Chapter 4

DISCUSSION

4.1 Some comments to the optimisation procedures

For the *in vitro* protein synthesis reaction based on the utilization of a cell-free extract (often termed as S30), not only the initial cell strain and growth media, but also the method for isolation of a cell lysate is of great value.

In our days modern instruments provide an advantage to obtain cell lysate of high yields in cell rupture with minimal processing (one pass) and easy recovery for large (up to 1 litre) and small (14 ml with 12 ml recovery) volumes of cell suspensions. Our choice felt for the homogeniser "Microfluidizer M-110L" from Microfluidics, which is a fast and efficient machine, easy to handle and use. Fast rapture is important while producing cell lysate, for the longer is this lysate in a liquid state (if not used directly), the faster do degrade its main components, it looses the so necessary activity for protein synthesis due to fast RNAs degradation (tRNAs and rRNAs). Also it is important to work with cell lysate in a cold room, where temperature should not exceed +4°C – simply keep the suspension and later lysate on ice (water). Another remark is concerning the concentration of the cell lysate – it has to be high for better activity of the later. It means, that the density of ribosomes should be high for the best protein synthesis. Thorough suspending of cell pellets in 1 ml/g (wet weight) of Tico buffer is sufficient enough to obtain high A_{260} measurements of S30 lysate.

Concerning the cell strain, it was expected that strains deficient in major RNases would allow for obtaining of high protein yields, since mRNA molecules are the least stable ones of all RNAs. In the case of an *in vitro* coupled transcription-translation system the level of mRNA is relatively stable and, in fact, does not affect the protein yield (Figure 3.6.2-1, page 68). With respect to this, our selection was an *E. coli* BL21 strain, which is deficient in the protease genes *lon* and *ompT*. This strain was used for preparation of cell lysate and fractionation analyses. Besides this strain we selected different protease deficient strains from ATCC and literature reports. Also, a Rosetta™ strain from

Novagen, which is enriched with some tRNAs isoacceptors species was of great interest to us for reasons, discussed bellow.

We used a batch system to speed up the optimisation procedure using small volumes down to 10 µl. This kind of an *in vitro* coupled transcription-translation system generally utilizes high-energy phosphate compounds (like PEP) to regenerate the adenosine triphosphate (ATP), necessary to drive protein synthesis, as well as adequate substrates (particularly nucleoside triphosphates and amino acids). This system also requires buffering, pH and proper ionic conditions, as well as catalyst stability and avoidance of inhibitory by-products. Though this system attempts to mimic "*in vivo* near" conditions, it contains some unnatural components, such as pH buffers and polyethylene glycol (PEG), and is far away from physiological environment of the cytoplasm. We optimised the final concentrations of main components of the batch cell-free system for *in vitro* transcription-translation reactions. This was done in order to ease up calculation of numbers for stock solutions and monitoring of the system components activity. It is easy to tackle which of the main mixes or stocks had lost the activity for protein synthesis reaction due to one of the following reasons: presence of RNases, expiry of one of the "energy" sources for the protein synthesis, *e.g.* ATP, ionic conditions changed, *e.g.* Mg^{2+} concentration, as well as T7 RNA polymerase inactivation.

Recent literature data about a promising *in vitro* translation system, which is designed to mimic the cytoplasm (thus named Cytomim), reports that most unnatural components like PEG were removed, and ionic composition of the reaction was altered to more closely replicate that of the cytoplasm (Jewett and Swartz, 2004a).

An essential point of the coupled transcription/translation system was to establish reliable quality criteria for the synthesis of the reporter product green fluorescent protein (GFP). With use of this protein we established a reliable method for judgement of GFP expression, based on the electrophoretic analysis in denaturing and native PAA gel conditions.

We measure relative band intensities of GFPcyc3 in native and SDS (denaturing) protein gels from one and the same reaction vial and correlate to those bands, which contain known amounts of GFP. Thus, we identify total yield and native fraction in it, which is for GFP in the range of 40-60%.

In the case we would like to learn more on the fragmentation level of a given protein we performed a reaction with [^{35}S]-Met incorporation. Operating with exact number of Met-amino acids in a single protein chain we can estimate the relative number of protein molecules synthesised per single ribosome (assuming that only 30% are active), the concentration of which is known from A_{260} measurements. According to our estimation about 7 molecules of GFPcyc3 were synthesised by one ribosome in the reaction mix during 120 min incubation at 37°C.

Table 4.2-1 The TCA precipitation result for GFPcyc3 synthesis

Sample	cpm	Average	Minus background	Met incorp. in GFP prot. synth.	Amount of GFP, pmol	GFP made by one ribosome in 120 min	~30% of 70S participate in GFP synth.
1. DNA -	1271						
2. DNA -	1661	1466	0				
3. DNA +	3811						
4. DNA +	4952	4382	2916	295.5	49.3	2	**6.6**

The specific activity of [^{35}S]-Met was about 10 cpm/pmol. GFP contains 6 methionine residues.

4.2 Synchronising the reactions of transcription and translation

One of the problems in cell-free systems is the uncoupling of the naturally coupled processes of transcription and translation. This happens due to utilization of the bacteriophage T7 RNA polymerase (RNAP) instead of *E. coli* RNAP T7 RNAP is 5-6 times faster that *E. coli* (Iost *et al.*, 1992), and this affects negatively not only protein synthesis initiation and elongation steps, but also *in vivo* assembly of the *E. coli* ribosomes, when T7 RNAP transcripts of rRNA are used. The negative effect was the observation that only a tiny fraction,

within a few percent of the transcribed product (either mRNA or rRNA), was used for translation and ribosome assembly, respectively (Iost and Dreyfus, 1995; Lewicki *et al.*, 1993). In *E. coli* cells processes of transcription and translation are tightly coupled. While mRNA is synthesised *E. coli* ribosomes initiate translation on the nascent chain of mRNA: the *E. coli* RNAP proceeds with a speed of ~60 nucleotides per second, and ribosomes proceed with a speed of ~20 amino acids per second (Bremer and Dennis, 1996). It follows that the first ribosome pursues immediately the transcription, leaving no room for a significant gap in-between. Therefore, the nascent mRNA chain cannot form a secondary structure that could hinder or block elongation (Iost and Dreyfus, 1995; Iost *et al.*, 1992). On the other hand, ribosomes protect mRNA from endonucleases that usually initiate the process of degradation from the 5' into 3' end direction. We tested two approaches, (i) applying slow mutants of T7 polymerases, and (ii) lowering the incubation temperature. The latter step seemed to be a promising trial, since lowering the growth temperature to 25°C improved dramatically from 15 to 60% the rRNA fraction used for ribosomal assembly, which indicated that the rate of T7 RNAP goes down faster than the assembly rate (Lewicki *et al.*, 1993).

Mark Dreyfus (Ecole Normale Superieure – CNRS, Paris) kindly supplied us with different T7 RNAP mutants, which were either slower than the wild type (WT) T7 RNAP, thus reaching the rate of *E. coli* RNAP, or were more processive than the WT T7 RNAP (Bonner *et al.*, 1994; Makarova *et al.*, 1995). From these mutants, according to our results based on the judgement of GFP expression, we selected a double-mutant I810N/P266L, which has a reduced rate but an improved processivity. Translation from the transcripts produced by this mutant resulted in almost 100% active GFP production (although the yield was reduced to 10%, see Figure 3.5.1-2C, page 61), indicating that not only translation, but also folding was affected positively in this case. This finding and literature reports were furthermore supported by results described in the following chapter.

In another set of experiments we also achieved almost 100% active GFP output just by reducing temperature of incubation from 37°C and 30°C to 25°C and 20°C, though the overall protein production was reduced two to three times (Figure 3.5.2-1A, page 63). Thus, in further attempts to optimise cell-free protein synthesis this approach was combined with others in order to increase the output of fully synthesised and fully active proteins of interest, as will be discussed below.

4.3 Trials to increase the yield of expressed proteins

The turnover of mRNA is defined in terms of the mRNA half-life, which is the time it takes for 50% of the mRNA molecules in the cell to be degraded. In general, half-lives of mRNA are shortest in species with the shortest replication cycles, also compared to the half-life of ribosomal RNAs (rRNAs) within one organism. The half-life of a bacterial mRNA measures on average one (most labile mRNAs) to seven and a half minutes (ribosomal proteins mRNA, Mohanty and Kushner, 1999), with two to three minutes in average for most of the mRNAs in *E. coli* (Selinger *et al.*, 2003). It is essential that rapidly replicating organisms adapt quickly to changes in their environment by stabilising or destabilising certain mRNAs. Ribonuclease III (RNase III) is one of the mRNA-degrading enzymes in *E. coli*. It attacks duplex regions in mRNAs and in rRNA precursors (Ross, 2001).

Liiv, *et al.* (1996) found earlier that the major stability determinant of rRNA is the presence of a stable helix bracketing mature rRNA. The stability determinant of pre-23S rRNA was used to increase the stability of mRNA (Liiv *et al.*, 1996). In a trial to increase the stability of mRNA molecules by the stability determinants of the pre-23S rRNA, we used a construct that was kindly provided by Jaanus Remme. These stability determinants are flanking the GFP gene and upon transcription result in the pseudo-circularisation of an mRNA molecule, thus, in theory, prolonging the half-life of it. On the contrary, another construct with an insert of 20 nucleotides that would rupture the

complementarity of the flanking regions was used as a negative control for mRNA stability. However, the level of GFP expression from this two constructs was same and in comparison to our standard GFP expression construct these levels were two times less, indicating again, that mRNA stability is not a limiting factor in cell-free protein synthesis system.

These conclusions were supported by the results of an assay addressed to picture the fate of the transcribed mRNA (Figure 3.6.2-1, page 68). We observed a significant decrease of mRNA in the middle of incubation time, and after three to five hours the mRNA levels went up again, showing that the parallel reduction of protein synthesis was not caused by a depletion of the energy supply (NTPs).

Another possibility of an impaired synthesis could be shortage of amino acids during the incubation period. According to Kim and Swartz, shortage of the building block for protein synthesis could be observed for arginine, cysteine and tryptophan in an *in vitro* system even in the absence of protein synthesis (Kim and Swartz, 2000). Later on, Jewett and Swartz identified eight out of twenty amino acids, concentrations of which changed dramatically from the starting 2 mM. Aspartic acid increased to 20 mM during the reaction. Alanine increased to 4 mM, and then decreased to 1.5 mM. Glutamate, which was initially 160 mM, decreased as well, though the corresponding value was not quantified precisely. The following amino acids were depleted during the first hours of reaction: cysteine, serine, threonine, glutamine and asparagine (Jewett and Swartz, 2004b).

Our results indicated that energy resources of the system are sufficient for mRNA synthesis. A close study of the amounts of GFPcyc3 mRNA indicates a slight reduction of its concentration in the middle of reaction. This takes place exactly at the same time, when the process of protein synthesis enters its maximal synthesis rate. As soon as protein synthesis reaches saturation phase, synthesis of mRNA bursts again – an indication of the presence of sufficient amounts of NTPs (Figure 3.6.2-1, page 68). All together, these pieces fit well

into a puzzle that shortage in amino acids rather than in NTPs is a limiting factor for increase of protein synthesis. Indeed, when a mixture of twenty amino acids was added after the fifth hour of incubation (saturation is reached normally at this time), and reaction was incubated up to 40 hours, the total yield of GFPcyc3 increased twice. But still the level of active fraction was too low (<30%). This led to combination of two approaches with a very satisfying result, namely a combination of a decrease of the incubation temperature to 20°C and the addition of amino acids that allowed the gain of 4 mg/ml of 93% (~100%) of active GFP (Figure 3.6.3-2, page 70).

4.4 Trials and considerations to increase the expression of eukaryotic proteins in the bacterial *E. coli* system

The term codon usage means a "snap shot" of codon abundance within the mRNAs present at a given state. The codon abundance is roughly related to that of the corresponding tRNAs. Highly expressed genes are, for example, those, which products are directly used in translation. Lowly expressed genes are, for example, those for RNases, proteases, and down regulated genes that may change due to temperature or amino acid variation.

Data published on tRNA species abundance in different organisms were the basis for this research, as well as data from a single organism at different growth rates (*E. coli*; Dong *et al.*, 1996). In *E. coli* there are 86 genes coding for 46 different tRNA species. A tRNA species is defined as a tRNA with a unique anticodon, and their gene copy number ranges from four to one. The copy number of tRNA genes for codons frequently appearing in the reading frames of highly expressed genes is quite often four. Thus, gene dosage is involved in the regulation of tRNA concentrations (Ueda T. and K., 2001).

Dong *et al.* showed earlier that in *E. coli* the co-variation of tRNA abundance and codon usage is also growth rate-dependent (Dong *et al.*, 1996). They found a strikingly reduced codon usage in highly expressed genes *versus* lowly expressed genes, which were switched on under bad nutrient conditions. It

follows that many more codons are used under slow growth conditions. Accordingly, a corresponding biased distribution of tRNA abundance at various growth rates were observed, which could be roughly correlated with the values of codon frequencies in the mRNA pools calculated for bacteria growing at different rates. Another observation they made was a correlation of tRNA concentrations with the gene dose. The mechanisms that support the growth-rate dependent co-variation of tRNA abundance and codon usage are not known. They identified the fraction of tRNAs out of total tRNA (%) for each isoacceptor species in *E. coli* grown at low growth rates.

We took as a control the known *E. coli* codon usage in highly and lowly expressed genes (Figure 4.4-1; Bulmer, 1988): In case of some codons there is a sharp difference for their abundance within highly expressed genes, while those that show up to be rare in highly expressed genes turned out to be often in lowly expressed genes. For example, in case of codons coding for aspartic acid, GAC is found frequent and GAU rarely in highly expressed genes, whereas just the opposite is true in lowly expressed genes. The same is valid for codons coding for Tyr, His, Asn and Phe. In cases of some amino acids, for example, arginine, glycine, leucine, serine, proline and threonine, some codons almost do not appear in highly expressed genes (violet line, least point marked by red square). This is a clear codon bias correlated to gene expression in the *E. coli* genome.

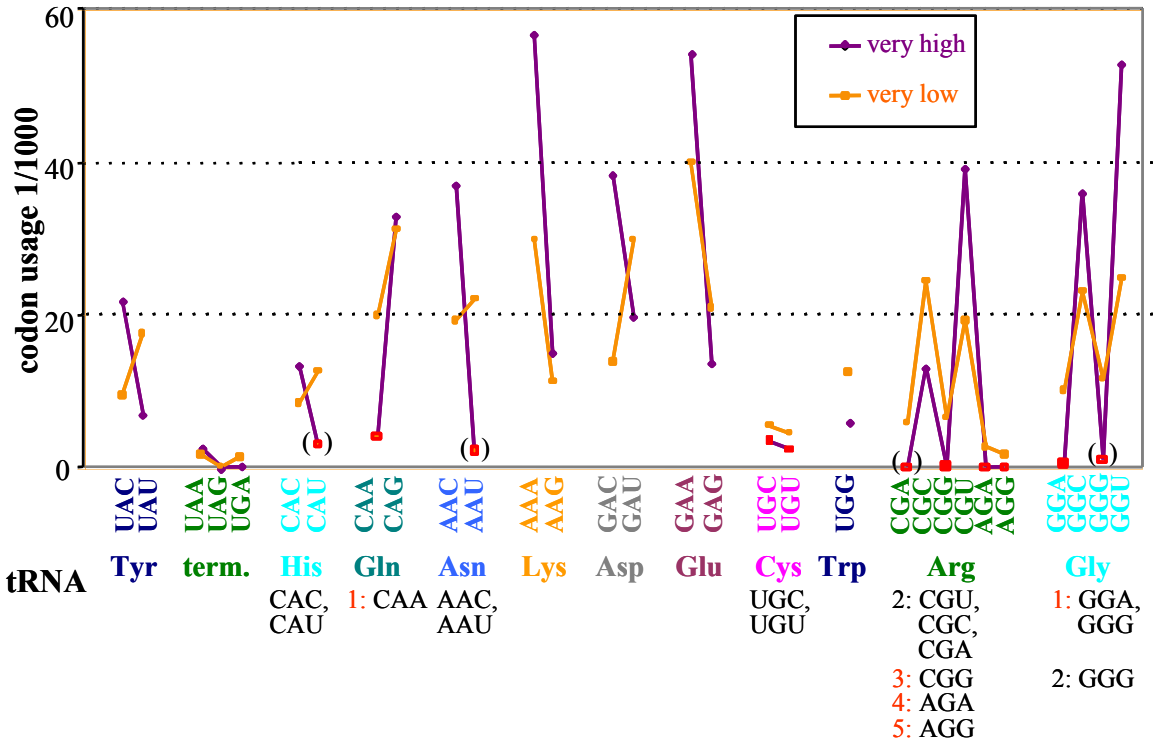

Figure 4.4-1 Distribution of different codons within E. coli genome. The genome was subdivided into two groups of genes corresponding to relatively **high** expressed (**violet**), and relatively **low** expressed (**orange**). Numbers stand for the tRNA isoacceptor species.

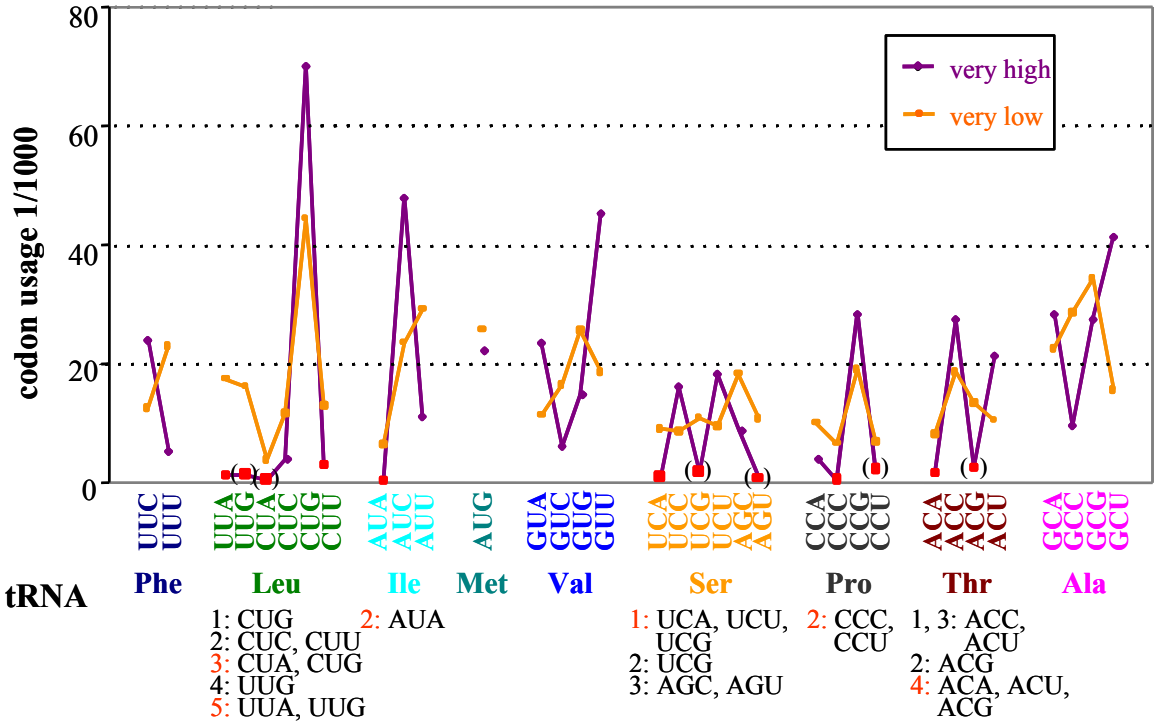

Figure 4.4-1 Continued…

When we performed a similar investigation in human genes, assuming that the ribosomal proteins and glycolysis enzymes are highly expressed genes, while genes for proteases and RNases are lowly expressed, none of such dramatic differences as seen in *E. coli* codons were detected. Also the total codon usage in the human genome does not deviate significantly from that in lowly or highly expressed genes. The probable reason is that in higher eukaryotes there is no need to give an overall preference to a certain codon within one or another group of genes due to homeostasis of the human cells, which are grouped in tissues and organs for better maintenance of their survival. In contrast, in *E. coli*, as in free-living bacteria, there is always a need to adapt the metabolism as fast as possible to an alteration in the environment.

The codon usage in human genes is shown in (Figure 4.4-2) and is, as explained, relatively stable and does not deviate from one group of genes to another, as well as for the total proteins group. This can be judged from the parallelism of most of the lines for both highly and lowly expressed genes, which directs to the fact that most codon/tRNA species for the same amino acid, for example, glutamine, lysine, leucine, valine, serine, threonine and alanine are represented equally within each group of genes, *i.e.*, if one codon is underrepresented in one group of genes (*e.g.,* ribosomal proteins), then it is underrepresented in all other groups of genes as well, with only slight difference in abundance, which is within the error bar.

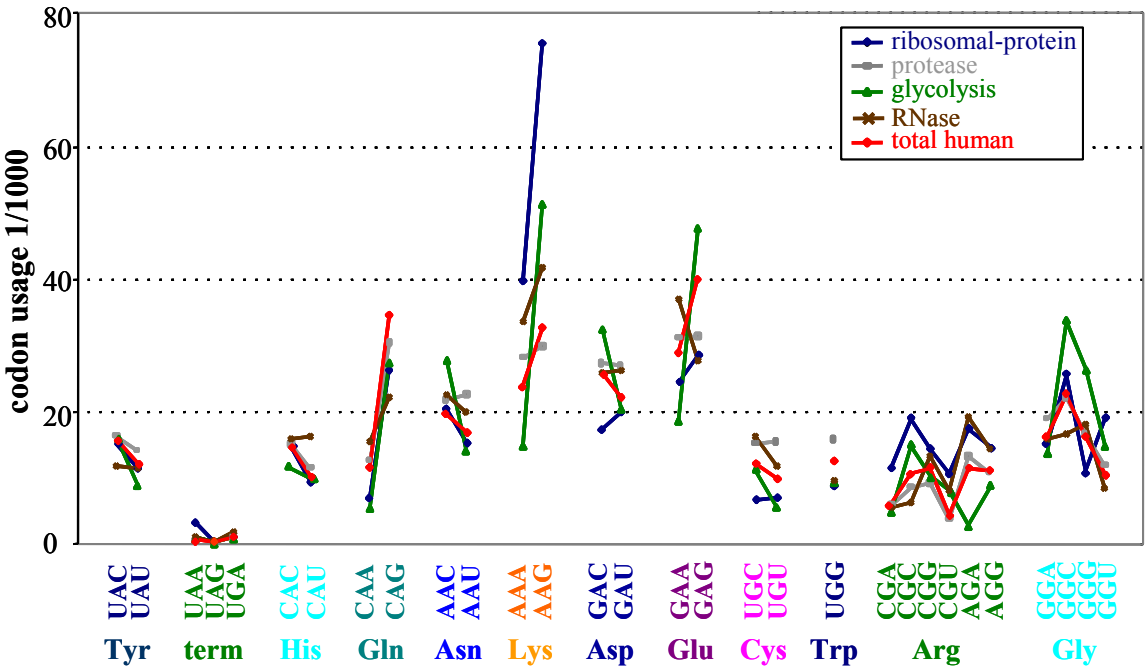

Figure 4.4-2 Distribution of different codons within human genome. Various gene families were selected from human genome and defined as highly expressed and lowly expressed ones. Codon distribution in overall human genes was also analysed.

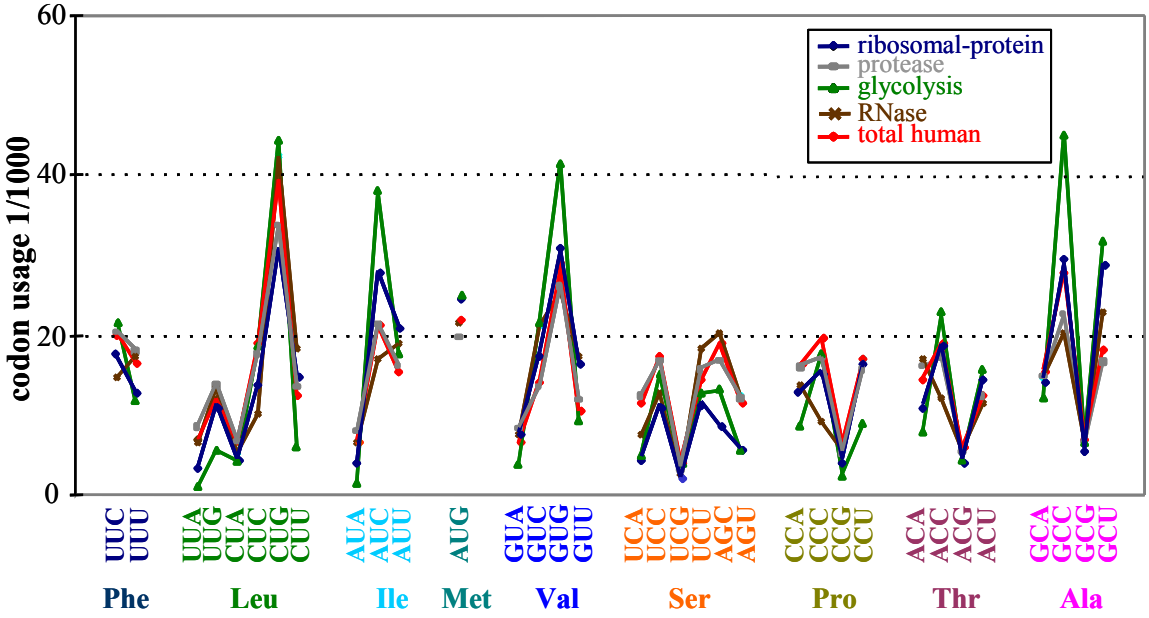

Figure 4.4-2 Continued…

It becomes clear from these results, that during expression of eukaryotic, particularly human genes in *E. coli* cells or cell-free extracts, there will be shortage of defined classes of tRNA species, which are underrepresented in *E. coli*.

We compared data from both analyses in order to identify those codons that are underrepresented in *E. coli*, but often used in the human genome, and we found out that there are at least 15 codons, which are of great importance for human protein synthesis and are almost not appearing in bacteria. Some of these codons were belonging to the same amino acid (*e.g.,* arginine, glycine, leucine, serine, proline and threonine group of codons; assigned as red squares in brackets on Figure 4.4-1). Since some tRNA species can recognize two to three different codons coding for the same amino acid, we determined the optimal set of tRNAs that should be enriched in the *E. coli* bulk tRNA to allow an easy translation for all human genes concerning the codon usage. See, for example, the Gly codons. The codons GGA and GGG are rare in *E. coli* (Figure 4.4-1) but relatively frequent in human genes (Figure 4.4-2). Since a single *E. coli* tRNA decodes both GGA and GGG, we only need to substitute the *E. coli* $tRNA^{bulk}$ with this $tRNA^{Gly}$ to decode both GGA and GGG in a human gene if expressed in an *E. coli* system (see orange number "1" below "Gly" in Figure 4.4-1). In this way we identified 11 tRNA species that should be enriched in *E. coli* $tRNA^{bulk}$ for a proper translation of a human gene in an *E. coli* system.

It is clear therefore that the tRNA concentrations in a bacterial system have to be adjusted to the eukaryotic codon usage.

Table 4.4-1 E. coli tRNAs and the codon recognition pattern (Dong et al., 1996).

tRNA	Anticodon (5'–3')	Codon recognition (5'–3')	№ of molecules per cell	Fraction of tRNA out of total tRNA (%)
Ala1B	UGC	GCU, GCA, GCG	3250(±223)	5.04
Ala2	GGC	GCC	617(±64)	0.95
Arg2	ACG	CGU, CGC, CGA	4752(±440)	7.37
Arg3	CCG	CGG	639(±63)	0.99
Arg4	UCU	AGA	867(±160)	1.34
Arg5	CCU	AGG	420(±69)	0.65
Asn	GUU	AAC, AAU	1193(±127)	1.85
Asp1	GUC	GAC, GAU	2396(±346)	3.72
Cys	GCA	UGC, UGU	1587(±126)	2.46
Gln1	UUG	CAA	764(±66)	1.18
Gln2	CUG	CAG	881(±94)	1.36
Glu2	UUC	GAA, GAG	4717(±411)	7.32
Gly1[a]	CCC	GGG		
Gly2	UCC	GGA, GGG	2137(±320)	3.31
Gly3	GCC	GGC, GGU	4359(±378)	6.76
His	GUG	CAC, CAU	639(±95)	0.99
Ile1	GAU	AUC, AUU	3474(±94)	5.39
Ile2[a]	CAU	AUA		
Leu1	CAG	CUG	4470(±346)	6.94
Leu2	GAG	CUC, CUU	943(±97)	1.46
Leu3	UAG	CUA, CUG	666(±94)	1.03
Leu4	CAA	UUG	1913(±190)	2.97
Leu5	UAA	UUA, UUG	1031(±117)	1.60
Lys	UUU	AAA, AAG	1924(±185)	2.97
Met f1	CAU	AUG	1211(±191)	1.88
Met f2	CAU	AUG	715(±107)	1.11
Met m	CAU	AUG	706(±96)	1.09
Phe	GAA	UUC, UUU	1037(±162)	1.60
Pro1	CGG	CCG	900(±150)	1.38
Pro2	GGG	CCC, CCU	720(±125)	1.11
Pro3	UGG	CCA, CCU, CCG	581(±95)	0.90
Sec	UCA	UGA	219(±73)	0.34
Ser1	UGA	UCA, UCU, UCG	1296(±94)	2.01
Ser2	CGA	UCG	344(±62)	0.53
Ser3	GCU	AGC, AGU	1408(±126)	2.18
Ser5	GGA	UCC, UCU	764(±127)	1.18
Thr1	GGU	ACC, ACU	104(±34)	0.16
Thr2	CGU	ACG	541(±94)	0.84
Thr3	GGU	ACC, ACU	1095(±62)	1.70
Thr4	UGU	ACA, ACU, ACG	916(±64)	1.42
Trp	CCA	UGG	943(±162)	1.46
Tyr1	GUA	UAC, UAU	769(±95)	1.19
Tyr2	GUA	UAC, UAU	1261(±126)	1.95
Val1	UAG	GUA, GUG, GUU	3840(±218)	5.96
Val2A	GAC	GUC, GUU	630(±98)	0.97
Val2B	GAC	GUC, GUU	635(±95)	0.98
4.5 S RNA			416(±63)	0.64

The number of tRNA molecules per cell and the fraction of tRNA out of total tRNA population in *E. coli* grown at 0.4 doublings per hour are shown as described in the text. ± stands for the standard deviations calculated from six independent measurements for each individual tRNA isoacceptor. The data on tRNA codon recognition patterns were obtained from (Björk, 1995; Garcia *et al.*, 1986; Ikemura, 1985; Ikemura and Ozeki, 1983; Komine *et al.*, 1990; Saxena and Walker, 1992).

[a] The tRNA isoacceptors Gly1 and Gly2 are treated collectively as are the data for Ile1 and Ile2. Highlighted with gold are the rows that correspond to our findings of the tRNA isoacceptors that are underrepresented in *E. coli* in respect of human genes expression.

Commercial suppliers have noticed a possible shortage already, and Novagen offers the Rosetta™ (DE3) strain that carries six tRNAs out of eight marked red in Table 4.4-2, in order to enable precise control of expression levels by adjusting the concentration of IPTG. The tRNAs introduced on a pRARE plasmid are *proL* tRNA2 (CCC), *leuW* tRNA3 (CUA), *argW* tRNA5 (AGG), *glyT* tRNA2 (GGA), *argU* tRNA4 (AGA and AGG), *ileX* tRNA2 (AUA).

Table 4.4-2. Rare codons in E. coli

Amino acid	Codon	Fraction in all genes	Fraction in Class II
Arg5	**AGG**	**0.022**	**0.003**
Arg4	**AGA**	**0.039**	**0.006**
Arg3	**CGG**	**0.098**	**0.008**
Arg2	**CGA**	**0.065**	**0.011**
Arg2	CGU	0.378	0.643
Arg2	CGC	0.398	0.330
Gly2	GGG	0.161	0.044
Gly2	**GGA**	**0.109**	**0.020**
Gly3	GGU	0.337	0.508
Gly3	GGC	0.403	0.428
Ile2	**AUA**	**0.073**	**0.006**
Ile1	AUU	0.507	0.335
Ile1	AUC	0.420	0.659
Leu5	UUG	0.129	0.034
Leu5	UUA	0.131	0.056
Leu3	CUG	0.496	0.767
Leu3	**CUA**	**0.037**	**0.008**
Leu2	CUU	0.104	0.056
Leu2	CUC	0.104	0.080
Pro3	CCG	0.525	0.719
Pro3	CCA	0.191	0.153
Pro2	CCU	0.159	0.112
Pro2	**CCC**	**0.124**	**0.016**

Codon usage is expressed here as a fraction of all possible codons for a given amino acid. "All genes" is the fraction represented in all 4,290 coding sequences in the *E. coli* genome (Nakamura *et al.*, 2000). "Class II" is the fraction represented in 195 genes highly and continuously expressed during exponential growth (Henaut and Danchin, 1996). Codons that are underrepresented in *E. coli* genes are marked in red. Numbers after amino acid represent tRNA isoacceptor species, according to Dong *et al.*

In order to homogenise the nomenclature of tRNA isoacceptors, I changed the names reported by Novagen to those used by Dong *et al.* This means that numbers after tRNAs in the following table and text correspond to the isoacceptor species, coding for a given amino acid.

We tested only two tRNAs as a simple check, $tRNA^{Leu}$ and $tRNA^{Ile}$. According to our analysis, only $tRNA^{Leu}$ of the two tRNAs that should be over-expressed proved the expectations: The $tRNA^{Leu}$ was charged to relatively higher levels as examined in an amino acylation assay, in comparison to $tRNA^{bulk}$ from standard non-induced *E. coli* cells, whereas $tRNA^{Ile}$ was not overexpressed (Table 3.7.1-1, page 72). Therefore, we are slow to use this strain for eukaryotic genes expression, whether *in vivo* or *in vitro*.

The important difference of our analysis to the tRNAs present in Rosetta™ (DE3) is the fact that 11 tRNAs should be supplied and not eight (highlighted red in Table 4.4-2, as suggested by Novagen; Henaut and Danchin, 1996; Nakamura *et al.*, 2000).

4.5 Design of an mRNA with an enhancer for high ribosome occupancy

Returning to initiation of translation, it is necessary to mention that 30S subunits, unlike eukaryotic 40S, are incapable of the process of scanning. A 30S subunit simply recognizes the ribosomal binding site (RBS). A sequence of the RBS of the mRNA, the Shine-Dalgarno sequence (SD), is known to be complementary to the sequence at 3' end of the 16S rRNA, known as anti-Shine-Dalgarno (anti-SD). The SD sequence is situated about 4 to 12 nucleotides upstream (in front) of the AUG start codon.

It is known that the very 3' end of 16S rRNA is in close proximity to the E-site of the 30S subunit, since the AUG start codon is positioned at the P-site – directed by the SD–anti-SD interactions, awaiting for IF2•GTP•Met-$tRNA_f^{Met}$ to come, and 50S subunit to assemble into the elongating 70S ribosome.

Imagine that RBS is “hidden” in a “forest” of secondary structures, which the 30S subunit alone cannot melt. As a result, we have no protein synthesis or one at very low levels. In fact, such a situation is much more usual for exogenous mRNAs, mRNAs of the genes with low expression levels, and for many eukaryotic mRNAs added to bacterial *in vitro* systems for protein synthesis. Even if the gene is introduced into *in vitro* system for transcription and translation, due to uncoupling of these processes by utilization of a T7 RNA polymerase, which is a fast molecule, mRNAs transcribed are still a hard task for 30S to initiate as well as for elongating 70S ribosomes. Therefore, we developed an mRNA with a cassette for an exogenous cistron that should allow high expression of this cistron. This construct shows the following features:

1: Weak secondary structures at the 5’-end

In early studies de Smit and van Duin presented data that expression is limited by mRNA structure or when mRNA has a low affinity for ribosomes. They also report experimental data on the energy of helix formation (ΔG^{0}_{f}) for mRNA, which in the range of –6.0 to –2.4 kcal/mol resulted in best expression levels, whereas a secondary structure of energy below –6.4 kcal/mol prevented ribosomal initiation (de Smit and van Duin, 1990). And another group reported few years earlier that the high preference for adenosine residues in true sites has indeed been ascribed to their low potential to form stable secondary structures (Looman *et al.*, 1987).

It is also known that both the eukaryotic and the prokaryotic ribosomes translate equally well artificial mRNAs such as poly-U and poly-A, without the aid of either a cap or IRES, which suggests that the basic RNA-binding properties of both kinds of ribosomes are the same. Another important feature for translation initiation is that an mRNA chain is single-stranded, which is a fundamental prerequisite for initiation to occur in both cell types: the unique ways in which such single-stranded molecule is ensured may suffice to account for most of the differences between prokaryotic and eukaryotic initiation (Londei, 2001).

It is important to consider the context of the 5'UTR sequences from natural enhancers and to include the starting coding sequence of the gene that follows. An example of a well-translated mRNA that does not have an SD sequence is the so-called epsilon sequence (UUAACUUUA) present in the leader region of *gene* 10 of the T7 bacteriophage. Is the absence of secondary structure responsible for its efficient translation? We analyzed the possible secondary structure of epsilon in the context of its WT mRNA by means offered on the mfold web server for nucleic acid folding and hybridization prediction developed by Zuker (Mathews *et al.*, 1999; Zuker, 2003), and found that the overall ΔG^0_f is relatively high (Figure 4.5-1), in comparison to the stability threshold identified as values below –6.0 kcal/mol.

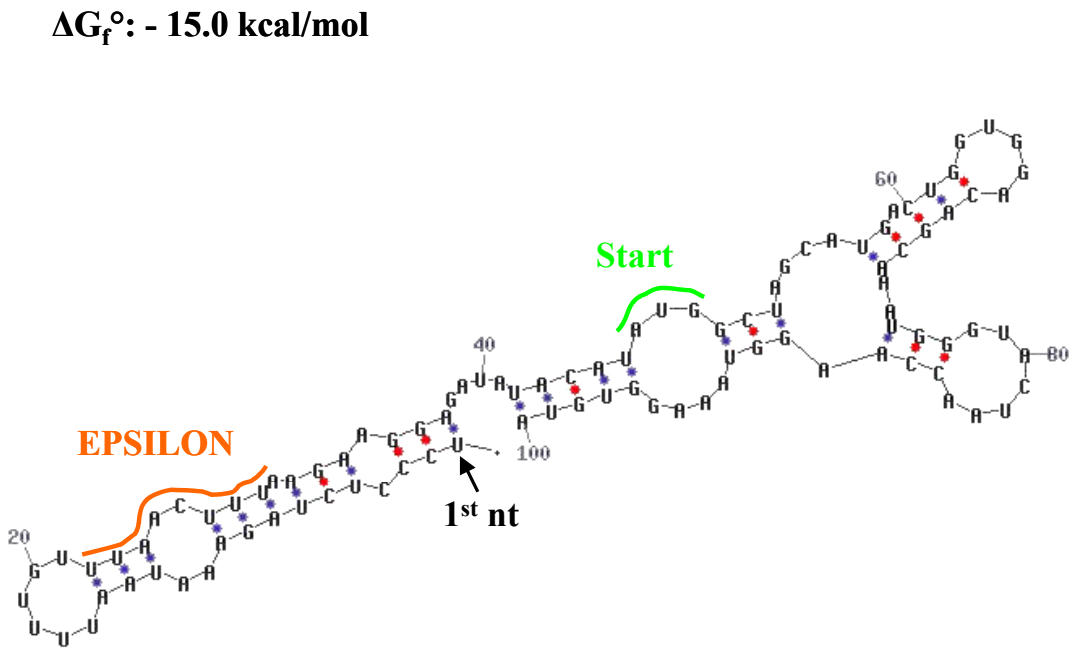

Figure 4.5-1 Epsilon sequence found in the leader region of gene 10 *of the T7 bacteriophage*. The $\Delta G_f°$ of the sequence that includes 5' untranslated region of *gene* 10, and about 50 nucleotides of sense codons. The 1st nucleotide (1st nt), the epsilon sequence and the AUG start codons are indicated.

However, secondary structures found in one mRNA molecule are not linked *via* stacking interactions. Therefore, we measured separately the stabilities of the secondary structures found in the case of *gene* 10 mRNA, and observed ΔG^0_f values, which can be solved by the 30S ribosomal subunit alone

(Figure 4.5-2A). Thus, a sequence, close to that of epsilon would be good if incorporated into the upstream region (5') of the mRNA molecule of interest.

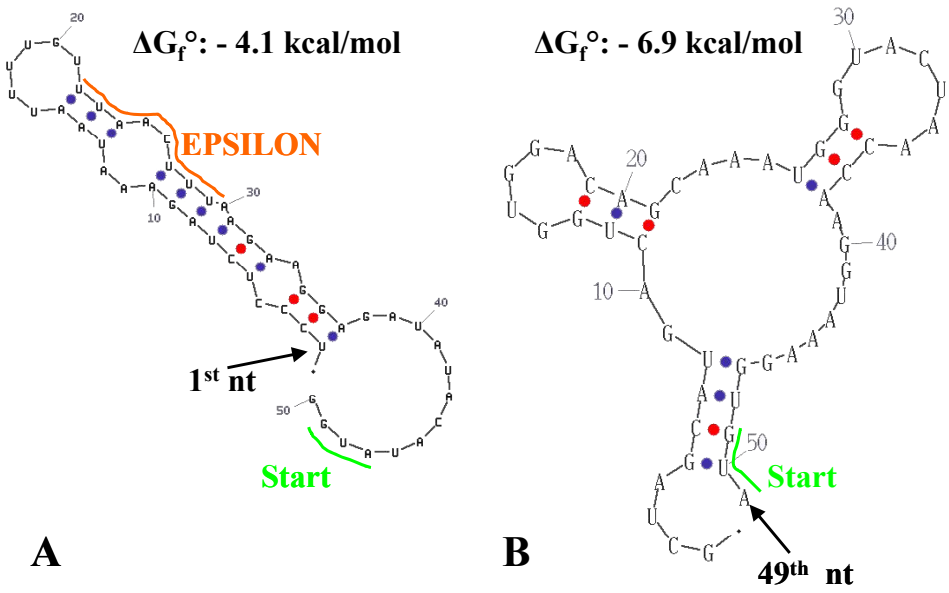

Figure 4.5-2 An individual secondary structures measurements of the epsilon gene 10 *T7 bacteriophage mRNA sequence*. (**A**) 5' untranslated region of *gene* 10, including AUG start codon; (**B**) secondary structure of AUG start codon followed by about 50 nucleotides of *gene* 10 coding sequence. The 1st nucleotide (1st nt), the epsilon sequence and the AUG start codons are indicated.

2: A short leader cistron preceding the exogenous cistron

We copied the codon sequence of the CAT leader peptide that precedes the chloramphenicol acetyl-transferase and is efficiently translated. This leader should warrant that the ribosomes easily initiate and thus lead to high ribosome occupancy of the mRNA.

3: Translation transition from the leader peptide to that of the exogenous cistron

The problem we would like to solve is that ribosomes present in high occupancies at the leader peptide should directly continue the translation of the following cistron without falling off the mRNA with a subsequent 30S *de novo* initiation. We applied a "trick" that is used during translation of polycistronic mRNAs coding for various ribosomal proteins. We included the transition site of the adjacent cistrons on a polycistronic mRNA coding for the ribosomal proteins

L29 and S17, where a short SD sequence of four nucleotides (for the downstream cistron) is followed by a stop codon UAA (for the upstream cistron) and the AUG-start codon under overlapping conditions: *GGU GCG* UAA **UG**. The stop codon UAA of the short leader is in red, the SD sequence of the following cistron is in italics, and the AUG start codon of the following cistron is in bold letters. The SD sequence is thus situated just four nucleotides upstream of this AUG-start codon and plays the role of an anchor that traps the empty (after termination and peptide release) ribosome exposing the AUG-codon near the P-site and reassuring the 70S initiation mode to occur. Thus, in front of the gene of interest we have a short open-reading frame that should facilitate the expression of the gene of interest.

The long row of sequence optimisation and energy comparisons of the respective secondary structures resulted finally in what we call *Berlin-sequence* (Figure 4.5-3 and Figure 4.5-4 for secondary structure), which we cloned into standard pET23c(+) vector, replacing the present T7 promoter with one carried by the insert.

Resuming everything mentioned above, this Berlin sequence can be described as a sequence based on successful examples from life, in order to create a good artificial non-SD enhancing sequence known for initiation of translation.

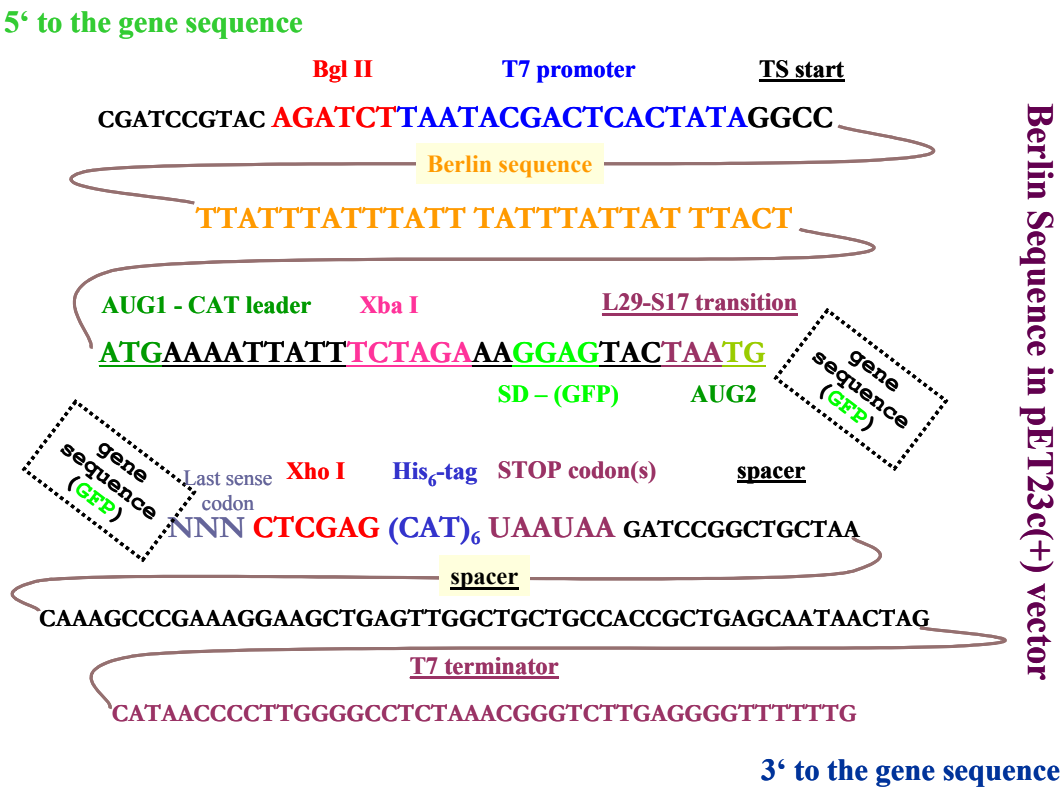

Figure 4.5-3 Sequence designed to enhance the initiation of translation. Marking of the important sites are given above and in colours.

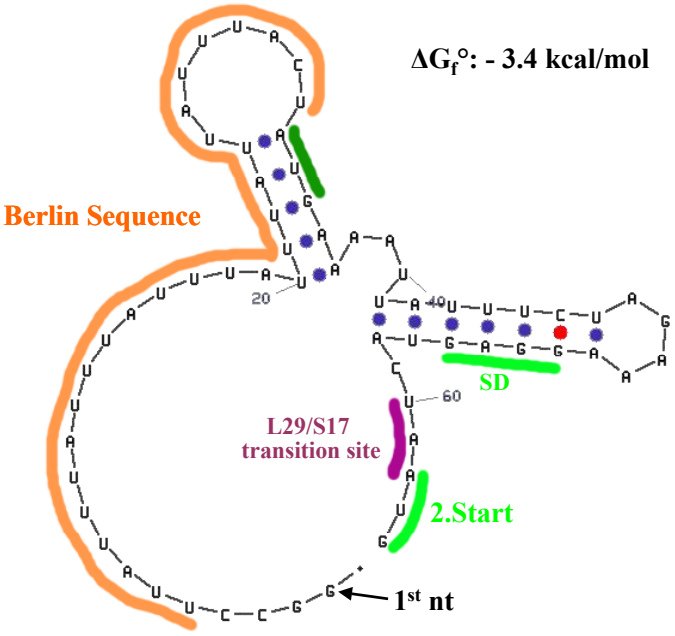

Figure 4.5-4 Secondary structure of the Berlin sequence designed to enhance initiation of translation. Marking are given according to the coloured regions that represent most significant sequences.

First results obtained by my colleague Witold Szaflarski in our group indicated that this construct triggers a super-expression *in vivo*, whereas the *in vitro* results are not yet satisfying. This part is an ongoing research program that includes further optimisation of the Berlin sequence.

4.6 Investigation of the fragmentation of a given protein

We aimed to reduce the level of fragmentation for *E. coli* translational elongation factor Tu (EF-Tu) by means that were already investigated and discussed in this work before. In our understanding the reduction of the incubating temperature may lead to the retardation or elimination of the unwanted fragments, which are observed during incubation of the reaction at 30°C. As a negative control for the temperature decrease effect we also ran a reaction with increased temperature that should have resulted in the increased level of fragmentation. The analysis was done for GFPcyc3 as well, assuming that this protein has almost zero level of fragmentation. We determined the *fragmentation index* (F.I.), in order to describe the level of fragmentation for EF-Tu. The calculation of this value was based on the relative intensities of the bands corresponding to full-length EF-Tu protein and two different and most pronounced fragments of EF-Tu, determined by radiolabelling (scanning of the SDS-PAAG), as indicated by the formula:

$$\text{F.I.} = (\text{frg. \#1} + \text{frg. \#2}) / (\text{frg. \#1} + \text{frg. \#2} + \text{EF-Tu})$$

According to this formula we calculated the F.I. of EF-Tu for different temperatures (Table 3.8-1, page 75) and found out that there is slight positive effect if at all.

In order to normalize the synthesis levels of full-length EF-Tu, we compared its band intensities to those of GFPcyc3, in kinetics at different incubation temperatures. We observed that the level of fragmentation towards the full-length protein at the beginning of each reaction was lower than at the end of the reactions. The other significant detail is that the total yield of EF-Tu at 37°C increased twice as compared to those at 20°C or 30°C (relative to

GFPcyc3 synthesis), and that the number and amount of fragments had increased, too (Table 3.8-1, page 75). We conclude that the reason for EF-Tu fragmentation is not related to the incubation temperature.

If the EF-Tu protein in the form of ternary complex is more stable towards fragmentation, then an extra addition of the amino acylated tRNA would stimulate the ternary complex formation and, as a result, reduce the fragmentation level. Because the previous results showed little affect of the incubation temperature on the EF-Tu fragmentation next reactions were performed at 30°C. Here, a three-fold excess of $tRNA^{Phe}$ and a twelve-fold excess of phenylalanine amino acid over the estimated synthesis of full-length EF-Tu protein was added either at the beginning of reaction, or one hour later after the incubation had already started. According to the estimation of F.I. and the overall analysis of the full-length EF-Tu protein synthesis (Table 3.8-2, page 77), the presence of $tRNA^{Phe}$:Phe mixture had again a little positive effect on the reduction of EF-Tu fragmentation when was added one hour later after incubation of the reaction had started.

In the presence of protease inhibitory mixes we observed in some cases a reduction of the EF-Tu fragments, but also an impairment of synthesis of the mature EF-Tu. This effect might be due to the components of the mixes, that could have acted as inhibitors of the protein synthesis. Besides this, the protease inhibitory mixes are recommended for storage of the preparative protein isolation from cell cultures.

Referring to the data obtained for incubation temperature affect on the EF-Tu synthesis, the major observation is that fragmentation of this protein is not much temperature dependent, but rather depends on the incubation time. The longer the incubation time, the more fragments are detected. According to our analysis, incubation up to two hours is recommended either at 30°C or at 20°C in order to obtain mainly mature EF-Tu protein. *In vivo*, additional factors might be involved. In fact, my colleague Yan Qin in our group has observed some

spectacular effects improving the output of fully active proteins by the addition of a universally conserved G-protein with a hitherto unknown function.

In the course of this thesis we have developed quality criteria that allow a critical evaluation of parameters important for the coupled transcription /translation system or improving the yield and quality of the synthesized protein. Improvements could be obtained by applying slower T7 polymerase and amino acid additions after half of the standard incubation time. Another example is that both lowering the incubation temperature to 25°C or 20°C plus the additional administration of amino acids improved the active fraction of the synthesized protein combined with a satisfying yield. We further identified 11 tRNAs that should be added to a bacterial system, *e.g.* from *E. coli*, for the optimal expression of eukaryotic genes. On the other hand, we could show that the shortage of NTPs or prolonging the half-life of mRNA does not improve the output of protein. But the optimisation coupled system as it stands after these analyses are not yet finished. This can be easily demonstrated by comparing the efficiency of the excellent RTS (Roche) with that of an *E. coli* cell. The reaction mix before synthesis contains about 40 mg per ml total proteins, and after 10 h an amount of GFP has been synthesized that comes to about 20% of the total proteins (8 mg/ml; Figure 3.4-3A, page 56). A continuation of this synthesis rate would lead to a doubling of the protein content in the reaction mixture (total proteins + synthesised GFP) after 50 h. *E. coli* has a doubling time of 20 min under reach medium conditions. It follows that the good RTS system is still 150-fold less efficient than protein synthesis *in vivo*. I have identified some ways to improve the system; others are under consideration of my colleagues in the Nierhaus group. Examples are a further optimisation of the Berlin sequence for *in vitro* expression and additions of newly found factors that improve the active fraction even at 30°C incubation temperature.