.

Eukaryotes

The major and most significant difference between prokaryotes and eukaryotes is the eukaryotic cells contain nucleus and membrane-bounded compartments in which specific metabolic activities take place. Nucleus contains eukaryotic cell's DNA. This kingdom also has organelles – small structures within cell that perform certain functions.

This thesis will deal with one form of Eukaryotes – Fungi. According to this, all described processes in this work refer to the eukaryotic cells.

2.2 DNA

The complete set of instructions for making an organism is called the genome. It consists of tightly coiled threads of *deoxyribonucleic acid* (DNA) and associated protein molecules, organized into structures called chromosomes. For each organism, the components of these threads encode all the information necessary for building and maintaining life, from simple bacteria to remarkably complex human beings. A DNA molecule consists of two strands that wrap around each other to resemble a twisted ladder whose sides are made of sugar and phosphate molecules and connected with nitrogen-containing chemicals called bases. Each strand is a linear arrangement of repeating similar units called nucleotides, which are each composed of one sugar, one phosphate, and a nitrogenous base. Four different bases are present in DNA: adenine (A), thymine (T), cytosine (C), and guanine (G). The particular order of the bases arranged along the sugar-phosphate backbone is called the DNA sequence; the sequence specifies the exact genetic instructions required to create a particular organism with its own unique traits. Special pairing exists between bases. Adenine pairs with thymine, while cytosine pairs with guanine. Weak bonds between the bases hold the two DNA strands together.

DNA is a directional molecule. It is always read and synthesized in the 5' to 3' direction (named after the 5' and 3' carbons in the carbon-ring of the sugar). Given this directionality of either strand, we can refer to sequences *upstream* (5') or *downstream* (3') of a particular nucleotide on the same strand. The two complementary strands run in opposite direction and are called anti-parallel; hence upstream in one strand is complementary to downstream on the opposite strand.

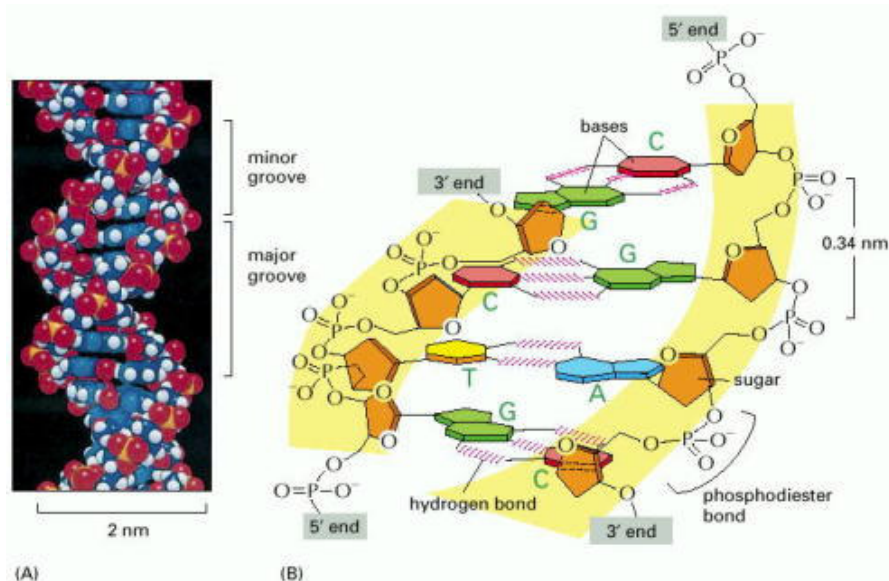


Figure 2-1 – DNA Structure

The DNA double helix. (A) A space-filling model of 1.5 turns of the DNA double helix. Each turn of DNA is made up of 10.4 nucleotide pairs and the center-to-center distance between adjacent nucleotide pairs is 3.4 nm. The coiling of the two strands around each other creates two grooves in the double helix. As indicated in the figure, the wider groove is called

the major groove, and the smaller the minor groove. (B) A short section of the double helix viewed from its side, showing four base pairs. The nucleotides are linked together covalently by phosphodiester bonds through the 3'-hydroxyl (-OH) group of one sugar and the 5'-phosphate (P) of the next. Thus, each polynucleotide strand has a chemical polarity; that is, its two ends are chemically different. The 3' end carries an unlinked -OH group attached to the 3' position on the sugar ring; the 5' end carries a free phosphate group attached to the 5' position on the sugar ring.

© 2002 by Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter.

2.2.1 Genes

A gene is a segment of a DNA, located in a particular position on a specific chromosome, whose base sequence contains the information necessary for protein synthesis.

Each DNA molecule contains many genes - the basic physical and functional units of heredity. A gene is a specific sequence of nucleotide bases, whose sequences carry the information required for constructing proteins, which provide the structural components of cells and tissues as well as enzymes for essential biochemical reactions.

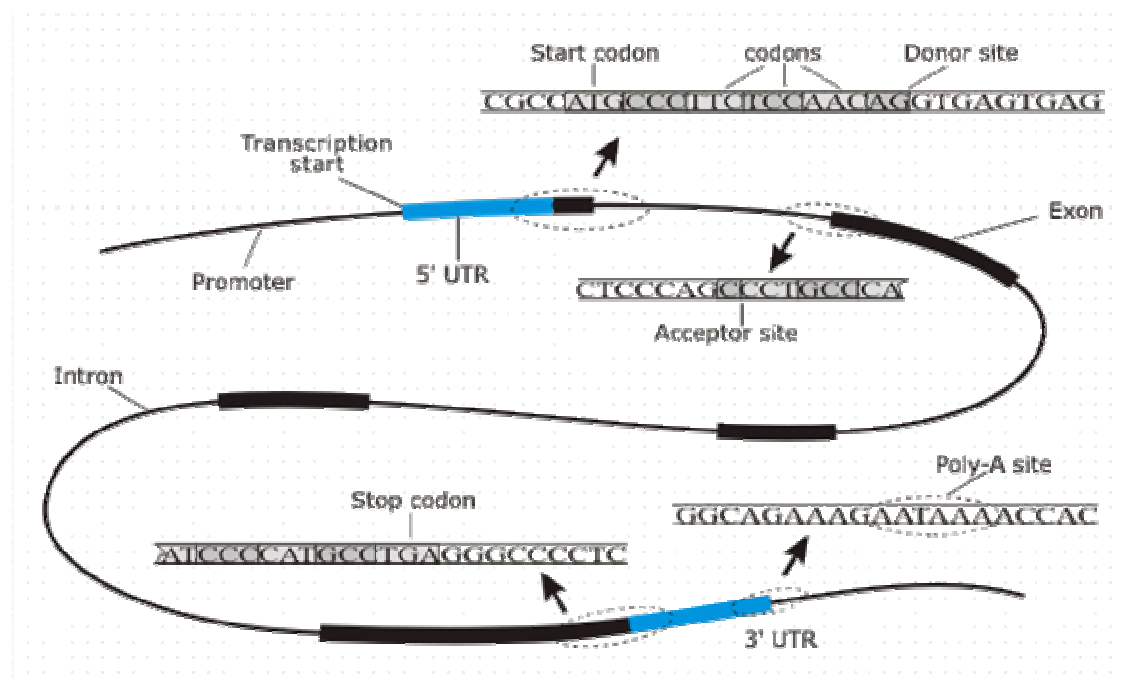


Figure 2-2 – Gene Structure

At the beginning of every gene we have promoter region, which is followed by a transcribed but non-coding region called *5' untranslated region* (5' UTR). Then the initial exon follows, which contains the start codon. Following the initial exon, there is an alternating series of introns and internal exons, followed by the terminating exon, which contains the stop codon.

It is followed by another non-coding region called the *3' untranslated region* (3'UTR). At the end of the eukaryotic gene, there is a polyadenylation (polyA) signal: the nucleotide Adenine repeating several times. The exon-intron boundaries (i.e., the splice sites) are signaled by specific short (2bp long) sequences. The 5'(3') end of an intron (exon) is called the *donor* site, and the 3'(5') end of an intron (exon) is called the *acceptor* site.

2.2.2 DNA replication

Each time a cell divides into two daughter cells, it duplicated its full genome. During cell division the DNA molecule unwinds and the weak bonds between the base pairs break, allowing the strands to separate. Position where DNA first opens is called replication origin. In yeast it is 100bp long, composed of DNA that attract the initiator proteins and stretches of DNA that is easy to open. A-T base pairing is usually found in replication origin because it is hold together with 2 hydrogen bonds (G-C pairing has 3 hydrogen bonds). Each strand directs the synthesis of a complementary new strand, with free nucleotides matching up with their complementary bases on each of the separated strands. Each daughter cell receives one old and one new DNA strand. The cells adherence to the base- pairing rules ensures that the new strand is an exact copy of the old one. This mechanism, know as proofreading minimizes the frequency of errors (mutations) that may greatly affect the resulting organism or its offspring.

2.3 From DNA to Protein

2.3.1 From DNA to RNA (transcription)

Ribonucleic Acids (RNA) consist of: ribose (a pentose - sugar with 5 carbons), phosphoric acid and nitrogen bases: purines (Adenine and Guanine) and pyrimidines (Cytosine and Uracil). Allowed base pairing is A-U and C-G. An RNA molecule is a single stranded, linear polymer in which the monomers (nucleotides) are linked together with phosphodiester bonds. These bonds link the 3' carbon in the ribose of one nucleotide to the 5' carbon in the ribose of the adjacent nucleotide. RNA molecule can fold in many different ways.

Several types of RNA exist.

- *messenger*RNA or mRNA - carries the genetic information out of the nucleus for protein synthesis
- *transfer*RNA or tRNA - decodes the information
- *ribosomal*RNA or rRNA - constitutes 50% of a ribosome, which is a molecular assembly involved in protein synthesis

An enzyme, RNA polymerase carries out transcription control. It moves along DNA molecule, unwinding it just one step ahead so one molecule at a time in 5' to 3' direction can be added. Role of RNA polymerase is to catalyze linkage of ribonucleotides. It doesn't have proofreading, because transcription doesn't have to be accurate as replication, since this RNA is just temporary.

In order to RNA becomes mRNA 2 steps are needed.

- RNA capping - at the 5' end, 7- methyl-guanosine cap (m7G) is attached
- Polyadenylation – at the 3' end, series of repeated Adenine nucleotides is attached, forming Poly(A)tail, few hundred nucleotides long

These two steps increase stability of mRNA.

First the whole length of gene is transcribed into a long RNA (primary transcript). After capping and polyadenylation, but before leaving the nucleus, introns are cut off and exons are joined together (RNA splicing). Now, RNA becomes functional mRNA ready to leave nucleus.

Primary transcript of many genes can be spliced in various ways to produce different mRNA, depending on the cell type in which the gene is being expressed. This allows different proteins to be produced from same gene.

2.3.2 From RNA to protein (translation)

All living organisms are composed largely of proteins - large, complex molecules made up of long chains of subunits called amino acids. Twenty different kinds of amino acids are usually found in proteins. Within the gene, each specific sequence of three DNA bases (codons) directs the cells protein- synthesizing machinery to add specific amino acids. The genetic code is thus a series of codons that specify which amino acids are required to make up specific proteins.

When mRNA is in cytoplasm it is ready to be translated to protein.

Ribosome plays major role in this process. It is a protein manufacture machine, made of 50 different proteins and rRNAs that catalyses translation of mRNA into amino acid sequence. Ribosome consists of 2 non-identical subunits: large (80S) and small (40S). The large subunit helps in fetching the amino acid molecules from tRNA and adding it to a polypeptide chain. The small subunit matches the tRNA to the codons of the mRNA. Each ribosome has 3 binding sites of tRNA: A-site, P-site and E-site. At one moment only 2 can be occupied.

Translation starts when tRNA carrying Methionine reach AUG codon and it is carried on in the 3 following steps:

- tRNA molecules binds to the A-site on ribosome
- new peptide bond is formed and small subunits shifts its position and moves the empty tRNA to E-site
- empty tRNA is ejected and small subunit moves back to original position

Translation ends when any of 3 stop codons UAA,UAG,UGA is reached. This process is regulated by protein called release factor. After reaching stop codon ribosome falls-a-part (large and small subunit separate).

Proteins are synthesized from the N terminus (encoded by the 5' part of the gene) to the C terminus (encoded by the 3' part of the gene).

Different combinations of four nucleotides can form $4^3 = 64$ triples, but there are only 20 amino acids to code for. As a result, more than one triplet encodes majority of amino acids. This redundancy results in the fact, that even though the mutation may occur, it doesn't necessary lead to the production of different protein. The position of the third base is particularly insensitive to such changes. It allows mutation to occur without an alteration of the amino acid product.

2.4 Gene Expression

Cells can control any of the processes that have been presented so far, transcription and translation, as well as other steps in protein synthesis. The genes it transcribes determine a cell's identity and fate. To ensure that gene expression is properly regulated, eukaryotic cells utilize a remarkable amount of protein machinery. Many genes important for cell growth and development are transcriptionally repressed by their packaging into DNA/protein structures called chromatin. In order to activate these genes chromatin is remodeled by the concerted action of sequence-specific DNA-binding transcriptional activator proteins and chromatin remodeling factors. These transcriptional activators also facilitate transcription by recruiting one or more of the 'basal' factors necessary for gene transcription. The production of an RNA transcript is also regulated later stages such as transcript elongation, termination, processing, stability and transport.

2.4.1 Posttranscriptional control

Posttranscriptional control is activated after RNA polymerase has bound to the gene's promoter and begun RNA synthesis. Here a variety of different processes are given.

1. attenuation of the RNA transcript by its premature termination
2. alternative RNA splice –site selection
3. control of 3'end formation by cleavage and Poly(A) addition
4. control of transport from nucleus to the cytosol
5. localization of mRNAs to particular parts of the cell
6. RNA editing
7. control of translational initiation
8. regulated mRNA degradation
9. translational recording

Most of these control processes require the recognition of specific sequence or structure in the RNA molecule being regulated. They can be recognized by either a regulatory protein or by a regulatory RNA molecule.

1° attenuation of the RNA transcript by its premature termination

Cell can control the degree of attenuation for particular genes. Proteins that are joint at the promoter can determine whether or not the polymerase will be able to pass through specific site of attenuation downstream.

2° alternative RNA splice –site selection

Genes are first transcribed as long mRNA, and then they are shorten to produce mature mRNA. One-way of doing this is RNA splicing or simpler removal of introns.

Alternative splicing is process that aloud the cell to splice in many different ways, making different polypeptide chains from the same gene.

Alternative RNA splicing can occurs because there is an intron sequence ambiguity: the standard spliceosome mechanism for removing intron sequences is unable to distinguish clearly between two or more alternative pairings of 5' and 3' splice sites, so that different choices are made by chance on different transcripts.

Regulated splicing is used to switch from the production of a non-functional protein to the production of functional ones.

Splicing can be regulated in 2 ways:

- negatively – regulatory molecule that prevents the splicing machinery from accessing particular splice site on the RNA
- positively – regulatory molecule that directs the splicing machinery to some other splicing site

3° 3' end cleavage and Poly(A) tail

3' end is determined by an RNA cleavage reaction. A cell can change the carboxyl terminus of the resultant protein (encoded by 3' end of mRNA)

4° RNA transport from nucleus to the cytosol

Primary RNA transcript is 10 times larger then the mature mRNA molecule.

Approximately 1/20 of the total mass of RNA leaves nucleus.

mRNA is ready to leave nucleus when cap at 5' end and poly(A) tail at 3' end are added. This will prevent introns of entering the cytosol.

5° localization of mRNAs

3'UTR contains signals responsible for directing mRNA to intracellular locations.

6° RNA editing

One way of RNA editing is when one or more Uracil nucleotides are inserted to selected region, changing the meaning of the message.

For some genes the editing is so extensive that over half of nucleotides in the mature mRNA are U nucleotides that were inserted during the editing process.

7° control of translational initiation

Another type of control found in eukaryotes uses one or more short open reading frames that lie between the 5' end of the mRNA and the beginning of the gene. Often, the amino acid sequences coded by these upstream open reading frames (uORFs) are not critical; rather the uORFs serve a purely regulatory function. A uORF present on an mRNA molecule will generally decrease translation of the downstream gene by trapping a scanning ribosome initiation complex and causing the ribosome to translate the uORF and dissociate from the mRNA before it reaches the protein coding sequences.

Another way that cells can initiate translation at positions distant from the 5' end of the mRNA. In these cases, translation is initiated directly at specialized RNA sequences, each of

which is called an internal ribosome entry site (IRES). IRES can occur in many different places in an mRNA. In some unusual cases, two distinct protein-coding sequences are carried in tandem on the same eucaryotic mRNA; translation of the first occurs by the usual scanning mechanism and translation of the second through IRES. IRESs are typically several hundred nucleotides in length and fold into specific structures that bind many, but not all, of the same proteins that are used to initiate normal cap-dependent translation. In fact, different IRESs require different subsets of initiation factors.

8° mRNA degradation

Histones are small, positively charged proteins that help binding of negatively charged sugar-phosphate back bone of DNA. They have a half-life of approximately 1h during DNA synthesis in S phase of cell cycle, but they become unstable and more degraded within a minute when synthesis stops.

The regulation of histone mRNA stability depends on short 3' stem-and-loop structure that replace the poly(A) tail. Special cleavage reaction creates this 3' end after the histone mRNA is synthesised by RNA polymerase II. Degradation rate for different types of mRNAs is strongly influenced by signals near 3' end (it is thought that degradation starts here).

9° translational recording

One form of translational recording is translational frameshifting – used by retro viruses, allowing more than one protein to be synthesized from a single mRNA.

2.5 Untranslated regions of mRNA

Untranslated regions (UTRs) are known to play crucial role in posttranscriptional regulation of gene expression, including modulation of the transport of mRNAs out of the nucleus and of translation efficiency, subcellular localization and stability.

5' end (the leader) is relatively short (fungi average length is 134bp), with the average length roughly the same among different organisms. The 5'UTR is the sequence between the 5' terminal cap structure and the initiation codon for protein synthesis. 3 elements are found in this region that have influence on translational control: hairpin loop, upstream open reading frames (uORFs) and internal ribosome entry sites (IRES). The leader can accurately regulate the amount of protein synthesised from a specific mRNA.

Length of 3' end (the trailer) can vary among different organisms. It is usually much longer than the leader (fungi average length is 237bp). The trailer connects stop codon (excluded) and the poly(A) tail. Different elements located on this part of the gene are responsible for subcellular localization and stability.

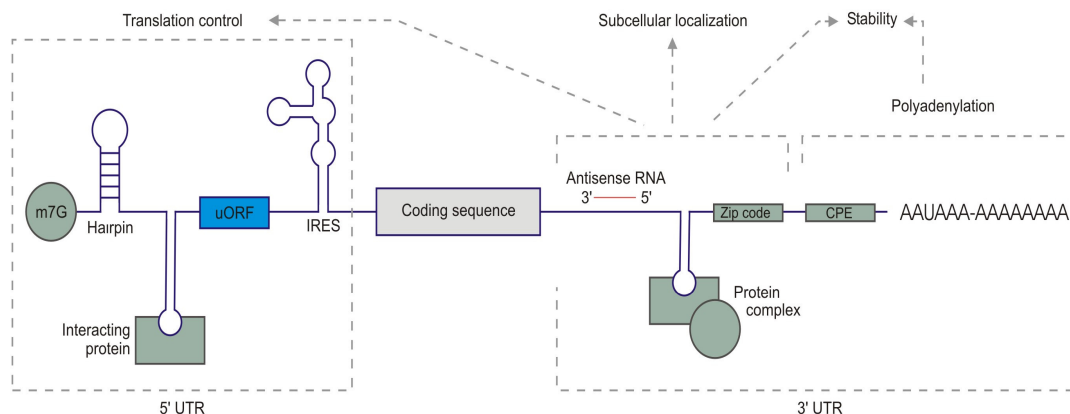


Figure 2-3 - mRNA structure

2.5.1 Structural characteristics of untranslated regions

The G+C content of 5'UTR sequence is greater than in 3'UTR sequence. Also, it has been shown [Pesole and al] that inverse correlation between G+C content and length of UTRs exist. However, this was expected due to the fact that GC-rich genome regions have much higher gene density than GC-poor genome regions. According to this, genes in GC-rich isochores (large regions of the genome that contain local similarities in base composition) tend to be shorter than genes in GC-poor isochores. Knowing these facts we can conclude that UTRs tend to be longer in GC-poor than in GC-rich isochores.

5'UTR has higher frequency of introns in the gene region, than 3'UTR. This implies that exons at the 5' end of a gene are much shorter than exons at the 3' end.

2.5.2 Upstream Open Reading Frames

In fungi 30% of 5'UTRs contain upstream AUG, suggesting that the 'first AUG rule' predicted by scanning model of ribosome start-site selection is disobeyed in a large number of cases. This implies that the 40S ribosomal subunit can sometime bypass the most upstream AUG codon, possibly because its sequence context makes it a poor initiation codon, to initiate translation at a more distal AUG. This mechanism is known as 'leaky scanning' and it allows that different proteins can be obtained from same mRNA. Presence of an upstream AUG is correlated with the long 5'UTR and weak start codon signal.

AUG codon found in 5'UTR can indicate existence of upstream ORF (uORF).

If before the main start codon, an upstream AUG is found followed by in-frame stop codon it creates an upstream Open Reading Frame (uORF). After translation of uORF and the detachment of large ribosomal subunit, small unit can hold onto mRNA, resume scanning, and reinitiate translation at a downstream AUG codon, or it may leave the mRNA, and as a result damage translation of the main ORF. The ability of a ribosome to reinitiate is limited in eukaryote by the stop codon context and by the length of the uORF. uORFs longer than 30 codons disable ribosome to reinitiate.

Although the complete genome of the yeast *Saccharomyces cerevisiae* has been sequenced, the number of mRNAs containing uORFs is not known.

The facts that mRNA start sites are not known, and that some genes have more than one promoter, constitute major problems in the identification of real uORFs.

2.5.3 Different modes of action of uORFs

A number of features of the 5' untranslated region can determine the impact of a uORF on posttranscriptional gene expression.

On the 'visibility' of the uAUG, two features can have influence. Those are: the length of the sequence upstream of uAUG and the sequence context of the uAUG. Highly efficient recognition of the uORF AUG by scanning ribosomes will mean that the main ORF will almost exclusively be translated by ribosomes that have reinitiated after translating the uORF. Example of this mechanism is GCN4 uORF1.

On the ability of the ribosome to reinitiate following characteristics are crucial: The length of the uORF and the context of the uORF stop codon. If the uORF AUG is not efficiently recognized, a high percentage of the ribosome will bypass the uORF via leaky scanning and initiate directly on the main ORF. The stop codon context prevents reinitiating (as in GCN4 uORF4), possibly by causing rapid release of ribosomal subunits from the mRNA.

Stalling of the ribosome is influenced by uORF encoded peptide. The uORF can encode a peptide that causes stalling of ribosome at the termination codon, thus blocking all further process along the mRNA. Where such a peptide is not encoded, or the modulating ligand (arginine in CPA1 gene) is at a low level, there is rapid clearance of the termination complex, thus allowing ribosomes to progress along the mRNA.

3 Data acquisition and catalogue

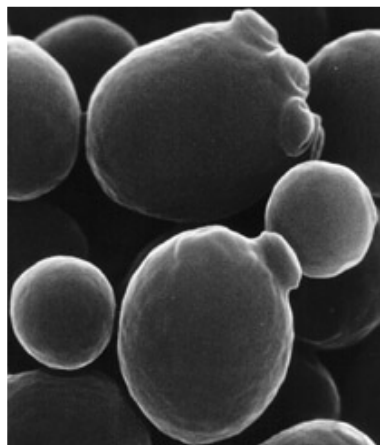
3.1 *Saccharomyces cerevisiae* as model organism

Saccharomyces cerevisiae (baker's yeast or budding yeast) is one of the simplest eukaryotic organisms. In the spring of 1996, the complete genome sequence of the *S. cerevisiae* was obtained, making yeast the first eukaryotic organism to be completely sequenced.

It is a small, single-cell fungus. Like other fungi, it has rigid cell wall and mitochondria but not chloroplast. It reproduces almost as fast as bacteria.

Most fundamental cellular processes are conserved from *S. cerevisiae* to humans and have first been discovered in yeast. *S. cerevisiae* cells divide in a similar manner to our own cells, and there are many other basic biological properties that are shared. About 20 per cent of human disease genes have counterparts in yeast. This suggests that such diseases result from the disruption of very basic cellular processes, such as DNA repair, cell division or the control of gene expression. It also means that we can use yeast to look at functional relationships involving these genes, and to test new drugs.

A yeast mutant that has lost the functional equivalent of a human disease gene can be screened with thousands of potential drugs in order to identify compounds that restore normal function to the yeast cell. These compounds might also be useful in humans. Due to the ease of handling and the highly developed genetics, functional genomics and cell biology yeast is an excellent model organism for training of candidates in cell biology, bioinformatics and Systems Biology.



Scanning electron micrograph of *S. cerevisiae*.
Alan Wheals, University of Bath, UK.

3.2 *Saccharomyces* Genome Database (SGD)

SGD is a scientific database of the molecular biology and genetics of the yeast *Saccharomyces cerevisia*.

The sequences and annotation for *S.cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. bayanus*, *S. castellii*, *S. kluyveri*, *S. Kudriavzevii* were obtained from the SGD.

We used TBlastN (as a part of Fungal Blast) to find the best orthologs to *S.Cerevisiae* genes.

3.3 InforMax

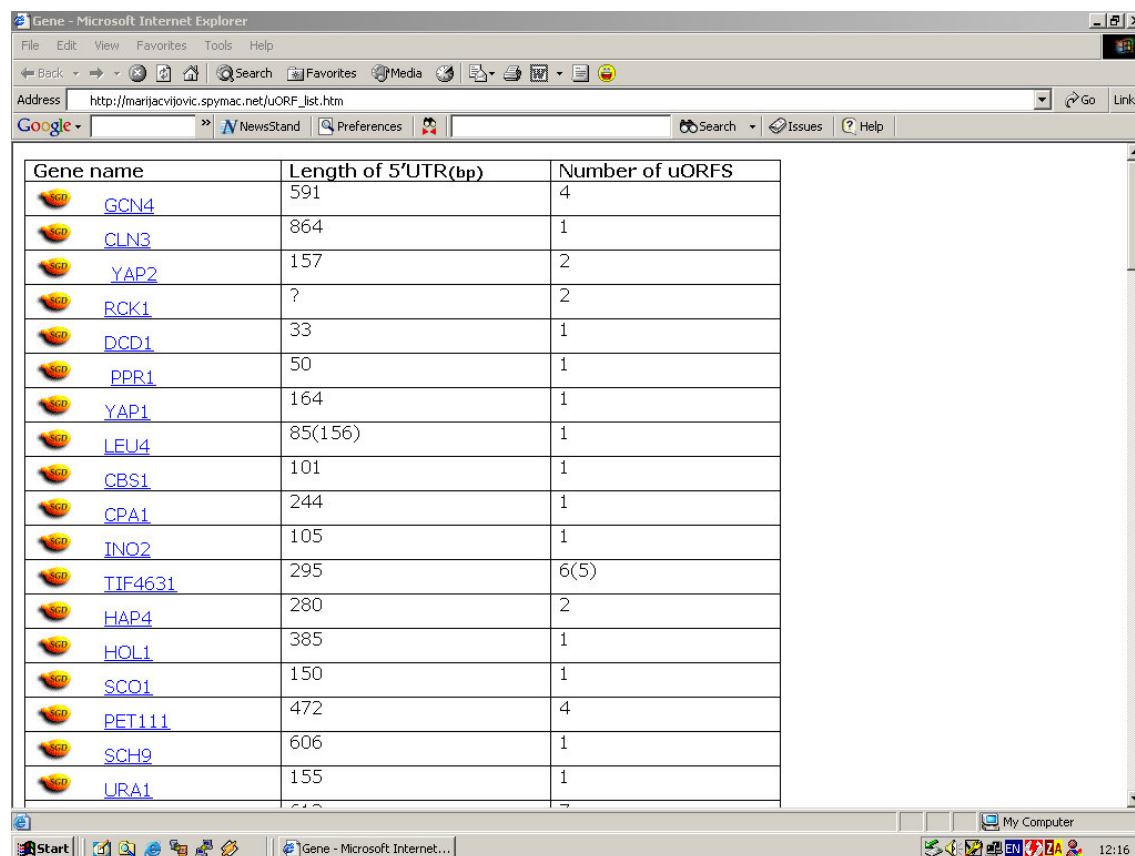
For storing and aligning the sequences, and making phylogenetic trees we used Vector NTI Suite, a product of InforMax Incorporated.

3.4 Making the catalogue

We constructed catalogue, which contains list of genes with identified uORFs. Extraction of 5'UTR sequences of *S.cerevisiae* was done using freely available UTRsource database (<http://bighost.area.ba.cnr.it/BIG/UTRHome/>). This database represents collection of data and analysis tool for the functional classification of 5' and 3'UTRs of eukaryotic mRNAs.

We made a collection of approximately 300 *S.cerevisiae* genes with know 5'UTR sequence. All set was analysed and in 90 genes upstream ORFs are detected. This was done using ORF Finder (<http://bioinformatics.org/>). It searches for open reading frames in the DNA sequence and then returns the range of each ORF, along with its protein translation.

Our catalogue contains 108 genes (90 new +18 - already known). Along with the gene name, it contains lengths of its 5'UTR sequence and number of uORFs. Also each gene has a link to the SGD web site and link to the page which contains its 5'UTR sequence, uORFs sequences and alignment with labelled uORFs for *S.cerevisiae* and 6 related species *S. paradoxus*, *S. mikatae*, *S. bayanus*, *S. castellii*, *S. kluyveri*, *S. kudriavzevii*.



Gene name	Length of 5'UTR(bp)	Number of uORFS
GCN4	591	4
CLN3	864	1
YAP2	157	2
RCK1	?	2
DCD1	33	1
PPR1	50	1
YAP1	164	1
LEU4	85(156)	1
CBS1	101	1
CPA1	244	1
INO2	105	1
TIF4631	295	6(5)
HAP4	280	2
HOL1	385	1
SCO1	150	1
PET111	472	4
SCH9	606	1
URA1	155	1

Figure 3-1 - Main page of the catalogue

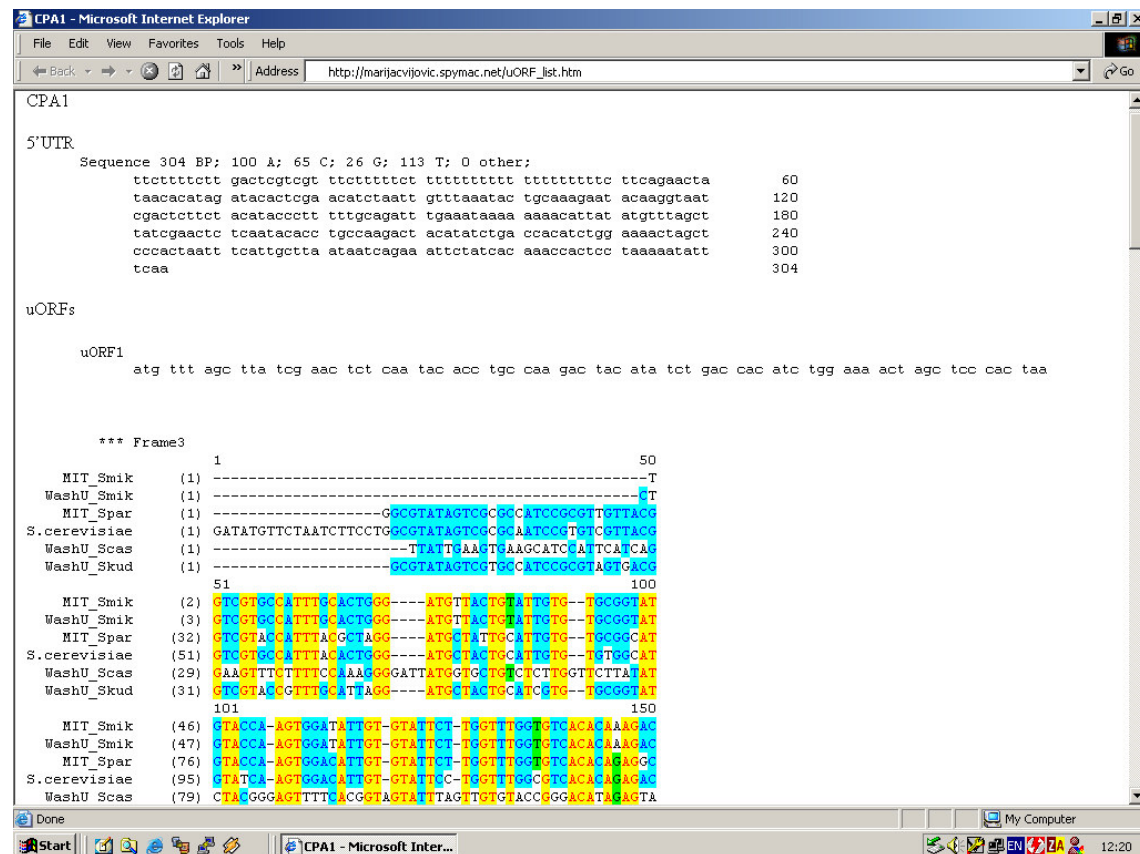


Figure 3-2 - CPA1 gene page

Each gene has its own page, which contains detailed overview of related information.

4 Study

In *S.cerevisia* only 18 genes have been documented to have uORFs. Idea of this project is to extract important parameters for conservation of uORFs between species using collection of studied genes.

Here, we give the list of genes used in this study.

Gene	Length of major 5'UTR (nt)	Number and size of uORFs (number of codons)	Gene product
CLN3	864	uORF1 (4)	G1 cyclin
GCN4	591	uORF1 (4), uORF2 (3), uORF3 (4), uORF4 (4)	Transcriptional activator of amino Biosynthetic pathway
INO2	105	uORF1 (20)	Transcriptional activator of phospholipids biosynthetic genes
PPR1	50	uORF1 (6)	Regulatory protein controlling transcription of two genes in pyrimidine biosynthesis pathway
SCO1	150	uORF1 (4)	PET gene involved in the accumulation of cytochrome c oxidase subunits I and II
URA1	155	uORF1 (24)	The enzyme catalyzes the conversion of dihydroorotic acid to orotic acid
CPA1	244	uORF1 (26)	Small subunit of cytosolic carbamoyl phosphate synthetase
HAP4	280	uORF1 (10), uORF2 (4)	Subunit of transcriptional activator complex binding CCAAT
LEU4	156	uORF1 (13)	A-Isopropylmalate synthase
RCK1	NA	uORF1 (6), uORF2 (12)	Serine/threonine protein kinase
TIF4631	295	uORF1 (12), uORF2 (20) uORF3 (15), uORF4 (8) uORF5 (22)	Translation initiation factor eIF4G1
YAP1	164	uORF1 (7)	Stress-related transcription factor
YAP2	157	uORF1 (6), uORF2 (23)	Stress-related transcription factor
CBS1	101	uORF1 (5)	PET gene involved in 5'end processing of the Cytochrome <i>b</i>
DCD1	33	uORF1 (4)	DCMP deaminase
HOL1	385	uORF1 (8)	Major facilitator family (drug resistance subfamily) of putative transport proteins
PET111	472	uORF1 (6), uORF2(31) uORF3 (11), uORF4(30)	Mitochondrial translational activator
SCH9	606	uORF1 (55)	Protein kinase that positively regulates the progression of yeast through G1 phase

Table 4-1 - List of studied genes

First step was to extract 5'UTRs from given genes and detect uORFs. All *S.cerevisiae* strains were downloaded from SGD web site and uORFs were identified using ORF Finder.

4.1 Distribution of uORFs

Analysing 18 *S.cerevisiae* genes it has been observed that uORFs can be spread along mRNA in three different ways.

Most of the uORFs follow the pattern found in GCN4 gene; they go after each other and last uORF always ends before coding sequence.

Another pattern is YAP2 example, where uORF overlaps with coding sequence. And the last and most difficult to observe is PET111 example where overlapping exist between uORFs and between uORFs and coding sequence.

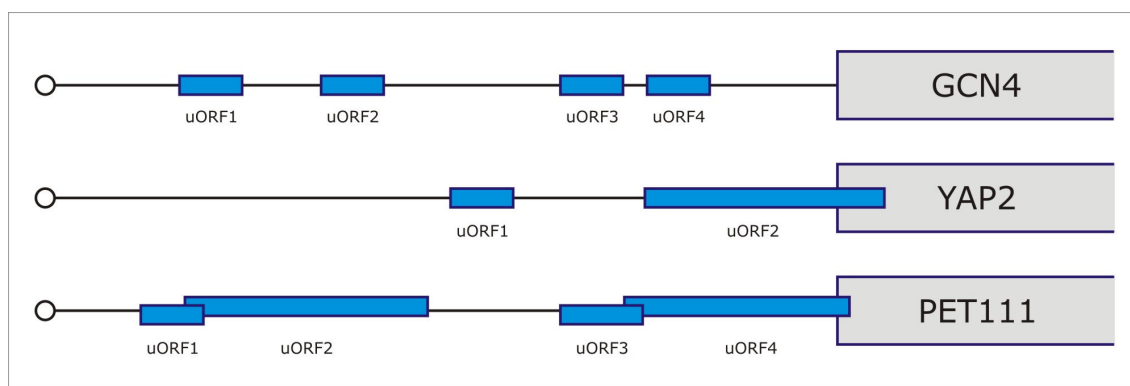


Figure 4-1 - Different types of uORF distribution (not drawn to scale)

4.2 Mechanism of uORFs

Since primary goal of this thesis is not biological explanation and study of principals involved in posttranscriptional control, only brief overview of *S.cerevisiae* three best studied uORF mechanisms will be given.

4.2.1 GCN4

This gene encodes a transcriptional activator that positively controls the expression of genes in response to amino acid starvation.

Upstream Open Reading Frames of GCN4 are among the best studied.

The 5'UTR of GCN4 is unique in many aspects:

- It is much longer (591bp) than most 5'UTR (only 6% of the genes have leader length over 150bp).
- It contains 4 AUG start codons followed closely by in frame termination codons.

uORFs mediate translational control via reinitiation mode. All uORFs are inhibitory to GCN4 translation.

uORF1 and uORF4 are critical for translational control response to amino acid limitation. They are sufficient to mediate the control mechanism.

They also have different properties:

- uORF1 translation allows efficient reinitiation
- uORF4 translation prevents reinitiation (has inhibitory function)

Behaviour of these two uORFs is not related to the specific peptide they encode. More relevant is the 3' ends of the uORFs.

The translation termination region around uORF1 is AU-rich and promotes reinitiation - UAA-ACCGAUUAUA.

Equivalent region around uORF4 is GC-rich and doesn't promote reinitiation - UAA-CGGUUACCU.

4.2.2 YAP1 and YAP2

Like GCN4, YAP1 and YAP2 are regulatory genes involved in the mechanism used by the yeast cell to protect itself in situation of stress. Also, both of them encode proteins that have significant homology to Gcn4.

YAP2 has 157bp long leader and two uORFs: uORF1 is 6 codons long and uORF2 has 23 codons. uORF2 overlaps with YAP2 ORF.

YAP1 leader is 164bp long containing only one 7 codons long uORF. This uORF is efficiently translated and promotes efficient reinitiation.

YAP2 uORFs act to block ribosomal scanning and also to accelerate mRNA decay, they also define a new type of mRNA destabilizing element.

Elimination of either one or both uORFs greatly enhances YAP2 translation.

4.3 Genome Structure and Evolution

Conservation of uORFs among related species might be a solution for finding functional uORFs. We assumed that if the uORF is conserved between *S.cerevisiae* sister species it has a function.

For cross-species study *S. paradoxus*, *S. mikatae*, *S. bayanus*, *S. kudriavzevii*, *S. castellii*, *S. kluyveri*, were selected.

First four species are separated from *S.cerevisiae* by estimate 15-20million years of evolution and belong to *Saccharomyces sensu stricto* group, while last two are more distant (approximately 150million years).

Selection of the species was based on several reasons: First, their genomes are sequenced and data are freely available through public sequence database and through the SGD maintained at Stanford. Second, sufficient sequence similarity to *S.cerevisiae* exist to allowed orthologous regions to be aligned reliably, but sufficient sequence divergence allow many functional elements to be recognized.

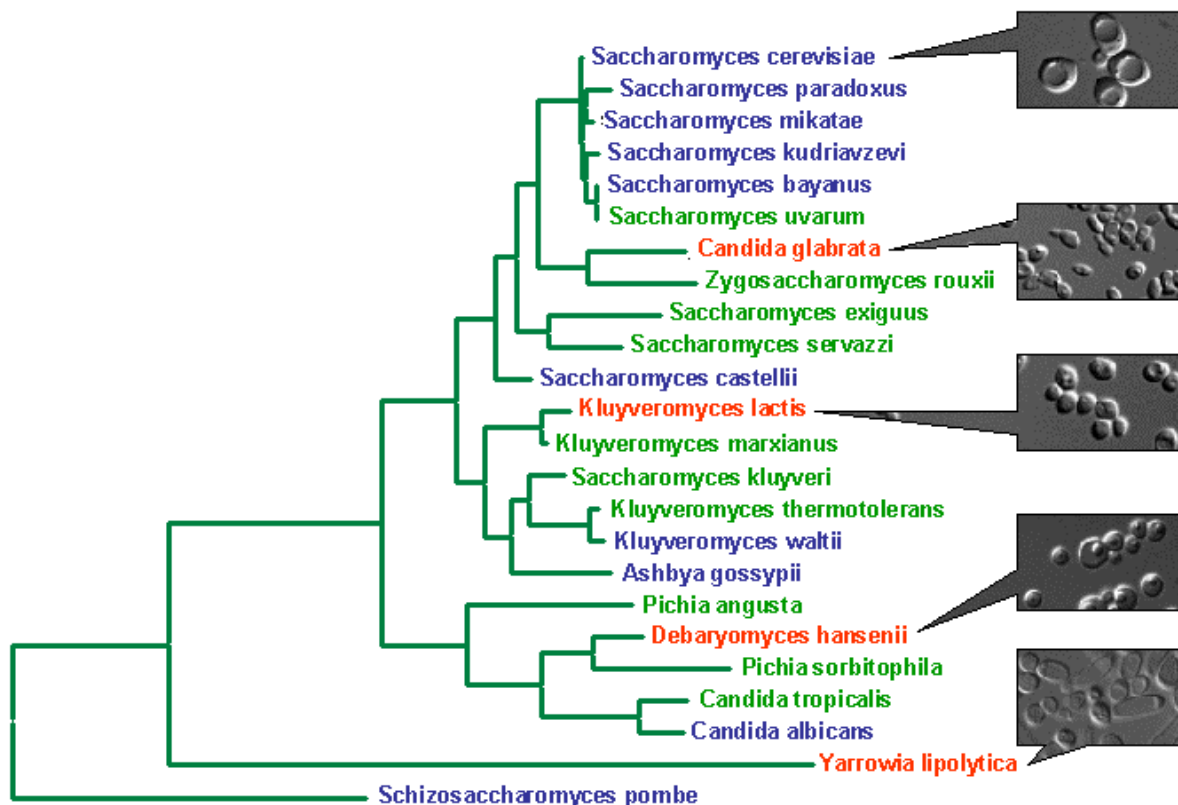


Figure 4-2 - Evolutionary tree
(<http://cbi.labri.fr/Genolevures/index.php>)

4.4 Genome alignment and synteny

4.4.1 Synteny

The six genomes show striking conservation of synteny and they can be well aligned at the nucleotide level.

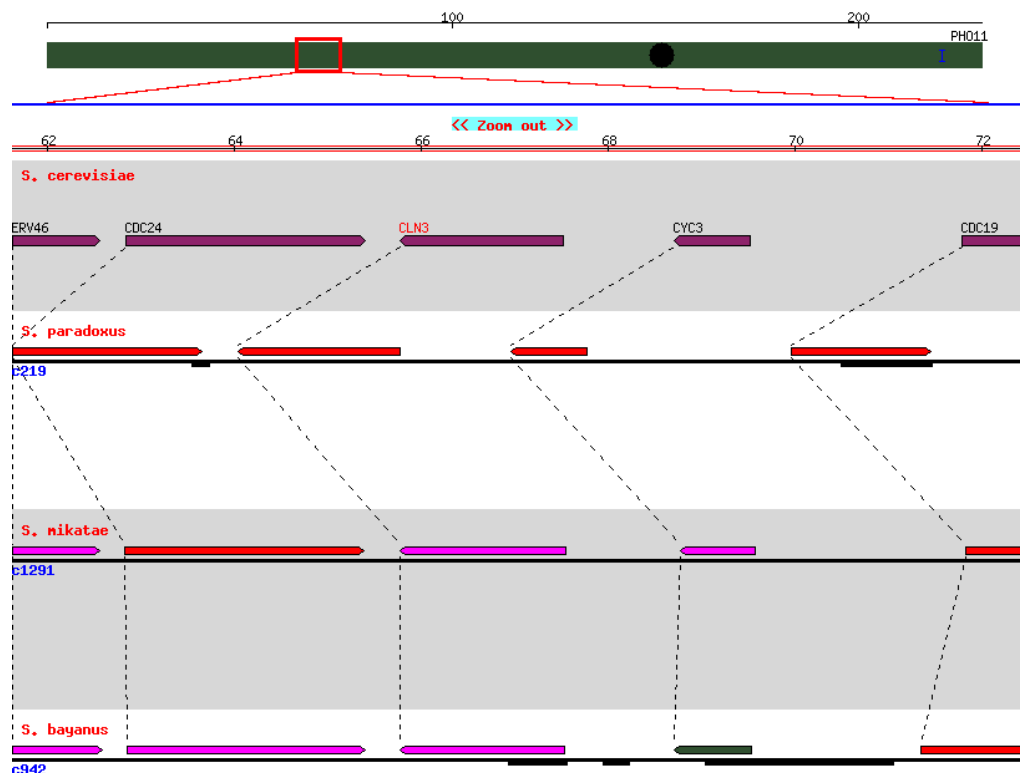


Figure 4-3 - Synteny viewer for CLN3 gene (neighbouring genes are also represented)

4.4.2 Alignment

We used ClustalW Multiple Alignment method for aligning the sequences. Due to the fact that 5'UTR sequences are not even known for all *S.cerevisiae* genes, we constructed alignments using 1000bp upstream region from initial AUG.

Orthologues genes sequences of chosen organism were downloaded and uORFs were identified. We compared upstream sequences using ClustalW Multiple Alignment method. Since, the idea was to make small database of tested genes, for alignment and sequence storing we used InforMax package. In this way all the alignments and results were stored and are instantly available.

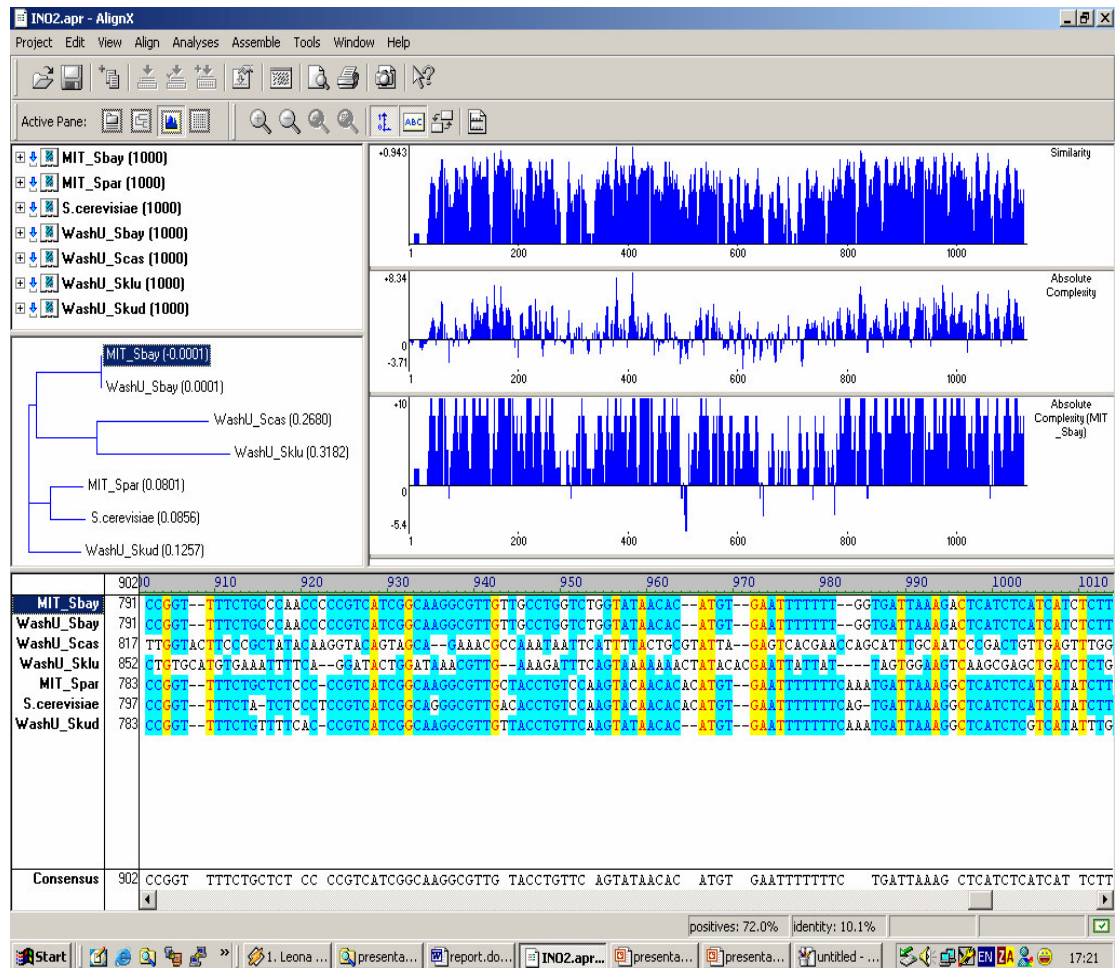


Figure 4-4 – InforMax screenshot

After this, the more time consuming part came – labelling all detected uORFs.

		701		750
MIT_Sbay	(594)	TTACAGAAAT	TAAAGATCAT	TGAAAA-TGGCTTGCTAAACCAATTATAT
WashU_Sbay	(555)	TTACAGAAAT	TAAAGATCAT	TGAAAA-TGGCTTGCTAAACCAATTATAT
MIT_Smik	(609)	TTACAGAAAT	TAAAGATCAT	TGAAAA-TGGCTTGCTAAACCAATTGTAT
WashU_Smik	(609)	TTACAGAAAT	TAAAGATCAT	TGAAAA-TGGCTTGCTAAACCAATTGTAT
MIT_Spar	(312)	TTACAGAAAT	TAAAGATCAT	TGAAAA-TGGCTTGCTAAACCAATTATAT
S.cerevisiae	(615)	TTAAA-AAAT	TAAAGATCAT	TGAAAA-TGGCTTGCTAAACCGATTATAT
WashU_Scas	(660)	TACCTCCAAAT	TATGACAGT-AAAAA	TGTTTGCTAAACCAATTATAT
WashU_Sklu	(625)	---AGAGTA	TCACTATTC	AAATGGCTTGCTAAACCAATTATAT
WashU_skud	(608)	T-----	TAAAGATCAT	TGAAAA-TGGCTTGCTAAACCAATTGTAT

Here, part of GCN4 gene alignment is given. *S.cerevisiae* uORF1 is shown in pink. Symbols +, - and * represent reading frame 1, 2 and 3, respectively.

4.5 Extended Analysis

Previously in this chapter we described role of GCN4 uORFs in translation efficiency. Since, it is the *S.cerevisiae* the best-studied example of uORFs mechanism; we extended analysis, finding orthologs in 13 different species.

We compared the amino acid sequence of GCN4 gene against the set of annotated genomes in SGD. This set includes:

3 Complete Genomes (*Ashbya gossypii* (UniBas), *Saccharomyces cerevisiae* (Reference Sequence), *Schizosaccharomyces pombe* (Sanger))

9 Annotated Genomes (*Aspergillus nidulans* (MIT), *Candida albicans* (Stanford), *Candida glabrata* (Genolevures II), *Debaryomyces hansenii* (GenolevuresII), *Encephalitozoon cuniculi* (GenoScope), *Kluyveromyces lactis* (GenolevuresII), *Magnaporthe grisea* (MIT), *Neurospora crassa* (MIT), *Yarrowia lipolytica* (GenolevuresII))

7 Fungal Genome Initiative (Assembled, Annotation In Progress) *Aspergillus terreus* (MIT), *Coprinus cinereus* (MIT) *Cryptococcus neoformans serotype A* (MIT), *Fusarium graminearum aka Gibberella zeae* (MIT), *Kluyveromyces waltii* (MIT), *Phanerochaete chrysosporium* (DOE JGI), *Ustilago maydis* (MIT).

Orthologous identification was done using WU-BLAST2 TBlastN method (part of Fungal Blast Tool at SGD).

From this set we identified orthologs in 13 species: *Ashbya gossypii*, *Aspergillus nidulans*, *Candida albicans*, *Candida glabrata*, *Debaryomyces hansenii*, *Kluyveromyces lactis*, *Magnaporthe grisea*, *Neurospora crassa*, *Yarrowia lipolytica*, *Cryptococcus neoformans*, *Gibberella zeae*, *Kluyveromyces waltii*, *Ustilago maydis*.

The next step was isolating upstream region of selected genes. Since the 5'UTR sequences are not known for these species, we took 1000bp upstream of initial AUG. Then, we identified all possible uORFs in these sequences, in all 3 different reading frames.

We aligned 20 sequences in total – 13 already mention, 6 *S.cerevisiae* sister species and *S.cerevisiae*. CLUSTALW alignment method was used.

We presented 450bp of each sequence, as the *S.cerevisiae* uORF are in this range and from alignment was clear that uORF from rest of the organisms will fit to this range. uORFs are represented as boxes and one-to-one correspondence between relevant ORFs was based upon multiple alignment.

This result gave us a good overview of uORF distribution and showed us that GCN4 uORFs are very well conserved trough evolution. Since the evolution distance among selected organisms is very far, this result can give us hope that only conserved uORFs are functional ones.

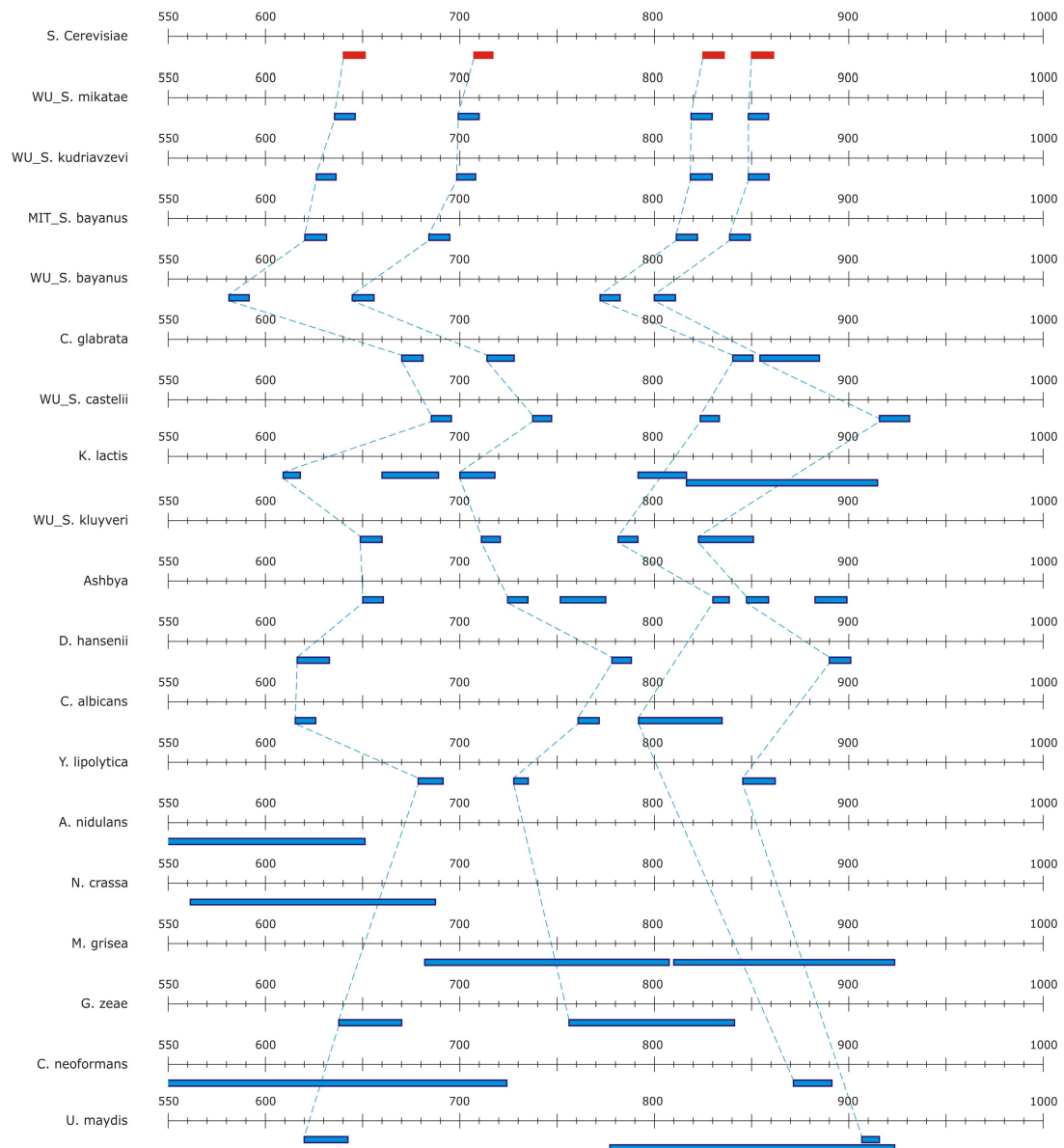


Figure 4-5 - Schematic representation of GCN4 uORFs in different organisms

In this figure uORFs of *S. cerevisiae* GCN4 gene are shown as red colour boxes and they are drawn to scale. Organisms are sorted in evolutionary order and their uORFs are shown as blue boxes (drawn to scale, also). Dashed lines represent uORFs correspondence.

5 Results

For 9 out of 18 genes it has been shown that uORFs are very well conserved (GCN4, CLN3, YAP2, PET111, HOL1, CPA1, HAP4, YAP1, TIF4631). By this we mean, that position and length of uORFs is almost identical as in *S.cerevisiae* corresponding genes.

For 2 genes, CBS1 and SCO1, it has been shown earlier that uORFs don't have influence on translation efficiency. Our analysis of these two genes showed that uORFs are not conserved, what supports our theory that only conserved uORFs have biological function.

Location of uORFs in these 2 genes is maybe the reason why are they without influence on translation. They are located closely to the initial AUG. CBS1 uORF is on position -50 in respect of main AUG, while SCO1 uORF is on position -48.

Results of our study are summarized in the following table:

Gene	uORF conservation yes/no	biological function yes/no	if without function, possible reason
CLN3	yes	yes	-
GCN4	yes	yes	-
INO2	no	no	uORF is too long
PPR1	no	no	uORF too close to main AUG or short 5'UTR
SCO1	no	no	uORF too close to main AUG (952-963)
URA1	no	no	uORF too close to main AUG (885-956) and too long
CPA1	yes	yes	-
HAP4	yes	yes	-
LEU4	no	no	uORF too close to main AUG (945-983)
RCK1	no	NA	uORFs too long
TIF4631	yes	yes	-
YAP1	yes	yes	-
YAP2	yes	yes	-
CBS1	no	no	uORF too close to main AUG (950-964)
DCD1	no	no	uORF too close to main AUG (973-984) or short 5'UTR
HOL1	yes	yes	-
PET111	yes	yes	-
SCH9	no	no	uORF is too long (55codons)

Table 5 - 1 - Results

Analysing set of 18 genes following conclusion can be made:

- smaller uORFs (4 to 6 codons) have greater impact on post-transcriptional control
- location of uORFs is important (i.e. if uORFs are too close to initial AUG they are not functional)
- peptide which uORF encodes is not relevant

Analysing whole collection it has been observed that:

- Length of the leader can varies between 30 and 2000bp (average length is ~370bp)
- A longer UTR doesn't necessarily correlate to a longer uORF or higher uORF frequency
- 46.5% of all studied genes have only 1 uORF
- 82% of all studied genes have up to 4 uORFs

6 Conclusions

This work showed that comparative genomic approach is a good way to extract important parameters for conservation of uORFs.

Even though, the set of genes we analysed is too small, we were able to identify essential properties of upstream ORFs and prove their functionality *in silico*.

6.1 Future work

We will use the derived rules and make a model, which will scan the entire *S. cerevisiae* genome to identify genes that are predicted by this method to be regulated on the translational level by uORFs.

References

- [1] Alberts, Bray, Johnson, Lewis, Raff, Roberts, Walter (1998) *The Cell*, Garland Publishing, New York
- [2] Pesole, G., et Al (2000) *The untranslated regions of eukaryotic mRNAs: Structure, function, evolution and bioinformatic tools for their analysis*, Briefings in Bioinformatics, Vol I, No3.236-249
- [3] Mignone, F., et Al (2002) *Untranslated regions of mRNAs*, Genome Biology, Vol 3 No3
- [4] Pesole, G., et Al (2002) *UTRdb and UTRsite: specialized database of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs*, Nucleic Acid Research, Vol 30, No 1 335-340
- [5] Pesole, G., et Al (2001) *Structural and functional features of eukaryotic mRNA untranslated regions*, Gene 276, 73-81
- [6] Morris, D. R., Geballe, A. P., (2000) Upstream Open Reading Frames as regulators of mRNA translation, Mol. and Cellular Biology, 8635-8642
- [7] Vilela, C., et Al (1998) The yeast transcription factor genes YAP1 and YAP2 are subject to differential control at the levels of both translation and mRNA stability, Nucleic Acid Research, Vol. 26, No5, 1150-1159
- [8] Dever, T., (2002) Gene-Specific Regulation by General Translation Factors, Cell, Vol. 108, 545-556
- [9] Thireos, G., et Al (1984), 5' untranslated regions are required for the translational control of a yeast regulatory genes, Proc. Natl. Acad. Sci., Vol.81, 5096-5100
- [10] Polymenis, M., Schmidt, E.V., (1997) Gene and Development, 11:2522-2531
- [11] Vilela, C., McCarthy, J.E.G, (2003) Regulation of fungal gene expression via short open reading frames in the mRNA 5' untranslated region, Molecular Microbiology, 49(4), 859-867
- [12] McCarthy, J.E.G, (1998), Posttranscriptional control of gene expression in yeast, Microbiology and Molecular Biology Review, 1492-1553
- [13] Kellis, M., et Al (2003), Sequencing and comparison of yeast species to identify genes and regulatory elements, Nature, Vol.423
- [14] Meijer, H., Thomas, A., (2002) Control of eukaryotic protein synthesis by upstream open reading frames in the 5' untranslated region of an mRNA, Biochemical Journal, BJ20011706
- [15] <http://opbs.okstate.edu/~melcher/MG/MGW2/MG237.html>
- [16] http://163.238.8.180/~davis/Bio_327/lectures/Post_Tx_Processes/RNA_stability.html

[17] <http://www.yeastgenome.org/>

[18] <http://bighost.area.ba.cnr.it/BIG/UTRHome/>

[19] <http://www.genedb.org>

Appendix A

Multiple Sequence Alignment

A multiple alignment arranges a set of sequences in a scheme where positions believed to be homologous are written in a common column. Although it would be biologically meaningful, the distinctions between global, local and other forms of alignment are rarely made in a multiple alignment. The reason for this is computational difficulties in computing multiple alignments.

Motivation

- A multiple alignment gives a stronger signal for sequence similarities than a pairwise alignment, i.e. in:
 - sequence assembly
 - molecular modelling
 - database search
 - primer design
- Ambiguities in the alignment may be resolved by a multiple alignment
- Looking into the relationship between the sequences leads to reconstructing a phylogenetic tree

We used ClustalW profile-based progressive multiple alignment method. The term “progressive” is used because ClustalW starts with a pairwise method to determine the most related sequences and then progressively adding less related sequences or groups of sequences to the initial alignment.

Here the algorithm for ClustalW is given:

1° Construct a distance matrix of all $N(N-1)/2$ pairs by pairwise dynamic programming alignment followed by approximate conversion of similarity scores to evolutionary distances using Kimura model

2° Construct a guided tree by a neighbouring-joining clustering algorithm by Saitou and Nei

3° Progressively align at nodes in order of decreasing similarity, using sequence-sequence, sequence-profile and profile-profile alignment [Bio.Seq. Analysis]

In order to compensate for biased representation in large subfamilies, ClustalW reduces the weight of related sequences.

To assign a weight to a sequence, ClustalW does the following:

1° It uses the values on the Neighbour – joining -tree from the sequence to the root of the tree.

2° If two or more sequences share a branch, which may indicate an evolutionary relationship, its value is split amongst the sequences.

To assign weights to the sequences in scoring multiple alignment:

- Each sequence has assigned a weight and the contribution of this sequence in the global alignment is scaled using this weight.
- We assume that the evolutionary tree computed in the second stage of the algorithm has an associated branch length. Then the weight of associated with the sequence is proportional to the sum of the branch length from the root to the corresponding leaf.

Neighbour – joining method

For constructing the tree Neighbour – joining (NJ) method was used. The NJ method works on a matrix of distances between all pairs of sequence to be analyzed. These distances are related to the degree of divergence between the sequences. The phylogenetic tree is calculated after the sequences are aligned. N-J Method produces an Unrooted, Additive tree.

Characteristics:

- The neighbour joining method is a heuristic method which joins at each step, the two closest sub-trees that are not already joined
- It is based on the minimum evolution principle
- One of the important concepts in the NJ method is *neighbours*, which are defined as two taxa that are connected by a single node in an unrooted tree

BLAST

BLAST (Basic Local Alignment Search Tool) is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. The scores assigned in a BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background hits. 5 different types of BLAST exist:

BLAST search type	Description
BLASTn	A type of BLAST search in which a nucleotide sequence is compared with the contents of a nucleotide sequence database to find sequences with regions homologous to regions of the original sequence.
BLASTp	A type of BLAST search in which an amino acid sequence is compared with the contents of an amino acid sequence database to find sequences with regions homologous to regions of the original sequence.

BLASTx	A type of BLAST search in which a nucleotide sequence is compared with the contents of an amino acid sequence database to find sequences with regions homologous to regions of the original sequence. The query sequence is translated in all six reading frames, and each of the resulting sequences is used to search the sequence database.
tBLASTn	A type of BLAST search in which an amino acid sequence is compared with the contents of a nucleotide sequence database to find sequences with regions homologous to regions of the original sequence. The sequences in the sequence database are translated in all six reading frames, and the resulting sequences are searched for regions homologous to regions of the query sequence.
tBLASTx	A type of BLAST search in which a nucleotide sequence is compared with the contents of a nucleotide sequence database to find sequences with regions homologous to regions of the original sequence. In a tBLASTx search, both the query sequence and the sequence database are translated in all six reading frames, and the resulting sequences are compared to discover homologous regions.

BLAST uses a heuristic algorithm which seeks local as opposed to global alignments and is therefore able to detect relationships among sequences which share only isolated regions of similarity (Altschul et al., 1990).

The initial search is done for a word of length "W" that scores at least "T" when compared to the query using a substitution matrix. Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of "S". The "T" parameter dictates the speed and sensitivity of the search. [NCBI]

In this study TBLASTN method was used for finding orthologs genes in various numbers of organisms.

The WordUP algorithm

This algorithm has been used for creating the UTRresource database.

The **WordUp** algorithm is aimed at the identification of statistically significant nucleotide strings which are shared or avoided in a set of sequences functionally equivalent but not evolutionary homologous (e.g. promoter regions, introns, etc.). The statistical significance of each oligonucleotide signal is simply determined through a chi-square test by comparing the actual and the expected number of sequences containing that given signal, calculated assuming that oligonucleotides are Poisson distributed and that their occurrence probability follows a first order Markov chain, i.e. depends on dinucleotide frequencies. The Poisson distribution is suitable for the description of rare events, such as the distribution of oligomers longer than w nucleotides in sequences quite shorter than 4^w nucleotides.

It must be stressed that statistical significance, even though it does not account for biological significance, can provide a substantial clue for this. The lists of statistically significant motifs,

i.e. those having a chi-square value above a given threshold constitute a motif vocabulary which is specific for the biological function shared by the analysed sequences.

The starting word length, w , to be used in the analysis, has to be defined as the shortest sequence allowing the validation that sequence oligomers are Poisson distributed (i.e. $L_{seq} \ll 4^w$). The best choice will be, thus, in general a string length six nucleotides long even if we do not know the actual length of the biological signals. In order to establish if there are significant oligomers of length $w+1$, $w+2$, ... we follow a dynamic elongation procedure, which considers pattern pairs overlapping by $w-1/w$ nucleotides. If the $w+1$ -mer is more significant than its two component w -mers, the $w+1$ -mer replaces the latter in the vocabulary of significant motifs. If this does not happen of the two overlapping w -mers the least significant is removed from the vocabulary, as its significance is likely to be due to the "overlapping effect". In general, this procedure can be used to determine the statistical significance of $w+k$ -mers considering overlapping $w+k-1$ -mers. [Pesole, G., et Al]