Universidade de Aveiro Departamento de Biologia
2008

Ana Catarina Batista Gomes

Evolução molecular de uma alteração ao código genético.

Molecular evolution of a genetic code alteration.

Universidade de Aveiro Departamento de Biologia
2008

Ana Catarina Batista Gomes

Evolução molecular de uma alteração ao código genético.

Molecular evolution of a genetic code alteration.

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Biologia, realizada sob a orientação científica do Prof. Doutor Manuel António da Silva Santos, Professor Associado do Departamento de Biologia da Universidade de Aveiro

Apoio financeiro da FCT e do FSE no âmbito do III Quadro Comunitário de Apoio.

o júri

presidente

Doutora Maria Celeste da Silva do Carmo
Professora Catedrática da Universidade de Aveiro

Doutor Amadeu Mortágua Velho da Maia Soares
Professor Catedrático da Universidade de Aveiro

Doutor Manuel António da Silva Santos
Professor Associado da Universidade de Aveiro

Doutor Francisco Manuel Lemos Amado
Professor Auxiliar da Universidade de Aveiro

Doutor Alexandre Akoulitchev
Senior Research Fellow, Universidade de Oxford

Doutor Lluis Ribas de Pouplana
Research Professor, ICREA, Universidade de Barcelona

agradecimentos

First and foremost, I would like to thank my supervisor, Manuel Santos, for the opportunity to work on this project and for his invaluable guidance from the very beginning of my scientific career. With you I have learnt to give the most of myself and to go for nothing but excellence: even if frustrating in the short term, it is always fruitful in the long term. Under your supervision, I have grown, both personal and professionally, and I have broadened my horizons and that is what I am most thankful for.

I am extremely grateful to Alexandre (Sasha) Akoulitchev for receiving me in his laboratory, giving me the opportunity for not only to perform all the Mass-Spectrometry experiments, which represent the core of this thesis, but also to experience the effervesce of science at Oxford. Thanks for your encouragement – especially when nothing worked. I am also grateful to Lluis Ribas de Pouplana for accepting me in his laboratory to carry out the aminoacylation experiments, it was a short but very pleasant stay.

As science is a highly collaborative activity, this work would not have been the same without the precious input of several people, to whom I would like to particularly thank: Isabel Miranda, for getting me into the biology and physiology of *C. albicans;* Benjamin Thomas, for teaching and helping so much on Mass-Spectrometry; Gabriela Moura, for all the help with the ANACONDA and for the most interesting lunches; Pedro Beltrão, for creating all the algorithms I needed, whenever I needed; Tatiana Lima-Costa, for all the help on SNPs detection; and Renaud Geslain for introducing me into the aminoacylation kinetics world (I just wish we would have any results).

I would also like to express my gratitude to all the past and present elements of Manuel's lab for their support, friendship and for making the lab such a pleasant place to work, and to everyone at Sasha's and Lluis' lab for making my staying abroad so enjoyable and less lonely.

I want to thank all my friends, in particular to Bruno, Ricardo and Diogo, who were always a phone call away.

I am most grateful to my family, for their support, in special to my parents, who always taught that without hard work nothing is possible, and stood by for all my choices, and to my sister.

Finally, I want to thank Zé, for your support, your love, your amazing ability of keeping me focus on my objectives, and for always being there.

To all of you: Thanks!

palavras-chave tRNAs, aminoacyl-tRNA synthetases, genetic code, *Candida albicans*

resumo

Durante os últimos anos, foram descritas alterações ao código genético, quer em procariotas, quer em eucariotas, quebrando o dogma de que o código genético é universal e imutável. Estudos recentes sugerem que a evolução de tais alterações requerem modificações ao nível da estrutura da maquinaria da tradução e são promovidas por mecanismos de descodificação ambígua. Em *C. albicans*, um organismo que é patogénico para o Homem, a alteração ao código genético é mediada por uma alteração na estrutura de um novo $tRNA_{CAG}$ de serina que descodifica o codão CUG de leucina como serina.

De forma a determinar se este tRNA, que é aminoacilado pelas Seryl- e Leucyl- tRNA sintetases, promove a descodificação ambígua do codão CUG, foi desenvolvido um sistema para a quantificar *in vivo*, por espectrometria de massa, os níveis de incorporação de serina e de leucina em codões CUG. Os resultados mostraram que em condições normais de crescimento leucina é incorporada a uma taxa de 3% e que serina é incorporada a uma taxa de 97%. No entanto, o nível de ambiguidade na descodificação de codões CUG aumentou para 5% em células crescidas em condições de stress, indicando que a incorporação de leucina em codões CUG é sensível a factores ambientais e é manipulada durante a tradução do mRNA. Tal, levanta a hipótese de que a incorporação de leucina poderá atingir níveis superiores aos determinados neste estudo. Para testar esta hipótese e determinar os níveis máximos de ambiguidade na descodificação do codão CUG tolerados pelas células, aumentou-se artificialmente a ambiguidade do codão CUG em *C. albicans*. Surpreendentemente, a incorporação de leucina subiu de 5% para 28%, o que representa um aumento na taxa de erro da tradução de 3500 vezes, relativamente ao descrito para o mecanismo de tradução.

Dado existirem 13.000 codões CUG no genoma de *C. albicans*, a sua descodificação ambígua expande de uma forma exponencial o proteoma deste fungo, criando assim um proteoma estatístico, resultante da síntese de um conjunto de moléculas diferentes para cada proteína a partir de um único RNA mensageiro (mRNA) que contenha codões CUG.

Os resultados obtidos demonstraram que o proteoma de *C. albicans* tem uma dimensão muito superior à prevista pelo seu genoma e demonstram um papel central da descodificação ambígua na evolução do código genético.

keywords tRNAs, aminoacyl-tRNA synthetases, genetic code, *Candida albicans*

abstract

Alterations to the standard genetic code have been found in both prokaryotes and eukaryotes, demolishing the dogma of an immutable and universal genetic code. Recent studies suggest that evolution of such alterations require structural change of the translation machinery and are driven through mechanisms that require codon decoding ambiguity. In the human pathogen *C. albicans*, a structural change in a novel ser-$tRNA_{CAG}$ allows for its recognition by both the LeuRS and SerRS *in vitro* and *in vivo*, providing such molecular device.

In order to determine whether this tRNA charging ambiguity results in ambiguous CUG decoding, we have developed a system for quantification of the level of serine and leucine at the CUG codon by Mass-Spectrometry. The data showed that 3.0% of leucine and 97.0% of serine are incorporated at CUG codons *in vivo* under standard growth conditions. Moreover, this ambiguity increases up to 5.0% under stress, indicating that it is sensitive to environmental change and raising the hypothesis that leucine incorporation may be higher than determine experimentally. In order to determine the scope of *C. albicans* tolerance to CUG ambiguity, we have created highly ambiguous *C. albicans* cell lines through tRNA engineering. These cell lines tolerated up to 28% leucine incorporation at CUGs, which represents an increase of 3500 fold in decoding error rate.

Since there are 13,000 CUG codons in *C. albicans* such ambiguity expands the proteome exponentially and creates a statistical proteome due to synthesis of arrays of protein molecules from mRNAs containing CUG codons.

The overall data showed that the dimension of the *C. albicans* proteome is far higher than that predicted from its genome and provides important new evidence for a pivotal role for codon ambiguity in the evolution of the genetic code.

Contents

Contents xi

List of Figures xiv

List of Tables xvi

1. Introduction 1

1.1. The genetic code 3
1.1.1. The standard genetic code 3
1.1.2. The origin and early evolution of the genetic code 4

1.2. Translation 8
1.2.1. Translation initiation 9
1.2.2. Translation elongation 10
1.2.3. Translation termination 13

1.3. The operational RNA code 14
1.3.1. Transfer RNAs 14
1.3.2. Aminoacyl-tRNA synthetases 21

1.4. Genetic code alterations 34
1.4.1. The mechanisms of evolution of genetic code alterations 34
1.4.2. Mitochondrial Genetic Code alterations 37
1.4.3. Cytoplasmic genetic code alterations 40
1.4.4. The Expansion of Genetic Code 40

1.5. The *Candida* spp. genetic code 46
1.6.1. The $tRNA_{CAG}^{Ser}$ 48
1.6.2. The evolution of CUG codon reassignment 49

1.6. Objectives of this study 51

2. Materials & Methods 53

2.1. Strains and Growth Conditions 55
2.1.1. Strains and genotypes 55
2.1.2. Growth and Maintenance of *E. coli*, *S. cerevisiae* and *C. albicans* 55

2.2. DNA Manipulation 56
2.2.1. Oligonucleotides 56
2.2.2. Plasmids 59
2.2.3. DNA amplification by PCR 62
2.2.4. PCR product purification 62
2.2.5. Agarose Gel electrophoresis 62

2.2.6. DNA extraction from agarose gel 63
2.2.7. DNA digestion with restriction enzymes 63
2.2.8. DNA Dephosphorylation and ligation 64
2.2.9. Transformation of *E. coli* 64
2.2.10. Site Directed Mutagenesis 66
2.2.11. Nucleic Acids precipitation and quantification 67
2.2.12. DNA sequencing 68
2.2.13. Transformation of *C. albicans* 68
2.2.14. *C. albicans* genomic DNA extraction 70

2.3. Protein Extraction, Purification and Analysis 70
2.3.1. Protein Extraction 70
2.3.2. Protein Purification 71
2.3.3. Protein Quantification 72
2.3.4. *P*oly*a*crylamide *g*el *e*lectrophoresis (PAGE) 73
2.3.5. Western-blotting analysis 73
2.3.6. *In gel* protein digestion 75
2.3.7. Mass-Spectrometry 75

2.4. Overexpression and purification of the *C. albicans* Ser-tRNA$_{CAG}$ 76
2.4.1. tRNA purification by affinity chromatography 77
2.4.2. High resolution tRNA electrophoresis 79

2.5. Aminoacylation kinetics assays 79

2.6. Bioinformatic tools and data mining 81
2.6.2. Protein and gene sequence alignments and phylogenetic analysis 82
2.6.3. Protein structure modelling 82

3. Quantification of CUG ambiguity in *C. albicans in vivo* by Mass-Spectrometry 83

3.1. Introduction 85

3.2. Results 90
3.2.1. Construction of a CUG mistranslation reporter system 90
3.2.2. Determination of leucine and serine incorporation at the CUG codon *in vivo* 101
3.2.3. *C. albicans* tolerates partial reversion of CUG identity 106

3.3. Discussion 108

4. The impact of CUG ambiguity in *C. albicans* biology 113

4.1. Introduction 115

4.2. Results 119
4.2.1. *C. albicans* has a statistical proteome 119
4.2.2. *C. albicans*' genome is optimized for CUG ambiguity 126
4.2.3. The CUG usage in *C. albicans* 129
4.2.4. The evolution of the CUG codon in *C. albicans'* genome 145

4.3. Discussion **150**

5. The role of the Leucyl- and Seryl- tRNA Synthetases in CUG ambiguity 153

5.1. Introduction **155**

5.2. Results **157**
5.2.1. Quantification of SerRS and LeuRS expression in *C. albicans* 157
5.2.2. The study of SerRS and LeuRS genes 160
5.2.3. Functional insights of the LeuRS and SerRS polymorphisms. 167
5.2.4. The aminoacylation of *C. albicans* $tRNA_{CAG}$ by the LeuRS and SerRS 172
5.2.5. Aminoacylation assays 177

5.3. Discussion **180**

6. General Discussion 183

6.1. The uniqueness of the *C. albicans* genetic code **185**

6.2. CUG ambiguity and the evolution of the *C. albicans* genome **187**

6.3. Hypothetical models for regulation of leucine incorporation at the CUG codon **188**

6.4. Conclusion **193**

6.5. Future work **195**

7. Annexes 197

Annexe A: Map of the Plasmids **199**

Annexe B: Reporter protein data **203**

Annexe C: MS-MS of both synthetic and reporter peptides **207**

Annexe D: Results from the clustering analysis **209**

Annexe E: Leucyl – tRNA synthetase **214**

Annexe F: Seryl – tRNA synthetase **217**

Annexe G: Sequencing of the promoter regions of Leucyl-tRNA synthetase **219**

8. References 221

List of Figures

Figure 1. 1 - The evolutionary map of the genetic code. 6
Figure 1. 2 – The structure of the eukaryotic ribosome by crio-electron microscopy. 8
Figure 1. 3 – The ribosome translation elongation cycle. 11
Figure 1. 4 – tRNA secondary and tertiary structure. 15
Figure 1. 5 – Distribution of identity elements over the tRNA structure. 17
Figure 1. 6 – Cloverleaf structure of tRNA with the localization of modified nucleotides. 19
Figure 1. 7 – The aminoacylation reaction. 22
Figure 1. 8 – General structure of Class I aminoacyl-tRNA synthetases. 23
Figure 1. 9 - Structure of the Class II aminoacyl-tRNA synthetases. 23
Figure 1. 10 - Interaction of the two distinct classes of aaRSs with tRNA. 24
Figure 1. 11 – The two classes of aminoacyl-tRNA synthetases and their sub-classes. 26
Figure 1. 12 – The antiparallel map of Class I versus Class II aminoacyl-tRNA synthetases. 27
Figure 1. 13 – The class I and II synthetases complexes. 28
Figure 1. 14 – Alternative pathways for tRNA aminoacylation. 30
Figure 1. 15 – Pre- and post-transfer editing of the aminoacylation reaction. 31
Figure 1. 16 – Sense codon reassignment. 36
Figure 1. 17 – The Gain-Loss model. 37
Figure 1. 18 – The mitochondrial genetic codes. 39
Figure 1. 19 - The nuclear/cytoplasmatic genetic code alterations. 39
Figure 1. 20 – The synthesis of selenoproteins. 42
Figure 1. 21 – Pyrrolysine incorporation pathways. 44
Figure 1. 22 – The evolution of the CUG codon reassignment in Candida spp. 46
Figure 1. 23 – The phylogenetic tree of CUG decoding in Hemiascomycetes. 47
Figure 1. 24 – The secondary structure of the $tRNA_{CAG}^{Ser}$. 49
Figure 1. 25 – The mutational pressure on C. albicans' genome. 50

Figure 3. 1 – Errors in translation. 86
Figure 3. 2 – The -1 and +1 frameshifting. 88
Figure 3. 3 – Scheme of the C. albicans CUG reporter gene. 91
Figure 3. 4 – Reporter protein. 92
Figure 3. 5 – Reporter protein purification. 93
Figure 3. 6 - Reporter protein re-purification by FPLC. 94
Figure 3. 7 – In gel digestion of the purified reporter protein. 94
Figure 3. 8 – HPLC-MS of the Serine-peptide. 96
Figure 3. 9 – The reporter protein was phosphorylated in vivo. 97
Figure 3. 10 - HPLC-MS of the Leucine peptide. 98
Figure 3. 11 – Spectrum of an equimolar mixture of serine and leucine peptides 99
Figure 3. 12– Mistranslation due to near-cognate decoding 101
Figure 3. 13– Candida albicans morphology. 103
Figure 3. 14 – Engineered $tRNA_{CAG}^{Leu}$ *gene from S. cerevisiae.* 107
Figure 3. 15 – Leucine incorporation at CUG codons in vivo in engineered C. albicans cells. 107
Figure 3. 16 - CUG ambiguity is sensitive to environmental cues. 109
Figure 3. 17 – Leucine incorporation on highly ambiguous cell lines. 109

Figure 4. 1 – The impact of mistranslation on the cell biology. 115
Figure 4. 2– CUG codon distribution over C. albicans genome. 119
Figure 4. 3 – CUG codon context analysis. 120
Figure 4. 4 – Probability of synthesis of proteins without leucine at CUG codons. 123
Figure 4. 5 – Novel proteins generated through the ambiguous CUG decoding. 126
Figure 4. 6 – Usage of C. albicans CUG codons in genes with different CAI values. 127
Figure 4. 7 – Usage of S. cerevisiae CUG codons in genes with different CAI values. 127

Figure 4. 8 – Novel proteins generated through the ambiguous CUG decoding in engineered S. cerevisiae. 129
Figure 4. 9 – C. albicans codon usage. 130
Figure 4. 10 – SCU_{CUG} correlation with ORF size, rare codons and GC content. 132
Figure 4. 11 – CUG usage in individual C. albicans chromosomes. 134
Figure 4. 12 –The SCU distribution of serine codons in various classes of enzymes. 136
Figure 4. 13 – Cluster analysis of the CUG and AGC codons usage in the different alleles. 149

Figure 5. 1– LeuRS and SerRS protein expression under different physiological conditions. 158
Figure 5. 2 – SerRS and LeuRS expression 158
Figure 5. 3– SerRS/LeuRS expression ratio. 158
Figure 5. 4 – Ratio between the cleaved and the native LeuRS. 159
Figure 5. 5 – Single Nucleotide Polymorphism analysis. 160
Figure 5. 6 – Polymorphisms identified in the LeuRS from different strains of Candida albicans. 161
Figure 5. 7 – Polymorphisms identified in SerRS gene from different strains of Candida albicans. 162
Figure 5. 8 – Polymorphisms identified in the C. albicans TrpRS gene. 163
Figure 5. 9 – Polymorphisms identified in SerRS gene of S. cerevisiae. 164
Figure 5. 10 – C. albicans has a naturally high SNPs rate 164
Figure 5. 11 – Impact of polymorphic variation on the 3D structure of LeuRS 166
Figure 5. 12 – Phylogeny of the LeuRS isoforms 167
Figure 5. 13 – Polymorphic amino acid residue localization on the structure of the complex LeuRS-$tRNA^{Leu}$. 168
Figure 5. 14 – Model of the amino acid substitutions and their phylogeny. 168
Figure 5. 15 – CUG localization on the C. albicans LeuRS primary structure. 170
Figure 5. 16 – CUG localization on the C. albicans SerRS primary structure 171
Figure 5. 17 – CUG localization on SerRS terciary structure 171
Figure 5. 18 – Purification of the recombinant LeuRS isoforms 173
Figure 5. 19 – Purification of the recombinant SerRS 173
Figure 5. 20 – Total tRNA extracts. 175
Figure 5. 21 – tRNA purification by chaplet column chromatography 177
Figure 5. 22 – Monitoring tRNA purification by denaturing TBE-Urea acrylamide gel 177
Figure 5. 23 – tRNA charging with LeuRS and SerRS 178
Figure 5. 24 – Amino acid activation by LeuRS and SerRS active sites. 179

Figure 6. 1 – Model for the transcriptional control of LeuRS expression 190
Figure 6. 2 – Interactome of LeuRS and SerRS. 192
Figure 6. 3 – The localization of the $tRNA_{CAG}^{Ser}$ in the genome 193

List of Tables

Table 1. 1 - The universal genetic code. 3
Table 1. 2 – Examples of tRNA identity anti-determinants. 18
Table 1. 3 – Natural occurring misacylations. 33
Table 1. 4 . Variations in the mitochondrial genetic code. 38

Table 2. 1 – List of the oligonucleotides used. 57
Table 2. 2 – Original plasmids used for obtaining the necessary DNA constructions. 59
Table 2. 3 – Plasmids constructed in this work. 60
Table 2. 4 – Primary antibodies. 75
Table 2. 5 – DNA probes used for tRNA purification. 78

Table 3. 1 – Leucine incorporation at the CUG codon on white cells. 102
Table 3. 2 – Leucine incorporation at the CUG codon on opaque cells. 103
Table 3. 3 – Leucine incorporation at the CUG codon on cells grown at 37ºC. 104
Table 3. 4 – Leucine incorporation at the CUG codon in cells grown at pH 4.0. 105
Table 3. 5 – Leucine incorporation at the CUG codon on cells grown in the presence of 1.5 mM H_2O_2. 106
Table 3. 6 – Leucine incorporation at the CUG codon on highly ambiguous cells. 108

Table 4. 1 – Expansion of the C. albicans proteome through CUG ambiguity. 121
Table 4. 2– Probabilistic decoding of a gene with 3 CUG codons. 123
Table 4. 3 - Novel proteins produced by ambiguous decoding of mRNAs whose genes have high CAI value. 125
Table 4. 4- Novel proteins produced by ambiguous decoding of mRNAs whose genes have low CAI value. 125
Table 4. 5 – Relative serine-Specific Codon Usage 131
Table 4. 6 – Pearson correlation matrix 131
Table 4. 7 – ORF distribution over C. albicans chromosomes 134
Table 4. 8 – ORF distribution for the six enzyme classes, and the respective SCU_{CUG} average 135
Table 4. 9 – The p values of Scheffe's test for the SCU_{CUG} distribution in the 6 enzyme classes. 135
Table 4. 10 – CUG and AGC codons SCU in the enzymes sub-classes. 137
Table 4. 11 CUG and AGC codons SCU in protein domains 140
Table 4. 12 – CUG and AGC codons SCUs in ORFs grouped according to their cellular localization 142
Table 4. 13 – CUG and AGC codons SCUs in ORFs grouped according to their cellular process 144
Table 4. 14 – Variation of AGC and CUG codons between alleles 146
Table 4. 15 –ORFs with higher d(S) score but lower d(N) score. 147

Table 5. 1 – Overview of the protein fractions purified 174
Table 5. 2– Pure tRNA obtained through the purification process 177
Table 5. 3 – K_{cat} of SerRS isoforms. 179

1. Introduction

1.1. The genetic code

1.1.1. The standard genetic code

The genetic code established in the 1960s defines the rules that govern the transfer of genetic information from nucleic acids to proteins (Crick, 1970). In the early studies, Nirenberg and co-workers incubated RNA samples in cell-free extracts containing bacterial ribosomes, enzymes, ATP, tRNAs and both cold and [^{14}C]-labelled amino acids. They started by programming the cell free lysates with poly-U oligonucleotides and were able to synthesize poly-Phe peptides, hence indicating that the UUU codon coded for phenylalanine. Similar experiments using different RNA templates unveiled the other codon assignments (Table 1. 1) (Nirenberg et al., 1966; Nirenberg and Matthaei, 1961; Nirenberg and Leder, 1964).

Table 1. 1 - The universal genetic code.

1st Base	2nd base: U		C		A		G		3rd Base
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	*Stop*	UGA	*Stop*	A
	UUG		UCG		UAG		UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

A close analysis of the distribution of amino acids over the genetic code table revealed biased allocation of codons associated to amino acids polar properties. For example, all codons with U at the second position code for hydrophobic amino acids (Phe, Leu, Ile, Met and Val), and amino acids that share similar chemical properties, namely Leu, Ile and Val are connected by a single base mutation at the first codon base. Six of the most hydrophilic amino acids – His, Gln, Asn, Lys, Asp and Glu - have an A at the second

codon position; Tyr, which is hydrophobic, is the exception to this rule. (Woese, 1965a; Woese, 1965b; Woese et al., 1966; Volkenstein, 1966). As a result, amino acids that are decoded by complementary anticodons tend to have opposite hydrophobicities (Volkenstein, 1966; Blalock and Smith, 1984). In line with these observations, codons encoding amino acids with similar chemical properties tend to be related. For example, the acidic amino acids Asp and Glu belong to a split codon family and their amine derivates Asn and Gln belong to codon families that only differ in the first codon position. It is not yet clear why the genetic code evolved in such a manner. However, it is likely that its biased codon organization and redundancy may minimize decoding error, since most errors occur through near cognate insertion of amino acids with similar chemical properties, hence causing a minimal impact on protein structure.

1.1.2. The origin and early evolution of the genetic code

With few exceptions (sections 1.4.2 and 1.4.3), the same genetic code is used in all organisms. Such uniformity suggests that the extant genetic code must have provided important selective advantages over other codes that may have existed before the last common ancestor (Woese, 2002). Since the origin of the genetic code remains poorly understood, one does not yet fully comprehend the establishment of the standard code. Nevertheless, several theories have been proposed to explain its evolution.

(i) The Adaptation of the Genetic Code

This theory postulates that the genetic code has been gradually refined to minimize the impact of codon decoding error. It sprung from a large scale analysis of the relationship between genetic code redundancy and amino acids chemical properties (Alf-Steinberger, 1969). In his work, the extant genetic code was compared with 200 alternative codes and the impact of point mutations at different positions was tested using Monte Carlo simulations. A statistical approach used to estimate the distribution of error values in a large sample of alternative codes

directly estimated the probability of evolution without selection of codes with better or as good performance than the natural code. The data showed that almost no random codes could minimize polarity changes better than the canonical code. Indeed, the 3rd codon position was highly optimized relative to random codes, followed by the 1st codon position, but there was no evidence for optimization in the 2nd codon position. This is consistent with the relative effects of translation error (Alf-Steinberger, 1969). These results were put aside for over 20 years, but were reviewed in 1990s to highlight the highly optimized nature of the genetic code for polar requirements, rather than other amino acid characteristics, such as hydropathy, molecular volume or isoelectric point (Haig and Hurst, 1991). This has functional meaning since changing a non-polar for a polar amino acid, or vice-versa, would most probably destroy protein folding and structure and could be lethal.

Nevertheless, those studies failed to address differences in decoding error associated to the different bases. Since both mutation and mistranslation are highly biased for the 4 bases (Collins and Jukes, 1994; Kumar, 1996; Moriyama and Powell, 1997; Morton and Clegg, 1995; Friedman and Weinstein, 1964; Parker, 1989; Woese, 1965b), the data had significant noise. To overcome this, Freeland and Hurst extended the Haig and Hurst's Monte Carlo approach by incorporating known biological biases that influence both mutational patterns and mistranslation. Their approached showed that in 1 million of randomly generated codes only 1 performed better than the natural genetic code, thus the *"genetic code is one in a million"* (Freeland and Hurst, 1998; Freeland et al., 2003).

(ii) Co-Evolution of the Genetic Code

This theory, proposed by Wong, postulates that the organization of the canonical genetic code reflects evolutionary pathways of amino acids biosynthesis (Wong, 1975). Thus, the earliest genetic code used a small subset of pre-biotically synthesized amino acids (such as Gly, Ala and Ser), which were coded by an

extremely degenerated code. Then, it expanded by incorporating new metabolic derivatives of these primordial amino acids (Figure 1. 1) (Wong, 1975; Wong and Bronskill, 1979; Di Giulio and Medugno, 1999). Wong carried out a correlation analysis between codons distribution and amino acids biosynthetic pathways and proved the existence of a precursor-product relationship between them. This study was latter strengthen by Di Giulio's work, who improved the robustness of the correlation algorithm (Di Giulio, 1999). Indeed, the existence of molecular fossils with ancient codon assignments, such as the Asp-tRNA^{Asn} →Asn-tRNA^{Asn} and the Glu-tRNA^{Gln} → Gln-tRNA^{Gln}, in most bacteria and in all archea, and Sep-tRNA^{Cys} → Cys-tRNA^{Cys}, in methanogenic archea (Section 1.3.2.3), strongly support the co-evolutionary theory (Di Giulio, 2001b).

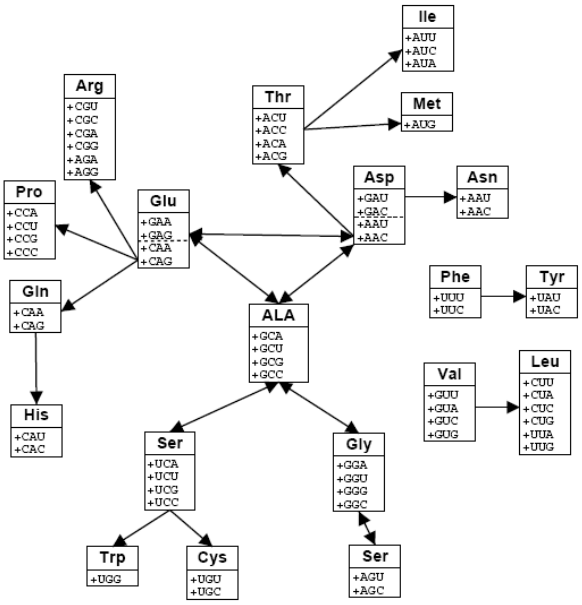


Figure 1. 1 - The evolutionary map of the genetic code.
Each box represents a single amino acid and its contemporary codons. The Glu and Asp enclosed in the dashed boxes were likely to be primitive codons assignments, required to create the relationships predicted by the coevolution theory. The single headed arrows show precursor-product relations, whereas double headed arrows indicate biosynthetic interconvertions. The arrow connected codons have a single base change (adapted from Wong, 1975).

(iii) The Steriochemical Origin of the Genetic Code

This hypothesis proposes that canonical codon assignments were originated through specific steric interacting ions between amino acids and their associated codons, so, primordial protein sequences were directly templated on base

sequences. Therefore, the actual complex translation mechanism, involving RNA and associated enzymes, is a late development (Yarus, 1998; Knight et al., 1999; Knight and Landweber, 2000).

The observation that led to this hypothesis came from *in vitro* selection amplification experiments (SELEX) using RNA-aptamers, which revealed that RNA molecules selected from random sequences that bind specific amino acids have more standard codons, anticodons or both for those amino acids than would be expected by chance. So far, a total of 43 RNA aptamers have been selected and isolated for specific binding of phenylalanine, isoleucine, histidine, leucine, glutamine, arginine, tryptophan and tyrosine (Caporaso et al., 2005; Yarus et al., 2005). Of these, research has been focused on the arginine binding aptamers because free arginine can mimic the natural interaction of HIV Tat peptides with TAR RNA (Tao and Frankel, 1992) and arginine aptamers have far more arginine codons at the binding site than the others (Knight and Landweber, 1998).

All these complementary theories focus on different characteristics of the genetic code, and they do provide important glimpses of the emergence and evolution of the standard genetic code (Knight et al., 1999; Di Giulio, 1999; Yarus et al., 2005). Nevertheless, the first theory explaining the origin of the genetic code was the *Frozen Accident Theory,* postulated by Crick in 1968 (Crick, 1968). This theory was a corner stone of the early days of molecular biology and postulated that the *"genetic code is universal because any change to it would be lethal or at least very strongly selected against"* (Crick, 1968). The theory assumed that once organisms with complex genomes encoding thousands of proteins were established, any change in the code would cause wide protein structure disruption, which would be lethal or highly detrimental. The robustness of this theory was shaken in 1979 (Barrell *et al.*, 1979) by the discovery of a genetic code change in human mitochondria, which involves decoding of the UGA stop codon as tryptophan. Since then, 16 alterations have been found in various organisms which put a definitive end to this theory.

1.2. Translation

The uprising of mRNA templated translation allowed for the transition from the "*RNA world*" into the "*Protein world*", which was an evolutionary breakthrough – as the 22 amino acids provided greater catalytic versatility than the 4 nucleic acids (Szathmary, 1999).

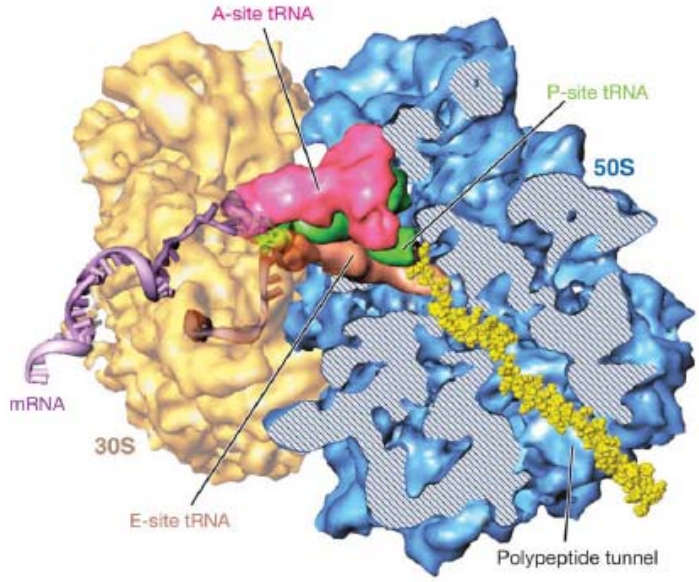


Figure 1. 2 – The structure of the eukaryotic ribosome by crio-electron microscopy.
Translation of the DNA/RNA genetic information into amino acid information is accomplished in the ribosome. The figure shows the small subunit (in orange) and the large subunit (blue) scanning an mRNA molecule (purple). The tRNAs are bound to the A-, P- and E-sites of the ribosome and the nascent polypeptide chain (yellow) is emerging through the polypeptide tunnel. Adapted from (Mitra and Frank, 2006).

The translational process, in particular the elongation and termination phases are rather conserved in the three kingdoms of life. This process relies on the existence of a translational machinery, composed by a large number of different molecules – mRNAs, tRNAs, amino acids, translational factors, rRNA, ribosomal proteins (RNP) and aminoacyl tRNA-synthetases (aaRS). Translation occurs at the ribosome (Figure 1. 2), a supramolecular complex composed of rRNA and proteins that contains three sites for binding tRNAs, namely the aminoacyl site (A site), peptidyl site (P site), and exit site (E site). It can be divided in three distinct stages: *initiation*, *elongation* and *termination*, which are briefly explained in this section.

1.2.1. Translation initiation

In the first stage of translation, the ribosome and mRNA are assembled in such a manner that the initiation codon (AUG) and the methionyl initiator tRNA bound are located in the P–site. This step requires help from initiation factors (IF). This step differs significantly between eukaryotes and prokaryotes (reviewed by Kapp and Lorsch, 2004), mainly because it is an important regulatory step of gene expression in the former, but not in the latter.

In prokaryotes, the 30S ribosomal subunit binds two initiation factors, IF1 and IF3. The IF1 binds over the A site of the 30S, thus preventing the initiator tRNA from binding to it, whereas the IF3 prevents the 30S and 50S subunits from premature assembly. The 30S-IF1-IF3 complex recruits the mRNA through base-pairing interactions between the 3'-end of the 16S rRNA and an mRNA sequence, named Shine-Delgano sequence, which is located 10 bases upstream the initiation codon. In the next step of initiation, the complex containing mRNA is joined by the ternary complex IF2•fMet-$tRNA_i^{fMet}$•GTP. Finally, this large complex combines with the 50S ribosomal subunit; and, simultaneously, the GTP bound to IF2 is hydrolyzed to GDP and Pi, which are released from the complex. Then, the three initiation factors are released and a functional 70S ribosome – the initiation complex, with the fMet-$tRNA^{fMet}$ in the P site and an empty A site – starts elongation.

In eukaryotes, translation initiation is more complex than in prokaryotes and archea. The translation initiation begins with formation of a eIF2•GTP•Met-$tRNA_i$ ternary complex, which binds to the 40S ribosomal subunit with help of eIF1, eIF1A and eIF3. This results in the formation of a 43S complex. Meanwhile, the eIF4F complex, which includes the factors eIF4E, eIF4G, and eIF4A, is assembled on the 5'-cap structure of the mRNA. In this complex, the eIF4A, which has RNA helicase activity, unwinds secondary structure found on the 5'-untranslated region (UTR), while eIF4G binds both the eIF4E and the poly(A) binding protein (PBP), which is bound to the 3'-poly(A) tail of the mRNA. Indeed, the eIF4F complex effectively ties together the 5'- and the 3'-ends of the mRNA (Gingras *et al.*, 1999). Then the 43S complex is loaded onto the mRNA, with the help of

eIF3, eIF4F and PBP, and starts scanning down the mRNA looking for the AUG initiation codon, which signals the beginning of the open reading frame (ORF). Once this codon is found, the GTP of the eIF2•GTP•Met-tRNA$_i$ ternary complex is hydrolysed, by eIF2 with the help of eIF5, hence promoting the release of the Met-tRNA$_i$ into the P-site and dissociation of eIF2•GDP along with other initiation factors. Then the complex eIF5B•GTP promotes the joining of the 60S ribosomal subunit to the Met-tRNA^i•mRNA•40S ternary complex, in a process that requires the GTP hydrolysis by eIF5B, which is subsequently released as an eIF5B•GDP complex (Pestova et al., 2000; Lee et al., 2002). So the 80S ribosome is assembled and ready to proceed with protein synthesis.

1.2.2. Translation elongation

In the second phase of translation (Figure 1. 3), the ribosome moves along the mRNA, towards its 3'-end, assembling amino acids into polypeptides by reading codons. It requires a group of proteins termed elongation factors (EF) – EF-Tu in prokaryotes, or eEF1A in eukaryotes – that participate both in recruitment of aminoacyl-tRNAs (aa-tRNAs) for ribosome decoding and in subsequent translocation of the ribosome as it moves along the mRNA. It is critical for the translational accuracy that only the tRNAs charged with their cognate amino acid are recognized by the elongation factors, which are able to discriminate. In prokaryotes, the EF-Tu•GTP binds all the correctly aminoacylated tRNAs with about the same affinity, hence obeying the thermodynamic compensation rule (LaRiviere et al, 2001).

At this stage, the ribosome selects aa-tRNAs that are delivered to its A-site as a ternary complex – EF-Tu•aa-tRNA•GTP or eEF1A•aa-tRNA•GTP – through cognate codon-anticodon interactions. This process represents a critical point in translation, and is achieved in two stages, separated by the irreversible hydrolysis of GTP from the ternary complex (Thompson and Stone, 1977; Ruusala et al., 1982).

During initial selection, a charged tRNA is presented to the ribosome A-site, where it is tested for cognate codon-anticodon pairing. At this stage, ternary complexes with noncognate anticodons rapidly dissociate without GTP hydrolysis (Pape et al., 1999; Pape et al., 2000). Cognate codon-anticodon pairing stabilizes the ternary complex on the ribosome and stimulates GTP hydrolysis, which promotes a conformational change and its subsequent dissociation, with the release of EF-Tu•GDP or eEF1A•GDP (Gromadski and Rodnina, 2004; Rodnina and Wintermeyer, 2001b; Rodnina and Wintermeyer, 2001a; Valle et al., 2003; Ogle and Ramakrishnan, 2005).

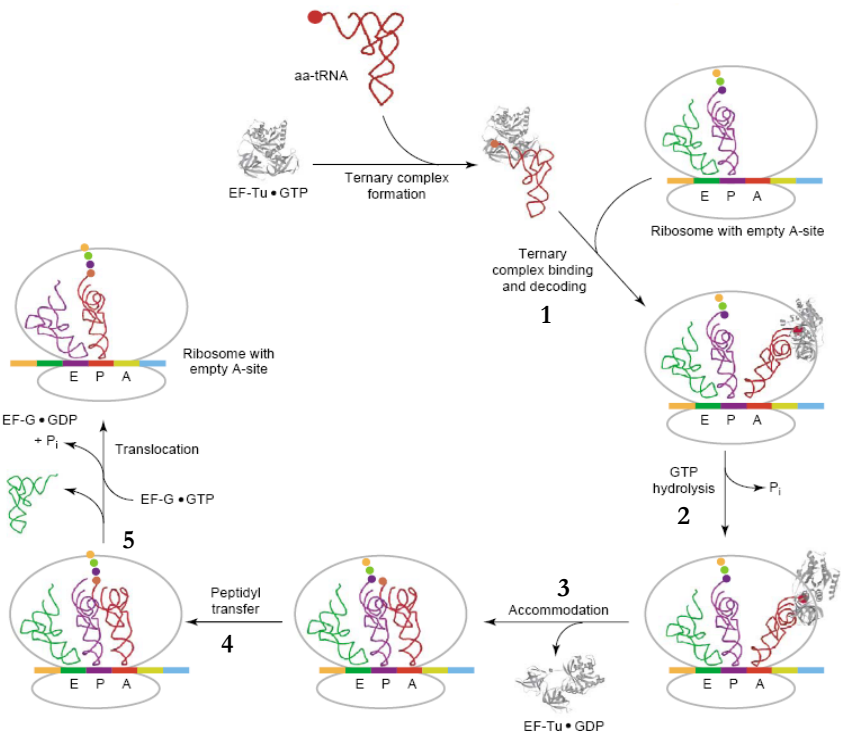


Figure 1. 3 – The ribosome translation elongation cycle.
The aa-tRNA forms a ternary complex with elongation factor Tu (EF-Tu) and GTP, and binds to the A-site of the ribosome (1). Correct codon–anticodon base pairing between the A-site mRNA codon and the tRNA anticodon activates the GTPase activity of EF-Tu and so the GTP hydrolysis occurs (2). Then a conformational change is induced in EF-Tu, resulting in its release from the aa-tRNA and enabling the acceptor end of the aa-tRNA to move into the A-site (3). After accommodation, the growing polypeptide esterified to the P-site-bound tRNA is transferred to the A-site-bound tRNA, elongating the peptide chain by one amino acid (4). With the aid of elongation factor G (EF-G), the deacylated P-site tRNA is then translocated to the E-site, and the A-site-bound tRNA is translocated to the P-site (5). The ribosomal A-site is then available for binding to the next ternary complex (Adapted from Dale and Uhlenbeck, 2005).

Positive discrimination of cognate aa-tRNA is further enhanced by a geometrical accommodation in the decoding site. As non-canonical codon-anticodon base pairing leads

to steric clashes, the geometry adopted by the ribosome is an effective criterion for positive discrimination of cognate aa-tRNA. During the aa-tRNA selection step, the ribosome changes its conformation from an open to a close state. In the open state, which is favoured when the A site is empty or bears a near-cognate anticodon, the ribosome is inactive for tRNA selection, whereas in the closed state, thus bearing the cognate anticodon, the rates of both GTPase activation and accommodation are accelerated. This geometric argument is reinforced by the finding that some antibiotics, such as paromomycin, force the ribosome to switch from the open to the closed conformation increasing error rate. The latter is due to increased acceptance of near-cognate aa-tRNAs. In other words, this conformational change is critical to maintain translational accuracy (Ogle et al., 2002; Ogle et al., 2003; Ogle and Ramakrishnan, 2005). This argument, also explains why the presence of a tRNA on the E-site lowers the affinity of the A-site and, consequently, increases the accuracy of selection of cognate anticodons (Nierhaus, 1990). Indeed the E-site tRNA makes contacts with both small and large ribosomal subunits and its presence increases the energetic cost of transition between the open and the closed states of the ribosome, increasing accuracy (Ogle *et al.*, 2002).

Once the aa-tRNA is accommodated, the ribosome peptidyl transferase center catalyses the formation of the peptide bond between the incoming aminoacyl residue, attached to the tRNA at the A-site, and the nascent peptidyl chain, which is attached to the tRNA at the P-site. At this stage, both tRNAs adopt an hybrid conformational state on the ribosome: the tRNA at the P-site is deacetylated, with its acceptor end at the E-site of the large subunit and its anticodon in the P-site of the small subunit; whereas the newly formed peptidyl-tRNA has its acceptor end in the P-site of the large subunit, while its anticodon is still in the A-site of the small subunit. Such movements of the acceptor ends of tRNA, on the large subunit of the ribosome, occur spontaneously and immediately after the formation of the peptide bond, and thus independently of the anticodon (Noller *et al.*, 2002).

The elongation cycle is completed by the movement of the mRNA–tRNA complex on the ribosome, in a process called translocation, catalyzed by the complex EF-G•GTP, in prokaryotes, or eEF2•GTP, in eukaryotes, at the expenses of the energy from the GTP hydrolysis. During translocation, the anticodon ends of the tRNAs and the mRNA move

along the small ribosome subunit, thus the deacetylated tRNA is displaced from the P-site to the E-site and then released from the ribosome; whereas the newly formed peptidyl-tRNA is displaced from the A-site to the P-site, hence resulting in an empty A-site, which is ready to accommodate a new aa-tRNA on the next round of elongation (Rodnina et al., 2002; Rodnina et al., 1999; Noller et al., 2002; Kapp and Lorsch, 2004).

1.2.3. Translation termination

Termination of protein synthesis is initiated when one of the three stop codons is present in the ribosome A-site. This step involves decoding of a STOP codon through an interaction between RNA (rRNA and mRNA) and proteins (release factors) and facilitates the hydrolytic release of the nascent polypeptide chain from the peptidyl-transferase centre of the ribosome. The release factors (RFs) are split in two classes: the class-I proteins recognize the STOP codons in the mRNA and the class-II proteins interact with class-I RFs and have GTPase activity. Prokaryotes have two class-I RFs with overlapping specificity: RF1 (specific for UAG and UAA) and RF2 (specific for UGA and UAA), whereas eukaryotes only have one factor, eRF1, which recognizes the three STOP codons. The class II RFs are RF3 and eRF3, in prokaryotes and eukaryotes, respectively (reviewed by Nakamura et al., 1996; Buckingham et al., 1997).

Several models have been proposed to explain the molecular mechanism of translation termination, and although there is a consensus about the termination elements, the order by which the events occurs is still open for debate (Freistroffer et al., 1997; Zavialov et al., 2001; Peske et al., 2005). In prokaryotes, the better accepted model proposed for termination posits that once a stop codon is recognized by RF1 or RF2, the ester bond between the nascent polypeptide and the tRNA at the P-site is hydrolysed, leading to the release of the polypeptide chain from the ribosome (Zavialov *et al.*, 2001). This originates a post-termination ribosome complex containing deacylated tRNA bound on the mRNA at the P-site and an empty A-site. Then, RF3 promotes rapid dissociation of RF1 or RF2 from the ribosome, in a GTP-dependent manner (Freistroffer *et al.*, 1997). Afterwards, the ribosomes, along with the tRNA and mRNA, are released from the post-termination complex by the concerted action of EF-G, RF3 and the ribosome recycling

factor (RRF), leaving these components available for a new round of translation (Peske *et al.*, 2005).

1.3. The operational RNA code

Accurate translation relies on the highly discriminating properties of the ribosome A-site. Most tRNAs that enter in the A-site fail to form three base pairs with the displayed codon and the tRNA rapidly dissociates. Therefore, in this process only cognate tRNAs are efficiently retained.

Nevertheless, the ribosome does not check whether tRNAs are correctly charged (Prather *et al.*, 1984) and, consequently, translation accuracy strongly relies on aminoacylation specificity. Indeed, the accuracy in the genetic code is ensured by an operational RNA code – the *"second genetic code"* – that correlates amino acids to specific structural features located in tRNAs structure and is imprinted in aaRSs structure (De, 1988; Schimmel et al., 1993).

1.3.1. Transfer RNAs

The existence of an adapter molecule that would carry an amino acid and interact with messenger RNA, playing a central role in translation, was first hypothesized by Crick (Crick, 1955): *"there would be 20 different kinds of adaptor molecules, one for each amino acid, and 20 different enzymes to join the amino acids to their adaptors"*. This theory proved to be correct, with the exception that there are more than 20 different tRNAs, which can be grouped in families of isoacceptors. Isoacceptors are tRNAs that, despite having different mRNA codon selectivity, are recognized by a single aaRS that charges them with their cognate amino acid. Since their discovery in the early 1970s, up to 5,800 different tRNA molecules have been identified in organisms belonging to the three domains of live (Sprinzl and Vassilenko, 2005). tRNAs have invariant and semi-invariant nucleotides (Figure 1. 4), though some tRNAs have atypical structures displaying variation at conserved positions.

1.3.1.1. Structure of tRNAs

The secondary structure of tRNAs was first predicted by Holley and co-workers (Holley, 1965). Comparative sequence analysis allowed them to identify invariant nucleotides and to define a cloverleaf secondary structure. The canonical cloverleaf (Figure 1. 4) consists of three stem-loop regions, a variable region, a terminal stem and a 3' single stranded N-C-C-A_{OH} end, to which the amino acids become attached. The tRNAs are clustered in two families – class-I and class-II, according to the length of their variable region. The class-I comprises the majority of tRNAs, which are characterised for having short variable loops of four or five nucleosides. Class-II tRNAs have longer variable arms of 10 to 24 bases and belong to leucine and serine amino acid families in eukaryotes and leucine, serine and tyrosine in bacteria and organelle translation systems (Dirheimer *et al.*, 1995a).

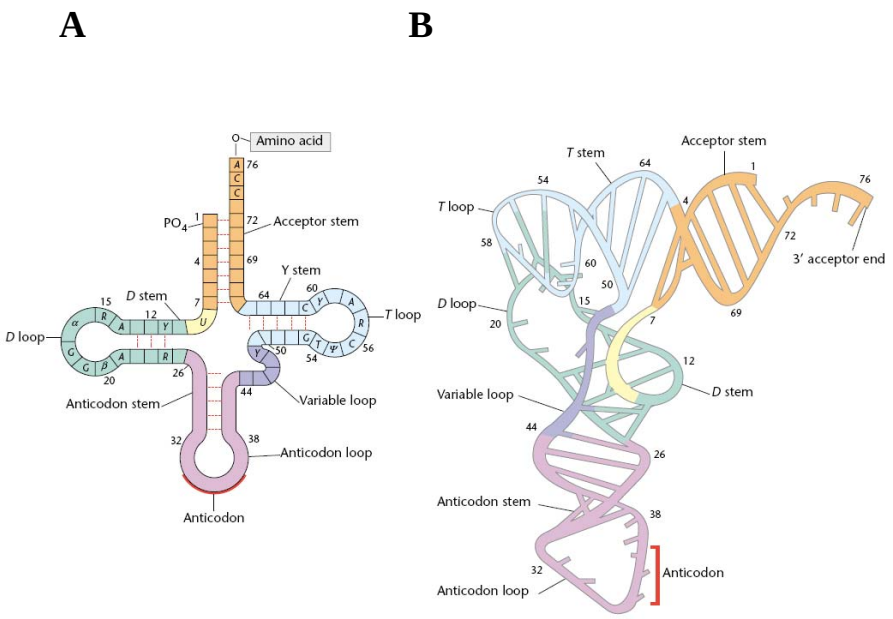


Figure 1. 4 – tRNA secondary and tertiary structure.
(A) Diagram showing the cloverleaf structure of tRNAs. The conserved nucleotides are indicated. The stems can be related to their different domains according to size : the acceptor stem is the longest with seven base pairs; both the TψC and the anticodon stems have five base pairs; and finally, the D stem has three or four base pairs, in class I and class II tRNAs, respectively. **(B)** L-shaped tertiary structure of tRNAs, representing the special location of its stems and loops.

An interesting feature of tRNA structure is the formation of non-canonical base pair interactions, of which the G•U wobble pairing is the most frequent, though there are more non-Watson-Crick interactions, such as A•A, C•C, C•U, G•A, U•U and U•Y (Grosjean *et*

al., 1982). The cloverleaf, in turn, assumes a L-shaped three-dimensional structure, where the D-arm is stacked onto the anticodon-arm and the TψC-arm is stacked onto the anticodon-arm and the acceptor stem, thus defining two distinct functional domains. The conserved and semi-conserved residues play a critical role in forming and maintaining the L-shaped structure, as the R15:Y48 tertiary interaction, known as *Levitt base pair*. This base pair stabilizes the stacking of the D-arm with the TψC- stem and keeps the D- and variable loops together (Levitt, 1969; Hou et al., 1993).

These distinct structural domains had independent origins. Indeed, they bind to different domains of aaRSs and the TψC-acceptor minihelix functions as an independent unit. In fact, this minihelix can be recognized and charged by aaRSs and recognized by the elongation factor EF-Tu (Schimmel and Ribas de, 1995). This suggests that the TψC-acceptor minihelix is an ancient structure, upon which the early genetic code might have relied on, whereas the D- and the anticodon arms are late acquisitions (Noller, 1993).

1.3.1.2. Identity Elements

There are twenty different aminoacylation systems, one for each amino acid and tRNA family. Since tRNAs are broadly similar in structure, the accurate discrimination between them is a challenge to the aminoacyl-tRNA synthetases. To overcome this problem, tRNAs contain certain structural elements, called identity determinants, which directly interact with the enzymes (Figure 1. 5). However, such identity determinants have varied slightly during evolution and the recognition system of tRNA families is sometimes different among different organisms.

In many cases, specific tRNA-protein interactions occur in the anticodon but in other cases the variable arm and the acceptor stem are also involved in tRNA recognition (Kim *et al.*, 2000). Since anticodon nucleotides interact directly with codon nucleotides during translation, they were the first to be considered as key elements for tRNA recognition by the aaRSs. Indeed, they play major roles in recognition of most of the tRNAs in both *E. coli* and *S. cerevisiae*. Actually, in *E. coli* only the $tRNA^{Leu}$, $tRNA^{Ser}$ and $tRNA^{Ala}$

families do not contain identity elements in the anticodon. These families decode six or four codons – the $tRNA^{Leu}$ decodes CUN and UUR codons, the $tRNA^{Ser}$ decodes AGY and UCN codons and $tRNA^{Ala}$ decodes GCN codons – therefore, have different isoacceptors tRNAs with different anticodons, which complicates recognition of the anticodons by the respective aaRSs.

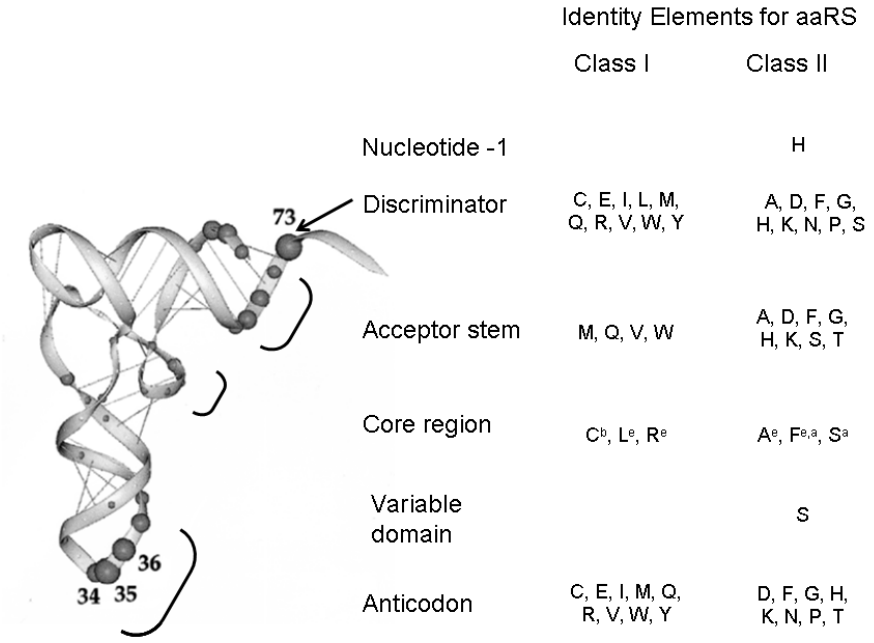


Figure 1. 5 – Distribution of identity elements over the tRNA structure.
The tRNA identity elements are distributed over four main features of the tRNA structure: the discriminator base, the acceptor stem, the core region and the anticodon-loop. The involvement of each feature in tRNA recognition by either class I or class II aaRSs is indicated. Apart from these, the variable arm is a key player for Ser identity, whereas the -1 nucleotide is important for His identity (adapted from Giege *et al.*, 1998).

The acceptor stem also contains a significant number of identity determinants, mainly in the first three base pairs – N_1-N_{72}, N_2-N_{71} and N_3-N_{70} – and the unpaired nucleotide N_{73} (Figure 1. 5). The latter is known as the "discriminator base", as it contributes to the identity of virtually every tRNA species (Normanly and Abelson, 1989; Lee et al., 1993; McClain et al., 1990; McClain, 1993). Each tRNA family has its own discriminator base and most tRNAs accepting chemically similar amino acids are characterized by an identical, phylogenetically well-conserved residue at this position (Crothers *et al.*, 1972). The importance of this base for tRNA recognition is highlighted in human leucine tRNAs where A_{73} to G_{73} mutation changes its identity to serine (Breitschopf and Gross, 1994).

The importance of the acceptor stem for tRNA aminoacylation has been extensively studied through aminoacylation of both acceptor-TψC stem minihelices and acceptor stem microhelices, which have proven to be, just by themselves, substrates for aminoacylation. For example, both minihelices and microhelices from alanine tRNAs are efficiently charged with alanine, provided that they contain the G_3-U_{70} base pair, which is the identity determinant for alanine (Francklyn *et al.*, 1992). The charging of specific RNA helices has been demonstrated with at least 11 different aminoacyl-tRNA synthetases, even for cases where the anticodon is known to play a significant role in the cognate tRNA recognition (Frugier et al., 1994; Hou et al., 1995; Quinn et al., 1995; Saks and Sampson, 1995), again, these studies demonstrate that there is an operational code embedded in the tRNA structure.

Table 1. 2 – Examples of tRNA identity anti-determinants.

Antideterminants	tRNA	aaRS Type
Lysidine 34 (modified C)	$tRNA^{Ile}$ (E. coli)	*MetRS*
U_{34}	$tRNA^{Ile}$ (E. coli)	*MetRS*
A_{36}	$tRNA^{Arg}$ (E. coli)	*TrpRS*
G_{37}	$tRNA^{Ser}$ *(yeast)*	*LeuRS*
m^1G_{37} *(methylated G)*	$tRNA^{Asp}$ *(yeast)*	*ArgRS*
A_{73}	$tRNA^{Ser}$ *(human)*	*SerRS*
G_3-U_{70}	$tRNA^{Ala}$ *(yeast)*	*ThrRS*
U_{30}-G_{40}	$tRNA^{Ile}$ *(yeast)*	*GlnRS, LysRS*

Interestingly, in addition to the positive identity elements present in a tRNA structure, which direct specific interactions with cognate synthetases, there are also negative elements, called anti-determinants, which contribute to the tRNA identity by blocking the recognition by other non-cognate synthetases (Table 1. 2). Such antideterminants can be modified or unmodified nucleotides at any structural domain of the tRNA. Several examples are known, but two of them are of special interest – *(i)* the lysidine residue (a modified C) at position 34 of the $tRNA^{Ile}$ acts as an anti-determinant for the MetRS, since the $tRNA^{Ile}$ recognizes AUA/U/C codons, whereas the $tRNA^{Met}$ recognizes the AUG codon (Muramatsu *et al.*, 1988); and *(ii)* the Leu/Ser recognition

system, where the A_{73} protects the $tRNA^{Leu}$ against the SerRS, whereas the G_{73} protects the $tRNA^{Ser}$ against the LeuRS (Breitschopf et al., 1995; Soma et al., 1996).

1.3.1.3. Modified bases

tRNAs are the most extensively modified nucleic acids in eukaryotes, prokaryotes and archea (Sprinzl et al., 1998; Woese et al., 1990). Base modifications are introduced post-transcriptionally and improve the specificity and efficiency of tRNAs, as they are involved in codon recognition and act as identity determinants for cognate aminoacylation (Yokoyama et al., 1985; Bjork, 1995; Agris, 2004).

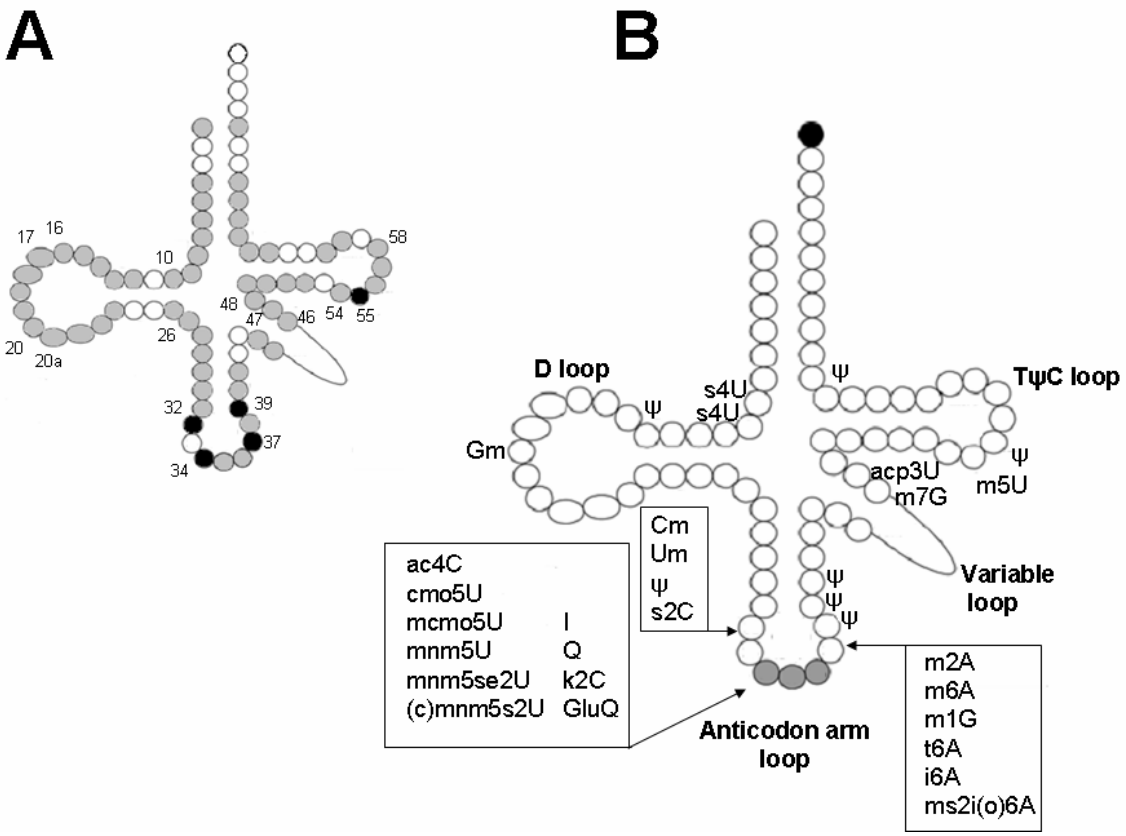


Figure 1. 6 – Cloverleaf structure of tRNA with the localization of modified nucleotides.
(A) Distribution of modified nucleotides in tRNAs from 546 tRNA sequences. In white are nucleotides for which no modification has yet been reported, in grey are nucleotides for which there is at least one modification in one tRNA, and finally, in black are those nucleotides for which more than 5 different modifications have been detected in the analysed tRNAs. The numbered positions are those where more than 25% of the nucleotides are modified. **(B)** Modified tRNA nucleotides found in *E. coli.* The positions 32, 34 and 37 are known to contain hypermodified nucleotides. Adapted from (Dirheimer et al., 1995b; Auffinger and Westhof, 1998)

Indeed, more than eighty modified nucleotides have been found in tRNAs and some of them are conserved in the 3 domains of life, as the dihydrouridine (D) in D-loops or ribothymidine in T-loops (Bjork *et al.*, 1999). The modified nucleotides can be found over 61 different positions on the tRNA (Figure 1. 6), however, the richest domain is the anticodon loop, especially the first anticodon position (N_{34}) and position 3´ to the anticodon triplet (N_{37}).The anticodon region is also the only structural domain that contains hypermodified bases, namely the guanosine derivatives wybutosine which is found at position 37 in almost all eukaryotic phenylalanine tRNAs, and queueosine (Q) at position 34 of Tyr, His, Asn and Asp tRNAs from prokaryotes and eukaryotes (Yokoyama *et al.*, 1985). Regarding minor modifications, such as methylation and acetylation, they are evenly distributed over the entire tRNA structure.

The modified bases at position 34 can either extend or restrict the decoding properties of tRNAs, for instance, inosine (I) (an adenosine derivate) permits base pairing with U, A and C; and the hypermodified Q pairs with all four nucleotides (A, U, C, G) (Yokoyama *et al.*, 1985). Concerning the modified bases at position 37, they seem to strengthen the base pairing between the last base of the anticodon (position 36) and the first base of the codon, as is the case of isopentenyl adenosine (i^6A) in tRNAs that read codons starting with U. In this case, i^6A improves A_{36}-UXX interaction and prevents base pairing of A_{36} with other bases (Bjork, 1995). Nevertheless, the most conserved modified residues in position 37 are m^1G in tRNAs that decode codons starting with C, and the t^6A in tRNAs that decode codons starting with A. The existence of these conserved modified residues points towards an important function for base modifications since they appeared early during the evolution of life (Bjork, 1995).

While modified bases in several positions do not have a significant influence on aminoacylation efficiency, certain modifications on the anticodon do lead to a change in tRNA conformation and play an important role in codon recognition (by both the aaRSs and the ribosomes) (Li *et al.*, 1997). For example, in *E. coli* the modification of cytidine to lysidine (k^2C) at position 34 in the two isoleucine tRNAs is sufficient for identity, and also prevents misacylation with methionine and alters decoding properties since k^2C pairs with A rather than G (Muramatsu *et al.*, 1988).

Finally, modified bases play an important role in the evolution of genetic code alterations. For example, decoding of the UGA stop codon as tryptophan in mitochondria is due to loss of its recognition by RF2 combined with a mutation in the anticodon of $tRNA^{Trp}$ that changed 5´-CCA-3´anticodon to 5´-U*CA-3´, where U* is the modified base 5- carboxymethyl-aminomethyl U ($cmnm^5U$). The 5´-U*CA-3´ anticodon pairs only with purines and hence it decodes both the tryptophan UGG by wobble and the stop UGA by Watson-Crick base pairing (Tomita *et al.*, 1999). In the case of animal mitochondria, the $tRNA^{Met}$ contains a modified base at position 34 – f^5C in vertebrates and nematodes, and $cmnm^5C$ in ascidia, and thus is able to decode both AUG and AUA. (Moriya et al., 1994; Watanabe et al., 1994; Kondow et al., 1999).

1.3.2. Aminoacyl-tRNA synthetases

The correct charging of tRNA with cognate amino acids is catalysed by aminoacyl-tRNA synthetases (aaRS), which recognize both the amino acid and the tRNA via its imprinted RNA code. In contrast to the standard genetic code, the operational RNA code is not degenerated, since there is only one aaRS for each amino acid. The aaRSs are enzymes from the 6.1.1 class, which have been exhaustively studied, so both their structure and mechanism are well documented.

The aminoacylation reaction is a highly specific two step reaction (Figure 1. 7). The first step involves the formation of aminoacyl adenylate, which is an enzyme-bound intermediate, resulting from the specific binding of the amino acid and its activation through a reaction with ATP:Mg^{2+}, with release of pyrophosphate. In the second step, the 3' terminal adenosine of the enzyme-bound tRNA reacts with the aminoacyl adenylate intermediate, leading to both esterification of the tRNA and the release of AMP.

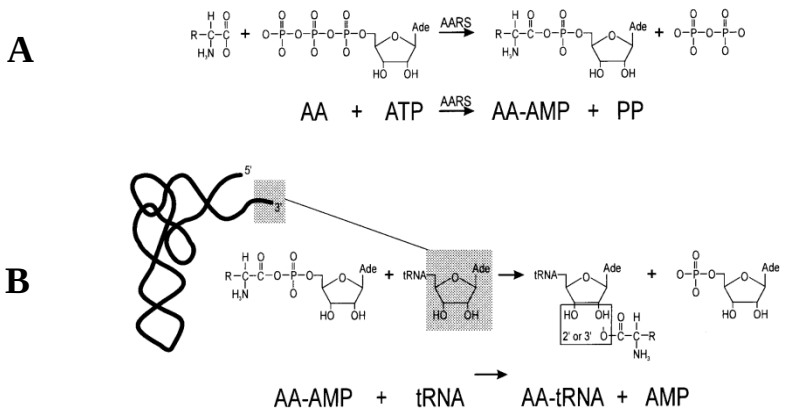


Figure 1. 7 – The aminoacylation reaction.
The aminoacylation reaction is achieved in two steps. **(A)** The amino acid is activated by attacking a molecule of ATP at the [alpha]- phosphate, giving rise to a mixed anhydride intermediate-aminoacyl-adenylate and inorganic pyrophosphate. **(B)** The amino acid moiety is transferred to the 3'-terminal ribose of the cognate tRNA, yielding an aminoacyl-tRNA and AMP

1.3.2.1. Classes of Aminoacyl-tRNA synthetases

The aaRS can be grouped in two classes – class I and class II – based in the conserved sequence motifs and structural architecture of the catalytic domains of the enzymes (Lenhard et al., 1997; Eriani et al., 1990; Cusack et al., 1990) (Figure 1. 8, Figure 1. 9, Figure 1. 11). This class division is very rigid, mutually exclusive (each enzyme can be classified as belonging to only one group) and inter-changes between classes are not possible. However, the lysyl-tRNA synthetase (LysRS) is an exception to this rule, since in some organisms it is a class I, while in others it is a class II enzyme. For example, in some archea, namely *Methanococcus maripaludis*, *Methanobacterium thermoautotrophicum* and *Methanococcus jannaschii,* and in some bacteria, namely in *Borrelia burgdorferi* and *Treponema pallidum*, belongs to the class I, whereas in all the other organisms from all the kingdoms of live it belongs to class II enzymes (Ibba et al., 1997b; Ibba et al., 1997a).

Class I enzymes comprise ArgRS, CysRS, GluRS, GlnRS, IleRS, LeuRS, MetRS, TrpRS, TyrRS and ValRS, and are characterized by a Rossman nucleotide-binding fold, consisting of alternating β-strands and α-helices, responsible for adenylate synthesis (Figure 1. 8). In these proteins the active site fold is divided in two halves linked by a polypeptide of variable length, designated as connective polypeptide 1 (CP1) (Starzyk *et*

al., 1987). Indeed, this insertion may form an editing domain and contains residues for binding the synthetase to the tRNA acceptor helix (Rould *et al.*, 1989). The Rossman fold is further characterized by two additional sequence motifs, namely an 11-amino acid element, which ends in the sequence His–Ile–Gly–His, known as the HIGH signature sequence, located in the first half of the nucleotide-binding fold, between the end of the first β-strand and the beginning of the first α-helix; and a KMSKS motif, located in the second half of the nucleotide-binding fold. (Delarue and Moras, 1993).

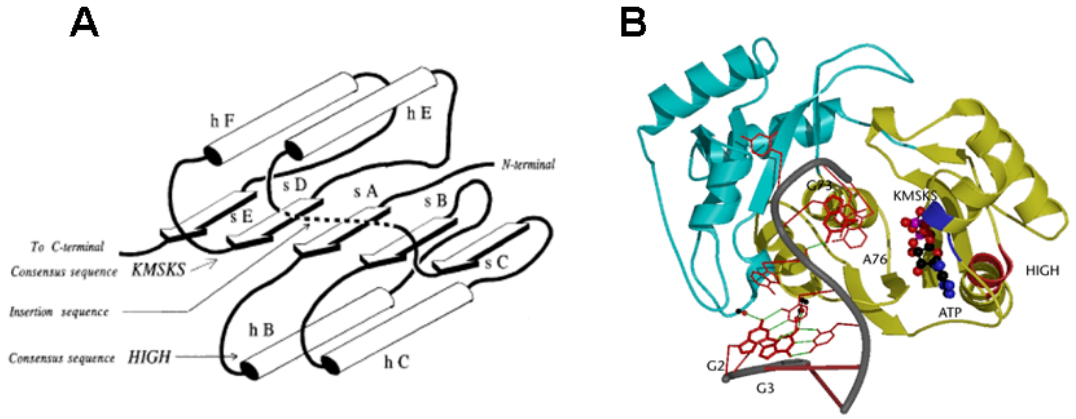


Figure 1. 8 – General structure of Class I aminoacyl-tRNA synthetases.
(A) Cartoon representing the structure of class I aaRSs, with the KMSKS and HIGH signatures. **(B)** The structure of the class I GluRS, complexed with the acceptor arm of its cognate tRNA. The Rossman fold is in yellow with the characteristic motifs HIGH and KMSKS, which are highlighted in red and dark blue, respectively. Adapted from (Moras, 1992; Arnez and Moras, 1997).

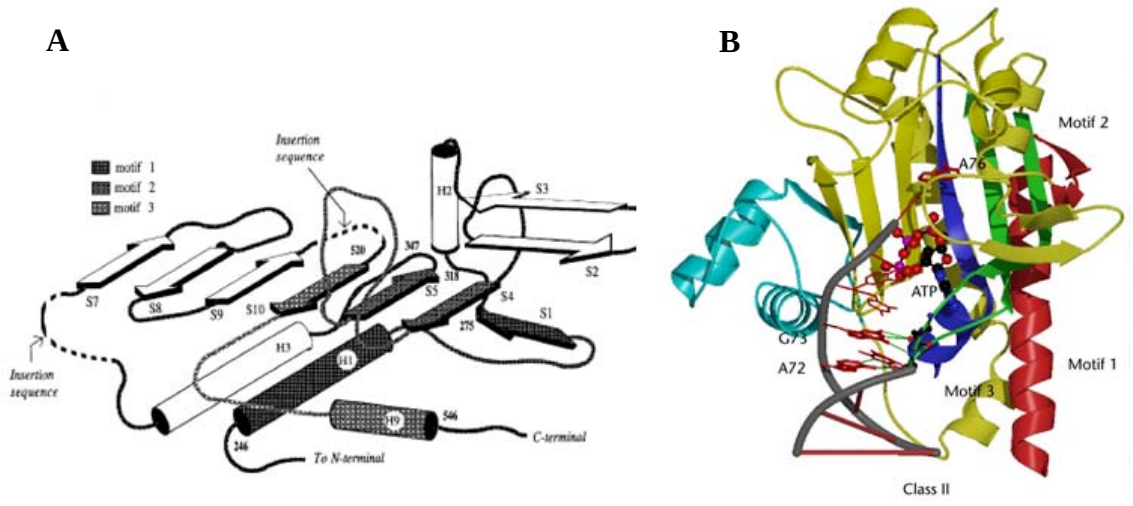


Figure 1. 9 - Structure of the Class II aminoacyl-tRNA synthetases.
(A) Cartoon representing the structure of class II aaRSs, with the motif 1, 2 and 3. **(B)** The structure of the class I AspRS, complexed with the acceptor arm of its cognate tRNA, with the characteristic motifs 1, 2 and 3 highlighted in red, green and dark blue, respectively. Adapted from (Moras, 1992; Arnez and Moras, 1997).

The class II enzymes are AlaRS, AsnRS, AspRS, GlyRS, HisRS, LysRS, PheRS, ProRS, SerRS and ThrRS (Mechulam et al., 1995; Woese et al., 2000), characterized by seven-stranded antiparallel β-sheet flanked by three α-helices (Figure 1. 9). The active site is formed by three conserved motifs known as motifs 1, 2, and 3, consisting of a N-terminal helix–loop–strand, a central strand–loop–strand, a C-terminal and strand–helix, respectively, whose sequence is highly degenerate (Eriani et al., 1990; Cusack et al., 1990).

However, the differences between the enzymes belonging to each class go beyond the secondary and tertiary structures. They also differ on their quaternary structure, as the class I synthetases are predominantly monomers, with the exception of TrpRS and TyrRS, while the class II synthetases are obligate homo or heterodimers, whose interface is established by the conserved motif 1 and is required for the integrity of their active site.

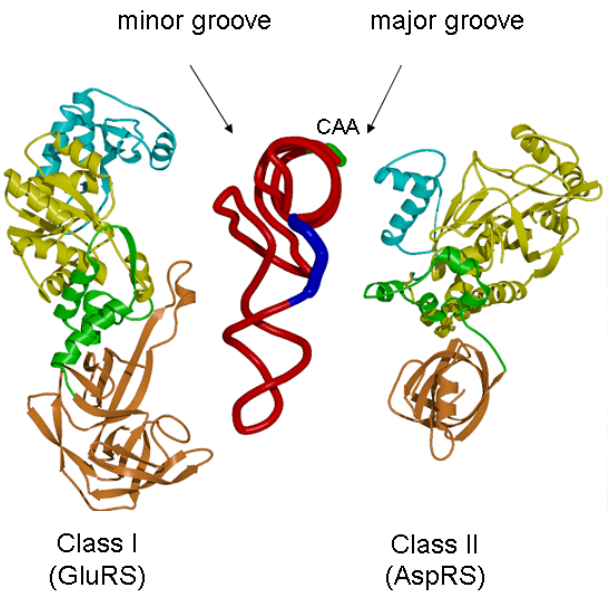


Figure 1. 10 - Interaction of the two distinct classes of aaRSs with tRNA.
A class I synthetase is represented on the left and a class II synthetase on the right. The mirror-symmetrical interaction with the tRNA (on the centre) is highlighted. Adapted from (Moras, 1992; Arnez and Moras, 1997).

The class partitioning is further manifested mechanistically in the two steps of the aminoacylation reaction. During the first step, the conformation of ATP bound to the class I and class II enzymes is different – in class I synthetases the ATP is in a straight

conformation, whereas in class II synthetases the ATP is positioned in a bent conformation. Also, during the second step of the reaction, while in class I enzymes, the aminoacyl group is transferred to the 2'-hydroxyl group of the terminal adenosine of the tRNA and then moved to the 3'-hydroxyl by a trans-esterification reaction; in class II enzymes the aminoacyl group is directly loaded on the 3'-hydroxyl of the terminal adenosine. These differences in the reaction mechanisms are a direct consequence of the manner that aaRSs use to bind tRNA. Class I aaRSs bind the tRNA minor groove, and class II aaRSs recognize its major groove (Figure 1. 10) (Ruff et al., 1991; Moras, 1992).

An analysis of the sequences and structures of synthetases have also shown that these enzymes can be further divided into three subclasses – *a*, *b*, and *c* – that share homologous anticodon binding modules (Figure 1. 11) (Cusack, 1995). So, synthetases of the same subclass are more similar to each other than to members of other subclasses. Class I*a* contains enzymes that recognize hydrophobic (Ile, Leu and Val) and sulphur-containing residues (Met and Cys) along with arginine; class I*b* enzymes recognize glutamic acid and glutamine; and class I*c* is formed by enzymes that recognize the aromatic tyrosine and tryptophan residues. Likewise, class II*a* enzymes recognize histidine, proline, serine, threonine, alanine and glycine residues; class II*b* enzymes recognize the charged aspartic acid and asparagine residues; and class II*c* recognize the aromatic phenylalanine. Interestingly, when the members of the two classes of synthetases are listed according to their subclasses, a *symmetry* emerges, both in terms of the number of members and in terms of the chemical properties of the amino acid. Such symmetry is particularly obvious between the members of subclasses I*b* and II*b*, as both recognize charged amino acids and their derivates; and between I*c* and II*c*, that recognize the aromatic amino acids (Moras, 1992; Cusack, 1995; Ribas and Schimmel, 2001b).

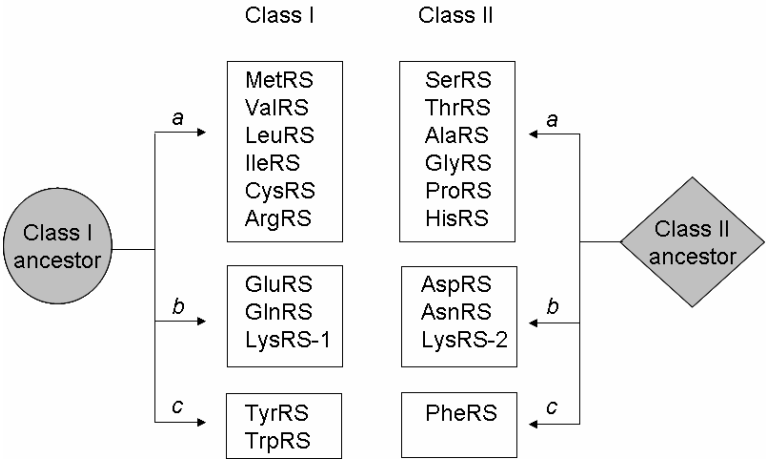


Figure 1. 11 – The two classes of aminoacyl-tRNA synthetases and their sub-classes.
The division if the aaRSs in classes I and II, and sub-classes *a, b* and *c*. The symmetry of the sub-classes is represented. Based on (Ribas and Schimmel, 2001a).

1.3.2.2. The evolution of aminoacyl-tRNA synthetases

The aaRSs are among the oldest proteins that appeared before the last common ancestor. Since aaRSs for a given amino acid are more related among different organisms than among other synthetases within the same organism (Nagel and Doolittle, 1991), their origin and evolutionary history reflects the history of life itself. For this reason, aaRSs can be regarded as potential markers for phylogenetic studies (Brown and Doolittle, 1995; Woese et al., 2000; Ribas et al., 2001). Interestingly, out of the 20 aminoacyl-tRNA synthetases, only 3 are not present in all organisms, namely the GluRS, AsnRS and CysRS. The first two are present in all eukaryotes, but only in some bacteria (Freist et al., 1997; Siatecka et al., 1998) and the latter is absent in the methanogenic archea *Methanocaldococcus jannaschii*, *Methanothermobacter thermautotrophicus* and *Methanopyrus kandleri* (Doolittle and Handy, 1998; Koonin and Aravind, 1998).

The existence of two classes of aaRSs containing 10 enzymes each, suggests that they have evolved from two ancestral molecules – the ancestors of the Rossman fold (class I) and of the antiparallel β-sheet (class II) (Eriani et al., 1995; Wolf et al., 1999). Similarly, each subclass is thought to have its own ancestor that arose after the progenitor of the entire class.

The class I and II enzymes high divergence, both at sequence and at mechanistic levels, is regarded as evidence for their independent origins in the archaic translational systems (Carter, Jr., 1993; Cavarelli and Moras, 1993). However, according to phylogenetic analysis of both classes of synthetases, they have about the same evolutionary age (Nagel and Doolittle, 1991), and it seems incongruous that in archaic systems two types of molecules would have independently emerged to perform the same catalytic function. This observation, led Rodin and Ohno to propose that the class division is intrinsic to the origin of translation itself and does not result from independent origins. According to them, the aaRSs arose from a primordial gene that encoded the ancestors of the two classes on opposite strands (Figure 1. 12) (Rodin and Ohno, 1995; Rodin and Rodin, 2006). This hypothesis was strengthen by two findings – *(i)* a gene of *Achlya klebsiana* encodes in the sense strand a glutamate dehydrogenase (GDH), and in the antisense strand a HSP70-like chaperonin (LeJohn *et al.*, 1994), and *(ii)* GDH has homology to class I aaRSs while the HSP70 ATP binding site has homology to motif 2 of class II SerRS (Carter and Duax, 2002).

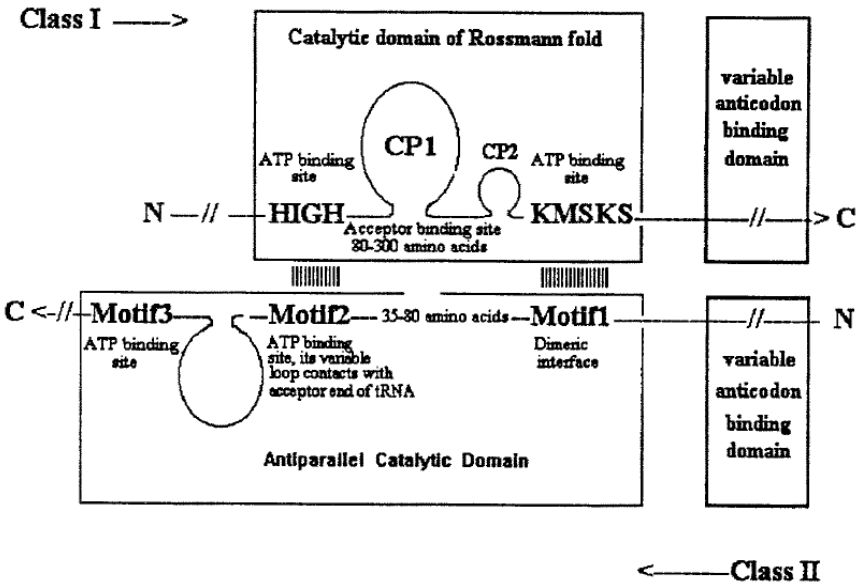


Figure 1. 12 – The antiparallel map of Class I versus Class II aminoacyl-tRNA synthetases.
The class I defining signature motif HIGH stands against the motif 2 of class II aaRSs, and the KMSKS against motif 1. Adapted from (Rodin and Ohno, 1995).

The analysis of the structure of aaRS-tRNA complexes suggests that catalytic domains of synthetases from opposite subclasses are able to bind to a single tRNA acceptor stem without any steric clashes, as they bind to opposite sides of the tRNA acceptor stem (Figure 1. 13). This symmetrical nature of the two classes suggests that their evolution was shaped under the same evolutionary pressure, and can be interpreted as evidence that primordial synthetases have developed a protection for the acceptor helix in a hostile environment, namely high temperature (Ribas and Schimmel, 2001b).

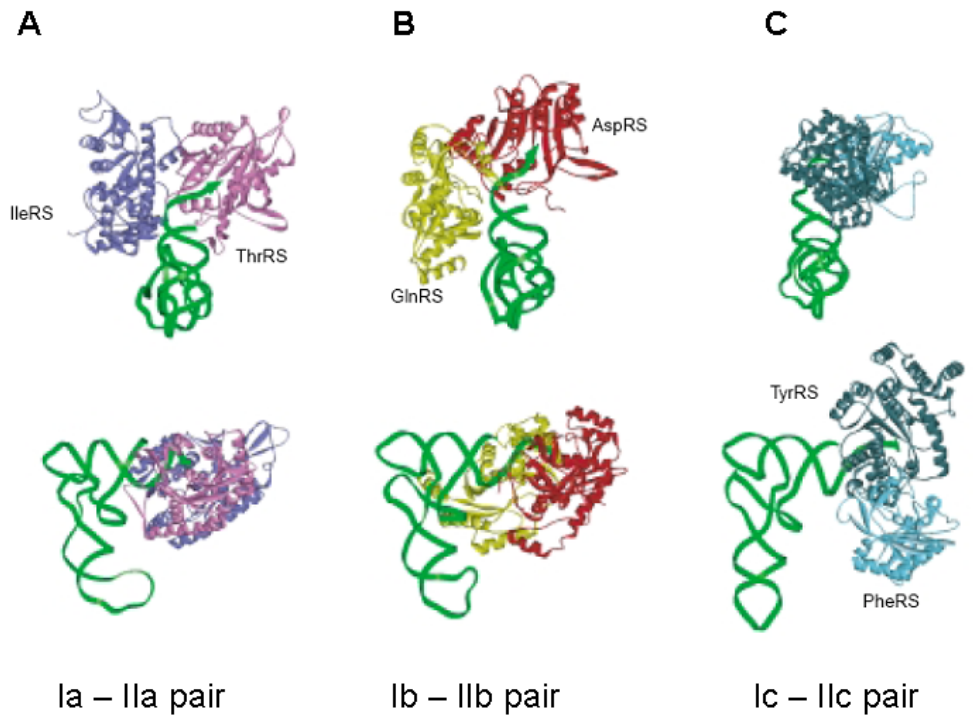


Figure 1. 13 – The class I and II synthetases complexes.
Model for the ternary complexes class I aaRS–class II aaRS–tRNA. On the top, the molecules are displayed along the axis of the anticodon stem loop, from the acceptor stem side, whereas in the bottom, the complexes are oriented with the plane defined by the axes of the tRNA acceptor stem and anticodon stem helices in parallel. **(A)** The IleRS-ThrRS-tRNA complex, both synthetases belong to the sub-class *a*. **(B)** The ternary complex formed with the sub-class *b,* GlnRS-AspRS-tRNA complex. **(C)** The sub-class *c* TyrRS-PheRS-tRNA complex. Adapted from (Ribas and Schimmel, 2001a).

Initially, these complexes of 2 synthetases and 1 tRNA may have been required for discrimination of closely related amino acids, namely valine *vs*. threonine in subclass *a*; glutamate *vs*. aspartate or glutamine *vs*. asparagine in subclass *b*; and tyrosine *vs*. phenylalanine in subclass *c*. The acquisition of the capacity to discriminate between similar

amino acids allowed the double aaRS complexes to separate and to evolve independently from each other (Ribas and Schimmel, 2001a; Ribas and Schimmel, 2001b).

At a later stage, a second aaRS domain was joined to the primordial catalytic site domain, which provided contacts with tRNA domains distal from the amino acid acceptor stem, namely the anticodon-domain in MetRS and GluRS and the variable loop in class II SerRS (Rould et al., 1991; Brunie et al., 1990; Cusack et al., 1996; Mosyak et al., 1995; Arnez et al., 1995). Thus, these two aaRS domains interact with different regions of the tRNAs – the catalytic domain interacts with the acceptor-TψC minihelix; while the second major domain interacts with other regions of the tRNA, such as the anticodon or the variable loop. The addition of the nonconserved domains possibly occurred when the D-arm and the anticodon domains of the tRNA emerged and became important for the translation process (Schimmel *et al.*, 1993).

These late domains of aaRSs were often recruited by other types of proteins and created novel functionalities. For example, the cytokine EMAPII (endothelial monocyte-activating polypeptide II) is homologous to the C-terminal domain of mammalian TyrRSs. Interestingly, this domain, which is not essential for aminoacylation, once cleaved by an elastase (an extracellular enzyme from polymorphonuclear leukocytes) has cytokine function (Wakasugi and Schimmel, 1999; Kleeman et al., 1997). Apart from this, aaRS like-domains are also involved in amino acid biosynthesis, DNA replication, RNA splicing and cell cycle control (reviewed in Francklyn et al., 2002; Martinis et al., 1999).

1.3.2.3. Ancient pathways for tRNA charging

The discovery of indirect synthesis of asparaginyl-, glutaminyl-, and cysteinyl-tRNAs has shed new light on the evolution of aaRSs (reviewed in Ibba and Soll, 2000) and provided valuable arguments for the co-evolution theory of the genetic code (Di Giulio, 2001a).

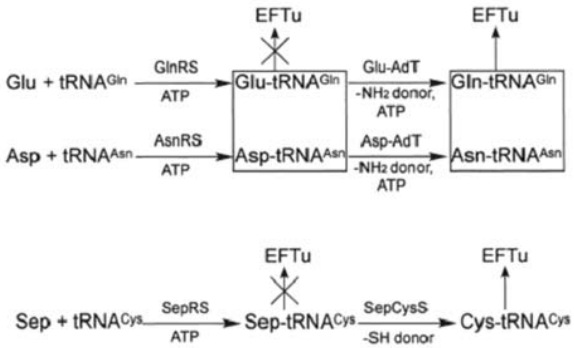


Figure 1. 14 – Alternative pathways for tRNA aminoacylation.
The ancient routes for the Gln-, Asn- and Cys-tRNAs charging. Both Gln- and Asp-tRNA charging is achieved by a transamidation reaction since $tRNA^{Gln}$ and $tRNA^{Asn}$ are firstly mischarged with Glu and Asp, respectively. These mischarged products are not recognized by the EF-Tu, and so are not used by the translational machinery. Then a transamidase transfers a $–NH_2$ group either to the Glu- or the Asp- residue on the tRNA, hence generating the Gln- and Asn-tRNA. The synthesis of the $Cys\text{-}tRNA^{Cys}$ undergoes a similar process, the tRNA is firstly mischarged with *O*-phospho-serine (Sep), by SepRS, and then the SepCysS catalysis the conversion of the Sep into Cys. Adapted from (Praetorius-Ibba and Ibba, 2003).

The synthesis of $Asn\text{-}tRNA^{Asn}$ and $Gln\text{-}tRNA^{Gln}$ in most bacteria and in all archea is accomplished by an indirect pathway that requires mischarging of those tRNAs by AspRS and GluRS, originating $Asp\text{-}tRNA^{Asn}$ and $Glu\text{-}RNA^{Gln}$ intermediates, respectively (Figure 1. 14) (Curnow et al., 1997; Curnow et al., 1996). However, the fidelity of translation is not compromised since the elongation factors do not recognize those mischarged tRNAs (Becker and Kern, 1998). Rather, the mischarged $Asp\text{-}tRNA^{Asn}$ and $Glu\text{-}tRNA^{Gln}$ are substrates for a tRNA-dependent aminotransferase (Asp/Glu-tRNA aminotransferase – AspAdT and GluAdT) (Curnow et al., 1997; Curnow et al., 1996; Ibba and Soll, 2000), that converts the attached aspartate to asparagine and the glutamate to glutamine, generating $Asn\text{-}tRNA^{Asn}$ and $Gln\text{-}tRNA^{Gln}$, respectively.

Another ancient indirect tRNA aminoacylation pathway is the formation of $Cys\text{-}tRNA^{Cys}$ in certain methanogenic archea lacking the CysRS (Figure 1. 14). In *Methanocaldococcus jannaschii*, *Methanothermobacter thermautotrophicus* and *Methanopyrus kandleri*, the $tRNA^{Cys}$ is charged with *O*-phosphoserine (Sep), a precursor of cystein, by a class II SepRS, forming the noncognate $Sep\text{-}tRNA^{Cys}$, which is converted to cognate $Cys\text{-}tRNA^{Cys}$ by the Sep-tRNA:Cys-tRNA synthase (SepCysS) (Sauerwald et al., 2005; O'Donoghue et al., 2005).

These ancient indirect aminoacylation pathways indicate that Cys, Asn, and Gln are recent acquisitions, and consequently, CysRS, AsnRS and GlnRS appeared more recently than other aaRSs, probably after the first split of the archeal and bacterial branches (Wong, 1975; Lamour et al., 1994; Becker et al., 2000; Stathopoulos et al., 2000; Sethi et al., 2005).

1.3.2.4. Editing

A central issue on protein synthesis is its high fidelity, which, in part, results from correct selection of both tRNA and amino acids by aaRSs. Since the latter is rather complex for chemically similar amino acids, namely leucine and isoleucine, aaRSs evolved an editing mechanism that prevents mischarged tRNA to reach protein synthesis (Nangle et al., 2002; Zhao et al., 2005).

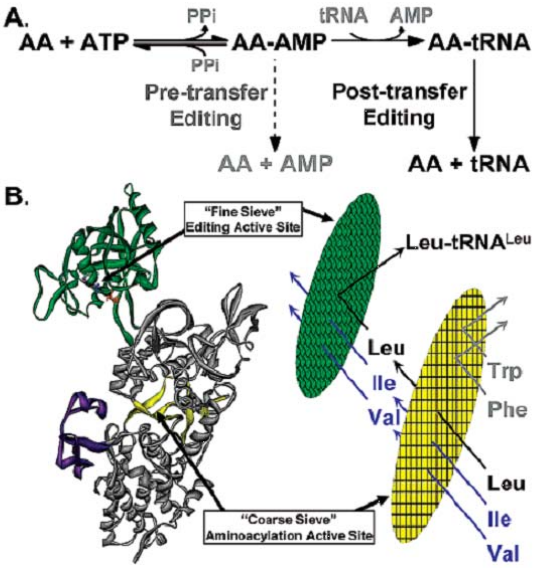


Figure 1. 15 – Pre- and post-transfer editing of the aminoacylation reaction.
(**A**) The aminoacylation reaction and the steps where pre- and the post- transfer editing occur. The pre-transfer editing is achieved immediately after the amino acid activation, whereas the post-transfer editing is only achieved after the aminoacylation of the tRNA. (**B**) Editing by *E. coli* LeuRS, which lacks the pre-transfer activity. The yellow filter represents the aminoacylation active site, in the Rossman fold, which in the LeuRS, besides Leu, activates Ile and Val, but discriminates against bulkier amino acids, namely Trp and Phe. The green filter represents the editing CP1 domain of LeuRS. Adapted from (Mursinna *et al.*, 2004).

Editing can occur at pre-transfer or post-transfer levels (Figure 1. 15 A). In the former, non-cognate amino acids, misactivated in the catalytic domain of aaRSs, are hydrolyzed before being transferred onto the tRNA (Fersht and Dingwall, 1979; Zhao et al., 2005). In the latter, the misactivated amino acids are transferred to the 3'-CCA end of tRNA, but are then hydrolysed before being transferred to the translation elongation factors, which transport aa-tRNA to the ribosome (EF-Tu or eEF-1A). However, some enzymes, namely IleRS, use both types of proofreading synergistically (Zhao *et al.*, 2005).

Editing in class-I aaRSs is exemplified by ValRS, LeuRS and IleRS, which have to differentiate amino acids that only differ by a methyl group, namely leucine, valine and isoleucine. These class I*a* aaRSs contain an highly efficient editing domain named CP1 domain, which is inserted in the Rossman fold (Nureki et al., 1998; Lin and Schimmel, 1996; Cusack et al., 2000). The CP1 domain has a threonine rich motif that is likely to participate in hydrolysis of the transiently misacylated tRNA (Nureki et al., 1998; Mursinna et al., 2001). Its core is a highly conserved beta-barrel fold, though its peripheral structures are quite variable (Zhao *et al.*, 2005).

In class II aaRSs, the editing mechanism is not yet fully understood, however it exists in AlaRS, ThrRS, ProRS, PheRS, and LysRS. The AlaRS hydrolyzes misactivated serine and glycine, the PheRS deacylates Ile-tRNA^{Phe}; the LysRS hydrolyzes misactivated homocysteine, homoserine, cysteine, threonine and alanine; the ProRS edits alanine; and the ThrRS edits serine (Tsui and Fersht, 1981; Beebe et al., 2003; Dock-Bregeon et al., 2000; Beuning and Musier-Forsyth, 2000; Beuning and Musier-Forsyth, 2001; Yarus, 1972; Jakubowski, 1997). These editing domains are diverse in structure and location and are unevenly distributed through the three domains of life. For instance, the ThrRS editing domain is located in the N-terminus fused to its catalytic core and has strong sequence homology among eukaryotes and bacteria, but is absent in archea. The AlaRS has an editing domain with similar architecture to the ThrRS editing domain and is present in all organisms. However, ThrRS and AlaRS are the only class II enzymes that have such similar editing domains (Sankaranarayanan et al., 1999; Dock-Bregeon et al., 2000). Conversely, the ProRS editing domain is formed by a large insertion within motifs 2 and 3,

though it is absent in higher eukaryotes (Beuning and Musier-Forsyth, 2000; Beuning and Musier-Forsyth, 2001).

1.3.2.5. tRNAs misacylation

Despite having highly refined quality control mechanisms, aaRSs misacylate tRNAs at a rate of 10^{-4} to 10^{-5} (reviewed in Jakubowski and Goldman, 1992). However, the rapid enzyme turnover and the kinetic proofreading by elongation factors (EF-Tu in prokaryotes and EF1α in eukaryotes) ensure that these misacylated tRNAs do not compromise the fidelity of protein synthesis. This explains why misacylated Asp-$tRNA^{Asn}$ and Glu-RNA^{Gln} (Section 1.3.2.3), do not compromise the fidelity of translation (Table 1. 3) (reviewed by Ibba and Soll, 2004). Also interesting is the initiation of prokaryotic protein synthesis with formyl-methionine, charged onto an initiator $tRNA_i^{fMet}$, which differs from the elongator $tRNA^{Met}$. The MetRS recognizes the anticodon of $tRNA_i^{fMet}$ and charges it with methionine (Schulman and Pelka, 1988). Formylation of methionine is catalysed by methionyl-tRNA formyltransferase (MTF), which specifically recognizes base pairs 2:71 and 3:70, in the acceptor stem of the $tRNA_i^{fMet}$ (Schmitt et al., 1998; Schulman and Her, 1973; Seong and RajBhandary, 1987).

Table 1. 3 – Natural occurring misacylations.
Examples of the tRNAs charged with non-cognate amino acids. These mischarged tRNA are not recognized by the elongation factors, or the initiator factor in the case of the $tRNA_i^{fMet}$. Only after a modification do they become correctly charged and, consequently, available for the translational machinery.

Amino Acid	tRNA	Mischarged tRNA (non recognized by EF / IF)	Correctly charged tRNA (recognized by EF)
Glu	$tRNA^{Gln}$	*Glu-*$tRNA^{Gln}$	*Gln-*$tRNA^{Gln}$
Asp	$tRNA^{Asn}$	*Asp-*$tRNA^{Asn}$	*Asn-*$tRNA^{Asn}$
Sep	$tRNA^{Cys}$	*Sep-*$tRNA^{Cys}$	*Cys-*$tRNA^{Cys}$
Ser	$tRNA^{Sec}$	*Ser-*$tRNA^{Sec}$	*Sec-*$tRNA^{Sec}$
Lys	$tRNA^{Pyl}$	*Lys-*$tRNA^{Pyl}$	*Pyl-*$tRNA^{Pyl}$
Met	$tRNA_i^{fMet}$	*Met-*$tRNA_i^{fMet}$	*fMet-*$tRNA_i^{fMet}$

1.4. Genetic code alterations

The discovery of genetic code alterations shows that the genetic code evolves, even in organisms with complex genomes and proteomes. However, most genetic code changes occur in mitochondria and cytoplasmic genetic code alterations are in fact a subset of the former, indicating that proteome size imposes significant constraints to the evolution of genetic code alterations. The diversity of genetic code alterations uncovered to date also shows that they occur in distinct phylogenetic lineages and evolve from the standard genetic code rather than from ancient alternative codes. Interestingly, certain codons are more prone to identity change that others. For example, codons starting with A or U often change their identity, while no genetic code change has yet been discovered involving codons starting with G. Interestingly, there are two genetic code alterations involving codons that start with C, namely CUN codons (Li and Tzagoloff, 1979), which are reassigned from leucine to threonine in yeast mitochondria and also the CUG codon which is reassigned from leucine to serine in various *Candida* species (Santos and Tuite, 1995). This strongly suggests that the strength of first position codon-anticodon base pairing limits codon identity alterations and supports the hypothesis that codon decoding efficiency is a key factor in the evolution of genetic code alterations. Finally, certain codons are rather unstable as they changed identity more than once. For example, the arginine AGG codons changed identity to Ser, Gly, and STOP (as reviewed in Knight *et al.*, 2001) and STOP-codons changed their identity to different amino acids, namely, tryptophan, tyrosine, glutamate, glutamine and cysteine (Osawa *et al.*, 1992).

1.4.1. The mechanisms of evolution of genetic code alterations

Two main theories have been proposed to explain the evolution of genetic code alterations, namely - the *"Codon Capture Theory"* and the "*Ambiguous Intermediate Theory*".

The *"Codon Capture Theory"* (Osawa and Jukes, 1989) postulates that code changes are the result of biased genome G+C pressure. Since the latter has a strong effect on codon

usage by modulating the frequency of the 3rd nucleotide position of codons (GC_3 pressure), the theory predicts that under strong G+C bias some codons may disappear altogether from the entire set of open reading frames of genomes (ORFeome) (Figure 1. 16a). The theory is supported by the finding that in *Mycoplasma capricolum*, whose genome has 25% G+C only, the CG rich arginine CGG codon disappeared from the ORFeome (is unassigned) and its cognate $tRNA^{Arg}$ has also been lost (Oba *et al.*, 1991); and in *Micrococcus luteus*, whose genome has 26% A+T, the A/T rich codons UUA, AUA and AGA are also unassigned (Ohama et al., 1990; Kano et al., 1991). The theory also proposes that rare codons are primary targets for identity change since these codons disappear from ORFeomes more easily than frequently used ones (Osawa *et al.*, 1992).

The *"Ambiguous Intermediate Theory"* (Schultz and Yarus, 1994), postulates that genetic code alterations are driven by selection and result from direct alteration of the translational machinery. In particular, mutations in tRNA genes that promote tRNA misreading are an important driving force of genetic code alterations. These mutations normally alter tRNA anticodons, translation release factors, tRNA modifying enzymes and aminoacyl-tRNA synthetases and create codons with more than one identity. That is, codons became decoded by either more than one tRNA or by both a release factor and a tRNA. This creates codon ambiguity and sets the stage for codon identity change, as the mutant tRNA may improve its decoding efficiency through additional mutations and become the main decoder of the codon undergoing the identity change (Figure 1. 16b). This theory does not require codon disappearance prior to reassignment and assumes that codon ambiguity is not deleterious. However, it does not explain what kind of selective advantage arises from codon ambiguity to allow for selection of genetic code alterations. This theory is supported by the existence of many natural suppressor tRNAs and a unique ambiguity status of CUG codons in many *Candida* species (Hanyu et al., 1986; Santos et al., 1997; Suzuki et al., 1997). This theory has been tested experimentally by engineering codon ambiguity in *E. coli* and yeast (Pezo et al., 2004; Bacher et al., 2003; Santos et al., 1996; Santos et al., 1999). Remarkably, cells are highly tolerant to codon ambiguity, but trigger a unique stress response which dramatically increases pre-adaptation potential. This suggests that codon ambiguity is advantageous under certain stress conditions (Bacher et al., 2003; Santos et al., 1999).

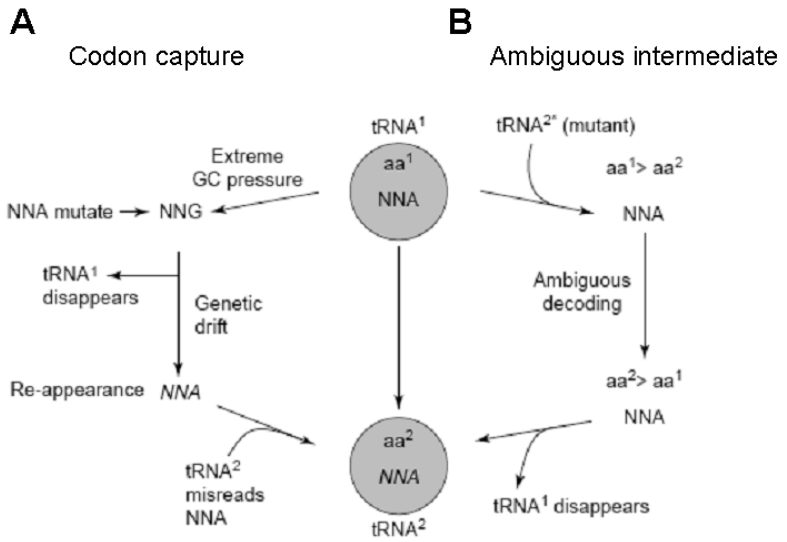


Figure 1. 16 – Sense codon reassignment.
(A) The codon capture theory. **(B)** The ambiguous intermediate theory. Adapted from (Santos *et al.*, 2004)**.**

Despite the differences between those two theories, they are not mutually exclusive, in fact, Sengupta and Higgs have recently proposed a generic unifying model for codon identity changes (Sengupta and Higgs, 2005) – the *Gain-Loss Model* – based on their observations that codon reassignments always involve both a gain and a loss (Figure 1. 17). They consider as *"gain"* the new tRNA for the reassigned codon or a gain of function of an existing tRNA (due to a mutation or a base modification); and as *"loss"* a deletion of tRNA or release factor genes, or loss of function of such gene, again, due to a mutation or a base modification. According to this model, the *Codon Capture Theory* and the *Ambiguous Intermediate Theory* have a synergistic action and it is the strength and the frequency of the loss or the gain that determines which mechanism is favoured – for instance, if a codon identity change requires a new modified base, a loss seems simpler than a gain, as it is easier to lose a tRNA gene than to gain a novel enzyme to create such modification, and hence the *Codon Capture* model would be favoured.

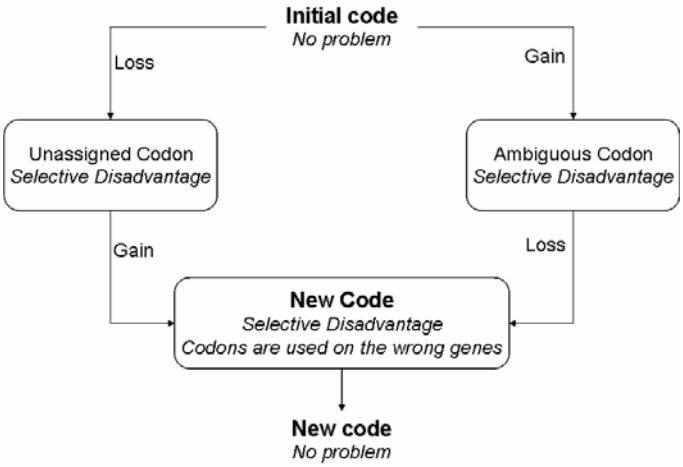


Figure 1. 17 – The Gain-Loss model.
The codon reassignment process under the gain-loss model. Adapted from (Sengupta and Higgs, 2005)**.**

1.4.2. Mitochondrial Genetic Code alterations

As mentioned above, the majority of genetic code alterations occur in mitochondria. This may be due to the smaller size and highly A+T biased genomes (reviewed in Knight *et al.*, 2001). Indeed, to date, 15 genetic code alterations have been reported in mitochondrial genomes, involving the reassignment of CUN, CGN, AGR, UGA, UAG, AUA, AAA and UAA codons (Table 1. 4). Such changes to the standard genetic code, albeit being widely spread, are not evenly distributed through the various phylogenetic groups, since many plant mitochondria do not have genetic code alterations while metazoan mitochondria are rather prone to them (Figure 1. 18).

Table 1. 4 . Variations in the mitochondrial genetic code.

Organism	UGA Stop	AUA Ile	AAA Lys	AGR Arg	CUN Leu	UAA Stop	UAG Stop	Example
Vertebrates	Trp	Met	–	Stop	–	–	–	Human, bovine, frog
Tunicates	Trp	Met	–	Gly	–	–	–	*Halocynthia roretzi*
Echinoderms	Trp	–	Asn	Ser	–	–	–	Starfish, sea urchin
Arthropods	Trp	Met	–	Ser*	–	–	–	*Drosophila* spp., mosquito, honeybee
Molluscs	Trp	Met	–	Ser	–	–	–	Squid, *Mytilus edulis*
Nematodes	Trp	Met	–	Ser	–	–	–	*Caenorhabditis elegans, Ascaris suum*
Platyhelminths	Trp	–	Asn	Ser	–	Tyr	–	*Fasciola hepatica, planaria*
Coelenterates	Trp	ND	ND	–	ND	ND	–	Hydra, *Metridium senile*
Yeasts	Trp	Met	–	–	Thr	–	–	*Saccharomyces cerevisiae, Torulopsis glabrata*
Green algae	Trp	–	–	–	–	–	Ala	*Hydrodictyon reticulatum*
	Trp	–	–	–	–	–	Leu	*Coelastrum microporum*
Euascomycetes	Trp	–	–	–	–	–	–	*Aspergillus nidulans, Neurospora crassa*
Protozoa	Trp	–	–	–	–	–	–	*Paramecium* spp.

The most ancient and common mitochondrial genetic code alteration involves the change of identity of the UGA stop codon to tryptophan (Yokobori et al., 2001; Inagaki et al., 1998). Though in green plants, namely in *Hydrodictyon reticulatum* and *Coelastrummicroporum*, the UAG stop changed its identity to alanine or serine (notes 14 and 15 from Figure 1. 18) (Hayashi-Ishimaru *et al.*, 1996). Apart from this, it is also surprising that some codons have changed identity to different amino acids in different organisms. For example, the arginine AGR codons have been reassigned to serine in platyhelminths, nematodes, annelids, arthropods, molluscs, echinoderms and hemichordates. Such new codons have further changed their identity from serine to glycine in urochordates and became stop codons in vertebrates. Finally, unassigned AGR codons were re-introduced in different species of *Brachiostoma* as glycine or serine (notes 3, 10, 11, 12 and 13 from Figure 1. 18). Another example of successive identity changes is the isoleucine AUA codon, which changed its identity to methionine in the metazoan clade, but in platyhelminths, echinoderms and hemichordates it has reverted its identity back to isoleucine (notes 2 and 4 from Figure 1. 18) (Castresana *et al.*, 1998).

Sense to sense identity changes have also occurred at leucine CUN and lysine AAA codons, which altered their identity to threonine (Pape *et al.*, 1985) and to asparagine in platyhelminths and echinoderms (Castresana *et al.*, 1998), respectively. Regarding the arginine CGN codon family, it has been unassigned in yeasts (Pape et al., 1985; Clark-Walker and Weiller, 1994).

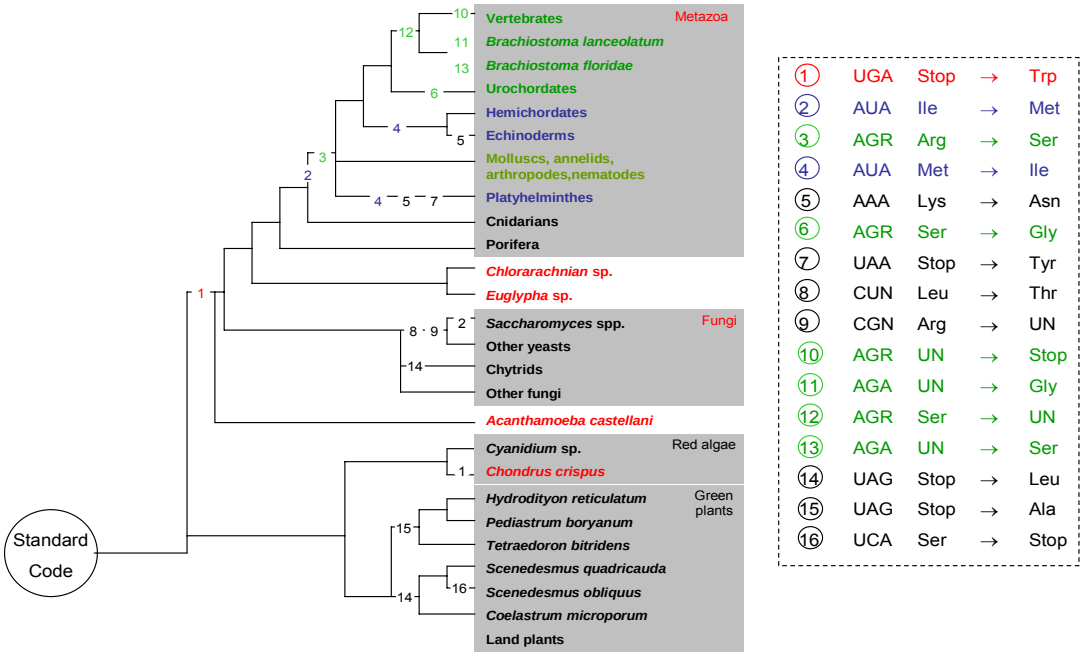


Figure 1. 18 – The mitochondrial genetic codes.
The phylogeny of the genetic code changes in mitochondrial genomes showing that some organisms have experienced consecutive alterations, highlighted in green and blue. Adapted from (Knight *et al.*, 2001)

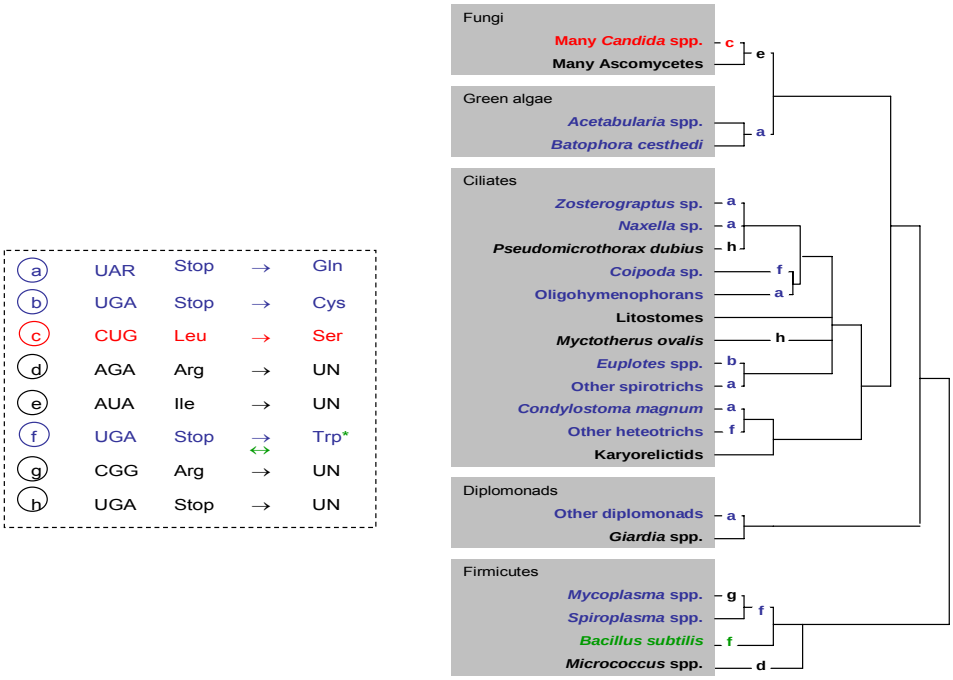


Figure 1. 19 - The nuclear/cytoplasmatic genetic code alterations.
The phylogenetic tree shows STOP codon reassignments in blue and STOP codon unassignments in black. The leucine to serine CUG codon change, in the *Candida* genus, is the only known sense to sense reassignment in eukaryotes. Adapted from (Knight *et al.*, 2001)

1.4.3. Cytoplasmic genetic code alterations

All the cytoplasmic genetic code alterations involve codons that have also experience a reassignemet in mitochondrial genomes (Figure 1. 19). These alterations involve mainly stop codons, but there are also unassigned codons in *Mycoplasma capricolum* and *Micrococcus luteus*, and the leucine CUG codon is reassigned to serine in many yeast species, in particular in species of the genus *Candida* (Ohama et al., 1993; Santos et al., 1993). The stop codons UAA/G (UAR) changed their identity to glutamine in some ciliates, in *Zosterograptus, Paramecium* and *Nexella* (Lozupone et al., 2001; Sanchez-Silva et al., 2003), in green algae of the genus *Acetaularia* (Schneider and de Groot, 1991), and in diplomonads (Keeling and Doolittle, 1997). In three peritrich ciliates – *Vorticella microstoma, Opisthonecta henneguyi* and *Opisthonecta matiensis* – only the UAA codon is decoded as glutamine (Sanchez-Silva *et al.*, 2003). On the other hand, the UGA stop codon changed its identity to cysteine, in the genus *Euplotes* (Lozupone *et al.*, 2001), and to tryptophan in some bacteria, namely in *Mycoplasma*, *Spiroplasma* and in *Bacillus subtilis* – interestingly in the later it remains ambiguous as it can be used to terminate protein synthesis or decoded as tryptophan (Lovett et al., 1991; Matsugi et al., 1998). Finally, in *Nyctotherus ovalis* the UGA has been unassigned (Lozupone *et al.*, 2001). Other unassigned codons are the arginine CGG codon in *Mycoplasma* species (Oba *et al.*, 1991) and the arginine AGA and the isoleucine AUA codons in *Micrococcus* (Ohama et al., 1990; Kano et al., 1991).

1.4.4. The Expansion of Genetic Code

As discussed above, there are only 20 primary amino acids specified in the genetic code, although at least 120 amino acids and amino acid derivatives have been identified as constituents of different proteins in different organisms (Crick, 1968; Uy and Wold, 1977), all of them are products of post-translational modifications. During the last 20 years, selenocysteine and pyrrolysine were also added to the genetic code since they are incorporated into proteins in response to UGA and UAG stop codons, respectively in various (not all) organisms (Zinoni et al., 1987; Hao et al., 2002). This expansion of the

genetic code from 20 to 22 amino acids confirmed the code flexibility and showed that codon identity can be reprogrammed through structural alteration of the protein synthesis machinery. It also showed that genetic code expansion brings about novel protein functionalities since these novel amino acids are located in the catalytic centre of the respective enzymes and participate directly in catalysis. Selenocysteine and pyrrolysine also suggest that additional non-standard amino acids may exist, however *in silico* strategies for genome mining have so far failed to identify the putative 23rd amino acid (Lobanov *et al.*, 2006).

1.4.4.1. Selenocysteine

Selenocysteine exists in all kingdoms of life. It is a cysteine analogue containing selenium instead of sulphur atoms and is critical for selenoprotein catalysis (reviewed by Hatfield and Gladyshev, 2002). Its translational insertion at UGA stop codons involves an alternative decoding mechanism mediated by several new translational factors, namely: *(i)* a unique tRNA ($tRNA_{Sec}$); *(ii)* a specific structure on the mRNA called selenocysteine insertion sequence (SECIS); *(iii)* a SECIS binding protein; and *(iv)* a new elongation factor (SelB) (Thanbichler and Bock, 2002; Namy et al., 2004). These novel translational elements are structurally different in prokaryotes, eukaryotes and archea.

In *E. coli*, the specific $tRNA^{Sec}$ is initially charged with serine, by the seryl-tRNA synthetase (SerRS), and then the selenocysteine synthase (SelA) converts the seryl-$tRNA^{Sec}$ into selenocysteyl-$tRNA^{Sec}$ using selenophosphate as a source for activated selenium. The selenophosphate is provided by selenophosphate synthetase (SelD) from selenide in an ATP-dependent reaction (Leinfelder et al., 1988; Forchhammer et al., 1991). Once the $tRNA^{Sec}$ is correctly charged with selenocysteine, it is captured by the SelB, which is a homologue of the elongation factor Ef-Tu, containing an extra C-terminal domain, which confers the ability to recognize the SECIS-element. The latter is an mRNA structure, located immediately downstream of selenocysteine-UGA codons that guides the elongation factor SelB to the ribosome. Thus, the decoding of selenocysteine UGA codons depends on a quaternary complex formed by selenocysteyl-$tRNA^{Sec}$, SelB, GTP and the

SECIS-element (Figure 1. 20) (Leinfelder et al., 1988; Forchhammer et al., 1991; Thanbichler and Bock, 2002; Hatfield and Gladyshev, 2002).

In both eukaryotes and archea the SECIS-element is located in the 3'-untranslated region (3'-UTR) of the mRNA (Berry et al., 1991; Rother et al., 2001). In eukaryotes, the SECIS-element is recognized by a SECIS binding protein (SBP2), which recruits a specific elongation factor (eEFSec) that recognizes the selenocysteyl-$tRNA_{Sec}$. Therefore, in eukaryotes, incorporation of selenocysteine requires a complex formed by selenocysteyl-$tRNA_{Sec}$, eEFSec, GTP, SBP2 and the SECIS-element (Figure 1. 20) (Tujebajeva *et al.*, 2000). In archea, the mechanism of selenocysteine incorporation is not yet fully understood, though a SECIS element in the 3'-UTR and an archeal specific elongation factor (aSelB) have been described (Rother *et al.*, 2001).

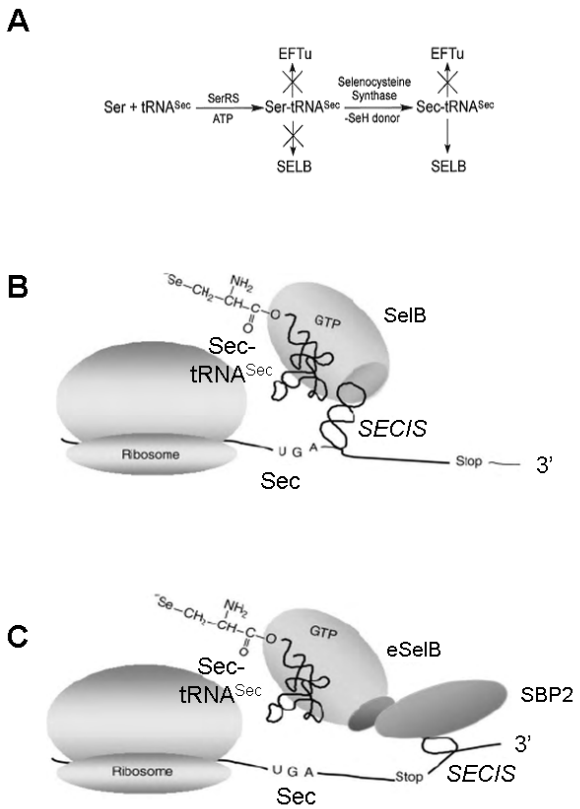


Figure 1. 20 – The synthesis of selenoproteins.
(**A**) The aminoacylation of $tRNA^{Sec}$. (**B**) Selenocysteine incorporation in prokaryotes is mediated by a SelB transcription factor and a structured SECIS element in the open reading frame. (**C**) In eukaryotes, the selenocysteine is inserted at the UGA codon by the eSelB translation factor, which interacts with the SBP2 that recognizes the SECIS element in the 3'-UTR of the mRNA.

1.4.4.2. Pyrrolysine

Pyrrolysine is translationally incorporated in methanogenic archea in response to UAG stop codons present in the monomethylamine methyltransferase, an enzyme of the catabolic route of methylamines. It is the most recent addition to the genetic code and is referred to as the 22nd amino acid. Its incorporation mechanism is not yet fully understood, though a suppressor tRNA with a CUA anticodon ($tRNA_{CUA}^{Pyl}$) is known to play a key role in pyrrolysine incorporation.

Two distinct pathways, namely a direct and an indirect pathway have been described for $tRNA_{CUA}^{Pyl}$ charging (Figure 1. 21). In the direct pathway, a cognate pyrrolysyl-tRNA synthetase (PylS) charges the cognate $tRNA_{CUA}$ with pyrrolysine (Blight et al., 2004; Korencic et al., 2004; Polycarpo et al., 2004). In the indirect pathway, the $tRNA_{CUA}^{Pyl}$ interacts with both class I (LysRS1) and class II (LysRS2) lysyl-tRNA synthetase, forming the ternary complex $tRNA_{CUA}^{Pyl}$:LysRS1:LysRS2. The $tRNA_{CUA}^{Pyl}$ is firstly charged with lysine, which is then converted to pyrrolysine by a not yet fully understood pathway (Polycarpo et al., 2003; Srinivasan et al., 2002). The existence of the indirect pathway to obtain Pyl-$tRNA_{CUA}^{Pyl}$, which is less efficient than the direct pathway (Krzycki, 2005), can be regarded as a backup mechanism to overcome pyrrolysine deficiency, and hence safeguards the biosynthesis of proteins that require pyrrolysine (Polycarpo *et al.*, 2004).

The Pyl-$tRNA_{CUA}^{Pyl}$ interacts *in vitro* with Ef-Tu and can be used by the *E. coli* translational machinery, indicating the requirement for a specific elongation factor (EF-Pyl) for its specific incorporation at specific UAG codons (Blight et al., 2004; Theobald-Dietrich et al., 2004). *In silico* analysis predicted the existence of a hairpin structure, called *py*rro*l*ysine *i*nsertion *s*equence (PYLIS), in the mRNA immediately after the reprogrammed UAG codon (Namy *et al.*, 2004), whose existence and structure have later been confirmed experimentally (Theobald-Dietrich *et al.*, 2004). With these data some authors have built a model for the Pyl incorporation at re-programmed UAG codons (Figure 1. 21) similar to the Sec incorporation. However, the Pyl incorporation is still puzzling the research community, indeed, a recent study demonstrated that Pyl was

efficiently inserted into proteins in an anonymous context and, apparently, did not depend on the presence of additional proteins (Ambrogelly *et al.*, 2007).

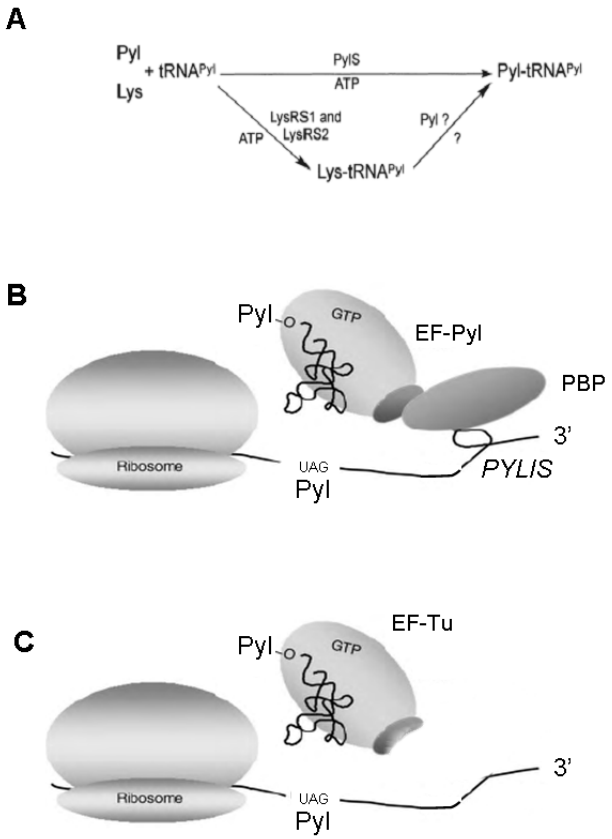


Figure 1. 21 – Pyrrolysine incorporation pathways.
(**A**) Charging of $tRNA^{Pyl}$ by both the direct and indirect pathways. (**B)** Model similar for Sec incorporation. According to this model there would be a PYLIS sequence, which would be recognized by the PLYIS-binding protein (PBP). Then the PBP would interact with a Pyl-specific elongation factor (EF-Pyl), and so the Pyl residue would be incorporated into the nascent polypeptide (Theobald-Dietrich *et al.*, 2004). (**C**) Alternatively, the Pyl residues can be inserted without the need for neither a EF-Pyl or PYLIS signal sequence (Ambrogelly *et al.*, 2007).

1.4.4.3. Artificial expansion of the genetic code

Modification of the 20 amino acids in living organisms indicates that proteins require additional chemical properties to carry out their natural functions, and that life with 20 amino acids is possible, but by no means optimal (Cropp and Schultz, 2004). Moreover, the extant alterations of the genetic code, together with its natural expansion, have unveiled an unforeseen malleability and an extraordinary adaptation capacity of living organisms. The current knowledge on the chemistry of life and recent biotechnology developments

have broaden the horizons for protein engineering, as it is now possible to genetically encode additional amino acids and hence enable evolution of novel proteins, or even entire organisms, with new or enhanced physical, chemical or biological properties. The array of possible applications is countless as the engineered proteins can be applied in fundamental research, for instance in crystallographic studies where methionine has been replaced by selenomethionine (Hendrickson *et al.*, 1990, reviewed by Hendrickson *et al.*, 2004), but also in applied research to create new pharmaceuticals, such as protease inhibitors used against HIV (Kiso, 1999; Mak et al., 2003) and *Candida albicans* infections (Bein et al., 2002; Bein et al., 2002).

Unnatural amino acids can be incorporated into proteins through both chemical and biosynthetic methodologies. The former is simple and straightforward, but only a limited number of residues can be modified with exogenous chemical reagents (Kent, 1988). Biosynthetic methods can be used *in vitro*, by introducing nonsense or frameshifting suppressor tRNAs, that are chemically misacylated with unnatural amino acids in cell free translation systems (Noren *et al.*, 1989); or *in vivo*, by engineering the translational apparatus of the living organisms, as has already been done in bacteriophages (Bacher *et al.*, 2003), *Escherichia coli* (Wang et al., 2001; Doring et al., 2001; Mehl et al., 2003) and *Saccharomyces cerevisiae* (Chin *et al.*, 2003). In summary, both theoretical and experimental approaches indicate that the genetic code is flexible and evolves. However, genetic code evolution is likely to introduce codon decoding ambiguity whose physiological and cellular consequences are not yet fully understood.

1.5. The *Candida* spp. genetic code

As discussed on the previous section, a number of alterations to the genetic code have been found in prokaryotic, non-plant mitochondrial and eukaryotic translation systems. However, the reassignment of the CUG codon from leucine to serine in *Candida albicans* and several other *Candida* species is unique, since it is the only nuclear genetic code change that involves a sense to sense reassignment (Santos et al., 1993; Santos et al., 1996; Suzuki et al., 1997).

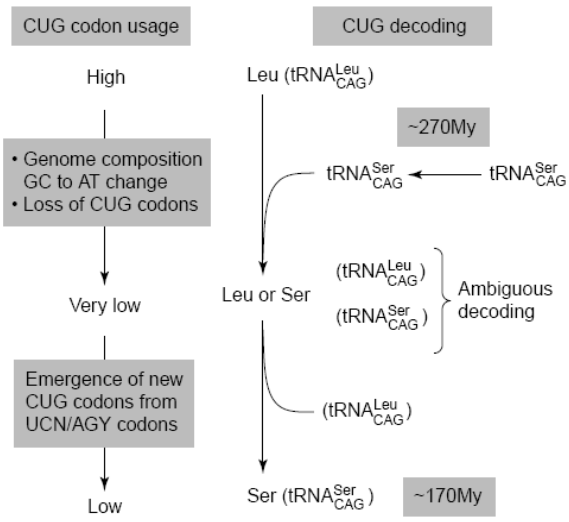


Figure 1. 22 – The evolution of the CUG codon reassignment in *Candida spp.*
The novel $tRNA_{CAG}^{Ser}$ appeared approximately 270 My ago and during 100 My competed with the cognate $tRNA_{CAG}^{Leu}$ for CUG decoding. This ambiguous CUG decoding was the main driving force responsible for decreasing CUG codon usage and consequent reorganization of CUGs in the genome. The 30,000 CUG codons present in the yeast ancestor disappeared, and "new" 16880 CUGs present in *C. albicans* genome evolved from UCN and AGY codons. For reasons not yet fully understood, the novel $tRNA_{CAG}^{Ser}$ was maintained while the cognate $tRNA_{CAG}^{Leu}$ was eliminated. Adapted from (Santos *et al,* 2004)

The evolution of this genetic code change can be regarded as a unifying model for the two theories of the evolution of genetic code changes – the *"Codon Capture Theory"* and the "*Ambiguous Intermediate Theory*" (Figure 1. 22). In one hand, it is mediated by an ambiguous tRNA, which introduces ambiguity at the CUG codon, and thus favours the "*Ambiguous Intermediate Theory*", but biased *Candida* genome A+T pressure also lowered CUG usage to low levels, thus favouring the *"Codon Capture Theory"*. The latter was

important to minimize the negative impact of CUG ambiguity in the proteome. Also, the appearance of the novel $tRNA_{CAG}^{Ser}$ played a critical role in the capture of many "new" CUG codons that mutated from codons coding for serine or amino acids with similar chemical properties (Massey *et al.*, 2003). Thus, the *"Codon Capture"* and the *"Ambiguous Decoding"* theories have synergistic effects on codon identity change.

This genetic code alteration is unevenly distributed through the *Candida* genus (Figure 1. 23) (Sugita and Nakase, 1999), thus indicating that reassignment of the CUG codon is at different evolutionary stages among its different species. Some *Candida* species translate the CUG codon exclusively as leucine, namely *C. glabrata* and *C. krusei*, while others like *C. cylindracea* decode it only as serine. However, in many species, such as *C. zeylanoides* and *C. albicans* the CUG codon is ambiguous, meaning that it is simultaneously translated as leucine and serine, because the $tRNA_{CAG}^{Ser}$ is charged with both serine (major) and leucine (minor) (Suzuki *et al.*, 1997) (this work).

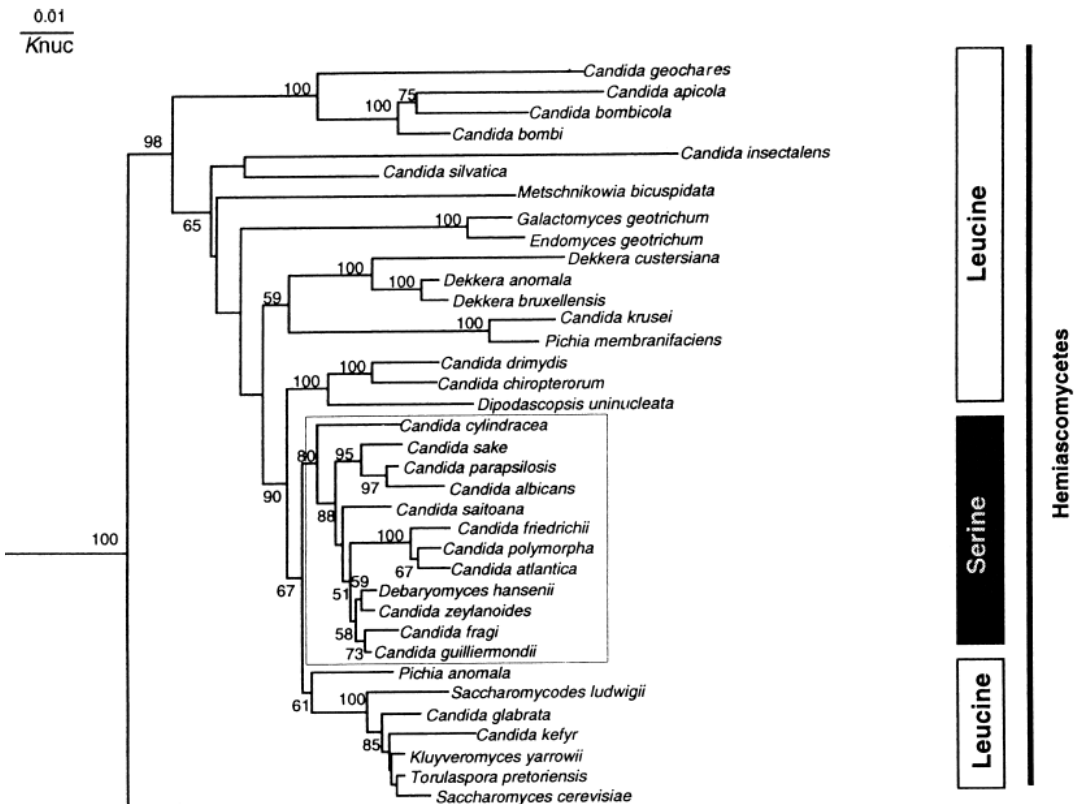


Figure 1. 23 – The phylogenetic tree of CUG decoding in Hemiascomycetes.
Those species that decode the CUG codon as serine are within the square box, all the other hemiascomycetes, including several *Candida* species, decode the CUG codon as the standard leucine. Adapted from (Sugita and Nakase, 1999).

1.6.1. The $tRNA_{CAG}^{Ser}$

The CUG reassignment from leucine to serine is mediated by a novel tRNA that has a hybrid nature. It has both leucine and serine identity elements (Figure 1. 24) that altogether are responsible for making this $tRNA_{CAG}^{Ser}$ an ambiguous molecule that is able to interact with both leucyl-tRNA synthetase (LeuRS) and seryl-tRNA synthetase (SerRS) (Santos et al., 1996; Perreau et al., 1999; Suzuki et al., 1997).

The discriminator base of this special tRNA is a Guanosine (G_{73}) which is an identity element for the serine tRNA-family. In *S. cerevisiae*, a single change of A_{73} to G_{73} of a $tRNA^{Leu}$ is sufficient to convert its identity to serine (Soma *et al.*, 1996). The other serine element of this tRNA is the variable arm, which contains a run of 3 conserved C-G pairs that is directly recognized by the SerRS. On the other hand, the anticodon arm of the $tRNA_{CAG}^{Ser}$ has leucine identity determinants, namely A_{35}, and m^1G_{37}, in the anticodon, which make direct contact with the LeuRS (Soma *et al.*, 1996). Interestingly, in *C. cylindracea*, the CUG codon is decoded as serine only because the $tRNA_{CAG}^{Ser}$ has a A_{37} instead of m^1G_{37} (Figure 1. 24) and the LeuRS is not able to recognize it (Suzuki *et al.*, 1997).

Another intriguing structural feature of this $tRNA_{CAG}^{Ser}$ is the presence of guanosine at position 33. All other eukaryotic elongation tRNAs have a highly conserved uridine at position 33 (U_{33}), which is required for the correct turn of the phosphate backbone (U-turn) and stacking of the anticodon bases (Ladner et al., 1975; Woo et al., 1980). The G_{33} mutation may have had an important role on CUG reassignment in *Candida* species (Suzuki et al., 1997; Santos et al., 1996; Santos et al., 1997), since it may have lowered the leucylation levels of the tRNA. Indeed, replacement of G_{33} with pyrimidines in *C. zeylanoids* $tRNA_{CAG}^{Ser}$ has increased its leucylation level (Suzuki *et al.*, 1997). But, it may have also played a role in lowering the decoding efficiency of the tRNA at the ribosome, since U_{33} stabilizes tRNA-rRNA interactions during translation (Ashraf *et al.*, 1999) and makes decoding more efficient.

In vitro, the $tRNA_{CAG}^{Ser}$ from *C. zeylanoids* can be charged with up to 30% with leucine (Suzuki *et al.*, 1997). However, the *in vivo* level of the mischarged $tRNA_{CAG}^{Ser}$ (Leu-$tRNA_{CAG}^{Ser}$) is only a 3%, thus showing that this mischarging event is repressed under physiological conditions.

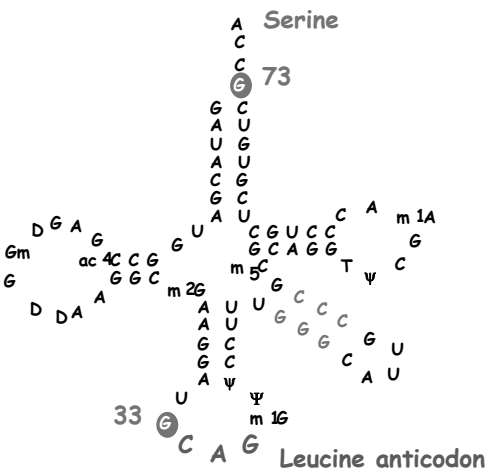


Figure 1. 24 – The secondary structure of the $tRNA_{CAG}^{Ser}$.
The *C. albicans* $tRNA_{CAG}^{Ser}$ is an hybrid tRNA with identity elements for both leu- and ser-tRNAs. Its anticodon arm is characteristic of the $tRNA^{Leu}$, whereas the acceptor stem and the variable region are characteristic of $tRNA^{Ser}$. Position 33 is highlighted as it was critical for the CUG reassignment from leucine to serine. The discriminator base (G_{73}) belongs to the serine family tRNAs (Santos *et al.*, 1993).

1.6.2. The evolution of CUG codon reassignment

Comparative genomics and molecular phylogeny studies have shown that the novel $tRNA_{CAG}^{Ser}$ appeared 272±25 million years ago, before the divergence of *Candida* and *Saccharomyces* genera. Therefore, the ancestor of yeasts was ambiguous and it is not yet clear why the mutant $tRNA_{CAG}^{Ser}$ was selected in the *Candida* lineage and lost in the *Saccharomyces* lineage. Furthermore, the existence of *Candida* species, namely *C. glabrata* and *C. krusei* that still decode the CUG codon as leucine, reinforces the idea that the evolution of CUG ambiguity is a special event that introduced some selective advantages in some, but not all, *Candida* species (Santos et al., 1993; Santos and Tuite, 1995; Suzuki et al., 1997; Yokogawa et al., 1992; Sugita and Nakase, 1999).

The complete pathway of the CUG identity alteration is not yet fully understood, however, molecular phylogeny studies, carried out by Massey *et al.* (2003), have revealed that the $tRNA_{CAG}^{Ser}$ originated from a serine rather than a leucine tRNA. This is in agreement with the proposal of Suzuki and colleagues (1994), who hypothesized that the CAG anticodon resulted from an insertion of an adenosine between the first two nucleotides of the CGA anticodon. The $tRNA_{CGA}^{Ser}$ gene has an intron located on the 3´-side of position 37 in the anticodon-loop and insertion of a single adenosine in the middle of the 5´-CGA-3´anticodon sequence would create the 5´-CAG-3´ anticodon (Suzuki *et al.*, 1994).

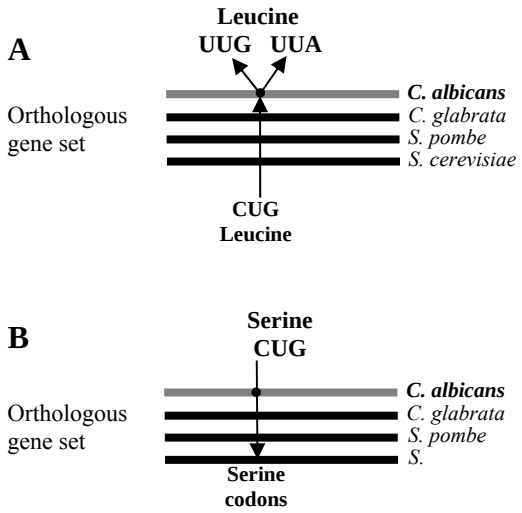


Figure 1. 25 – The mutational pressure on *C. albicans'* genome.
(**A**) Ambiguous CUG decoding forced a change of "old" CUG codons to leucine codons UUG and UUA. (**B**) Simultaneously, "new" CUG codons appeared via mutation of UCN serine codons (Massey *et al.*, 2003)

The appearance of such mutant $tRNA_{CAG}^{Ser}$ creates an ambiguous decoding of the CUG codon, since there were two distinct tRNA species, and so any CUG codon could be decoded as leucine by the cognate $tRNA_{CAG}^{Leu}$ and as serine by the mutant $tRNA_{CAG}^{Ser}$. (Santos et al., 1996; Massey et al., 2003). This ambiguous decoding of the CUG codon decreased CUG usage (Figure 1. 25) (Massey *et al.*, 2003). Indeed, only 2% of CUG codons existent in the ancestor of yeasts are still present in *C. albicans* and most likely in the genomes of other *Candida* species. The 13,074 CUG codons now present in the *C. albicans* haploid genome evolved after the appearance of the $tRNA_{CAG}^{Ser}$, over the last

272±25 MY, from codons coding for serine or amino acids with similar chemical properties and not from codons coding for leucine (Massey *et al.*, 2003). Therefore, the CUG codons that actually exist in the genome of *C. albicans* are "new" and have no relationship with CUG codons present in non-ambiguous yeasts, such as *S. pombe* or *S. cerevisiae*.

1.6. Objectives of this study

Despite the important progress made to date on the study of genetic code alterations we are still far from understanding their evolution at the molecular level. The uniqueness of CUG identity change and the availability of molecular biology tools, robust genome analysis methods and the availability of the *C. albicans* genome sequence make this fungus an interesting model system to study the evolution of genetic code alterations. The aim of this work was to contribute to better understand how the CUG codon changed identity from leucine to serine and shed new light on how this unusual event shaped the evolution of the genus *Candida.* Finally, we hoped that this study would contribute to shed new light on the evolution of the genetic code, in particular on its expansion during the early stages of its development and on the evolution of tRNA and aminoacyl-tRNA synthetases. In order to achieve this, we have defined the following objectives for this project:

1) To investigate whether the CUG codon is decoded as both serine and leucine *in vivo* in *C. albicans*. In other words, does the translational machinery discriminate between Ser-$tRNA_{CAG}$ and Leu-$tRNA_{CAG}$ or does the mischarged leu-$tRNA_{CAG}$ participate in protein synthesis?
2) To quantify misincorporation of leucine *in vivo* under different physiological conditions.
3) To increase CUG ambiguity *in vivo* in *C. albicans*.
4) To evaluate the impact of the ambiguous CUG decoding event.
5) To study the mechanism of interaction between the $tRNA_{CAG}^{Ser}$ and both the Leucyl- and Seryl-tRNA synthetases.

2. Materials & Methods

2.1. Strains and Growth Conditions

2.1.1. Strains and genotypes

- *Escherichia coli*

JM109, genotype: *recA1 SupE44 endA1 hsdR17* (rk–, mk+) *gyrA96 relA1 thi Δ(Lac-proAB) [F', traD36, proAB, lacI*^{q}*lacZ ΔM15*]

BL21-Codon Plus®, from Stratagene, genotype: *E. coli* B F$^-$, *ompT, hsd*$S_\beta(r_\beta^- m_\beta^-)$, *dcm*$^+$, Tet^r, *gal* λ(DE3) *endA* Hte [*argU ileY leuW* Cam^r]

XL1, genotype: *recA1 endA1 gyrA96 thi-1 hsdR17 SupE44 relA1 lac [F' proAB lacI*^{q}*ΔM15* (Tet^r)]

- *Candida albicans*

CAI-4 (*ura3Δ::imm434/ura3::imm434*).

Strains *2005*, *1006*, *C316* and *IGC* were obtained by Santos (Santos *et al.*, 1994). All of them are wild type stains, *C316* is a clinical isolate, and the IGC strain was isolated from tree leaves.

- *S cerevisiae*

CEN-PK2 (MAT a/α, ura3-52/ura3-52, trp1-289/trp1-289, leu2-3 112/leu2-3 112, his31/his31).

W303 (mat alpha ade 2-1 can1-100 his3-11-15 trp1-1 ura3-1)

J940557 (MAS5) – wild type, clinical isolate

J940610 (MAS4) – wild type, clinical isolate

2.1.2. Growth and Maintenance of *E. coli*, *S. cerevisiae* and *C. albicans*

Escherichia coli strains were grown at 37ºC on LB broth [1% (w/v) peptone from casein, 0.5% (w/v) yeast extract, 1% (w/v) sodium chloride; (Merck)] or on LB/2% (w/v) agar. Transformed *E. coli* strains were grown in the LB-Amp [LB, 50 µg/ml ampicillin

sodium (Duchefa, Haarlem)]. Strains were stored at -80ºC in 0.5 LB-Amp/20% (v/v) glycerol.

Wild types *S. cerevisiae* strains were grown at 30°C on YEPD (2% glucose; 1% yeast extract, 1% peptone).

Wild type *Candida albicans* strains were grown at 30°C on YEPD (2% glucose; 1% yeast extract, 1% peptone), whereas transformed strains were grown in MM-URA (0.67% yeast nitrogen base without amino acids, 2% glucose, 2% agar and 100µg/ml of each required amino acids, without uracil). Transformed strains were stored at -80ºC in 0.5 MM-URA/40% (v/v) glycerol.

For the measurement of ambiguous CUG decoding under different physiological conditions, slight changes were made to the growth conditions, namely:

- Opaque cells were grown at 25ºC on MM-URA
- Heat stress: growth was on MM-URA at 37ºC.
- Oxidative stress: growth was at 30ºC on MM-URA with 1.5 mM H_2O_2.
- Low pH: growth was at 30ºC on MM-URA buffered with citrate buffer (sodium citrate – acid citric) pH 4.0

2.2. DNA Manipulation

Generally, unless otherwise stated, all DNA manipulations were performed as described in Sambrook *et al.* (1989).

2.2.1. Oligonucleotides

Oligonucleotides were purchased from MWG-Biotech AG (Germany) and were resuspended in ultra pure milliQ (mQ) water to a final concentration of 100 mM.

Table 2. 1 – List of the oligonucleotides used.

Oligo	Sequence (5' → 3')	Tm (ºC)
Construction of the reporter protein		
oUA201	*ATTAGGAAGCTTAGTGTTGCGTGTGTGTCAG*	*58*
oUA202	*TTATCCCTCGAGACCGTTTGGTCTACCCAAG*	*58*
oUA204	*AATTTTCTGCAGCCTTTTGGTGTACGAGAG*	*54*
oUA205	*CTCAACTCGCGAGCTAGTTGAATATTATGTAAGATCTG*	*68*
oUA215	*ACTAGACCGCGGGATTATAAAGATGATGATGATAAGAACGACAAATACTCATTAGC*	*54*
oUA216	*ATTAGATCGCGATTAGTGATGGTGATGGTGATGGTTTTTGTTGGAAAGAGCAAC*	*58*
oUA217	*TCCAGTTGTCTGGAATACC*	*56*
oUA224	*TTCCAACTCAATTCACTCCTC*	*60*
oUA225	*ACCCAAAATGGCCAAGAATGG*	*60*
Sequencing of *C. albicans* LeuRS gene		
oUA711	*GTGCGAGTAGGAGTGCC*	*50*
oUA712	*GGTGTCTTGCACGCCG*	*50*
oUA713	*CTAGAGTTGATTGGAGACG*	*48*
oUA714	*GATGCTGGTAATGGTGAC*	*48*
oUA715	*GTGCAGTTGGCCAACGC*	*48*
oUA716	*GTCGAATCTTTGTCAGATTC*	*48*
oUA717	*GGAGCTGATGCCTCTAG*	*52*
oUA718	*GCCGAATACCTTTACAGAG*	*52*
oUA719	*AAAGCCAGGGCTCATAG*	*48*
oUA721	*GAATCTGACAAAGATTCGAC*	*48*
oUA723	*CAGCATCTTCAGTTGCC*	*52*
oUA724	*GGAGAATCTGATGGAACAC*	*52*
oUA740	*GGGAATAATGCTCTCATACC*	*50*
oUA741	*CCATAATCCACTTTTC*	*50*

Sequencing of *S. cerevisiae* LeuRS gene		
oUA749	*ATAAATCATAATCACGTAAAGC*	*46*
oUA750	*CATCTAATAAAGGCATCG*	*46*
oUA751	*CTTCACTTGGGGTCG*	*46*
oUA752	*TTGTTTTTGGCTTGTTCG*	*46*
oUA753	*AAGGAAGATTACTACACTG*	*48*
oUA754	*TAGCAGCATTAGCGTTAG*	*48*
oUA755	*TTGCGTTTGCCGATGCG*	*48*
oUA756	*TTCTGGTTGCTGTTTATTG*	*48*
Sequencing of *C. albicans* SerRS gene		
oUA705	*CGA TCC AGA AAG AGG GG*	*54*
oUA733	*GATTTTCTTTTTTTCTGATACAT*	*52*
oUA734	*CCCACCACCACAACCC*	*52*
oUA736	*ATTAGTGCTTACCATGCCGG*	*60*
oUA738	*CCGGCATGGTAAGCACTAAT*	*60*
Sequencing of *C. albicans* TrpRS gene		
oUA759	*TACAAAATGGTTACAAGAAG*	*52*
oUA760	*GCCCAAGAATGAGTGAGAC*	*52*
oUA761	*GCAAAGCATAGAGGGGTC*	*52*
oUA762	*GGGGTCTTTGGTGGTAATC*	*52*
Site directed mutagenesis of LeuRS and SerRS		
oUA261	*CTGTTTTTCTAAAGCTCCTGCTGATGACGAAGATGCAG*	
oUA262	*CTGCATCTTCGTCATCAGCAGGAGCTTTAGAAAAATCAG*	
oUA263	*GAACTTTTCAAGAAAGAGAGTCTCGATGTGAAGGAGAA*	
oUA264	*GTTCTCCTTCACATCGAGACTCTCTTTCTTGAAAAGTTTC*	
oUA265	*AACCAAGACAAGTTAAGAACTGGTGACTACGATTCCTTC*	

oUA266	*GAAGGAATCGTAGTCACCAGTTTCTTAACTTTGTCTTGGTT*	
oUA267	*GCTATTCTTGATGCTCTGGAATATGTCAGAAGCCTTACC*	
oUA268	*GGTAAGGCTTCTGACATATTCCAGAGCATCAAGAATAGC*	
Construction of tRNA overexpression system		
oUA218	*AATTTCAAGCTTACTAGTTGAAACACC*	*52*
oUA219	*CTCAATCTCGAGCCCACAGATGATTGAC*	*48*
oUA220	*ATAGGACTGCAGACTAGTTGAAACACC*	*52*
oUA221	*TTATCCAAGCTTCCCACAGATGATTGAC*	*48*

2.2.2. Plasmids

2.2.2.1. Original plasmids

Table 2. 2 – Original plasmids used for obtaining the necessary DNA constructions.

Name	Description	References/ Supplier
pSL1190	*E. coli* vector containing the Amp^R gene, thus allowing for selection of transformants in media containing ampicillin. Unique cleavage sites, on the multicloning site of the plasmid were used to insert the desired DNA fragments, namely the *Hind* III, *Xho* I, *Nru* I and *Pst* I sites.	Pharmacia
pUA12	Constructed by Miranda using the *C. albicans* pRM1 vector , which is an autoreplicative shuttle vector constructed by Pla and colleagues. It contains two Autonomously Replicating Sequences (ARS): ARS2 and ARS 3 and two auxotrophic *C. albicans* markers (URA3 and LEU2) for replication and selection in yeasts. It also has an ampicillin resistant marker to allow for DNA manipulation in *E. coli*. The pUA12 has a multiple cloning site, inserted in the LEU2 promoter region at the *NruI/EcoRV* cleavage sites.	(Pla et al., 1995; Miranda, 2007)

pUA15	Constructed by Miranda and is based on pUA12. It contains a copy a *S. cerevisiae* $tRNA_{UAG}^{Leu}$, whose 5´-UAG-3´anticodon was mutated to the 5´-CAG-3´ anticodon for cognate decoding of CUG codons.	(Miranda, 2007)
pUKC701	Constructed by Santos and colleagues. It is based on pSL315, which contains the LEU2 auxotrophic marker for *S. cerevisiae and* the Amp^R marker for selection of E. coli in ampicillin media. This is a single copy plasmid in *S. cerevisiae*. The *C. albicans* $tRNA_{CAG}^{Ser}$ was cloned in the *Sma* I and *Spe* I sites of this vector's multicloning site.	(Santos et al., 1999; Sikorski and Hieter, 1989)
pUKC1710	Constructed by O'Sullivan for overexpression of the *C. albicans* LeuRS (strain 2005). This plasmid is based on the pET-15 expression system, from Novagen.	(O'Sullivan *et al.*, 2001b)
pUKC1722	Constructed by O'Sullivan for the overexpression of the *C. albicans* SerRS (strain 2005). This plasmid is based on the pET-15 expression system, from Novagen.	(O'Sullivan *et al.*, 2001a)
pUA301	Plasmid constructed by Santo, resulting from site directed mutagenesis of the CUG codon of the SerRS cloned into plasmid pUKC1722. The CUG codon was mutated to the serine TCG codon.	Unpublished

2.2.2.2. Constructed plasmids

Table 2. 3 – Plasmids constructed in this work.

Name	Description
pUA61	*E. coli* plasmid based on the pSL1190 vector. This plasmid was used to assemble the CUG reporter system used for measuring CUG ambiguity in *C. albicans*. For this, the reporter gene was assembled in three sequential steps, using the restriction sites *Hind* III, *Xho* I, *Nru* I and *Pst* I.
pUA63	*C. albicans* plasmid, based on the pUA12 shuttle vector. The whole reporter gene was extracted from pUA61, using the *Hind* III and *Pst* I restriction sites, and was inserted at the same restriction sites of pUA12.

pUA65	*C. albicans* plasmid based on pUA15. Contains copy the *S. cerevisiae* $tRNA_{UAG}^{Leu}$ gene. Again, for this plasmid, the whole reporter gene was transferred from pUA61 as a *Hind* III and *Pst* I fragment and inserted in pUA15.
pUA74	Plasmid based on pUKC1710 for overexpression of *C. albicans* LeuRS in *E. coli.* This plasmid contains the isoform-A of the *Ca*LeuRS from CAI-4 strain, created by mutation of the CUG codon to the serine replaced by a serine TCG-codon by site directed mutagenesis.
pUA81	Plasmid based on pUA74. It was used for the overexpression of isoform-B of the CAI-4 *C. albicans* LeuRS in *E. coli.* The LeuRS isoform-B was obtained by site directed mutagenesis that altered all the non-silent SNPs. As in pUA74, its CUG codon is replaced by the serine TCG-codon.
pUA82	Plasmid based on pUA74. It was used for overexpression of *C. albicans* LeuRS in *E. coli.* This plasmid contains the isoform-A of the *Ca*LeuRS, with a CUG codon which is decoded as leucine in *E. coli*.
pUA83	Plasmid based on pUA81. It was used for overexpression of *C. albicans* LeuRS in *E. coli.* This plasmid contains the isoform-B of the *Ca*LeuRS, with a CUG codon which is decoded as leucine in *E. coli*.
pUA72	Intermediate plasmid with two copies of the *C. albicans* wild type $tRNA_{CAG}^{Ser}$ gene. It was constructed using pUKC701 as the base plasmid. The second $tRNA_{CAG}^{Ser}$ was inserted at its *Hind* III and *Xho* I restriction sites.
pUA73	Intermediate plasmid with three copies of the $tRNA_{CAG}^{Ser}$ gene constructed upon the pUA72. The third $tRNA_{CAG}^{Ser}$ was inserted at the *Pst* I and *Hind* III restriction sites.
pUA77	Plasmid constructed for the overexpression of the $tRNA_{CAG}^{Ser}$ *in vivo* in *C. albicans*, based on pUA12. A *Xho* I and *Apa* I DNA fragment containing three copies of the $tRNA_{CAG}^{Ser}$ gene in tandem, was extracted from the pUA73 and it was then inserted in the same restriction sites of the multicloning site of pUA12.

2.2.3. DNA amplification by PCR

DNA fragments were amplified by polymerase chain reaction (PCR) from plasmid or genomic DNA templates. Reactions were carried out using 5.0 ng.μL^{-1} of template DNA, in a mixture of 1mM dNTPs, 10µM of forward primer, 10µM of reverse primer, 2.0 mM $MgCl_2$, 10 mM Tris-HCl (pH 9.0), 50 mM KCl, 0.1% TritonX-100 and 0.05 U.μL^{-1} of *Taq* Polymerase (Fermentas or Bioron).

PCR reactions were performed in a Mastercycler (Eppendorf), for 25 cycles of 30s at 92ºC for DNA melting, 30s at the desired Tm, to promote the template-primer annealing, and finally 30s-90s at 72ºC for DNA elongation (the duration of this step was dependent of the length of the PCR product). An additional initial melting step for 2 min at 92ºC and a final elongation step for 3 min at 72ºC were also carried out.

The Tm was set according the primers melting temperature, which is indicated on the 3rd column of Table 2. 1.

2.2.4. PCR product purification

After PCR reactions, primers, nucleotides, enzymes and salts, were removed from the amplified DNA using the QIAquick PCR Purification Kit (Qiagen), as described by the manufacturer.

2.2.5. Agarose Gel electrophoresis

DNA molecules were fractionated on agarose gels. Multi-Purpose agarose (Boehringer Mannheim) was melt using a microwave oven in TAE [40 mM Tris-acetic acid, 10 mM *E*thylene*D*iamine*T*etr*A*cetic acid (EDTA), pH 8.0] at concentrations ranging from 0.8 to 1.0% (w/v). Ethidium bromide (EtBr) (Invitrogen) was added to the melted agarose to a final concentration of 0.2 µg.mL^{-1}, and gels were then casted on BioRad casting systems. DNA samples were mixed with 6x loading buffer [0.25% (w/v) of bromophenol blue, 0.25% (w/v) of xylene cyanol, 30% (v/v) glycerol] in 1:6 ratio, loaded

into the wells and finally eletrophoresed at 70 V (Power Pac 3000, Bio-Rad) for one hour in submerged horizontal electrophoresis systems (Mini-Sub Cell GT, Bio-Rad).

For DNA visualization, the eletrophoresed samples gels were exposed to U. V. light using a Gel Doc 2000 Gel Documentation System (BioRad) coupled to a PC. The images were acquired and analyzed with the Quantity One software (Bio-Rad).

2.2.6. DNA extraction from agarose gel

For DNA purification from agarose, 0.8% low-melt SeaKem® Gold agarose (Flowgen) gels were used. Bands corresponding to desired DNA fragments were removed from the gel with the QIAEX II Kit (Qiagen), as describe by the manufacturer, with slight adaptations. Briefly, gels were prepared without EtBr and DNA samples were electrophoresed for 60 min at 70V. After electrophoresis, gels were stained for 10 min in 100 mL of TAE containing 0.5 $\mu g.ml^{-1}$ EtBr and were then washed in dH_2O for 10 min. DNA was visualised by UV light, excised using a clean scalpel and transferred to a clean microcentrifuge tube. Agarose slices were weighted and QX1- Buffer was added (3 volumes buffer : 1 volume of gel). Gels were disrupted with pipette tips and buffer [10 µL of 3M Sodium Acetate (NaOAc)], pH 4.5, was added. The sample was incubated at 50ºC until the gel was completely melted. Afterwards, 10 µL of QIAEX II resin was added and samples were kept at 50ºC for more 5 min. with gently vortexing. Samples were then centrifuged for 30 sec, the supernatant discarded and the pellet washed twice with 500 µL of PE-Buffer. Finally, the pellet containing the resin with the DNA was air-dried for 15 min and incubated with 50 µL of mQ dH_2O for 5 min at room temperature for DNA elution. The DNA was recovered after a centrifugation step at 16 000*g* for 30 sec.

2.2.7. DNA digestion with restriction enzymes

Restriction digestions were performed to prepare DNA for cloning and to screen positive clones for confirming the DNA ligation and insertion into the cloning vectors. Digestions of up to 5 µg of DNA were performed in 20 µL reactions with the required enzymes (Fermentas) and appropriate buffer, for periods of time ranging 3 h to overnight,

at 37°C. DNA digestion was verified using agarose gel electrophoresis (as described above).

2.2.8. DNA Dephosphorylation and ligation

To prevent self-ligation DNA vectors were treated with alkaline phosphatase. 20 µL reactions were prepared with 2 µg of digested vector DNA, 2 Units of shrimp alkaline phosphatase (SAP) (Roche), 2 µL of 10x dephosphorylation buffer (0.5 M Tris-HCl, 50 mM $MgCl_2$, pH 8.5). Reactions were carried out at 37°C for 1 h and then SAP was inactivated at 65°C for 15 min. For DNA cloning, ligations of digested DNA fragments were performed with T4 DNA ligase (Gibco BRL or Fermentas). Routinely, 10-30 fmol of vector DNA were mixed with 30-90 fmol of insert DNA fragments, in four independent ligation reactions, with different vector: insert molar ratios, namely 1:0 (negative control), 1:1, 1:2 and 1:5. The reactions were carried out in a 1.5 ml microcentrifuge tube containing 4 µL of 5x Ligase Reaction Buffer [250 mM Tris-HCl (pH 7.6), 50 mM MgCl2, 5 mM ATP, 5 mM dithiothreitol (DTT), 25% (w/v) polyethylene glycol-8000], 5 Units T4 DNA Ligase. Reactions volumes were adjusted to 20 µL with H_2OmQ and ligations were incubated overnight at 12°C.

2.2.9. Transformation of *E. coli*

E. coli cells were routinely used as hosts for manipulation of recombinant DNAs. For preparation of competent cells, a fresh *E. coli* colony was inoculated in 5 ml of LB and was grown at 37°C, overnight, with vigorous shaking (200 rpm). Fresh 5 ml LB cultures were then inoculated, with 200 µL of the overnight culture, and were grown, at 37°C to an OD_{550} of 0.3. 100 ml LB cultures were then inoculated with 4 ml of the previous cultures and allowed to grow to an OD_{550} of 0.3 at 37°C, with shaking. At this point, cultures were incubated on ice for 5 min and centrifuged at 500 *g*, for 5 min at 4°C. Pellets were resuspended gently in 40 ml of cold TFB I [100 mM RbCl, 50 mM $MnCl_2.4H_2O$, 30 mM potassium acetate (KOAc), 10 mM $CaCl_2.2H_2O$, 15% (w/v) glycerol, pH 5.8], and cells were collected by centrifugation at 500 *g* for 5 min at 4°C. Finally, pellets were resuspended in 5 ml TFB II [10 mM 4-Morpholinepropanesulfonic acid (MOPS), 10 mM

RbCl, 75 mM $CaCl_2$, 15% (w/v) glycerol, pH 6.8] and were distributed in 200 µL aliquots into ice cooled microcentrifuge tubes. Cells were then directly used for transformation or flash frozen in dry ice and stored at -80°C.

For transformations, 200 µL of "competent" cells were incubated on ice, for 30 min, with 10-100 ng of DNA (or 10 µL of ligation reaction). Cells were submitted to heat shock at 42°C, for 90s and immediately incubated on ice for 2 min. Afterwards, cells were allowed to regenerate in 800 µL of SOC medium [20 mM glucose, 2% (w/v) tryptone, 0.5 % (w/v) yeast extract, 0.05% (w/v) NaCl, 2.5 mM KCl, pH 7.0], which was added to the mixture and were incubated at 37°C for 1 h with 200 rpm agitation. Cells were centrifuged for 20s at 500*g,* and 800 µL of the supernatant was discarded. Pellets were resuspended, with the remaining supernatant, and plated on LB/Amp agar. Plates were incubated overnight at 37°C.

2.2.9.1. Plasmid DNA preparation

Rapid plasmid mini preparations were carried out from colonies picked up from LB-agar-Amp plates. For this, colonies were inoculated into 5 ml LB-Amp and allowed to grow overnight at 37°C with agitation (200 rpm). 1.5 mL were transferred to microcentrifuge tubes, cells were centrifuged at 15,000 *g* for 5 min at room temperature and pellets were resuspended in 100 µL of solution I (50 mM glucose, 25 mM Tris pH 8.0, 10 mM EDTA, pH 8.0), and then in 200 µL of solution II [0.2 M NaOH, 1% (w/v) Sodium dodecyl sulphate (SDS)]. After mixing, 150 µL of cold solution III [3 M KOAc, pH 5.0] was added, mixed by inverting tubes, and then incubating them on ice for 5 min. Samples were centrifuged at 15,000 *g*, for 5 min at 4°C to remove cell debris, and DNA containing supernatants were recovered into clean 1.5 ml microcentrifuge tubes. The DNA was precipitated with 1 volume of isopropanol at room temperature for 10 min and then centrifuged at 15,000 *g*, for 5 min at 10°C. Pellets were washed with 1 ml of cold 70% (v/v) ethanol and centrifuged at 16,000*g* for 5 min at 4ºC. DNA pellet was dried at 37°C and resuspended in 20 µL of sterile mQ H_2O.

For high quality DNA mini preparations,, QIAprep Miniprep Kits (Qiagen) were used as described by the manufacturer's instructions. The optional wash with Buffer PB was always done.

Large scale DNA plasmid preparation (*Maxi-Prep*) was carried out using the GenElute™ Plasmid Maxiprep Kit (Sigma), according to the manufacturer's instructions, with minor changes. Briefly, cells from 200 mL overnight cultures were harvested by centrifugation at 5,000*g* for 10 min at room temperature. Pellets were resuspended with 6.0 mL of the Resuspension Solution with RNase A. Cells were then lysed with 6.0 mL of Lysis Solution and lysis was allowed to proceed for 5 min, and then 8.0 mL of Neutralization/Binding Solution was added. Cellular debris were pelleted by centrifugation at 15,000*g* for 20 min at 4ºC and supernatants were loaded into the GenElute Maxiprep binding column, which was then centrifuged at 5,000*g* for 1 min at 4ºC and the flow-through was discarded. The column was washed with 8.0 mL of the Optional Wash Solution and centrifuged at 5,000*g* for 1 min at 4ºC and the flow-through discarded. The final column wash was done with 15 mL of Wash Solution and was centrifuged at 5,000*g* for 5 min at 4ºC, the flow-through was discarded and the column was again centrifuged for 1 min to dry up the resin. Finally, DNA was eluted by adding 5.0 mL of sterile mQ dH_2O and centrifugation at 5,000*g* for 5 min at 4ºC.

2.2.10. Site Directed Mutagenesis

In vitro site directed mutagenesis was carried out with the QuikChange Site-Directed Mutagenesis Kit from Stratagene, according to the manufacturer instructions. However, we usually used 25 µL rather than 50 µL reactions which are recommended by the manufacturer. Two synthetic oligonucleotides primers complementary to both strands of the plasmid, containing the desired mutation in the middle, were used to extent the plasmid during amplification with PfuTurbo™ DNA polymerase. Primers contained 35 and 40 bases in length and melting temperature higher than 80ºC. PCR reactions were performed in 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 20 mM Tris-HCl, 2 mM $MgSO_4$, 0.1% Triton® X-100, 0.1 $mg.mL^{-1}$ Bovine Serum Albumin (BSA), pH 8.8 with 0.2 mM of each dNTPs. The reactions contained 5 to 10 ng of DNA template, 60 - 70 ng of each primer and *PfuTurbo*

DNA polymerase at a final concentration of 0.5 U.μL^{-1}.The PCR programs consisted of a first cycle at 95ºC for 30 s, followed by 18 cycles of 95ºC for 30 s, 55°C for 1 min and 68°C for 20 min. The amplification was checked by 0.8% agarose gel electrophoresis. After visualizing bands in gels, the original DNA templates were digested with 5 U of *Dpn* I for 1h at 37ºC.

Dpn I treated-DNA was transferred to 50 µL of XL1-Blue competent cells (supplied by the manufacturer) and gently mixed. Transformations proceeded with 30 min incubation on ice, followed by a heat pulse of 45 s at 42ºC and were then cooled on ice for 2 min. Cells were allowed to recover in 0.5 mL of NZY^+ broth [1%(w/v) NZ amine, 0.5% (w/v) yeast extract, 0.5% (w/v) NaCl, 0.4% (w/v) glucose, 12.5 mM $MgCl_2$, 12.5 mM $MgSO_4$, pH adjusted to 7.5 with NaOH], preheated at 42ºC, for 1h at 37ºC with shaking at 180 rpm. Cells were then centrifuged for 20s at 500*g* and 300 µL of supernatant were discarded and cell pellets resuspended, with the remaining supernatant. Finally, cells were plated on LB-Amp agar and incubated overnight at 37°C. From each transformation four colonies were isolated and their plasmid DNA was extracted and sequenced, as described in sections 2.2.9.1 and 2.2.12, respectively

2.2.11. Nucleic Acids precipitation and quantification

Nucleic acids were precipitated with NaOAc and ethanol. To the DNA or RNA solutions 0.1 volumes of 3 M NaOAc, pH 4.6 and 3 volumes of ethanol were added, so that it would have a final concentration of 0.3 M NaOAc, pH 4.6 and 70% of ethanol. Solutions were routinely incubated at -30ºC for periods ranging from 2h to overnight, after which samples were spun at 16,000*g* for 15 min at 4ºC. Supernatants were discarded and pellets were washed with 500 µL of 70% (v/v) ethanol and spun again for 16,000*g* for 10 min at 4ºC. The pellets were then dried, resuspended in water and quantified by UV Spectrometry, at wave-lengths of 260 nm and 280 nm, considering that 1 unit of absorbance at 260 nm corresponds to 50 $\mu g.mL^{-1}$ of dsDNA and 40 $\mu g.mL^{-1}$ of RNA. Since proteins have maximal UV absorbance at 280 nm, the A_{260}/A_{280} was used as a

measure of nucleic acid solutions quality. DNA and RNA preparation with ratios between 1.7 and 2.2 were used in further manipulations.

2.2.12. DNA sequencing

DNA samples were prepared for sequencing following the ABI PRISM® BigDye™ Terminator Cycle Sequencing Ready Reaction Kit protocol, with AmpliTaq® DNA Polymerase, FS (*PE* Applied Biosystems). Briefly, 20 µL sequencing reactions were prepared with 200-500 ng of template DNA, 3.2 pmol of primer and 4 µL of Terminator Reaction Mix. PCR programs had an initial step of 2 min at 96ºC, followed by 25 cycles of heating at 96°C for 10 s, 50°C for 5 s and 60°C for 4 min. The extension products were purified by precipitation, the pellets were dried at room temperature and resuspended in 20-25 µL of Template Suppression reagent. Samples were heated at 95°C for 2 min, to allow for denaturation and were kept on ice until loading on the ABI Prism 377 DNA Sequencer (*PE* Applied Biosystems), according to the ABI Prism 310 Genetic Analyzer User's Manual.

2.2.13. Transformation of *C. albicans*

The transformation protocol for *C. albicans* was based on the protocol described in the "Manual for the Preparation and Transformation of *Pichia pastoris* Spheroplasts" Version A from Invitrogen. 200 ml of *C. albicans* CAI-4 cultures were routinely prepared overnight in YEPD, at 30°C, with 180 rpm agitation. Cells were harvested, when the culture reached an OD_{600} between 0.2 and 0.3, by centrifugation at 3,200*g* for 10 min at room temperature. The pellets were washed in 20 ml of sterile distilled water resuspended in 20 ml of fresh SED [19 ml of SE (1 M sorbitol, 25 mM EDTA, pH 8.0) with 1ml of 1 M DTT], and centrifuged at 3,200*g* for 5 min at room temperature. They were then washed with 20 ml of 1 M sorbitol and centrifuged at 3,200*g* for 5 min at room temperature. Cells were finally resuspended in 20 ml of SCE buffer (1 M sorbitol, 1 mM EDTA, 1 mM sodium citrate, pH 5.8). The cell suspensions were divided into two tubes containing 10 ml each. One tube was used to monitor spheroplast formation and the other was kept at room temperature and later used for transformation.

C. albicans spheroplasts were then prepared by adding 60 µg of Zymoliase 100 T (Seikagaku Corp) to the 10 mL of previously treated *C. albicans* cells and incubating at 30ºC until 80% of spheroplasts were obtained. For this, a primary time course assay was performed with 10 mL of cells. Spheroplasts were quantified by collecting fractions of 200 µL of cell suspension at different time points and adding 800 µL of 5% SDS (v/v) to each fraction. Their absorbance at 800 nm was immediately measured and the spheroplasts were quantified using the following equation:

$$\% \text{ Spheroplasts} = 100-[(OD_{800} \text{ of fraction } t_X/OD_{800} \text{ of fraction } t_0) \times 100]$$

Where t_X corresponded to fractions collected at 2, 4, 6, 8, 10, 15 min and so on, until the percentage of spheroplasts reached the needed 80%. The blank control used consisted of 200 µL of SCE buffer mixed with 800 µL of 5% (v/v) SDS. For transformation, spheroplasts prepared as above were harvested by centrifugation at 750 *g* for 10 min at room temperature and resuspended in 10 mL of 1 M sorbitol, they were again harvested, and washed with 10 mL of CaS buffer (1 M sorbitol, 10 mM $CaCl_2$, 10 mM TrisCl, pH 7.5). Finally, spheroplasts were harvested and resuspended 0.6 mL of CaS. They were immediately used for transformation. For each transformation, 100 µL of spheroplasts were dispensed into 1.5 mL microcentrifuge tubes, and to each aliquot the plasmid DNA was added, in quantities ranging from 3 to 12 µg. Additionally, herring YeastMarker DNA carrier (Clonotech) was added. The DNA was allowed to get into the cells for 10 min at room temperature. Afterwards, each reaction was gently mixed with 1 mL of fresh PEG/CaT [20% (w/v) Polyethylene Glycol (PEG) 3350, 10 mM $CaCl_2$, 10 mM Tris, pH 7.5] and centrifuged at 750*g* for 10 min at room temperature. The supernatant was discarded and the pellet resuspended in 150 µL of SOS medium (1 M sorbitol, 0.3xYPD, 10 mM $CaCl_2$) and incubated for 30 min at room temperature. Finally, cells were plated on MM-Ura agar plates and incubated at 30ºC, for 5-7 days to allow for colony formation

2.2.14. *C. albicans* genomic DNA extraction

Genomic DNA of all strains and species was extracted using the Wizard Genomic DNA Purification Kit (Promega), according to the manufacturers' instructions.

2.3. Protein Extraction, Purification and Analysis

2.3.1. Protein Extraction

Candida albicans proteins were extracted from cultures grown to OD_{600} of 0.3. For this, cells were collected by centrifugation, for 5 minutes at 4000*g*, and lysed in a lysis solution containing 6 M Urea, 100 mM NaH_2PO_4, 10 mM Tri-Cl, 0.01 % Triton X-100, 7.5 % Glycerol, 2.0mM phenylmethanesulphonylfluoride (PMSF), pH 8.0, and a cocktail of EDTA-free protease inhibitors (Roche). Lysis was carried out in a BeadBeater (BioSpec Products) with 15 cycles shaking for 1 minute with 3 minutes resting on ice.

Recombinant LeuRS and SerRS overexpressed in *E. coli* BL21-CodonPlus® were prepared from 750 ml cultures. For this, 10 ml-LB/Amp overnight cultures were used to inoculate 50 mL-LB/Amp which were allowed to grow at 37ºC to an OD_{600} of 0.6. Then, 4 ml of these fresh cultures were used to inoculate 750 mL-LB/Amp cultures which were allowed to grow to an OD_{600} of 0.6. Protein overexpression was then induced by the addition of isopropyl-beta-D-thiogalactopyranoside (IPTG) to a final concentration of 0.5 mM. Cultures were incubated for 5h at 30 ºC with shaking (180 rpm). Once the induction was over, cells were harvested by centrifugation at 3,200*g* for 10 min at room temperature. The pellet was resuspended in 37 mL of Lysis Buffer (50 mM Na_2PO_4, 500 mM NaCl, 0.05% Triton X-100, 0.1mM PMSF, 10 mM Imidazol, 10% Glycerol, pH 8.0) supplemented with 50 mg of Lysozyme (Sigma). The suspension was frozen and stored at -20ºC. For protein extraction, cells were lysed by sonication, using five pulses of 10 sec. at 100W with 10 sec resting on ice between each pulse. The lysates were cleared by centrifugation at 10,000*g* for 20 min at 4ºC. The supernatants were further centrifuged at 18,000*g* for 15 min at 4ºC. The supernatant was collected and the purification of the overexpressed protein was immediately started, as described below.

2.3.2. Protein Purification

The reporter protein was tagged at the C-terminus with a $(His)_6$-Tag and was purified using nickel affinity chromatography. For this, protein extracts were incubated in batch with 1.0 mL of Ni-NTA Agarose (Qiagen), overnight with gentle agitation. The extracts were then centrifuged at 3,500*g* for 10 min at 4ºC, and supernatants were collected and frozen. Routinely, 5 ml of supernatant were used to resuspend the NiNTA-agarose prior to loading into a Poly-Prep Chromatography Column (BioRad). Both column washing and protein elution were performed with Buffer A_1 (6 M Urea, 100 mM NaH_2PO_4, 10 mM Tri-Cl, 0.01 % Triton X-100, 7.5 % Glycerol), at different pH. The washes were performed as follows: firstly with 5.0 mL of Buffer A_1, pH 7.2; then with 5.0 mL of Buffer A, pH 6.8; and finally with 10 mL of Buffer A_1, pH 6.3. Fractions of 5.0 mL were collected. The elution of the reporter protein was carried out with Buffer A at pH 5.8. A total of 7.5 mL were collected in 10 fractions of 0.75mL each. A final wash with 5 mL of Buffer A_1 at pH 4.5 was also done. The presence of the reporter protein in each fraction was monitored by SDS-PAGE and Western-Blotting.

The fractions enriched in the reporter protein were then subjected to Fast Protein Liquid Chromatography (FPLC). For this, the pH was restored to 7.0 and the samples loaded into an AKTApurifier system coupled with a HiTrap Chelating HP column (Amersham Biosciences), chelated with $NiCl_2$. The column was washed with 5 mL of buffer A_1, and the protein eluted with a gradient of imidazol from 0 to 0.5M in 10 mL of buffer B_1 (buffer A with 1.0 M Imidazol). Fractions of 0.50 mL were collected and reporter protein purification was monitored by SDS-PAGE and Western-Blotting.

The recombinant SerRS was also purified by nickel affinity chromatography, as described above, except that 1 mL of Ni Sepharose High Performance (Amersham) was routinely used. The resin was incubated for 1 h at 4ºC with gentle agitation, and then centrifuged at 3,000*g* for 5 min at 4ºC. As before, almost all supernatant was removed and frozen. Sedimented agarose was resuspended in the remaining supernatant and loaded into a Poly-Prep Chromatography Column (BioRad). Column washing and protein elution were done with Buffer A_2 (50 mM Na_2HPO_4, 500 mM NaCl, pH 8.0) supplemented with

Imidazol, at different concentrations. The column was firstly washed with 15 mL of Buffer A_2 + 20 mM Imidazol, then with 15 mL of Buffer A_2 + 40 mM Imidazol. The protein was eluted with a step gradient of 60 mM, 100 mM and 150 mM Imidazol in Buffer A_2. Two fractions of 7.5 mL of each elution step were collected. A final column wash with 15 mL of Buffer A_2 + 500 mM Imidazol was also carried out.

Purification of the recombinant *C. albicans* LeuRS overexpressed in *E. coli* was carried out as described for the SerRS with the following alterations. The 1mL Ni Sepharose High Performance (Amersham) column was washed with 15 mL of Buffer A_2 + 20 mM Imidazol, and the protein was eluted with a step gradient of 40 mM, 60 mM and 100 mM Imidazol in Buffer A_2. Two fractions of 7.5 mL of each elution step were collected. A final column wash was of 15 mL of Buffer A_2 + 500 mM Imidazol was also carried out.

2.3.3. Protein Quantification

The purified proteins were quantified using the BCA Protein Assay Reagent Kit (Pierce), which is based on bicinchoninic acid (BCA). Quantifications were carried out according to the manufacturer's instructions, with minor changes. Briefly, the working reagent (WR) was prepared by mixing BCA Reagent-A with BCA Reagent-B in a 50:1 ratio. In general, 0.9 ml of WR was prepared for each quantification. Protein samples were prepared in 100 µL. Blanks, containing no protein were also prepared. The samples used to build standard curves contained 100, 75, 50 or 25 µg of protein. To all protein samples, 0.9 mL of WR was added and mixed, and samples were incubated at 37ºC for 30 min and were then cooled down to room temperature and their absorbance at 562nm measured. The absorbance values of the standards were plotted against their respective amount of protein and a linear regression was determined to build the equation for protein quantification. Only regressions with R^2 values above 0.97 were considered.

2.3.4. Polyacrylamide gel electrophoresis (PAGE)

Proteins were fractionated on 10% or 12% PAGE prepared with 29:1 acrylamide/bis-acrylamide as indicated on the Roche Molecular Biochemicals Lab FAQS. Protein samples were diluted in 2 or 3 µl of 6x sample buffer (30 % glycerol, 10 % SDS, 0.6 M DTT and 0.012 % bromophenol blue in 0.5 M Tris-Cl / 0.4 % SDS, pH 6.8), to a final volume of 12 or 18 µl, and boiled for 1 minute before loading onto the gel. Low Molecular Weight (Amersham) and Pre-stained markers (SIGMA) were used for stained and blotted gels, respectively. Gels were run on BioRad mini-gel apparatus, at 50 V for about 1 hour and then at 100-150 V for about 2 hours, until the front of the migration reached the bottom of the gel. Electrophoresis buffer contained 25 mM Tris, 192 mM Glycine and 0.2 % SDS. After electrophoresis, gels were stained or blotted as described below. The gel images were acquired using a densitometer and analysed with the QuantityOne software (BioRad).

Coomassie Blue stain was prepared as a solution of 0.25 % Brilliant Blue R in 50 % methanol and 10 % acetic acid. This solution was filtered before use. Gels were stained by immersion in the solution for 5 to 10 minutes, with low agitation. After staining, gels were destained in 10 % ethanol and 7.5 % acetic acid with agitation, until the protein bands were visible, and stored in distilled water.

When gels were used for *in gel* protein digestion and peptide Mass Spectrometry assays, NuPAGE Bis-Tris 10% pre-cast gels (Invitrogen) were used. These gels were run in NuPAGE MOPS SDS running buffer (Invitrogen) for 2h at 120V. Gels were stained with SimplyBlue SafeStain (Invitrogen), for 1 hour, and destained with milliQ water for either 3 hours or overnight.

2.3.5. Western-blotting analysis

After electrophoresis, proteins were electroblotted onto nitrocellulose membranes (Hybond ECL, Amersham) prior to immunodetection. For this, six sheets of 3MM paper (Whatman) and blotting membranes were cut to gel dimensions. Membranes were pre-hydrated in distilled water and then hydrated in transfer buffer, TGM (20 mM Tris-Cl, 150

mM glycine and 20 % methanol), for 5 minutes. Gels were also equilibrated in TGM for 5 minutes. 3 sheets of 3MM paper hydrated in TGM were placed on the anode of the transfer system, and a "sandwich" was assembled by laying down the membrane on top of the paper sheets. The gel was then added on top and was covered with 3 additional sheets of 3MM paper hydrated with TGM. Air bubbles were avoided by rolling a glass pipette over the gel/paper sandwich before placing the cathode plate on the semi-dry blotter (BioRad). Transfers were carried out at 0.8 mA/cm^2 of gel (approximately 12V for standard sized gels) for 20 minutes. After the transfer, membranes were washed in TBS-T (140 mM NaCl, 1 mM KCl, 19 mM Na_2HPO_4, 2 mM K_2HPO_4 pH7.4, with 0.1 % (v/v) Tween-20) for 15 minutes and blocked at room temperature for 2 hours with 5 % (w/v) skimmed milk powder (Molico, Nestlé) in TBS-T.

Membranes from above were washed twice with TBS-T, for 5 minutes prior to addition of the primary antibody. Incubations with primary antibodies were specific for each antibody used (Table 2. 4). After this, membranes were washed 3 times for 20 minutes each time in TBS-T. Incubation with secondary antibodies was carried out in TBS-T with 1 % skimmed milk, for 2 hours at room temperature. The secondary antibodies were chosen according to specification of the primary antibodies used; they were either anti-Mouse (Amersham) or anti-Rabbit (Sigma), but both of them were diluted 1:5000 in TBS-T. Finally membranes were washed 3 times with TBS-T, for 15 minutes each time. Antibody incubations were carried out inside sealed plastic bags in order to reduce reaction volumes.

Immunodetection was performed by chemiluminescence, using the ECL kit from Amersham, according to the manufacturer's instructions. For this, detection reagent A and detection reagent B from the ECL kit were mixed in the dark in a 1:40 ratio and this mixture was applied onto the membranes surface, ensuring that the entire surface was covered. After 5-minute incubation, the mixture was removed with a pipette and the membranes were covered with clingfilm, avoiding air bubbles. Membranes were then exposed to X-ray film (Kodak) for a suitable period of time and the film was developed and fixed using Kodak reagents.

Table 2. 4 – Primary antibodies.

Primary Antibody	Source	Dilution	Incubation Conditions	Obs.
Anti-FLAG	*Rabbit*	*1:3000*	*Overnight at 4ºC*	*Polyclonal, from Sigma*
Anti-PhosphoSerine	*Mouse*	*1:2000*	*Overnight at 4ºC*	*Polyclonal, from Qiagen*
Anti-LeuRS	*Rabbit*	*1:1000*	*1h at room temperature*	*Whole serum, kind gift from M. Tuite at U. Kent*
Anti-SerRS	*Rabbit*	*1:1000*	*1h at room temperature*	*Whole serum, kind gift from M. Tuite at U. Kent*
Anti-Actin (H-300, sc-*10731)*	*Rabbit*	*1:500*	*Overnight at 4ºC*	*Polyclonal, from Santa Cruz Biotech.*

2.3.6. *In gel* protein digestion

(Adapted from Chapman, 2000).

Bands corresponding to the CUG reporter protein were cut from gels and *in gel* protein-digestions were performed. For this, gel slices were washed twice for 20 minutes in 100 mM ammonium bicarbonate with 50% acetonitrile, and afterwards for 15 minutes with acetonitrile and then air dried. Gel pieces were rehydrated with 30 µL of cleavage solution [20 mM Tri-Cl pH 7.6, 0.15 M NaCl, 2.5 mM $CaCl_2$, 2 U of Enterokinase and 2 U of Thrombin (both from Novagen)] and incubated for 36 hours at room temperature. The digested peptides were removed from the gel slices by washing with 50 % acetonitrile at 37 ºC for 1 hour. Supernatants were collected and concentrated by speed-vacuum. Immediately prior to mass spectrometry analysis, formic acid to a final concentration of 0.1 % was added to concentrated samples.

2.3.7. Mass-Spectrometry

Peptide samples were loaded onto a Q-ToF Micro (Micromass) system equipped with a nanoelectrospray ion source coupled to a nanoflow High Performance Liquid Chromatography (HPLC) system (CappLC, Micromass) for mass spectrometry. The instrument was operated in positive ion mode. The capillary voltage was maintained at

3500 V and the sample cone at 35 V. The ion source temperature was 100 ºC. The cone gas flow was set at 130 L/hour and the nebulizing gas flow was maintained at 2 psi. A PepMap C18 pre-column cartridge (5 µm particles, 100 Å pores, 300 µm x 5 mm) was used to trap and desalt the peptides and a PepMap C18 analytical column (3 µm particles, 100 Å pores, 75 µm x 15 cm) was used to separate them. The flow rate through the column was 250 nL/min. Pre- and analytical columns were equilibrated with aqueous phase (2% acetonitrile and 0.1% formic acid) for 15 minutes. The digested peptides were bound to pre-columns and desalted with aqueous 0.1% formic acid at 30 µL/min for 3 minutes. The organic phase (98% acetonitrile and 0.1% formic acid) was increased from 5% to 40% during 27 minutes and increased from 40% to 90% during 5 minutes. Finally, it was held at 90% for 5 minutes, reduced to 5% over 5 minutes and maintained at 5% for 15 minutes. Data were analyzed with Masslynx software from Micromass.

2.4. Overexpression and purification of the *C. albicans* $tRNA_{CAG}^{Ser}$

C. albicans CAI-4 was transformed with pUA77, which contained 3 copies of the $tRNA_{CAG}^{Ser}$ gene. Cells were grown at 30ºC to an OD_{600} of 2.5 – 3.0 in cultures of 750 mL in MM-Ura. Cells were harvested by centrifuging at 3,500*g*, for 15 min at 4ºC. Several cultures were prepared to obtain 120g of cell pellet (wet weight). Cell pellets were frozen and stored at -80ºC until further use. Total RNAs were extracted in several successive steps in 250 mL bottles (Nalgen). For this, 30 g of cell pellet were resuspended in 60 mL of tRNA Extraction Buffer (0.1 M NaCl, 5 mM magnesium acetate (MgOAc), 2 mM DTT, 1.5% SDS (w/v), 10 mM Tris-Cl, pH 7.0) and then 1 *vol.* of phenol, equilibrated with Tris-Cl with a final pH of 6.4 (Sigma), was added. The mixture was shaken overnight at 200 rpm at 25 ºC and then incubated in a water bath at 65 ºC for 1h. The two phases were separated by centrifugation at 3,200*g* for 20 min, at 4º C, and the upper aqueous phase was transferred to a new 250 mL bottle. To remove contaminant proteins present in the aqueous phase, the RNAs were re-extracted with 1 vol of phenol, equilibrated with Tris-Cl, pH of 6.4 (Sigma), and the mixture was shaken for 1h at 200 rpm at 25 ºC. The aqueous phase containing RNAs was separated from the organic phase by centrifugation at 3,200 g, for 20 min, at 4° C, and collected as 15 mL fractions in 50 mL tubes. The crude RNA was then

precipitated overnight at -20 ºC with 2 *vol.* of absolute ethanol and then harvested at 10,000*g* for 30 min at 4ºC. The pellet was washed with absolute ethanol, the sample centrifuged at 10,000*g* for 30 min at 4ºC, the supernatant discarded and the pellet was air dried for 15-20 min. Finally the pellet was resuspended in 10 mL of 0.1 M NaOAc. pH 4.5 and all fractions were pulled.

The total RNA extracts were cleaned from contaminating rRNAs, mRNAs and proteins using 90 ml DEAE-52 (Sigma) columns equilibrated with 0.1 M NaOAc. pH 4.5. The columns were successively washed with 100 mL of each of the following buffers: 0.1 M NaOAc. pH 4.5; 0.1 M NaOAc. pH 4.5 + 0.1 M NaCl; 0.1 M NaOAc. pH 4.5 + 0.2 M NaCl; and finally, 0.1 M NaOAc. pH 4.5 + 0.3 M NaCl. The tRNAs were eluted from the column with 90 mL of 0.1 M NaOAc. pH 4.5 + 1 M NaCl and were precipitated overnight with 2 *vol.* of absolute ethanol. Pellets were collected by centrifugation and dried as described above. The tRNA preparations were then de-acylated in 1M Tris-Cl, 1 mM EDTA pH 8.0, for 1h at 37ºC. These tRNA preparations were precipitated overnight with 2 *vol.* of absolute ethanol and 0.1 *vol.* of NaOAc, as descried in section 2.2.11, and resuspended in CCC Binding Buffer (1.2 M NaCl, 15 mM EDTA, 30 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) - KOH, pH 7.5)

2.4.1. tRNA purification by affinity chromatography
(Adapted from Tsurui et al., 1994; Suzuki et al., 1996)

Individual tRNAs were purified by affinity chromatography using DNA probes (Table 2. 5). For this, biotinylated DNA probes (MWG) were immobilized on Streptavidin agarose as follows: DNA probes were resuspended in 100 mM Tris-Cl, pH 7.5 to a final concentration of 3 units of absorbance at 260 nm (A_{260}) per 100 µL, and in a 1.5 mL tube, 100 µL of DNA probe was incubated with 300 µL of Streptavidin Sepharose™ slurry (Amersham) for 2 h at room temperature on a rotating wheel.

Table 2. 5 – DNA probes used for tRNA purification

tRNA	DNA probe sequence (5' → 3')	Biotin
$tRNA_{CAG}^{Ser}$	*CGC GGG CAA TGC CCA AAG GAA CCT GCA TCC*	*3' –*
$tRNA_{AGA}^{Ser}$	*CGA CAC GAG CAG GGT TCG AAC CTG CGC GG*	*5' –*
$tRNA_{CAA}^{Leu}$	*TGA CAC CAA GGA GAT TCG AAC TCC TGC AT*	*5' –*

Then, it was centrifuged at 3,000g and the A_{260} of the supernatant was measured to assess the incorporation efficiency of the probes by determining the ratio between the amount of free probe in solution after and before incubation. When ligation efficiency was above 85% supernatants were discarded and the resin was resuspended in CCC Binding Buffer. The slurry was finally packed into 0.5 mL columns (Pierce).

Chromatography columns were assembled in an oven TCC-100 (Dionex) and connected to a low pressure liquid chromatography system (BioRad). The crude tRNA extracts, in CCC Binding Buffer, were circulated overnight in a closed circuit at a flow rate of 0.3 $mL.min^{-1}$. The following program cycle was used: 10 min at 80 ºC, to denature tRNAs, 35 ºC for 30 min. for renaturation. tRNA binding to the column was at 65 ºC for 90 min, then at 50 ºC for 500 min, and then the temperature was restored to 35 ºC for 60 min, for column washing.

Columns were washed at 35ºC with 15 mL of CCC Washing Buffer (0.6 M NaCl, 7.5 mM EDTA, 15 mM HEPES-KOH, pH 7.5). Once the absorbance at 258 nm (A_{258}) of the sample stabilized, the buffer was changed to CCC Low Salt-Buffer (20 mM NaCl, 0.25 mM EDTA, 0.5 mM HEPES-KOH, pH 7.5) and columns were washed until the A_{258} was stable. Then, tRNA elution started by increasing the column temperature to 50ºC and stopping the buffer flow. Once the temperature stabilized, the buffer flow was restored, at 0.3 $mL.min^{-1}$ and fractions of 0.3 mL were collected until the A_{258} became stable (linear). Then, the buffer flow was stopped again and the oven temperature was increased to 65 ºC. After a 5 min incubation at this temperature, the flow buffer was restored and new fractions were collected. The efficiency of tRNA purification was monitored by electrophoresis in semi-denaturing 12.5% (w/v) polyacrylamide mini-gels [12.5%

Acril:Bisacrylamide (19:1), 4 M Urea, TBE (0.09 M Tris, 0.09 M boric acid pH 8.3, 2.5 mM EDTA), 1% (w/v) APS, 0.1% (w/v) TEMED].

2.4.2. High resolution tRNA electrophoresis

For high resolution tRNA fractionation, 25 cm x 40 cm gels were assembled between clean glass plates separated using 0.5 mm thick spacers. Plates were held together using steel clips. The gel moulds were then filled with the gel solution [12.5% Acril:Bisacrylamide (19:1), 4 M Urea, TBE (0.09 M Tris, 0.09 M boric acid pH 8.3, 2.5 mM EDTA), 1% (w/v) APS, 0.1% (w/v) TEMED], and the slot former introduced. Gels were allowed to polymerize at room temperature, wrapped in Clingfilm and stored at 4 ºC for later use. Gels were then inserted in an adjustable vertical running system (ADJ3, Anagene), the slot formers removed and the buffer tanks were filled with TBE. Prior to sample loading, a pre-run of 1 h at 500 V was performed. The tRNA samples were diluted in 2x Loading Buffer [10 mM NaOAc pH 5.0, 8 M urea, 0.05% (w/v) bromophenol blue, 0.05% (w/v) xylene cyanol] and loaded onto the gel with a 50 µl Hamilton microsyringe. Electrophoresis was done at 700 V, at 4 ºC overnight and fractioned tRNAs were stained in TBE-EtBr and visualized under a UV light.

2.5. Aminoacylation kinetics assays

Aminoacylation reactions were carried out in a buffer (100 µL) containing 100 mM Tris-Cl, pH 7.6, 15 mM $MgCl_2$, 4mM DTT, 250 mM NaCl, 10 mM KCl, 40 µM amino acid (either [^{3}H]leucine or [^{3}H]serine) (400 Ci/mol), 0.01% BSA and 2 mM ATP. In these reactions the concentration of both enzyme and tRNA were varied. For this, tRNAs were re-folded before use by heating to 85 ºC for 4 min in re-folding buffer (60 mM Tris, pH=7.8, 2 mM MgCl2) followed by slow cooling to room temperature. Reactions were initiated by adding the enzyme and, at varying time intervals, 20 µL aliquots were quenched by spotting on Whatman No. 3MM disks soaked with 5% trichloroacetic acid (TCA). The filters were washed 3 times for periods of 5 min each in 5% TCA. Then they were washed in 96% ethanol and counted in a liquid scintillation counter (Beckman).

Amino acid activation assays were based on the amino acid dependent ATP-PPi exchange reaction which can be used to determine the kinetics of activation of amino acids by aaRSs. This reaction was used in this study to determine the functionality of the active site of aaRSs, with an excess of both the enzyme and the amino acid. The reactions were carried out in 100 µL of 100 mM Tris-Cl, pH 7.8, 15 mM $MgCl_2$, 4mM DTT, 250 mM NaCl, 10 mM KCl, 4 mM amino acid (either leucine or serine), 0.01% BSA, 2 mM ATP and 2 mM [γ^{32}P]PPi (2TBq.mol^{-1}) (Amersham). The enzyme concentrations were of 0.1 µM SerRS or 1.7 µM LeuRS. Aliquots (20 µL) were removed from the reaction solution at various time points and quenched into 250 µL of buffer solution containing 1.6% w/v activated charcoal, 4.46% Na-PPi and 3.5% w/v $HClO_4$. The 270 µL charcoal suspension was then filtered on a Whatman filter, assembled on vacuum filtering system, and the filter washed once with 4 mL of 40 mM Na-PPi, 1.4% $HClO_4$, followed by a wash of 4 mL with distilled water and of a last wash with 4 mL 96% ethanol. The filters were then placed into scintillation vial and 4 mL of scintillation liquid was added. The [^{32}P]-labelled ATP absorbed on the charcoal was quantified by liquid scintillation (Geslain *et al.*, 2006).

To determine the number of catalytic active sites, we carried out an active site titration reaction (Fersht *et al.*, 1975). This method is based on the stoichiometric depletion of 1 mol of ATP for the formation of both 1 mol of pyrophosphate and 1 mol of complex aminoacyl-adenylte•enzyme (AA~AMP•E). In this reaction there is an initial linear decrease in the ATP concentration, consequence of the rapid burst of AA~AMP•E formation. The active site titrations were carried out at 30ºC in 150 µL reactions, with enzyme concentrations in the range of 0.25µM to 1µM in the presence of 100 mM Tris-Cl, pH 7.8, 15 mM $MgCl_2$, 2mM DTT, 250 mM NaCl, 10 mM KCl, 1 mM amino acid (either leucine or serine), PPase 2 mU.µL^{-1}, 10 µM ATP and 1000 cpm.µL^{-1} of [γ^{32}P]ATP (Amersham). The enzyme was added at time 0 and ATP depletion was monitored at the time points 0.25, 0.5, 1, 2, 5, 15, 30 min. For each time point, an aliquot of 20 µL was taken out and the reaction was stopped by mixing with a 200 µL suspension of 7% percloric acid and 2% activated charcoal, to capture [γ^{32}P]ATP. This 220 µL suspension was then filtered using Whatman filters, assembled on a vacuum filtering system, and the filter washed with 4 mL of 0.5% percloric acid, 54 mL of water and finally with 4 mL of 96% ethanol. Filters were then placed into scintillation vials and 4 mL of scintillation

liquid was added. The amount of [γ^{32}P]ATP present in each sample was measured on a scintillation counter (Beckman).

2.6. Bioinformatic tools and data mining

2.6.1. Analysis of the genome and the proteome of *C. albicans*

The *C. albicans* genome (assembly 19; haploid version), containing 6438 annotated Open Reading Frames (ORFs), was downloaded from the *Candida* Genome Database (www.candidagenome.org), and analyzed with ANACONDA (Moura *et al.*, 2005). This in house built software package counted all codons present in the annotated ORFs and calculated the CAI values for each gene. The probability of generating different proteins from genes containing CUGs, due to serine or leucine insertion at those CUG positions was calculated by the binomial distribution: $b_{(i,n,p)} = \frac{n!}{i!(n-1)!} p^i (1-p)^{n-i}$, were n is the total number of CUG codons per gene, p is the probability of leucine incorporation at CUG positions for different percentages of ambiguity, and i is the number of CUGs decoded as leucine (Ex: For genes containing 3 CUGs; n =3 and i = 0, 1, 2 or 3).

The total number of novel proteins in the proteome of *C. albicans* was estimated based on the studies of Ghaemmaghami et al (2003), who discovered that the abundance of proteins is co-related to the codon adaptation index (CAI) and that it ranges from 50 up to more than 10^6 molecules per cell. In our calculations we assumed that *i)* all the genes are expressed and *ii)* the abundance of proteins (N_{total}) is of 5,000 molecules for the 10% of genes with the lowest CAI values; of 50,000 molecules for the 10% of genes with the highest CAI values; and of 20,000 molecules for the remaining 80% of genes. The number of novel proteins arising (N_{novel}) for each gene is given by: $N_{novel} = N_{total} \times (1-b_{(0,n,p)})$.

2.6.2. Protein and gene sequence alignments and phylogenetic analysis

Both the gene and protein sequences used in this study were obtained from public databases. The NCBI database (http://www.ncbi.nlm.nih.gov) was routinely used in this study. The gene sequences from *Candida lusitaniae, Candida guilliermondii* and *Candida tropicalis* were extracted from their whole genome sequences, which are available at the Broad Institute, http://www.broad.mit.edu/annotation/fgi/. Finally, the gene sequences from *S. bayanus* and *S. paradoxus* were obtained at http://cbi.labri.fr/Genolevures/index.php.

The BLASTP online server (http://www.ncbi.nlm.nih.gov/BLAST/) was used to find homologs of the *C. albicans* LeuRS and SerRS proteins. The multiple sequence alignments of both gene and protein sequences were carried out with ClustalW (Thompson *et al.*, 1994) and displayed with either the BioEdit or the ESPript software packages (Gouet *et al.*, 2003) (similarity score matrix: BLOSUM62). Phylogenetic analysis were carried out using Mega3.1 (Kumar *et al.*, 2004) and were obtained with the neighbour-joining algorithm and a bootstrap of 1000 replications.

2.6.3. Protein structure modelling

Structural templates for the *C. albicans* LeuRS and SerRS were obtained from the Protein Data Bank (PDB) of the Research Collaboratory for Structural Bioinformatics (RCSB). The selected structural templates of the LeuRS were from *P. horikoshii* and *T. thermophilus*, with the accession numbers 1WZ2 and 1OBC, respectively. The structural template of SerRS was from *T. thermophilus*, with the accession number 1SES. The theoretical models for the *C. albicans* proteins were generated by comparative protein modeling using the automated SWISS-MODEL servers (Arnold *et al.*, 2006) and were displayed and analyzed with the Pymol or the Rasmol software.

3. Quantification of CUG ambiguity in *C. albicans in vivo* by Mass-Spectrometry

The results presented in this chapter were published in the following paper:

Gomes, A.C., Miranda, I., Silva, R. M, Moura, G.R, Thomas, B., Akoulitchev, A. and Santos, M.A.S. (2007) "A Genetic Code Alteration Generates a Proteome of High Diversity in the Human Pathogen *Candida albicans*" *Genome Biology* **8:**R206; doi:10.1186/gb-2007-8-10-r206.

3.1. Introduction

Life maintenance and perpetuation is dependent on accurate flow of genetic information. The DNA replication error ranges from 10^{-10} to 10^{-11}, whereas the transcription error is in the order of 10^{-4} to 10^{-6} (Edelmann and Gallant, 1977). The translational errors arise from both wrong aminoacylation and codon misreading, and are in the order of 10^{-4} to 10^{-5} (Fersht and Dingwall, 1979; Freist et al., 1985; Loftfield and Vanderjagt, 1972, reviewed in Parker, 1989; Ogle and Ramakrishnan, 2005). This suggests that there is no evolutionary pressure for the ribosome to increase decoding accuracy above that of aminoacylation levels. In fact, hyperaccurate ribosomes slow down growth rate indicating that protein synthesis accuracy is a compromise between decoding fidelity and decoding speed (Parker, 1989).

Ribosome decoding errors are of 3 main types, namely *(i)* *missense errors*, which result in substitution of one amino acid for another; *(ii)* *processivity errors* that can be due to frameshifting and *(iii)* *non-sense suppression*, which result in readthrough of termination codons (Farabaugh and Bjork, 1999). Such errors in protein synthesis result always in the production of aberrant proteins (Figure 3. 1), although their impact on the cell physiology may be variable.

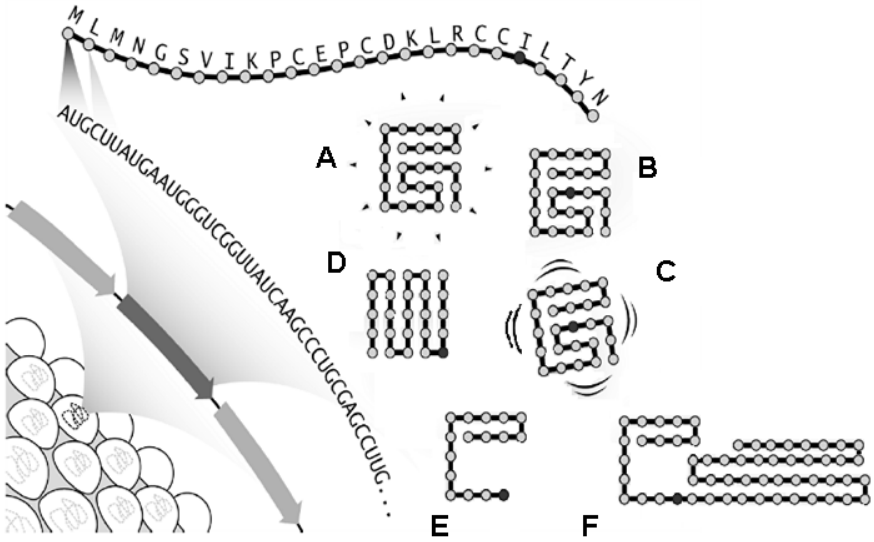


Figure 3. 1 – Errors in translation.
(A) Proteins that are correctly translated and fold properly are fully functional. However, erroneous codon decoding produces aberrant proteins (B-E). **(B)** Mistranslated proteins can retain the wild type tertiary structure, and maintain activity. **(C)** Some mistranslated proteins can still fold, but become unstable and less active. **(D)** Missense errors disrupt protein structure. For example, frameshifting (E-F) can result in synthesis of truncated proteins because stop codons may appear just downstream of the frameshift site **(E)**, alternatively readthrough proteins are synthesized if premature stop codons resulting from frameshifting are not recognized **(F)**. Readthrough proteins are also synthesized by non-sense suppression of wild type (in frame) stop codons. The misread residue is represented as a black dot. Adapted from (Drummond *et al.*, 2005)

Missense errors are the most frequent translation errors under general growth conditions (Kurland and Gallant, 1996). These errors can arise from tRNA mischarging or due to incorrect tRNA selection by the ribosome. Missense error rate is in the order of 10^{-4} to 10^{-5}, which is in agreement with global translation error rates. In *E. coli* different amino acids substitutions rates have been measured. Leucine misincorporation in poly(Phe) peptides is in the order of $4x10^{-4}$ (Wagner *et al.*, 1982) and phenylalanine incorporation in recombinant mEGF, which does not have any Phe codon, is $6x10^{-4}$ (Scorer *et al.*, 1991). Some missense errors are not deleterious since most amino acid substitutions involve chemically similar amino acids that do not disrupt protein structure and function (Kurland and Gallant, 1996). However, most missense errors decrease the activity of produced proteins and do have an impact on cell physiology and fitness (Ehrenberg and Kurland, 1984; Kurland and Ehrenberg, 1984). Also, missense errors may increase during stress conditions, namely amino acid starvation (Parker, 1989) and may decrease growth rate

(Nangle *et al.*, 2002). Interestingly under strong stress conditions it increases adaptation and is selectively advantageous (Santos et al., 1996; Santos et al., 1999).

When a mRNA is being translated by the ribosome, the maintenance of the mRNA reading frame, after translocation, is of utmost importance, as any ribosomal slippage precludes synthesis of full length proteins, not only because the decoded message does not correspond to that expected from the mRNA open reading frame, but also because the ribosome usually encounters termination codons during out-of-frame reading. The latter is due to the fact that stop codons can arise from single base changes of several different codons. The ribosome, itself, has developed mechanisms to maintain the reading frame during decoding by positioning of the 3 tRNAs in the decoding centre and by stabilizing the complex formed between the anticodon and the mRNA codon at the P site (Li et al., 2001; Hansen et al., 2003). Frameshifting errors, either -1 or +1 (Figure 3. 2), occur at a frequency of 10^{-5} (Kurland and Gallant, 1996). This basal error rate may increase at particular mRNA sequences or under certain physiological conditions (Fu and Parker, 1994; Barak et al., 1996; Stahl et al., 2004). For example, mRNA sequences prone to -1 frameshifting are the heptameric sequences X-XXY-YYZ, where X and Z can be any nucleotide and Y is either a A or a U. (Jacks et al., 1988; Dinman et al., 1991; Curran, 1993). Two models explain such frameshifting, the first proposes that it occurs before translocation and is induced by simultaneous slippage of the two tRNAs present in the P- and A-sites of the ribosome (Jacks *et al.*, 1988). The second proposes that it occurs after translocation, when the codons of the heptameric sequence occupy the E- and P-sites of the ribosome (Horsfield *et al.*, 1995). In both cases the tRNA at the P-site has always a central role in the frameshifting (Baranov *et al.*, 2004).

Frameshift errors also occur in a sequence independent manner, when the ribosome stalls at “hungry codons”, which may arise due to aa-tRNA limitation. In this case, +1 frameshifting is caused by slow entry of the cognate aa-tRNA into the A-site. This creates a ribosomal pause and induces peptidyl-tRNA to shift in the P-site (Farabaugh, 1996; Gallant and Lindsley, 1992; Lindsley et al., 2005). Nevertheless, under this circumstance the frameshift might be regarded as a safeguard of translation, as it allows the ribosome to continue and facilitates its recycling. Likewise, the peptidyl-tRNA can also change the

reading frame at nonsense codons because of slow decoding of stop codons by release factors (Weiss *et al.*, 1990).

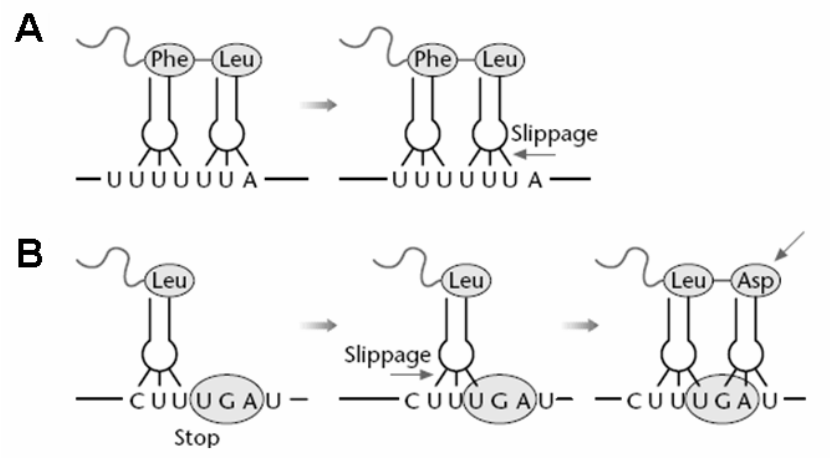


Figure 3. 2 – The -1 and +1 frameshifting.
(A) The -1 frameshift results from a slippage of the mRNA in the 5' direction. **(B)** The +1 frameshift is caused by a slippage of one base of the mRNA towards the 3' end.

Modified bases in the anticodon loop of tRNAs play an important role in reading frame maintenance. For example, m^1G_{37} and $ms^2io^6A_{37}$ prevent +1 frameshifts, but apparently have no role in preventing −1 frameshifts (Urbonavicius et al., 2001; Urbonavicius et al., 2003). On the other hand, it has been reported that the ψ_{39} modification may induce frameshifting, as it destabilizes the interaction between the tRNA and the E-site of the ribosome, inducing a higher frequency of release of the tRNA from the E-site, thus promoting slippage in the P-site (Bekaert and Rousset, 2005).

Non-sense suppression happens when the stop codons, namely UAA, UAG and UGA, are recognized by near-cognate tRNAs (nonsense suppressors), leading to the synthesis of readthrough proteins. Natural nonsense suppression occurs at a frequency of 10^{-3} to 10^{-5}, but each stop codon is suppressed with different efficiency. In bacteria, suppression of the UGA codon ranges from 10^{-2} to 10^{-5}; of the UAG from 10^{-3} to 10^{-4}; and of UAA from 10^{-4} to less than 10^{-5} (reviewed in Parker, 1989). Indeed, suppression efficiency is influenced by a variety of factors, namely stop codon context and presence of stimulatory elements downstream the stop codon (Bertram *et al.*, 2001).

Despite the negative impacts of mistranslation, which are discussed in chapter 4, it plays an important role in the evolution of genetic code expansions and alterations because these evolve gradually through codon decoding ambiguity (Knight et al., 2001; Schultz and Yarus, 1994; Santos et al., 2004), at least in some cases. For example, both selenocysteine and pyrrolysine incorporation is achieved through re-programming of UGA and UAG stop codons, respectively, representing a context-dependent non-sense suppression event (section 1.4.4). Selenocysteine is incorporated in both prokaryotic and eukaryotic selenoproteins at UGA stop codons by novel translation elongation factors (SelB-prokaryotes; EF-sec and SBP2-eukaryotes), a new tRNA ($tRNA^{Sec}$) and a selenocysteine mRNA insertion element (SECIS) (Namy *et al.*, 2004), whereas pyrrolysine, is inserted at the UAG-stop codon using a pyrrolysine insertion sequence (PYLIS), in the mRNA of methylamine methyltransferases (Theobald-Dietrich *et al.*, 2004). Also, the artificial expansion of the genetic code to incorporate non-natural amino acids (Anderson et al., 2004; Santoro et al., 2002) is achieved either through non-sense suppression or frameshifting (section 1.4.4.3).

Indeed, most alterations and expansions of the genetic code are mediated by structural changes in the protein synthesis machinery, in particular in tRNAs, aminoacyl-tRNA synthetases, elongation and termination factors (Yokobori et al., 2001; Santos et al., 1996; Santos et al., 2004). Nevertheless, *per se* they do not provide any insight into evolutionary forces that drive codon identity redefinition. Neither do they help to evaluate the impact of the code changes on proteome and genome stability, gene expression, adaptation and ultimately on evolution of new phenotypes. In order to address these questions, *C. albicans* was chosen as a well studied model system (Santos et al., 1993; Santos and Tuite, 1995; Santos et al., 1996; Santos et al., 1997). This fungal species has changed the identity of the leucine CUG codon to serine through an ambiguous codon decoding mechanism that affected approximately 30,000 CUG codons in more than 50% of its ancestor genes (Massey *et al.*, 2003). The CUG reassignment from leucine to serine in *Candida spp.*, is the only known sense to sense codon identity alteration in eukaryotic cytoplasmic translation systems. This genetic code change has evolved gradually over 272±25 My, through an ambiguous codon decoding mechanism that arouse from leucine mischarging of a $tRNA_{CAG}^{Ser}$ (Massey et al., 2003; Suzuki et al., 1997; Sugiyama et al.,

1995). In *C. zeylanoids* the $tRNA_{CAG}^{Ser}$ can be charged *in vitro* with leucine and *in vivo* it is charged with 3% leucine (Suzuki *et al.*, 1997).

The connection between ambiguous charging of the $tRNA_{CAG}^{Ser}$ and ambiguous CUG decoding remains to be established. In here, this question was dissected using a reporter protein engineered to allow for quantification of leucine and serine insertion at CUG positions by mass spectrometry. We show that direct mass spectrometry is a powerful methodology to quantify mRNA decoding error. The latter has been poorly characterized and overlooked over the years due to lack of robust methodologies to quantify peptide mixtures arising from translation of single mRNA molecules. Our methodology opens the door for quantification of mistranslation under different physiological conditions.

3.2. Results

3.2.1. Construction of a CUG mistranslation reporter system

A reporter protein to quantify CUG ambiguity (CUG-reporter system) (Figure 3. 3, Figure 3. 4) was constructed using the *C. albicans* phosphoglycerate kinase (*CaPGK1*) gene as a backbone system for assembly of a chimeric gene. The *CaPGK1* gene has a high CAI value of 0.829 and does not contain CUG codons (Annexe B), indicating that it is a highly expressed gene and that it is not affected by ambiguous CUG decoding. We have inserted an N-terminal reporter cassette containing a single CUG codon to quantify leucine and serine incorporation at this position. This cassette peptide was flanked by thrombin and enterokinase cleavage sites, which were used to cleave the reporter peptide from the recombinant protein for mass spectrometry analysis.

The chimeric gene was constructed in three sequential steps. Firstly, the promoter and a DNA fragment encoding the N-terminal 69 amino acids of the protein were cloned into the multicloning site of the pSL1190 vector (*Hind III* and *Xho I* sites). These restriction sites were included in the tail of the 5' and 3' primers, oUA201 and oUA202, respectively. Secondly, the fragment containing the CUG codon and the coding sequences

of thrombin and enterokinase were introduced using a long oligonucleotide containing *Xho* I and *Sac* II restriction sites (oUA215). This oligonucleotide was used as primer to re-amplify the *CaPGK1* backbone of the reporter. The other PCR primer (3´ end primer; oUA216) that hybridized to the 3´end of the *CaPGK1* Open Reading Frame (ORF) contained a tail of six histidines to aid protein purification by affinity chromatography. This primer also contained a stop codon and an *Nru* I restriction site. This second fragment was cloned into the plasmid containing the first fragment (see above) into the *Xho* I and *Nru* I restriction sites. Thirdly, the 3'UTR sequence of eEF1-α was also inserted in the chimeric gene at the *Nru* I and *Pst* I restriction sites. Again, the restriction sites were added in the tail of both 5' and 3' primers, oUA205 and oUA204, respectively. This reporter gene assembled into the pSL1190 vector was then removed from this vector as a single DNA fragment containing *Hind* III and *Pst* I ends and was subcloned into the *C. albicans* pRM1 shuttle vector at identical restriction sites (Figure 3. 3).

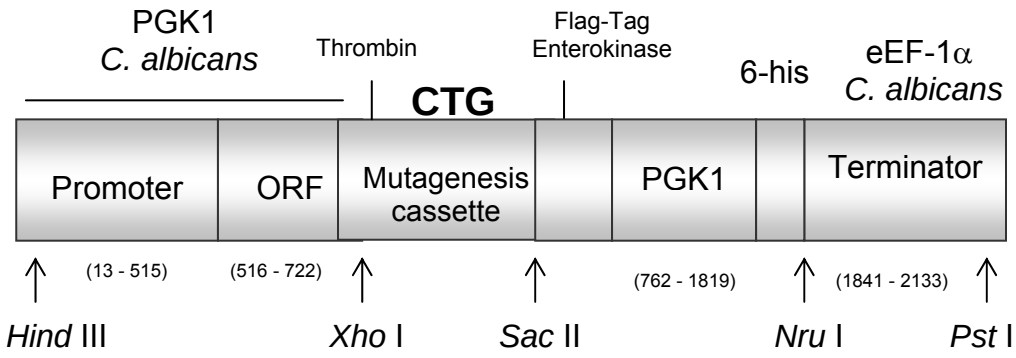


Figure 3. 3 – Scheme of the *C. albicans* CUG reporter gene.
The CTG codon was introduced into a mutagenesis cassette which was fused to the *CaPGK1* gene as shown in the diagram. Two tags containing, the Flag- and 6xHis-epitopes were added to allow for the detection of the protein by Western blot and for its purification by affinity chromatography. The mutagenesis cassette was engineered to permit its easy replacement at the *Xho I* and *Sac II* restriction sites and is flanked by the sequence encoding both proteases cleavage sites.

The reporter peptide of interest contained 17 amino acids (Figure 3. 4), and its sequence was LVPR↓GS**X**PRDYKDDDDK↓, where **X** indicates the residue encoded by the CUG codon. This peptide contains thrombin and enterokinase cleavage sites. Additionally, two tags were added to the protein to allow for its detection and purification, namely, the FLAG-Tag, which was added in the mutagenesis cassette and was used for

protein immunodetection; and a 6-histidine tag, which was added at the C-terminus and used for protein purification.

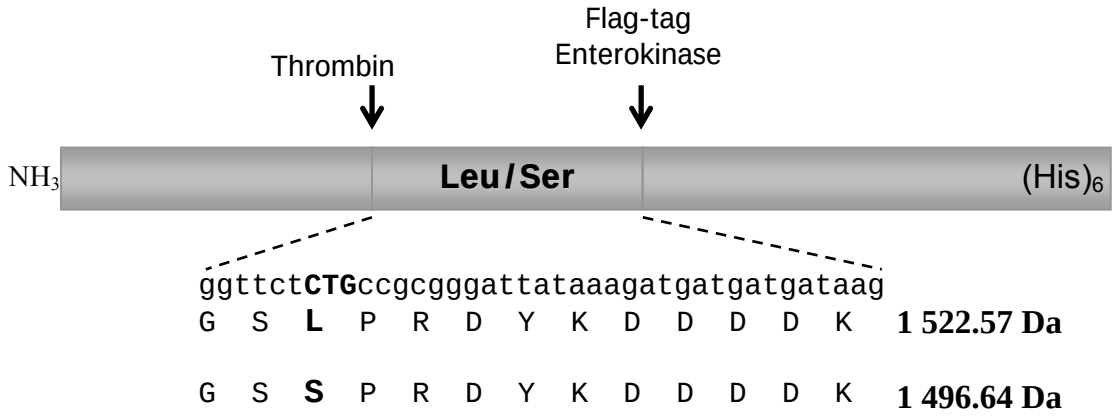


Figure 3. 4 – Reporter protein.
Diagram showing the primary structure of the reporter protein used to quantify Leucine and Serine incorporation at the CUG codon position. The FLAG-Tag epitope was used for immunodetection and the 6-His Tag was used for protein purification by nickel affinity chromatography. The sequence of the peptides that were detected by mass-spectrometry and their respective molecular weights are also indicated.

3.2.1.1. Reporter protein purification and processing.

C. albicans CAI-4 cells were transformed with the pUA63 plasmid, as described in materials and methods (Section 2.2.13), and were grown on MM-URA at 30ºC to an OD_{600} of 0.6 – 0.9. Cells were collected by centrifugation and lysed under denaturing conditions (section 2.3.1). The reporter protein was then purified, firstly in batch system, using a Ni-NTA agarose slurry, which was incubated with the crude protein extract. The protein was eluted by decreasing the pH of the elution buffer and its purity was monitored by SDS-PAGE (Figure 3. 5).

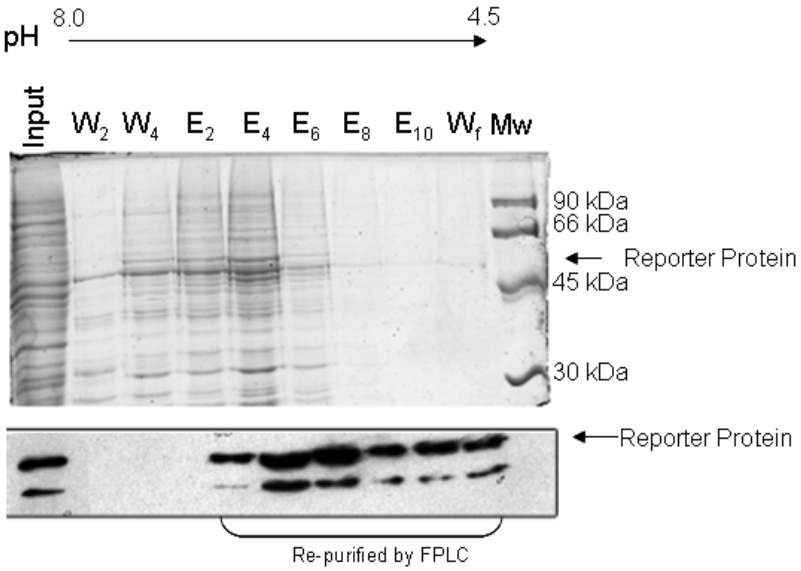


Figure 3. 5 – Reporter protein purification.
SDS-PAGE of the fractions collected from the Ni-NTA agarose used to purify the protein. The column washes and the protein elution were performed with a decreasing pH gradient. W_2 – wash at pH of 7.0; W_4 – wash at pH of 6.3; E – elutions at pH of 5.8: 10 fractions of 0.75 mL were collected and the even fractions were loaded on the gel. W_f – final wash at pH of 4.3. Mw – Molecular weight. The reporter protein position on the gel is indicated by the arrow and its apparent MW on the gel corresponds to 47.0 kDa.

The fractions containing the reporter protein (E_2 - E_{10} and W_f in Figure 3. 5) were pulled together and re-purified by FPLC. The protein was then eluted with increasing concentration of imidazol (Figure 3. 6). Once purified, the reporter protein was electrophoresed on NuPAGE 10% pre-cast gels (Figure 3. 7) and its band was cut and *in gel* digested with both thrombin and enterokinase (see methods 2.3.6). The peptides were then eluted from the gel by washing with a 50 % acetonitrile solution at 37 ºC for 1 hour and analyzed by mass-spectrometry using a Q-ToF Micro (Micromass) system.

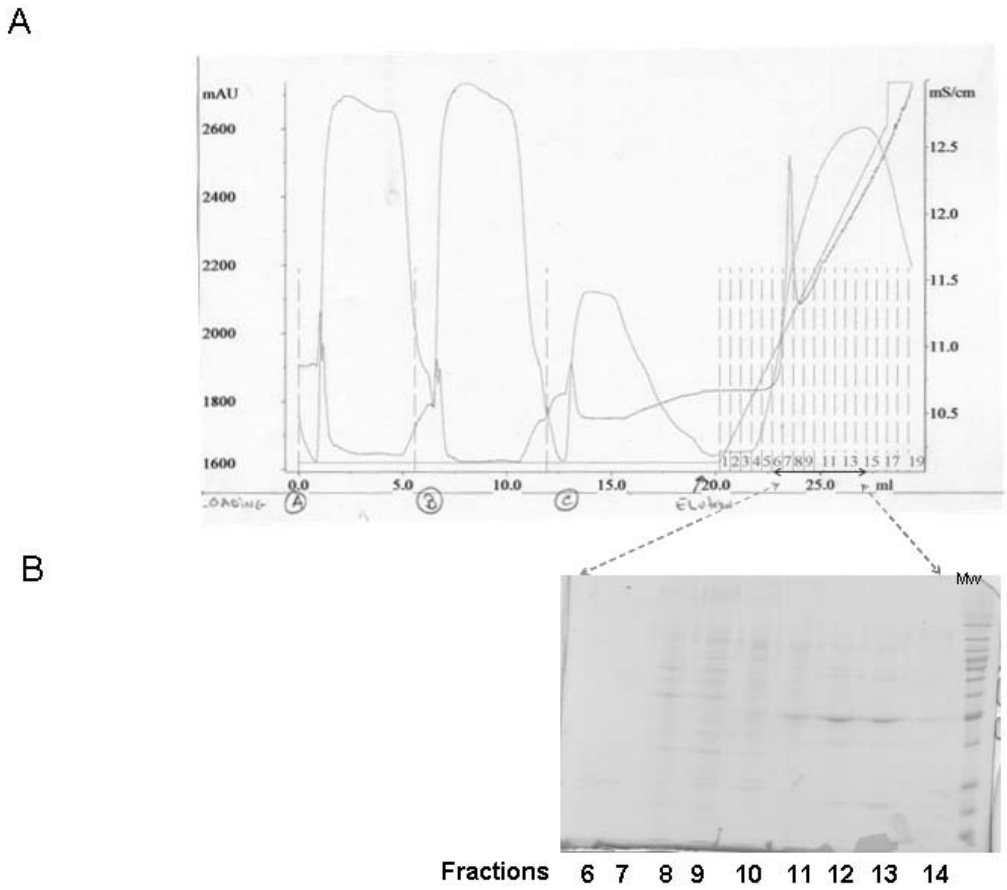


Figure 3. 6 - Reporter protein re-purification by FPLC.
(A) The figure shows purification of the reporter from a 5.0 mL fraction obtained from a batch purification using Ni-NTA agarose. This fraction was loaded twice onto a HiTrap Chelating HP column, chelated with $NiCl_2$. The protein was eluted with a linear gradient of imidazol starting after 20.0 mL of wash. The fractions collected were numbered as indicated in the panel. **(B)** SDS-PAGE showing proteins present in fractions 6-14. The reporter protein is visible in fractions 11, 12, 13 and 14.

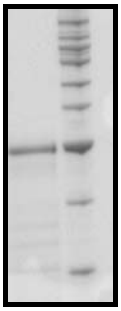


Figure 3. 7 – *In gel* digestion of the purified reporter protein.
Once purified, the protein was electrophoresed on 10% SDS-PAGE and stained with SimplyBlue SafeStain (Invitrogen), for 1 hour, and destained with milliQ water overnight. The band containing the protein was then cut using a clean scalpel. Finally, the protein was *in gel* digested with thrombin and enterokinase for 36 hours at room temperature.

3.2.1.2. Mass-Spectrometry data analysis.

After cleavage, the reporter peptide containing the sequence GS**X**PRDYKDDDDK, where **X** is serine or leucine, was eluted from the gel as described above and its mass was determined by mass-spectrometry using Q-ToF Micro (Micromass) system. For this, we have taken into consideration the mass difference between serine and leucine containing peptides, which is 26 Da, and also the chemical differences between these amino acids. Since serine is polar (with a hydropathy index of -0.8) and leucine is apolar (with a hydropathy index of 3.8), and serine has a hydroxyl group (-OH) which is chemically reactive and can be phosphorylated, the two peptides behave differently on the HPLC-MS system and such differences were exploited to quantify serine and leucine incorporation at the CUG position. Indeed, the serine peptide had a low retention time on the C18-HPLC column (10.08 minutes) (Figure 3. 8A) due to its hydrophilic nature and was found in three different forms with different molecular weights, namely *i*) with the unmodified serine residue (Ser-OH) with a molecular mass of 1496.6 Da. Its peak appeared at a mass/charge ratio of 499.88 and 749.33, for charges of +3 and +2, respectively (Figure 3. 8B); *ii*) with a covalently linked phosphate group to the serine's hydroxyl group (Ser-O-PO_3) with a molecular weight of 1576.5 Da. Its peak appeared at a mass/charge ratio of 526.5, for a charge of +3 (Figure 3. 8C); and *iii*) with the ester bond between the phosphate and the serine's hydroxyl group broken (Ser-H), which may have arisen due to the high voltage of the mass-spectrometer cone. The molecular mass of this peptide was 1478.4 Da and its peak appeared at a mass/charge ratio of 493.8, for a charge of +3 (Figure 3. 8D).

In order to confirm whether the reporter protein was phosphorylated, a Western blot against phophoserine was carried out (Figure 3. 9). As a negative control, the phosphate groups were removed from the reporter protein with calf intestinal alkaline phosphatase (CIP). For this, 5 µg of the reporter protein were incubated with 10 units of CIP (New England Biolabs) for 60 minutes at 37°C in a 50 µL reaction, containing 10 mM NaCl, 1 mM $MgCl_2$, 0.1 mM dithiothreitol and 5 mM Tris-HCl, pH 7.9.

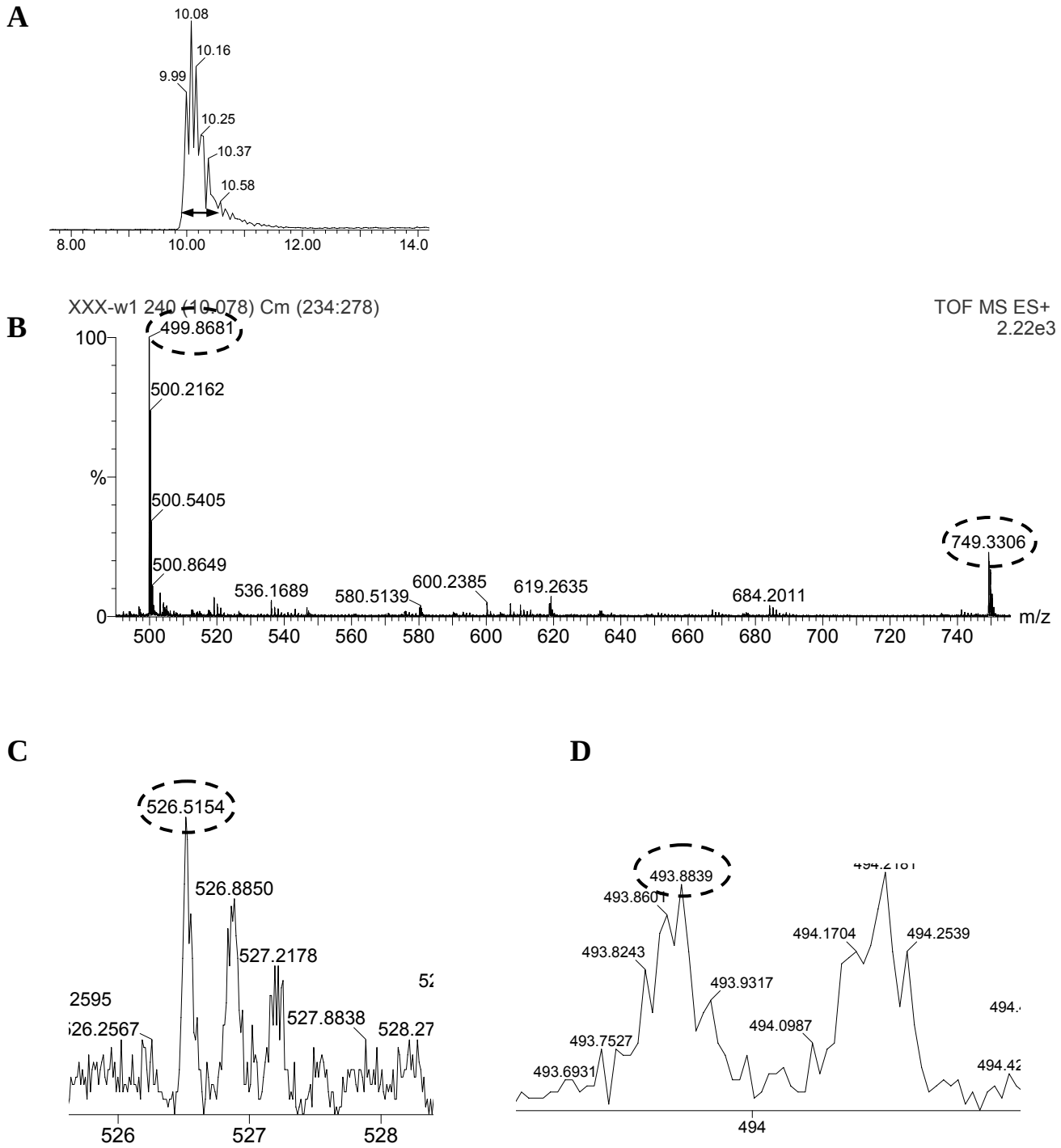


Figure 3. 8 – HPLC-MS of the Serine-peptide.
Spectra of the reporter peptide obtained after digestion of the reporter protein with enterokinase and thrombin. **(A)** HPLC fractionation prior to mass determination, showing the elution of the serine peptide at 10.08 minutes. The arrow indicates the interval of time used to obtain the combined spectra. **(B)** The major peaks of the combined spectra are 499.86 and 749.33, corresponding to the unmodified Ser-OH-peptide, with a charge of +3 and +2, respectively. (**C**) and (**D**) Detail (zoom) of the previous spectra showing the 526.5 and 493.8 regions, respectively. Multiple peaks correspond to ^{13}C isotopic forms of the amino acids. The right upper corner of panel-B shows the total number of counts for the major peak (499.8).

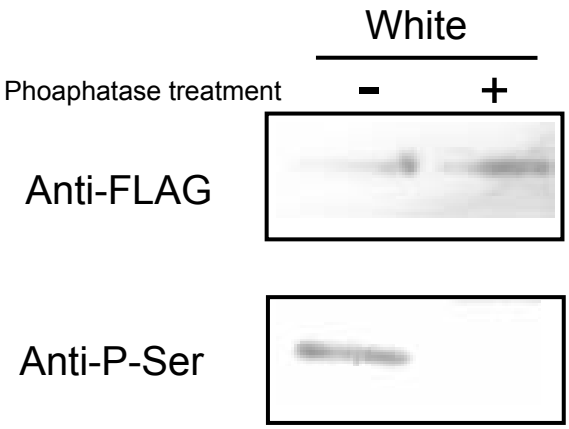


Figure 3. 9 – The reporter protein is phosphorylated *in vivo.*
Detection of the reporter protein with anti-FLAG and anti-P-Ser antibodies. The protein fractions were loaded into a 10% Nu-PAGE gel (Invitrogen) and run for 2h at 120 V. Then proteins were transferred to nitrocellulose membranes and Western-Blots were carried out. 5 µg of the reporter protein were treated with 10 units of calf alkaline phosphatase for 60 minutes at 37°C.

The Western blot result confirmed that the protein was phosphorylated. Since the reporter contained the sequence Ser-Ser-Pro, which is a strong phosphorylation signal (Blom *et al.*, 1999), and the corresponding phospho-peptide was detected in the mass-spectrum, it is reasonable to assume that the phospho-serine found in the reporter protein is present in the reporter peptide.

These results were further confirmed using synthetic peptides of identical amino acid sequences to the reporter peptides (Annexe C). The synthetic Ser-peptide that produced spectra in the interval of 9.90 and 10.58 minutes were taken into account, giving combined spectra which were then analysed. Also, MS-MS analyses were carried out with both synthetic peptides and the reporter protein (Annexe C), which were compared and proved that the analysed peptide corresponded to the designed reporter peptide. The peaks corresponding to the mass/charge ratio of the three species of serine peptides were screened, the baseline subtracted and the number of counts for each species added to obtain the abundance of the serine peptide.

A similar approach was followed for quantification of the peptide containing leucine at the CUG position. This peptide had a molecular mass of 1522.57 Da, showed higher

retention time (12.8 minutes) on the HPLC, caused by its stronger hydrophobicity. Its mass spectrum showed peaks at a mass/charge ratio of 508.57 and of 762.35, for charges of +3 and +2, respectively (Figure 3. 10);

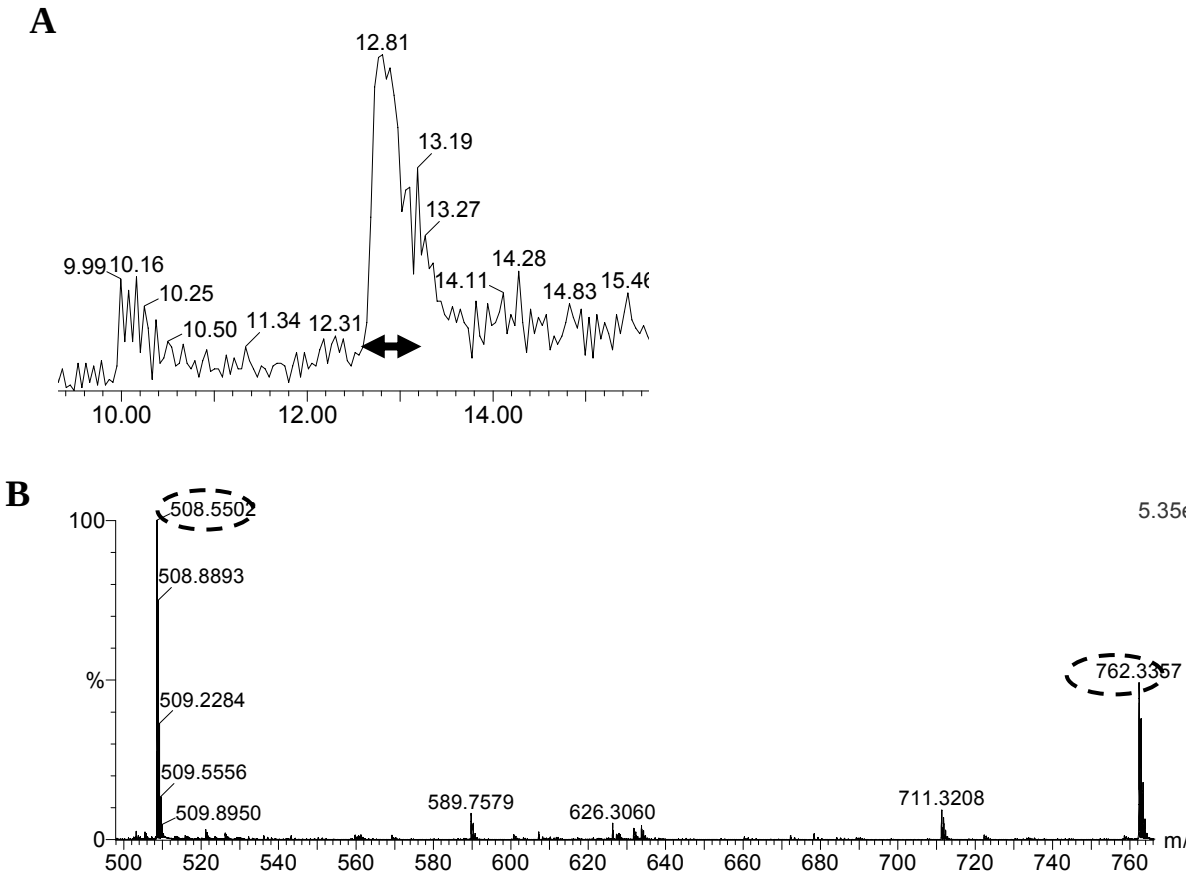


Figure 3. 10 - HPLC-MS of the Leucine peptide.
Spectra of the reporter peptide containing leucine at the CUG position. The peptide was obtained by digestion of the reporter protein with enterokinase and thrombin. **(A)** The peptide had a retention time on HPLC of 12.8 minutes. The arrow shows the interval of time used to obtain the combined spectra. **(B)** Peak corresponding to the leucine peptide, with a mass/charge ratio of 508.5. Multiple peaks correspond to ^{13}C isotopic forms.

Likewise, MS-MS spectra were obtained for both the synthetic and the reporter peptides, and then compared (Annexe C), to ensure that peaks analysed corresponded to the reporter peptide. The spectra of the leucine peptide were obtained in the interval of 12.7 and 12.9 minutes, giving combined spectra which were then analysed (Figure 3. 10). The peaks corresponding to the mass/charge ratio of the leucine peptide were screened and its number of counts quantified.

3.2.1.2.1. Data normalization

In order to measure accurately serine and leucine incorporation at the CUG codon position, the ionization of leucine- and serine- peptides was monitored. This was important to ensure that putative differences in ionization efficiency were not interfering with the quantification of both peptides. Synthetic peptides of both forms were prepared in an equimolar solution and were analyzed by mass-spectrometry (Figure 3. 11).

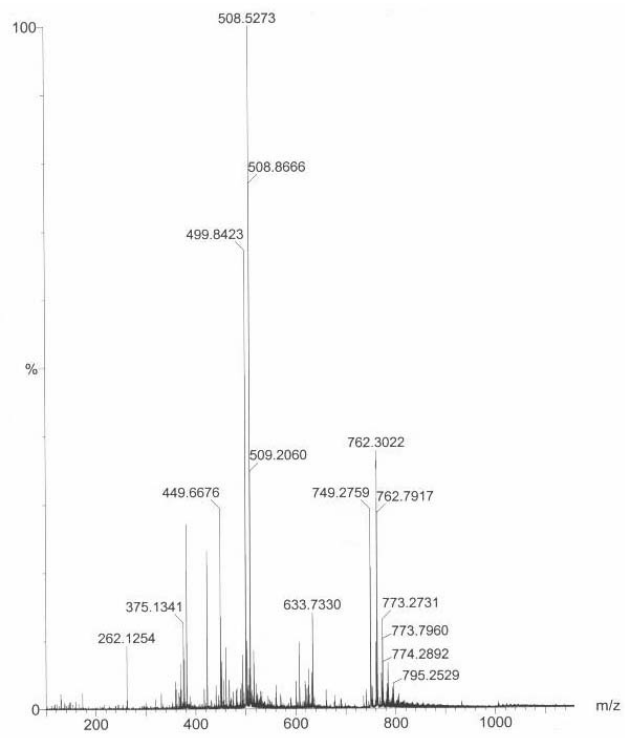


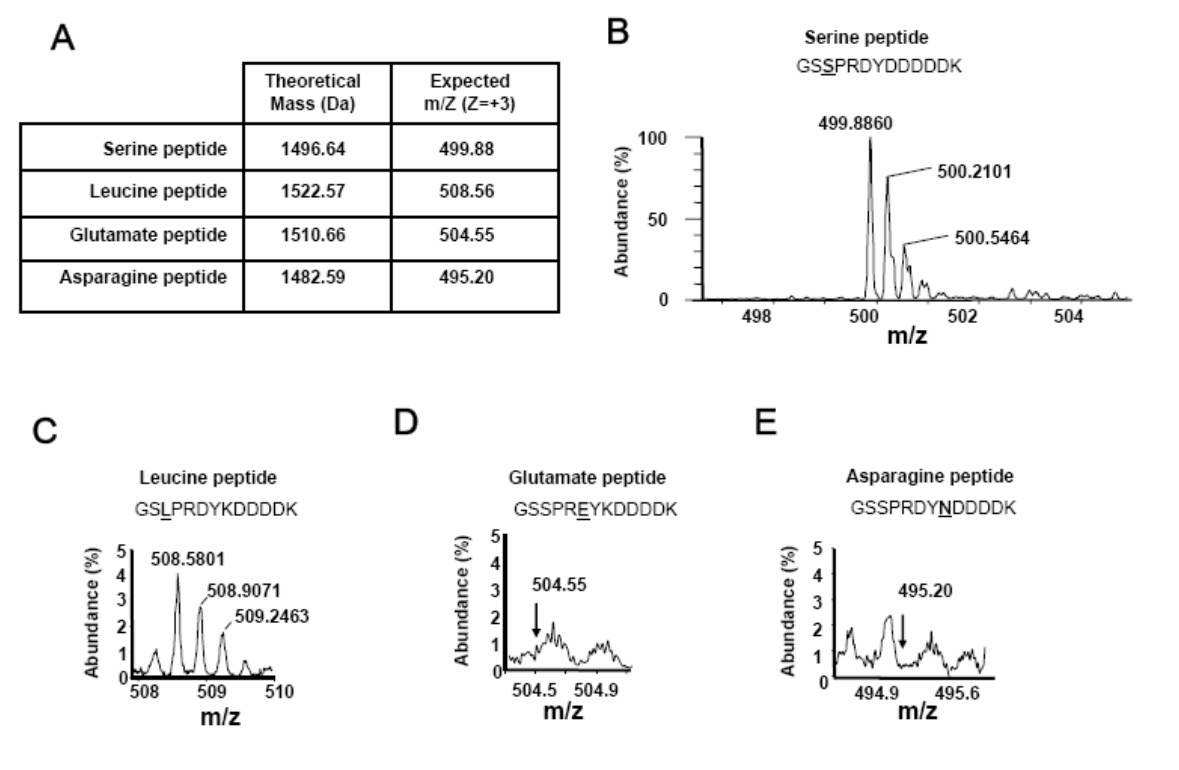
Figure 3. 11 – Spectrum of an equimolar mixture of serine and leucine peptides
An equimolar mixture of the leucine and serine peptides was prepared and applied to the HPLC-MS. The spectrum shows that the peak intensity of the leucine peptide is higher than that of the serine peptide, indicating that the former did indeed ionize more efficiently than the latter.

The synthetic peptide spectra showed that the serine peptide had a weaker signal than the leucine peptide – both for the +2 and +3 m/Z. In other words, the serine containing peptide ionises less efficiently than the leucine peptide. This allowed us to normalize the data considering the relative ionization efficiencies of the leucine peptide as 100% and that

of the serine peptide as 70%. Therefore, the mass-spectrometry data was normalized by correcting the number of counts obtained for the leucine peptide by a factor of 0.7.

3.2.1.2.2. Amino acid misincorporation

In order to ensure that leucine misincorporation at the CUG position could be detected above background noise, the amino acid misincorporation at near-cognate codons was also monitored. The near-cognate misreading is the most frequent mistranslation error since it involves misreading at the wobble position by near cognate tRNAs (Kurland and Gallant, 1996). This error has been monitored in yeast *in vivo* and is in the order of 0.001% (Stansfield *et al.*, 1998). Since the aspartate GAU and lysine AAA codons encoded by the reporter peptide (Figure 3. 4) could be misread by near-cognate $tRNA^{Glu}$ and $tRNA^{Asn}$, respectively, the mass of these aberrant peptides containing glutamate at the aspartate-GAU position or asparagine at the lysine-AAA position was determined (Figure 3. 12 A). The peptides resulting from correct serine incorporation and leucine misincorporation at the CUG position were clearly visible in the mass-spectrum (Figure 3. 12 B,C), while the peptides containing serine at the CUG position plus glutamate at the aspartate-GAU or asparagine at the lysine-AAA positions were not detected (Figure 3. 12 D, E), confirming that our methodology was robust for accurate quantification of mistranslation of the *C. albicans* serine CUG codon as leucine.



	Theoretical Mass (Da)	Expected m/Z (Z=+3)
Serine peptide	1496.64	499.88
Leucine peptide	1522.57	508.56
Glutamate peptide	1510.66	504.55
Asparagine peptide	1482.59	495.20

Figure 3. 12– Mistranslation due to near-cognate decoding
(A) Table with the theoretical mass and the expected m/Z peaks of the peptides that were screened in the MS experiments. The serine peptide is the product of correct translation of the recombinant gene and was the most abundant. The leucine peptide corresponded to a peptide synthesized by ambiguous decoding of the CUG codon by the *C. albicans* $tRNA_{CAG}^{Ser}$. The glutamate peptide was the product of decoding of the aspartate-GAU codon as glutamate by the near-cognate tRNA that decodes the glutamate GAA and GAG codons. Likewise, the lysine-AAA and AAG codons could be decoded by the near-cognate tRNAs that decode the asparagines AAU and AAC codons. **(B)** Mass spectrum of the serine peptide. **(C)** Mass spectrum of the leucine peptide. **(D)** Mass spectrum showing the region where the peak corresponding to the peptide containing glutamate at the aspartate position was expected (arrow). **(E)** Mass spectrum showing the region where the peak corresponding to the peptide containing asparagines in the position of the lysine AAA codons was expected (arrow).

3.2.2. Determination of leucine and serine incorporation at the CUG codon *in vivo*

Leucine incorporation at the CUG codon position was initially quantified in the most abundant type of *Candida albicans* cells, i.e., white cells, grown at 30ºC. The abundance of each peak was determined as described above, and is summarized in Table 3. 1.

Table 3. 1 – Leucine incorporation at the CUG codon on white cells.
The abundance of each peptide species, obtained from independent HPLC-MS experiments. The % of leucine incorporation was obtained as explained in the text below. (n.d. – not detectable).

File	Serine - Peaks					Leucine - Peaks				%Leu
	Ser		Ser-OH	Ser-P				Total		
	Z=+3	Z=+2	Z=+3	Z=+3	Total	Z=+3	Z=+2			
	499,88	**749,32**	**526,53**	**493,86**		**508,56**	**762,35**		Correc.	
A9	1920	346	49	n.d.	2315	52	31	83	58,1	**2,45**
A10	1790	303	62	53	2208	67	29	96	67,2	**2,95**
A15	1260	177	47	n.d.	1484	48	20	68	47,6	**3,11**
A17	2820	429	185	187	3621	110	37	147	102,9	**2,76**
A19	2740	1006	113	n.d.	3859	80	35	115	80,5	**2,04**
A13c	1890	291	129	170	2480	86	44	130	91	**3,54**
A23	9570	2110	461	606	12747	428	203	631	441,7	**3,35**
V-1w	1170	190	12	n.d.	1372	71	n.d.	71	49,7	**3,50**
V-2w	4170	1000	175	n.d.	5345	176	78	254	177,8	**3,22**
V-3w	760	161	n.d.	118	1039	51,3	n.d.	51,3	35,91	**3,34**
XXX-2	663	92	18	n.d.	773	34	n.d.	34	23,8	**2,99**
XXX-1	2220	510	40	30	2800	93	n.d.	93	65,1	**2,27**

The total number of counts of all spectra collected for the serine (*Sp*) or leucine (*Lp*) peaks, were used to determine the relative frequency of leucine incorporation at the CUG codon position, by applying the expression: %Leucine = [*Lp* / (*Lp* + *Sp*)] x 100, and was 3.0 % ± 0.49 in white cells. These results unequivocally showed that the CUG codon is ambiguous in *C. albicans*, which is surprising because such high misincorporation levels are forbidden by genetic code accuracy rules.

3.2.2.1. CUG ambiguity in opaque cells

Since *C. albicans* is polymorphic, we have also quantified leucine incorporation at CUG positions in different cell types, namely in opaque cells. These cells result from low frequency (10^{-4}) switching of white cells (Lan *et al.*, 2002) and are the mating competent form of *C. albicans*. Opaque cultures of *C. albicans* are normally unstable, but it is

possible to maintain them at low temperature or by re-plating in fresh medium (Figure 3. 13). In these conditions, cultures containing more than 90% of opaque cells can be maintained. This type of cells are known to be morphologically and physiologically distinct from the white cells (Lan *et al.*, 2002), and we wondered whether these physiological differences would have implications for CUG ambiguity.

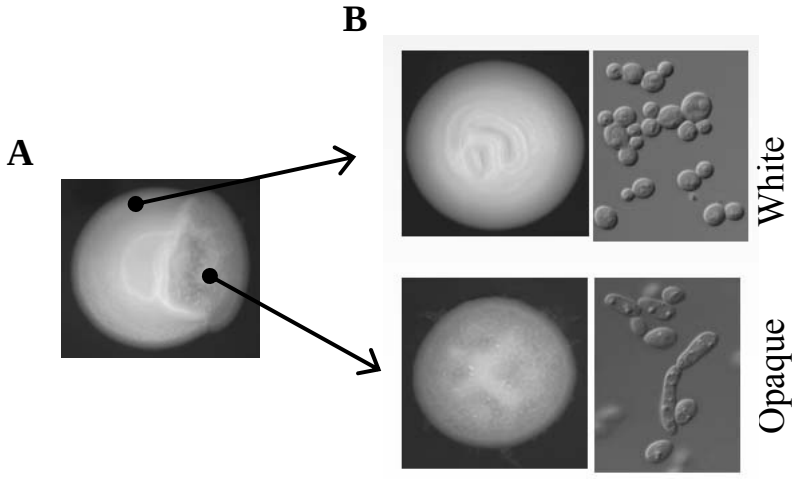


Figure 3. 13– Candida albicans morphology.
(A) From *C. albicans CAI-4* strain, transformed with pUA63, a colony with a white/opaque sector was screened. From this sector, an opaque cell line was isolated and maintained in fresh agar medium. **(B)** Details of colony and cellular morphological differences between white and opaque cells.

Table 3. 2 – Leucine incorporation at the CUG codon on opaque cells.
The abundance of serine and leucine peptides isolated from the reporter protein by thrombin/enterokinase digestion was determined by HPLC-MS experiments. The % of leucine incorporation was obtained as explained above (n.d. – not detectable)

File	Serine - Peaks					Leucine - Peaks				%Leu
	Ser		Ser-OH	Ser-P				Total		
	Z=+3	Z=+2	Z=+3	Z=+3	Total	Z=+3	Z=+2			
	499,88	**749,32**	**526,53**	**493,86**		**508,56**	**762,35**		Correc.	
XXIX-4	6890	2310	249	365	9814	159	n.d.	159	111,3	**1,12**
XXIX-2	6160	4550	110	123	10943	66	n.d.	66	46,2	**0,42**
XXIX-5	2730	635	165	65	3595	25	n.d.	25	17,5	**0,48**

An opaque cell line, expressing the reporter protein, was selected from a white colony by successive plating on agar plates until a culture containing more than 90% of opaque cells was obtained. Then, the reporter protein was purified and analyzed, as

previously described (Sections 2.3.2, 2.3.6 and 2.3.7). And the HPLC-MS data obtained (Table 3. 2) allowed for determination of leucine incorporation at CUG codon in these opaque cells.

The leucine incorporation at the CUG codon in opaque cells was 0.66 % ± 0.28, which was significantly different from that determined in white cells (3%), ($p = 10^{-8}$). This result indicated that *C. albicans* is, somehow, able to manipulate the levels mistranslation of CUG codons.

3.2.2.2. CUG ambiguity in different physiological conditions.

The surprising discovery that CUG ambiguity varied between white and opaque cells prompted us to investigate whether such variation could also be observed in different physiological conditions. For this, the CUG reporter protein was expressed in *C. albicans* grown at 37ºC, under oxidative stress and in low pH. The protein was purified, cleaved and analyzed by mass-spectrometry, as described above. Quantification of leucine incorporation at the CUG position was also carried out using synthetic peptides to calibrate the mass-spectrometer (as above). Interestingly, at 37ºC, which is the optimal growth temperature for *C. albicans*, leucine and serine incorporation at the CUG position was 3.7 % ± 0.41 and 96.3% ± 0.41, respectively (Table 3. 3). Therefore, there was a slight increase in CUG ambiguity at 37ºC in relation to 30ºC, but this small increase is within the standard deviation and one should be careful to give it a real physiological significance.

Table 3. 3 – Leucine incorporation at the CUG codon on cells grown at 37ºC.
The abundance of leucine and serine peptide species, obtained from independent HPLC-MS experiments. The % of leucine incorporation was obtained as described above.

File	Serine - Peaks					Leucine - Peaks				%Leu
	Ser		Ser-OH	Ser-P				Total		
	Z=+3	Z=+2	Z=+3	Z=+3	**Total**	Z=+3	Z=+2			
	499,88	**749,32**	**526,53**	**493,86**		**508,56**	**762,35**		Correc.	
C12	3160	1050	291	265	4766	123	105	228	159,6	**3,24**
C16	11800	5490	932	786	19008	785	332	1117	781,9	**3,95**
C21	2010	440	203	236	2889	112	58	170	119	**3,99**

Next, CUG ambiguity was evaluated in cells grown on MM-Ura at pH 4.0. For this, the minimal growth medium was buffered with 50 mM of citrate buffer (Abaitua *et al.*, 1999) and the culture was incubated at 30ºC overnight. Again, preparation of the reporter protein and its analysis by mass-spectrometry was carried out as described before (Table 3. 4). Most surprisingly, leucine incorporation at the CUG codon was measured 4.9 % ± 1.1.

This is a significant increase in decoding error and may be of physiological significance. Indeed, *C. albicans* is mainly destroyed by macrophages forming lyzosomes at low pH (Watanabe *et al.*, 1991), but *C. albicans* survives such acidic environment and is able to escape the macrophage (Kaposzta *et al.*, 1999). This data does not allow us to establish a link between CUG ambiguity and macrophage survival, however increased CUG ambiguity increases synthesis of new proteins (see below; chapter 4) which may be secreted or exposed on the surface of the *C. albicans* cell. In other words, can *C. albicans* sense the presence of macrophages and somehow use CUG ambiguity to generate antigenic diversity? This would allow it to escape the immune system. If so, it shows how an apparently chaotic molecular event can generate important selective advantages.

Table 3. 4 – Leucine incorporation at the CUG codon in cells grown at pH 4.0.
The abundance of serine and leucine peptide species was obtained from independent HPLC-MS experiments. The % of leucine incorporation was obtained as explained above. (n.d. – not detectable)

File	Serine - Peaks					Leucine - Peaks				%Leu
	Ser		Ser-OH	Ser-P				Total		
	Z=+3	Z=+2	Z=+3	Z=+3	Total	Z=+3	Z=+2			
	499,88	**749,32**	**526,53**	**493,86**		**508,56**	**762,35**		Correc.	
D19	1960	680	236	338	3214	192	95	287	200,9	**5,88**
D11	1000	474	58	107	1639	65	66	131	91,7	**5,30**
D9	1660	314	n.d.	104	2078	63	50	113	79,1	**3,67**

Following the above line of thought on the relationship between CUG ambiguity and pathogenesis, we have quantified leucine incorporation at the CUG codon under oxidative stress (Table 3. 5). As for acidic pH, the immune system uses oxidative stress as an important weapon against invading pathogens (Vazquez-Torres and Balish, 1997; Miller and Britigan, 1997). To simulate such condition, cells were grown in 1.5 mM of H_2O_2. Leucine and serine incorporation was 4.0 % ± 0.71 and 96.0 % ± 0.71, respectively.

Table 3. 5 – Leucine incorporation at the CUG codon on cells grown in the presence of 1.5 mM H_2O_2. The abundance of serine and leucine peptide species, obtained from independent HPLC-MS experiments. The % of leucine incorporation was obtained as explained above.

File	Serine - Peaks					Leucine - Peaks				%Leu
	Ser		Ser-OH	Ser-P				Total		
	Z=+3	Z=+2	Z=+3	Z=+3	Total	Z=+3	Z=+2			
	499,88	**749,32**	**526,53**	**493,86**		**508,56**	**762,35**		Correc.	
E13	1060	213	46	69	1388	43	42	85	59,5	**4,110535**
E16	2610	536	256	292	3694	174	86	260	182	**4,695562**
E17	1250	450	61	128	1889	27	58	85	59,5	**3,053631**
E18	1100	237	63	100	1500	43	37	80	56	**3,598972**

The above results showed unequivocally that CUG ambiguity varies between cell type and between physiological conditions. This is surprising because it suggests that somehow charging of the $tRNA_{CAG}^{Ser}$ is regulated and sensitive to the surrounding environment. The molecular mechanism underlying such regulation is still unknown, however it will be most interesting to unravel it and establish a link between such regulatory system and *C. albicans* adaptation (see below). More importantly, these data raised the questions of "how much CUG ambiguity can be tolerated by *C. albicans*?" and can CUG identity be reverted from serine back to leucine?" In order to answer these new questions CUG ambiguity was artificially increased *in vivo* in *C. albicans*, as described below.

3.2.3. *C. albicans* tolerates partial reversion of CUG identity

To engineer increased CUG ambiguity *in vivo* in *C. albicans*, a *S. cerevisiae* $tRNA_{CAG}^{Leu}$ gene, which was derived from the *S. cerevisiae* $tRNA_{UAG}^{Leu}$ gene through mutation of the first anticodon wobble base (U to G), was used (Figure 3. 14). The reporter protein gene was also inserted in the same plasmid already containing the mutated $tRNA_{CAG}^{Leu}$ (pUA15). This resulted in a new plasmid, named pUA65. The new leucine tRNA gene was cloned between the *Xho I* and *Ava III* restriction sites and the CUG reporter gene between the sites *Hind III* and *Pst I* (see section 2.2.2.2).

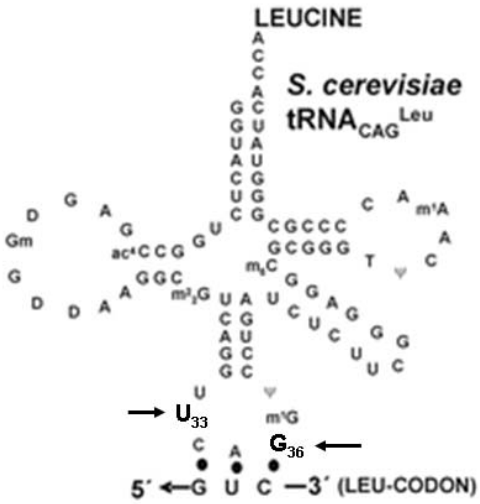


Figure 3. 14 – Engineered tRNA$_{CAG}^{Leu}$ gene from *S. cerevisiae.*
The wobble base of tRNA$_{CAG}^{Leu}$ gene from *S. cerevisiae* was mutated to G in order to increase the decoding efficiency of this tRNA (Miranda, 2007).

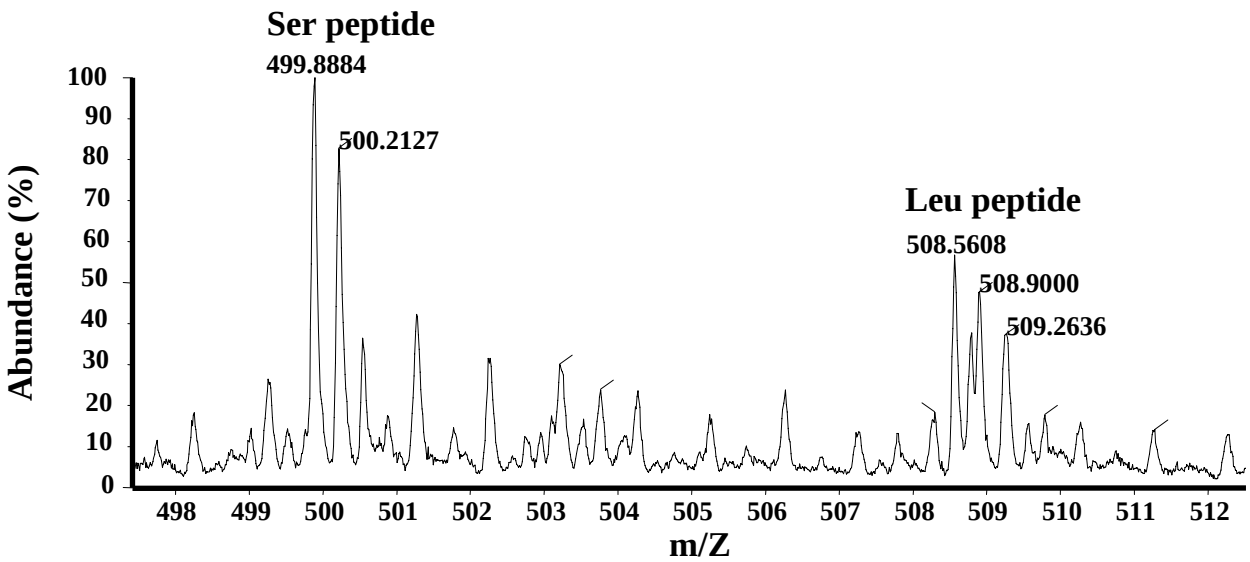


Figure 3. 15 – Leucine incorporation at CUG codons in vivo in engineered *C. albicans* cells.
HPLC-MS analysis of the CUG reporter protein expressed in the pUA65 recombinant *C. albicans* cell lines. There was a sharp increase of the relative amount of the peptide containing leucine at the CUG-codon position.

The pUA65 vector was then transformed into *C. albicans* CAI-4 and positive clones were then used to purify the reporter protein. For this, transformed cells were grown at 30ºC in liquid cultures, overnight to an OD_{600} of 1.5. The reporter protein was purified, digested with thrombin and enterokinase and analyzed by mass-spectrometry as described above. Remarkably, the mass-spectra showed a dramatic increase in the abundance of the leucine peptide (Figure 3. 15, Table 3. 6).

Table 3. 6 – Leucine incorporation at the CUG codon on highly ambiguous cells.
The abundance of serine and leucine peptide species, obtained from independent HPLC-MS experiments.
The % of leucine incorporation was obtained as explained above.

File	Serine - Peaks					Leucine - Peaks				%Leu
	Ser		Ser-OH	Ser-P				Total		
	Z=+3	Z=+2	Z=+3	Z=+3	Total	Z=+3	Z=+2			
	499,88	**749,32**	**526,53**	**493,86**		**508,56**	**762,35**		Correc.	
F8	1420	315	83	132	1950	932	199	1131	791,7	**28,88**
F9	1900	365	116	161	2542	1200	217	1417	991,9	**28,07**
F12	607	131	70	82	890	392	76	468	327,6	**26,91**

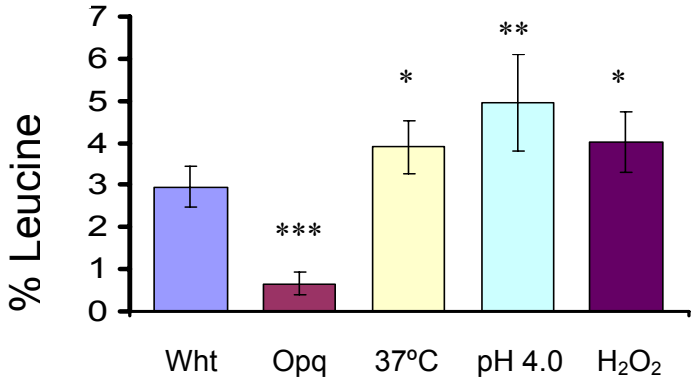
In these cells, the measured leucine and serine incorporation was of 27.9% ± 1.0 and 72.1% ± 1.0, respectively. This unanticipated result provided the first unequivocal evidence for dual identity of the CUG codon in *C. albicans*. In other words, *C. albicans* tolerates partial reversion of identity of the CUG codon without apparent decrease of fitness. It will now be most interesting to further increase CUG ambiguity and determine whether the identity of the CUG codon can be completely reversed in *C. albicans*. Interestingly, this data is in line with the above results showing that the CUG codon is ambiguous in wild type cells and suggests that the *C. albicans* proteome is not disrupted by serine or leucine insertion at CUG positions.

3.3. Discussion

Genetic code alterations pose important new biological questions whose answers remain elusive, especially about the mechanisms by which they evolve, their potential selective advantage and their physiological acceptability. We have chosen the *Candida*

genetic code change as a model to elucidate such questions. The studies described in this chapter indicate that *C. albicans* decodes the CUG codon ambiguously, that such ambiguity changes between cell types, physiological conditions (Figure 3. 16) and, moreover, that leucine incorporation at CUG positions can be sharply increased up to 28% (Figure 3. 17).

A



B

	% Serine. incorp.	SD.
White	**97.04**	**± 0.49**
Opaque	**99.34**	**± 0.28**
37ºC	**96.10**	**± 0.64**
pH 4.0	**95.05**	**± 1.14**
H_2O_2	**95.97**	**± 0.71**

Figure 3. 16 - CUG ambiguity is sensitive to environmental cues.
In order to determine whether the level of leucine **(A)** and serine **(B)** incorporation *in vivo* was sensitive to environmental change, *C. albicans* cells were grown at 37ºC, in 50 mM citrate buffer at pH 4.0 and in presence of 1.5 mM H_2O_2. To determine the level of ambiguity of the CUG codon in opaque cells, an opaque cell line was selected from a white colony by successive plating on agar plates until a culture containing more than 90% of opaque cells was obtained. *$p < 0.05$; ** $p < 0.001$; *** $p << 0.001$.

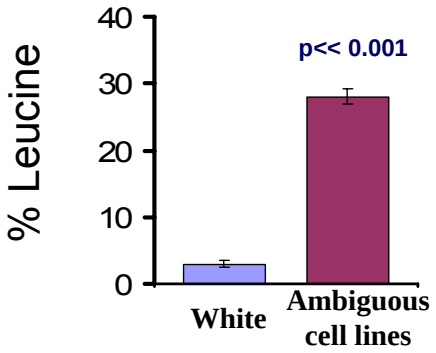


Figure 3. 17 – Leucine incorporation on highly ambiguous cell lines.
Peptide quantification shows that 27.9% ± 1.00 of the peptides incorporate leucine and 72.1% ± 1.00 incorporate serine at the CUG codon corresponding to an increase in decoding error of 2800 fold above standard mRNA decoding error (10^{-4}). P-value is of p<0,001.

These data clearly support the "*Ambiguous Intermediate Theory*" for the evolution of the genetic code, discussed in section 1.4.1 (Santos and Tuite, 1995; Schultz and Yarus, 1994), because it demonstrated that the *Candida* spp. genetic code alteration evolved through codon decoding ambiguity. This data also supports previous data obtained in our and other laboratories showing that organisms can tolerate high levels of codon ambiguity (Bacher et al., 2003; Pezo et al., 2004; Santos et al., 1996; Santos et al., 1999; Chin et al., 2003). However, codon decoding ambiguity is toxic, decreases fitness and may ultimately lead to cell death, as is the case in multicellular organisms (Lee et al., 2006; Nangle et al., 2002). For these reasons, evolution of genetic code alterations, through such codon ambiguity mechanisms, is most interesting.

The partial reversion of CUG codon identity from serine back to leucine, which was demonstrated by 27.9% of leucine incorporation, was carried out to expose the malleability of the genetic code in *C. albicans* and to reconstruct the high level of CUG ambiguity existent in the *Candida* ancestor. Surprisingly, these highly ambiguous cell lines were very heterogeneous in both cell and colony morphologies. Colonies were characterised by the formation of aerial hyphae and white-opaque sectoring, whereas its cells were larger and often formed very long filaments (Miranda, 2007). Indeed, morphological variation, growth at high temperature and yeast-hypha transition, as well as proteinase and lipase secretion and various adhesins, all play important roles in infection (Calderone and Fonzi, 2001; Berman and Sudbery, 2002). The phenotypic diversity induced by CUG ambiguity exposed some of these virulence traits and suggests that increasing CUG ambiguity under stress may be relevant to pathogenesis (Miranda, 2007). Furthermore, morphological variation alters cell surface antigens which are a safeguard against the immune system.

Considering that the basal mRNA decoding error in yeast is in the order of 10^{-5} (Stansfield et al., 1998) the measured leucine misincorporation rates in *C. albicans* represents an 66- up to 490- fold increase in decoding error in opaque cells and in cells grown under oxidative stress, respectively. Moreover, such increase in the mRNA decoding error can be as high as 2790-fold when compared to the typical error of translation. These results also unequivocally show that the $tRNA_{CAG}^{Ser}$ is charged *in vivo* with both serine and leucine, and that the mischarged leu-$tRNA_{CAG}^{Ser}$ is neither edited by

the LeuRS nor discriminated by translation elongation factor 1A (eEF1A). This event results in a wide proteome destabilization, which is likely to trigger morphogenesis, and raises intriguing questions about the complexity of the *C. albicans* proteome. These issues are addressed in the following chapters, which focus on the calculation of the number of different proteins that can be generated from the *C. albicans* gene set and on how *C. albicans* manipulates leucine misincorporation at CUG positions.

4. The impact of CUG ambiguity in *C. albicans* biology

The results presented in this chapter are part of the work published in the following papers:

Gomes, A.C., Miranda, I., Silva, R. M, Moura, G.R, Thomas, B., Akoulitchev, A. and Santos, M.A.S. (2007) "A Genetic Code Alteration Generates a Proteome of High Diversity in the Human Pathogen *Candida albicans*" *Genome Biology* **8:**R206; doi:10.1186/gb-2007-8-10-r206.

Silva, R. M., Paredes, J. A., Moura, G., Manadas, B., Costa, T.L., Miranda, I., Gomes, A.C, Koerkamp, M. J. G., Perrot, M., Holstege, F., Boucherie, H., Santos, M.A.S. (2007) **"**Critical roles for a genetic code alteration in the evolution of the genus *Candida*." *The EMBO Journal* **26,** 4555–4565; doi:10.1038/sj.emboj.7601876

4.1. Introduction

Living systems have evolved highly accurate translational machineries, however, protein synthesis is not an error free process. The mistranslation of mRNA results in synthesis of aberrant proteins, which either have amino acid substitutions or are truncated, and most of them are unable to fold properly. Therefore, the ultimate consequence of mistranslation is the production of misfolded proteins. The presence of such aberrant proteins can be deleterious or even lethal to the cell (Figure 4. 1).

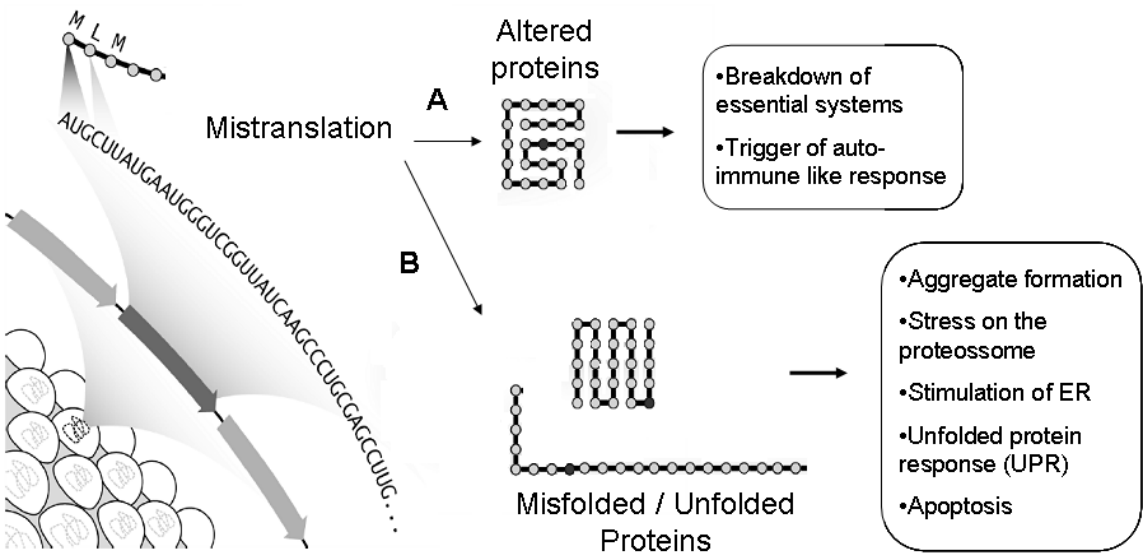


Figure 4. 1 – The impact of mistranslation on the cell biology.
Mistranslation results in formation of either **(A)** altered or **(B)** misfolded and unfolded proteins, which impose a burden on cell physiology. The misread residue is represented as a black dot. Adapted from (Drummond et al., 2005; Nangle et al., 2006).

The precise impact of mistranslation in the cell physiology is still poorly understood. However, mistranslation is receiving increased attention because misfolded proteins promote the formation of aggregates, stress the endoplasmatic reticulum (ER) and trigger the unfolded protein response (UPR), and ultimately lead to cell death by apoptosis. These responses are associated to several pathologies, namely formation of cataracts (Ikesugi *et al.*, 2006), alcoholic liver disease (Kaplowitz and Ji, 2006), diabetes (Harding and Ron, 2002), mitochondrial encephalomyopathies MELAS and MERRF (Yasukawa *et al.*, 2000), cancer (Ma and Hendershot, 2004), and several neurodegenerative diseases, namely

Alzheimer's, Huntington's and Parkinson's diseases (Lindholm et al., 2006; Rao and Bredesen, 2004, reviewed in Chiti and Dobson, 2006; Zhao and Ackerman, 2006).

Most of those diseases are multifactorial and the agent that causes the misfolding of proteins is unknown. Despite this, some of them were directly linked with mutations on ribosomal protein genes, on translation elongation factor genes, tRNA genes and on aminoacyl synthetase genes, which induce mistranslation. For instance the MELAS and MERRF diseases are caused by mutant tRNAs (Yasukawa et al., 2000; Yasukawa et al., 2001), and a tRNA mutation is associated to hypertension and dyslipidemia, which are risk factors for cardiovascular diseases (Wilson *et al.*, 2004). A mutant form of the translation factor eIF2B is also associated to leukoencephalopathy with vanishing white matter, which is a neurological disease (Leegwater *et al.*, 2001). Further, mutations in the ribosomal proteins S19 and S24, which result in abnormal processing of ribosomal RNA, are responsible for a congenital anaemia, known as the Diamond-Blackfan anaemia (Flygare and Karlsson, 2007; Draptchinskaia et al., 1999; Gregory et al., 2007). Finally, mutant TyrRS and GlyRS are involve in Charcot-Marie-Tooth neuropathies (Jordanova et al., 2006; Seburn et al., 2006) and a mutant AlaRS is involved in cerebellar Purkinje cell loss and ataxia (Lee *et al.*, 2006).

Organisms have evolved mechanisms to minimize mRNA mistranslation. For example, the universally conserved heat-shock response, the proteasome and molecular chaperones, which refold various misfolded proteins, form a safety network against aberrant proteins (reviewed in Lindquist and Craig, 1988; Pickart and Cohen, 2004). The cytosolic and nuclear proteins targeted for degradation are covalently modified at lysine residues with ubiquitin, which is a small (76 amino acids), but highly conserved polypeptide (Thrower et al., 2000; Weissman, 2001). These tagged misfolded proteins are targeted for degradation by the proteasome, which also degrades many other correctly folded proteins. Indeed, besides protein quality control the proteasome is also involved in many diverse cellular processes, namely regulation of cell cycle progression, signal transduction or antigen processing (reviewed in Kostova and Wolf, 2003; Pickart and Cohen, 2004).

In eukaryotic cells, a wide range of proteins are synthesized in ribosomes attached to the ER, namely secreted and membrane proteins, and the accumulation of misfolded proteins imposes stress on the ER, which activates the UPR signal transduction pathway, causing temporary remodelling of the ER (Schroder and Kaufman, 2005). The balance of ER resident proteins is shifted to remove aberrant substrates and to restore the ER capacity to efficiently mature resident and exported proteins. The UPR pathway functions as a tripartite signal that involves *(1)* increasing the expression of housekeeping proteins that can work toward properly folding the misfolded proteins, *(2)* attenuating the secretory pathway load by decreasing the expression of secretory cargo, and *(3)* increasing the capacity for *ER*-*a*ssociated protein *d*egradation (ERAD) (Oyadomari et al., 2006; Pearse and Hebert, 2006, reviewed in Bernales *et al.*, 2006).

Despite the negative effects described above, mistranslation plays an important role in cell physiology. For instance, 30% of the newly synthesized proteins in HeLa, lymph node and dendritic cells are *d*efective *ri*bosomal *p*roducts (DRiPs) that arise from missense, frameshifting and ribosome drop off at mRNA pausing sites. This is important for the surveillance of the immune system because the peptides resulting from proteasome degradation of DRiPs are a major source of peptides for MHC class I molecules (Princiotta et al., 2003; Eisenlohr et al., 2007; Yewdell and Nicchitta, 2006). Also, mistranslation can have positive evolutionary roles, in particular when cells are submitted to stress, namely starvation (Parker and Precup, 1986). In this case, increased mistranslation results in synthesis of arrays of altered proteins that provide a selective advantage for the cell. For example, *Saccharomyces cerevisiae* has evolved a system to exploit hidden genetic variation via conformational alteration of the translation termination factor Sup35p, namely the [PSI^+] prion. Strains harbouring the [PSI^+] prion experience generalized stop codon readthrough of genes and pseudogenes, which induces global proteome disruption and results in morphological variation (Uptain and Lindquist, 2002; True et al., 2004). These [PSI^+] strains have a short-term survival advantage, when grown under stress conditions, over strains that lack it [psi^-], since increased stop codon readthrough generates beneficial phenotypes, though [PSI^+] and [psi^-] strains have identical growth rate under normal growth conditions (Tuite and Lindquist, 1996; Eaglestone et al., 1999).

As demonstrated in the previous chapter, *C. albicans* mistranslates constitutively and tolerates amino acid mis-incoropration at rates 2790 fold higher than the typical error rate. However, such increase in CUG ambiguous decoding results in genome instability, increases morphogenesis and generates new phenotypes (Miranda, 2007). Nevertheless, such mistranslation did not decrease growth rate (Miranda, 2007). This unanticipated result shows that *C. albicans* responds in unique ways to mistranslation because in all other studied cases similar mistranslation had a strong impact on growth rate (Bacher and Ellington, 2001; Pezo et al., 2004; Santos et al., 1996). This raises the hypothesis that *C. albicans* may have evolved unique mechanisms to circumvent the deleterious effects of mistranslation.

Therefore, to better understand the impact of ambiguous decoding of the CUG codon, and to obtain a full picture of its global effect on *C. albicans* biology, the genomic distribution and the usage of the CUG codon were studied in detail. This was achieved by determining Specific Codon Usage (*SCU*) values for the CUG codon. As the genetic code is degenerated, and one amino acid can be coded by more than one codon, the *SCU* is a simple measure of non-uniform usage of synonymous codons in coding sequences (Sharp and Li, 1986). Indeed, the pattern of codon usage in genes reflects a complex balance among biases generated by mutation, selection and random genetic drift, such biases are due to *(i)* diversity in the (% G+C) at the third codon position (Alvarez *et al.*, 1994); *(ii) (ii)* abundance of tRNAs (Ikemura, 1985); *(iii)* overall base composition of genes (Ellis and Morrison, 1995); and differences in both *(iv)* gene expression level (Pouwels and Leunissen, 1994) and *(v)* the location of the genes in the genome (Chiapello *et al.*, 1999).

The impact of CUG ambiguity on protein synthesis was also evaluated using mathematical models. For this, the number of CUG codons per gene was correlated to its Codon Adaptation Index (CAI). The latter is a measure of the relative adaptiveness of the codon usage of a gene towards the codon usage of highly expressed genes and predicts the expression level (Sharp and Li, 1987; Sharp et al., 1986). All these studies were carried out using the *C. albicans* genome assembly 19, of 17/08/2005, which represents its haplotype and contains 6438 genes (Braun *et al.*, 2005) (http://candida.bri.nrc.ca/candida/).

4.2. Results

4.2.1. *C. albicans* has a statistical proteome

The *Candida albicans* genome, which contains 6438 genes, was analysed using the ANACONDA software built by the Bioinformatics group of Aveiro (Moura *et al.*, 2005). Analysis of the codon content of each gene revealed that the *C. albicans*' genome contains 13,074 CUG codons, distributed over 66% of its genes, at a frequency of 1 to 38 CUGs *per* gene (Figure 4. 2), though most of them (57.7%) have between 1 to 5 CUG codons.

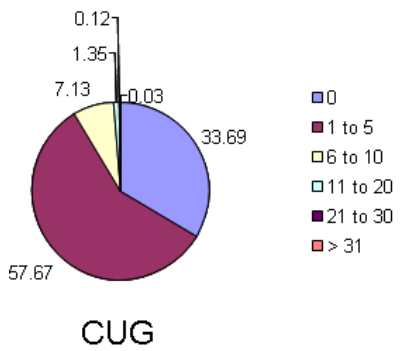


Figure 4. 2– CUG codon distribution over *C. albicans* genome.
In the genome of *C. albicans* one third of its genes do not have CUG codons. The majority of its genes, 57.7%, contain between 1 to 5 CUG codons, while 7.1% of its genes have between 6 and 10, and only a rather small fraction of genes have more than 10 CUG codons.

Since codon-pair context influences mRNA decoding accuracy (Berg and Silva, 1997; Murgola et al., 1984; Bossi and Ruth, 1980), a genome wide codon-context survey of the CUG codon was carried out for the genome of *C. albicans.* Similar analysis were also carried out for *S. cerevisiae, S. pombe, A. fumigatus, S. bayanus, S. mitikae, S. paradoxus, C. glabrata, D. hansenii, K. lactis* and *Y. lipolytica*. These analyses were also performed with the ANACONDA software by taking advantage of its statistical methodologies for codon-context analysis, namely contingency tables and residual analysis (Moura *et al.*, 2005). The data obtained was displayed using a colour coded map that represented the 3' and 5' contexts of CUG codons from the genomes analysed (Figure 4. 3). This study failed to identify any particular context bias for the CUG codon in *C.*

albicans, indicating that leucine and serine are randomly inserted at CUG positions. This indicates that CUG ambiguity has a global impact on the *C. albicans* proteome.

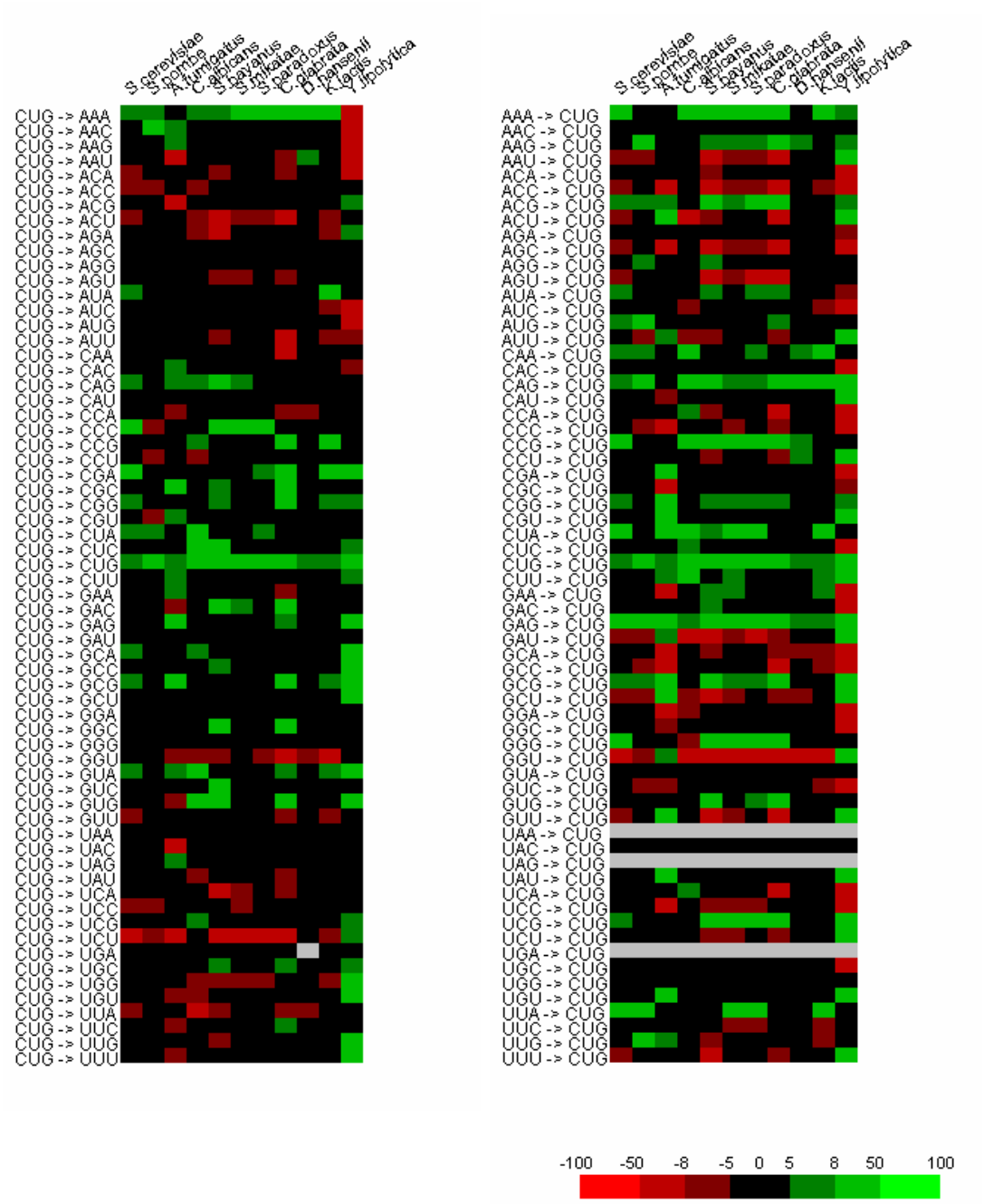


Figure 4. 3 – CUG codon context analysis.
The 5'- and 3'- context CUG codons from the 11 genomes tested. Red represents rejected contexts and green represents preferred contexts. The neutral contexts are in black.

Table 4. 1 – Expansion of the *C. albicans* proteome through CUG ambiguity. Determination of the total number of combinatorial proteins encoded in *C. albicans* genome.

CUG *per* gene ***n***	No. Genes with ***n*** CUG **(A)**	No. Possible proteins (2^n) **(B)**	Total number of proteins (**A** x **B**)
0	2169	1	2169
1	1439	2	2878
2	953	4	3812
3	609	8	4872
4	423	16	6768
5	289	32	9248
6	174	64	11136
7	103	128	13184
8	93	256	23808
9	44	512	22528
10	45	1024	46080
11	22	2048	45056
12	22	4096	90112
13	15	8192	122880
14	6	16384	98304
15	7	32768	229376
16	6	65536	393216
17	3	131072	393216
18	2	262144	524288
19	2	524288	1048576
20	2	1048576	2097152
21	3	2097152	6291456
22	0	4194304	0
23	1	8388608	8388608
24	2	16777216	33554432
25	0	33554432	0
26	0	67108864	0
27	2	134217728	268435456
28	0	268435456	0
29	0	536870912	0
30	0	1073741824	0
31	0	2147483648	0
32	0	4294967296	0
33	1	8589934592	8589934592
34	0	17179869184	0
35	0	34359738368	0
36	0	68719476736	0
37	0	1.37439E+11	0
38	1	2.74878E+11	2.7488E+11
Total:	**6438**		**2.8379E+11**

The data supported the hypothesis that leucine misincorporation at CUG codons in *C. albicans* is not dependent on CUG context and raised the opportunity to quantify the consequences of CUG ambiguity on the *C. albicans* proteome. For this, the theoretical number of novel proteins generated by CUG ambiguity was determined by the expression 2^n, where n is the total number of CUGs per gene. The data showed that the *C. albicans* proteome expands exponentially with the increase in the number of CUG codons per gene and that the 6438 protein encoding genes of *C. albicans* have the potential to produce the staggering number of 2.8379×10^{11} different proteins through CUG ambiguity (Table 4. 1).

Therefore, genes containing more than 2 CUGs produce arrays of related protein molecules containing leucine or serine inserted randomly at CUG positions. This is of profound biological significance and implies that the *C. albicans* proteome has a statistical nature, because each cell has a unique combination of proteins. Considering that the rates of leucine incorporation at CUG codons vary with different physiological conditions, the impact of such variation on the *C. albicans* proteome can be calculated by determining the associated probability of one gene, with n CUG codons, to have i leucines incorporated at these CUG positions, under each growth condition. This was calculated by expanding the binomial distribution (Equation 4. 1):

$$b_{(i,n,p)} = \frac{n!}{i!(n-1)!} p^i (1-p)^{n-i}$$

(Equation 4. 1)

Where n is the total number of CUG codons per gene, p is the probability of leucine incorporation at CUG positions in different growth conditions, and i is the number of CUGs being decoded as leucine. As a working example, the probability of synthesis of different proteins with 0, 1, 2, or 3 leucines incorporated during translation of mRNA containing 3 CUG codons, for the ambiguity levels determined experimentally under different growth conditions, were calculated (Table 4. 2).

Table 4. 2– Probabilistic decoding of a gene with 3 CUG codons.
The probability of synthesis of proteins with 0, 1, 2 or 3 leucines incorporated at CUG codons, in cells growing in different physiological conditions. $q_{(Ser)}$ and $p_{(Leu)}$ are the measured serine and leucine incorporation rates, respectively. The number of proteins was determined by expanding the binomial distribution (Equation 4. 1), with n=3, i=0, 1, 2 or 3 and p=$p_{(Leu)}$ for each physiological condition.

	$q_{(Ser)}$	$p_{(Leu)}$	P(L=0)	P(L=1)	P(L=2)	P(L=3)
White	0.9704	0.0296	9.14E-01	8.36E-02	2.55E-03	2.59E-05
Opaque	0.9934	0.0066	9.80E-01	1.95E-02	1.29E-04	2.85E-07
37ºC	0.9610	0.0390	8.87E-01	1.08E-01	4.39E-03	5.94E-05
pH 4.0	0.9505	0.0495	8.59E-01	1.34E-01	6.99E-03	1.21E-04
H_2O_2	0.9597	0.0403	8.84E-01	1.11E-01	4.68E-03	6.55E-05
pUA65	0.7193	0.2807	3.72E-01	4.36E-01	1.70E-01	2.21E-02
Typ. Error	0.9999	0.0001	1.00E+00	3.00E-04	3.00E-08	1.00E-12

The same methodology can be used to determine the probability of a mRNA with n CUGs to generate proteins with only serines at CUG positions (i=0), under the studied physiological conditions ($b_{(0,n,p)}$). Again, the differences in leucine misincorporation between 2.96% and 28.1% have significant consequences for the synthesis of combinatorial proteins (Table 4. 2). For instance, in the highly ambiguous cell lines (pUA65), the probability of synthesis of proteins with serines only at CUG positions is 0.5 for genes with 2 CUGs, but it is 0.01 for genes with 14 CUGs. Likewise, for all other conditions, such probability decreases as the number of CUGs per gene increases. This effect is strongly affected by small increases of leucine misincorporation (Figure 4. 4).

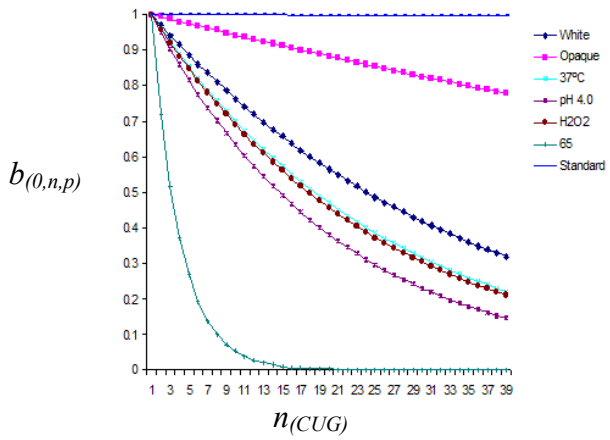


Figure 4. 4 – Probability of synthesis of proteins without leucine at CUG codons.
Probability of synthesis of *C. albicans* proteins with 100% serine incorporated at CUG positions for genes with n CUGs. This probability is high for genes with 1 CUG codon, but decreases sharply as the number of CUGs per gene increases. Each *C. albicans* protein is composed by a statistical mixture of molecules that may contain leucine or serine at CUG positions. This data was obtained using (Equation 4. 1, with the *p(leu)* of each tested condition, n is the number of CUG codons and i = 0.

From these analyses one can infer, 1) the associated probabilities of a given CUG codon to be decoded as serine or leucine; and 2) the total number of different proteins that can be generated from the ambiguous CUG decoding event. However, these analyses only provide a theoretical framework to understand the potential of *C. albicans* to generate new proteins from ambiguous CUG decoding and do not quantify the number of proteins present in a *C. albicans* cell. For this, one has to take into consideration the number of molecules *per* cell for each protein encoded by the *C. albicans* genome. Recent studies carried out by Ghaemmaghami and colleagues (Ghaemmaghami *et al.*, 2003) demonstrated that protein abundance in yeast ranges from 50 up to more than 10^6 molecules per cell. Since *C. albicans* and *S. cerevisiae* are phylogenetically related one can assume that overall protein expression values are similar and use the *S. cerevisiae* data set as a reference for *C. albicans*. If so, one should also assume that 1) all *C. albicans* genes are expressed and 2) the abundance of proteins (N_{total}) is 5,000 molecules/cell for the 10% of genes with lowest CAI values, 3) of 50,000 molecules/cell for the 10% of genes with highest CAI values (Ghaemmaghami *et al.*, 2003), and 4) of 20,000 molecules/cell for the remaining 80% of genes. These assumptions allow one to estimate the number of different protein molecules that are present within a *C. albicans* cell and the number of novel proteins that are generated (N_{novel}) (Equation 4. 2) from each mRNA and from the entire set of mRNAs (transcriptome).

$$N_{novel} = N_{total} \times (1 - b_{(0,n,p)})$$

(Equation 4. 2)

Interestingly, the impact of ambiguous CUG decoding is very strong for highly expressed genes, but is weaker for genes whose expression is low. This effect is highlighted in Table 4. 3 and Table 4. 4, where the number of different proteins arising from ambiguous CUG decoding of genes with high and low expression levels are displayed. The selected genes for this analysis were *CDC3,* which has a CAI of 0.694, and *RAD17,* which has a CAI of 0.448 (see CAI values in the next section, p.126). These genes have 3 CUG codons, and each of them belongs to the group of the 10% most and 10% least expressed *C. albicans* genes, respectively.

Table 4. 3 - Novel proteins produced by ambiguous decoding of mRNAs whose genes have high CAI value.

		Native	Novel Proteins							
Condition	$p_{(Leu)}$	**SSS**[1]	**SSL**[2]	**SLS**[2]	**LSS**[2]	**LSL**[3]	**SLL**[3]	**LLS**[3]	**LLL**[4]	**Total**
White	0.0296	45691	1393	1393	1393	42	42	42	1	4306
Opaque	0.0066	49020	324	324	324	2	2	2	0	978
37ºC	0.0390	44374	1801	1801	1801	73	73	73	2	5624
pH 4.0	0.0495	42938	2235	2235	2235	116	116	116	6	7059
H_2O_2	0.0403	44194	1856	1856	1856	77	77	77	3	5802
pUA65	0.2807	18604	7261	7261	7261	2834	2834	2834	1106	31391
Typ. Error	0.0001	49986	4	4	4	0	0	0	0	12

[1] N= 50,000 x $b_{(0,3,p)}$; [2] N= [50,000 x $b_{(1,3,p)}$] / 3; [3] N= [50,000 x $b_{(2,3,p)}$] / 3; [4] N= 50,000 x $b_{(3,3,p)}$

Table 4. 4- Novel proteins produced by ambiguous decoding of mRNAs whose genes have low CAI value.

		Native	Novel Proteins							
Condition	$p_{(Leu)}$	**SSS**[1]	**SSL**[2]	**SLS**[2]	**LSS**[2]	**LSL**[3]	**SLL**[3]	**LLS**[3]	**LLL**[4]	**Total**
White	0.0296	4569	139	139	139	4	4	4	0	429
Opaque	0.0066	4901	32	32	32	0	0	0	0	96
37ºC	0.0390	4437	180	180	180	7	7	7	0	561
pH 4.0	0.0495	4293	223	223	223	11	11	11	0	702
H_2O_2	0.0403	4419	185	185	185	7	7	7	0	576
pUA65	0.2807	1860	726	726	726	283	283	283	110	3137
Typ. Error	0.0001	4998	0	0	0	0	0	0	0	0

[1] N= 5000 x $b_{(0,3,p)}$; [2] N= [5000 x $b_{(1,3,p)}$] / 3; [3] N= [5000 x $b_{(2,3,p)}$] / 3; [4] N= 5000 x $b_{(3,3,p)}$

Although CUG ambiguity generates approximately 10% of new molecules of both Cdc3p and Rad17p, it is clear that the total number of new molecules is much higher for Cdc3p (4306 for 3% ambiguity) than for Rad17p (429 for 3% ambiguity) (Table 4. 3 and Table 4. 4). Also, if one focus on the leucine incorporation values of 2.96% and 4.95%, in cells grown at 30ºC and neutral pH and in cells grown at pH 4.0, respectively, one can observe that there is 1.67 fold increase in decoding ambiguity in the latter. However, this corresponds to 2.7 fold increase of proteins containing 2 leucines and 6 fold increase of proteins containing 3 leucines. Finally, by applying this analysis to all the ORFs of *C. albicans*' genome, it was possible to determine the total number of novel proteins *per* cell, which ranges from 1.56 x10^6 in opaque cells to 42.8x10^6 in pUA65 transformed cells,

whereas under standard growth conditions (30 ºC) the number of novel protein molecules is $6.7x10^6$ (Figure 4. 5).

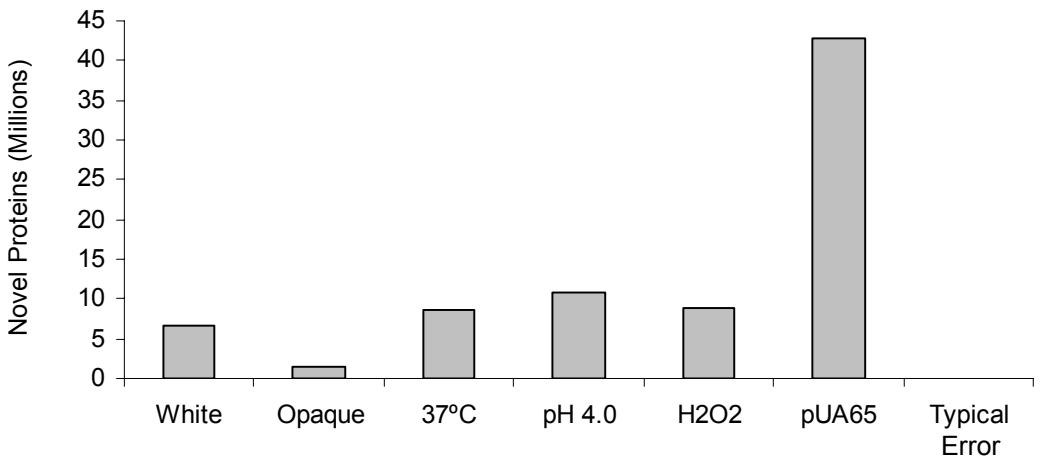


Figure 4. 5 – Novel proteins generated through the ambiguous CUG decoding.
The number of novel proteins generated through CUG ambiguity is correlated with protein expression levels, indicating that the impact of ambiguous CUG decoding is higher in highly expressed proteins. This analysis assumed that protein expression levels in *C. albicans* and *S. cerevisiae* are identical and considered the values of protein expression determined by Ghaemmaghami (2003). This graph was generated by determining the number of novel proteins (Equation 4. 2) arising in each physiological condition for each gene, and then summing up all of them.

The above results illustrate the malleability of the *C. albican*s proteome and indicate that this organism has evolved a novel mechanism to generate protein diversity. Interestingly, if one considers the 6.7 million novel proteins in cells growing under the optimal conditions, this number is quite far from the 283,760 million of potential combinatorial proteins that are encoded by its genome (Table 4. 1). This shows that the complexity of *C. albicans* proteome is not fully exploited under normal growth conditions.

4.2.2. *C. albicans*' genome is optimized for CUG ambiguity

In order to shed new light on the impact of CUG ambiguity of the *C. albicans* proteome a survey of CUG codons was carried out, taking into consideration protein expression levels. This study was complemented with a similar CUG usage study in *S. cerevisiae*, which was used as a reference for this analysis (Figure 4. 6 and Figure 4. 7).

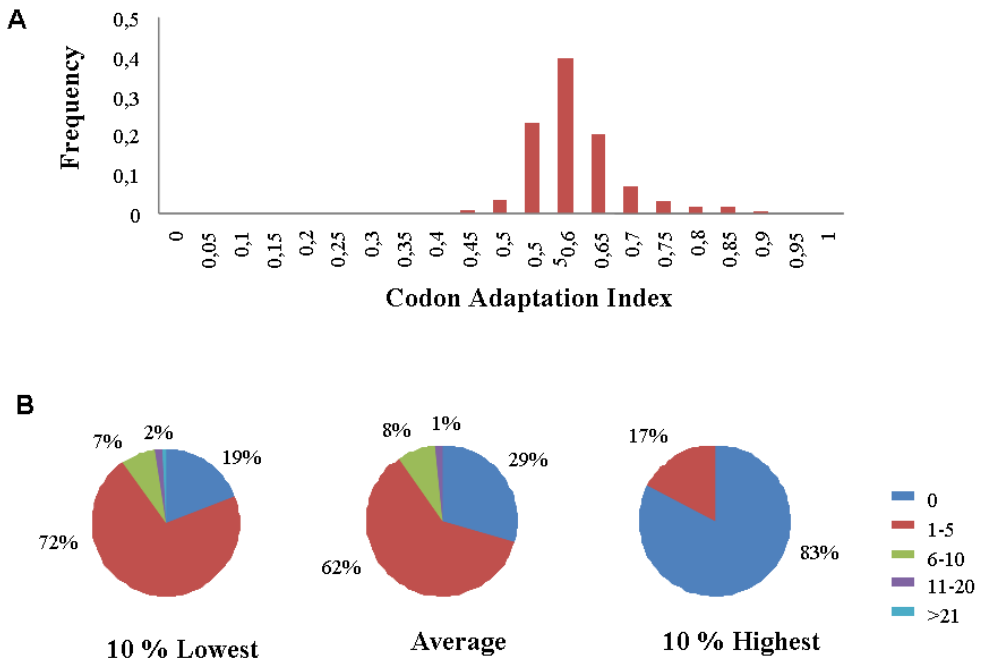


Figure 4. 6 – Usage of *C. albicans* CUG codons in genes with different CAI values.
(A) The CAI values of the *C. albicans* genes were determined using the ANACONDA algorithm (Moura *et al.*, 2005) **(B)** The distribution of CUG codons *per* gene according to their CAI ranking order. In *C. albicans*, CUG codons are strongly repressed in the 10% of genes with highest CAI values. Data obtained from the analysis of *C. albicans* genome (assembly 19) with ANACONDA.

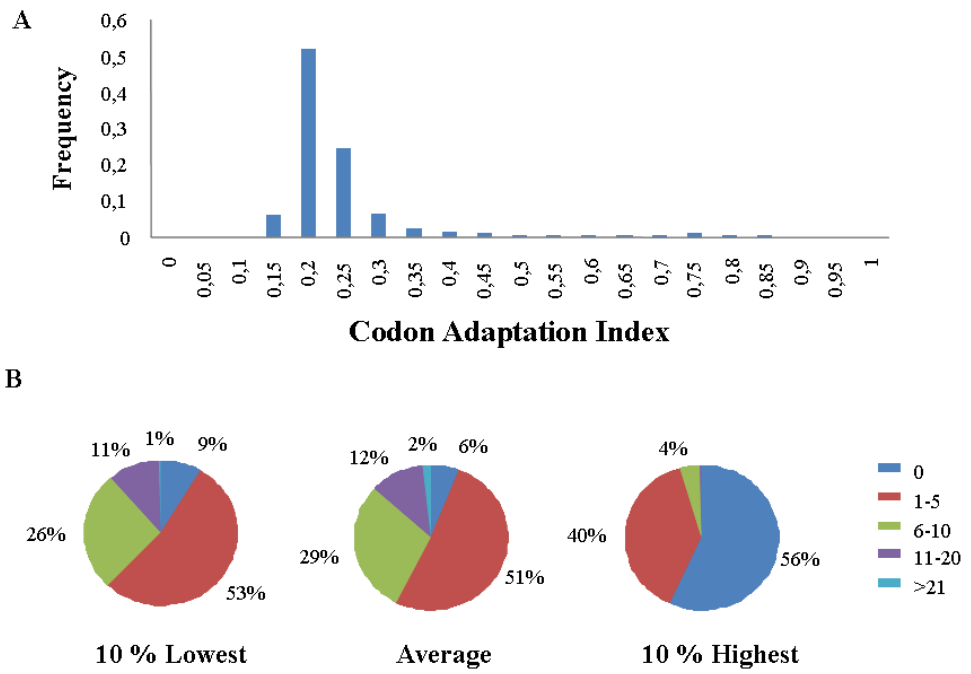


Figure 4. 7 – Usage of *S. cerevisiae* CUG codons in genes with different CAI values.
(A) The CAI values of *S. cerevisiae* genes were determined using the ANACONDA algorithm. **(B)** Distribution of CUG codons per gene according to their CAI ranking order.

Interestingly, *C. albicans* strongly represses CUG usage in the 10% of genes whose expression is highest (higher CAI values) and accumulates them in the 10% of genes whose expression is at the bottom of the CAI scale. Indeed, while 83% of the most expressed genes do not have CUG codons, 81% of genes whose expression is low have at least 1 CUG codon. This is in sharp contrast with CUG usage in *S. cerevisiae*, where only 56% of the highly expressed genes do not have CUG codons. Furthermore, the accumulation of CUG codons is more frequent in the *S. cerevisiae* genome, where, with the exception of the most expressed genes, approximately one third of the genes have more than 5 CUG codons. These observations go in line with the studies made by Massey and colleagues, who have investigated the evolution of CUG codons in both *C. albicans* and *S. cerevisiae*. Those studies were based on alignments of orthologous genes and showed that 98% of the CUG codons in *S. cerevisiae* were reassigned to leucine codons in *C. albicans* (Massey *et al.*, 2003).

The impact of such CUG codon distribution according to protein expression levels becomes clearer if one determines the number of novel proteins synthesized in artificially ambiguous *S. cerevisiae* cells and compares it with the novel proteins arising in white *C. albicans* cells. Conversely to *C. albicans*, transformation of *S. cerevisiae* cells with wild-type the *C. albicans*' $tRNA_{CAG}^{Ser}$ (G_{33}) decreased growth rate by 47.9% (Santos et al., 1996). The mass-spectrometry methodology and the reporter system, described in the previous chapter, was used out in *S. cerevisiae* cells transformed with the *C. albicans* $tRNA_{CAG}^{Ser}$ (G_{33}) and with an engineered U_{33}-$tRNA_{CAG}^{Ser}$, and showed that serine incorporation was of 1.4% and 2.31% for in G_{33} and U_{33} $tRNA_{CAG}^{Ser}$ cell lines, respectively (Silva *et al.*, 2007). The measured serine mis-incorporation in those cells represents 1400 and 2310 fold increase in decoding error (considering $1x10^{-4}$ the typical error) (Stansfield *et al.*, 1998). Note that these values of serine mis-incorporation are below the natural CUG ambiguity in *C. albicans* (2.96%). The number of novel proteins in the U_{33} and G_{33} *S. cerevisiae* cell lines was determined as described above, and in cells with 2.31% of serine misincorporation was of $12.5x10^6$ and in the cells 1.40% with serine misincorporation was of $7.9\ x10^6$ (Figure 4. 8).

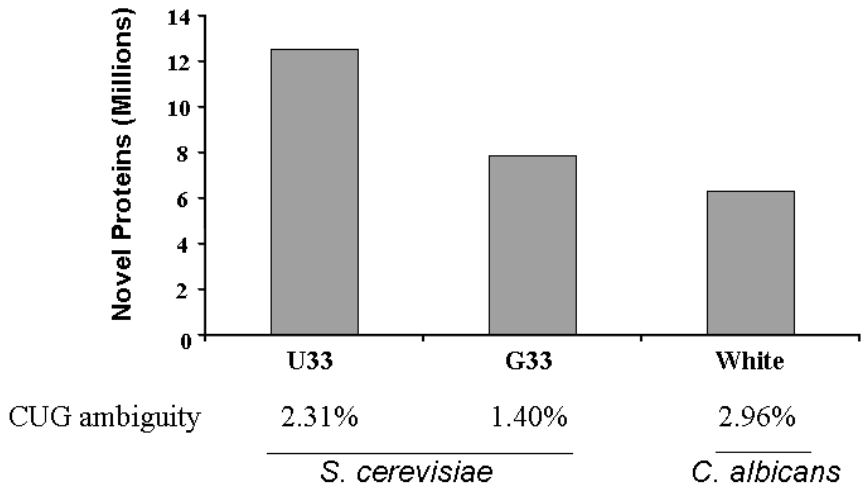


Figure 4. 8 – Novel proteins generated through the ambiguous CUG decoding in engineered *S. cerevisiae.*

In other words, in *S. cerevisiae*, 2.31% of serine mis-incorporation at CUGs resulted in the generation of 12.5 million proteins, whereas as in *C. albicans* 2.96% of leucine mis-incorporation at CUGs resulted in the production of "only" 6.3 million novel proteins. Therefore, similar CUG ambiguity levels resulted in the production of twice the number of novel proteins in *S. cerevisiae*. Furthermore, such mistranslation induced the general stress response in *S. cerevisiae* but did not do so in *C. albicans* (Silva et al., 2007; Enjalbert et al., 2003).

4.2.3. The CUG usage in *C. albicans*

Another important question regarding CUG usage refers to its distribution in the *C. albicans* genome. Its usage frequency is 0.43% (Figure 4. 9), and, therefore, belongs to the category of rare codons.

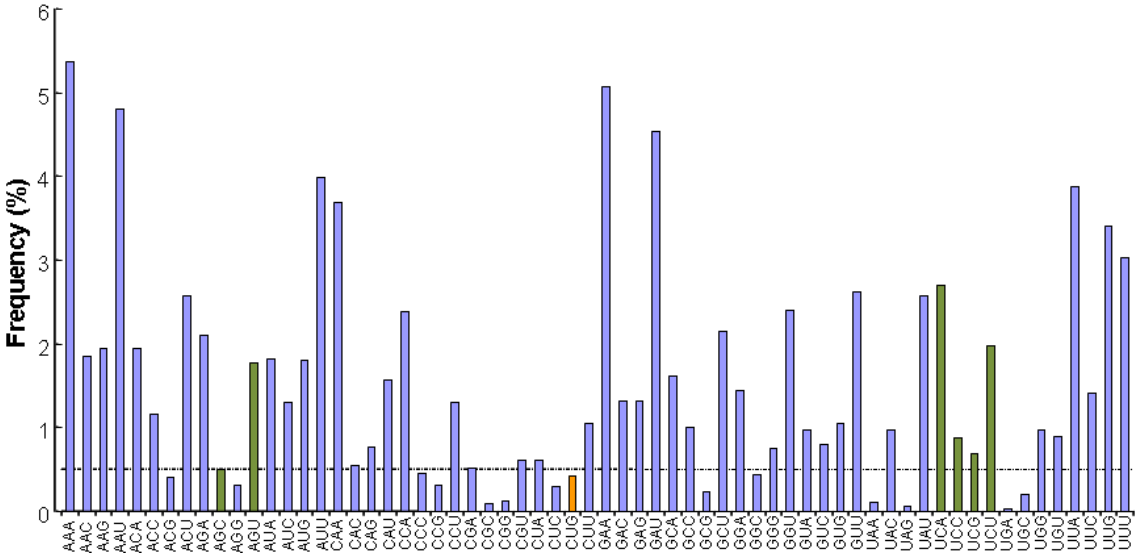


Figure 4. 9 – *C. albicans* codon usage.
The CUG codon is represented in orange, and the other serine codons are in green. The dashed line defines the threshold value that separates rare and non-rare codons, considering that rare codons are used below 0.5% of the time. The total codon count was obtained from the *C. albicans* genome Assembly-19 using ANACONDA.

The contribution of each codon for the entire set of amino acids was measured as the Specific Codon Usage (*SCU*), which reflects their relative usage, and can be calculated for each codon, as follows:,

$$SCU_{(NNN)} = \frac{n_{NNN}}{n_{aa}}$$

(Equation 4. 3)

where the $SCU_{(NNN)}$ is the SCU of a given codon, $n_{(NNN)}$ is the number of times that such codon appears in the genome and $n_{(aa)}$ is the total number of amino acid residues in the entire genome, which are coded by such codon.

The CUG codon, in *Candida* spp, belongs to the serine codon-family, along with the other 6 codons, namely AGC, AGU, UCU, UCA, UCC and UCG. The total usage of each serine codon was determined using ANACONDA, and *SCU* values of each ORF were calculated as described in (Equation 4. 3). The CUG codon is the least used codon of the serine family, but its usage is rather similar to AGC codon (Table 4. 5).

Table 4. 5 – Relative serine-Specific Codon Usage

	AGC	*AGU*	*CUG*	*UCA*	*UCC*	*UCG*	*UCU*
no. Codons	15292	53914	**13074**	81648	26676	20954	60096
SCU	0.056	0.198	**0.048**	0.301	0.098	0.077	0.221

To characterize the distribution of CUGs in the ORFeome, the *SCU* was used as the unit of measurement of the amount of CUG codons, since it allows for data normalization in terms of serine abundance. This choice is based on the fact that the SCU_{CUG} is more informative than the absolute number of CUGs within an ORF. For instance, the SCU_{CUG} of a gene containing a single serine residue, which is encoded by a CUG, is 1.0, whereas the SCU_{CUG} of a gene with 2 CUGs, but with 20 serine residues is 0.1. Therefore, by comparing the *SCU* of the CUG codon one takes into account the relative amount of serine residues that it encodes.

Firstly, CUGs distribution was studied by taking in consideration ORF size, GC content and presence of rare codons. This allowed one to rule out these secondary effects on CUG accumulation in specific functional categories. For this, the *Pearson* correlation coefficient, which is the most common measure for linear associations, was used. It varies between –1 and +1, and a *Pearson* correlation of 0 indicates that there is no correlation between the variables. The coefficients (Table 4. 6) did not show correlation between the *SCU* of the CUG codon and the above tested variables, thus ruling them out.

Table 4. 6 – Pearson correlation matrix
A Pearson correlation analysis was carried out to test the correlation between SCU_{CUG} and the tested parameters for the 6437 ORFs of *C. albicans* genome.

	No. Codons	*% Rare Codons*	*% GC*
SCU_{CUG}	-0.037	-0.122	0.052

SCU_{CUG} values were plotted against each of the tested variables (Figure 4. 10), to visualise the relationship between the analysed parameters, as well as to identify outliers. For this, and to reduce background noise, only the ORFs that had at least one CUG codon were used (4,269 in total), and the data was divided in four groups (coloured

differentially), corresponding to each of the four quartiles, to allow for an easier interpretation of the data.

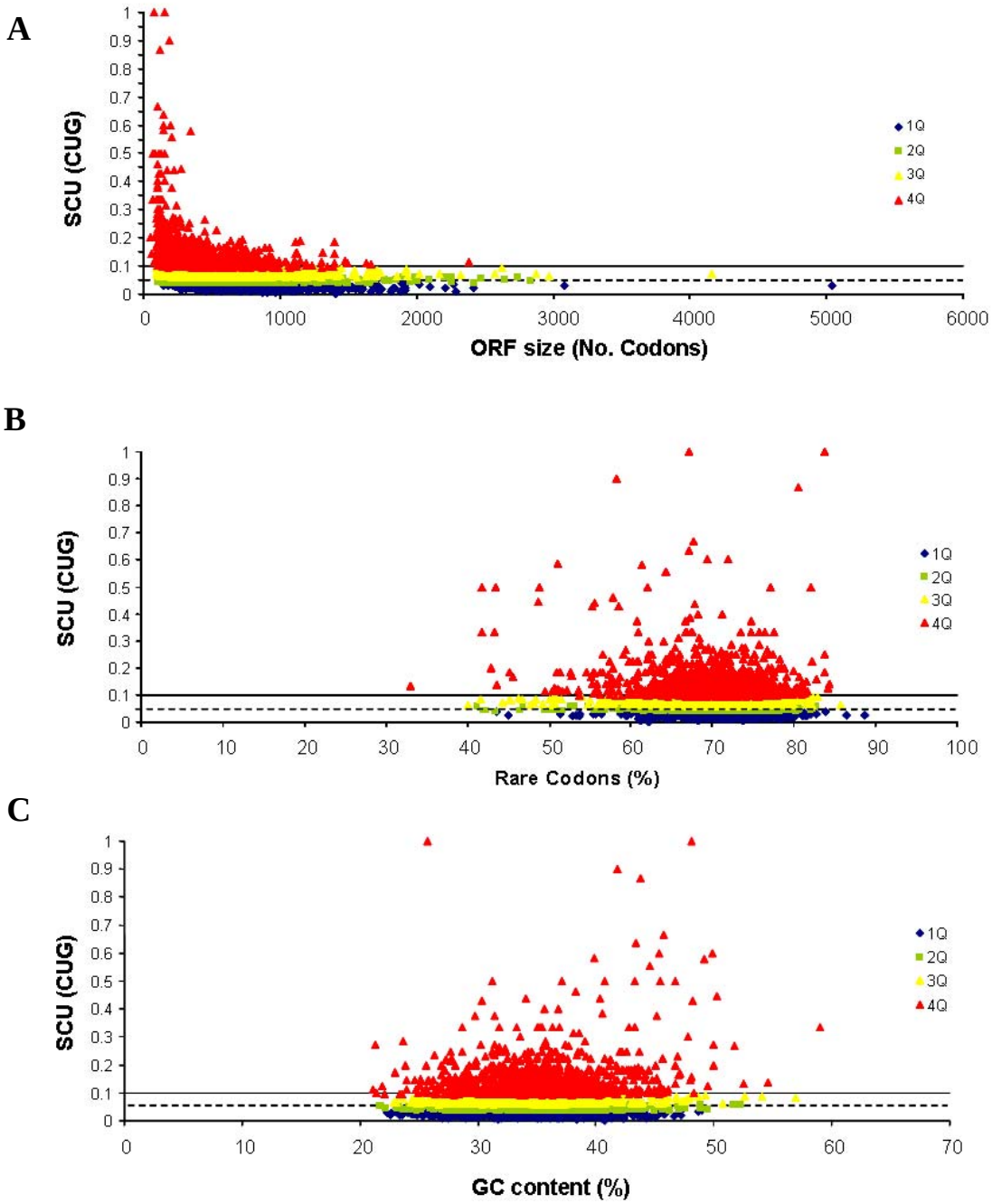


Figure 4. 10 – SCU_{CUG} correlation with ORF size, rare codons and GC content.
(A) Small ORFs tend to have higher levels of CUGs. **(B)** Correlation of SCU_{CUG} with the presence of rare codons in ORFs. **(C)** The SCU_{CUG} and the GC content of ORFs. The SCU_{CUG} correlation with both rare codons and the GC content is very homogenous, and significant trends/bias were not observed. The dashed line indicates the average usage of CUGs and the solid line indicates values that are twice the average. In red are the 25% of the ORFs with the highest usage of CUGs (the 4th quartile), in yellow the 3rd quartile, in green the 2nd quartile and in blue are the 25% of ORFs with the lowest CUG usage.

Therefore, the most surprising result was the accumulation of CUG codons in small ORFs (Figure 4. 10 A), which have evolved recently (Beltrão, P., personal communication). This indicates a fast accumulation of CUGs in genes that are evolving rapidly and that are specific of *C. albicans*. Interestingly 41 ORFs accumulate CUG codons since at least 1 out of 3 serines are coded by a CUG. These ORFs correspond to 10% of ORFs with highest CUG usage. However, little can be said about the function of these genes because most of them are annotated as hypothetical proteins (17) or are not annotated in *C. albicans* genome assembly 19 (16). Nevertheless, orf19.3774 encodes an ubiquitin-like protein, which contains one serine residue coded by a CUG codon (SCU_{CUG} = 1) and also orf19.5761, which contains 38 serines coded by CUG codons only. This ORF is annotated as a hypothetical protein.

Since ORF size, GC content and rare codon bias did not influence CUG usage, one wondered whether particular features of CUG usage could be uncovered by analysing its distribution in the *C. albicans* genome. For this, ORFs were grouped in functional categories and CUG distribution in these ORFs was studied using SCU_{CUG} values and the ANOVA statistical test. In order to ensure that the ANOVA analysis was reliable the data sets were pre-tested for normality and homogeneity of variances, using the *Kolmogorov-Smirnov*'s and *Levene*'s tests, respectively. The data sets did not pass the Kolmogorov-Smirnov test for normality, indicating that the SCU_{CUG} did not follow the normal distribution. However, the ANOVA analysis would still be reliable if the data passed the *Levene*'s test for the homogeneity of variances (Brownie and Boos, 1994). Whenever the ANOVA indicated that there were differences between the groups the *post hoc Scheffe*'s test was carried out to identify the outlier group. Finally, when both the normality and the Levene's tests failed the mean and the standard deviation of codon usage were plotted. All the following statistical tests were carried out using *Statistica 7.0* from StatSoft, Inc, according to the software instructions.

4.2.3.1. The CUG codon distribution in individual chromosomes

In the annotated genome of *C. albicans* one of the analysed features is the physical mapping of each ORF (Braun *et al.*, 2005). There are 8 chromosomes, namely Chr.1, Chr.2, Chr.3, Chr.4, Chr.5, Chr.6, Chr.7 and Chr.R (Table 4. 7) and the gene content of each chromosome varies between 422 (Chr. 7) to 1309 (Chr. 1), and from all the ORFs, only 373 were not allocated to any chromosome. The serine UCA codon was most frequently used in all chromosomes and CUG and AGC were the least used. CUG usage was rather similar between chromosomes (Figure 4. 11) and its distribution pattern was also similar to that of other least used serine codons, namely AGC, UCC and UCG.

Table 4. 7 – ORF distribution over *C. albicans* chromosomes

	Chr. 1	Chr. 2	Chr. 3	Chr. 4	Chr. 5	Chr. 6	Chr. 7	Chr. R	Non allocated
No. Genes	*1309*	*1011*	*748*	*663*	*519*	*423*	*422*	*944*	*373*

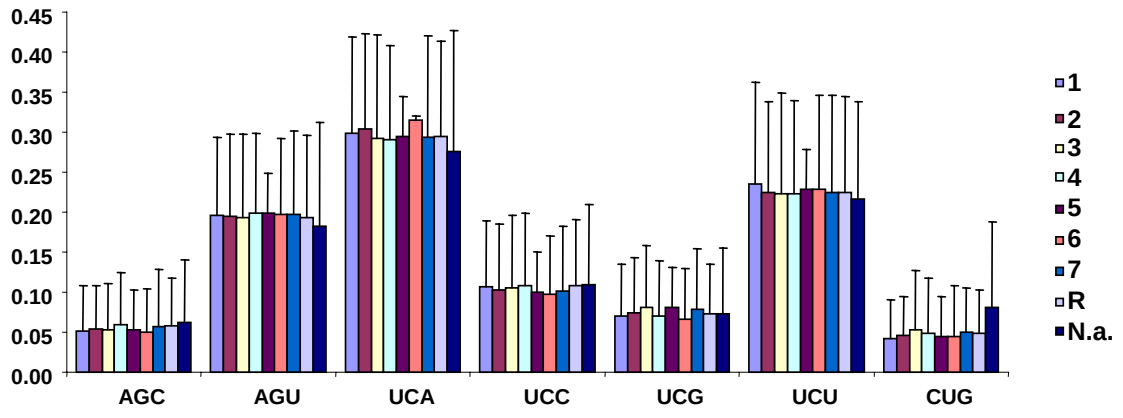


Figure 4. 11 – CUG usage in individual *C. albicans* chromosomes.
The *SCU* values of each serine codon of the ORFs belonging to the 8 chromosomes are equal to its genome average. Only the SCU_{CUG} of non-annotated ORFs (N.a.) showed a slightly different SCU_{CUG} values.

4.2.3.2. The CUG codon distribution in different classes of enzymes

The above analysis was then extended to functional categories, namely *C. albicans* enzymes. There are 1503 ORFs annotated as encoding enzymes (Table 4. 8), which are grouped according the chemical nature of the reaction that they catalyse. The enzyme classification system (EC) groups the enzymes in six major classes, namely: Oxidoreductases (EC 1); Transferases (EC 2); Hydrolases (EC 3); Lyases (EC 4); Isomerases (EC 5); and Ligases (EC 6).

Table 4. 8 – ORF distribution for the six enzyme classes, and the respective SCU_{CUG} average

	EC 1	*EC 2*	*EC 3*	*EC 4*	*EC 5*	*EC 6*
No. Genes	*345*	*473*	*468*	*80*	*60*	*108*
***Average* SCU_{CUG}**	*0.0286*	*0.0389*	*0.0394*	*0.0300*	*0.0251*	*0.0332*

This data set passed the Levene's test for the homogeneity of variances, allowing one to perform an ANOVA to test the hypothesis that CUG usage is not biased in the different classes of enzymes. Indeed, the tested hypothesis failed with $p<0.05$, meaning that the probability of having at least one group whose CUG usage is different from the others is higher than 95%. In order to identify the groups that have biased CUG usage the *Scheffe*'s test was carried out (Table 4. 9).

Table 4. 9 – The *p* values of *Scheffe*'s test for the SCU_{CUG} distribution in the 6 enzyme classes.

	EC1	EC2	EC3	EC4	EC5
EC2	*0.034098*				
EC3	*0.020756*	0.999981			
EC4	0.999907	0.684268	0.623579		
EC5	0.996823	0.341934	0.295967	0.993249	
EC6	0.961812	0.905307	0.864445	0.998143	0.921859

The transferases and hydrolases (EC 2 and EC 3) were the only enzyme classes that showed significant CUG usage bias. This bias was more significant when compared with the Oxireductases (EC 1). Interestingly, CUG usage in enzymes genes was lower than its

average usage in the whole genome (0.048), but the same was also observed for the ACG codon, which is another rare serine codon (Figure 4. 12). This suggests that CUG is not repressed due to its ambiguous decoding, but most likely due to the abundance of its cognate tRNA, as is the case for the other rare codons.

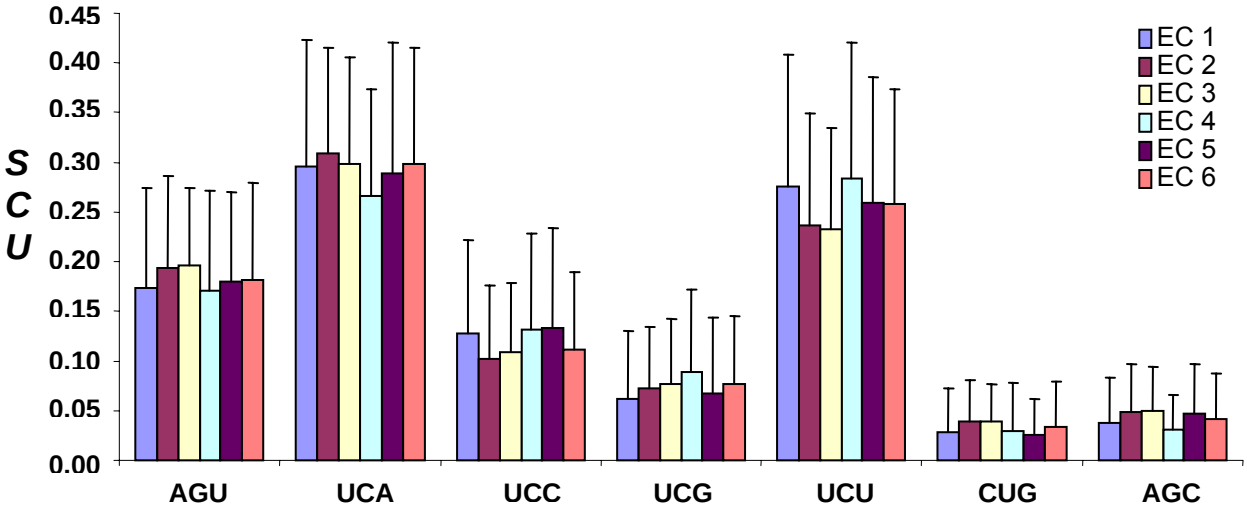
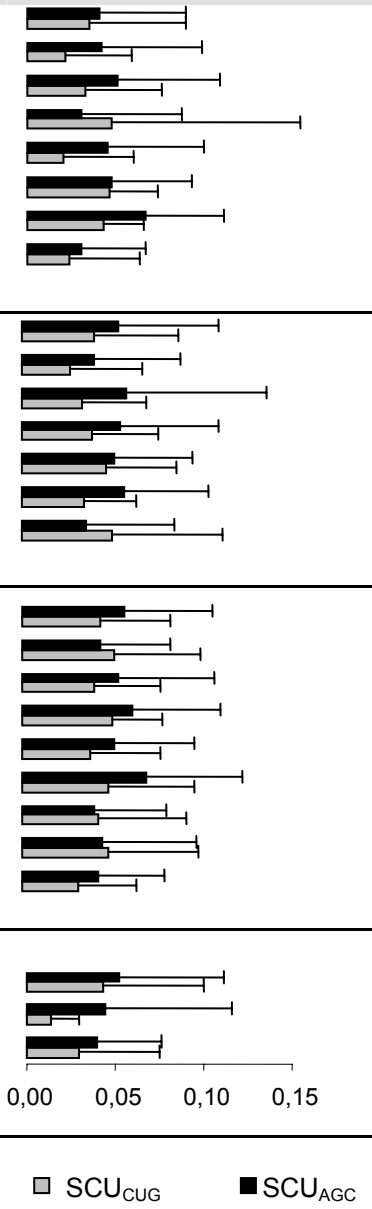


Figure 4. 12 –The *SCU* distribution of serine codons in various classes of enzymes.

For a deeper analysis of the CUG codon distribution, each class of enzymes was further sub-grouped, according to the specific reaction catalysed. Enzymes are organized in classes and sub-classes, but this analysis did not use such strict division criteria. Rather a more general criterion was used, which considered groups with more than 5 elements only. The remaining enzymes that formed groups with fewer enzymes were put together in a major group, named *"other"* (Table 4. 10). The EC4 and EC5 classes contained less than 100 elements and their sub-groups had fewer than 10 elements and, for these reasons, they were not subjected to this analysis.

Table 4. 10 – CUG and AGC codons *SCU* in the enzymes sub-classes.

		No. Genes	Means SCU_{CUG}	Means SCU_{AGC}
EC1	Other	113	0,0323	0,0378
	Superoxide Dismutase	7	0,0204	0,0390
	Redutase	74	0,0311	0,0473
	Peroxidase	7	0,0443	0,0286
	Oxidase	25	0,0189	0,0423
	Monooxygenase	10	0,0430	0,0444
	Dioxygenase	8	0,0399	0,0621
	Dehydrogenase	99	0,0220	0,0283
EC2	Other	206	0,0377	0,0504
	Polymerase	37	0,0254	0,0378
	Methyltransferase	24	0,0318	0,0550
	Mannosyltransferase	13	0,0367	0,0515
	Kinase	148	0,0444	0,0481
	Aminotransferase	15	0,0329	0,0540
	Acetyltransferase	30	0,0473	0,0337
EC3	Other	202	0,0406	0,0534
	Phosphatase	25	0,0479	0,0411
	Ribonuclease	68	0,0370	0,0505
	Protease	15	0,0472	0,0570
	Phospholipase	64	0,0357	0,0475
	Lipase	8	0,0452	0,0643
	Deaminase	9	0,0395	0,0380
	Deacetylase	7	0,0443	0,0413
	ATPase	52	0,0294	0,0393
EC6	Other	41		
	Ubiquitin Ligase	25	0,0427	0,0523
	Acetyl-CoA	7	0,0138	0,0447
	Aminoacyl-tRNA Synthetase	35	0,0296	0,0393



The data failed to meet the required homogeneity of variance and it was not possible to reach meaningful conclusions from this analysis, indicating that CUG usage in each of the sub-groups is very similar. However, in some groups, CUG usage was below its usage in the overall genome, namely in Superoxide Dismutases (EC1), Dehydrogenases (EC1), Polymerases (EC2), and ATPases (EC3). But, in these enzymes AGC usage was also below the average, suggesting that such codon repression is related to effects of rare

codons rather than CUG ambiguous decoding. Interestingly, in the Acetyl-coenzyme A synthetases (EC6) the usage of the CUG codon is very low – it is one third of the whole genome, while the usage of the AGC is not, suggesting that, at least in this sub-group, CUG usage may be repressed.

Deacetylases (EC3), Phosphatases (EC3), Acetyltransferases (EC2) and Peroxidases (EC1) showed SCU_{CUG} values in the same range of the genome's SCU_{CUG}, but SCU_{AGC} values were lower, suggesting that these genes may repress rare codons, but not CUGs.

Finally, CUG usage was also analysed in the leucyl- and seryl-aminoacyl tRNA synthetase genes (*CaCDC30* and *CaSES1*, respectively), which are directly involved in the genetic code change as they both charge the $tRNA_{CAG}^{Ser}$; and in ubiquitin ligases and in proteases, as they are involved on the recognition and degradation of aberrant proteins. However, the behaviour of the CUG codon was similar to that of the overall genome, hence indicating that it does not play a particular important role in these enzymes.

4.2.3.3. The CUG distribution in protein domains

PFAM is a comprehensive collection of protein domains and families containing 7973 protein families (Finn *et al.*, 2006). It was developed and is hosted by the Sanger Institute (http://www.sanger.ac.uk/Software/Pfam/). The availability of this information allowed a detailed characterization of the distribution of CUG codons in the gene parts that encode those protein domains. The ORFs present in the assembly 19 of *C. albicans'* genome contain 2919 known protein motifs (corresponding to 45% of all the ORFs), corresponding to 962 different motifs.

For the analysis of the CUG codon distribution on protein domains, all the annotated ORFs were grouped according to their predominant protein motif and only those groups with more than 9 elements were considered. Therefore, 956 ORFs distributed over 40 PFAM domains were analysed (Table 4. 11). Again, the data failed to meet the required

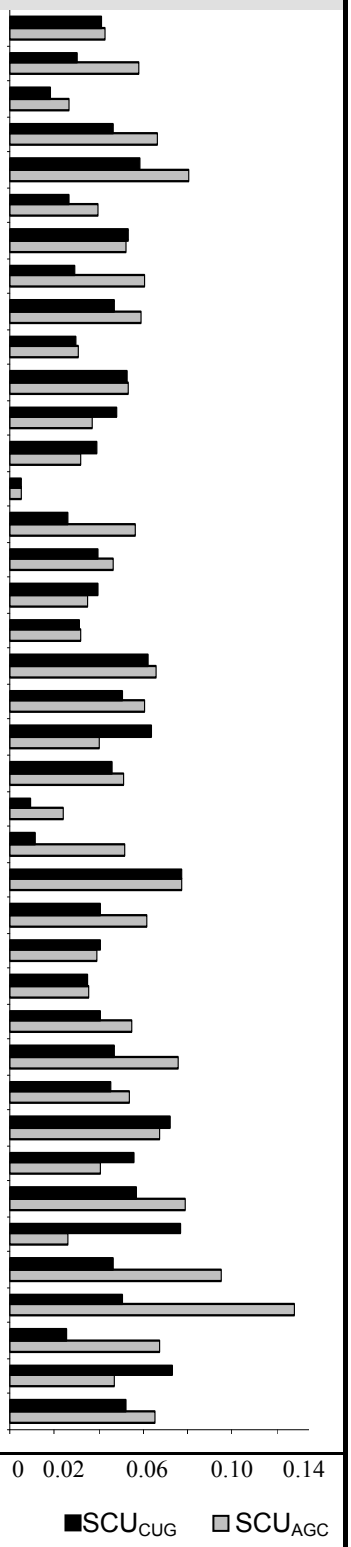
assumptions for the ANOVA analysis and the results below are merely indicative as no significant bias could be detected. Interestingly, in three out of the four most abundant domains, namely on PF00069 (Protein kinase), PF00172 (Fungal Zn(2)-Cys(6) cluster) and PF00083 (Sugar transporter), CUG codon usage was equal to AGC usage, thus indicating that ambiguous CUG decoding does not affect these proteins, as it is not repressed.

Another abundant domain is the leucine rich repeats (PF00560), which are short sequence motifs, present in a number of proteins with diverse functions and cellular locations. It is rather interesting that in these repeats CUG usage is twice as high as its usage on the whole genome, and is higher than AGC usage. One can not exclude that some of these ORFs are misannotated, as the CUG codon in the PFAM motif-search engine has its standard meaning as leucine. However, this bias is interesting because these repeats are usually involved in protein-protein interactions.

Comparison of CUG and AGC usage identified 3 groups: *i)* ORFs whose AGC and CUG usages are similar to their genome's usage; *ii)* ORFs whose CUG usage is than AGC usage; and *iii)* ORFs whose CUG usage is higher than AGC usage. The first group includes the Histone Core domain (PF 00125), which has the lowest usage and shows a strong bias against rare codons. Regarding to the second group, where the CUG codon is repressed in comparison to the AGC codon, it is composed by the following domains: Dehalogenase-like hydrolase (PF00702), PHD-finger (PF00628), F-box domain (PF00646), Aldo/keto reductase family (PF00248), Proteasome A-type and B- type (PF00227), Cyclin, N-terminal domain (PF00134) and ABC transporter (PF00005). Finally, in the PX domain (PF00787), the Acetyltransferase (GNAT) family (PF00583) and the Ubiquitin-conjugating enzymes (PF00179), CUG usage is positively discriminated.

Table 4. 11 CUG and AGC codons *SCU* in protein domains

			Means	
		No. Genes	SCU_{CUG}	SCU_{AGC}
PF00004	ATPase family	27	0.0414	0.0430
PF00005	ABC transporter	21	0.0301	0.0579
PF00009	Elongation factor GTP binding	14	0.0185	0.0268
PF00018	SH3 (Src homology 3)	16	0.0463	0.0663
PF00023	Ankyrin repeat	10	0.0586	0.0807
PF00026	Eukaryotic aspartyl protease	15	0.0267	0.0397
PF00069	Protein kinase	90	0.0534	0.0523
PF00071	Ras family	21	0.0292	0.0604
PF00076	RNA recognition motif.	41	0.0469	0.0589
PF00083	Sugar transporter	57	0.0296	0.0309
PF00096	Zinc finger, C2H2 type	39	0.0530	0.0533
PF00106	short chain dehydrogenase	24	0.0481	0.0370
PF00107	Zinc-binding dehydrogenase	24	0.0391	0.0317
PF00125	Core histone H2A/H2B/H3/H4	11	0.0051	0.0051
PF00134	Cyclin, N-terminal domain	12	0.026	0.0563
PF00149	Calcineurin-like phosphoesterase	17	0.0400	0.0465
PF00153	Mitochondrial carrier protein:	31	0.0398	0.0348
PF00171	Aldehyde dehydrogenase family	11	0.0314	0.0318
PF00172	Fungal Zn(2)-Cys(6) cluster	68	0.0623	0.0660
PF00176	PSNF2 family N-terminal domain	17	0.0506	0.0605
PF00179	Ubiquitin-conjugating enzyme	14	0.0638	0.0400
PF00226	DnaJ domain	24	0.0460	0.0514
PF00227	Proteasome A-type and B- type	13	0.0096	0.0243
PF00248	Aldo/keto reductase family	16	0.0112	0.0517
PF00249	Myb-like DNA-binding domain	9	0.0773	0.0771
PF00270	DEAD/DEAH box helicase	33	0.0406	0.0615
PF00271	Helicase C- terminal domain	10	0.0409	0.0392
PF00324	Amino acid permease	32	0.0349	0.0356
PF00400	WD domain, G-beta repeat	78	0.0408	0.0550
PF00443	Ubiquitin hydrolase family	13	0.0473	0.0757
PF00515	TPR Domain	15	0.0455	0.0540
PF00560	Leucine Rich Repeat	35	0.0722	0.0671
PF00561	alpha/beta hydrolase fold	16	0.0560	0.0406
PF00566	TBC domain	10	0.0571	0.0791
PF00583	Acetyltransferase (GNAT) family	13	0.0767	0.0262
PF00628	PHD-finger	10	0.0464	0.0951
PF00646	F-box domain	12	0.0505	0.1280
PF00702	Dehalogenase-like hydrolase	13	0.0254	0.0673
PF00787	PX domain	10	0.0730	0.0468
PF01794	Ferric reductase like transmembrane component	14	0.0522	0.0655



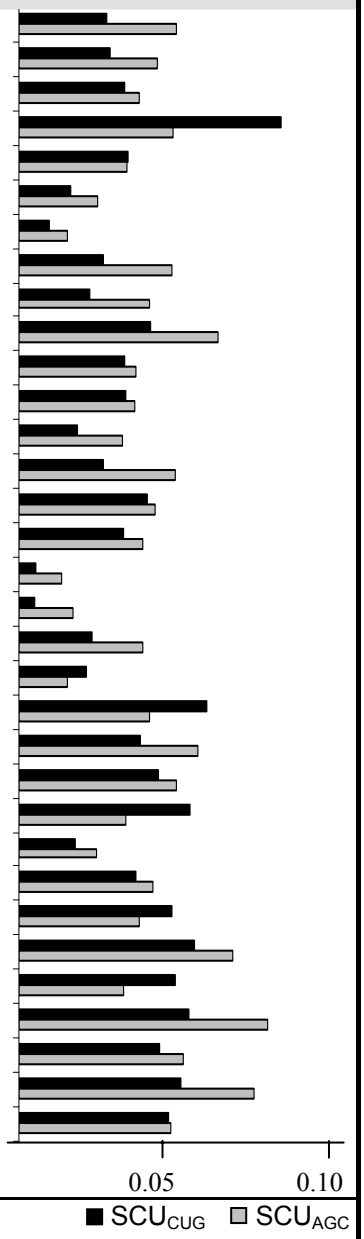
4.2.3.4. The CUG codon usage and the gene ontology

The availability of several sequenced genomes and the discovery that most genes of core biological functions are shared by all eukaryotes prompted the uniformization of the cellular terminology. The Gene Ontology (GO) terms have then arisen with the objective of standardizing gene terminology. GO terms are split into three related ontologies – the **molecular function** of gene products; their role in multi-step **biological processes**; and their localization to **cellular components** (Ashburner *et al.*, 2000) (http://www.geneontology.org). The genome of *C. albicans* was also annotated using GO terms, thus allowing one to study CUGs distribution in different ontologies. To carry out such a study, *C. albicans* ORFs were grouped according to their gene ontology categories, and only those groups with 10 or more elements were analysed.

Interestingly, genes that belong to the cellular location ontology (Table 4. 12), in particular those bound to the membrane surface, but not integrated into the hydrophobic region, differ the most in terms of CUG usage. Indeed, CUG usage in this category is 2 fold higher than genome average of CUG usage. This may be of biological relevance because these proteins are directly exposed to the immune system and are used as antigens. In other words, leucine/serine ambiguity at CUGs in these genes may help *C. albicans* to escape the immune system. Similar CUG usage bias was found in nuclear membrane genes and in genes of the SAGA-complex, which is a large multiprotein complex that possesses histone acetyltransferase activity and is involved in regulation of transcription (ex: Gcn5p; (Grant *et al.*, 1998)).

Table 4. 12 – CUG and AGC codons *SCUs* in ORFs grouped according to their cellular localization

	No. Genes	Means SCU_{CUG}	Means SCU_{AGC}
C:cell wall	47	0.0315	0.0566
C:membrane	61	0.0326	0.0498
C:plasma membrane	138	0.0378	0.0433
C:extrinsic membrane protein	12	0.0940	0.0552
C:integral membrane protein	54	0.0390	0.0387
C:hydrogen-transporting ATPase	23	0.0186	0.0282
C:respiratory chain complex	16	0.0109	0.0174
C:mitochondrial ribosome	65	0.0304	0.0550
C:peroxisome	40	0.0254	0.0469
C:endosome	25	0.0473	0.0715
C:lipid particle	15	0.0379	0.0418
C:vacuole (sensu Fungi)	47	0.0383	0.0415
C:proteasome	43	0.0209	0.0370
C:COP vesicle coat	26	0.0303	0.0562
C:Golgi apparatus	76	0.0459	0.0488
C:endoplasmic reticulum	152	0.0377	0.0442
C:cytosolic ribosome (sensu Eukarya)	94	0.0061	0.0153
C:ribosome	21	0.0057	0.0194
C:eukaryotic translation initiation complex	10	0.0262	0.0444
C:RNase complex	18	0.0243	0.0174
C:nuclear membrane	11	0.0673	0.0467
C:nuclear pore	29	0.0435	0.0640
C:transcription factor	50	0.0501	0.0565
C:SAGA complex	17	0.0613	0.0384
C:DNA-directed RNA polymerase	35	0.0203	0.0280
C:histone complex	19	0.0421	0.0482
C:anaphase-promoting complex	11	0.0547	0.0430
C:DNA replication factor complex	10	0.0629	0.0768
C:mRNA processing	16	0.0559	0.0374
C:CCR4-NOT complex	11	0.0609	0.0890
C:mediator complex	15	0.0505	0.0588
C:spindle pole	15	0.0579	0.0844
C:cellular_component unknown	511	0.0534	0.0544



Conversely, CUG usage is repressed in ribosomal protein genes, but these genes also use AGC codons (rare codons) less frequently than expected indicating that such repression is probably related to rare codon bias. This is in agreement with the high expression level of ribosomal proteins and their biased codon usage. Other genes coding

for abundant proteins also show reduced CUG usage, namely respiratory chain complex and RNA polymerase genes.

Interestingly, several genes showed negative CUG usage bias and positive AGC usage bias, suggesting that CUG usage may be under negative selection. Among these are genes that code for proteins of the spindle pole, which are involved in the organization of the cytoskeleton, and genes of the CCR4-NOT complex, which is involved in several different cellular pathways, namely transcription regulation, mRNA degradation and post-transcriptional modifications (Panasenko *et al.*, 2006).

In the biological processes gene ontology (Table 4. 13), a strong repression of CUG usage was observed in several classes, specially in genes coding for proteins involved in ATP synthesis coupled to proton transport, carbohydrate metabolism, heme biosynthesis, ubiquitin-dependent protein and in fatty acid catabolism. It is also repressed in genes of the NAD(+) biosynthesis, aging processes, processing of 20S pre-rRNA, drug susceptibility/resistance, endocytosis, G1/S transition of mitotic cell cycle, DNA replication and amino acid metabolism. Conversely, CUG usage is positively biased in genes of the cyclin catabolism, chromatin modification, Golgi to endosome transport, cell growth and/or maintenance, budding, mRNA processing and the DNA repair.

In general, CUG usage is repressed in genes related to translational processes and metabolic pathways. Such repression may be explained by the need for accurate synthesis of proteins that play critical roles in cell functioning. On the other hand, genes with the highest CUG usage code for proteins involved in Golgi to endosome transport and on DNA repair.

Table 4. 13 – CUG and AGC codons *SCUs* in ORFs grouped according to their cellular process

		Means	
	No. Genes	SCU_{CUG}	SCU_{AGC}
P:ATP synthesis coupled proton transport	10	0.0063	0.0248
P:carbohydrate metabolism	10	0.0154	0.0602
P:cyclin catabolism	10	0.0543	0.0357
P:protein complex assembly	10	0.0362	0.0424
P:regulation of redox homeostasis	10	0.0681	0.0748
P:chromatin modification	11	0.0546	0.0396
P:Golgi to endosome transport	11	0.0901	0.0748
P:heme biosynthesis	11	0.0264	0.0412
P:NAD (+) biosynthesis	11	0.0357	0.0594
P:signal transduction	11	0.0512	0.0482
P:glycogen metabolism	12	0.0249	0.0296
P:nitrogen metabolism	12	0.0391	0.0354
P:aging	13	0.0320	0.0456
P:chromatin assembly/disassembly	13	0.0161	0.0199
P:axial budding	14	0.0465	0.0571
P:deubiquitination	15	0.0455	0.0546
P:Golgi to vacuole transport	15	0.0637	0.0731
P:RNA elongation from Pol II promoter	15	0.0420	0.0477
P:cell growth and/or maintenance	16	0.0505	0.0390
P:cytokinesis	16	0.0563	0.0661
P:Golgi to plasma membrane transport	16	0.0571	0.0625
P:G1 phase of mitotic cell cycle	17	0.0519	0.0759
P:processing of 20S pre-rRNA	17	0.0306	0.0602
P:autophagy	18	0.0503	0.0638
P:drug susceptibility/resistance	18	0.0312	0.0519
P:endocytosis	18	0.0255	0.0467
P:histone	18	0.0367	0.0628
P:amino acid transport	19	0.0307	0.0346
P:budding	19	0.0529	0.0333
P:chromatin modeling	19	0.0498	0.0630
P:vesicle-mediated transport	19	0.0486	0.0691
P:cell cycle	21	0.0456	0.0577
P:chromosome segregation	21	0.0492	0.0651
P:chromatin silencing	22	0.0623	0.0663
P:G1/S transition of mitotic cell cycle	22	0.0387	0.0674
P:protein folding	22	0.0356	0.0276
P:protein amino acid phosphorylation	23	0.0503	0.0590
P:ergosterol biosynthesis	24	0.0193	0.0276

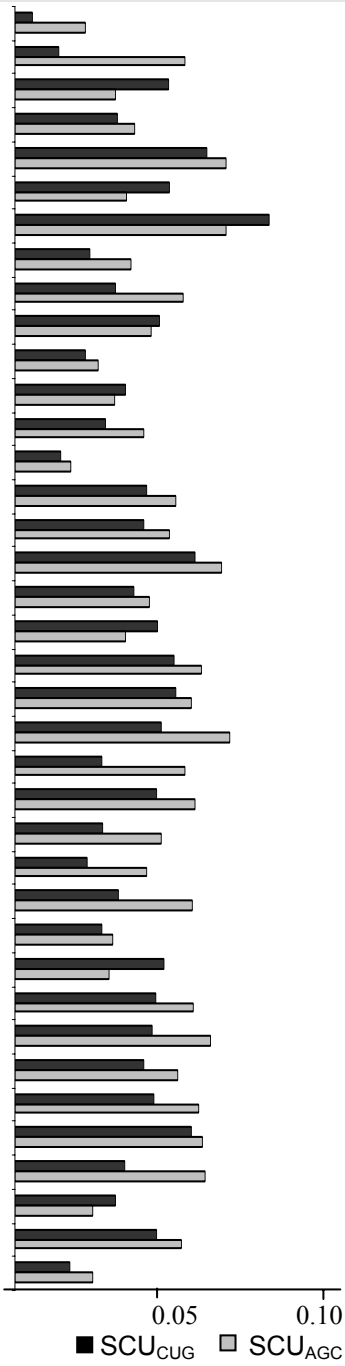
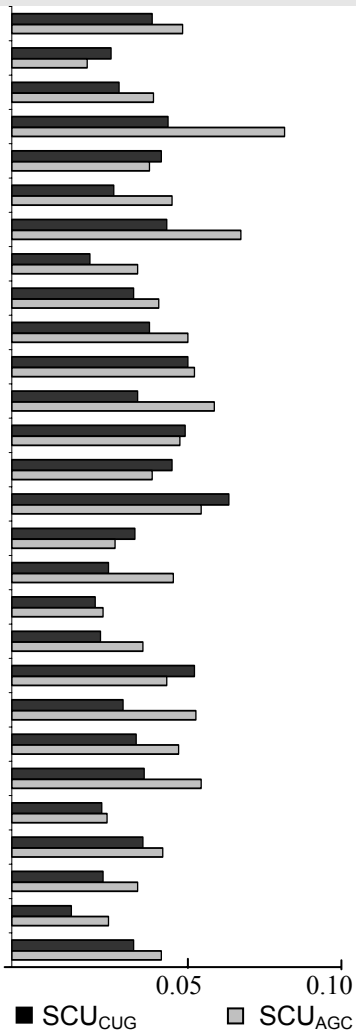


Table 4. 13 – (cont.)

	No. Genes	Means SCU_{CUG}	Means SCU_{AGC}
P:NLS-bearing substrate-nucleus import	25	0.0453	0.0551
P:ribosome biogenesis	26	0.0320	0.0243
P:rRNA processing	26	0.0347	0.0458
P:meiosis	27	0.0504	0.0882
P:negative regulation of transcription	27	0.0483	0.0445
P:actin filament organization	28	0.0330	0.0516
P:regulation of transcription	32	0.0501	0.0741
P:ubiquitin-dependent protein catabolism	32	0.0251	0.0405
P:homeostasis	34	0.0394	0.0474
P:tRNA processing	34	0.0445	0.0569
P:mitosis	36	0.0568	0.0590
P:DNA replication	37	0.0404	0.0652
P:protein targeting	37	0.0559	0.0541
P:mRNA processing	41	0.0515	0.0455
P:DNA repair	42	0.0701	0.0610
P:glycosylation	43	0.0398	0.0335
P:aerobic respiration	44	0.0311	0.0523
P:35S primary transcript processing	46	0.0268	0.0294
P:fatty acid metabolism	48	0.0287	0.0424
P:mRNA splicing	54	0.0590	0.0502
P:cell wall organization and biogenesis	59	0.0358	0.0592
P:ER to Golgi transport	62	0.0400	0.0539
P:response to stress	62	0.0427	0.0613
P:translation	78	0.0292	0.0308
P:transcription	100	0.0424	0.0486
P:amino acid metabolism	102	0.0293	0.0408
P:protein biosynthesis	158	0.0191	0.0313
P:biological_process unknown	232	0.0392	0.0482



4.2.4. The evolution of the CUG codon in *C. albicans'* genome

The *C. albicans* genome is highly unstable. This has been considered as the only means of generating genetic variation for this organism because it does not have a sexual cycle. It can only mate through a parasexual mechanism but there is no nuclear fusion and consequently meiotic recombination is impaired (Rustchenko et al., 1994; Rustchenko et al., 1997). *C. albicans* is also a diploid yeast and its genome is composed by 16 chromosomes that frequently rearrange. Various strains loose or gain chromosomes generating stable aneuploids (as strains WO-1 and SGY-243, with 19 and 21

chromosomes, respectively (reviewed in Rustchenko, 2007) Interestingly, some chromosomes are more stable than others. For example, the chromosome R, which contains tandem repeats of genes encoding the rRNA (Barton *et al.*, 1995), is highly unstable due to intragenic rearrangements at the repeated sequences. Genome instability is also linked to the *C. albicans* parasexual life-cycle, because, after mating, the tetraploid cells do not undergo meiosis and use a yet poorly understood mechanism of chromosome loss to reduce genome size (Hull et al., 2000; Magee and Magee, 2004; Magee and Magee, 2000; Bennett and Johnson, 2003; Soll, 2004; Bennett and Johnson, 2005).

The *C. albicans* genome is also highly heterozygotic (Forche *et al.*, 2004). Indeed, homologous chromosomes are substantially divergent, and many of its genes are present as two distinct alleles (Braun *et al.*, 2005). For this reason, the *C. albicans* genome assembly 20 includes an indication of the different alleles for most of the open reading frames. This allows one to draw a tentative model for CUG codon evolution in some of the genes for which two alleles are known.

In order to study CUG evolution using alleles information, we used the approach described above for CUG analysis in different ontology classes. Again, the rare AGC codon was used as a control codon, but this time the comparison took into consideration codon usage differences between the two alleles of each gene (Table 4. 14). A preliminary genome analysis showed that 76.6% of genes had 2 heterozygotic alleles, and, for this reason, CUG and AGC usage was analysed and the difference between the two alleles was calculated, d(CUG) and d(AGC), respectively. Most genes did not show differences in CUG and AGC usage between the 2 alleles. However, this analysis showed that AGC usage varies more frequently between alleles than the CUG codon (Table 4. 14), suggesting a stabilization of the CUG content on both allelic forms.

Table 4. 14 – Variation of AGC and CUG codons between alleles

	ORFs without codon difference	ORFs with codon difference	ORFs without alleles
CUG	4711	220	1507
AGC	4588	343	1507

This data prompted the question of whether the CUG codon is under negative selection in the *C. albicans* genome. In order to answer this, a set of 185 genes whose AGC and CUG codon usages are altered in the two alleles and have homologues in *S. cerevisiae*, was selected for phylogenetic analysis. For this, both alleles were aligned with the *S. cerevisiae* homologues. This allowed one to determine which of the 2 alleles had higher similarity to the *S. cerevisiae* homologue. This analysis was carried out using the software PALM (*v* 3.14) (Yang, 1997), and each allele was scored for *i*) all nucleotide substitutions (*d(S)*); and *ii)* the neutral nucleotide substitutions (*d(N)*). Indeed, these values permitted determining which allele diverges the most from *S. cerevisiae* both in terms of neutral and overall substitutions, and thus is evolving faster.

Not surprisingly, most of the allelic forms that had a higher *d(S)* score also had a higher *d(N)* score, which reinforces the robustness of the approach, and indicates that those genes with more neutral substitutions also have more non-neutral substitutions. Interestingly, from the 185 ORFs analysed only in 4 of them the allele with higher number of substitutions was not the allele with higher number of non-neutral substitutions (Table 4. 15). Still, within this group of 4 genes only *PAC2,* which is a non-essential gene involved in the tubulin heterodimer formation (Fleming *et al.*, 2000), showed difference between CUG and AGC usage.

Table 4. 15 –ORFs with higher *d(S)* score but lower *d(N)* score.

Ca ORF with more neutral substitutions (*d(N)*)	*Ca* ORF more substitutions (*d(S)*)	*S. cerevisiae* homologue	Gene name
orf19.4335	orf19.11811	S000003492	TNA1
orf19.3954	orf19.11436	S000003402	PSD2
orf19.8292	orf19.675	S000000265	YEL077C
orf19.2921	orf19.10438	S000000809	PAC2

In order to exploit the behaviour of CUG codon usage in the remaining 181 genes, and thus infer the direction of the allelic evolution in terms of CUG usage, the difference of both SCU_{CUG} and SCU_{AGC} between the ancestral and the most recent allelic form was determined, by applying Equation 4. 4 and Equation 4. 5, respectively.

$$d(CUG) = (SCU_{CUG})_{recent} - (SCU_{CUG})_{ancestral}$$

Equation 4. 4

$$d(AGC) = (SCU_{AGC})_{recent} - (SCU_{AGC})_{ancestral}$$

Equation 4. 5

If this difference is higher than 0, there is a preference for the usage of the analysed codon in the alleles that are evolving faster, but if it is lower than 0 there is repression of its usage. Then, the values of *d(CUG)* and *d(AGC)* were submitted to a clustering analysis using MeV 4.0 from the TM4 Software package (Saeed *et al.*, 2003). A complete linkage cluster with a bootstrap of 100 and a similarity matrix using the Uncentered Pearson correlation was generated. This analysis showed 8 clusters among the 181 genes, with a distance threshold of -0.59 (Figure 4. 13).

Those clusters allowed one to analyse the behaviour of CUG and AGC codons usage in the selected genes (Figure 4. 13, Annexe D). The most interesting clusters are 1 and 6, as CUG usage behaviour is opposite to that observed for AGC. Indeed, in group-1, CUG usage in the fast evolving ORF is increased and AGC usage is decreased, suggesting an evolutionary gain associated to CUG ambiguity. On the other hand, in group-6, there is repression of CUG usage, when compared with AGC usage, thus indicating that ambiguous CUG decoding of such ORFs might be detrimental to the organism.

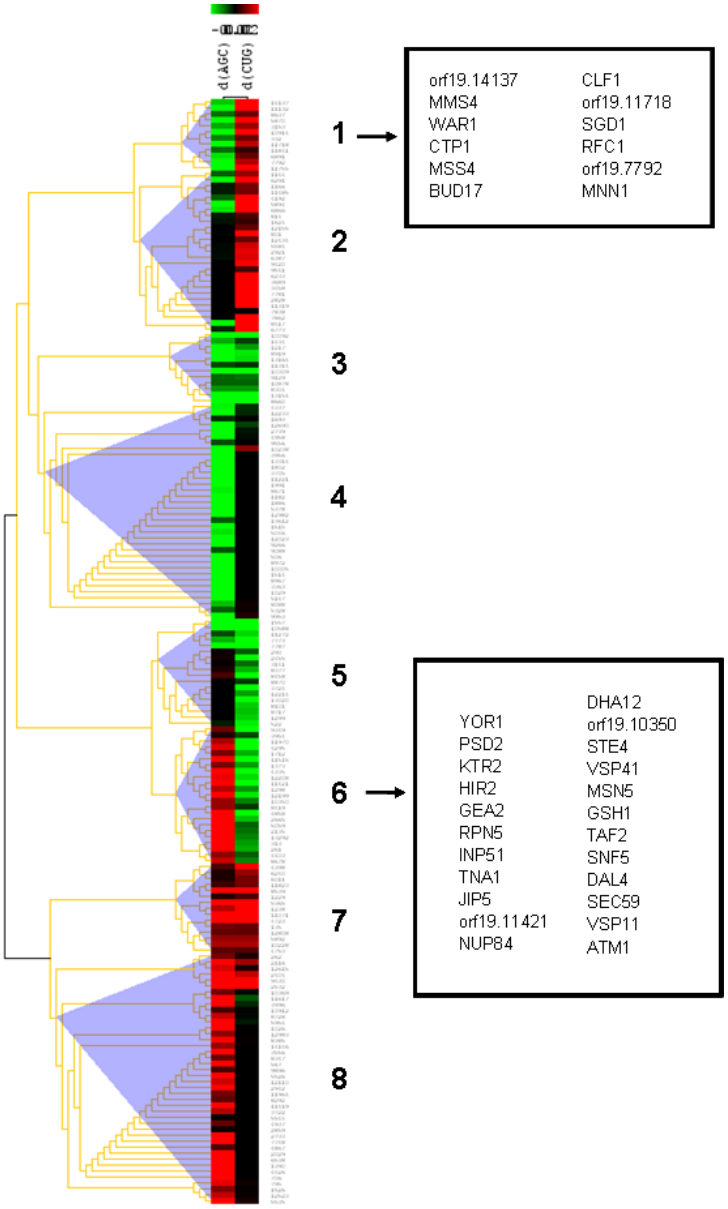


Figure 4. 13 – Cluster analysis of the CUG and AGC codons usage in the different alleles.
The cluster analysis has unveiled the existence of 8 clusters. In clusters 1 and 6 the behaviour of CUG usage is opposite to that of AGC usage. The ORFs belonging to those clusters are in the text boxes. Red indicates an increase in codon usage in fast evolving alleles, green indicates a decrease in codon usage in the fast evolving allele, in black are those alleles without difference in codon usage.

In group-1 genes, *MNN1* is rather interesting because it encodes an alpha-1,3-mannosyltransferase, which is an integral membrane glycoprotein of the Golgi complex, required for addition of alpha1,3-mannose linkages to N-linked and O-linked oligosaccharides (Yip *et al.*, 1994); *BUD17*, which is involved in the maintenance of the bipolar budding pattern (Ni and Snyder, 2001). Also, the *MSS4* gene, which encodes a

phosphatidylinositol-4-phosphate 5-kinase, is associated with hyphal growth and is repressed by macrophages (Hairfield *et al.*, 2002).

In group-6 genes, CUG ambiguity may have been detrimental. Indeed, these genes are mainly involved in transcription – as *TAF2, HIR2, SNF5*; or are ABC transporters (*ATM1* and *YOR1*) and permeases (*TNA1, DAL4*). In this group there is also a mannosyltransferase (*KTR2*) and a phosphatidylinositol 4,5-bisphosphate 5-phosphatase (*INP51*).

4.3. Discussion

In this chapter, a comprehensive analysis of the CUG codon usage in the *C. albicans* genome was carried out. The data showed that CUG ambiguity expands the *C. albicans* proteome exponentially. This is of profound biological significance as arrays of proteins are generated from single mRNAs creating a statistical proteome. Indeed, the 6,438 genes in the *C. albicans* genome have the potencial to produce 2.83×10^{11} different proteins. Moreover, CUG codon context biases were not detected indicating that the CUG codon is randomly decoded as either serine or leucine. This implies that *C. albicans* proteins are quasi-species (Freist *et al.*, 1998) and that the probability of finding two identical cells in a population is extremely small. Such exponential increase of the size of the *C. albicans* proteome may ultimately be the main factor contributing to its morphological variation (Miranda, 2007).

The data also showed that the *C. albicans* genome evolved to tolerate CUG ambiguous decoding and that its genome is optimized to cope with it. Indeed, a CUG ambiguity rate of 2.96% in the wild-type white *C. albicans* results in the production of 6.5×10^6 novel proteins, while in engineered *S. cerevisiae* cells 2.31% and 1.40% of CUG ambiguity results in 12.5×10^6 and 7.9×10^6 novel proteins, respectively. However, the hidden malleability of *C. albicans*' proteome is extremely high, as this organism tolerates at least 28.0% of CUG ambiguity, which results in 39.5×10^6 novel proteins.

In order to identify the genes that are more affected by CUG ambiguous decoding in *C. albicans*, the CUG codon usage was investigated in 6 enzyme classes, in its 8 chromosomes, in protein domains and in gene ontologies. CUG codon usage is repressed in enzymes genes, as are all the other rare codons, hence indicating that such bias does not result from CUG ambiguous decoding. Also, CUG usage bias in individual chromosomes or in protein domains were not observed. Conversely, the CUG codon usage was repressed in genes that code for proteins involved in translation, which can be regarded as a safeguard for correct protein synthesis. On the other hand, a CUG usage bias was detected in genes coding for proteins bound to the membrane surface. Considering that *C. albicans* is a pathogen and that these proteins are recognized by the host immune system, such CUG codon usage may be important for pathogenesis, as an increase in the ambiguous CUG decoding results in an alteration of the surface antigens, which would be an elegant strategy to escape the immune system.

Taken together, these data highlight novel features of CUG ambiguity, in particular in proteome expansion and diversity. Recent studies from our laboratory have shown that CUG ambiguity in *C. albicans* generates phenotypic diversity (Gomes et al., 2007). Indeed, the highly ambiguous cell lines, expressing the $tRNA_{CAG}^{Leu}$, displayed highly variable morphologies characterized by formation of aerial hyphae, white-opaque sectoring and hypha that penetrated deeply into agar and produced opaque sectors (Gomes et al., 2007). CUG ambiguity also induces karyotype alterations and remodels gene expression and cell physiology (Miranda, 2007). Moreover, in *S. cerevisiae* CUG ambiguous decoding also resulted in transcriptome and proteome alterations and in a ploidy variation. Further, the partial redefinition of CUG identity in *S. cerevisiae* blocked lateral gene transfer and imposed a immediate genetic barrier to sexual reproduction, by decreasing sporulation efficiency, fertility and mating (Silva et al., 2007). All these studies clearly show that organisms with large proteomes can tolerate very high levels of codon ambiguity, confirming previous synthetic biology studies on the artificial expansion of the genetic code (Chin et al., 2003), and that the *C. albicans* proteome has a statistical nature of high complexity.

5. The role of the Leucyl- and Seryl- tRNA Synthetases in CUG ambiguity

5.1. Introduction

The identity of CUG codons is variable in the genus *Candida* (Section 1.5): *C. glabrata* decodes CUGs as leucine, *C. cylindracea* has totally reassigned them to serine and several other *Candida* species decode them ambiguously (Sugita and Nakase, 1999; Santos et al., 1993; Santos et al., 1996; Suzuki et al., 1997). Such differences in the CUG codon decoding are due to structural differences in the $tRNA_{CAG}$ – the only cognate tRNA for CUG codons in the *Candida* genus. Indeed, the ambiguous CUG decoding in *C. albicans* results from mischarging of the $tRNA_{CAG}^{Ser}$ (Sugita and Nakase, 1999; Santos et al., 1993; Santos et al., 1996; Suzuki et al., 1997). This mischarging is very interesting from a structural perspective, since it is not yet clear how this novel tRNA is recognized by the LeuRS and why this enzyme fails to edit the mischarged leu-$tRNA_{CAG}^{Ser}$.

In *E. coli,* recognition of $tRNA^{Leu}$ by the cognate LeuRS is achieved through interactions with the A_{73} – the discriminator base – and tertiary structural elements, namely the position of the invariant $G_{18}G_{19}$ sequence in the D-arm, the semi-invariant $R_{15}{\cdot}Y_{48}$ tertiary base-pair, the base R_{59} in the TψC loop, and the unpaired nucleotides present at the base of the variable-arm (Asahara *et al.*, 1993). In archea and in most eukaryotes the LeuRSs recognize the long variable-arm of cognate $tRNA^{Leu}$ (Fukunaga and Yokoyama, 2005), whereas in yeast the LeuRS makes direct contact with both the A_{73} discriminator base and the methyl group of m^1G_{37} and with A_{35} in the anticodon-loop. It also makes non-specific contacts with the phosphate backbone of the anticodon-stem (Soma et al., 1996; Soma and Himeno, 1998). Interestingly, like the canonical $tRNA^{Leu}$, *C. albicans* $tRNA_{CAG}^{Ser}$ contains A_{35} and m^1G_{37} in its anticodon-loop, but not the A_{73} discriminator base, which is G_{73} (Sugita and Nakase, 1999; Santos et al., 1993; Santos et al., 1996; Suzuki et al., 1997). Such difference in the discriminator base is important because changing A_{73} to G_{73} in both yeast (Soma *et al.*, 1996) and human $tRNA^{Leu}$ (Breitschopf et al., 1995; Breitschopf and Gross, 1994) changes the tRNAs identity from leucine to serine. In the *Pyrococcus horikoshii* LeuRS-$tRNA^{Leu}$ complex, A_{73} is recognized by the amino acid residue 504 of the editing domain and the interaction is disrupted when A_{73} is replaced by G_{73} (Fukunaga and Yokoyama, 2005). Whether or not the *C. albicans* LeuRS evolved a

novel mechanism for recognizing both G and A at position 73 is yetknown. Regarding the failure of LeuRS to edit mischarged leu-$tRNA_{CAG}^{Ser}$, the LeuRS binds its cognate amino acid (leucine), activates it (as normal) and transfers it to the $tRNA_{CAG}^{Ser}$. In other words, both leucine and $tRNA_{CAG}^{Ser}$ are cognate substrates for the LeuRS and consequently the post-transfer editing mechanism is not activated. This is supported by the high degree of amino acid conservation between LeuRS of *C. albicans* and other yeasts, particularly within the editing domain. Functionally, the *S. cerevisiae CDC60* (LeuRS) gene could be also complemented by its *C. albicans* homologue (O'Sullivan *et al.*, 2001b).

Concerning the recognition of the serine tRNAs by the cognate SerRS, in *E. coli* it is achieved through interactions with the variable-arm, whose length and tertiary structure are crucial for serylation (Himeno et al., 1990; Asahara et al., 1994). In yeast *in vitro* aminoacylations and footprinting experiments revealed that the discriminator base is not crucial and that the variable-arm functions as the major identity element (Dock-Bregeon et al., 1990; Soma et al., 1996; Himeno et al., 1997). Indeed, the role of the discriminator G_{73} varies within the different organisms: *i)* it acts as an identity antideterminant against LeuRS in bacteria (Asahara *et al.*, 1993) and lower eukaryotes (Himeno et al., 1990; Soma et al., 1996); and *ii)* it is an essential identity requirement for human $tRNA^{Ser}$ (Breitschopf et al., 1995; Breitschopf and Gross, 1994; Achsel and Gross, 1993). However, in *E. coli,* apart from the discriminator base, the SerRS recognizes directly acceptor-stem bases from positions 1-72 to 5-68, with the major recognition cluster located between positions 2-71 and 4-69 (Saks and Sampson, 1996; Normanly et al., 1992; Sampson and Saks, 1993). The crystal structure of the SerRS–$tRNA^{Ser}$ complex from *Thermus thermophilus* supported the notion that recognition of $tRNA^{Ser}$ depends on specific contacts with the variable-arm of $tRNA^{Ser}$, which is achieved through the sugar–phosphate backbone interactions by the N-terminal α-helical arm of SerRS (Cusack *et al.*, 1996). Such discrimination mechanism was also reported in yeast (Lenhard *et al.*, 1999), in archeal (Bilokapic *et al.*, 2004) and in human SerRS (Achsel and Gross, 1993), indicating that the recognition of the variable arm of $tRNA^{Ser}$ by SerRS is evolutionarily conserved in the three kingdoms of life. With respect to the *C. albicans* $tRNA_{CAG}^{Ser}$, its extra variable arm and its G_{73} discriminator base are both characteristic of the serine-family of tRNAs.

An interesting feature of the *C. albicans* $tRNA_{CAG}^{Ser}$ is the presence of a unique guanosine in the turn of its anticodon-loop (G-turn) – a conserved position occupied by U_{33} (U-turn), which reduces the leucylation efficiency of the $tRNA_{CAG}^{Ser}$ (Suzuki *et al.*, 1997), therefore, the G_{33} can be regarded as an anti-determinant of leucine identity. It is not yet clear how and when this G_{33} appeared and why it was kept, though it is possible that such change decreased the toxicity of the mutant $tRNA_{CAG}^{Ser}$, by lowering its leucylation in the early stages of CUG identity alteration (Perreau et al., 1999; Santos et al., 1996).

As demonstrated in chapter-3, within the *C. albicans* cytoplasm there are two charged forms of the $tRNA_{CAG}^{Ser}$: the leu-$tRNA_{CAG}^{Ser}$ and the ser-$tRNA_{CAG}^{Ser}$, and both compete for the CUG codon decoding at the ribosome A-site. However, it has also been demonstrated that the leucine incorporation at CUG codons varies under different physiological conditions. Therefore, it is important to unveil the regulatory mechanisms that modulate the levels of CUG ambiguity *in vivo*. In an attempt to understand how such CUG decoding ambiguity can be regulated and shed new light on the evolutionary pathway of CUG reassignment, a study of the enzymes responsible for the ambiguous charging of the $tRNA_{CAG}$, namely the LeuRS and the SerRS, was carried out.

5.2. Results

5.2.1. Quantification of SerRS and LeuRS expression in *C. albicans*

A putative regulatory mechanism for CUG ambiguity is the differential expression of the SerRS and LeuRS. In order to test this hypothesis, the expression of LeuRS and SerRS was monitored by Western-blot in cells grown in conditions for which leucine misincorporation was quantified by mass-spectrometry (Figure 5. 1). A total of three independent experiments were carried out for each condition, and for each experiment Western-blots were done in duplicate.

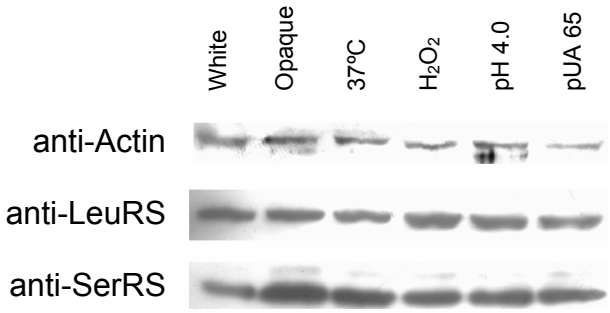


Figure 5. 1– LeuRS and SerRS protein expression under different physiological conditions.
Western-blot against LeuRS and SerRS of whole protein extracts, from cells grown in the different physiological conditions indicated. Actin was used to normalize the data

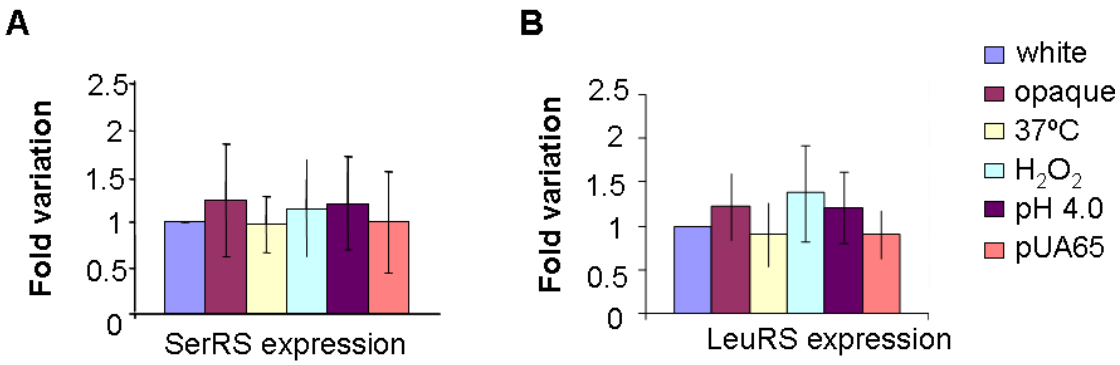


Figure 5. 2 – SerRS and LeuRS expression
Fold variation of protein expression of either **(A)** SerRS or **(B)** LeuRS in the different physiological conditions, in relation to their levels in white cells.

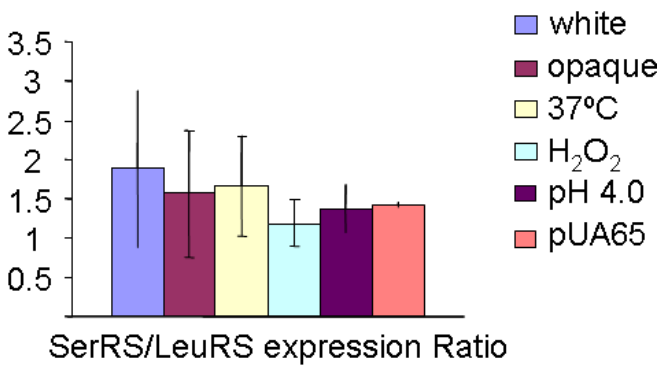


Figure 5. 3– SerRS/LeuRS expression ratio.
The ratio between the SerRS and the LeuRS expression for the different physiological conditions is indicated.

The intensity of each signal was determined using the *Quantity One* software (BioRad) and the variation between each condition was then assessed (Figure 5. 1). Since it was not possible to use the anti-LeuRS and anti-SerRS antibodies in the same Western-blot, actin, whose expression fluctuates very little, was used to normalize the data. Fold-variation of expression for both SerRS and LeuRS were determined relative to their expression levels in white cells (Figure 5. 3).

No significant variation in LeuRS and SerRS expression was detected, suggesting that CUG ambiguity is not regulated by differential expression of the SerRS and LeuRS. Indeed, the SerRS/LeuRS ratio (Figure 5. 3) did not show significant difference between the tested conditions, although SerRS expression is apparently higher than that of the LeuRS.

Interestingly, the LeuRS antibody detected 2 bands while the SerRS antibody detected a single band. Since in *S. cerevisiae* the LeuRS enzyme is cleaved by the *yscB* protease (Larrinoa and Heredia, 1991) it is possible that the two bands detected in the *C. albicans* extracts also resulted from protease cleavage of the full length LeuRS enzyme, raising the possibility that the cleaved LeuRS is not active or is less active than the full length enzyme. However, no significant variation in the ratio between the full length and the cleaved protein (Figure 5. 4), among the tested growth conditions was detected, indicating that LeuRS processing is not involved in regulation of CUG decoding ambiguity.

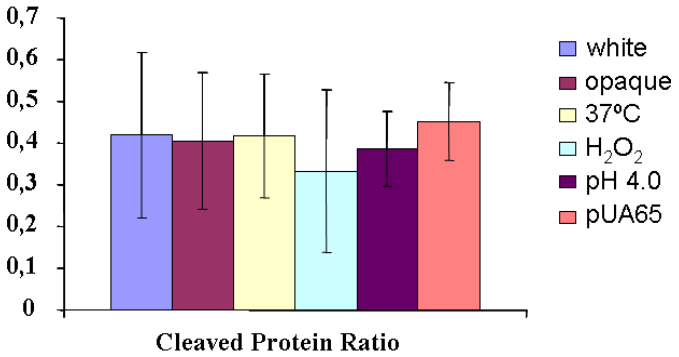


Figure 5. 4 – Ratio between the cleaved and the native LeuRS
The ratios between the intensity of bands corresponding to the full length and the cleaved protein, measured by Western-blot, in different physiological conditions.

5.2.2. The study of SerRS and LeuRS genes

As the published sequences for both synthetase genes deposited in the NCBI genebank database were obtained from *C. albicans* strain *2005* (Annexes E and F), which was not the strain used in the above studies (it was the used *CAI-4* strain); the genes of both synthetases were re-sequenced and the results compared with the sequence available on the NCBI genebank database. Surprisingly, the sequence of the LeuRS gene of *CAI-4* strain had several polymorphisms when compared with the sequence deposited in genebank. For this reason, a single nucleotide polymorphism (SNP) screen was carried out using five *C. albicans* strains. For this, the LeuRS gene (Ca*CDC60*) was amplified from genomic DNA, by PCR, and then sequenced. SNPs were detected by analysis of the sequencing output, using the BioEdit software.

5.2.2.1. SNPs analysis

The SNPs were either heterozygous, when two different nucleotides for the same position were present within the same strain; or homozygous, when there was a clear alteration of nucleotide relative to the reference strain (Figure 5. 5).

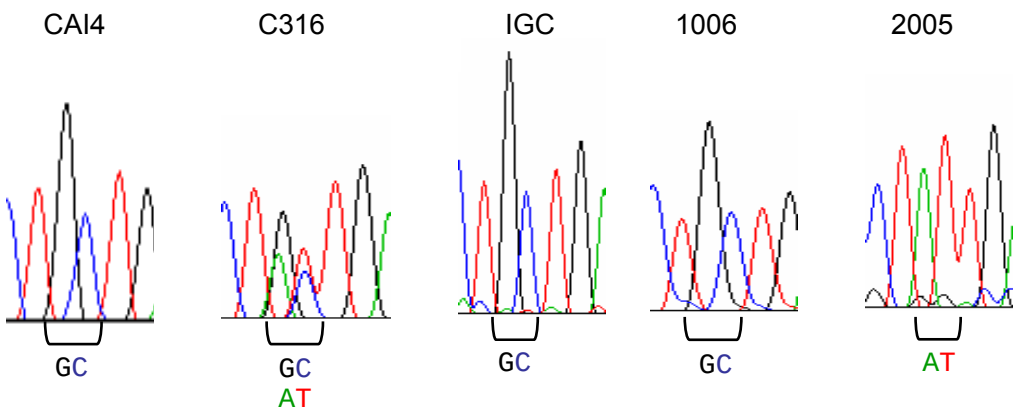


Figure 5. 5 – Single Nucleotide Polymorphism analysis.
Polymorphism detection in the gene encoding the LeuRS. Example of its nucleotides 394 and 395. Strain C316 is heterozygous, while the other strains are homozygous at these positions.

To standardize the analysis, whenever a polymorphic position was found, it was assigned to one of the two alleles of the *CaCDC60* gene (alleles *a* and *b*). For such

attribution, the sequence in question was always compared with the respective sequence deposited in the NCBI genebank database. As the public sequences were obtained after cloning the respective genes (O'Sullivan et al., 2001b; O'Sullivan et al., 2001a), they were always designated as allele *a*, whereas the alternative nucleotides where considered as allele *b*.

5.2.2.2. SNP analysis of the *C. albicans* LeuRS gene

For SNP detection, the coding sequence of LeuRS gene was amplified from genomic DNA of 5 different strains of *C. albicans: CAI-4, C316, IGC, 1006* and *2005*. Afterwards, the amplification products were sequenced and the polymorphic sites were detected by analysis of the sequencing output.

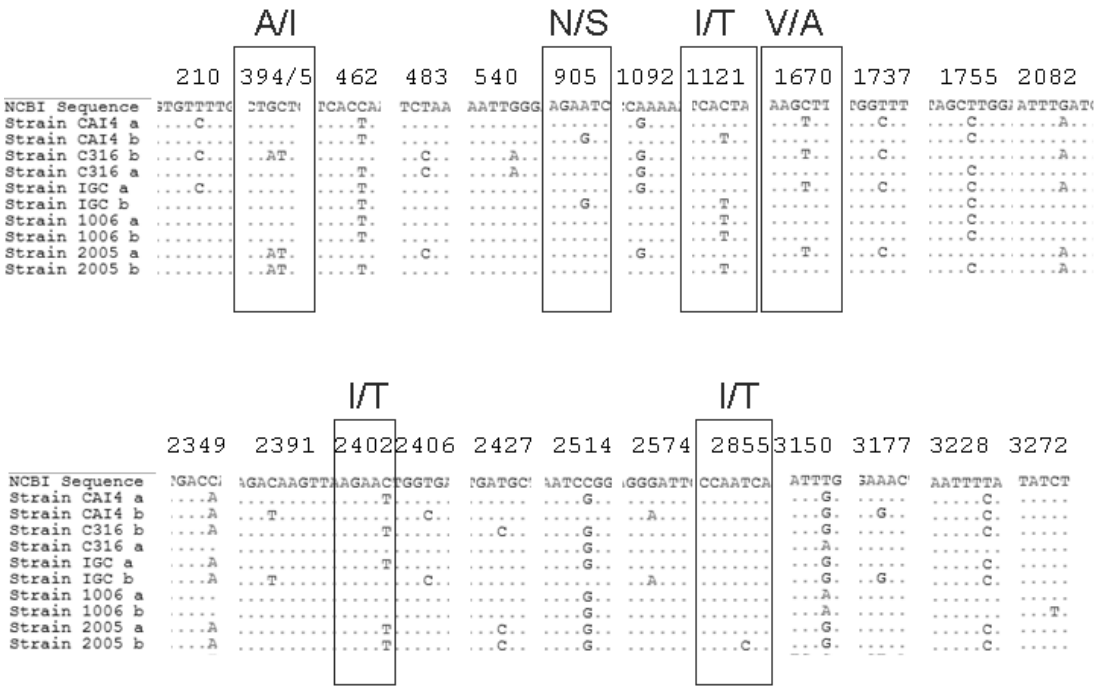


Figure 5. 6 – Polymorphisms identified in the LeuRS from different strains of *Candida albicans*.
Alignment of the LeuRS gene sequence from different *C. albicans* strains. The dots correspond to the identities of the first sequence (from the NCBI). The non-silent polymorphisms are framed, and the corresponding amino acid alterations are indicated. From all SNPs identified on the gene sequence, only six of them are not silent and lead to a change of the protein sequence.

The SNPs were distributed over 24 different positions of the LeuRS gene (Figure 5. 6). The SNP distribution is uneven, as both the number and localization of the polymorphic sites were not the same in all the tested strains, even though some of them were common to

different strains. Interestingly, there is a single neutral SNP in strain 1006. The impact of the SNPs in aaRS structure is softened because of the 24 polymorphisms discovered only 6 of them involve amino acid changes (Figure 5. 6), in other words, only a quarter of such alterations are non-synonymous.

These data clearly show that, with the exception of strain 1006, there are two isoforms of the LeuRS protein in the strains tested. Moreover, there is an intrinsic variety between strains: while the allele *a* is common to strains *CAI-4*, *IGC* and *C316*, the allele *b* varies in strain *C316*; and none of the isoforms of strain *2005* were found in the other strains. That is, there are at least 6 different LeuRS isoforms and only 2 strains encoded the same isoforms. Moreover, the promoter of the LeuRS gene is different in both alleles (Annexe G) suggesting the existence of a control mechanism in the transcription of the different alleles.

5.2.2.3. SNP analysis of the *C. albicans* SerRS gene

SNP screening of the *C. albicans* SerRS gene (Ca*SES1*) was carried out as described above for the LeuRS. Several SNPs were detected for the SerRS gene, which are distributed over 9 positions, within the analysed strains (Figure 5. 7). Again, the pattern of SNPs distribution was uneven.

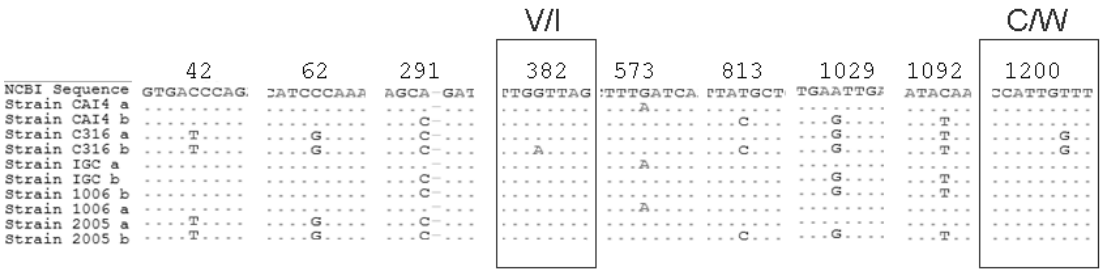


Figure 5. 7 – Polymorphisms identified in SerRS gene from different strains of *Candida albicans*. Alignment of the DNA sequence of SerRS gene of different *C. albicans* strains. Only in the C316 strain did the SNPs lead to a change of the amino acid in the protein sequence. The non-silent polymorphisms are framed, and the corresponding amino acid alterations are indicated.

The majority of the SNPs detected in the SerRS gene are neutral, the only exception was found in strain *C316* (Figure 5. 7), where amino acid changes were detected at the heterozygotic nucleotide 382, and at the homozygotic nucleotide 1200.

5.2.2.4. The natural variability of aminoacyl-tRNA synthetases

To assess whether polymorphic variation was an intrinsic characteristic of LeuRS and SerRS genes only, another synthetase gene was screened for the existence of SNPs. This control SNP screen was carried out using the tryptophanyl-tRNA synthetase gene.

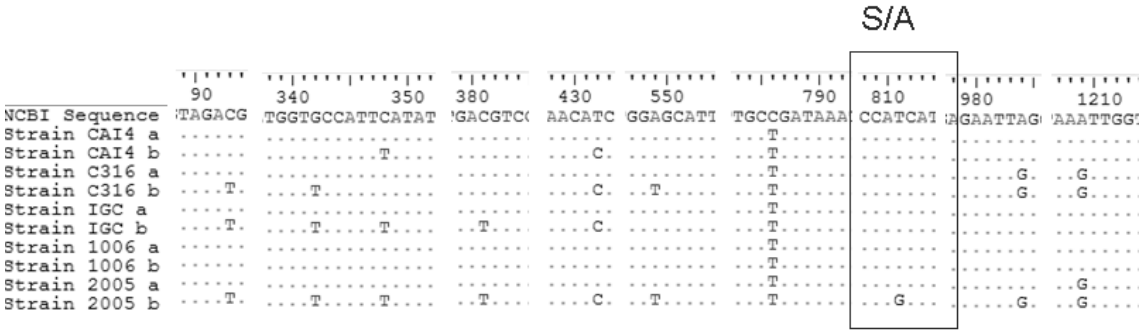


Figure 5. 8 – Polymorphisms identified in the *C. albicans* TrpRS gene.
Alignment of the DNA sequence of the TrpRS gene from different *C. albicans* strains. Only in the *2005* strain did the SNPs lead to a change of the amino acid in the protein sequence. The non-silent polymorphisms are framed, and the corresponding amino acid alterations are indicated.

SNPs were found among the five strains screened, distributed over 9 positions. However, of all the polymorphisms found in the gene sequence, only one corresponded to a change of the encoded amino acid and, this change appeared in strain 2005 only. This heterozygotic change in position 810 changed the serine TCA to the alanine GCA codon (Figure 5. 8).

Finally, to evaluate whether the high genetic diversity observed in the *C. albicans* aminoacyl-tRNA synthetase genes was specific of this fungus or a common feature in the fungal world, SNPs were also screened in LeuRS genes of four strains of *S. cerevisiae*: two of them were laboratory strains – *CEN-PK2* and *W303*, and the other two were clinical isolates – *MAS-4* and *MAS-5*.

The number of SNPs detected in *S. cerevisiae* was lower than that observed in *C. albicans*. Nucleotide changes were found at only 8 different positions, all of them on the pathogenic *S. cerevisiae* strains. However, in terms of protein sequence, only one of such SNPs found lead to an amino acid alteration, namely the heterozygotic change at nucleotide position 720 which results in substitution of an alanine for a threonine (Figure

5. 9).

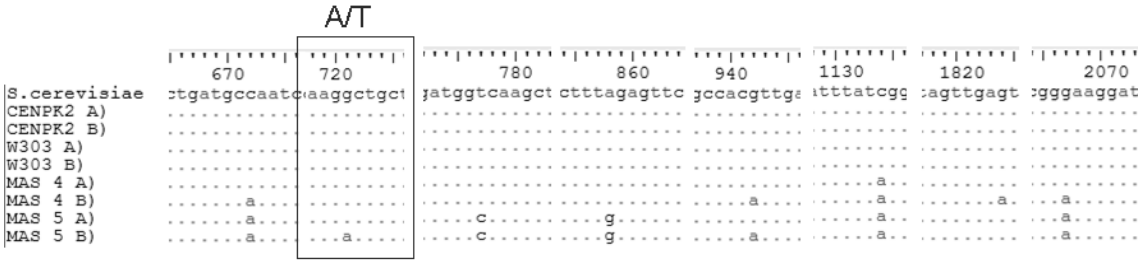


Figure 5. 9 – Polymorphisms identified in SerRS gene of *S. cerevisiae.*
Alignment of the sequenced LeuRS gene from different *S. cerevisiae* strains. Interestingly, none of the laboratory strains had SNPs, but the clinical *MAS-4* and *MAS-5* strains were polymorphic. In *MAS-5* the SNPs lead to an amino acid change in the protein sequence. The non-silent polymorphisms are framed, and the corresponding amino acid alterations are indicated.

In order to access the natural variability among the synthetases in *C. albicans*, the number of SNPs found for each gene was normalized for gene length and then compared. The *C. albicans* genes had similar number of polymorphic rates, namely 7.3, 6.0 and 7.0 per 1000 bp for the LeuRS, SerRS and TrpRS, respectively. All strains displayed several SNPs, indicating high mutation rate in this pathogenic fungus (Figure 5. 10 A). However, in *S. cerevisiae* the LeuRS gene from pathogenic/clinical strains was 2.4 / 1000 bp (Figure 5. 10 A). It was also interesting that the two non-pathogenic *S. cerevisiae* strains did not have SNPs.

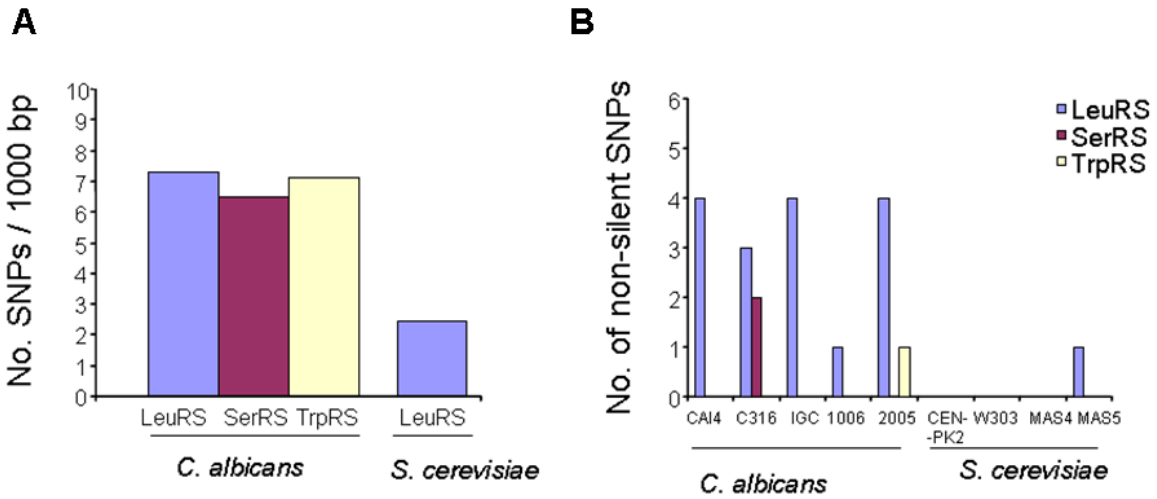


Figure 5. 10 – *C. albicans* has a naturally high SNPs rate
(A) Rate of SNPs in the sequenced genes of both *C. albicans* and *S. cerevisiae*. **(B)** The non-silent SNPs in LeuRS, SerRS and TrpRS. Each strain of *C. albicans* has its characteristic LeuRS isoform.

Nevertheless, most of those mutations were neutral because a base change did not correspond to a change on the encoded amino acid. From the genes analysed, the one encoding the LeuRS in *C. albicans* was, by far, the one with higher number of non-synonymous polymorphisms (Figure 5. 10 B). In fact, all but *1006* strain had two isoforms of the LeuRS protein. Conversely, only one of the strains of *S. cerevisiae* screened had a SNP that lead to an amino acid change, namely the *MAS-5* strain. In the case of both SerRS and TrpRS genes only the *C. albicans* strains *C316* and *2005*, respectively, showed two isoforms of the proteins (Figure 5. 10 B).

5.2.2.5. Structural analysis of the non-synonymous SNPs in the LeuRS

The LeuRS is a class I*a* aminoacyl-tRNA synthetase with a molecular mass of 133kDa and contains the highly conserved HA(I)HG, TLRPET and KMSKS signature motifs. In order to have a global view of the impact of the non-synonymous SNPs on the LeuRS structure, the amino acid substitutions resulting from the SNPs identified were located in the tertiary structure of the protein (Figure 5. 11).

Briefly, the substitution of base 132, from alanine to isoleucine, is located downstream of the HAGH motif; both the 302 and 374 substitutions are located on the editing domain; the amino acid alterations at positions 557 and 952 are on regions of the protein that apparently do not have specific functional roles and, finally, the alteration of residue 801 is located on the tRNA binding domain of the synthetase.

Regarding the chemical properties of the altered amino acids introduced by polymorphic variation, none are conservative. This indicates that the side chains of these amino acids are chemically different, which may cause distortion of the structure of the protein and consequently alter its activity. The non-conservative substitutions found in residue positions 374, 801 and 952 in the *C. albicans* LeuRS, involved isoleucine, which is hydrophobic, and threonine, which is hydrophilic. At position 132 alanine was replaced with isoleucine. Semi-conservative changes were identified in residues 302 and 557, where an asparagine and an alanine are replaced by a serine and a valine, respectively.

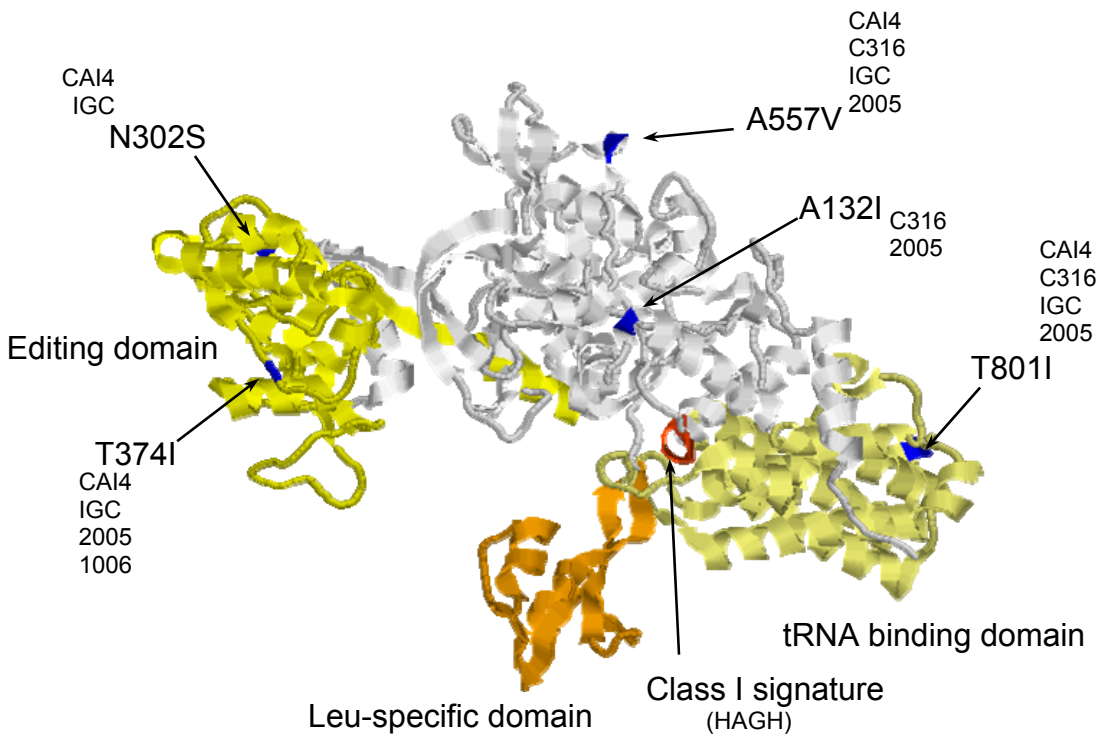


Figure 5. 11 – Impact of polymorphic variation on the 3D structure of LeuRS
Identification of the positions of non-synonymous polymorphisms (blue) on the 3D structure of the LeuRs was carried out using molecular modelling techniques. For this, the structure of the *C. albicans* LeuRS was modelled by RasMol using the crystal structure of the *T. thermophilus* LeuRS. The strains where each SNP was identified are indicated below each polymorphic residue. Some of the most important domains of LeuRS, such as Class I signature (red), the editing domains (yellow), the Leucine specific domains (orange) and the tRNA binding domain (pale yellow), are represented. Amino acid change in position 952 is not seen in this figure because it is located in a region of the LeuRS of *C. albicans* that does not exist in the *T. thermophilus* LeuRS.

5.2.2.6. The phylogeny of the different alleles of the *C. albicans* LeuRS

The polymorphic variability of the *C. albicans* LeuRS gene prompted one to carry out a phylogenetic analysis of the different alleles. With this, one hoped to highlight relationships between the alleles (Figure 5. 12).

Interestingly, this phylogenetic analysis revealed a division between both allelic forms *a* and *b*, as each of them form a cluster, suggesting that there is an evolutionary relation between these forms. The only exception is the 1006 strain, where alleles *a* and *b* are very close because this strain showed only one polymorphic position.

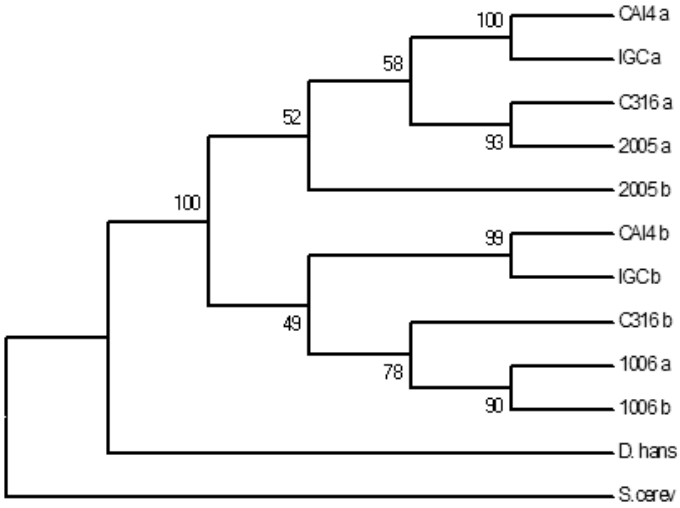


Figure 5. 12 – Phylogeny of the LeuRS isoforms
The phylogenetic tree was constructed using the Mega3 and is based on the LeuRS isoforms. The sequences were aligned and their phylogenetic relationship was analysed using the NJ algorithm. A bootstrap analysis with 100 repetitions was also carried out and its values are shown at key nodes.

5.2.3. Functional insights of the LeuRS and SerRS polymorphisms.

The existence of non-silent SNPs in the *C. albicans* LeuRS prompted the question of whether such polymorphisms have an impact on the kinetics of aminoacylation of the cognate tRNAs. Further, as both the SerRS and LeuRS genes contain one CUG codon, which can be ambiguously decoded as serine or leucine, 2 or 4 forms of the SerRS and LeuRS, respectively, are present in *C. albicans*. Therefore, a comprehensive analysis of the LeuRS and SerRS isoforms was carried out.

5.2.3.1. The LeuRS from *C. albicans*

Since the above studies on CUG ambiguity were carried out using the *C. albicans* strain *CAI-4*, the LeuRS isoforms of this strain were chosen for characterization. In order to predict the impact of the amino acid changes on the global activity of the enzyme, its structure was modelled, based on the known structure of the LeuRS of *Pyrococcus horikoshii* (Fukunaga and Yokoyama, 2005) and the putative structural changes induced by the amino acid substitutions were analysed. This analysis showed a superficial location of

the polymorphic amino acids in structural domains that do not interact with the tRNA (Figure 5. 13).

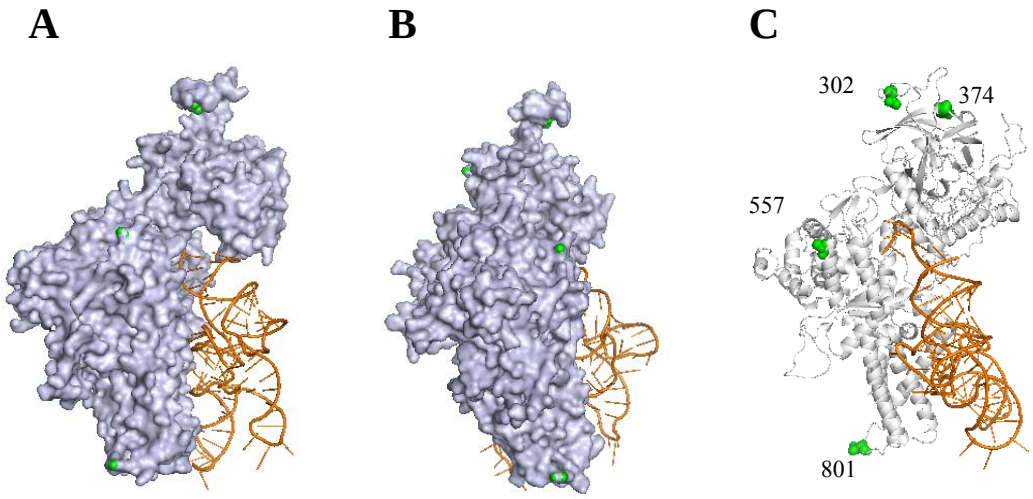


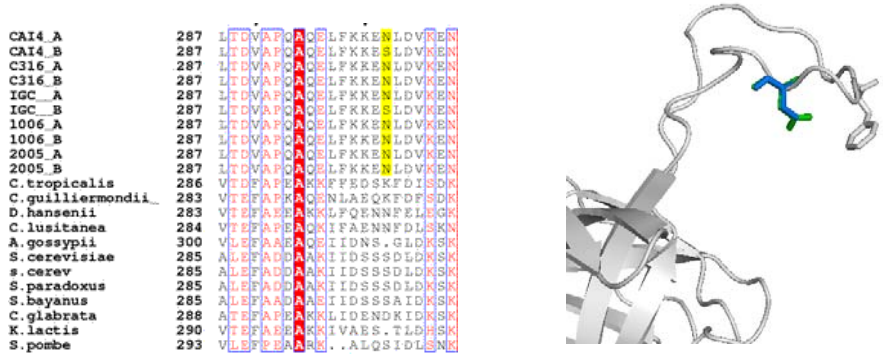
Figure 5. 13 – Polymorphic amino acid residue localization on the structure of the complex LeuRS-tRNA^{Leu}.
The polymorphic amino acid residues are represented as green spheres, on both the surface of the LeuRS **(A and B)** and on the cartoon of the LeuRS tertiary structure **(C)**. The tRNA interacting with the protein is represented by the orange ribbon. The structures were modelled with Pymol, based on the structure of the LeuRS from *Pyrococcus horikoshii,* deposited on the PDB under the code 1WZ2 (Fukunaga and Yokoyama, 2005).

Also, the alignment of the amino acid sequence of LeuRS from several yeasts showed that the threonine residue at position 801 is highly conserved, thus indicating that its change to isoleucine may have a stronger negative impact on the LeuRS structure than the other four amino acid changes in strain *CAI-4* (Figure 5. 14).

Figure 5. 14 – Model of the amino acid substitutions and their phylogeny.
(At right)
(A) The asparagine/serine polymorphism at position 302, is located in a non-conserved position. Likewise, **(B)** the threonine/isoleucine polymorphism is on position 374, and **(C)** the valine/alanine polymorphism is on position 557. They are both located in non-conserved regions of the protein. **(D)** Conversely, the threonine residue at position 801 is highly conserved. The cartoons for each amino acid substitution are on the right panels. The residues of allele *a* are represented in green, and those of allele *b* in blue. The protein structures were obtained with Pymol and the protein sequence alignments were obtained using ESPript, with the Blosum62 algorithm (Gouet *et al.*, 2003).

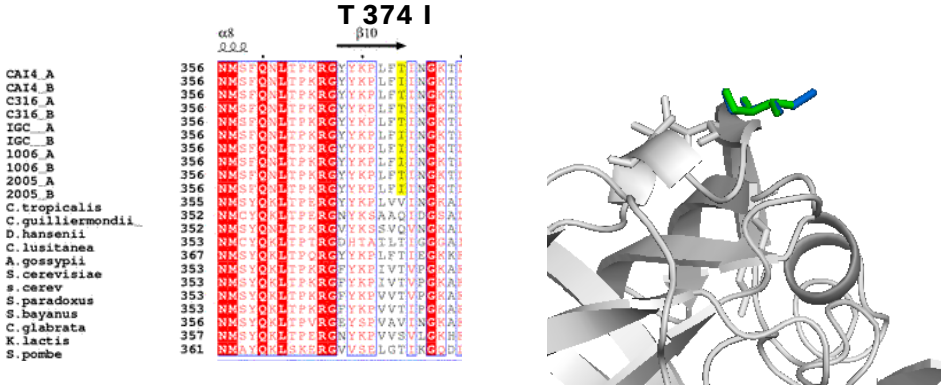
A

N 302 S



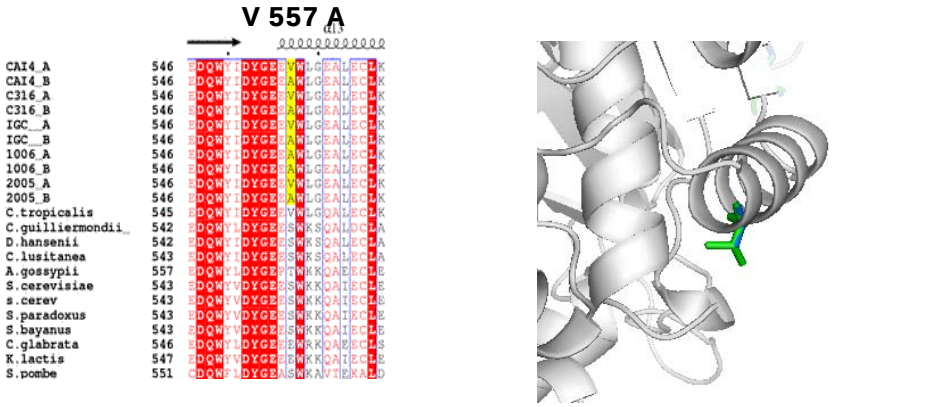
B

T 374 I



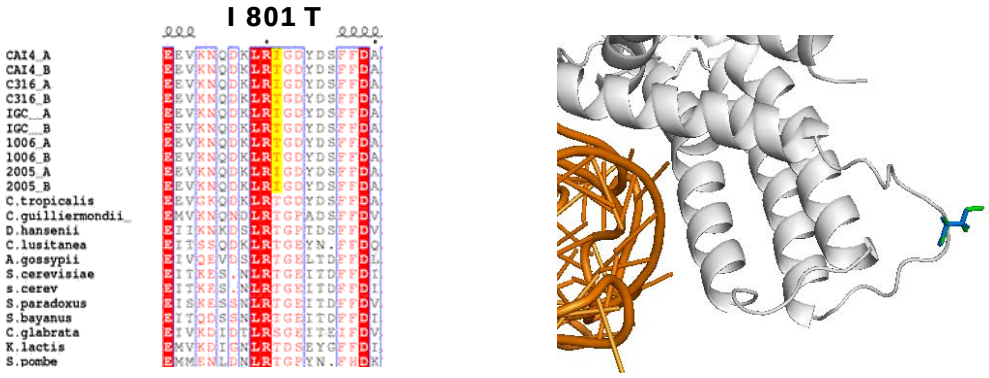
C

V 557 A



D

I 801 T



Concerning the amino acid residue encoded by the CUG codon (Figure 5. 15), it is located on the C- terminal protein domain, which was not possible to model from the structure of the LeuRS of *P. horikoshii* due to sequence divergence in this domain. Still, the C- terminal domain is rather similar among the yeasts, so it was possible to perform a primary sequence alignment of this region of the protein. Indeed, this analysis revealed that the residues encoded by the CUG codon in *C. albicans* are on a rather conserved region of the C-terminal domain of the LeuRS and, surprisingly, even those species whose CUG codon is decoded as serine have a leucine at this position, encoded by a standard leucine codon.

Figure 5. 15 – CUG localization on the *C. albicans* LeuRS primary structure
The amino acid residue number 919 of the LeuRS is a highly conserved leucine. However, in *C. albicans* and *C. tropicalis* this residue is encoded by a CUG codon, which is decoded mainly as serine (yellow). Interestingly, in both *D. hansenii, C. guilliermondii* and *C. lusitanea*, whose CUG codon has also changed its identity, the leucine at this position is conserved, as it is encoded by other leucine codons. The protein sequence alignments were carried out using ESPript with the Blosum62 algorithm (Gouet *et al.*, 2003).

5.2.3.2. The SerRS from *C. albicans*

The *C. albicans* SerRS also has one CUG codon in its coding sequence, indicating that there are two forms of this protein *in vivo*, due to the ambiguous CUG decoding. A multispecies alignment showed that this codon is located in a non-conserved region, thus there is some flexibility in this residue (Figure 5. 16). Concerning the tertiary and quaternary structure, it was modelled on the basis of the crystal structure of the *T. thermophilus* SerRS (Cusack *et al.*, 1996). The serine encoded by the CUG codon is inserted in a highly structurally conserved region of the protein, namely in the interface of

the dimmer where there are a number of direct interactions between both subunits (Figure 5. 17). This raises the hypothesis that the SerRS may be unstable in *C. albicans* due to CUG ambiguity.

Figure 5. 16 – CUG localization on the *C. albicans* SerRS primary structure
The residue encoded by the CUG codon in *C. albicans* is highlighted in yellow, for the species that have undergone the CUG codon reassignment. This is a non-conserved residue among the other yeasts SerRSs.

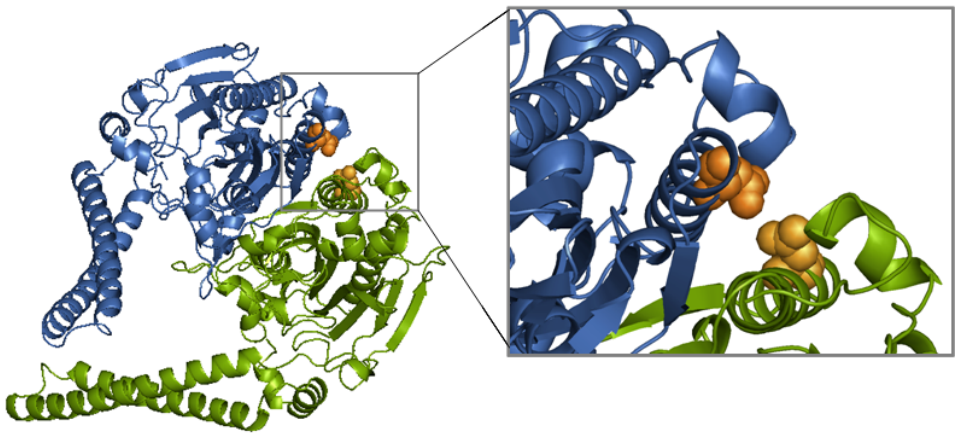


Figure 5. 17 – CUG localization on SerRS tertiary structure
The structure of the protein dimmer was modelled with Pymol, from the structure deposited in the PDB under the 1SES code (Belrhali *et al.*, 1994). The offset is a zoom in of the region of the protein with the residue encoded by the CUG codon, which is in orange. The two molecules of the dimmer are represented in blue and green.

5.2.4. The aminoacylation of *C. albicans* $tRNA_{CAG}$ by the LeuRS and SerRS

In order to study the activity of the LeuRS and SerRS from *C. albicans*, both genes were overexpressed in *E. coli*. The LeuRS protein was overexpressed using the plasmid pUKC1710, which was previously constructed by O'Sullivan (O'Sullivan *et al.*, 2001b). In order to facilitate purification of the recombinant LeuRS, a 6 histidine tag was inserted immediately after the initiation codon of the LeuRS gene from strain *2005*. This gene sequence was available in the laboratory and was mutated by site-directed mutagenesis to remove polymorphisms and reconstruct the LeuRS gene sequence from strain *CAI-4*, which was used in the CUG ambiguity experiments described in Chapter-3. Also, the CUG codon was mutated to the TCA-serine codon to ensure that serine and not leucine was inserted at the CUG position in the recombinant protein in *E. coli*. In total, 4 different plasmids, coding for 4 LeuRS isoforms were constructed: *i)* pUA74 encoded the most abundant allele *a*; *ii)* pUA81 encoded the most abundant allele *b,* both pUA74 and pUA81 had the CUG codon changed to the serine-TCA codon; *iii)* pUA82 encoded the least abundant allele *a* and *iv)* pUA83 encoded the least abundant allele *b,* where both pUA82 and pUA83 retained the CUG codon, which is decoded as leucine in *E. coli* (Figure 5. 18 and Table 5. 1).

Regarding the *C. albicans* SerRS, a plasmid (pUKC1722) containing the Ca*SES1* gene was also constructed by O'Sullivan (O'Sullivan *et al.*, 2001a). As before, the *CaSES1* gene sequence was from *C. albicans* strain *2005*. This plasmid was used to overexpress the SerRS with a leucine residue at the CUG codon position, the minor isoform of the SerRS protein in *C. albicans*. In order to produce the major form of the SerRS in *E. coli* containing serine at the CUG position, this codon was altered to the TCA-serine codon by site directed mutagenesis, resulting in plasmid pUA301.

Overexpression of the various isoforms of the *C. albicans* LeuRS and SerRS was carried out in *E. coli* BL21-CodonPlus® cells. These cells contain a plasmid that encodes

extra copies of the *argU* and *proL* tRNA genes, which enhances the expression in *E. coli* of proteins encoded by genes with high content of rare codons. The protein expression was induced by the addition of IPTG to a final concentration of 0.1 mM, for 5h at 30ºC, then the protein extracts were prepared and the proteins purified (Figure 5. 18 and Figure 5. 19), as described in Material and Methods.

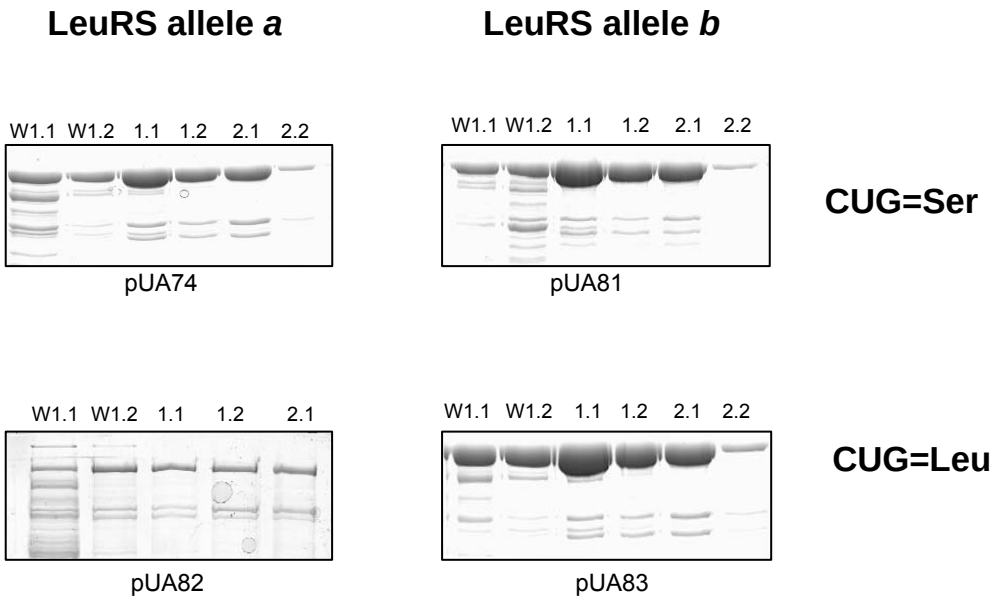


Figure 5. 18 – Purification of the recombinant LeuRS isoforms
Overexpression in *E. coli* and subsequent purification of *C. albicans* LeuRS isoforms, with a nickel chelating resin (Ni-NTA, Qiagen). The protein purification process was monitored by 10% SDS-PAGE, stained with coomassie-blue. The W1.1 and W1.2 fractions refer to column washing with 20 mM Imidazol. Fractions 1.1 and 1.2 refer to the protein elution with 20 mM Imidazol, and fractions 2.1 and 2.2 refer to the protein elution with 40 mM Imidazol.

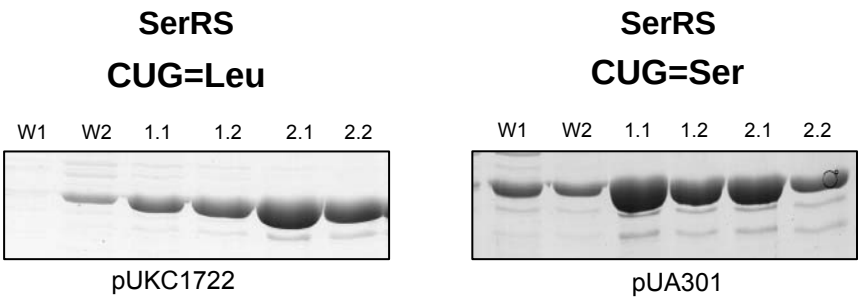


Figure 5. 19 – Purification of the recombinant SerRS
Overexpression, in *E. coli,* and subsequent purification of *C. albicans* SerRS, with a nickel chelating resin (Ni-NTA, Qiagen). The protein purification process was monitored by 12% SDS-PAGE, stained with coomassie-blue. The W1 and W2 fractions refer to column washing with 20 mM and 40 mM Imidazol, respectively. Fractions 1.1 and 1.2 refer to the protein elution with 60 mM Imidazol, and fractions 2.1 and 2.2 refer to the protein elution with 100 mM Imidazol.

The efficiency of the protein purification process was monitored by SDS-PAGE. The purest fractions were selected for the aminoacylation kinetics assays (Figure 5. 1), and the protein present in the selected fractions was quantified using the BCA assay (from Pierce).

Table 5. 1 – Overview of the protein fractions purified

	Plasmid	CUG decoding	Isoform	Selected Fraction	Volume (mL)	Protein concentration ($\mu g.\mu L^{-1}$)
SerRS	pUA301	Serine	- -	2.2	4.5	0.40
SerRS	pUKC1722	Leucine	- -	2.1	4.5	0.34
LeuRS	pUA74	Serine	*a*	1.1	4.5	1.05
LeuRS	pUA82	Leucine	*a*	1.1	4.5	0.21
LeuRS	pUA81	Serine	*b*	2.1	4.5	1.06
LeuRS	pUA83	Leucine	*b*	1.2	4.5	0.56

5.2.4.1. tRNA purification

For aminoacylation assays, the *C. albicans* $tRNA_{CAG}^{Ser}$ was also purified to near homogeneity. The determination of the kinetics of tRNA aminoacylation reactions is normally carried out using *in vitro* transcribed tRNAs (Sampson and Uhlenbeck, 1988), however, the modified bases in the anticodon of the $tRNA_{CAG}^{Ser}$ are very important to maintain its structure (Santos *et al.*, 1996). For example, the *C. albicans'* $tRNA_{CAG}$ has m^1G_{37} which is directly recognized by the LeuRS and influences the aminoacylation kinetics (Suzuki *et al.*, 1997). For this reason, it was important to obtain a fully modified fraction of the tRNA to ensure its correct aminoacylation. For these assays, two positive controls were used, namely the abundant leucine-CAA and serine-AGA tRNAs.

Since the abundance of the $tRNA_{CAG}^{Ser}$ is low in wild type *C. albicans* cells due to the low copy of its gene (1 copy/haploid genome), the latter was cloned into plasmid pUA12 as a single fragment containing 3 of its copies in tandem, this yielded the plasmid pUA77 (Section 2.2.2.2). The $tRNA_{CAG}^{Ser}$ was then purified from *C. albicans* cells transformed with the pUA77. (Figure 5. 20 A). Regarding the $tRNA_{AGA}^{Ser}$ and the $tRNA_{CAA}^{Leu}$, they are very abundant tRNAs, with 4 and 6 genome copies, respectively, and

their purification was carried out directly from *C. albicans* total tRNA preparations (Figure 5. 20 B).

Purification of the tRNAs was carried out by affinity chromatography as described by Tsurui and colleagues (Tsurui et al., 1994; Suzuki et al., 1996). For this, total tRNA extracts prepared as described in 2.4 were hybridized with a specific solid-phase DNA probe immobilized on agarose beads, as described in section 2.4.1. In the first step, 120 mg and 90 mg of total tRNA were extracted from both 180 g of wild type and from 120 g of pUA77 transformed *C. albicans* cells, respectively (Figure 5. 20).

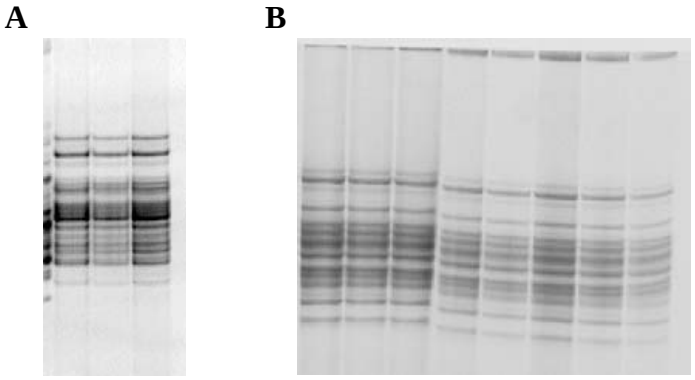
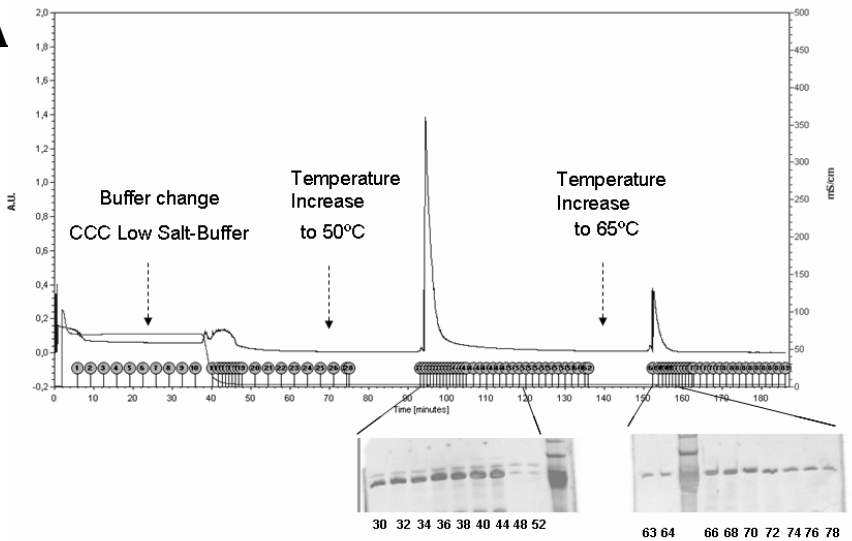


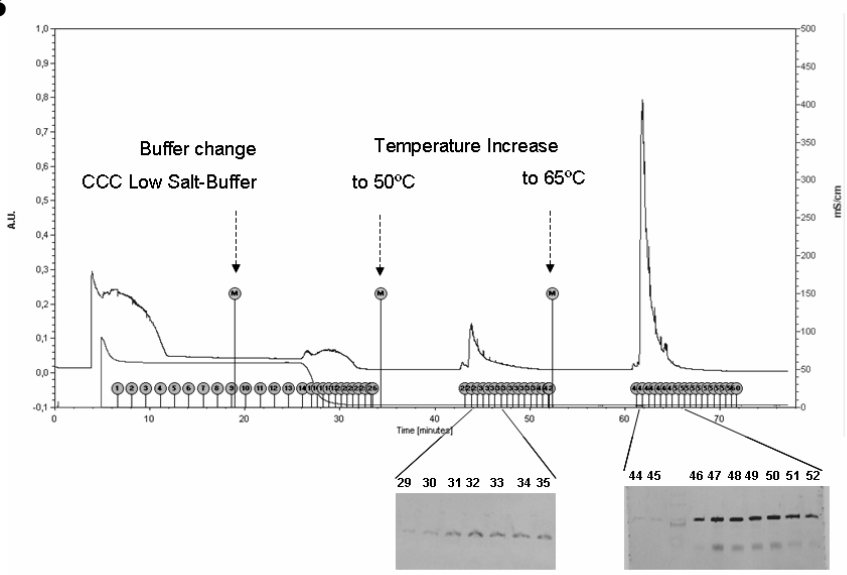
Figure 5. 20 – Total tRNA extracts.
Total tRNA extracts from **(A)** *CAI-4*-pUA77 and **(B)** *CAI-4* wild type cells. The integrity of total tRNA extracts was analysed by electrophoresis on a denaturing 8M Urea-TBE 10% acryl:bisacrylamide gel. Gels were stained with ethidium bromide for tRNA visualization stained with ethidium bromide by UV.

The $tRNA_{AGA}^{Ser}$ and the $tRNA_{CAA}^{Leu}$ were purified from the same total tRNA extract, which was divided in two batches. One was used for two independent purifications of the $tRNA_{AGA}^{Ser}$ (Figure 5. 21 B) and the other for the purification of the $tRNA_{CAA}^{Leu}$ (Figure 5. 21 C). The $tRNA_{CAG}^{Ser}$ was purified from 90 mg of total tRNAs from the *C. albicans*-pUA77 cells. Again, this total preparation was divided into two batches of 45 mg each, that were used twice (Figure 5. 21A). At the end, the fractions containing the purified tRNAs were pulled together, and run in denaturing 8M urea:TBE 10% Acryl:Bisacrylamide (19:1) gels, in order to access tRNA purity and integrity (Figure 5. 22).

A


Buffer change
CCC Low Salt-Buffer
Temperature
Increase
to 50ºC
Temperature
Increase
to 65ºC
A.U.
mS/cm
Time [minutes]
30 32 34 36 38 40 44 48 52
63 64
66 68 70 72 74 76 78

B


Buffer change
CCC Low Salt-Buffer
Temperature Increase
to 50ºC
to 65ºC
A.U.
mS/cm
Time [minutes]
29 30 31 32 33 34 35
44 45
46 47 48 49 50 51 52

C

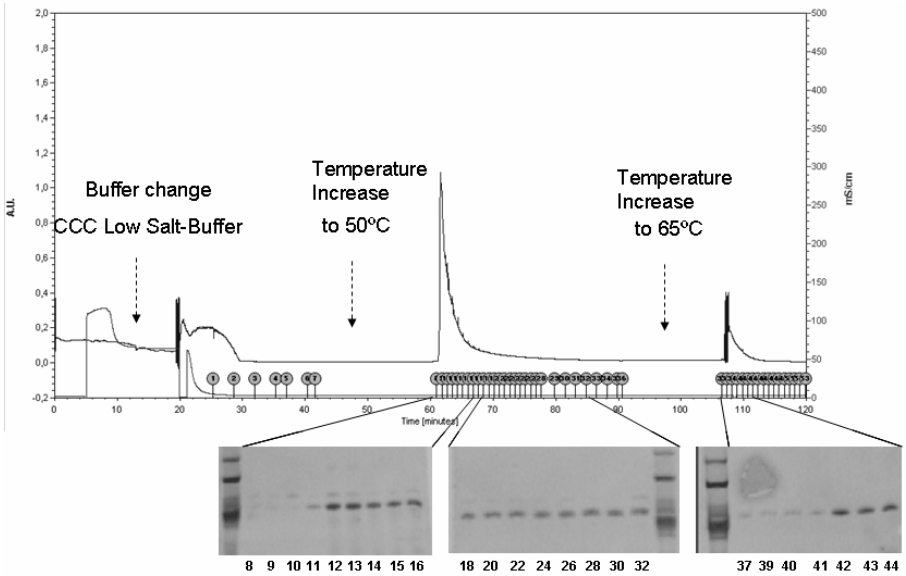

Buffer change
CCC Low Salt-Buffer
Temperature
Increase
to 50ºC
Temperature
Increase
to 65ºC
A.U.
mS/cm
Time [minutes]
8 9 10 11 12 13 14 15 16
18 20 22 24 26 28 30 32
37 39 40 41 42 43 44

Figure 5. 21 – tRNA purification by chaplet column chromatography
(At left)
(A) $tRNA_{CAG}^{Ser}$. **(B)** $tRNA_{AGA}^{Ser}$. **(C)** $tRNA_{CAA}^{Leu}$. The purification was performed using the BioLogic LP chromatography system (BioRad), coupled with a TCC-100 column oven (Dionex). The purification procedure is indicated on the chromatogram. The elution products were run on semi-denaturing TBE-4M Urea 15% acryl:bisacrylamide mini-gels, which were stained with ethidium bromide and visualized with an UV lamp source. The fractions containing the purified tRNAs were pulled together and precipitated.

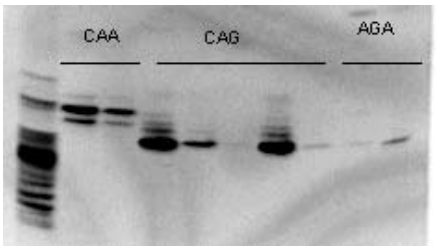


Figure 5. 22 – Monitoring tRNA purification by denaturing TBE-Urea acrylamide gel
The final purified fractions were applied onto a TBE-8M Urea 15% acryl:bisacrylamide gel, which was left to run overnight at 700 V.

The purified tRNA fractions were precipitated overnight with 0.1 vol of 3M NaOAc. and 2.5 vol of absolute ethanol, and then resuspended in TE buffer. Their concentration was determined by measuring their optical density at 260 nm (Table 5. 2). Pure tRNAs were then frozen at -80ºC for later use in the aminoacylation kinetics assays.

Table 5. 2– Pure tRNA obtained through the purification process

	Mw (Da)	[$tRNA_{NNN}$] (µM)	Volume (µL)
$tRNA_{CAG}^{Ser}$	28,384	97.8	15
$tRNA_{AGA}^{Ser}$	28,340	135	60
$tRNA_{CAA}^{Leu}$	29,293	112	65

5.2.5. Aminoacylation assays

The aminoacylation of tRNAs by their cognate aminoacyl-tRNA synthetases undergoes two steps (Section 1.3.2), whose kinetics can be independently measured by two different approaches. In the first step, the cognate amino acid is activated by the active site of the protein and in the second step the activated amino acid is loaded onto the acceptor stem of the cognate tRNA (Figure 1. 7). The first step of the reaction can be monitored by

the addition of radio-labelled [γ^{32}P]PPi. In the absence of tRNAs and in the presence of an excess of PPi, the reverse reaction is favoured, hence the γ-PO_4 of the PPi is transferred to the AMP, resulting in the formation of radiolabelled [γ^{32}P]ATP, which can be detected by scintillation counting methods (Section 2.5). The second step of tRNA charging reaction can be monitored by the addition of radiolabelled amino acids. In the presence of an excess of amino acid charging of the tRNA is favoured and the radiolabelled aa-tRNA can also be detected by scintillation counting methods (Section 2.5).

Surprisingly, initial studies with the purified LeuRS showed that neither total tRNA nor purified $tRNA_{CAA}^{Leu}$ or $tRNA_{CAG}^{Ser}$ could be *in vitro* charged with leucine. Conversely, serine charging by the SerRS was efficient (Figure 5. 23).

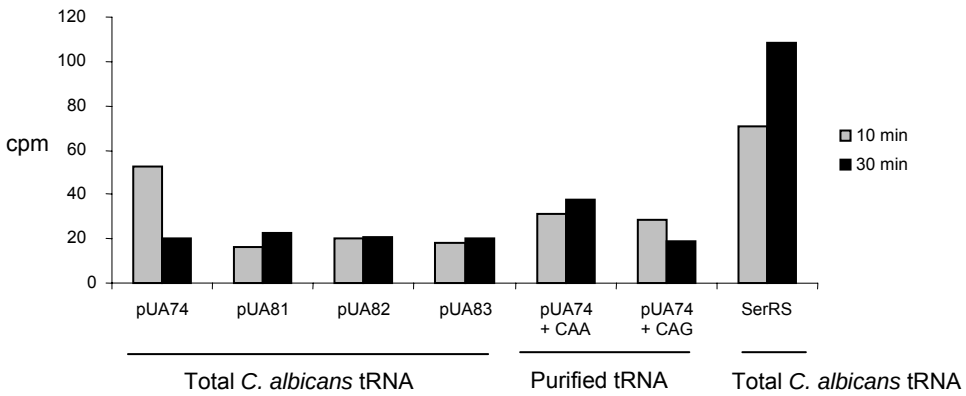


Figure 5. 23 – tRNA charging with LeuRS and SerRS
The overall aminoacylation mechanism was studied by tRNA charging assays with either [^{3}H]Leucine or [^{3}H]Serine. The tested LeuRSs (pUA74, pUA81, pUA82 and pUA83) failed to charge the leucine tRNAs. Both total tRNA extracts of *C. albicans* and purified tRNAs ($tRNA_{CAA}$ and $tRNA_{CAG}$) were tested. Conversely the SerRS was able to charge the serine tRNAs from the total tRNA extracts.

In order to clarify the failure of the LeuRS to aminoacylate the leucine tRNAs, the activity of the enzyme was tested by studying the first step of the aminoacylation reaction. That is, the ability of the active site of the protein to activate the amino acids (Figure 5. 24). Indeed, these studies showed that the active sites of all the four LeuRS isoforms were fully active, hence indicating that whatever was affecting tRNA charging with leucine, it was not related with a possible loss of activity during the process of protein purification. Considering that the protein was cloned into an *E. coli* expression vector with a 6-Histidine

tag at the N-terminus, which was not removed prior to the kinetics assays, it is probable that the tag interfered with binding of the tRNA to the enzyme.

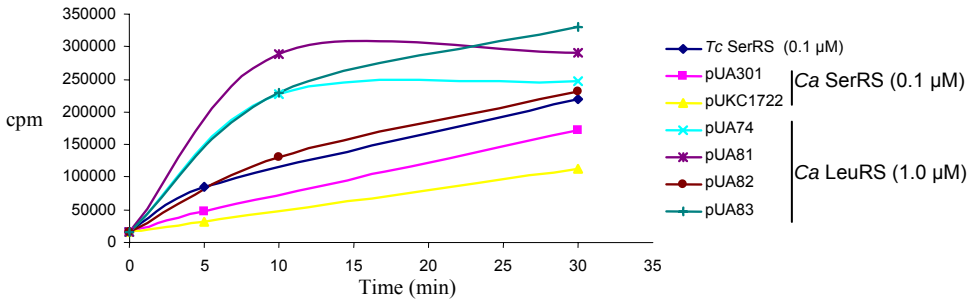


Figure 5. 24 – Amino acid activation by LeuRS and SerRS active sites.
The amino acid activation assays were performed to assess the activity of the active site of both SerRS and LeuRS. The SerRS from *T. cruzi* was used as a positive control. As the goal of this assay was to determine whether or not the LeuRS was active, an excess of this protein was used. The LeuRSs had fully active sites, as the amino acids were readily activated.

Since the two isoforms of the *C. albicans* SerRS were fully functional, it was possible to determine their aminoacylation kinetic parameters. The k_{cat} values for each protein isoform were obtained through tRNA charging assays with increasing concentration of the proteins, with tRNA concentration kept constant. For these experiments, the exact concentration of the tRNA and the enzyme were determined. The tRNA was titrated by a tRNA charging assay in the presence of an excess of the enzyme, so that the aminoacylation reaction was limited by the amount of tRNA and reached a plateau, which indicated the total amount of tRNA present in the assay. The exact concentration of the enzyme was determined by an active site titration assay. Once the exact concentrations of $tRNA_{AGA}$, $tRNA_{CAG}$, $SerRS_{pUA301}$ and $SerRS_{pUKC1722}$ were known, the k_{cat} values were calculated (Table 5. 3).

Table 5. 3 – K_{cat} of SerRS isoforms.

	k_{cat} (s^{-1})	
	$SerRS_{pUA301}$ **(CUG=Ser)**	***$SerRS_{pUKC1722}$*** **(CUG=Leu)**
$tRNA_{CAG}^{Ser}$	0.16±0.04	0.69±0.03
$tRNA_{AGA}^{Ser}$	0.16±0.04	0.89±0.01

Interestingly, the k_{cat} of the SerRS isoform with leucine at the CUG position was higher than that of the isoform with serine (the more abundant isoform in *C. albicans*). Furthermore, the k_{cat} of the SerRS isoform with leucine at the CUG position for the $tRNA_{AGA}^{Ser}$ is higher than for the $tRNA_{CAG}^{Ser}$. However, the k_{cat} only provides information of the reaction turn over and does not permit taking conclusions about the kinetics of the serylation of both $tRNA_{CAG}$ and $tRNA_{AGA}$ by the two isoforms of SerRS. Indeed, one can only conclude that $tRNA_{AGA}^{Ser}$/$SerRS_{(CUG=Leu)}$ pair had the highest turn-over. In order to fully characterise the serylation kinetics it is necessary to determine its k_m, that is, the enzyme affinity for the substrate, which was not possible in this project due to lack of time.

5.3. Discussion

In this chapter one tried to elucidate how CUG ambiguity is regulated *in vivo* in *C. albicans*, and clarify whether the CUG codon evolved to tolerate its ambiguity. Initially, one assessed whether the expression level of both enzymes was correlated to leucylation rates under different physiological conditions. For this, the amount of LeuRS and SerRS in cells grown at different physiological conditions was determined by Western-blot. However, no significant variation in the amount of both proteins or on the ratio LeuRS/SerRS was detected. These results indicate that the differential expression of these proteins is not responsible for cellular regulation of CUG ambiguity. Further, since it has been described that in *S. cerevisiae* the LeuRS enzyme is cleaved by the *yscB* protease (Larrinoa and Heredia, 1991), it is likely that the two bands detected in the *C. albicans* extracts also result from protease cleavage of the full length LeuRS enzyme, raising the hypothesis that the cleaved LeuRS is not active or is less active than the full length enzyme. However, the ratio full length/cleaved enzyme was the same in the growth conditions tested, indicating that regulation of CUG ambiguity does not result from a partial post-translational inactivation of the LeuRS by proteolytic cleavage.

The aminoacyl-tRNA synthetases are the enzymes responsible for the covalent bond between the amino acids and their cognate tRNA, therefore, it is necessary high specificity

in the selection of their cognate substrates in order to ensure faithful protein translation. Indeed, accumulation of error in the aminoacylation process will decrease the fidelity of protein synthesis and, eventually, lead to cell death. For this reason, this class of enzymes is highly conserved and is under an evolutionary pressure to maintain their sequence. Surprisingly, while studying the gene coding for the LeuRS (*CDC60*), in *C. albicans*, several SNPs were discovered. As some of them were non-silent nucleotide changes *C. albicans* cells contain various LeuRS isoforms.

Furthermore, a screen of the *CDC60* sequence in five different strains of *C. albicans* revealed that such polymorphisms were widely spread because 6 LeuRS isoforms were detected in 5 strains. This observation raised the question of whether or not protein diversity generated through polymorphic variation is a unique characteristic of the LeuRS, in *C. albicans*. To clarify this question the genes coding for the SerRS and TrpRS were also screened for presence of SNPs. Interestingly, SNPs were detected but they were silent, indicating that they do not generate protein isoforms. Also, the *CDC60* gene from 4 strains of *S. cerevisiae* was screened for the existence of SNPs, and again, all but one where silence. Taken together, the data collected for the SerRS and TrpRS gene from *C. albicans*, and for the LeuRS from *S. cerevisiae*, go in line with the existence of an evolutionary pressure for maintaining the protein sequence. However, this does not apply to the LeuRS gene in *C. albicans*, which raises the hypothesis of a more active role of this enzyme in the regulation of ambiguous CUG decoding. Interestingly, the only SNP found in the *S. cerevisiae* LeuRS gene was from a pathogenic strain, raising the intriguing hypothesis that pathogenic strains might have increased levels of mistranslation to generate phenotypic diversity (Miranda, 2007).

Moreover, the different LeuRS isoforms had divergent promoters, which suggests that the expression of the different alleles is regulated by transcription. In the general discussion (section 6.3), a model for such regulation is further exploited. These observations prompted the study of the aminoacylation reaction by both LeuRS and SerRS, as different affinities for the tRNA among the protein isoforms could be responsible for the regulation of CUG ambiguous decoding.

In *C. albicans,* the nucleotide polymorphisms are not the only source of protein variation, because ambiguous CUG decoding generates protein diversity. Therefore, in *C. albicans* cells there are 4 different LeuRS proteins and 2 different SerRS proteins. To clarify the functional role of such diversity one determined the substrate affinity (k_m) and the reaction turn-over (k_{cat}) of these enzymes with $tRNA_{CAG}^{Ser}$, $tRNA_{AGA}^{Ser}$ and $tRNA_{CAA}^{Leu}$. For this, overexpression plasmids, encoding all these proteins isoforms, were built, and the proteins were expressed in *E. coli* and purified. Similarly, an overexpression system for $tRNA_{CAG}^{Ser}$ in *C. albicans* was built, and the native $tRNA_{CAG}^{Ser}$, $tRNA_{AGA}^{Ser}$ and $tRNA_{CAA}^{Leu}$ were purified. The purified proteins and tRNAs were used to determine the aminoacylation kinetics, but it was not possible to complete these studies in the time frame of this thesis. Nevertheless, the hypothesis that each pair enzyme-substrate has different aminoacylation kinetic parameters and that it is possible to regulate the expression of each LeuRS isoform is as a good model for regulation of CUG ambiguous decoding under different physiological conditions. So, it is important to obtain the kinetics parameters for these aminoacylation reactions. Also, if there are differences in the aminoacylation kinetics of the proteins that contain leucine and serine at the CUG position, suggesting a feed-back regulation mechanism of CUG ambiguous decoding, this should be exploited.

6. General Discussion

6.1. The uniqueness of the *C. albicans* genetic code

In order to explain the evolutionary mechanism of genetic code alterations, two distinct theories have emerged – the *Codon Capture Theory* (Osawa and Jukes, 1989); and the *Ambiguous Intermediate Theory* (Schultz and Yarus, 1994). The *Codon Capture Theory* postulates that under a strong CG- or AT- pressure, codons poor in AT- or CG-, respectively, tend to disappear, which allows for the loss of the tRNAs that decode them. These lost codons can be re-assigned at later stages, by mutant tRNAs from different isoacceptor families. Such tRNAs direct codon reassignment. This theory is supported by the unassignment of CGG codons in *Mycoplasma capricolum*, and of AGA and AUA codons in *Micrococcus luteus*, whose CG- genome content is of 25% and 75%, respectively (Osawa *et al.*, 1992). On the other hand, the *Ambiguous Intermediate Theory* postulates that a structural change in the translational machinery is the key element in a genetic code change. Such structural change could occur on a tRNA molecule, allowing it to recognize near-cognate codons and creating codon ambiguity. This theory is strongly supported by reassignment of the leucine CUG codon to serine in some species of the *Candida* genus (Sugita and Nakase, 1999; Santos and Tuite, 1995).

Candida albicans is an excellent model system to study the evolution of genetic code alterations, and, in particular to test the *Ambiguous Intermediate* theory. In *Candida spp,* the CUG codon is decoded by a $tRNA_{CAG}^{Ser}$, which has appeared due to altered splicing of a $tRNA_{IGA}$, about 272 My ago – prior to the divergence between the *Saccharomyces* and *Candida* genus. This tRNA competed for approximately 100 My with the wild type $tRNA_{CAG}^{Leu}$ for CUG decoding (Massey et al., 2003; Yokogawa et al., 1992). However, when the *Saccharomyces* and the *Candida* genus diverged, the $tRNA_{CAG}^{Ser}$ was lost in the ancestral lineage of *Saccharomyces spp*, hence these organisms reverted CUG identity to its original meaning due to the presence of a cognate $tRNA_{CAG}^{Leu}$, while the ancestors of *Candida spp* lost the $tRNA_{CAG}^{Leu}$ and retained the mutant $tRNA_{CAG}^{Ser}$. This CUG codon identity change imposed a negative pressure on the CUG codon usage, which triggered massive mutational change of CUG codons to UUG or UUA leucine codons. This mutational force was so intense, that 98% of the CUG

codons of the *Candida* ancestor mutated. Simultaneously, the $tRNA_{CAG}^{Ser}$ has also created a positive selective pressure for the capture of new CUG codons, from the serine UCN codon family (Massey *et al.*, 2003). Altogether, the appearance of the novel $tRNA_{CAG}^{Ser}$ and the massive CUG codon redistribution in the genome of the *Candida* ancestor, strongly corroborate the synergistic effects of the *Ambiguous Intermediate* and *Codon Capture* theories in genetic code alterations.

Interestingly, CUG decoding in the *Candida* genus is highly heterogeneous. *C. glabrata* maintained the standard CUG decoding as leucine, *C. cylindracea* fully reassigned the CUG decoding from leucine to serine, and other *Candida* species decode it ambiguously (Sugita and Nakase, 1999; Suzuki et al., 1997; Santos et al., 1997). Such heterogeneity in CUG decoding has been explained by specific changes in the structure of the $tRNA_{CAG}^{Ser}$ in the different *Candida* species (Santos *et al.*, 2004). Considering that several species of the *Candida* genus, namely *Candida albicans* and *Candida tropicalis* are major fungal human pathogens (De Backer *et al.*, 2000), it is of utmost importance to fully understand their fundamental molecular biology. Therefore, the aim of this thesis was to study the decoding properties of the CUG codon in *C. albicans,* and to understand both the mechanism of genetic code alterations and the biology and physiology of *C. albicans.*

The studies presented in this thesis proved unequivocally that CUG codons are ambiguously decoded *in vivo* in *C. albicans*. Such ambiguity results from random insertion of 97% serine and 3% leucine at CUG positions. This data is in agreement with previous *in vitro* data which showed that the $tRNA_{CAG}^{Ser}$ could be charged with both leucine and serine (Suzuki *et al.*, 1997). Also, this study revealed that the levels of CUG ambiguity go beyond the basal cell's physiology, and is dynamically manipulated in response to the external stimuli. The leucine incorporation rate at the CUG codons varies between 0.66% to 4.95%, in opaque cells and in cells grown at pH 4.0, respectively. However, one should not exclude the hypothesis that in other physiological conditions the

leucine incorporation might be even higher. Indeed, we showed in this study that *C. albicans* tolerates up to 30% CUG ambiguity without visible effects in growth rate.

6.2. CUG ambiguity and the evolution of the *C. albicans* genome

The double identity of the CUG codon implies that each *C. albicans* protein is represented by a mixture of molecules containing leucine or serine at CUG positions. This indicates that proteome complexity is much greater than that expected for the 6438 *C. albicans* genes. The 13,074 CUG codons in of haploid genome of *C. albicans*, distributed over 66% of its genes, at a frequency of 1 to 38 CUGs per gene have the potential to generate 2^n polypeptides (*n* = total number of CUGs per gene), thus increasing the size of the *C. albicans* proteome exponentially. This work unveiled that the 6438 protein encoding genes of *C. albicans* have the potential to produce 283,000 million of combinatorial proteins. In other words, the *C. albicans* proteome has a statistical nature. So, when considering the biology of *C. albicans*, one should think in terms of probability rather than absolute numbers. For instance, the probability of a protein encoded by a gene with 3 CUGs to contain 1 leucine in cells grown at 30ºC, 37ºC, pH 4.0 and H_2O_2 is 8.36%, 10.8%, 13.4% and 11.1%, respectively; whereas in the engineered highly ambiguous cells, 43% of the proteins have at least 1 leucine incorporated at one of the CUG positions.

The real impact of CUG ambiguity in protein diversity can only be determined by taking into consideration both the number of CUG codons and the expression level of each gene. In this work, a tentative model to determine the number of different proteins in a cell was built and, although it is based on three important assumptions, provides an approximate estimate of the real impact of CUG ambiguity on the proteome. According to this model, the number of novel proteins encoded by *C. albicans* for CUG ambiguity levels of 2.9% is of 6.7×10^6, and it ranges from 1.56×10^6 up to 10.7×10^6, in opaque cells and in cells grown at pH 4.0, respectively. Still, these numbers are below the 42.8×10^6 novel

proteins present in highly ambiguous cells (28% ambiguity), which illustrate the plasticity of *C. albicans* proteome.

Previous studies have shown that only 2% of the original CUG codons are still present in the genes of *S. cerevisiae* and *S. pombe* and that the remaining 98% are located in positions that correspond to serine or amino acids with similar chemical properties to serine (Massey *et al.*, 2003). The most recent assembly of the *C. albicans* genome permitted a detailed study of CUGs distribution according to chromosome localization, gene ontology, protein domains and gene evolution. For this, a comprehensive analysis of the usage of both CUG and AGC codons in the ORFs of *C. albicans* was carried out, and compared with each other. However, this approach did not unveil classes of genes with a unique CUG usage. Despite this, it highlighted a group of genes with high potential level of interest on further studies (Group 1 and 6, Annexe D).

6.3. Hypothetical models for regulation of leucine incorporation at the CUG codon

The ambiguous CUG decoding in *C. albicans*, resulting from $tRNA_{CAG}^{Ser}$ mischarging is rather interesting from a structural perspective because it is not yet clear how this novel tRNA is recognized by the LeuRS. Archeal and most eukaryotic LeuRSs recognize the long variable arm of cognate $tRNA^{Leu}$ (Fukunaga and Yokoyama, 2005), while the yeast LeuRS makes direct contact with the methyl group of m^1G_{37} and with A_{35} in the anticodon-loop and non-specific contacts with the phosphate backbone of the anticodon-stem (Soma et al., 1996; Suzuki et al., 1997). Like canonical $tRNA^{Leu}$, $tRNA_{CAG}^{Ser}$ contains A_{35} and m^1G_{37} in its anticodon-loop. However, the discriminator base is G_{73} (as in other $tRNA^{Ser}$) and not A_{73} (as in $tRNA^{Leu}$), which should prevent its recognition by the *C. albicans* LeuRS, as its role as an anti-determinant for leucylation has been shown in both yeast (Soma *et al.*, 1996) and human $tRNA^{Leu}$ (Breitschopf et al., 1995; Breitschopf and Gross, 1994). It is possible that the *C. albicans* LeuRS evolved a novel mechanism for recognizing both G and A at position 73. Another unique feature of

this $tRNA_{CAG}^{Ser}$ is the nature of the nucleotide at the position 33 – in all tRNAs there is a U at this position, but in this $tRNA_{CAG}^{Ser}$ there is a G. Such G_{33} has been described to play an important role in decreasing the affinity of the LeuRS, hence lowering its leucylation (Suzuki *et al.*, 1997).

Once leucine incorporation at CUG positions varies under different physiological conditions and the *C. albicans*' genome was proved to be extremely malleable, characterization of the charging mechanism of the $tRNA_{CAG}^{Ser}$ by SerRS and LeuRS is very important. Also, the LeuRS gene sequence of strain *CAI-4* was different from the published sequence from strain *2005*. Such differences were non-synonymous and it is likely that the isoforms have different aminoacylation properties. Further, 4 SNPs were identified in the *CAI-4* strain, indicating that the *C. albicans* LeuRS is highly polymorphic, as confirmed by further sequencing of the LeuRS genes from strains *1006*, *C316* and *IGC*. This diversity in protein sequences is unique for the LeuRS in *C. albicans*, which might correlate with the observed differences in leucine incorporation at the CUG codons, especially as the obtained preliminary data suggest the existence of two distinct promoters for each allele. Those SNPs and the divergent promoters of the LeuRS suggest that leucine incorporation at the CUG codons may be modulated by LeuRs-tRNA affinity differences and by different expression levels of each LeuRS isoform (Figure 6. 1).

According to the above model, one of the isoforms would have a higher affinity for the $tRNA_{CAG}^{Ser}$ and its expression would be controlled by a transcription factor sensitive to external stimuli. So, in a stress condition it would become more expressed, leading to a higher amount of the leu-$tRNA_{CAG}^{Ser}$, which would compete with the ser-$tRNA_{CAG}^{Ser}$ for CUG decoding at the ribosome, hence the leucine incorporation at CUG codons would be increased.

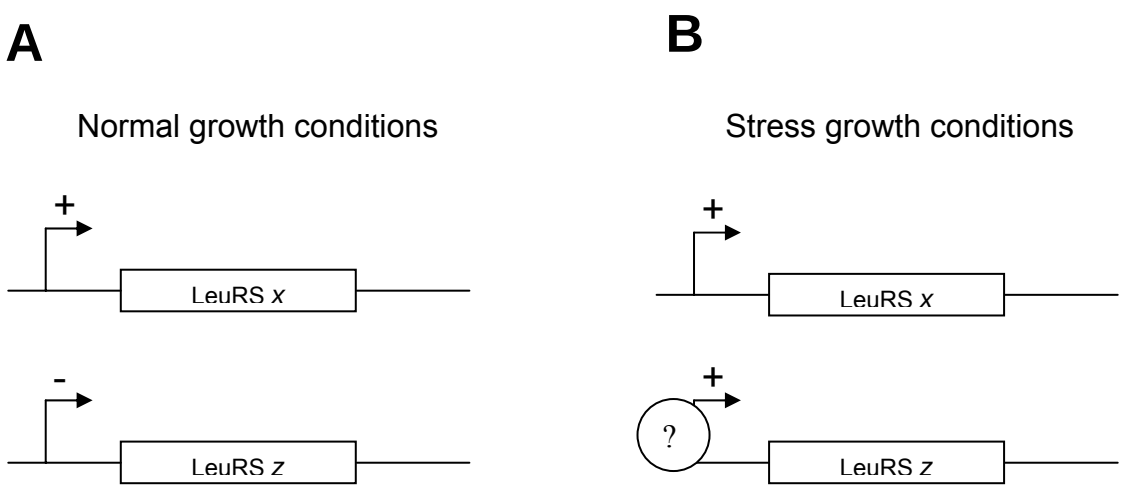


Figure 6. 1 – Model for the transcriptional control of LeuRS expression
Under normal growth conditions **(A)** one LeuRS isoform may be expressed, but not the other; whereas under stress conditions **(B)** the promoter of the latter isoform would be activated, so that it would became expressed.

In order to test this model, the various LeuRS isoforms were overexpressed in *E. coli*, and the $tRNA_{CAG}^{Ser}$, the $tRNA_{AGA}^{Ser}$ and the $tRNA_{CAA}^{Leu}$ were purified to determine the kinetics of the aminoacylation reaction. Unfortunately, the time available for these experiments did not allow one to finish them. Nevertheless, it would be important to further exploit this model by determining the kinetic parameters of the reaction and study the expression levels of each isoform under different physiological conditions by RT-qPCR. Also, it could be very interesting to study the activation of the LeuRS divergent promoters. For this, it will be important to sequence them from various strains and then analyse the sequences *in silico*. This may uncover specific enhancers that control LeuRS transcription. If such elements are identified, the promoters could be fused to the green fluorescent protein (GFP) gene to monitor promoter activation *in vivo* under different physiological conditions.

Both the SerRS and LeuRS genes contain a CUG codon, however, these genes complement *S. cerevisiae* SerRS and LeuRS gene knockouts without significant decrease in growth rate (O'Sullivan et al., 2001b; O'Sullivan et al., 2001a). This is probably due to the localization of the CUG codon in the non-conserved positions, so that such complementation is possible because both leucine and serine can be accommodated in the

position without major structural disruption. Nevertheless, one can not exclude some alterations in aminoacylation kinetics in the leucine isoforms. This could in fact provide an important regulatory mechanism of CUG ambiguity because the SerRS could work as a sensor for leucine incorporation levels through a negative feed-back mechanism. This would be possible if the SerRS isoform containing leucine at the CUG codon had higher affinity for the tRNA. Again, to test this hypothesis it is necessary to determine the kinetic parameters of the aminoacylation reaction for each enzyme, which was not achieved in the present work.

Another interesting feature of the polymorphic variation observed in the LeuRS and SerRS was the superficial location of the amino acid residues encoded by the SNPs. This may indicate that the SNPs do not affect the aminoacylation kinetics, but, as aminoacyl-tRNA synthetases form macrocomplexes by interacting with translational and non-translational factors, the SNPs may compromise protein-protein interactions and affect cellular networks. This hypothesis is also very interesting and should be tested experimentally. A preliminary analysis of the interactome of both LeuRS and SerRS, in *S. cerevisiae,* unveiled some interesting interactions.

According to the Database of Interacting Proteins (DIP, http://dip.doe-mbi.ucla.edu/dip/Main.cgi) (Salwinski *et al.*, 2004) the LeuRS interacts with 4 proteins (Figure 6. 2 A), namely with an arginase (Car1p), responsible for arginine degradation; with a RNA polymerase subunit, common to RNA polymerase I and III (Rpc40p); with the translation initiation factor eIF1 (Sui1p); and with a phosphoprotein phosphatase type 2C (Ptc6p). On the other hand, the SerRS interacts with 9 different proteins (Figure 6. 2 B): with Sen15p, a subunit of the tRNA splicing endonuclease; with Rvs167p, which is involved in regulation of actin cytoskeleton; with Air2p, a RING finger protein that interacts a methyltransferase and hence may regulate methylation of some genes; with Hrr25p, which is a casein kinase that binds the C-terminal domain (CTD) of RNA polymerase II, and is involved in regulating diverse events including gene expression, DNA repair and chromosome segregation; with Lys14p, a transcriptional activator involved in the regulation of genes of the lysine biosynthesis pathway; with Hrp1p, which is a nuclear ribonucleoprotein, involved in the cleavage and polyadenylation of pre-

mRNA 3' ends; with YBL036C, a racemase; the Sir3p, which is involved in the establishment of the transcriptionally silent chromatin state; and with YOL087C, an hypothetical protein. Nevertheless, the yeast interactome is not yet fully established, indeed, a comparison of the protein interaction networks between different public databases (*e.g.* the DIP *vs.* the BioGRID (http://www.thebiogrid.org/index.php) *vs.* iHOP (http://www.ihop-net.org/UniPub/iHOP/)) shows that there are divergences among them.

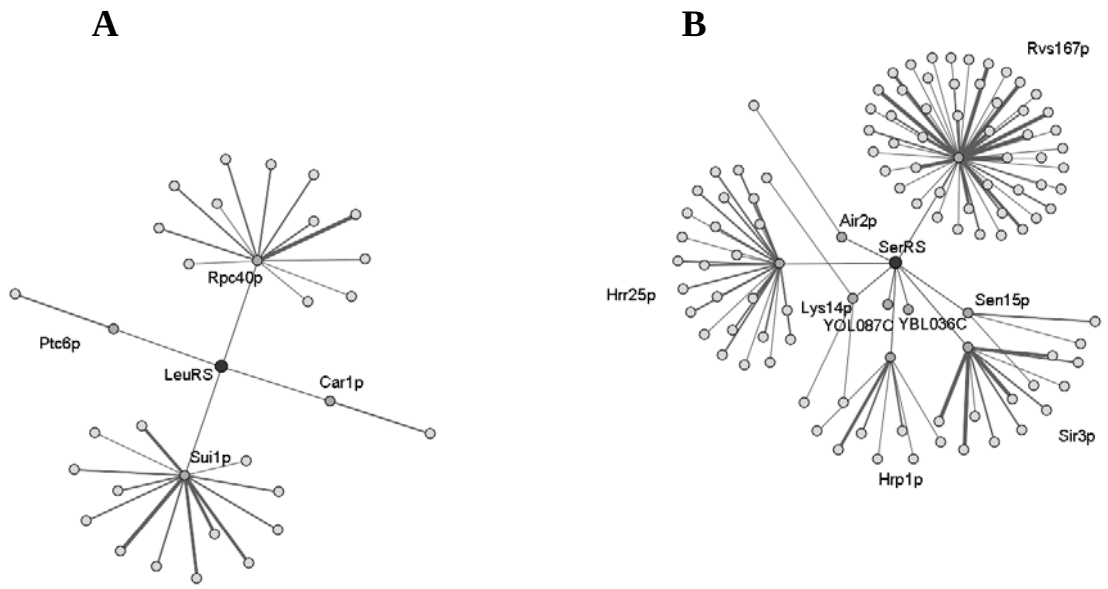


Figure 6. 2 – Interactome of LeuRS and SerRS
The interactome of **(A)** LeuRS and **(B)** SerRS, obtained from the Database of Interacting Proteins (DIP; http://dip.doe-mbi.ucla.edu/dip/Main.cgi)

Also, regulated expression of the $tRNA_{CAG}^{Ser}$ may provide an additional mechanism for controlling CUG ambiguity. A preliminary analysis of the $tRNA_{CAG}^{Ser}$ gene showed that it is located in the promoter region of the orf19.954, a homolog of the *S. cerevisiae YDJ1* gene, which encodes a protein of the DnaJ/Hsp40 family (Figure 6. 3). These proteins are chaperones involved in protein translation, folding, unfolding, translocation, and degradation, primarily by stimulating the ATPase activity of chaperone proteins, namely Hsp70 and Hsp90 (Qiu *et al.*, 2006).

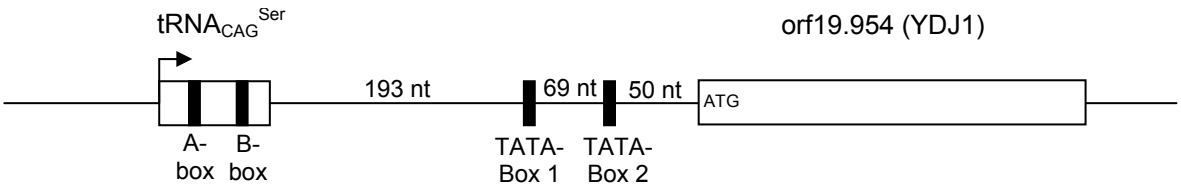


Figure 6. 3 – The localization of the tRNA$_{CAG}^{Ser}$ in the genome
The localization of the gene encoding the tRNA$_{CAG}^{Ser}$, in the chromosome 5 of *C. albicans*, as predicted in the assembly 20 (http://www.candidagenome.org/).

Such proximity of the tRNA$_{CAG}^{Ser}$ gene to the promoter of a protein of the DnaJ/Hsp40 family is interesting, because in *S. cerevisiae* strains expressing the [*PSI*$^{+}$] prion the overexpression of Ydj1 has a prion-curing effect (Kryndushkin *et al.*, 2002), hence it has been related with mistranslation events. Further, transcription of tRNA genes can suppress transcription of nearby RNA polymerase II genes (Wang et al., 2005; Hull et al., 1994). Therefore, it would be very interesting to study the expression of this tRNA$_{CAG}^{Ser}$/YDJ1 system, by using reporter genes, such as GFP or β-Galactosidase.

6.4. Conclusion

Apart from the mechanistic aspects of CUG ambiguity, this work provides new insights into the evolution of the genetic code. In yeasts, codon ambiguity successfully induces the stress response and increases tolerance to high temperature, lethal doses of heavy metals and drugs (Santos *et al.*, 1999). Previous work from the laboratory has shown that high ambiguity levels of CUG codons results in the generation of phenotypic diversity (Miranda, 2007), illustrating the positive effects of genetic code ambiguity and its negative effects on the proteome. Also, inactivation of the Hsp90 molecular chaperone in *Drosophila melanogaster* and *Arabidopsis thaliana*, allowed the expression of polymorphic proteins involved in cell signalling pathways and generated phenotypic diversity (Queitsch et al., 2002; Rutherford and Lindquist, 1998; Sollars et al., 2003; True and Lindquist, 2000). In *S. cerevisiae* and *C. albicans*, Hsp90 has a critical role in drug resistance by maintaining mutant drug resistance genes in a functional state (Cowen and

Lindquist, 2005). Yet, in another published study, generalized stop codon readthrough of genes and pseudogenes by the yeast [*PSI*$^+$] prion, disrupted the proteome, but resulted in morphological variation (Tuite and Lindquist, 1996).

All the above cases, genetic code ambiguity, Hsp90 inhibition and [*PSI*$^+$] prion induction, have similar destabilizing impacts on the proteome - all lead to large scale synthesis/accumulation of aberrant proteins - and increased phenotypic variation. Indeed, these data clearly indicate that the negative effect of codon ambiguity on the proteome may be overcome by its capacity to generate novel adaptive traits. Recent experiments on introduction of non-natural amino acids into the genetic code confirm the hypothesis that organisms can be highly tolerant to genetic code changes and readily adapt to genetic code ambiguity (Bacher et al., 2003; Bacher and Ellington, 2001; Balashov and Humayun, 2002; Ren et al., 1999; Slupska et al., 1996).

This thesis shows how genetic code ambiguity generates unanticipated proteome expansion. The data supports the hypothesis that earlier expansion of the genetic code from a small number of amino acids existent in primordial life forms, to the 22 encoded by extant organisms, could have been driven by selection through codon ambiguity. Further, the statistical proteome described herein for *C. albicans* supports the hypothesis that gradual codon identity changes create genetic barriers, such as the decrease in the sporulation and mating efficiency in *S. cerevisiae* lineages carrying the *C. albicans* $tRNA_{CAG}^{Ser}$ (Silva et al., 2007), resulting in the evolution of new species. This is confirmed by the inability to express heterologous genes in *C. albicans*. In other words, the *Candida* genus should have arisen as a direct consequence of this genetic code alteration. Indeed, the exponential expansion of the *C. albicans* proteome is of profound biological significance as arrays of proteins are generated from single mRNAs creating a statistical proteome. It implies that the probability of finding identical *C. albicans* cells in nature is extremely small.

6.5. Future work

This thesis showed that *C. albicans* has an ambiguous genetic code, which is sensitive to external stimuli. Such ambiguity expands the proteome on an unforeseen scale. But, how can the cell regulate such ambiguous decoding? Does it have an impact on the structure and function of the *C. albicans* proteins? These are still unanswered questions that are important to clarify. Therefore, the results described in this thesis define three future working lines, as follows:

1) To further study the LeuRS and SerRS proteins. It is imperative to determine the kinetic parameters of the aminoacylation reaction and clarify the existence of different promoters of the LeuRS. Their activation under different physiological conditions should also be studied. It is also important to study the interactome of the LeuRS and SerRS.

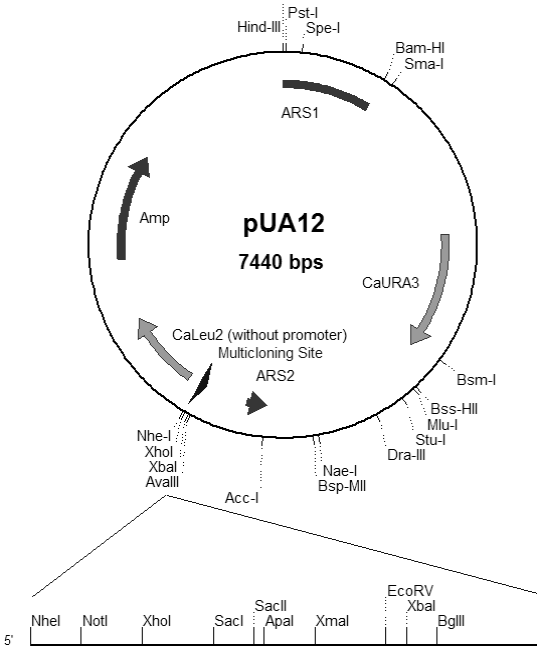
2) To study the expression of the $tRNA_{CAG}^{Ser}$ under the different growth conditions, and the co-expression of the $tRNA_{CAG}^{Ser}$/YDJ1.

3) To expand the SNPs screen, not only to more *C. albicans* strains, but also to more pathogenic strains of *S. cerevisiae*, and evaluate the impact of the polymorphisms in the LeuRS and SerRS structure by crystallizing both proteins.

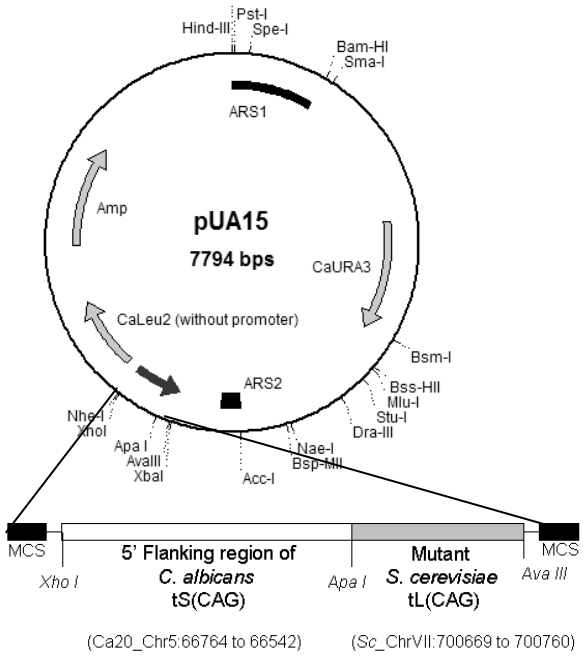
7. Annexes

Annexe A: Map of the Plasmids

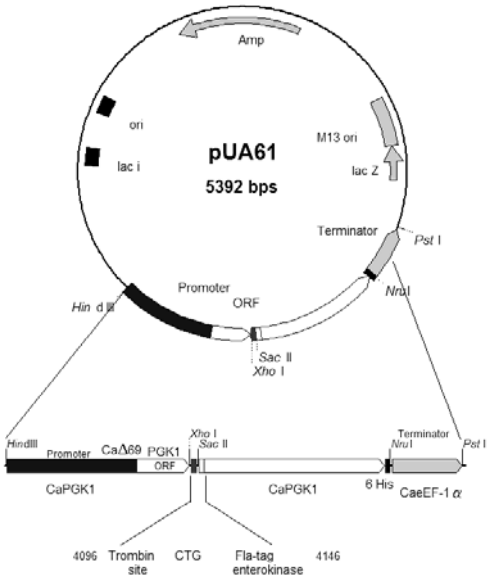
A.1 – Plasmids used in the chapter 3, for the *in vivo* determination of leucine incorporation at the CUG codon in *C. albicans*



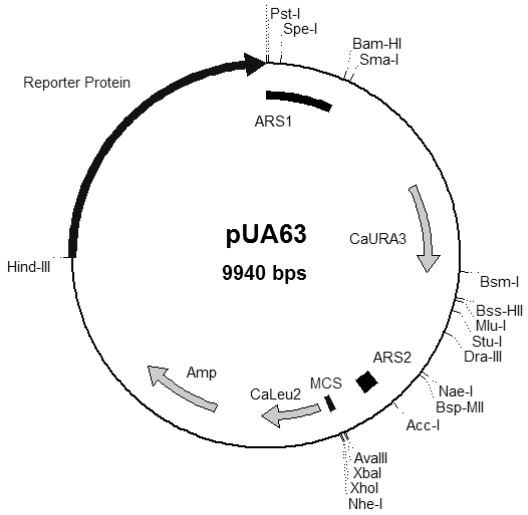
pUA12, based on pRM1, constructed by Miranda, I (2007).



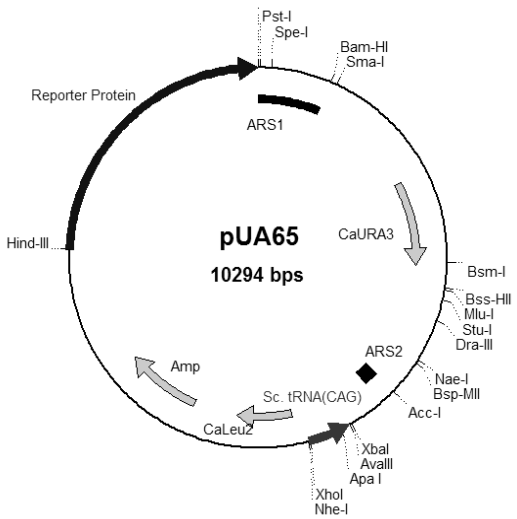
pUA15, plasmid bearing a mutant $tRNA_{CAG}^{Leu}$ from *S. cerevisiae,* built to increase the leucine incorporation in *C. albicans.* Based on pUA12, constructed by Miranda, I (2007).



pUA61, *E. coli* plasmid based on the pSL1190 vector. This plasmid was used to assemble the CUG reporter system used for measuring CUG ambiguity in *C. albicans*. For this, the reporter gene was assembled in three sequential steps, firstly, the promoter was cloned using the *Hind* III and *Xho* I restriction sites, the sequence coding for the reporter peptide, the core of the *CaPGK1* and the 6-Histidines Tag on the C- terminal, using the *Xho* I and *Nru* I restriction sites, and finally, the terminator sequence, from the *CaeEF-1α* gene at the *Nru* I and *Pst* I. restriction sites.

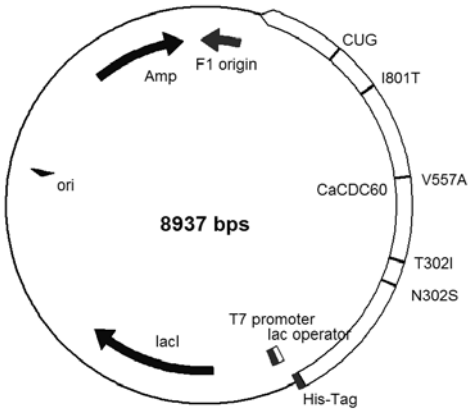


pUA63, *C. albicans* plasmid, based on the pUA12 shuttle vector. The whole reporter gene was extracted from pUA61, using the *Hind* III and *Pst* I restriction sites, and was inserted at the same restriction sites of pUA12.



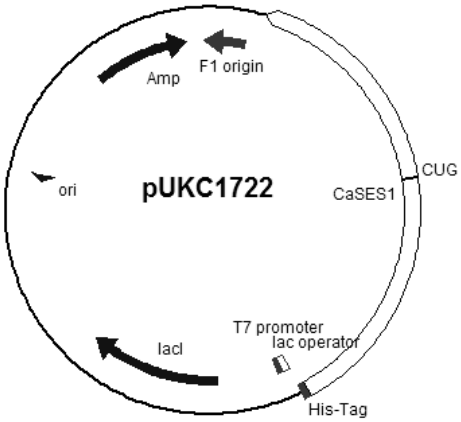
pUA65*, C. albicans* plasmid based on pUA15. Contains copy the *S. cerevisiae* $tRNA_{UAG}^{Leu}$ gene. Again, for this plasmid, the whole reporter gene was transferred from pUA61 as a *Hind* III and *Pst* I fragment and inserted in pUA15.

A.2 – Plasmids used in the chapter 5, for protein overexpression in *E. coli*

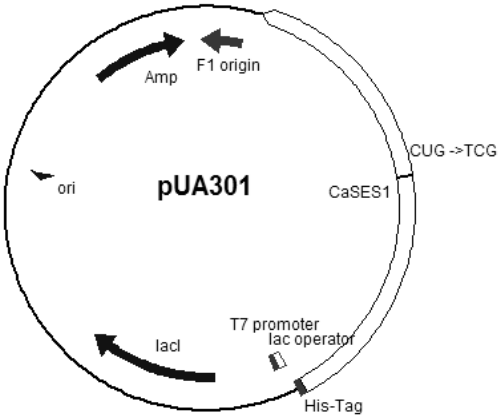


pUA74, pUA81, pUA82 and pUA83.
These plasmids were used to overexpress the 4 different LeuRS isoforms in *E. coli*. They are based on pUKC1710, which was built by O'Sullivan (2001). Several site directed mutagenesis were carried out in order to obtain the 4 isoforms of the protein. Below is a scheme of the differences between these plasmids.

					CUG	
					Ser	Leu
CAI-4	302	374	557	801		
allele a	KKENLDVI	PLFTING	YGEEAWL	LRTGDYD	pUA74	pUA82
allele b	...S....	...I...	....V...	..I....	pUA81	pUA83

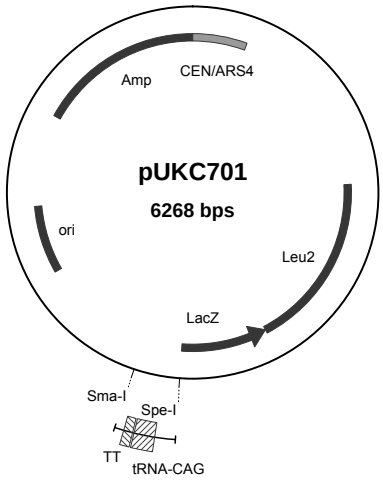


pUKC1722, plasmid for the overexpression of the SerRS protein in *E. coli*, built by O'Sullivan (2001).

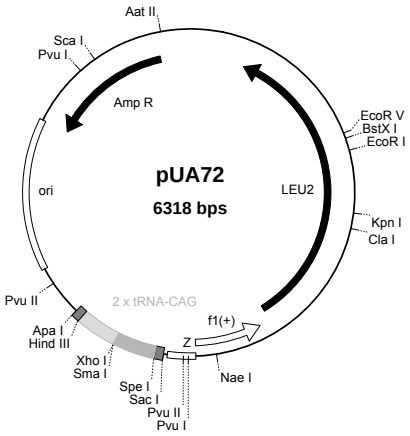


pUA301, plasmid for the overexpression of the SerRS protein in *E. coli*, with a serine at the CUG-position.

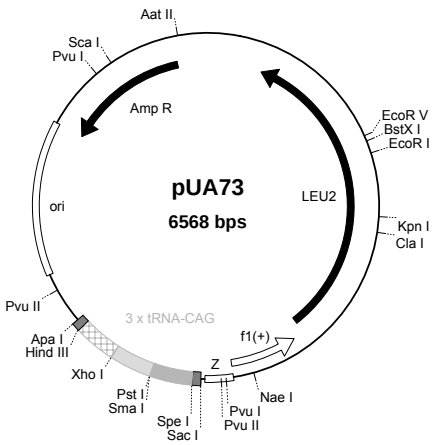
A.3 – Plasmids used in the chapter 5, for $tRNA_{CAG}^{Ser}$ overexpression in *C. albicans*



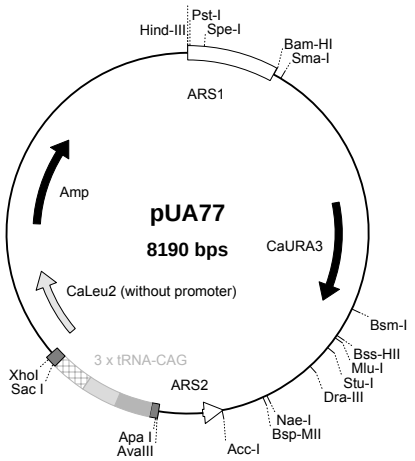
pUKC701, constructed by Santos (1999). It is based on pRS315, which contains the LEU2 auxotrophic marker for *S. cerevisiae* and the Amp^R marker for selection of E. coli in ampicillin media. This is a low copy plasmid in *S. cerevisiae*. The *C. albicans* $tRNA_{CAG}^{Ser}$ was cloned in the *Sma* I and *Spe* I sites of this vector's multicloning site.



pUA72, plasmid based on the pUKC701, where an extra copy of the *C. albicans* $tRNA_{CAG}^{Ser}$ was cloned in the *Hind* III and *Xho* I sites



pUA73, plasmid built on pUA72, where a third copy of the *C. albicans* $tRNA_{CAG}^{Ser}$ was cloned in the *Xho* I and *Pst* I sites



pUA77, plasmid built on pUA12, for the $tRNA_{CAG}^{Ser}$ overexpression in *C. albicans*. The DNA fragment containing the 3 copies of $tRNA_{CAG}^{Ser}$ was extracted from the pUA73, using *Xho* I and *Apa* III restriction sites and then cloned in the same restriction sites of pUA12.

Annexe B: Reporter protein data

B.1 - Sequence of the designed reporter protein, with primers annotations, as well as indications of both endonucleases and proteases cleavage sites.

```
                oUA 201
     attaggaagcttagtgttgcgtgtgtgtcaggaggcatgcgaaatatggagcgttcttcc
   1 ---------+---------+---------+---------+---------+---------+   60
           HindIII

      ccgggataaactcaatcgaggtattttttttcttcacttgaacacgagacgctttgac
  61 ---------+---------+---------+---------+---------+---------+   120

     attgaccaaatatgacgcaaacattatgatttggtgattgtttagctatagggaaggcaa
 121 ---------+---------+---------+---------+---------+---------+   180

     ttggaataggtgtcatatgatctagggctacaattgatcaatttgagcaaaaataataga
 181 ---------+---------+---------+---------+---------+---------+   240

     cgtagacactaattattaatcttgatttcattatacgtatgttcaaatttgttttctctc
 241 ---------+---------+---------+---------+---------+---------+   300

     tcttccagttgtgcgtctactgactcctgattgaattttttttgttcactttctacaa
 301 ---------+---------+---------+---------+---------+---------+   360

                                   oUA 224
     aaatatataaatcctcttttaattccaactcaattcactcttcatttcattcaattgaat
 361 ---------+---------+---------+---------+---------+---------+   420

     tttttttctttcttctttctttttatttcttatattttctaattcaatttttcaatc
 421 ---------+---------+---------+---------+---------+---------+   480

     aagcgattatcaattgcagaagaataactatcaaa**atg**tcattatctaacaaattatcag
 481 ---------+---------+---------+---------+---------+---------+   540
                                         MetSerLeuSerAsnLysLeuSerV

     tcaaagacttagacgttgctggaaagagagtctttatcagagttgacttcaacgtcccat
 541 ---------+---------+---------+---------+---------+---------+   600
     alLysAspLeuAspValAlaGlyLysArgValPheIleArgValAspPheAsnValProL

     tggacggtaagactatcaccaacaaccaaagaattgttgctgctttgccaaccatcaaat
 601 ---------+---------+---------+---------+---------+---------+   660
     euAsnGlyLysThrIleThrAsnAsnGlnArgIleValAlaAlaLeuProThrIleLysT
                oUA 234                                  oUA 202
     acgttgaagaacataaaccaaaatacattgtcttggcttcccacttgggtagaccaaacg
 661 ---------+---------+---------+---------+---------+---------+   720
     yrValGluGluHisLysProLysTyrIleValLeuAlaSerHisLeuGlyArgProAsnG

     gtctcgagctagttccaagaggttct**CTG**ccgcgggattataaagatgatgatgataag
 721 ---------+---------+---------+---------+---------+---------+   779
       XhoI  Trombin site      CTG SacII  Flag-Tag Enterokinase
     lyLeuGluLeuValProArgGlySerLeuProArgAspTyrLysAspAspAspAspLys
```

```
     aacgacaaatactcattagctccagttgctactgaattggaaaaattgttgggtcaaaaa
 780 ---------+---------+---------+---------+---------+---------+   839
     AsnAspLysTyrSerLeuAlaProValAlaThrGluLeuGluLysLeuLeuGlyGlnLys

     gtcaccttcttgaacgattgtgttggtccagaagtcaccaaggctgttgaaaacgccaaa
 840 ---------+---------+---------+---------+---------+---------+   899
     ValThrPheLeuAsnAspCysValGlyProGluValThrLysAlaValGluAsnAlaLys

     gatggtgaaatctttttgttggaaaacttgagataccacattgaagaagaaggttcttcc
 900 ---------+---------+---------+---------+---------+---------+   959
     AspGlyGluIlePheLeuLeuGluAsnLeuArgTyrHisIleGluGluGluGlySerSer

     aaagacaaggatggtaagaaagtcaaggctgatccagaagccgttaagaaattcagacaa
 960 ---------+---------+---------+---------+---------+---------+   1019
     LysAspLysAspGlyLysLysValLysAlaAspProGluAlaValLysLysPheArgGln

     gaattgacttcattggctgatgtctacattaacgatgcctttggtactgctcacagagcc
1020 ---------+---------+---------+---------+---------+---------+   1079
     GluLeuThrSerLeuAlaAspValTyrIleAsnAspAlaPheGlyThrAlaHisArgAla

     cactcctctatggttggtctcgaagttccacagagagctgctggtttcttaatgtccaaa
1080 ---------+---------+---------+---------+---------+---------+   1139
     HisSerSerMetValGlyLeuGluValProGlnArgAlaAlaGlyPheLeuMetSerLys

                                                      oUA 225
     gaattggaatactttgctaaggctttggaaaacccagaaagaccattcttggccattttg
1140 ---------+---------+---------+---------+---------+---------+   1199
     GluLeuGluTyrPheAlaLysAlaLeuGluAsnProGluArgProPheLeuAlaIleLeu

     ggtggtgctaaagtttctgacaagattcaattgattgacaacttgttggacaaggttgat
1200 ---------+---------+---------+---------+---------+---------+   1259
      GlyGlyAlaLysValSerAspLysIleGlnLeuIleAspAsnLeuLeuAspLysValAsp

     atgttgattgttggtggtggtatggccttcactttcaagaaaatcttgaacaaaatgcca
1260 ---------+---------+---------+---------+---------+---------+   1319
     MetLeuIleValGlyGlyGlyMetAlaPheThrPheLysLysIleLeuAsnLysMetPro

     attggtgattctcttttcgatgaagccggtgctaaaaacgttgaacacttggttgaaaaa
1321 ---------+---------+---------+---------+---------+---------+   1379
     IleGlyAspSerLeuPheAspGluAlaGlyAlaLysAsnValGluHisLeuValGluLys

     gctaagaaaaacaatgttgaattgatcttgccagttgattttgtcactgctgataaattc
1380 ---------+---------+---------+---------+---------+---------+   1439
     AlaLysLysAsnAsnValGluLeuIleLeuProValAspPheValThrAlaAspLysPhe

                                                  oUA 217
     gacaaagatgccaaaacttcttctgctactgatgctgaaggtattccagacaactggatg
1440 ---------+---------+---------+---------+---------+---------+   1499
     AspLysAspAlaLysThrSerSerAlaThrAspAlaGluGlyIleProAspAsnTrpMet
```

```
      ggtttggactgtggtccaaaatctgtcgaattgttccaacaagctgttgccaaagctaag
1500  ---------+---------+---------+---------+---------+---------+   1559
      GlyLeuAspCysGlyProLysSerValGluLeuPheGlnGlnAlaValAlaLysAlaLys

      accattgtttggaacggtccaccaggtgttttcgaatttgaaaaattcgccaacggtact
1560  ---------+---------+---------+---------+---------+---------+   1619
      ThrIleValTrpAsnGlyProProGlyValPheGluPheGluLysPheAlaAsnGlyThr

      aaatccttattggatgctgctgtcaaatctgctgaaaatggtaacattgttatcattggt
1620  ---------+---------+---------+---------+---------+---------+   1679
      LysSerLeuLeuAspAlaAlaValLysSerAlaGluAsnGlyAsnIleValIleIleGly

      ggtggtgatactgctactgttgctaagaaatacggtgtcgttgaaaaattatcgcacgtt
1680  ---------+---------+---------+---------+---------+---------+   1739
      GlyGlyAspThrAlaThrValAlaLysLysTyrGlyValValGluLysLeuSerHisVal

      tctactggtggtggtgcttcattggaattattagaaggtaaagacttgccaggtgtagtt
1740  ---------+---------+---------+---------+---------+---------+   1799
      SerThrGlyGlyGlyAlaSerLeuGluLeuLeuGluGlyLyaAspLeuProGlyValVal

                                                   oUA 205
          oUA 216                 oUA 207
      gctctttccaacaaaaaccatcaccatcaccatcactaaTCGCGAgctagttgaatatta
1800  ---------+---------+---------+---------+---------+---------+   1859
                                6 His           NruI
      AlaLeuSerAsnLysAsnHisHisHisHisHisHisEnd

      tgtaagatctgttagagtttttattttgtattcatttatttagtttattttcttattata
1860  ---------+---------+---------+---------+---------+---------+   1919

      tttgaaatataatatatatttaaaaaaaaaatttacagtgtagaatttttggtagtccg
1920  ---------+---------+---------+---------+---------+---------+   1979

      gtttgttttgaaatcagtggtggtattcaatatttgattaaattttggtatgaatttgtg
1980  ---------+---------+---------+---------+---------+---------+   2039

      ttgaaaaataaaaaataagcgagaaatttgcgtggcatattatttgtaatgttcgaatat
2040  ---------+---------+---------+---------+---------+---------+   2099

          oUA 204
      tctctcgtacaccaaaaggctgcagccaatt
2100  ---------+---------+---------+-                                2131
                          PstI
```

B.2 -: Count codon of the reporter protein

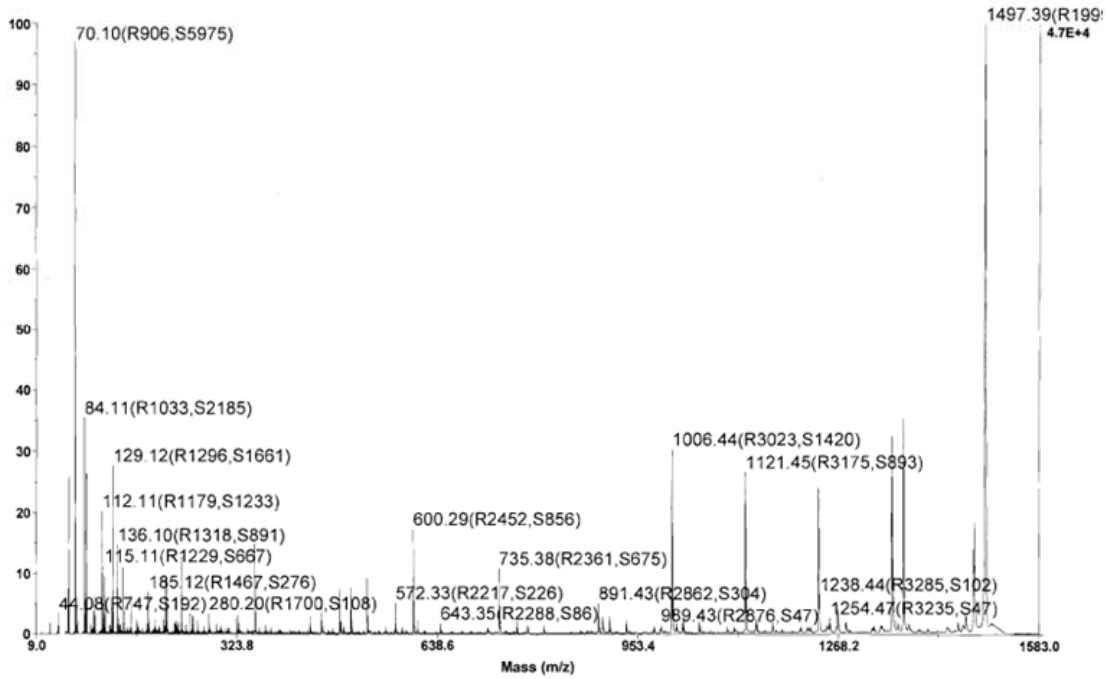
(440 codons)

fields: [triplet] [frequency: **per thousand**] ([number])

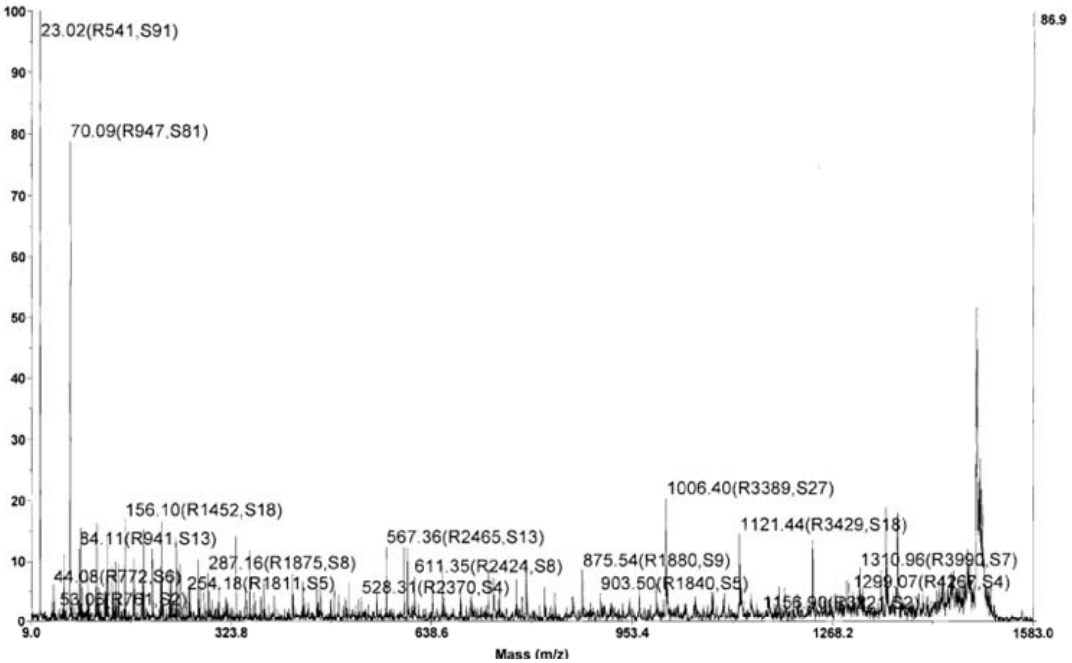
```
UUU 13.6(     6)  UCU 25.0(    11)  UAU  2.3(     1)  UGU  4.5(     2)
UUC 27.3(    12)  UCC 13.6(     6)  UAC 15.9(     7)  UGC  0.0(     0)
UUA 20.5(     9)  UCA 11.4(     5)  UAA  0.0(     0)  UGA  0.0(     0)
UUG 68.2(    30)  UCG  2.3(     1)  UAG  0.0(     0)  UGG  4.5(     2)

CUU  4.5(     2)  CCU  0.0(     0)  CAU  9.1(     4)  CGU  0.0(     0)
CUC  4.5(     2)  CCC  0.0(     0)  CAC 20.5(     9)  CGC  0.0(     0)
CUA  2.3(     1)  CCA 40.9(    18)  CAA 13.6(     6)  CGA  0.0(     0)
CUG  2.3(     1)  CCG  2.3(     1)  CAG  2.3(     1)  CGG  2.3(     1)

AUU 29.5(    13)  ACU 27.3(    12)  AAU  4.5(     2)  AGU  0.0(     0)
AUC 15.9(     7)  ACC 11.4(     5)  AAC 47.7(    21)  AGC  0.0(     0)
AUA  0.0(     0)  ACA  0.0(     0)  AAA 72.7(    32)  AGA 22.7(    10)
AUG 15.9(     7)  ACG  0.0(     0)  AAG 34.1(    15)  AGG  0.0(     0)

GUU 59.1(    26)  GCU 65.9(    29)  GAU 45.5(    20)  GGU 79.5(    35)
GUC 27.3(    12)  GCC 22.7(    10)  GAC 29.5(    13)  GGC  0.0(     0)
GUA  2.3(     1)  GCA  0.0(     0)  GAA 68.2(    30)  GGA  2.3(     1)
GUG  0.0(     0)  GCG  0.0(     0)  GAG  2.3(     1)  GGG  0.0(     0)
```

Annexe C: MS-MS of both synthetic and reporter peptides

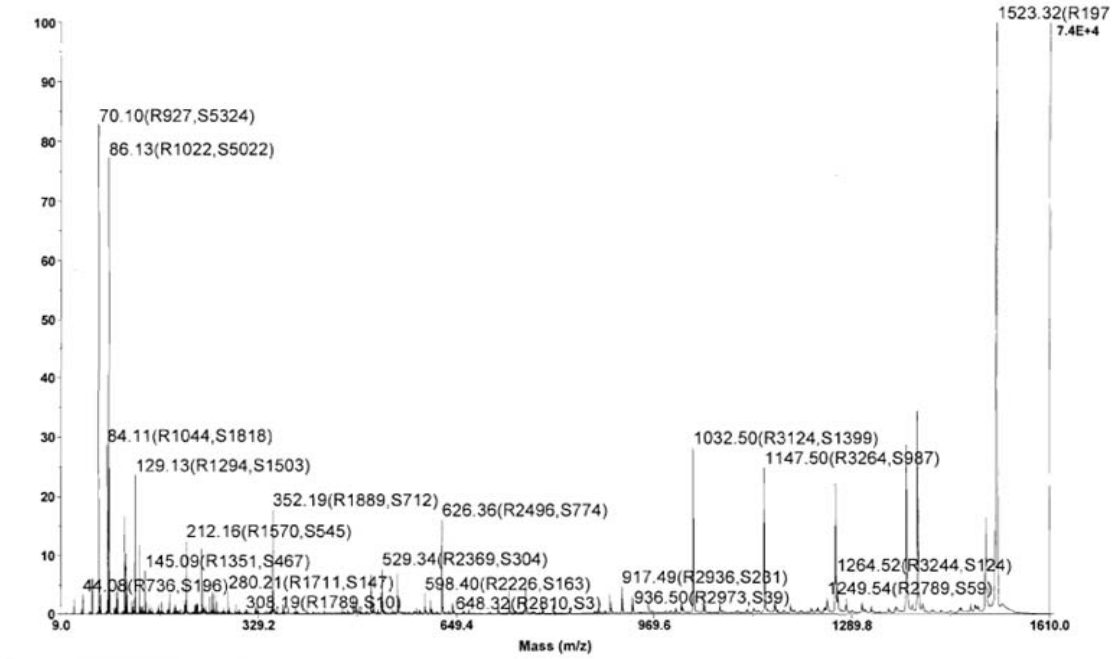
C1. MS-MS spectra of the synthetic serine peptide.



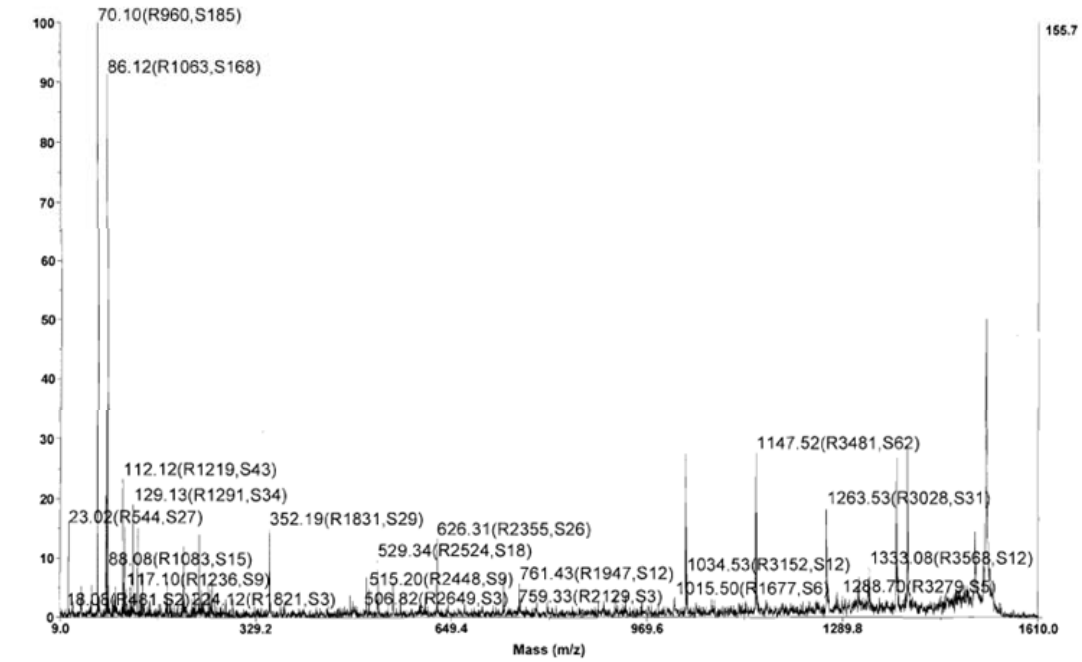
C.2- MS-MS spectra of the serine peptide purified from *C. albicans* cell extracts.



C3. - MS-MS spectra of the synthetic leucine peptide



C.4 - MS-MS spectra of the leucine peptide purified from *C. albicans* cell extracts.



Annexe D: Results from the clustering analysis

ORF (assembly 19)	*S. cerevisiae* gene	Function
Group 1		
orf19. 14137		
orf19. 11132	MMS4	putative transcriptional co-activator
orf19. 8637	WAR1	transcription factor activity
orf19. 5870	CTP1	citrate transport protein
orf19. 3153	MSS4	phosphatidylinositol 4-phosphate kinase
orf19. 10914	BUD17	involved in bud site selection
orf19. 332	CLF1	pre-mRNA splicing factor
orf19. 11718		
orf19. 11841	SGD1	suppressor of glycerol defect
orf19. 6891	RFC1	replication factor C subunit 1 | processivity factor for DNA polymerase delta and epsilon
orf19. 7792		
orf19. 11755	MNN1	mannosyltransferase

ORF (assembly 19)	*S. cerevisiae* gene	Function
Group 2		
orf19. 1144		
orf19. 6291	FUN30	helicase of the Snf2/Rad54 family
orf19. 1166		
orf19. 11485	DUN1	DNA damage response
orf19. 4192	CDC14	protein phosphatase required for mitosis
orf19. 5894		
orf19. 6866	SNP1	U1 small nuclear ribonucleoprotein
orf19. 814	SSY1.5	transcriptional regulator of multiple amino acid permeases
orf19. 1624	MAK10	glucose-repressible protein
orf19. 12155		
orf19. 801	TBF1	telomere TTAGGG repeat-binding factor
orf19. 12434	KEM1	multifunctional nuclease
orf19. 5584	PEP3	vacuolar membrane protein
orf19. 2921	PAC2	tubulin folding cofactor E
orf19. 6387	HSP104	heat shock protein 104
orf19. 9420		
orf19. 9541	SNX4	Sorting NeXin
orf19. 6233		
orf19. 3689		
orf19. 3458	VSP68	conserved protein involved in vacuolar targeting
orf19. 7781		
orf19. 2828	ALF1	alpha-tubulin foldin, cofactor B
orf19. 11319	EFB1	translation elongation factor eEF1beta | GDP/GTP exchange factor for Tef1p/Tef2p
orf19. 7838		
orf19. 7862	MED7	RNA polymerase II holoenzyme/mediator subunit
orf19. 8417	BAT2	branched-chain amino acid transaminase

orf19. 6773	ECM29	protein Involved in cell wall biogenesis and architecture

ORF (assembly 19)	*S. cerevisiae* gene	Function
Group 3		
orf19. 10092	CDC60	cytosolic leucyl tRNA synthetase
orf19. 1434	DPB11	DNA polymerase II complex component
orf19. 1217		
orf19. 8919	CPY1	serine carboxypeptidase Y precursor
orf19. 13164	ALS9-1	misassembled agglutinin-like protein 9
orf19. 11314	TFC3	RNA polymerase III transcription factor
orf19. 10309	BBC1	associates with the Bee1p-Vrp1p-Myo3/5p complex
orf19. 9129	SAC3	involved in processes affecting the actin cytoskeleton and mitosis | leucine permease transcriptional regulator
orf19. 10878	UBP10	ubiquitin-specific protease
orf19. 8304		
orf19. 13154	NOG2	nuclear/nucleolar GTP-binding protein 2
orf19. 8660		

ORF (assembly 19)	*S. cerevisiae* gene	Function
Group 4		
orf19. 4337	ESBP6	monocarboxylate permease
orf19. 12233	IPT1	necessary for synthesis of mannose-(inositol-P)2-ceramide (M(IP)2C) | inositolphosphotransferase 1 | mannosyl diphosphorylinositol ceramide synthase
orf19. 1693	TAO3	transcriptional activator
orf19. 12690	PKH2	ser/thr protein kinase, phosphorylates, activates YPK1
orf19. 2739	RLF2	chromatin assembly complex, subunit p90
orf19. 4958	EMC25	protein involved in cell wall biogenesis and architecture
orf19. 9556	REG1	RNA processing | negative regulator of glucose- repressible genes | regulatory subunit for protein phosphatase Glc7p
orf19. 10238	HIT1	required for growth at high temperature
orf19. 3956		
orf19. 13314	RIP1	component of ubiquinol cytochrome- c reductase complex
orf19. 1802		
orf19. 3705		
orf19. 11221	KAR4	transcription factor similar to pheromone-induced protein
orf19. 1991	PTM1	member of the major facilitator superfamily
orf19. 8671	RPN4	Regulatory Particle Non-ATPase
orf19. 1182	VAM7	vacuolar morphogenesis protein
orf19. 1886	RCL1	RNA 3 -terminal phosphate cyclase
orf19. 5378	SCL1	20S proteasome subunit YC7ALPHA/Y8
orf19. 12982	AMD3	putative amidase
orf19. 13612	NPR1	nitrogen permease reactivator protein
orf19. 1515	CHT4	chitinase
orf19. 5046	RAM1	protein farnesyltransferase, beta subunit
orf19. 12023	MAK32	necessary for structural stability of L-A dsRNA-containing particles
orf19. 9266	BZZ1	cortical patch protein involved in actin organization
orf19. 9088	FAB1	phosphatidylinositol 3-phosphate 5-kinase

orf19. 506	YDJ1	dnaJ homolog and heat shock protein
orf19. 8972		
orf19. 10005	TMT1	trans-aconitate methyltransferase 1
orf19. 1514	UBP1	ubiquitin-dependent protease
orf19. 8967	SGN1	poly(A) RNA binding protein
orf19. 3363	VTC4	polyphosphate synthetase
orf19. 1029	RPP1	nuclear ribonuclease P subunit (RNase P)| required for processing of tRNA and 35S rRNA
orf19. 5147		
orf19. 8088	UBP2	ubiquitin-specific protease
orf19. 5328	GCN1	translational activator of GCN4
orf19. 9953	MSE1	glutamyl-tRNA synthetase, mitochondrial

ORF (assembly 19)	*S. cerevisiae* gene	Function
Group 5		
orf19. 1557		
orf19. 10588	TVP15	conserved hypothetical protein
orf19. 11272	PAT1	topoisomerase II- associated protein
orf19. 7773	VPS15	vacuolar protein sorting protein kinase
orf19. 7787	YAK2	serine-threonine protein kinase, PKA suppressor
orf19. 290	KRE5	UDPglucose- glycoprotein glucose phosphotransferase
orf19. 2455		
orf19. 3141	SMY2	related to kinesins
orf19. 8377		
orf19. 8458		
orf19. 8870	MEC1	cell cycle checkpoint protein
orf19. 3724		
orf19. 12214	RSM25	mitochondrial ribosome small subunit component
orf19. 13020		
orf19. 8101	TAF12	TFIID and SAGA subunit
orf19. 8717	MTR10	involved in nuclear protein import
orf19. 1299	RPN6	proteasome regulatory particle subunit
orf19. 522	PIM1	mitochondrial ATP-dependent protease

ORF (assembly 19)	*S. cerevisiae* gene	Function
Group 6		
orf19. 9349	YOR1	oligomycin resistance ATP-dependent permease | ABC transporter
orf19. 3954	PSD2	phosphatidylserine decarboxylase
orf19. 11970	KTR2	mannosyltransferase
orf19. 4295	HIR2	histone transcription regulator
orf19. 1712	GEA2	GDP/GTP exchange factor for ARF
orf19. 11515	RPN5	non-ATPase unit of 26S proteasome complex
orf19. 1373	INP51	phosphatidylinositol phosphate 5-phosphatase
orf19. 4335	TNA1	high affinity nicotinic acid plasma membrane permease
orf19. 12208	JIP5	Jumonji Interacting Protein

orf19. 11421		
orf19. 1298	NUP84	nuclear pore complex subunit
orf19. 12199	DHA12	membrane transporter of the MFS-MDR family
orf19. 10350		
orf19. 8419	STE4	beta subunit of heterotrimeric G protein
orf19. 4858	VSP41	vacuolar protein sorting
orf19. 2665	MSN5	supressor of snf1 mutation
orf19. 5059	GSH1	gamma-glutamylcysteine synthetase
orf19. 2135	TAF2	component of TFIID complex
orf19. 13292	SNF5	component of SWI/SNF transcription activator complex
orf19. 313	DAL4	allantoin permease
orf19. 261	SEC59	dolichol kinase required for core glycosylation
orf19. 4403	VSP11	vacuolar peripheral membrane protein
orf19. 8678	ATM1	mitochondrial ABC transporter

ORF (assembly 19)	*S. cerevisiae* gene	Function
Group 7		
orf19. 4398		
orf19. 6240	CYK3	involved in CYtoKinesis
orf19. 6011	SIN3	transcription regulatory protein
orf19. 11823	SEC16	multidomain vesicle coat protein
orf19. 8539	THR1	homoserine kinase
orf19. 1229	CSE1	specific exportin for Srp1p
orf19. 5365		
orf19. 1238	TUB4	gamma tubulin
orf19. 11071	REC12	required for chromosome pairing
orf19. 4723	FAD1	flavin adenine dinucleotide (FAD) synthetase
orf19. 135	EXO84	exocyst complex component and pre-mRNA splicing factor
orf19. 12808	TPS3	alpha,alpha-trehalose-phosphate synthase, regulatory subunit
orf19. 5892	HUL4	ubiquitin-protein ligase
orf19. 10228	MSH5	meiosis-specific mutS homolog
orf19. 4753	FRK26	6-phosphofructose-2-kinase

ORF (assembly 19)	*S. cerevisiae* gene	Function
Group 8		
orf19. 262	SMC3	chromosome condensation and segregation protein
orf19. 2116	NAT2	N-acetyltransferase for N- terminal methionine
orf19. 12615	CDC35	adenylate cyclase
orf19. 2404	POP1	nuclear RNase P and RNase MRP component
orf19. 9430	MEK1	serine/threonine protein kinase
orf19. 2532	PRS3	prolyl-tRNA synthetase, cytoplasmic
orf19. 10369		
orf19. 11617		
orf19. 3996	GPI10	glycosyl phosphatidylinositol (GPI) synthesis

orf19.	10912	SED4	involved in vesicle formation at the endoplasmic reticulum
orf19.	8728	CKU70	Ku family DNA binding and repair protein
orf19.	5954	AMA1	activator of meiotic anaphase promoting complex
orf19.	1026	CSL4	exosome 3'->5exonuclease | involved in kinetochore-related function
orf19.	12983	WSC2	cell wall integrity, stress response
orf19.	8385	SCY1	conserved protein
orf19.	14146	NUP145	nucleoporin
orf19.	3556	KAP104	karyopherin beta 2 | transportin
orf19.	8347	TSC11	TOR binding protein
orf19.	567	TFB3	TFIIH subunit
orf19.	9896	URA2	multifunctional pyrimidine biosynthesis protein
orf19.	5526	SEC20	secretory pathway protein
orf19.	12110	PWP1	beta-transducin superfamily with periodic tryptophan residues
orf19.	2942	DIP52	dicarboxylic amino acid permease
orf19.	11964	SWI3	general RNA polymerase II transcription factor
orf19.	8292		
orf19.	11419	SDF1	Sporulation DeFiciency
orf19.	3722	FAP1	FKBP12-associated protein | transcription factor homolog
orf19.	5544	SAC6	actin filament bundling protein - fibrim homolog
orf19.	4937	CHS3	chitin-UDP acetyl-glucosaminyl transferase 3
orf19.	2859	SRP40	nonribosomal protein of the nucleolus and coiled bodies
orf19.	2733	VPS30	involved in vacuolar protein sorting and autophagy
orf19.	7748	RIM9	low similarity to a regulator of sporulation
orf19.	4867	SWE1	serine/tyrosine dual-specificity protein kinase that inhibits G2/M transition
orf19.	2029	RFC5	DNA replicationn factor C | leading strand elongation mismatch repair (ATPase)
orf19.	6538	TFP3	hydrogen-transporting ATPase
orf19.	1390	PMI1	mannose-6-phosphate isomerase
orf19.	4426	PEX3	peroxisomal integral membrane protein
orf19.	706	NMD3	nonsense mRNA degradation; ribosomal assembly
orf19.	795	VSP36	defective in vacuolar protein sorting | regulator of G-protein signaling activity
orf19.	1526	SNF2	component of SWI/SNF global transcription activator complex
orf19.	12523	APC10	anaphase promoting complex component
orf19.	5535	FEN2	member of allantoate permease family

Annexe E: Leucyl – tRNA synthetase

E.1 - Sequence of the Leucy-tRNA synthetase in the genebank

```
LOCUS     AF293346    3987 bp    DNA        PLN     20-AUG-2000
DEFINITION  Candida albicans cytosolic leucyl-tRNA synthetase (CDC60)
gene,
ACCESSION   AF293346
VERSION     AF293346.1  GI:9858189
KEYWORDS    .
SOURCE      Candida albicans.
ORGANISM    Candida albicans
            Eukaryota; Fungi; Ascomycota; Saccharomycotina;
            Saccharomycetes; Saccharomycetales; mitosporic
            Saccharomycetales; Candida.
REFERENCE   1  (bases 1 to 3987)
  AUTHORS   O'Sullivan,J.M., Mihr,M.J. and Tuite,M.F.
  TITLE     Candida albicans leucyl-tRNA synthetase
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 3987)
  AUTHORS   O'Sullivan,J.M., Mihr,M.J. and Tuite,M.F.
  TITLE     Direct Submission
  JOURNAL   Submitted (03-AUG-2000) Department of Biosciences, University
            of Kent, Giles Lane, Canterbury, Kent CT2 7NJ, UK
FEATURES             Location/Qualifiers
     source          1..3987
                     /organism="Candida albicans"
                     /strain="2005E"
                     /db_xref="taxon:5476"
     mRNA            <617..>3910
                     /gene="CDC60"
                     /product="cytosolic leucyl-tRNA synthetase"
     gene            <617..>3910
                     /gene="CDC60"
     CDS             617..3910
                     /gene="CDC60"
                     /codon_start=1
                     /transl_table=12
                     /product="cytosolic leucyl-tRNA synthetase"
                     /protein_id="AAG01037.1"
                     /db_xref="GI:9858190"
                     /translation="MSGPVTFEKTFRRDALIDIEKKYQKVWAEEKVFEVDAPTFE
      ECPIEDVEQVQEAHPKFFATMAYPYMNGVLHAGHAFTLSKVEFATGFQRMNGKRALFPLGFHCTGMPIK
      AAADKIKREVELFGSDFSKAPIDDEDAEESQQPAKTETKREDVTKFSSKKSKAAAKQGRAKFQYEIMMQ
      LGIPREEVAKFANTDYWLEFFPPLCQKDVTAFGARVDWRRSMITTDANPYYDAFVRWQINRLRDVGKIK
      FGERYTIYSEKDGQACLDHDRQSGEGVGPQEYVGIKIRLTDVAPQAQELFKKENLDVKENKVYLVAATL
      RPETMYGQTCCFVSPKIDYGVFDAGNGDYFITTERAFKNMSFQNLTPKRGYYKPLFTINGKTLIGSRID
      APYAVNKNLRVLPMETVLATKGTGVVTCVPSDSPDDFVTTRDLANKPEYYGIEKDWVQTDIVPIVHTEK
      YGDKCAEFLVNDLKIQSPKDSVQLANAKELAYKEGFYNGTMLIGKYKGDKVEDAKPKVKQDLIDEGLAF
      VYNEPESQVISRSGDDCCVSLEDQWYIDYGEEVWLGEALECLKNMETYSKETRHGFEGVLAWMKNWAVT
      RKFGLGTKLPWDPQYLVESLSDSTVYMAYYTIDRFLHSDYYGKKAGKFDIKPEQMTDEVFDYIFTRRDD
      VETDIPKEQLKEMRREFEYFHPLDVRVSGKDLIPNHLTFFIYTHVALFPKRFWPRGVRANGHLLLNNAK
      MSKSTGNFMTLEQIIEKFGADASRIAMADAGDTVEDANFDEANANAAILRLTTLKDWCEEEVKNQDKLR
      IGDYDSFFDAAFENEMNDLIEKTYQQYTLSNYKQALKSGLFDFQIARDIYRESVNTTGIGMHKDLVLKY
      IEYQALMLAPIAPHFAEYLYREVLGKNGSVQTSKFPRASKPVSKAILDASEYVRSLTRSIREAEGQALK
      KKKGKSDVDGSKPISLTVLVSNTFPEWQDNYIELVRELFEQNKLDDNNVIRQKVGKDMKRGMPYIHQIK
      TRLATEDADTVFNRKLTFDEIDTLKNVVEIVKNAPYSLKVEKLEILSFNNGETKGKNIISGEDNIELNF
      KGKIMENAVPGEPGIFIKNVE"
```

```
BASE COUNT     1348 a    635 c    830 g   1174 t
ORIGIN
        1 ttatacggca gaagattttg atggtgtgaa aacaattggg ctttgtggat tggacccttc
       61 attaaatttc cagcaagtga agacaatttt tgaggaaagg tttgggaagg ttgctaaagt
      121 tttgttgttc ccagaagaca aacaagcttt ggtagagttt gtgaaacctg gagatgctgg
      181 caagacgagt atgagtaata attttgtgaa actaggcgaa agcgaagcaa aaatagtgac
      241 caaggaagag ataacaatgg gtaaaagtgc tatttcaaat acaactacaa ctccatcggc
      301 acctcttaca atgattccca ccacagttag acgaaaaaaa tccaaaaaat agaccaaaga
      361 tgtaattagt gattagtcac ttaacaaccc taaatagttt tgaaacctcc cgtaatagtt
      421 attctaattg tctcgttagt gcgagtagga gtgcctcacg taatataaat tgttgtgata
      481 caatcaagtt ctcttaataa aaaaaaaata cagagagcga gagagtttgt gtgagagtaa
      541 gaaaaagaaa atttttcact tcttgagtca ttcttgtaac cataatccac ttttgtttcc
      601 aacaaactat aaaatcatga gtggtcctgt tacttttgaa aagacatttc gtagagatgc
      661 cttaatcgat atagaaaaga aatatcaaaa ggtatgggca gaagagaaag tttttgaagt
      721 tgatgcccca acttttgaag aatgtcctat tgaagatgtt gaacaagttc aagaagcaca
      781 tccaaaattc tttgccacta tggcttatcc ttacatgaat ggtgtcttgc acgccggtca
      841 tgcctttaca ttgtctaaag ttgaatttgc aactgggttc caaagaatga atggtaagag
      901 agcattattc ccattgggtt tccattgtac gggtatgcca attaaagcag ctgccgataa
      961 aatcaaaaga gaagttgaat tgtttggatc tgatttttct aaagctccta ttgatgacga
     1021 agatgcagaa gaaagccaac aaccagctaa aaccgaaact aaaagagaag atgtcaccaa
     1081 attctcttcc aaaaaatcca aggctgctgc caaacaaggt agagccaagt tccaatatga
     1141 gatcatgatg caattaggaa tcccaagaga agaagttgcc aagtttgcta acaccgacta
     1201 ctggttagag tttttcccac cattgtgtca aaaagatgta actgcttttg gggctagagt
     1261 tgattggaga cgttctatga tcacaaccga tgctaatcct tattatgatg catttgttag
     1321 atggcaaatt aatagattga gagatgttgg taaaattaag tttggtgaaa gatataccat
     1381 ttattctgaa aaggatggcc aagcatgttt ggatcacgat agacaatctg gtgaaggtgt
     1441 tggtccacaa gaatatgttg gtataaaaat cagattaact gatgtagcac cacaagcaca
     1501 agaacttttc aagaaagaga atctcgatgt gaaggagaac aaagtttact tggttgctgc
     1561 aactttaaga ccagaaacta tgtatggtca aacttgttgt tttgtgagtc caaaaattga
     1621 ttatggtgtt tttgatgctg gtaatggtga ctatttcatt accactgaac gtgctttcaa
     1681 aaatatgtct ttccaaaact tgactccgaa aagaggatat tataaaccac ttttcactat
     1741 caatggtaag acattgattg gatctcgaat tgatgctcca tatgctgtca acaaaaactt
     1801 gagagttttg cctatggaaa cagttcttgc aaccaaaggt actggtgtgg tcacttgtgt
     1861 tccatcagat tctccagatg attttgttac cacaagagac ttggccaata aaccagagta
     1921 ctatggaatt gaaaaagact gggtacaaac agatattgtt cctattgtcc ataccgaaaa
     1981 atacggtgat aagtgtgctg agtttttggt taatgatttg aagatacagt caccaaaaga
     2041 ttctgtgcag ttggccaacg ccaaggaatt ggcttataaa gaaggttttt acaatggtac
     2101 tatgcttatt ggtaaataca aaggtgataa agttgaagac gccaagccta aagtcaaaca
     2161 agacttaatt gatgaaggtc ttgcttttgt ttacaatgaa ccagaatccc aagttatttc
     2221 tagatctggt gatgattgtt gtgtatcatt ggaagatcaa tggtatattg attatggtga
     2281 agaagtttgg ttgggtgaag ccttagaatg tcttaagaac atggaaacat actccaagga
     2341 aaccagacat ggcttcgaag gtgttttagc ttggatgaag aactgggctg tcaccagaaa
     2401 atttggtttg ggtactaaat tgccttggga tcctcaatat ttggtcgaat ctttgtcaga
     2461 ttctactgtc tatatggctt attatactat tgatcgtttc ttgcattcag attattacgg
     2521 taagaaggca ggtaagttcg acattaagcc agagcaaatg actgatgaag tatttgatta
     2581 catctttact cgtcgtgatg acgttgaaac tgacattcca aaggaacaat tgaaggaaat
     2641 gagaagagag tttgaatatt ttcacccatt agacgtcaga gtttcaggaa aagatttaat
     2701 cccaaatcat ttgacattct tcatctatac ccatgtcgcc ttgttcccaa aaagattttg
     2761 gccaagaggt gttagagcca acggacattt gttgttgaac aatgctaaga tgtccaaatc
     2821 aactggtaac tttatgactt tagaacaaat cattgaaaaa ttcggagctg atgcctctag
     2881 aattgctatg gccgatgcag gtgacactgt tgaagatgcc aactttgacg aagccaatgc
     2941 taatgctgca atcttgagat tgacaacttt gaaagattgg tgtgaagaag aagtgaaaaa
     3001 ccaagacaag ttaagaattg gtgactacga ttccttcttt gacgctgctt ttgaaaatga
     3061 aatgaatgat ttgattgaaa agacttacca acaatacact ttgagtaatt acaaacaagc
     3121 attgaaatcg ggattgtttg atttccaaat cgccagagat atttatagag aaagtgtaaa
     3181 cacaacaggg attggtatgc acaaggatct tgttttgaaa tacattgaat accaagcatt
     3241 gatgttagct ccaattgctc ctcattttgc cgaatacctt tacagagaag ttttaggtaa
```

```
3301 aaatggaagt gttcaaacta gcaagttccc aagagcctca aagcctgttt ccaaagctat
3361 tcttgatgct ctggaatatg tcagaagcct taccagatct atccgtgaag cagaaggtca
3421 agctttgaaa aagaagaaag gaaagtctga tgttgatggg tcaaaaccaa tcagcttgac
3481 agttttggtt tccaacactt tcccagaatg gcaagataac tatattgaac ttgtcagaga
3541 attgtttgaa caaaacaagt tggacgacaa taatgttata agacaaaagg ttggcaagga
3601 catgaaacgt ggtatgccat acatccacca aattaaaact agattggcaa ctgaagatgc
3661 tgacactgtt ttcaacagaa aattgacttt tgatgaaatc gatacattga aaaatgttgt
3721 tgaaattgtc aagaatgccc catactctct taaagttgaa aaattggaga ttcttagttt
3781 caataacggt gaaactaagg ggaagaatat tattagtggt gaagacaata ttgagctcaa
3841 tttcaagggt aaaataatgg aaaatgctgt acctggtgag cctggtatct ttattaaaaa
3901 tgtcgaataa atagagtctt gtttaggttg cttttaatac ataacttttt gtttagagat
3961 atcaataata ctatgagccc tggcttt
//
```

Annexe F: Seryl – tRNA synthetase

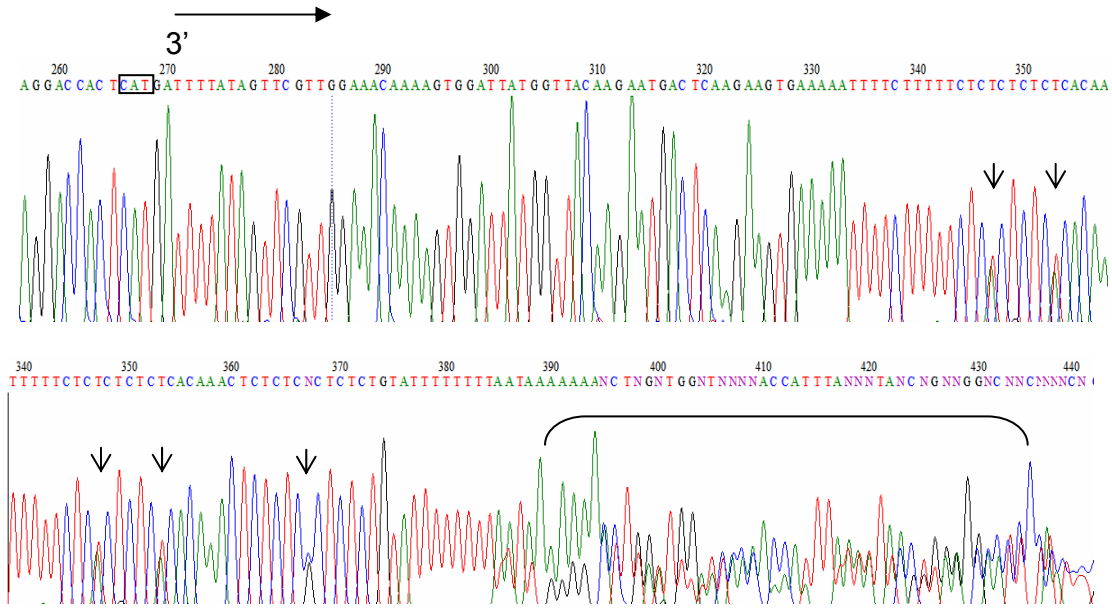
F.1 - Sequence of the Seryl-tRNA synthetase in the genebank

```
LOCUS       AF290915     3098 bp    DNA             PLN       19-MAR-2001
DEFINITION  Candida albicans seryl-tRNA synthetase (SES1) gene, complete
cds.
ACCESSION   AF290915
VERSION     AF290915.1  GI:9931531
KEYWORDS    .
SOURCE      Candida albicans.
  ORGANISM  Candida albicans
            Eukaryota; Fungi; Ascomycota; Saccharomycotina;
Saccharomycetes;
            Saccharomycetales; mitosporic Saccharomycetales; Candida.
REFERENCE   1  (bases 1 to 3098)
  AUTHORS   O'Sullivan,J.M., Mihr,M.J., Santos,M.A.S. and Tuite,M.F.
  TITLE     Seryl-tRNA synthetase is not responsible for the evolution of
CUG
            codon reassignment in Candida albicans
  JOURNAL   Yeast 18 (4), 313-322 (2001)
   PUBMED   11223940
REFERENCE   2  (bases 1 to 3098)
  AUTHORS   O'Sullivan,J.M., Mihr,M.J. and Tuite,M.F.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-JUL-2000) Biosciences, University of Kent,
Giles
            Lane, Canterbury, Kent CT2 7NJ, UK
FEATURES             Location/Qualifiers
     source          1..3098
                     /organism="Candida albicans"
                     /strain="CBS 5736"
                     /db_xref="taxon:5476"
                     /chromosome="3"
     mRNA            1841
                     /gene="SES1"
                     /product="seryl-tRNA synthetase"
     gene            1841
                     /gene="SES1"
     CDS             453..1841
                     /gene="SES1"
                     /note="aminoacyl-tRNA synthetase"
                     /codon_start=1
                     /transl_table=12
                     /product="seryl-tRNA synthetase"
                     /protein_id="AAG02209.1"
                     /db_xref="GI:9931532"
                     /translation="MLDINAFLVEKGGDPEIIKASQKKRGDSVELVDEIIAE
      YKEWVKLRFDLDEHNKKLNSVQKEIGKRFKAKEDAKDLIAEKEKLSNEKKEIIEKEAEADKNLRSKI
      NQVGNIVHESVVDSQDEENNELVRTWTPENYKKPEQIAAATGAPAKLSHHEVLLRLDGYDPERGVRI
      VGHRGYFLRNYGVFLNQALINYGLSFLSSKGYVPLQAPVMMNKEVMAKTAQLSQFDEELYKVIDGED
      EKYLIATSEQPISAYHAGEWFESPAEQLPVRYAGYSSCFRREAGSHGKDAWGIFRVHAFEKIEQFVL
      TEPEKSWEEFDRMIGCSEEFYQSLGLPYRVVGIVSGELNNAAAKKYDLEAWFPFQQEYKELVSCSNC
      TDYQSRNLEIRCGIKQQNQQEKKYVHCLNSTLSATERTICCILENYQKEDGLVIPEVLRKYIPGEPE
      FIPYIKELPKNTTSVKKAKGKN"
```

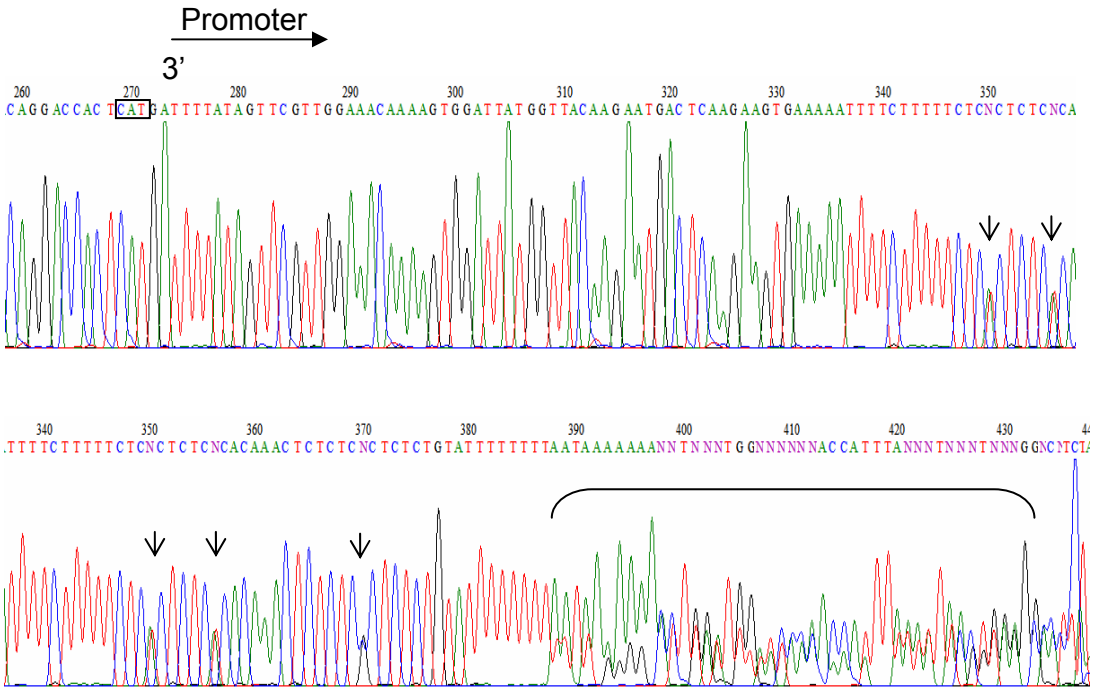
```
BASE COUNT     1107 a    515 c    545 g    931 t
ORIGIN
        1 atagctgttt cctacatata aaccattcct aaggaaatgg ttgtcgcact ttgtcgcact
       61 ttgtctcttt gtttgttaat cgaattgaat tgaatgaaaa tagtgaaaaa aaaaaaaaat
      121 tacaggtcgt aaagaataga aaaatttttt tgttccacgt aataatcacc atacaaattt
      181 aaaccaaacc caccaccaca acccccctaag ttacattcta gatacatagc tgtttcctac
      241 atataaacca ttcctaagga aatggttgtc agcactttgt cgcactttgt ctctttgttt
      301 gttaatcgaa ttgaattgaa tgaaaatagt gaaaaaaaaa aaaaattaca ggtcgtaaag
      361 aatagaaaaa tttttttgtt ccacgtaata atcaccatac aaatttaaac caaacccacc
      421 accacaaccc cctaagttac attctagata ccatgttaga cattaatgca tttctcgttg
      481 aaaagggagg tgacccagaa attattaaag catcccaaaa gaaaagaggt gactccgtcg
      541 aattagttga tgaaatcatc gccgaatata aagaatgggt taaattaaga ttcgatttag
      601 atgaacacaa caagaaattg aattcagtac aaaaagaaat tggtaaaaga ttcaaagcta
      661 aagaagatgc taaagattta attgctgaaa aggaaaaatt gagtaatgaa aaaaaggaaa
      721 ttattgaaaa agaagctgaa gcagataaga atttacgtag taaaatcaat caagttggta
      781 acatcgttca tgaatcagtt gttgattctc aagatgaaga aaacaatgaa ttggttagaa
      841 cctggactcc agaaaattac aaaaaaccag aacaaattgc tgcagctact ggtgcaccag
      901 ccaaattatc tcatcatgaa gtattgttaa gattagatgg ttacgatcca gaaagagggg
      961 ttagaattgt tggtcatcgt ggttatttct taagaaacta tggggtattt ttgaaccaag
     1021 ctttaatcaa ctacggttta ctgtttttga gtagcaaagg atacgttcca ttgcaagcac
     1081 cagttatgat gaataaagaa gtcatggcta aaaccgcaca attgtctcaa tttgacgaag
     1141 aattgtataa agtcattgat ggtgaagatg aaaaatattt aattgccact tcagaacaac
     1201 caattagtgc ttaccatgcc ggtgaatggt ttgaatcacc agcagaacaa ttgccagttc
     1261 gttatgctgg ttattcatca tgtttcagaa gagaagctgg atcacacggt aaagatgctt
     1321 ggggtatttt ccgtgtccat gcttttgaaa agattgaaca atttgttttg actgaaccag
     1381 aaaaatcatg ggaagaattt gatagaatga ttggatgttc agaagaattt tatcaatcat
     1441 taggattgcc atacagagtt gttggtattg tttcaggtga attgaacaat gctgctgcta
     1501 agaaatacga tttggaagct tggttcccat tccaacaaga atacaaagaa ttggtttcat
     1561 gttcaaactg tactgattat caatcaagaa atttggaaat cagatgtggt ataaaacaac
     1621 aaaaccaaca agaaaagaaa tacgtccatt gtttgaactc aactttaagt gctactgaaa
     1681 gaactatctg ttgtatttta gaaaactacc aaaaggaaga tgggttggtt attcctgaag
     1741 tattgagaaa atacattcct ggtgaaccag aatttattcc atacattaag gaattgccaa
     1801 aaaacaccac ttctgttaaa aaagctaaag gtaagaatta gatgtttata gtgcatgtta
     1861 tactccattt tattaaaaca ttataatagt atgtcatttt ctttttatct ttgtattctt
     1921 aaaaactggt gatatgtatc agaaaaagga aaatcacacg acacgtcatg aatggatgga
     1981 tggtgcacag ccttgccttg tatgcaatac tgacatcatc gccacttagt gctactacca
     2041 ccgaccccac cacaaccaac aaaaccgcat tgacgacccg aaattcaaac gaaaacgtag
     2101 gcaaaacatc aacaactcaa tcgaagaagg aaaaaaaaaa taaccttaaa aatatatttt
     2161 gcaaaacaaa tatccacatt atccattatt cgatagctga atatttgttt tcattcgagc
     2221 ttccactacc acgtttattt aaatatacta tctagcaagc catgtcaaat atggaatcat
     2281 cccatgtgaa taatgtggaa tcaccaccag aatatgtatc tcaaccacca ccaaaatatg
     2341 tacctcgaca gtcatcatcg tcgtcatctt caatatcaga tcaggaatca gatattcaca
     2401 atccaccaca aagggccagt gaaaatcaat tatcgacttg ttgttctgat tgttggtgta
     2461 attgttttga tagttgttct ggcactaatt gtactgcttc cgataagaat atttgtggca
     2521 gtatactagt tgttttgtgt tgtggaagca caattgggta tgcccactaa tgcttgaata
     2581 acagctttat tgtcccgttg tcaaccatag aattattaat atctattcat tcaaatgtat
     2641 atagtattgt ggggtttgtt aggatatgtt cttttaatga attaatggtt tttatgttat
     2701 ccctctgtaa gttataggaa atcttctgat tcgaattatt cagtactgtg agcaatacgt
     2761 atatagcgaa gtcgcattaa agtgccgaga cactagaagc agtaaaaatt tgattgctac
     2821 taaaatacag catgacatag ttaacacttt tagtgtatac cattgttaac ctgaaacgat
     2881 cctacattag agctctacaa ctgattggct tctttagttt tctattgttt tgtaaattgc
     2941 aattggggga gaggtcccgg tccagctcac aagaaacaat agtccattgt ttttctgggg
     3001 aacataaaat cgcgggagct tcctaattaa gtttataccc gaaagaaata taaggaatta
     3061 aagctgatat gcaggtattg ctaacactaa taagattt
```

Annexe G: Sequencing of the promoter regions of Leucyl-tRNA synthetase

G.1 – *C. albicans* CAI4 strain



G.2 – *C. albicans* IGC strain



The sequence polymorphisms found on these strains are indicated by the arrows.

8. References

Abaitua,F., Rementeria,A., San Millan,R., Eguzkiza,A., Rodriguez,J.A., Ponton,J., and Sevilla,M.J. (1999). In vitro survival and germination of Candida albicans in the presence of nitrogen compounds. Microbiology *145 (Pt 7)*, 1641-1647.

Achsel,T. and Gross,H.J. (1993). Identity determinants of human tRNA(Ser): sequence elements necessary for serylation and maturation of a tRNA with a long extra arm. EMBO J. *12*, 3333-3338.

Agris,P.F. (2004). Decoding the genome: a modified view. Nucleic Acids Res. *32*, 223-238.

Alf-Steinberger,C. (1969). The genetic code and error transmission. Proc. Natl. Acad. Sci. U. S. A *64*, 584-591.

Alvarez,F., Robello,C., and Vignali,M. (1994). Evolution of codon usage and base contents in kinetoplastid protozoans. Mol. Biol. Evol. *11*, 790-802.

Ambrogelly,A., Gundllapalli,S., Herring,S., Polycarpo,C., Frauer,C., and Soll,D. (2007). Pyrrolysine is not hardwired for cotranslational insertion at UAG codons. Proc. Natl. Acad. Sci. U. S. A *104*, 3141-3146.

Anderson,J.C., Wu,N., Santoro,S.W., Lakshman,V., King,D.S., and Schultz,P.G. (2004). An expanded genetic code with a functional quadruplet codon. Proc. Natl. Acad. Sci. U. S. A *101*, 7566-7571.

Arnez,J.G., Harris,D.C., Mitschler,A., Rees,B., Francklyn,C.S., and Moras,D. (1995). Crystal structure of histidyl-tRNA synthetase from Escherichia coli complexed with histidyl-adenylate. EMBO J. *14*, 4143-4155.

Arnez,J.G. and Moras,D. (1997). Structural and functional considerations of the aminoacylation reaction. Trends Biochem. Sci. *22*, 211-216.

Arnold,K., Bordoli,L., Kopp,J., and Schwede,T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics. *22*, 195-201.

Asahara,H., Himeno,H., Tamura,K., Hasegawa,T., Watanabe,K., and Shimizu,M. (1993). Recognition nucleotides of Escherichia coli tRNA(Leu) and its elements facilitating discrimination from tRNASer and tRNA(Tyr). J. Mol. Biol. *231*, 219-229.

Asahara,H., Himeno,H., Tamura,K., Nameki,N., Hasegawa,T., and Shimizu,M. (1994). Escherichia coli seryl-tRNA synthetase recognizes tRNA(Ser) by its characteristic tertiary structure. J. Mol. Biol. *236*, 738-748.

Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., Harris,M.A., Hill,D.P., Issel-Tarver,L., Kasarskis,A., Lewis,S., Matese,J.C., Richardson,J.E., Ringwald,M., Rubin,G.M., and Sherlock,G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25-29.

Ashraf,S.S., Ansari,G., Guenther,R., Sochacka,E., Malkiewicz,A., and Agris,P.F. (1999). The uridine in "U-turn": contributions to tRNA-ribosomal binding. RNA. *5*, 503-511.

Auffinger,P. and Westhof,E. (1998). Location and Distribution of Modified Nucleotides in tRNA. In Modification and Editing of RNA, H.Grosjean and R.Benne, eds. (Washington: American Society for Microbiology).

Bacher,J.M., Bull,J.J., and Ellington,A.D. (2003). Evolution of phage with chemically ambiguous proteomes. BMC. Evol. Biol. *3*, 24.

Bacher,J.M. and Ellington,A.D. (2001). Selection and characterization of Escherichia coli variants capable of growth on an otherwise toxic tryptophan analogue. J. Bacteriol. *183*, 5414-5425.

Balashov,S. and Humayun,M.Z. (2002). Mistranslation induced by streptomycin provokes a RecABC/RuvABC-dependent mutator phenotype in Escherichia coli cells. J. Mol. Biol. *315*, 513-527.

Barak,Z., Gallant,J., Lindsley,D., Kwieciszewki,B., and Heidel,D. (1996). Enhanced ribosome frameshifting in stationary phase cells. J. Mol. Biol. *263*, 140-148.

Baranov,P.V., Gesteland,R.F., and Atkins,J.F. (2004). P-site tRNA is a crucial initiator of ribosomal frameshifting. RNA. *10*, 221-230.

Barrell,B.G., Bankier,A.T., and Drouin,J. (1979). A different genetic code in human mitochondria. Nature *282*, 189-194.

Barton,R.C., van,B.A., and Scherer,S. (1995). Stability of karyotype in serial isolates of Candida albicans from neutropenic patients. J. Clin. Microbiol. *33*, 794-796.

Becker,H.D. and Kern,D. (1998) Thermus thermophilus: A link in evolution of the tRNA-dependent amino acid amidation pathways. Proc. Natl. Acad. Sci. U. S. A *95*: 12832-12837.

Becker,H.D., Roy,H., Moulinier,L., Mazauric,M.H., Keith,G., and Kern,D. (2000). Thermus thermophilus contains an eubacterial and an archaebacterial aspartyl-tRNA synthetase. Biochemistry *39*, 3216-3230.

Beebe,K., Ribas,d.P., and Schimmel,P. (2003). Elucidation of tRNA-dependent editing by a class II tRNA synthetase and significance for cell viability. EMBO J. *22*, 668-675.

Bein,M., Schaller,M., and Korting,H.C. (2002). The secreted aspartic proteinases as a new target in the therapy of candidiasis. Curr. Drug Targets. *3*, 351-357.

Bekaert,M. and Rousset,J.P. (2005). An extended signal involved in eukaryotic -1 frameshifting operates through modification of the E site tRNA. Mol. Cell *17*, 61-68.

Belrhali,H., Yaremchuk,A., Tukalo,M., Larsen,K., Berthet-Colominas,C., Leberman,R., Beijer,B., Sproat,B., ls-Nielsen,J., Grubel,G., and . (1994). Crystal structures at 2.5 angstrom resolution of seryl-tRNA synthetase complexed with two analogs of seryl adenylate. Science *263*, 1432-1436.

Bennett,R.J. and Johnson,A.D. (2003). Completion of a parasexual cycle in Candida albicans by induced chromosome loss in tetraploid strains. EMBO J. *22*, 2505-2515.

Bennett,R.J. and Johnson,A.D. (2005). Mating in Candida albicans and the search for a sexual cycle. Annu. Rev. Microbiol. *59*, 233-255.

Berg,O.G. and Silva,P.J. (1997). Codon bias in Escherichia coli: the influence of codon context on mutation and selection. Nucleic Acids Res. *25*, 1397-1404.

Berman,J. and Sudbery,P.E. (2002). Candida Albicans: a molecular revolution built on lessons from budding yeast. Nat. Rev. Genet. *3*, 918-930.

Bernales,S., Papa,F.R., and Walter,P. (2006). Intracellular signaling by the unfolded protein response. Annu. Rev. Cell Dev. Biol. *22*, 487-508.

Berry,M.J., Banu,L., Chen,Y.Y., Mandel,S.J., Kieffer,J.D., Harney,J.W., and Larsen,P.R. (1991). Recognition of UGA as a selenocysteine codon in type I deiodinase requires sequences in the 3' untranslated region. Nature *353*, 273-276.

Bertram,G., Innes,S., Minella,O., Richardson,J., and Stansfield,I. (2001). Endless possibilities: translation termination and stop codon recognition. Microbiology *147*, 255-269.

Beuning,P.J. and Musier-Forsyth,K. (2000). Hydrolytic editing by a class II aminoacyl-tRNA synthetase. Proc. Natl. Acad. Sci. U. S. A *97*, 8916-8920.

Beuning,P.J. and Musier-Forsyth,K. (2001). Species-specific differences in amino acid editing by class II prolyl-tRNA synthetase. J. Biol. Chem. *276*, 30779-30785.

Bilokapic,S., Korencic,D., Soll,D., and Weygand-Durasevic,I. (2004). The unusual methanogenic seryl-tRNA synthetase recognizes tRNASer species from all three kingdoms of life. Eur. J. Biochem. *271*, 694-702.

Bjork,G.R. (1995). Genetic dissection of synthesis and function of modified nucleosides in bacterial transfer RNA. Prog. Nucleic Acid Res. Mol. Biol. *50*, 263-338.

Bjork,G.R., Durand,J.M., Hagervall,T.G., Leipuviene,R., Lundgren,H.K., Nilsson,K., Chen,P., Qian,Q., and Urbonavicius,J. (1999). Transfer RNA modification: influence on translational frameshifting and metabolism. FEBS Lett. *452*, 47-51.

Blalock,J.E. and Smith,E.M. (1984). Hydropathic anti-complementarity of amino acids based on the genetic code. Biochem. Biophys. Res. Commun. *121*, 203-207.

Blight,S.K., Larue,R.C., Mahapatra,A., Longstaff,D.G., Chang,E., Zhao,G., Kang,P.T., Green-Church,K.B., Chan,M.K., and Krzycki,J.A. (2004). Direct charging of tRNA(CUA) with pyrrolysine in vitro and in vivo. Nature *431*, 333-335.

Blom,N., Gammeltoft,S., and Brunak,S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J. Mol. Biol. *294*, 1351-1362.

Bossi,L. and Ruth,J.R. (1980). The influence of codon context on genetic code translation. Nature *286*, 123-127.

Braun,B.R., van Het,H.M., d'Enfert,C., Martchenko,M., Dungan,J., Kuo,A., Inglis,D.O., Uhl,M.A., Hogues,H., Berriman,M., Lorenz,M., Levitin,A., Oberholzer,U., Bachewich,C., Harcus,D., Marcil,A., Dignard,D., Iouk,T., Zito,R., Frangeul,L., Tekaia,F., Rutherford,K., Wang,E., Munro,C.A., Bates,S., Gow,N.A., Hoyer,L.L., Kohler,G.,

Morschhauser,J., Newport,G., Znaidi,S., Raymond,M., Turcotte,B., Sherlock,G., Costanzo,M., Ihmels,J., Berman,J., Sanglard,D., Agabian,N., Mitchell,A.P., Johnson,A.D., Whiteway,M., and Nantel,A. (2005). A human-curated annotation of the Candida albicans genome. PLoS. Genet. *1*, 36-57.

Breitschopf,K., Achsel,T., Busch,K., and Gross,H.J. (1995). Identity elements of human tRNA(Leu): structural requirements for converting human tRNA(Ser) into a leucine acceptor in vitro. Nucleic Acids Res. *23*, 3633-3637.

Breitschopf,K. and Gross,H.J. (1994). The exchange of the discriminator base A73 for G is alone sufficient to convert human tRNA(Leu) into a serine-acceptor in vitro. EMBO J. *13*, 3166-3169.

Brown,J.R. and Doolittle,W.F. (1995). Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc. Natl. Acad. Sci. U. S. A *92*, 2441-2445.

Brownie,C. and Boos,D. (1994). Type I Error Robustness of ANOVA and ANOVA on Ranks When the Number of Treatments is Large. Biometrics *50*, 542-549.

Brunie,S., Zelwer,C., and Risler,J.L. (1990). Crystallographic study at 2.5 A resolution of the interaction of methionyl-tRNA synthetase from Escherichia coli with ATP. J. Mol. Biol. *216*, 411-424.

Buckingham,R.H., Grentzmann,G., and Kisselev,L. (1997). Polypeptide chain release factors. Mol. Microbiol. *24*, 449-456.

Calderone,R.A. and Fonzi,W.A. (2001). Virulence factors of Candida albicans. Trends Microbiol. *9*, 327-335.

Caporaso,J.G., Yarus,M., and Knight,R. (2005). Error minimization and coding triplet/binding site associations are independent features of the canonical genetic code. J. Mol. Evol. *61*, 597-607.

Carter,C.W., Jr. (1993). Cognition, mechanism, and evolutionary relationships in aminoacyl-tRNA synthetases. Annu. Rev. Biochem. *62*, 715-748.

Carter,C.W. and Duax,W.L. (2002). Did tRNA synthetase classes arise on opposite strands of the same gene? Mol. Cell *10*, 705-708.

Castresana,J., Feldmaier-Fuchs,G., and Paabo,S. (1998). Codon reassignment and amino acid composition in hemichordate mitochondria. Proc. Natl. Acad. Sci. U. S. A *95*, 3703-3707.

Cavarelli,J. and Moras,D. (1993). Recognition of tRNAs by aminoacyl-tRNA synthetases. FASEB J. *7*, 79-86.

Chapman,J.R. (2000). Mass spectrometry of proteins and peptides. (Totowa, NJ: Humana Press).

Chiapello,H., Ollivier,E., Landes-Devauchelle,C., Nitschke,P., and Risler,J.L. (1999). Codon usage as a tool to predict the cellular location of eukaryotic ribosomal proteins and aminoacyl-tRNA synthetases. Nucleic Acids Res. *27*, 2848-2851.

Chin,J.W., Cropp,T.A., Chu,S., Meggers,E., and Schultz,P.G. (2003). Progress toward an expanded eukaryotic genetic code. Chem. Biol. *10*, 511-519.

Chiti,F. and Dobson,C.M. (2006). Protein misfolding, functional amyloid, and human disease. Annu. Rev. Biochem. *75*, 333-366.

Clark-Walker,G.D. and Weiller,G.F. (1994). The structure of the small mitochondrial DNA of Kluyveromyces thermotolerans is likely to reflect the ancestral gene order in fungi. J. Mol. Evol. *38*, 593-601.

Collins,D.W. and Jukes,T.H. (1994). Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics *20*, 386-396.

Cowen,L.E. and Lindquist,S. (2005). Hsp90 potentiates the rapid evolution of new traits: drug resistance in diverse fungi. Science *309*, 2185-2189.

Crick,F. (1970). Central dogma of molecular biology. Nature *227*, 561-563.

Crick, F. H. A note for the tRNA tie club. 1955. Ref Type: Report

Crick,F.H. (1968). The origin of the genetic code. J. Mol. Biol. *38*, 367-379.

Cropp,T.A. and Schultz,P.G. (2004). An expanding genetic code. Trends Genet. *20*, 625-630.

Crothers,D.M., Seno,T., and Soll,G. (1972). Is there a discriminator site in transfer RNA? Proc. Natl. Acad. Sci. U. S. A *69*, 3063-3067.

Curnow,A.W., Hong,K., Yuan,R., Kim,S., Martins,O., Winkler,W., Henkin,T.M., and Soll,D. (1997). Glu-tRNAGln amidotransferase: a novel heterotrimeric enzyme required for correct decoding of glutamine codons during translation. Proc. Natl. Acad. Sci. U. S. A *94*, 11819-11826.

Curnow,A.W., Ibba,M., and Soll,D. (1996). tRNA-dependent asparagine formation. Nature *382*, 589-590.

Curran,J.F. (1993). Analysis of effects of tRNA:message stability on frameshift frequency at the Escherichia coli RF2 programmed frameshift site. Nucleic Acids Res. *21*, 1837-1843.

Cusack,S. (1995). Eleven down and nine to go. Nat. Struct. Biol. *2*, 824-831.

Cusack,S., Berthet-Colominas,C., Hartlein,M., Nassar,N., and Leberman,R. (1990). A second class of synthetase structure revealed by X-ray analysis of Escherichia coli seryl-tRNA synthetase at 2.5 A. Nature *347*, 249-255.

Cusack,S., Yaremchuk,A., and Tukalo,M. (2000). The 2 A crystal structure of leucyl-tRNA synthetase and its complex with a leucyl-adenylate analogue. EMBO J. *19*, 2351-2361.

Cusack,S., Yaremchuk,A., and Tukalo,M. (1996). The crystal structure of the ternary complex of T.thermophilus seryl-tRNA synthetase with tRNA(Ser) and a seryl-adenylate analogue reveals a conformational switch in the active site. EMBO J. *15*, 2834-2842.

Dale,T. and Uhlenbeck,O.C. (2005). Amino acid specificity in translation. Trends Biochem. Sci. *30*, 659-665.

De Backer,M.D., Magee,P.T., and Pla,J. (2000). RECENT DEVELOPMENTS IN MOLECULAR GENETICS OF CANDIDA ALBICANS. Ann. Rev. Microbiol. *54*, 463-498.

De,D.C. (1988). Transfer RNAs: the second genetic code. Nature *333*, 117-118.

Delarue,M. and Moras,D. (1993). The aminoacyl-tRNA synthetase family: modules at work. Bioessays *15*, 675-687.

Di Giulio,M. (2001b). The non-universality of the genetic code: the universal ancestor was a progenote. J. Theor. Biol. *209*, 345-349.

Di Giulio,M. (1999). The coevolution theory of the origin of the genetic code. J. Mol. Evol. *48*, 253-255.

Di Giulio,M. (2001a). A blind empiricism against the coevolution theory of the origin of the genetic code. J. Mol. Evol. *53*, 724-732.

Di Giulio,M. and Medugno,M. (1999). Physicochemical optimization in the genetic code origin as the number of codified amino acids increases. J. Mol. Evol. *49*, 1-10.

Dinman,J.D., Icho,T., and Wickner,R.B. (1991). A -1 ribosomal frameshift in a double-stranded RNA virus of yeast forms a gag-pol fusion protein. Proc. Natl. Acad. Sci. U. S. A *88*, 174-178.

Dirheimer,G., Baranowski,W., and Keith,G. (1995a). Variations in tRNA modifications, particularly of their queuine content in higher eukaryotes. Its relation to malignancy grading. Biochimie *77*, 99-103.

Dirheimer,G., Keith,G., Dumas,P., and Westhof,E. (1995b). Primary, Secundary and Terciary Structures of tRNAs. In tRNA: Structure, Biosynthesis and Function, D.Soll and U.L.RajBhandary, eds. (Washington: American Society for Microbiology).

Dock-Bregeon,A., Sankaranarayanan,R., Romby,P., Caillet,J., Springer,M., Rees,B., Francklyn,C.S., Ehresmann,C., and Moras,D. (2000). Transfer RNA-mediated editing in threonyl-tRNA synthetase. The class II solution to the double discrimination problem. Cell *103*, 877-884.

Dock-Bregeon,A.C., Garcia,A., Giege,R., and Moras,D. (1990). The contacts of yeast tRNA(Ser) with seryl-tRNA synthetase studied by footprinting experiments. Eur. J. Biochem. *188*, 283-290.

Doolittle,R.F. and Handy,J. (1998). Evolutionary anomalies among the aminoacyl-tRNA synthetases. Curr. Opin. Genet. Dev. *8*, 630-636.

Doring,V., Mootz,H.D., Nangle,L.A., Hendrickson,T.L., Crecy-Lagard,V., Schimmel,P., and Marliere,P. (2001). Enlarging the amino acid set of Escherichia coli by infiltration of the valine coding pathway. Science *292*, 501-504.

Draptchinskaia,N., Gustavsson,P., Andersson,B., Pettersson,M., Willig,T.N., Dianzani,I., Ball,S., Tchernia,G., Klar,J., Matsson,H., Tentler,D., Mohandas,N., Carlsson,B., and Dahl,N.

(1999). The gene encoding ribosomal protein S19 is mutated in Diamond-Blackfan anaemia. Nat. Genet. *21*, 169-175.

Drummond,D.A., Bloom,J.D., Adami,C., Wilke,C.O., and Arnold,F.H. (2005). Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. U. S. A *102*, 14338-14343.

Eaglestone,S.S., Cox,B.S., and Tuite,M.F. (1999). Translation termination efficiency can be regulated in Saccharomyces cerevisiae by environmental stress through a prion-mediated mechanism. EMBO J. *18*, 1974-1981.

Edelmann,P. and Gallant,J. (1977). Mistranslation in E. coli. Cell *10*, 131-137.

Ehrenberg,M. and Kurland,C.G. (1984). Costs of accuracy determined by a maximal growth rate constraint. Q. Rev. Biophys. *17*, 45-82.

Eisenlohr,L.C., Huang,L., and Golovina,T.N. (2007). Rethinking peptide supply to MHC class I molecules. Nat. Rev. Immunol. *7*, 403-410.

Ellis,J.T. and Morrison,D.A. (1995). Schistosoma mansoni: patterns of codon usage and bias. Parasitology *110 (Pt 1)*, 53-60.

Enjalbert,B., Nantel,A., and Whiteway,M. (2003). Stress-induced gene expression in Candida albicans: absence of a general stress response. Mol. Biol. Cell *14*, 1460-1467.

Eriani,G., Cavarelli,J., Martin,F., Ador,L., Rees,B., Thierry,J.C., Gangloff,J., and Moras,D. (1995). The class II aminoacyl-tRNA synthetases and their active site: evolutionary conservation of an ATP binding site. J. Mol. Evol. *40*, 499-508.

Eriani,G., Delarue,M., Poch,O., Gangloff,J., and Moras,D. (1990). Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. Nature *347*, 203-206.

Farabaugh,P.J. (1996). Programmed translational frameshifting. Annu. Rev. Genet. *30*, 507-528.

Farabaugh,P.J. and Bjork,G.R. (1999). How translational accuracy influences reading frame maintenance. EMBO J. *18*, 1427-1434.

Fersht,A.R., Ashford,J.S., Bruton,C.J., Jakes,R., Koch,G.L., and Hartley,B.S. (1975). Active site titration and aminoacyl adenylate binding stoichiometry of aminoacyl-tRNA synthetases. Biochemistry *14*, 1-4.

Fersht,A.R. and Dingwall,C. (1979). Evidence for the double-sieve editing mechanism in protein synthesis. Steric exclusion of isoleucine by valyl-tRNA synthetases. Biochemistry *18*, 2627-2631.

Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R., Eddy,S.R., Sonnhammer,E.L., and Bateman,A. (2006). Pfam: clans, web tools and services. Nucleic Acids Res. *34*, D247-D251.

Fleming,J.A., Vega,L.R., and Solomon,F. (2000). Function of tubulin binding proteins in vivo. Genetics *156*, 69-80.

Flygare,J. and Karlsson,S. (2007). Diamond-Blackfan anemia: erythropoiesis lost in translation. Blood *109*, 3152-3154.

Forche,A., Magee,P.T., Magee,B.B., and May,G. (2004). Genome-wide single-nucleotide polymorphism map for Candida albicans. Eukaryot. Cell *3*, 705-714.

Forchhammer,K., Leinfelder,W., Boesmiller,K., Veprek,B., and Bock,A. (1991). Selenocysteine synthase from Escherichia coli. Nucleotide sequence of the gene (selA) and purification of the protein. J. Biol. Chem. *266*, 6318-6323.

Francklyn,C., Perona,J.J., Puetz,J., and Hou,Y.M. (2002). Aminoacyl-tRNA synthetases: versatile players in the changing theater of translation. RNA. *8*, 1363-1372.

Francklyn,C., Shi,J.P., and Schimmel,P. (1992). Overlapping nucleotide determinants for specific aminoacylation of RNA microhelices. Science *255*, 1121-1125.

Freeland,S.J. and Hurst,L.D. (1998). The genetic code is one in a million. J. Mol. Evol. *47*, 238-248.

Freeland,S.J., Wu,T., and Keulmann,N. (2003). The case for an error minimizing standard genetic code. Orig. Life Evol. Biosph. *33*, 457-477.

Freist,W., Gauss,D.H., Ibba,M., and Soll,D. (1997). Glutaminyl-tRNA synthetase. Biol. Chem. *378*, 1103-1117.

Freist,W., Pardowitz,I., and Cramer,F. (1985). Isoleucyl-tRNA synthetase from bakers' yeast: multistep proofreading in discrimination between isoleucine and valine with modulated accuracy, a scheme for molecular recognition by energy dissipation. Biochemistry *24*, 7014-7023.

Freist,W., Sternbach,H., Pardowitz,I., and Cramer,F. (1998). Accuracy of protein biosynthesis: quasi-species nature of proteins and possibility of error catastrophes. J. Theor. Biol. *193*, 19-38.

Freistroffer,D.V., Pavlov,M.Y., MacDougall,J., Buckingham,R.H., and Ehrenberg,M. (1997). Release factor RF3 in E.coli accelerates the dissociation of release factors RF1 and RF2 from the ribosome in a GTP-dependent manner. EMBO J. *16*, 4126-4133.

Friedman,S.M. and Weinstein,I.B. (1964). Lack of fidelity in the translation of synthetic polyribonucleotides. Proc. Natl. Acad. Sci. U. S. A *52*, 988-996.

Frugier,M., Florentz,C., and Giege,R. (1994). Efficient aminoacylation of resected RNA helices by class II aspartyl-tRNA synthetase dependent on a single nucleotide. EMBO J. *13*, 2218-2226.

Fu,C. and Parker,J. (1994). A ribosomal frameshifting error during translation of the argI mRNA of Escherichia coli. Mol. Gen. Genet. *243*, 434-441.

Fukunaga,R. and Yokoyama,S. (2005). Aminoacylation complex structures of leucyl-tRNA synthetase and tRNALeu reveal two modes of discriminator-base recognition. Nat. Struct. Mol. Biol. *12*, 915-922.

Gallant,J.A. and Lindsley,D. (1992). Leftward ribosome frameshifting at a hungry codon. J. Mol. Biol. *223*, 31-40.

Geslain,R., Aeby,E., Guitart,T., Jones,T.E., de Moura,M.C., Charriere,F., Schneider,A., and de Pouplana,L.R. (2006). Trypanosoma Seryl-tRNA Synthetase Is a Metazoan-like Enzyme with High Affinity for tRNASec. J. Biol. Chem. *281*, 38217-38225.

Ghaemmaghami,S., Huh,W.K., Bower,K., Howson,R.W., Belle,A., Dephoure,N., O'Shea,E.K., and Weissman,J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737-741.

Giege,R., Sissler,M., and Florentz,C. (1998). Universal rules and idiosyncratic features in tRNA identity. Nucleic Acids Res. *26*, 5017-5035.

Gingras,A.C., Raught,B., and Sonenberg,N. (1999). eIF4 initiation factors: effectors of mRNA recruitment to ribosomes and regulators of translation. Annu. Rev. Biochem. *68*, 913-963.

Gomes,A.C., Miranda,I., Silva,R.M., Moura,G.R., Thomas,B., Akoulitchev,A., and Santos,M.A. (2007). A genetic code alteration generates a proteome of high diversity in the human pathogen Candida albicans. Genome Biol. *8*, R206.

Gouet,P., Robert,X., and Courcelle,E. (2003). ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. Nucleic Acids Res. *31*, 3320-3323.

Grant,P.A., Schieltz,D., Pray-Grant,M.G., Steger,D.J., Reese,J.C., Yates,J.R., III, and Workman,J.L. (1998). A subset of TAF(II)s are integral components of the SAGA complex required for nucleosome acetylation and transcriptional stimulation. Cell *94*, 45-53.

Gregory,L.A., guissa-Toure,A.H., Pinaud,N., Legrand,P., Gleizes,P.E., and Fribourg,S. (2007). Molecular basis of Diamond Blackfan anemia: structure and function analysis of RPS19. Nucl. Acids Res. *35*, 5913-5921.

Gromadski,K.B. and Rodnina,M.V. (2004). Kinetic determinants of high-fidelity tRNA discrimination on the ribosome. Mol. Cell *13*, 191-200.

Grosjean,H., Cedergren,R.J., and McKay,W. (1982). Structure in tRNA data. Biochimie *64*, 387-397.

Haig,D. and Hurst,L.D. (1991). A quantitative measure of error minimization in the genetic code. J. Mol. Evol. *33*, 412-417.

Hairfield,M.L., Westwater,C., and Dolan,J.W. (2002). Phosphatidylinositol-4-phosphate 5-kinase activity is stimulated during temperature-induced morphogenesis in Candida albicans. Microbiology *148*, 1737-1746.

Hansen,J.L., Moore,P.B., and Steitz,T.A. (2003). Structures of five antibiotics bound at the peptidyl transferase center of the large ribosomal subunit. J. Mol. Biol. *330*, 1061-1075.

Hanyu,N., Kuchino,Y., Nishimura,S., and Beier,H. (1986). Dramatic events in ciliate evolution: alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two Tetrahymena tRNAs. EMBO J. *5*, 1307-1311.

Hao,B., Gong,W., Ferguson,T.K., James,C.M., Krzycki,J.A., and Chan,M.K. (2002). A new UAG-encoded residue in the structure of a methanogen methyltransferase. Science *296*, 1462-1466.

Harding,H.P. and Ron,D. (2002). Endoplasmic reticulum stress and the development of diabetes: a review. Diabetes *51 Suppl 3*, S455-S461.

Hatfield,D.L. and Gladyshev,V.N. (2002). How selenium has altered our understanding of the genetic code. Mol. Cell Biol. *22*, 3565-3576.

Hayashi-Ishimaru,Y., Ohama,T., Kawatsu,Y., Nakamura,K., and Osawa,S. (1996). UAG is a sense codon in several chlorophycean mitochondria. Curr. Genet. *30*, 29-33.

Hendrickson,T.L., de Crecy-Lagard,V., and Schimmel,P. (2004). Incorporation of nonnatural amino acids into proteins. Annu. Rev. Biochem. *73*, 147-176.

Hendrickson,W.A., Horton,J.R., and LeMaster,D.M. (1990). Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. EMBO J. *9*, 1665-1672.

Himeno,H., Hasegawa,T., Ueda,T., Watanabe,K., and Shimizu,M. (1990). Conversion of aminoacylation specificity from tRNA(Tyr) to tRNA(Ser) in vitro. Nucleic Acids Res. *18*, 6815-6819.

Himeno,H., Yoshida,S., Soma,A., and Nishikawa,K. (1997). Only one nucleotide insertion to the long variable arm confers an efficient serine acceptor activity upon Saccharomyces cerevisiae tRNA(Leu) in vitro. J. Mol. Biol. *268*, 704-711.

Holley,R.W. (1965). Structure of an alanine transfer ribonucleic acid. JAMA *194*, 868-871.

Horsfield,J.A., Wilson,D.N., Mannering,S.A., Adamski,F.M., and Tate,W.P. (1995). Prokaryotic ribosomes recode the HIV-1 gag-pol-1 frameshift sequence by an E/P site post-translocation simultaneous slippage mechanism. Nucleic Acids Res. *23*, 1487-1494.

Hou,Y.M., Sterner,T., and Bhalla,R. (1995). Evidence for a conserved relationship between an acceptor stem and a tRNA for aminoacylation. RNA. *1*, 707-713.

Hou,Y.M., Westhof,E., and Giege,R. (1993). An unusual RNA tertiary interaction has a role for the specific aminoacylation of a transfer RNA. Proc. Natl. Acad. Sci. U. S. A *90*, 6776-6780.

Hull,C.M., Raisner,R.M., and Johnson,A.D. (2000). Evidence for mating of the "asexual" yeast Candida albicans in a mammalian host. Science *289*, 307-310.

Hull,M.W., Erickson,J., Johnston,M., and Engelke,D.R. (1994). tRNA genes as transcriptional repressor elements. Mol. Cell Biol. *14*, 1266-1277.

Ibba,M., Bono,J.L., Rosa,P.A., and Soll,D. (1997a). Archaeal-type lysyl-tRNA synthetase in the Lyme disease spirochete Borrelia burgdorferi. Proc. Natl. Acad. Sci. U. S. A *94*, 14383-14388.

Ibba,M., Morgan,S., Curnow,A.W., Pridmore,D.R., Vothknecht,U.C., Gardner,W., Lin,W., Woese,C.R., and Soll,D. (1997b). A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases. Science *278*, 1119-1122.

Ibba,M. and Soll,D. (2000). Aminoacyl-tRNA synthesis. Annu. Rev. Biochem. *69*, 617-650.

Ibba,M. and Soll,D. (2004). Aminoacyl-tRNAs: setting the limits of the genetic code. Genes Dev. *18*, 731-738.

Ikemura,T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. *2*, 13-34.

Ikesugi,K., Yamamoto,R., Mulhern,M.L., and Shinohara,T. (2006). Role of the unfolded protein response (UPR) in cataract formation. Exp. Eye Res. *83*, 508-516.

Inagaki,Y., Ehara,M., Watanabe,K.I., Hayashi-Ishimaru,Y., and Ohama,T. (1998). Directionally evolving genetic code: the UGA codon from stop to tryptophan in mitochondria. J. Mol. Evol. *47*, 378-384.

Jacks,T., Madhani,H.D., Masiarz,F.R., and Varmus,H.E. (1988). Signals for ribosomal frameshifting in the Rous sarcoma virus gag-pol region. Cell *55*, 447-458.

Jakubowski,H. (1997). Aminoacyl thioester chemistry of class II aminoacyl-tRNA synthetases. Biochemistry *36*, 11077-11085.

Jakubowski,H. and Goldman,E. (1992). Editing of errors in selection of amino acids for protein synthesis. Microbiol. Rev. *56*, 412-429.

Jordanova,A., Irobi,J., Thomas,F.P., Van,D.P., Meerschaert,K., Dewil,M., Dierick,I., Jacobs,A., De,V.E., Guergueltcheva,V., Rao,C.V., Tournev,I., Gondim,F.A., D'Hooghe,M., Van,G., V, Callaerts,P., Van Den,B.L., Timmermans,J.P., Robberecht,W., Gettemans,J., Thevelein,J.M., De,J.P., Kremensky,I., and Timmerman,V. (2006). Disrupted function and axonal distribution of mutant tyrosyl-tRNA synthetase in dominant intermediate Charcot-Marie-Tooth neuropathy. Nat. Genet. *38*, 197-202.

Kano,A., Andachi,Y., Ohama,T., and Osawa,S. (1991). Novel anticodon composition of transfer RNAs in Micrococcus luteus, a bacterium with a high genomic G + C content. Correlation with codon usage. J. Mol. Biol. *221*, 387-401.

Kaplowitz,N. and Ji,C. (2006). Unfolding new mechanisms of alcoholic liver disease in the endoplasmic reticulum. J. Gastroenterol. Hepatol. *21*, S7-S9.

Kaposzta,R., Marodi,L., Hollinshead,M., Gordon,S., and da Silva,R.P. (1999). Rapid recruitment of late endosomes and lysosomes in mouse macrophages ingesting Candida albicans. J. Cell Sci. *112 (Pt 19)*, 3237-3248.

Kapp,L.D. and Lorsch,J.R. (2004). The molecular mechanics of eukaryotic translation. Annu. Rev. Biochem. *73*, 657-704.

Keeling,P.J. and Doolittle,W.F. (1997). Widespread and ancient distribution of a noncanonical genetic code in diplomonads. Mol. Biol. Evol. *14*, 895-901.

Kent,S.B. (1988). Chemical synthesis of peptides and proteins. Annu. Rev. Biochem. *57*, 957-989.

Kim,H.S., Kim,I.Y., Soll,D., and Lee,S.Y. (2000). Transfer RNA identity change in anticodon variants of E. coli tRNA(Phe) in vivo. Mol. Cells *10*, 76-82.

Kiso,Y. (1999). Protease inhibitors. Biopolymers *51*, 1.

Kleeman,T.A., Wei,D., Simpson,K.L., and First,E.A. (1997). Human tyrosyl-tRNA synthetase shares amino acid sequence homology with a putative cytokine. J. Biol. Chem. *272*, 14420-14425.

Knight,R.D., Freeland,S.J., and Landweber,L.F. (2001). Rewiring the keyboard: evolvability of the genetic code. Nat. Rev. Genet. *2*, 49-58.

Knight,R.D., Freeland,S.J., and Landweber,L.F. (1999). Selection, history and chemistry: the three faces of the genetic code. Trends Biochem. Sci. *24*, 241-247.

Knight,R.D. and Landweber,L.F. (1998). Rhyme or reason: RNA-arginine interactions and the genetic code. Chem. Biol. *5*, R215-R220.

Knight,R.D. and Landweber,L.F. (2000). Guilt by association: the arginine case revisited. RNA. *6*, 499-510.

Kondow,A., Suzuki,T., Yokobori,S., Ueda,T., and Watanabe,K. (1999). An extra tRNAGly(U*CU) found in ascidian mitochondria responsible for decoding non-universal codons AGA/AGG as glycine. Nucleic Acids Res. *27*, 2554-2559.

Koonin,E.V. and Aravind,L. (1998). Genomics: re-evaluation of translation machinery evolution. Curr. Biol. *8*, R266-R269.

Korencic,D., Polycarpo,C., Weygand-Durasevic,I., and Soll,D. (2004). Differential modes of transfer RNASer recognition in Methanosarcina barkeri. J. Biol. Chem. *279*, 48780-48786.

Kostova,Z. and Wolf,D.H. (2003). For whom the bell tolls: protein quality control of the endoplasmic reticulum and the ubiquitin-proteasome connection. EMBO J. *22*, 2309-2317.

Kryndushkin,D.S., Smirnov,V.N., Ter-Avanesyan,M.D., and Kushnirov,V.V. (2002). Increased expression of Hsp40 chaperones, transcriptional factors, and ribosomal protein Rpp0 can cure yeast prions. J. Biol. Chem. *277*, 23702-23708.

Krzycki,J.A. (2005). The direct genetic encoding of pyrrolysine. Curr. Opin. Microbiol. *8*, 706-712.

Kumar,S. (1996). Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. Genetics *143*, 537-548.

Kumar,S., Tamura,K., and Nei,M. (2004). MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief. Bioinform. *5*, 150-163.

Kurland,C. and Gallant,J. (1996). Errors of heterologous protein expression. Curr. Opin. Biotechnol. *7*, 489-493.

Kurland,C.G. and Ehrenberg,M. (1984). Optimization of translation accuracy. Prog. Nucleic Acid Res. Mol. Biol. *31*, 191-219.

LaRiviere, F. J., Wolfson, A. D., and Uhlenbeck, O. C. (2001). Uniform binding of aminoacyl-tRNAs to elongation factor Tu by thermodynamic compensation. Science 294, 165-168.

Ladner,J.E., Jack,A., Robertus,J.D., Brown,R.S., Rhodes,D., Clark,B.F., and Klug,A. (1975). Structure of yeast phenylalanine transfer RNA at 2.5 A resolution. Proc. Natl. Acad. Sci. U. S. A *72*, 4414-4418.

Lamour,V., Quevillon,S., Diriong,S., N'Guyen,V.C., Lipinski,M., and Mirande,M. (1994). Evolution of the Glx-tRNA synthetase family: the glutaminyl enzyme as a case of horizontal gene transfer. Proc. Natl. Acad. Sci. U. S. A *91*, 8670-8674.

Lan,C.Y., Newport,G., Murillo,L.A., Jones,T., Scherer,S., Davis,R.W., and Agabian,N. (2002). Metabolic specialization associated with phenotypic switching in Candidaalbicans. Proc. Natl. Acad. Sci. U. S. A *99*, 14907-14912.

Larrinoa,I.F. and Heredia,C.F. (1991). Yeast proteinase yscB inactivates the leucyl tRNA synthetase in extracts of Saccharomyces cerevisiae. Biochim. Biophys. Acta *1073*, 502-508.

Lee,C.P., Mandal,N., Dyson,M.R., and RajBhandary,U.L. (1993). The discriminator base influences tRNA structure at the end of the acceptor stem and possibly its interaction with proteins. Proc. Natl. Acad. Sci. U. S. A *90*, 7149-7152.

Lee,J.H., Pestova,T.V., Shin,B.S., Cao,C., Choi,S.K., and Dever,T.E. (2002). Initiation factor eIF5B catalyzes second GTP-dependent step in eukaryotic translation initiation. Proc. Natl. Acad. Sci. U. S. A *99*, 16689-16694.

Lee,J.W., Beebe,K., Nangle,L.A., Jang,J., Longo-Guess,C.M., Cook,S.A., Davisson,M.T., Sundberg,J.P., Schimmel,P., and Ackerman,S.L. (2006). Editing-defective tRNA synthetase causes protein misfolding and neurodegeneration. Nature *443*, 50-55.

Leegwater,P.A., Vermeulen,G., Konst,A.A., Naidu,S., Mulders,J., Visser,A., Kersbergen,P., Mobach,D., Fonds,D., van Berkel,C.G., Lemmers,R.J., Frants,R.R., Oudejans,C.B., Schutgens,R.B., Pronk,J.C., and van der Knaap,M.S. (2001). Subunits of the translation initiation factor eIF2B are mutant in leukoencephalopathy with vanishing white matter. Nat. Genet. *29*, 383-388.

Leinfelder,W., Zehelein,E., Mandrand-Berthelot,M.A., and Bock,A. (1988). Gene for a novel tRNA species that accepts L-serine and cotranslationally inserts selenocysteine. Nature *331*, 723-725.

LeJohn,H.B., Cameron,L.E., Yang,B., and Rennie,S.L. (1994). Molecular characterization of an NAD-specific glutamate dehydrogenase gene inducible by L-glutamine. Antisense gene pair arrangement with L-glutamine-inducible heat shock 70-like protein gene. J. Biol. Chem. *269*, 4523-4531.

Lenhard,B., Filipic,S., Landeka,I., Skrtic,I., Soll,D., and Weygand-Durasevic,I. (1997). Defining the active site of yeast seryl-tRNA synthetase. Mutations in motif 2 loop residues affect tRNA-dependent amino acid recognition. J. Biol. Chem. *272*, 1136-1141.

Lenhard,B., Orellana,O., Ibba,M., and Weygand-Durasevic,I. (1999). tRNA recognition and evolution of determinants in seryl-tRNA synthesis. Nucleic Acids Res. *27*, 721-729.

Levitt,M. (1969). Detailed molecular model for transfer ribonucleic acid. Nature *224*, 759-763.

Li,J., Esberg,B., Curran,J.F., and Bjork,G.R. (1997). Three modified nucleosides present in the anticodon stem and loop influence the in vivo aa-tRNA selection in a tRNA-dependent manner. J. Mol. Biol. *271*, 209-221.

Li,M. and Tzagoloff,A. (1979). Assembly of the mitochondrial membrane system: sequences of yeast mitochondrial valine and an unusual threonine tRNA gene. Cell *18*, 47-53.

Li,Z., Stahl,G., and Farabaugh,P.J. (2001). Programmed +1 frameshifting stimulated by complementarity between a downstream mRNA sequence and an error-correcting region of rRNA. RNA. *7*, 275-284.

Lin,L. and Schimmel,P. (1996). Mutational analysis suggests the same design for editing activities of two tRNA synthetases. Biochemistry *35*, 5596-5601.

Lindholm,D., Wootz,H., and Korhonen,L. (2006). ER stress and neurodegenerative diseases. Cell Death. Differ. *13*, 385-392.

Lindquist,S. and Craig,E.A. (1988). The heat-shock proteins. Annu. Rev. Genet. *22*, 631-677.

Lindsley,D., Gallant,J., Doneanu,C., Bonthuis,P., Caldwell,S., and Fontelera,A. (2005). Spontaneous ribosome bypassing in growing cells. J. Mol. Biol. *349*, 261-272.

Lobanov,A.V., Kryukov,G.V., Hatfield,D.L., and Gladyshev,V.N. (2006). Is there a twenty third amino acid in the genetic code? Trends Genet. *22*, 357-360.

Loftfield,R.B. and Vanderjagt,D. (1972). The frequency of errors in protein biosynthesis. Biochem. J. *128*, 1353-1356.

Lovett,P.S., Ambulos,N.P., Jr., Mulbry,W., Noguchi,N., and Rogers,E.J. (1991). UGA can be decoded as tryptophan at low efficiency in Bacillus subtilis. J. Bacteriol. *173*, 1810-1812.

Lozupone,C.A., Knight,R.D., and Landweber,L.F. (2001). The molecular basis of nuclear genetic code change in ciliates. Curr. Biol. *11*, 65-74.

Ma,Y. and Hendershot,L.M. (2004). The role of the unfolded protein response in tumour development: friend or foe? Nat. Rev. Cancer *4*, 966-977.

Magee,B.B. and Magee,P.T. (2000). Induction of mating in Candida albicans by construction of MTLa and MTLalpha strains. Science *289*, 310-313.

Magee,P.T. and Magee,B.B. (2004). Through a glass opaquely: the biological significance of mating in Candida albicans. Curr. Opin. Microbiol. *7*, 661-665.

Mak,C.C., Brik,A., Lerner,D.L., Elder,J.H., Morris,G.M., Olson,A.J., and Wong,C.H. (2003). Design and synthesis of broad-based mono- and bi- cyclic inhibitors of FIV and HIV proteases. Bioorg. Med. Chem. *11*, 2025-2040.

Martinis,S.A., Plateau,P., Cavarelli,J., and Florentz,C. (1999). Aminoacyl-tRNA synthetases: a new image for a classical family. Biochimie *81*, 683-700.

Massey,S.E., Moura,G., Beltrao,P., Almeida,R., Garey,J.R., Tuite,M.F., and Santos,M.A. (2003). Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the CTG codon in Candida spp. Genome Res. *13*, 544-557.

Matsugi,J., Murao,K., and Ishikura,H. (1998). Effect of B. subtilis TRNA(Trp) on readthrough rate at an opal UGA codon. J. Biochem. (Tokyo) *123*, 853-858.

McClain,W.H. (1993). Rules that govern tRNA identity in protein synthesis. J. Mol. Biol. *234*, 257-280.

McClain,W.H., Foss,K., Jenkins,R.A., and Schneider,J. (1990). Nucleotides that determine Escherichia coli tRNA(Arg) and tRNA(Lys) acceptor identities revealed by analyses of mutant opal and amber suppressor tRNAs. Proc. Natl. Acad. Sci. U. S. A *87*, 9260-9264.

Mechulam,Y., Meinnel,T., and Blanquet,S. (1995). A family of RNA-binding enzymes. the aminoacyl-tRNA synthetases. Subcell. Biochem. *24*, 323-376.

Mehl,R.A., Anderson,J.C., Santoro,S.W., Wang,L., Martin,A.B., King,D.S., Horn,D.M., and Schultz,P.G. (2003). Generation of a bacterium with a 21 amino acid genetic code. J. Am. Chem. Soc. *125*, 935-939.

Miller,R.A. and Britigan,B.E. (1997). Role of oxidants in microbial pathophysiology. Clin. Microbiol. Rev. *10*, 1-18.

Miranda,I. (2007). A Genetic Code Alteration Is a Phenotype Diversity Generator in the Human Pathogen Candida albicans. PLoS ONE *2*, e996.

Mitra,K. and Frank,J. (2006). Ribosome dynamics: insights from atomic structure modeling into cryo-electron microscopy maps. Annu. Rev. Biophys. Biomol. Struct. *35*, 299-317.

Moras,D. (1992). Structural and functional relationships between aminoacyl-tRNA synthetases. Trends Biochem. Sci. *17*, 159-164.

Moriya,J., Yokogawa,T., Wakita,K., Ueda,T., Nishikawa,K., Crain,P.F., Hashizume,T., Pomerantz,S.C., McCloskey,J.A., Kawai,G., and . (1994). A novel modified nucleoside found at the first position of the anticodon of methionine tRNA from bovine liver mitochondria. Biochemistry *33*, 2234-2239.

Moriyama,E.N. and Powell,J.R. (1997). Synonymous substitution rates in Drosophila: mitochondrial versus nuclear genes. J. Mol. Evol. *45*, 378-391.

Morton,B.R. and Clegg,M.T. (1995). Neighboring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. J. Mol. Evol. *41*, 597-603.

Mosyak,L., Reshetnikova,L., Goldgur,Y., Delarue,M., and Safro,M.G. (1995). Structure of phenylalanyl-tRNA synthetase from Thermus thermophilus. Nat. Struct. Biol. *2*, 537-547.

Moura,G., Pinheiro,M., Silva,R., Miranda,I., Afreixo,V., Dias,G., Freitas,A., Oliveira,J.L., and Santos,M.A. (2005). Comparative context analysis of codon pairs on an ORFeome scale. Genome Biol. *6*, R28.

Muramatsu,T., Yokoyama,S., Horie,N., Matsuda,A., Ueda,T., Yamaizumi,Z., Kuchino,Y., Nishimura,S., and Miyazawa,T. (1988). A novel lysine-substituted nucleoside in the first position of the anticodon of minor isoleucine tRNA from Escherichia coli. J. Biol. Chem. *263*, 9261-9267.

Murgola,E.J., Pagel,F.T., and Hijazi,K.A. (1984). Codon context effects in missense suppression. J. Mol. Biol. *175*, 19-27.

Mursinna,R.S., Lee,K.W., Briggs,J.M., and Martinis,S.A. (2004). Molecular dissection of a critical specificity determinant within the amino acid editing domain of leucyl-tRNA synthetase. Biochemistry *43*, 155-165.

Mursinna,R.S., Lincecum,T.L., Jr., and Martinis,S.A. (2001). A conserved threonine within Escherichia coli leucyl-tRNA synthetase prevents hydrolytic editing of leucyl-tRNALeu. Biochemistry *40*, 5376-5381.

Nagel,G.M. and Doolittle,R.F. (1991). Evolution and relatedness in two aminoacyl-tRNA synthetase families. Proc. Natl. Acad. Sci. U. S. A *88*, 8121-8125.

Nakamura,Y., Ito,K., and Isaksson,L.A. (1996). Emerging understanding of translation termination. Cell *87*, 147-150.

Namy,O., Rousset,J.P., Napthine,S., and Brierley,I. (2004). Reprogrammed genetic decoding in cellular gene expression. Mol. Cell *13*, 157-168.

Nangle,L.A., De,C.L., V, Doring,V., and Schimmel,P. (2002). Genetic code ambiguity. Cell viability related to the severity of editing defects in mutant tRNA synthetases. J. Biol. Chem. *277*, 45729-45733.

Nangle,L.A., Motta,C.M., and Schimmel,P. (2006). Global effects of mistranslation from an editing defect in mammalian cells. Chem. Biol. *13*, 1091-1100.

Ni,L. and Snyder,M. (2001). A Genomic Study of the Bipolar Bud Site Selection Pattern in Saccharomyces cerevisiae. Mol. Biol. Cell *12*, 2147-2170.

Nierhaus,K.H. (1990). The allosteric three-site model for the ribosomal elongation cycle: features and future. Biochemistry *29*, 4997-5008.

Nirenberg,M., Caskey,T., Marshall,R., Brimacombe,R., Kellogg,D., Doctor,B., Hatfield,D., Levin,J., Rottman,F., Pestka,S., Wilcox,M., and Anderson,F. (1966). The RNA code and protein synthesis. Cold Spring Harb. Symp. Quant. Biol. *31*, 11-24.

Nirenberg,M. and Leder,P. (1964). RNA CODEWORDS AND PROTEIN SYNTHESIS. THE EFFECT OF TRINUCLEOTIDES UPON THE BINDING OF SRNA TO RIBOSOMES. Science *145*, 1399-1407.

Nirenberg,M. and Matthaei,J. (1961). The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. Proc. Natl. Acad. Sci. U. S. A *47*, 1588-1602.

Noller,H.F. (1993). Peptidyl transferase: protein, ribonucleoprotein, or RNA? J. Bacteriol. *175*, 5297-5300.

Noller,H.F., Yusupov,M.M., Yusupova,G.Z., Baucom,A., and Cate,J.H. (2002). Translocation of tRNA during protein synthesis. FEBS Lett. *514*, 11-16.

Noren,C.J., nthony-Cahill,S.J., Griffith,M.C., and Schultz,P.G. (1989). A general method for site-specific incorporation of unnatural amino acids into proteins. Science *244*, 182-188.

Normanly,J. and Abelson,J. (1989). tRNA identity. Annu. Rev. Biochem. *58*, 1029-1049.

Normanly,J., Ollick,T., and Abelson,J. (1992). Eight base changes are sufficient to convert a leucine-inserting tRNA into a serine-inserting tRNA. Proc. Natl. Acad. Sci. U. S. A *89*, 5680-5684.

Nureki,O., Vassylyev,D.G., Tateno,M., Shimada,A., Nakama,T., Fukai,S., Konno,M., Hendrickson,T.L., Schimmel,P., and Yokoyama,S. (1998). Enzyme structure with two catalytic sites for double-sieve selection of substrate. Science *280*, 578-582.

O'Donoghue,P., Sethi,A., Woese,C.R., and Luthey-Schulten,Z.A. (2005). The evolutionary history of Cys-tRNACys formation. Proc. Natl. Acad. Sci. U. S. A *102*, 19003-19008.

O'Sullivan,J.M., Mihr,M.J., Santos,M.A., and Tuite,M.F. (2001a). Seryl-tRNA synthetase is not responsible for the evolution of CUG codon reassignment in Candida albicans. Yeast *18*, 313-322.

O'Sullivan,J.M., Mihr,M.J., Santos,M.A., and Tuite,M.F. (2001b). The Candida albicans gene encoding the cytoplasmic leucyl-tRNA synthetase: implications for the evolution of CUG codon reassignment. Gene *275*, 133-140.

Oba,T., Andachi,Y., Muto,A., and Osawa,S. (1991). CGG: an unassigned or nonsense codon in Mycoplasma capricolum. Proc. Natl. Acad. Sci. U. S. A *88*, 921-925.

Ogle,J.M., Carter,A.P., and Ramakrishnan,V. (2003). Insights into the decoding mechanism from recent ribosome structures. Trends Biochem. Sci. *28*, 259-266.

Ogle,J.M., Murphy,F.V., Tarry,M.J., and Ramakrishnan,V. (2002). Selection of tRNA by the ribosome requires a transition from an open to a closed form. Cell *111*, 721-732.

Ogle,J.M. and Ramakrishnan,V. (2005). Structural insights into translational fidelity. Annu. Rev. Biochem. *74*, 129-177.

Ohama,T., Muto,A., and Osawa,S. (1990). Role of GC-biased mutation pressure on synonymous codon choice in Micrococcus luteus, a bacterium with a high genomic GC-content. Nucleic Acids Res. *18*, 1565-1569.

Ohama,T., Suzuki,T., Mori,M., Osawa,S., Ueda,T., Watanabe,K., and Nakase,T. (1993). Non-universal decoding of the leucine codon CUG in several Candida species. Nucleic Acids Res. *21*, 4039-4045.

Osawa,S. and Jukes,T.H. (1989). Codon reassignment (codon capture) in evolution. J. Mol. Evol. *28*, 271-278.

Osawa,S., Jukes,T.H., Watanabe,K., and Muto,A. (1992). Recent evidence for evolution of the genetic code. Microbiol. Rev. *56*, 229-264.

Oyadomari,S., Yun,C., Fisher,E.A., Kreglinger,N., Kreibich,G., Oyadomari,M., Harding,H.P., Goodman,A.G., Harant,H., Garrison,J.L., Taunton,J., Katze,M.G., and Ron,D. (2006). Cotranslocational degradation protects the stressed endoplasmic reticulum from protein overload. Cell *126*, 727-739.

Panasenko,O., Landrieux,E., Feuermann,M., Finka,A., Paquet,N., and Collart,M.A. (2006). The yeast Ccr4-Not complex controls ubiquitination of the nascent-associated polypeptide (NAC-EGD) complex. J. Biol. Chem. *281*, 31389-31398.

Pape,L.K., Koerner,T.J., and Tzagoloff,A. (1985). Characterization of a yeast nuclear gene (MST1) coding for the mitochondrial threonyl-tRNA1 synthetase. J. Biol. Chem. *260*, 15362-15370.

Pape,T., Wintermeyer,W., and Rodnina,M. (1999). Induced fit in initial selection and proofreading of aminoacyl-tRNA on the ribosome. EMBO J. *18*, 3800-3807.

Pape,T., Wintermeyer,W., and Rodnina,M.V. (2000). Conformational switch in the decoding region of 16S rRNA during aminoacyl-tRNA selection on the ribosome. Nat. Struct. Biol. *7*, 104-107.

Parker,J. (1989). Errors and alternatives in reading the universal genetic code. Microbiol. Rev. *53*, 273-298.

Parker,J. and Precup,J. (1986). Mistranslation during phenylalanine starvation. Mol. Gen. Genet. *204*, 70-74.

Pearse,B.R. and Hebert,D.N. (2006). Cotranslocational degradation: utilitarianism in the ER stress response. Mol. Cell *23*, 773-775.

Perreau,V.M., Keith,G., Holmes,W.M., Przykorska,A., Santos,M.A., and Tuite,M.F. (1999). The Candida albicans CUG-decoding ser-tRNA has an atypical anticodon stem-loop structure. J. Mol. Biol. *293*, 1039-1053.

Peske,F., Rodnina,M.V., and Wintermeyer,W. (2005). Sequence of steps in ribosome recycling as defined by kinetic analysis. Mol. Cell *18*, 403-412.

Pestova,T.V., Lomakin,I.B., Lee,J.H., Choi,S.K., Dever,T.E., and Hellen,C.U. (2000). The joining of ribosomal subunits in eukaryotes requires eIF5B. Nature *403*, 332-335.

Pezo,V., Metzgar,D., Hendrickson,T.L., Waas,W.F., Hazebrouck,S., Doring,V., Marliere,P., Schimmel,P., and Crecy-Lagard,V. (2004). Artificially ambiguous genetic code confers growth yield advantage. Proc. Natl. Acad. Sci. U. S. A *101*, 8593-8597.

Pickart,C.M. and Cohen,R.E. (2004). Proteasomes and their kin: proteases in the machine age. Nat. Rev. Mol. Cell Biol. *5*, 177-187.

Pla,J., Perez-Diaz,R.M., Navarro-Garcia,F., Sanchez,M., and Nombela,C. (1995). Cloning of the Candida albicans HIS1 gene by direct complementation of a C. albicans histidine auxotroph using an improved double-ARS shuttle vector. Gene *165*, 115-120.

Polycarpo,C., Ambrogelly,A., Berube,A., Winbush,S.M., McCloskey,J.A., Crain,P.F., Wood,J.L., and Soll,D. (2004). An aminoacyl-tRNA synthetase that specifically activates pyrrolysine. Proc. Natl. Acad. Sci. U. S. A *101*, 12450-12454.

Polycarpo,C., Ambrogelly,A., Ruan,B., Tumbula-Hansen,D., Ataide,S.F., Ishitani,R., Yokoyama,S., Nureki,O., Ibba,M., and Soll,D. (2003). Activation of the pyrrolysine suppressor tRNA requires formation of a ternary complex with class I and class II lysyl-tRNA synthetases. Mol. Cell *12*, 287-294.

Pouwels,P.H. and Leunissen,J.A. (1994). Divergence in codon usage of Lactobacillus species. Nucleic Acids Res. *22*, 929-936.

Praetorius-Ibba,M. and Ibba,M. (2003). Aminoacyl-tRNA synthesis in archaea: different but not unique. Mol. Microbiol. *48*, 631-637.

Prather,N.E., Murgola,E.J., and Mims,B.H. (1984). Nucleotide substitution in the amino acid acceptor stem of lysine transfer RNA causes missense suppression. J. Mol. Biol. *172*, 177-184.

Princiotta,M.F., Finzi,D., Qian,S.B., Gibbs,J., Schuchmann,S., Buttgereit,F., Bennink,J.R., and Yewdell,J.W. (2003). Quantitating protein synthesis, degradation, and endogenous antigen processing. Immunity. *18*, 343-354.

Qiu,X.B., Shao,Y.M., Miao,S., and Wang,L. (2006). The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones. Cell Mol. Life Sci. *63*, 2560-2570.

Queitsch,C., Sangster,T.A., and Lindquist,S. (2002). Hsp90 as a capacitor of phenotypic variation. Nature *417*, 618-624.

Quinn,C.L., Tao,N., and Schimmel,P. (1995). Species-specific microhelix aminoacylation by a eukaryotic pathogen tRNA synthetase dependent on a single base pair. Biochemistry *34*, 12489-12495.

Rao,R.V. and Bredesen,D.E. (2004). Misfolded proteins, endoplasmic reticulum stress and neurodegeneration. Curr. Opin. Cell Biol. *16*, 653-662.

Ren,L., Rahman,M.S., and Humayun,M.Z. (1999). Escherichia coli cells exposed to streptomycin display a mutator phenotype. J. Bacteriol. *181*, 1043-1044.

Ribas,d.P., Brown,J.R., and Schimmel,P. (2001). Structure-based phylogeny of class IIa tRNA synthetases in relation to an unusual biochemistry. J. Mol. Evol. *53*, 261-268.

Ribas,d.P. and Schimmel,P. (2001a). Aminoacyl-tRNA synthetases: potential markers of genetic code development. Trends Biochem. Sci. *26*, 591-596.

Ribas,d.P. and Schimmel,P. (2001b). Two classes of tRNA synthetases suggested by sterically compatible dockings on tRNA acceptor stem. Cell *104*, 191-193.

Rodin,S.N. and Ohno,S. (1995). Two types of aminoacyl-tRNA synthetases could be originally encoded by complementary strands of the same nucleic acid. Orig. Life Evol. Biosph. *25*, 565-589.

Rodin,S.N. and Rodin,A.S. (2006). Partitioning of Aminoacyl-tRNA Synthetases in Two Classes Could Have Been Encoded in a Strand-Symmetric RNA World. DNA and Cell Biology *25*, 617-626.

Rodnina,M.V., Daviter,T., Gromadski,K., and Wintermeyer,W. (2002). Structural dynamics of ribosomal RNA during decoding on the ribosome. Biochimie *84*, 745-754.

Rodnina,M.V., Savelsbergh,A., Matassova,N.B., Katunin,V.I., Semenkov,Y.P., and Wintermeyer,W. (1999). Thiostrepton inhibits the turnover but not the GTPase of elongation factor G on the ribosome. Proc. Natl. Acad. Sci. U. S. A *96*, 9586-9590.

Rodnina,M.V. and Wintermeyer,W. (2001a). Fidelity of aminoacyl-tRNA selection on the ribosome: kinetic and structural mechanisms. Annu. Rev. Biochem. *70*, 415-435.

Rodnina,M.V. and Wintermeyer,W. (2001b). Ribosome fidelity: tRNA discrimination, proofreading and induced fit. Trends Biochem. Sci. *26*, 124-130.

Rother,M., Resch,A., Wilting,R., and Bock,A. (2001). Selenoprotein synthesis in archaea. Biofactors *14*, 75-83.

Rould,M.A., Perona,J.J., Soll,D., and Steitz,T.A. (1989). Structure of E. coli glutaminyl-tRNA synthetase complexed with tRNA(Gln) and ATP at 2.8 A resolution. Science *246*, 1135-1142.

Rould,M.A., Perona,J.J., and Steitz,T.A. (1991). Structural basis of anticodon loop recognition by glutaminyl-tRNA synthetase. Nature *352*, 213-218.

Ruff,M., Krishnaswamy,S., Boeglin,M., Poterszman,A., Mitschler,A., Podjarny,A., Rees,B., Thierry,J.C., and Moras,D. (1991). Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexed with tRNA(Asp). Science *252*, 1682-1689.

Rustchenko,E. (2007). Chromosome instability in Candida albicans. FEMS Yeast Res. *7*, 2-11.

Rustchenko,E.P., Howard,D.H., and Sherman,F. (1994). Chromosomal alterations of Candida albicans are associated with the gain and loss of assimilating functions. J. Bacteriol. *176*, 3231-3241.

Rustchenko,E.P., Howard,D.H., and Sherman,F. (1997). Variation in assimilating functions occurs in spontaneous Candida albicans mutants having chromosomal alterations. Microbiology *143 (Pt 5)*, 1765-1778.

Rutherford,S.L. and Lindquist,S. (1998). Hsp90 as a capacitor for morphological evolution. Nature *396*, 336-342.

Ruusala,T., Ehrenberg,M., and Kurland,C.G. (1982). Is there proofreading during polypeptide synthesis? EMBO J. *1*, 741-745.

Saeed,A.I., Sharov,V., White,J., Li,J., Liang,W., Bhagabati,N., Braisted,J., Klapa,M., Currier,T., Thiagarajan,M., Sturn,A., Snuffin,M., Rezantsev,A., Popov,D., Ryltsov,A., Kostukovich,E., Borisovsky,I., Liu,Z., Vinsavich,A., Trush,V., and Quackenbush,J. (2003). TM4: a free, open-source system for microarray data management and analysis. Biotechniques *34*, 374-378.

Saks,M.E. and Sampson,J.R. (1996). Variant minihelix RNAs reveal sequence-specific recognition of the helical tRNA(Ser) acceptor stem by E.coli seryl-tRNA synthetase. EMBO J. *15*, 2843-2849.

Saks,M.E. and Sampson,J.R. (1995). Evolution of tRNA recognition systems and tRNA gene sequences. J. Mol. Evol. *40*, 509-518.

Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U., and Eisenberg,D. (2004). The Database of Interacting Proteins: 2004 update. Nucl. Acids Res. *32*, D449-D451.

Sambrook,J., Maniatis,T., and Fritsch,E.F. (1989). Molecular cloning a laboratory manual. (Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory).

Sampson,J.R. and Saks,M.E. (1993). Contributions of discrete tRNA(Ser) domains to aminoacylation by E.coli seryl-tRNA synthetase: a kinetic analysis using model RNA substrates. Nucleic Acids Res. *21*, 4467-4475.

Sampson,J.R. and Uhlenbeck,O.C. (1988). Biochemical and Physical Characterization of an Unmodified Yeast Phenylalanine Transfer RNA Transcribed in vitro. PNAS *85*, 1033-1037.

Sanchez-Silva,R., Villalobo,E., Morin,L., and Torres,A. (2003). A new noncanonical nuclear genetic code: translation of UAA into glutamate. Curr. Biol. *13*, 442-447.

Sankaranarayanan,R., Dock-Bregeon,A.C., Romby,P., Caillet,J., Springer,M., Rees,B., Ehresmann,C., Ehresmann,B., and Moras,D. (1999). The structure of threonyl-tRNA synthetase-tRNA(Thr) complex enlightens its repressor activity and reveals an essential zinc ion in the active site. Cell *97*, 371-381.

Santoro,S.W., Wang,L., Herberich,B., King,D.S., and Schultz,P.G. (2002). An efficient system for the evolution of aminoacyl-tRNA synthetase specificity. Nat. Biotechnol. *20*, 1044-1048.

Santos,M.A., Cheesman,C., Costa,V., Moradas-Ferreira,P., and Tuite,M.F. (1999). Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in Candida spp. Mol. Microbiol. *31*, 937-947.

Santos,M.A., el-Adlouni,C., Cox,A.D., Luz,J.M., Keith,G., and Tuite,M.F. (1994). Transfer RNA profiling: a new method for the identification of pathogenic Candida species. Yeast *10*, 625-636.

Santos,M.A., Keith,G., and Tuite,M.F. (1993). Non-standard translational events in Candida albicans mediated by an unusual seryl-tRNA with a 5'-CAG-3' (leucine) anticodon. EMBO J. *12*, 607-616.

Santos,M.A., Moura,G., Massey,S.E., and Tuite,M.F. (2004). Driving change: the evolution of alternative genetic codes. Trends Genet. *20*, 95-102.

Santos,M.A., Perreau,V.M., and Tuite,M.F. (1996). Transfer RNA structural change is a key element in the reassignment of the CUG codon in Candida albicans. EMBO J. *15*, 5060-5068.

Santos,M.A. and Tuite,M.F. (1995). The CUG codon is decoded in vivo as serine and not leucine in Candida albicans. Nucleic Acids Res. *23*, 1481-1486.

Santos,M.A., Ueda,T., Watanabe,K., and Tuite,M.F. (1997). The non-standard genetic code of Candida spp.: an evolving genetic code or a novel mechanism for adaptation? Mol. Microbiol. *26*, 423-431.

Sauerwald,A., Zhu,W., Major,T.A., Roy,H., Palioura,S., Jahn,D., Whitman,W.B., Yates,J.R., III, Ibba,M., and Soll,D. (2005). RNA-dependent cysteine biosynthesis in archaea. Science *307*, 1969-1972.

Schimmel,P., Giege,R., Moras,D., and Yokoyama,S. (1993). An operational RNA code for amino acids and possible relationship to genetic code. Proc. Natl. Acad. Sci. U. S. A *90*, 8763-8768.

Schimmel,P. and Ribas de,P.L. (1995). Transfer RNA: from minihelix to genetic code. Cell *81*, 983-986.

Schmitt,E., Panvert,M., Blanquet,S., and Mechulam,Y. (1998). Crystal structure of methionyl-tRNAfMet transformylase complexed with the initiator formyl-methionyl-tRNAfMet. EMBO J. *17*, 6819-6826.

Schneider,S.U. and de Groot,E.J. (1991). Sequences of two rbcS cDNA clones of Batophora oerstedii: structural and evolutionary considerations. Curr. Genet. *20*, 173-175.

Schroder,M. and Kaufman,R.J. (2005). The mammalian unfolded protein response. Annu. Rev. Biochem. *74*, 739-789.

Schulman,L.H. and Her,M.O. (1973). Recognition of altered E. coli formylmethionine transfer RNA by bacterial T factor. Biochem. Biophys. Res. Commun. *51*, 275-282.

Schulman,L.H. and Pelka,H. (1988). Anticodon switching changes the identity of methionine and valine transfer RNAs. Science *242*, 765-768.

Schultz,D.W. and Yarus,M. (1994). Transfer RNA mutation and the malleability of the genetic code. J. Mol. Biol. *235*, 1377-1380.

Scorer,C.A., Carrier,M.J., and Rosenberger,R.F. (1991). Amino acid misincorporation during high-level expression of mouse epidermal growth factor in Escherichia coli. Nucleic Acids Res. *19*, 3511-3516.

Seburn,K.L., Nangle,L.A., Cox,G.A., Schimmel,P., and Burgess,R.W. (2006). An active dominant mutation of glycyl-tRNA synthetase causes neuropathy in a Charcot-Marie-Tooth 2D mouse model. Neuron *51*, 715-726.

Sengupta,S. and Higgs,P.G. (2005). A unified model of codon reassignment in alternative genetic codes. Genetics *170*, 831-840.

Seong,B.L. and RajBhandary,U.L. (1987). Mutants of Escherichia coli formylmethionine tRNA: a single base change enables initiator tRNA to act as an elongator in vitro. Proc. Natl. Acad. Sci. U. S. A *84*, 8859-8863.

Sethi,A., O'Donoghue,P., and Luthey-Schulten,Z. (2005). Evolutionary profiles from the QR factorization of multiple sequence alignments. Proc. Natl. Acad. Sci. U. S. A *102*, 4045-4050.

Sharp,P.M. and Li,W.H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. *15*, 1281-1295.

Sharp,P.M. and Li,W.H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. *24*, 28-38.

Sharp,P.M., Tuohy,T.M., and Mosurski,K.R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. *14*, 5125-5143.

Siatecka,M., Rozek,M., Barciszewski,J., and Mirande,M. (1998). Modular evolution of the Glx-tRNA synthetase family--rooting of the evolutionary tree between the bacteria and archaea/eukarya branches. Eur. J. Biochem. *256*, 80-87.

Sikorski,R.S. and Hieter,P. (1989). A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. Genetics *122*, 19-27.

Silva,R.M., Paredes,J.A., Moura,G.R., Manadas,B., Lima-Costa,T., Rocha,R., Miranda,I., Gomes,A.C., Koerkamp,M.J., Perrot,M., Holstege,F.C., Boucherie,H., and Santos,M.A. (2007). Critical roles for a genetic code alteration in the evolution of the genus Candida. EMBO J. *26*, 4555-4565.

Slupska,M.M., Baikalov,C., Lloyd,R., and Miller,J.H. (1996). Mutator tRNAs are encoded by the Escherichia coli mutator genes mutA and mutC: a novel pathway for mutagenesis. Proc. Natl. Acad. Sci. U. S. A *93*, 4380-4385.

Soll,D.R. (2004). Mating-type locus homozygosis, phenotypic switching and mating: a unique sequence of dependencies in Candida albicans. Bioessays *26*, 10-20.

Sollars,V., Lu,X., Xiao,L., Wang,X., Garfinkel,M.D., and Ruden,D.M. (2003). Evidence for an epigenetic mechanism by which Hsp90 acts as a capacitor for morphological evolution. Nat. Genet. *33*, 70-74.

Soma,A. and Himeno,H. (1998). Cross-species aminoacylation of tRNA with a long variable arm between Escherichia coli and Saccharomyces cerevisiae. Nucleic Acids Res. *26*, 4374-4381.

Soma,A., Kumagai,R., Nishikawa,K., and Himeno,H. (1996). The anticodon loop is a major identity determinant of Saccharomyces cerevisiae tRNA(Leu). J. Mol. Biol. *263*, 707-714.

Sprinzl,M., Horn,C., Brown,M., Ioudovitch,A., and Steinberg,S. (1998). Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. *26*, 148-153.

Sprinzl,M. and Vassilenko,K.S. (2005). Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. *33*, D139-D140.

Srinivasan,G., James,C.M., and Krzycki,J.A. (2002). Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. Science *296*, 1459-1462.

Stahl,G., Salem,S.N., Chen,L., Zhao,B., and Farabaugh,P.J. (2004). Translational accuracy during exponential, postdiauxic, and stationary growth phases in Saccharomyces cerevisiae. Eukaryot. Cell *3*, 331-338.

Stansfield,I., Jones,K.M., Herbert,P., Lewendon,A., Shaw,W.V., and Tuite,M.F. (1998). Missense translation errors in Saccharomyces cerevisiae. J. Mol. Biol. *282*, 13-24.

Starzyk,R.M., Webster,T.A., and Schimmel,P. (1987). Evidence for dispensable sequences inserted into a nucleotide fold. Science *237*, 1614-1618.

Stathopoulos,C., Li,T., Longman,R., Vothknecht,U.C., Becker,H.D., Ibba,M., and Soll,D. (2000). One polypeptide with two aminoacyl-tRNA synthetase activities. Science *287*, 479-482.

Sugita,T. and Nakase,T. (1999). Non-universal usage of the leucine CUG codon and the molecular phylogeny of the genus Candida. Syst. Appl. Microbiol. *22*, 79-86.

Sugiyama,H., Ohkuma,M., Masuda,Y., Park,S.M., Ohta,A., and Takagi,M. (1995). In vivo evidence for non-universal usage of the codon CUG in Candida maltosa. Yeast *11*, 43-52.

Suzuki,T., Ueda,T., and Watanabe,K. (1996). A new method for identifying the amino acid attached to a particular RNA in the cell. FEBS Lett. *381*, 195-198.

Suzuki,T., Ueda,T., and Watanabe,K. (1997). The 'polysemous' codon--a codon with multiple amino acid assignment caused by dual specificity of tRNA identity. EMBO J. *16*, 1122-1134.

Suzuki,T., Ueda,T., Yokogawa,T., Nishikawa,K., and Watanabe,K. (1994). Characterization of serine and leucine tRNAs in an asporogenic yeast Candida cylindracea and evolutionary implications of genes for tRNA(Ser)CAG responsible for translation of a non-universal genetic code. Nucleic Acids Res. *22*, 115-123.

Szathmary,E. (1999). The origin of the genetic code: amino acids as cofactors in an RNA world. Trends Genet. *15*, 223-229.

Tao,J. and Frankel,A.D. (1992). Specific binding of arginine to TAR RNA. Proc. Natl. Acad. Sci. U. S. A *89*, 2723-2726.

Thanbichler,M. and Bock,A. (2002). The function of SECIS RNA in translational control of gene expression in Escherichia coli. EMBO J. *21*, 6925-6934.

Theobald-Dietrich,A., Frugier,M., Giege,R., and Rudinger-Thirion,J. (2004). Atypical archaeal tRNA pyrrolysine transcript behaves towards EF-Tu as a typical elongator tRNA. Nucleic Acids Res. *32*, 1091-1096.

Thompson,J.D., Higgins,D.G., and Gibson,T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673-4680.

Thompson,R.C. and Stone,P.J. (1977). Proofreading of the codon-anticodon interaction on ribosomes. Proc. Natl. Acad. Sci. U. S. A *74*, 198-202.

Thrower,J.S., Hoffman,L., Rechsteiner,M., and Pickart,C.M. (2000). Recognition of the polyubiquitin proteolytic signal. EMBO J. *19*, 94-102.

Tomita,K., Ueda,T., and Watanabe,K. (1999). The presence of pseudouridine in the anticodon alters the genetic code: a possible mechanism for assignment of the AAA lysine codon as asparagine in echinoderm mitochondria. Nucleic Acids Res. *27*, 1683-1689.

True,H.L., Berlin,I., and Lindquist,S.L. (2004). Epigenetic regulation of translation reveals hidden genetic variation to produce complex traits. Nature *431*, 184-187.

True,H.L. and Lindquist,S.L. (2000). A yeast prion provides a mechanism for genetic variation and phenotypic diversity. Nature *407*, 477-483.

Tsui,W.C. and Fersht,A.R. (1981). Probing the principles of amino acid selection using the alanyl-tRNA synthetase from Escherichia coli. Nucleic Acids Res. *9*, 4627-4637.

Tsurui,H., Kumazawa,Y., Sanokawa,R., Watanabe,Y., Kuroda,T., Wada,A., Watanabe,K., and Shirai,T. (1994). Batchwise purification of specific tRNAs by a solid-phase DNA probe. Anal. Biochem. *221*, 166-172.

Tuite,M.F. and Lindquist,S.L. (1996). Maintenance and inheritance of yeast prions. Trends Genet. *12*, 467-471.

Tujebajeva,R.M., Ransom,D.G., Harney,J.W., and Berry,M.J. (2000). Expression and characterization of nonmammalian selenoprotein P in the zebrafish, Danio rerio. Genes Cells *5*, 897-903.

Uptain,S.M. and Lindquist,S. (2002). Prions as protein-based genetic elements. Annu. Rev. Microbiol. *56*, 703-741.

Urbonavicius,J., Qian,Q., Durand,J.M., Hagervall,T.G., and Bjork,G.R. (2001). Improvement of reading frame maintenance is a common function for several tRNA modifications. EMBO J. *20*, 4863-4873.

Urbonavicius,J., Stahl,G., Durand,J.M., Ben Salem,S.N., Qian,Q., Farabaugh,P.J., and Bjork,G.R. (2003). Transfer RNA modifications that alter +1 frameshifting in general fail to affect -1 frameshifting. RNA. *9*, 760-768.

Uy,R. and Wold,F. (1977). Posttranslational covalent modification of proteins. Science *198*, 890-896.

Valle,M., Zavialov,A., Li,W., Stagg,S.M., Sengupta,J., Nielsen,R.C., Nissen,P., Harvey,S.C., Ehrenberg,M., and Frank,J. (2003). Incorporation of aminoacyl-tRNA into the ribosome as seen by cryo-electron microscopy. Nat. Struct. Biol. *10*, 899-906.

Vazquez-Torres,A. and Balish,E. (1997). Macrophages in resistance to candidiasis. Microbiol. Mol. Biol. Rev. *61*, 170-192.

Volkenstein,M.V. (1966). The genetic coding of protein structure. Biochim. Biophys. Acta *119*, 421-424.

Wagner,E.G., Jelenc,P.C., Ehrenberg,M., and Kurland,C.G. (1982). Rate of elongation of polyphenylalanine in vitro. Eur. J. Biochem. *122*, 193-197.

Wakasugi,K. and Schimmel,P. (1999). Highly differentiated motifs responsible for two cytokine activities of a split human tRNA synthetase. J. Biol. Chem. *274*, 23155-23159.

Wang,L., Brock,A., Herberich,B., and Schultz,P.G. (2001). Expanding the genetic code of Escherichia coli. Science *292*, 498-500.

Wang,L., Haeusler,R.A., Good,P.D., Thompson,M., Nagar,S., and Engelke,D.R. (2005). Silencing near tRNA genes requires nucleolar localization. J. Biol. Chem. *280*, 8637-8639.

Watanabe,K., Kagaya,K., Yamada,T., and Fukazawa,Y. (1991). Mechanism for candidacidal activity in macrophages activated by recombinant gamma interferon. Infect. Immun. *59*, 521-528.

Watanabe,Y., Tsurui,H., Ueda,T., Furushima,R., Takamiya,S., Kita,K., Nishikawa,K., and Watanabe,K. (1994). Primary and higher order structures of nematode (Ascaris suum) mitochondrial tRNAs lacking either the T or D stem. J. Biol. Chem. *269*, 22902-22906.

Weiss,R.B., Huang,W.M., and Dunn,D.M. (1990). A nascent peptide is required for ribosomal bypass of the coding gap in bacteriophage T4 gene 60. Cell *62*, 117-126.

Weissman,A.M. (2001). Themes and variations on ubiquitylation. Nat. Rev. Mol. Cell Biol. *2*, 169-178.

Wilson,F.H., Hariri,A., Farhi,A., Zhao,H., Petersen,K.F., Toka,H.R., Nelson-Williams,C., Raja,K.M., Kashgarian,M., Shulman,G.I., Scheinman,S.J., and Lifton,R.P. (2004). A cluster of metabolic defects caused by mutation in a mitochondrial tRNA. Science *306*, 1190-1194.

Woese,C.R. (2002). On the evolution of cells. Proc. Natl. Acad. Sci. U. S. A *99*, 8742-8747.

Woese,C.R. (1965b). Order in the genetic code. Proc. Natl. Acad. Sci. U. S. A *54*, 71-75.

Woese,C.R. (1965a). On the evolution of the genetic code. Proc. Natl. Acad. Sci. U. S. A *54*, 1546-1552.

Woese,C.R., Dugre,D.H., Saxinger,W.C., and Dugre,S.A. (1966). The molecular basis for the genetic code. Proc. Natl. Acad. Sci. U. S. A *55*, 966-974.

Woese,C.R., Kandler,O., and Wheelis,M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. U. S. A *87*, 4576-4579.

Woese,C.R., Olsen,G.J., Ibba,M., and Soll,D. (2000). Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol. Mol. Biol. Rev. *64*, 202-236.

Wolf,Y.I., Aravind,L., Grishin,N.V., and Koonin,E.V. (1999). Evolution of aminoacyl-tRNA synthetases--analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res. *9*, 689-710.

Wong,J.T. (1975). A co-evolution theory of the genetic code. Proc. Natl. Acad. Sci. U. S. A *72*, 1909-1912.

Wong,J.T. and Bronskill,P.M. (1979). Inadequacy of prebiotic synthesis as origin of proteinous amino acids. J. Mol. Evol. *13*, 115-125.

Woo,N.H., Roe,B.A., and Rich,A. (1980). Three-dimensional structure of Escherichia coli initiator tRNAfMet. Nature *286*, 346-351.

Yang,Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. *13*, 555-556.

Yarus,M. (1972). Phenylalanyl-tRNA synthetase and isoleucyl-tRNA Phe : a possible verification mechanism for aminoacyl-tRNA. Proc. Natl. Acad. Sci. U. S. A *69*, 1915-1919.

Yarus,M. (1998). Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin. J. Mol. Evol. *47*, 109-117.

Yarus,M., Caporaso,J.G., and Knight,R. (2005). Origins of the genetic code: the escaped triplet theory. Annu. Rev. Biochem. *74*, 179-198.

Yasukawa,T., Suzuki,T., Ishii,N., Ohta,S., and Watanabe,K. (2001). Wobble modification defect in tRNA disturbs codon-anticodon interaction in a mitochondrial disease. EMBO J. *20*, 4794-4802.

Yasukawa,T., Suzuki,T., Ueda,T., Ohta,S., and Watanabe,K. (2000). Modification defect at anticodon wobble nucleotide of mitochondrial tRNAs(Leu)(UUR) with pathogenic mutations of mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes. J. Biol. Chem. *275*, 4251-4257.

Yewdell,J.W. and Nicchitta,C.V. (2006). The DRiP hypothesis decennial: support, controversy, refinement and extension. Trends Immunol. *27*, 368-373.

Yip,C.L., Welch,S.K., Klebl,F., Gilbert,T., Seidel,P., Grant,F.J., O'Hara,P.J., and MacKay,V.L. (1994). Cloning and Analysis of the Saccharomyces cerevisiae MNN9 and MNN1 Genes Required for Complex Glycosylation of Secreted Proteins. PNAS *91*, 2723-2727.

Yokobori,S., Suzuki,T., and Watanabe,K. (2001). Genetic code variations in mitochondria: tRNA as a major determinant of genetic code plasticity. J. Mol. Evol. *53*, 314-326.

Yokogawa,T., Suzuki,T., Ueda,T., Mori,M., Ohama,T., Kuchino,Y., Yoshinari,S., Motoki,I., Nishikawa,K., Osawa,S., and . (1992). Serine tRNA complementary to the nonuniversal serine codon CUG in Candida cylindracea: evolutionary implications. Proc. Natl. Acad. Sci. U. S. A *89*, 7408-7411.

Yokoyama,S., Watanabe,T., Murao,K., Ishikura,H., Yamaizumi,Z., Nishimura,S., and Miyazawa,T. (1985). Molecular mechanism of codon recognition by tRNA species with modified uridine in the first position of the anticodon. Proc. Natl. Acad. Sci. U. S. A *82*, 4905-4909.

Zavialov,A.V., Buckingham,R.H., and Ehrenberg,M. (2001). A posttermination ribosomal complex is the guanine nucleotide exchange factor for peptide release factor RF3. Cell *107*, 115-124.

Zhao,L. and Ackerman,S.L. (2006). Endoplasmic reticulum stress in health and disease. Curr. Opin. Cell Biol. *18*, 444-452.

Zhao,M.W., Zhu,B., Hao,R., Xu,M.G., Eriani,G., and Wang,E.D. (2005). Leucyl-tRNA synthetase from the ancestral bacterium Aquifex aeolicus contains relics of synthetase evolution. EMBO J. *24*, 1430-1439.

Zinoni,F., Birkmann,A., Leinfelder,W., and Bock,A. (1987). Cotranslational insertion of selenocysteine into formate dehydrogenase from Escherichia coli directed by a UGA codon. Proc. Natl. Acad. Sci. U. S. A *84*, 3156-3160.