Distribution and phylogeny of the bacterial translational GTPases and the Mqsr/YgiT regulatory system

Table of Contents

LIST OF ORIGINAL PUBLICATIONS

I. **Margus, T**., Remm, M. and Tenson, T. (2007) Phylogenetic distribution of translational GTPases in bacteria. BMC genomics, 8, 15.

II. Kasari, V., Kurg, K., **Margus, T**., Tenson, T. and Kaldalu, N. (2010) The Escherichia coli mqsR and ygiT genes encode a new toxin-antitoxin pair. J Bacteriol, 192, 2908-2919.

III. **Margus, T**., Remm, M. and Tenson, T. (2011) A computational study of elongation factor G (EFG) duplicated genes: diverged nature underlying the innovation on the same structural template. PLoS One, 6, e22789.

My contribution to the articles:

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

Ref. II: performed *in silico* analysis;

Ref. III: conceived the project, performed the analysis and wrote the original draft.

LIST OF ABBREVIATIONS

BI	Bayesian inference
BLAST	basic local alignment search tool
DDC	duplication–degeneration–complementation
d_N	substitutions per non-synonymous site
d_S	substitutions per synonymous site
EM	electron microscopy
FRET	fluorescence resonance energy transfer
G1,G2...G5	short conserved motifs of GTPase domain
GAP	GTPase activating protein
GDPCP	5'-guanosyl-methylene-triphosphate
GEF	guanine nucleotide exchange factor
HGT	horizontal gene transfer – same as LGT
HMM	hidden Markov model
InterPro	integrated database of predictive protein signatures
LGT	lateral gene transfer – same as HGT
LUCA	the last universal common ancestor
ML	maximum likelihood method for constructing phylogeny
MP	maximum parsimony method for constructing phylogeny
MSA	multiple sequence alignment
P-loop	structural loop defined from crystal structure of P-loop GTPases
Pfam	a database of protein families, their annotations, and MSA

	generated using hidden Markov models
POST	post-translocational state
PRE	pre-translocational state
PSSM	position-specific scoring matrix
RefSeq	non-redundant and curated subset of Genbank databases
rRNA	ribosomal RNA
SIMBI	a class of P-loop GTPases
SRL	sarcine-ricine loop, structural loop of 23S rRNA
TA	toxin-antitoxin
TADB	Database of type-II toxin-antitoxin systems
TBLASTN	BLAST family program - searches translated nucleotide databases using a protein query
TRAFAC	class of P-loop GTPases named according to translation factors
trGTPases	translational GTPases
ω	number of substitutions per non-synonymous site divided by number of substitutions per synonymous site
Walker A	conserved motif known as G1 motif
Walker B	conserved motif known as G3 motif

INTRODUCTION

Protein synthesis is a fundamental function of cells. The molecular machinery of protein synthesis is highly conserved. Most of its components have clearly recognizable homologs in Archaea, Bacteria and Eukaryota. The machinery involved in this process consists of the ribosome (70S in bacteria and 80S in eukaryotes) and its attendant molecules (e.g. translation factors, RNA, mRNA). The general translation cycle comprises initiation, elongation, termination and recycling phases. The translation factors assist the ribosome in each of these phases. Translation factors that utilize GTP are called translational GTPases (trGTPases). Four large families of trGTPases - IF2/eIF5B, SelB/eIF2γ, EF-Tu/EF-1α and EFG/EF-2 - can be distinguished (Leipe et al. 2002). For each of those families one ancestral gene existed in the last universal common ancestor (LUCA) (Leipe et al. 2002). Additional trGTPase families appeared later. These additional families, which have diverse biological roles in bacteria, are: LepA, TypA, RPP(tetR), RF3 and ATPS2 (CysN). Considering protein domain order and sequence similarities, the LepA, TypA, RPP(tetR) and RF3 genes probably arose after duplications of one ancestral gene from the EFG/EF2 family (Caldon et al. 2001; Inagaki et al. 2002; Connell et al. 2003; Owens et al. 2004; Qin et al. 2006). This suggests that during bacterial evolution an ancient branch of the EFG/EF2 family was a source for protein synthesis-related GTPases with new functional roles.

Analyses of microorganisms with complete genome sequences reveal remarkable variation of protein synthesis machinery among bacteria. We used data from complete genomes to characterize the phylogenetic distribution of trGTPases and to investigate the evolution of elongation factor G in greater detail. We describe the dynamics of gene evolution in terms of duplication, pseudogenization and fixation.

Bacteria have several response systems to rapid changes in the environment. One class of these systems includes the toxin-antitoxin (TA) modules. TA systems have important roles in the physiology of cells in their natural habitats. They are involved in biofilm formation, quorum sensing and multidrug resistance (Gerdes and Wagner 2007; Yamaguchi and Inouye 2011). Several toxins of the TA systems of *Escherichia coli* target protein synthesis. The toxin of the mqsR/ygiT TA system affects protein synthesis by cleaving mRNA. The phylogenetic distribution of the mqsR/ygiT toxin-antitoxin system in bacteria is another topic studied within this dissertation.

1. REVIEW OF LITERATURE

1.1. Protein synthesis

1.1.1. Introduction

Protein synthesis is vital for all living cells, being the last phase of expression of information stored in protein-coding genes. It is performed by the ribosome, a highly conserved RNA-protein complex. The prokaryotic ribosome consists of two asymmetric subunits: 30S and 50S. The small (30S) subunit of the *E. coli* ribosome is formed from 16S rRNA and approximately 20 proteins. The large (50S) subunit is assembled from 23S and 5S rRNA and over 30 proteins. The ribosome is not the sole component of the protein synthesis system. Messenger RNA (mRNA) brings coded information to the ribosome, transfer RNAs (tRNAs) supply the ribosome with amino acids, and translation factors assist the ribosome through the different phases of protein synthesis.

Despite differences in ribosome composition and the number of translation factors among the three domains of life (Archaea, Bacteria and Eukaryota), the basic reactions and translation factors are conserved in all of them (Caldon et al. 2001; Caldon and March 2003). The conserved core set of genes indicates that protein synthesis already existed in the last universal common ancestor (LUCA), a hypothetical life form that was the ancestor of all three domains (Leipe et al. 2002). The variety of functions in present-day organisms is mostly caused by gene duplication(s) followed by the acquisition of a new function by a duplicate – evolution by gene duplication.

1.1.2. Phases of protein synthesis and "classical" translation factors

The protein synthesis cycle comprises four phases: initiation, elongation, termination, and recycling. In the first step, the initiation complex is assembled from the 30S and 50S subunits, mRNA and initiator tRNA (Figure 1). In the elongation phase of protein synthesis, the ribosome decodes the mRNA sequence in discrete steps (codons) using tRNAs as substrates. During elongation the ribosome actively synthesizes proteins through three sequential steps: (i) decoding, (ii) peptide bond formation, and (iii) translocation (Figure 1). Translation enters the termination phase when the stop codon in mRNA reaches the A site. In this phase the synthesized peptide is released from the ribosome, yielding the post-termination ribosomal complex.

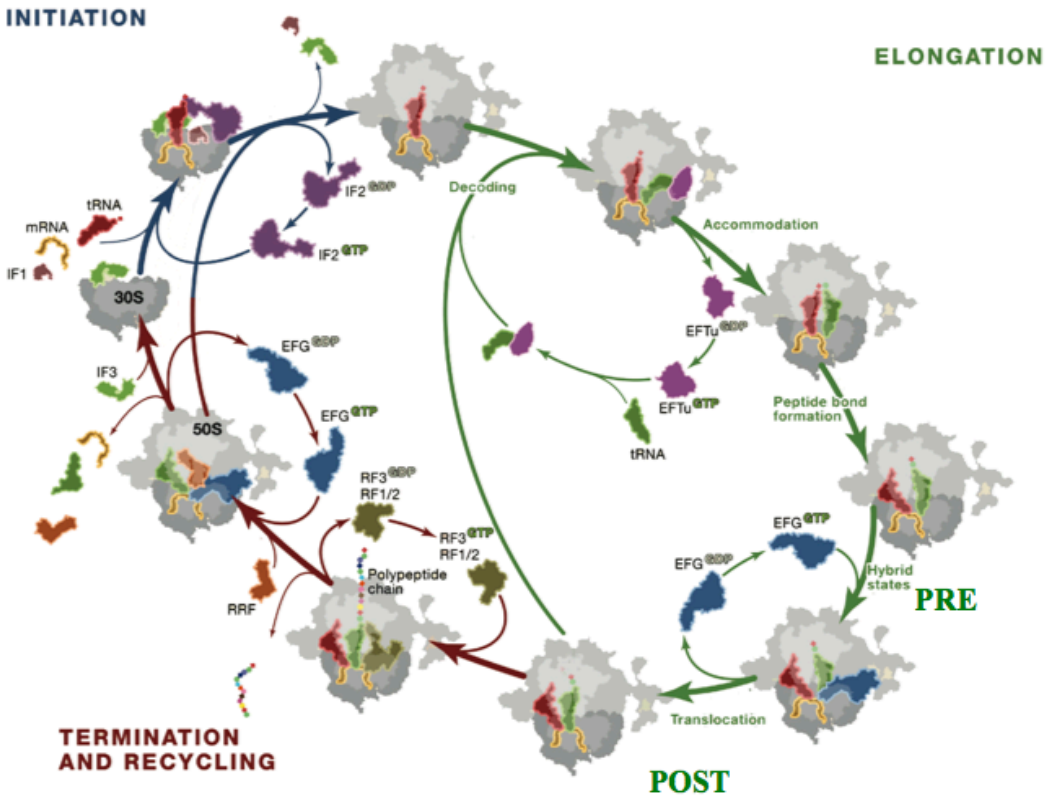

Figure 1. The four phases of protein synthesis: initiation, elongation, termination, and recycling. Modified from Sohmen et al. (2009) (Sohmen et al. 2009).

In the recycling phase, the post-termination ribosome is dissociated into its subunits. tRNA and mRNA also leave the ribosome, thus preparing it for another round of initiation.

Throughout the protein synthesis cycle, the ribosome is assisted by a large number of accessory proteins called translation factors. The protein synthetic machinery is assembled during the initiation of translation – a multistep process that in bacteria is controlled by initiation factors IF1, IF2 and IF3 (Gualerzi and Pon 1990). In the elongation phase, elongation factor Tu (EF-Tu) assists the A-site occupation by an aminoacyl-tRNA (aa-tRNA) (Rodnina et al. 1995), and elongation factor G (EFG) facilitates translocation (Agrawal et al. 1998). To keep the EF-Tu pool charged with GTP, the elongation factor Ts (EF-Ts) is also required. The termination phase is facilitated by three release factors - RF1, RF2 and RF3. RF1 and RF2 recognize a stop codon in an empty A-site, thereby releasing the peptide chain from the

ribosome, whereas RF3 is required for release of RF1 and RF2 from the ribosome (Freistroffer et al. 1997). The recycling phase is carried out by the ribosome recycling factor (RRF) and EFG (Hirashima and Kaji 1973; Karimi et al. 1999).

1.2. P-loop GTPases

1.2.1. Introduction

Proteins that bind and hydrolyze GTP are called G proteins (GTPases). P-loop GTPases and related ATPases share the P-loop fold, which is one of the most common protein folds constituting 10-18% of all protein-coding gene products synthesized by the cell (Koonin et al. 2000). Structurally, P-loop NTPases are α/β proteins comprising a central part consisting of β-sheets (mostly parallel) surrounded by α-helices. The P-loop itself is a relatively small loop – a structural element determined from its crystal structure (Figure 3). At the sequence level, the P-loop NTPases contain a characteristic set of conserved motifs: G1 (also referred to as Walker A motif), G2, G3 (also referred to as Walker B), G4 and G5 (Walker et al. 1982). The G1 motif (Walker A) is located in the P-loop. The P-loop GTPases are divided into two major classes: TRAFAC and SIMBI (Leipe et al. 2002). The TRAFAC class contains enzymes involved in the four phases of protein synthesis (initiation, elongation, termination, recycling), signal transduction, cell motility, and intracellular transport (Leipe et al. 2002).

1.2.2. GTPase cycle

All G proteins go through the same cycle of reactions. Binding and hydrolysis of GTP drive transitions through three conformational states: OFF (GDP-bound), 'empty', and ON (GTP-bound) (Bourne et al. 1991). Hydrolysis of GTP triggers conformational changes. These changes are confined primarily to two segments, called the “switch regions” (Figure 2) (Milburn et al. 1990). The transition between the ON and OFF states is usually induced by the binding of a GTPase-activating protein (GAP) or association of the G protein with a particular conformational state of its cognate target or effector (Figure 2). After GTP hydrolysis, the G protein is in the OFF (GDP-bound) state and needs to be recharged with GTP. Guanine nucleotide exchange factor (GEF) stimulates release of the bound GDP, which is followed by GTP binding to the GTPase.

Figure 2. Schematic representation of GTPase cycle and its regulation. GAP and GEF regulate the GTPase cycle of a G protein by adapting it to cellular needs.

1.2.3. GTPase domain

The GTP binding domains, also known as G domains, share a common and well conserved structural core (Sprang 1997; Vetter and Wittinghofer 2001). This core has the proper nucleotide-binding structure and can be characterized at the sequence level by five conserved motifs: G1-G5 (Figure 3).

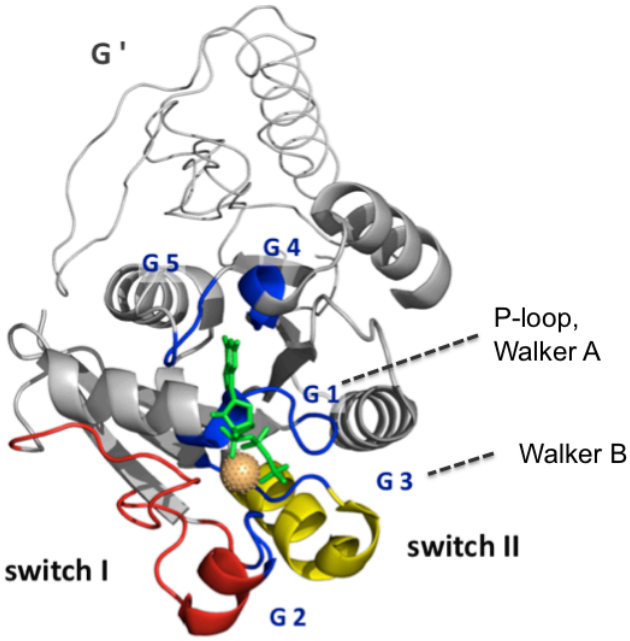

Figure 3. The GTPase domain of EFG. The conserved motifs, G1, G2, G3, G4, and G5, of EFG (PDB code 1WDT) are shown in blue. Structural elements such as the P-loop, switch I and switch II are blue, red and yellow, respectively. G' insertion (between G4 and G5) is shown as a ribbon, with the rest of the structure shown as a cartoon. Walker A and Walker B are early names for conserved motifs G1 and G3, respectively.

Some of these motifs (G1, G2, G3) scan phosphates, discriminating between the tri- and bi-phosphate forms of the bound nucleotide. Motifs G4 and G5 scan the nucleoside part of nucleotide (Table 1)(Bourne et al. 1991).

Table 1. The consensus motifs of the GTPase domain in trGTPases (Bourne et al. 1991)

Motif	Consensus	Function
G1	GXXXXGKT/ST	Interactions with α- and β-phosphates
G2	RGITI	Binding of γ-phosphate and Mg^{2+}
G3	DXPGH	Indirect Mg^{2+} binding
G4	NKXD	Recognition of G nucleotide
G5	GSAL/K	Binding of nucleotide

In Ras proteins it has been shown that GAP interacts with the G2 motif during GTPase activation (Bourne et al. 1991). Since GAP differs among GTPase families, the G2 motif has also evolved to adapt to these changes. For trGTPases, the GAP role is carried out by the large ribosomal subunit (Ramakrishnan 2002; Nilsson and Nissen 2005). There the trGTPases bind to overlapping sites on the ribosome (Ramakrishnan 2002; Nilsson and Nissen 2005). In the three domains of life, the exclusively conserved consensus sequence of the G2 motif is "RGITI".

1.2.4. Towards the identification of the GTPase activation mechanism of trGTPases

The ribosome is a large macromolecular complex. Several parts of the ribosome have been discussed as the candidate GAP for the trGTPases (more on trGTPases in section 1.3.). In their early study, Hamel et al. (1972) showed that the ribosome loses its GTP-inducing property when 50S subunits are incubated in high salt conditions (Hamel et al. 1972). The protein fraction removed by this treatment is primarily the L7/L12 stalk of the 50S ribosomal subunit. EM studies have shown that this part of the ribosome interacts with the negatively charged region of the G' subdomain of EFG (Diaconu et al. 2005; Nechifor et al. 2007). Later studies have confirmed that L7/L12 dimers are necessary for stimulating the GTPase activity of the translation factors, particularly EF-Tu and EFG (Savelsbergh et al. 2000; Mohr et al. 2002). The L7/L12 stalk is important for the recognition of IF2*GTP during initiation of translation (Huang et al. 2010). However, L12 is not a GTPase activating protein (GAP) for trGTPases (Huang et al. 2010). In the absence of L12, the binding of EF-Tu to the ribosome is severely impaired (Kothe et al. 2004) and the reduction

of GTPase activity is probably related to reduced affinity between the ternary complex and the ribosome. Savelsberg et al. (2005) demonstrated that mutating conserved amino acids on the surface of the L7/L12 C-terminal domain (CTD) leads to strong inhibition of EFG turnover, with little effect on rapid single-round GTP hydrolysis and translocation (Savelsbergh et al. 2005).

Recently, two high-resolution (3.2 - 3.6 Å) X-ray structures of the ribosome-bound trGTPases have been determined (Gao et al. 2009; Voorhees et al. 2010). In the first structure, EFG was trapped in the post-translocational state of the ribosome (Gao et al. 2009). In the second structure, EF-Tu was bound to the ribosome with aa-tRNA and a non-hydrolysable GTP analog (Voorhees et al. 2010) (more detail in section 1.3.4.). Voorhees et al. (2010) suggested that A2662 (part of the sarcin-ricin loop [SRL]) of the 23S RNA corresponds to the GAP (Voorhees et al. 2010). They reported that A2662 interacts with His84 (numeration according *E. coli* EF-Tu) and suggested that His84 acts as a general base, which activates the water molecule that attacks the γ-phosphate and hydrolyses GTP (Voorhees et al. 2010). The suggestion that His84 is a general base was criticized by Liljas et al. (2011). They considered it unlikely on several grounds, arguing that in the particular protein environment the His residue is most likely to be positively charged, making it unable to act according to the mechanism proposed (Liljas et al. 2011). In addition, replacing His84 with Ala84 reduces the rate of GTP hydrolysis (in ribosome-bound ternary complex) by six orders of magnitude (Daviter et al. 2003), whereas mutation to Gln84 has a moderate effect (Daviter et al. 2003).

1.3. Translational GTPases (trGTPases)

1.3.1. Introduction

Traditionally, trGTPases are defined as proteins in which the GTPase activity is induced by the large ribosomal subunit (Ramakrishnan 2002; Nilsson and Nissen 2005). Alternatively, computational methods that analyze information hidden in the protein sequence and structural data can be used to determine the relationship between different proteins and their families. Phylogenetic methods and profile-based algorithms extend the set of trGTPases by incorporating members that are evolutionarily related. Bacterial trGTPases consist of the families IF2, EF-Tu, SelB, EFG, LepA(EF4), RF3, RPP(tetR), TypA(BipA), and ATPS2(CysN). Each protein family carries specific function(s) of which some are irreplaceable (vital) to the cell whereas others have effects under specific conditions or environments. Translational

GTPases carrying the same functions in archaea and eukaryotes are usually designated by the prefixes “a” and “e”, respectively (Table 2).

Table 2. Translational GTPases of bacteria, archaea and eukaryotes

Bacteria	Archaea	Eukaryota
IF2	aIF5B	eIF5B
-	aIF2	eIF2
EF-Tu	aEF-1A	eEF-1A
SelB	aSelB	eSelB
EFG	aEF2	eEF2
RF3*	-	eRF3*
LepA(EF4)	-	-
RPP(tetR)	-	-
TypA(BipA)	-	-
ATPS2(CysN)**	-	-
-	-	Hbs1p
-	-	Ski7p
-	-	Snu114p
-	-	Ria1p

* RF3 originated from EFG in bacteria, whereas eRF3 came from eEF1-1A in eukaryotes

** ATPS2(CysN) was acquired laterally and it functions independently of the ribosomes. (This table is based on data from an article by Leipe et al. (2002) (Leipe et al. 2002) and the thesis of Atkinson (Atkinson 2008)).

Some proteins that carry a clear signature of trGTPases have acquired a new function, which is not (directly) related to protein synthesis. For example, ATPS2 (CysN) is known to function as a large subunit of ATP sulfurylase in bacteria; Snu114p in eukaryotes is a part of the eukaryotic spliceosome. The full list of trGTPases in all three domains is shown in Table 2. I use the term trGTPases throughout this work to refer to bacterial trGTPases, unless otherwise indicated.

1.3.2. Three essential sets of trGTPases

Most of our knowledge about protein synthesis has come from a few well-studied model organisms. It is natural that the classical set of trGTPases is based on protein synthesis in *E. coli*. These trGTPases include IF2, EF-Tu, EFG, and RF3, which together cover the four phases of protein synthesis (Figure 1).

An overlapping but slightly different set of trGTPases emerges when ancestral branches of GTPases are identified. Analyzing evolutionary relationships of P-loop GTPases led to the definition of four groups of trGTPases traceable to LUCA (Leipe et al. 2002). These big families are: IF2/eIF5B; SelB/eIF2γ; EF-Tu/EF-1α; and EFG/EF-2 (Leipe et al. 2002). Unexpectedly, SelB/eIF2γ was detected in LUCA, but RF3 was not. Does this mean that the function catalyzed by SelB is more conserved in bacteria than the function catalyzed by RF3? SelB brings selenocystein tRNA to the ribosome by recognizing the stop codon UGA in a specific context (Bock et al. 1991). However, SelB has a patchy distribution across the tree of life and only 20% of bacteria have it (Romero et al. 2005; Margus et al. 2007).

With the completion of sequencing of the first bacterial genome (*Haemophilus influenzae*) in 1995, biology entered the genomic era. By reading the "DNA book" written in a four-letter alphabet we can determine most building blocks, pathways, regulators and other vital components essential for the living cell. Using the entire genome sequence it is also possible to determine which genes are absent from the genome of a given species. Comparing the repertoire of complete genomes enables us to see the whole picture from another perspective than is prescribed by studying a model organism or a single system. This was the approach we took in determining the distribution of trGTPases in bacteria (Margus et al. 2007). One of the results that emerged was a definition of the core set of trGTPases in bacteria, which comprises IF2, EF-Tu, EFG and LepA(EF4) (Margus et al. 2007). LepA is almost ubiquitous among bacteria (Margus et al. 2007). Eukaryotic LepA originated in chloroplasts or mitochondria. A back-translocase function has been assigned to LepA (Qin et al. 2006), but its exact effect(s) are still debatable (Liu et al. 2011).

1.3.3. Domain architecture of trGTPases

Domains are the basic building blocks of protein structure and they are also the basic evolutionary units. Most domains have conserved and specific "signatures" that can be converted to sequence models and stored in specific motif databases, e.g. Pfam or InterPro (Hunter et al. 2009; Punta et al. 2011). These models can be used to assign functional annotation to novel protein sequences.

Translational GTPases are multi-domain proteins comprising at least three different domains. All trGTPases have two domains in common - the GTPase domain and domain II. Additional domains are characteristic of a specific family and/or shared between closely related families (Figure 4). The primary sequence of the GTPase domain is well conserved. Domain II structure is conserved, but the primary

sequence can differ considerably among families.

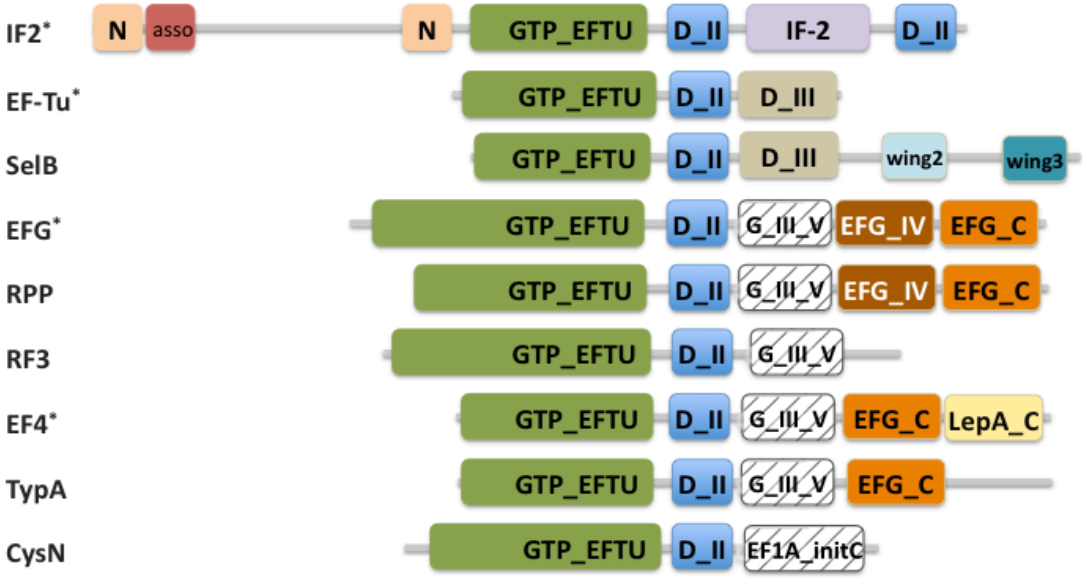

Figure 4. Domain architecture of bacterial trGTPases. Colored boxes indicate domains defined in Pfam; dashed line boxes represent InterPro domains. Domains are given as they are in Pfam/InterPro: N – IF2_N; asso – IF2_assoc; GTP_EFTU – GTP_EFTU; D_II - GTP_EFTU_D2; D_III – GTP_EFTU_D3; IF-2 – IF-2; wing2 – SelB-wing 2; wing3 – SelB-wing 3; EFG_IV – EFG_IV; EFG_C – EFG_C; LepA_C – Lepa_C. Domain names in the figure and InterPro are: G_III_V – Elongation fac G/III/V; and EF1A-initC – Transl elong EF1A/init IF2. Asterisks denote members of the core set of trGTPases in bacteria.

Additional domains can be specific to a family (such as IF2_N in IF2 or Wing domains in SelB) or several families. Family-specific domains are usually located in either the N or C terminus and carry a specific function for the family. For example, SelB-wing domains recognize mRNA loop structures (SECIS element). The SECIS element specifies the UGA stop codon that is used for incorporating selenocystein (Soler et al. 2007). The LepA C terminal domain (LepA_C) has a unique structure with currently unknown function (Evans et al. 2008).

The shared presence of additional domains can predict relationships among these families. The third domain of EF-Tu (GTP_EFTU_D3 in Pfam) is involved in binding of charged tRNA and EF-Ts (Wang et al. 1997). The same domain is seen in another elongation factor, SelB. Its function is similar to EF-Tu, but is restricted to a specific case – incorporating selenocystein. Another universally conserved family is the

EFG/EF-2 family (Leipe et al. 2002). The EFG and RPP(tetR) domain structure is identical, but their functions are different. While EFG catalyzes translocation, RPP(tetR) helps to overcome translation arrest caused by the antibiotic tetracycline (Chopra and Roberts 2001; Roberts 2005). There are three more families (RF3, TypA, and LepA) among the trGTPases that contain one or both of the additional domains first described in EFG. These domains are G_III_V and EFG_C.

1.3.4. Structures of trGTPases and their functional complexes

One of the first trGTPases whose structure was determined at high resolution (2.7Å) was EF-Tu (1EFM) (Jurnak 1985). It took almost 10 years to resolve the structure of another elongation factor, EFG (AEvarsson et al. 1994; Czworkowski et al. 1994). Comparison of the EF-Tu and EFG structures revealed similarities between the GTPase domain and the second domain, but also pointed to differences. The part of the structure formed by EFG domains III, IV, and V is absent from EF-Tu (AEvarsson et al. 1994; Czworkowski et al. 1994). However, when the EF-Tu structure with bound aa-tRNA and nucleotide was determined, similarities between the overall shape of the ternary complex and EFG became evident (Nissen et al. 1995). Thus, three domains (III, IV, V) of the protein EFG mimic the tRNA part of the ternary complex (Figure 5) (Nissen et al. 1995; Nyborg et al. 1997).

From the EFG structure it was also proposed that a conformational change in EFG, coupled with GTP hydrolysis, drives the translocation by physically chasing the newly formed peptidyl-tRNA from the ribosomal A site to the P site (Abel and Jurnak 1996; Nyborg et al. 1997).

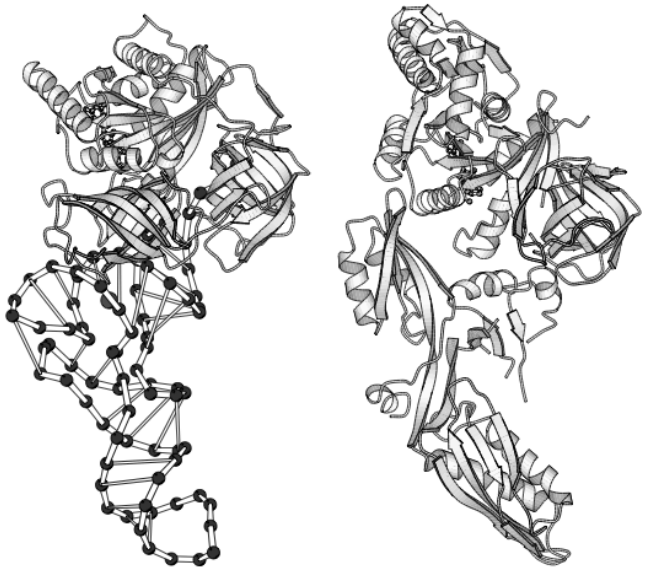

Figure 5. Macromolecular mimicry between the ternary complex and EFG. The ternary complex is to the left and EFG is to the right. In EFG, domain III (not fully resolved) is just below domain II. Domain V is to the left of domain III, while the elongated domain IV is at the bottom. The figure is adapted from Nyborg et al. (1997) (Nyborg et al. 1997).

The finding that the structure of the ternary complex is similar to the structure of EFG led to the molecular mimicry hypothesis (Ito et al. 1996; Nakamura 2001). This proposes that different translation factors evolved independently, but acquired similar structures determined by the nature of their overlapping binding sites on the ribosome (Nakamura 2001; Ito et al. 2002).

More structures of bacterial trGTPases have gradually become available. The structure of EFG-2 of *T. thermophilus* appeared in PDB in 2005. Also, the X-ray structures of ATPS (Cys N), RF3, SelB, LepA, and TypA/BipA have been resolved during the last seven years (Table 3) (Mougous et al. 2006; Gao et al. 2007a; Soler et al. 2007; Evans et al. 2008; Nocek et al. 2008).

Table 3. Structures of trGTPases and their complexes with the ribosome

PDB code	Year	Description	Reference
1efm	1985	EF-Tu with GDP	Jurnak, F. et al., Science 1985
1efg*	1994	EFG with GDP	Czworkowski, J. et al., EMBO J 1994
1elo*	1994	EFG without nucleotide	Aevarsson, A. et al., EMBO J 1994
1ttt	1995	EF-Tu*Pht-tRNA*GDPNPN	Nissen, P. et al., Science 1995
1wdt**	2005	EFG with GTP	Connell, S.R. et al., Mol. Cell 2007
1zun	2006	ATPS (CysN) heterodimer	Mougous ,J.D. et al., Mol. Cell 2006
2h5e	2007	RF3*GDP	Gao, H. et al., Cell 2007
2ply	2007	SelB*SECIS-RNA	Soler, N. et al., JMB 2007
3cb4	2008	LepA(EF4)	Evans, R.N. et al., PNAS 2008
3e3x	2008	TypA/BipA C-terminal part	PDB entry
2wri, 2wrj	2009	70S*EFG*GDP*FA***	Gao et al., Science 2009
2xqd, 2xqe	2010	70S*EF-Tu*GDPCP****	Voorhees et al., Science 2010
3sfs, 3sgf	2012	70S*RF3*GTP*****	Zhou et al., RNA 2012

(*) structures of EFG representing the EFG I subfamily

(**) structures of EFG representing the EFG II subfamily

(***) 70S ribosome complex with EFG and fusidic acid (FA)

(****) 70S ribosome complex with EF-Tu and un-cleavable GTP analog (GDPCP)

(*****) 70S ribosome complex with RF3 and GTP

High-resolution crystal structures of both the large and small ribosomal subunits have led to an invaluable framework for studies of different phases of protein synthesis (Ramakrishnan 2002; Schmeing and Ramakrishnan 2009). Combining X-ray structures and EM reconstructions provided a structural explanation of translocation. A model was proposed in which tRNA movements are facilitated by head-swivel ratcheting and unratcheting motions of the ribosome (Gao et al. 2009; Ratje et al. 2010). Resolving the structure of the 70S ribosome with the ternary complex (EF-Tu*aa-tRNA*GDPCP) deepens our understanding of GTP hydrolysis by the trGTPases (Voorhees et al. 2010).

1.3.5. Evolutionary relationship of trGTPases

In their study of the classification and evolution of P-loop GTPases, Leipe et al. (2002) defined four superfamilies of trGTPases, which can be traced back to LUCA (Leipe et al. 2002). However, the whole set of trGTPases extends to nine families, indicating that some of them appeared later during bacterial evolution (Margus et al. 2007).

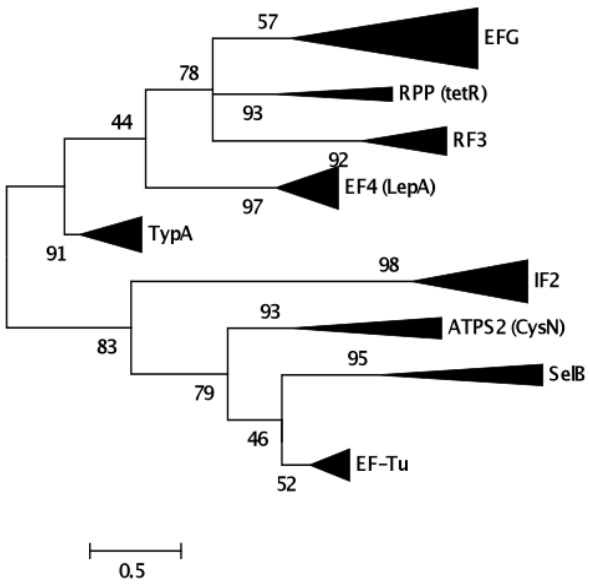

Figure 6. Unrooted tree of bacterial trGTPases. The bootstrap consensus tree inferred from 100 replicates is taken to represent the evolutionary history of the taxa analyzed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) is shown next to the branches (Felsenstein 1985). The analysis involved 85 amino acid sequences. All positions containing gaps and missing data were eliminated. There was a total of 208 positions in the final dataset. Evolutionary analyses were conducted in MEGA5 (Abdulkarim and Hughes 1996; Tamura et al. 2011).

Constructing phylogeny reveals closely related proteins and enables one to define a set of families sharing the same ancestral composition. For example, RPP, RF3 and EFG are neighbor branches (Figure 6). The relatedness of these three families is well supported (bootstrap value 78) but branching order is not resolved. When interpreting such trees we must consider that they have been built on the basis of domains shared among all families, in this case the GTPase domain and domain II. Information about possible additional similarities between subsets of families is not reflected on the tree (Figure 6). Although EFG shares three additional domains with RPP(tetR) and only one additional with RF3, this information is not used for building the sequence-based tree and the branching order is not reliably resolved (Figure 6). LepA(EF4) and TypA actually share two additional domains (G_III_V and EFG_C) with EFG, despite being evolutionarily rather distant from it. The phylogenetic tree of trGTPases (Figure 6) does not conflict with the composition of domains; it rather lacks the power to resolve branching order.

There have been numerous examples where gene duplication and a following acquisition of new function have been shown to be the most parsimonious explanation for the appearance of additional families (Hughes 1994; Force et al. 1999; Van de Peer 2004; Wojtowicz and Tiuryn 2007). Usually, such additional families carry out some auxiliary function and are needed in specific phases of life or under certain environmental conditions. Phylogenetic profiling, where non-uniform and/or patchy distribution has been found to be characteristic of additional trGTPases, supports this assumption (Margus et al. 2007). In this context, the presence of LepA in almost all bacterial genomes is remarkable. Another surprising observation was the rare presence and patchy distribution of SelB – a member of an ancient trGTPase family (Leipe et al. 2002; Margus et al. 2007). The key components of the Sec-decoding trait are SelA, SelB, SelD, and YbbB. This trait is preferred by bacteria that inhabit high temperature and anaerobic environments and is rare in bacteria living at low temperatures and under aerobic conditions (Zhang et al. 2006). The rare presence of SelB today could be therefore related to the bias in choosing bacterial species for sequencing. Another reason for the rarity of SelB could be general geological changes on Earth – the appearance of oxygen and cooling of the planet's mantle.

ATPS2 (CysN) is an unusual trGTPase. The gene for CysN evolved from an archaeal or eukaryotic elongation factor 1α (EF-1α) by LGT, followed by a change in

the function of the gene (Inagaki et al. 2002). Bacterial CysN retained its GTPase activity, which regulates production of APS (adenosine-5'- phosphosulfate), but it lost the requirement for the ribosome to trigger GTP hydrolysis. CysN probably has no function in translation (Mougous et al. 2006).

1.3.6. trGTPase functions under debate

The primary functions of universally conserved trGTPases are well known and have been discussed above. They also appear to have "moonlighting" functions – additional activities unrelated to their main role in the cell. For example, acting as a chaperone by mediating protein folding might be an additional function of IF2, EF-Tu and EFG (Caldas et al. 1998; Caldas et al. 2000).

In some cases the primary function is still (or again) debated. One such protein is the classical trGTPase RF3. RF3 catalyzes a GTPase-dependent release of type I release factor (RF1 or RF2) from the ribosome indicating a function related to termination (Freistroffer et al. 1997; Zavialov et al. 2001). However, Zaher and Green (2011) showed that RF3 maintains a post-peptidyl-transfer quality-control (PT QC) mechanism by which mistakes are assessed retrospectively, i.e. after formation of the peptide bond (Zaher and Green 2011). The key event is the induction of RF3-dependent termination – induced by the end of translation cycle or by mistakes made during translation.

The elongation cycle in protein synthesis is characterized by oscillation of the ribosome between the pre-translocation (PRE) and post-translocation (POST) complexes (Figure 1). Qin et al. (2006) showed that LepA can catalyze reverse translocation *in vitro*, i.e. LepA binds to the POST state and back-translocates stalled ribosomes under high Mg^{2+} concentration (Qin et al. 2006). They proposed that the primary effect, increased activity of the reporter protein, is caused by increased fidelity under an elevated Mg^{2+} concentration. However, Shoji et al. (2010) demonstrated that the *ΔLepA* strain does not show increased frequency of miscoding or frameshifting errors under normal or stress conditions, which indicates that LepA does not contribute to the fidelity of translation (Shoji et al. 2010). LepA function is probably related to proper protein folding by decreasing the rate of synthesis (Shoji et al. 2010; Liu et al. 2011). The observed effects are higher under suboptimal and/or stress conditions when membrane-bound LepA is released into the cytoplasm (Pech et al. 2011). Thus the mechanism enables the cell to respond quickly to sudden and dramatic changes in the environment, which explains why LepA is so well conserved in bacteria.

The fact that some bacteria have multiple genes coding for EFG has been known for some time, but it has been unclear whether the copies have similar or different functions. Connell et al. (2007) showed that EFG-2 in *T. thermophilus* is active in poly(U) synthesis, i.e. it does not differ significantly from EFG-1 (Connell et al. 2007). Suematsu et al. (2010) demonstrated that in the spirochaete *Borrelia burgdorferi* EF-G1 is a translocase, whereas EF-G2 is exclusively a recycling factor (Suematsu et al. 2010). In this context, the absence of any link between protein synthesis and EFG-2 in the actinobacterium *Mycobacterium smegmatis* was somewhat unexpected. They performed several experiments and demonstrated that: (a) MsEFG2 knockout had no effect under several growth conditions; (b) MsEFG2 did not complement MsEFG1; (c) MsEFG2 bound GTP, but GTP hydrolysis was not induced by the ribosome (Seshadri et al. 2009). The results obtained from the *M. smegmatis* system suggested a novel (unknown) function and therefore testing it and/or finding an adequate assay proved to be complicated. Which route the different EFG paralogs had taken, and which processes have shaped the EFG family during evolution, remain intriguing questions.

1.4. Toxin-antitoxin system in bacteria

Toxin–antitoxin (TA) operons are common among free-living bacteria. The toxin products of TA operons target various cellular functions that regulate cell growth and death (Gerdes et al. 2005). TA systems have important roles in the physiology of cells in their natural habitats, including biofilm formation, quorum sensing, formation of persistors, and multidrug resistance (Gerdes and Wagner 2007; Yamaguchi and Inouye 2011). In *E. coli*, cellular targets of the TA system toxins include the protein synthesis machinery (mRNA, tRNA, 30S, and 50S ribosome subunits), DNA replication and the cytoskeleton (Tan et al. 2011). The main target of TA systems in *E. coli* is protein synthesis. The same is probably true for other bacteria.

A toxin-antitoxin system usually consists of two closely linked genes that together encode both a stable toxic protein and a short-lived inhibitor of the toxin. On the basis of the function of the antitoxin, all TA systems have been classified into three groups: types I, II and III. In type I, toxin expression is inhibited by binding of an antisense antitoxin RNA to the toxin-coding transcript (Gerdes and Wagner 2007). The type II TA system utilizes a protein antitoxin to keep the toxin inactivated via protein-protein interaction. In type III, RNA binds to the toxin protein, resulting in a

non-toxic RNA-toxin complex (Fineran et al. 2009). Most of the known TA systems belong to type I or type II.

Inactivation of the antitoxin in response to stressful changes in the environment activates the toxin. Chromosome-encoded TA systems might act as bacterial programmed cell death executioners. In *E. coli* the MazE-MazF system leads to cell death (Hazan et al. 2004) under a wide range of stressful conditions. Other workers have shown that TA toxins are activated in response to stress and starvation, but cell death does not seem to follow, i.e. the toxins induce reversible growth arrest (Christensen-Dalsgaard et al. 2010). The RelB-RelE TA system's involvement in response to amino acid starvation is one of the best-studied examples. RelE toxin is activated by proteolysis of the RelB antitoxin, which leads to cleavage of ribosome-associated mRNA, followed by overall shutdown of translation and an increase in the concentration of aa-tRNAs (Christensen and Gerdes 2004). Adjustment of nutrient consumption and increased translational fidelity allow bacteria to survive starvation. Thus, TA toxins seem to be global regulators of metabolism, growth and division.

TA operons are commonly described as mobile genetic elements (Sevin and Barloy-Hubler 2007). Owing to their mobility, TA systems show a patchy distribution among prokaryotic genomes. Some genomes contain tens of TA systems whereas others have none (Sevin and Barloy-Hubler 2007; Shao et al. 2011). For example, there are eight well-characterized TA systems (Yamaguchi and Inouye 2011) and 29 putative TA systems in *E. coli* (Sevin and Barloy-Hubler 2007; Shao et al. 2011). Approximately 60 putative TA systems have been predicted in the genome of *Mycobacterium tuberculosis*, whereas only two have been detected in the genome of its non-pathogenic counterpart, *M. smegmatis* (Pandey and Gerdes 2005). This indicates that the TA systems are also related to bacterial pathogenicity.

Identification and annotation of TA systems is problematic due to the small size of the toxin and antitoxin genes. Moreover, most of these genes may have atypical GC content and codon usage. To overcome these obstacles, specialized software for identifying TA gene pairs has been developed (Sevin and Barloy-Hubler 2007; Guglielmini et al. 2008). These tools use the information from already-characterized TA families and are useful for detecting missing ORFs in two-gene TA operons. A more complex task was undertaken by Makarova et al. (2009), who analyzed 750 completed genomes of bacteria and archaea and predicted 12 new families of toxins and 13 families of antitoxins (Makarova et al. 2009). All these predictions, results of related experimental work and extensive literature information from PubMed were

gathered into one database - TADB (http://bioinfo-mml.sjtu.edu.cn/TADB/) (Shao et al. 2011). TADB is an integrated database that provides comprehensive information about Type II toxin–antitoxin (TA) loci (Shao et al. 2011). It contains information about 10,753 Type II TA gene-pairs identified within 1240 prokaryotic genomes (Shao et al. 2011). However, the function is unknown for a strikingly large fraction of TA systems (or TA-like systems) and many more cellular targets will be identified for TA systems that have yet to be characterized.

1.5. Evolution by gene duplication

1.5.1. Introduction

To define evolution briefly, I have chosen to cite Arthur Lesk who wrote “Evolution is the change over time in the world of living things” (Lesk 2008). An efficient way to create something new in this world is often to modify something that already exists, i.e. by duplicating and modifying genetic material. One of the earliest observations of duplication of genetic material was made by Bridges in 1936. He reported the doubling of a chromosomal band in a mutant fruit fly that had extremely small eyes (Bridges 1936). A potential role of gene duplication in evolution was suggested and various scenarios of duplicate gene evolution were proposed later (Stephens 1951; Nei 1969). In his influential book “Evolution by gene duplication”, Susumo Ohno popularized this idea further (Ohno 1970). He reasoned that a single copy is enough for the gene to function and therefore extra copies would be redundant (Nei 1969; Ohno 1970). A new copy accumulates mutations more freely and most often becomes a pseudogene (in the process of pseudogenization). Ohno suggested that during the accumulation of neutral mutations, a new gene function can occasionally appear that will be maintained by selection (the process of neofunctionalization) (Ohno 1970). His ideas started to flourish from the late 1990s, when the first genome sequences were completed and the prevalence and importance of gene duplication was clearly demonstrated. However, empirical data also suggested that many more gene duplicates are preserved than predicted by the neofunctionalization model. To explain this conundrum, Hughes (1994) and later Force et al. (1999) proposed models that introduced the idea of splitting the functions of the original gene between paralogs (the process of subfunctionalization) (Hughes 1994; Force et al. 1999). Since then, many models of gene duplication have been proposed. However, because of the lack of a comprehensive framework, it is tedious to discriminate among these different models.

1.5.2. Classification of gene duplication models

The aim of this section is to give a short overview of the classification of gene duplication models, based on phases leading to the stable preservation of a duplicated gene according to Innan and Kondrashov (2010) (Innan and Kondrashov 2010). It provides the common framework for discussing gene duplication models and brings out the main differences among the categories. It does not discuss each model in depth.

In competing for evolutionary preservation, all genetic changes undergo three main stages: (a) origin through mutation, (b) fixation phase, and (c) preservation phase. Gene duplications follow this scenario with one addition: the acquisition of differences between the copies can alter the chance that both copies will be preserved. Approximately a dozen models of gene duplication have been proposed over the years. Many of them describe the phase of acquisition of differences between gene copies as critical for the preservation of a new gene. This phase is referred to as the fate-determination phase (Figure 7). Figure 7 is based on the neofunctionalization model, but with small modifications it can be generally applicable.

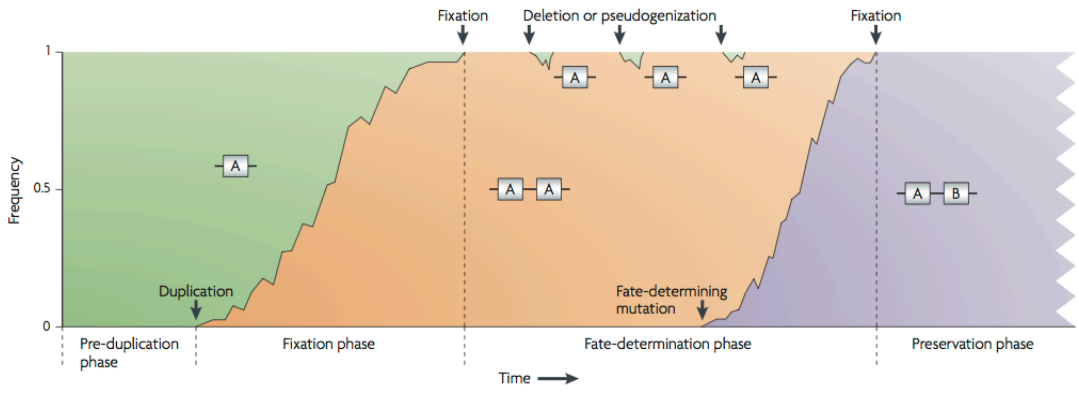

Figure 7. Phases leading to the stable preservation of a duplicated gene. Adapted from Innan and Kondrashov (2010) (Innan and Kondrashov 2010).

Focusing on the selective forces and evolutionary events at different stages in the life history of the duplication, Innan and Kondrashov (2010) claimed there is substantial overlap in the descriptions and predictions of different models (Innan and Kondrashov 2010). They grouped several models in the same category (Innan and Kondrashov 2010). The scenario in which a new duplicate gene pair (A–A) will be fixed in the population of a diploid organism with probability 1/2N over an average of 4N generations defines the models belonging to category I (e.g. popular neo- and

subfunctionalization models). Models in this category assume that duplication would not affect fitness (fixation of the copy being a neutral process). As a consequence, gene duplication must go rapidly through the fate-determination phase. When it does not, one of the copies becomes pseudogenized, i.e. a race takes place between pseudogenization and the appearance of an advantageous mutation and its selection. This category contains three models: (a) the neofunctionalization model of Ohno (Ohno 1970), (b) the duplication–degeneration–complementation (DDC) model of Force et al. (1999)(Force et al. 1999), and (c) the specialization models (or EAC) of Hughes (1994) (Hughes 1994).

The models in categories II and III involve positive selection. In these cases the fixation probability is higher and the fixation time is shorter than in the neutral case of category I (Innan and Kondrashov 2010). For models under category II, the duplication itself is advantageous. Reasons for this type of adaptation can be: (a) masking a deleterious mutation (Kondrashov et al. 2002), (b) a beneficial increase in gene dosage (Clark 1994), and (c) the possibility of the immediate appearance of a new function (Lynch and Katju 2004). Category III comprises models in which duplication occurs in a gene for which population-genetic variation exists. When polymorphisms become immediately fate-determining mutations they promote fixation of the duplicated copy. Duplication and fixation of a fate-determining mutation is almost instantaneous. Therefore, these models do not have a fate-determining phase. Models in this category are: (a) the adaptive radiation model, (b) the permanent heterozygote model and (c) the multi-allelic diversifying selection model (Innan and Kondrashov 2010). Finally, the dosage balance model is classified as the sole member of category IV. There is no fixation phase in the dosage balance model because the fixation of a duplicated copy occurs simultaneously with other events, e.g. large scale or whole genome duplication (Papp et al. 2003).

1.5.3. Gene duplication models and functional state of a new gene copy

The aim of this section is to create a bridge between gene duplication models and the “final” (functional) states of gene copies. I will also illustrate the difference between these two terms.

There are many more models describing the fate of genes after duplication than there are functional states of a new gene copy after it becomes fixed in a population. Considering the function of the original and the function of its copy, the models described above can be reduced to a few “final states” (insofar as “final state” makes sense in the context of evolution) (Innan and Kondrashov 2010). These possibilities

include: (a) the function of the original is retained and its copy has a new function (e.g. neofunctionalization); (b) the two functions of the original gene are split between paralogs (e.g. subfunctionalization); (c) both copies have the same function (as in positive dosage); (d) both copies have multiple functions (diversifying selection).

Gene duplication models describe the path that starts from the event of gene duplication and ends with fixation, i.e. “final state”. As we can see, there are more different gene duplication models than “final states”. To determine a specific model one needs to test whether natural selection has influenced the fate of the duplicated gene. There is a good theory for measuring selection in protein coding genes. According to this theory, synonymous substitutions are considered neutral and non-synonymous substitutions are considered not neutral. Therefore, most of these models estimating substitutions per synonymous site (d_S) and substitutions per non-synonymous site (d_N) estimate the presence or absence of selection from the ratio of d_N to d_S (Suyama et al. 2006). Selective pressure is measured by the ratio $\omega = d_N/d_S$. When non-synonymous substitutions occur at the same rate as synonymous ones and $\omega = 1$, substitution has no effect on fitness, suggesting neutral evolution. If an amino acid change is deleterious then $\omega << 1$ (purifying selection). When a change offers a selective advantage, non-synonymous changes are fixed at a higher rate than synonymous and $\omega > 1$ (positive selection). For example, in the case of Ohno’s classical neofunctionalization model, the expected selective pressures for the original and a copy in the fate-determining phase will be $\omega_{original} << 1$ and $\omega_{copy} = 1$, respectively. There is asymmetry in a pair (original gene and its copy) in this phase. When a new gene copy reaches the preservation phase, purifying selection is applied to both and $\omega_{original} = \omega_{copy} << 1$.

Substitutions per synonymous and non-synonymous site can reliably be determined when the corresponding sites are unsaturated. This condition is satisfied for most gene families in higher eukaryotes. For bacteria, the same is true only for a tiny fraction of the genes that resulted from recent duplication(s) and are shared among closely related species. For most gene families in bacteria (phyla/class level), synonymous sites are saturated. This makes it impossible to estimate d_S and d_N and to use models of gene evolution. When estimating selection of a gene becomes complicated, the amino acid sequence can be used instead. Protein sequences are presented as 20 symbols (amino acids) and saturation is reached much later than for gene sequences (4 symbols). Proteins with more divergent sequences can be used for analysis – they still contain information. The problem is that there is no good general model for protein sequences, in contrast to gene codon sequences. The root

of the problem is that protein evolution and the relationship of primary sequence to structure and function are poorly understood.

However, when synonymous sites in a new gene copy become saturated, it is likely that this gene/protein has reached the preservation phase. Consequently, the problem can, at least partially, be reduced to discriminating among four functional states ("final states"). These functional states are: (a) the function of original is retained while the new copy has a novel function; (b) two functions of the original gene are split between paralogs; (c) both copies have the same function; (d) both copies have multiple functions.

1.5.4. Positions related to functional change/shift

The aim of this section is to elucidate the evolutionary dynamics of a new gene copy and how it is related to the amino acid residues that are involved in functional changes in the protein sequence.

An amino acid residue is functionally important if it is evolutionarily conserved. Two types of conservation changes have been associated with functional change (Figure 8B). Type I conserved changes result in a shift of a group-specific amino acid property (Lichtarge et al. 1996; Gu 2001). Such divergence is exemplified by a radical shift in the physico-chemical property of an amino acid. Type I conserved positions are also known as cluster-specific residues (Lichtarge et al. 1996; Madabushi et al. 2004), "constant-but-different" (Gribaldo et al. 2003), and type-II functionally divergent positions (Gu 2006).

Another class of conservation changes result in a site-specific rate shift (Gu 1999; Knudsen and Miyamoto 2001; Gaucher et al. 2002). A typical case is an amino acid residue that is highly conserved in a subset of homologous genes but becomes variable in another subset of homologous genes. There are two scenarios leading to similar site-specific rate shifts. According to the first scenario, selection will be lost in a position that is under selection in the original copy, i.e. before duplication. Alternatively, a position with weak (or missing) selection that evolves under purifying selection results in conservation in this position. Typically, it is difficult to determine the pre-duplication pattern of selection and therefore no distinction can be made between these two scenarios. However, when the original copy of a gene has retained its original function(s) and selection pattern, it is possible to make the distinction. In other words we can determine which positions become conserved or relaxed in a new gene (Figure 8B). In terms of further functional characterization, such discrimination is very helpful.

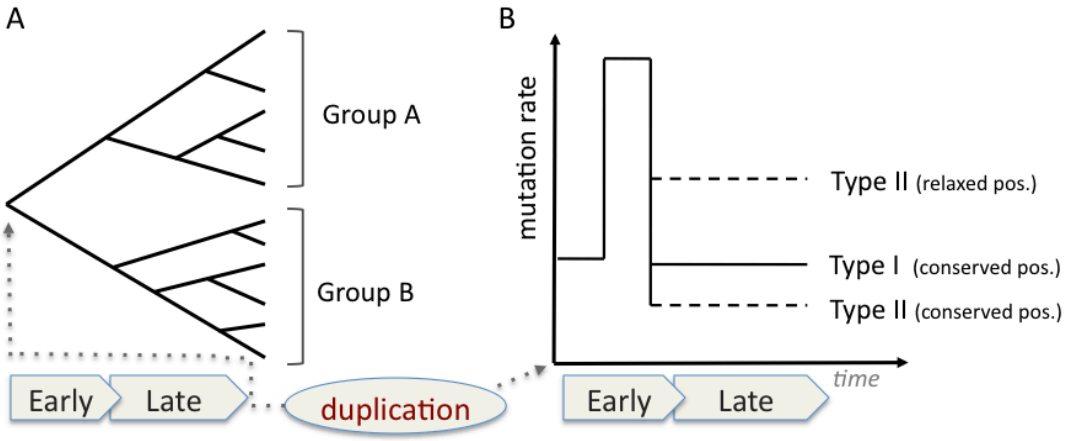

Figure 8. (A) Two groups of genes formed after duplication. Early and late designate the corresponding stages of gene evolution. (B) The mutation rate in the early and late stages of protein evolution after duplication. The evolutionary rate can increase after the gene duplication event for a functional shift-related change, resulting in changed functional constraints between groups A and B. Modified from Gu (1999) (Gu 1999).

It is commonly believed that after a gene duplication event, the evolutionary rate can increase (Li 1997). This phase is called fate-determination by Innan and Kondrashov (2010) (Innan and Kondrashov 2010) or the early phase by Gu et al. (1999) (Gu 1999) (Figures 7 and 8, respectively). During this phase, mutations carrying the essence of new/changed function will appear. These changes lead a new copy to the preservation phase (late phase in Figure 8). On an evolutionary time-scale, it helps us to estimate when a specific function or property appears in a group of organisms. The importance of a preserved gene is proportional to the depth of duplication events in the universal tree of life. Being close to LUCA means longer survival on the stage of evolution and is also proportional to the importance of the gene.

1.6. Bioinformatician's basic toolbox for studying protein families

1.6.1. Molecular data and data quality

In computational biology, one of the main types of data is *sequence data* (sequence of DNA or protein). Another type of data is knowledge about sequences – what they are doing, what is their function, and how their expression is regulated – also referred

to as *annotation*. Nowadays, most annotations of new sequences are transferred from those whose functions are determined experimentally to novel sequences using sequence similarity as the criterion.

Unfortunately, the available sequence data do not represent the entire complexity of living organisms. Only a tiny fraction of all organisms have been sequenced. Comparison of phyla distribution of the completed bacterial genomes reveals that from 1,740 genomes in the database 46% belong to the *Proteobacteria* (795) and 25% to *Firmicutes* (435), leaving ~30% to the other 18 phyla described to date (NCBI 2012). This indicates that fully sequenced genomes are highly biased towards a few common phyla. The diversity of 16S rRNA sequences obtained directly from different environments suggests that our current knowledge about bacteria describes only a small fraction of the diversity (Wu et al. 2009). Therefore, computational biology must deal with highly biased sequence data where reliable functional annotation is relatively rare. A bioinformatics approach enables one to extend functional annotation among homologous sequences to a certain degree.

1.6.2. Sequence alignment and database searching

Many different algorithms have been created to solve sequence alignment problems. Various criteria can also be used to classify these algorithms, e.g. by performing tasks, the methods can be divided into database searching algorithms, multiple sequence alignment algorithms and many other types of alignment algorithms.

The most commonly used program for similarity searches is BLAST (Basic Local Alignment Search Tool) (Altschul et al. 1997). BLAST scans a query sequence against a sequence database. As a measure of the significance of each “match”, the alignment between query and database sequence is given a score (measured in bits) and an E-value, which is the number of expected matches with the same or better bit-score, but without biological significance. The ability of BLAST to detect distant homologs is restricted by the information residing in the sequences compared. The “rule of thumb” states that it is safe to consider sequences homologous when the proportion of identical positions in alignment is >70% for DNA/RNA and >30% for proteins. However, in many cases the real homologs are beyond this safe threshold and cannot be reliably determined. More sensitive methods use models instead of single sequences to detect homology. The models are built from multiple sequence alignments of homologous sequences and include position-specific information about variation for a specific protein family. These methods are slower because they need more steps than a BLAST search. This multistep procedure is included in the

program PSI-BLAST, where the search begins with a simple BLAST and subsequent searches are performed by an algorithm utilizing a position-specific scoring matrix (PSSM) (Altschul et al. 1997). A search is iterative: when new sequences are identified they are added to the model and the next search iteration is performed with an updated PSSM until no more sequences are found. PSSM does not allow gaps (insertions and deletions) to be introduced into the model. Therefore it is best to use PSSMs for sequence families with limited numbers of insertions and deletions. However, during evolution, newly appearing insertions and deletions are quite common and therefore a searching strategy that considers such events is required. The program package called HMMER has been developed to overcome these restrictions (Eddy 1998). This model is based on states of probabilities associated with each position of alignment, and, in addition to amino acids, it contains insertion or deletion as an additional state for each position (Eddy 1998). Because of this feature, HMM models are more sensitive than PSSMs for finding distant homologs, and have been widely used to detect functional domains and to annotate sequences with unknown function (Sonnhammer et al. 1997). HMM models of functional protein domains are collected into the Pfam database, which is based on manually curated and often structure-based alignments of homologous sequences (Bateman et al. 2004).

1.6.3. Multiple sequence alignment

Multiple sequence alignment (MSA) is one of the most widely used methods for simultaneous comparison of protein or nucleic acid sequences (Edgar and Batzoglou 2006). To build an MSA makes sense when a collection of evolutionarily related sequences has been assembled, and one wants to identify features shared by these sequences.

Exact algorithms for calculating optimal MSA require a significant amount of computer memory and computational time. The time and memory requirement increases exponentially with the number of sequences in MSA. These algorithms are able to align up to 10 sequences. Most MSA computing programs are based on heuristics – simplifications to split this complex problem into smaller tasks. One such simplification is known as a progressive alignment algorithm – computing pairwise alignments between all sequences and then constructing one big multiple alignment by progressively joining them. The best-known implementation of a progressive alignment algorithm is CLUSTALW (Thompson et al. 1994), which gained its popularity because it was one of the first user-friendly heuristic MSA algorithms (Thompson et al. 1994). However, it does not refine an already computed alignment

when new sequences are added, so there is concern about readjusting gaps (insertions/deletions in the alignment). A number of powerful algorithms and their implementations have recently been developed. Iterative methods have been implemented in MAFFT (Katoh et al. 2005) and MUSCLE (Edgar 2004), where the progressive alignment step is followed by an iterative procedure to improve the overall alignment. MAFFT scales well in multiprocessor architecture, making it a useful tool for calculating high quality alignments from a large number (400-800) of protein sequences. Consistency-based methods such as PROBCONS (Do et al. 2005) and T-COFFEE (Notredame et al. 2000) combine progressive alignment with a different scoring system. T-COFFEE is probably the most accurate consistency-based program (Edgar and Batzoglou 2006). Early versions of T-COFFEE could align up to 50 sequences when run in accurate mode, but new implementations (version 8.6) have enhanced its performance for an input of up to 200 protein sequences. The T-COFFEE package also contains template-based methods (Expresso and PSI-Coffee) for MSA. A template-based method uses external information, such as X-ray/NMR structures, to improve MSA accuracy. Use of such methods depends on the availability of external information, e.g. on protein structure.

1.6.4. Estimating conservation

A properly constructed MSA is the prerequisite and cornerstone for detecting residue conservation in a protein family. MSA helps to detect the most important amino acids required for proper functioning of proteins in that family. Conserved positions/regions can be estimated visually by inspecting MSA with user-friendly MSA viewers such as JALVIEW (Clamp et al. 2004) or BioEdit (Hall 1998). Consensus sequences are often used to generalize large alignments. It is much easier to compare consensus sequences than alignments. However, consensus sequences have many flaws. As a result, biologically relevant signals are often missed. Information theory provides a mathematically robust way of presenting sequence conservation quantitatively in bits of information using sequence logo graphics (Schneider and Stephens 1990). Sequence logos concentrate on the order of predominance of the residues, their relative frequencies, and information for each specific amino acid at every position in a single graphic. Web Logos is the web interface for constructing sequence logos using MSA as the input (Crooks et al. 2004).

All these tools help to extract signals from sequence alignments and to interpret the results. Shannon's information theory states that the information content of an event is inversely proportional to its expectation, i.e. it increases with unexpectedness (Shannon 1948). Therefore, conserved positions in otherwise highly divergent

backgrounds (30% conserved positions) are more likely to be functionally important residue(s) than those in less divergent backgrounds (80% conserved positions).

1.6.5. Tree-inferring algorithms

Nowadays, there are hundreds of different programs for inferring phylogenetic trees on the basis of four or five different algorithms. The most important algorithms are: distance based (neighbor joining – NJ and UPGMA), maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI). The distance-based algorithms (NJ and UPGMA) are the simplest and also the quickest for inferring a tree. They are able to deal with more than 10,000 sequences. Pairwise distances are computed for the whole set of sequences from which a tree is to be computed. The problem with distance-based algorithms is that the richness of information gathered in sequences is reduced into a single value – distance. MP, ML, and BI are discrete data methods. Basically, they construct trees for every column in the alignment and choose the one that fits best with most columns. An MP algorithm searches for the tree that explains the data with a minimal number of amino acid or nucleotide substitutions. MP algorithms are useful for inferring trees from DNA and coding regions, but they cannot use amino acid substitution matrices and are therefore not used for protein sequences. An ML algorithm weighs the probability of all possible substitutions (amino acid or nucleotide) according to various models of evolution. The likelihood is then the probability of the data, given a tree and the model. The original MP and ML algorithms were relatively slow and were able to compute trees from approximately 50 sequences. Modern ML algorithms take advantage of improved tree-searching heuristics and parallel architecture. For example, the program RAxML (version 7.2.8) is able to compute a phylogenetic tree for 25,000 sequences within two weeks (Stamatakis 2006).

2. RESULTS

2.1. Aims of the study

We have investigated families related to the protein synthesis machinery with our main focus on classical GTP-hydrolyzing translation factors – trGTPases – taking an evolutionary perspective.

The specific foci of the work presented are:

1. Analysis of phylogenetic distribution of trGTPases in bacteria
    a. Develop a reliable methodology for detecting trGTPases from data of completed bacterial genome sequences
    b. Determine the phylogenetic distribution of trGTPases in bacteria
    c. Define the core set of trGTPases in bacteria

2. Evolutionary and functional characterization of EFG paralogs in bacteria
    a. Determine phylogenetic relationships of EFG paralogs
    b. Determine phylogenetic distribution of EFG subfamilies in bacteria
    c. Characterize the EFG II subfamily in terms of its evolution, distribution, and conserved positions most probably related to functional changes

3. Analysis of phylogenetic distribution of mqsR and ygiT, the new toxin-antitoxin system in bacteria
    a. Adapt phylogenetic profiling methodology to analysis of mqsR and ygiT families in bacteria
    b. Determine the phylogenetic distribution of mqsR and ygiT

2.2. Phylogenetic distribution of trGTPases in bacteria (I)

2.2.1. Elaborating methodology for detecting trGTPases

Completed genome sequences, associated predictions and annotation of open reading frames (ORFs) serve as a valuable source of information for bioinformatics studies. However, the quality of annotation in public databases is often unreliable. For example, genes can have different names in different bacteria. Often, the starting position of a gene has not been determined correctly. To overcome these shortcomings, a methodology that can deal with errors of these types was developed. Our methodology integrates analyses of protein and genome sequences.

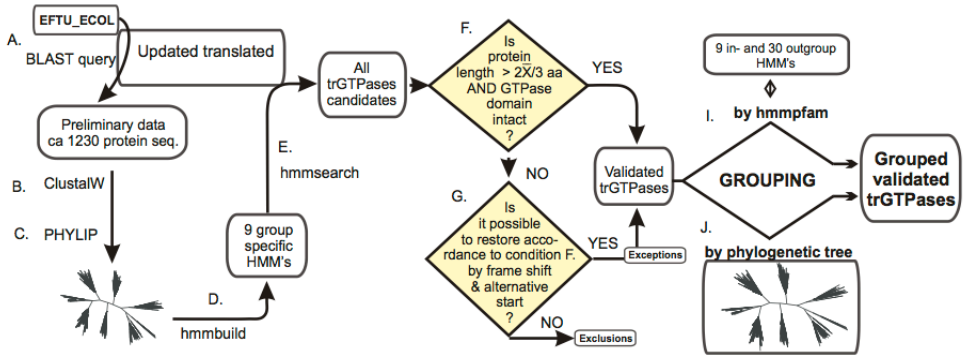

Figure 9. Workflow scheme of the methodology for detecting trGTPases. Activities and data flow are shown by arrows; data are shown in boxes and decision schemes are in rhombi.

At the protein level, the key features are the sensitivity and selectivity of homology detection. This is achieved by using hidden Markov models (HMM) for searching and grouping, and validating the results using tree-based methods (Figure 9). At the genome level, TBLASTN searches ensured that un-annotated ORFs are not missed. This methodology is universal and can be adapted to the analysis of any protein family.

2.2.2. The phylogenetic profiling of trGTPases

The phylogenetic profiling of trGTPases consists of the following steps: (a) determining trGTPases for each genome, (b) grouping these trGTPases into families, (c) computing 16S rRNA-based species tree for bacteria, (d) mapping trGTPase families into a species tree, (f) deriving conclusions based on the distribution of trGTPase families and associated data such as genome size and/or rRNA operon copy number. Assuming that a given completed genome sequence is correct, our methodology is able to determine whether protein families are present or absent in a genome. This type of analysis covered 191 bacteria with completed genome

sequences. One of the main results of this work was a definition of the core set of trGTPases for bacteria, comprising IF-2, EF-Tu, EFG and LepA(EF4). Unexpectedly, this set does not contain RF3. We discovered that RF3 occurs in 62% of the bacteria we examined. The absence of RF3 did not correlate with either genome size or copy-number of rRNA operons. The presence of additional trGTases (RPP[tetR], RF3, SelB, TypA, CysN/NodQ) correlates with genome size. Additional GTPases are rare in small genomes (<1.8Mb). Another interesting finding was related to duplications of the core set trGTPases – EF-Tu and EFG. When EF-Tu copies were almost identical, EFG paralogs appeared to be substantially divergent. The wide distribution of divergent paralogs raises many interesting questions, some of which have been addressed in our later study.

2.3. Phylogenetic distribution of mqsR and ygiT, the new toxin-antitoxin system in bacteria (II)

To detect the ability of a genome to encode the MqsR-ygiT toxin-antitoxin system, sequence similarity searches by BLAST were carried out against sequences from the completed bacterial genomes database (NCBI RefSeq) using MqsR as a query. Among the 914 genomes examined, 40 were found to contain sequences homologous to *E. coli* MqsR. Most hits were found in gamma- and betaproteobacteria, but some putative MqsR toxins were also detected in alphaproteobacteria, deltaproteobacteria, *Chlorobi* and a species of *Acidobacteria*. Most of the genomes contained a single gene for MqsR. Interestingly, three copies of the MqsR gene were found in *Geobacter uraniireducens*.

2.4. Evolutionary and functional characterization of EFG paralogs in bacteria (III)

The EFG gene family in bacteria is highly divergent. The EFG paralogs share only 30–40% identity at the protein level. Furthermore, no experimental data relating to characterization of EFG paralogs were available at the time of the study.

2.4.1. Identification and characterization of EFG subfamilies

Phylogenetic trees for determining EFG subfamilies were constructed using Bayesian inference (BI) and maximum likelihood (ML) methods. We showed that EFG duplicate genes form four subfamilies within the phylogenetic tree: EFG I, spdEFG1, spdEFG2, and EFG II (Figure 10).

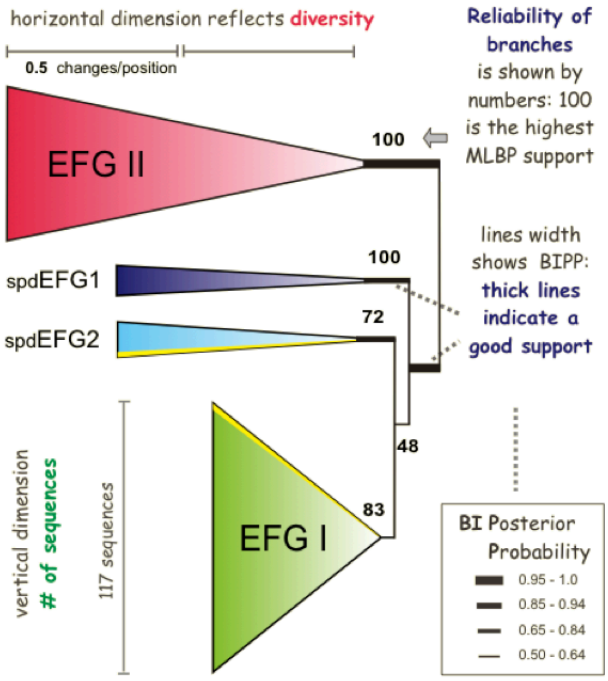

Figure 10. Phylogenetic tree of the four subfamilies of EFG duplications. The tree was constructed using MrBayes (BI) and RaxML (ML). Sequences from the same subfamily are compressed and shown as triangles.

To characterize the evolutionary processes that shape the EFG family, we analyzed the genome context of EFG paralogs and evolutionary events such as recent duplications, lateral gene transfer (LGT) and gene losses via pseudogenization. The results of these analyses and EFG subfamily information were mapped on to the species tree of bacteria, thereby giving a comprehensive picture of the events associated with the evolution of this gene family (Margus et al. 2011).

2.4.2. Analysis of the EFG II subfamily

We characterized the EFG II subfamily in terms of sequence conservation, appearance of insertions and deletions (indels) in multiple sequence alignment, and the evolutionary relationships among members of the EFG II subfamily. Comparison of the EFG II subfamily with the well-studied EFG I revealed some differentially conserved amino acids, which are good candidates for addressing questions about the functional properties of EFG II.

2.4.2.1. Phylogenetic structure of the EFG II subfamily

The distribution of EFG among bacteria showed that each phylum contains at least one group with EFG II as an additional EFG. The phylogenetic tree of EFG II reveals clearly distinguishable phyla/class specific groups called sub-subgroups of EFG II. However, this tree has two peculiarities. First, sequences from two different sub-

subgroups are separated by a large distance in most cases. The approximate estimate by MrBayes is 1.5 changes per position. Second, EFG II sequences from *α-proteobacteria* and *Cyanobacteria* formed one well-supported branch, i.e. EFG II sequences originating from two different phyla belong to the same sub-subgroup. Most of these sub-subgroups are also supported by specific indels determined from multiple sequence alignment.

2.4.2.2. Comparison of the EFG I and EFG II subfamilies

Comparison of these subfamilies was based on sequence conservation information extracted from multiple sequence alignments and expressed as sequence logos (Figure 11). Comparison was made at four different levels: (a) overall conservation, (b) conservation of domains, (c) conservation of GTPase domain motifs G1-G5 and (d) EFG II-specific conserved positions differing from EFG I.

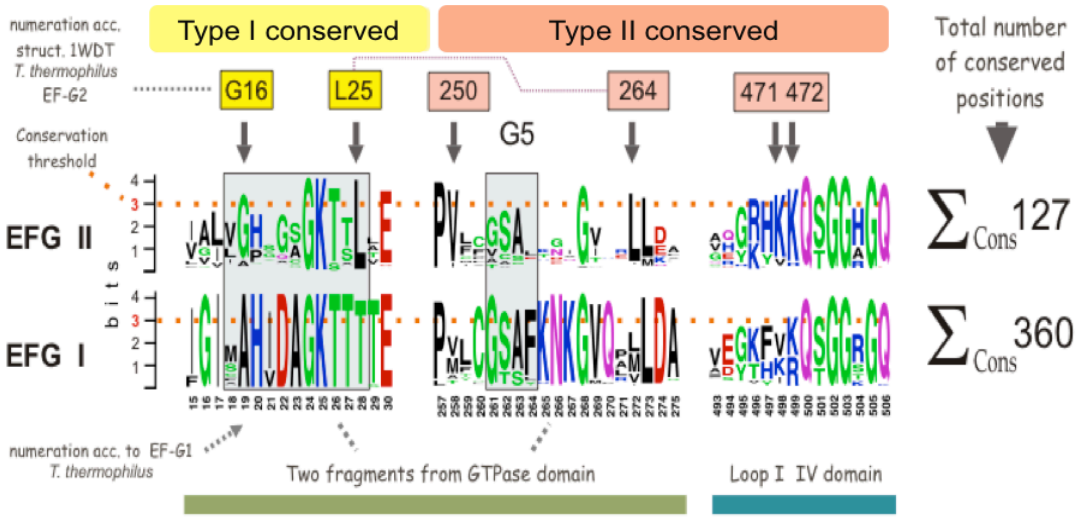

Figure 11. Sequence logos of the EFG I and EFG II subfamilies. Conservation threshold was set to 3 bits. Position conservations of 4.3, 3.3, and 2.8 bits correspond to 100%, 50-50%, and 33-33-33% of specific amino acid(s) conservation in a given position. As an example, two type I conserved positions are indicated by yellow boxes and four type II conserved positions by pink boxes. The bit-score also depends on the number of sequences; when fewer than 20 sequences are used even 100% conservation does not give a reasonable bit-score.

The high diversity within the EFG II subfamily is predominantly caused by the high variation within the GTPase domain, domain II and domain III. Such a low conservation indicates that the first three domains have been evolving under principally different constraints, favoring divergence in the EFG II and homogeneity in

the EFG I subfamilies. Analysis of the GTPase domain showed that short motifs forming the GTP binding pocket, G1, G3, G4, and G5, are conserved. Intriguingly, the trGTPase-specific consensus RGITI in the G2 motif is relaxed. Instead of RGITI, sub-subgroup-specific variants of the G2 motif could be detected (Margus et al. 2011).

EFG II-specific conserved positions were split into two categories. The first consists of five positions at which different amino acids are conserved in the EFG I and EFG II subfamilies (type I conserved positions). Each of these five positions is associated with substantial changes in physico-chemical properties. The second category consists of seven positions that are relaxed in EFG I, but are under stronger selection in the EFG II subfamily (type II conserved positions). Detecting higher conservation in some positions of EFG II than EFG I was unexpected, especially considering the greater divergence of the EFG II subfamily. The locations of type I conserved positions are restricted to the first two domains, the GTPase domain and domain II, whereas type II conserved positions are more uniformly distributed throughout EFG.

All these EFG II-specific positions (types I and II) were mapped on to the EFG structure and the EFG–ribosome complex structure. The importance and possible functional consequences of these EFG II-specific conservations will be discussed in the following section.

3. DISCUSSION

3.1. Bioinformatics methodologies, data quality and presumptions

Reliable methodology is an important part of any scientific study. This is the case for bioinformatics research where custom-made solutions are a common practice. When a protein family can be detected easily and distinguished from other homologous families, a simple BLAST search will be sufficient. A relevant case is searching for the MqsR homologs in bacteria. However, when there are many homologous protein families and the borders between those families are undefined, more sophisticated methodology is required (for example in characterizing trGTPases). Such methodologies often use out-groups so they can be more sensitive without sacrificing selectivity. They are also able to find traces of pseudogenes and can deal with wrong annotations. A methodology of this kind is not a single program; it is a combination of programs and filters as illustrated in Figure 9. Even then, the validity of our conclusions about the gene repertoire in a genome relies on certain presumptions. One presumption is that a given genome sequence is complete and correct. In most genomic regions this is true, but not for toxic and non-clonable DNA. Kimelman et al. (2012) analyzed 393 microbial genomes and mapped >15,000 genes residing in cloning gaps, many of which were toxins (Kimelman et al. 2012). This indicates that the number of toxin genes in bacteria had been underestimated. It is beneficial that the next generation sequencing methods (NGS) do not need the cloning step, so toxic genes/DNA can be discovered.

3.2. Phylogenetic distribution of trGTPase genes

Analyzing the distribution of a specific gene is directly linked to the estimation of its importance to the cell. Widely distributed genes are more important for a wider spectrum of organisms than those with a patchy distribution. By reporting the presence or absence of a gene in a genome and arranging the data according to the species tree, clade-specific genes can be revealed. In the case of trGTPases, IF2, EF-Tu, and EFG are universally conserved. Those proteins are present in all three domains of life. LepA appears to be bacteria-specific; it is not present in archaea or eukaryotes. The question is: what is the specific feature of bacterial ribosomes that distinguishes them from archaeal/eukaryotic ribosomes so that they need a specific GTPase factor for proper functioning? LepA's function as a back-translocase was questioned by Liu et al. (2011), who demonstrated that LepA competes with EFG for

binding to the PRE complex and not to the POST complex (Liu et al. 2011) (see also 1.3.5. on debated trGTPase functions). LepA is located in membranes and is released into the cytoplasm under suboptimal and/or stress-conditions (Pech et al. 2011).

Existing models of protein synthesis have armed us with an understanding of its mechanisms, which take place in a logical order to produce a complete protein (Figure 1). Since many of the processes involved in protein synthesis are conserved, factors that catalyze them are also expected to be conserved. We have demonstrated that the RF3 coding gene was found in only 60% of the genomes analyzed (Margus et al. 2007), leaving 40% of bacteria without it. RF3 is involved in the termination of translation. This “canonical” function was recently challenged by Zaher and Green (2011). They demonstrated that RF3 maintains a post-peptidyl-transfer-quality-control mechanism, evaluating mistakes retrospectively after the peptide bond has formed (Zaher and Green 2011). One might argue that the difference between these two mechanisms is small. Indeed, translation is terminated almost by the same scheme, but there is a substantial difference in the state of the ribosome that induces RF3-dependent termination (see also the debate about trGTPase functions).

3.3. Evolutionary and functional characterization of EFG paralogs

We demonstrated that EFG paralogs form four subfamilies within the phylogenetic tree: EFG I, spdEFG1, spdEFG2 and EFG II (Margus et al. 2011). From an evolutionary perspective, it would be intriguing to ask at which evolutionary stage the EFG gene was duplicated. We proposed the hypothesis that the four EFG subfamilies are the result of ancient duplication (Margus et al. 2011). Our hypothesis is supported by three independent observations. First, spdEFGs appeared at approximately the same time as modern eukaryotic cells carrying mitochondria (Atkinson and Baldauf 2010). Based on phylogenetic tree of EFG and the wide distribution of the EFG II subfamily (second and third observations), the duplication event that gave rise to the EFG II subfamily occurred early in prokaryotic evolution (Margus et al. 2011). Relying on the ancient origin and distinct separation of the EFG subfamilies from one another, it is reasonable to assume that each of them has been on the stage of evolution long enough to acquire its specialized function.

What are these subfamily-specific functions and how have they evolved? There are many models describing gene fate after duplication, which can be reduced to a few

“final states” (insofar as this term makes sense in the context of evolution) (Innan and Kondrashov 2010). These “final states” or possibilities are: (a) where the function of the original is retained and a new copy has a novel function (as in neofunctionalization); (b) two functions of the original gene have split between paralogs (as in subfunctionalization); (c) both copies have the same function (e.g. positive dosage); (d) both copies have multiple functions (diversifying selection) (Innan and Kondrashov 2010). When dealing with ancient duplication(s), the first thing is to determine the “final state” to which a given duplication belongs. The easiest way to test the possibility of splitting function between paralogs is phylogenetic profiling. For a bifunctional protein, the original gene is replaced by two paralogs.

Indeed, functional tests performed by Suematsu et al. (2010) have demonstrated that the functions of bacterial EFG I in *Borrelia burgdorferi* are split between EFG paralogs (in this work, spdEFG1and 2)(Suematsu et al. 2010). Cryo-EM mapping of the *E. coli* EFG complex with RRF predicts a total of five sets of contact points on EFG domains III and IV (Gao et al. 2007b). Atkinson and Baldauf (2010) examined spdEFGs in more detail from an evolutionary perspective, but comparison of spdEFG2-specific conserved positions with the proposed RRF contact points failed to show any correlation (Atkinson and Baldauf 2010). EFG and RRF work as a pair, and changing only one component in this pair can lead to a non-functional complex. For instance, *Thermus thermophilus* RRF is non-functional in *E. coli* (Fujiwara et al. 1999), but it becomes functional upon co-expression of *T. thermophilus* EFG (Ito et al. 2002). We can argue that when a specific RRF works only with a specific set (compatible set) of EFGs, the RRF contact points on EFG are also RRF-specific, i.e. could differ between compatible sets of EFGs. If this is true, then conservation in contact points can be detected by analyzing one compatible set of EFGs. When three or more sets of EFGs are examined together, no conservation in contact points can be detected. Another possibility is that RRF contact points on EFG are not accurate because of limitations in the cryo-EM methodology.

Splitting function between paralogs releases part of a gene from selection and enables new mutations to accumulate. These mutations can lead to the appearance of a new function – neofunctionalization in the shadow of subfunctionalization (Rastogi and Liberles 2005). It has been proposed that this type of evolutionary scenario is more frequent in small/closed populations where the probability of losing a gene copy is high and a gene copy will be preserved owing to subfunctionalization (Rastogi and Liberles 2005). In bacteria, populations are much larger than in

animals, but many bacteria live under strong pressure to minimize the genome and therefore minimize the amount of duplicated genetic material. Atkinson and Baldauf (2010) proposed that spdEFG evolution probably does not follow the simple subfunctionalization model (Atkinson and Baldauf 2010). Whether it follows neofunctionalization in the shadow of subfunctionalization requires further study.

3.4. The EFG II subfamily

Among the trGTPases, the EFG II subfamily is peculiar in several ways. First, it consists of sequences that are highly divergent, much more than the EFG I subfamily. Second, phylogenetic analysis reveals relatively distantly related phyla/class-specific sub-subgroups, an unusual inner-structure of a subfamily. Third, EFG II is widely distributed; ~40% of bacteria contain EFG II as an additional EFG. The divergent nature of the EFG II subfamily encourages us to ask what role this protein really performs. What biochemical functions are common to EFG I and EFG II? Which protein regions/domains carry functions specific to the EFG II subfamily? Is the EFG II subfamily functionally homogeneous? We believe that the set of 12 EFG II-specific conserved positions is the key to answering these questions in future.

SUMMARY AND CONCLUSIONS

The following conclusions can be drawn from this thesis:

1. The core set of trGTPases in bacteria comprises IF2, EF-Tu, EFG, and LepA(EF4). While IF2, EF-Tu and EFG are universally conserved in all domains of life, LepA is a bacteria-specific translation factor.
2. RF3 does not belong to the core set of bacterial trGTPases and therefore the function assigned to it is probably not universal for the bacterial translation system.
3. The mqsR/ygiT TA-system is widespread among bacterial genomes.
4. A divergent set of EFG paralogs form four subfamilies within the phylogenetic tree: EFG I, spdEFG1, spdEFG2 and EFG II.
5. The deep branches on the EFG phylogenetic tree, the wide distribution of EFG I and II and the monophyly of spdEFG1 with mtEFG1 all support the hypothesis that the EFG I and EFG II subfamilies resulted from an ancient duplication of a common ancestor.
6. Twelve distinctive positions are characteristic of the EFG II subfamily. Functional interpretation based on comparison with the EFG I subfamily enables us to propose that:
    a. Positions 16Gly, 25Leu, 61Ser, 216Asp, 250Val, 264Leu are related to modifying the GTPase activity.
    b. Position 352Lys/Arg and increased charges in positions 469..472 are probably related to the interaction of the factor with the ribosome.
7. The phylogenetic tree of EFG II has phyla/class-specific sub-subgroups. These sub-subgroups are characterized by:
    a. A sub-subgroup-specific G2 motif consensus, which differs from the trGTPase-specific RGITI consensus.
    b. Sub-subgroup-specific insertions and deletions.

REFERENCES

Abdulkarim F, Hughes D. 1996. Homologous recombination between the tuf genes of Salmonella typhimurium. *J Mol Biol* **260**(4): 506-522.

Abel K, Jurnak F. 1996. A complex profile of protein elongation: translating chemical energy into molecular movement. *Structure* **4**(3): 229-238.

AEvarsson A, Brazhnikov E, Garber M, Zheltonosova J, Chirgadze Y, al-Karadaghi S, Svensson LA, Liljas A. 1994. Three-dimensional structure of the ribosomal translocase: elongation factor G from Thermus thermophilus. *Embo J* **13**(16): 3669-3677.

Agrawal RK, Penczek P, Grassucci RA, Frank J. 1998. Visualization of elongation factor G on the Escherichia coli 70S ribosome: the mechanism of translocation. *Proc Natl Acad Sci U S A* **95**(11): 6134-6138.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**(17): 3389-3402.

Atkinson GC. 2008. Evolution of the translational GTPase superfamily. In *Department of Biology*, Vol PhD, p. 154. University of York, York.

Atkinson GC, Baldauf SL. 2010. Evolution of elongation factor G and the origins of mitochondrial and chloroplast forms. *Mol Biol Evol*.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL et al. 2004. The Pfam protein families database. *Nucleic Acids Res* **32**(Database issue): D138-141.

Bock A, Forchhammer K, Heider J, Leinfelder W, Sawers G, Veprek B, Zinoni F. 1991. Selenocysteine: the 21st amino acid. *Mol Microbiol* **5**(3): 515-520.

Bourne HR, Sanders DA, McCormick F. 1991. The GTPase superfamily: conserved structure and molecular mechanism. *Nature* **349**(6305): 117-127.

Bridges CB. 1936. The Bar "Gene" a Duplication. *Science* **83**(2148): 210-211.

Caldas T, Laalami S, Richarme G. 2000. Chaperone properties of bacterial elongation factor EF-G and initiation factor IF2. *J Biol Chem* **275**(2): 855-860.

Caldas TD, El Yaagoubi A, Richarme G. 1998. Chaperone properties of bacterial elongation factor EF-Tu. *J Biol Chem* **273**(19): 11478-11482.

Caldon CE, March PE. 2003. Function of the universally conserved bacterial GTPases. *Curr Opin Microbiol* **6**(2): 135-139.

Caldon CE, Yoong P, March PE. 2001. Evolution of a molecular switch: universal bacterial GTPases regulate ribosome function. *Mol Microbiol* **41**(2): 289-297.

Chopra I, Roberts M. 2001. Tetracycline antibiotics: mode of action, applications, molecular biology, and epidemiology of bacterial resistance. *Microbiol Mol Biol Rev* **65**(2): 232-260 ; second page, table of contents.

Christensen SK, Gerdes K. 2004. Delayed-relaxed response explained by hyperactivation of RelE. *Molecular Microbiology* **53**(2): 587-597.

Christensen-Dalsgaard M, Jorgensen MG, Gerdes K. 2010. Three new RelE-homologous mRNA interferases of Escherichia coli differentially induced by environmental stresses. *Mol Microbiol* **75**(2): 333-348.

Clamp M, Cuff J, Searle SM, Barton GJ. 2004. The Jalview Java alignment editor. *Bioinformatics (Oxford, England)* **20**(3): 426-427.

Clark AG. 1994. Invasion and maintenance of a gene duplication. *Proc Natl Acad Sci U S A* **91**(8): 2950-2954.
Connell SR, Takemoto C, Wilson DN, Wang H, Murayama K, Terada T, Shirouzu M, Rost M, Schuler M, Giesebrecht J et al. 2007. Structural basis for interaction of the ribosome with the switch regions of GTP-bound elongation factors. *Mol Cell* **25**(5): 751-764.
Connell SR, Trieber CA, Dinos GP, Einfeldt E, Taylor DE, Nierhaus KH. 2003. Mechanism of Tet(O)-mediated tetracycline resistance. *Embo J* **22**(4): 945-953.
Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome research* **14**(6): 1188-1190.
Czworkowski J, Wang J, Steitz TA, Moore PB. 1994. The crystal structure of elongation factor G complexed with GDP, at 2.7 A resolution. *Embo J* **13**(16): 3661-3668.
Daviter T, Wieden HJ, Rodnina MV. 2003. Essential role of histidine 84 in elongation factor Tu for the chemical step of GTP hydrolysis on the ribosome. *J Mol Biol* **332**(3): 689-699.
Diaconu M, Kothe U, Schlunzen F, Fischer N, Harms JM, Tonevitsky AG, Stark H, Rodnina MV, Wahl MC. 2005. Structural basis for the function of the ribosomal L7/12 stalk in factor binding and GTPase activation. *Cell* **121**(7): 991-1004.
Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome research* **15**(2): 330-340.
Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics (Oxford, England)* **14**(9): 755-763.
Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 1-19.
Edgar RC, Batzoglou S. 2006. Multiple sequence alignment. *Curr Opin Struct Biol* **16**(3): 368-373.
Evans RN, Blaha G, Bailey S, Steitz TA. 2008. The structure of LepA, the ribosomal back translocase. Vol 105, pp. 4673-4678.
Felsenstein J. 1985. Confidence-Limits on Phylogenies - an Approach Using the Bootstrap. *Evolution* **39**(4): 783-791.
Fineran PC, Blower TR, Foulds IJ, Humphreys DP, Lilley KS, Salmond GP. 2009. The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. *Proc Natl Acad Sci U S A* **106**(3): 894-899.
Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**(4): 1531-1545.
Freistroffer DV, Pavlov MY, MacDougall J, Buckingham RH, Ehrenberg M. 1997. Release factor RF3 in E.coli accelerates the dissociation of release factors RF1 and RF2 from the ribosome in a GTP-dependent manner. *Embo J* **16**(13): 4126-4133.
Fujiwara T, Ito K, Nakayashiki T, Nakamura Y. 1999. Amber mutations in ribosome recycling factors of Escherichia coli and Thermus thermophilus: evidence for C-terminal modulator element. *FEBS Lett* **447**(2-3): 297-302.
Gao H, Zhou Z, Rawat U, Huang C, Bouakaz L, Wang C, Cheng Z, Liu Y, Zavialov A, Gursky R et al. 2007a. RF3 induces ribosomal conformational changes

responsible for dissociation of class I release factors. *Cell* **129**(5): 929-941.
Gao N, Zavialov AV, Ehrenberg M, Frank J. 2007b. Specific interaction between EF-G and RRF and its implication for GTP-dependent ribosome splitting into subunits. *J Mol Biol* **374**(5): 1345-1358.
Gao YG, Selmer M, Dunham CM, Weixlbaumer A, Kelley AC, Ramakrishnan V. 2009. The structure of the ribosome with elongation factor G trapped in the posttranslocational state. *Science* **326**(5953): 694-699.
Gaucher EA, Gu X, Miyamoto MM, Benner SA. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem Sci* **27**(6): 315-321.
Gerdes K, Christensen SK, Lobner-Olesen A. 2005. Prokaryotic toxin-antitoxin stress response loci. *Nat Rev Microbiol* **3**(5): 371-382.
Gerdes K, Wagner EG. 2007. RNA antitoxins. *Curr Opin Microbiol* **10**(2): 117-124.
Gribaldo S, Casane D, Lopez P, Philippe H. 2003. Functional divergence prediction from evolutionary analysis: a case study of vertebrate hemoglobin. *Mol Biol Evol* **20**(11): 1754-1759.
Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol* **16**(12): 1664-1674.
-. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol Biol Evol* **18**(4): 453-464.
-. 2006. A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. *Mol Biol Evol* **23**(10): 1937-1945.
Gualerzi CO, Pon CL. 1990. Initiation of mRNA translation in prokaryotes. *Biochemistry* **29**(25): 5881-5889.
Guglielmini J, Szpirer C, Milinkovitch MC. 2008. Automated discovery and phylogenetic analysis of new toxin-antitoxin systems. *BMC Microbiol* **8**: 104.
Hall T. 1998. BioEdit. Biological sequence alignment editor for Windows.
Hamel E, Koka M, Nakamoto T. 1972. Requirement of an Escherichia coli 50 S ribosomal protein component for effective interaction of the ribosome with T and G factors and with guanosine triphosphate. *J Biol Chem* **247**(3): 805-814.
Hazan R, Sat B, Engelberg-Kulka H. 2004. Escherichia coli mazEF-mediated cell death is triggered by various stressful conditions. *J Bacteriol* **186**(11): 3663-3669.
Hirashima A, Kaji A. 1973. Role of elongation factor G and a protein factor on the release of ribosomes from messenger ribonucleic acid. *J Biol Chem* **248**(21): 7580-7587.
Huang C, Mandava CS, Sanyal S. 2010. The ribosomal stalk plays a key role in IF2-mediated association of the ribosomal subunits. *J Mol Biol* **399**(1): 145-153.
Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* **256**(1346): 119-124.
Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L et al. 2009. InterPro: the integrative protein signature database. *Nucleic Acids Res* **37**(Database issue): D211-215.

Inagaki Y, Doolittle WF, Baldauf SL, Roger AJ. 2002. Lateral transfer of an EF-1alpha gene: origin and evolution of the large subunit of ATP sulfurylase in eubacteria. *Curr Biol* **12**(9): 772-776.
Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* **11**(2): 97-108.
Ito K, Ebihara K, Uno M, Nakamura Y. 1996. Conserved motifs in prokaryotic and eukaryotic polypeptide release factors: tRNA-protein mimicry hypothesis. *P Natl Acad Sci USA* **93**(11): 5443-5448.
Ito K, Fujiwara T, Toyoda T, Nakamura Y. 2002. Elongation factor G participates in ribosome disassembly by interacting with ribosome recycling factor at their tRNA-mimicry domains. *Mol Cell* **9**(6): 1263-1272.
Jurnak F. 1985. Structure of the Gdp Domain of Ef-Tu and Location of the Amino-Acids Homologous to Ras Oncogene Proteins. *Science* **230**(4721): 32-36.
Karimi R, Pavlov MY, Buckingham RH, Ehrenberg M. 1999. Novel roles for classical factors at the interface between translation termination and initiation. *Molecular Cell* **3**(5): 601-609.
Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**(2): 511-518.
Kimelman A, Levy A, Sberro H, Kidron S, Leavitt A, Amitai G, Yoder-Himes DR, Wurtzel O, Zhu Y, Rubin EM et al. 2012. A vast collection of microbial genes that are toxic to bacteria. *Genome research* **22**(4): 802-809.
Knudsen B, Miyamoto MM. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc Natl Acad Sci U S A* **98**(25): 14512-14517.
Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. *Genome Biol* **3**(2): RESEARCH0008.
Koonin EV, Wolf YI, Aravind L. 2000. Protein fold recognition using sequence profiles and its application in structural genomics. *Adv Protein Chem* **54**: 245-275.
Kothe U, Wieden HJ, Mohr D, Rodnina MV. 2004. Interaction of helix D of elongation factor Tu with helices 4 and 5 of protein L7/12 on the ribosome. *J Mol Biol* **336**(5): 1011-1021.
Leipe DD, Wolf YI, Koonin EV, Aravind L. 2002. Classification and evolution of P-loop GTPases and related ATPases. *J Mol Biol* **317**(1): 41-72.
Lesk AM. 2008. *Introduction to Bioinformatics*.
Li W-H. 1997. *Molecular Evolution*. Sinauer Associated, Inc.,Publisher.
Lichtarge O, Bourne HR, Cohen FE. 1996. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* **257**(2): 342-358.
Liljas A, Ehrenberg M, Aqvist J. 2011. Comment on "The mechanism for activation of GTP hydrolysis on the ribosome". *Science* **333**(6038): 37; author reply 37.
Liu H, Chen C, Zhang H, Kaur J, Goldman YE, Cooperman BS. 2011. The conserved protein EF4 (LepA) modulates the elongation cycle of protein synthesis. *Proc Natl Acad Sci U S A* **108**(39): 16223-16228.
Lynch M, Katju V. 2004. The altered evolutionary trajectories of gene duplicates. *Trends Genet* **20**(11): 544-549.
Madabushi S, Gross AK, Philippi A, Meng EC, Wensel TG, Lichtarge O. 2004. Evolutionary trace of G protein-coupled receptors reveals clusters of

residues that determine global and class-specific functions. *J Biol Chem* **279**(9): 8126-8132.
Makarova KS, Wolf YI, Koonin EV. 2009. Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. *Biol Direct* **4**: 19.
Margus T, Remm M, Tenson T. 2007. Phylogenetic distribution of translational GTPases in bacteria. *BMC genomics* **8**: 15.
-. 2011. A computational study of elongation factor G (EFG) duplicated genes: diverged nature underlying the innovation on the same structural template. *PLoS One* **6**(8): e22789.
Milburn MV, Tong L, deVos AM, Brunger A, Yamaizumi Z, Nishimura S, Kim SH. 1990. Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic ras proteins. *Science* **247**(4945): 939-945.
Mohr D, Wintermeyer W, Rodnina MV. 2002. GTPase activation of elongation factors Tu and G on the ribosome. *Biochemistry* **41**(41): 12520-12528.
Mougous JD, Lee DH, Hubbard SC, Schelle MW, Vocadlo DJ, Berger JM, Bertozzi CR. 2006. Molecular basis for G protein control of the prokaryotic ATP sulfurylase. *Mol Cell* **21**(1): 109-122.
Nakamura Y. 2001. Molecular mimicry between protein and tRNA. *Journal of Molecular Evolution* **53**(4-5): 282-289.
NCBI. 2012. Bacterial sequence database. Vol 2012. NCBI.
Nechifor R, Murataliev M, Wilson KS. 2007. Functional interactions between the G' subdomain of bacterial translation factor EF-G and ribosomal protein L7/L12. *J Biol Chem* **282**(51): 36998-37005.
Nei M. 1969. Gene duplication and nucleotide substitution in evolution. *Nature* **221**(5175): 40-42.
Nilsson J, Nissen P. 2005. Elongation factors on the ribosome. *Curr Opin Struct Biol* **15**(3): 349-354.
Nissen P, Kjeldgaard M, Thirup S, Polekhina G, Reshetnikova L, Clark BF, Nyborg J. 1995. Crystal structure of the ternary complex of Phe-tRNAPhe, EF-Tu, and a GTP analog. *Science* **270**(5241): 1464-1472.
Nocek B, Mulligan R, Duggan E, Clancy S, Joachimiak A. 2008. The C-terminal part of BipA protein from Vibrio parahaemolyticus RIMD 2210633. Protein Data Bank.
Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**(1): 205-217.
Nyborg J, Nissen P, Kjeldgaard M, Thirup S, Polekhina G, Clark BF, Reshetnikova L. 1997. Macromolecular mimicry in protein biosynthesis. *Fold Des* **2**(3): S7-11.
Ohno S. 1970. *Evolution by Gene Duplication*. Springer-Verlag, Berlin,
Heidelberg,
New York.
Owens RM, Pritchard G, Skipp P, Hodey M, Connell SR, Nierhaus KH, O'Connor CD. 2004. A dedicated translation factor controls the synthesis of the global regulator Fis. *Embo J* **23**(16): 3375-3385.
Pandey DP, Gerdes K. 2005. Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Res* **33**(3): 966-976.

Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**(6945): 194-197.
Pech M, Karim Z, Yamamoto H, Kitakawa M, Qin Y, Nierhaus KH. 2011. Elongation factor 4 (EF4/LepA) accelerates protein synthesis at increased Mg2+ concentrations. *Proc Natl Acad Sci U S A* **108**(8): 3199-3203.
Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J et al. 2011. The Pfam protein families database. *Nucleic Acids Res*.
Qin Y, Polacek N, Vesper O, Staub E, Einfeldt E, Wilson DN, Nierhaus KH. 2006. The highly conserved LepA is a ribosomal elongation factor that back-translocates the ribosome. *Cell* **127**(4): 721-733.
Ramakrishnan V. 2002. Ribosome structure and the mechanism of translation. *Cell* **108**(4): 557-572.
Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol* **5**: 28.
Ratje AH, Loerke J, Mikolajka A, Brunner M, Hildebrand PW, Starosta AL, Donhofer A, Connell SR, Fucini P, Mielke T et al. 2010. Head swivel on the ribosome facilitates translocation by means of intra-subunit tRNA hybrid sites. *Nature* **468**(7324): 713-716.
Roberts MC. 2005. Update on acquired tetracycline resistance genes. *FEMS Microbiol Lett* **245**(2): 195-203.
Rodnina MV, Pape T, Fricke R, Wintermeyer W. 1995. Elongation factor Tu, a GTPase triggered by codon recognition on the ribosome: mechanism and GTP consumption. *Biochem Cell Biol* **73**(11-12): 1221-1227.
Romero H, Zhang Y, Gladyshev VN, Salinas G. 2005. Evolution of selenium utilization traits. *Genome Biol* **6**(8): R66.
Savelsbergh A, Mohr D, Kothe U, Wintermeyer W, Rodnina MV. 2005. Control of phosphate release from elongation factor G by ribosomal protein L7/12. *Embo J* **24**(24): 4316-4323.
Savelsbergh A, Mohr D, Wilden B, Wintermeyer W, Rodnina MV. 2000. Stimulation of the GTPase activity of translation elongation factor G by ribosomal protein L7/12. *J Biol Chem* **275**(2): 890-894.
Schmeing TM, Ramakrishnan V. 2009. What recent ribosome structures have revealed about the mechanism of translation. *Nature* **461**(7268): 1234-1242.
Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**(20): 6097-6100.
Seshadri A, Samhita L, Gaur R, Malshetty V, Varshney U. 2009. Analysis of the fusA2 locus encoding EFG2 in Mycobacterium smegmatis. *Tuberculosis (Edinb)* **89**(6): 453-464.
Sevin EW, Barloy-Hubler F. 2007. RASTA-Bacteria: a web-based tool for identifying toxin-antitoxin loci in prokaryotes. *Genome Biol* **8**(8): R155.
Shannon CE. 1948. A mathematical theory of communication. *Bell System Technical Journal* **27**: 379-423
623-656.
Shao Y, Harrison EM, Bi D, Tai C, He X, Ou HY, Rajakumar K, Deng Z. 2011. TADB: a web-based resource for Type 2 toxin-antitoxin loci in bacteria and archaea. *Nucleic Acids Res* **39**(Database issue): D606-611.

Shoji S, Janssen BD, Hayes CS, Fredrick K. 2010. Translation factor LepA contributes to tellurite resistance in Escherichia coli but plays no apparent role in the fidelity of protein synthesis. *Biochimie* **92**(2): 157-163.

Sohmen D, Harms JM, Schlunzen F, Wilson DN. 2009. Enhanced SnapShot: Antibiotic inhibition of protein synthesis II. *Cell* **139**(1): 212-212 e211.

Soler N, Fourmy D, Yoshizawa S. 2007. Structural insight into a molecular switch in tandem winged-helix motifs from elongation factor SelB. *Journal of Molecular Biology* **370**(4): 728-741.

Sonnhammer EL, Eddy SR, Durbin R. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**(3): 405-420.

Sprang SR. 1997. G protein mechanisms: insights from structural analysis. *Annu Rev Biochem* **66**: 639-678.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)* **22**(21): 2688-2690.

Stephens SG. 1951. Possible significances of duplication in evolution. *Adv Genet* **4**: 247-265.

Suematsu T, Yokobori SI, Morita H, Yoshinari S, Ueda T, Kita K, Takeuchi N, Watanabe YI. 2010. A bacterial elongation factor G homolog exclusively functions in ribosome recycling in the spirochaete Borrelia burgdorferi. *Mol Microbiol*.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**(Web Server issue): W609-612.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* **28**(10): 2731-2739.

Tan QA, Awano N, Inouye M. 2011. YeeV is an Escherichia coli toxin that inhibits cell division by targeting the cytoskeleton proteins, FtsZ and MreB. *Molecular Microbiology* **79**(1): 109-118.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**(22): 4673-4680.

Van de Peer Y. 2004. Computational approaches to unveiling ancient genome duplications. *Nat Rev Genet* **5**(10): 752-763.

Vetter IR, Wittinghofer A. 2001. The guanine nucleotide-binding switch in three dimensions. *Science* **294**(5545): 1299-1304.

Voorhees RM, Schmeing TM, Kelley AC, Ramakrishnan V. 2010. The mechanism for activation of GTP hydrolysis on the ribosome. *Science* **330**(6005): 835-838.

Walker JE, Saraste M, Runswick MJ, Gay NJ. 1982. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *Embo J* **1**(8): 945-951.

Wang Y, Jiang Y, Meyering-Voss M, Sprinzl M, Sigler PB. 1997. Crystal structure of the EF-Tu.EF-Ts complex from Thermus thermophilus. *Nat Struct Biol* **4**(8): 650-656.

Wojtowicz D, Tiuryn J. 2007. Evolution of gene families based on gene duplication, loss, accumulated change, and innovation. *J Comput Biol* **14**(4): 479-495.

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**(7276): 1056-1060.

Yamaguchi Y, Inouye M. 2011. Regulation of growth and death in Escherichia coli by toxin-antitoxin systems. *Nat Rev Microbiol* **9**(11): 779-790.

Zaher HS, Green R. 2011. A primary role for release factor 3 in quality control during translation elongation in Escherichia coli. *Cell* **147**(2): 396-408.

Zavialov AV, Buckingham RH, Ehrenberg M. 2001. A posttermination ribosomal complex is the guanine nucleotide exchange factor for peptide release factor RF3. *Cell* **107**(1): 115-124.

Zhang Y, Romero H, Salinas G, Gladyshev VN. 2006. Dynamic evolution of selenocysteine utilization in bacteria: a balance between selenoprotein loss and evolution of selenocysteine from redox active cysteine residues. *Genome Biol* **7**(10): R94.

SUMMARY IN ESTONIAN

Valgud on raku ehituskivideks ja eluks vajalike reaktsioonide katalüüsijateks. Bioinformaatika on meid varustanud võimsate järjestuste analüüsi vahenditega. Järjestuse sarnasuse alusel grupeeruvad valgud perekondadeks. Valguperekonna moodustavad homoloogsed järjestused ehk siis järjestused, mis pärinevad samast eellasjärjestusest. Tihti omavad samasse perekonda kuuluvad valgud ka sama või üksteisele lähedast funktsiooni. Meie teadmised valkude funktsioonidest pärinevad üksikutelt mudelorganismidelt. Tihti huvitab teadlasi kui universaalne või spetsiifiline on üks või teine kirjeldatud funktsioon. Kuidas ja millal evolutsiooni käigus tekib olemasolevast materjalist uute omadustega (uue funktsiooniga) valk läbi geeniduplikatsiooni? Kui tihti on sellised sündmused evolutsioonilises ajaskaalas aset leidud?

Oma töös olen ma analüüsinud bakterite translatsioonilisi GTPaase (trGTPaas) ja mqsR/ygiT toksiin-antitoksiin (TA) süsteemi valke. Ühiseks nimetajaks mõlemale on valgusünteesi aparaat – mõlemad on seotud ribosoomiga ja sealtkaudu raku võimega sõltuvalt vajadusele toota valke.

Küsimused, mida selles kontekstis on küsitud, saab laias laastus jagada kaheks: a) valguperekonna esindatusega seotud ja b) valguperekonna evolutsiooni ja funktsionaalse innovatsiooniga seotud. Translatsiooniliste GTPaaside puhul bakterites saame rääkida üheksast erinevast perekonnast – üheksast erinevast funktsioonide komplektist. Täisgenoomidele põhinev analüüs näitas, et üheksast trGTPaaside perekonnast on bakterites konserveerunud neli: IF2, EF-Tu, EFG ja LepA(EF4). Vaatamata sellele, et RF3'e on omistatud klassikalise valgusünteesi mudeli valguses kanooniline roll translatsiooni lõpetamisel, puudus RF3 geen ligikaudu 40% analüüsitud bakteri genoomides. Samas aga ebaselge funktsiooniga LepA osutus bakterite spetsiifiliseks trGTPaasiks.

Eelnev analüüs tõi ka välja EFG paraloogide laia esinemise – paljud bakterigenoomid sisaldasid 2-3 üksteisest küllaltki erinevat (divergeerunud) EFG geeni. Lähem analüüs tõi välja, et kogu varieeruvuse EFG perekonnas võib jagada neljaks alamperekonnaks: EFG I, spdEFG1, spdEFG2 ja EFG II. Eksperimentaalselt on hästi iseloomustatud EFG I. Uuritud on ka spdEFG'sid ja leitud, et esimene neist omab translokaasi aktiivsust translatsioonil ja teine osaleb ribosoomide retsükleerimisel. Laialt levinud EFG II alamperekond on aga halvasti uuritud.

Fülogeneetiline analüüs võimaldab püstitada hüpoteesi nelja EFG alamperekonna iidsest päritolust, st. nad on tekkinud ajalises skaalas enne (või samaaegselt) eukarüootse rakuvormi lahknemist arhedest ja bakteritest. Funktsionaalse innovatsiooni kandjaks EFG II valgus võib pidada eelkõige 12 positsiooni, mis on spetsiifiliselt konserveerunud just EFG II alamperekonnal. EFG II'e iseloomulikus kõrge divergentsuse taustal tõusevad need positsioonid esile GTPaasi domäänis, domäänis II ja neljandas domäänis. Konserveerunud muutused GTPaasi domäänis, millest osad on GTP'd siduvas G1 motiivis, võimaldavad teha järeldusi muutunud GTP sidumise ja hüdrolüüsi tingimuste kohta. Suurenenud laeng neljanda domääni lingu otsas, mis *E. coli* EFG'l siseneb A-saiti, võimaldab spekuleerida muutuse üle translokatsiooni keskkonnas. Konserveerunud muutused domään II piirkonnas viitavad muutunud interaktsioonile ribosoomi, domään I ja domään III vahel.

EFG II alamperekonna fülogeneetiline ja järjestuste analüüs näitab selgelt hõimkonna/klassi spetsiifiliste alam-alamgruppide olemasolu. Need alam-alamgrupid erinevad teineteisest G2 motiivi konserveeruvuse ja insertsioonide/deletsioonide mustri alusel. See teine tase kirjeldab EFG II kui hõimkonna/klassi spetsiifilist faktorit.

Mis on EFG II roll tegelikult ja kuidas ning millistes tingimustes ta komplementeerib EFG I, ootab alles vastuseid. Antud töö on loonud raamistiku tulevaste eksperimentide tarvis.

CURRICULUM VITAE

Tõnu Margus

Date and place of birth:	9th October, 1962, Tartu
Address:	Department of Bioinformatics, Institute of Molecular and Cell Biology, University of Tartu 23 Riia str, 51010, Tartu, Estonia
Phone:	+372 7374036
E-mail:	tonu.margus@ut.ee

Education and professional employment

1981 – 1986	University of Tartu, Faculty of Biology and Geography (specialty of genetics and teacher of biology and chemistry)
2007 – …	PhD studies in bioinformatics/gene-technology
1986 – 1997	engineer at Institute of Chemical and Biological Physics
1997 – 2005	bioinformatics consultant at Estonian Biocentre
2005 – 2007	computer specialist at IMCB and bioinformatics specialist at University of Tartu, Institute of Technology (UTIT)
2007 – …	programmer at Estonian Biocentre

Scientific work

My main research projects have been related to protein synthesis. During my PhD studies I focused on evolutionary and functional aspects of translational GTPases of bacteria. I have also participated in projects of human fertility and extracting evolutionary signal from sequences of mitochondria.

List of publications

1. **Margus, T.**; Remm, M.; Tenson, T. (2011). A Computational Study of Elongation Factor G (EFG) Duplicated Genes: Diverged Nature Underlying the Innovation on the Same Structural Template. PLoS ONE, 6, e22789
2. Kasari, V.; Kurg, K.; **Margus, T.**; Tenson, T.; Kaldalu, N. (2010). The Escherichia coli mqsR and ygiT genes encode a new toxin-antitoxin pair. Journal of Bacteriology, 192(11), 2908 - 2919.

3. Loogväli, E-L.; Kivisild, T.; **Margus, T**.; Villems, R. (2009). Explaining the imperfection of the molecular clock of hominid mitochondria. PLoS ONE, 4(12), e8260
4. Rull, K.; Nagirnaja, L.; Ulander, V.-M.; Kelgo, P.; **Margus, T**.; Kaare, M.; Aittomäki, K.; Laan, M. (2008). Chorionic Gonadotropin Beta gene variants are associated with recurrent miscarriage in two European populations. Journal of Clinical Endocrinology and Metabolism, 93(12), 4697 - 4706.
5. **Margus, T**.; Remm, M.; Tenson, T. (2007). Phylogenetic distribution of translational GTPases in bacteria. BMC Genomics, 8, 15
6. Hallast, P.; Nagirnaja, L.; **Margus, T**.; Laan, M. (2005). Segmental duplications and gene conversion: Human luteinizing hormone/chorionic gonadotropin beta gene cluster. Genome Research, 15(11), 1535 - 1546.
7. Lewicki, BTU.; **Margus, T**.; Remme, J.; Nierhaus, KH. (1993). Coupling of ribosomal-RNA transcription and ribosomal assembly invivo - formation of active ribosomal-subunits in escherichia-coli requires transcription of ribosomal-RNA genes by host RNA- polymerase which cannot be replaced by bacteriophage-t7 RNA- polyme. Journal of Molecular Biology, 231(3), 581 - 593.
8. Remme, J.; **Margus, T**.; Villems, R.; Nierhaus, KH. (1989). The 3rd ribosomal transfer RNA- binding site, the e-site, is occupied in native polysomes. European Journal of Biochemistry, 183(2), 281 - 284.
9. Marju K.; Jaanus R.; Simon YWH.; John D.; Egle T.; Igor LT.; Alexander PS.; Peep Mn.; Ilpo K.; Alexei VA et al. (2012). Complete mitochondrial genomes and a novel spatial genetic method reveal cryptic phylogeographical structure and migration patterns among brown bears in north-western Eurasia. *Journal of Biogeography*,
10. Rull K, Christiansen OB, Nagirnaja L, Steffensen R, **Margus T**, Laan M. 2013. A modest but significant effect of CGB5 gene promoter polymorphisms in modulating the risk of recurrent miscarriage. *Fertil Steril*.

ELULOOKIRJELDUS

Tõnu Margus

Sünni aeg ja koht:	9. oktoober, 1962, Tartu
Aadress:	Bioinformaatika õppetool, Molekulaar- ja Rakubioloogia Instituut, Tartu Ülikool Riia 23, 51010, Tartu, Eesti
Telefon:	+372 7374036
E-post:	tonu.margus@ut.ee

Haridus ja erialane teenistuskäik

1981 – 1986	Tartu Ülikool, bioloogia (eriala geneetika ja bioloogia ning keemia õpetaja)
2007 – …	õpingud doktorantuuris bioinformaatika/geenitehnoloogia erialal
1986 – 1997	insener, Keemilise ja Bioloogilise Füüsika Instituut
1997 – 2005	bioinformaatika konsultant, Eesti Biokeskus
2005 – 2007	arvuti- ja bioinformaatika spetsialist TÜMRI ja TÜTI
2007 – …	programmeerija, Eesti Biokeskus

Teadustegevus

Minu peamised teadushuvid on olnud seotud valgusünteesi uurimisega. Doktorantuuri ajal fokuseerusin translatsiooniliste GTPaaside evolutsiooniliste ja funktsionaalsete aspektide uurimisele ja valguperekonna iseloomustamisele. Minu teadustegevus on olnud seotud ka inimese viljakuse uurimise projektiga ja evolutsioonilise signaali tuvastamisega mitokondri järjestustest.

Publikatsioonid

1. **Margus, T.**; Remm, M.; Tenson, T. (2011). A Computational Study of Elongation Factor G (EFG) Duplicated Genes: Diverged Nature Underlying the Innovation on the Same Structural Template. PLoS ONE, 6, e22789
2. Kasari, V.; Kurg, K.; **Margus, T.**; Tenson, T.; Kaldalu, N. (2010). The Escherichia coli mqsR and ygiT genes encode a new toxin-antitoxin pair. Journal of Bacteriology, 192(11), 2908 - 2919.

3. Loogväli, E-L.; Kivisild, T.; **Margus, T**.; Villems, R. (2009). Explaining the imperfection of the molecular clock of hominid mitochondria. PLoS ONE, 4(12), e8260
4. Rull, K.; Nagirnaja, L.; Ulander, V.-M.; Kelgo, P.; **Margus, T**.; Kaare, M.; Aittomäki, K.; Laan, M. (2008). Chorionic Gonadotropin Beta gene variants are associated with recurrent miscarriage in two European populations. Journal of Clinical Endocrinology and Metabolism, 93(12), 4697 - 4706.
5. **Margus, T**.; Remm, M.; Tenson, T. (2007). Phylogenetic distribution of translational GTPases in bacteria. BMC Genomics, 8, 15
6. Hallast, P.; Nagirnaja, L.; **Margus, T**.; Laan, M. (2005). Segmental duplications and gene conversion: Human luteinizing hormone/chorionic gonadotropin beta gene cluster. Genome Research, 15(11), 1535 - 1546.
7. Lewicki, BTU.; **Margus, T**.; Remme, J.; Nierhaus, KH. (1993). Coupling of ribosomal-RNA transcription and ribosomal assembly invivo - formation of active ribosomal-subunits in escherichia-coli requires transcription of ribosomal-RNA genes by host RNA- polymerase which cannot be replaced by bacteriophage-t7 RNA- polyme. Journal of Molecular Biology, 231(3), 581 - 593.
8. Remme, J.; **Margus, T**.; Villems, R.; Nierhaus, KH. (1989). The 3rd ribosomal transfer RNA- binding site, the e-site, is occupied in native polysomes. European Journal of Biochemistry, 183(2), 281 - 284.
9. Marju K.; Jaanus R.; Simon YWH.; John D.; Egle T.; Igor LT.; Alexander PS.; Peep Mn.; Ilpo K.; Alexei VA et al. (2012). Complete mitochondrial genomes and a novel spatial genetic method reveal cryptic phylogeographical structure and migration patterns among brown bears in north-western Eurasia. *Journal of Biogeography*,
10. Rull K, Christiansen OB, Nagirnaja L, Steffensen R, **Margus T**, Laan M. 2013. A modest but significant effect of CGB5 gene promoter polymorphisms in modulating the risk of recurrent miscarriage. *Fertil Steril*.

ACKNOWLEDGEMENTS

I would like to thank our departments collective for creating friendly atmosphere for scientific research. I thank Tanel Tenson and Maido Remm for initiating this work, helping to guide through the jungle of problems and leaving me considerable amount of freedom. They were invaluable opponents and helped to shape my manuscripts to mature products. My special thanks are going to my wife Kärt for warm support throughout this way.

PUBLICATIONS