TARTU ÜLIKOOL
BIOLOOGIA-GEOGRAAFIA TEADUSKOND
Molekulaar- ja rakubioloogia instituut
Molekulaarbioloogia õppetool

Age Tats

BAKTERIVALKUDE TRANSLATSIOONI MÕJUTAVATE MOTIIVIDE BIOINFORMAATILINE ANALÜÜS

Keskastme lõputöö

Juhendajad Tanel Tenson, PhD
prof. Maido Remm, PhD

Tartu 2003

SISUKORD

SISUKORD 1
KASUTATUD LÜHENDID 3
1. TRANSLATSIOON PROKARÜOOTIDES 5
1.1 Bakteriaalse ribosoomi ehitus 5
1.2 Initsiatsioon 5
1.3 Elongatsioon 6
1.4 Terminatsioon 6
2. VALGU N-TERMINAALSETE SIGNAALIDE BIOLOOGILINE TÄHTSUS 7
2.1 Amino-terminaalse metioniini töötlemine 7
2.2 Translatsiooni efektiivsuse mõjutamine 7
2.3 Lokalisatsioonisignaalid 8
2.4 Degradatsioonisignaalid 9
3. TRANSLATSIOONI MÕJUTAVAD MOTIIVID TRANSLATSIOONI INITSIATSIOONI REGIOONIS 10
3.1 Shine-Dalgarno järjestus 11
3.2 Vahemik Shine-Dalgarno järjestuse ja startkoodoni vahel 11
3.3 Enhanser ja S1 valk 11
3.4 Startkoodonist positsioonis -1 asuv nukleotiid 12
3.5 Startkoodon 12
3.6 Startkoodonist downstream asuvad koodonid 13
4. KOODONKASUTUS *E.COLI*'S 13
4.1 Kõrgelt ekspresseerunud geenide ennustusmeetodid 13
4.1.1 CAI 14
4.1.2 CU 15
4.2 Koodonkasutus *E.coli* erinevates geeniklassides 16
4.3 Koodonkasutus kõrgelt ekspresseerunud geenides 16
II PRAKTILINE OSA 18
1. EESMÄRK 18
2. MEETODID 18
2.1 Andmed 18
2.2 Kõrgelt ekspresseerunud geenide valimine 19
2.3 Koodonkasutuse analüüs 19

2.4 Aminohapete kasutuse analüüs 20
2.5 Nukleotiidide sageduse analüüs 20
2.6 Dipeptiidide sageduse analüüs 20
2.7 HMM 21
2.8 Hüdrofoobsusprofiilide arvutamine 22
3. TULEMUSED 22
3.1 Analüüsimetoodika ülevaade. 22
3.2 Nukleotiidide sageduse jaotus valkude N-terminuses 24
3.3 Koodonkasutus positsioonide kaupa 29
3.4 Aminohapete kasutus positsioonide kaupa 30
3.5 Dipeptiidide kasutamine N-terminuses 32
3.6 HMM analüüsi põhimõte, eesmärk ja tulemused 32
3.7 Hüdrofoobsuse jaotus valkude N-terminuses 33
4. ARUTELU 36
5. KOKKUVÕTE 37
6. SUMMARY 38
7. VIITED 39

KASUTATUD LÜHENDID

A	alaniin
Ala	alaniin
Arg	arginiin
A-sait	aminoatsüülsait
Asn	asparagiin
Asp	aspartaat
C	tsüsteiin
CAI	koodoni adaptatsiooni ndex (*codon adaptation index*)
CU	koodonkasutus (*codon usage*)
C-terminaalne	karboksüülterminaalne
Cys	tsüsteiin
D	aspartaat
DR	allavoolu asuv *(downstream)* regioon
E	glutamaat
EF	elongatsioonifaktor
E-sait	väljumissait (*exit*-)
F	fenüülalaniin
fMet-tRNA^{fMet}	N-formüül-metionüül-tRNA
G	glütsiin
Gln	glutamiin
Glu	glutamaat
Gly	glütsiin
H	histidiin
His	histidiin
HMM	*Hidden Markov Model*
IF	initsiatsioonifaktor
I	isoleutsiin
Ile	isoleutsiin
K	lüsiin
L	leutsiin
Leu	leutsiin
Lys	lüsiin

M	metioniin
MAP	metioniini aminopeptidaas
Met	metioniin
mRNA	informatsiooni *(messenger*-) RNA
MSF	*multiple sequence file* (formaat)
N	asparagiin
N-formüül	amino-formüül
nt	nukleotiid
N-terminaalne	aminoterminaalne
P	proliin
Phe	fenüülalaniin
Pro	proliin
P-sait	peptidüülsait
PT	peptidüültransferaasi tsenter
Q	glutamiin
R	arginiin
RNA	ribonukleiinhape
rRNA	ribosomaalne RNA
S	seriin
SD	Shine-Dalgarno
Ser	seriin
SRP	signaali äratundev partikkel *(signal recognition particle)*
T	treoniin
TAT	kaksikarginiini translokatsioon
Thr	treoniin
TIR	translatsiooni initisiatsiooni regioon
Trp	trüptofaan
Tyr	türosiin
UTR	mittetransleeritav regioon (*untranslated region*)
V	valiin
Val	valiin
W	trüptofaan
Y	türosiin

I KIRJANDUSE ÜLEVAADE

1. TRANSLATSIOON PROKARÜOOTIDES

1.1 Bakteriaalse ribosoomi ehitus

Bakteriaalne ribosoom sedimentatsiooni koefitsendiga 70S koosneb suurest 50S subühikust ja väikesest 30S subühikust. 50S subühik koosneb 23S rRNA-st, 5S rRNA-st ja 31 valgust. 30S subühiku koostises on 16S rRNA ja 21 valku. Väike subühik dekodeerib peamiselt geneetilist informatsiooni, suur subühik vastutab peptiidi elongatsiooni ja valgu vabanemise eest. Tunnel, mille kaudu tekkiv peptiid väljub, algab suures subühikus asuva PT tsentri lähedalt, läbib suure subühiku ja jõuab suure subühiku tagaküljele. Eksperimentide tulemused on näidanud, et see pole sugugi ainus võimalik tee tekkiva peptiidi väljumiseks. Näiteks on leitud tõendeid, et globiin voltub ja seondub oma heemi rühmaga juba PT tsentris (Komar *et al*., 1997). Kuna väljumistunnel on väike, ei saa globiin ilmselt traditsioonilist teed pidi ribosoomist lahkuda ning võib seda teha läbi kahe subühiku liitumiskoha (Tenson and Ehrenberg, 2002). Samuti on näidatud, et mõned bakterite tekkivad peptiidahelad ristseonduvad 16S rRNA-ga, milleni ei saa jõuda väljumistunneli kaudu (Choi *et al*., 1998). Erinevate teede valiku aluseks on pakutud tekkiva polüpeptiidi N-terminuse aminohappe järjestust (Tenson and Ehrenberg, 2002).

1.2 Initsiatsioon

Translatsiooni initsiatsioon algab ribosoomi 30S subühiku 16S rRNA 3' otsa seondumisega mRNA-l asuva Shine-Dalgarno järestusega (SD). (Shine and Dalgarno, 1974). Järgnevalt siseneb N-formüül-metionüül-tRNA (fMet-tRNA^{fMet}) ribosoomi P-saiti. Mõlemat protsessi soodustavad kolm initsiatsioonifaktorit (IF): IF1, IF2, IF3. Moodustunud 30S initsiatsioonikompleks koosneb 30S subühikust, fMet-tRNA^{fMet}-st ja kolmest initsiatsioonifaktorist. 30S initsiatsioonikompleksi moodustumine on arvatavasti valgusünteesi taset määrav etapp ning on seetõttu peamine translatsiooni efektiivsuse määraja

(de Smit and van Duin, 1994; Draper, 1996). Elongatsioonifaktori EF-Tu abil toimetatakse ribosoomi A-saiti mRNA teise koodoniga komplementaarne aminoatsüül-tRNA.

1.3 Elongatsioon

Elongatsioon algab peptiidsideme moodustumisega, mida katalüüsib PT. P-saidis oleva initsiaator-tRNA aktiveeritud formüül-metioniin üksus viiakse üle A-saidis asuva aminoatsüül-tRNA aminorühmale. Järgneb ribosomaalne translokatsioon, mille käigus ribosoom liigub mRNA-l edasi kolme aluse võrra. Selle tulemusena liigub deatsetüleeritud tRNA E-saiti, peptidüül-tRNA P-saiti ning A-sait jääb tühjaks. Tühi A-sait on valmis järgmisele koodonile vastava aminoatsüül-tRNA sisenemiseks. Kirjeldatud tsüklit korratakse niikaua, kuni uus valk on valmis, st jõutakse stoppkoodonini.

1.4 Terminatsioon

Valgu sünteesi lõpetavad kolm koodonit: UAA, UAG, UGA. Ükski neist pole esindatud tRNA poolt. Vastavad koodonid tuntakse A-saidis ära terminatsioonifaktorite poolt, mis aktiveerivad ribosoomi hüdrolüüsima peptidüül-tRNA-d P-saidis (Ganoza, 1966; Capecchi, 1967; Scolnick *et al.*, 1968). Vabanemise reaktsioon on analoogne peptidüüli üleviimisele elongatsioonifaasis, kuid aktseptoriks on nüüd aminoatsüül-tRNA asemel vesi. Peale polüpeptiidahela vabastamist vabastatakse ribosoomist ka tRNA ning ribosoom dissotseerub 30S ja 50S subühikuteks (Bade *et al.*, 1969).

Terminatsiooni mõjutavad järjestusepõhiselt viimased tekkiva peptiidi aminohapped ja terminatsioonikoodonile järgnev nukleotiid (Björnsson *et al.*, 1996; Mottagui-Tabar *et al.*, 1994; Zhang and Li, 1987). Koodonkontekst stoppkoodoni ümbruses on mittejuhuslik ning terminatsiooni efektiivsus on korrelatsioonis koodonkasutusega (Brown *et al.*, 1990; Mottagui-Tabar and Isakssson, 1997). Peamine mõju terminatsiooni efektiivsusele on terminaalsel dipeptiidil ning sõltub dipeptiidi -2 aminohappe laengust (seotuna van der Waalsi jõududega) ning -1 aminohappe α-heeliksi soodustamise astmest ehk teisisõnu hüdrofoobsusest (Mottagui-Tabar and Isaksson, 1997).

2 VALGU N-TERMINAALSETE SIGNAALIDE BIOLOOGILINE TÄHTSUS

2.1 Amino-terminaalse metioniini töötlemine

Eubakteriaalse valgusünteesi alguses moodustub polüpeptiidile fMet ots (Kozak, 1983; Meinnel *et al*., 1993; Schmitt *et al*., 1996), mis ei ole valgu initsiatsiooniks aboluutselt vajalik ning seda võib vaadelda kui initsiatsiooni optimeerimise strateegiat (Mazel *et al*., 1994). fMet töödeldakse järgneva elongatsiooni käigus. Kõigepealt eemaldatakse N-formüül osa deformülaasiga, mille tulemusena jääb aminoterminusse metioniin. Märkimisväärse hulga valkude puhul eemaldatakse järgnevalt ka aminoterminaalne metioniin (ülevaade Sherman *et al*., 1985). Erinevates töödes on näidatud, et vastav ensüüm – metioniini aminopeptidaas (MAP) – lõikab metioniini ära juhul, kui metioniin esineb selliste aminohapete ees nagu Ala, Gly, Pro, Ser, Thr või Val, ning lõikamist ei toimu, kui metioniinile vahetult järgnev aminohape on Arg, Asn, Asp, Gln, Glu, Ile, Leu, Lys või Met (Tsunasawa *et al*., 1985; Ben-Bassat *et al*., 1987; Miller *et al*., 1987; Moerschell *et al*., 1990). Tavaliselt eelistatakse lõikusel metioniinile järgnevat jääki, mille külgahel on lühike: pika või koguka külgahela puhul pole MAP võimeline metioniini eemaldama (Hirel *et al*., 1989; Dalbogue *et al*., 1990; Schmitt *et al*., 1996).

MAP ei suuda eemaldada deformüleerimata metioniini ning seetõttu on alati vajalik eelnev N-formüül-metioniini deformülatsioon (Solbiati *et al*., 1999).

2.2 Translatsiooni efektiivsuse mõjutamine

Translatsiooni limiteerivaks faasiks enamikul juhtudel on initsiatsioon. Lisaks võtmeetapiks oleva 30S initsiatsioonikompleksi moodustumisele, kus on kriitilise tähtsusega paardumine SD ja 30S subühiku vahel, võivad rolli mängida ka enhanser piirkond (Boni *et al*., 1991; Zhang and Deutscher, 1992), mRNA (sekundaar)struktuur startsaidi juures (Dreyfus, 1988; de Smit and van Duin, 1994) ja koodoneelistus geenide alguses (Bulmer, 1988). Translatsioon on vähemal määral mõjutatav ka kahes ülejäänud etapis. Elongatsioonifaas võib peatuda tänu mRNA struktuuridele (Plat, 1988) ja haruldaste koodonite kasutusele (Varenne *et al*., 1989; Curran and Yarus, 1989). Viimasel juhul saab määravaks vastavate tRNA-de limitatsioon. Terminatsiooni efektiivsus sõltub terminatsioonikoodoni efektiivsusest (Jin *et al*., 2002).

Lisaks mängivad rolli ka terminatsioonifaktorite ja ribosoomi vahelise seondumise tugevus ja kiirus. Ebakorrektne terminatsioon ei vii mitte ainult defektse valguni, vaid on ka ressurside raiskamine (Rocha *et al*., 1999).

2.3 Lokalisatsioonisignaalid

N-terminaalsed järjestused ehk liiderjärjestused suunavad bakterites tekkivad või täielikult valmis valgud tsütosoolist plasmamembraani (Schatz and Beckwith, 1990). Signaaljärjestused võivad asuda ka valgu sees või C-terminaalses otsas (Kutay *et al*., 1995). Valkude ekspordiks on viis peamist rada. Kõigi radade signaaljärjestused on struktuurilt sarnased ning koosnevad positiivselt laetud N-terminaalsest domäänist (n-regioon), hüdrofoobsest keskmisest domäänist (h-regioon) ja polaarsest C-terminaalsest domäänist (c-regioon), mis ümbritseb lõikesaiti (von Heijne, 1988; Izard and Kendall, 1994) (joonis 1). Järjestuste üldises pikkuses on leitud varieerumist 15 kuni 50 aminohappejäägini (von Heijne, 1986). Variatsioonide põhjustajaks on erineva pikkusega n-regioon.

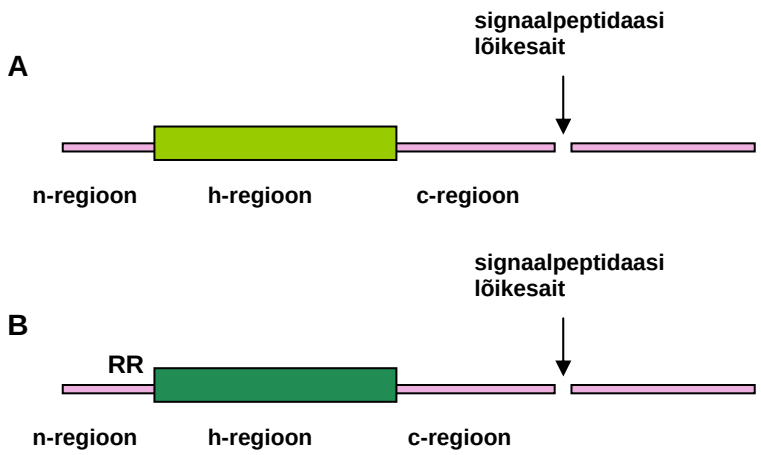
Kaks peamist ekspordirada on posttranslatsiooniline Sec rada ja TAT ehk kaksik-arginiini rada (*twin-arginine translocation*). TAT-raja substraatidele on iseloomulikud ebatavaliselt pikad kaksikarginiini sisaldavad signaalpeptiidid konsensusjärjestusega S/T-R-R-X-¤-¤ (¤ on hüdrofoobne jääk) N-terminuses, kus kaksikarginiin on täielikult konserveerunud (Berks, 1996). Lisaks omavad enamik kaksikarginiini signaaljärjestusi C-terminaalses regioonis vahetult enne signaalpeptidaasi lõikesaiti ühte või enamat positiivselt laetud jääki, nn „Sec-vältimise“ signaali (Bogsch *et al*., 1997). TAT raja signaalpeptiidid on märksa vähem hüdrofoobsed kui Sec raja signaalpeptiidid ning kui tõsta TAT raja signaaljärjestuse h-regiooni hüdrofoobsust, suunatakse vastavat signaali kandev valk Sec rajasse (Cristóbal *et al*., 1999).

Kolmas, SRP-sõltuv rada on osaliselt kattuv Sec-sõltuva rajaga. Bakteriaalne SRP transpordib mõned substraadid Sec rajasse (Valent *et al*., 1998).

Esineb ka valke, mis sisenevad membraani spontaanselt prootongradiendi osalusel. Kõik sellisel viisil translokeeritud bakteriaalsed valgud järgivad reeglit, mis määrab, et peptiide, millel esineb palju positiivseid jääke, ei translokeerita läbi membraani (Duong *et al*., 1997).

ΔpH translokatsiooni poolt transporditud valkude liiderjärjestused on väga sarnased Sec-sõltuva raja substraatide signaaljärjestustele.

Peale membraani sisenemist lõigatakse üldjuhul kõik signaaljärjestused hoolimata rajast prekursor-valgu küljest lahti membraaniseoselise signaalpeptidaasiga. Lõikamise toimumine või mittetoimumine sõltub signaaljärjestuse positsioonides -3 ja -1 lõikesaidist N-terminaalses suunas paiknevatest aminohapetest: lõikamine toimub, kui -1 positsioonis on väikese külgahelaga aminohape ja positsioonis -3 on laenguta aminohape (von Heijne, 1985). Lisaks mõjutavad lõikamist ka h-regiooni pikkus ja n-regiooni omadused (Nilsson *et al*., 1994).



Joonis 1. Bakteriaalsed signaalpeptiidid. Sec raja signaalpeptiidi struktuur (A). Kaksikarginiin motiivid (B) suunavad valgud TAT rajasse.

2.4 Degradatsioonisignaalid

N-terminaalne reegel (*N-end rule*) seob valgu eluea tema N-terminaalsete aminohappejääkidega. Valkude omadusi, mis tingivad molekuli lagundamise, kutsutakse degradatsioonisignaalideks ehk N-degroniteks. Peamine degradatsioonisignaali komponent on valgu destabiliseeriv N-terminaalne aminohappejääk. Nimelt teatud aminohapete esinemine valgu N-terminaalses otsas suunab valgu degradatsiooni. Destabiliseerivad aminohappejäägid jagatakse primaarseteks, sekundaarseteks ja tertsiaarseteks. *E.coli*'s tertsiaarseid destabiliseerivaid jääke ei esine (joonis 2). Olemas on ka stabiliseerivad jäägid – Gly, Val, Met – ,millele N-terminaalse reegli raja sihtmärkkomponendid ei seondu piisavalt efektiivselt või ei modifitseeri neid piisavalt efektiivselt (Varshavsky, 1996).

.

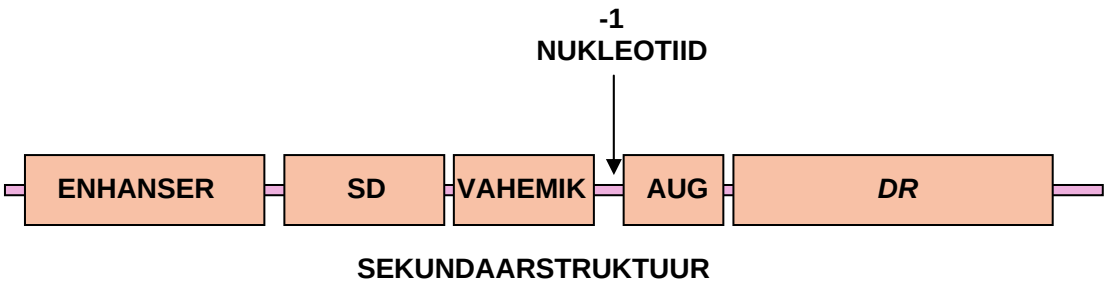
.

Primaarsed	**F L W Y**
Sekundaarsed	**R K**
Stabiliseerivad	**H I N Q D E C A S R G V P[?] M**

Joonis 2. Destabiliseerivad aminohappejäägid *E.coli*'s. (Küsimärk proliini juures tähistab ebakindlat kuuluvust stabiliseerivate järjestuste hulka.)

3 TRANSLATSIOONI MÕJUTAVAD MOTIIVID TRANSLATSIOONI INITSIATSIOONI REGIOONIS

Tähtsaks translatsiooni initsiatsiooni taseme määrjaks nii pro- kui ka eukarüootidel on initsiatsioonikoodoni kontekst ehk translatsiooni initsiatsiooni regioon (TIR). Lugedes initsiatsioonikoodoni esimeseks positsiooniks +1, paiknevad motiivid, mis on olulised prokarüootse mRNA TIR-i äratundmiseks, positsioonide -20 ja +15 vahel (Stormo *et al.*, 1982). Nendeks on 7 translatsiooni efektiivsuse seisukohalt olulist motiivi: TIR-i sekundaarstruktuur, Shine-Dalgarno järjestus (SD), enhanser, vahemaa SD ja startkoodoni vahel, startkoodonist -1 positsioonis asuv nukleotiid, startkoodonile järgnev regioon ning loomulikult startkoodon ise (joonis 3).



Joonis 3. Translatsiooni efektiivsust määravad motiivid translatsiooni initsiatsiooni regioonis.

3.1 Shine-Dalgarno järjestus

Shine-Dalgarno järjestus on startkoodonile eelnev lühike puriinirikas järjestus, mis paardub komplementaarse järjestusega (anti-SD) ribosoomi väikese subühiku 16S rRNA 3' otsa lähedal (5' – UCACCUCCUUA – 3') (Shine and Dalgarno 1974; Steitz and Jakes 1975). Seondumise peamine tähtsus seisneb ribosoomi paigutamises õige startkoodoni juurde. De Boer *et al* (1983) poolt uuritud pikkade SD järjestuste inhibitoorne efekt translatsioonile leidis taas käsitlust Komarova *et al* (2002) poolt. Selgus, et pikendatud SD järjestus (10 nt) on palju vähem efektiivne kui lühemad SD järjestused (efektiivseim 6 nt). Sakai *et al* (2000) poolt leiti, et SD motiivide järjestused on kõige sagedamini 7-meersed; mõõdeti vabaenergia väärtuseid mRNA 5'UTR otsa ja 7-meerse SD motiivi paardumisel (*gap*'e ja *mismatch*'e ei lubatud), mis andsid informatsiooni SD järjestuse konserveerumise kohta - madal vabaenergia väärtus viitas kõrgele konserveerumisele. Kõige sagedamini esinev SD järjestus *E.coli*'s on TCAGGAG (Sakai *et al*, 2000). SD pole siiski translatsiooni initsiatsiooniks hädavajalik, kuna initsiatsioon toimub, kuigi väiksema efektiivsusega, ka mRNA-lt, mis algab koheselt startkoodoniga (vanEtten and Janssen, 1998).

3.2 Vahemik Shine-Dalgarno järjestuse ja startkoodoni vahel

Vahemik SD järjestuse ja startkoodoni vahel on korrelatsioonis nii SD järjestuse konserveerumise kui mustriga (Shultzaberger *et al*., 2001) ning seega ka translatsiooni efektiivsusega. Optimaalseks vahemiku pikkuseks on pakutud 5 nukleotiidi (Chen *et al*., 1994), mis vastab ka 16S rRNA 3' otsa ja ribosoomi P-saidi vahelise vahemaa pikkusele.

3.3 Enhanser ja S1 valk

Enhanser on A/U rikas järjestus enne SD järjestust. Enhanser seondub mRNA-le seonduva ribosomaalse valgu S1-ga (Boni *et al*., 1991; Zhang and Deutscher, 1992). S1 on suurim ribosomaalne valk, mis paikneb 30S subühikus. S1 tagab translatsiooni ka sellistelt *E.coli* mRNA-delt, millel puudub või on nõrk SD järjestus (Roberts and Rabinovitz, 1989; Farwell *et al*., 1992; Tzareva *et al*., 1994). Hiljuti on näidatud, et S1 on olulise tähtsusega ka pikki SD järjestusi omavate mRNA-de efektiivseks translatsiooniks, moduleerides SD-aSD

interaktsiooni tugevust (Komarova *et al*., 2002). Seega muudab S1 translatsiooni 10 nt SD järjestustega mRNA-delt võimalikuks hoolimata teoreetilisest eeldusest, et väga pikk SD-aSD dupleks peaks ribosoomi täielikult lõksu püüdma.
Kuna S1 seondumine enhanserile suudab translatsiooni efektiivsust märkimisväärselt tõsta, pakub Komarova *et al* (2002), et S1 interaktsiooni mRNA-ga võib vaadelda esimese etapina translatsiooni initsiatsioonil ja seda isegi siis, kui mRNA kannab pikka SD järjestust.

3.4 Startkoodonist positsioonis -1 asuv nukleotiid

Kolmandaks initsiatsioonikoodoni kontekstis esinevaks initsiatsioonitaseme mõjutajaks on algselt oletatud ja hiljem ka *in vivo* tõestust leidnud seos vahetult startkoodonile eelneva nukleotiidi (-1 positsioon) ja $tRNAf^{Met}$ 37. positsiooni vahel (Esposito *et al*., 2003). 37. positsioonis konserveerunud adeniin ning -1 positsioonis asuv uridiin soodustavad translatsiooni. Vastav nukleotiidide kasutus pole erinevates prokarüootides universaalne ning võimalik on ka GC variant. Kuigi see 5' suunas pikendatud koodon-antikoodon interaktsioon ei ole initsiatsiooniks absoluutselt vajalik, võib see anda lisaspetsiifilisuse mittestandardsete initsiatsioonikoodonitega transkriptidele ja/või säilitada translatsiooni initsiatsiooni spetsiifilisust teatud keskkonna poolt põhjustatud stressitingimustes või juhul, kui mRNA või tRNA hulgad on limiteeritud (Esposito *et al*., 2003).

3.5 Startkoodon

Eelistatuimaks startkoodoniks eubakterites on AUG. *E.coli*'s kasutatakse järgmisi startkoodoneid: AUG (~90%), GUG (~8%), UUG (~1%) ja AUU (väga harva) (Schneider *et al*., 1986). Genoomis esinevate startkoodonite suhteline sagedus ja genoomi G+C sisaldus ei ole omavahel korrelatsioonis. AUG eelistuse põhjuseks on tema komplementaarsus initsiaator-tRNA antikoodoniga.

3.6 Startkoodonist downstream asuvad koodonid

Algselt arvati, et startkoodonist allavoolu asuv regioon (DR) mõjutab translatsiooni initsiatsiooni mRNA-rRNA komplementaarse aluspaardumise kaudu. Mutatsiooniliste muutuste sisseviimisel nii mRNA-sse kui 16S rRNA-sse selline teooria aga kinnitust ei leidnud (Firpo and Dahlberg, 1998; O'Connor *et al*., 1999). Siiski mõjutab see regioon translatsiooni initsiatsiooni kontekstipõhiselt (O'Connor *et al*., 1999). Erinevate +2 koodonite kasutusel on saavutatud 15-20 kordne translatsiooni efektiivsuse muutus (Stenström *et al*., 2001). Siiski pole see ilmselt limiteeritud vaid +2 koodoniga. Kuna on leitud, et sama tRNA poolt dekodeeritud isokoodonid annavad ekspressioonitasemelt enamikel juhtudel sarnase efekti, võib oletada, et tähtsust omab ka tRNA struktuur, kuid see pole täpsemat kinnitust leidnud. Kuni 5 koodoni pikkuste erinevate DR regioonide uurimisel on selgunud, et nn. tugev DR regioon võib anda translatsiooni tasemel kõrge efekti ka ilma sobiva Shine-Dalgarno järjestuseta (Stenström *et al*., 2001).

4 KOODONKASUTUS *E.COLI*'S

4.1 Kõrgelt ekspresseerunud geenide ennustusmeetodid

Kõrgelt ekspresseerunud geenidel nii pro- kui eukarüootidel esineb tugev koodoneelistus. Kaheks laiemalt levinud arvuliseks näitajaks geenide ekspressioonitaseme ennustamisel on koodoni adaptatsiooni indeks (*codon adaptation index* (CAI)) ja *codon usage* (CU). Kolmanda variandina on hiljuti pakutud lineaarset mudelit, mis põhineb lineaarsel regressioonil (Jansen *et al*., 2003). (Kuna vastav mudel on suhteliselt uus ja pole veel laia kasutamist leidnud, siis käesolevas töös sellel pikemalt ei peatuta). Kõik kolm mudelit annavad sarnase küllaltki kõrge korrelatsiooni (0.69-0.72) tegelike ekspressioonitaseme andmetega, mis saadi geenikiibilt ja 2D-geelelektroforeesilt (Jansen *et al*., 2003).

4.1.1 CAI

CAI on kvantitatiivne meetod geeni ekspressioonitaseme ennustamiseks tema koodonjärjestuse baasil (Sharp and Li, 1987). Algselt loodi ta pärmi (*Saccharomyces cerevisiae*) põhjal, kuid on kasutatav ka teistes organismides. CAI omistab igale koodonile (va stoppkoodonid) parameetri, milleks on relatiivne adaptiivsus (w, relative adaptivness). Koodoni relatiivne adaptiivsus on defineeritud kui tema sageduse suhe kõige rohkem kasutatavasse sünonüümsesse koodonisse. Parameetri arvutamise aluseks on kõrgelt ekspresseerunud geenide hulk (24 geeni), kuhu kuuluvad ribosoomivalke (ca pooled), välismembraanivalke, elongatsioonifaktoreid, kuumaehmatuse valke ja RNA polümeraasi subühiku valke kodeerivad geenid. Relatiivne adaptiivsus väljendub järgnevalt:

$$w_{aa,i}(G) = \frac{f_{aa,i}(G)}{f_{aa,\max}(G)},$$

kus $f_{aa,i}$ on koodoni i (mis kodeerib aminohapet aa) sagedus ja $f_{aa,max}$ on kõige rohkem kõrgelt ekspresseerunud geenides G aminohappe aa kodeerimiseks kasutatava koodoni sagedus. Relatiivne adaptiivsus varieerub 0-st 1-ni. Vastavalt 0, kui koodonit ei kasutata rühmas G üldse, ja 1, kui koodon esineb antud aminohappe puhul rühmas G kõige rohkem.

Geeni g CAI on kõigi geenijärjestuses olevate koodonite relatiivsete adaptiivsuste geomeetriline keskmine:

$$CAI_g = \prod_{i=1}^{N} w_i^{1/N},$$

kus w_i on i-nda koodoni relatiivne adaptiivsus geenis, milles on N koodonit. Selle valemi saab muuta järgnevaks valemiks:

$$CAI_g = \prod_{k=1}^{61} w_k^{X_{k,g}},$$

kus w_k tähistab nüüd geneetilise koodi 61 koodonist k-nda koodoni relatiivset adaptiivsust. $X_{k,g}$ on koodoni k fraktsioon geenis g olevate koodonite koguarvus:

$$X_{k,g} = \frac{C_{k,g}}{\sum_{i=1}^{61} C_{i,g}},$$

kus $C_{k,g}$ on koodoni k esinemise arv geenis g. Nagu relatiivne adaptiivsuski, varieerub CAI 0-st 1-ni. Kõrgemad CAI väärtused näitavad, et vastavad geenid on tõenäoliselt kõrgelt ekspresseeruvad (Sharp and Li, 1987).

Kontrastina CU mudelile ja lineaarsele mudelile on CAI parameetrid normaliseeritud sünonüümsete koodonite kasutusega, kuid sellest hoolimata annab CAI mudel sama häid tulemusi (Jansen *et al*., 2003).

Algselt vähene andmehulk (kõigest 24 geeni) on tegelikult ettenägelikult valitud ning annab adekvaatse tulemuse. Hoolimata suurest hulgast praegu kättesaadavatest ekspressiooniandmetest ei ole õnnestunud hiljem korrelatsiooni CAI ja tegeliku ekspressioonitaseme vahel tõsta rohkem kui 0.01 punkti võrra (Jansen *et al*., 2003). Sellise tulemuse korral kuulus vastavasse algandmete gruppi, mille pealt arvutati iga koodoni relatiivne adaptiivsus, 21 geeni, so vaid 3 geeni vähem kui algselt loodud mudelis.

4.1.2 CU

CU (*codon usage*) indeks avaldati kui alternatiivne võimalus CAI-le (Karlin and Mrazek, 2000). See mudel kasutab samuti koodonite esinemist geenis, et ennustada geeniekspressiooni taset, kuigi esitus on erinev. Ekspressioonimõõt E(g) defineerub järgmiselt:

$$E(g) = \frac{B(g \mid C)}{\frac{1}{2}B(g \mid RP) + \frac{1}{4}B(g \mid Ch) + \frac{1}{4}B(g \mid Tf)},$$

kus geenikomplekt C sisaldab kõiki geene genoomis, RP ribosomaalseid valke, Ch *chaperon*'e ja Tf transkriptsioonifaktoreid. Kõrgelt ekspresseerunud geenid omavad vastavalt kõrgemat E(g) väärtust kui madalalt ekspresseeruvad geenid (Karlin and Mrazek. 2000).

4.2 Koodonkasutus *E.coli* erinevates geeniklassides

Samuel Karlin (Karlin *et al.,* 1998) on uurinud koodonkasutust erinevatel alustel moodustatud geenigruppides. Laialt aktsepteeritud on kõrge korrelatsiooni esinemine nn optimaalsete koodonite ja geeni ekspressioonitaseme vahel. Geenid jagati ka kolmeks võrdseks osaks: amino-, keskmiseks ja karboksüülosaks ning analüüsiti iga gruppi kui erinevate „geenide“ gruppi. Ilmnes, et geenide keskmine ja viimane kolmandik on koodonkasutuselt sarnasemad kui kumbki neist võrrelduna esimese kolmandikuga. Geenide grupeerimisel funktsiooni järgi on üldteada ribosomaalseid valke kodeerivate geenide koodonkasutuse suur erinevus, mis on ilmselt omakorda seotud ka vastavate geenide kõrge ekspressiooniga. Samas *E.coli* eksponentsiaalses kasvufaasis teised olulised kõrgelt ekspresseeruvad geenid on aminoatsüül-tRNA süntetaasi geenid, mille koodoneelistus pole aga sugugi nii ekstreemne kui ribosoomivalke kodeerivatel geenidel. Siiski kõige sarnasem koodonkasutus keskmise geeni koodonkasutusega esineb geenides, mis kodeerivad transpordivalke ja nukleotiidide ja aminohapete bisosünteesiga seotud valke. Seevastu kõige erinevamad on translatsiooni, transkriptsiooni ja energia metabolismiga seotud valke kodeerivad geenid. Koodoneelistus *E.coli*’s võib olla seotud ka geenide pikkusega. Lühematel geenidel on tavaliselt tugevam koodoneelistus. Lisaks on Karlin *et al* poolt veel näidatud, et hiljuti imporditud geenidel ehk nn tulnukgeenidel ilmneb väga erinev koodonkasutus peremehe geenide koodonkasutusest. Need geenid on üldiselt tundmatu funktsiooniga ning kas AT- või GC-rikkad.

4.3 Koodonkasutus kõrgelt ekspresseerunud geenides

E.coli ja paljud teised bakterid kasutavad sünonüümseid koodoneid väga erineva sagedusega. Koodoneelistus on märkimisväärselt tugevam kõrgelt ekspresseerunud geenides ning on näidatud selge korrelatsioon sünonüümsete koodonite kasutuse ning valgu hulga vahel (Ikemura, 1985). Selline eelistuse erinevus on translatsiooni efektiivsusele suunatud selektsiooni tagajärg. Kõrgelt ekspresseerunud geenide spetsiifilise koodoneelistuse tekke peamiseks põhjuseks on mitmete autorite poolt pakutud tRNA-de kättesaadavuse varieerumine (näit Andersson *et al*., 1990).

Sakai *et al* (2002) poolt on näidatud *E. coli*’s selge korrelatsioon Shine-Dalgarno järjestuse konserveerumise ja CAI väärtuste vahel: geenid, millel on kõrgem CAI väärtus omavad sagedamini konserveerunud SD järjestust kui vastavate organismide madalama CAI-ga

geenid; seega kuna CAI väärtus on korrelatsioonis geeni ekspressioonitasemega, järeldub, et SD järjestuse konserveerumise ulatus on korrelatsioonis translatsiooni efektiivsusega; põhjuseks ilmselt asjaolu, et sagedasti ekspresseeruvate geenide puhul peab ribosoom SD järjestuse efektiivsemalt ära tundma kui harvem ekspresseeruvate geenide puhul (Sakai *et al*., 2001). Samas Komarova *et al* (2002) märgib vastuoluliselt, et kõrgelt ekspresseeruvatel mRNA-del ei ilmne translatsiooni efektiivsuse ja SD-aSD dupleksis esinevate komplementaarsete aluspaaride arvu vahel korrelatsiooni, kuid nad jagavad sarnaseid omadusi: neil on üsna pikad 5' UTR-id efektiivsuse kindlustamiseks ning mõnel juhul ka translatsiooni reguleerimiseks.

Sakai *et al* (2002) poolt leiti ka, et kõrgema CAI-ga geenid kasutavad kõige sagedamini startkoodonina koodonit AUG. AUG koodon interakteerub tugevamalt initsiaator-tRNA-ga ja on seetõttu kõrgelt ekspresseerunud geenides eelistatud, et saavutada kõrget translatsiooni efektiivsust (Sakai *et al*., 2001).

Sünonüümsete koodonite mittejuhuslik kasutus *E.coli* kõrgelt ekspresseeruvate geenide alguses viitab sellele, et initsiatsioonikoodonist allavoolu asuvad koodonid on lisaks SD järjestusele ja startkoodonile samuti tähtsad initsiatsiooni efektiivsuse seisukohalt. Allavoolu koodonite puhul on kinnitust leidnud +2 positsioonis asuva koodoni kõrge adeniini sisalduse seos kõrge ekspressioonitasemega ning vastupidiselt, madala ekspressiooniga on seotud G ja C rikkad koodonid; on võimalik, et A-rikkad koodonid hoiavad mRNA selles piirkonnas struktureerumata ja sellega soodustavad initsiatsiooni (Stenström *et al*., 2001).

SD ja DR järjestused toimivad koos translatsiooni efektiivsusele, kuid nad võivad ka iseseisvalt anda kõrge geeniekspressiooni (Stenström *et al*., 2002).

II PRAKTILINE OSA

1 EESMÄRK

Käesoleva töö eesmärgiks on järjestuse analüüsi abil selgitada, millised bioloogilised signaalid esinevad valkude N-terminuses, võrrelduna teiste valgu piirkondadega ning millised neist signaalidest on iseloomulikud kõrgelt ekspresseerunud valkudele.

2 MEETODID

2.1 Andmed

Töös on kasutatud 7 erineva bakteri geeni- ja valgujärjestusi: *Escherichia coli* (tüvi K-12, 4290 valku kodeerivat geeni), *Bacillus subtilis* (4100 geeni), *Helicobacter pylori* (tüvi J99, 1491 geeni), *Haemophilus influenzae* (1709 geeni), *Rickettsia prowazekii* (834 geeni), *Mycobacterium tuberculosis* (tüvi H37Rv, 3918 geeni) ja *Treponema pallidum* (1031 geeni). Kõik järjestused pärinevad Genbank'i andmebaasist.
(ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Escherichia_coli_K12/,
järjestuse nr. Genbank'is U00096.
ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Bacillus_subtilis/, järj. nr.AL001926.
ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Helicobacter_pylori_J99/,
järj. nr. AE001439.
ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Haemophilus_influenzae/, järj. nr. L42023.
ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Rickettsia_prowazekii/, järj. nr. AJ235265.
ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Mycobacterium_tuberculosis_H37Rv/, järj. nr. AL123456.
ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Treponema_pallidum/, järj. nr. AE000520)

2.2 Kõrgelt ekspresseerunud geenide valimine

E.coli kõrgelt ekspresseerunud geenide andmehulgad valisime programmi CodonW (John Peden, versioon 1.3, http://www.molbiol.ox.ac.uk/cu/) abil arvutatud CAI (*codon adaptation index*) järgi. 40 kõrgelt ekspresseerunud geenist koosnevasse andmehulka kuulus vastavalt 40, 60 geenist koosnevasse 60 ja 80 geenist koosnevasse 80 esimest kõrgeima CAI'ga geeni. Ülejäänud kuue bakteri kõrgelt ekspresseerunud geenideks võtsime nende ortoloogid *E.coli* 40 kõrgelt ekspresseerunud geeniga. Ortoloogid leidsime kahe bakterigenoomi võrdlusel BLASTiga (http://www.ncbi.nlm.nih.gov/blast/) kõik valgud kõigi vastu ja valisime sellised paarid, mis andsid mõlemas suunas parima skoori.

2.3 Koodonkasutuse analüüs

Koodonkasutuse analüüs sisaldas kolme etappi:

1) eelistatud ja mitte-eelistatud koodonite leidmine geeni positsioonides 1-10 (võrrelduna ülejäänud geenipiirkonnaga) üle kõigi *E.coli* geenide.

2) eelistatud ja mitte-eelistatud koodonite leidmine geeni positsioonides 1-10 (võrrelduna ülejäänud geenipiirkonnaga) üle *E.coli* 40 kõrgelt ekspresseerunud geeni.

3) tulemuste võrdlemine.

Eelistatud ja mitte-eelistatud koodonite leidmine põhines χ^2 testil, mille abil saab kontrollida, kas vaatlustulemused on kooskõlas teooria poolt ennustatavatega või ei kinnita vaatlustulemused teooriat. χ^2 leitakse järgneva valemi põhjal:

$$\chi^2 = \sum \frac{(A-E)^2}{E},$$

kus A (*actual,* kasutusel on ka termin *observed* (O)) on vaadeldud tulemus ja E (*expected*) on teooria põhjal oodatud tulemus. Mida enam vaadeldud tulemus erineb teooriast, seda suuremaks saab ka antud statistiku väärtus. Antud juhul seega, mida suurem on χ^2, seda rohkem vastavat koodonit vastavas positsioonis eelistatakse $(A-E>0)$ või välditakse $(A-E<0)$ võrrelduna geeni ülejäänud osaga. χ^2 –jaotuse täiendkvantiilide h väärtuseks ehk nn. kriitiliseks väärtuseks antud töös on 6,635 (usaldusnivool 0,001 ja vabadusastmete arvul 1). Vaatluse alt on välja jäetud esimene koodon (initsiatsioonikoodon).

2.4 Aminohapete kasutuse analüüs

Aminohapete kasutuse analüüsi tegime analoogselt koodonkasutuse analüüsile, (vt. punkt 2.3). Vaatluse alt on välja jäetud esimene aminohape.

2.5 Nukleotiidide sageduse analüüs

Nukleotiidide sageduse leidsime kõigile seitsmele eespool nimetatud bakterile. Välja on jäetud esimene positsioon, mis on startkoodon ja antud kontekstis tähtsust ei oma. Analüüs põhines A, U, C ja G nukleotiidide arvu kokkulugemisel ja protsentuaalse sageduse leidmisel igas koodonis geenide esimeses 15 positsioonis. Analüüsisime nii kõigi geenide kui ka 40 kõrgelt ekspresseerunud geeni nukleotiidide sagedust kolmes erinevas grupis:

1) nukleotiidide üldine sagedus, kus pole arvestatud nukleotiidi täpset asukohta koodoni sees.
2) sünonüümsetes positsioonides (tavaliselt 3. positsioon, Leu ja Arg puhul 1. ja 3. positsioon ning Ser puhul kõik kolm positsiooni) asuvate nukleotiidide sagedus.
3) mittesünonüümsetes positsioonides asuvate nukleotiidide sagedus.

Ekstreemsete punktide juhuslikkust kontrollisime χ^2-testiga.

2.6 Dipeptiidide sageduse analüüs

Dipeptiidide analüüsil võrdlesime kõigi valkude esimeses 10 positsioonis esinevate dipeptiidide sagedusi ülejäänud positsioonides esinevate dipeptidiide sagedustega. Selleks jagasime vaadeldud dipeptiidi sageduse eeldatava dipeptiidi sagedusega ning saadud jagatise esitasime kahendlogaritmi kujul.

2.7 HMM

Kolme erineva HMM (*Hidden Markov Model*) aluseks võtsime aminohapped positsioonides 2-12

1) 40 kõrgelt ekspresseerunud valgust,
2) 60 kõrgelt ekspresseerunud valgust,
3) 25 kõrgelt ekspresseerunud valgust.

FASTA formaadist HMM jaoks vajalikku MSF formaati teisendamiseks kasutasime programmi belvu (http://www.sacs.ucsf.edu/Documentation/seqsoftware/belvu.html). Käsurealt:

```
belvu –o MSF file.fas > file.msf
```

Järgnevalt ehitasime HMM mudelid, nii lokaalse kui globaalse:

```
hmmbuild –g globaalne_mudel.hmm file.msf
hmmbuild –f lokaalne_mudel.hmm file.msf
```

Globaalne mudel püüab joondada järjestused kogu nende ulatuses, lokaalne sisaldab ainult kõige sarnasemat lokaalset regiooni.
Mudelid kalibreerisime, et muuta otsingut tundlikumaks ja saada täpsemad E-väärtused (*expectation values*):

```
hmmcalibrate mudel.hmm
```

Saadud mudelitega otsisime nii kogu genoomi kui vastava arvu kõrgelt ekspresseerunud valkude vastu. Näiteks otsing kogu genoomi vastu näeb käsurealt välja järgmine:

```
hmmsearch mudel.hmm m52pep.fas
```

Saadud mudelitest genereerisime 100 juhuslikku järjestust, millele arvutasime hiljem ka hüdrofoobsusprofiili:

```
hmmemit –n 100 mudel.hmm
```

2.8 Hüdrofoobsusprofiilide arvutamine

Hüdrofoobsusprofiilide arvutamise aluseks võtsime Kyte-Doolittle hüdrofoobsuse skaala (Kyte and Doolittle, 1982) (tabel 1), kus aminohappel on seda positiivsem väärtus, mida hüdrofiilsem ta on, ja vastupidi, seda negatiivsem väärtus, mida hüdrofoobsem ta on. Uuritavaks piirkonnaks olid valkude esimesed 15 positsiooni. Analüüsi teostasime nii ühese kui kahese aknaga, mõlemal juhul ühese sammuga. Tulemusena saime graafilise pildi valgu N-terminuse hüdrofoobsusest. Erinevalt eespool kirjeldatud analüüsidest leiti hüdrofoobsusprofiilid lisaks kõikide valkude rühmale ja 40 kõrgelt ekspresseerunud valgu rühmale ka 60 ja 80 kõrgelt ekspresseerunud valgust koosnevatele rühmadele. Võrdluse eesmärgil arvutasime hüdrofoobsusprofiili veel *E.coli* juhuslikult võetud 40 valgule ja kuue bakteri (*Bacillus subtilis, Rickettsia prowazekii, Haemophilus influenzae, Mycobacterium tuberculosis, Helicobacter pylori* ja *Treponema pallidum*) 40 kõrgelt ekspresseerunud valgule.

A	-1.8	**L**	-3.8	**E**	3.5	**N**	3.5	**S**	0.8	**F**	-2.8	**V**	-4.2	**H**	3.2	**Q**	3.5	**W**	0.9
C	-2.5	**M**	-1.9	**G**	0.4	**P**	1.6	**T**	0.7	**I**	-4.5	**D**	3.5	**K**	3.9	**R**	4.5	**Y**	1.3

Tabel 1. Kyte Doolittle hüdrofoobsuse skaala (Kyte and Doolittle, 1982).

3 TULEMUSED

3.1 Analüüsimetoodika ülevaade.

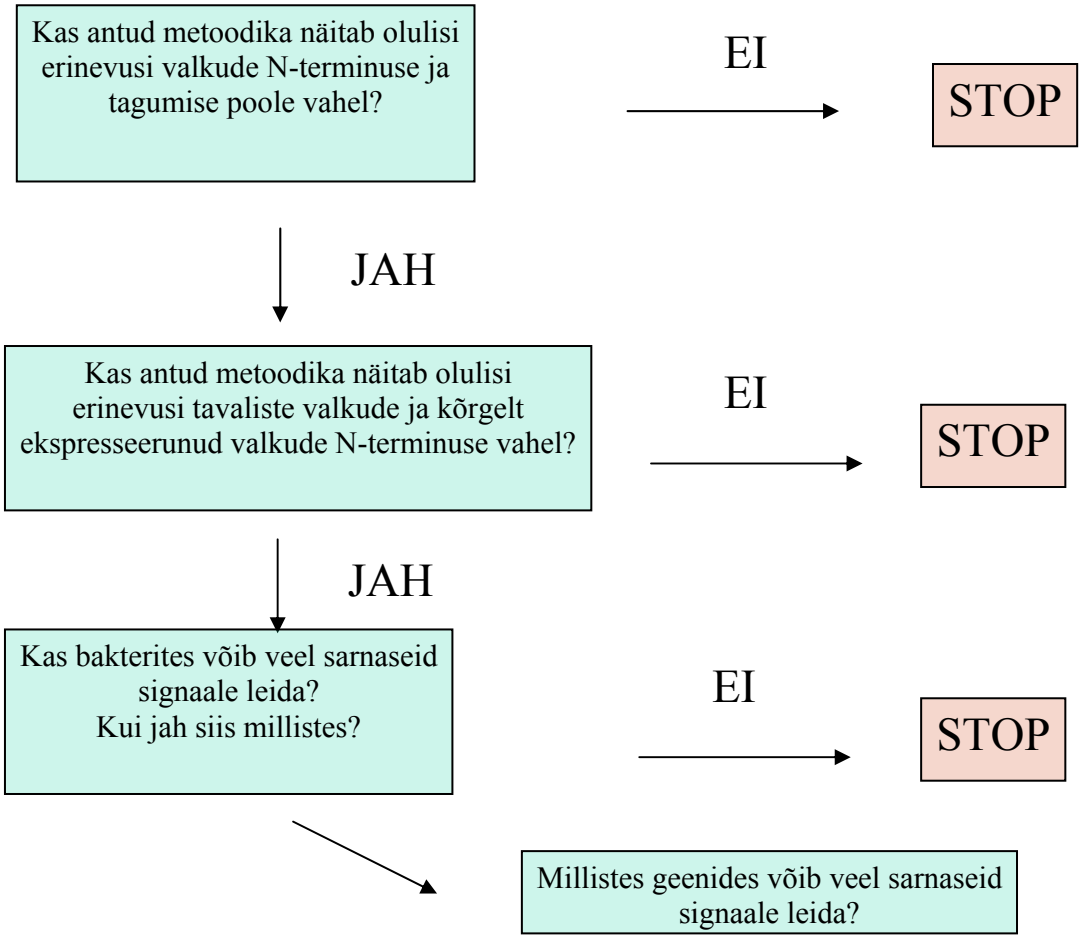
Töö ülesehitus põhines järgmistel loogilistel alustel (joonis 4):

Alustuseks otsida erinevusi valkude N-terminuse ja ülejäänud osa vahel. Erinevuste leidumise korral kontrollida erinevuste esinemist omakorda tavaliste valkude ja kõrgelt ekspresseerunud valkude N-terminuste vahel. Kui ka see uuring annab positiivse tulemuse, uurida, kas sarnaseid signaale esineb ka teistes bakterites peale *E. coli* ning kas geene on nende sarnaste signaalide alusel võimalik mingil viisil grupeerida.

Analüüsitud näitajateks olid:

1) nukleotiidide sagedus
2) sünonüümsete/mittesünonüümsete nukleotiidide sagedus

3) koodonite sagedus
4) aminohapete sagedus
5) dipeptiidide sagedus
6) HMM spetsiifiline muster
7) hüdrofoobsusprofiil

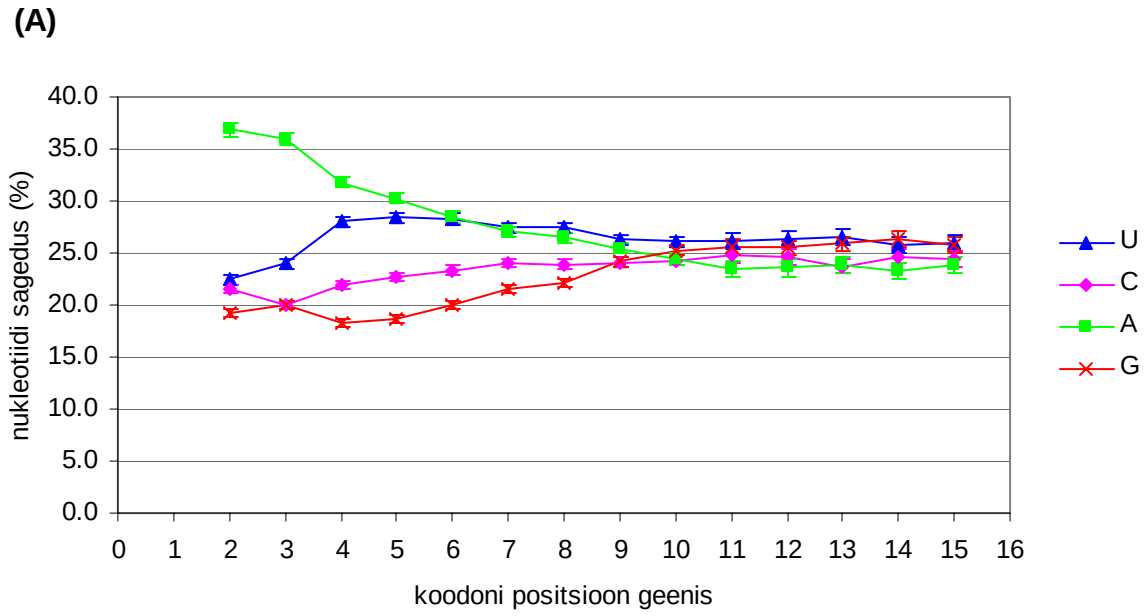


Joonis 4. Analüüsimetoodika ülevaade.

3.2 Nukleotiidide sageduse jaotus valkude N-terminuses

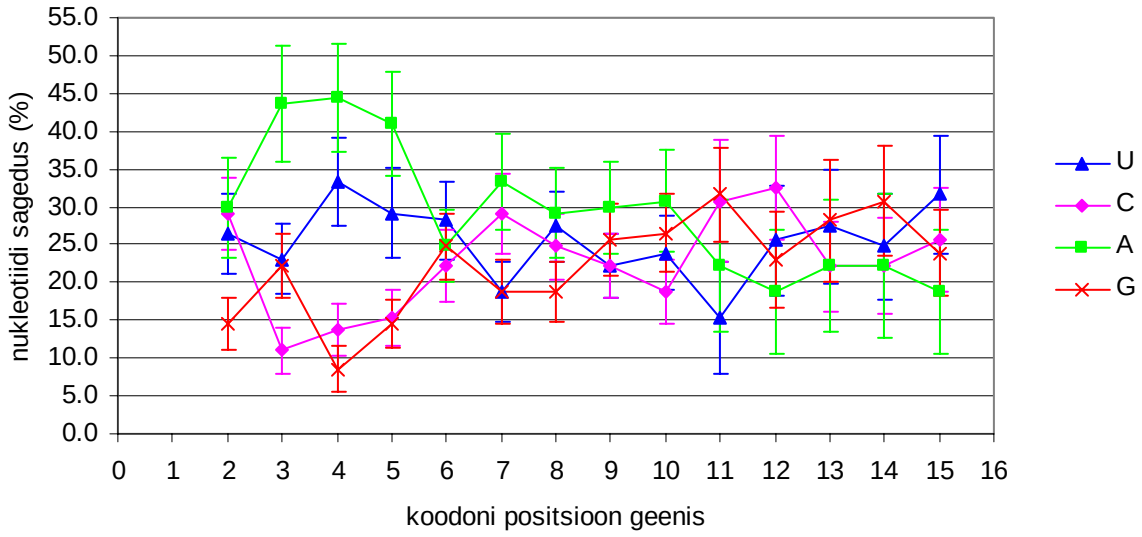
Esmalt analüüsisime nukleotiidide sagedust kõigi geenide (joonis 5A) ja kõrgelt ekspresseerunud geenide (joonis 5B) N-terminuses.

Kõrgelt ekspresseerunud geenide üldine nukleotiidide sagedus geeni esimeses 15 koodonis ei erine märkimisväärselt nukleotiidide sagedusest kõigi *E.coli* geenide puhul. Mõlemal juhul on enimesindatud nukleotiidideks A ja U ning vähemesindatuteks G ja C. Kõige sagedamini esineb A nukleotiid. Alates 10. positsioonist on juba märgata nukleotiidide sageduse ühtlustumist. Graafiku pikendamisel 50 positsioonini ühtlane sageduse jaotus jätkub (joonist pole näidatud), mis viitab sellele, et erinev nukleotiidide sageduse jaotus on iseloomulik just valgu N-terminusele.



Joonis 5A. Nukleotiidide üldine sagedus kõigis *E. coli* geenides.

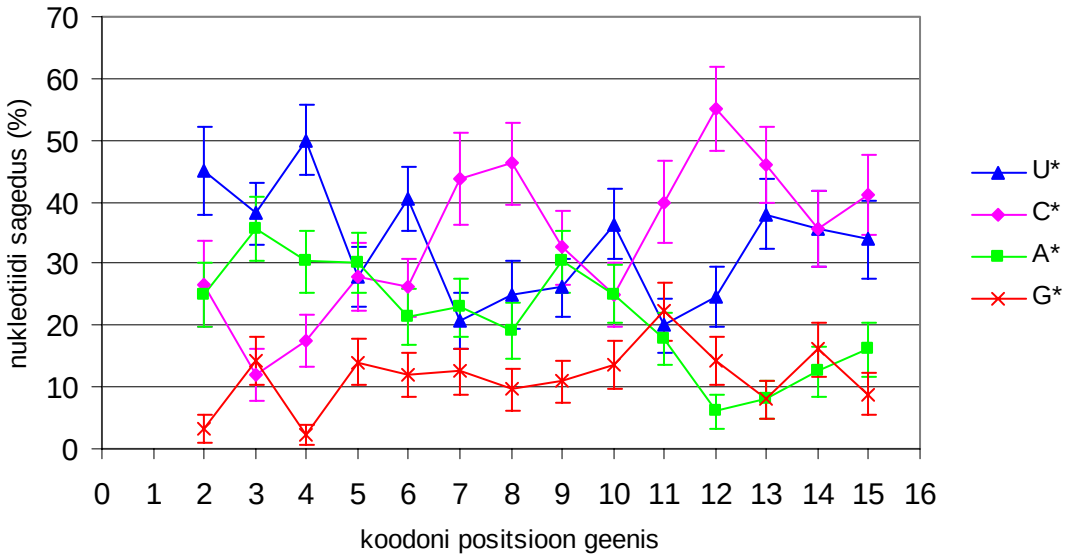
(B)



Joonis 5B. Nukleotiidide üldine sagedus *E.coli* 40 kõrgelt ekspresseerunud geenis.

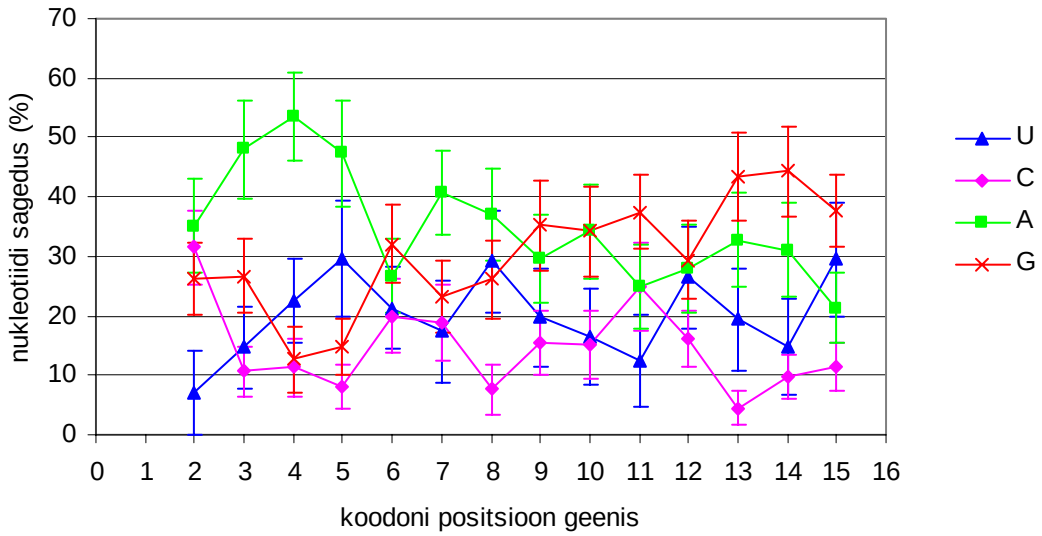
Sarnane analüüs eraldi kõrgelt ekspresseerunud geenides mittesünonüümsetes ja sünonüümsetes positsioonides asuvate nukleotiidide puhul andis A, G ja C nukleotiidide puhul sarnase mustriga graafiku. Erineb vaid nukleotiidi U sageduse jaotus, mis on sünonüümsete positsioonides enameelistatud kui A nukleotiid. Samas mittesünonüümsetes positsioonides kehtib sama seaduspärasus, mis nähtus üldisest nukleotiidide sageduse analüüsist: eelistatuim on A, mitte U (joonis 6A, joonis 6B). See tulemus lubab eeldada, et eelistatakse mitte võimalikult A-rikast N-terminust, vaid teatud aminohappeid, mida kodeerivad A-rikkad koodonid.

(A)



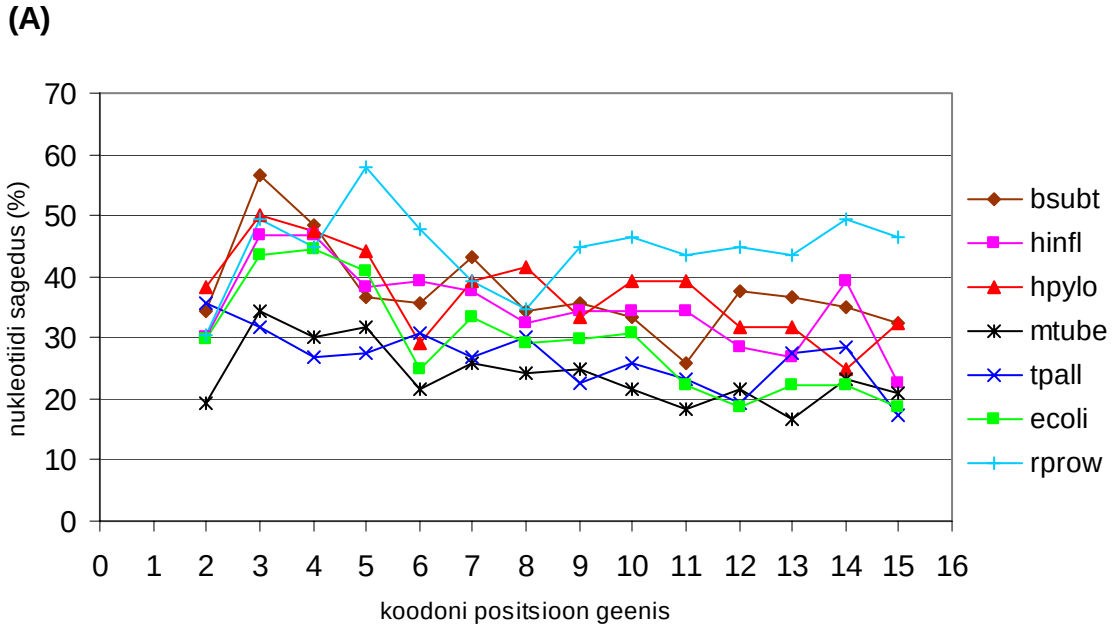
Joonis 6A. *E. coli* sünonüümsetes positsioonides asuvate nukleotiidide sagedus kõrgelt ekspresseerunud geenides.

(B)



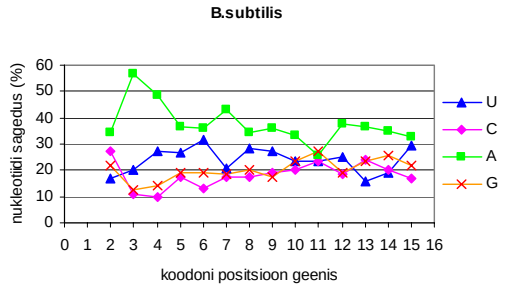
Joonis 6B. *E. coli* mittesünonüümsetes positsioonides asuvate nukleotiidide sagedus kõrgelt ekspresseerunud geenides.

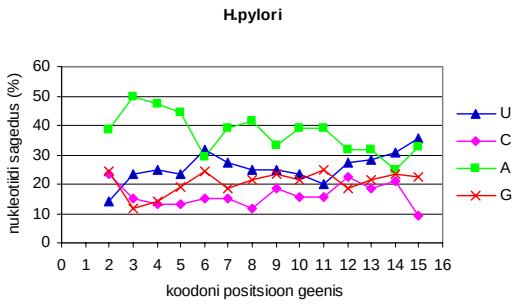
Üldine A nukleotiidi sagedus teistel bakteritel näitas sarnast jaotust *E. coli* nukleotiidide sageduse jaotusele *B.subtilis*'e, *H.pylori* ja *H. influenzae* puhul. Seda eelkõige esimeses 5-6 positsioonis (joonis 7A). Graafiku pikendamisel 50 koodonini sellist märkimisväärset sarnasust kaugemates positsioonides enam ei esinenud. Kõigil neil bakteritel on sarnaselt *E.coli*'le N-terminuses eelistatuim just nimelt A nukleotiid, millele järgneb sageduselt U nukleotiid. Eelistus on märgatavalt suurem esimeses 5 kuni 6 positsioonis, hiljem langeb ja jääb teatud vahemikku, nagu nähtus taas graafiku pikendamisest. *R. prowazekii*'l esineb samuti A eelistus, kuid jaotus on erinev A sageduse jaotusest eespool mainitud bakterites. *M.tuberculosis*'e ja *T.pallidum*'i puhul A nukleotiidi üleesindatust ei esine (joonis 7B). Kuna tegemist tundub olevat aminohappe eelistusega, võib joonisel 7A olev sarnasus tekkida antud geenide aminohappelisest konserveerumisest positsioonides 2-5.

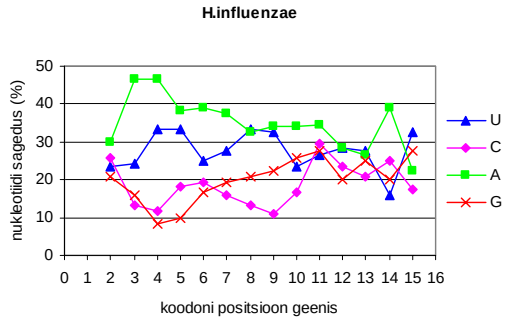


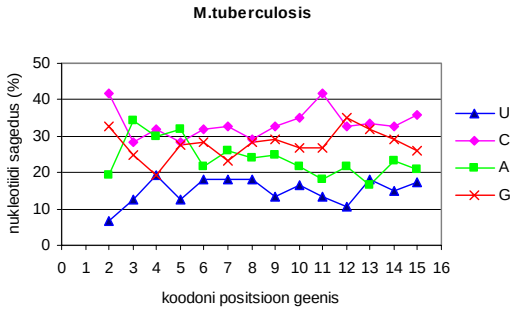
Joonis 7A. A nukleotiidi sagedus teistes bakterites kõrgelt ekspresseerunud geenides. (bsubt – *B.subtilis*, hinfl – H. *influenzae*, hpylo – *H. pylori*, mtube – *M. tuberculosis*, tpall – *T. pallidum*, ecoli – *E.coli*, rprow – *R.prowazekii*).

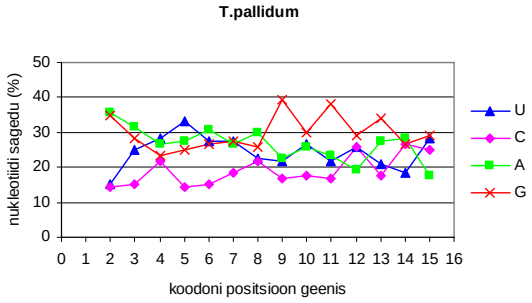
(B)

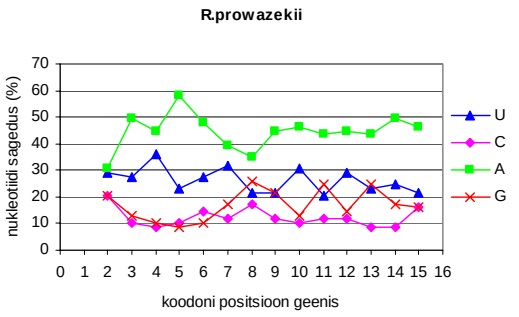












Joonis 7B. Nukleotiidide üldine sagedus teistes bakterites kõrgelt ekspresseerunud geenides

3.3 Koodonkasutus positsioonide kaupa

Joonisel 8 on välja toodud eelistatud ja mitte-eelistatud koodonid kõigis *E.coli* geenides geeni esimeses 10 positsioonis. Eelistatakse koodoneid, mis on A ja U nukleotiidide rikkad. Koodoni kolmest positsioonist vähemalt kahes paiknevad just need nukleotiidid. Samas mitte-eelistatud koodonid on G ja C rikkad. Kõrgelt ekspresseerunud geenides esines sama olukord (tabelit pole näidatud). Kriitilist väärtust 6.635 ületanud koodonite üldarv on kõrgelt ekspresseerunud geenides väiksem, mis on põhjustatud ilmselt valimi väiksusest, kuid esindatud koodonid ei erine eelistatud ja mitte-eelistatud koodonitest kõigis geenides. Samamoodi on eelistatud A ja U rikkad ning mitte-eelistatud G ja C rikkad koodonid.

2		3		4		5		6	
E	**ME**	**E**	**ME**	**E**	**ME**	**E**	**ME**	**E**	**ME**
AAA	CUG	AAG	CUG	UUA	GCG	AUA	GGC	AGA	GGC
AGU	GGC	AGA	GGC	AUA	GGU	AGA	GCG	AUU	GCG
ACA	GUG	ACA	CCG	ACA	CUG	ACA	GGU	UUA	GUG
AAU	GCC	AAA	GCG	AAA	GGC	UUA	CUG	AUA	CCG
UCU	GGU	AUA	GUG	AGA	GUG	ACU	CCG	ACA	GGU

7		8		9		10	
E	**ME**	**E**	**ME**	**E**	**ME**	**E**	**ME**
AUA	GCG	UUA	GCG	ACA	GUG	AGG	CCG
UUA	GUG	AGA	GGC	AGA	GCG	AUA	GAA
AGA	GGC	ACA	CCG	AUA	UAC	UUA	GAG
AGG	CUG	AUA	CAG	CUA	GAG	CUU	GCG
CUU	GGG	CUU	GAA	CUU	GGC	ACA	GUG

Joonis 8. *E. coli* koodonkasutus positsioonides 2-10 kõigis geenides. (E – eelistatud koodon, ME – mitte-eelistatud koodon). Kõigi geenide koodonite esinemissagedust vastavas positsioonis võrreldi koodonite esinemise sagedusega kõigi geenide ülejäänud positsioonides.

3.4 Aminohapete kasutus positsioonide kaupa

Joonistel 9A ja 9B on välja toodud eelistatud ja mitte-eelistatud aminohapped *E. coli* valkude esimeses 10 positsioonis. Kõigis valkudes on positsioonides 2, 3 ja 4 ülekaalukalt enimesindatud aminohape lüsiin. Nii suurt ülekaalu ei esinenud üheski teises positsioonis. Eelistatud aminohapete hulgas on rohkem hüdrofiilseid (Lys, Ser, Thr) ja mitte-eelistatute hulgas rohkem hüdrofoobseid aminohappeid. (Ala, Gly, Val, Leu), samuti Asp, Glu, mis on negatiivse laenguga aminohapped (joonis 9A). Võrreldes kõigi valkude aminohapete kasutust kõrgelt ekspresseerunud valkude aminohappe kasutusega, on näha, et üldjuhul eelistatakse samu aminohappeid (joonis 9B). Erinev on ainult kõrgelt ekspresseerunud valkude 6. positsioonis eelistatud aminohape tsüsteiin. Tekib küsimus, kas antud aminohapped on nendes positsioonides üle-esindatud ainult seetõttu, et nad on hüdrofiilsed või täidavad nad mingit muud biokeemilist või struktuurset funktsiooni. Kuna sünonüümsed koodonid ei näidanud A-eelistust, on tõenäoline, et surve on ikkagi aminohapetele, mitte koodonitele ja nukleotiididele. Seega võib teatud ettevaatlikkusega öelda, et A-üleesindatus on tagajärg, mitte põhjus.

Samas hüdrofiilsed aminohapped Ser ja Arg ei oma koodonite teises positsioonis A nukleotiidi ja ei ole A-rikkad, kuid on sellegipoolest üleesindatud. Alates 5. positsioonist kuulub eelistatud aminohapete hulka ka hüdrofoobne Ile. Seega otsustasime kontrollida, kas N-terminuses on võimalik leida a) kindlat aminohapete mustrit või b) hüdrofiilsusprofiili.

(A)

2		3		4		5		6	
E	**ME**	**E**	**ME**	**E**	**ME**	**E**	**ME**	**E**	**ME**
Lys	Gly	Lys	Gly	Lys	Ala	Ile	Gly	Ile	Gly
Ser	Val	Thr	Ala	Ile	Gly	Thr	Ala	Lys	Ala
Asn	Leu	Asn	Val	Ser	Asp	Lys	Glu	Thr	Asp
Thr	Tyr	Gln	Leu	Asn	Glu	Phe	Asp	Leu	Val
	Asp	Ile	Pro	Phe	Val	Asn	TER	Phe	Glu
7		**8**		**9**		**10**			
E	**ME**	**E**	**ME**	**E**	**ME**	**E**	**ME**		
Ile	Gly	Leu	Gly	Leu	Glu	Leu	Glu		
Leu	Glu	Ile	Ala	Ile	Tyr	Ser	TER		
Lys	Asp	Arg	Glu	Ser	TER		Asp		
Ser	Met	Lys	TER		Gly				

Joonis 9A. *E.coli* aminohapete kasutus positsioonides 2-10 kõigis valkudes (E – eelistatud aminohape, ME – mitte-eelistatud aminohape). Kõigi valkude aminohapete sagedust vastavas positsioonis võrreldi aminohapete sagedusega kõigi valkude ülejäänud positsioonides.

(B)

2	3	4	5	6
E	**E**	**E**	**E**	**E**
Ser	Lys	Ile	Lys	Cys
7	**8**	**9**	**10**	
E	**E**	**E**	**E**	
-	Leu	-	-	

Joonis 9B. *E.coli* aminohapete kasutus positsioonides 2-10 kõrgelt ekspresseeruvates valkudes (E - eelistatud aminohape). Kõrgelt ekspresseerunud valkude aminohapete sagedust vastavas positsioonis võrreldi aminohapete sagedusega kõrgelt ekspresseerunud geenide teistes positsioonides. Põhjus, miks kõrgelt ekspresseerunud valkudes vaid väga vähesed aminohapped ületasid meie poolt seatud 99 % tõenäosuse piiri, on ilmselt väheses andmehulgas.

3.5 Dipeptiidide kasutamine N-terminuses

Dipeptiidide sagedust mõõtsime valgu esimeses 10 positsioonis võrrelduna ülejäänud valgu positsioonidega.

Eelistatuimaks dipeptiidiks esimeses 10 positsioonis kõigis *E.coli* valkudes oli ülekaalukalt Cys-Cys. Sellele järgnesid eelistuselt paarid Trp-Trp, His-Cys, Met-Ser ja Met-Lys. Alaesindatud dipeptiidideks N-terminuses võrrelduna valgu ülejäänud osaga olid Trp-Thr, Met-Trp, Cys-Thr, Met-Cys ja Trp-Asp. Kuigi mitte-eelistatute paarides on esindatud vähemalt üks hüdrofoobne aminohape, on eelistatuimaks paariks kaks trüptofaani, mis on samuti hüdrofoobne aminohape. Seega otsest hüdrofiilsete dipeptiidide eelistust ja hüdrofoobsete mitte-eelistust N-terminuses ei saa väita.

3.6 HMM analüüsi põhimõte, eesmärk ja tulemused

Erinev hüdrofiilsete ja hüdrofoobsete aminohapete kasutus sai aluseks võimaliku aminohappelise mustri otsimiseks Markovi mudeli (HMM) abil valgu N-terminaalses osas (kasutasime programmi HMMER 2.3, http://hmmer.wustl.edu/). HMM kasutab aminohapete jaoks (või nukleotiidide jaoks) positsioonispetsiifilisi skoore ning modelleerib ka insertsioone ja deletsioone järjestuses. Meetodil on siiski omad limitatsioonid. Nimelt eeldab HMM, et kindla positsiooni samasus on sõltumatu kõigi teiste positsioonide samasusest.

Mudelite aluseks olid aminohapped kõrgelt ekspresseerunud valkude (vastavalt 25, 40 ja 60 valgu) positsioonides 2-12. Mudelid joondasime nii lokaalselt kui globaalselt.

Saadud mudelite alusel otsisime sarnaseid mustreid nii mudelite enda aluseks olnud 25, 40 ja 60 kõrgelt ekspresseerunud valgu seast kui ka kõigi *E. coli* valkude seast. Mudeli E väärtus näitab, mitu antud skooriga järjestust võiks leida antud suurusega andmebaasist tänu juhuslikule kokkusattumisele, st ilma bioloogilise tähenduseta. Antud andmete puhul oli nii globaalse joondamisega mudelil kui ka lokaalse joondamisega mudelil parimaks skoori E-väärtuseks 8. Oluliseks loetakse tavaliselt E-väärtused, mis on väiksemad kui 0.01. Tulemus viitab sellele, et kuigi kindlad aminohapped on valkude N-terminuses üle-esindatud, ei ole need aminohapped positsioonide kaupa konserveerunud. Seetõttu ei saa väita bioloogiliselt olulise mustri esinemist.

Samuti ei sarnanenud mudelist genereeritud 100 juhusliku järjestuse hüdrofoobsusprofiil teiste saadud profiilidega (vt ptk 3.7).

3.7 Hüdrofoobsuse jaotus valkude N-terminuses

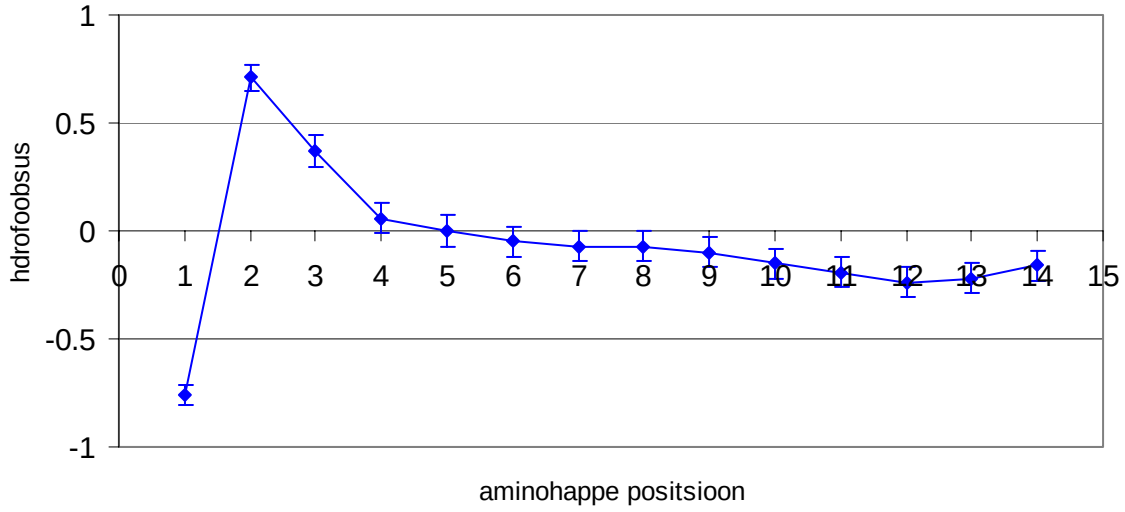
Lisaks HMM-idele arvutasime ka tavaliste ja kõrgelt ekspresseerunud valkude hüdrofoobsusprofiilid.

Kõigi valkude profiililt ilmneb, et valkude esimesed aminohapped on tõepoolest hüdrofiilsemad kui peale 10. positsiooni esinevad aminohapped (joonis 10A). 40 kõrgelt ekspresseerunud valgu hüdrofoobsusprofiil on täiesti erinev kõigi valkude hüdrofoobsusprofiilist. Nimelt esineb kõrgelt ekspresseerunud valkudel teatud tsüklilisus iga 3-4 aminohappe kaupa. See tsüklilisus säilub nii ühese kui ka kahese aknaga analüüsis (joonis 10B). Sarnane profiil esineb ka 60 kõrgelt ekspresseerunud valgu puhul ning on aimatav ka 80 kõrgelt ekspresseerunud valgu puhul, kuid kipub viimatinimetatud grupis juba hajuma ja võtma kõigi valkude profiili kuju (jooniseid pole näidatud). Algne oletus, et tegu võiks olla α-heeliksiga ei pea ilmselt paika, sest profiili pikendamisel 100 aminohappeni piikide esinemine jätkub, kuid mitte enam kindla sammuga. 40 juhusliku valgu puhul sarnast mustrit ei moodustunud (joonis 10B).

Samade bakterite kõrgelt ekspresseerunud valgud, mis andsid nukleotiidide sageduse puhul sarnase jaotuse *E. coli* nukleotiidide jaotusega, annavad ka sarnase hüdrofoobsusprofiili *E.coli* hüdrofoobsusprofiiliga. Mustri kokkulangevus on eriti märgatav esimeses 4-5 positsioonis. Saadud tulemus annab aluse oletada, et erinevatel bakteritel esinevad sarnased profiilid on sarnaste aminohapete, mida kodeerivad A-rikkad koodonid, eelistuse tagajärg (joonis 11).

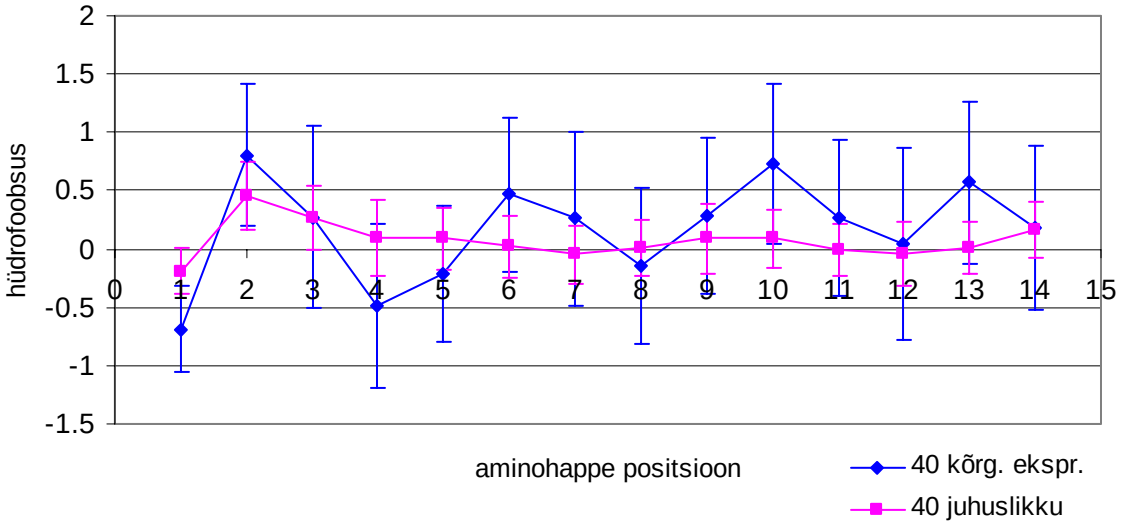
Ühese aknaga saadud profiilide standardvea analüüs jätab paraku võimaluse, et saadud profiil on juhuslik. Teiste sõnadega, meil ei ole piisavalt suur andmehulk, et tõestada tsüklilise hüdrofoobsusprofiili olemasolu.

(A)

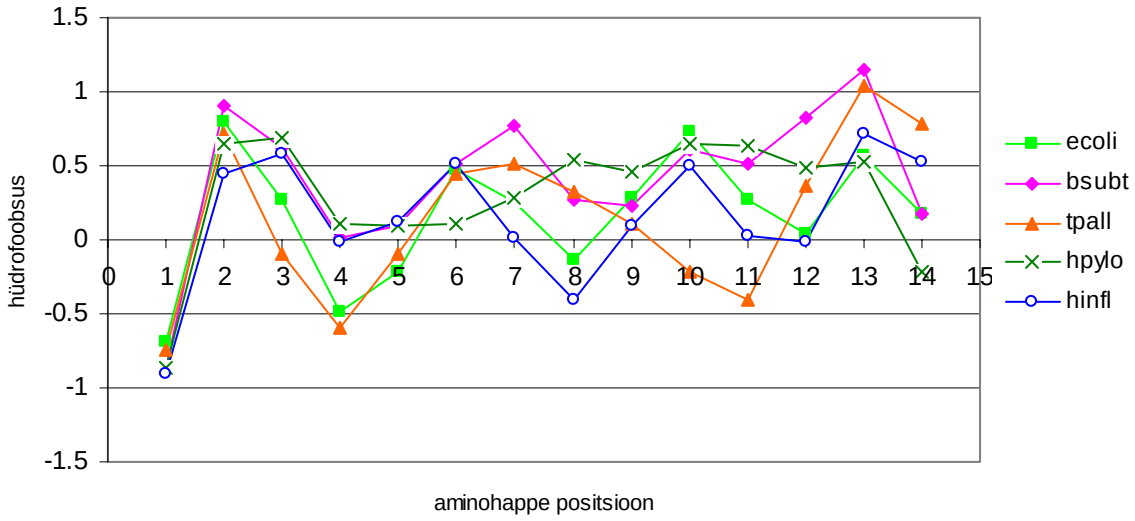


Joonis 10A. *E.coli* kõigi valkude hüdrofoobsusprofiil (kahene aken).

(B)



Joonis10B. *E. coli* 40 kõrgelt ekspresseerunud valgu hüdrofoobsusprofiil (kahene aken) ja juhuslikult valitud 40 *E.coli* valgu hüdrofoobsusprofiil (kahene aken).



Joonis 11. Hüdrofoobsusprofiilid teistes bakterites kõrgelt ekspresseerunud valkudes (kahene aken). (Joonise selguse huvides on välja jäetud *M. tuberculosis* ja *R. prowazekii*, mis andsid teistest liialt erineva profiili).

4. ARUTELU

Nukleotiidide analüüs nii kõigis kui ka kõrgelt ekspresseerunud geenides näitas A-nukleotiidi üleesindatust geenide algustes. Eriti tugev on see eelistus kõrgelt ekspresseerunud geenide 3-5 koodonis ning on lisaks *E.coli*'le konserveerunud ka paljudes teistes bakterites. Kuna sünonüümsed koodonid ei näidanud erilist A-eelistust, võib arvata, et see eelistus ei ole tingitud mitte survest nukleiinhapete tasandil, vaid surve toimib aminohapete tasandil.

HMM analüüsist nähtus, et esimesed 12 N-terminaalset aminohapet ei ole konserveerunud erinevate bakterite kõrgelt ekspresseerunud valkude vahel. Järelikult puudub selgelt eristuv primaarjärjestuse motiiv.

Kõrgelt ekspresseerunud valkude hüdrofoobsusprofiilil võib näha kindlaid mustreid, kuid pole piisavalt andmeid, et väita, et need on olulised. Samas on need mustrid konserveerunud erinevate vabalt elavate bakterite vahel, mis annab täiendavaid tõendeid profiili olulisusest.

Järgnevalt tuleks uurida, kas valke saab hüdrofoobsusprofiilide järgi klasterdada. Kuna HMM sai tehtud esimesele 12 aminohappele, kuid eeldatav aminohappeline konserveerumine hõlmas vaid 4 positsiooni, võis meie mudel seda konserveerumist mitte märgata. Seetõttu tuleks edaspidi proovida HMM-i teha ka aminohapetele positsioonides 2-5. Mustreid tuleks otsida ka joondamata kujul. Sobivad programmid selleks on:

TEIRESIAS (http://cbcsrv.watson.ibm.com/Help/aboutTspd.htm),

SPEXS (http://ep.ebi.ac.uk/EP/SPEXS/),

MEME (http://meme.sdsc.edu/meme/website/intro.html).

Kõrvale ei saa jätta võimalust, et ka valgu sekundaarstruktuuril võib siin teatud roll olla.

Dipeptiide analüüsi saab samuti detailsemalt läbi viia, kuna antud juhul ei olnud dipeptiidide sagedus arvutatud positsioonispetsiifiliselt, vaid esimeses 10 positsioonis kokku. Lisaks tuleks dipeptiidide sagedust uurida ka kõrgelt ekspresseerunud valkudes. Tähelepanu tuleks pöörata eelkõige konserveerunud positsioonidele 2-5.

5. KOKKUVÕTE

Käesolevas töö eesmärgiks oli järjestuse analüüsi abil selgitada, millised bioloogilised signaalid esinevad valkude N-terminuses, võrrelduna teiste valgu piirkondadega ning millised neist signaalidest on iseloomulikud kõrgelt ekspresseerunud valkudele.

Analüüsisime *E.coli* geenide nukleotiidset järjestust startkoodonist allavoolu ja valkude aminohappelist järjestust valgu N-terminuses esimeses 15 positsioonis. Tulemuste võrdlemine geeni ülejäänud positsioonidega näitas A nukleotiidide üleesindatust koodonites 2-5 ning see eelistus oli eriti tugev kõrgelt ekspresseerunud geenide positsioonides 3-5. Lisaks *E.coli*'le on selline eelistus konserveerunud ka teistes bakterites. Sünonüümsetes positsioonides esinevate nukleotiidide analüüs ei näidanud aga A eelistust. Seega on võimalik, et bioloogiline surve on suunatud aminohapetele, mitte koodonitele ja nukleotiididele. Kõrgelt ekspresseerunud valkude hüdrofoobsusprofiilid olid samuti erinevates bakterites sarnased, mis annab tõendeid konserveerunud hüdrofoobsusprofiili esinemisest, kuigi me ei saa andmete vähesuse tõttu kindlalt väita, et konserveerumine on statistiliselt oluline.

6. SUMMARY

The aim of the current paper was to examine which biological signals appear in the N-terminus compared to other areas of proteins, and which of those signals are specific to highly expressed proteins. The comparison of the results with other positions of the gene showed the preference of A nucleotide in codons 2-5. This preference was particularly strong in positions 3-5 in highly expressed genes. This kind of preference was also conserved in other bacteria in addition to *E.coli.* The analysis of nucleotides in synonymous positions didn't show the preference of A. Therefore is possible biological pressure towards amino acids instead of codons and nucleotides. The hydrophobicity profiles of highly expressed proteins were also similar between different bacteria, giving the proof of conserved hydrophobicity profile. Neverthless we can't declare the statistical importance of this conservation because of the insufficiency of the data.

7. VIITED

Andersson, S.G.E. and Kurland, C.G. (1990) Codon preferences in free-living microorganisms. *Microbiol. Rev*., **54**, 198-210.

Bade EG, Gonzalez NS, Algranati IS. (1969) Dissociation of 70S ribosomes: some properties of the dissociating factor from Bacillus stearothermophilus and Escherichia coli. *Proc Natl Acad Sci U S A*, **64,** 654-60.

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

Berks, B.C. (1996) A common export pathway for proteins binding complex redox cofactors? *Mol. Microbiol*., **22**, 393-404.

Bogsch, E., Brink, S., Robinson, C. (1997) Pathway specificity for a ΔpH dependent precursor thylakoid lumen protein is governed by a 'Sec-avoidance' motif in the transfer peptide and a 'Sec-incompatible' mature protein. *EMBO J*., **16**, 3851-3859.

Capecchi, M. R. (1967) Polypeptide chain termination in vitro: isolation of a release factor. *Proc. Natl. Acad. Sci. U. S. A.* **58,** 1144-1151.

Chen, H., Bjerkness, M., Kumar, R., Jay, E. (1994) Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translational initiation codon of Escherichia cli mRNAs. Nucl. Acid Res., **22**, 4953-4957.

Cristóbal, S., de Gier, J.-W., Nielsen, H., von Heijne, G. (1999) Competition between Sec- and TAT-dependent protein translocation in *Escherichia coli*. *EMBO J.,* **18**, 2982-2990.

Dalbey, R.E. **and Robinson, C.** (1999) Protein translocation into and across the bacterial plasma membrane and the plant thylakoid membrane. *Trends Biochem. Sci*., **24**, 17-22.

Dalboge, H.S., Bayne, S., Pedersen, J. (1990) *In vivo* processing of N-terminal methionine in *Escherichia coli*. *FEBS Letters*, **266**, 1-3.

Draper, D.E. (1996) In Curtiss, R., Ingraham, J.L., Lin, E.C.C., Brooks Low, K., Magasanik, B., Reznikoff, W.S., Riley, M., Schaechter, M., Umbarger, H.E. (eds), *Escherichia coli* and S*almonella*: Cellular and Molecular Biology. ASM Press, Washington, DC, 902-908.

Duong, F., Eichler, J., Price, A., Leonard, M.R., Wickner, W. (1997) Biogenesis of the gram-negative bacterial envelope. *Cell*, **91**, 567-573.

Esposito, D, Fey, J.P., Eberhard, S., Hicks, A.J., Stern, D.B. (2003) *In vivo* evidence for the prokaryotic model of extended codon-anticodon interaction in translation initiation. *EMBO J.*, **22**, 651-656.

Firpo, M.A. **and Dahberg, A.E.** (1998) The importance of base pairing in the penultimate stem of *Escherichia coli* 16S rRNA for ribosomal subunit association. *Nucl. Acid Res*., **26**, 2156-2160.

Ganoza, M. C. (1966) Polypeptide chain termination in cell-free extracts of *E. coli*. *Cold Spring Harbor Symp. Quant. Biol.* **31,** 273-278

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

Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol*., **1**, 13-34.

Izard, J.W. **and Kendall, D.A**., (1994) Signal peptides: exquisitely designed transport promoters. *Mol. Microbiol*., **13**, 765-773.

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

Jin H, Bjornsson A, Isaksson LA (2002) Cis control of gene expression in E.coli by ribosome queuing at an inefficient translational stop signal. *EMBO J*., **21,** 4357-4367.

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

Karlin, S. **and Mrazek, J.** (2000) Predicted Highly Expressed Genes of Diverse Prokaryotic Genomes. *Journal of Bacteriology*, **182**, 5238-5250.

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

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

Kutay, U, Ahnert-Hilger, G., Hartmann, E., Wiedenmann, B., Rapoport, T.A. (1995) Transport route for synaptobrevin via a novel pathway of insertion into the endoplasmic reticulum membrane. *EMBO J,* **14**, 217-223.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol*., **157**, 105-132.

Mazel, D., Pochet, S., Marliere, P. (1994) Genetic characterization of polypeptide deformylase, a distinctive enzyme of eubacterial translation. *EMBO J*., **13**, 914-923.

Meinnel, T., Mechulam, Y., Blanquet, S. (1993) Methionine as translationstart signal: a review of the enzymes of the pathway in the *Escherichia coli. Biochimie*, **75**, 1061-1075.

Miller, C.G., Strauch, K.L., Kukral, A.M., Miller, J.L., Wingfield, P.T., Mazzei, G.J., Weren, R.C., Graber, P., Movva, N.R. (1987) N-terminal methionine-specific peptidase in Salmonella typhimurium. *Proc. Natl. Acad. Sci. USA*, **84**, 2718-2722.

Moerschell, R.P., Hosokawa, Y., Tsunasawa, S., Sherman, F. (1990) The specificities of yeast methionine aminopeptidase and acetylation of amino-terminal methionine *in vivo. J. Biol. Chem*., **265**, 19638-19643.

Mottagui-Tabar, S. and Isaksson, L.A. (1997) Only the last amino acids in the nascent peptide influence translation termination in Escherichia coli genes. *FEBS Letters*, **414**, 165-170.

Nilsson, I., Whitley, P., von Heijne, G. (1994) The COOH-terminal ends of internal signal and signal-anchor sequences are positioned differently in the ER translocase. *J. Cell Biol*., **126**, 1127-1132.

O'Connor, M., Asai, T., Squires, C.L., Dahlberg, A.E. (1999) Enhancement of translation by the downstream box does not involve base pairing of mRNA with penultimate stem sequence of 16S rRNA. *Proc. Natl. Acad. Sci*. USA, **96**, 8973-8978.

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

Sakai, H., Imamura, C., Osada, Y., Saito, R., Washio, T., Tomita, M. (2001) Correlation Between Shine-Dalgarno Sequence conservation and Codon Usage of Bacterial Genes. *J. Mol. Evol*., **52**, 164-170.

Schatz, P.J., **and Beckwith, J.** (1990) Genetic analysis of protein export in *Escherichia coli. Annu. Rev. Genet*., **24**, 215-248.

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

Schneider, T.D., Stormo, G.D., Gold, l., Ehrenfeucht, A. (1986) Information content of binding sites on nucleotide sequences. J. Mol. Biol., **188**, 415-431.

Scolnick, E., Tompkins, R., Caskey, T., and Nirenberg, M. (1968) Release factors differing in specificity for terminator codons. *Proc. Natl. Acad. Sci. U. S. A.*, **61,** 768-774.

Sharp,P.M. **and Li,W.H**. (1987) The codon adaptation index – a measure of directional synonymous codon usage bias and its potential applications., *Nucleic Acids Res*., **15**, 1281-1295.

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

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

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

Solbiati, J., Chapman-Smith, A., Miller, J.L., Miller, C.G., Cronan Jr, J.E. (1999) Processing of the N Termini of Nascent Polypeptide Chains Requires Deformylation Prior to Methionine Removal. *J. Mol. Biol*., **290**, 607-614.

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

Stenström, C. M., Holmgren, E., Isaksson, L.A. (2001) Cooperative effects by the initiation codon and its flanking regions on translation initiation. *Gene*, **273**, 259-265.

Stenström, C.M. **and Isaksson, L.A.** (2002) Influences on translation iniiation and early elongation by the messenger RNA region flanking the initiation codon at the 3' side. *Gene*, **288**, 1-8.

Stormo, G.D., Schneider, T.D., Gold, L.M. (1982) Characterization og translational initiation sites in E.coli. Nucl. Acid Res., **10**, 2971-2996.

Tenson, **T. and Ehrenberg, M.** (2002) Regulatory Nascent Peptides in the Ribosomal Tunnel. *Cell*, **108**, 591-594.

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

Valent, Q.A., Scotti, P.A., High, S., de Gier, J.W., von Heijne, G., Lentzen, G., Wintermayer, W., Oudega, B., Luiring, J. (1998) The *Escherichia coli* SRP and SecB targeting pathways converge at the translocon. *EMBO J*., **17**, 2504-2512.

vanEtten, W.J. **and Janssen, G.R.** (1998) An AUG initiation codon, not codon-anticodon complementarity, is required for the translation of unleadered mRNA in *Escherichia coli*. *Mol. Microbiol*., **27**, 987-1001.

Varshavsky, A. (1996) The N-end rule: Functions, mysteries, uses. *Biochemistry*, **93**, 12142-12149.

von Heijne, G. (1985) Structural and thermodynamic aspects of the transfer of proteins into and across membranes, *Academic Press*.

von Heijne, G. (1986) Towards a comparative anatomy of N-terminal topogenic protein sequences. *J. Mol. Biol*., **189**, 2239- 242.

von Heijne, G. (1988) Transcending the impenetrable: how proteins come to terms with membranes. *Biochim. Biophys. Acta*, **947**, 307-333.