AGE TATS

Sequence motifs influencing the efficiency of translation

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original publications which will be referred to in the text by their Roman numerals:

I. Vimberg, V., **Tats, A**., Remm, M. and Tenson, T. (2007) Translation initiation region sequence preferences in Escherichia coli. *BMC Molecular Biology*, 8:100.

II. **Tats, A**., Remm, M. and Tenson, T. (2006) Highly expressed proteins have an increased frequency of alanine in the second amino acid position. *BMC Genomics*, 7:28.

III. **Tats, A**., Tenson, T. And Remm, M. (2008) Preferred and avoided codon pairs in three domains of life. *BMC Genomics*, 9:463.

My contributions to the articles:

Ref I: performed *in silico* analysis and participated in the preparation of the manuscript;

Ref II: performed data analysis and participated in the writing of the manuscript;

Ref III: performed data analysis and participated in the writing of the manuscript.

LIST OF ABBREVIATIONS

antiSD	anti Shine-Dalgarno
CAI	codon adaptation index
COA	correspondence analysis
fMet	formyl-methionine
HEG	highly expressed genes
HGT	horizontal gene transfer
MAP	methionine aminopeptidase
N-terminal	amino-terminal
ORF	open reading frame
ORFeome	all protein coding sequences of an organism
RDCU	relative dicodon usage
RSCU	relative synonymous codon usage
SD	Shine-Dalgarno

INTRODUCTION

Proteins participate in every process of a cell. Those essential macromolecules are formed during translation process where amino acids are joined into proteins based on the information encoded within mRNA. Translation as the last phase of gene expression takes place in a large complex – the ribosome. The general structure and function of the ribosome is highly conserved in all living organisms. The conservation of basic mechanisms allows presuming that signals responsible for the regulation of the translation could share common motifs. Therefore, the discovery of new conserved motifs could suggest aspects of translational regulation not known so far.
The growing number of sequenced genomes in recent years has provided invaluable resource for comparative genomics studies. By using computational methods, all this data can be analysed to shed new light to the regulation of translation starting from the thorough analysis of the regulatory and coding sequences of one genome extending to the search for conserved motifs in genomes belonging to different domains of life.

In the literature part of current thesis short introduction to the mechanism of translation and the known important sequence elements responsible for its efficiency is made. Secondly, amino-terminally located motifs responsible for the stabilization and the degradation of a protein are discussed. The third part of literature review gives an overview of the codon usage and codon context biases reflecting the selection for efficient translation in coding sequences of a genome.

In research part of current thesis I introduce some new aspects of protein synthesis mechanism we have discovered by using bioinformatical methods. Two of the three articles in this thesis are belonging entirely to the field of comparative genomics. The research part is focused on: 1) the relationship between Shine-Dalgarno sequence base pairing potential and gene expression levels in *Escherichia coli*; 2) characterization of conserved sequence motifs at the beginning of highly expressed genes and 3) characterization of universally biased codon pairs in coding sequences of different genomes.

1. REVIEW OF LITERATURE

1.1 The mechanism of translation

The machinery of translation – the ribosome – consists of two unequal subunits with three tRNA binding sites called E-, P- and A-sites based on the type of binding tRNA: deacylated tRNA, peptidyl-tRNA and aminoacyl-tRNA (Yusupov *et al.*, 2001; Selmer *et al.*, 2006). The correct functioning of those three sites is one of the main aspects responsible for the accuracy of protein synthesis.

Initiation

The efficiency of translation depends heavily on the initiation stage of translation. During initiation two ribosomal subunits are joined on mRNA and initiator-tRNA binds to initiation codon in the P-site of the ribosome with the help of initiation factors.

For effective recognition of the translation initiation region by the ribosome this region includes several determinants for the location and the efficiency of translational start. Interestingly, despite of the evolutionary conservation of the translation the initiation stage has extensive differences between bacteria and eukaryotes. This is reflected also in different translation initiation region determinants.

In bacteria the small subunit of the ribosome in complex of several initiation factors directly recognizes the translation initiation region. Upstream of the initiation codon is located a ribosomal binding site containing Shine-Dalgarno (SD) sequence (Shine and Dalgarno, 1974; Shultzaberger *et al.*, 2001). The SD sequence base pairs with the anti Shine-Dalgarno (antiSD) sequence on the 16S rRNA 3' terminal end (Shine and Dalgarno, 1974). The length of SD:antiSD duplex can vary. There was no full-scale analysis of SD region length in *Escherichia coli* genes, but the average number of paired nucleotides in 1159 *E. coli* genes was shown to be 6.3 (Schurr *et al.*, 1993). SD sequences longer than six nucleotides are not very efficient, probably because more time is needed for clearance of translation initiation region (de Boer *et al.*, 1983; Komarova *et al.*, 2002). Indeed, the average SD sequence length in highly expressed ribosomal protein genes is 4.4 nucleotides (Komarova *et al.*, 2002). A significant positive correlation between the presence of SD sequence and the predicted expression level of a gene was reported in 30 prokaryotic genomes analysed *in silico* (Ma *et al.*, 2002). Unfortunately, the influence of the SD:antiSD interaction strength to the expression level was not analysed. Only a weak correlation between free energy of SD:antiSD interaction and translational efficiency was found in experimental analysis (Lee *et al.*, 1996).

The distance between the SD sequence and initiation codon (the spacing) has large effect on the efficiency of translation. Too long or too short spacer region may inhibit the efficient translation (Shine and Dalgarno, 1975; Chen *et al.*, 1994). The optimal spacing varies from 5 to 13 nucleotides (Ringquist *et al.*, 1992; Chen *et al.*, 1994). Some studies have used the term 'aligned spacing' which defines the region between the reference SD sequence (5' – UAAGGAGGU – 3') and the initiation

codon (Ringquist *et al.*, 1992; Chen *et al.*, 1994). Aligned spacing of 5 nucleotides is shown to be the most optimal (Chen *et al.*, 1994).

Another important element, A/U rich enhancer sequence in front of the SD sequence contributes to the effectiveness of translation (Komarova *et al.*, 2002; Komarova *et al.*, 2005). This sequence can act as a standby binding site for the small ribosomal subunit (de Smit and van Duin, 2003; Studer and Joseph, 2006).

There is no SD-sequence in eukaryotes. Instead, in eukaryotes the small ribosomal subunit first binds with the help of numerous additional proteins to the 5' end of the mRNA and then scans towards the 3' end until the initiation codon is encountered (Kozak, 1989). The efficiency of translation is reduced if the sequence surrounding the AUG codon deviates significantly from certain preferred nucleotides. For example, the nucleotide context at the beginning of *Saccharomyces cerevisiae* HEG is shown to be AUGUC(U/C) (Hamilton *et al.*, 1987; Miyasaka, 1999; Fuglsang, 2004). The so-called Kozak consensus sequence GCC(A/G)CCAUGG was obtained from 699 vertebrate genes (Kozak, 1987, 1997). Later it was revealed that preferred nucleotide sequences around initiation codon are quite diverse among different eukaryotes (Cavener and Ray, 1991). However, the G nucleotide following the initiation codon (+4G) was still present in most of the studied eukaryotic species. In addition, the bias for C nucleotide at position +5 has been described in eukaryotic genomes (Nakagawa *et al.*, 2008).

Archaeal translation initiation shares characteristic features to both – bacterial and eukaryotic translation initiation. Archaeal translation initiation factors are homologous to those of eukaryotes (Kyrpides and Woese, 1998). Some archaea, *e.g Sulfolobus solfataricus*, use two distinct mechanisms for translation initiation: SD-dependent initiation operates on distal cistrons of polycistronic mRNAs, whereas 'leaderless' initiation operates on monocistronic mRNAs and on opening cistrons of polycistronic mRNAs which start directly with the initiation codon (Benelli *et al.*, 2003). In addition to archaea, leaderless mRNAs which are lacking entirely 5'-untranslated region have been identified in bacteria and eukaryotes (Jannsen, 1993). In case of leaderless initiation, codon-anticodon interaction between initiator-tRNA and the initiation codon appears to be necessary for efficient binding of small ribosomal subunit to the 5' extremity of the leaderless mRNA (Grill *et al.*, 2000; Benelli *et al.*, 2003).

Elongation

Elongation stage is very similar in prokaryotes and eukaryotes. During the elongation ribosome moves along the mRNA being assisted by several elongation factors for incorporating amino acids into the growing polypeptide chain. Elongation starts with peptidyl-tRNA (in case of first elongation step it is initiator-tRNA carrying methionine or formyl-methionine) in the P-site. Aminoacyl-tRNA binds to its complimentary codon of mRNA in vacant A-site. During the peptidyl transferase reaction the polypeptide chain from the peptidyl-tRNA in P-site is transferred to the amino acid of the aminoacyl-tRNA in the A-site and the polypeptide is extended by one amino acid. Former peptidyl-tRNA becomes deacylated. During translocation when ribosome moves ahead on the mRNA by one codon, tRNA carrying the nascent peptide is moved from the A-site to the P-site leaving A-site free for the next aminoacyl-tRNA. Deacylated-tRNA leaves through the E-site from the ribosome.

The binding of deacylated-tRNA to the E-site plays fundamental role in maintaining the reading frame. Ribosomes where deacylated-tRNA binding to E-site is compromised by mutations have increased frameshifting frequencies (Sergiev *et al.*, 2005). Correct reading frame is achieved through codon-anticodon binding and allosteric linkage with the A-site. Namely, an occupied E-site induces a low-affinity A-site and an occupation of the A-site triggers the release of the E-site tRNA (Geigenmuller and Nierhaus, 1990; Marquez *et al.*, 2004; Trimble *et al.*, 2004).

Important role in achieving the accurate and efficient translation lies on the translated mRNA sequence itself. The impact of codon usage and codon context usage to the translational efficiency and accuracy is surveyed in Chapters 1.3.4 and 1.4.

Termination

Elongation continues until ribosome reaches a stop codon. In bacteria the most frequently used stop codon is UAA (Sharp and Bulmer, 1988). The stop codon usage bias exists also in eukaryotes. In lower eukaryotes like fungi and invertebrates UAA is preferred while the most over-represented stop codon in plants and mammals is UGA (Sun *et al.*, 2005). Correct recognition of stop codon is another critical stage of protein synthesis. Stop codon read-through leads to the translation beyond the natural end of the coding sequence. This will lead to the non-functional protein product which in most cases is harmful to the cell due to its misfolding and aggregation with other misfolded proteins. In addition, the degradation of such non-functional proteins wastes the energetic resources of the cell. On the other hand, stop codon read through might be used for regulatory purposes. In case of yeast [PSI+] phenotype the change in the conformation and function of the translation termination factor has led to the increased read-through of stop codons creating the phenotypes more tolerant in certain ecological niches (Uptain and Lindquist, 2002; True *et al.*, 2004).

The sequence context around stop codon plays important role in efficient translation termination. Numerous studies have assigned the sequence immediately following the stop codon as the most crucial determinant of accurate translation termination creating so-called extended stop signal (Poole *et al.*, 1995; Tate *et al.*, 1996). Specific context varies in different organisms, but A or G as nucleotides 3' from the stop codon are shown to be preferred (Brown *et al.*, 1990; Sun *et al.*, 2005). In prokaryotes interaction between 3' nucleotide and release factor 2 is shown (Poole *et al.*, 1998). The effect of 3' sequence can involve even much longer region. In *S. cerevisiae* up to six nucleotides after stop codon can determine the stop codon read-through efficiency (Namy *et al*., 2001).

Upstream sequences have weaker role. Although in bacteria and baker's yeast the nature of the last amino acids in synthesized protein has been related to the termination efficiency (Mottagui-Tabar *et al.*, 1998), later analysis in *S. cerevisiae* and *Neurospora grassa* did not find significant bias in 5' codons from the stop codon (Williams *et al.*, 2004).

Stop codon is recognized by the release factor which terminates the translation. In bacteria stop codons UAG and UAA are recognized by the release factor 1; release factor 2 recognizes UGA and UAA stops (Caskey, 1977; Kisselev and Buckingham, 2000; Kisselev *et al.*, 2003). In eukaryotes one release factor recognizes all three stop codons (Konecki *et al.*, 1977; Frolova *et al.*, 1994). As a result, the polypeptide chain is released from the tRNA and leaves the ribosome.

1.2 N-terminal signals in protein sequences

Several signals important for influencing protein half-life and functionality are located at the beginning of proteins. These include longer N-terminal signal peptides determining the subcellular location of proteins but also smaller signals. Among posttranslational modifications the N-terminal modifications are the most common processing events. The identity of amino acid residue following the starting methionine is important determinant of methionine removal and the stability of the protein. Such signals could influence the conservation level of the beginning of protein coding genes and resulting proteins.

1.2.1 Signals determining the cleavage on amino-terminal methionine residue

During the start of bacterial protein synthesis the fMet is incorporated to amino-terminus of the polypeptide (Kozak, 1983; Meinnel *et al.*, 1993; Schmitt *et al.*, 1996). During the following elongation cycle it is processed. Firstly, N-formyl part is removed with deformylase resulting with the methionine in the amino-terminus. In large number of proteins this methionine is also removed (Sherman *et al.*, 1985). The cleavage of the amino-terminal methionine depends on the identity of the following amino acid residue. The corresponding enzyme, methionine aminopeptidase (MAP), cleaves methionine in case it is situated in front of Ala, Gly, Pro, Ser, Thr or Val; the methionine remains intact in case the following amino acid is Arg, Asn, Asp, Gln, Glu, Ile, Leu, Lys or Met (Tsunasawa *et al.*, 1985; Ben-Bassat *et al.*, 1987; Miller *et al.*, 1987; Moerschell *et al.*, 1990). Usually the cleavage promoting residues have short side-chain; MAP is not able to remove the methionine in case of residues with long or bulky side-chains (Hirel *et al.*, 1989; Dalboge *et al.*, 1990; Schmitt *et al.*, 1996).

The cleavage of amino-terminal methionine occurs also in eukaryotes and archaea. *S. cerevisiae* have two different MAPs with varied cleavage specificity against the same substrates but still making the cleavage only in case of amino acids with small side chains (Chen *et al.*, 2002). Archaeal MAPs are evolutionarily located in the borderline between bacteria and eukaryotes and also having similar substrate specificity for small amino acids (Falb *et al.*, 2006).

1.2.2 Protein stabilization and destabilization signals

Certain residues at the N-terminal part of the protein direct the protein into degradation. Those residues are described by the N-end rule and define the life span of the protein (Varshavsky, 1996). This regulated proteolysis is conserved from bacteria to mammals. Despite distinct proteolytic machineries, the recognition of the substrate shares common principles. Also, prokaryotes and eukaryotes have a common set of amino acids acting as stabilizing or destabilizing N-terminal residues (Table 1).

Table 1. Eukaryotic and bacterial N-end rules (based on (Varshavsky, 1996; Tasaki and Kwon, 2007)). ○ – stabilizing residue; ● – destabilizing residue.

	F	L	W	Y	R	K	H	I	N	Q	D	E	C	A	S	T	G	V	P	M
E.coli	●	●	●	●	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○
S.cerevisiae	●	●	●	●	●	●	●	●	●	●	●	●	○	○	○	○	○	○	○	○
Mammals	●	●	●	●	●	●	●	●	●	●	●	●	●	○	○	○	○	○	○	○

1.3 Codon usage bias

The genetic information within the DNA is transferred to the protein sequences via mRNA. The rules by which codons in mRNA are translated into amino acids in protein are specified by the genetic code. One of the main characteristics of the genetic code is degeneracy – more than one codon is specifying the same amino acid. Codons coding for the same amino acid are called synonymous codons. According to standard genetic code 18 out of the 20 amino acids have synonymous codons, only methionine and tryptophan are coded by one codon. Because of the degeneracy it would be expected that all synonymous codons coding for the same amino acid are distributed randomly and equally in protein coding sequences. In fact, this is not the case. Some synonymous codons are used more frequently and preferred over others. This phenomenon is called synonymous codon usage bias or simply codon usage bias.

A simple measure for evaluating the codon usage bias is the relative synonymous codon usage index (RSCU) (Sharp *et al.*, 1986). The RSCU is the observed frequency of a codon divided by the frequency expected if all synonymous codons for a specific amino acid were used equally. RSCU value 1.0 indicates the lack of codon usage bias. A codon that is used more frequently than expected has RSCU value larger than 1.0 and a codons that is used less frequently than expected has RSCU value smaller than 1.0.

Biased codon usage can be the result of different factors like genomic GC-content, strand specific mutational bias, horizontal gene transfer and translational selection and it varies among genomes, among genes and within genes.

1.3.1 GC-content related codon usage bias

Prokaryotes present wide variations in genomic GC-content. Among sequenced bacterial genomes it varies from 16.5% in *Carsonella ruddii* (Nakabachi *et al.*, 2006) to 74.9% in *Anaeromyxobacter dehalogenans* (Sanford *et al.*, 2002). This inter-species variation has been related mainly to the mutation driven process (Lobry, 1997; Singer and Hickey, 2000) but the adaptation to environmental conditions (mainly in termophilic bacteria) have also been suggested (Bernardi and Bernardi, 1986; Musto *et al.*, 2004). However, although the GC-content of structural RNAs (tRNAs, rRNAs) and growth temperature are highly correlated (Hurst and Merchant, 2001; Das *et al.*, 2006) the genomic GC-content as a whole does not correlate with the growth temperature (Hurst and Merchant, 2001). In hyperthermophilic archaea

Nanoarchaeum equitans the protein coding sequences have obtained the overrepresentation of purines (Das *et al.*, 2006).

The wide variation of GC-content is reflected in codon usage as well. Organisms with high genomic GC-content show clear preference for G or C ending synonymous codons and *vice versa*, protein coding sequences in AT-rich genomes have codon usage biased towards A and T ending synonymous codons (Table 2).

1.3.2 Replicational-transcriptional selection

The majority of the genes in bacterial genomes are located on the leading strand (Rocha, 2002). In addition, the leading strand contains more HEG than lagging strand (Nomura and Morgan, 1977; Brewer, 1988). Such strand biases are suggested to be related with the maintenance of the speed of the replication fork and reduced interruptions of gene expression. Namely, on leading strand the replication and transcription occur in the same direction and this minimizes the collisions of DNA and RNA polymerases (Nomura and Morgan, 1977; Brewer, 1988; Rocha, 2002; Price *et al.*, 2005).

The difference of nucleotide compositions between leading and lagging DNA strands could also create variation of codon usage. Genes on the leading strand are often more GT-rich. Such strand specific codon bias is observed especially in spirochaetes (Lobry, 1996; Lafay *et al.*, 1999). Since leading and lagging strand are replicated by different mechanisms (Kornberg and Baker, 1992), the structure of the replication fork creates the situation where lagging strand is longer in a single-stranded structure than leading strand (Marians, 1992) and thus more exposed to the possible DNA damage. Similarly, during transcription coding strand is transiently exposed and more sensitive to certain mutations such as C to T deamination (Beletskii and Bhagwat, 1996).

1.3.3 Horizontal gene transfer and codon usage bias

During horizontal gene transfer (HGT) the genetic material can be passed from one organism to another independent of their phylogenetic distance (Akiba *et al.*, 1960). HGT is common between bacteria and thought to be the main mechanism creating the increased drug resistance. The examples of HGT are also known in eukaryotes (Hall *et al.*, 2005).

Atypical nucleotide compositions and species specific differences of codon usage allow discriminating between horizontally transferred genes and the genes of the host genomes (Kaplan and Fine, 1998; Moszer *et al.*, 1999; Garcia-Vallve *et al.*, 2000; Ochman *et al.*, 2000). Horizontally transferred genes might have different codon usage from the host since they descend from a different background. Thus the sequence can provide a clue about their origin. However, in case of very similar GC-content of donor and acceptor genomes or already adjusted codon usage between transferred genes and host (process called ‘amelioration’) (Lawrence and Ochman, 1997) the detection of horizontally transferred genes could be quite complicated.

Table 2. The RSCU values of two genomes with different GC-content. The most preferred codon for each amino acid is highlighted. *Mycoplasma capricolum* as AT-rich genome prefers AT-rich synonymous codons and *Frankia alni* as GC-rich genome prefers GC-rich synonymous codons.

Mycoplasma capricolum GC = 24%	TTT	Phe	1.89	TCT	Ser	1.58	TAT	Tyr	1.81	TGT	Cys	1.76
	TTC		0.11	TCC		0.03	TAC		0.19	TGC		0.24
	TTA	Leu	4.46	TCA		2.11	TAA	Stop	0.50	TGA	Stop	2.35
	TTG		0.31	TCG		0.05	TAG	Stop	0.15	TGG	Trp	1.00
	CTT	Leu	0.52	CCT	Pro	1.54	CAT	His	1.60	CGT	Arg	0.73
	CTC		0.01	CCC		0.09	CAC		0.40	CGC		0.08
	CTA		0.68	CCA		2.30	CAA	Gln	1.91	CGA		0.12
	CTG		0.03	CCG		0.06	CAG		0.09	CGG		0.00
	ATT	Ile	2.17	ACT	Thr	2.45	AAT	Asn	1.70	AGT	Ser	1.96
	ATC		0.14	ACC		0.11	AAC		0.30	AGC		0.26
	ATA		0.69	ACA		1.42	AAA	Lys	1.81	AGA	Arg	4.92
	ATG	Met	1.00	ACG		0.02	AAG		0.19	AGG		0.15
	GTT	Val	2.63	GCT	Ala	2.42	GAT	Asp	1.82	GGT	Gly	1.83
	GTC		0.08	GCC		0.11	GAC		0.18	GGC		0.10
	GTA		1.14	GCA		1.42	GAA	Glu	1.85	GGA		1.88
	GTG		0.14	GCG		0.06	GAG		0.15	GGG		0.18

Frankia alni GC=72%	TTT	Phe	0.10	TCT	Ser	0.11	TAT	Tyr	0.20	TGT	Cys	0.25
	TTC		1.90	TCC		1.85	TAC		1.80	TGC		1.75
	TTA	Leu	0.01	TCA		0.16	TAA	Stop	0.13	TGA	Stop	2.35
	TTG		0.23	TCG		1.99	TAG	Stop	0.53	TGG	Trp	1.00
	CTT	Leu	0.16	CCT	Pro	0.13	CAT	His	0.34	CGT	Arg	0.43
	CTC		2.12	CCC		1.36	CAC		1.66	CGC		2.46
	CTA		0.06	CCA		0.16	CAA	Gln	0.10	CGA		0.31
	CTG		3.41	CCG		2.35	CAG		1.90	CGG		2.55
	ATT	Ile	0.13	ACT	Thr	0.10	AAT	Asn	0.15	AGT	Ser	0.22
	ATC		2.82	ACC		2.47	AAC		1.85	AGC		1.68
	ATA		0.05	ACA		0.12	AAA	Lys	0.15	AGA	Arg	0.05
	ATG	Met	1.00	ACG		1.31	AAG		1.85	AGG		0.21
	GTT	Val	0.12	GCT	Ala	0.12	GAT	Asp	0.25	GGT	Gly	0.46
	GTC		2.20	GCC		2.12	GAC		1.75	GGC		2.30
	GTA		0.07	GCA		0.16	GAA	Glu	0.28	GGA		0.28
	GTG		1.62	GCG		1.60	GAG		1.72	GGG		0.96

1.3.4 Translational selection

The selection for translational efficiency (or simply translational selection) related to codon usage bias has particularly attracted the attention of researchers. Due to the energetic cost of protein synthesis, inaccurate and inefficient translation is a very pricy event for the cellular resources. The codon usage bias can reduce that cost by creating sequences consisting of optimal codons.

The synonymous codon usage in bacteria *E. coli*, *Bacillus subtilis* as well as in eukaryotes *S. cerevisiae*, *Drosophila melanogaster* and *Caenorhabditis elegans*, is in correlation with the amount of tRNA isoacceptors – more frequently occurring codons are read by the more abundant isoacceptors (Ikemura, 1981, 1985; Dong *et al.*, 1996;

Moriyama and Powell, 1997; Percudani *et al.*, 1997; Kanaya *et al.*, 1999; Duret, 2000). In other words, an organism prefers to use codons which are more rapidly translated because the ribosome does not have to pause for waiting the tRNAs. In addition, such bias lowers more the chance for incorrect tRNA attachment than in the case when frequencies of synonymous codons and concentrations of tRNA isoacceptors are more evenly distributed (Ehrenberg and Kurland, 1984). Indeed, it was identified in *E. coli* that the usage of alternative synonymous codons could be biased in order to reduce the costs of energy and resources resulting from the nonsense and missense errors during translation. It appeared that longer protein coding genes had more biased codon usage (Stoletzki and Eyre-Walker, 2007). Since synthesis of longer proteins spends more resources, the selection for optimal codons in longer genes has important effect.

In addition, translational selection appears as different usage of synonymous codons in genes with high and low expression levels. HEG have usually more biased codon usage while lowly expressed genes have more uniform codon usage (Sharp and Li, 1987) (Figure 1).

The observations of biased codon usage in HEG have led to the creation of codon adaptation index (CAI). CAI is a numerical value which characterizes the similarity of synonymous codon usage in a given gene to that in the HEG (Sharp and Li, 1987). The group of HEG consists of genes coding for *e.g* ribosomal proteins, outer membrane proteins, elongation factors, heat-shock proteins and RNA polymerase subunits. Therefore, CAI can be used for predicting the gene expression level and identifying the HEG in a given genome. CAI values vary between 0 and 1. A CAI value of 1 is achieved when all amino acids in a given protein are coded by the best codon in each synonymous codon family. The correlation between CAI and experimental gene expression level is well documented (Futcher *et al.*, 1999; Coghlan and Wolfe, 2000; dos Reis *et al.*, 2003; Jansen *et al.*, 2003; Lithwick and Margalit, 2003; Jia and Li, 2005).

Selection for translational efficiency usually exists in fast growing prokaryotes and eukaryotes (Sharp *et al.*, 1986; Shields and Sharp, 1987; Stenico *et al.*, 1994) but is also described in plants (Fennoy and Bailey-Serres, 1993; Chiapello *et al.*, 1998). In human genome the evidence is less clear as the large scale variation of GC-content or so-called isochoric structure of the human genome appears to be the main influence of codon composition (Vinogradov, 2003). However, weak positive correlation between gene expression levels and the frequency of optimal codons has also been found in humans (Kotlar and Lavner, 2006).

Translational selection might act differently along the protein coding genes. In *E. coli* it is shown that the first part of protein coding sequences has more biased codon usage than the middle and final part, independently of sequence length (Karlin *et al.*, 1998). The influence of the +2 codon to the translational efficiency has been measured (initiation codon being marked as +1). It appeared, that 15-20-fold effect can be obtained by varying this codon in the mRNA coding sequence; in *E. coli* AAA is the most common and most expression promoting codon in position +2 (Stenstrom et al., 2001). Conversely, NGG codons in positions +2, +3 or +5 give strongly reduced gene expression (Gonzalez de Valdivia and Isaksson, 2004).

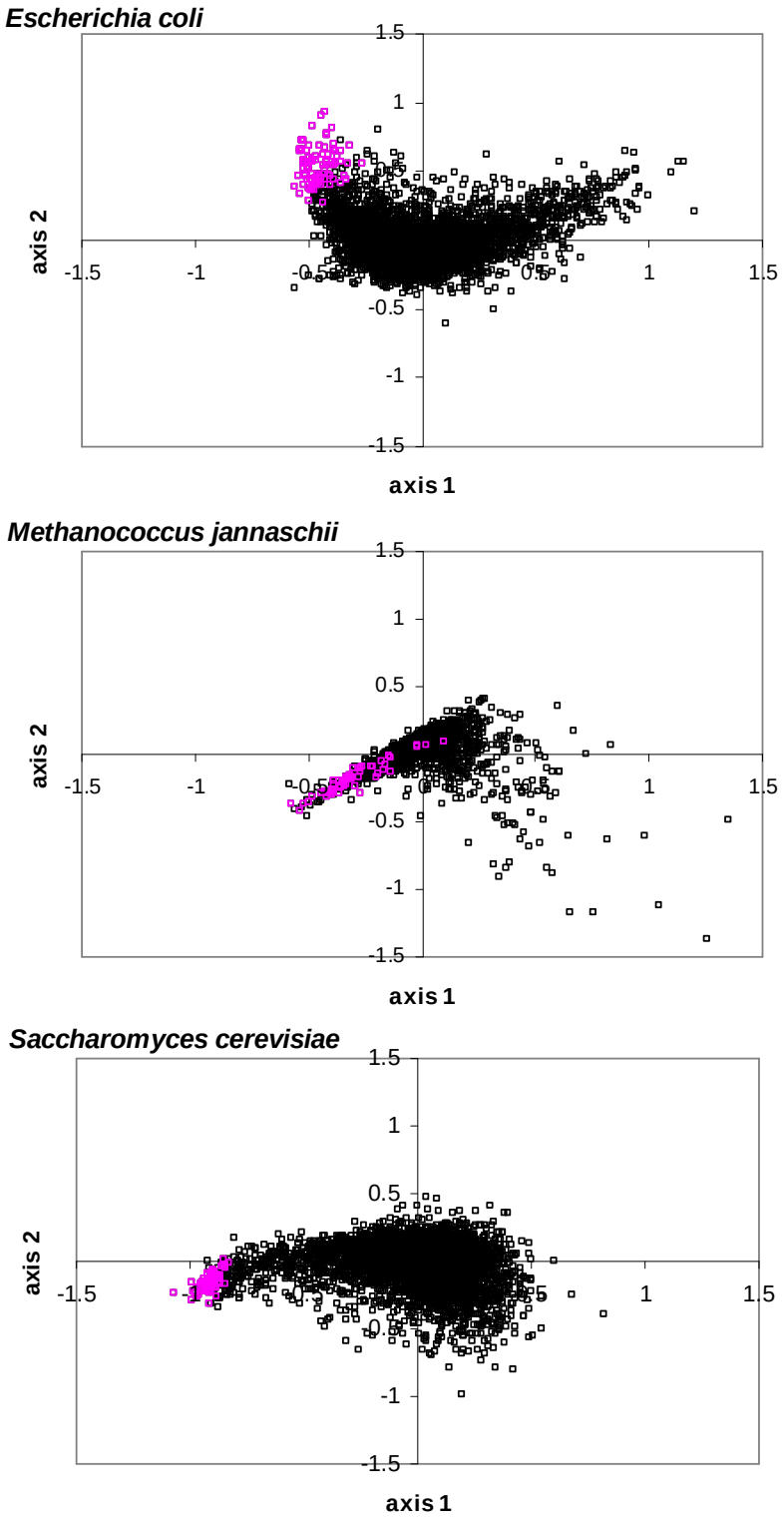

Figure 1. Positions of 80 most highly expressed genes according to their location of two main axes of COA of RSCU. Highly expressed genes group together because of the similarity of synonymous codon usage.

In addition to the AAA preference as +2 codon (Stenstrom *et al.*, 2001), the preference for A exists in about 20-30 nucleotide positions at the beginning of *E. coli* genes (Rocha *et al.*, 1999). This is suggested to be influenced by the need to decrease the stability of mRNA secondary structure in the initiation site (Rocha *et al.*, 1999; Stenstrom *et al.*, 2001).

Previous studies have showed that in many bacteria so-called minor codons (*e.g* AGG, AGA), which are otherwise very rare in a genome, are used preferentially near the initiation codon (Chen and Inouye, 1990; Ohno *et al.*, 2001). Such minor codons should reduce the translational efficiency due to the limited amount of corresponding tRNAs and should not be favoured in HEG. However, the preference for those codons near the translational start exists even in HEG suggesting some kind of regulatory role in response to changes in the tRNA pool size (Ohno *et al.*, 2001).

It is shown that rare arginine codons AGA and AGG in *E. coli* are prone for peptidyl-tRNA drop-off (Cruz-Vera *et al.*, 2003; Cruz-Vera *et al.*, 2004). Peptidyl-tRNA drop-off is peptidyl-tRNA dissociation from the ribosome before the correct end of the translation, resulting in an erroneous protein synthesis product (Menninger, 1976, 1978; Menez *et al.*, 2000). In addition, if drop-off occurs very frequently, it would lead to the saturation of peptidyl-tRNA hydrolase – an enzyme responsible for recycling peptidyl-tRNAs for new deaminoacylated tRNAs. As a result of enzyme saturation the pool of deaminoacylated tRNAs becomes limiting and does not allow efficient translation (Hernandez-Sanchez *et al.*, 1998; Tenson *et al.*, 1999; Heurgue-Hamard *et al.*, 2000; Menez *et al.*, 2000).

Importantly, the rate of drop-off is influenced by the length of nascent peptide – peptidyl-tRNAs with nascent peptides shorter than seven amino acids are more prone for drop-off than longer versions (Heurgue-Hamard *et al.*, 2000). This suggests that the preference for otherwise rare codons at the beginning of protein coding genes could be related to the regulation of protein synthesis via translation inhibition by peptidyl-tRNA drop-off mechanism.

The peptidyl-tRNA drop-off rates can be increased by mutations in peptidyl transferase centre of the ribosome leading to weaker interaction between tRNA and ribosomal A-site (Maivali *et al.*, 2001). Interestingly, the rates differ during the starvation for different amino acids (Caplan and Menninger, 1979). The peptidyl-tRNA drop off efficiency does not correlate with codon frequency. For example, as a result of drop-off the peptidyl-tRNAs reading codons decoding amino acids lysine, threonine and asparagine accumulate fastest and those reading codons decoding leucine, glycine and cysteine accumulate slowest (Menninger, 1978). In general, all codons beginning with A nucleotide and/or having A as the second nucleotide in the codon are more prone for drop-off (Cruz-Vera *et al.*, 2003).

1.4 Codon context bias

The properties and functionality of every base pair and codon are influenced by the surrounding sequence – the context (Yarus and Folley, 1985; Shpaer, 1986; Gouy, 1987). This influence acts through the functional interactions involving the tRNAs and the ribosome. Similarly to codon usage the codon context usage is also biased and influences the translational efficiency. Experimental results support the suggestion that codon context is even more strongly related to translational efficiency than single codon usage (Irwin *et al.*, 1995). Codon pair biases are directional, *e.g* in *E. coli* the

ACCCUG and CUGACC pairs are translated at markedly different rates, although both codons are frequently used (Irwin *et al.*, 1995).

There have been several studies analyzing codon pair biases in different species (Gutman and Hatfield, 1989; Berg and Silva, 1997; Fedorov *et al.*, 2002; Boycheva *et al.*, 2003; Moura *et al.*, 2005; Buchan *et al.*, 2006; Moura *et al.*, 2007a). The main selective effects on codon context are found in the nucleotides following the codon in the 3' direction (Berg and Silva, 1997; Fedorov *et al.*, 2002; Buchan *et al.*, 2006; Moura *et al.*, 2007a). However, the specific avoided or preferred patterns differ among species (Buchan *et al.*, 2006; Moura *et al.*, 2007a). The only universal context rule discovered is avoidance of type nnUAnn codon pairs (Moura *et al.*, 2007a). It was suggested that the codon context in eukaryotes is biased because target sequences for DNA methylation and trinucleotide repeats are present at high frequencies, while in bacteria and archaea the codon context is influenced mainly by the translational machinery (Moura *et al.*, 2007a).

The exact mechanism through which codon context functions has remained obscure. It is suggested that the interaction between tRNAs in the ribosome might influence the sequence context effects in protein coding genes. (Smith and Yarus, 1989; Buchan *et al.*, 2006). Since ribosome has three sites for tRNA binding, contexts involving as much as three codons (codon-triplets) were analysed in 11 fungal species (Moura *et al.*, 2007b). Despite of the close phylogenetic relationships of studied organisms the codon-triplet context varied, although certain common trends were observed. For example, nCC, nCG and nGn codons were associated with the codon-triplets which were not simply under-represented but entirely absent in ORFeomes (Moura *et al.*, 2007).

It is important to keep in mind that special amino acid motifs essential in formation of protein 3D structures influence the frequencies of codon contexts. For example, the membrane associated proteins contain regions of hydrophobic amino acids. This leads to biased frequencies of dipeptides which in turn influences the codon context frequencies. This aspect should be taken into consideration when calculating the over- or under-representation of codon contexts. Unfortunately this approach has not been very continual in context studies so far being used in *E. coli* codon pair analysis (Gutman and Hatfield, 1989) and later in larger analysis of 16 genomes (Buchan *et al.*, 2006).

1.4.1 Frameshifting promoting sequence contexts

Certain sequence contexts have been shown to be more prone to generate ribosomal frameshifts. Such contexts are for example mononucleotide repeats, which may cause translational (Gurvich *et al.*, 2003) or also transcriptional slippage (Wagner *et al.*, 1990; Baranov *et al.*, 2005). So called 'hungry' codons for which aminoacyl-tRNA is in short supply in starvation conditions could also increase the frequency of frameshifts errors if located in specific nucleotide contexts (Lindsley and Gallant, 1993).

Ribosomal frameshifting is also used as a gene expression regulating mechanism. Several programmed frameshifting sites have been described in the coding regions of mRNAs from different organisms (*e.g.* (Licznar *et al.*, 2003; Jacobs *et al.*, 2007)). Such sites are used for regulating gene expression through recoding.

Ribosomal frameshifts in these cases do not result in incorrect polypeptide but a polypeptide with a different biological function. For example, bacterial release factor 2 expression regulation operates through the frameshifting (Craigen *et al.*, 1985; Craigen and Caskey, 1986). In the early region of release factor 2 gene is located stop codon UGA in correct reading frame. In case of sufficient amount of release factor 2 in a cell this stop is effectively recognized by release factor 2 and the translation is terminated. In case of release factor 2 shortage the +1 frameshifting occurs and the ribosome continues in new frame synthesizing release factor 2 protein in full length. In all bacteria (except *Chlorobium tepidum*), where such programmed frameshifting is used for release factor 2 expression regulation, CUU UGA is the promoting context (Baranov *et al.*, 2002).

Nevertheless, frameshifting errors are rare events, occurring with a frequency less than once every 10,000 codons (Kurland, 1992). This means that sequences that are prone to frameshifting are successfully avoided in protein coding sequences. Using this as an assumption, additional putative programmed frameshifting sites have been predicted in *S. cerevisiae* protein coding genes by computational methods. Among the significantly under-represented heptanucleotides were found previously known frameshifting promoting contexts CUU-AGG-C and CUU-AGU-U; several other significantly under-represented contexts were experimentally proved to be prone for translational frameshifting (*e.g* GGU-CAG-A) (Shah *et al.*, 2002).

2. RESULTS AND DISCUSSION

2.1 Aims of the present study

The aim of this thesis was to shed new light on translational efficiency regulation by using computational and comparative genomics methods. The specific aims of the present study were as following:

1. To analyse the relationship between SD sequence and antiSD sequence base pairing strength and gene expression levels in *Escherichia coli* genes;

2. To describe the preferred and avoided motifs which are conserved at the beginning of highly expressed open reading frames of different organisms belonging into different domains of life;

3. To describe the universally preferred and avoided codon pairs in all three domains of life; to analyse discovered patterns in order to characterize the possible forces behind the differential usage of codon pairs.

2.2 Shine-Dalgarno sequence length and predicted expression level (I)

In bacteria, mRNA region called Shine-Dalgarno sequence is one of the key regions in binding of small ribosomal subunit to the mRNA. As described in Ref I, the selection of SD sequence is influenced by the growth temperature and not influenced by the growth rate; in addition, the SD:antiSD interaction efficiency is considerably related to the enhancer sequence. As a part of this systematic study of SD selection preferences in *E. coli* we made *in silico* analysis for studying the correlation between SD length and predicted expression levels of protein coding sequences. This is the first *in silico* study comparing the SD:antiSD interaction length and the CAI for all *E. coli* genes.

Experimental research of co-authors showed that the highest translation level at 37ºC was achieved in case of six paired nucleotides between SD region and 16S rRNA 3' end. The most effective SD sequence at 37ºC was AGGAGG. The free energy of complete binding of AGGAGG with antiSD is -7.7 kcal/mol. Experimental results raised a series of questions: Are the most optimal SD sequences also most commonly used in *E. coli* mRNAs? Do the SD:antiSD interaction length and strength correlate with the gene expression level? To answer those questions we conducted computational analysis of SD:antiSD interactions for all *E. coli* genes using program hybrid-min from UNAFold package (Markham and Zuker, 2005). With this program we calculated the minimal free energy values for rRNA-mRNA duplexes and at the same time collected the information about the length and location of SD:antiSD pairing.

We found that the average number of paired nucleotides in protein coding genes in *E. coli* was 5.8 which is in good agreement with our experimental results (6 nt). However, the average predicted interaction for all those sequences was weaker than

for the experimentally found most effective SD sequence – -6 kcal/mol compared to -7.7 kcal/mol. A closer look at the paired regions showed that this was the result of 'non-optimal' SD sequences involved in base-pairing. Namely, the SD:antiSD interaction was often shifted to more A/U rich regions in SD sequence and contained mismatches giving also the weaker interaction. The SD:antiSD constructs in experiments were continuous stretches of paired nucleotides without mismatches. It has to be mentioned that it is impossible to calculate the exact energetic effect of mismatches in this context. Specifically, we are dealing with the situation where the rRNA and mRNA duplexes are not drifting freely in the solution but stabilized by contacts with ribosomal proteins and RNA.

Plotting the expression level represented as CAI against the number of paired nucleotides in SD:antiSD region showed that the average CAI values and thus the gene expression levels were the same regardless of the interaction length (Figure 5 in Ref I). We propose that SD sequences and the enhancer sequences are acting co-operatively in stimulating the translation. Our observations also explain the previous reports that in some cases the strength of the SD:antiSD interaction does not determine the efficiency of translation initiation region (Ringquist *et al.*, 1992; de Smit and van Duin, 1994; Lee *et al.*, 1996). Still, it is important to note, that in our experimental study the spacing between SD sequence and start codon was not varied. Spacer region used in experiments has been reported to direct efficient translation initiation in *E. coli* (Barrick *et al.*, 1994). However, in computational analysis the SD:antiSD base pairing strength in *E. coli* genes was calculated as the minimal free energy between the 21 nucleotides upstream from the start codon and 13 nucleotides from the 3' end 16S rRNA where the distance between start codon and SD:antiSD interaction was not fixed. Therefore, the possible effect to the efficiency of translation by the spacing was not considered. It would be interesting to analyse the different spacer regions in *E. coli* genes with different expression levels and SD:antiSD interaction lengths in future.

2.3 Alanine preference in the second amino acid position of highly expressed proteins (II)

mRNA regions downstream from the initiation codon influence the efficiency of translation both in prokaryotes and eukaryotes. Studies on different motifs in this region have usually covered all protein coding genes in the genome. Surely, this can give an idea about the most frequently or rarely used sequence patterns important in regulation of the translation initiation event. However, the strongest effects concerning the effectiveness of translation should emerge if a group of HEG were analysed. HEG include the genes that encode proteins which are needed for the basic survival of the cell, *e.g* genes coding ribosomal proteins, elongation factors, RNA polymerase subunits, outer membrane proteins and heat shock proteins. Hence, the translation of those proteins has to be constantly quick and effective.

We compiled a representative set of different unicellular organisms covering a wide variety of different genomes to discover conserved patterns in all three domains of life. Our dataset included organisms with very small, average and very large genome sizes, different GC%, free living organisms, obligatory parasites and extremophiles. For all those genomes we compared the downstream regions in two datasets – the HEG and the all genes dataset.

The first surprising result came from the comparison of nucleotide frequencies. Previous studies had shown that A-rich sequence after the initiation codon was most expression promoting sequence and respectively, AAA as the second codon in protein coding sequences was the most expression promoting second codon (Stenstrom *et al.*, 2001). However, our results showed that although the third, fourth and fifth codon in HEG of some organisms is indeed more A-rich than in all genes, the second codon had significantly lower A-nucleotide frequency in HEG than in all genes (Figure 3 in Ref II). Further analysis revealed the increased frequency of G- and C-nucleotides in the second codon of HEG which in turn was the result of significant over-representation of GCN codons in that position. The GCN preference as second codon existed in 11 of the 15 organisms studied (Table 1 in Ref II). Interestingly, our results support the part of Kozak consensus (Kozak, 1987, 1997) with the G nucleotide following the start codon, suggesting that the signals around translational start are more conserved between bacteria and eukaryotes than previously thought. Our dataset did not contain higher eukaryotes but recent studies have supported our conclusions also in vertebrates and plants where the preference for C as the 5^{th} nucleotide at the beginning of HEG was discovered (Niimura *et al.*, 2003; Nakagawa *et al.*, 2008).

GCN family encodes the amino acid alanine. In all studied organisms except *Mycoplasma genitalium* the frequency of alanine as the second residue in highly expressed proteins was increased (Table 2 in Ref II). Is this universal preference caused by the codon based or amino acid based selection? The comparison of GCN codon usage in the second position of HEG and in the whole HEG sequences did not show any preference for a specific alanine codon in the second position (Table 3 in Ref II). This suggests that the selection is acting on amino acid level. Still, we cannot altogether rule out the codon based regulation. One possible explanation supporting the GCN codon selection is related to peptidyl-tRNA drop-off – premature peptidyl-tRNA dissociation from the ribosome not allowing the normal protein synthesis (Menninger, 1976, 1978; Menez *et al.*, 2000). It is shown, that the rate of drop-off event depends on two factors: the nascent peptide length and the specificity of the codon. Firstly, the shorter the peptide, the more efficient the drop-off (Heurgue-Hamard *et al.*, 2000). Hence, first few steps of translation and the region at the beginning of the protein coding sequence are most critical. Secondly, those peptidyl-tRNAs that read codons with A-nucleotides in the first or second position are more prone to drop-off (Cruz-Vera *et al.*, 2003). As our analysis showed, A-nucleotide frequency was significantly reduced in second codon of HEG. Therefore it is possible that GCN codons are used to avoid frequent drop-off events and to stabilize the dipeptidyl-tRNA on the ribosome.

The amino acid based selection could be related to the stability of proteins. The first few N-terminal amino acids modulate the stability of proteins (Varshavsky, 1996) and determine the cleavage of N-terminal formyl-methionine (or methionine in eukaryotes) (Tsunasawa *et al.*, 1985; Ben-Bassat *et al.*, 1987; Solbiati *et al.*, 1999). The rules for formyl-methionine or methionine removal are similar in bacteria and eukaryotes (Hirel *et al.*, 1989; Moerschell *et al.*, 1990): the initiating amino acid is cleaved in case the second residue is alanine, glycine, proline, serine, threonine or valine. In addition, all those six amino acids are stabilizing in bacteria and also in eukaryotes (Table 1) (Varshavsky, 1996; Tasaki and Kwon, 2007). We discovered that the genes coding for proteins with cleavage determining and stabilizing amino acids in the second position are highly enriched within HEG (Table 5 in Ref II). Therefore it is possible that the observed nucleotide and codon preferences in HEG are caused by the functional constraint of amino acid residues. Still, as alanine

showed the most universal and most strong preference it remains unclear why alanine has been chosen from the set of six amino acids with similar properties. It is possible that alanine is more efficient in directing the removal of the initiating amino acid and promoting protein stability.

Altogether, despite of the significant differences in translation initiation mechanism in bacteria and eukaryotes the discovered preferences at the beginning of HEG seem to be universal.

2.4 Universally preferred and avoided codon pairs (III)

Although translation initiation stage is the most important step influencing the efficiency of translation, during entire elongation cycle the efficiency and accuracy of protein synthesis is also regulated to a large extent. This is achieved by the ribosome maintaining the correct reading frame and incorporating correct amino acids as well as by the use of optimal synonymous codons and sequence contexts in mRNA. The existence of codon context bias is widely known. Similarly to codon bias studies it has been accepted that the context bias is species specific. Still, it has been proposed that codon context is even more important than codon usage for translational efficiency (Irwin *et al.*, 1995).

To analyze whether single codon preference or codon pair preference is more conserved on the evolutionary scale, we compared different bacteria according to relative synonymous codon usage (RSCU) and relative dicodon usage (RDCU). Similarity was measured by calculating the correlation of RSCU and RDCU values between each pair of bacteria. All pairs of bacteria analyzed were divided into nine groups according to the evolutionary distance separating each pair. The average correlation coefficients of RSCU and RDCU were calculated for each group. Comparison of RSCU and RDCU correlations showed that overall codon pair usage is indeed less conserved than single codon usage (Figure 5 in Ref III). However, these findings do not rule out that a set of universally avoided or preferred contexts still exists.

To study the common rules of codon context bias we looked for codon pairs that are significantly preferred or avoided in three domains of life. These conserved cases of biased codon pairs could shed light on the mechanisms shaping the genes and genomes. To ensure that the effects observed at the codon pair level were not caused by avoidances or preferences of dipeptides, the expected codon pair values were normalized to the dipeptide frequencies.

In addition to neighbouring codons we tested the conservation of more distant (1-3, 1-4, 1-5) codon pairs. However, for codons 1-3 we found only one codon pair with significant under-representation – GGUnnnGGU. No conserved biases were found for more distant codon pairs.

In order to differentiate between the effects resulting from the DNA based selection and translation based selection the strength of biases was compared in ORFeomes and in genomes. We found that conserved patterns result mainly from translational effects not from DNA-related mechanisms since the biases are stronger in ORFeomes than in genomes.

It was previously claimed that codon pair preference is primarily determined by a tetranucleotide combination including the last nucleotide of the first codon and all three nucleotides of the second codon (Buchan *et al.*, 2006). However, after dividing

our datasets based on sub-patterns we discovered that different patterns ranging from dinucleotides to hexanucleotides could explain conserved biased codon pair usage. Still, with only one exception, all discovered patterns contained fixed nucleotide in the last position of the first codon in the codon pair.

Among codon pairs that were significantly avoided in more than half the organisms studied the most frequently avoided codon pairs contained following patterns: nnUAnn, nnGGnn, nnGnnC, nnCGCn, GUCCnn, CUCCnn, nnCnnA or UUCGnn. Avoidance of nnUAnn was described previously (Moura *et al.*, 2007) where part of this avoidance was related to the avoidance of TpA dinucleotides in genomic sequences. Our genome and ORFeome comparisons indicated that the avoidance of such codon pairs is mainly related to the translational mechanisms since the avoidance was stronger in ORFeomes than in genomes. Since many of the type nnUAnn codon pairs contained out-frame UAA and UAG stop codons this avoidance could be the result of reducing premature translation termination after frameshifting events. Interestingly, type nnUAnn codon pairs contained those out-frame stops also in antisense frame – the fact that we cannot explain based on mechanisms known today. Furthermore, none of the avoided codon pairs contained out-frame UGA suggesting that the UA dinucleotide could have important role separately from the out-frame stops.

Among the conserved avoided codon pairs occurred also CUUAGU – part of previously known programmed frameshifting site in yeast telomerase subunit EST3 (CUU-AGU-U) (Morris and Lundblad, 1997).

In addition, mononucleotide repeats known to cause frameshifting were discovered among avoided codon pairs being more avoided in ORFeomes than in genomes.

The number of codon pairs significantly preferred in more than half of organisms studied was smaller than the number of avoided codon pairs (81 compared to 207). Although the effect sizes were similar in both groups, this suggests that it is more important for the cell to avoid possibly harmful contexts during protein synthesis than to enhance the number of optimal codons. The most frequently preferred codon pairs contain the patterns nnGCnn, nnCAnn or nnUnCn.

Codon frequencies correlate with tRNA concentrations, suggesting that this is a major selective force on codon usage patterns (Ikemura, 1981; Dong et al., 1996; Elf et al., 2003). The codon pair preferences can be shaped by several different molecular mechanisms. One is the possible decrease of frameshifting errors through avoidance of mononucleotide repeats (Wagner et al., 1990; Gurvich et al., 2003; Baranov et al., 2005). In addition, it has been suggested that codon context might be influenced by certain structural constraints imposed by two tRNAs occupying the ribosomal P- and A-sites (Smith and Yarus, 1989; Irwin et al., 1995; Buchan et al., 2006). Unfortunately, we currently have very limited information about the details of interaction between different tRNAs with the ribosome (Korostelev et al., 2006; Selmer et al., 2006; Dunham et al., 2007), which precludes further extension of this hypothesis.

In addition, it is possible that codon pair preferences help to distinguish actual reading frames from noncoding sequences similarly to codon preferences in some species. However this question would need much longer analysis and at the moment we can only speculate that our observations could add some predictive power to gene prediction algorithms in future.

CONCLUSIONS

By using bioinformatical methods we have characterized several aspects of regulatory and coding sequences related to the efficiency of protein synthesis. The most important conclusions of current thesis are:

1. In *Escherichia coli* the base pairing potential of the SD sequence and the expression level of a gene is not correlated suggesting the importance of enhancer sequences acting co-operatively with SD sequence in translation stimulation.

2. Strong alanine preference exists at the beginning of highly expressed proteins in different organisms possibly related to the translational efficiency and/or protein stabilization

3. A universal set of similarly biased codon pairs exists in different genomes from three domains of life. Most of the codon pairs have stronger bias on the ORFeome level than the corresponding hexanucleotides have on the whole genome level, suggesting that translation has a greater influence on codon pair biases than molecular mechanisms that shape the genomic DNA in general.

REFERENCES

Akiba, T., Koyama, K., Ishiki, Y., Kimura, S. and Fukushima, T. (1960) On the mechanism of the development of multiple-drug-resistant clones of Shigella. *Jpn J Microbiol*, **4,** 219-227.

Baranov, P.V., Gesteland, R.F. and Atkins, J.F. (2002) Release factor 2 frameshifting sites in different bacteria. *EMBO Rep*, **3,** 373-377.

Baranov, P.V., Hammer, A.W., Zhou, J., Gesteland, R.F. and Atkins, J.F. (2005) Transcriptional slippage in bacteria: distribution in sequenced genomes and utilization in IS element gene expression. *Genome Biol*, **6,** R25.

Barrick, D., Villanueba, K., Childs, J., Kalil, R., Schneider, T.D., Lawrence, C.E., Gold, L. and Stormo, G.D. (1994) Quantitative analysis of ribosome binding sites in E.coli. *Nucleic Acids Res*, **22,** 1287-1295.

Beletskii, A. and Bhagwat, A.S. (1996) Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli. *Proc Natl Acad Sci U S A*, **93,** 13919-13924.

Ben-Bassat, A., Bauer, K., Chang, S.Y., Myambo, K., Boosman, A. and Chang, S. (1987) Processing of the initiation methionine from proteins: properties of the Escherichia coli methionine aminopeptidase and its gene structure. *J Bacteriol*, **169,** 751-757.

Benelli, D., Maone, E. and Londei, P. (2003) Two different mechanisms for ribosome/mRNA interaction in archaeal translation initiation. *Mol Microbiol*, **50,** 635-643.

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.

Bernardi, G. and Bernardi, G. (1986) Compositional constraints and genome evolution. *J Mol Evol*, **24,** 1-11.

Boycheva, S., Chkodrov, G. and Ivanov, I. (2003) Codon pairs in the genome of Escherichia coli. *Bioinformatics*, **19,** 987-998.

Brewer, B.J. (1988) When polymerases collide: replication and the transcriptional organization of the E. coli chromosome. *Cell*, **53,** 679-686.

Brown, C.M., Stockwell, P.A., Trotman, C.N. and Tate, W.P. (1990) The signal for the termination of protein synthesis in procaryotes. *Nucleic Acids Res*, **18,** 2079-2086.

Buchan, J.R., Aucott, L.S. and Stansfield, I. (2006) tRNA properties help shape codon pair preferences in open reading frames. *Nucleic Acids Res*, **34,** 1015-1027.
Caplan, A.B. and Menninger, J.R. (1979) Tests of the ribosomal editing hypothesis: amino acid starvation differentially enhances the dissociation of peptidyl-tRNA from the ribosome. *J Mol Biol*, **134,** 621-637.
Caskey, C.T. (1977) Peptide chain termination, pp. 443-456. Weissbach, H. and Pestka, S., eds, *Molecular mechanisms of protein synthesis*. Academic Press.
Cavener, D.R. and Ray, S.C. (1991) Eukaryotic start and stop translation sites. *Nucleic Acids Res*, **19,** 3185-3192.
Chen, G.F. and Inouye, M. (1990) Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within the first 25 codons of the Escherichia coli genes. *Nucleic Acids Res*, **18,** 1465-1473.
Chen, H., Bjerknes, M., Kumar, R. and Jay, E. (1994) Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. *Nucleic Acids Res*, **22,** 4953-4957.
Chen, S., Vetro, J.A. and Chang, Y.H. (2002) The specificity in vivo of two distinct methionine aminopeptidases in Saccharomyces cerevisiae. *Arch Biochem Biophys*, **398,** 87-93.
Chiapello, H., Lisacek, F., Caboche, M. and Henaut, A. (1998) Codon usage and gene function are related in sequences of Arabidopsis thaliana. *Gene*, **209,** GC1-GC38.
Coghlan, A. and Wolfe, K.H. (2000) Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. *Yeast*, **16,** 1131-1145.
Craigen, W.J., Cook, R.G., Tate, W.P. and Caskey, C.T. (1985) Bacterial peptide chain release factors: conserved primary structure and possible frameshift regulation of release factor 2. *Proc Natl Acad Sci U S A*, **82,** 3616-3620.
Craigen, W.J. and Caskey, C.T. (1986) Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature*, **322,** 273-275.
Cruz-Vera, L.R., Hernandez-Ramon, E., Perez-Zamorano, B. and Guarneros, G. (2003) The rate of peptidyl-tRNA dissociation from the ribosome during minigene expression depends on the nature of the last decoding interaction. *J Biol Chem*, **278,** 26065-26070.
Cruz-Vera, L.R., Magos-Castro, M.A., Zamora-Romo, E. and Guarneros, G. (2004) Ribosome stalling and peptidyl-tRNA drop-off during translational delay at AGA codons. *Nucleic Acids Res*, **32,** 4462-4468.

Dalboge, H., Bayne, S. and Pedersen, J. (1990) In vivo processing of N-terminal methionine in E. coli. *FEBS Lett*, **266,** 1-3.
Das, S., Paul, S., Bag, S.K. and Dutta, C. (2006) Analysis of Nanoarchaeum equitans genome and proteome composition: indications for hyperthermophilic and parasitic adaptation. *BMC Genomics*, **7,** 186.
de Boer, H.A., Comstock, L.J., Hui, A., Wong, E. and Vasser, M. (1983) Portable Shine-Dalgarno regions; nucleotides between the Shine-Dalgarno sequence and the start codon affect the translation efficiency. *Gene Amplif Anal*, **3,** 103-116.
de Smit, M.H. and van Duin, J. (1994) Translational initiation on structured messengers. Another role for the Shine-Dalgarno interaction. *J Mol Biol*, **235,** 173-184.
de Smit, M.H. and van Duin, J. (2003) Translational standby sites: how ribosomes may deal with the rapid folding kinetics of mRNA. *J Mol Biol*, **331,** 737-743.
Dong, H., Nilsson, L. and Kurland, C.G. (1996) Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *J Mol Biol*, **260,** 649-663.
dos Reis, M., Wernisch, L. and Savva, R. (2003) Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Res*, **31,** 6976-6985.
Dunham, C.M., Selmer, M., Phelps, S.S., Kelley, A.C., Suzuki, T., Joseph, S. and Ramakrishnan, V. (2007) Structures of tRNAs with an expanded anticodon loop in the decoding center of the 30S ribosomal subunit. *Rna*, **13,** 817-823.
Duret, L. (2000) tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet*, **16,** 287-289.
Ehrenberg, M. and Kurland, C.G. (1984) Costs of accuracy determined by a maximal growth rate constraint. *Q Rev Biophys*, **17,** 45-82.
Elf, J., Nilsson, D., Tenson, T. and Ehrenberg, M. (2003) Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science*, **300,** 1718-1722.
Falb, M., Aivaliotis, M., Garcia-Rizo, C., Bisle, B., Tebbe, A., Klein, C., Konstantinidis, K., Siedler, F., Pfeiffer, F. and Oesterhelt, D. (2006) Archaeal N-terminal protein maturation commonly involves N-terminal acetylation: a large-scale proteomics survey. *J Mol Biol*, **362,** 915-924.
Fedorov, A., Saxonov, S. and Gilbert, W. (2002) Regularities of context-dependent codon bias in eukaryotic genes. *Nucleic Acids Res*, **30,** 1192-1197.

Fennoy, S.L. and Bailey-Serres, J. (1993) Synonymous codon usage in Zea mays L. nuclear genes is varied by levels of C and G-ending codons. *Nucleic Acids Res*, **21,** 5294-5300.

Frolova, L., Le Goff, X., Rasmussen, H.H., Cheperegin, S., Drugeon, G., Kress, M., Arman, I., Haenni, A.L., Celis, J.E., Philippe, M. and et al. (1994) A highly conserved eukaryotic protein family possessing properties of polypeptide chain release factor. *Nature*, **372,** 701-703.

Fuglsang, A. (2004) Bioinformatic analysis of the link between gene composition and expressivity in Saccharomyces cerevisiae and Schizosaccharomyces pombe. *Antonie Van Leeuwenhoek*, **86,** 135-147.

Futcher, B., Latter, G.I., Monardo, P., McLaughlin, C.S. and Garrels, J.I. (1999) A sampling of the yeast proteome. *Mol Cell Biol*, **19,** 7357-7368.

Garcia-Vallve, S., Romeu, A. and Palau, J. (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res*, **10,** 1719-1725.

Geigenmuller, U. and Nierhaus, K.H. (1990) Significance of the third tRNA binding site, the E site, on E. coli ribosomes for the accuracy of translation: an occupied E site prevents the binding of non-cognate aminoacyl-tRNA to the A site. *Embo J*, **9,** 4527-4533.

Gonzalez de Valdivia, E.I. and Isaksson, L.A. (2004) A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in Escherichia coli. *Nucleic Acids Res*, **32,** 5198-5205.

Gouy, M. (1987) Codon contexts in enterobacterial and coliphage genes. *Mol Biol Evol*, **4,** 426-444.

Grill, S., Gualerzi, C.O., Londei, P. and Blasi, U. (2000) Selective stimulation of translation of leaderless mRNA by initiation factor 2: evolutionary implications for translation. *Embo J*, **19,** 4101-4110.

Gurvich, O.L., Baranov, P.V., Zhou, J., Hammer, A.W., Gesteland, R.F. and Atkins, J.F. (2003) Sequences that direct significant levels of frameshifting are frequent in coding regions of Escherichia coli. *Embo J*, **22,** 5941-5950.

Gutman, G.A. and Hatfield, G.W. (1989) Nonrandom utilization of codon pairs in Escherichia coli. *Proc Natl Acad Sci U S A*, **86,** 3699-3703.

Hall, C., Brachat, S. and Dietrich, F.S. (2005) Contribution of horizontal gene transfer to the evolution of Saccharomyces cerevisiae. *Eukaryot Cell*, **4,** 1102-1115.

Hamilton, R., Watanabe, C.K. and de Boer, H.A. (1987) Compilation and comparison of the sequence context around the AUG startcodons in Saccharomyces cerevisiae mRNAs. *Nucleic Acids Res*, **15,** 3581-3593.

Hernandez-Sanchez, J., Valadez, J.G., Herrera, J.V., Ontiveros, C. and Guarneros, G. (1998) lambda bar minigene-mediated inhibition of protein synthesis involves accumulation of peptidyl-tRNA and starvation for tRNA. *Embo J*, **17,** 3758-3765.

Heurgue-Hamard, V., Dincbas, V., Buckingham, R.H. and Ehrenberg, M. (2000) Origins of minigene-dependent growth inhibition in bacterial cells. *Embo J*, **19,** 2701-2709.

Hirel, P.H., Schmitter, M.J., Dessen, P., Fayat, G. and Blanquet, S. (1989) Extent of N-terminal methionine excision from Escherichia coli proteins is governed by the side-chain length of the penultimate amino acid. *Proc Natl Acad Sci U S A*, **86,** 8247-8251.

Hurst, L.D. and Merchant, A.R. (2001) High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proc Biol Sci*, **268,** 493-497.

Ikemura, T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol*, **146,** 1-21.

Ikemura, T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol*, **151,** 389-409.

Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol*, **2,** 13-34.

Irwin, B., Heck, J.D. and Hatfield, G.W. (1995) Codon pair utilization biases influence translational elongation step times. *J Biol Chem*, **270,** 22801-22806.

Jacobs, J.L., Belew, A.T., Rakauskaite, R. and Dinman, J.D. (2007) Identification of functional, endogenous programmed -1 ribosomal frameshift signals in the genome of Saccharomyces cerevisiae. *Nucleic Acids Res*, **35,** 165-174.

Jannsen, G. (1993) Eubacterial, archaebacterial and eukaryotic genes that encode leaderless mRNA, p. 59-67. R. H. Baltz, G. D. Hegeman, and P. L. Scatrud (ed.), *Industrial microorganisms: basic and applied molecular genetics*. ASM Press. Washington, D.C.

Jansen, R., Bussemaker, H.J. and Gerstein, M. (2003) Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res*, **31,** 2242-2251.

Jia, M. and Li, Y. (2005) The relationship among gene expression, folding free energy and codon usage bias in Escherichia coli. *FEBS Lett*, **579,** 5333-5337.

Kanaya, S., Yamada, Y., Kudo, Y. and Ikemura, T. (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, **238,** 143-155.

Kaplan, J.B. and Fine, D.H. (1998) Codon usage in Actinobacillus actinomycetemcomitans. *FEMS Microbiol Lett*, **163,** 31-36.

Karlin, S., Mrazek, J. and Campbell, A.M. (1998) Codon usages in different gene classes of the Escherichia coli genome. *Mol Microbiol*, **29,** 1341-1355.

Kisselev, L., Ehrenberg, M. and Frolova, L. (2003) Termination of translation: interplay of mRNA, rRNAs and release factors? *Embo J*, **22,** 175-182.

Kisselev, L.L. and Buckingham, R.H. (2000) Translational termination comes of age. *Trends Biochem Sci*, **25,** 561-566.

Komarova, A.V., Tchufistova, L.S., Supina, E.V. and Boni, I.V. (2002) Protein S1 counteracts the inhibitory effect of the extended Shine-Dalgarno sequence on translation. *Rna*, **8,** 1137-1147.

Komarova, A.V., Tchufistova, L.S., Dreyfus, M. and Boni, I.V. (2005) AU-rich sequences within 5' untranslated leaders enhance translation and stabilize mRNA in Escherichia coli. *J Bacteriol*, **187,** 1344-1349.

Konecki, D.S., Aune, K.C., Tate, W. and Caskey, C.T. (1977) Characterization of reticulocyte release factor. *J Biol Chem*, **252,** 4514-4520.

Kornberg, A. and Baker, T.A. (1992) DNA replication, WH Freeman & Co. New York.

Korostelev, A., Trakhanov, S., Laurberg, M. and Noller, H.F. (2006) Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements. *Cell*, **126,** 1065-1077.

Kozak, M. (1983) Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. *Microbiol Rev*, **47,** 1-45.

Kozak, M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res,* **15,** 8125-8148.
Kozak, M. (1989) The scanning model for translation: an update. *J Cell Biol*, **108,** 229-241.
Kozak, M. (1997) Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. *Embo J,* **16,** 2482-2492.
Kotlar, D. and Lavner, Y. (2006) The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids. *BMC Genomics*, **7,** 67.
Kurland, C.G. (1992) Translational accuracy and the fitness of bacteria. *Annu Rev Genet*, **26,** 29-50.
Kyrpides, N.C. and Woese, C.R. (1998) Archaeal translation initiation revisited: the initiation factor 2 and eukaryotic initiation factor 2B alpha-beta-delta subunit families. *Proc Natl Acad Sci U S A,* **95,** 3726-3730.
Lafay, B., Lloyd, A.T., McLean, M.J., Devine, K.M., Sharp, P.M. and Wolfe, K.H. (1999) Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res,* **27,** 1642-1649.
Lawrence, J.G. and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol*, **44,** 383-397.
Lee, K., Holland-Staley, C.A. and Cunningham, P.R. (1996) Genetic analysis of the Shine-Dalgarno interaction: selection of alternative functional mRNA-rRNA combinations. *Rna*, **2,** 1270-1285.
Licznar, P., Mejlhede, N., Prere, M.F., Wills, N., Gesteland, R.F., Atkins, J.F. and Fayet, O. (2003) Programmed translational -1 frameshifting on hexanucleotide motifs and the wobble properties of tRNAs. *Embo J,* **22,** 4770-4778.
Lindsley, D. and Gallant, J. (1993) On the directional specificity of ribosome frameshifting at a "hungry" codon. *Proc Natl Acad Sci U S A,* **90,** 5469-5473.
Lithwick, G. and Margalit, H. (2003) Hierarchy of sequence-dependent features associated with prokaryotic translation. *Genome Res,* **13,** 2665-2673.
Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol,* **13,** 660-665.
Lobry, J.R. (1997) Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene,* **205,** 309-316.

Ma, J., Campbell, A. and Karlin, S. (2002) Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol*, **184,** 5733-5745.
Maivali, U., Saarma, U. and Remme, I. (2001) [Mutations in the Escherichia coli 23S rRNA increase the rate of peptidyl-tRNA dissociation from the ribosome]. *Mol Biol (Mosk)*, **35,** 666-671.
Marians, K.J. (1992) Prokaryotic DNA replication. *Annu Rev Biochem*, **61,** 673-719.
Markham, N.R. and Zuker, M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res*, **33,** W577-581.
Marquez, V., Wilson, D.N., Tate, W.P., Triana-Alonso, F. and Nierhaus, K.H. (2004) Maintaining the ribosomal reading frame: the influence of the E site during translational regulation of release factor 2. *Cell*, **118,** 45-55.
Meinnel, T., Mechulam, Y. and Blanquet, S. (1993) Methionine as translation start signal: a review of the enzymes of the pathway in Escherichia coli. *Biochimie*, **75,** 1061-1075.
Menez, J., Heurgue-Hamard, V. and Buckingham, R.H. (2000) Sequestration of specific tRNA species cognate to the last sense codon of an overproduced gratuitous protein. *Nucleic Acids Res*, **28,** 4725-4732.
Menninger, J.R. (1976) Peptidyl transfer RNA dissociates during protein synthesis from ribosomes of Escherichia coli. *J Biol Chem*, **251,** 3392-3398.
Menninger, J.R. (1978) The accumulation as peptidyl-transfer RNA of isoaccepting transfer RNA families in Escherichia coli with temperature-sensitive peptidyl-transfer RNA hydrolase. *J Biol Chem*, **253,** 6808-6813.
Miller, C.G., Strauch, K.L., Kukral, A.M., Miller, J.L., Wingfield, P.T., Mazzei, G.J., Werlen, R.C., Graber, P. and Movva, N.R. (1987) N-terminal methionine-specific peptidase in Salmonella typhimurium. *Proc Natl Acad Sci U S A*, **84,** 2718-2722.
Miyasaka, H. (1999) The positive relationship between codon usage bias and translation initiation AUG context in Saccharomyces cerevisiae. *Yeast*, **15,** 633-637.
Moerschell, R.P., Hosokawa, Y., Tsunasawa, S. and Sherman, F. (1990) The specificities of yeast methionine aminopeptidase and acetylation of amino-terminal methionine in vivo. Processing of altered iso-1-cytochromes c created by oligonucleotide transformation. *J Biol Chem*, **265,** 19638-19643.
Moriyama, E.N. and Powell, J.R. (1997) Codon usage bias and tRNA abundance in Drosophila. *J Mol Evol*, **45,** 514-523.

Morris, D.K. and Lundblad, V. (1997) Programmed translational frameshifting in a gene required for yeast telomere replication. *Curr Biol*, **7,** 969-976.
Moszer, I., Rocha, E.P. and Danchin, A. (1999) Codon usage and lateral gene transfer in Bacillus subtilis. *Curr Opin Microbiol*, **2,** 524-528.
Mottagui-Tabar, S., Tuite, M.F. and Isaksson, L.A. (1998) The influence of 5' codon context on translation termination in Saccharomyces cerevisiae. *Eur J Biochem*, **257,** 249-254.
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.
Moura, G., Pinheiro, M., Arrais, J., Gomes, A.C., Carreto, L., Freitas, A., Oliveira, J.L. and Santos, M.A. (2007) Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure. *PLoS ONE*, **2,** e847.
Moura, G.R., Lousado, J.P., Pinheiro, M., Carreto, L., Silva, R.M., Oliveira, J.L. and Santos, M.A. (2007) Codon-triplet context unveils unique features of the Candida albicans protein coding genome. *BMC Genomics*, **8,** 444.
Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valin, F. and Bernardi, G. (2004) Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Lett*, **573,** 73-77.
Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H.E., Moran, N.A. and Hattori, M. (2006) The 160-kilobase genome of the bacterial endosymbiont Carsonella. *Science*, **314,** 267.
Nakagawa, S., Niimura, Y., Gojobori, T., Tanaka, H. and Miura, K. (2008) Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res*, **36,** 861-871.
Namy, O., Hatin, I. and Rousset, J.P. (2001) Impact of the six nucleotides downstream of the stop codon on translation termination. *EMBO Rep*, **2,** 787-793.
Niimura, Y., Terabe, M., Gojobori, T. and Miura, K. (2003) Comparative analysis of the base biases at the gene terminal portions in seven eukaryote genomes. *Nucleic Acids Res*, **31,** 5195-5201.
Nomura, M. and Morgan, E.A. (1977) Genetics of bacterial ribosomes. *Annu Rev Genet*, **11,** 297-347.

Ochman, H., Lawrence, J.G. and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405,** 299-304.

Ohno, H., Sakai, H., Washio, T. and Tomita, M. (2001) Preferential usage of some minor codons in bacteria. *Gene*, **276,** 107-115.

Percudani, R., Pavesi, A. and Ottonello, S. (1997) Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. *J Mol Biol*, **268,** 322-330.

Poole, E.S., Brown, C.M. and Tate, W.P. (1995) The identity of the base following the stop codon determines the efficiency of in vivo translational termination in Escherichia coli. *Embo J*, **14,** 151-158.

Poole, E.S., Major, L.L., Mannering, S.A. and Tate, W.P. (1998) Translational termination in Escherichia coli: three bases following the stop codon crosslink to release factor 2 and affect the decoding efficiency of UGA-containing signals. *Nucleic Acids Res*, **26,** 954-960.

Price, M.N., Alm, E.J. and Arkin, A.P. (2005) Interruptions in gene expression drive highly expressed operons to the leading strand of DNA replication. *Nucleic Acids Res*, **33,** 3224-3234.

Ringquist, S., Shinedling, S., Barrick, D., Green, L., Binkley, J., Stormo, G.D. and Gold, L. (1992) Translation initiation in Escherichia coli: sequences within the ribosome-binding site. *Mol Microbiol*, **6,** 1219-1229.

Rocha, E. (2002) Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol*, **10,** 393-395.

Rocha, E.P., Danchin, A. and Viari, A. (1999) Translation in Bacillus subtilis: roles and trends of initiation and termination, insights from a genome analysis. *Nucleic Acids Res*, **27,** 3567-3576.

Sanford, R.A., Cole, J.R. and Tiedje, J.M. (2002) Characterization and description of Anaeromyxobacter dehalogenans gen. nov., sp. nov., an aryl-halorespiring facultative anaerobic myxobacterium. *Appl Environ Microbiol*, **68,** 893-900.

Schmitt, E., Guillon, J.M., Meinnel, T., Mechulam, Y., Dardel, F. and Blanquet, S. (1996) Molecular recognition governing the initiation of translation in Escherichia coli. A review. *Biochimie*, **78,** 543-554.

Schurr, T., Nadir, E. and Margalit, H. (1993) Identification and characterization of E.coli ribosomal binding sites by free energy computation. *Nucleic Acids Res*, **21,** 4019-4023.

Selmer, M., Dunham, C.M., Murphy, F.V.t., Weixlbaumer, A., Petry, S., Kelley, A.C., Weir, J.R. and Ramakrishnan, V. (2006) Structure of the 70S ribosome complexed with mRNA and tRNA. *Science*, **313,** 1935-1942.

Sergiev, P.V., Lesnyak, D.V., Kiparisov, S.V., Burakovsky, D.E., Leonov, A.A., Bogdanov, A.A., Brimacombe, R. and Dontsova, O.A. (2005) Function of the ribosomal E-site: a mutagenesis study. *Nucleic Acids Res*, **33,** 6048-6056.

Shah, A.A., Giddings, M.C., Parvaz, J.B., Gesteland, R.F., Atkins, J.F. and Ivanov, I.P. (2002) Computational identification of putative programmed translational frameshift sites. *Bioinformatics*, **18,** 1046-1053.

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.

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 Bulmer, M. (1988) Selective differences among translation termination codons. *Gene*, **63,** 141-145.

Sherman, F., Stewart, J.W. and Tsunasawa, S. (1985) Methionine or not methionine at the beginning of a protein. *Bioessays*, **3,** 27-31.

Shields, D.C. and Sharp, P.M. (1987) Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases. *Nucleic Acids Res*, **15,** 8023-8040.

Shine, J. and Dalgarno, L. (1974) The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A*, **71,** 1342-1346.

Shine, J. and Dalgarno, L. (1975) Terminal-sequence analysis of bacterial ribosomal RNA. Correlation between the 3'-terminal-polypyrimidine sequence of 16-S RNA and translational specificity of the ribosome. *Eur J Biochem*, **57,** 221-230.

Shpaer, E.G. (1986) Constraints on codon context in Escherichia coli genes. Their possible role in modulating the efficiency of translation. *J Mol Biol*, **188,** 555-564.

Shultzaberger, R.K., Bucheimer, R.E., Rudd, K.E. and Schneider, T.D. (2001) Anatomy of Escherichia coli ribosome binding sites. *J Mol Biol*, **313,** 215-228.

Singer, G.A. and Hickey, D.A. (2000) Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol*, **17,** 1581-1588.

Smith, D. and Yarus, M. (1989) tRNA-tRNA interactions within cellular ribosomes. *Proc Natl Acad Sci U S A*, **86,** 4397-4401.

Solbiati, J., Chapman-Smith, A., Miller, J.L., Miller, C.G. and Cronan, J.E., Jr. (1999) Processing of the N termini of nascent polypeptide chains requires deformylation prior to methionine removal. *J Mol Biol*, **290,** 607-614.

Stenico, M., Lloyd, A.T. and Sharp, P.M. (1994) Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. *Nucleic Acids Res*, **22,** 2437-2446.

Stenstrom, C.M., Jin, H., Major, L.L., Tate, W.P. and Isaksson, L.A. (2001) Codon bias at the 3'-side of the initiation codon is correlated with translation initiation efficiency in Escherichia coli. *Gene*, **263,** 273-284.

Stoletzki, N. and Eyre-Walker, A. (2007) Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Mol Biol Evol*, **24,** 374-381.

Studer, S.M. and Joseph, S. (2006) Unfolding of mRNA secondary structure by the bacterial translation initiation complex. *Mol Cell*, **22,** 105-115.

Sun, J., Chen, M., Xu, J. and Luo, J. (2005) Relationships among stop codon usage bias, its context, isochores, and gene expression level in various eukaryotes. *J Mol Evol*, **61,** 437-444.

Tasaki, T. and Kwon, Y.T. (2007) The mammalian N-end rule pathway: new insights into its components and physiological roles. *Trends Biochem Sci*, **32,** 520-528.

Tate, W.P., Poole, E.S., Dalphin, M.E., Major, L.L., Crawford, D.J. and Mannering, S.A. (1996) The translational stop signal: codon with a context, or extended factor recognition element? *Biochimie*, **78,** 945-952.

Tenson, T., Herrera, J.V., Kloss, P., Guarneros, G. and Mankin, A.S. (1999) Inhibition of translation and cell growth by minigene expression. *J Bacteriol*, **181,** 1617-1622.

Trimble, M.J., Minnicus, A. and Williams, K.P. (2004) tRNA slippage at the tmRNA resume codon. *Rna*, **10,** 805-812.

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.

Tsunasawa, S., Stewart, J.W. and Sherman, F. (1985) Amino-terminal processing of mutant forms of yeast iso-1-cytochrome c. The specificities of methionine aminopeptidase and acetyltransferase. *J Biol Chem*, **260,** 5382-5391.

Uptain, S.M. and Lindquist, S. (2002) Prions as protein-based genetic elements. *Annu Rev Microbiol*, **56,** 703-741.

Wagner, L.A., Weiss, R.B., Driscoll, R., Dunn, D.S. and Gesteland, R.F. (1990) Transcriptional slippage occurs during elongation at runs of adenine or thymine in Escherichia coli. *Nucleic Acids Res*, **18,** 3529-3535.

Varshavsky, A. (1996) The N-end rule: functions, mysteries, uses. *Proc Natl Acad Sci U S A*, **93,** 12142-12149.

Williams, I., Richardson, J., Starkey, A. and Stansfield, I. (2004) Genome-wide prediction of stop codon readthrough during translation in the yeast Saccharomyces cerevisiae. *Nucleic Acids Res*, **32,** 6605-6616.

Vinogradov, A.E. (2003) Isochores and tissue-specificity. *Nucleic Acids Res*, **31,** 5212-5220.

Yarus, M. and Folley, L.S. (1985) Sense codons are found in specific contexts. *J Mol Biol*, **182,** 529-540.

Yusupov, M.M., Yusupova, G.Z., Baucom, A., Lieberman, K., Earnest, T.N., Cate, J.H. and Noller, H.F. (2001) Crystal structure of the ribosome at 5.5 A resolution. *Science*, **292,** 883-896.

SUMMARY IN ESTONIAN

Translatsiooni efektiivsust mõjutavad järjestuse motiivid

Valke, igas rakuprotsessis osalevaid makromolekule, sünteesitakse translatsiooni käigus vastavalt informatsioonile mRNAs. Translatsioon, geeniekspressiooni viimane etapp, toimub ribosoomis. Ribosoomi üldine struktuur ja funktsioon on konserveerunud kogu elusloodes, mis viitab omakorda ka translatsiooni regulatsioonisignaalide konserveerumisele. Samas toimub translatsiooni initsiatsioon kui kõige olulisem valgusünteesi efektiivsust määrav etapp bakterites ja eukarüootides erineva mehhanismi alusel. Eukarüootides alustab ribosoom mRNA skaneerimist mRNA 5’ otsast kuni jõuab startkoodonini. Bakterites seondub ribosoom enne startkoodonit asuvale nn. Shine-Dalgarno (SD) järjestusele aluspaardumise kaudu SD järjestuse ja 16S rRNA 3’ otsa nukleotiidide vahel (antiSD järjestus). Eelnevad eksperimentaalsed tööd on näidanud vaid nõrka korrelatsiooni SD:antiSD aluspaardumise tugevuse ning geeni ekspressioonitaseme vahel.

Käesoleva doktoritöö esimeses tulemuste osas on kirjeldatud SD:antiSD interaktsiooni pikkuse ning geenide ekspressioonitasemete vahelise seose *in silico* analüüsi tulemused *Escherichia coli*’s. Selgus, et keskmine SD:antiSD aluspaardumise pikkus oli sõltumata geeni ekspressioonitasemest sama. See tulemus

ning kaasautorite poolt tehtud eksperimentaalsed analüüsid lubavad oletada, et SD järjestus toimib koostöös enhanser järjestusega, milleks on enne SD järjestust asuv A/U nukleotiidide rikas regioon.

Startkoodoni efektiivseks äratundmiseks esineb bakterite valke kodeerivate järjestuste startkoodoni ümbruses veel teisigi olulisi motiive. Näiteks *Escherichia coli* valke kodeerivates järjestustes on sagedamini esinevaks startkoodonile järgnevaks koodoniks AAA, mida on seostatud kõrge ekspressioonitasemega. Lisaks on kõrge ekspressioonitasemega geenide alguses leitud muidu harvade koodonite sagedast esinemist. Nende üleesindatus kodeeriva järjestuse alguses viitab võimalikule regulatsioonimehhanismile. Eukarüootide startkoodoni ümbruses on kirjeldatud nn. Kozaki consensus GCC(A/G)CCAUGG, mis on vajalik efektiivseks translatsiooniks.

Käesoleva doktoritöö teiseks eesmärgiks oli võrdleva genoomika abil otsida ning kirjeldada võimalikke konserveerunud motiive valke kodeerivate järjestuste alguses. Uuringute tulemusena leiti, et nii bakterite, arheate kui eukarüootide kõrge ekspressioonitasemega geenides esineb mitte AAA koodonite, vaid GCN koodonite tugev üleesindatus vahetult startkoodonile järgneva koodonina. See tulemus langeb kokku osaga Kozaki konsensusest ning viitab sellele, et hoolimata erinevatest translatsiooni initsiatsioonimehhanismidest on translatsiooni initsiatsiooniregiooni järjestus üle elusloodus rohkem konserveerunud kui seni arvatud. Üheks võimalikuks seletuseks GCN koodonite üleesindatusele teise koodonina kõrge ekspressioonitasemega geenides on peptidüül-tRNA ärakukkumise mehhanism, mille käigus peptidüül-tRNA vabaneb ribosoomist enneaegselt ning katkestab normaalse valgusünteesi. Sage peptidüül-tRNA ärakukkumine takistab efektiivset translatsiooni. Peptidüül-tRNA ärakukkumine toimub sagedamini väga lühikeste peptiidide puhul ning koodonitelt, mis sisaldavad A nukleotiidi esimeses või teises positsioonis. Seega võib GCN koodonite eelistamine teise koodonina olla seotud eesmärgiga vähendada võimalikke peptidüül-tRNA ärakukkumise sündmuseid. GCN koodonperekond kodeerib aminohapet alaniin. Uuritud organismide kõrge ekspressioonitasemega valkudes esines tõepoolest ka tugev alaniini üleesindatus teise aminohappena. Alaniin on üks kuuest valke stabiliseerivast aminohappest nii bakterites kui eukarüootides. Kuigi alaniini üleesindatus oli kõige universaalsem ja tugevam, selgus, et kõrge ekspressioonitasemega valkude teises positsioonis leidus kõiki stabiliseerivaid aminohappeid oluliselt rohkem kui organismi kõigi valkude teises positsioonis. Seetõttu on võimalik, et avastatud nukleotiidi ja koodonieelistused kõrge ekspressioonitasemega geenide teises koodonpositsioonis võivad olla seotud ka aminohapete selektsiooniga.

Käesoleva doktoritöö kolmandaks eesmärgiks oli uurida koodonkonteksti konserveerumist elusloodustes. Nimelt lisaks koodonkasutuse eelistustele, kus erinevaid sünonüümseid koodoneid kasutatakse kodeerivates järjestustes erineva sagedusega, on ka koodonpaaride sagedused erinevad. Eksperimentaalsed tulemused on aga näidanud, et koodonkontekst võib translatsiooni efektiivsuse ja täpsusega olla veelgi tugevamini seotud kui koodonkasutus. Siiski on seni arvatud, et erinevates organismides on koodonkonteksti eelistused erinevad. Doktoritöö raames tehtud analüüsid näitasid, et hoolimata üldisest koodonkonteksti spetsiifilisusest erinevates organismides, esineb siiski teatud hulk koodonpaare, mis on üle kogu elusloodus valke kodeerivates järjestustes sarnaselt välditud või eelistatud. Enamus sellistest koodonpaaridest olid tugevamini välditud või eelistatud valke kodeerivates järjestustes võrreldes vastavate heksanukleotiididega genoomides. See viitab translatsioonilisele selektsioonile koodonpaaride kasutuses. Samas ei ole tegu

dipeptidiide kasutusest tulenevate mustritega, kuna nende mõju oli tulemustest välja taandatud.

Kõige sagedamini välditud koodonpaarid sisaldasid mustreid nnUAnn, nnGGnn, nnGnnC, nnCGCn, GUCCnn, CUCCnn, nnCnnA ja UUCGnn. Kõige sagedamini eelistatud koodonpaarid sisaldasid mustreid nnGCnn, nnCAnn, nnUnCn. Välditud nnUAnn koodonpaaridest sisaldasid paljud paarid väljaspool õiget lugemisraami asuvaid UAG ja UAA stoppkoodoneid. Seetõttu võib nnUAnn koodonpaaride vältimine tuleneda enneaegse translatsiooni terminatsiooni vältimisest raaminihke tagajärjel. Seni teadmata põhjustel sisaldasid nnUAnn tüüpi välditud koodonpaarid raamist väljas asuvaid stoppkoodoneid ka antisense ahelal. Samas ei sisaldanud ükski universaalselt välditud koodonpaar UGA stoppkoodonit, mis viitab, et UA dinukleotiidil võib olla stoppkoodonitest eraldiseisev roll.

Üheks koodonkonteksti eelistuste ja vältimiste põhjustajaks on pakutud ribosoomis paikneva kahe tRNA omavaheline struktuuriline sobivus. Kahjuks on tänaseks hetkeks olemas väga vähe informatsiooni erinevate tRNAde omavaheliste interaktsioonide kohta ribosoomis, mis lubaks seda hüpoteesi testida avastatud koodonkonteksti konserveerumise suhtes.

ACKNOWLEDGEMENTS

Firstly, I am grateful to my supervisors prof. Tanel Tenson and prof. Maido Remm. You have taught, guided and supported me since I was undergraduate student. Thanks to your encouragement I have made it this far.

Many thanks to Ülo Maiväli from Department of Molecular Biology for finding the time to read and criticise my manuscripts over and over again.

I would also like to thank my colleagues form the Institute of Technology and Department of Bioinformatics, especially
Vladimir Vimberg as the co-author of the paper;
Triinu Kõressaar for the help with hybridization energy calculations;
Reidar Andreson for the useful suggestions about writing the thesis;
Tõnu Margus for new ideas, friendly attitude and the help with computer programs;
Tõnu Möls and Märt Möls for the help in statistics;
Hedi Peterson – your enthusiasm, motivation and energy to do the science are infectious;
Katre Palm for the help with English and for bringing the joy and warmth during your short stay in our department;
Oliivika Zeiger for doing all those 'little things' we would be in trouble with.

My gratitude belongs to my parents for letting me find my own way and make my own choices, and brother Priit for boosting my confidence.
My friends, especially Christopher Kohver – thank you for the support and humour in life;

Last but not least, Timo – you have been beside me during all my university years sharing the good and sometimes not so good times. Thank you from all of my heart for your love, patience and sacrifices you have made. I would have not been able to do this without you.