DIVERSITÄTSGENERIERENDE RETROELEMENTE - IDENTIFIKATION, KLASSIFIZIERUNG, PHYLOGENIE UND *IN VITRO*-FUNKTIONSANALYSEN

vom Fachbereich Biologie der Universität Kaiserslautern
zur Verleihung des akademischen Grades
„Doktor der Naturwissenschaften“ genehmigte Dissertation

von
Thomas Schillinger
aus Völklingen

Betreuerin: Frau Jun.-Prof. Dr. rer. nat. Nora Zingler
Korreferentin: Frau Prof. Dr. rer. nat. Regine Hakenbeck

Datum der wissenschaftlichen Aussprache: 8. August 2013

D 386

erschienen 2013 in Kaiserslautern

Die vorliegende Arbeit wurde in der Zeit von September 2009 bis März 2013 in der Arbeitsgruppe für Molekulare Genetik der Technischen Universität Kaiserslautern unter der Leitung von Frau Juniorprofessor Dr. rer. nat. Nora Zingler angefertigt.

Vorsitzender der Prüfungskommission: Herr Professor Dr. rer. nat. Stefan Kins

Erste Gutachterin: Frau Junior-Professor Dr. rer. nat. Nora Zingler

Zweite Gutachterin: Frau Professor Dr. rer. nat. Regine Hakenbeck

Für meine Eltern

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

(v/v)	Volumenkonzentration
(w/v)	Massenkonzentration
*	Stop-Codon
°C	Grad Celsius
µg	Mikrogramm
µL	Mikroliter
µM	mikromolar
A	Adenin/Alanin
APS	Ammoniumpersulfat
Atd	Accessory tropism domain
Avd	Accessory variation domain
B	Cytosin, Guanin oder Thymin ("*nicht Adenin*")
BLAST	Basic Local Alignment Search Tool
bp	Basenpaar
C	Cytosin/Cystein
CBD	Chitin Binding-Domain
Ci	Curie
cpm	Counts per Minute
D	Aspartat
DGR	Diversitätsgenerierende Retroelemente
DiGReF	Diversity-generating retroelement finder
DNA	Deoxyribonucleinsäure
dNTP	Deoxyribonucleotidtriphosphat
ds	doppelsträngig
DSP	Di (N-succinimidyl)-3,3'-dithiopropionat
DTT	Dithiothreitol
E	Glutamat
EDTA	Ethylendiamintetraacetat
E-Wert	Erwartungswert, *expect value*
F	Phenylalanin
FGE	Formyglycin-generierendes Enzym
G	Guanin/Glycin
g	Gramm oder Erdbeschleunigung
GST	Glutathion-S-Transferase
h	Stunden
His	Histidin
HIV	Humanes Immundefizienzvirus
I	Isoleucin
IMH	Initiator of Mutagenic Homing
IPTG	Isopropyl-β-D-galactopyranosid
K	Lysin
kbp	Kilobasenpaare
kDa	Kilodalton
L	Leucin oder Liter
LB	Lysogeny Broth
LINE1/L1	Long Interspersed Element 1
M	Methionin oder Molar
mCi	Millicurie
mg	Milligramm

min	Minuten
mL	Milliliter
mM	Millimolar
M-MLV	Moloney Murines Leukämievirus
Mtd	Major tropism determinant
N	Adenin, Cytosin, Thymin oder Guanin ("*any base*") oder Asparagin
ng	Nanogramm
Ni-NTA	Nickel-Nitrilotriessigsäure
NJ	Neighbour-Joining
nM	Nanomolar
nm	Nanometer
NTP	Ribonucleotidtriphosphat
OD	Optische Dichte
ORF	Open Reading Frame
P	Prolin
PAA	Polyacrylamid
PAGE	Polyacrylamidgelelektrophorese
PCR	Polymerasekettenreaktion
PHI-BLAST	Pattern Hit Initiated BLAST
PNK	Polynucleotidkinase
PSI-BLAST	Position-Specific Iterated BLAST
Q	Glutamin
R	Arginin
RNA	Ribonucleinsäure
RT	Reverse Transkriptase oder Raumtemperatur
s	Sekunden
S	Serin
SCL	Subcellular Localization
SDS	Natriumdodecylsulfat
ss	Einzelsträngig
T	Thymin/Threonin
TAE	Tris/Acetat/EDTA
TB	Terrific Broth
TBE	Tris/Borat/EDTA
TEMED	N,N,N',N'-Tetramethylethylendiamin
TPRT	Target-Primed Reverse Transcription
TR	Template Repeat
Tris	2-Amino-2-hydroxymethyl-propane-1,3-diol
U	Uracil
V	Valin oder Volt
VR	Variable Region
W	Tryptophan oder Watt
x	Beliebige Aminosäure
Y	Tyrosin

Zusammenfassung

Diversitätsgenerierende Retroelemente (DGRs) stellen einen neuen Typus Retroelement dar, die gezielt einen Teil einer codierenden Sequenz des Wirtsgenoms über einen Copy and Replace-Mechanismus hypermutieren und somit zur Erzeugung biologischer Diversität beitragen können. Trotz dieser einzigartigen Eigenschaften und dem potentiellen Wert dieser Elemente für Industrie und Forschung konzentrierten sich seit der Beschreibung des ersten DGRs vor über zehn Jahren die meisten Publikationen auf mechanistische Eigenschaften des Prototypen aus dem *Bordetella* Bakteriophagen. Allerdings sind zahlreiche Fragen zur Funktionsweise dieser Elemente noch immer ungeklärt. Ebenso wurden bisher extensivere, vergleichende Studien, die weitere Vertreter dieser Elemente berücksichtigen, noch nicht durchgeführt.

Die vorliegende Dissertation leistet einen wichtigen Beitrag zum tieferen Verständnis diversitätsgenerierender Retroelemente. Das eigens für diesen Zweck konzipierte Programm DiGReF erlaubte eine umfassende Analyse der Bestände öffentlicher Datenbanken auf DGRs in sequenzierten Genomen. Mit Hilfe dieser Daten konnten weitere Aspekte dieser Elemente aufgeklärt werden, die eine Analyse ihrer Verteilung, ihrer phylogenetischen Beziehungen, ihrer Struktur und eine Charakterisierung der einzelnen Elemente einer DGR-Kassette umfassten. So konnte gezeigt werden, dass das zuvor für wenige Elemente beschriebene Merkmal der Adeninsubstitution eine gemeinsame Eigenschaft aller DGRs ist, während keine C-, T- oder G-Substitionen auftreten. Ebenso fanden sich erste Belege dafür, dass die beiden essentiellen Elemente Template Repeat und reverse Transkriptase nicht notwendigerweise ein gemeinsames Transkript besitzen. Außerdem konnte erstmalig die Gruppe der weitgehend uncharakterisierten akzessorischen Proteine umfassender beschrieben und ein Consensusmotiv ermittelt werden. Für künftige Studien werden die DiGReF-Software und die Ergebnisse dieser Arbeit von grundlegender Bedeutung sein.

Der zweite Teil dieser Arbeit fokussierte sich auf die experimentelle Charakterisierung zweier Kernkomponenten von DGRs, der reversen Transkriptase und den akzessorischen Proteinen. Während die Aufreinigung einer DGR-assoziierten reversen Transkriptase noch weitere experimentelle Arbeiten erfordern wird, konnte das akzessorische Protein Alr3496 aus der Blaualge *Nostoc* sp. PCC 7120 erfolgreich in rekombinanter Form aufgereinigt werden. Es konnte weiterhin gezeigt werden, dass Alr3496 diverse Nucleinsäuresubstrate bindet, und in der Lage ist, die Hybridisierung von komplementären DNA-Strängen zu katalysieren. Dies legt nahe, dass akzessorische Proteine aus DGR-Elementen eine Rolle als Nucleinsäurechaperone übernehmen.

Summary

Diversity-generating retroelements (DGRs) constitute a new type of retroelement, which is able to diversify a distinct section of a host open reading frame by using a copy and replace mechanism; biological diversity is created in the process. On the contrary, classic retroelements utilize a copy and paste mechanism and can potentially disrupt coding or regulatory sequences of the host genome by insertion.

The first description of a DGR was published about one decade ago; since then, several studies have focused on mechanistical properties of this prototypical element from the *Bordetella* bacteriophage. However, various questions concerning these elements are still unanswered. Moreover, more comprehensive studies that compare elements from various organisms have not been conducted so far.

This thesis provides a significant contribution to a deeper understanding of diversity-generating retroelements. An exclusively-designed program called Diversity-Generating Retroelement Finder (DiGReF) allowed for a comprehensive analysis of public databases for the presence of DGRs in sequenced genomes. Using these data, more advanced studies were conducted that included analysis of distribution, phylogeny, structure and characterization of every component in DGR cassettes. For instance, it was revealed that adenine substitution, which was reported previously for several elements, is a common hallmark of all DGRs. Also, first evidence was found that the two core components of DGRs, template repeat and reverse transcriptase, do not necessarily share a common RNA transcript. Additionally, the largely uncharacterized group of accessory proteins was more closely investigated, leading to the identification of a previously unreported consensus motif. Future studies will benefit from the DiGReF software, and the results described herein.

In the second part of this thesis, two key components of DGRs – the reverse transcriptase and an accessory protein – were assessed experimentally for further characterization. While more work has to be invested in the purification of a DGR reverse transcriptase, successful recombinant expression and purification of the accessory protein Alr3496 from the cyanobacterium *Nostoc* sp. PCC 7120 was achieved, and binding of various nucleic acid substrates was observed in filter binding assays. Additionally, further experiments demonstrated that Alr3496 is able to catalyze hybridization of complementary DNA strands, which suggests that DGR accessory proteins might act as nucleic acid chaperones.

I. Einleitung

Bis weit in das 20. Jahrhundert hinein herrschte in der Biologie Uneinigkeit darüber, ob die genetische Information über Deoxyribonucleinsäure (*deoxyribonucleic acid*, DNA) oder Proteine weitergegeben wird. Schließlich konnte in einer Reihe von Experimenten zwischen 1944 und 1961 demonstriert werden, dass DNA Träger der Erbinformation ist, die Weitergabe an nachfolgende Generationen jeweils semikonservativ erfolgt, und ein Triplettcode die Information über die Primärstruktur von Proteinen beinhaltet (Avery et al., 1944; Hershey and Chase, 1952; Nirenberg and Matthaei, 1961; Watson and Crick, 1953). Dass das Genom eines Organismus nicht notwendigerweise einen starren, unveränderlichen Zustand besitzt, sondern vielmehr diversen Umbauten und Modifikationen unterworfen ist, war bereits zu diesem Zeitpunkt mit der Theorie und Entdeckung des Crossing Overs durch Thomas Hunt Morgan und Barbara McClintock nicht unbekannt (Creighton and McClintock, 1931). McClintock war es auch, die in den 50er Jahren Studien vorstellte, in denen sie die Existenz mobiler Genomabschnitte postulierte (McClintock, 1956), was gleichbedeutend war mit tiefgreifenden topologischen Veränderungen des Erbgutträgers. Es dauerte jedoch weitere 20 Jahre, bis ihre Arbeit Anerkennung fand, und das Konzept von mobilen genetischen Elementen (MGEs) von der gängigen Lehrmeinung akzeptiert wurde (Starlinger and Saedler, 1972). Heute ist bekannt, dass MGEs wichtige biologische Funktionen erfüllen, beispielsweise in der Erzeugung genetischer Vielfalt und der Entstehung neuer Gene (Chenais et al., 2012; Feschotte, 2008), oder in der Beteiligung an der epigenetischen Regulation von Wirtsgenomabschnitten (Chueh et al., 2009).

Bei MGEs handelt es sich um DNA-Abschnitte von wenigen hundert bis mehreren tausend Basenpaaren (bp) Länge, die die Fähigkeit besitzen, ihre Position und/oder ihre Zahl im Genom zu verändern. Ein markantes Merkmal dieser Elemente ist, dass der Transpositionsprozess die Integrität des Wirtsgenoms weitgehend ignoriert; mobile genetische Elemente sind sozusagen nur an der eigenen Weiterverbreitung interessiert, weswegen sie im Englischen oft als „selfish“ bezeichnet werden (Doolittle and Sapienza, 1980; Orgel and Crick, 1980; Werren, 2011). Im besten Falle erfolgt der Einbau des Elements an einer neutralen Stelle im Wirtsgenom, die keine codierende oder regulierende Aufgabe besitzt. Ist jedoch ein funktionaler Genomabschnitt Ziel eines MGEs, so hat dies in der Regel negative Auswirkungen für das Wirtsgenom, und somit für den beherbergenden Organismus als Ganzes (Kazazian, 1998; Lee et al., 2012). Ebenso wurde beobachtet, dass in höheren Eukaryoten selbst die Insertion in Genomabschnitte, welche nur in der Nähe von aktiven Abschnitten liegen, eine Störung dieser Abschnitte verursachen kann (Kuehnen et al., 2012). Grund hierfür sind epigenetische Mechanismen, welche durch die Präsenz eines MGEs eine geringere Transkriptionsrate

der umliegenden Sequenzen bewirken können. Grundsätzlich zeigen die bisher bekannten MGEs eher eine geringe Spezifität hinsichtlich der Wirtsgenomsequenzen, in die sie sich integrieren. Lediglich einzelne Elemente wie R1 (*Drosophila melanogaster*) und R2 (*Bombyx mori*) inserieren spezifisch in Loci, die für ribosomale RNAs codieren (Jakubczak et al., 1990). Andere Elemente wiederum, wie z. B. TRAS1/SART1 sind spezifisch für Telomersequenzen in *B. mori* (Takahashi et al., 1997), während Prophagen oft in der Nähe oder innerhalb von tRNA-Genen inserieren (Canchaya et al., 2004). Die Mehrzahl mobiler genetischer Elemente jedoch stellt weit weniger definierte Ansprüche an ihre jeweiligen Zielsequenzen. L1-Elemente benötigen lediglich die Sequenz TTTA (Feng et al., 1996), während einige DNA-Transposons wie das Tc1/*mariner*-Element in der Regel nur noch ein AT-Dinucleotid im Wirtsgenom zur Insertion nutzen (Plasterk et al., 1999). Noch geringere Bedingungen erfordern scheinbar Alu-Elemente: hier vollzieht sich der Einbau weitgehend zufällig, und gleichmäßig über das Wirtsgenom verteilt, mit einer gewissen Bevorzugung von Sequenzen in der Nähe ähnlicher Elemente oder genischer Regionen (Wagstaff et al., 2012).

Diversitätsgenerierende Retroelemente (DGRs) stellen eine faszinierende Ausnahme unter MGEs dar (Doulatov et al., 2004; Liu et al., 2002; Medhekar and Miller, 2007). Sie können einen Teil eines codierenden Abschnitts der Wirts-DNA durch eine hypermutierte Variante ersetzen, indem sie eine exakt definierte Zielsequenz verwenden. Das resultierende Protein wird somit in seinen biochemischen Eigenschaften verändert. Dies kann mit einem unmittelbaren Selektionsvorteil für den Wirt verbunden sein, da es dem Zielprotein beispielsweise ermöglicht wird, durch DGR-vermittelte Hypermutation schneller auf einen externen Anpassungsdruck zu reagieren. Da lediglich eine einzige, definierte Sequenz ersetzt wird, legt das Element auch keine neuen Kopien von sich selbst an; es agiert vielmehr als Teil des Wirtsorganismus, und nicht mehr als eigenständiges, auf die eigene Weiterverbreitung ausgerichtetes Element. Im bisher am detailliertesten beschriebenen Element, dem *Bordetella* Bakteriophagen-DGR, ähnelt dieses System auf verblüffende Weise der Technik des Phagendisplays, einer Methode zur Identifizierung von Bindungspartnern für einen gewählten Liganden (Smith, 1985).

Trotz dieser bemerkenswerten Eigenschaften ist über die Elemente selbst, ihre Verbreitung, ihren Funktionsmechanismus und ihre Genese bislang nicht viel bekannt. In der vorliegenden Arbeit wurden daher bioinformatische und experimentelle Techniken eingesetzt, um den Kenntnissstand über diese Elemente beträchtlich zu erweitern, und ein aktualisiertes Bild von diesen speziellen Vertretern mobiler genetischer Elemente zu definieren.

1.1 Diversitätsgenerierende Retroelemente (DGRs)

Der erste Vertreter dieser neuen Art mobiler genetischer Elemente wurde im *Bordetella* Bakteriophagen entdeckt (Liu et al., 2002). Hierbei handelt es sich um einen 42,5 kbp großen dsDNA-Phagen, der Bakterien der Gattung *Bordetella* infizieren kann. Die erfolgreiche Infektion einer Zielzelle erfordert die Bindung des Phagen an das Oberflächenprotein Pertactin (Leininger et al., 1991; Liu et al., 2002), was durch ein Protein an den Spitzen der Schwanzfasern des Phagen erfolgt; dieses wird vom major tropism determinant-Locus (*mtd*) codiert. Die Expression von Pertactin ist – wie zahlreiche andere Proteine in *Bordetella* (Akerley et al., 1995; Cotter and DiRita, 2000; Cotter and Miller, 1997) – abhängig vom Lebenszyklus des Bakteriums. Während es in der Plus-Phase auf der Zelloberfläche exprimiert vorliegt, fehlt es in der Minusphase; die Infektion ist dem Phagen in diesem Stadium somit normalerweise nicht möglich. Es wurde jedoch beobachtet, dass eine geringe Zahl von Phagen (etwa einer von 10^6 Phagen) zur Infektion von Minus-Phasenzellen in der Lage ist, indem sie alternative Oberflächenproteine benutzen (Liu et al., 2002). Dieses Phänomen wird Tropismenwechsel (*tropism switching*) genannt. Die Mehrheit der Phagen, die aus einer infizierten Minus-Phasenzelle hervorgingen, konnte ebenso nur andere Minus-Phasenzellen infizieren. Mit einer Frequenz von einem von 10^3 Phagen wurde der Tropismus jedoch wieder zu Plus-Phasenzellen revertiert. Bei beiden Tropismenwechseln konnte eine dritte Phagenspezies isoliert werden, die in der Lage war, Zellen in beiden Phasen zu infizieren. Die Phagenvarianten wurden entsprechend ihrer Tropismen Bvg plus-tropic Phage (BPP-1), Bvg minus-tropic Phage (BMP-1) und Bvg indiscriminant Phage (BIP-1) genannt, wobei sich Bvg auf das Signaltransduktionssystem BvgAS von *Bordetella* Bakterien bezieht, das den Wechsel zwischen diesen beiden Lebensphasen und einer intermediären Phase vermittelt (Deora et al., 2001; Stockbauer et al., 2001; Uhl and Miller, 1996).

Sequenzanalysen der verschiedenen Phagenvarianten zeigten, dass jeweils das *mtd*-Gen in einer definierten Region an seinem 3‘-Ende stark mutiert war (Liu et al., 2002). Dieser Sequenzabschnitt erhielt die Bezeichnung ‚variable Region‘ (VR). Auffällig war hierbei, dass Hypervariation nur an einzelnen, distinkten Positionen auftrat. Bei einer genaueren Betrachtung der benachbarten Sequenzabschnitte fand sich downstream vom *mtd*-Locus eine Kopie der VR, in der die hypervariablen Reste ausschließlich Adeninen entsprachen. Diese Region erhielt später die Bezeichnung ‚Template Repeat‘ (TR). Darüber hinaus wurden zwei offene Leserahmen identifiziert: einer codiert ein unbekanntes Proteinprodukt, das man zunächst mit ‚*accessory tropism determinant*‘ (Atd), und später ‚*accessory variation determinant*‘ (bAvd) benannt hat., Der zweite Leserahmen codiert füreine reverse Transkriptase (*brt*), was für ein DNA-Virus äußerst ungewöhnlich ist (s. Abbildung 1).

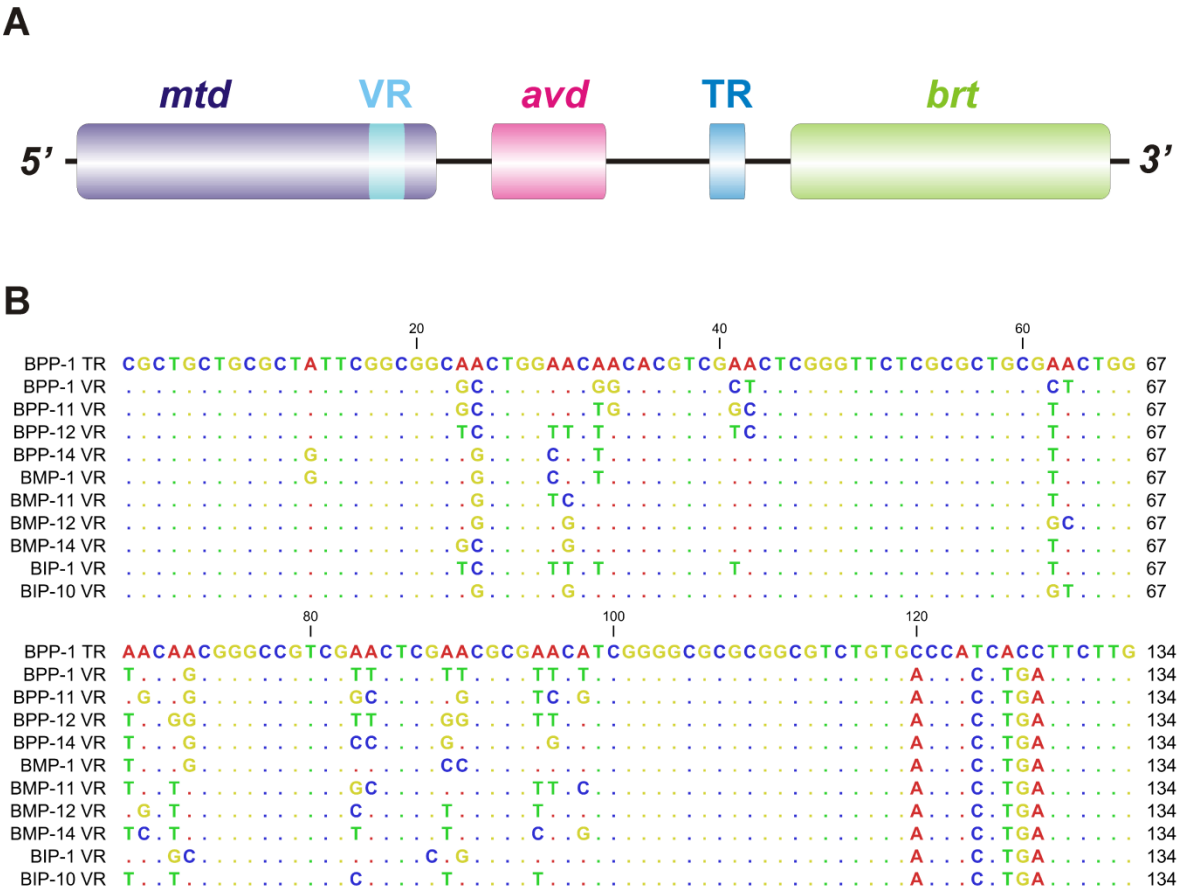


Abbildung 1: Das DGR-Element des *Bordetella* Bakteriophagens. (A) Struktur des *Bordetella* Bakteriophagen-DGRs. In der variablen Region (VR) des *mtd*-Locus tritt massive Sequenzvariation auf, die einen Tropismenwechsel des Phagens ermöglicht. Downstream befinden sich Leserahmen für die Accessory Variability Determinant (*avd*) und für eine reverse Transkriptase (*brt*). Zwischen *avd* und *brt* liegt der Template Repeat, der eine nahezu identische Kopie der variablen Region des *mtd*-Locus darstellt; in ihm tritt keine Sequenzvariation auf, hypervariable Positionen in der VR entsprechen hier ausschließlich Adeninen. (B) Alignment der BPP-1 TR und den variablen Regionen mehrerer unterschiedlicher Tropismenvarianten (nach Liu et al., 2002). Zur besseren Übersichtlichkeit sind nur veränderte Positionen dargestellt.

Deletionsmutationen des *brt*-Locus sowie der Repeatregion führten zu einer signifikant verminderten Fähigkeit zum Tropismenwechsel (Liu et al., 2002). In einem weiteren Experiment konnte ein Transfer von stummen Mutationen in der Repeatregion in die entsprechende Region des *mtd*-Gens in tropismenveränderten Phagennachkommen demonstriert werden. Diese einzelnen Beobachtungen führten zur Entwicklung des Konzepts der diversitätsgenerierenden Retroelemente.

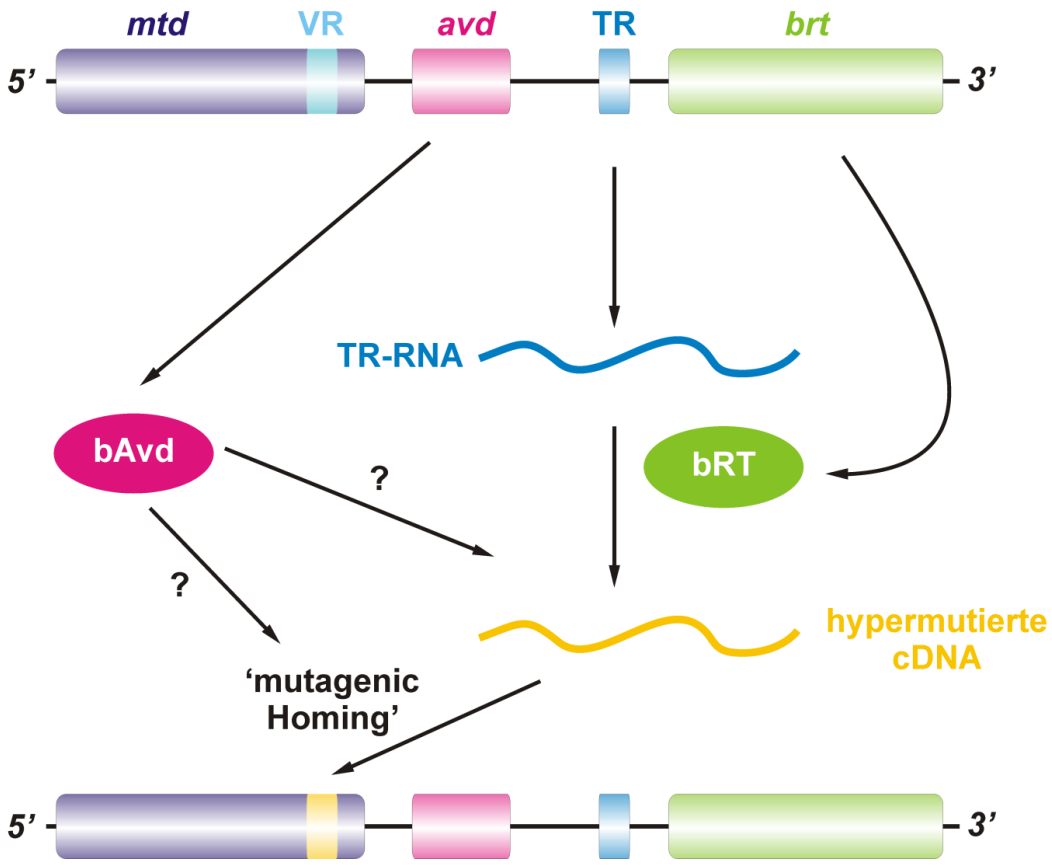


Abbildung 2: Schema des DGR-Mechanismus. Der TR-Locus wird transkribiert, und das Transkript über die assoziierte reverse Transkriptase in cDNA umgeschrieben. Hierbei kommt es zu Fehlern bei der Basenkomplementation von Adeninen, so dass in die resultierende cDNA an entsprechenden Positionen Nicht-Thymine eingebaut werden. In einem bisher nicht geklärten Vorgang, der die Bezeichnung ‚mutagenic Homing' erhalten hat, ersetzt die mutagenisierte cDNA den entsprechenden Parentalstrang in der variablen Region des Zielgens. Durch Reparaturmechanismen o. Ä. wird die Synthese des Zweitstranges durchgeführt, der nun Nicht-Adenine an Positionen enthält, die Adeninen im TR entsprechen. Die Rolle des Proteins bAvd ist noch nicht geklärt; denkbar ist eine Beteiligung am RT-Prozess oder im 'mutagenic homing'.

Der heute als am wahrscheinlichsten angenommene Mechanismus dieser Elemente ist in Abbildung 2 skizziert. Er beginnt mit der Transkription des Template Repeats. Die entstandene TR-RNA wird anschließend als Matrize zur cDNA-Synthese verwendet, was durch die reverse Transkriptase Brt erfolgt (Guo et al., 2008). Hierbei werden Fehlpaarungen in den entstehenden DNA-Strang eingefügt, die zu den beobachteten Adeninaustauschen führen; diese werden im Nachfolgenden gemäß IUPAC-Nomenklatur als A->N-Mutationen (‚*Adenine to aNy nucleotide*') bezeichnet. Die mutierte cDNA wird anschließend in den Parentalstrang eingefügt, wobei die entsprechende Region am 3'-Ende des *mtd*-Gens ersetzt wird. Dieser Vorgang wird als ‚mutagenic homing' bezeichnet, und ist noch immer weitgehend unverstanden. Es ist möglich, dass das akzessorische Protein bAvd hierbei eine Rolle spielt, da Deletionsmutanten des zugehörigen Gens *avd* ebenfalls eine deutlich verminderte Fähigkeit zum Tropismenwechsel zeigten (Doulatov et al., 2004). Interessanterweise wurde beobachtet, dass Mutationen präferentiell an den ersten beiden Positionen eines Codons erfolgen (Medhekar and Miller, 2007; Minot et al., 2012). Während

Mutationen in der dritten Codonposition, der sogenannten Wobble-Position, häufig zu synonymen Substitutionen führen, bewirken Veränderungen in den ersten beiden Positionen meist auch Veränderungen auf Aminosäureebene. Dies führt zu einer effektiven lokalen Variation des Mtd-Proteins, was gleichbedeutend ist mit einer beträchtlichen Erweiterung des Wirtszellspektrums für den Phagen, und einem deutlichen Selektionsvorteil. Zuvor wurde bereits angesprochen, dass dieses System als eine Art natürliches Phagendisplay beschrieben werden könne. Im Unterschied zu der Labortechnik, bei der in der Regel nur einer der Bindungspartner variiert wird, erfolgt hier die Anpassung des einen Bindungspartners über massive Sequenzvariation an ein Set aus potentiellen Bindungspartnern. Die Möglichkeit einer Adaption des *Bordetella* Bakteriophagen-DGRs an alternative Zielproteine konnte bereits demonstriert werden: Guo et al. konnten den defekten offenen Leserahmen eines Kanamycinresistenzfaktors durch DGR-Aktivität reparieren, und hierdurch dem Träger dieses Gens ein Überleben auf kanamycinhaltigem Medium ermöglichen (Guo et al., 2011). Analoge Variationen dieses Systems für den Einsatz in komplexeren Versuchsanordnungen sollten demnach ebenso möglich sein.

Die Existenz weiterer DGRs konnte zunächst von Doulatov et al. gezeigt werden, die in ersten Datenbankanalysen acht weitere Vertreter identifizieren und gleichzeitig zeigen konnten, dass diese Elemente auch in Prokaryoten vorliegen können, und sich phylogenetisch von anderen Retroelementen wie LTR-Retrotransposons oder Gruppe II-Introns klar abgrenzen (Doulatov et al., 2004). Bis zum Jahr 2008 wurden ca. 40 dieser Elemente in zumeist prokaryotischen Genomen identifiziert (Simon and Zimmerly, 2008), eine genauere Beschreibung fand bisher jedoch nicht statt. Nahezu sämtliche Studien, die seit der Erstbeschreibung der DGR-Elemente im Jahr 2002 erschienen sind, konzentrierten sich auf das DGR-Element aus dem *Bordetella* Bakteriophagen, und somit stellt auch heute noch dieser Prototyp das am besten untersuchte System dieser Gruppe von Retroelementen dar. Die folgenden Abschnitte fassen die Ergebnisse dieser Studien zusammen, und geben einen Überblick darüber, was heute über Struktur und Mechanismus dieser Elemente bekannt ist.

1.1.1 Template Repeat und variable Region

Die bisher identifizierten diversitätsgenerierenden Retroelemente zeigten stets Repeatstrukturen, welche sich lediglich in Adeninpositionen voneinander unterschieden; dies muss jedoch nicht zwangsläufig bedeuten, dass nicht auch DGRs mit einer anderen Basenpräferenz existieren, und bisher lediglich nicht gefunden worden sind. Während in den bekannten DGRs variable Regionen

stets Teil eines Zielgens sind, wurden Template-Repeats meist downstream der Zielgene identifiziert. Interessanterweise können Template Repeats Teil des offenen Leserahmens der reversen Transkriptase sein (so z. B. im Falle des *Vibrio harveyi* VHML-Phagen), und befinden sich auch sonst unmittelbar up- oder downstream dieses Leserahmens (Medhekar and Miller, 2007). Dies legt nahe, dass TR und RT ein gemeinsames Transkript verwenden und die frisch translatierte RT ihre eigene mRNA als Template zur cDNA-Synthese verwendet, wie es von einigen Retroelementen wie beispielsweise dem LINE1-Element bekannt ist (Wei et al., 2001).

Eine der interessantesten Eigenschaften diversitätsgenerierender Retroelemente ist die Fähigkeit, theoretisch unendliche viele Zyklen der Diversifizierung durchlaufen zu können, und somit ein Protein so lange zu verändern, bis es den jeweils geltenden Anforderungen am besten entspricht. Hierzu muss jedoch die Urkopie der variablen Region – also der Template Repeat – unbedingt erhalten bleiben; wäre auch er Ziel des mutagenic Homings, würden sukzessive alle Adenine zu Nichtadeninen mutiert, bis keine Variation mehr möglich wäre. Da das mutagenic Homing jedoch höchstwahrscheinlich auf einem Homologiemechanismus beruht, muss es ein Merkmal geben, das TR von VR unterscheidet. Sequenzvergleiche dieser beiden Regionen zeigen eine C/G-reiche Region am 3‘-Ende der Repeats, die gefolgt wird von einem Sequenzabschnitt, der in sämtlichen untersuchten VRs unverändert ist, sich aber in fünf Nucleotiden vom entsprechenden TR-Abschnitt unterscheidet (s. Abbildung 1B). Doulatov et al. konnten zeigen, dass diese Sequenz, die sie *initiator of mutagenic homing* (IMH) benannt haben, für die Unterscheidung zwischen TR und VR verantwortlich ist (Doulatov et al., 2004). In Experimenten, in denen die IMH-Region des Template Repeats (IMH*) die entsprechende Sequenz in der variablen Region ersetzte, konnte keine Hypermutation der VR beobachtet werden, während ein umgekehrter Austausch (also von IMH* zur IMH der VR) Variabilität im Template Repeat bewirkte. Ob diese Region auch in anderen DGRs zu finden ist oder es sich um ein spezielles Merkmal des *Bordetella* Bakteriophagen-Elements handelt, ist in den bisherigen Studien noch nicht geklärt worden. Die Beobachtung, dass die IMH-Sequenz der variablen Region sich nie ändert, ist zugleich ein Hinweis darauf, dass die Stelle, an der die Insertion der hypermutierten cDNA erfolgt, upstream dieses Elements, also in der G/C-reichen Region liegen muss. Dies konnte 2008 von Guo et al. auch experimentell nachgewiesen werden (Guo et al., 2008). In der gleichen Arbeit wurden zudem die Minimalanforderungen an Template Repeats für eine erfolgreiche Variation und anschließendes Homing bestimmt. Am 3‘-Ende der TR sind lediglich die G/C-reiche Region sowie die IMH* wichtig, während am 5‘-Ende 10 bis 19 Basen vorliegen müssen, die Homologie zur VR aufweisen; Sequenzen, die dazwischen liegen, können ausgetauscht oder – im Falle des *Bordetella* Bakteriophagen DGRs – um bis zu 200 Nucleotide verlängert werden. Dieses Resultat ist besonders für zukünftige Anwendungen von Bedeutung, in denen DGRs als Diversitätsgeneratoren für heterologe Zielproteine eingesetzt werden könnten, da es zeigt, dass

grundsätzlich jedes Protein auf diese Weise hypermutiert werden könnte, und mechanistisch nur geringe Anforderungen an TR/VR-Sequenzen gestellt werden.

Zusätzlich zu IMH und IMH*, die für die Diskriminierung zwischen TR und VR verantwortlich sind, wurde ein weiteres Element am 3‘-Ende der *Bordetella* Bakteriophagen-VR identifiziert, welches eine Rolle im Homing-Prozess zu besitzen scheint. Guo et al. zeigten 2011, dass dort eine Stem-Loop-Struktur auf DNA-Ebene ausgebildet werden kann, deren Deletion zu einer drastisch verminderten Homing-Aktivität führt (Guo et al., 2011); da der komplementäre Strang logischerweise ebenfalls zur Ausbildung dieser Struktur fähig ist, könnte es sich sogar um eine Kreuz-Struktur (‚*cruciform*‘) handeln. Auch in anderen DGR-Elementen konnten Sequenzen gefunden werden, die gleichartige Strukturen ausbilden könnten. Stämme wiesen stets eine Länge von 7 bis 10 Basenpaaren auf und waren G/C-reich, zeigten sonst jedoch keinen hohen Konservierungsgrad. Ein anderer Befund ergab sich bei den Loopstrukturen: hierbei handelte es sich stets um Tetranucleotidsequenzen, welche dem Consensus 5‘-GRNA-3‘ entsprachen. Interessanterweise kann ein ähnliches Motiv (5‘-GNRA-3‘) häufig in RNA-Molekülen, die einen hohen Grad an Sekundär- und Tertiärstrukturierung aufweisen, als stabilisierendes bzw. verankerndes Element gefunden werden (Abramovitz and Pyle, 1997; Costa and Michel, 1995; Woese et al., 1990). Ob das VR-assoziierte Motiv tatsächlich eine Entsprechung des RNA-Motivs auf DNA-Ebene darstellt, oder welche Aufgabe im speziellen der Stem Loop/Kreuzform-Struktur im Homing-Mechanismus zukommt, konnte bisher jedoch noch nicht geklärt werden.

1.1.2 Die reverse Transkriptase

Reverse Transkriptasen (RTs) sind Proteine, die DNA mithilfe einer RNA-Matrize synthetisieren können. Seit ihrer Entdeckung durch Howard Temin (Temin and Mizutani, 1970) und David Baltimore (Baltimore, 1970) im Jahr 1970 konnten RTs in einer Vielzahl biologischer und oftmals medizinisch relevanter Zusammenhänge identifiziert werden, wie beispielsweise in Retroviren (Gallo et al., 1984; Poiesz et al., 1981), oder in der Reparatur von eukaryotischen Chromosomenenden durch das Enzym Telomerase, bei dem es sich ebenfalls um eine reverse Transkriptase handelt (Meyerson et al., 1997; Nakamura et al., 1997). Einige reverse Transkriptasen, wie beispielsweise die des humanen Immundefizienzvirus (HIV), weisen eine relativ hohe Fehlerrate bei der cDNA-Synthese auf, und führen somit zu zahlreichen Punktmutationen in den resultierenden Proteinen (Preston et al., 1988; Roberts et al., 1988; Smyth et al., 2012). Dies erschwert retrovirale Therapieansätze erheblich, da die molekularen Ziele der Wirkstoffe ihre Strukturmerkmale stetig verändern, und hoch-spezifisch wirkende Agenzien wirkungslos werden lassen (Wainberg et al., 1993). Die Fehler dieser RTs sind allerdings nicht auf einen bestimmten Basentypus beschränkt; Mutationen können an jeder Position der neu synthetisierten cDNA auftreten.

Bisher beschriebene reverse Transkriptasen weisen jeweils eine ähnliche strukturelle Organisation auf. Generell werden Polymerase-Strukturen gerne mit denen einer rechten Hand verglichen, bestehend aus den Domänenelementen Daumen (*thumb*), Handfläche (*palm*), Fingern (*fingers*) sowie einer RNase H-Domäne und einer Verbindungsdomäne (s. Abbildung 3). Das katalytische Zentrum der Polymeraseaktivität, ein Tetrapeptidmotiv der Form YxDD, ist dabei in der Handfläche lokalisiert, während Finger und Daumen an der Substraterkennung beteiligt sind und mit Oligonucleotiden interagieren.

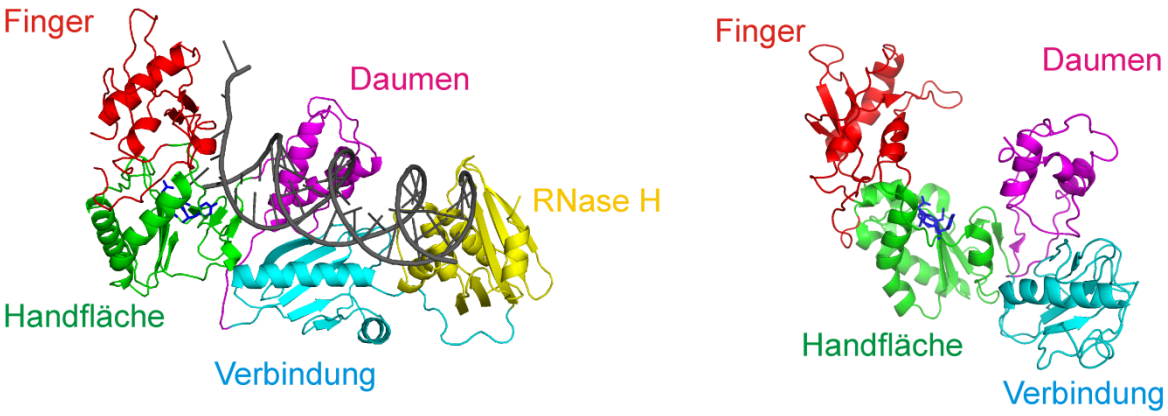


Abbildung 3: Reverse Transkriptasen und ihre Domänenstruktur. Trotz einiger Unterschiede in den Details weisen die reverse Transkriptase des humanen Immundefizienzvirus HIV (links, Untereinheit p66) und das homologe Enzym des murinen Leukämievirus Moloney-MLV (rechts) eine ähnliche Organisation auf, die sich in Finger-, Handflächen- und Daumendomäne sowie eine Verbindungsdomäne gliedert. Nicht abgebildet ist die RNase H-Domäne aus M-MLV, da bisher noch keine Struktur des vollständigen Enzyms vorliegt. Zusätzlich abgebildet ist ein Oligonucleotid im Komplex mit HIV-RT (grau), das zwischen Daumen und Fingern zur RNaseH-Domäne verläuft. Katalytische Zentren (YxDD) in den Handflächen sind in beiden Molekülen dunkelblau dargestellt. RCSB-Codes 1RTD und 1RW3 (Das and Georgiadis, 2004; Huang et al., 1998).

Zusätzlich zu den in Abbildung 3 dargestellten Strukturen liegt heute lediglich eine weitere Kristallstruktur einer reversen Transkriptase vor, die der katalytisch aktiven TERT-Untereinheit der Telomerase aus dem Rotbraunen Reismehlkäfer (*Tribolium castaneum*) (Gillis et al., 2008). Diese geringe Datenlage spiegelt ein zentrales Problem in der Strukturaufklärung dieser Enzymklasse wider. Neben der Erzeugung stabiler, wohlgeordneter Kristalle stellt bereits die Expression eines löslichen, aktiven Proteins in ausreichenden Mengen eine gewisse Herausforderung dar. Zahlreiche Veröffentlichungen aus der Anfangszeit der HIV-Forschung demonstrieren dies anschaulich; Versuche, das RT-Protein aus HIV rekombinant herzustellen und in aktiver Form aufzureinigen waren zunächst nur von begrenztem Erfolg (Flexner et al., 1988; Hizi et al., 1988; Larder et al., 1987), und erst eine Mutagenese des Leserahmens mit der damals noch jungen Technik der Polymerasekettenreaktion (*polymerase chain reaction*, PCR) führte zu einer Optimierung der

Ausbeute an löslichem Protein (D'Aquila and Summers, 1989), und zuletzt zur Lösung der Struktur durch Kohlstaedt et al. (Kohlstaedt et al., 1992). Gleichermaßen mussten Löslichkeitsprobleme bei der Strukturanalyse der M-MLV-RT durch Punktmutation einiger Aminosäuren und über gezielte Entfernung ungeordnet vorliegender Bereiche überwunden werden, um schließlich eine Kristallstruktur erzeugen und auswerten zu können (Das and Georgiadis, 2001, 2004).

Der *Bordetella* Bakteriophage ist ein DNA-Virus, und daher war die Anwesenheit eines Gens, das für eine reverse Transkriptase codiert, zunächst überraschend. Inzwischen konnte durch Deletions- und Nonsensemutationsversuche eindrucksvoll demonstriert werden, dass das bRT-Enzym des *Bordetella* Bakteriophagen-DGR für die Funktion des Elements essentielle Bedeutung besitzt (Guo et al., 2008; Liu et al., 2002). Da durch Intron-Tagging belegt werden konnte, dass die TR-Region tatsächlich transkribiert wird (Guo et al., 2008), kann außerdem davon ausgegangen werden, dass das TR-Transkript in einem weiteren Schritt von der elementcodierten reversen Transkriptase in cDNA umgeschrieben wird. Doch wie entstehen die Mutationen, die schließlich in der VR identifiziert werden können?

Wie bereits erwähnt, zeigen einige retrovirale RTs eine hohe Fehlerrate bei der cDNA-Synthese. Es liegt daher nahe anzunehmen, dass DGR-RTs während der reversen Transkription ebenfalls Mutationen in die cDNA einfügen, und zwar mit einer Spezifität für Adeninpositionen. Es muss jedoch darauf hingewiesen werden, dass dies bisher nur eine Theorie ist, und ein eindeutiger experimenteller Nachweis noch aussteht. Zwar wurde bisher die heterologe Expression und Aufreinigung der RT aus dem *Bordetella* Bakteriophagen-DGR beschrieben (Liu et al., 2002), allerdings wurde noch keine ausreichende biochemische, strukturelle und funktionelle Charakterisierung des Proteins vorgenommen. Die Gründe hierfür sind nicht bekannt, allerdings weist eine kürzlich erschienene Publikation darauf hin, dass das Wildtypenzym keine hohe *in vitro*-Stabilität aufweist, und zudem nur über eine Aufreinigung aus Inclusion Bodies zu isolieren ist, was strukturelle und funktionelle Analysen deutlich erschwert (Alayyoubi et al., 2012).

1.1.3 Der akzessorische offene Leserahmen

In einigen der bisher beschriebenen DGRs befindet sich ein zusätzlicher offener Leserahmen zwischen VR und TR (Medhekar and Miller, 2007). Wie durch Deletionsexperimente gezeigt wurde, wird dieser Leserahmen für den Tropismenwechsel – und somit für DGR-Aktivität – im *Bordetella* Bakteriophagen benötigt (Doulatov et al., 2004). Die genaue Funktion des entsprechenden Genprodukts, welches im *Bordetella* Bakteriophagen zunächst ‚accessory tropism determinant' (Atd) (Doulatov et al., 2004) und später ‚accessory variation determinant' (bAvd) (Alayyoubi et al., 2012)

genannt wurde, ist unbekannt. Eine jüngst erschienene Publikation konnte lediglich zeigen, dass bAvd *in vitro* unspezifisch Nucleinsäuren bindet, und mit der reversen Transkriptase des BPP1-Elements einen Komplex bildet, wobei dieser Komplex höchstwahrscheinlich eine Bedeutung für das mutagene Homing im DGR-Mechanismus besitzt (Alayyoubi et al., 2012). Des weiteren konnte durch sukzessive Deletionen gezeigt werden, dass der 3‘-Teil des *avd*-Leserahmens nicht nur eine codierende Funktion besitzt, sondern einen Teil des Template Repeats bildet, der sich unmittelbar downstream anschließt. Welche Aufgabe diesem Teil zukommt, ist nicht erläutert worden. Möglicherweise befinden sich regulatorische Elemente wie ein Promoter in der codierenden Sequenz von bAvd, die in den Deletionsmutanten verloren gingen und daher keine Transkription des TR mehr erlauben. Von Alayyoubi et al. stammt außerdem eine erste Kristallstruktur von bAvd, die eine homopentamere Quartärstruktur vorhersagt, welche wiederum aus einem Vier-Helix-Monomer gebildet wird (Alayyoubi et al., 2012).

Es wurde außerdem darauf hingewiesen, dass in einigen DGRs, die keinen akzessorischen offenen Leserahmen aufweisen, die RT-Gene mit einem Locus überlappen, der möglicherweise eine *Helicase and RnaseD C-terminal* Domäne (HRDC) codiert (Medhekar and Miller, 2007). Hierbei handelt es sich um Domänen, die in anderen Proteinen an der Bindung von Nucleinsäuren beteiligt sind (Liu et al., 1999). Im Hinblick auf den noch immer größtenteils unverstandenen Homing-Prozess, in dem der parentale DNA-Strang geöffnet und aufgewunden werden muss, liegt es nahe zu vermuten, dass das bAvd-Protein und seine Homologen eine den HRDC-Domänen analoge Funktion übernehmen, und an der Bindung von Nucleinsäuren beteiligt sind und dass HRDC-Domänen das Fehlen dieser Faktoren kompensieren. Interessanterweise werden HRDC-Domänen ebenfalls ausschließlich aus Alpha-Helices gebildet, diese weisen allerdings im Vergleich zur Struktur von bAvd eine unterschiedliche Orientierung zueinander auf (Bernstein and Keck, 2005; Killoran and Keck, 2008; Kim and Choi, 2010).

1.1.4 Das Zielprotein

Neben den akzessorischen Proteinen stellen DGR-Zielproteine eine weitere Komponente dar, über die nur wenig bekannt ist. Im prototypischen DGR des *Bordetella* Bakteriophagens wird das *mtd*-Gen des Phagen durch DGR-Aktivität hypermutiert, um Proteinvarianten zu generieren, die dem Phagen die Bindung an alternative Oberflächenproteine erlauben (Liu et al., 2002). Diese Element-Wirtsbeziehung ist zurzeit die einzige, die hinsichtlich ihrer Funktion verstanden ist. Ein weiteres Zielprotein (TvpA aus Treponema denticola) wurde 2011 von Le Coq und Ghosh näher betrachtet, allerdings ohne Rückschlüsse auf seine biologische Funktion ziehen zu können (Le Coq and Ghosh, 2011). In jener Arbeit lag der Fokus auf strukturellen Eigenschaften von TvpA; es wurde gezeigt, dass das Protein – ebenso, wie Mtd (McMahon et al., 2005) – ein Faltungsmuster aufweist, das zuerst bei

C-Typ-Lektinen beobachtet und nach diesen benannt wurde (Weis et al., 1991). Dieses Muster erlaubt extensive Variationen einzelner Aminosäurereste, die eingebettet zwischen konservierten, invarianten Aminosäuren liegen. Letztere stellen eine Art Gerüst dar, durch das die Gesamtstruktur der Domäne erhalten bleibt, während der biochemische Charakter durch Hypermutation der an der Oberfläche exponierten Aminosäuren variiert werden kann (s. Abbildung 4).

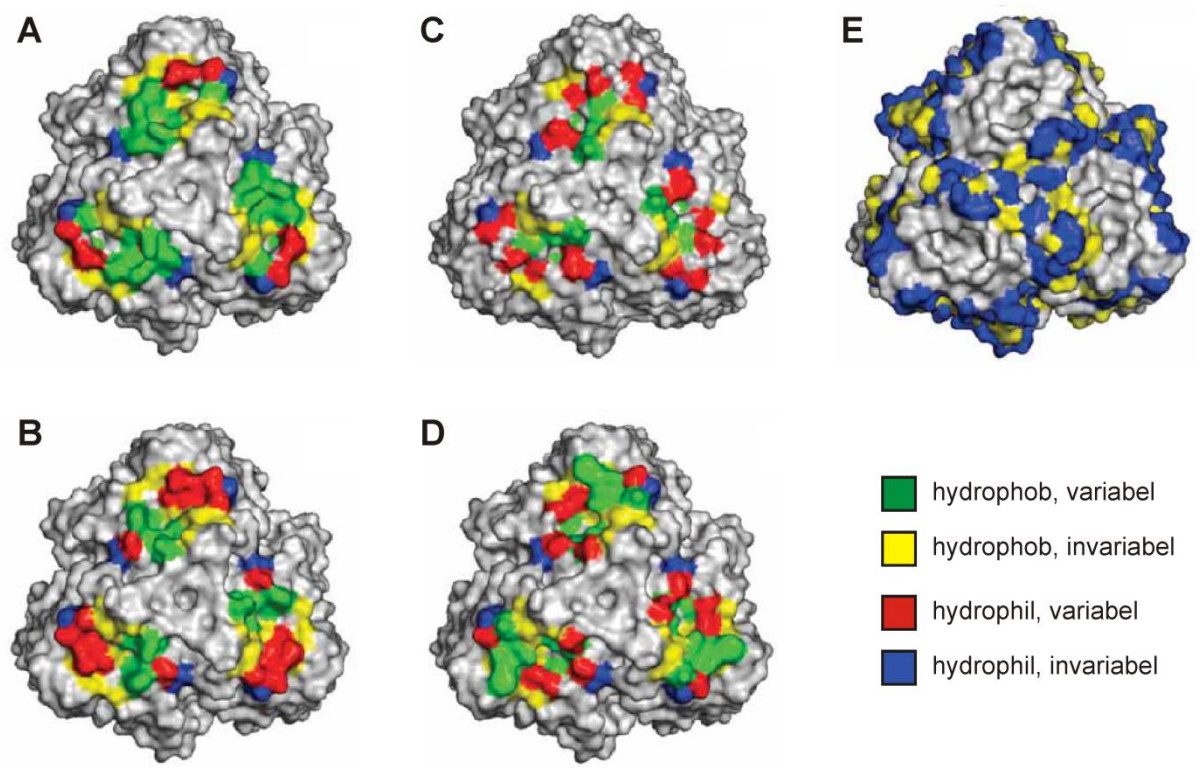


Abbildung 4: Topologie der hypervariablen Reste des Mtd-Proteins. Dargestellt sind die dreidimensionalen Strukturen der Rezeptorbindestellen von vier verschiedenen Mtd-Trimeren, die durch DGR-Aktivität hypermutiert wurden (A bis D). Grün und rot hervorgehobene Reste geben variable Positionen an, gelbe und blaue Markierungen die Aminosäuren, die nicht mutiert werden. Zusätzlich zeigt Teilabbildung E die invarianten Aminosäuren, die die Rezeptorbindestelle umgeben (nach McMahon et al., 2005).

Dies stellt somit eine elegante Lösung eines grundsätzlichen Problems DGR-vermittelter Mutagenese dar: wie kann die Bindungsspezifität eines Proteins variiert werden, ohne dass es zum Kollaps der Struktur, und somit zu einem vollständigen Funktionsverlust kommt? So können beispielsweise Antikörper – das einzige bisher bekannte Konzept in der Natur, das sich mit DGRs vergleichen lässt – unter bestimmten Voraussetzungen aggregieren, und zum Krankheitsbild der AL-Amyloidose (*Amyloid Light Chain Amyloidosis*) führen. Im Gegensatz zu DGR-Zielproteinen mit C-Typ-Lektinfaltungsmuster liegen variable Aminosäuren dort konzentriert an einer Stelle der Polypeptidkette vor, weshalb ungünstige Sequenzvariationen nicht durch zwischenliegende,

stabilisierend-wirkende Reste kompensiert werden können. Die Gefahr eines Strukturverlusts und einer damit einhergehenden Aggregation über freiliegende, hydrophobe Regionen ist somit höher als bei Proteinen mit C-Typ-Lektinfaltungsmuster. Weitere Unterschiede zeigen sich im Variationspotential dieser beiden Konzepte. Antikörper und T-Zellrezeptoren erreichen eine Sequenzvielfalt von ~ 10^{14} bis 10^{16} Proteinen; im Gegensatz hierzu können DGRs 10^{13} (im Fall von Mtd) bis 10^{24} (im Fall von TvpA) verschiedene Proteinvarianten generieren, und somit das erstgenannte System deutlich übertreffen (Chothia and Lesk, 1987; Davis and Bjorkman, 1988).

Mangels experimentell bestimmter Referenzdaten ist es schwierig, eine Aussage über die biologischen Funktionen der Zielproteine, und somit auch über den Einfluss der DGR-vermittelten Hypervariation zu treffen. Im Falle von TvpA konnte ein spirochaetenspezifisches Lipopeptidsignal ermittelt werden (Medhekar and Miller, 2007), was eine Membranständigkeit des Proteins nahelegt. Dies stünde auch im Einklang mit der bisherigen Annahme, dass DGRs eine Möglichkeit bieten, schnell und effizient auf äußeren Selektionsdruck zu reagieren und eine Anpassung des Wirts an die veränderten Bedingungen vorzunehmen.

1.2 Zielsetzung

Diversitätsgenerierende Retroelemente stellen eine neuartige Klasse von mobilen genetischen Elementen mit bisher einzigartigen Eigenschaften dar. Im Gegensatz zu klassischen Retroelementen gehen sie eine Art symbiotische Beziehung mit ihrem Wirtsgenom ein, ohne sich zu vervielfältigen und hierüber wirtseigene Genstrukturen zu zerstören. Der Nutzen, den diese Elemente für wissenschaftliche und industrielle Anwendungen bereitstellen könnten, ist offensichtlich. Diversitätsgeneratoren könnten in der Entwicklung neuartiger pharmakologischer Wirkstoffe eingesetzt werden, in der Erzeugung neuer Enzymklassen zur katalytischen Umsetzung von Metaboliten oder Umweltgiften, oder – in Analogie zum *Bordetella* Bakteriophagen – als Evolutionsbeschleuniger von Phagen, die gegen medizinisch-relevante bakterielle Erreger eingesetzt werden und damit eine Strategie gegen die zunehmende Verbreitung von Antibiotikaresistenzen darstellen könnten. Hierzu ist ein genaueres Studium dieser Elemente notwendig.

Einige wegweisende Veröffentlichungen haben bereits elegante Experimente beschrieben, in denen erste Details über die molekularen Mechanismen der DGR-vermittelten Mutagenese im *Bordetella* Bakteriophagen ermittelt werden konnten. Wenngleich angenommen werden kann, dass DGR-Elemente in anderen Phagen und Prokaryoten ähnliche Eigenschaften aufweisen, wurden diese Elemente bisher weder experimentell noch in deskriptiven Studien behandelt. Darüber hinaus sind noch viele Details ungeklärt: wie erfolgt die Hypermutation an Adeninpositionen? Was sind die molekularen Grundlagen des Homing-Prozesses? Was ist die Funktion des bAvd-Proteins und seiner Homologe? Wie weit verbreitet sind DGRs in der Natur, und wie werden sie weitergegeben?

Die vorliegende Arbeit widmet sich einigen dieser bisher offenen Fragen. Ein erstes Ziel stellte die Entwicklung des Programms DiGReF in Zusammenarbeit mit der Abteilung für Genetik der Technischen Universität Kaiserslautern dar, welches eine weitgehend automatisierte Analyse von Nucleotidsequenzen auf DGR-typische Repeatstrukturen erlaubt. Mit diesem Programm könnten in einem nächsten Schritt die erhaltenen DGRs auf gemeinsame und unterschiedliche Merkmale hin untersucht werden. Neue Einblicke in die Natur und Genese dieser Elemente könnten hierdurch ermöglicht werden, und Rückschlüsse auf molekulare Eigenschaften dieser Elemente erlauben, auf die weitere Experimente aufbauen können.

Ein anderer Teil dieser Arbeit konzentrierte sich auf zwei essentielle Faktoren von DGRs. Eines der Ziele war die rekombinante Expression und Reinigung einer reversen Transkriptase zur eingängigeren strukturellen und biochemischen Charakterisierung. Desweiteren sollte die Rolle der bisher kaum beschriebenen akzessorischen Proteine ermittelt werden, die sie im DGR-Mechanismus erfüllen. Hierzu sollte versucht werden, einen Vertreter dieser Proteine rekombinant zu erzeugen, zu isolieren und in einer Reihe von Funktionsanalysen zu untersuchen.

II. Material & Methoden

2.1 Material

2.1.1 Chemikalien

Alle Chemikalien wurden von der Applichem GmbH, Darmstadt, bezogen. Wenn möglich, wurde ein molekularbiologiegeeignter Reinheitsgrad gewählt. Chemikalien, die nicht von der Applichem GmbH stammten, sind in Tabelle 1 genannt. Der Reinheitsgrad war stets *pro analysi* (p.A.) oder höher, sofern nicht anders angegeben.

Tabelle 1: Nicht von Applichem bezogene Chemikalien

Substanz	*Hersteller*
40% Acrylamid:Bisacrylamidlösung (37,5:1)	Carl Roth GmbH
Agarose	Serva
Ammoniumperoxodisulfat	Carl Roth GmbH
Complete Protease Inhibitor Cocktail	Roche
Coomassie Brilliant Blue R250	Carl Roth GmbH
Di (N-succinimidyl)-3,3'-dithiopropionat	Sigma-Aldrich
Eisessig	Carl Roth GmbH
Ethanol 95%	Sigma-Aldrich
Glycerin 86%	Chemikalienausgabe TU Kaiserslautern
Glycin (Elektrophoresegrad)	Carl Roth GmbH
Imidazol (Ultrol Grade)	Calbiochem/Merck-Millipore
Isopropanol	JT Baker
N,N,N',N'-Tetramethylethylendiamin (TEMED)	Carl Roth GmbH
NDSB 195/201/256	Calbiochem/Merck-Millipore
Orange G	Sigma-Aldrich
Salzsäure 37%	JT Baker
Zinkchlorid	Sigma-Aldrich
Für RNase-freie Arbeiten (jeweils Molecular Biology Grade)	
Ethylendiamintetraacetat -Lösung (0,5 M)	Calbiochem/Merck-Millipore
Magnesiumchlorid-Lösung (1 M)	Sigma-Aldrich
Nucleotidtriphosphat-Set	Carl Roth GmbH

2.1.2 Lösungen und Puffer

Nachfolgend genannt sind die Lösungen und Puffer, die in dieser Arbeit verwendet wurden.

Arbeiten mit DNA

Agarosegele
1 % (w/v) Agarose
40 mM Tris
20 mM Essigsäure
1 mM EDTA
0,03 µg/mL Ethidiumbromid

Tris/Acetat/EDTA-Puffer (TAE, 50x)
2 M Tris (Base)
1 M Essigsäure
50 mM EDTA

5x KCM-Lösung für Transformationen
500 mM KCl
150 mM $CaCl_2$
250 mM $MgCl_2$

Annealing-Puffer (10x)
200 mM Tris-Cl (pH 7,5)
500 mM NaCl

Alkalische Lyse - Lösung 1
50 mM Glucose
10 mM EDTA (pH 8,0)
25 mM Tris-Cl (pH 8,0)

Alkalische Lyse - Lösung 2
0,2 N NaOH
1 % (w/v) SDS

Alkalische Lyse - Lösung 3
3 M Kaliumacetat (pH 5,5)

Tris/Borat/EDTA-Puffer (TBE, 10x)
890 mM Tris (Base)
890 mM Borsäure
20 mM EDTA

Polyacrylamidgele
89 mM Tris
89 mM Borsäure
2 mM EDTA
5 % (w/v) Acrylamid
8 M Harnstoff (nur bei denaturierenden Gelen)
0,05 % (w/v) Ammoniumpersulfat
0,05 % (v/v) TEMED

Elutionspuffer für Aufreinigungen aus PAA-Gelen
10 mM MOPS (pH 6,0)
300 mM NaCl
1 mM EDTA (pH 8,0)

Arbeiten mit RNA

***in vitro*-Transkriptionspuffer (10x)**
400 mM Tris-Cl (ph 8,0)
100 mM NaCl
220 mM MgCl2
20 mM Spermidin
0,1 % (v/v) Triton X-100

Arbeiten mit Proteinen

Polyacrylamidgele - Trenngel
375 mM Tris-Cl (pH 8,8)
8 bis 15 % (w/v) Acrylamid
0,1 % (w/v) SDS
0,1 % (w/v) Ammoniumpersulfat
0,1 % (v/v) TEMED

Polyacrylamidgele - Sammelgel
190 mM Tris-Cl (pH 6,8)
5 % (w/v) Acrylamid
0,2 % (w/v) SDS
0,1 % (w/v) Ammoniumpersulfat
0,1 % (v/v) TEMED

Coomassie-Färbelösung
0,25 % (w/v) Coomassie Brilliant Blue R-250
45 % (v/v) Methanol
10 % (v/v) Essigsäure

Entfärbelösung
55 % (v/v) Ethanol
10 % (v/v) Eisessig

SDS-PAGE-Laufpuffer (10x)
250 mM Tris
1,92 M Glycin
1 % SDS (w/v)

Aufreinigung His-Alr3497

Lysepuffer
50 mM Tris-Cl (pH 8,0)
10 % (v/v) Glycerin
350 mM NaCl
0,2 % (v/v) NP-40
10 mM Imidazol
15 mM β-Mercaptoethanol

Waschpuffer
10 mM Tris-Cl (pH 7,5)

Elutionspuffer
10 mM Tris (pH 7,5)
750 bis 5 M NaCl

Lagerpuffer
40 mM Tris pH 8,0
50 % (v/v) Glycerin
10 mM $MgCl_2$
2 mM DTT

Aufreinigung Alr3496-His

Lysepuffer
50 mM Tris-Cl (pH 8,0)
10 % (v/v) Glycerin
1 M NaCl
10 mM Imidazol

Waschpuffer
50 mM Tris-Cl (pH 8,0)
10 % (v/v) Glycerin
1 M NaCl
100 mM Imidazol

Elutionspuffer
50 mM Tris-Cl (pH 8,0)
10 % (v/v) Glycerin
1 M NaCl
500 mM Imidazol

Lagerpuffer
10 mM Tris-Cl (pH 8,0)
50 % (v/v) Glycerin
150 mM NaCl

Aufreinigung T7 RNA Polymerase

Lysepuffer
25 mM Tris-Cl (pH 8,0)
500 mM NaCl
10 % (v/v) Glycerin
10 mM Imidazol
5 mM β-Mercaptoethanol

Waschpuffer A
25 mM Tris-Cl (pH 8,0)
500 mM NaCl
10 % (v/v) Glycerin
25 mM Imidazol
5 mM β-Mercaptoethanol

Waschpuffer B
25 mM Tris-Cl (pH 8,0)
300 mM NaCl
10 % (v/v) Glycerin
25 mM Imidazol
5 mM β-Mercaptoethanol

Elutionspuffer
25 mM Tris-Cl (pH 8,0)
300 mM NaCl
10 % (v/v) Glycerin
200 mM Imidazol
5 mM β-Mercaptoethanol

Lagerpuffer
40 mM Tris pH 8,0
50 % (v/v) Glycerin
100 mM NaCl
10 mM DTT
1 mM EDTA
0,1 % (v/v) Triton X-100

Aufreinigung RNase-Inhibitor

Lysepuffer
25 mM Tris-Cl (pH 7,5)
500 mM NaCl
10 % (v/v) Glycerin
10 mM Imidazol
5 mM β-Mercaptoethanol

Waschpuffer A
25 mM Tris-Cl (pH 7,5)
500 mM NaCl
10 % (v/v) Glycerin
20 mM Imidazol
5 mM β-Mercaptoethanol

Waschpuffer B
25 mM Tris-Cl (pH 7,5)
300 mM NaCl
10 % (v/v) Glycerin
20 mM Imidazol
5 mM β-Mercaptoethanol

Elutionspuffer
25 mM NaH_2PO_4 (pH 8,0)
300 mM NaCl
10 % (v/v) Glycerin
200 mM Imidazol
5 mM β-Mercaptoethanol

Lagerpuffer
25 mM Tris-Cl (pH 7,5)
50 % (v/v) Glycerin
100 mM KCl
10 mM DTT

Aufreinigung M-MLV Reverse Transkriptase

Lysepuffer
50 mM NaH_2PO_4 (pH 8,0)
300 mM NaCl
10 % (v/v) Glycerin
10 mM Imidazol

Waschpuffer
50 mM NaH_2PO_4 (pH 8,0)
300 mM NaCl
10 % (v/v) Glycerin
20 mM Imidazol

Elutionspuffer
25 mM NaH_2PO_4 (pH 8,0)
300 mM NaCl
10 % (v/v) Glycerin
200 mM Imidazol

Lagerpuffer
25 mM NaH_2PO_4 (pH 8,0)
150 mM NaCl
50 % (v/v) Glycerin

Aufreinigung *Taq* DNA-Polymerase

Lysepuffer
50 mM Tris-Cl (pH 8,0)
10 % (v/v) Glycerin
50 mM KCl
0,1 % (v/v) NP-40
1 mM EDTA (pH 8,0)
5 mM β-Mercaptoethanol

Lagerpuffer
20 mM Tris-Cl (pH 8,0)
50 % (v/v) Glycerin
100 mM NaCl
0,1 mM EDTA (pH 8,0)
1 mM DTT

ATPase-Assays

2x Malachitgrünlösung
0,03 % Malachitgrün-Oxalat
10 mM Natriummolybdat
0,05 % (v/v) Triton X-100
mit 0,7 M Salzsäure auf Endvolumen gebracht

ATPase-Assaypuffer (10x)
200 mM Tris-Cl (pH 7,5) oder MOPS (pH 6,5)
20 mM $MgCl_2$
60, 600 oder 1500 mM KCl oder NaCl

Unwinding-Assays & Nucleinsäurechaperonassays

Reaktionspuffer (10x)
200 mM HEPES (pH 7,6)
10 mM EDTA (pH 8,0)
10 mM $MgCl_2$
10 mM DTT
1 % (v/v) Triton X-100

Filter Binding-Assays

Reaktionspuffer (10x)
200 mM Tris-Cl (pH 7,5)
60 mM NaCl
50 mM β-Mercaptoethanol
10 mM EDTA

Chemische Quervernetzung

Cross Linking-Puffer (5x)
50 mM NaH_2PO_4 (pH 8,0)
750 mM NaCl
50 % (v/v) Glycerin

Alle autoklavierfähigen Lösungen wurden für 20 min bei 121 °C und 2 bar, oder ggf. durch Filtration (Porengröße 0,2 µm) sterilisiert.

2.1.3 Medien

Die folgenden Medien wurden in dieser Arbeit verwendet.

LB-Medium (nach Miller)
10 g/L Trypton
5 g/L Hefeextrakt
10 g/L NaCl

TB-Medium
12 g/L Trypton
24 g/L Hefeextrakt
0,4 % (v/v) Glycerin
170 mmol/L KH_2PO_4
720 mmol/L K_2HPO_4

Autoinduzierendes Medium (nach Studier, 2005)
10 g/L Trypton
5 g/L Hefeextrakt
5 g/L NaCl
25 mmol/L Na_2HPO_4
25 mmol/L KH_2PO_4
50 mmol/L NH_4Cl
5 mmol/L Na_2SO_4
0,5 % (v/v) Glycerin
0,05 % (w/v) Glucose
0,2 % (w/v) Lactose

Alle Medien wurden für 20 min bei 121 °C und 2 bar durch Autoklavieren sterilisiert. Die Zusätze zu TB-Medium und zum autoinduzierendem Medium wurden entweder ebenfalls autoklaviert oder durch Filtration (Porengröße 0,2 µm) sterilisiert, und erst kurz vor Verwendung zum Medium hinzugegeben.

2.1.4 Antibiotika

Die verwendeten Antibiotika sowie ihre jeweiligen Stammlösungs- und Arbeitskonzentrationen sind in Tabelle 2 angegeben.

Tabelle 2: Verwendete Antibiotika

Antibiotikum	*Konzentration der Stammlösung [mg/mL]*	*Arbeitskonzentration [µg/mL]*
Ampicillin	100	100
Kanamycin	50	50
Chloramphenicol	30	30
Gentamycin	20	20

2.1.5 Stämme

In Tabelle 3 dargestellt sind die in dieser Arbeit verwendeten Stämme, die Bezugsquelle sowie relevante Angaben zum Genotyp.

Tabelle 3: Verwendete *E.coli*-Stämme

Stamm	*Genotyp*	*Bezugsquelle*
DH5α T1^{R}	F- ϕ80lacZΔM15 Δ (lacZYA-argF)U169 recA1 endA1 hsdR17 (rk-, mk+) phoA supE44 λ-thi-1 gyrA96 relA1 tonA	Invitrogen/Life Technologies GmbH, Darmstadt
Top10	F– *mcr*A Δ (*mrr-hsd*RMS-*mcr*BC) Φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ (*ara leu*) 7697 *gal*U *gal*K *rps*L (StrR) *end*A1 *nup*G	Invitrogen/Life Technologies GmbH, Darmstadt
BL21 (DE3)	fhuA2 [lon] ompT gal (λ DE3) [dcm] ΔhsdS λ DE3 = λ sBamHIo ΔEcoRI-B int:: (lacI::PlacUV5::T7 gene1) i21 Δnin5	New England Biolabs GmbH, Frankfurt
T7 Express	fhuA2 lacZ::T7 gene1 [lon] ompT gal sulA11 R (mcr-73::miniTn10--Tet^{S})2 [dcm] R (zgb-210::Tn10--Tet^{S}) endA1 Δ (mcrC-mrr)114::IS10	New England Biolabs GmbH, Frankfurt
Rosetta II (DE3)	F^{-} ompT hsdS$_{B}$ (r_{B}^{-} m_{B}^{-}) gal dcm (DE3) pRARE2 (Cam^{R})	Novagen/Merck Millipore KGaA, Darmstadt
ArcticExpress (DE3)	B F^{-} ompT hsdS (r_{B}^{-} m_{B}^{-}) dcm^{+} Tet^{r} gal λ (DE3) endA Hte [cpn10 cpn60 Gent^{r}]	Agilent Technologies Deutschland GmbH, Böblingen

2.1.6 Enzyme

Alle Enzyme wurden von New England Biolabs (NEB) bezogen. Ausnahmen können Tabelle 4 entnommen werden

Tabelle 4: Nicht von NEB bezogene Enzyme

Enzym	*Hersteller*
HybriPol DNA Polymerase	Bioline GmbH, Luckenwalde
Moloney Murine Leukemia Virus Reverse Transkriptase	Abt. Molekulare Genetik, TU Kaiserslautern
MonsterScript Reverse Transcriptase	Epicentre/Biozym Scientific GmbH, Hessisch Oldendorf
My*Taq* DNA Polymerase	Bioline GmbH, Luckenwalde
Phusion High Fidelity DNA Polymerase	Finnzymes/Thermo Fisher Scientific GmbH, Dreieich
Rat Lung RNase Inhibitor	Abt. Molekulare Genetik, TU Kaiserslautern
T7 RNA Polymerase	Abt. Molekulare Genetik, TU Kaiserslautern

Details zu den selbstaufgereinigten Enzymen können dem Abschnitt 2.2.3.3 entnommen werden.

2.1.7 Antikörper

Für immunbiologische Arbeiten wurden Antikörper benutzt, die in Tabelle 5 aufgeführt sind.

Tabelle 5: Verwendete Antikörper

Antikörper	*Hersteller*
goat anti-mouse IgG-HRP	Santa Cruz Biotechnology, Inc, Heidelberg
mouse anti-CDB mAb	New England Biolabs, Frankfurt
mouse anti-GST mAb	Santa Cruz Biotechnology, Inc, Heidelberg
mouse anti-His Epitop-Tag mAb	Dianova GmbH, Hamburg

2.1.8 Kits

Folgende Kits wurden in dieser Arbeit verwendet:

Tabelle 6: Verwendete Kits

Kit	*Hersteller*
HiYield Plasmid Mini-Kit	Süd-Laborbedarf GmbH, Gauting
NucleoSpin® Gel and PCR Clean-up Kit	Macherey-Nagel GmbH & Co. KG, Düren
PureLink® HiPure Plasmid Filter Maxiprep Kit	Invitrogen/Life Technologies GmbH, Darmstadt
GeneArt Seamless Cloning & Assembly Kit	Invitrogen/Life Technologies GmbH, Darmstadt
QIAquick Nucleotide Removal Kit	Qiagen NV, Hilden

2.1.9 Radiochemikalien

Tabelle 7 können die Radiochemikalien entnommen werden, die in dieser Arbeit benutzt wurden.

Tabelle 7: Verwendete Radiochemikalien

Radiochemikalie	*Hersteller*
[α-^{32}P]-Adenosintriphosphat	Hartmann Analytic GmbH, Braunschweig
[α-^{32}P]-Deoxycytosintriphosphat	Hartmann Analytic GmbH, Braunschweig
[γ-^{32}P]-Adenosintriphosphat	PerkinElmer Inc, Rodgau

2.1.10 Marker und Größenstandards

Eine Übersicht zu den verwendeten Markern und Größenstandards kann Tabelle 8 entnommen werden.

Tabelle 8: Marker und Größenstandards

	Hersteller
DNA Ladder 100 bp	New England Biolabs GmbH, Frankfurt
DNA Ladder 1 kb	New England Biolabs GmbH, Frankfurt
ColorPlus Prestained Protein Ladder (10-230 kDa)	New England Biolabs GmbH, Frankfurt
ColorPlus Prestained Protein Marker, Broad Range	New England Biolabs GmbH, Frankfurt
PageRuler Prestained Protein Ladder	Thermo Fisher Scientific/Fermentas, St. Leon-Rot
Gel Filtration Calibration Kit LMW	GE Healthcare/Life Sciences GmbH, Freiburg

2.1.11 Software

Für bioinformatische Arbeiten wurden die Programme verwendet, die in Tabelle 9 verzeichnet sind.

Tabelle 9: Verwendete Software

Software	*Hersteller*	*Referenz*
DiGReF (Diversity-generating retroelement finder)	Abt. Genetik, TU Kaiserslautern	(Schillinger et al., 2012)
BLAST (Basic Local Alignment Search Tool)	National Center for Biotechnology Information	(Altschul et al., 1990)
PSI-BLAST (Position-specific iterated BLAST)	National Center for Biotechnology Information	(Altschul et al., 1997)
PHI-BLAST (Pattern-Hit initiated BLAST)	National Center for Biotechnology Information	(Zhang et al., 1998)
CLC Sequence Viewer v6.4	CLC bio, Aarhus/Dänemark	
Prophage Finder	Department of Biological Sciences, University of Wisconsin-Parkside	(Bose and Barber, 2006)
ACLAME/Prophinder	Laboratoire de Bioinformatique des Génomes et des Réseaux, Université Libre de Bruxelles	(Leplae et al., 2010)
PyMOL Molecular Graphics System, Version 1.5.0.4	Schrödinger, LLC, New York	(Schrodinger, 2010)
ImageJ	National Institutes of Health, Bethesda	(Schneider et al., 2012)

2.1.12 Geräte

Eine Auswahl der in dieser Arbeit verwendeten Geräte ist in Tabelle 9 aufgeführt.

Tabelle 10: Verwendete Geräte

Gerät	*Hersteller*
Cyclone Phosphor Imager	PerkinElmer Inc, Rodgau
Dot Blot 96 System	Biometra/Analytic Jena AG, Jena
Gel iX Imager Geldokumentationssystem	Intas Science Imaging Instruments GmbH, Göttingen
Innova Incubation Shaker 44/44R	New Brunswick/Eppendorf AG, Hamburg
Qubit® 1.0 Fluorometer	Invitrogen/Life Technologies GmbH, Darmstadt
Kühlbodenzentrifuge RC5B	Sorvall/Heraeus/Thermo Scientific, Dreieich
Kühlmikrozentrifuge 5415R	Eppendorf AG, Hamburg
Kühltischzentrifuge Z383K	Hermle Labortechnik GmbH, Wehingen
MaxQ4000 Incubation Shaker	Thermo Scientific, Dreieich
Membranpumpe MZ 2C NT +2AK	Vacuubrand GmbH & Co. KG, Wertheim
Mini Sub-Cell GT Horizontalgelektrophoresesystem	Bio-Rad Laboratories GmbH, München
Mini-PROTEAN® Tetra Cell Vertikalgelelektrophoresesystem	Bio-Rad Laboratories GmbH, München
MyCycler Thermocycler	Bio-Rad Laboratories GmbH, München
peqSTAR 2x Thermocycler	Peqlab Biotechnologie GmbH, Erlangen
Sonopuls HD 2200 Ultraschallhomogenisator	Bandelin Electronic GmbH & Co. KG, Berlin

2.1.13 Sonstiges

Alle weiteren Materialien sind in Tabelle 10 aufgeführt.

Tabelle 11: Sonstige Materialien

	Hersteller
Hybond ECL Nitrocellulosemembran	Amersham/GE Healthcare Life Science GmbH, Freiburg
Hybond N+ Nylonmembran	Amersham/GE Healthcare Life Science GmbH, Freiburg
Protino Ni-NTA-Agarose	Macherey-Nagel GmbH & Co. KG, Düren
Glutathione Superflow Resin	Qiagen NV, Hilden
Chitin Beads	New England Biolabs GmbH, Frankfurt
Grade DE81 Ion Exchange Paper	GE Healthcare Life Science GmbH, Freiburg

2.2 Methoden

2.2.1 Arbeiten mit DNA

2.2.1.1 Agarosegelelektrophorese

Zur analytischen und präparativen Auftrennung von DNA-Gemischen wurden standardmäßig 1 % (w/v) Agarosegele in 1x Tris/Acetat/EDTA-Puffer verwendet. Hierzu wurde 1 g Agarose in 100 mL 1x TAE-Puffer in einer Mikrowelle erhitzt, bis der Feststoff komplett gelöst war. Nach kurzem Abkühlen wurden 3 µL Ethidiumbromid (10 mg/mL, Endkonzentration 0,3 µg/mL) zur Lösung hinzugefügt; die Lösung wurde in eine abgedichtete Kammer gegossen, ein Kamm gesteckt, und die Elektrophorese nach Auspolymerisierung in 1x TAE bei 70 bis 100 V durchgeführt. Hierzu wurde die DNA-Probe zuvor mit einem Ladepuffer gemischt, der einen oder mehrere inerte Farbstoffe zur normaloptischen Kontrolle des Laufverhaltens enthielt. Die Visualisierung der DNA nach ausreichender Auftrennung erfolgte unter UV-Anregung.

2.2.1.2 Denaturierende Polyacrylamidgelelektrophorese (PAGE)

Zur hochauflösenden analytischen und präparativen Auftrennung von DNA-Gemischen wurden denaturierende Polyacrylamidgele zwischen 5 und 12 % (v/v) in 1x Tris-Borat-EDTA-Puffer benutzt. Es wurde eine 20 %ige Stammlösung aus 40 %igem Acrylamid/Bisacrylamid-Lösung (29 : 1), 10x TBE-

Pufferkonzentrat und Harnstoff (Endkonzentration 8 M) verwendet, die mit einer identisch zusammengesetzten Lösung ohne Acrylamid auf den gewünschten Acrylamidgehalt verdünnt wurde. Die Lösung wurde unmittelbar nach Zugabe von Ammoniumpersulfat (Endkonzentration 0,05 %) und TEMED (Endkonzentration 0,05 %) luftblasenfrei zwischen zwei Glasplatten gespritzt, ein Kamm gesteckt, und die Elektrophorese nach Auspolymerisierung in 1x TBE je nach Gelgröße bei 12 bis 24 Watt durchgeführt. Bei präparativen Gelen erfolgte die Visualisierung über UV-Shadowing, bei analytischen Gelen über radioaktive Markierung der Proben und Phosphor-Imaging.

Zur Rückgewinnung von DNA aus Polyacrylamidgelen wurde die Bande ausgeschnitten, zerkleinert und mit 3 Volumen PAGE-Elutionspuffer für mindestens 4 Stunden bei 4 °C inkubiert. Anschließend wurde die Lösung über einen Spritzenfilter (Porengröße 0,2 µm) gegeben und die DNA durch Zugabe von Ethanol herausgefällt. Das Pellet aus der anschließenden Zentrifugation wurde noch einmal kurz gewaschen und schließlich in einem geeigneten Volumen Wasser bzw. TE-Puffer aufgenommen, und die Konzentration über die Absorption der Probe bei 260 nm bestimmt.

2.2.1.3 Präparation von Plasmid-DNA

Für die Präparation von Plasmid-DNA aus *E.coli*-Zellen wurden wahlweise kommerzielle Kits verwendet, die in Tabelle 6 genannt sind, oder eine alkalische Lyse der Zellen mit anschließender Fällung und ggf. Phenol-Chloroform-Extraktion durchgeführt. Bei der Verwendung von Kits wurden stets die Herstellerangaben befolgt, während die DNA-Isolierung über alkalische Lyse nach einem leicht modifizierten Protokoll von Birnboim und Doly durchgeführt wurde (Birnboim and Doly, 1979; Sambrook and Russell, 2001). Zusammengefasst wurden Übernachtkulturen von transformierten DH5α-Zellen mittels Zentrifugation pelletiert und die Zellen in Lösung 1 gründlich resuspendiert. Die Lyse der Zellen erfolgte durch Zugabe von Lösung 2, die Neutralisierung mit Lösung 3. Denaturierte chromosomale DNA, Proteine und Membranbestandteile wurden mittels Zentrifugation von der renaturierten Plasmid-DNA abgetrennt. Die Zugabe von 0,7 Volumen Isopropanol bewirkte die Präzipitation der Plasmid-DNA, welche anschließend mit 70 % Ethanol gewaschen wurde und ggf. über Zugabe eines Volumens Phenol-Chloroform-Isoamylalkohol (25:24:1, pH 7,5-8,0) weiter aufgereinigt wurde. Die Quantifizierung der DNA erfolgte spektroskopisch bei 260 nm oder fluorometrisch mit dem Qubit® Fluorometer-System.

2.2.1.4 Transformation von E.coli

Zur Präparation von kompetenten Zellen wurde eine Übernachtkultur eines *E.coli*-Stammes 1:100 in LB-Medium verdünnt und bis zu einer optischen Dichte bei 600 nM (OD_{600}) von ca. 0,5 bis 0,7 bei

37 °C und 200 rpm wachsen gelassen (= logarithmische Wachstumsphase). Die Zellen wurden bei 6000 x g und 4 °C vom Kulturmedium abgetrennt, in eiskaltem TSB-Medium resuspendiert und für mindestens 1 Stunde auf Eis inkubiert. Anschließend wurden die Zellen aliquotiert, in flüssigem Stickstoff schockgefroren und bis zur weiteren Verwendung bei –80 °C gelagert. Die Transformationseffizienz wurde mit einem Standardvektor (pBS) bestimmt und betrug für Klonierungsstämme mindestens 1 x 10^7 Transformanden (*colony forming units*, CFU) pro µg DNA, für Expressionsstämme mindestens 1 x 10^5 CFU/µg DNA.

Zur Transformation wurden 10 bis 50 ng Plasmid-DNA in 1x KCM-Lösung (Gesamtvolumen 100 µL) zu 100 µL kompetenter Zellen gegeben, für 20 min auf Eis und weitere 10 min bei RT inkubiert. Bei Verwendung eines bakteriostatischen Selektionsantibiotikums wurden die Zellen direkt auf Agarplatten mit entsprechendem Antibiotikum ausplattiert, bei einem bakterioziden Selektionsantibiotikum zuvor für 1 Stunde mit 1 mL vorgewärmtem LB-Medium bei 37 °C und 200 rpm inkubiert, um eine Expression des Resistenzfaktors zu ermöglichen.

2.2.1.5 Klonierungen

Klonierungen wurden in dieser Arbeit größtenteils über den enzymatischen Verdau von Vektor und Insert und anschließender Ligation der Fragmente durchgeführt. Grundsätzlich wurden nur Restriktionsenzyme verwendet, welche keine glatten Enden, sondern 5‘-Überhänge produzieren, was „sticky end“-Klonierungen erlaubt. Hierzu wurden Vektor und Insert mit einem oder idealerweise zwei Restriktionsenzymen nach Herstellerangaben verdaut (mit einem mindestens zehnfachen Enzymüberschuss), mit dem NucleoSpin® Gel and PCR Clean-up Kit (Macherey-Nagel) von Enzymen und Salzen gereinigt und in einem molaren Insert-zu-Vektor-Verhältnis von 5:1 mit T4 DNA-Ligase (NEB) für mindestens 2,5 Stunden bei RT in 1x Ligasepuffer inkubiert. Anschließend wurde die Ligase für 20 min bei 65 °C inaktiviert, und das Ligationsprodukt für eine Transformation von *E.coli* DH5α-Zellen verwendet. Positive Transformanden wurden über Kolonie-Polymerasekettenreaktion (Kolonie-PCR) bestimmt (s. Abschnitt 2.2.1.6) und für die Inokulation einer Übernachtkultur verwendet. Plasmid-DNA wurde aus diesen Kulturen nach den Angaben aus Abschnitt 2.2.1.3 isoliert und über einen analytischen Verdau mittels geeigneter Restriktionsenzyme untersucht. Die Sequenz positiver DNA-Klone wurde abschließend über Sanger-Sequenzierung durch die Firma Seq-it GmbH, Kaiserslautern oder LGC Genomics GmbH, Berlin verifiziert.

Für einige Konstrukte, die in dieser Arbeit erzeugt wurden, wurde das GeneArt Seamless Cloning & Assembly Kit der Life Technologies GmbH verwendet, welches ein nicht näher erläutertes Rekombinationsverfahren nutzt, um ein Insert mit einem Vektorbackbone zu verknüpfen. Hierzu muss der jeweilige Zielvektor zunächst mit wahlweise ein oder zwei Restriktionsenzymen linearisiert

und von Enzymen und Puffern gereinigt werden. Das Insert wird mittels PCR erzeugt; hierbei werden an den Termini des Inserts Abschnitte von 15 bp angefügt, die homolog zu den Enden des linearisierten Zielvektors sind. Insert- und Vektor-DNA werden anschließend in einem molaren Verhältnis von 2:1 mit einem nicht näher beschriebenen Enzymgemisch für 30 Minuten bei RT inkubiert, und das Produkt für die Transformation von *E.coli* Top10-Zellen verwendet. Positive Transformanden wurden wie vorstehend beschrieben analysiert und weiterverarbeitet.

2.2.1.6 Polymerasekettenreaktionen (PCR)

Polymerasekettenreaktionen wurden in dieser Arbeit mit mehreren Enzymen durchgeführt. Für analytische PCR-Reaktionen wurde entweder die rekombinant erzeugte DNA-Polymerase aus *Thermus aquaticus* (*Taq*) oder das Enzym My*Taq* der Firma Bioline GmbH verwendet, während für präparative Arbeiten die Enzyme Phusion der Firma Thermo Scientific oder Hybripol von der Bioline GmbH eingesetzt wurden. Kommerzielle Enzyme wurden stets mit ihren zugehörigen Puffern nach Herstellerangaben verwendet; Reaktionen mit der rekombinant erzeugten *Taq*-Polymerase enthielten den in 2.1.2 genannten Puffer in 1x Konzentration. Weiterhin enthielten die Reaktionen standardmäßig alle vier Deoxyribonucleotide in einer Konzentration von jeweils 200 µM sowie die beiden Primer in einer Konzentration von jeweils 200 nM, sofern bei kommerziellen Enzymen nicht anders angegeben. Die Thermoprofile wurden ebenfalls an die Herstellerangaben angepasst; für Reaktionen mit der *Taq*-Polymerase wurde eine initiale Denaturierungsphase von 5 Min bei 98 °C gewählt, gefolgt von üblicherweise 25 Zyklen aus Denaturierung (10 s, 98 °C), Annealing (15 s, variable Temperatur) und Elongation (1 min/kbp, 72 °C). Nach einer finalen Elongationsphase von 10 min bei 72 °C wurden die Proben bei 4 °C bis zur weiteren Verwendung gelagert.

Als Template wurden standardmäßig 30 ng eines Plasmids oder 1 µg genomische DNA eingesetzt. Einen Sonderfall stellt die Kolonie-PCR dar, die als Screeningmethode bei Klonierungen Anwendung fand (s. 2.2.1.5). Hierbei wurden Bakterien mit einem sterilen Zahnstocher oder einer Pipettenspitze von einer Kolonie auf einer Agarplatte entnommen, in einem 0,2 mL-PCR-Gefäß abgestreift und anschließend mit dem entsprechenden PCR-Mix gemischt.

2.2.1.7 Fluorometrische Quantifizierung von DNA-Proben

Zur fluorometrischen Bestimmung geringer DNA-Konzentrationen (< 1 µg/µL) wurde die Qubit®-Plattform der Firma Invitrogen/Life Technologies GmbH, Darmstadt, mit dem Qubit® DNA BR Assay-Kit verwendet. Zu 199 µL einer Arbeitslösung, bestehend aus einem Puffer und einem Farbstoff,

wurde 1 µL einer DNA-Lösung gegeben, gemischt, für 10 Minuten bei RT im Dunkeln inkubiert und die Konzentration mit dem Qubit® Fluorometer 1.0 bestimmt.

2.2.1.8 Radioaktive Markierung von DNA-Oligonucleotiden (Endlabeling)

Die radioaktive Markierung von DNA-Oligonucleotiden wurde über Endlabeling durchgeführt. Hierzu wurde das Oligonucleotid (Endkonzentration ~0,66 µM, entspricht 10 pmol 5'-Termini) in einem Reaktionsvolumen von 15 µL mit 15 Richardson-Units T4 Polynucleotidkinase und 150 µCi [γ-^{32}P]-ATP für 1 Stunde bei 37 °C inkubiert. Anschließend wurden nicht-inkorporierte Nucleotide über Größenausschlussfiltration mit einer MobiSpin S-300-Säule oder mit dem QIAquick Nucleotide Removal Kit entfernt. Die Aktivität des markierten Oligonucleotids wurde per Szintillationsmessung bestimmt, die Lagerung erfolgte bei -20 °C.

2.2.1.9 Erzeugung von DNA:DNA-Duplices

Zur Hybridisierung zweier komplementärer DNA-Stränge wurden diese äquimolar oder ggf. mit einem zwei- bis dreifachen Überschuss eines der Stränge in 20 mM Tris-Cl, pH 7,5 und 50 mM NaCl für 1 Minute bei 95 °C denaturiert und anschließend langsam auf Raumtemperatur abgekühlt. Die Duplex wurde bei -20 °C gelagert.

2.2.2 Arbeiten mit RNA

2.2.2.1 Allgemeines

Für die Handhabung von RNA galten im Wesentlichen die gleichen Arbeitsvorschriften und Vorgehensweisen wie bei Arbeiten mit DNA. Um eine Kontamination mit ubiquitären Ribonukleasen (RNasen) zu vermeiden, wurden zudem folgende Punkte beachtet:

- Glaswaren wurden für 6 Stunden bei 220 °C gebacken.
- Alternativ wurden Glaswaren und Plastikwaren für mindestens 20 Minuten mit 3 % Wasserstoffperoxid behandelt.
- Zum Ansetzen von wässrigen Lösungen wurde Wasser verwendet, das mit Diethylpyrocarbonat (DEPC) vorbehandelt wurde. Hierzu wurde DEPC in einer Endkonzentration von 0,1 % zu Wasser gegeben, gemischt und für mindestens 16 Stunden bei 37 °C inkubiert. Anschließend wurde das Wasser autoklaviert, um restliches DEPC zu inaktivieren.

- Es wurden entweder sterile, nukleasenfreie serologische Einwegpipetten oder Filterspitzen für Pipettierungen verwendet.
- Bei sämtlichen Arbeitsschritten wurden Handschuhe getragen.
- Zur akuten Dekontaminierung von Arbeitsflächen, Pipetten und weiterem Arbeitsmaterial wurde RNase Away (Molecular Bioproducts/Thermo Fisher Scientific) verwendet.

Desweiteren wurde RNA soweit möglich bei – 80 °C gelagert, um thermisch induzierter Degradierung entgegen zu wirken.

2.2.2.2 in vitro-Transkription

Zur Erzeugung von RNA im präparativen Maßstab wurden *in vitro*-Transkriptionen mit rekombinant erzeugter T7 RNA-Polymerase durchgeführt. Hierzu wurde zunächst Plasmid-DNA, die die gewünschte Sequenz hinter einem T7-Promoter enthält, downstream der Sequenz linearisiert, um eine *run off*-Transkription zu ermöglichen. Der gereinigte Vektor wurde zusammen mit T7 Polymerase, RNase Inhibitor sowie Ribonucleotiden in 1x Transkriptionspuffer mit 10 mM DTT für mindestens 2 Stunden bei 37 °C inkubiert. Anschließend wurde der während der Reaktion gebildete Magnesium-Pyrophosphat-Komplex kurz abzentrifugiert und die RNA mit 3 Volumen Ethanol bei -80 °C aus dem Überstand gefällt. Anschließend erfolgte eine weitere Reinigung der RNA über präparative PAGE (s. Abschnitt 2.2.1.2). Die Konzentration der RNA-Präparationen wurde mittels UV-Spektrometrie bestimmt.

2.2.2.3 Fluorometrische Quantifizierung von RNA-Proben

Zur fluorometrischen Bestimmung geringer RNA-Konzentrationen (< 1 µg/µL) wurde die Qubit®-Plattform der Firma Invitrogen/Life Technologies GmbH, Darmstadt, mit dem Qubit® RNA HS Assay-Kit oder dem Qubit® RNA Assay-Kit verwendet. Zu 199 µL einer Arbeitslösung, bestehend aus einem Puffer und einem Farbstoff, wurde 1 µL einer RNA-Lösung gegeben, gemischt, für 10 Minuten bei RT im Dunkeln inkubiert und die Konzentration mit dem Qubit® Fluorometer 1.0 bestimmt.

2.2.2.4 Radioaktive Markierung von RNA (Bodylabeling)

Die radioaktive Markierung von RNA wurde in dieser Arbeit über Bodylabeling durchgeführt. Hierzu wurde eine *in vitro*-Transkriptionsreaktion wie in 2.2.2.2 beschrieben angesetzt, und die Reaktion mit [α-^{32}P]-ATP supplementiert (ca. 6 x 10^7 cpm pro 150 µL). Die Reaktion wurde für 2 Stunden bei 37 °C inkubiert und anschließend über Größenausschlussfiltration mit einer MobiSpin S-300-Säule von

nicht-inkorporierten Nucleotiden gereinigt. Die Konzentration der synthetisierten RNA wurde fluorometrisch bestimmt, die Aktivität über Szintillationsmessung. Die markierte RNA wurde bis zur weiteren Verwendung bei -20 °C gelagert.

2.2.2.5 Erzeugung von RNA:DNA-Duplices

Für die Erzeugung von Heteroduplices wurde der DNA-Strang in einem dreifachen Überschuss zum komplementären RNA-Strang gegeben und in 20 mM Tris-Cl, pH 7,5 und 50 mM NaCl für 1 Minute bei 95 °C denaturiert; anschließend wurde die Reaktion langsam auf Raumtemperatur abgekühlt. Die Duplex wurde bei -20 °C gelagert.

2.2.3 Arbeiten mit Proteinen

2.2.3.1 Denaturierende Polyacrylamidgelelektrophorese (SDS-PAGE)

Zur quantitativen und qualitativen Analyse von Proteinproben wurden Polyacrylamidgelektrophoresen unter denaturierenden Bedingungen durchgeführt; als denaturierendes Agens wurde den Gelen Natriumdodecylsulfat (SDS) zugefügt (Laemmli, 1970). Die verwendeten PA-Gele bestanden aus einem Trenngel mit variablem Acrylamidgehalt (zwischen 8 und 15 %) sowie einem darübergeschichteten Sammelgel, welches einen niedrigeren Acrylamidgehalt und einen niedrigeren pH-Wert aufweist. Die Zusammensetzungen der Gellösungen können Abschnitt 2.1.2 entnommen werden, die Polymerisierung wurde durch Zugabe von 0,08 % APS und 0,08 % TEMED initiiert. Zur verbesserten Auflösung wurde den Proteinproben (falls nicht anders angegeben) 100 mM β-Mercaptoethanol in SDS-Ladepuffer zugegeben, um intermolekülare Schwefelbrücken zu reduzieren. Proteinproben wurden mit Ladepuffer für 5 Minuten bei 95 °C denaturiert und sofort auf das Gel geladen; die elektrophoretische Auftrennung erfolgte nach Sicht bei 120 bis 160 Volt. Anschließend erfolgte die Visualisierung der Proteine über Färbung mit Coomassie Brilliant Blue R-250. In einigen Fällen wurde eine Schnellfärbung durchgeführt (Ortiz et al., 1992). Hierzu wurde das Gel je nach Prozentigkeit für 10 bis 20 Minuten zunächst in 0,2 M Imidazol, und anschließend für 30 bis 60 Sekunden in 0,3 M Zinkchlorid inkubiert. Dieses bildet einen weißlichen Komplex mit SDS und Imidazol, während Proteinbanden durchscheinend bleiben. Anschließend wurde das Gel mit destilliertem Wasser gewaschen und mit Coomassie gefärbt.

2.2.3.2 Aufkonzentration von Proteinproben

Für gering konzentrierte Proteinproben wurde vor Auftragung auf ein Polyacrylamidgel eine Acetonpräzipitation durchgeführt. Hierzu wurde eiskaltes Aceton in dreifachem Volumen zur Probe gegeben und gemischt, und für mindestens 30 Minuten bei -20 °C inkubiert. Nach Zentrifugation wurde der Überstand verworfen und das Pellet in einem geringen Volumen SDS-Ladepuffer aufgenommen.

2.2.3.3 Überexpressionen und Aufreinigungen rekombinanter Proteine

Für die Erzeugung und Isolierung rekombinanter Proteine aus *E.coli* wurde in der Regel das nachfolgende Protokoll verwendet. Die Variationen dieses Protokolls, welche sich für jedes Protein unterschieden, können Tabelle 12 entnommen werden. Die zugehörigen Puffer sind in Abschnitt 2.1.2 aufgeführt.

Tabelle 12: Variationen des Aufreinigungsprotokolls für einzelne Proteine

Protein	*Konstrukt*	*Expressionsstamm*	*IPTG-Konzentration*	*Expressionsbedingungen*
Alr3496	pTS31	*E.coli* T7 Express	0,1 mM	3h, 37 °C / ÜN, 16 °C
M-MLV-RT	pSD08	*E.coli* BL21 (DE3)	0,7 mM	3h, 37 °C
RNase Inhibitor	pSD91	*E.coli* Rosetta II (DE3)	0,25 mM	ÜN, 16 °C
T7 Polymerase	pSD84	*E.coli* BL21 (DE3)	0,4 mM	4h, 37 °C
***Taq* DNA Polymerase**	pSD20	*E.coli* BL21 (DE3)	0,5 mM	ÜN, 37 °C

Zunächst wurde ein geeigneter Expressionsstamm mit 10 bis 50 ng des jeweiligen Konstrukts transformiert und auf einem LB-Nährboden mit passendem Antibiotikum ausplattiert. Von dieser Platte wurden mehrere Kolonien zur Inokulation einer Vorkultur aus LB-Medium und Antibiotikum verwendet. Diese wurde nach Übernachtinkubation bei 37 °C und 200 rpm 1:100 verdünnt, unter gleichen Bedingungen inkubiert, und die Bakteriendichte regelmäßig über die Absorption bei 600 nm bestimmt. Zwischen einer optischen Dichte von 0,4 bis 0,7 wurde die Kultur ggf. auf Expressionstemperatur abgekühlt und die Expression durch Zugabe von IPTG gestartet. Zuvor wurde 1 mL der Kultur entnommen, die Bakterien pelletiert, in 2x SDS-Ladepuffer resuspendiert und bis zur weiteren Verwendung bei -20 °C gelagert (= Probe der uninduzierten Kultur). Nach Expression wurde die Dichte der Kultur erneut gemessen, die gleiche Zellzahl wie zuvor entnommen und weiterverarbeitet (= Probe der induzierten Kultur). Der Rest der Expressionskultur wurde bei 4 °C und 6000 x g pelletiert und entweder frisch weiterverarbeitet, oder bei -20 °C bzw. -80 °C gelagert. Die Lyse der Zellen erfolgte auf Eis in einem geeigneten Volumen Lysepuffer mit dem Bandelin HD 2200 Ultraschallhomogenisator, ausgerüstet mit der KE76-Sonotrode mit insgesamt 10 Pulsen zu je 10

Sekunden bei ca 50 % Output. Zwischen den einzelnen Pulsen lagen Kühlpausen von mindestens 20 Sekunden, bzw. von 5 bis 10 Minuten zwischen dem fünften und sechsten Puls. Lösliche Bestandteile wurden von unlöslichen über Zentrifugation bei 12.000 bis 14.000 x g für 30 min bei 4 °C abgetrennt. Von beiden Fraktionen wurden Proben genommen, mit 2x SDS-Ladepuffer versetzt und bis zur weiteren Verwendung bei -20 °C gelagert. Der Überstand wurde mit einer geeigneten Matrix inkubiert, die zuvor mit mindestens 5 Säulenvolumen (CV) Lysepuffer equilibiert worden war. Anschließend wurde das Gemisch auf eine Säulenvorrichtung gegeben, der Durchfluss verworfen, nachdem zuvor eine Probe entnommen und wie zuvor beschrieben weiterverarbeitet worden war, und die Matrix unter geeigneten Bedingungen gewaschen. Von der Matrix wurde eine Probe genommen und mit 2x SDS-Probenpuffer versetzt, und gebundenes Protein zu mehreren Fraktionen eluiert. Abschließend wurde eine weitere Probe der Matrix genommen. Die Fraktionen wurden entweder über SDS-PAGE analysiert, oder – bei etablierten Aufreinigungsprotokollen und meist hoher Reinheit des Proteins – über ein Schnellverfahren, bei dem 10 µL jeder Fraktion auf eine Nitrocellulosemembran aufgetragen und getrocknet wurde. Anschließend wurde die Membran in 1 % Ponceau S inkubiert, überschüssiger Farbstoff mit destilliertem Wasser weggewaschen und Fraktionen, die eine Färbung zeigten, für die weitere Verwendung gesammelt. In der Regel schloss sich eine Dialyse gegen einen geeigneten Lagerpuffer an; hierzu wurden die Fraktionen in einen Dialyseschlauch passender Größe gefüllt und über Nacht bei 4 °C gegen das mindestens 200-fache Probenvolumen dialysiert. Anschließend wurden eventuell vorhandene Präzipitate durch Zentrifugation bei 4 °C für 30 min bei 12.100 x g entfernt, die Proteinkonzentration des Überstandes bestimmt und die dialysierte Probe bei geeigneter Temperatur gelagert.

Die Aufreinigung der *Taq* DNA-Polymerase erfolgte nicht über Affinitätschromatographie. Stattdessen wurde nach Lyse und Separation der löslichen von den unlöslichen Bestandteilen der Überstand für 1h bei 75 °C inkubiert. Hierbei wurde der Großteil der Proteine aus *E.coli* denaturiert, während die thermostabile Polymerase in Lösung blieb. Nach einem weiteren Zentrifugationsschritt (30 min bei 14.000 x g und 4 °C) wurden 30 % (w/v) Ammoniumsulfat zum Überstand hinzugefügt, erneut zentrifugiert, das Pellet in einem geringen Volumen Lysepuffer resuspendiert und über Nacht gegen Lagerpuffer dialysiert.

2.2.3.4 Fluorometrische Quantifizierung von Proteinproben

Zur fluorometrischen Bestimmung der Proteinkonzentration wurde die Qubit®-Plattform der Firma Invitrogen/Life Technologies GmbH, Darmstadt, mit dem Qubit® Protein Assay-Kit verwendet. Zu 199 µL einer Arbeitslösung, bestehend aus einem Puffer und einem Farbstoff, wurde 1 µL einer

Proteinlösung gegeben, gemischt, für 30 Minuten bei RT im Dunkeln inkubiert und die Konzentration mit dem Qubit® Fluorometer 1.0 bestimmt.

2.2.3.5 Immunologische Analysen mittels Western Blots

Zur eindeutigen Bestimmung von Fusionsproteinen wurden Immunoblots (Western Blots) durchgeführt. Hierzu wurden Proteingemische zunächst per SDS-PAGE wie in Abschnitt 2.2.3.1 beschrieben aufgetrennt und anschließend mit dem Tank Blot-Modul des Bio-Rad Mini-PROTEAN® Tetra Cell-Systems auf eine Nitrocellulose- oder Polyvinylidenfluoridmembran (PVDF) übertragen. Die Transfereffizienz wurde über Färbung der Membranen mit Ponceau S überprüft. Bei PVDF-Membranen erfolgte vor dem Transfer und vor Inkubation mit Ponceau S eine kurze Aktivierung der Oberfläche mit Methanol. Da die Transfereffizienz der DGR-assoziierten Proteine zunächst sehr gering ausfiel, wurde der Transferpuffer für die Übertragung von Proteinen mit hoher Basizität angepasst(Szewczyk and Kozloff, 1985). Unspezifische Bindungsstellen wurden mit einem Gemisch aus 3 % Bovinem Serum Albumin (BSA) und 5 % fettfreier Milch in phosphatgepufferter Kochsalzlösung (PBS) entweder für 1h bei RT oder über Nacht bei 4 °C abgeblockt. Die Inkubation mit den jeweiligen Primärantikörpern erfolgte in geeigneter Verdünnung im Blockingreagenz bei RT für 1h. Die Membranen wurden anschließend mehrmals mit PBS gewaschen und anschließend wiederum für 1h bei RT mit einem Meerrettich-Peroxidase-gekoppelten Sekundärantikörper inkubiert. Nach erneutem Waschen erfolgte die Visualisierung mit dem ECL plus-Kit von Amersham/GE Lifescience auf Röntgenfilmen.

2.2.3.6 Chemische Quervernetzung (Cross-Linking) von Proteinen

Um den *in vitro*-Oligomerisierungszustand des akzessorischen Proteins Alr3496 zu bestimmen, wurden Quervernetzungsexperimente mit Di (N-succinimidyl)-3,3'-dithiopropionat (DSP) analog zu Willmund et al. durchgeführt (Willmund et al., 2008). Hierzu wurden variierende Mengen rekombinant erzeugten Proteins (zwischen 0,5 und 13,5 µg) in 1x Cross-Linking Puffer mit variierenden Konzentrationen (zwischen 20 und 1800 µM) des Cross-Linkers für 30 min bei 25 °C inkubiert und die Quervernetzung durch Zugabe von je 1 µL 4,4 M ε-Aminocapronsäure und weitere Inkubation für 15 min bei 25 °C gestoppt. Zu den Ansätzen wurde ein gleiches Volumen 2x SDS-Ladepuffer ohne β-Mercaptoethanol gegeben und die Probe ggf. bis zur Analyse per SDS-PAGE bei -20 °C gelagert.

2.2.3.7 Bestimmung der Proteinstabilität

Um sicherzustellen, dass die Reaktionsbedingungen der hier durchgeführten Versuche (d. h. die Zusammensetzungen der verwendeten Reaktionspuffer) nicht zu einer vorzeitigen Präzipitation der getesteten Enzyme führen, wurden diese zuvor auf äquivalente Weise mit den Reaktionspuffern inkubiert, und die Proteinkonzentrationen nach kurzer Zentrifugation bei 12.100 x g fluorometrisch bestimmt. Reaktionsbedingungen wurden als geeignet angesehen, wenn die gemessene Konzentration um weniger als 10 % unter der initialen Konzentration lag.

2.2.4 Sonstige Methoden

2.2.4.1 RT-Aktivitätsassays

Zur Bestimmung der Aktivität rekombinant erzeugter und aufgereinigter reverser Transkriptasen wurde in dieser Arbeit der Assay nach Matsuura et al. in modifizierter Variante eingesetzt(Matsuura et al., 1997). Gereinigte Proteine wurden in einer 10 µL-Reaktion mit einem RNA:DNA-Duplex (1 µM:3 µM), Deoxyribonucleotiden (jeweils 500 µM) und [α-^{32}P]-Deoxycytosintriphosphat (ca. 0,25 µCi pro 10 µL-Reaktion) mit einem variablen Reaktionspuffer für 90 min bei 25 °C inkubiert. Als Positivkontrolle diente entweder eine kommerzielle reverse Transkriptase (MonsterScript von Epicentre) oder die RT aus dem Moloney Murine Leukemia Virus (M-MLV-RT). Diese wurden bei 37 °C für 90 min inkubiert. Als Negativkontrollen wurden jeweils identische Ansätze verwendet, bei denen das Enzym gegen ein gleiches Volumen Wasser ausgetauscht worden war. Außerdem wurde den Reaktionen RNase Inhibitor zugegeben, um die Template-RNA vor Degradation durch RNase-Kontaminationen zu schützen. Anschließend wurden 5 µL der Reaktion auf DE81-Filterscheiben aufgetragen und diese nach einer Trockenzeit von mindestens 1 Stunde mehrmals mit 100 mM Na_2HPO_4 und abschließend einmal mit absolutem Ethanol gewaschen(Kuchta, 1996). Das verwendete Filtermaterial bindet höhermolekulare Nucleinsäuren mit sehr hoher Affinität, während kleinere Fragmente und Einzelnucleotide nicht gebunden werden. Somit kann durch Szintillationsmessung der Aktivität inkorporierter dCTP-Nucleotide ein direkter Rückschluss auf die cDNA-Syntheserate bzw RT-Aktivität in den Reaktionsansätzen gezogen werden.

2.2.4.2 Filter-Binding Assays

Die Fähigkeit des Proteins Alr3496, verschiedene Nucleinsäuresubstrate zu binden, wurde über den modifizierten Filter-Binding Assays nach Wong & Lohman untersucht(Wong and Lohman, 1993). Variierende Proteinmengen wurden in 12 µL-Reaktionen zu radioaktiv-markierten Nucleinsäuresubstraten in Reaktionspuffer gegeben, die in mindestens zehnfachem Unterschuss zur

niedrigsten eingesetzten Proteinkonzentration vorlagen (üblicherweise 10 nM). Untersucht wurde die wahrscheinlich native TR-RNA des *Nostoc* sp. PCC 7120-DGRs sowie die kontextfremde UnaL2 3‘-UTR-RNA, ssDNA, dsDNA sowie ein Heteroduplex aus dem 3‘-Ende der TR-RNA, der eine ausgeprägte Hairpinstruktur sowie die wahrscheinliche IMH beinhaltet, und einer hierzu vollständig komplementären DNA, sowie einer DNA, die die Mismatches der korrespondierenden variablen Region aufweist (s. Abbildung 29). Inkubationstemperaturen und –zeiten wurden wie angegeben variiert, und je 10 µL der Reaktionen auf ein Membransandwich in einer Dot-Blot-Apparatur gegeben. Das Sandwich bestand aus einer oberen Nitrocellulosemembran, die für Proteine und proteingebundene Nucleinsäuren selektiv ist, sowie einer unteren Nylonmembran, die Nucleinsäuren bindet (s Abbildung 5).

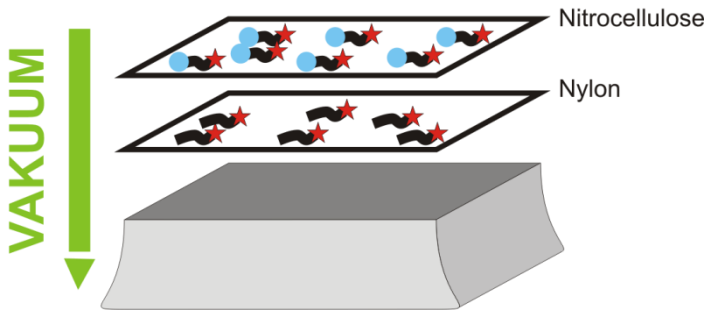


Abbildung 5: Schematischer Aufbau des Membransandwiches, modifiziert nach Wong & Lohman. Proteingebundene und freie Nucleinsäuren werden auf eine obere Nitrocellulosemembran appliziert und vom angelegten Vakuum angesaugt. Proteingebundene Nucleinsäuren verbleiben auf der Nitrocellulose, während freie Nucleinsäuren von der unteren Nylonmembran gebunden werden.

Anschließend wurden die Spots mit einem ca. hundertfachen Volumen Reaktionspuffer gewaschen, die Membranen getrocknet und radioaktive Signale über eine Storage Phosphor-Imagerplatte detektiert. Diese wurde mit dem Cyclone Storage Phosphor-Imager ausgelesen, und die erhaltenen Autoradiogramme mit der Software ImageJ ausgewertet. Die Intensitäten der radioaktiven Signale korrelieren direkt mit der Menge der gebundenen Nucleinsäure; die Intensität der Nitrocellulosemembranen geben somit Aufschluss über die Menge an proteingebundener Nucleinsäure, während die Signale der Nylonmembran die Menge der freien Nucleinsäure wiedergibt. Die Dissoziationskonstante des Proteins für das jeweilige Nucleinsäuresubstrat ergab sich aus der Proteinkonzentration, bei der 50 % des Signals proteingebunden vorlag.

2.2.4.3 Size-Exclusion-Chromatographie

Zur Bestimmung des nativen *in vitro*-Oligomerisierungszustands von Alr3496 wurde eine Size-Exclusion-Chromatographie des Proteins in der Abteilung für Molekulare Biophysik der Technischen Universität Kaiserslautern durchgeführt. Ein Probenvolumen von 200 µL einer Proteinpräparation (1 mg/mL) in 50 mM NaH_2PO_4 (pH 8,0), 500 mM NaCl und 20 % Glycerol wurde mit Hilfe eines Äkta Purifiers 10 (GE Healthcare) über eine Superdex 75 10/300 GL-Säule mit einer Flußrate von 0,5 mL/min gegeben, und Fraktionen zu 300 µL gesammelt. Über die Absorption bei 280 nm konnten proteinhaltige Fraktionen bestimmt, und ein Elutionsprofil ermittelt werden. Das Molekulargewicht des nativen Komplexes wurde über Extrapolation mit dem Gel Filtration Calibration LMW Standard (GE Healthcare) ermittelt.

2.2.4.4 ATPase-Assays

In diesen Experimenten sollte ermittelt werden, ob Alr3496 eine ATPase-Aktivität aufweist. Hierzu wurde ein ATPase-Assay nach Carter et al. durchgeführt, welcher das bei der Hydrolyse von ATP freigesetzte Phosphat für eine Nachweisreaktion nutzt(Carter and Karl, 1982). Dieses kann im Komplex mit Molybdat an Malachitgrün unter Bildung eines schwer löslichen, grünen Komplexes binden. Die Intensität der Färbung kann mit einem ELISA-Reader bei 620 nm gemessen werden, und ist direkt proportional zur Menge des freien Phosphats im Reaktionsansatz; zur rechnerischen Bestimmung der Phosphatkonzentration wurde jeweils eine Eichgerade aus den Messwerten von vier Standardlösungen bekannter Phosphatkonzentrationen erstellt.

Variiert wurden die Reaktionsbedingungen hinsichtlich Enzymkonzentration, Pufferzusammensetzung und der Zugabe verschiedener Nucleinsäurespezies. Enzymkonzentrationen von 3, 10 und 30 µM wurden mit den in Abschnitt 2.1.2 aufgeführten Reaktionspuffern in 1x Konzentration sowie ggf. Nucleinsäuren in – soweit möglich – äquimolaren Konzentrationen in einem Reaktionsvolumen von 19 µL für 10 min bei 37 °C vorinkubiert. Anschließend wurde den Reaktionen ATP zugegeben (Endkonzentration 1 mM) und für 1 Stunde bei 37 °C inkubiert. Anschließend wurden jeweils 4 µL der Reaktion zu 200 µL 1x Malachitgrünlösung gegeben, und diese Ansätze für 10 min bei RT inkubiert. Die Menge des gebildeten Phosphats wurde durch Messung der Absorption bei 620 nm und Abgleich mit der Eichkurve bestimmt. Von diesen Werten wurden Eigenbeiträge der ATP-Lösung und der Proteinpräparationen allein subtrahiert (s. Tabelle 20).

2.2.4.5 Unwinding-Assays

Um eine mögliche Helicaseaktivität nachzuweisen, wurde zusätzlich zu den ATPase-Assays untersucht, ob Alr3496 in der Lage ist, einen DNA-Doppelstrang in Einzelstränge aufzutrennen.

Hierzu wurde Oligonucleotid NZ116 mit dem in Abschnitt 2.2.1.8 beschriebenen Protokoll radioaktiv markiert und mit einem zweifachen Überschuss des komplementären Oligonucleotids NZ117 wie in Protokoll 2.2.1.9 beschrieben hybridisiert. Dieser Duplex wurde mit variierenden Konzentrationen Alr3496 und 1 mM ATP in 1x Unwinding-Reaktionspuffer in einem Endvolumen von 10 µL für 10 min bei 37 °C inkubiert und die Reaktion durch Zugabe von 10 µL 2x Ladepuffer für native PAGE gestoppt. Ein Teil der Reaktionen wurde auf ein 12 %iges Polyacrylamidgel aufgetragen und unter nativen Bedingungen elektrophoretisch aufgetrennt. Das Gel wurde getrocknet, die Detektion radioaktiver Signale erfolgte über eine Storage Phosphor-Imagerplatte. Diese wurde mit dem Cyclone Storage Phosphor-Imager ausgelesen. Als Negativkontrollen wurden in dem Assay Wasser und BSA statt Protein eingesetzt, zur Kontrolle des Laufverhaltens wurden markierter Einzelstrang und Duplex aufgetragen.

2.2.4.6 Nucleinsäurechaperon-Assays

Es wurde untersucht, ob Alr3496 in der Lage ist, die Hybridisierung zweier komplementärer DNA-Stränge zu katalysieren. Hierzu wurde das Oligonucleotid NZ116 wie in Abschnitt 2.2.1.8 erläutert radioaktiv markiert und in einem Assay analog zu Martin et al. mit dem komplementären Oligonucleotid NZ117 und variierenden Konzentrationen Alr3496 in 1x Unwinding-Reaktionspuffer in einem Volumen von 10 µL für 10 min bei 37 °C inkubiert (Martin and Bushman, 2001). Anschließend wurde die Reaktion durch Zugabe von 2x Ladepuffer für native PAGE gestoppt, die Reaktionen auf ein 12 %iges Polyacrylamidgel aufgetragen und eine Elektrophorese durchgeführt. Abschließend wurde das Gel getrocknet und eine Storage Phosphor-Imagerplatte aufgelegt. Das Autoradiogramm wurde mit dem Cyclone Storage Phosphor-Imager ausgelesen. Auch hier dienten Wasser und BSA als Negativkontrollen, Duplex und Einzelstrang als Referenz.

III. Ergebnisse

Zahlreiche Aspekte diversitätsgenerierender Retroelemente sind noch immer ungeklärt, da bisher lediglich zu einem bestimmten Element, dem aus dem *Bordetella* Bakteriophagen, detailliertere Studien veröffentlicht worden sind. Aus diesem Grunde sollten in dieser Arbeit einerseits bioinformatische Strategien eingesetzt werden, um eine systematische Suche nach DGRs in öffentlichen Datenbanken durchzuführen und mittels vergleichender Genomanalysen (comparative genomics) Gemeinsamkeiten, Unterschiede und darüber hinausgehende Informationen zu diesen Elementen herauszuarbeiten. In einem zweiten Vorhaben sollten zwei essentielle Faktoren von DGRs, die reverse Transkriptase und ein akzessorisches Protein, rekombinant erzeugt, aufgereinigt und biochemisch charakterisiert werden.

3.1 Bioinformatische Analysen

3.1.1 DiGReF, ein bioinformatisches Tool zur Analyse diversitätsgenerierender Retroelemente

Diversitätsgenerierende Retroelemente können anhand von zwei Merkmalen identifiziert werden: zum einen durch eine reverse Transkriptase, zum anderen durch Repeatstrukturen in der genomischen Nachbarschaft des RT-Gens, die sich ausschließlich in ihrem Adeningehalt unterscheiden dürfen. Beide Merkmale müssen zusammenkommen, damit von einem DGR gesprochen werden kann, da es durchaus denkbar ist, dass Repeatstrukturen dieser Art auch natürlichweise im Genom von Prokaryoten oder Phagen vorkommen können. Umgekehrt wurde gezeigt, dass reverse Transkriptasen essentieller Bestandteil von DGRs sind, und nicht zuletzt die namensgebenden Faktoren dieser Retroelemente darstellen. Während reverse Transkriptasen relativ einfach durch BLAST-Suchen aufgefunden werden können, ist die manuelle Suche nach Repeatstrukturen äußerst arbeitsintensiv und fehleranfällig. Aus diesem Grunde wurde in Zusammenarbeit mit der Abteilung Genetik der TU Kaiserslautern das PERL-basierte Programm DiGReF entwickelt, das genomische Sequenzen auf DGR-typische Repeatstrukturen untersuchen kann. Dieses Programm wurde von Professor Dr. John A. Cullum, Dipl.-Biol. Mohamed Lisfi und

Jingyun Chi, MSc. konzipiert und geschrieben; Teil dieser Arbeit war zunächst die Evaluierung des Programms hinsichtlich Effizienz, Fehleranfälligkeit und Vollständigkeit der Suchergebnisse.

Erste Programmversionen implementierten Suchstrategien, die nach identischen Repeats einer definierten Länge (20 nt) in der genomischen Umgebung von RT-Genen suchten. Wurde ein solcher Repeat gefunden, wurden die Sequenzen und zusätzliche 50 Nucleotide up- und downstream extrahiert. Diese Sequenzen wurden in einem zweiten Schritt mit einem Programm wie ClustalW alignt, wodurch substitutierte Positionen sichtbar gemacht werden konnten. Bei manueller Durchsicht der Ergebnisse zeigte sich jedoch, dass dieses Verfahren eine geringe Sensitivität in der Identifizierung von DGRs aufwies. Häufig finden sich in genuinen DGRs keine zusammenhängenden Repeatsequenzen einer Länge von 20 Nucleotiden, so dass diese Elemente vom Programm nicht als DGR erkannt wurden. Ebenso suchten diese frühen Programmversionen lediglich nach Repeat*paaren*, was DGRs mit zwei oder mehreren VRs effektiv ausschloss. Über die manuelle Kontrolle des Outputs war es somit möglich, die Spezifität und Sensitivität des Programms sukzessive zu verfeinern, und die Analysestrategie zu optimieren.

In der finalen Version erhält DiGReF als Input eine oder mehrere GI-Nummern von putativen DGR-RTs. Auf genomischer Ebene werden jeweils 5000 Nucleotide up- und downstream der zugehörigen offenen Leserahmen extrahiert, und die Gesamtsequenz aus 10.000 Nucleotiden zzgl. RT-ORF mittels eines Sliding Windows von 50 nt Länge auf VR/TR-Strukturen untersucht. Dies bedeutet, das ein Substring von 50 nt Länge gewählt und mit dem Rest der Sequenz abgeglichen wird. Zuvor werden sämtliche Adeninpositionen des Substrings durch „Wildcards“ ersetzt, das heißt, sie dürfen beim Abgleich mit dem Rest der Sequenz jeder der vier Basen A, T, C oder G entsprechen. Gibt es eine Übereinstimmung, wird das Suchfenster inkrementell um 1 nt erweitert und mit dem ermittelten Substring verglichen, bis ein Mismatch auftritt; das Sequenzpaar wird als TR/VR ausgegeben. Liegt keine derartige Übereinstimmung vor, wird ein weiterer 50 nt-Substring als Suchmotiv gewählt, der relativ zum vorigen um 1 nt versetzt ist.

Grundsätzlich ist es möglich, mit DiGReF ganze Genome auf TR/VR-Sequenzen zu durchsuchen. Der Rechenaufwand, der nötig wäre um sämtliche Genome der öffentlich zugänglichen NCBI-Datenbank zu durchsuchen, wäre jedoch beträchtlich. Aus diesem Grunde wurde für eine systematische Suche nach DGRs vor Einsatz des DiGReF-Programms eine Vorauswahl von Kandidatengenomen und der in Frage kommenden Sequenzabschnitte vorgenommen. Kandidaten-RTs wurden über eine PSI-BLAST-Suche ermittelt, die die Sequenzen acht bekannter DGR-RTs als Query verwendete. Anschließend wurden über DiGReF Repeatstrukturen identifiziert und die Ergebnisse einer manuellen Qualitätskontrolle unterzogen (s. Abbildung 6).

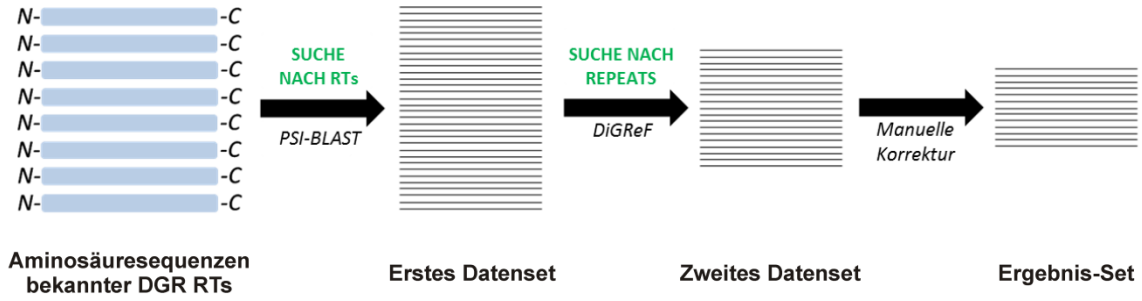


Abbildung 6: Schematische Darstellung der DGR-Suche. Über eine PSI-BLAST-Suche mit bekannten DGR-RTs als Query wird ein erstes Datenset erzeugt. Mit Hilfe des Programms DiGReF werden DGR-typische VR/TR-Strukturen in einer Umgebung von 5000 bp gesucht und in einem zweiten Datensatz ausgegeben. Eventuelle Redundanzen werden in einem letzten Schritt manuell entfernt.

3.1.1.1 Auswahl von Kandidatengenomen mittels PSI-BLAST

In den Veröffentlichungen von Doulatov et al., Medhekar & Miller sowie Simon & Zimmerly wurde bis heute zumindest die Existenz von etwa 40 DGRs erwähnt(Doulatov et al., 2004; Medhekar and Miller, 2007; Simon and Zimmerly, 2008), wenngleich die Elemente selbst – bis auf das prototypische *Bordetella* Bakteriophagen-DGR – nie genauer beschrieben worden sind. Von acht Vertretern dieser Elemente wurden die Primärsequenzen der reversen Transkriptasen ausgewählt. Wie bereits zuvor für andere Retroelemente extensiv beschrieben (Simon and Zimmerly, 2008), ist die jeweilige RT sehr gut geeignet, das zugehörige Element spezifisch zu identifizieren und phylogenetisch zu charakterisieren. Darüber hinaus stellt die RT das konstanteste Element eines DGRs dar: während die Zielproteine eine starke Variabilität aufweisen und natürlich auch nicht notwendigerweise immer mit einem DGR verbunden sein müssen, ist die Anwesenheit eines akzessorischen Proteins offenbar nicht in allen Fällen erforderlich.

Um ein möglichst umfassendes Set von DGRs zu erhalten, wurden die acht RTs so ausgewählt, dass sie ein breites Herkunftsspektrum aufweisen (s. Tabelle 13). Somit sollte ausgeschlossen werden, ausschließlich nah verwandte Elemente aus z. B. einem ökologisch isolierten Habitat zu erhalten.

Tabelle 13: Query-RTs der PSI-BLAST-Suche

RT	*GI-Nummer*	*Organismus*	*Klasse*	*Erstmals beschrieben*
Npun_F4892	186684985	*Nostoc punctiforme* PCC 73102	Cyanobakterien	(Doulatov et al., 2004)
Dred_1227	134299090	*Desulfotomaculum reducens* MI-1	Firmicutes	(Simon and Zimmerly, 2008)
RUMOBE_01080	149833092	*Ruminococcus obeum* ATCC 29174	Firmicutes	(Simon and Zimmerly, 2008)
LPC_1855	148359926	*Legionella pneumophila* str. Corby	Gammaproteobakterien	(Simon and Zimmerly, 2008)
Tery_1035	113474819	*Trichodesmium erythraeum* IMS101	Cyanobakterien	(Doulatov et al., 2004)
TDE2266	42527768	*Treponema denticola* ATCC 35405	Spirochaeten	(Doulatov et al., 2004)
VAS14_16052	90580666	*Photobacterium angustum* S14	Gammaproteobakterien	(Simon and Zimmerly, 2008)
bbp5	41179367	*Bordetella* phage BPP-1	Phagen	(Liu et al., 2002)

Die acht RTs wurden jeweils als Query für eine PSI-BLAST-Suche verwendet. PSI-BLAST ist eine Variation des BLAST-Algorithmus, bei der nach einem ersten Suchlauf, der einer normalen BLASTp-Suche entspricht, eine Scoring-Matrix aus vom Benutzer ausgewählten Treffern erstellt wird. Diese Matrix dient in der darauffolgenden Suche (Iteration) als Query, um die Suche weniger über die Querysequenz als vielmehr über das Profil der besten Resultate zu beeinflussen. Mit Hilfe dieses Verfahrens können weiter entfernte Mitglieder einer Proteinfamilie besser gefunden werden als mit einer konventionellen BLASTp-Suche. Es wurden zwei Iterationen mit den jeweils besten 30 Treffern als Querys durchgeführt, nach diesen veränderte sich das erhaltene Set nicht mehr in relevanter Weise. Um die Zahl der Kandidaten-RTs weiter einzugrenzen wurde manuell geprüft, ob es eine Grenze in den BLASTp-Ergebnissen gibt, die genuine DGR-RTs von den restlichen trennt. BLAST-Ergebnisse erhalten stets einen E-Wert, der die Wahrscheinlichkeit beschreibt, mit dem verwendeten Query rein durch Zufall einen Treffer in der vorliegenden Datenbank zu erhalten. Es zeigte sich, dass alle erhaltenen RTs mit einem E-Wert von 0,005 oder kleiner genuine DGR-RTs darstellten, so dass alle Treffer mit einem schlechteren E-Wert in der weiteren Analyse nicht betrachtet wurden. Das erste Datenset bestand an diesem Punkt aus 2651 potentiellen DGR-RTs, basierend auf der NCBI-Datenbankversion vom November 2011.

3.1.1.2 DiGReF-Analyse

In einem nächsten Schritt wurde das Programm DiGReF benutzt, um unter den erhaltenen 2651 RTs tatsächliche DGR-RTs, und somit die gesamten DGR-Elemente, zu identifizieren. Als Mindestkriterien

für TR/VR-Paare wurden nach verschiedenen Testläufen 50 Basenpaare Gesamtlänge mit 10 mutierbaren Adeninen im Template Repeat und 7 Adeninaustauschen in der variablen Region gewählt. Eine Lockerung dieser Kriterien verminderte die Spezifität des Programms und führte zu einer deutlichen Anhäufung von anderen Retroelementen wie Gruppe-II-Introns; lediglich 6 wahrscheinliche DGRs, die eine relativ geringe Mutageneseaktivität aufwiesen, konnten zusätzlich ermittelt werden.

Über DiGReF konnten 171 DGRs identifiziert werden. Nach manueller Bearbeitung des Sets, bei der redundante Einträge entfernt wurden, blieben 155 DGRs, von denen 126 zuvor noch nicht beschrieben worden waren. Dieses Set wurde im Anschluss extensiven Analysen unterzogen, die in den Abschnitten 3.1.2.1 bis 3.1.2.7 beschrieben werden. Eine vollständige Auflistung der Elemente findet sich in Anhang A1.

3.1.1.3 Suche nach atypischen DGRs

Die bisher publizierten DGRs wiesen ausschließlich Adeninaustausche auf. Um der Frage nachzugehen, ob es auch Elemente mit einer anderen Basenspezifität gibt, wurde das Programm so modifiziert, dass entweder nach Cytosin-, Guanin- oder Thyminaustauschen gesucht wurde, wobei die Kandidatensequenzen aus 3.1.1.1 mit ansonsten gleichen Suchparametern analysiert wurden. Es wurden keine entsprechenden Elemente gefunden, allerdings förderte diese Suche drei Elemente zutage, bei denen TR und VR auf dem relativ zur RT komplementären Strang liegen, und somit als Repeats mit Thyminaustauschen erkannt wurden. Bezogen auf die TR-RNA handelt es sich also wiederum um Adeninaustausche. Auf diese Elemente wird in Abschnitt 3.1.2.1 noch einmal eingegangen werden.

Ebenso wurde geprüft, ob es DGRs mit RT-Proteinen gibt, die eine geringere Sequenzidentität zu den Query-RTs aufweisen, und mit ihnen somit auch weniger nah verwandt sind. Es wurden einige RTs mit höheren E-Werten aus den Ergebnissen aus 3.1.1.2 ausgewählt und wiederum als Query für eine BLASTp-Suche verwendet. Hierbei wurden keine neuen Treffer im Vergleich zu den Ergebnissen aus 3.1.1.1 erhalten. Somit kann davon ausgegangen werden, dass das erhaltene Set von DGRs vollständig ist.

3.1.2 Auswertung des Datensets aus 155 DGRs

3.1.2.1 Strukturvielfalt der DGRs

Mithilfe eines geeigneten Programms, wie beispielsweise Artemis (Rutherford et al., 2000), ist es möglich, den Output, der von DiGReF generiert wurde, graphisch darzustellen. Eine Umwandlung des Outputs in ein Artemis-kompatibles Format wurde mit dem zusätzlichen Programm *output_artemis.pl*, das von Dipl.-Biol. Mohamed Lisfi ebenfalls in PERL geschrieben wurde, durchgeführt. Der Quelltext dieses Programms findet sich im Anhang dieser Arbeit. Abbildung 7 zeigt beispielhaft einen Screenshot von einem DGR, der mit Artemis visualisiert wird.

Abbildung 7: Visualisierung der DGR-Struktur über Artemis. Der von DiGReF erzeugte Output kann in ein Artemis-kompatibles Format umgewandelt werden, was eine übersichtliche Darstellung der DGR-Komponenten ermöglicht. Man erkennt VR (rot), TR (grün, unten in Detailansicht angewählt) und den RT-ORF (hellblau). Es ist zudem möglich, offene Leserahmen in der Sequenz anzeigen zu lassen (cyan), was eine einfache Zuordnung der variablen Region zu einem Zielgen ermöglicht.

Wie bereits in vorigen Studien berichtet wurde(Doulatov et al., 2004; Medhekar and Miller, 2007), können diversitätsgenerierende Retroelemente eine Vielzahl von Arrangements der Einzelkomponenten aufweisen. Dies wird in den Ergebnissen dieser Arbeit bestätigt und darüber hinaus durch bisher unbeschriebene DGR-Typen ergänzt (s. Abbildung 8). So können TR und RT, wie im Falle des DGRs aus *Rhodomicrobium vannielii* ATCC 17100 (GI 312115534) durch ein Zielgen räumlich voneinander getrennt vorliegen; ähnlich verhält es sich bei den Elementen aus *Photobacterium angustum* S14 (GI 90580666), *Shewanella baltica* OS155 (GI 126090247) und *Vibrio* sp. RC586 (GI 262403399). Hier liegen TR und RT-ORF sogar auf jeweils unterschiedlichen Strängen. Diese vier Fälle, in denen die jeweiligen variablen Regionen der Zielgene lediglich Adeninsubstitutionen aufweisen und die Elemente somit aktiv sind, zeigen, dass TR und RT nicht

notwendigerweise zusammen transkribiert werden, wie es bei einigen anderen Retroelementen der Fall ist.

Weiterhin fällt auf, dass ~15% aller identifizierten DGRs mehrere Zielgene über einen einzigen Template Repeat hypermutieren können. In der Standardanalyse wurden bis zu drei Zielgene pro DGR beobachtet, während bei einer extensiveren Suche bis zu vier VRs gefunden wurden, die sich über 300 Kilobasenpaare vom DGR entfernt befinden können (*Pseudogulbenkiania* sp. NH8B, GI 347538767). Denkbar ist, dass sich bei genomweiter Anwendung von DiGReF noch mehr Zielgene finden lassen, was hier aus Gründen der Rechenkapazität nicht überprüft wurde.

Die Vielfalt der DGR-Strukturen macht eine systematische Nomenklatur der Elemente notwendig, die idealerweise offen gehalten wird für zukünftige Ergänzungen des hier beschriebenen Sets und sich gleichzeitig an einem eindeutigen Merkmal der Elemente orientiert. In dieser Arbeit wurde ein System aus zunächst vier Gruppen eingeführt, die sich über die relative Anordnung von TR und RT zueinander definieren. Gruppe 1 bilden somit alle DGRs, bei denen der TR upstream von der RT liegt, während er bei den Elementen der Gruppe 2 downstream von dieser lokalisiert ist. Gruppe 3 umfasst DGRs, bei denen der offene Leserahmen der RT partiell oder vollständig mit dem TR überlappt. Gruppe 4 beinhaltet schließlich den hier erstmalig beschriebenen DGR-Typus, bei dem TR und RT-ORF auf getrennten Strängen vorliegen. Als sekundäres Klassifizierungsmerkmal, angezeigt durch kleine lateinische Buchstaben, die den Gruppennummern nachgestellt werden, dient die Anzahl und Position der Zielgene, und somit der variablen Regionen. Einzelheiten zu den Strukturtypen sowie der Repräsentation in unserem Ergebnis-Set können Abbildung 8 entnommen werden.

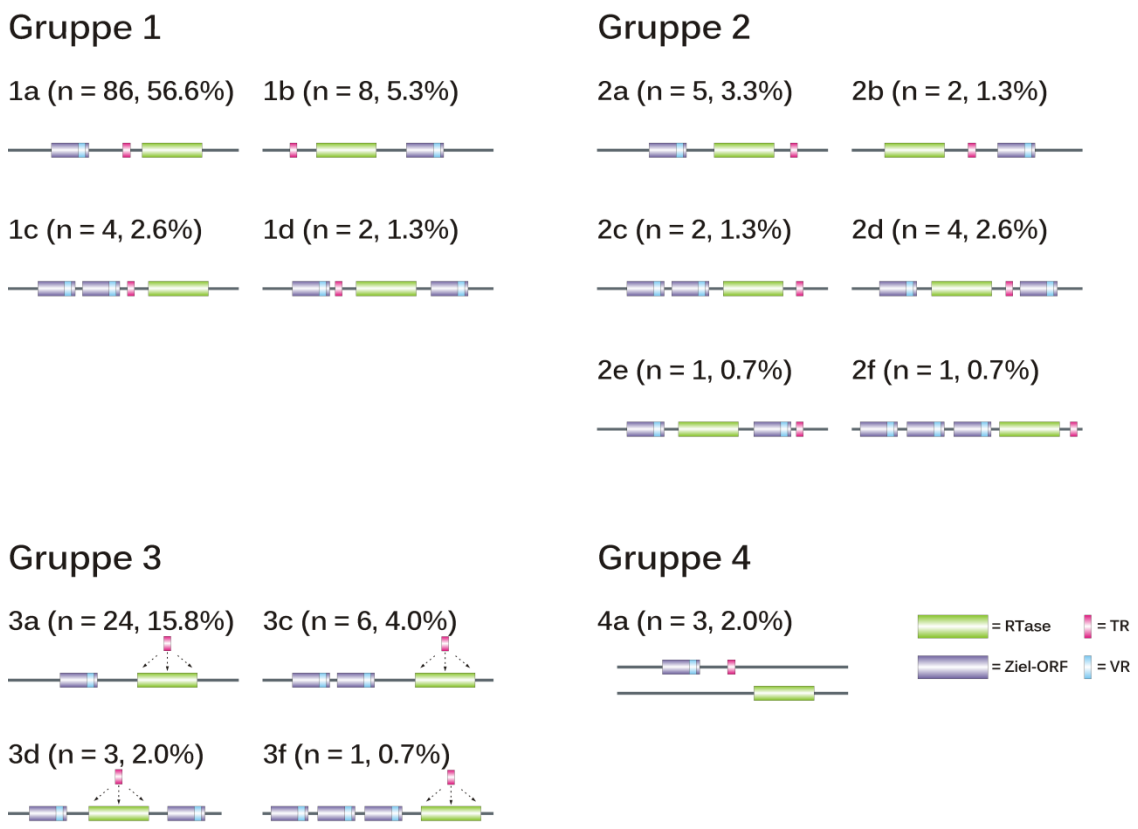


Abbildung 8: Vorgeschlagenes Klassifikationssystem für DGRs. Die im Rahmen dieser Arbeit identifizierten Elemente weisen eine Vielzahl von strukturellen Arrangements auf. Zur vereinfachten Beschreibung wird ein System vorgeschlagen, welches DGR-Elemente anhand der relativen Position von Template Repeat und RT-ORF in vier Gruppen einteilt: Elemente, mit einer TR upstream (1) oder downstream (2) des RT-ORFs, und Elemente, bei denen RT-ORF und TR teilweise oder vollständig überlappen (3). Eine vierte Gruppe wird durch die in dieser Arbeit erstmalig beschriebenen Elemente gebildet, bei denen TR und RT-ORF auf unterschiedlichen Strängen lokalisiert sind. Das zweite Identifikationsmerkmal bilden Position und Zahl der mutierten Zielgene, angegeben durch kleine lateinische Buchstaben.

Trotz zahlreicher Strukturvariationen entfällt der größte Anteil der hier beschriebenen DGRs auf das Arrangement des prototypischen Elements aus dem *Bordetella*-Bakteriophagen (Strukturtyp 1a, 56,6%). Eine weitere umfangreiche Gruppe wird vom Strukturtyp 3a gebildet (15,8%), bei dem die TR teilweise oder ganz mit dem RT-ORF überlappt. Hier muss jedoch angemerkt werden, dass automatische Annotationen von Genomen häufig Fehler bei der Festlegung der 5‘-Termini von Genen machen; die Leserahmen werden artifiziell verlängert, da weiter upstream gelegene ATG-Codons, die *in frame* mit dem Rest des Gens sind, als Startcodon angenommen werden. Dies kann in diesem Fall dazu führen, dass TR und RT-ORF nur scheinbar überlappen, und einige Elemente des Typs 3a eigentlich auch zu Typ 1a gezählt werden müssten (natürlich gilt dasselbe für die Typen 3b, 3c, usw.). Die hier angegebene Mächtigkeit der Gruppe 1a bildet also somit nur die Untergrenze, und es kann davon ausgegangen werden, dass weitere Elemente hinzugezählt werden können.

3.1.2.2 DGRs können mit mehreren Zielgenen verbunden sein

Bereits im vorangegangen Abschnitt wurde erwähnt, dass ein einziges DGR mehrere Zielgene, die fern im Genom liegen können, hypermutieren kann. Da dies über einen einzigen Template Repeat geschieht, ist es plausibel anzunehmen, dass auch die Zielgene selbst eine hohe Homologie zueinander aufweisen. Die Sequenzidentitäten wurden über das Programm *needle*, das Teil des EMBOSS-Softwarepakets ist, ermittelt. Sie betrugen zwischen 42,6% und 81,3%, und erfüllten somit die Kriterien für Paralogie, die von Blattner et al. definiert wurden als mindestens 30% Sequenzidentität über mindestens 60% der Nucleotidsequenz. Ebenso wurde untersucht, ob sich genomweit noch weitere Paraloge der Zielgene finden lassen, die nicht vom DGR hypermutiert werden. Hierzu wurde eines der Zielgene als Query für eine BLASTp-Suche gegen das jeweilige Genom verwendet, und die erhaltenen Proteine mit *needle* auf ihre Sequenzidentität überprüft. In allen Fällen bis auf zwei (*Burkholderia glumae* BGR1, GI 238027140 und *Desulfobacter postgatei* 2ac9, GI 355363092) fanden sich tatsächlich Leserahmen, die diese Bedingungen erfüllten. Die mögliche biologische Bedeutung dieses Befunds wird in Kapitel 4 dieser Arbeit näher erläutert.

3.1.2.3 TR/VR-Paare und Adeninaustausche

Das Mutagenesepotential eines DGRs hängt unmittelbar vom Adeningehalt des zugehörigen Template-Repeats ab: je mehr mutierbare Adenine vorhanden sind (besonders, wenn sie sich an erster oder zweiter Stelle des Codons befinden), desto eher können auf Proteinebene Aminosäureaustausche stattfinden, und die biochemischen Eigenschaften des Zielproteins verändert werden. Zudem erscheint es plausibel anzunehmen, dass kritische Ober- und Untergrenzen für die Längen von TR und VR existieren: sind sie zu kurz, kann nur ein kleiner Bereich des resultierenden Proteins verändert werden; sind sie zu lang, werden u. U. zu große Teile des Proteins mutiert und seine Funktion beeinträchtigt. Bisher beschriebene TRs und VRs zeigten unterschiedliche Längen: im Falle des *Bordetella* Bakteriophagen lag sie bei 134 bp, während andere DGRs mit teilweise unter 100 Nucleotiden deutlich kürzere Repeats aufwiesen. In unserem Ergebnis-Set fanden sich Längen zwischen der willkürlich gewählten Mindestlänge der Suchsequenz von 50 bp (*Clostridium symbiosum* WAL-14163, GI 323485235) und 184 bp (*Bacteroides fragilis* 638R, GI 301163046), mit einer mittleren Länge von 100±50bp. Mithilfe des PERL-Scripts *output_graph.pl*, das von Dipl.-Biol. Mohamed Lisfi geschrieben wurde, können graphische Darstellungen von VR und TR erzeugt werden, die zudem die Basengehalte beider Sequenzen beinhalten (s. Abbildung 9, PERL-Script siehe Anhang).

```
>TR1/35255--35305/          AATGTCACTAAATTAGGTGCTAATTGGAATAATGGCTTGAATACCAGTGC          A: 17 C: 7  G: 11 T: 16
                            xx||||x||xx|||x|||||||x||||x||x||||||||||||x||x||||
>VR1/35027--35077/          CGTGTCGCTGTATTGGGTGCTAGTTGGTATTATGGCTTGAATGCCGGTGC          A: 7  C: 8  G: 18 T: 18
```

Abbildung 9: Graphische Darstellung von VR/TR-Paaren. Das PERL-Script output_graph.pl extrahiert VR/TR-Paare und erlaubt einen Direktvergleich der Sequenzen, hier am Beispiel des DGRs aus Clostridium symbiosum WAL-14163, GI 323485235. Mismatches werden mit einem ‚X' markiert, identische Basen mit ‚|'. Außerdem werden die Basengehalte der Sequenzen ausgegeben, was eine Berechnung der Adeninsubsti-tutionen und der Hypermutationsrate ermöglicht.

Aus diesen konnten wiederum die Gesamtzahl der mutierbaren Adenine in der VR, die Zahl der veränderten Adenine in der VR sowie die Basen, zu denen diese mutiert wurden und schließlich eine Hypermutationsrate ermittelt werden, die sich aus dem Verhältnis der mutierten Adenine zur Gesamtzahl der Adenine ergibt. Die Ergebnisse können Abbildung 10 sowie Tabelle A1 entnommen werden.

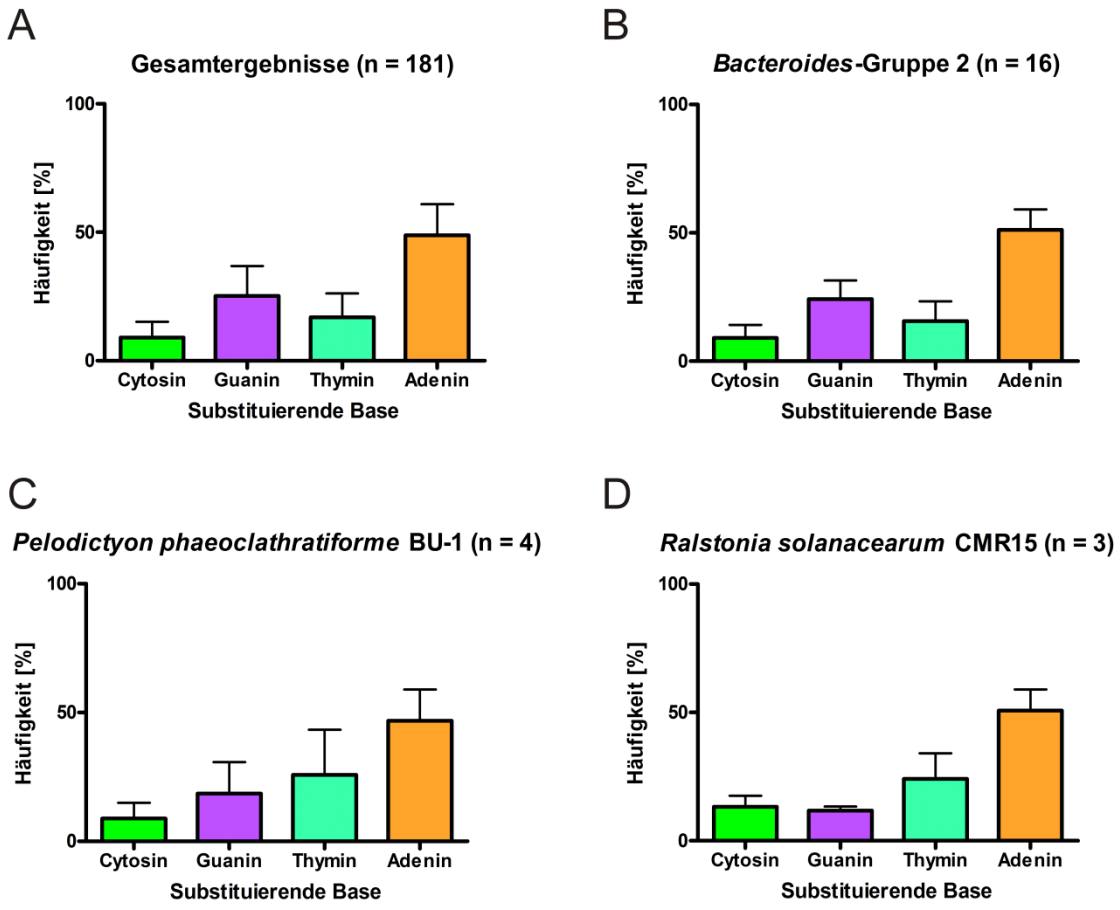


Abbildung 10: Basensubstitutionen in DGRs. (A) Die Template Repeats und variablen Regionen aller 155 ermittelten DGRs wurden auf A->N-Mutationen untersucht, und die Häufigkeiten ermittelt, mit denen Cytosin, Guanin, Thymin und Adenin an einer adeninkorrespondierenden Position in der VR auftraten. (B) Um einen eventuellen Einfluss von Wirtsfaktoren herauszufiltern, wurden die Substitutionsmuster einer phylogenetisch definierten Gruppe (des *Bacteroides*-Clusters 2 aus Abbildung 12) untersucht. (C)/ (D) Gleichermaßen wurden einzelne Organismen mit multiplen VRs analysiert. Dargestellt sind die Mittelwerte der relativen Häufigkeiten einer Base in n variablen Regionen, die Fehlerbalken geben die zugehörige Standardabweichung wieder. Man erkennt, dass letztere meist so hoch ausfällt, dass keine eindeutige Aussage hinsichtlich eines signifikant stärker vertretenen Substitutionsmusters gemacht werden kann.

Es wurden stark variierende Hypermutationsraten zwischen 0,21 und 0,86 beobachtet, der Median liegt bei 0,5. Hier ist jedoch zu beachten, dass ein Adeninrest in der VR entweder durch fehlerfreie reverse Transkription entstanden sein kann, und es somit nie ein Mutageneseereignis an dieser Position gab, oder durch präferentielle Selektivität des DGR-Mechanismus für Adenine. Diese Selektivität könnte eine intrinsische Eigenschaft der beteiligten RT sein, oder eventuellen Wirtsfaktoren zuzuschreiben, die bislang jedoch noch nicht bekannt sind. Bei Fehlen einer Selektivität würde man einen proportionalen Anstieg der A->B-Mutationen mit der Zahl der Adenine im entsprechenden TR erwarten, und ein Basenverhältnis von annähernd 1:1:1:1. Wie man Abbildung 10A und Tabelle A1 jedoch entnehmen kann, entsprechen die vorliegenden Substitutionsmuster diesen Erwartungen nicht. Es wurden insgesamt bevorzugt G-Substitutionen, gefolgt von T- und schließlich C-Substitutionen beobachtet, wenngleich ohne statistisch-signifikante Ausprägungen (s. Fehlerbalken). Ebenso konnte ein χ^2-Test zeigen, dass zumindest im Falle von G- und T-Substitutionen eine Abweichung von einer reinen Normalverteilung, und somit ein möglicher Einfluss von Wirtsfaktoren vorliegt ($p = 0.00008$ bzw. 0.0057). Als nächstes wurde eine Untergruppe von DGRs betrachtet, die jeweils aus *Bacteroides* sp. stammten. Hier sollten eventuelle Wirtsfaktoren jeweils identisch sein und eine eventuelle Basenpräferenz in den Ergebnissen deutlicher herauszulesen sein. Wie Abbildung 10B allerdings zeigt, lässt sich auch hier keine statistisch gesicherte Aussage hinsichtlich einer bestimmten Basenpräferenz treffen; ebenso verhält es sich für Basenaustausche eines einzigen DGRs mit mehreren VRs (Abbildung 10C und D). Zusammenfassend konnte an dieser Stelle gezeigt werden, dass Wirtsfaktoren zumindest keinen besonders starken Einfluss auf die Basenpräferenz eines DGRs haben. Ebenso weisen die vorliegenden Daten darauf hin, dass die Substitutionsmuster nicht ausschließlich einer Normalverteilung folgen, und ein reiner Zufallsmechanismus ebenfalls ausgeschlossen werden kann.

Es ist weiterhin fraglich, ob eine solche Art der Analyse überhaupt zu einem auswertbaren Ergebnis führen kann. Es muss berücksichtigt werden, dass durch einen Selektionsdruck, der zwar nicht unter Kultivierungsbedingungen, jedoch in der offenen Natur vorherrscht, ein Bias in das Substitutionsmuster eingebracht wird und dieses verzerrt. Um dieses Problem zu lösen, wäre eine dynamische Betrachtung eines oder mehrerer einzelner DGR-Systeme notwendig, welche zudem auch ohne einen Anpassungsdruck aktiv sind.

3.1.2.4 DGR-RTs

Die Proteinsequenzen der reversen Transkriptasen aus dem Ergebnis-Set wurden extrahiert und mit MAFFT alignt. Von Simon & Zimmerly wurde für einige wenige DGR-RTs eine Struktur aus sieben

Domänen beschrieben, die auch in den reversen Transkriptasen anderer Elemente gefunden werden können; der bemerkenswerteste Unterschied ist das Fehlen einer RNaseH-Domäne im Falle der DGR-RTs. Diese Architektur findet sich auch bei sämtlichen RT-Proteinen, die in dieser Arbeit ermittelt wurden; somit wird diese Einteilung als ein generelles Strukturmerkmal dieser RT-Klasse bestätigt. Konservierte Bereiche innerhalb dieser Domänen sind das für reverse Transkriptasen typische YxDD-Motiv in Domäne 5, das hier in 86% der Fälle entweder als YMDD oder YVDD enthalten ist, und ein Motiv aus sieben Aminosäuren in Domäne 4 des Proteins: (L/I/V)GxxxSQ. Dieses Motiv wurde zuvor bereits in fast allen DGR-RTs identifiziert, aufgrund des geringen Datenbestandes war es jedoch nicht klar, ob es sich hierbei tatsächlich um ein exklusives Merkmal handelt, oder lediglich phylogenetisch nah beieinander liegende RTs untersucht wurden. In dieser Arbeit wiesen über 90% dieses Motiv auf, wobei Variationen zumeist an den beiden letzten Positionen (SQ) vorliegen. In 9,7% der RTs finden sich die Kombinationen VQ, NQ, AQ, SP, SH oder PA. Darüber hinaus wurde bei Vergleich mit den PSI-BLAST-Ergebnissen, die als Input für DiGReF dienten (s. Abschnitt 3.1.1.1) und mit bekannten RTs aus anderen Retroelementen wie Retroviren, Gruppe-II-Introns und Non-LTR-Retrotransposons festgestellt, dass das Consensus-Motiv, wie es dem Sequenzlogo in Abbildung 11 entnommen werden kann, DGR-RTs deutlich von anderen Retroelementen abgrenzt, die zumeist ein QGxxxSP-Motiv an dieser Stelle aufweisen.

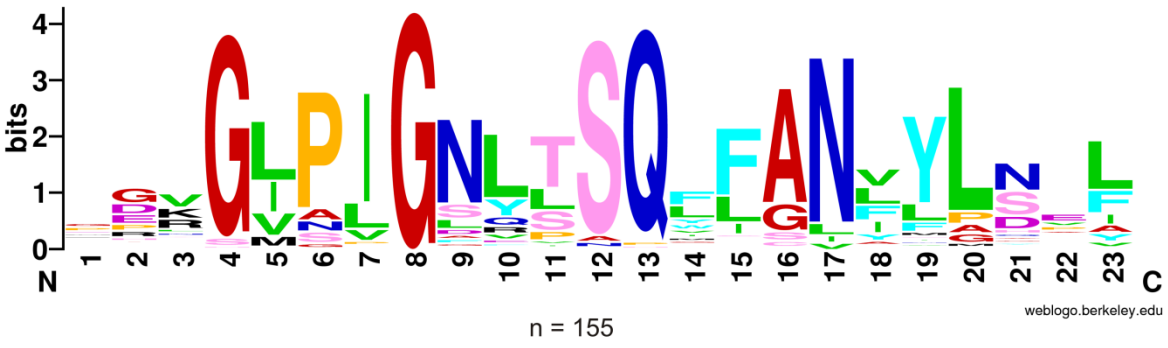


Abbildung 11: Sequenzlogo der DGR-RTs. Das Logo veranschaulicht graphisch die relativen Häufigkeiten von bestimmten Aminosäuren an den dargestellten Positionen 1 bis 23 aus Domäne 4 in 155 DGRs, gemessen in Bits. Im Falle des DGRs aus dem *Bordetella* Bakteriophagen entspricht das Motiv den Aminosäurepositionen 175 bis 197. Chemisch-ähnliche Aminosäuren besitzen g leiche Farben. Man erkennt eine hohe Konservierung aliphatischer Reste an Position 7 sowie des Glycins an Position 8, und das DGR-typische SQ-Motiv an den Positionen 12 und 13.

Aus früheren Arbeiten zur HIV-RT ist bekannt, dass der Prolinrest, der in DGR-RTs meist durch Glutamin ersetzt ist, an der Koordination von dNTPs und dem Template bei cDNA-Synthese beteiligt ist, während Mutationen an dieser Position zu einer generell verminderten Synthesegenauigkeit des

Enzyms führt. Es ist daher möglich, dass dieses DGR-spezifische Motiv für das prominenteste Merkmal dieser Elemente, nämlich adeninspezifische Hypermutation, verantwortlich sein könnte. Hierbei scheinen auch die anderen beobachteten Kombinationen für die beiden terminalen Reste zulässig zu sein, da nach Durchsicht der TR/VR-Sequenzen dieser Elemente weder eine geringere Hypermutationsrate, noch eine Anhäufung von B->N-Mutationen festgestellt werden konnte. Dies wäre zu erwarten bei inaktivem Status dieser Elemente aufgrund der natürlichen Mutationsvorgänge in Genomen.

Umgekehrt fanden sich 28 reverse Transkriptasen mit einem intakten SQ-Motiv in Domäne 4, in deren Nähe DiGReF keine VR/TR-Strukturen ermitteln konnte. Bei Durchsicht dieser Elemente stellte sich heraus, dass die zugehörigen Leserahmen entweder auf sehr kurzen Contigs lagen, oder nahe am Ende eines Contigs. Somit konnten nicht auf beiden Seiten 5000bp extrahiert und durchsucht werden. Die wahrscheinlichste Erklärung ist hier also, dass es sich um echte DGRs handelt, sie als solche aufgrund der ungenügenden Sequenzierfragmentlängen allerdings nicht erkannt werden können. In 6 Fällen (GIs 212705095, 212705094, 227485655, 227486150, 260439164, 338175397) handelte es sich nicht um intakte RT-ORFs, so dass es sich wahrscheinlich um defekte DGRs handelt. In mindestens zwei Fällen (GIs 17232506 und 189468311) war zunächst nicht klar, warum keine TR/VR-Sequenzen entdeckt wurden: die Leserahmen lagen nicht auf kurzen Contigs, oder nahe an den 5‘- oder 3‘-Enden eines Sequenzierfragments, und zeigten auch keine auffälligen Aberrationen, die zu einer Inaktivierung der RTs geführt hätte. Bei manueller Durchsicht der flankierenden Sequenzen fanden sich jedoch TR/VR-ähnliche Strukturen, die jedoch zahlreiche nah beieinanderliegende B->N-Mutationen aufwiesen, und somit DiGReFs Analysefensterweite von 50bp unterliefen. Es ist unklar, ob diese Elemente erst kürzlich inaktiviert wurden, oder ob es sich um Elemente handelt, die eine intrinsisch-höhere Fehleranfälligkeit für alle Basenspezies, und nicht nur für Adenine im Speziellen, aufweisen.

3.1.2.5 Phylogenetische Analyse

Als nächstes wurde ermittelt, in welchen Bakterienklassen DGRs auftauchen, ob es eine signifikante Anreicherung in bestimmten Klassen gibt, und ob Auffälligkeiten in der Phylogenie der Elemente auftreten. Zunächst wurde gezählt, wie häufig ein Element in einer bestimmten Bakterienklasse gefunden werden kann; hierzu wurde die Einteilung der Taxonomy-Datenbank von NCBI verwendet, das Ergebnis ist in Tabelle 14 dargestellt.

Tabelle 14: Phylogenetische Verteilung der DGRs

Taxonomy-Klassifizierung	*Sequenzierte Genome auf NCBI [%]*	*Hits in unserem Ergebnis-Set [%]*
Actinobacteria	7,8	5,2
Bacteroidetes/Chlorobi group	3,9	27,7
Cyanobacteria	2,3	5,8
Deinococcus-Thermus	0,6	0,6
Firmicutes	23,6	31,0
Nitrospirae	0,1	0,6
Alphaproteobacteria	9,0	2,6
Betaproteobacteria	6,0	7,7
Gammaproteobacteria	21,4	12,3
Delta/Epsilonproteobacteria	3,1	2,6
Spirochaetes	5,6	0,6
unclassified Bacteria	0,2	0,6
Chlamydiae/Verrumicrobia group	1,0	0,6
Phages	11,7	1,9
Other Bacteria	3,8	

Man erkennt drei prokaryotische Klassen, die in unserem Set überrepräsentiert sind: Bacteroidetes, Firmicutes sowie Gammaproteobakterien. Zusammengenommen beinhalten sie 71 % aller DGRs, die in dieser Arbeit identifiziert werden konnten. Die restlichen Klassen sind mit zumeist < 10 % deutlich unterrepräsentiert, häufig konnte nur ein einziges Element (0,6 %) in diesen Klassen ermittelt werden. Diese Verteilung muss jedoch nicht notwendigerweise die natürliche Verteilung dieser Elemente wiederspiegeln; dies wäre nur dann der Fall, wenn alle Klassen in gleicher Zahl in der Datenbank repräsentiert wären. Um mögliche Ungleichverteilungen in den Datenbanken zu kompensieren, wurde Taxonomy daher zudem die Anzahl aller komplett sequenzierter Genome der jeweiligen Klassen entnommen und mit denen der DGRs verglichen. Während für Firmicutes und Gammaproteobakterien sich auch in der NCBI-Datenbank generell eine Überrepräsentation findet, und wir in unserem Ergebnis-Set somit lediglich eine Korrelation mit diesem Sachverhalt beobachten, scheinen Bacteroidetes-Elemente tatsächlich vermehrt in unseren Ergebnissen enthalten zu sein. Da Bacteroidetes-Vertreter in der Humanmedizin eine besondere Bedeutung als Überträger von Resistenzen und als Nosokomialkeime besitzen, besteht ein hohes Interesse an der Sequenzierung vieler klinischer Isolate. Somit stammen zahlreiche Bacteroidetes-DGRs entweder aus unterschiedlichen klinischen Isolaten derselben Bacteroides-Spezies, oder aus nicht näher klassifizierten Vertretern, die somit ebenfalls nur unterschiedliche Isolate darstellen könnten; folgerichtig weisen sie eine enorm hohe Sequenzidentität zueinander auf, wie man auch dem zugehörigen phylogenetischen Baum (s. u.) entnehmen kann. Somit erklärt sich, warum bei relativ geringem Anteil der Bacteroidetes-Gruppe an den sequenzierten Prokaryotengenomen eine

scheinbare Überrepräsentation in unseren Ergebnissen zu finden ist: es handelt sich hierbei um je ein einzelnes Element, das – nur geringfügig verändert – natürlich auch in anderen Isolaten derselben Spezies auftritt. Man kann hieran erkennen, dass eine quantitative Interpretation der Präsenz von DGRs in bestimmten prokaryotischen Klassen schwierig ist, da der Datenbankbestand keineswegs einen Blick auf die natürliche Verteilung in der Natur erlaubt.

Ein wichtiges Ergebnis jedoch ist die qualitative Beschreibung der Verteilung von DGRs: in nahezu jeder prokaryotischen Klasse ließen sich Elemente finden, und es ist wahrscheinlich, dass mit fortlaufender Entwicklung der Genomdatenbanken auch in den bisher fehlenden Klassen DGRs identifiziert werden können. Die Verbreitung von DGRs beschränkt sich somit keineswegs auf Mitglieder ökologischer Habitate oder auf bestimmte, durch phänotypische Merkmale wie aerobe oder anaerobe Lebensweise, Trophie oder Extremophilie definierte Gruppen von Prokaryoten, sondern umfasst nahezu alle Klassen und eine Vielzahl von Habitaten, was ihren universellen Nutzen demonstriert. Interessanterweise ist ihre Verbreitung innerhalb einzelner Klassen hingegen sporadisch: sie treten nur bei einzelnen Mitgliedern auf, und durchsetzen nie eine komplette Klasse. Ebenso ist die absolute Zahl der Elemente, die identifiziert werden konnten, vergleichsweise gering: in über 6000 sequenzierten Genomen wurden lediglich 155 Elemente gefunden, was einer Inzidenz von < 3% entspricht. Es stellt sich somit die Frage, warum DGRs nicht weiter verbreitet sind, wenn sie ihren Wirten einen so großen Anpassungsvorteil bieten könnten. Überlegungen zu dieser Frage werden in einem eigenen Abschnitt in Kapitel 4 diskutiert.

Eine weitere Frage, die sich stellt, betrifft die Inzidenz der DGRs in Phagen. Über jedes zehnte sequenzierte Genom in der NCBI-Datenbank stammt aus einem Phagen, nicht zuletzt, weil seit jeher ein besonderes Interesse an Phagen und ihrer Anwendung in der Forschung bestand, und weil ihre in der Regel kleinen Genome relativ leicht zu sequenzieren sind. Dennoch ließen sich nur drei Phagen-DGRs in unseren Ergebnissen finden, was einer Inzidenz von 1,9% entspricht. Die Gründe hierfür sind unklar, und werden ebenfalls in Kapitel 4 noch einmal angesprochen.

Mithilfe der Proteinsequenzen der DGR-RTs, die den höchsten Konservierungsgrad aller DGR-Elemente aufweisen, wurde ein phylogenetischer Neighbour Joining-Baum, basierend auf einem ClustalW-Alignment, mit 1000 Bootstrapping-Replikaten erstellt (s. Abbildung 12). Als Wurzel des Baums wurde die reverse Transkriptase des Gruppe II-Introns aus *Bacillus halodurans* (GI 47076650) gewählt. Um algorithmenbasierte Fehlgruppierungen besser zu erkennen, wurde außerdem ein ebensolcher Baum unter Verwendung des Maximum Likelihood-Verfahrens erstellt und die

Ergebnisse miteinander verglichen. Es zeigten sich keine relevanten Unterschiede zwischen beiden Bäumen, und im Folgenden wurde mit dem NJ-basierten Baum weitergearbeitet.

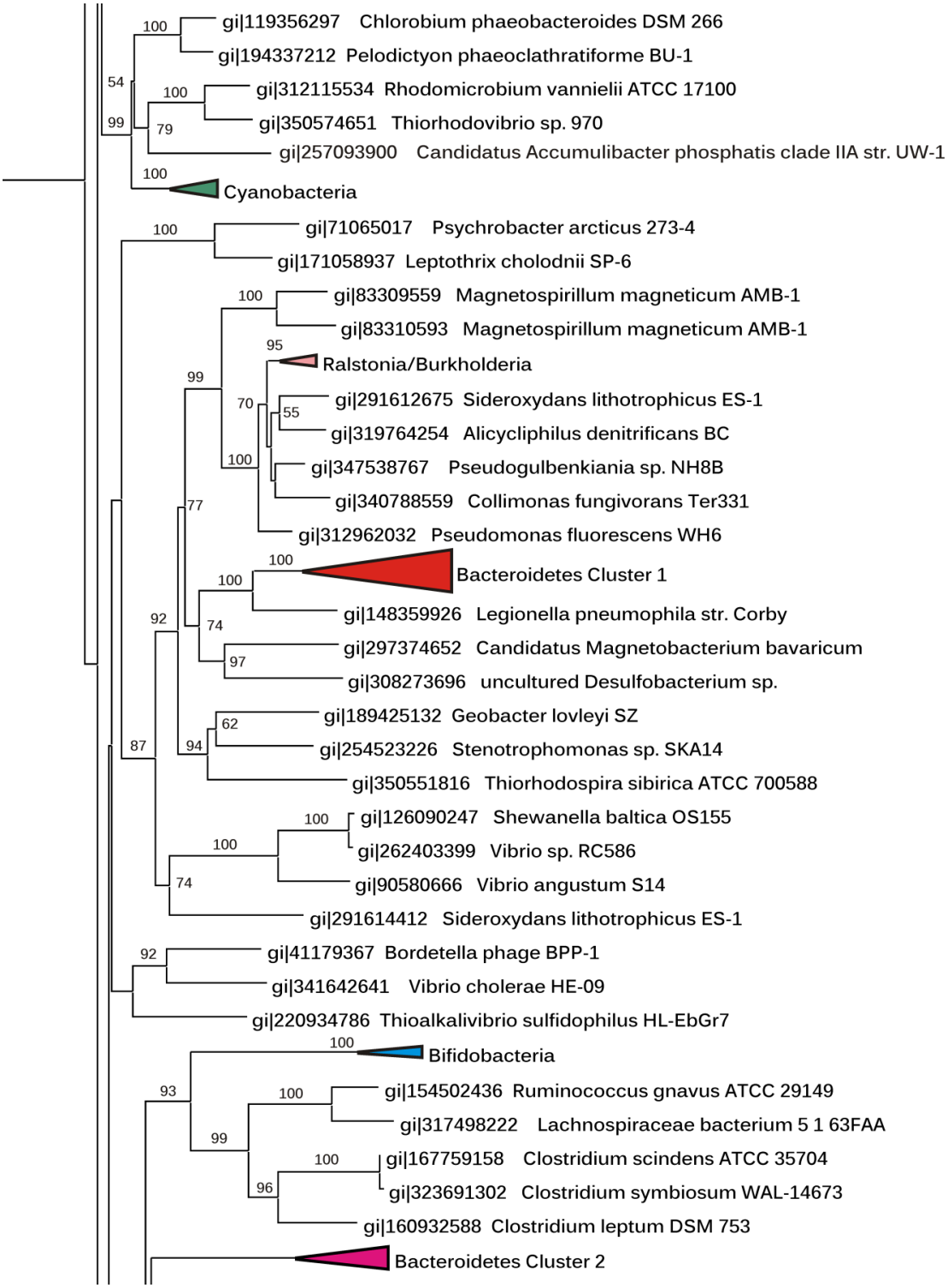


Abbildung 12: Phylogenetischer Baum der DGR-RTs. Ein phylogenetischer Baum, basierend auf einem ClustalW-Alignment der Aminosäuresequenzen von DGR-RTs, wurde mit dem Neighbour-Joining-Algorithmus und 1000 Bootstrapping-Replikaten erstellt. Dargestellt ist ein repräsentativer Ausschnitt. Große phylogenetische Gruppen mit Bootstrappingwerten von 100 sind kollabiert dargestellt (farbige Dreiecke).

Zusätzlich wurde ein NJ-Baum erstellt, der auf 16S RNA-Sequenzen der Organismen aus unserem Set beruht, und somit die Verwandtschaftsverhältnisse dieser Organismen beschreibt. Beide Bäume wurden unter der Annahme miteinander verglichen, dass diversitätsgenerierende Retroelemente ebenso wie das jeweilige Wirtsgenom der natürlichen Mutationsrate unterworfen sind und die Phylogenie somit hinsichtlich beider Merkmale ähnlich sein sollte. Bei der Betrachtung wurden nur Gruppierungen berücksichtigt, die jeweils über hohe Bootstrapping-Werte (> 70) verfügen, und somit eine hohe Robustheit der Gruppierung anzeigen. In der Mehrheit der Fälle wurde die obige Annahme bestätigt; auf RT-Ebene lassen sich Cluster aus Proteinen mit hoher Homologie zueinander erkennen, die sich ebenfalls auf Ebene der korrespondierenden 16S RNA-Sequenzen finden lassen. Es kann somit also davon ausgegangen werden, dass DGRs in diesen Fällen von einem gemeinsamen Vorläuferorganismus über vertikalen Gentransfer an evolutionär jüngere Spezies weitergegeben wurden. Die hier definierten Cluster können häufig bestimmten prokaryotischen Klassen zugeordnet werden, wie Cyanobakterien, Clostridiales oder Bacteroidetes. Interessanterweise lassen sich die meisten Elemente aus der Gruppe der Bacteroidetes in zwei unterschiedlichen Clustern finden, so dass hier möglicherweise von zwei verschiedenen Vorläuferelementen ausgegangen werden kann. In einigen Fällen wurden jedoch RT-Proteine miteinander gruppiert, deren korrespondierende 16S RNA-Sequenzen deutlich weniger Homologie zueinander aufwiesen, und zwischen denen somit eine größere Distanz im 16S RNA-Baum besteht. Anders ausgedrückt bedeutet dies, dass zwei Organismen über DGR-Elemente verfügen können, die zueinander einen höheren Verwandtschaftsgrad aufweisen als die Organismen selbst. Eine wahrscheinliche Erklärung für diesen Befund könnte ein horizontaler Gentransfer zwischen zwei relativ geringfügig verwandten Organismen gewesen sein, bei dem das DGR-Element ebenfalls weitergegeben wurde. Die betreffenden Elemente und ihre genomische Lokalisation wurden daraufhin näher untersucht. Es stellte sich heraus, dass DGRs offensichtlich, in Ermangelung eines eigenen Systems zur Mobilisierung, eine Vielzahl von anderen mobilen genetischen Elementen benutzen, um sich horizontal, also von einer Art zu einer anderen, zu verbreiten.

a) Phagen/Prophagen: Drei DGRs (GIs 27311204, 291621968 und 41179367) aus unserem Set wurden in Phagen gefunden, hiervon ist eines das prototypische *Bordetella* Bakteriophagen-DGR. Um weitere potentiell prophagenassoziierte DGRs zu ermitteln, wurde die ACLAME-Datenbank benutzt, in der sequenzierte Genome bereits auf Prophagenregionen hin untersucht wurden. In neun Fällen (GIs 150390313, 134299090, 171058937, 83309559, 83310593, 15000754, 71065017, 146277368, 120599269) konnten DGRs aus unserem Set solchen Regionen zugeordnet werden, alle Fälle weisen im RT-Baum eine vom 16S RNA-Baum abweichende Eingruppierung auf. Als nächstes wurde versucht, mit dem Prophagefinder-Tool die Genomsequenzen der Organismen aus unseren Ergebnissen auf

Prophagenregionen zu untersuchen. Dieses Programm übersetzt mittels BLASTx eine Nucleinsäuresequenz von 5 Kilobasenpaaren bis 10 Megabasenpaaren (was in der Größenordnung der meisten bakteriellen Genome liegt) in allen sechs Leserahmen und vergleicht die erhaltenen Proteine mit einer Phagenproteindatenbank. In einem weiteren Schritt werden die Treffer nach Sequenzidentität und topologischen Aspekten, also Ballungen von Treffern in einem bestimmten Bereich der Inputsequenz, analysiert, und potentielle Prophagenregionen ausgegeben. Hierüber kann dann manuell ein Vergleich mit den genomischen Koordinaten der DGRs erfolgen. Die Verlässlichkeit dieses Programms kann jedoch in Frage gestellt werden; so werden zwar Prophagenregionen gefunden, die teilweise mit DGR-Regionen überlappen, allerdings ist es fraglich, ob es sich hierbei um „echte“ Prophagen handelt, oder nur Ballungen von offenen Leserahmen mit hoher Homologie zu Phagenproteinen. Häufig fehlen essentielle Proteine wie Integrasen, Capsidproteine oder Polymerasen in den ermittelten Regionen. Dies kann natürlich einerseits auf fehlende Referenzproteine in der Datenbank zurückzuführen sein, andererseits auf einen „defekten/gestrandeten“ Phagen im Genom, dessen Gene durch die natürliche Mutationsrate und die Tendenz, genetischen Ballast zu elimineren, mit der Zeit verloren gehen. Es ist somit nie eindeutig entscheidbar, ob ein DGR sich in einer echten, mobilisierbaren Prophagenregion aufhält, oder einmal tatsächlich Teil eines Phagen war und nun ebenso gestrandet ist bzw. vom Wirt assimiliert wurde, oder die genetische Nachbarschaft lediglich aus Leserahmen gebildet wird, die eine gewisse Homologie zu Phagenproteinen besitzen. Mit Prophinder wurde ein weiteres Tool verwendet, welches ein ähnliches Prinzip wie Prophagefinder implementiert: hier wird eine BLASTp-Suche gegen eine Phagenproteindatenbank durchgeführt, wobei als Query-Sequenzen die Zielgene der DGRs aus unserem Set gewählt wurden. Mit der strengsten Sucheinstellung, die einen E-Wert von 0,0001 als Cut-Off verwendet, konnten ca. 43,3 % aller Zielgene aus unserem Ergebnis-Set zwischen 1 und 20 Treffern in der ACLAME-Datenbank zugeordnet werden. Eine Vergleichsabfrage mit demselben Cut-Off, die eine Queryliste aus zufällig ausgewählten Proteinen aus denselben Organismen benutzte, ergab allerdings ebenso in 21,7 % aller Queryproteine eine gute Übereinstimmung in ACLAME. Dies ist für eine Negativkontrolle ein relativ hoher Wert, so dass auch hier keine verlässliche Aussage über die Zielproteine und eine mögliche Assoziation mit Phagen möglich ist. Die genaue Zahl von DGRs, die über Phagen verbreitet werden, kann somit aufgrund fehlender geeigneter Analysetools noch nicht gegeben werden. Mit den neun ermittelten Fällen und den drei Phagen-DGRs wurden allerdings bereits 7,7 % der DGRs aus unserem Set als phagen- oder prophagenassoziiert erkannt.

b) Transposons/Konjugative Transposons: Das Element aus *Vibrio* sp. RC586 (GI 262403399) weist in seiner unmittelbaren genomischen Nachbarschaft mehrere Leserahmen auf, die mit dem Tn7-Transposon assoziiert werden; da diese jedoch alle auf einem relativ kleinen Contig von etwa 14,2 kbp liegen, ist eine genauere Analyse dieses Sequenzfragments nicht möglich. Ein weiteres Beispiel fand sich innerhalb des Bacteroides-Clusters 1 (s. Abbildung 12), und somit nicht direkt erkennbar bei einem Vergleich der phylogenetischen Bäume. Im Falle von *Bacteroides* sp. 1_1_14 (GI 298387225) und *Bacteroides ovatus* 3_8_47FAA (GI 336413338) konnte durch einen MEGABLAST-Vergleich der Contigs ein Fragment von ca. 42,6 kbp ermittelt werden, auf dem die jeweiligen DGR-Elemente liegen (s. Abbildung 13).

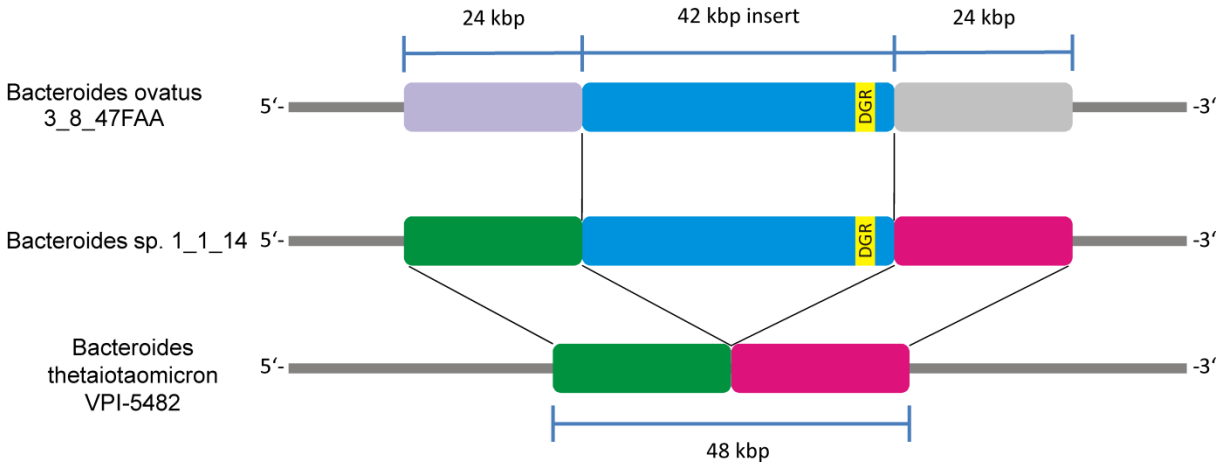


Abbildung 13: Nutzung eines 42 kbp-Elements als DGR-Shuttle. Durch Vergleich der Contigs von *Bacteroides* sp. 1_1_14 und *Bacteroides ovatus* 3_8_47FAA, die DGR-Elemente mit hoher Homologie zueinander beinhalten, konnte ein 42 kbp großes Element (blau) identifiziert werden, an dessen 3'-Ende das DGR (gelb) liegt. Die flankierenden Sequenzen (grün und rot) in *Bacteroides* sp. 1_1_14 unterscheiden sich von denen in *Bacteroides ovatus* 3_8_47FAA (violett und grau), und können ohne das Insert in einer dritten Spezies, *Bacteroides thetaiotaomicron* VPI-5482, gefunden werden.

Die das 42 kbp-Insert jeweils flankierenden Sequenzen (in der Abbildung violett/grau bzw. rot/grün) weisen eine deutlich geringere Sequenzidentität untereinander auf, sodass davon ausgegangen werden kann, dass es sich um ein austauschbares und entweder mobiles oder zumindest mobilisierbares Element handelt, das vom DGR-Element zur Verbreitung genutzt wird. Gestützt wird diese Annahme durch einen Vergleich mit *Bacteroides thetaiotaomicron* VPI-5482, in dem sich die upstream und downstream flankierenden Sequenzen aus *Bacteroides* sp. 1_1_14 finden lassen, jedoch ohne 42 kbp –Element. Es handelt sich somit wahrscheinlich um die Präinsertionsstelle des Elements. Worum es sich bei diesem Element handelt, ist unklar. Eine BLASTp-Suche der größten enthaltenen offenen Leserahmen (mind.

900 bp) identifiziert einige Phagenproteine, allerdings auch ein DNA-Transpositionsprotein, eine Clp-Protease sowie ein Radical SAM-Domänenprotein. Es könnte sich somit um ein komplexeres, mobilisierbares DNA-Element handeln, das Module von Phagen und Transposons besitzt.

c) Plasmide: Das DGR-Element von *Shewanella baltica* OS155 (GI 126090247) ist auf einem Plasmid lokalisiert, und somit das erste Element, das auf dieser Klasse von mobilen Nucleinsäuren gefunden wurde.

3.1.2.6 Zielgene

Um mehr über die biologische Bedeutung von DGRs herauszufinden, ist es notwendig, die Proteine genauer zu beschreiben, die von ihnen hypermutiert werden. Hierzu wurden die Zielproteine der DGRs über die Zuordnung der VR-Regionen zu offenen Leserahmen in Artemis bestimmt, extrahiert, und ihre Datenbankannotationen in einem ersten Schritt ermittelt (s. Tabelle 15).

Tabelle 15: Datenbankannotationen der Zielproteine

Annotation	*Anzahl*	*Anteil [%]*
Hypothetical/predicted protein, Unknown function	82	50,0
FGE Enzym/NACHT/NTPase-Domäne	45	27,4
DUF1566	7	4,3
Major tropism determinant	5	3,0
Concanavalin A-Typ Lektin/Glucanase Superfamilie	5	3,0
DUF3988	2	1,2
Andere	18	11,0

Die meisten Proteine werden mit „hypothetical protein“, „unkown protein“ oder ähnlichem annotiert (55,5%), was an dieser Stelle keine weiteren Rückschlüsse auf ihre Funktion zulässt. Eine weitere große Gruppe wird – unter anderem – mit „Formylglycine generating sulfatase enzyme“ beschrieben (ca. 27 %); auf diese unter DGR-Zielproteinen recht häufig anzutreffende Annotation wurde bereits von Le Coq & Ghosh hingewiesen. Der Name ist etwas irreführend, tatsächlich handelt es sich hier um eine Gruppe von Enzymen, die die Aktivierung von Sulfatasen katalysiert, und nicht um Sulfatasen selbst; dies geschieht durch die Methylierung eines Cysteinrests im aktiven Zentrum

von Sulfatasen, wodurch Formylglycin gebildet wird. Eine weitere häufiger auftretende Annotation von Zielproteinen aus unseren Ergebnissen ist „major tropism determinant“ (3 %), also Proteine mit hoher Homologie zum Zielprotein des *Bordetella* Bakteriophagen-DGRs.. Interessanterweise ist eben dieses Zielprotein das einzige mit dieser Annotation, das nach bisherigem Kenntnisstand (vgl. Abschnitt 3.1.2.6) eindeutig mit einem Phagen assoziiert werden kann. Daneben gibt es einige konkrete Annotationen der Zielproteine, wie beispielsweise „Serralysin“, „6-Phosphofructokinase“, „NADH-dependent Glyceratdehydrogenase“ oder „DNA Topoisomerase IV“. Diese wurden manuell mit genuinen Vertretern dieser Enzyme verglichen, und können als Fehlannotationen vernachlässigt werden.

Die subzelluläre Lokalisation (subcellular localization, SCL) könnte ebenfalls Aufschluss darüber geben, in welche Prozesse Zielproteine involviert sind; so erscheint es am sinnvollsten, Proteine zu hypermutieren, die einem sich verändernden Umfeld präsentiert werden, und somit eher membranständig und mit der hypermutierten Region extrazellulär orientiert erwartet werden könnten. Automatische Annotationsvorgänge schließen teilweise SCL-Vorhersagen mit ein, allerdings scheint die Verlässlichkeit solcher Routinen nicht besonders hoch zu sein; so wird laut NCBI *Protein*-Eintrag beispielsweise für das Zielprotein aus *Clostridium bolteae* ATCC BAA-613 (GI 160938972) eine mitochondriale Assoziation vorhergesagt. Aus diesem Grunde wurden die DGR-Zielproteine aus unserem Set einer separaten SCL-Analyse unterzogen. Hierzu wurden die Programme PSortb v3.0.2 und in Ergänzung hierzu SVMTM verwendet: während ersteres eine detailliertere Vorhersage einer cytosolischen, extrazellulären oder membranassoziierten Lokalisation eines Proteins erlaubt und eine Kombination aus einer BLASTp-Suche, Support Vector-Machines, Transmembranhelicesvorhersage sowie Motiv- und Lokalisationssignalanalyse implementiert, überprüft SVMTM die Primärstruktur lediglich auf mögliche Transmembranbereiche. Eine Kurzübersicht der Ergebnisse kann Tabelle 16 entnommen werden, detailliertere Angaben finden sich in der Anhangtabelle A3.

Tabelle 16: SCL-Vorhersage von DGR-Zielproteinen

	Zahl	*Anteil [%]*
Gram-positiv		
Zytosol	7	13,2
Membran (intrazellulär)	3	5,7
Extrazellulär	5	9,4

Unbekannt	38	71,7
Transmembranregion	3	5,7
Gram-negativ		
Zytosol	17	15,7
Membran (intrazellulär)	1	0,9
Periplasma	1	0,9
Membran (extrazellulär)	7	6,5
Extrazellulär	13	12,0
Unbekannt	69	63,9
Transmembranregion	13	12,0

Zuletzt wurde ein ClustalW-Alignment der extrahierten Zielproteine durchgeführt, und das Ergebnis in einem phylogenetischen Baum dargestellt (s. Abbildung 14).

Abbildung 14: Phylogenetischer Baum der DGR-Zielproteine. Die Aminosäuresequenzen von DGR-Zielproteinen wurden mit ClustalW alignt und ein phylogenetischer Baum erstellt. Grob lassen sich DGR-Zielproteine in drei Hauptäste einteilen, von denen einer sämtliche Proteine, die über eine FGE-Enzym-Annotation verfügen, beinhaltet, sowie sämtliche Proteine mit Major Tropism Determinant-Annotation.

Man erkennt drei Hauptäste; während zwei der Äste ausschließlich aus Proteinen mit unbekannter Funktion bestehen, finden sich sämtliche FGE-Enzyme sowie „major tropism determinants“ und eine dritte Gruppe mit der Annotation „Concanavalin A-Typ Lektin“ im dritten Ast. Somit scheint zwischen diesen Proteinen ein gewisses Verwandtschaftsverhältnis zu bestehen, was die FGE-Enzyme noch deutlicher in die Nähe der durch den *Bordetella* Bakteriophagen bereits gut beschriebenen Major Tropism-Determinanten rückt.

Um einen besseren Eindruck über das absolute Verwandtschaftsverhältnis der Proteine zu erhalten, wurden innerhalb eines Astes sowie zwischen Vertretern unterschiedlicher Äste E-Werte über BLASTp ermittelt. In Einklang mit früheren Beobachtungen besitzen die meisten Proteine – selbst innerhalb eines Astes – keine besonders große Sequenzidentität zueinander, wie man an E-Werten von oft 0,01 und schlechter erkennt. Dies bedeutet entweder, dass DGRs recht unterschiedliche Zielproteine akquiriert haben, oder umgekehrt, dass sich aus wenigen Ahnenproteinen, die mit einem DGR verbunden waren, eine Vielfalt von Proteinen entwickelt hat; einen genaueren

Rückschluss auf das relative Alter dieser Elemente lassen die derzeitigen Ergebnisse nicht zu, allerdings kann in Anbetracht dieser Ergebnisse angenommen werden, dass DGRs keine jungen Elemente sind.

3.1.2.7 Akzessorische Proteine

In einigen bisher beschriebenen DGRs wurden akzessorische Proteine identifiziert, die eine unbekannte, jedoch essentielle Rolle im Mechanismus der Hypermutation des Zielgens besetzen. Eine präzisere Definition dieser Proteingruppe ist bislang nicht erfolgt; so gibt es beispielsweise keineHinweise auf konservierte Motive, mit denen eine systematische Suche nach Homologen möglich wäre. Daher wurden die von DiGReF extrahierten genomischen Sequenzen mit dem EMBOSS-Programm *getorf* auf offene Leserahmen untersucht, und die ermittelten Gene in die jeweiligen Aminosäuresequenzen übersetzt. Um die Mächtigkeit der Ergebnisse und somit die erforderliche Rechenleistung der nachfolgenden Schritte zu reduzieren, wurde ein Cut-Off von minimal 270 Nucleotiden/90 Aminosäuren gesetzt, da die bisher publizierten akzessorischen Proteine ähnlich kurz sind (~ 100 Aminosäuren); durchsucht wurden jeweils der Strang, auf dem der RT-ORF liegt sowie der komplementäre Strang. Es wurden 5821 Leserahmen extrahiert. Diese wurden mit ClustalW alignt und auf auffällige Motive hin untersucht. Neben den DGR-RTs mit ihrem charakteristischen SQ-Motiv in Domäne 4 und dem YxDD-Motiv in Domäne 5 fand sich eine Gruppe von 82 Proteinen, die mit dem bAvd-Protein aus dem *Bordetella*-Phagen alignte und ein Consensus-Motiv in der Nähe des C-Terminus aufwies. Diese Gruppe wurde noch einmal extrahiert und alignt, und ein Sequenzlogo des Motivs erstellt (s. Abbildung 15)

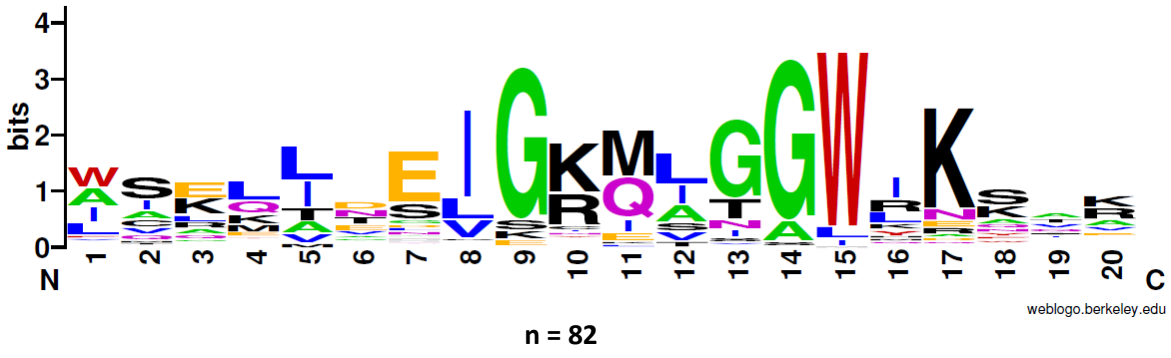


Abbildung 15: Sequenzlogo putativer akzessorischer Proteine. Das Consensus-Motiv von 82 extrahierten Proteinen wurde auf relative Aminosäurehäufigkeiten pro Position untersucht und ein Sequenzlogo erstellt. Die Positionen korrespondieren mit den Aminosäureresten 102 bis 121 des akzessorischen Proteins bAvd aus dem *Bordetella* Bakteriophagen. Man erkennt eine hohe Konservierung aliphatischer Aminosäuren an Position 8, basischer Aminosäuren an den Positionen 10 und 17 sowie der Aminosäure Tryptophan an Position 15.

Aus dem Sequenzlogo lässt sich das Consensus-Motiv

[L/I/V] – G – [K/R] – x – x – G – G – W – x – K

ableiten. In 92% der Proteine fanden sich Aminosäuren mit aliphatischen Seitenketten (Leucin, Isoleucin und Valin) an Position 8 des Sequenzmotivs in Abbildung 15, während 84% ein Glycin an Position 9 aufweisen. Die nächste Position wird in 78% der Sequenzen mit einer basischen Aminosäure (Lysin oder Arginin) besetzt. An Position 12 lassen sich in 81% der Sequenzen wiederum aliphatische Aminosäuren identifizieren, und in ca. 15% Serin oder Threonin. Position 13 ist weniger stark konserviert und wird in 60% der Sequenzen mit Glycin besetzt, in knapp einem Fünftel (19,5%) mit Serin oder Threonin. Es folgen wiederum zwei stark konservierte Positionen: Glycin ist in ca. 88% der Fälle an Position 14 vertreten, und Tryptophan zu 89%. Weiterhin auffällig ist, dass das Motiv in ebenfalls 89% aller Sequenzen von mindestens einer oder bis zu drei basischen Aminosäuren an den Positionen 16 bis 18 gefolgt wird.

Als nächstes wurde überprüft, an welcher Position innerhalb der DGR-Kassette diese Proteine zu finden sind. Bisher beschriebene DGRs konnten *bAvd*-artige Leserahmen stets unmittelbar vor dem Leserahmen der reversen Transkriptase lokalisieren. Durch die hier angewandte Suchstrategie sollte es möglich sein, auch Leserahmen zu finden, die sich auf abweichenden Positionen, oder auf dem Gegenstrang befinden könnten. Es zeigte sich, dass auch hier die Leserahmen stets unmittelbar upstream der RT lokalisiert sind. Dieses Ergebnis könnte darauf schließen lassen, dass akzessorisches Protein und reverse Transkriptase ein gemeinsames RNA-Transkript teilen, möglicherweise, weil beide Proteine als Komplex vorliegen müssen.

3.2 Aufreinigung und Funktionsanalyse der DGR-RT Alr3497 aus *Nostoc* sp. PCC 7120

Für die adeninspezifische Hypermutation, die ein einzigartiges Merkmal von DGRs darstellt, ist nach heutigem Kenntnisstand die elementeigene reverse Transkriptase wahrscheinlich von zentraler Bedeutung. Um diese Enzymklasse mechanistisch und strukturell eingehender zu charakterisieren, ist eine erfolgreiche rekombinante Synthese eines oder mehrerer Vertreter dieser Enzyme eine notwendige Voraussetzung. Allerdings stellt die rekombinante Synthese von reversen Transkriptasen noch immer eine Herausforderung dar; bis heute konnten lediglich drei Kristallstrukturen dieser Enzymklasse erhalten und gelöst werden (s. Abschnitt 1.1.2). Im Folgenden werden die Schritte

beschrieben, die in dieser Arbeit unternommen wurden, um eine lösliche und funktionsfähige reverse Transkriptase eines DGR-Elements zu erzeugen und zu charakterisieren.

3.2.1 Klonierungen

In dieser Arbeit wurden mehrere Konstrukte erzeugt, deren Details Tabelle 16 entnommen werden können.

Tabelle 17: Konstrukte aus DGR RT-Expressionsstudien

Konstrukt	*Donororganismus der RT*	*Vektor*	*Tag*
pTS1	*Treponema denticola*	pCR4	-
pTS2	*Bacteroides thetaiotaomicron*	pCR4	-
pTS3	*Treponema denticola*	pTWIN1	*Mxe* GyrA Intein/CBD, C-terminal
pTS4	*Bacteroides thetaiotaomicron*	pTWIN1	*Mxe* GyrA Intein/CBD, C-terminal
pTS5	*Treponema denticola*	pGEX-6P-1	GST-Tag, N-terminal
pTS6	*Bacteroides thetaiotaomicron*	pGEX-6P-1	GST-Tag, N-terminal
pTS8	*Bacteroides thetaiotaomicron*	pET15b	6xHis, N-terminal
pTS9	*Treponema denticola*	pET21a	6xHis, C-terminal
pTS10	*Bacteroides thetaiotaomicron*	pET21a	6xHis, C-terminal
pTS11	*Treponema denticola*	pTWIN1	*Ssp* DnaB Intein/CBD, N-terminal
pTS12	*Bacteroides thetaiotaomicron*	pTWIN1	*Ssp* DnaB Intein/CBD, N-terminal
pTS13	*Treponema denticola*	pETGEX-CT[§]	GST-Tag, C-terminal
pTS14	*Bacteroides thetaiotaomicron*	pETGEX-CT[§]	GST-Tag, C-terminal
pTS15	*Thermus aquaticus*	pCR2.1	-
pTS16	*Vibrio* Phage VHML	pCR2.1	-

pTS17	*Nostoc* sp. PCC 7120	pUC57	-
pTS18	*Thermus aquaticus*	pET15b	6xHis, N-terminal
pTS19	*Vibrio* Phage VHML	pET15b	6xHis, N-terminal
pTS20	*Nostoc* sp. PCC 7120	pET15b	6xHis, N-terminal
pTS21	*Thermus aquaticus*	pET21a	6xHis, C-terminal
pTS22	*Vibrio* Phage VHML	pET21a	6xHis, C-terminal
pTS23	*Nostoc* sp. PCC 7120	pET21a	6xHis, C-terminal
pTS24	*Thermus aquaticus*	pET11a	-
pTS25	*Vibrio* Phage VHML	pET11a	-
pTS26	*Nostoc* sp. PCC 7120	pET11a	-

[$] (siehe Sharrocks, 1994)

Es wurden die reversen Transkriptasen aus *Treponema denticola* ATCC 35405 (TDE2266, GI 42527768), *Bacteroides thetaiotaomicron* VPI-5482 (BT2313, GI 29347723), *Thermus aquaticus* Y51MC23 (TaqDRAFT_5432, GI 218295563), *Vibrio* Phage VHML (ORF37, GI 27311204) sowie *Nostoc* sp. PCC 7120 (Alr3497, GI 17230989) ausgewählt, um ein breites Spektrum von Organismen abzudecken. Für Konstrukte mit den RTs aus *Treponema denticola* ATCC 35405 und *Bacteroides thetaiotaomicron* VPI-5482 wurde als Ausgangsmaterial genomische DNA verwendet, welche von der Deutschen Sammlung für Mikroorganismen und Zellkulturen (DSMZ) bezogen wurde. Die Amplifikation der codierenden Sequenzen erfolgte mittels PCR. Die Leserahmen der RTs aus *Thermus aquaticus* Y51MC23, *Vibrio* Phage VHML und *Nostoc* sp. PCC 7120 wurden hingegen per Gensynthese assembliert; zusätzlich wurden die Sequenzen an die Codon Usage in *E.coli* angepasst, um die heterologe Expression in diesem System zu optimieren. Darüber hinaus wurden mehrere Tags an die reversen Transkriptasen fusioniert; dies ermöglichte einerseits eine Aufreinigung über Affinitätschromatographie, andererseits ist bekannt, dass größere Tags wie Glutathion-S-Transferase die Löslichkeit des Fusionsproteins erhöhen können. Um mögliche Interferenzen der Tags mit der nativen Struktur der Proteine zu umgehen, wurde, soweit möglich, jeder Tag jeweils C- und N-terminal fusioniert. Durch analytische PCR, Restriktionsfragmentlängenvergleiche und Sequenzierungen wurde jeweils die Integrität der Konstrukte verifiziert.

3.2.2 Rekombinante Expression in *E.coli* und Aufreinigung

Mehrere Stämme von *Escherichia coli* wurden mit den oben genannten Konstrukten transformiert und Überexpressionen wie in Kapitel 2.2.3.3 beschrieben durchgeführt. Es wurden hierbei Stämme gewählt, die für rekombinante Proteinexpression optimiert wurden, und somit geringere Mengen endogener Proteasen aufweisen.

In initialen Versuchen wurde die Expression des Zielproteins bei 37 °C mit einer IPTG-Konzentration von 1 mM induziert und der Löslichkeitsstatus wie in Abschnitt 2.2.3.3 beschrieben bestimmt. Falls die RT in der unlöslichen Fraktion vorlag, wurden die Expressions- und Lysebedingungen variiert; getestet wurden Inkubationen bei Raumtemperatur und 16 °C, verminderte IPTG-Konzentrationen bis 0,1 mM sowie Lysepuffer mit variierenden pH-Werten, Salzkonzentrationen und Zusätzen wie reduzierenden oder löslichkeitsunterstützenden Agenzien und Detergenzien. Es zeigte sich, dass die RT-Proteine aus *Treponema denticola* und *Bacteroides thetaiotaomicron* zwar teilweise als Fusionsproteine in E.coli exprimiert wurden, jedoch stets als unlösliche Proteinaggregate (Inclusion Bodies). Da weder durch unterschiedliche Tags, noch durch Variation der Expressionsbedingungen die Löslichkeit dieser Proteine erhöht werden konnte, wurden andere Vertreter der DGR-RTs für weitere Expressionsstudien ausgewählt. Die Leserahmen dieser Proteine wurden zusätzlich an die Codon Usage in E.coli angepasst. Für einige dieser Proteine konnte tatsächlich eine Expression in löslicher Form beobachtet werden.

Die Ergebnisse der Expressions- und Löslichkeitstests dieser Fusionsproteine sind in den Abbildungen 16 bis 18 dargestellt, eine Zusammenfassung aller Ergebnisse der Expressionsstudien kann Tabelle 18 entnommen werden.

Tabelle 18: Expression und Aufreinigung von DGR-RTs

Quellorganismus der RT	*Tag*	*Erwartete Größe [kDa]*	*Expression*	*Löslichkeitsstatus*	*Aufreinigung*
Treponema denticola **ATCC 35405**	*Mxe* GyrA Intein/CBD, C-terminal	60,3	ja	unlöslich	nicht möglich
	GST, N-terminal	64,5	ja	unlöslich	nicht möglich
	6xHis,	39,7	nein	-	-

	C-terminal				
	Ssp DnaB Intein/CBD, N-terminal	69,9	nein	-	-
	GST, C-terminal	64,5	ja	nicht bestimmt	-
***Bacteroides thetatiotaomicron* VPI-5482**	*Mxe* GyrA Intein/CBD, C-terminal	61,6	ja	unlöslich	nicht möglich
	GST, N-terminal	65,8	ja	unlöslich	nicht möglich
	6xHis, N-terminal	41,0	ja	unlöslich	nicht durchge-führt
	6xHis, C-terminal	41,0	ja	unlöslich	nicht möglich
	Ssp DnaB Intein/CBD, N-terminal	71,2	ja	nicht bestimmt	-
	GST, C-terminal	65,8	ja	nicht bestimmt	-
***Thermus aquaticus* Y51MC23**	6xHis, N-terminal	36,9	ja	löslich	nicht möglich
	6xHis, C-terminal	36,9	nein	-	-
	-	36,1	ja	unlöslich	nicht möglich
***Vibrio* Phage VHML**	6xHis, N-terminal	36,1	ja	löslich	nicht möglich
	6xHis, C-terminal	36,1	nein	-	-
	-	35,3	nein	-	-

***Nostoc* sp. PCC 7120**	6xHis, N-terminal	42,6	ja	löslich	über Heparin, jedoch nicht reproduzier-bar
	6xHis, C-terminal	42,5	Ja	löslich	nicht möglich
	-	41,7	Ja	unlöslich	nicht möglich

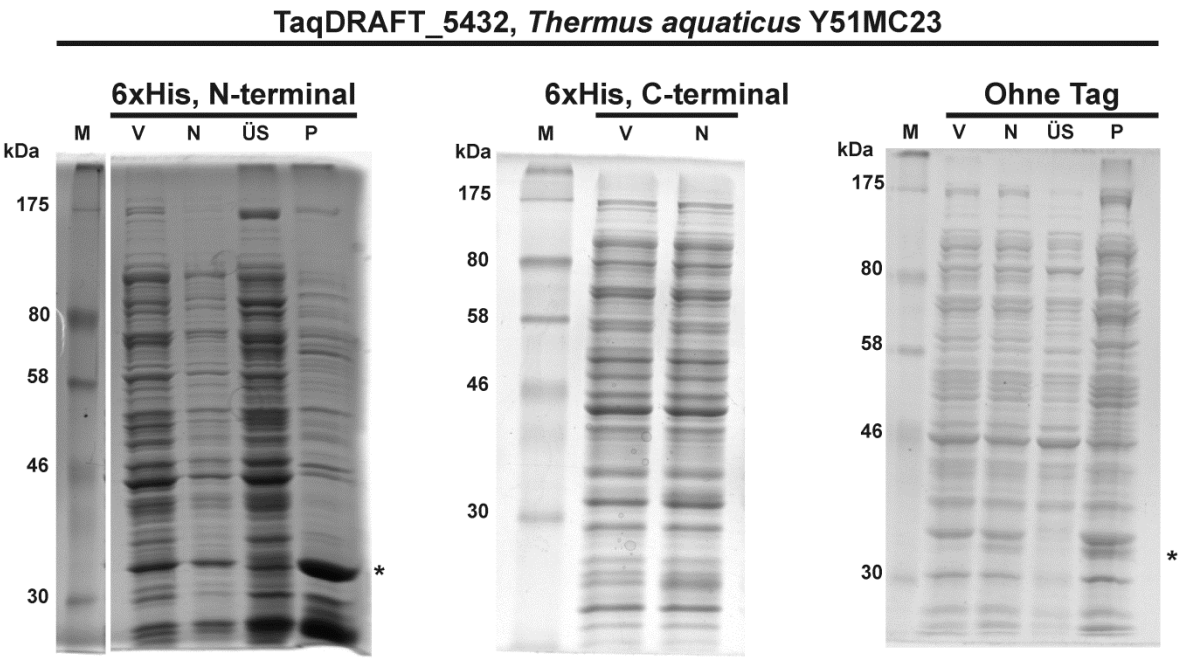


Abbildung 16: Expression und Löslichkeit der DGR-RT aus *Thermus aquaticus* Y51MC23. Die Expression der Proteine erfolgte in *Escherichia coli* T7 Express (New England Biolabs) bei 16 °C mit 0,5 mM IPTG, weitere Details können Abschnitt 2.2.3.3 entnommen werden. Aufgetragen sind Gesamtzellextrakte vor (V) und nach (N) Expression des Proteins, sowie Proben der löslichen Proteinfraktion im Überstand (ÜS) bzw. der unlöslichen Bestandteile im Pellet (P). Die Laufhöhe der RT-Proteine ist mit (*) gekennzeichnet. Expression des Fusionsproteins mit N-terminalem 6xHis-Tag ist nicht erkennbar, allerdings weist die Fraktion der löslichen Proteine eine Anreicherung eines Proteins der erwarteten Größe auf. Deutlichere Expression lag im Falle des Proteins ohne Tag vor, allerdings nur in unlöslicher Form. 12 % SDS-PAGE, Coomassie Brillant Blue R250. M = ColorPlus Prestained Protein Marker Broad Range, New England Biolabs

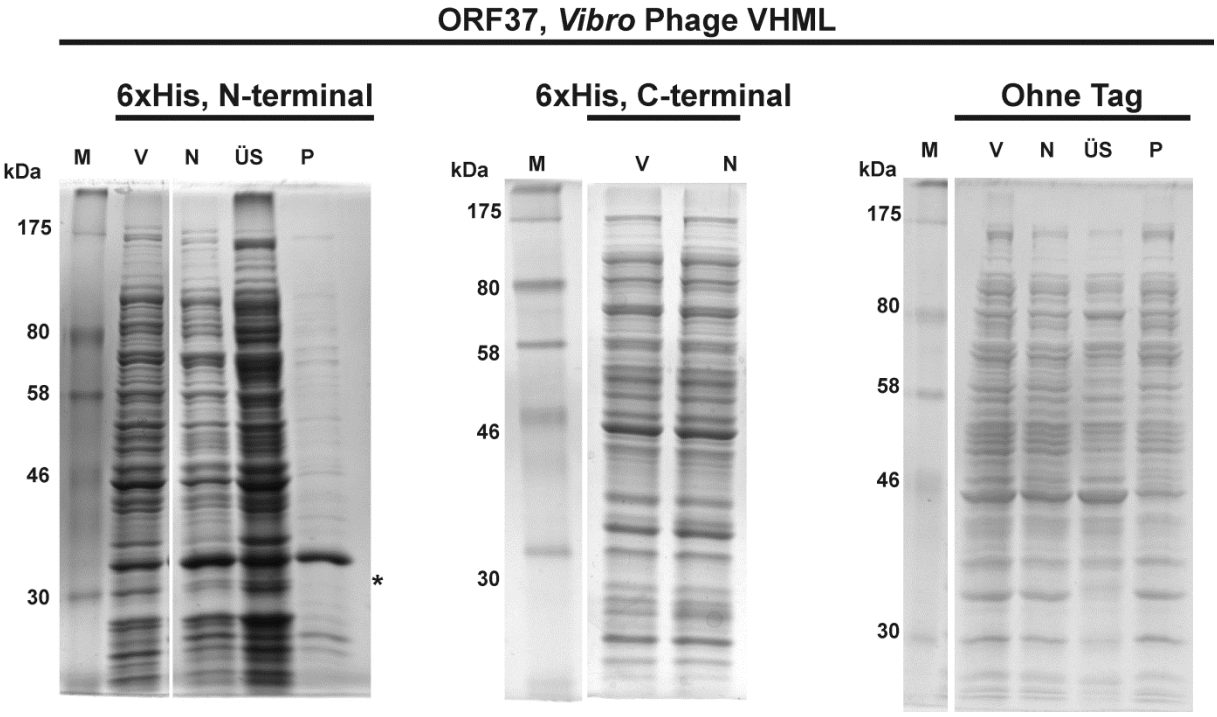


Abbildung 17: Expression und Löslichkeit der DGR-RT aus *Vibrio* Phage VHML. Details wie in Abbildung 16 und Abschnitt 2.2.3.3. Nur im Falle einer N-terminalen Fusion des 6xHis-Tags ist eine schwache Expression des Proteins in löslicher Form sichtbar. 12 % SDS-PAGE, Coomassie Brillant Blue R250. M = ColorPlus Prestained Protein Marker Broad Range, New England Biolabs

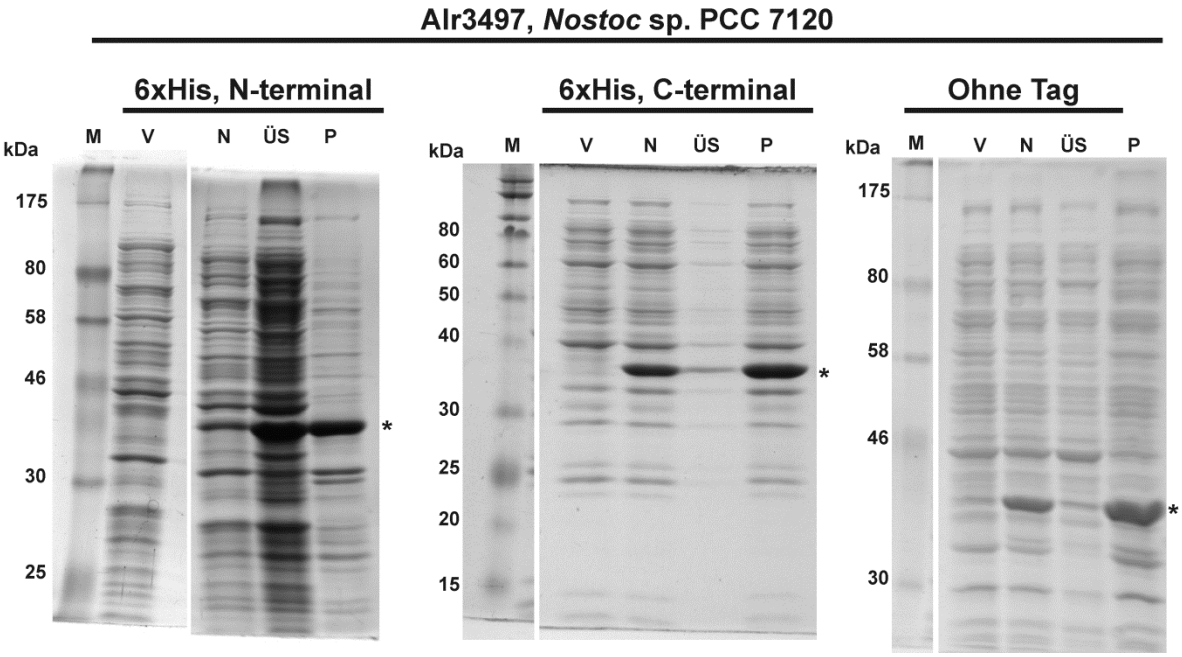


Abbildung 18: Expression und Löslichkeit der DGR-RT aus *Nostoc* sp. PCC 7120. Details wie in Abbildung 16 und Abschnitt 2.2.3.3. Expression der RT war mit und ohne 6xHis-Tag möglich, in löslicher Form jedoch nur als Fusionsprotein. 12 % SDS-PAGE, Coomassie Brillant Blue R250. M = ColorPlus Prestained Protein Marker Broad Range bzw. Prestained Protein Ladder, Broad Range , New England Biolabs

Lagen detektierbare Mengen löslichen Proteins vor, wurde versucht, diese mittels Affinitätschromatographie von endogenen *E.coli*-Proteinen zu isolieren. Eine Bindung der Proteine an Ni-NTA über das 6xHis-Tag war jedoch nur unter denaturierenden Bedingungen möglich (nicht gezeigt), so dass davon ausgegangen werden kann, dass der N-Terminus des Fusionsproteins nach Expression nicht zugänglich ist. Interessanterweise zeigte sich hingegen, dass im Falle von His-Alr3497 eine Bindung des Proteins unter nativen Bedingungen an eine Heparin-Cellulose-Matrix möglich war (s. Abbildung 19). Dieser Versuch wurde von Maike Gieseke aus der Arbeitsgruppe für Molekulare Genetik der TU Kaiserslautern durchgeführt.

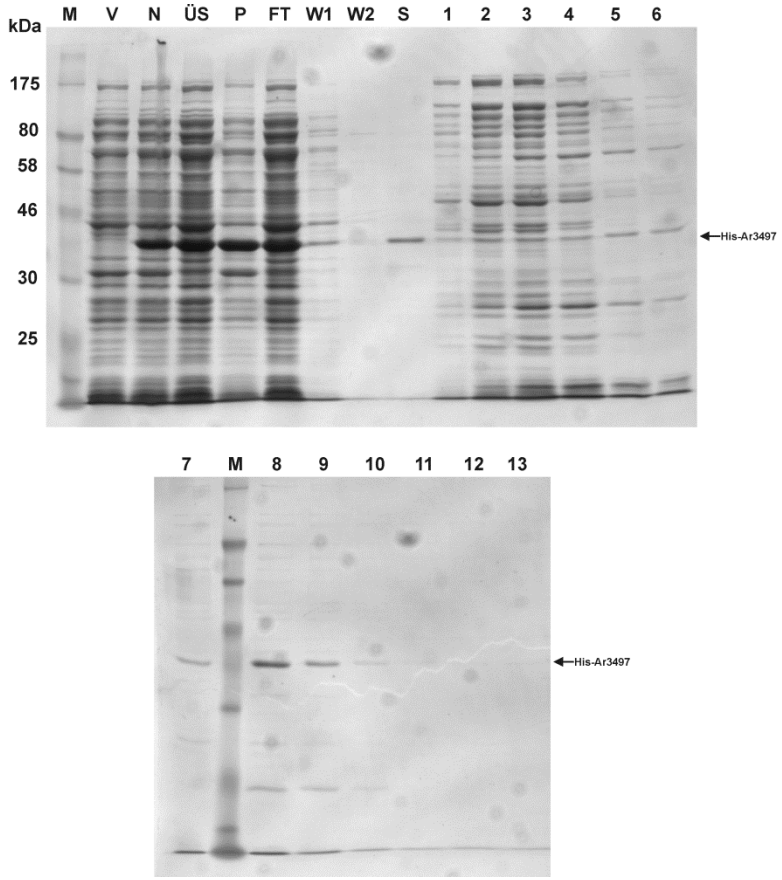


Abbildung 19: Aufreinigung von His-Alr3497 über Heparin-Cellulose. Lysate von T7 Express-Zellen, die His-Alr3497 in 4h bei 16 °C mit 0,5 mM IPTG exprimiert haben, wurden wie in Abschnitt 2.2.3.3 beschrieben hergestellt und mit äquilibrierter Heparin-Cellulose inkubiert. Ein Teil des löslichen Fusionsproteins konnte an die Matrix binden und durch Elution mit ansteigenden NaCl-Konzentrationen teilweise eluiert werden. FT = Durchfluss nach Inkubationdes Überstandes mit Säulenmaterial, W1 und W2 = Durchfluss nach Waschen des Säulenmaterials, S = Säulenmaterial nach Elution, 1 bis 13 = Elutionen. 12 % SDS-PAGE, Coomassie Brillant Blue R250. M = ColorPlus Prestained Protein Marker Broad Range, New England Biolabs. Durchgeführt von Dipl.-Biol. Maike Gieseke von der Abt. Molekulare Genetik der TU Kaiserslautern.

Heparin-Cellulose wird häufig zur Affinitätsreinigung von DNA-/RNA-bindenden Proteinen eingesetzt, was in diesem Falle darauf hinweist, dass His-Alr3497 in seiner nativen Faltung vorliegt. Im Folgenden konnte jedoch das Ergebnis, das in Abbildung 19 dargestellt ist, mit neueren Chargen Heparin-Cellulose nicht reproduziert werden. Das hier isolierte Protein wurde, ebenso wie Rohlysate ohne

weitere Aufreinigungsschritte, in einem Assay auf Aktivität überprüft, dessen Ergebnisse im folgenden Abschnitt dargestellt sind.

3.2.3 RT-Aktivitätsassays

Um die Aktivität der reversen Transkriptasen in dieser Arbeit nachzuweisen, wurde ein Versuchsaufbau analog zu Matsuura et al. gewählt (Matsuura et al., 1997). Hierbei wird eine cDNA-Synthese mit einer reversen Transkriptase in Gegenwart eines Reaktionspuffers durchgeführt, der neben variablen Pufferionen, Salzen und Detergenzien dNTPs enthält, von denen eine Spezies zusätzlich radioaktiv markiert vorliegt. Der Reaktionsansatz wird nach Inkubation auf DE81-Filter aufgetragen, die lediglich höhermolekulare Nucleinsäuren, jedoch keine Einzelnucleotide binden. Nichtinkorporierte Nucleotide werden in mehreren Waschschritten entfernt und die verbliebene Aktivität über Szintillationszählung bestimmt. Der Assay konnte in der Form, wie er in Abschnitt 2.2.4.1 ausführlich beschrieben wird, etabliert werden. Als Positivkontrolle für aufgereinigte RTs wurde anfangs eine kommerzielle reverse Transkriptase (MonsterScript, Epicentre/Biozym GmbH) verwendet, später die Moloney Murine Leukemia Virus-RT (M-MLV-RT), die wie in Abschnitt 2.2.3.3 beschrieben aus *E.coli* aufgereinigt wurde und somit einen vergleichbareren Hintergrund ergibt. Ebenso wurden Rohlysate von *E.coli* nach Expression des jeweiligen Fusionsproteins getestet, um auszuschließen, dass die reverse Transkriptase erst durch den Aufreinigungsprozess an Aktivität verliert. Hier wurden Rohlysate von *E.coli* nach Expression der M-MLV-RT als Positivkontrolle verwendet, als Negativkontrolle wurden *E.coli*-Zellen mit dem Vektor pBS transformiert und im Folgenden ebenso behandelt wie die Zellen der Positivkontrolle. Variiert wurden jeweils die Reaktionspuffer hinsichtlich pH-Wert, Salzkonzentration, Additive, und divalenten Metallkationen; Experimente wurden bis zu drei Mal durchgeführt, und jede Probe stets doppelt bestimmt. Abbildung 20 zeigt beispielhaft ein Ergebnis dieser Versuche; eine Gesamtübersicht über die in den Assays getesteten Bedingungen wird in Tabelle 19 dargestellt.

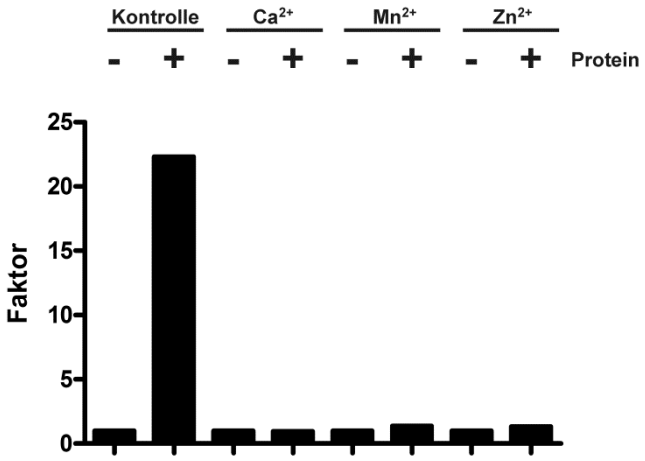


Abbildung 20: Ergebnis eines RT-Aktivitätsassays. Aufgereinigtes His-Alr3497 wurde wie in Kapitel 2.2.4.1 beschrieben mit einem RNA-Template und einem DNA-Primer inkubiert. Der Reaktionspuffer enthielt zudem die angegebenen Metallkationen in 0,5 mM-Konzentration. Negativkontrollen bestanden aus einem identischen Reaktionsansatz, in dem His-Alr3497 gegen ein gleiches Volumen Wasser ausgetauscht worden war. Nach Reaktion wurden die Ansätze auf DE81-Filter aufgetragen, gewaschen und die verbliebene Aktivität über Szintillationsmessung bestimmt. Faktoren wurden berechnet aus dem Quotienten der Aktivitäten einer Probe und der zugehörigen Negativkontrolle; Aktivitäten der Negativkontrollen wurden mit sich selbst in Relation gesetzt und berechneten sich daher stets zu 1.

Tabelle 19: Variationen der RT-Aktivitätsassays

Puffersystem	*pH-Wert*	*Salz*	*Divalentes Ion*	*Sonstiges*
50 mM MES	4,5	10 mM KCl	10 mM $MgCl_2$	
50 mM MES	5	10 mM KCl	10 mM $MgCl_2$	
50 mM MES	5,5	10 mM KCl	10 mM $MgCl_2$	
50 mM MES	6,5	10 mM KCl	10 mM $MgCl_2$	
50 mM MOPS	6	10 mM KCl	10 mM $MgCl_2$	
50 mM MOPS	7	10 mM KCl	10 mM $MgCl_2$	
50 mM MOPS	7,5	10 mM KCl	10 mM $MgCl_2$	
50 mM Tris	6,5	10 mM KCl	10 mM $MgCl_2$	
50 mM Tris	7,5	10 mM KCl	10 mM $MgCl_2$	
50 mM Tris	7,5	100 mM KCl	10 mM $MgCl_2$	
50 mM Tris	7,5	1 mM KCl	10 mM $MgCl_2$	
50 mM Tris	7,5	10 mM KCl	-	
50 mM Tris	7,5	10 mM KCl	0,5mM $CaCl_2$	
50 mM Tris	7,5	10 mM KCl	0,5mM $MnCl_2$	
50 mM Tris	7,5	10 mM KCl	0,5mM $ZnCl_2$	
50 mM Tris	7,5	10 mM KCl	10 mM $MgCl_2$	0,05% NP-40
50 mM Tris	8	10 mM KCl	10 mM $MgCl_2$	
50 mM Tris	8,5	10 mM KCl	10 mM $MgCl_2$	

Es konnte unter den getesteten Pufferbedingungen bisher keine Aktivität detektiert werden.

3.3. Aufreinigung und Funktionsanalysen des akzessorischen Proteins Alr3496 aus *Nostoc* sp. PCC 7120

3.3.1 Klonierung

Im bioinformatischen Teil der vorliegenden Arbeit wurden Analysen der akzessorischen Proteine, die mit einigen DGRs assoziiert sind, durchgeführt. Es konnte eine Gruppe von Proteinen identifiziert werden, die über ein neuartiges Consensus-Motiv verfügen und jeweils relativ hohe isoelektrische Punkte aufweisen (Mittelwert ~ pH 9,73 $\pm$ 0,58). Eine naheliegende Vermutung ist daher, dass es sich um nucleinsäurebindende Faktoren handelt, denen im DGR-Mechanismus mehrere, bisher ungeklärte Rollen zukommen könnten. Da bioinformatische Methoden allein nicht ausreichen, um dieser Theorie nachzugehen, wurde ein Vertreter dieser Proteine – Alr3496 aus *Nostoc* sp. PCC 7120 – ausgewählt, um seine Funktion experimentell näher zu bestimmen. Der offene Leserahmen des Alr3496-Proteins konnte über ein rekombinationsbasiertes Verfahren (GeneArt Seamless Cloning & Assembly Kit, Invitrogen/Life Technologies) erfolgreich in den Vektor pET24b kloniert werden, das Konstrukt erhielt die Bezeichnung pTS31. Über dieses ist eine Expression des Proteins mit einem C-terminalen 6xHis-Tag möglich. Die Sequenz des Inserts und der flankierenden Plasmidbereiche wurde durch Sequenzierung verifiziert, die zugehörige Vektorkarte kann dem Anhang entnommen werden.

3.3.2 Rekombinante Expression in *E.coli* und Aufreinigung

3.3.2.1 Expression bei 16 °C vs. Expression bei 37 °C

Die Expression erfolgte zunächst im Stamm T7 Express in LB-Medium, mit 0,1 mM IPTG als Inducer und bei 16 °C für etwa 6,5 Stunden. Wie man Abbildung 21 entnehmen kann, haben diese Bedingungen bereits zu einer guten Expressionsrate des Proteins geführt; darüber hinaus zeigte sich, dass ein großer Teil des Proteins nach Lyse in der löslichen Fraktion lokalisiert war. Im Anschluss hieran wurde von Lena Conrad, B. Sc., (Abt. für Molekulare Genetik der TU Kaiserslautern) untersucht, wie stark die Expression in 5 Stunden bei 37 °C ist, und wie sich diese Inkubationstemperatur auf die Löslichkeit des Proteins auswirkt. Abbildung 21 zeigt, dass auch hier eine deutliche Produktion des Proteins in löslicher Form erfolgt.

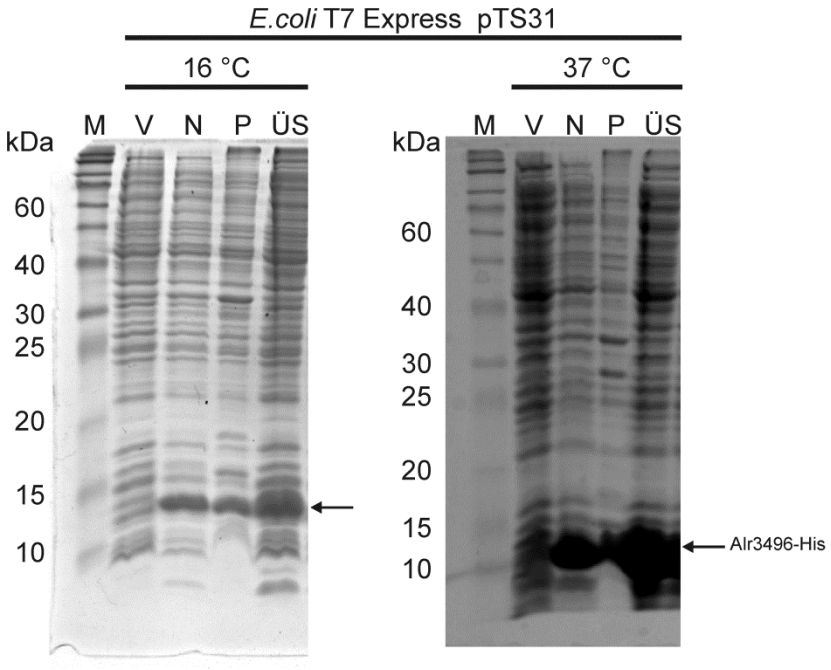


Abbildung 21: Expression und Löslichkeitsstatus von Alr3496-His. *E.coli* T7 Express-Zellen wurden wie zuvor beschrieben transformiert und eine Expression der Zielproteine bei 16 °C für 6,5 h (links) bzw. 37 °C für 5h (rechts) mit je 0,1 mM IPTG durchgeführt. Deutliche Expression eines größtenteils löslichen Proteins von ca. 14 kDa ist in beiden Fällen erkennbar. 15 % SDS-PAGE, Coomassie Brillant Blue R250. M = ColorPlus Prestained Protein Ladder, Broad Range, New England Biolabs

3.3.2.2 Aufreinigung des Proteins Alr3496

Die Isolierung und Reinigung des Proteins erfolgte über eine Ni-NTA-Matrix. Die Proben, die während der Aufreinigung genommen und per SDS-PAGE analysiert wurden, sind in Abbildung 22 dargestellt.

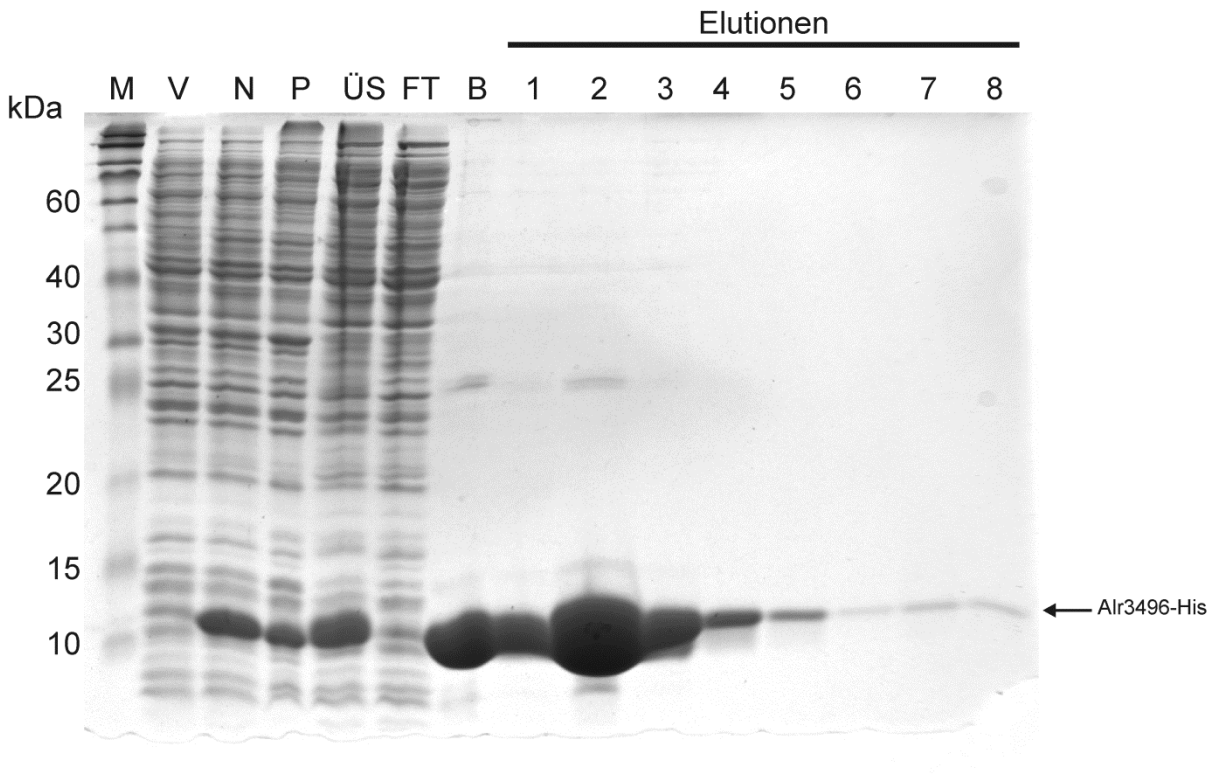


Abbildung 22: Aufreinigung von Alr3496-His. Die Expression des Proteins erfolgte wie zuvor beschrieben in *E.coli* T7 Express-Zellen bei 16 °C. Die lösliche Fraktion der Zelllysate wurde mit Ni-NTA inkubiert. An der Durchflussprobe nach Inkubation (FT) und der Probe der gewaschenen Ni-NTA-Agarosebeads (B) erkennt man, dass das Protein vollständig an die Matrix gebunden werden konnte. Das Protein konnte mit 500 mM Imidazol vom Säulenmaterial eluiert werden. 15 % SDS-PAGE, Coomassie Brillant Blue R250. M = ColorPlus Prestained Protein Ladder, Broad Range, New England Biolabs

Es zeigte sich, dass Reinheit und Stabilität des Proteins stark von der verwendeten Salz-, Glycerol- und Imidazolkonzentration des Puffers abhängen, während das Puffersystem wahlweise Tris oder Phosphat sein kann. Moderatere Salzkonzentrationen von 300 mM NaCl führten zwar zu Löslichkeit des Proteins und zu einer (schwächeren) Bindung an das Säulenmaterial, eine Elution des Proteins war allerdings nur in geringen Mengen möglich. Der Großteil verblieb an die Matrix gebunden, was häufig ein Zeichen von Präzipitation des Proteins ist. Mit einer höheren Salzkonzentration von 1 M NaCl war hingegen eine effiziente Matrixbindung und nachfolgende Elution möglich. Aufgrund der starken Bindung des Proteins an das Säulenmaterial können zudem sehr stringente Waschbedingungen (100x Säulenvolumen mit 100 mM Imidazol) verwendet werden, um unspezifisch gebundene Proteine oder endogene Proteine mit mehreren aufeinanderfolgenden Histidinen effektiv zu entfernen. Detergenzien wurden nicht verwendet, die Aufreinigung erfolgte zudem unter nicht-reduzierenden Bedingungen. Das Protein eluiert erst bei hohem Imidazolkonzentrationen von 500 mM vollständig von der Ni-NTA-Matrix als Protein mit einem apparenten Molekulargewicht von 12 bis 14 kDa. In einigen Fällen sind weitere Banden bei ca. 28 kDa, 42 kDa, 56 kDa und auch höher zu erkennen. Hierbei handelt es sich sehr wahrscheinlich um Oligomere des Proteins.

3.3.2.3 Lagerung und Langzeitstabilität

Die Umdialysierung des Proteins in den Lagerpuffer wurde über ein kurzes Pufferscreening optimiert: hierbei wurde jeweils das gleiche Volumen eluierten Proteins in vier verschiedenen Dialysepuffern über Nacht inkubiert, am nächsten Tag zentrifugiert und der Proteingehalt des Überstands gemessen (s. Tabelle 18).

Tabelle 20: Abhängigkeit der Stabilität von Alr3496 von der Glycerol- und NaCl-Konzentration

Puffer	*NaH_2PO_4, pH 8,0 [M]*	*Glycerol [%]*	*NaCl [M]*	*Gesamtausbeute [mg]*
1	0,01	50	0,15	1,2
2	0,01	50	0	1,5
3	0,01	5	0,15	0,6
4	0,01	5	0	0,2

Es zeigte sich, dass die Stabilität stark von der verwendeten Glycerolkonzentration abhängt; Puffer 1 und 2, die jeweils 50 % Glycerol beinhalteten, zeigten die größte Ausbeute löslichen Proteins. Die Natriumchloridkonzentration scheint zunächst keinen großen Einfluss auf die Stabilität des Proteins zu haben, im Direktvergleich von Puffer 1 und 2 sogar einen möglicherweise negativen Effekt auszuüben. Da jedoch bereits in der Aufreinigung des Proteins beobachtet wurde, dass hohe Natriumchloridkonzentrationen benötigt werden, um eine optimale Bindung des Proteins an das Säulenmaterial zu gewährleisten, und auch die Ergebnisse der Dialyse gegen Puffer 3 und 4 möglicherweise einen stabilisierenden Effekt nahelegen, wurden für die standardmäßige Lagerung 50% Glycerol und 150 mM NaCl eingesetzt; der abweichende Befund für Puffer 2 könnte auf eine Messungenauigkeit zurückzuführen sein. Das Protein ist unter diesen Bedingungen mindestens 6 Monate bei 4 °C stabil.

3.3.3 Strukturanalysen

3.3.3.1 In Silico-Modellierung

Die Primärstruktur des Alr3496-Proteins wurde verwendet, um mit zwei *in silico*-Modellierungsmethoden (Rosetta und i-TASSER) die potentielle Tertiärstruktur des Proteins zu bestimmen. Die jeweils besten Modelle sind in Abbildung 23 dargestellt.

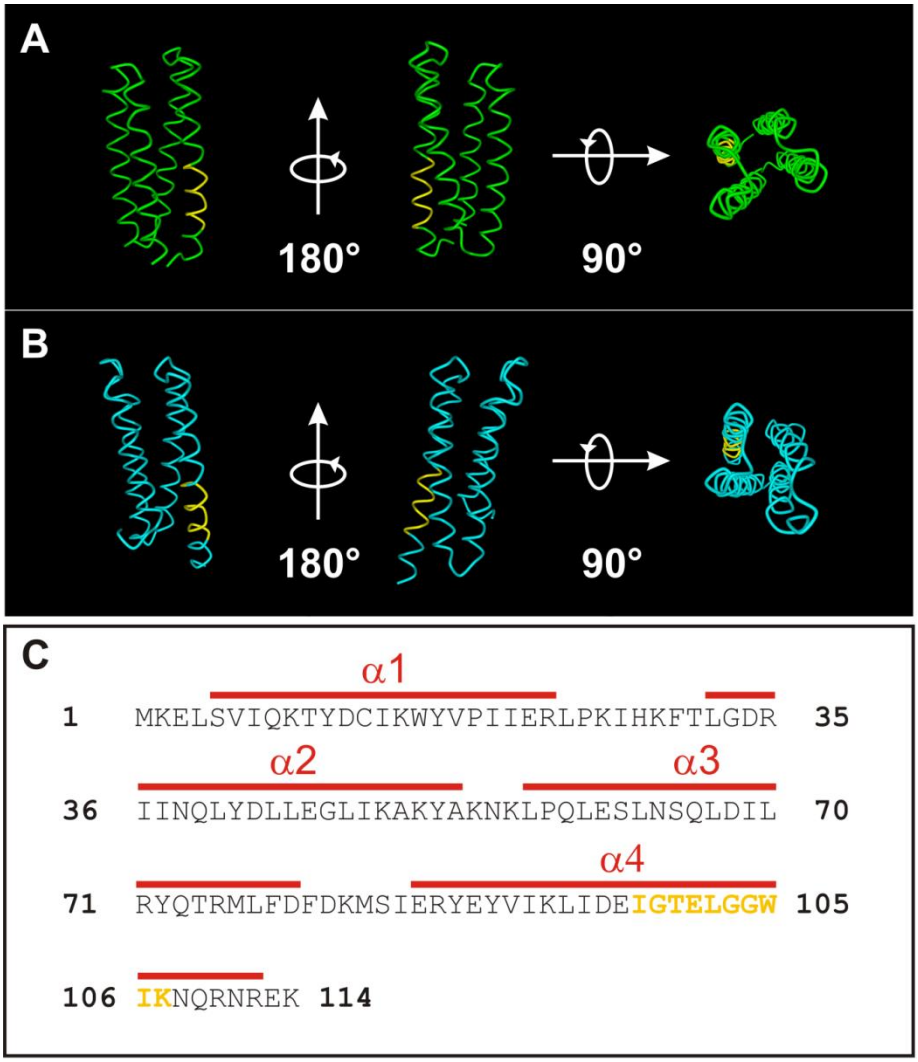


Abbildung 23: *in silico*-Strukturaufklärung von Alr3496. (A) ROSETTA-Struktur, (B) iTASSER-Struktur, (C) Primärstruktur mit eingezeichneten Helixbereichen. Das in Abschnitt 3.1.2.8 ermittelte Motiv ist in den Teilabbildungen jeweils gelb markiert. Die Berechnung der ROSETTA-Struktur wurde von Dipl-Biol. Meik Walther aus der Abteilung für Genetik der TU Kaiserslautern durchgeführt.

Beide Programme errechneten jeweils eine sehr ähnliche Struktur, die eines Vier-Helix-Bündels. Das Consensus-Motiv, das in dieser Arbeit identifiziert wurde (s. 3.1.2.7), befindet sich in Helix α4. Weitere bAvd-Homologe wurden mit i-Tasser analysiert, und es wurden jeweils ähnliche Vier-Helix-Strukturen erhalten. Außerdem wurde über eine BLASTp-Suche eine Homologie zu den sogenannten Ribosomal_S23p-Proteinen festgestellt.

3.3.3.2 Gelfiltration

Um den Oligomerisierungszustand von Alr3496 zu bestimmen, wurde eine Gelfiltration des Proteins durchgeführt. Hierzu wurde in einem ersten Versuch eine 200 µL-Probe des Proteins in 50 mM NaH_2PO_4 (pH 8,0), 150 mM NaCl und 10% Glycerol auf eine äquilibierte Superdex 75-Säule (GE Healthcare) geladen. Während des Filtriervorganges präzipitierte das Protein aufgrund von Wechselwirkungen mit dem Säulenmaterial, daher wurde in einem zweiten Versuch ein alternativer Laufpuffer gewählt, der 500 mM NaCl und 20% Glycerol enthielt. Die Stabilität des Proteins in diesem Laufpuffer wurde zuvor durch Beobachtung der löslichen Proteinkonzentration über mehrere Tage hinweg sichergestellt. Es ergab sich folgendes Elutionsprofil:

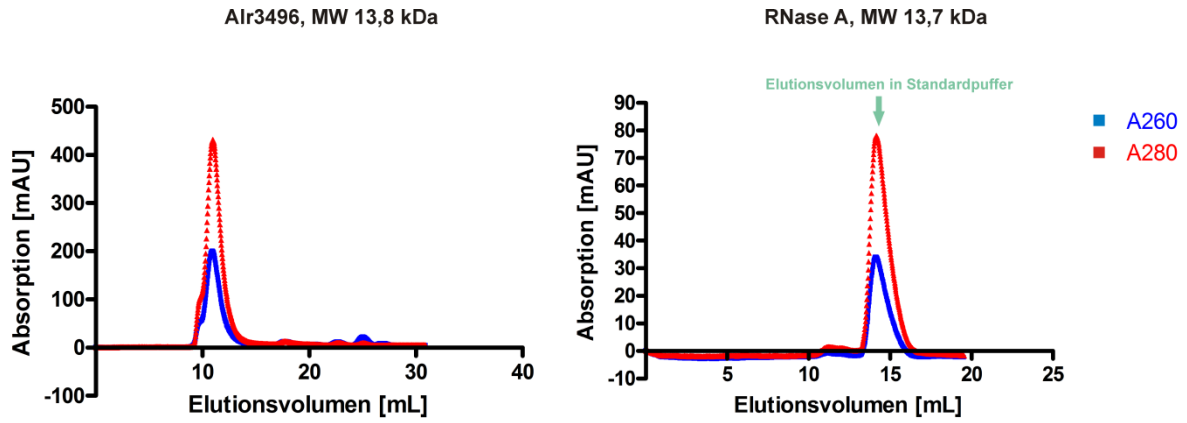


Abbildung 24: Gelfiltration von Alr3496. Zweihundert Mikroliter einer Alr3496-Präparation (1 mg/mL in 10 mM NaH_2PO_4, 20 % Glycerol und 500 mM NaCl) wurden über eine Superdex-75-Säule (GE Healthcare) gegeben und das Elutionsprofil des Proteins ermittelt. Dargestellt sind die Absorptionen bei 280 nm (rot) und 260 nm (blau). Man erkennt einen singulären Peak, was darauf hindeutet, dass das Protein als uniforme Spezies von der Säule eluiert wird. Ein Absorptionsmaximum ist bei 10,97 mL zu erkennen, was einem Molekulargewicht von 56,7 kDa entspricht. Ein Protein vergleichbarer Größe (RNase A) zeigt im selben Puffer ein deutlich unterschiedliches Elutionsverhalten, mit einem Absorptionsmaximum bei 14,15 mL. Zudem entspricht das Elutionsvolumen von RNase A dem unter Standardbedingungen (grüner Pfeil), so dass ausgeschlossen werden kann, dass der verwendete Puffer einen signifikanten Einfluss auf das Laufverhalten hat.

Man erkennt einen einzelnen Peak bei 10,97 mL, was nach Abschätzung über einen kommerziellen Größenstandard einem Molekulargewicht von 56,7 kDa entspricht. Dies würde auf eine Tetramerisierung des Proteins hinweisen. Darüber hinaus scheint das Protein in nur einem Oligomerisierungszustand vorzuliegen, da es sich im Profil um einen einzelnen, diskreten Peak handelt. Bei der zuvor beobachteten zusätzlichen Bande bei ca. 28 kDa scheint es sich um eine dimere Erscheinungsform unter den Bedingungen der denaturierenden SDS-PAGE zu handeln.

Die Werte des Größenstandards wurden in einem Puffer ermittelt, der weniger als 20 % Glycerin enthielt, was eine direkte Vergleichbarkeit mit dem hier ermittelten Elutionsvolumen erschwert. Um das Retentionsverhalten des Proteins in dem modifizierten Laufpuffer besser abschätzen zu können, wurde im selben Puffer RNase A ebenfalls über die Säule gegeben; RNase A gehört zu den Proteinen des Größenstandards und besitzt ebenso wie Alr3496 ein Molekulargewicht von ca. 14 kDa, und sollte somit ein ähnliches Elutionsprofil aufweisen. Das Ergebnis ist in Abbildung 24 dargestellt.

Im Vergleich zum Kalibrierpunkt für RNase A (14,21 mL) wird in dem modifizierten Laufpuffer mit 14,15 mL nur eine geringe Abweichung beobachtet, so dass – zumindest auf diesem einzelnen Messwert basierend – das berechnete Molekulargewicht des Alr3496-Komplexes zutreffen kann.

3.3.3.3 Nicht-reduzierende SDS-PAGE

Das bAvd-Homolog Alr3496 beinhaltet lediglich einen Cysteinrest (Cys13). Im Falle einer Oligomerisierung des Proteins über Disulfidbrücken könnte somit maximal eine Dimerisierung erwartet werden. Eine relativ simple und schnelle Methode, diese Frage zu beantworten, besteht in der SDS-PAGE unter nicht-reduzierenden Bedingungen. Üblicherweise wird dem Laemmli-Ladepuffer ein reduzierendes Agens zugesetzt, um Cysteine im reduzierten Zustand zu halten und somit intermolekulare Disulfidbrückenbindungen aufzulösen. Hier wurden jedoch oxidierende Bedingungen gewählt, um mögliche native Oligomerisierungen, die durch Disulfidbrückenbindungen stabilisiert werden, zu erhalten und in der Coomassiefärbung sichtbar zu machen. Eine Verdünnung des Proteins wurde geteilt und mit reduzierendem oder nicht-reduzierendem Laemmli-Puffer versetzt, und gleiche Volumina der Proben mehrmals auf ein 15% SDS-Polyacrylamidgel aufgetragen, um Pipettierungenaugkeiten beim Laden zu kompensieren. Abbildung 25 zeigt das Ergebnis dieses Versuchs in der Coomassie-Färbung.

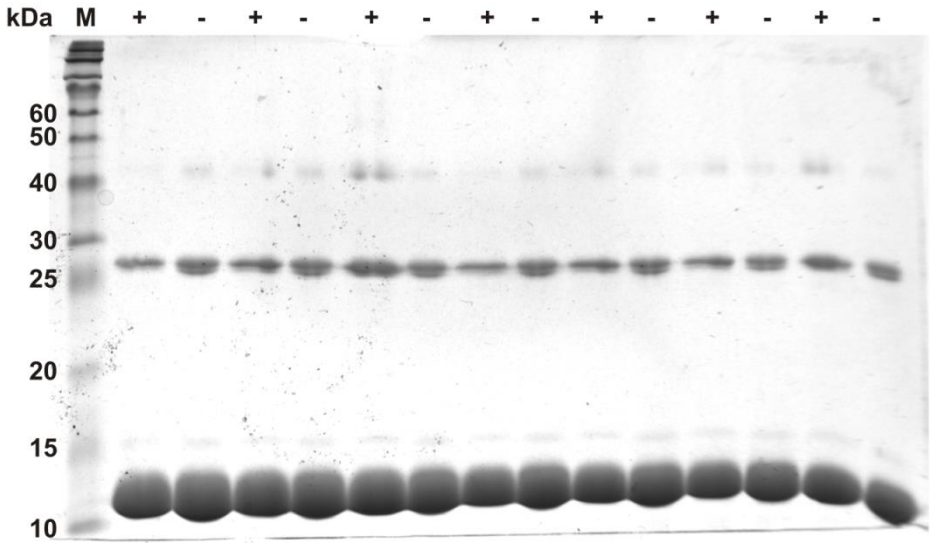


Abbildung 25: Einfluss des Redoxzustands von Alr3496 auf die Quartärstruktur. Alternierend wurden je ~ 2,5 µg des Proteins in einem Ladepuffer mit 2,15 M β-Mercaptoethanol (+) bzw. ohne reduzierendes Agens (-) auf ein 15 % SDS-Polyacrylamidgel aufgetragen. Man erkennt keine signifikante Zunahme der potentiellen Dimerbande bei ~ 28 kDa unter nicht-reduzierenden Bedingungen. 15 % SDS-PAGE, Coomassie Brillant Blue R250. M = ColorPlus Prestained Protein Ladder (10-230 kDa)

Man erkennt keine Veränderung des Bandenmusters bei Vergleich der Proben mit und ohne reduzierendem Agens.

3.3.3.4 Chemische Quervernetzung (Cross-Linking) von Alr3496

Zur weiteren Bestimmung der nativen Stöchiometrie des Alr3496-Proteins wurde eine chemische Quervernetzung des Proteins mit Di (N-succinimidyl)-3,3'-dithiopropionat (DSP) wie in Abschnitt 2.2.3.6 beschrieben durchgeführt. Variierende Proteinmengen zwischen 0,5 µg und 13,5 µg wurden mit DSP-Konzentrationen zwischen 0 und 1,8 mM in einem Volumen von 10 µL inkubiert. Die Produkte wurden mit nicht-reduzierendem SDS-Ladepuffer versetzt und auf ein 15% SDS-Polyacrylamidgel geladen. Die Ergebnisse sind in Abbildung 26 dargestellt. Es war nicht möglich, eine bestimmte Oligomerisierungsform des Proteins über Quervernetzung mit DSP zu isolieren. Beobachtet werden konnte lediglich eine Zunahme unspezifischer Vernetzung mit zunehmender DSP-Konzentration.

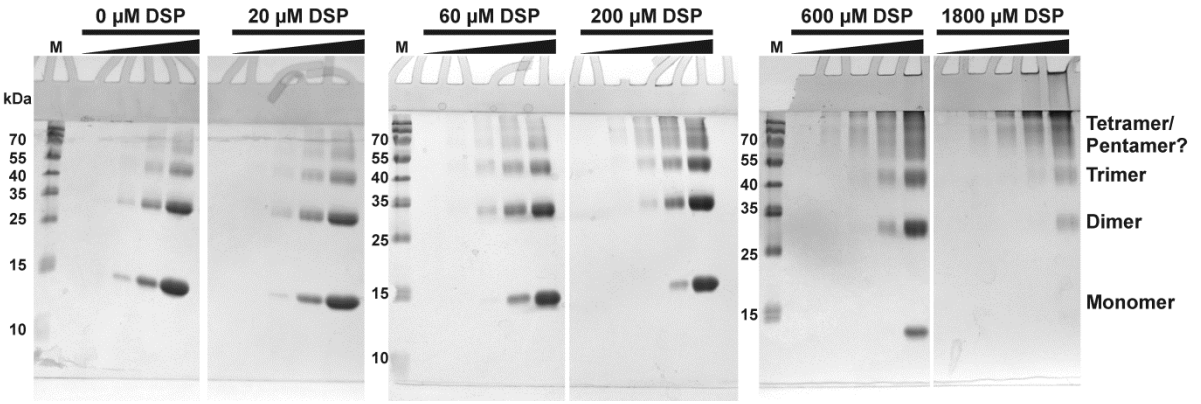


Abbildung 26: Chemische Quervernetzung mit DSP. Variierende Mengen Alr3496 (0 µg, 0,5 µg, 1,5 µg, 4,5 µg und 13,5 µg, durch Keilform über Gelbildern dargestellt) wurden mit den angegebenen Konzentrationen DSP wie beschrieben inkubiert und nach Abstoppen der Reaktion mit ε-Aminocapronsäure ohne Reduktionsmittel über 15 % SDS-PAGE aufgetrennt. Man erkennt keine Zunahme einer bestimmten Oligomerisierungsform bei ansteigenden DSP-Konzentrationen. Deutlich wird hingegegen eine gleichmäßige Abnahme der Banden-intensitäten von Monomeren, Dimeren und schließlich auch Trimeren in Abhängigkeit der DSP-Konzentration. Weiterhin erkennt man unspezifische Vernetzung durch Zunahme nicht-diskreter Bandenmuster im höhermolekularen Bereich, bis hin zu Komplexen, die nicht bzw. kaum in das Sammelgel eingewandert sind (z. B. 1800 µM DSP, höchste Proteinkonzentration).

3.3.4 Biochemische und funktionelle Charakterisierung

3.3.4.1 Bestimmung der Affinitätskonstanten für Nucleinsäuresubstrate

Die Funktion des bAvd-Proteins aus dem *Bordetella* Bakteriophagen und seiner Homologe ist bisher nicht geklärt worden. Ein erster Hinweis auf eine potentielle Funktion kam 2008 in der Publikation von Guo et al. (Guo et al., 2008), in der von einer Bindung des bAvd-Proteins an RNA berichtet wurde; allerdings beruhte diese Aussage auf unveröffentlichten Daten. Die relativ hohen isoelektrischen Punkte der ermittelten bAvd-Homologe (Mittelwert ~pH 9,73 $\pm$ 0,58) würden ebenfalls zu dieser Beobachtung passen. Um das Bindeverhalten von Alr3496 genauer zu charakterisieren, wurde ein Assay nach Wong & Lohman gewählt, in dem das Protein mit einer radioaktiv markierten Nucleinsäure inkubiert wird. Die Anteile der radioaktiven Sonde, die entweder an Protein gebunden oder frei vorliegen, werden durch Abtrennung der Probe über ein Membransandwich aus Nitrocellulose und Nylon und anschließende Quantifizierung des Signals bestimmt (Details s. Abschnitt 2.2.4.2). Abbildung 27 zeigt beispielhaft das Ergebnis eines solchen Versuchs.

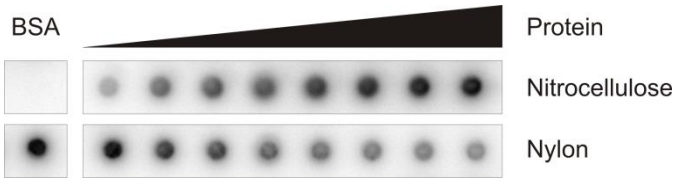


Abbildung 27: Filter Binding-Assays nach Wong und Lohman (Wong and Lohman, 1993). Beispielhaftes Ergebnis eines Filter Binding-Assays.Es wurden Alr3496-Konzentrationen zwischen 0,1 und 0,8 µM in Abständen zu 0,1 µM eingesetzt. Man erkennt eine Zunahme des radioaktiven Signals auf Nitrocellulose mit ansteigener Proteinkonzentration, während die Signalintensität auf Nylon eine reziproke Entwicklung zeigt. Als Negativkontrolle wurden 1,5 µM BSA gewählt.

In ersten Experimenten wurden zunächst Temperatur- und Zeitabhängigkeiten der Reaktion bestimmt, hierbei stellte sich heraus, dass bereits nach ca. 2 Minuten ein reproduzierbarer Endzustand erreicht wird, der sich auch bei längeren Inkubationszeiten nicht ändert (s. Abbildung 28A).

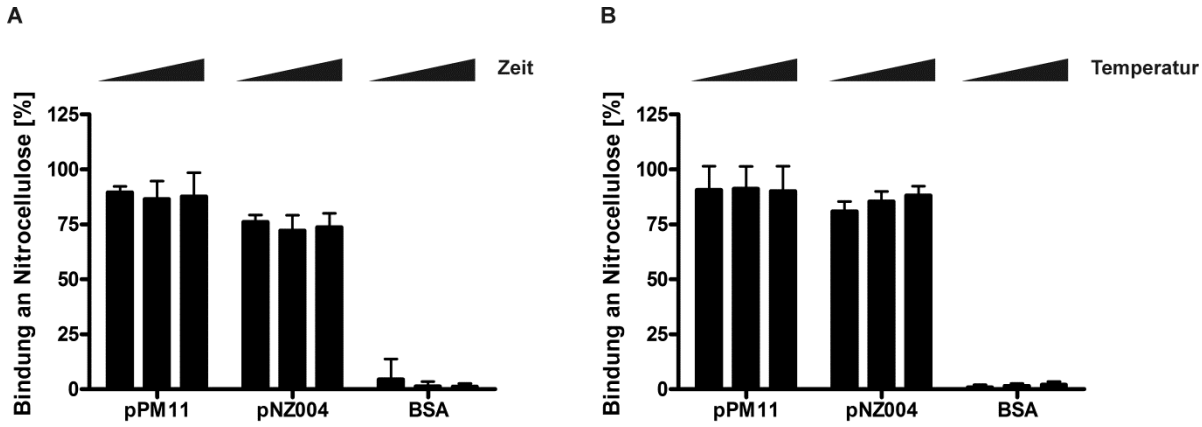


Abbildung 28: Zeit- und Temperaturabhängigkeit der Substratbindung durch Alr3496. Das bAvd-Homolog Alr3496 aus *Nostoc* sp. PCC 7120 wurde mit der TR-RNA des gleichen DGR-Elements (pPM11) oder mit einer kontextfremden RNA (pNZ004) inkubiert und der Anteil des proteingebundenen Signals wie im Text beschrieben bestimmt. Variiert wurden die Inkubationsdauer (A, jeweils 2, 10 und 30 Minuten) sowie die Inkubationstemperatur (B, 0 °C, 4 °C und Raumtemperatur). Als Negativkontrolle wurde die TR-RNA aus *Nostoc* sp. PCC 7120 mit BSA unter jeweils identischen Bedingungen inkubiert. Die Säulen stellen jeweils Mittelwerte aus 3 (A) bzw. 2 (B) unabhängigen Experimenten dar, die Fehlerbalken die zugehörigen Standardfehler.

Um sicherzustellen, dass auch bei variierenden Versuchsbedingungen ein stabiler Endzustand erreicht wird, wurde dennoch die Inkubationszeit jeweils auf 30 Minuten erhöht. Desweiteren konnten keine Unterschiede in den Ergebnissen unter verschiedenen Inkubationstemperaturen (0 °C, 4 °C, Raumtemperatur) festgestellt werden (s. Abbildung 28B); die Versuche wurden zur Vereinfachung daher auf Eis (0 °C) und bei Raumtemperatur (21 – 22 °C) durchgeführt. Es wurde außerdem bestätigt, dass das Alr3496-Protein unter den gewählten Versuchsbedingungen stabil ist;

hierzu wurden entsprechende Verdünnungen des Proteins in 1x Reaktionspuffer für 30 Minuten auf Eis oder bei Raumtemperatur inkubiert, anschließend kurz zentrifugiert, die Proteinkonzentration der Überstände gemessen und mit der Ausgangskonzentration verglichen.

Im Folgenden wurden mehrere Nucleinsäurespezies (jeweils 10 nM) mit 100 nM bis 6 µM Alr3496 inkubiert. Getestet wurden die wahrscheinliche TR-RNA des DGRs aus *Nostoc* sp. PCC 7120, wie sie von P. Möller der Abteilung Molekulare Genetik/TU Kaiserslautern experimentell bestimmt wurde (pPM11-Transkript), eine kontextfremde RNA (die 3‘-UTR des UnaL2 non-LTR-Retrotransposons, pNZ004-Transkript), zwei DNA:RNA-Hybride aus der wahrscheinlichen IMH-Region der TR-RNA (pLM1-Transkript) und einem hierzu vollständig komplementären DNA-Strang bzw. einem DNA-Strang mit IMH*-Mismatches, sowie ebenfalls kontextfremde ssDNA und dsDNA (NZ116 und NZ116:NZ117). Eine Übersicht der Substrate ist in Abbildung 29 gegeben.

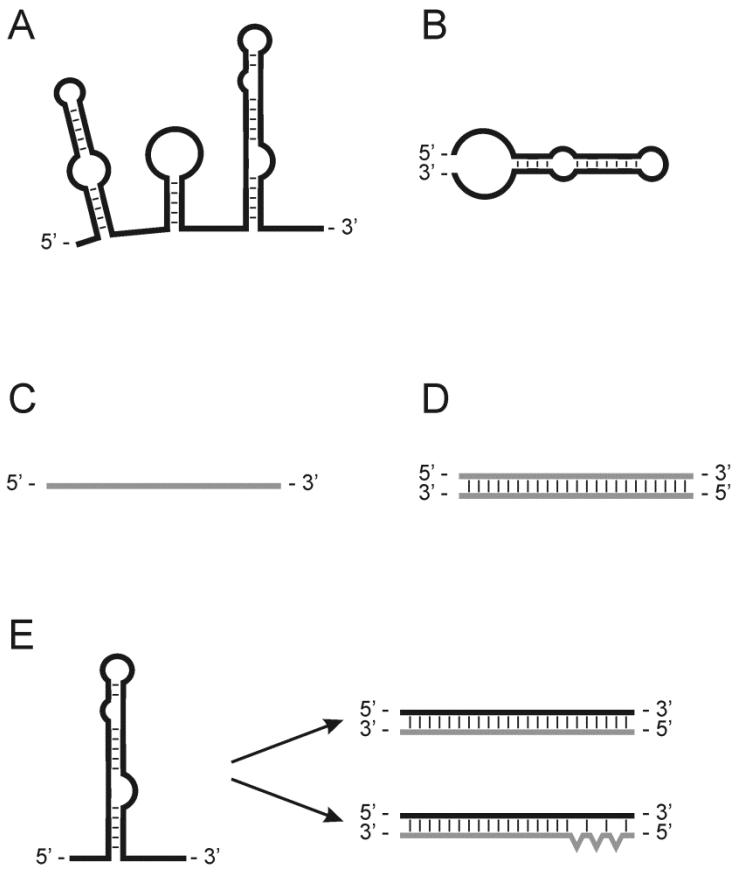


Abbildung 29: Nucleinsäuresubstrate für Filter-Binding Assays. RNA-Spezies sind schwarz, DNA-Spezies grau dargestellt. Getestete Substrate umfassten (A) die wahrscheinliche TR-RNA des DGR-Elements aus *Nostoc* sp. PCC 7120, (B) die 3‘-UTR des UnaL2-Elements, (C) ssDNA, (D) dsDNA, und (E) einen RNA:DNA-Heteroduplex aus der 3‘-Struktur der RNA aus (A), die die IMH beinhaltet (pLM1-Transkript), sowie einem perfekt-komplementären DNA-Strang bzw. einem DNA-Strang mit drei Mismatches, der der homologen Sequenz der zugehörigen variablen Region entspricht.

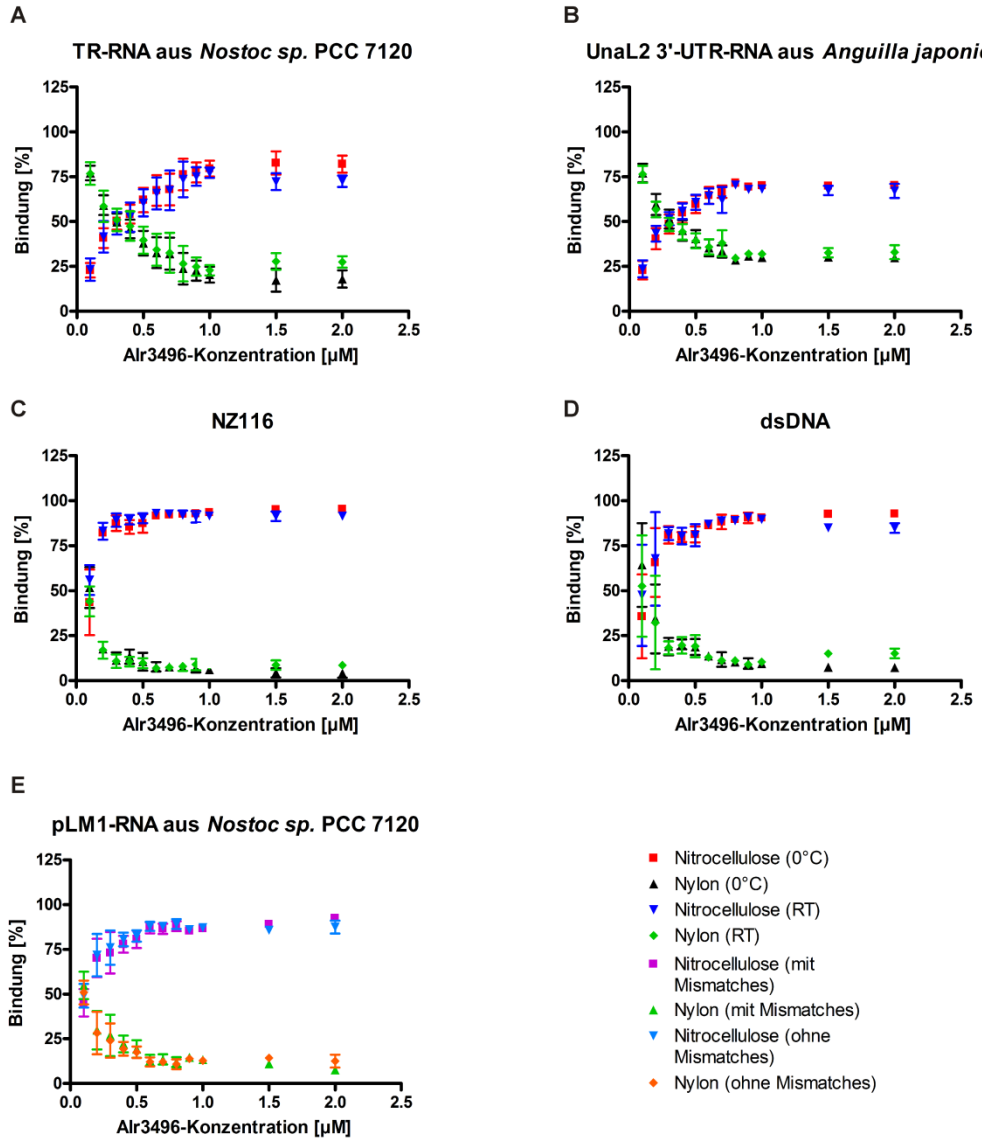


Abbildung 30: Filter Binding-Assays mit variierenden Alr3496-Konzentrationen. Variierende Alr3496-Konzentrationen wurden mit 10 nM RNA (A und B), ssDNA (C), dsDNA (D) und zwei RNA:DNA-Hybriden (E) für 30 min bei 0°C und Raumtemperatur inkubiert (in E nur RT) und die Anteile proteingebundener sowie freier Nucleinsäure wie zuvor beschrieben bestimmt. Dargestellte Punkte sind Mittelwerte aus jeweils drei unabhängigen Experimenten mit Doppelbestimmungen, Fehlerbalken geben die Standardabweichung an.

Man erkennt jeweils eine Bindung des Substrates, wobei ansteigende Proteinkonzentrationen ebenso einen Anstieg des proteingebundenen Substratanteils bewirken, der schließlich in ein Plateau übergeht. Wie aus den Vorversuchen (s. Abbildung 28B) bereits bekannt, wirkt sich die Inkubationstemperatur nicht auf detektierbare Weise auf die Bildung des Enzym-Substratkomplexes aus, die Kurven der Bindung bei 0 °C und RT überlagern jeweils einander. Im Vergleich der einzelnen Substratspezies treten jedoch Unterschiede hervor. Während die wahrscheinliche TR-RNA des *Nostoc* sp. PCC 7120-DGRs (Abbildung 30A) zu maximal ~85 % gebunden vorliegt, wird für eine

kontextfremde RNA (Abbildung 30B) lediglich ~75 % Bindung erreicht. Für DNA-Substrate liegt der gebundene Anteil noch höher (85 bis 95 %, Abbildung 30C und D). Interessanterweise handelt es sich hierbei um kontextfremde DNA-Substrate, d. h. es kann ausgeschlossen werden, dass ein komplexeres Sequenzmuster von Alr3496 erkannt wird. Die getesteten RNA-DNA-Hybride wurden hier so gewählt, wie sie im Verlauf der DGR-vermittelten Hypervariation auftreten können: der RNA-Strang bildete das vermutliche 3'-Endes der TR-RNA, inklusive einer ausgeprägten Hairpinregion und der vermutlichen IMH-Region. Als DNA-Stränge wurden Oligos gewählt, die entweder die nicht perfekt-komplementäre Sequenz der IMH-Region beinhalten (TS77), oder die perfekt-komplementäre Sequenz der IMH*-Region (TS78). Man erkennt in Abbildung 30E für beide Varianten einen nahezu identischen Kurvenverlauf, mit einem Maximum von ~ 98 % Bindung. Aus den Bindungskurven kann direkt die Dissoziationskonstante abgelesen werden. Sie ist ein Maß für die Affinität zweier Bindungspartner, und liegt hier bei der Proteinkonzentration vor, die eine Bindung von 50 % der Substrate bewirkt. Die Ergebnisse sind in Tabelle 21 dargestellt.

Tabelle 21: Ermittelte Bindungsparameter von Alr3496 mit variierenden Substraten

Substrat	*Inkubationstemperatur*	*K_D [µM]*	*Maximal gebundene Substratmenge [%]*
pPM11-RNA	0 °C	0,3	87,5 ± 2,1
	RT	0,3 - 0,4	84,4 ± 3,0
pNZ004-RNA	0 °C	0,3 - 0,4	75,7 ± 0,9
	RT	0,3	74,8 ± 0,8
NZ116	0 °C	< 0,2	95,3 ± 2,0
	RT	< 0,2	91,4 ± 1,9
NZ116:NZ117	0 °C	< 0,2	92,8 ± 2,1
	RT	< 0,2	84,7 ± 2,5
pLM1:TS77	0 °C	0,1 - 0,2	97,9 ± 0,9
pLM1:TS78	0 °C	0,1	97,9 ± 1,1

3.3.4.2 ATPase-Aktivitätsassay mit Alr3496

Über die in Abschnitt 3.3.4.1 beschrieben Filter Binding Assays konnte gezeigt werden, dass Alr3496 Nucleinsäuren binden kann. Als nächstes wurde untersucht, ob es sich bei dem Protein um eine

Helicase und somit um eine ATPase handelt. Der Malachitgrün-Assay ist eine äußerst sensitive Methode zum Nachweis einer ATPase-Aktivität, der auf der Bildung eines Komplexes aus Molybdat und freiem Phosphat, das bei der Spaltung von ATP gebildet wird, beruht(Carter and Karl, 1982). Dieser Phosphat/Molybdatkomplex kann wiederum mit Malachitgrün unter Bildung eines grünen Komplexes reagieren, der über die Absorption bei 640 nm spektrometisch erfasst werden kann.

Die Versuche wurden von Lena Conrad, BSc., im Rahmen ihrer Masterarbeit in der Abteilung Molekulare Genetik/TU Kaiserslautern durchgeführt. Zunächst wurde sichergestellt, dass das Protein im verwendeten Reaktionspuffer ausreichend stabil ist; dies verlief analog zur Stabilitätsbestimmung, wie sie bereits für die Filter Binding-Assays zuvor beschrieben wurde. Anschließend wurden die Beiträge der verwendeten ATP-Lösung bzw. der Proteinlösung selbst zu einer eventuellen Farbreaktion bestimmt, indem ein äquivalentes Volumen der Lösungen direkt zur Malachitgrünfärbelösung gegeben und die Intensität des Farbumschlags bestimmt wurde. Das Ergebnis ist in Tabelle 22 dargestellt.

Tabelle 22: Beiträge der ATP-Lösung und der Proteinpräparation zur Farbreaktion

Lösung	*Phosphatkonzentration [µM]*
ATP	1,4
Alr3496, 3 µM	0,3
Alr3496, 10 µM	0,7
Alr3496, 30 µM	2,2

Man erkennt, dass sowohl die ATP-Lösung als auch die Proteinlösung bereits eine gewisse Menge freien Phosphats in die Reaktion mit einbringen. Daher wurden in der nachfolgenden Auswertung der Proben die Beiträge dieser beiden Komponenten von den ermittelten Konzentrationen abgezogen.

Es wurden mehrere Konzentrationen des Proteins mit variierenden Reaktionspuffern und unterschiedlichen Nucleinsäurezusätzen inkubiert und anschließend mit Malachitgrünlösung gemischt, und die Absorption bei 640 nM gemessen. Als Negativkontrollen wurden die jeweiligen Storagepuffer der Proteinpräparationen in identischer Weise mit den restlichen Reaktionskomponenten inkubiert und gemessen. Bei Experimenten mit Nucleinsäurezugabe wurden die Proben zusätzlich mit und ohne Zugabe von ATP gemessen. Ein repräsentatives Ergebnis dieser Versuche ist in Abbildung 31A dargestellt.

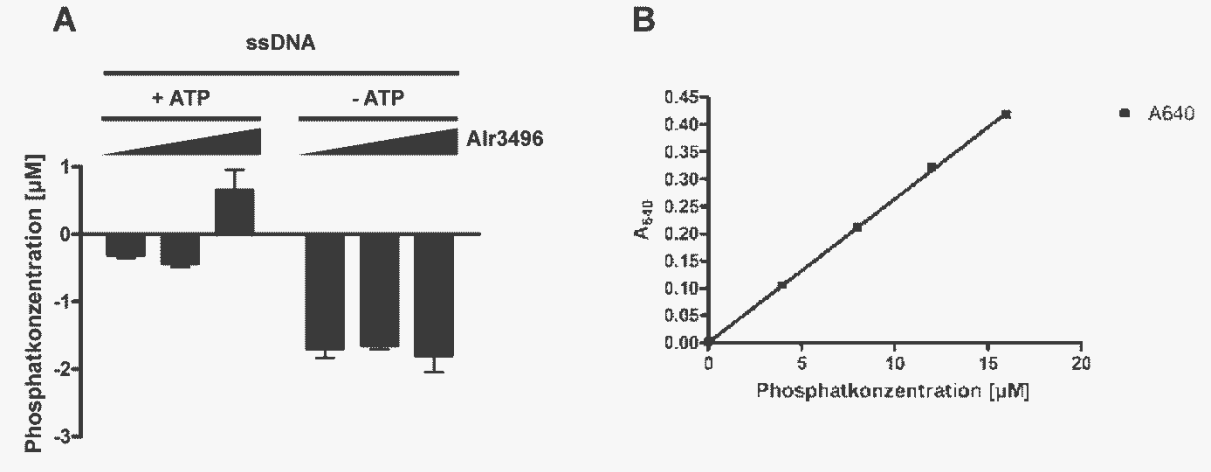


Abbildung 31: ATPase-Assays mit Malachitgrün. (A) Repräsentatives Ergebnis eines Assays. Einzelsträngige DNA wurde mit 3, 10 und 30 µM Alr3496 inkubiert, mit Malachitgrünlösung versetzt und die Farbintensität über die Absorption bei 640 nm gemessen. Die Phosphatkonzentration wurde mit einer Eichgeraden bestimmt, wie sie in (B) abgebildet ist. Außerdem wurden die Eigenbeiträge der Proteinpräparation und ggf. der ATP-Lösung wie in Tabelle 22 angegeben von den Konzentrationen abgezogen. (B) Für jedes Einzelexperiment wurde eine Eichgerade erstellt mit 0, 4, 8, 12 und 16 µM Phosphat. Über lineare Regression ergab sich eine Formel, mit der die Phosphatkonzentrationen der Proben wie in (A) berechnet werden konnten. Man beachte, dass erst ab einer Phosphatkonzentration von 4 µM eine Absorption von 0,1 gemessen wird; dies bedeutet, dass eine exakte Bestimmung von Phosphatkonzentrationen < 4 µM mit dieser Methode nicht möglich ist, da sich die Absorption außerhalb des verlässlichen photometrischen Messbereichs befindet.

Abbildung 31A kann zudem entnommen werden, dass nach Abzug der Beiträge der ATP-Lösung und der Proteinpräparation zum Gesamtsignal teilweise negative Werte errechnet werden. Hierbei handelt es sich um „Konzentrationen“ im Bereich von bis zu ±2 µM; wie man der Eichkurve in Abbildung 31B entnehmen kann, befinden sich diese Konzentrationen außerhalb des verlässlichen photometrischen Messbereichs, so dass sie als irrelevant angesehen werden können. In keinem der getesteten Fälle konnte eine ATPase-Aktivität beobachtet werden.

3.3.4.3 Electrophoretic Mobility Shift-Assay mit Alr3496

Zusätzlich zu den Filter Binding-Assays wurde ein Gel Shift-Experiment durchgeführt, um die Bindung von Alr3496 an ein Nucleinsäuresubstrat zu visualisieren. Auch dieses Experiment wurde von Lena Conrad durchgeführt. Das Ergebnis ist in Abbildung 32 dargestellt.

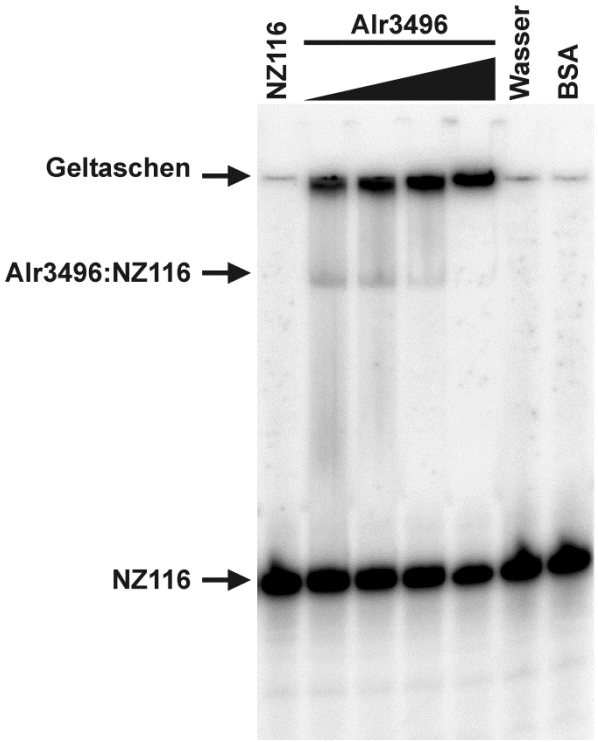


Abbildung 32: Electrophoretic Mobility Shift Assay mit Alr3496. Eine radioaktiv markierte DNA-Sonde (NZ116, 18 nM) wurde für 10 min mit variierenden Konzentrationen Alr3496 (0,66, 2, 6 oder 14 µM) bei 37 °C inkubiert und die Ansätze auf ein natives 12%-Polyacrylamidgel geladen. Man erkennt eine deutliche Verschiebung des Signals relativ zum Signal der freien Sonde bei den drei niedrigsten Konzentrationen, während bei der höchsten Konzentration kein Komplex zu sehen ist. Für jede der getesteten Konzentrationen lässt sich jedoch ein starkes Signal auf Höhe der Geltaschen beobachten, was auf Bildung eines höhermolekularen Komplexes aus Protein und Sonde schließen lässt, und das Fehlen einer Komplexbande bei 14 µM Alr3496 erklären könnte. Als Negativkontrollen wurden den Ansätzen Wasser oder BSA statt Protein zugefügt.

3.3.4.4 Unwinding-Assay mit Alr3496

Trotz der negativ verlaufenen ATPase-Assays mit Malachitgrün könnte es sich bei Alr3496 dennoch um eine Helicase handeln. So ist es z. B. von DEAD-Box-Proteinen bekannt, dass diese teilweise nur dann eine Helicaseaktivität aufweisen, wenn das passende Substrat in der Reaktion vorliegt. Für weitere Untersuchungen wurden Unwinding-Assays mit Alr3496 durchgeführt. Hierbei wurde eine Duplex-DNA, in der einer der Stränge radioaktiv markiert vorlag, mit variierenden Proteinkonzentrationen inkubiert, und die Reaktionen auf ein natives Polyacrylamidgel aufgetragen. Nach elektrophoretischer Auftrennung und Trocknung des Gels wurde eine Aufnahme per Phosphor Imaging erstellt und ausgewertet. Dieser Versuch wurde ebenfalls von Lena Conrad aus der Abteilung für Molekulare Genetik der TU Kaiserslautern durchgeführt.

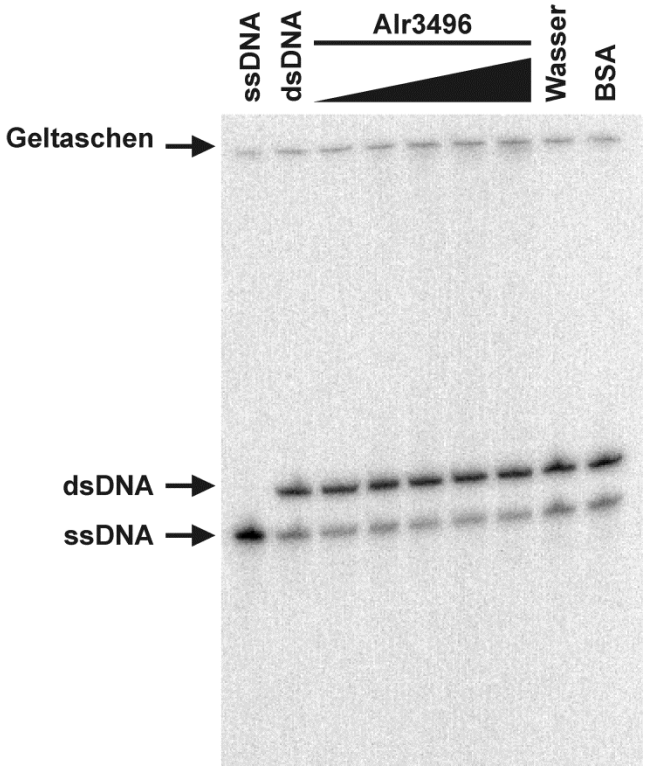


Abbildung 33: Unwinding-Assay mit Alr3496. Variierende Konzentrationen des akzessorischen Proteins Alr3496 wurden mit einem Duplex aus den komplementären Oligonucleotiden NZ116 und NZ117 (je 1,8 nM) für 10 min bei 37 °C inkubiert, und die Reaktionen auf ein natives 12%-Polyacrylamidgel aufgetragen. Getestete Proteinkonzentrationen waren 24 nM, 74 nM, 222 nM, 666 nM und 2 µM Alr3496. Kontrollen wie im Text beschrieben.

Alle Proben enthielten zusätzlich 1 mM ATP. Als Kontrollen wurden jeweils der Einzelstrang und der NZ116:NZ117-Duplex aufgetragen. Außerdem wurde Alr3496 durch Wasser bzw 2 µM BSA ersetzt und im Weiteren wie die restlichen Proben behandelt. Auch hier konnte kein Unwinding der Stränge beobachtet werden.

3.3.4.5 Nucleinsäurechaperonassay mit Alr3496

Als nächstes wurden Versuche durchgeführt, in denen Alr3496 auf eine potentielle Chaperonfunktion für Nucleinsäuren untersucht wurde. Die Versuche wurden analog zu Martin et al. (Martin and Bushman, 2001)konzipiert: zwei zueinander komplementäre Nucleinsäurestränge (NZ116 + NZ117), von denen einer radioaktiv markiert ist, wurden mit dem Protein inkubiert und anschließend auf ein natives Polyacrylamidgel geladen. Nach Elektrophorese wurde über Storage Phosphor Imaging eine Aufnahme des Gels erstellt. Dieser Versuch wurde ebenfalls von Lena Conrad aus der Abteilung für Molekulare Genetik der TU Kaiserslautern durchgeführt.

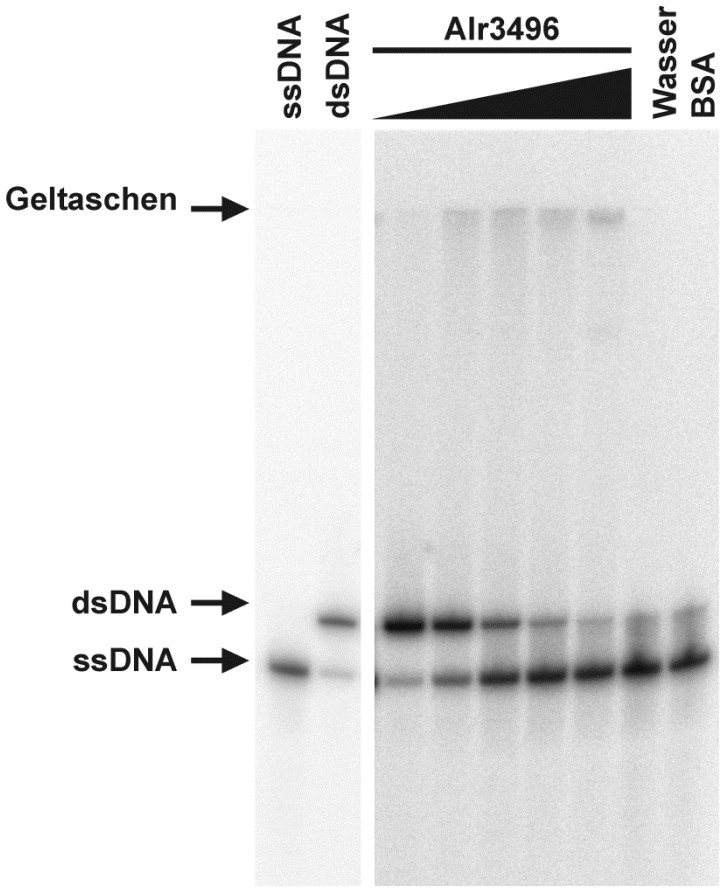


Abbildung 34: Nucleinsäurechaperonassay mit Alr3496. Variierende Konzentrationen des akzessorischen Proteins Alr3496 wurden mit den komplementären Oligonucleotiden NZ116 und NZ117 für 20 min bei 37 °C inkubiert, und die Reaktionen auf ein natives 12%-Polyacrylamidgel aufgetragen. Getestete Proteinkonzentrationen waren 24 nM, 74 nM, 222 nM, 666 nM und 2 µM Alr3496. Man erkennt eine proteinabhängige Hybridisierung der beiden Nucleinsäurestränge, wobei höhere Proteinkonzentrationen den Effekt eher verringern. Auch hier finden sich allerdings mit zunehmenden Proteinkonzentrationen stärkere Signale in den Geltaschen, was auf Bildung höhermolekularer Komplexe hindeutet. Als Laufkontrolle wurden Einzel- und Doppelstrang ohne Protein aufgetragen; diese Proben stammen vom selben Gel, wurden allerdings mit einem anderen Phospho-Imagerscreen belichtet, was den unterschiedlichen Hintergrund erklärt. Sonstige Kontrollen wie im Text beschrieben.

Als Kontrollen wurden jeweils der Einzelstrang und der NZ116:NZ117-Duplex aufgetragen. Außerdem wurde Alr3496 durch Wasser bzw 2 µM BSA ersetzt und im Weiteren wie die restlichen Proben behandelt. Wie man Abbildung 36 entnehmen kann, konnte in der Tat eine proteinkonzentrationsabhängige Hybridisierung der Stränge beobachtet werden. Man erkennt eine deutliche Hybridisierung der beiden Stränge bei 24 nM, die mit zunehmenden Konzentrationen geringer ausfällt. Ein Grund hierfür könnte die Bildung von höhermolekularen Komplexen aus Proteinen und Sonde sein, die bereits in den Gel Shift-Assays in Abschnitt 3.3.4.3 zu erkennen war. Durch diese Komplexbildung ist es dem Protein nicht mehr möglich, eine effiziente Hybridisierung der beiden Stränge zu gewährleisten, weshalb es zu dem hier beobachteten konzentrationsabhängigen Effekt kommt.

IV. Diskussion

4.1 Eine aktualisierte Sicht auf diversitätsgenerierende Retroelemente

4.1.1 155 DGRs können mit DiGReF in öffentlichen Datenbanken ermittelt werden

Diversitätsgenerierende Retroelemente sind eine faszinierende, neuartige Gruppe von Retroelementen, die zahlreiche Merkmale aufweisen, die für Retroelemente und MGEs im Allgemeinen sehr untypisch sind. Zu Beginn dieser Arbeit waren diese Elemente bereits seit 7 Jahren bekannt, doch war unser Bild von ihnen im Allgemeinen noch sehr lückenhaft. Einige weitere Vertreter dieser Elemente waren bereits gefunden worden, doch gab es weder Angaben über ihre genaue Zahl oder ihre Verbreitung, noch eine extensivere Beschreibung einzelner Elemente. Ebenso war nicht geklärt, inwieweit die Erkenntnisse, die aus den Forschungen am *Bordetella* Bakteriophagen-DGR gewonnen wurden, Allgemeingültigkeit besitzen und auf analoge Elemente übertragen werden können, oder ob es sich um individuelle Ausprägungen eines Einzelfalls handelt. Zudem gestalteten sich manuelle Suchen nach diversitätsgenerierenden Retroelementen als äußerst zeitaufwendig und fehleranfällig, und die Sensitivität der Suchstrategie war schwer zu bewerten. In dieser Arbeit wurde das PERL-basierte Programm DiGReF entwickelt, das die Identifizierung von DGRs weitgehend automatisiert. DiGReF konnte insgesamt 155 DGRs in der nr-Datenbank von NCBI ermitteln, von denen der größte Teil (126) zuvor noch nicht beschrieben wurde (Schillinger et al., 2012). Umgekehrt waren in den Ergebnissen sämtliche DGRs enthalten, die bisher publiziert worden sind. Mit diesen Daten wurden zum ersten Mal vergleichende Analysen dieser Elemente ermöglicht, deren Ergebnisse im Folgenden diskutiert werden, und die das Bild von DGRs umfassender definieren als frühere Studien.

4.1.1.1 DiGReF-Analysen sind zuverlässig und vollständig

Eine manuelle Überprüfung der Qualität der Suchergebnisse zeigte, dass es sich jeweils um „echte“ DGRs mit TR/VR-Sequenzen handelt, wie sie bereits für andere DGRs beschrieben wurden. Ebenso konnte beobachtet werden, wie sich die verwendeten Suchparameter auf die Sensitivität bzw. Spezifität auswirken. In dieser Arbeit wurde nach Repeatstrukturen mit einer Mindestlänge von 50 Nucleotiden gesucht, die mindestens 10 Adenine und 7 Adeninaustausche aufweisen sollten. Wurde die Zahl der geforderten Adeninaustausche auf 5 gesenkt, wurden 47 weitere Treffer ausgegeben. Eine genauere Untersuchung dieser Treffer ergab allerdings, dass nur sechs von ihnen potentielle DGRs mit eher geringer Mutageneserate darstellen. Von den restlichen 41 scheinen 13 ein Element in

hoher Kopienzahl in *Arthrospira platensis* NIES-39 darzustellen, mit hoher Homologie zu zwei weiteren Elementen in *Arthrospira maxima* CS-328. Es ist fraglich, ob ein DGR-Element in so hoher Kopienzahl pro Genom vorliegen könnte, da beispielsweise bezüglich der reversen Transkriptase eine enorme Redundanz auf genomischer Ebene bestünde, und andere Konstellationen, wie beispielsweise multiple VRs, die mit einem einzigen DGR-Element verbunden sind, eine effizientere Lösung darstellen. Von den übrigen 28 Elementen lassen sich sieben eindeutig als Gruppe II-Introns identifizieren (Dai and Zimmerly, 2002; Smith and Lee, 2009). Dies zeigt, dass es auch im Umfeld dieser Elemente Repeats mit wenigen „pseudo-adeninspezifischen" Variationen geben kann. Über die restlichen 21 Elemente kann zwar keine exakte Aussage gemacht werden, doch ist es mit den vorstehenden Beobachtungen wahrscheinlich, dass es sich auch hier größtenteils um neuartige Gruppe II-Introns handelt, welche noch nicht in Datenbanken verzeichnet worden sind.

Dieses Ergebnis demonstriert, dass die initial gewählten Parameter geeignet waren, um eine hinreichende Sensitivität bei guter Spezifität in der Suche nach DGRs zu erzielen, und diese Elemente zu ihren nächsten Verwandten, den Gruppe II-Introns, gut abzugrenzen. Interessanterweise gab es RTs mit DGR-typischem Consensusmotiv (s. 3.1.2.4), für die von DiGReF keine TR/VR-Repeats ermittelt werden konnten. Hier handelte es sich meist um Leserahmen auf relativ kleinen Contigs, so dass es DiGReF nicht möglich war, jeweils 5000 Nucleotide up- und downstream des Leserahmens zu extrahieren und zu durchsuchen. Ebenso wurden zwei DGRs durch manuelle Durchsicht gefunden, in denen B->N-Mutationen gehäuft auftraten, und es somit kein Fenster von 50 Nucleotiden gab, in denen ausschließlich Adeninaustausche vorlagen. Eine Identifizierung mit DiGReF war in diesen Fällen daher nicht möglich. Insgesamt sind diese beiden Problemfälle – kurze Contigs und Anhäufungen von B->N-Mutationen – als vernachlässigbar anzusehen. Das Problem lückenhafter Bakteriengenome wurde andernorts bereits angesprochen (Klassen and Currie, 2012), und es existieren erste Ansätze, echte „End-to-End"-Sequenzierungen zu erzeugen, und zudem Lücken, die häufig durch stark veränderliche Sequenzen und repetitive Abschnitte gebildet werden, selbst in bereits sequenzierten Genomen retrospektiv zu schließen (Ribeiro et al., 2012; Shendure and Lieberman Aiden, 2012). Es kann davon ausgegangen werden, dass künftige Sequenzierroutinen eine deutliche Verbesserung der Genomqualität mit sich bringen werden, und eine Analyse durch DiGReF problemlos erfolgen kann. Das Auftreten von zwei DGRs mit B->N-Mutationen in dieser Studie kann hingegen zweierlei bedeuten. Entweder handelt es sich hierbei um Elemente, die erst vor kurzem inaktiviert wurden, und nun der normalen Mutationsrate des Genoms unterworfen sind, was sie für eine weitere Betrachtung eher uninteressant erscheinen lassen würde, da sie in wenigen Generationen höchstwahrscheinlich aus dem Genom elimiert werden. Andererseits könnte es sich um eine DGR-Untergruppe handeln, deren RT-Enzyme keine ausgeprägte Präferenz von Adeninen (oder einer anderen Base) aufweisen. Gleichartige DGR-Typen könnten zwar durch eine Modifikation

des DiGReF-Algorithmus identifiziert werden, mit der lediglich nach Repeatstrukturen in der Nähe DGR-artiger RTs gesucht werden würde, die eine gewisse Zahl von Mismatches beinhalten, andererseits scheint diese Gruppe zahlenmäßig eher gering vertreten zu sein, und kann daher vernachlässigt werden. Insgesamt konnte also demonstriert werden, dass über DiGReF-Analysen DGR-Elemente, welche in den Genomen von derzeit sequenzierten Organismen vorliegen, nahezu quantitativ erfasst werden können. Die vorstehend ausführlich besprochenen Ausnahmen sind größtenteils auf momentane technische Limitationen zurückzuführen, so dass die Abdeckung durch DiGReF in zukünftigen Analysen mit abnehmendem Fragmentierungsgrad sequenzierter Genome eher noch effizienter ausfallen wird.

Um auch potentielle DGR-RTs zu erfassen, die nur eine geringe Ähnlichkeit zu den Query-RTs besitzen, wurde außerdem ein weiteres Set von Kandidaten-RTs über eine zweite PSI-BLAST-Suche erzeugt; als Queries dienten hierbei validierte DGR-RTs aus dem ersten erhaltenen Set, die nur geringe Homologie zu den ersten Query-RTs aufwiesen. Dieses Set wurde ebenfalls mit DiGReF analysiert, allerdings ergaben sich keine neuen Treffer. Hieraus lässt sich schließen, dass bereits über den ersten Suchlauf alle DGR-RTs erfasst werden konnten, und die Ergebnisse in dieser Arbeit ein vollständiges Bild dieser Elemente wiedergeben.

4.1.1.2 DiGReF kann leicht an individuelle Fragestellungen angepasst werden

Im letzten Abschnitt wurde bereits erwähnt, dass die Suchparameter „Repeatlänge“, „Adeningehalt“ und „Adeninaustausche“ in DiGReF einfach zu modifizieren sind. Darüber hinaus kann das Script von DiGReF leicht mit weiteren Scripten verknüpft und an individuelle Fragestellungen angepasst werden. So kann der Output, den das Programm erzeugt, mit einem weiteren Script beispielsweise in ein Format umgewandelt werden, das von einem Programm wie Artemis (Rutherford et al., 2000) geöffnet werden kann. Dies ermöglicht eine übersichtliche visuelle Darstellung der einzelnen DGR-Elemente und ihrer relativen Position zueinander, was in dieser Arbeit entscheidend zur Entwicklung einer Nomenklatur beigetragen hat. Ebenfalls ist es möglich, sich über das Programm Artemis die Leserahmen anzeigen zu lassen, in denen variable Regionen liegen, und eine BLASTp-Suche direkt aus dem Programm heraus zu starten; dies diente in dieser Arbeit der Identifizierung und Extraktion der Zielproteine. Mit einem weiteren Zusatzscript konnten Template Repeats und die entsprechenden variablen Regionen aus den DGR-Outputdateien extrahiert und so dargestellt werden, dass ein direkter visueller Vergleich möglich ist. Darüber hinaus wurde eine Auswertung der Basenaustausche jedes einzelnen Elements deutlich vereinfacht, da das gleiche Script die Zahl jeder Basenspezies in den beiden Repeats zählt und ausgibt. Weitere Programmerweiterungen in Form simpler PERL-Scripte sind denkbar, beispielsweise zur bequemen Umformatierung des Outputs in Formate, die von

sekundären Analyseprogrammen gelesen werden können, oder zur Automatisierung einer komplexeren Auswertung, die auf den Ergebnissen dieser Arbeit beruht. So könnten in zukünftigen Programmversionen neue Datenbankeinträge analysiert, DGRs extrahiert und sofort auf Strukturtyp, Basenaustausche, Anwesenheit von akzessorischen Proteinen und weitere Merkmale hin analysiert werden.

Es konnte gezeigt werden, dass DiGReF auf zuverlässige und quantitative Weise in der Lage ist, DGRs in öffentlichen Datenbanken zu erfassen; die erhaltenen 155 Elemente bilden ein vollständiges Set von genuinen diversitätsgenerierenden Retroelementen. Falsch-positive Ergebnisse lagen dort vor, wo die Parameter zugunsten einer sensitiveren Suche weniger stringent gewählt wurden, während falsch-negative Ergebnisse lediglich im Falle von kurzen Contigs und Anhäufungen von B->N-Mutationen auftraten. Die Verwendung einer weit verbreiteten Programmiersprache erlaubt es, durch zusätzliche Programmodule die Auswertung von DGRs leicht an spezifische Forschungsfragen anzupassen oder ganze Analyseketten zu entwerfen, an deren Beginn DiGReF implementiert wird.

Die Analyse struktureller und mechanistischer Merkmale von diversitätsgenerierenden Retroelementen beschränkte sich bisher auf eine eher geringe, überschaubare Zahl von Elementen (Guo et al., 2008; Guo et al., 2011; Le Coq and Ghosh, 2011). So konnten beispielsweise zwar adeninexklusive Mutationen für alle bisher betrachteten Elemente bestätigt werden, nur war unklar, ob dies tatsächlich für alle DGRs galt, oder nur für eine bestimmte Untergruppe. Genereller ausgedrückt: beobachtete man tatsächlich allgemeingültige Eigenschaften, oder bezogen sich die Gemeinsamkeiten nur auf das verfügbare Set von DGRs, die gerade eben aufgrund einer hohen Homologie zueinander gefunden worden waren? Durch DiGReF ist es erstmals möglich gewesen, ein vollständiges Set von DGRs zu generieren, und dieses auf strukturelle Gemeinsamkeiten und Unterschiede hin zu untersuchen (Schillinger et al., 2012). Die Ergebnisse dieser Analysen werden in den folgenden Abschnitten dargestellt.

4.1.2 Mechanistische Aspekte von DGRs

4.1.2.1 Adeninspezifität ist ein allgemeines Merkmal diversitätsgenerierender Retroelemente

Template Repeats und variable Regionen mit A->B-Mutationen sind das Hauptidentifikationsmerkmal von diversitätsgenerierenden Retroelementen, und wurden für die automatisierte Suche nach DGRs mit DiGReF als Suchkriterium verwendet. Dieses Kriterium wurde gewählt, da bisher publizierte

Studien über DGRs ausschließlich Adeninaustausche in den korrespondierenden variablen Regionen beobachtet haben. Dies muss jedoch nicht zwangsläufig bedeuten, dass nicht auch andere DGR-Varianten existieren, die eine andere Basenpräferenz in ihrem Mutageneseverhalten aufweisen. Aus diesem Grund wurde DiGReF derart modifiziert, dass in analoger Weise nach Sequenzpaaren mit Cytosin-, Guanin- oder Thyminaustauschen gesucht wurde; es konnten allerdings keine derartigen Elemente identifiziert werden, so dass die Adeninpräferenz als ein eindeutiges Identifikationsmerkmal von DGRs bestätigt werden konnte. Diese Suchstrategie kann darüber hinaus auch als Indikator für die Spezifität von DiGReF angesehen werden. Im Zuge der massiven Sequenzierung von prokaryotischen Organismen wurde festgestellt, dass repetitive Sequenzen auch in den Genomen von Bakterien in hoher Zahl zu finden sind, und diversen genetischen Elementen wie MITEs (miniature inverted-repeat transposable elements), REPs (repetitive extragenic palindromic sequences) oder CRISPRs zugeordnet werden können (Delihas, 2011). Sollten repetitive Sequenzen von mindestens 50 Basenpaaren Länge, die sich jeweils an den Positionen einer bestimmten Nucleotidart unterscheiden, schon natürlicherweise in prokaryotischen Genomen vorkommen, könnte keine Aussage darüber getroffen werden, ob diese Variationen tatsächlich durch DGR-Aktivität erzeugt wurden. Die in diesem Abschnitt beschriebenen DiGReF-Analysen ergaben jedoch – bis auf die in Abschnitt 4.1.1.1 genannten Ausnahmen – keine Treffer. Ebenso wurden im Zuge der Programmevaluation von DiGReF zufällig ausgewählte Proteine analog zu den Kandidaten-RTs als Input gewählt. Auch hier wurden von DiGReF keine DGR-typischen Repeatstrukturen ermittelt. Dies zeigt, dass Repeats dieser Art tatsächlich nur in Verbindung mit den Leserahmen für DGR-typische RTs gefunden werden können.

Warum exklusiv Adenine Ziel der DGR-vermittelten Mutagenese sind, ist noch nicht hinreichend untersucht worden. Das plausibelste Modell, das den DGR-Mechanismus erklärt, geht von einer fehleranfälligen cDNA-Synthese aus, die aus strukturellen Besonderheiten der reversen Transkriptase resultiert (Medhekar and Miller, 2007). Denkbar ist, dass Templat-Nucleotide sich während des Synthesevorgangs an einer Position des katalytischen Zentrums befinden, die Adenine mit niedrigerer Stringenz koordiniert als andere Nucleotide, woraus der Einbau nicht-komplementärer Basen in die cDNA folgt. Es stellt sich hier die Frage, ob die Wahl der nicht-komplementären Basen zufällig erfolgt, oder ob Substitutionen in den variablen Regionen einem gewissen Muster folgen, d. h. ob es gewisse Präferenzen in der Substitution von Adenin gegen eine bestimmte Base gibt. Wie in dieser Arbeit gezeigt wurde, ist eine Aussage hierzu schwierig. Zwar können vermehrt A->G-Substitutionen beobachtet werden, während A->T- und A->C-Substitutionen weniger zahlreich gefunden werden, allerdings besitzt diese Beobachtung keine statistische Signifikanz. Es konnte hingegen gezeigt werden, dass G- und T-Substitutionen keiner Normalverteilung folgen, und entweder die beteiligte reverse Transkriptase oder wirtseigene Faktoren die Selektivität

beeinflussen. Eine gesonderte Betrachtung des Substitutionsmusters in einer Gruppe von *Bacteroides*-Elementen mit n = 16 (um den evt. Beitrag von Wirtsfaktoren zu subtrahieren) und des Musters zweier einzelner Organismen mit drei bzw. vier Zielgenen (um den evt. Beitrag der reversen Transkriptase zu subtrahieren) ergab ebensowenig ein klareres Bild. Ob eine Beteiligung von Wirtsfaktoren gänzlich ausgeschlossen werden kann, oder möglicherweise eine Verzerrung der Substitutionsmuster aufgrund biologischer Selektion oder unterschiedlicher Codon Usage der jeweiligen Organismen vorliegt, wird Gegenstand künftiger Betrachtungen und entsprechender Laborexperimente sein.

4.1.2.2 Der SQ-Consensus ist diagnostisch für DGR-RTs

Als DGR-Komponente mit dem höchsten Konservierungsgrad wurden die reversen Transkriptasen in dieser Arbeit am eingehendsten untersucht. Es wurde gezeigt, dass das Consensusmotiv [L/I/V]GxxxSQ in Domäne 4 der Proteine sie deutlich von den reversen Transkriptasen anderer Retroelemente unterscheidet, und somit diagnostisch ist für DGRs. Der markanteste Unterschied zu den entsprechenden Motiven anderer Retroelemente ist das SQ-Dipeptid, welches sonst einem SP entspricht. Allerdings wurden in den hier besprochenen Ergebnissen Variationen dieses Dipeptids beobachtet, d. h. das es bietet ein hinreichendes, aber nicht notwendiges Kriterium, um ein DGR-Element als solches zu bestimmen. Es liegt nahe, das ermittelte Consensusmotiv aufgrund seines exklusiven Auftretens in DGRs mechanistisch mit den charakteristischen Adeninaustauschen zu assoziieren, und es wurde bereits für die reverse Transkriptase des HI-Virus gezeigt, dass Domäne 4 (dort QGxxxSP), und hier besonders Prolin an Position 157, das dem Glutaminrest im DGR-Consensus entspricht, für die korrekte Ausrichtung von Template-Nucleotiden und dNTPs zuständig ist (Klarmann et al., 2000; Smith et al., 1999). In Abschnitt 4.1.2.1 wurde bereits spekuliert, dass die Proteinarchitektur möglicherweise Bedingungen schafft, die die korrekte Koordination von Cytosinen, Guaninen und Uracilen mit dNTPs erlaubt, Adeninen jedoch einen gewissen Bewegungsspielraum lässt, und es somit zu unspezifischer Basenpaarung kommt. Hierfür scheint der Glutaminrest jedoch nicht allein verantwortlich zu sein, da auch Alanin, Histidin und sogar Prolin in den hier untersuchten DGRs an dieser Position auftreten können, und die zugehörigen TR/VR-Sequenzen DGR-typische Substitutionen von Adeninen aufweisen. Es könnte somit das Zusammenspiel des gesamten Motivs, oder weiterer, bisher nicht beachteter Aminosäureseitenketten sein, die zum adeninspezifischen Phänotyp der DGR-RTs führen. Um diese Frage zu beantworten, reichen bioinformatische Analysen allein nicht aus; hilfreich wäre hierzu eine Kristallstruktur einer DGR-RT, im Komplex mit Template und dNTPs, doch gestaltet sich die heterologe Expression und Aufreinigung eines funktionsfähigen Proteins als schwierig (s. Abschnitt 4.2.1). Alternativ könnte zunächst untersucht werden, ob eine Chimären-RT, bestehend aus dem

Korpus einer gut aufzureinigenden reversen Transkriptase (z. B. der M-MLV-RT) und einer DGR-typischen Domäne 4 eine erhöhte Fehleranfälligkeit an Adeninpositionen des Templates aufweist. Diese und weitere Strategien zur Strukturaufklärung der DGR-RTs werden in Abschnitt 4.2 weiter besprochen.

4.1.2.3 Akzessorische DGR-Proteine weisen ein neuartiges Consensusmotiv auf

Über die akzessorischen Proteine der diversitätsgenerierenden Retroelemente ist bislang wenig bekannt. Von den bisher publizierten DGRs zeigt nur eine Teilgruppe zusätzliche Leserahmen, die für akzessorische Proteine codieren (Medhekar and Miller, 2007). Allerdings konnte eine Mutation des *avd*-Leserahmens im *Bordetella* Bakteriophagen-DGR bereits früh zeigen, dass das Genprodukt für die Funktion des DGRs eine essentielle Rolle spielt (Doulatov et al., 2004). Die vorliegende Arbeit trägt einen erheblichen Teil dazu bei, die Funktion dieser Proteinklasse und ihren Beitrag zum DGR-Mechanismus besser zu verstehen, worauf insbesondere in Abschnitt 4.3 noch einmal eingegangen wird, der sich mit den experimentellen Ergebnissen zum akzessorischen Protein Alr3496 aus *Nostoc* sp. PCC 7120 befasst.

Von den 155 DGRs, die im Rahmen dieser Arbeit identifiziert wurden, wiesen 82 (~ 53 %) einen offenen Leserahmen auf, der Homologien mit dem bereits beschriebenen akzessorischen Protein bAvd aus dem *Bordetella* Bakteriophagen-DGR aufweist. Hierüber konnte ein Consensusmotiv erstellt werden, das unter anderem eine hohe Konservierung basischer Reste an seinem C-Terminus beinhaltet. Zusammen mit den relativ hohen isoelektrischen Punkten dieser Proteine (Mittelwert pH 9,73 $\pm$ 0,58, s. Tabelle A2) legt dies die Vermutung nahe, dass akzessorische Proteine eine nucleinsäurebindende Funktion besitzen. Diese Schlussfolgerung diente als Ausgangspunkt für die experimentelle Charakterisierung des Proteins Alr3496 im zweiten Teil dieser Arbeit.

Interessant ist, dass in lediglich der Hälfte der hier identifizierten Elemente akzessorische Proteine gefunden werden konnten. Dies kann einerseits mit der Suchstrategie selbst zusammenhängen, die das bAvd-Protein des *Bordetella* Bakteriophagen-DGRs als Hilfe zur Mustererkennung im Alignment benutzte, doch zeigte auch ein erneutes Alignment der extrahierten offenen Leserahmen, nach Entfernung der zuvor identifizierten 82 Proteine, keine auffälligen Proteingruppen mit hoher Homologie. Dies kann wiederum bedeuten, dass zwar weitere akzessorische Proteine existieren, sie aber recht gering konserviert sind und somit nicht über ein Alignment erfasst werden können, oder dass die verbliebenen DGRs keine akzessorischen Proteine benötigen, weil sie alternative Funktionsmechanismen entwickelt haben. Andererseits ist es ebenfalls möglich, dass das DGR-Element in manchen Fällen kein akzessorisches Protein codieren muss, weil bereits vom Wirtsorganismus selbst ein homologes Protein zur Verfügung gestellt wird, welches das DGR-Element

rekrutiert. Eine einfache BLASTp-Suche zeigt, dass akzessorische DGR-Proteine Homologien zu den ‚ribosomal 23S rRNA'-Proteinen aufweisen, die in einigen Organismen gefunden wurden und eine bislang ungeklärte Funktion besitzen (Afseth et al., 1995; Ralph and McClelland, 1993); hierbei könnte es sich um eben jene endogenen Substitute handeln, die einem potentiell mobilen genetischen Element die Möglichkeit bieten, auf überflüssigen genetischen Ballast zu verzichten. Ein experimenteller Nachweis dieser These könnte über einen Rescue-Versuch in einem *in viv*o DGR-System erfolgen. In Analogie zu den Deletionsversuchen, die von Doulatov *et al.* durchgeführt wurden, könnte diesem Versuchsaufbau zusätzlich ein ‚ribosomal 23S rRNA'-Protein *in trans*, also über ein Expressionsplasmid, hinzugegeben werden. Lässt sich hierdurch die DGR-Funktion wiederherstellen, wäre dies ein deutlicher Hinweis darauf, dass einige DGR-Elemente Wirtszellproteine für ihre eigene Funktion rekrutieren müssen, während andere selbst für die benötigten Faktoren codieren.

4.1.3 Strukturelle Aspekte von DGRs

Die hier ermittelten 155 DGRs zeigen eine große strukturelle Variabilität. Nahezu alle Arrangements der Elemente wurden beobachtet, sowie eine variable Zahl von Zielgenen, die bei einer ausgedehnteren Analyse der Wirtsgenome möglicherweise noch erhöht werden könnte. Diese Beobachtungen lassen mehrere Schlussfolgerungen zu. Einige Retroelemente zeigen eine bevorzugte Nutzung ihrer eigenen mRNA als Template für die reverse Transkription. Dies scheint bei diversitätsgenerierenden Retroelementen nicht unbedingt der Fall zu sein, da TR und RT-ORF in einigen DGRs durch ein weiteres Element wie z. B. ein Zielgen räumlich voneinander getrennt sind (s. Abbildung 8, Strukturtyp 2e). In drei DGRs liegen beide TR und RT sogar auf verschiedenen DNA-Strängen (s. Abbildung 8, Strukturgruppe 4). Diese Beobachtung könnte für zukünftige Experimente von großer Bedeutung sein. Ebenso verrät die relativ kompakte Struktur von DGRs mit einer Gesamtlänge von selten mehr als 2000 bis 3000 bp, dass die Zielgene höchstwahrscheinlich mit dem DGR zusammen weitergegeben werden, da man sonst größere Distanzen zwischen den einzelnen Komponenten der DGRs erwarten würde.

Die Zahl der sequenzierten Organismen ist in den letzten Jahren rapide gestiegen, und wird weiterhin steigen. Somit werden auch neue DGRs entdeckt und veröffentlicht werden. Es ist daher sinnvoll, bereits frühzeitig über eine Klassifikation der DGRs nachzudenken. In dieser Arbeit wurde eine Nomenklatur vorgeschlagen, die auf der relativen Anordnung von TR und RT-ORF zueinander basiert. Das System wurde bewusst offen gehalten, um eine Anpassung an neue Strukturtypen, die in dieser Arbeit nicht beobachtet wurden, zu gewährleisten.

4.1.4 Funktionelle Aspekte von DGRs

4.1.4.1 DGRs weisen eine erstaunlich geringe Inzidenz auf

In der vorliegenden Arbeit konnte zum ersten Mal ermittelt werden, wieviele DGRs es tatsächlich in sequenzierten Organismen gibt, und wie hoch ihre Inzidenz ist. Eskonnten 155 DGR-Elemente in über 6000 Prokaryoten- und Phagengenomen ermittelt werden; somit verfügen ca. 2,6% aller sequenzierten Prokaryoten und Phagen über ein solches Element. Diese Zahl überrascht: warum tritt ein Element, das seinem Wirt eine so potente Anpassungsmöglichkeit bietet, nicht in weitaus höherer Zahl in der Natur auf? Zur Beantwortung dieser Frage ist ein tieferes Verständnis der eigentlichen biologischen Funktion dieser Elemente notwendig, da ein DGR-Element nur dann von seinem Wirt assimiliert werden kann, wenn es sich als für ihn vorteilhaft erweist; ist dem nicht so, stellt es genetischen Ballast dar und wird aus dem Wirtsgenom entfernt. Das Beispiel des *Bordetella* Bakteriophagen-DGRs demonstriert eindrucksvoll, wie Wirtsorganismen DGRs als Werkzeug zur eigenen Evolution einsetzen und sich hierüber einen entscheidenden Selektionsvorteil verschaffen können. Hinweise darauf, welche Funktionen die Zielproteine in anderen Organismen erfüllen, sind bisher jedoch nicht erbracht worden. Die Analysen, die in dieser Arbeit durchgeführt wurden, und deren Ergebnisse in Abschnitt 4.1.4.2 diskutiert werden, können ebensowenig eine befriedigende Antwort geben. Zum gegenwärtigen Zeitpunkt fehlen somit entscheidende Informationen, die Aufschluss über die Bedeutung von DGR-Elementen in diesen Organismen, und somit auch zur Beantwortung der Frage, warum sie eine so geringe Inzidenz aufweisen. Mehrere Theorien wurden allerdings aufgestellt, die diesen Befund erklären könnten (Schillinger and Zingler, 2012), und werden im Folgenden erläutert.

a) *Der Nutzen eines DGRs ist geringer als bisher angenommen.*

   Bisherige Betrachtungen gingen vom Prototyp-DGR des *Bordetella* Bakteriophagen aus, in dem das Element ein Protein hypermutiert, das für das Attachment an Wirtszellen verantwortlich ist (Liu et al., 2002). Durch stetige Hypermutation kann das Protein einem sich ebenso stetig verändernden Wirtszellspektrum angepasst werden, was dem Phagen wiederum mehr Möglichkeiten bietet, sich zu vermehren. Allein von der vorteilhaften Natur eines einzelnen DGRs auf die Gesamtheit zu schließen, ist jedoch möglicherweise ein Trugschluss. Damit ein DGR nutzbringend eingesetzt werden kann, sind mehrere Voraussetzungen nötig. Zunächst einmal muss ein Organismus ein Protein, das eine Hypermutation dieses Umfangs durchläuft, sinnvoll in sein Proteinrepertoire integrieren

können. In dieser Arbeit konnte gezeigt werden, dass DGRs stets eine recht kompakte Struktur aufweisen, und meist nur wenige Kilobasenpaare umfassen. Dies spricht für die Theorie, dass die Zielgene ebenfalls zur DGR-Kassette gehören und mit weitergegeben werden, da sonst keine mechanistischen Gründe gegen eine dezentralisierte Lokalisierung von Zielgenen und den restlichen DGR-Komponenten sprechen würden. Sollte also das Produkt des Zielgens für den neuen Organismus entweder nicht von Nutzen oder gar schädlich sein, wird das Element schnell inaktiviert und deletiert werden. Es kann somit sein, dass die DGR-assoziierten Zielproteine nur für eine bestimmte Gruppe von Organismen von Nutzen sind, wogegen allerdings der Befund aus Abschnitt 3.1.2.5 spricht, dass niemals komplette Klassen von Organismen DGRs aufweisen, sondern pro Klasse nur in vereinzelten, individuellen Organismen Elemente gefunden wurden. Andererseits muss, falls das DGR-Element erst ein neues Zielgen im Wirtsorganismus akquiriert, die Insertion des Elements an einer „passenden“ Stelle erfolgen, und ein Template-Repeat entstehen, der das Zielprotein an einer geeigneten Stelle hypermutiert; ebenso müssen auch nach Hypermutation essentielle Aminosäuren, die für die korrekte Faltung des Proteins erforderlich sind, erhalten bleiben. Nicht zuletzt muss ein DGR-Element, das beispielsweise über horizontalen Gentransfer weitergegeben wurde, den Defensivmechanismen der neuen Wirtszelle entkommen, mit denen die Infiltration durch Fremd-DNA verhindert wird. In Anbetracht dieser Vielzahl von Bedingungen, die ein DGR-Element erfüllen muss, könnte die beobachtete Inzidenz von 2,6% durchaus erklärbar sein. Zum Vergleich können CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeat-Elemente) herangezogen werden. Hierbei handelt es sich ebenfalls um genetische Elemente, die erst in den letzten Jahren vermehrt in prokaryotischen Genomen identifiziert wurden (Jansen et al., 2002). CRISPRs umfassen ein System aus DNA-Repeatstrukturen, die nach Transkription im Komplex mit Proteinen der Abwehr von Fremd-DNA dienen, und somit eine Art prokaryotisches Immunsystem darstellen, das gewisse Ähnlichkeiten zum RNAi-Mechanismus in Eukaryoten aufweist (Brouns et al., 2008; Carte et al., 2008). Diese deutlich autonomeren Elemente sind mit einer Inzidenz von ca. 40 % allerdings ebenfalls nicht so zahlreich vertreten, wie man es aufgrund des potentiellen Nutzens vermuten könnte (van der Oost et al., 2009).

b) *Datenbanken spiegeln keine natürlichen Verteilungen wider.*

Wie bereits in Abschnitt 3.1.2.5 aufzeigt, geben Datenbanken nicht notwendigerweise ein Bild wieder, das den natürlichen Gegebenheiten entspricht. Und so, wie es Verzerrungen in der Verteilung der DGRs in den einzelnen Prokaryotenklassen geben kann, könnte die hier

beobachtete absolute Zahl von 155 DGRs in > 6000 sequenzierten Organismen auf einen nicht-repräsentativen Datenbankbestand zurückzuführen sein. Viele der sequenzierten Organismen stammen aus mikrobiologischen Sammlungen, in denen sie unter konstanten, artifiziellen Bedingungen kultiviert wurden. Sollten DGRs von Organismen eingesetzt werden, um sich an wechselnde äußere Bedingungen schneller anpassen zu können, würde dieser Anpassungsdruck unter Kultivierungsbedingungen fehlen. In der Folge könnte es daher zum Verlust des Elements kommen, da es dem Wirt keinen Vorteil mehr bietet, und genomischen Ballast darstellt. Diese Beobachtung wurde bereits an anderer Stelle gemacht, beispielsweise beim Verlust von Virulenzfaktoren unter Kultivierung (Elena and Lenski, 2003; Fux et al., 2005). Ein Indiz für die Plausibilität dieser Theorie lieferte eine PHI-BLAST-Suche in der env_nr-Datenbank, in der Sequenzen aus nicht näher klassifizierten Umweltisolaten, beispielsweise aus Metagenomsequenzierungsprojekten, deponiert sind. Im Unterschied zu einer BLASTp-Suche verwendet PHI-BLAST neben der Querysequenz noch ein zusätzliches Muster (pattern), mit dem die Treffer gefiltert werden können. Wie in dieser Arbeit gezeigt werden konnte, ist der [LIV]GxxxSQ-Consensus charakteristisch für DGR-RTs, und kann somit als diagnostisches Merkmal bei einer Suche mit PHI-BLAST benutzt werden. Die Suche förderte weitere 106 Sequenzen zutage, die potentiellen DGR-RTs entsprechen; eine genauere Analyse mit DiGReF ist jedoch aufgrund der meist kurzen Contiglänge nicht möglich Wenngleich die Identität dieser Treffer somit noch nicht vollständig geklärt werden kann, weist dieses Ergebnis jedoch darauf hin, dass in Umweltisolaten deutlich mehr DGRs gefunden werden können als in langjährig kultivierten Organismen, und die geringe Inzidenz auf die Qualität des sequenzierten DNA-Materials zurückzuführen ist. Eine genauere Bestimmung der natürlichen Häufigkeiten von DGRs ist somit zum gegenwärtigen Zeitpunkt nicht möglich.

c) *DGRs sind hauptsächlich phagenassoziiert.*

Die dritte Erklärung für die unerwartet niedrige Zahl von DGR-Elementen zieht in Erwägung, dass DGRs hauptsächlich von Phagen genutzt werden, und möglicherweise analoge Funktionen zum *Bordetella* Bakteriophagen-DGR aufweisen. Die in Prokaryoten beobachteten DGRs könnten somit zu einem großen Teil Komponenten eines Prophagen sein; die beobachtete geringe Inzidenz würde daher auf die relativ geringe Zahl von Sonderfällen zurückzuführen sein, in denen im Genom eines Prokaryoten ein Prophage vorliegt, der zudem Träger eines diversitätsgenerierenden Retroelementen sein muss. Eine Verifizierung dieser Theorie könnte über eine verlässliche Prophagenanalyse der

Prokaryotengenome erfolgen; eine solche wurde wie in Abschnitt 3.1.2.6 beschrieben zwar versucht, allerdings bestehen Zweifel an der Aussagekraft der Ergebnisse. So wurden zwar Regionen, in denen DGR-Elemente liegen, vom Programm ProphageFinder als Prophagenregion markiert, allerdings sind diese Regionen häufig deutlich kürzer als ein kompletter Prophage es erfordern würde; zudem fehlen den potentiellen Prophagen häufig Leserahmen, die für essentielle Faktoren wie beispielsweise Integrasen, Polymerasen oder Capsidproteine codieren. Eine hinreichend gesicherte Aussage bzgl. Phagen- oder Prophagenassoziation kann daher aufgrund mangelnder Spezifität nicht gegeben werden. Außerdem wurde die mit der BLAST-Funktion der ACLAME-Datenbank geprüft, ob, sich Homologe zu den Zielgenen der DGR-Elemente aus dem in dieser Arbeit beschriebenen Set in dieser Datenbank finden lassen. Es zeigte sich, dass 43,3 % der Zielproteine einem Phagenprotein zugeordnet werden konnte. Um die Qualität dieses Ergebnises und die Spezifität des Programms bewerten zu können, wurde außerdem eine Liste mit Proteinen analysiert, die aus den gleichen Organismen wie die DGR-Zielgene stammten, mit diesen jedoch nicht assoziiert sind. Es zeigten 21,7 % dieser Kontrollproteine eine Homologie zu einem Phagenprotein. Dieses Resultat lässt zwei Interpretationen zu: entweder sind prokaryotische Genome tatsächlich stark durchsetzt von Prophagen, so dass bei zufälliger Auswahl eines Proteins stets eine Grundwahrscheinlichkeit von ca. 20% gegeben ist, dass es sich um ein Phagenelement handelt, oder die Qualität der Referenzdatenbank ist – ähnlich wie bei ProphageFinder – noch nicht auf einem Stand, der eine zufriedenstellende Zuordnung erlauben könnte, und lässt an dieser Stelle keinen Aufschluss über die Zielproteine der DGRs zu. Mit 43,3 % versus 21,7 % könnte sich allerdings eine leichte Tendenz abzeichnen, dass es deutlich mehr phagenassoziierte Elemente unter den 155 identifizierten DGRs gibt, als bisher erkannt. Dennoch stellt sich weiterhin die Frage, wieso nur drei DGRs direkt in sequenzierten Phagen ermittelt wurden, wenn der Anteil der Virengenome an den hier durchsuchten Sequenzen immerhin 11 % beträgt.

d) *Das entwicklungsgeschichtliche Alter der DGRs beeinflusst deren Verbreitung*

Weitere Überlegungen zogen in Betracht, dass DGRs relativ junge oder relativ alte Elemente sein könnten. Bei entwicklungsgeschichtlich jungen Elementen könnte eine weitreichende Verbreitung noch nicht stattgefunden haben, was die geringe Zahl der in dieser Arbeit identifizierten Elemente erklären würde. Hiergegen spricht jedoch, dass unter diesen Bedingungen DGRs eine eher monophyletische Verteilung zeigen müssten, und die Zielproteine – sofern sie Teil des Elements sind, das weitergegeben wird – hohe Homologie zueinander besitzen sollten (Hennig, 1966). Beides ist jedoch nicht der Fall; DGRs zeigen eine

paraphyletische Verteilung, d. h. sie werden in nahezu allen prokaryotischen Klassen gefunden, wobei keine Klasse eine vollständige Durchsetzung aufweist (Benachenhou et al., 2009; Chen et al., 2011; Nardi et al., 2003). Die Sequenzkonservierung bei den Zielproteinen fällt hingegen äußerst gering aus, wie die hohen E-Werte in BLAST-Suchen der Proteine gegeneinander zeigen (s. Abschnitt 3.1.2.6). Andererseits könnte es sich bei diversitätsgenerierenden Retroelementen wie bei ihren phylogenetisch nächsten Verwandten, den Gruppe II-Introns (Doulatov et al., 2004), um entwicklungsgeschichtlich alte Elemente handeln, die nur noch für einzelne Organismen, die unter einem hohen Anpassungsdruck stehen, von Nutzen sind. Ein zusätzliches Indiz für diese Theorie ist die paraphyletische Verteilung der Elemente in den Klassen.

Zu beachten ist, dass die vorstehenden Erklärungsansätze sich nicht gegenseitig ausschließen, und keine der genannten Theorien für sich alleine die beobachtete Verteilung von diversitätsgenerierenden Retroelementen erklären kann. Es kann vielmehr davon ausgegangen werden, dass der tatsächliche Grund für die geringe Zahl der DGRs in einer Kombination der hier beschriebenen Theorien besteht. So könnte es sein, dass die Akquisition eines nutzbringenden DGRs zunächst relativ aufwendig ist, und Organismen, nachdem sie eine ökologische Nische besetzt haben, keinen Bedarf mehr an einer Mikroevolution eines bestimmten Proteins haben, und das Element nach einigen Generationen verlieren. Lediglich in vereinzelten Fällen erweist sich eine Hypervariation eines Proteins noch immer als Selektionsvorteil, so dass diese Elemente bis heute erhalten geblieben sind. Dass die derzeit sequenzierten Genome keine hinreichend genaue Abbildung der natürlichen Gegebenheiten darstellen, erschwert in einem solchen Szenario zusätzlich eine quantitative Beschreibung dieser Elemente.

4.1.4.2 DGRs hypermutieren eine Vielzahl neuartiger, unbekannter Proteine

Um die biologische Bedeutung eines DGRs verstehen zu können, muss die Funktion des jeweiligen Zielgens bekannt sein. Daher wurden die Zielgene der 155 DGRs aus dieser Arbeit extrahiert und auf ihre Annotation, ihren phylogenetischen Verwandtschaftsgrad, ihre subzelluläre Lokalisation und eine mögliche Phagen- bzw. Prophagenassoziation überprüft. Es konnte gezeigt werden, dass die Zielgene der DGRs sich phylogenetisch in drei größere Gruppen einordnen lassen; während zwei Gruppen sich nicht weiter definieren lassen, umfasst die dritte u. a. alle DGR-Zielproteine, die über automatische Annotationsprozesse die Annotation „Formylglycin-generierendes Enzym“ (FGE) erhalten haben. FGE-Enzyme sind für die Aktivierung von Sulfatasen verantwortlich, und katalysieren die Methylierung eines Cysteinrests im aktiven Zentrum dieser Enzyme, wobei ein Formylglycinrest

gebildet und die Sulfatase aktiviert wird (Dierks et al., 1999; Schmidt et al., 1995). Sulfatasen selbst sind wiederum an diversen Synthese- und Degradationsvorgängen in Eukaryoten und Prokaryoten beteiligt, indem sie die Hydrolyse von Sulfatestern katalysieren (Dierks et al., 2005). Arylsulfatase A beispielsweise ist für die Degradation von Sulfatiden in Nervenzellgewebe verantwortlich; die Degradation dieser Stoffe ist notwendig, da bei Fehlfunktionen des Enzyms die Substrate akkumulieren und Krankheitsbilder wie die metachromatische Leukodystrophie verursachen (Gieselmann et al., 1991; Gieselmann et al., 1994; Polten et al., 1991). Arylsulfatase A ist wiederum Substrat von SUMF1, einem FGE-Homolog im Menschen. Ist dieses mutiert, können beide Enzyme (neben weiteren Substraten) nicht aktiviert werden, und es kommt zum Krankheitsbild der multiplen Sulfatasedefizienz (Cosma et al., 2003; Dierks et al., 2003; Schlotawa et al., 2011).

Warum eine Gruppe von Aktivatorproteinen Ziel der DGR-vermittelten Hypermutation sein soll, erscheint zunächst unklar. Le Coq und Ghosh haben hierzu bereits angemerkt, dass einige Organismen – wie *Treponema denticola* – sich als obligate Anaerobier in einer Umgebung befinden, die die sauerstoffabhängige Umwandlung von Cystein zu Formylglycin nicht zulässt (Le Coq and Ghosh, 2011). Darüber hinaus konnten sie zeigen, dass die variable Region des Zielproteins in *Treponema denticola* (TvpA) der Region in humanem FGE (hFGE) entspricht, in der die beiden katalytisch aktiven Cysteine und ein Serin lokalisiert sind (Dierks et al., 2005). Die entsprechenden Aminosäurereste in TvpA unterliegen der Hypermutation, was einen Funktionsverlust eines FGE-Enzyms zur Folge hätte. Es kann somit davon ausgegangen werden, dass zumindest im Falle von TvpA kein FGE-Enzym vorliegt. Da sich die variablen Regionen der in dieser Arbeit identifizierten DGRs nahezu ausschließlich am 3‘-Ende der Zielgene befinden, werden in den meisten Fällen die äquivalenten Cysteine und Serine ebenfalls hypermutiert werden, was auch für diese Proteine eine Funktion als FGE-Enzym unwahrscheinlich macht. Näherliegender ist es, dass DGR-Zielproteine und FGE-Enzyme entweder einen gemeinsamen Vorläufer besitzen, oder dass eine konvergente Evolution erfolgt ist, an deren Ende eine Enzymklasse steht, deren Architektur für Zielproteine und FGE-Enzyme gleichermaßen vorteilhaft ist. Le Coq und Ghosh versuchten in ihrer Veröffentlichung ebenfalls eine Deutung, und schlugen eine Analogie der Zielproteine zu Abzymen vor, katalytisch aktiven monoklonalen Antikörpern, die spezifisch den Übergangszustand einer Reaktion erkennen und stabilisieren können (Tramontano et al., 1986). Ein solches Konzept wäre durchaus auf DGR-Zielproteine übertragbar. Es könnte sich um Zelloberflächenproteine oder sekretierte Proteine handeln, die die Verwertung von Metaboliten in der Umgebung ermöglichen, die aufgrund ihrer Größe oder Struktur nicht aufgenommen werden können. Über eine solche externe „Vorverdauung“ wäre es für das Bakterium in Zeiten der Energieknappheit möglich, neue Nahrungsquellen zu erschließen. Ebenso denkbar ist eine noch stärkere Analogie zu Antikörpern in Form von Faktoren, die ihrerseits an Proteine auf den Oberflächen feindlicher Bakterienzellen oder von Phagen in der

Umgebung binden und diese binden und/oder lysieren könnten. Durch DGR-vermittelte Hypervariation könnte somit ein Mikrobizidarsenal erzeugt werden, welches auf eine Vielzahl von Zellen oder Phagen in der Umgebung einwirken und schnell an neue Ziele angepasst werden kann.

Eine weitere Untergruppe des „FGE-Enzym“-Astes trägt die Annotation „Concanavalin/A-Typ Lektin“. Lektine sind eine Klasse von Proteinen, die häufig an der Outside-In-Signaltransduktion beteiligt sind (Hebert, 2000). Für einige DGR-Zielproteine wurde bereits ein konserviertes Lektinfaltungsmuster der variablen Regionen experimentell nachgewiesen (Le Coq and Ghosh, 2011; McMahon et al., 2005); interessanterweise scheint die Anlage von Template Repeats dahingehend selektioniert zu werden, dass dieses Faltungsmuster als eine Art Gerüst stets erhalten bleibt, und keiner Hypermutation unterworfen ist. Die Variation hingegen erfolgt an Aminosäurepositionen, die nicht unmittelbar an der Faltung beteiligt sind. Diese Anordnung wirft umgekehrt natürlich die Frage auf, ob die anderen beiden Äste aus Proteinen gebildet werden, die nicht dem Lektinfaltungsmuster folgen, und möglicherweise eigene Lösungen für das Problem der Strukturkonservierung bei gleichzeitiger Hypermutation entwickelt haben. Die derzeit vorliegenden Daten können in dieser Hinsicht noch keinen Aufschluss geben, so dass die Beantwortung dieser Frage Gegenstand künftiger Betrachtungen sein wird.

4.2 Die reverse Transkriptase Alr3497 aus *Nostoc* sp. PCC 7120

Die Aufreinigung einer rekombinant erzeugten, biologisch aktiven reversen Transkriptase stellt nach wie vor eine Herausforderung in den Biowissenschaften dar. Ein Blick in die RCSB Datenbank (Berman et al., 2000) illustriert dies: es existieren 260 Einträge, die reverse Transkriptasen oder mit ihnen assoziierte Strukturen beinhalten. Hiervon entfallen 76 % auf Strukturen zur RT des humanen Immundefizienzvirus (HIV), 9 % auf die Struktur der RT des murinen Leukämievirus und 2 % auf Telomerase-assoziierte Strukturen. Die restlichen 13 % werden größtenteils aus Struktureinträgen gebildet, die nur eine periphere Assoziation mit reversen Transkriptasen haben, beispielsweise RNase H-Domänen oder Integrasen, die aus einem gemeinsamen Vorläuferpolypeptid prozessiert werden. Man erkennt, dass tatsächlich von nur drei unterschiedlichen reversen Transkriptasen Strukturen vorliegen, während weitere Enzyme dieser Klasse bisher noch nicht erfolgreich kristallisiert und analysiert werden konnten.

Die reversen Transkriptasen diversitätsgenerierender Retroelemente unterscheiden sich hauptsächlich in drei Punkten von den bisher beschriebenen RT-Enzymen: sie besitzen keine

RNase H-Domäne, sie weisen ein klassenspezifisches Consensusmotiv in Domäne 4 auf, und ihre Primärstruktur ist mit durchschnittlich 378 Aminosäuren relativ kurz. Da sich kleine Proteine in der Regel aufgrund höherer Löslichkeit besser aufreinigen lassen als größere (siehe u. a. Koschorreck et al., 2005; Palomares et al., 2004; Schlieker et al., 2002), lag die Vermutung nahe, dass auch im Falle der DGR-RTs eine gute *in vitro*-Löslichkeit vorliegen könne, was ein entscheidender Faktor für nachfolgende Aufreinigungsschritte und Funktionsanalysen wäre. Eine Bestätigung dieser Annahme geben bioinformatische Tools wie http://www.biotech.ou.edu/, welche eine Vorhersage über die *in vitro*-Löslichkeit eines Proteins treffen können, das in *E.coli* überexprimiert wird (Davis et al., 1999; Wilkinson and Harrison, 1991). Hierbei werden besonders das Verhältnis von negativ geladenen Resten (Glu und Asp) zu positiv geladenen Resten (Lys und Arg) sowie die Zahl von Aminosäuren, die vermehrt in Turn-Regionen auftreten (Gly, Ser, Asn und Pro), für die Vorhersage berücksichtigt. So wird für die RT des DGR-Elements aus *Treponema denticola* ATCC 35405 eine Wahrscheinlichkeit von 71,8 % auf Löslichkeit ausgegeben, für die RT aus *Bacteroides thetaiotaomicron* VPI-5482 68,6 % und für die RT aus *Nostoc* sp. PCC 7120 zumindest 57,2 %. Wie in dieser Arbeit jedoch gezeigt wurde, erwiesen sich sämtliche getesteten RTs als schwer bzw. gar nicht löslich. Auch die Fusion an größere Tags wie GST oder CBD/Inteine, welche im Allgemeinen einen löslichkeitsfördernden Effekt vermitteln, wirkte sich nicht begünstigend aus, ebensowenig wie eine Herabsenkung der Expressionstemperatur oder Variationen in der Inducer-Konzentration bzw. Medienzusammensetzung. Lediglich im Falle der codonoptimierten RTs aus dem Cyanobakterium *Nostoc* sp. PCC 7120 eine nennenswerte Löslichkeit beobachtet werden, welche schließlich die Isolation des Proteins über Affinitätschromatographie ermöglichte. Nachfolgende Tests ergaben jedoch, dass im Vergleich zu kommerziellen RTs oder einer selbstaufgereinigten M-MLV-RT auch unter variierenden Pufferbedingungen keine detektierbare Aktivität des aufgereinigten Proteins festzustellen war. Somit sind entweder noch keine adäquaten Reaktionsbedingungen gefunden worden, oder die isolierte Form des Proteins ist generell inaktiv, was auf eine zwar lösliche, jedoch falsch gefaltete Variante zurückzuführen sein könnte. Erschwerend kommt hinzu, dass der Aufreinigungserfolg bisher nicht reproduzierbar war, da sich Alr3497 mit anderen Chargen Heparincellulose als der in Abbildung 19 verwendeten nicht isolieren ließ. Somit muss auch die Möglichkeit in Betracht gezogen werden, dass der Aufreinigungseffekt in diesem Versuch nicht aufgrund einer intrinsischen Affinität einer reversen Transkriptase zu Heparin erfolgte, sondern lediglich eine unspezifische Interaktion eines – richtig oder falsch gefalteten – Proteins mit einem Bestandteil der Heparincellulose darstellte. Diese Befunde reihen sich in die Beobachtungen ein, die bereits für die RT aus dem *Bordetella* Bakteriophagen-DGR (bRT) gemacht wurden. Hier erfolgte zunächst eine Aufreinigung aktiven Proteins aus Inclusion Bodies mit anschließender Rückfaltung (Liu et al., 2002); später konnte die Isolation unter nativen Bedingungen durchgeführt werden, doch

scheint die Stabilität dieses Proteins eher gering zu sein, da die Autoren darauf hinweisen, dass das Protein innerhalb von 24 Stunden nach Aufreinigung für Assays verwendet wurde (Alayyoubi et al., 2012). Eine Aufklärung der Proteinstruktur fand bisher nicht statt. Interessanterweise konnte eine Mutante des bRT-Proteins gefunden werden, welche an Position 138 statt eines Aspartatrests ein Glutamin beinhaltet, und welche eine erhöhte Stabilität gegenüber dem Wildtypenzym aufweist (Alayyoubi et al., 2012). Eine analoge Mutante könnte für die RT aus *Nostoc* sp. PCC 7120 (Alr3497) über Mutagenese von D135 erzeugt werden. Sequenzvergleiche zeigen, dass es sich hierbei um das homologe Aspartat zum katalytisch aktiven D110 in der RT des HI-Virus handelt. Auch im Falle von bRT zeigte sich, dass D138 für die katalytische Aktivität von Bedeutung zu sein scheint, da die Mutante lediglich ein Fünftel der Wildtypenzymaktivität aufweist, jedoch könnte selbst eine schwach-aktive Variante von Alr3497 bereits bedeutend zur weiteren Charakterisierung dieser Enzymklasse beitragen.

Es ist offensichtlich, dass klassische Versuche zur Erhöhung der Löslichkeit eines rekombinant erzeugten Proteins, wie verminderte Syntheserate über Absenkung der Inkubationstemperatur oder geringere Inducerkonzentration, Fusion an löslichkeitsfördernde Peptide/Proteine oder Variation des Expressionsstammes im Falle der DGR-RTs nur wenig Erfolg zeigten. In folgenden Arbeiten sollten daher alternative Strategien implementiert werden, von denen einige im Folgenden kurz erläutert werden:

a) *Erhöhung der intrinsischen Löslichkeit durch ‚directed evolution'*

   Eine Möglichkeit, die intrinsische Löslichkeit eines Proteins *in vitro* oder in einem heterologen Expressionssystem zu erhöhen, besteht zunächst in der Erzeugung einer Mutantenbibliothek über DNA Shuffling (Stemmer, 1994) oder fehlerbehaftete PCR-Methoden (Cadwell and Joyce, 1992) und Fusion der Mutantengene an die Leserahmen eines geeigneten Reportergens, beispielsweise GFP oder eines Resistenzfaktors gegen ein Antibiotikum (Fisher et al., 2006; Jiang et al., 2007; Seitz et al., 2010; Waldo et al., 1999). Fluoreszenz oder Überleben in einem antibiotikahaltigen Medium sind somit ein Indikator für Löslichkeit des Fusionsproteins, und der Mutante allein.

b) *‚Rational design' des Proteins über gezielte Mutagenese und limitierte Proteolyse*

   Proteine aggregieren in der Regel über intermolekulare Wechselwirkungen freiliegender hydrophober Seitenketten. In manchen Fällen sind diese Seitenketten für die Funktion des Proteins an sich nicht wichtig, und können deletiert werden. Beispielsweise konnte durch limitierte Proteolyse und Strukturvergleiche der M-MLV-RT mit der HIV-RT festgestellt werden, dass ein aminoterminaler Bereich von 24 Aminosäuren unstrukturiert vorliegt und

keine Entsprechung in der HIV-RT besitzt; dieser wurde daraufhin deletiert (Georgiadis et al., 1995; Najmudin et al., 2000). Einige gezielte Mutagenesen aufeinanderfolgender hydrophober Reste der M-MLV-RT erhöhten die Löslichkeit bei gleichzeitiger Beibehaltung der biologischen Aktivität darüber hinaus beträchtlich (Georgiadis, 2001). Jedoch sind Vergleiche zwischen der HIV-RT, der M-MLV-RT und den DGR-RTs nur von begrenzter Aussagekraft, da wie eingangs erwähnt sich letztgenannte Klasse teilweise beträchtlich von anderen RTs unterscheidet, und die Bestimmung eines „Core"-Enzyms mit großen Unsicherheiten belegt ist. Erfolgversprechender wäre die Nutzung der in dieser Arbeit gewonnen Daten über DGR-RTs. Über eingehendere Sequenzvergleiche könnten stark konservierte Regionen (abgesehen von den bereits besprochenen Domänen 4 und 5) identifiziert werden, während andere Regionen, die eine weniger starke Konservierung aufweisen, entweder deletiert oder in ihrem hydrophoben Charakter abgeschwächt werden könnten.

c) *Erzeugung eines Chimärenproteins*

Zur Erzeugung eines löslichen und aktiven Proteins könnte man sich einer bereits vorhandenen löslichen RT, wie beispielsweise der in dieser Arbeit verwendeten M-MLV-RT, bedienen. Wie gezeigt werden konnte, unterscheiden sich DGR-RTs von anderen RTs unter anderem in dem Motiv in Domäne 4 ([L/I/V]GxxxSQ statt QGxxxSP). Analog zu Versuchen von Kanda & Saigo (Kanda and Saigo, 1993), in denen Domäne 5 der M-MLV-RT gegen die homologe Domäne eines eukaryotischen Retrotransposons ausgetauscht wurde, oder zu Isaka et al. (Isaka et al., 1998), die Teile der RT des Simian Immundeficiency Virus und der des Human Immundeficiency Virus gegeneinander ausgetauscht haben, könnte hier Domäne 4 der M-MLV-RT gegen die einer DGR-RT ersetzt, und das Substitutionsverhalten des erzeugten Chimärenproteins gegenüber dem Wildtypprotein untersucht werden. Nachteilig ist, dass hierüber nur die Analyse der Domäne 4 erlaubt wird, während die Interaktion mit Bindungspartnern wie TR-RNA oder Alr3496 nicht untersucht werden kann. Sollte das so entstandene Protein außerdem keine adeninspezifischen Mutationen erzeugen, wäre unklar, ob dies auf das Fehlen eines zusätzlichen Faktors zurückzuführen ist, oder lediglich darauf, dass die Struktur der M-MLV-RT dies nicht erlaubt.

4.3 Das akzessorische Protein Alr3496 aus *Nostoc* sp. PCC 7120

Für einige DGRs konnte ein zusätzlicher offener Leserahmen identifiziert werden, der sich zwischen Zielgen und Template Region befindet, und dessen Proteinprodukt offenbar essentiell für die Funktion des Elements ist; das Homolog des *Bordetella* Bakteriophagen-DGRs wurde zunächst Atd (für *accessory tropism determinant*), später bAvd (für *accessory variability determinant*) genannt (Alayyoubi et al., 2012; Doulatov et al., 2004; Medhekar and Miller, 2007). BLAST-Analysen belegen eine Homologie einiger akzessorischer Proteine zu den bisher nicht eingehender charakterisierten ribosomalen S23-Proteinen. Im bioinformatischen Teil dieser Arbeit wurde eine systematische Suche nach akzessorischen Proteinen in den ermittelten 155 DGRs durchgeführt. Hierbei wurde eine Gruppe von Proteinen identifiziert, die über ein zuvor unbekanntes Consensusmotiv am 3‘-Ende und eine generell positive Ladung charakterisiert wird. Ein Vertreter dieser Gruppe, Alr3496 aus *Nostoc* sp. PCC 7120, wurde ebenfalls als Teil dieser Arbeit erfolgreich kloniert, aufgereinigt und auf strukturelle sowie biochemische Eigenschaften hin untersucht.

4.3.1 Alr3496 kann rekombinant erzeugt und aufgereinigt werden, und besitzt hervorragende *in vitro*-Stabilität

Im Gegensatz zu den Versuchen, die reverse Transkriptase des DGR-Elements aus *Nostoc* sp. PCC 7120 rekombinant zu erzeugen und zu isolieren, verliefen die Experimente zur Aufreinigung des akzessorischen Proteins erfolgreicher. Es wurden Reinigungsprotokolle erstellt, mit denen hohe Ausbeuten eines reinen Proteins erhalten werden können (bis zu 8 mg Protein pro Liter Kultur). Bei optimalen Lagerbedingungen blieb das Protein über 6 Monate hinweg bei 4 °C stabil. Über das fusionierte Hexahistidintag ist es darüber hinaus möglich, das Protein für einige Anwendungen an einer Ni-NTA-Matrix zu immobilisieren. Denkbar sind beispielsweise Pulldown-Assays mit lysierten *Nostoc* sp. PCC 7120-Bakterien, in denen endogene Bindungspartner an Alr3496 binden und in den nachfolgenden Waschschritten mit aufgereinigt werden; eine Identifizierung der copräzipitierten Proteine könnte über SDS-PAGE und massenspektrometrische Methoden erfolgen. Der wahrscheinlichste Bindungspartner ist die reverse Transkriptase, wie es in Copräzipitationsexperimenten von Alayyoubi et al. gezeigt wurde (Alayyoubi et al., 2012), es ist jedoch nicht ausgeschlossen, hierüber auch bisher unbekannte, wirtseigene Faktoren zu bestimmen, die für die Aktivität des DGRs benötigt werden und Alr3496 binden.

Zeitgleich zu den hier beschriebenen Versuchen wurde von Alayyoubi et al. das homologe Protein bAvd aus dem *Bordetella* Bakteriophagen-DGR aufgereinigt, und seine Struktur bestimmt (Alayyoubi et al., 2012). Im Gegensatz zu Alr3496 gestaltete sich dort die Isolation des Proteins deutlich

schwieriger, da keine native Löslichkeit vorlag und die Aufreinigung aus Inclusion Bodies mit anschließender Rückfaltung des Proteins und zusätzlicher Größenausschlussfiltration erfolgen musste. Über die Langzeitstabilität des Proteins wird keine Aussage gemacht. Aufgrund dieser Ergebnisse kann davon ausgegangen werden, dass Alr3496 aufgrund seiner problemlosen Aufreinigung, seiner hohen Langzeitstabilität und der Stabilität in einer Reihe von Puffern, die in dieser Arbeit für Funktionstests benutzt wurden, gegenüber dem Homolog bAvd deutliche Vorteile und gute Perspektiven für weitere Funktionsanalysen bietet, die sich an diese Arbeit anschließen werden.

4.3.2 Alr3496 nimmt *in vitro* eine Tetramer- oder Pentamerstruktur ein

Für das Protein bAvd aus dem *Bordetella* Bakteriophagen-DGR und die beiden ribosomalen S23-Proteine Xcc0516 und Bt_0352, die keine Assoziation mit einem DGR-Element aufweisen, liegen bereits Strukturen aus kristallographischen Experimenten vor. Ihnen gemeinsam ist eine homopentamere Quartärstruktur, während das Monomer aus einem Vier-Helix-Bündel gebildet wird. Das in dieser Arbeit ermittelte Consensusmotiv findet sich dabei an der Außenseite des Pentamers, wo es möglicherweise mit Substraten oder anderen Faktoren interagieren kann.

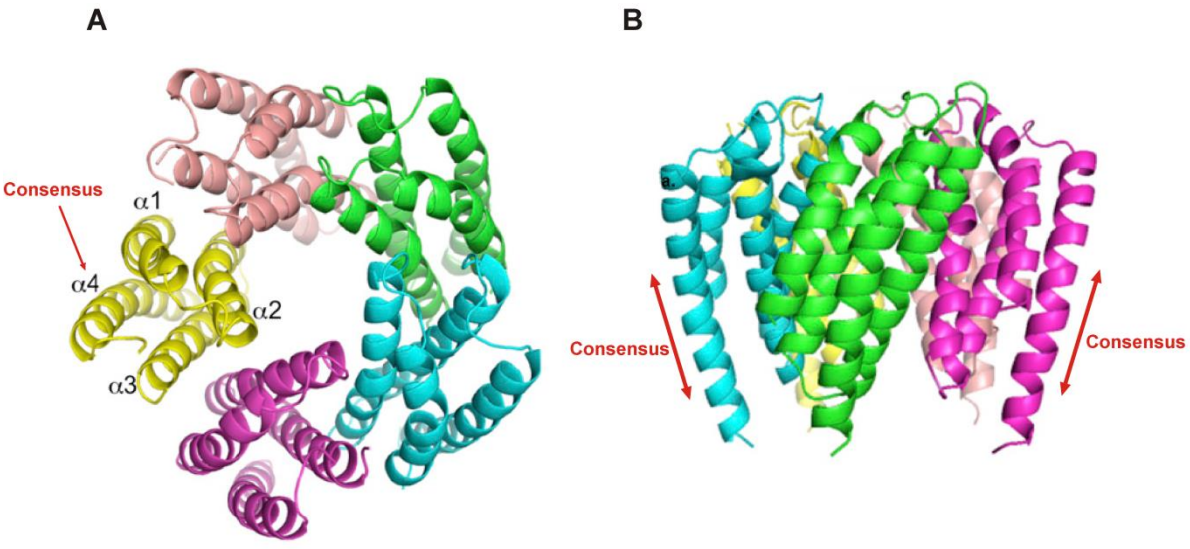


Abbildung 35: Pentamerstruktur des akzessorischen Proteins bAvd. (A) Aufsicht auf das Homopentamer von bAvd. Die verschiedenen Protomere sind unterschiedlich farbig dargestellt. Das Consensusmotiv befindet sich in Helix 4, an der Außenseite des Pentamers (roter Pfeil). (B) Seitenansicht der Struktur aus (A). Auch hier ist die Position des Consensus-Motivs mit roten Pfeilen markiert. Abbildung adaptiert nach Alayyoubi et al., 2012.

Die *in silico*-Strukturvorhersage des Proteins Alr3496 ergab mit zwei verschiedenen Programmen ebenfalls ein Vier-Helix-Bündel; dies stützt die Vermutung, dass dieses Protein, ebenso wie seine Homologe, eine ähnliche Faltung und Oligomerisierung annimmt. Da recht hohe und reine Ausbeuten mit den hier ermittelten Aufreinigungsbedingungen erzielt werden können, sind gute Voraussetzungen für eine Kristallisation und anschließende Strukturaufklärung des Proteins gegeben. Zunächst wurden jedoch andere Techniken eingesetzt, um Informationen über den Oligomerisierungszustand des Proteins zu erhalten. Aus den Gelfiltrationsexperimenten ergab sich, dass Alr3496 wie ein Protein von etwa 57 kDa von einer Sephadex-Säule eluiert wird; dieses Molekulargewicht würde einer Tetramerisierung des Proteins entsprechen, während ein Pentamer ein Molekulargewicht von ca. 70 kDa besitzen und ein deutlich anderes Elutionsverhalten zeigen würde. Weitere Proteinpeaks wurden jedoch nicht detektiert, was bedeutet, dass in der verwendeten Proteinpräparation nur eine Oligomerisierungsform vorliegt. Quervernetzungsexperimente mit DSP erlaubten hingegen keine Aussage über die Stöchiometrie des *in vitro*-Zustands des Proteins, da mit keinem der getesteten DSP/Protein-Verhältnisse eine deutliche Zunahme einer Oligomerisierungsform beobachtet werden konnte (s. Abbildung 26). Eine mögliche Erklärung für dieses Ergebnis könnte in einer ungeeigneten Wahl des Quervernetzers DSP gelegen haben, welcher mit seiner Größe (*spacer arm*) von 12 Å im mittleren Bereich der üblicherweise zur Quervernetzung von Proteinen eingesetzten Agenzien liegt. Jedoch könnten die Protomere enger voneinander als 12 Å entfernt sein, und den Einsatz alternativer Quervernetzer erfordern. Denkbar wären Agenzien mit einem kürzeren Spacer-Arm, wie beispielsweise Glutaraldehyd (5 Å) (Robin et al., 2009) oder 1,5-Difluoro-2,4-dinitrobenzen (3 Å) (Keinan et al., 2010), falls die Untereinheiten des Alr3496-Oligomers nur einen geringen Abstand zueinander haben, und DSP aus sterischen Gründen nicht reagieren konnte. Umgekehrt könnten die sterischen Bedingungen zwischen den Protomeren dergestalt sein, dass eine Quervernetzung erst durch Agenzien wie Bis-Succinimidylpolyethylenglykole (21 bis 35 Å) (Lay et al., 2012) erreicht werden könnte.

Es stellt sich somit die Frage, ob Alr3496 tatsächlich eine tetramere Quartärstruktur aufweist, was das Protein deutlich von bAvd und den nicht-DGR-assoziierten Homologen Xcc0516 und Bt_0352 unterscheiden würde. Ein Alignment von Alr3496 und bAvd zeigt, dass nicht alle Reste, die an den hydrophoben Wechselwirkungen zwischen den Protomeren des *Bordetella* Bakteriophagen-Homologs teilnehmen, konserviert vorliegen (s. Abbildung 36); so finden sich beispielsweise an den Positionen der aliphatischen Reste V40, L50 und I88 jeweils aromatische Aminosäuren in Alr3496. Diese besitzen zwar ebenfalls hydrophoben Charakter, könnten jedoch aufgrund ihrer Größe eine Pentamerisierung verhindern. Denkbar ist, dass statt einer 72°-Anordnung der Protomere zueinander eine 90°-Anordnung angenommen wird, was zu einer Interaktion alternativer Reste und einer Tetramerisierung führen. Für V40, A41, I88 und M77 existieren in Alr3496 keine homologen Reste, so

dass eine Pentamerisierung möglicherweise nicht erfolgen kann, und eine Tetramerisierung bevorzugt wird.

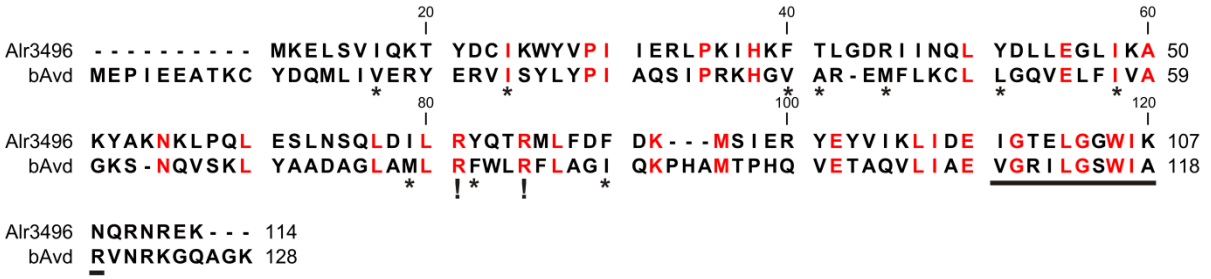


Abbildung 36: Clustal-Alignment von Alr3496 und bAvd. Konservierte Aminosäuren sind in Rot hervorgehoben; Reste, die laut Alayyoubi et al. an der hydrophoben Interaktion zwischen Protomeren teilnehmen, sind mit (*) gekennzeichnet, während Reste, die für das mutagene Homing essentiell zu sein scheinen, mit (!) markiert sind. Das für akzessorische Proteine charakteristische Motiv, das in dieser Arbeit identifiziert wurde, ist unterstrichen.

Ebenso ist allerdings nicht auszuschließen, dass die Abschätzung des Molekulargewichts unter dem verwendeten Laufpuffer nicht ganz korrekt war, und lediglich bei globulären Proteinen wie RNase A eine gute Übereinstimmung mit den zuvor ermittelten Elutionsvolumina des kommerziellen Standards ergibt. Weitere Gelfiltrationen mit einem breiteren Spektrum von Proteinen, die in 20 % Glycerol vorliegen, könnten genauere Werte liefern und eine präzisere Bestimmung des tatsächlichen Molekulargewichts des Alr3496-Oligomers erlauben. Als Alternativmethode zur Gelfiltration könnte der Oligomerisierungszustand über eine analytische Ultrazentrifugation bestimmt werden (Jose et al., 2012; Yamamoto et al., 2008). Eine andere Möglichkeit, die nicht außer Betracht gelassen werden sollte, ist ein unterschiedliches Oligomerisierungsverhalten unter *in vitro*-Bedingungen. Dieses Phänomen wurde bereits andernorts beobachtet (Barranco-Medina et al., 2008; Garvey et al., 2011), und auch hier könnte lediglich die Wahl des Puffers, des Tags oder des Expressionssystems *E.coli* zur Bildung eines Tetramers geführt haben, während unter anderen Bedingungen möglicherweise ein Pentamer beobachtet werden könnte. Umgekehrt muss beachtet werden, dass weder Alayyoubi et al., noch Lin et al. alternative Methoden zur Überprüfung der Struktur eingesetzt haben, und die beobachteten Homopentamere lediglich Kristallisationsartefakte sein könnten (Alayyoubi et al., 2012; Lin et al., 2006).

4.3.3. Das bAvd-Homolog Alr3496 ist ein Nucleinsäurechaperon

Bisher ist die Funktion der akzessorischen Proteine nicht geklärt worden, ebensowenig die ihrer Homologe, der ribosomalen S23-Proteine. Wie Abbildung 2 zu entnehmen ist, bestehen mögliche Funktionen für akzessorische Proteine einerseits in einer Beteiligung am RT-Prozess, beispielsweise in einer Art Adapterprotein, welches TR-Transkripte und die elementeigene reverse Transkriptase zusammenführt. Andererseits könnte das Protein eine Rolle im bisher weitgehend unverstandenen Mutagenic Homing-Vorgang spielen. Für diesen Prozess haben Bob Medhekar und Jeff F. Miller vier verschiedene Modelle vorgeschlagen, die in Abbildung 37 skizziert sind (Medhekar and Miller, 2007).

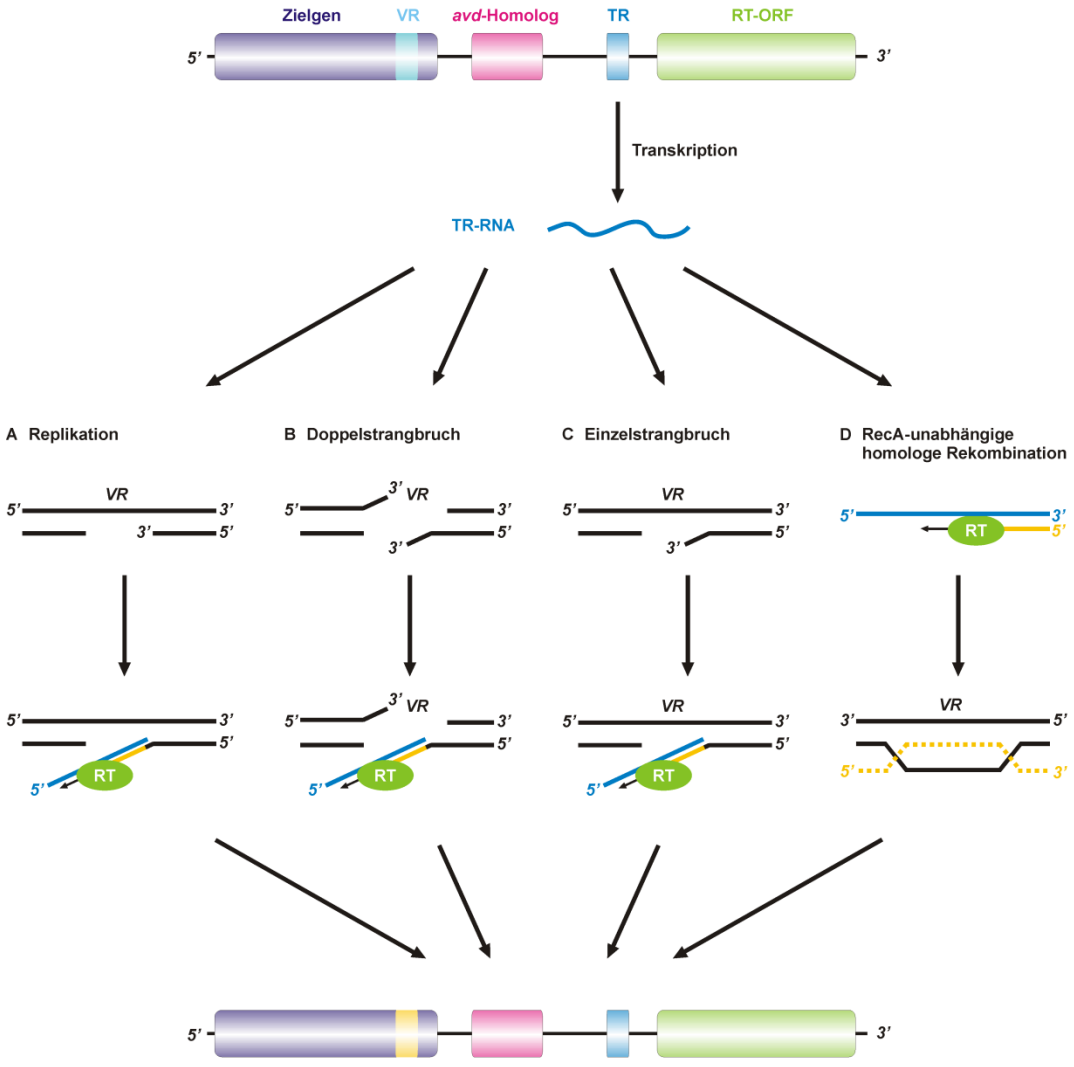


Abbildung 37: Mögliche Mechanismen des Mutagenic Homing. Ein TR-Transkript wird in A bis C über einen Target Primed Reverse Transcription (TPRT)-Prozess in cDNA umgeschrieben. Als Primer dienen jeweils freie 3'OH-Enden von (A) Okazakifragmenten während der Replikation oder von (B) Doppelstrangbrüchen bzw. (C) Einzelstrangbrüchen, die jeweils durch noch unbekannte Faktoren generiert werden. Modell D geht hingegen von einem RT-Prozess aus, der unabhängig vom Homing-Prozess ist, und durch einen unbekannten Primer initiiert wird. Der Einbau erfolgt anschließend durch homologe Rekombination der cDNA mit dem Antisense-Strang der VR, wobei eine Beteiligung von RecA bereits ausgeschlossen werden konnte (Guo et al., 2008). Abbildung adaptiert von Medhekar und Miller, 2007

Das erste Modell nimmt an, dass DGR-Aktivität während der Replikation vorliegt. Okazakifragmente, auf dem Folgestrang dienen hierbei als Primer einer reversen Transkription, weswegen dieser Prozess auch Target Primed Reverse Transcription (TPRT) genannt wird. Anschließend wird der generierte Antisense-Strang durch zelleigene Faktoren mit den Downstreamfragmenten ligiert, und der Sense-Strang durch eine komplementäre Variante ersetzt. Dieser Prozess würde somit analog zum Insertionsmechanismus des Gruppe II-Introns RmtInt1 aus *Sinorhizobium meliloti* verlaufen (Martinez-Abarca et al., 2004). In zwei weiteren Modellen werden freie 3‘-OH-Enden, die aus Doppel- bzw. Einzelstrangbrüchen durch bisher unidentifizierte Faktoren wie Nucleasen oder Stress resultieren, wiederum in einer TPRT-Reaktion als Primer für die cDNA-Synthese des Antisense-Strangs verwendet (Christensen et al., 2006; Gasior et al., 2006). Das letzte Modell, das Medhekar und Miller postuliert haben, trennt RT- und Insertionsreaktion voneinander (Storici et al., 2007). Die neu synthetisierte cDNA ersetzt in einem homologen Rekombinationsprozess den Antisense-Strang der alten VR, während ein komplementärer Gegenstrang auch hier über zelleigene DNA-Reparaturmechanismen synthetisiert wird. In jedem dieser Modelle könnten akzessorische Proteine eine essentielle Funktion besitzen. In einer kürzlich veröffentlichten Publikation konnte gezeigt werden, dass bAvd *in vitro* an diverse Nucleinsäurespezies und an die reverse Transkriptase desselben Elements bindet (Alayyoubi et al., 2012). Welchen Hintergrund diese Bindung hat, oder ob sie nur ein Versuchsartefakt ist, konnte nicht beantwortet werden. Über die *in vivo*-Funktion des bAvd-Proteins ist lediglich bekannt, dass sie essentiell ist für die Funktion des DGRs (Doulatov et al., 2004); zusätzlich konnte beobachtet werden, dass offenbar auch die codierende Sequenz des bAvd-Proteins eine wichtige Funktion besitzt, da 5‘-Deletionen des Leserahmens bei gleichzeitiger Supplementierung des Proteins *in trans* eine verminderte Aktivität des DGR-Elements bewirkten (Alayyoubi et al., 2012). In dieser Arbeit konnte ebenfalls gezeigt werden, dass Alr3496 sowohl RNA als auch DNA und RNA:DNA-Hybride binden kann (s. Abbildung 30). Die Affinität für die wahrscheinliche TR-RNA und eine fremde RNA war vergleichbar hoch, während zu RNA:DNA-Hybriden und DNA offenbar eine höhere Affinität besteht, die in weiteren Versuchen näher untersucht werden muss. Zudem konnte reproduzierbar gezeigt werden, dass Alr3496 in der Lage ist, die Bildung von Duplex-DNA zu katalysieren und somit als Nucleinsäurechaperon zu agieren (s. Abbildung 34); ein Aufwinden von Duplex-DNA in Einzelstränge wurde hingegen nicht beobachtet (s. Abbildung 33). Dies steht auch im Einklang mit den Ergebnissen aus Experimenten zum Nachweis einer ATPase-Aktivität, die ebenfalls negativ verlaufen sind (s. Abschnitt 3.3.4.2). Es bleibt jedoch nicht auszuschließen, dass eine sequenzspezifische Helicaseaktivität von Alr3496 vorliegt, und weitere mögliche native Bindungspartner getestet werden sollten. Ein Problem stellen hierbei die relativ hohen Eigenbeiträge der ATP- und Proteinlösungen dar, die für die Malachitgrünassays eingesetzt werden. Künftige Experimente zum Nachweis einer ATPase-Aktivität könnten über

Methoden erfolgen, welche nicht unmittelbar von der Menge des vorhandenen Phosphats abhängen. Eine solche Methode beruht auf der Ausnutzung einer mehrstufigen metabolischen Reaktion, in deren Verlauf NADH zu NAD+ oxidiert wird. Die Rate, mit der diese Oxidation verläuft, kann über die Abnahme des Absorptionssignals bei 340 nm bestimmt werden, und ist direkt proportional zur ATPase-Aktivität des betrachteten Proteins (Norby, 1971).

Mit dem hier beobachteten enzymkatalysierten Hybridisieren zweier Stränge liegt der erste Hinweis auf die Funktion der bAvd-Homologe innerhalb des DGR-Mechanismus vor. Alr3496 und homologe Proteine könnten somit als Nucleinsäurechaperone an der postulierten TPRT-Reaktion beteiligt sein und ein Annealing von TR-RNA VR-DNA katalysieren. Da außerdem eine Bindung an die reverse Transkriptase des DGR-Elements für bAvd gezeigt werden konnte (Alayyoubi et al., 2012), ist es denkbar, dass das Protein eine zweite Funktion als Lotse besitzt, welcher RT und Template:Primer-Komplex in räumliche Nähe zueinander bringt, und somit eine Voraussetzung für eine cDNA-Synthese schafft. Es ist sogar möglich, dass es sich um ein Adapterprotein handelt, das durch die Bindung der RT eine spezifische reverse Transkription der TR-RNA einleitet, während andere, wirtseigene RNA-Ziele ignoriert werden. Die Dissoziationskonstanten des Proteins mit verschiedenen Nucleinsäuresubstraten betrugen zwischen 100 und 400 nM, wobei niedrigere Dissoziationskonstanten in den hier durchgeführten Versuchen teilweise nicht gut erkennbar waren und noch einmal separat bestimmt werden sollten. Diese Konstanten reihen sich im Mittelfeld der Affinitätskonstanten für andere nucleinsäurebindende Proteine ein, die in Tabelle 23 gegeben sind.

Tabelle 23: Affinitätskonstanten einiger nucleinsäurebindender Proteine

Protein	*Spezies*	*Art*	*K_D [nM]*	*Referenz*
Sp1	*H. sapiens*	Transkriptionsfaktor	0,41 - 0,53	(Letovsky and Dynan, 1989)
Egr-1	*H. sapiens*	Transkriptionsfaktor	0,01 - 0,014	(Nalefski et al., 2006)
hGABPa	*H. sapiens*	Transkriptionsfaktor	1,5	(Suzuki et al., 1998)
UvrD	*E. coli*	Helicase	1400	(Ratcliff and Erie, 2001)
RecQ	*E. coli*	Helicase	3,8 - 21,6	(Zhang et al., 2006)
NS3	Hepatitis C-Virus	Helicase	2-4	(Levin and Patel, 2002)
TraI36	Plasmid	Helicase/Relaxasedomäne	3300	(Harley and Schildbach, 2003)
p7	Humanes Immundefizienz-Virus 1	Nucleocapsidprotein/ Nucleinsäurechaperon	123 - 233	(Stewart-Maynard et al., 2008)
p12	Rous Sarcoma-Virus	Nucleocapsidprotein/ Nucleinsäurechaperon	200 - 660	(Stewart-Maynard et al., 2008)
ORF1p	Non-LTR Retrotransposon L1	Nucleinsäurechaperon	0,7 – 4,6	(Kolosha and Martin, 2003)

Die hier beobachteten Bindungen sind somit weder besonders stark, im Vergleich zu beispielsweise den Transkriptionsfaktoren in Tabelle 21, noch besonders schwach, da andere Proteine, wie die Relaxasedomäne des TraI-Proteins oder die UvrD-Helicase jeweils um den Faktor 10 schwächer binden. Hinzugefügt werden sollte, dass die schwächeren Konstanten in Tabelle 23 in der Regel aus Experimenten stammen, in denen generell die Bindung des Proteins an Nucleinsäuren, und nicht unbedingt seines nativen Ziels, getestet wurde. Analog hierzu könnten mit den nativen Bindungspartnern von Alr3496 deutlich höhere Affinitäten festgestellt werden. Da die Affinität des Proteins zur wahrscheinlichen TR-RNA vergleichbar war wie die zu einer Fremd-RNA, und die Affinitäten niedriger als z. B. für DNA:RNA-Hybride, scheint die TR-RNA allein keine Erkennungssequenz zu beinhalten, die zu einer verstärkten Bindung des Proteins führt. In weiteren Experimenten sollten kontextfremde DNA:RNA-Hybride untersucht werden, um zu ermitteln, ob möglicherweise ein doppelsträngiges Motiv in den hier durchgeführten Versuchen erkannt wurde, und daher eine niedrigere Dissoziationskonstante bestimmt wurde, oder ob dieses Verhalten generell bei Heteroduplices beobachtet werden kann. Nicht geklärt in diesem Zusammenhang ist die Frage nach einer Beteiligung des Proteins an der Diskriminierung von VR und TR. In dieser Arbeit wurden RNA:DNA-Hybride untersucht, welche entweder der Bindung der TR-RNA an die homologe Sequenz der TR-DNA entspricht, oder der Bindung an die entsprechende VR-DNA; beide Substrate wurden von Alr3496 mit gleicher Affinität gebunden. Somit scheint unter den hier getesteten Bedingungen Alr3496 nicht am Erkennungsprozess beteiligt zu sein, der einen Einbau der mutagenisierten TR-RNA am TR-Locus verhindert. Möglich ist, dass hierfür eine komplexere Interaktion der Stem-Loop-Motive der TR-RNA oder der VR-DNA mit dem Protein notwendig ist, und eine IMH-Region allein nicht ausreicht; dies würde jedoch den Ergebnissen von Doulatov et al. widersprechen, denen zufolge ein simpler Austausch der TR-IMH* gegen die VR-IMH ausreicht, um auch den TR-Locus für mutagenes Homing empfänglich zu machen (Doulatov et al., 2004).

4.4 Ausblick

Diversitätsgenerierende Retroelemente weisen selbst eine Dekade nach ihrer erstmaligen Beschreibung noch eine Reihe ungeklärter Fragen auf hinsichtlich ihrer Ursprünge, ihrer Funktion und der zugrundeliegenden molekularen Mechanismen. Diese Arbeit leistet einen Beitrag dazu, das Bild dieser Elemente vollständiger zu definieren, als es bisherige Studien vermocht haben. Zukünftige Analysen neu hinzukommender Elemente sollten mit dem Rahmenwerk, das mit dieser Arbeit geschaffen wird, deutlich einfacher, schneller und effizienter zu bewerkstelligen sein, und damit weitere wertvolle Beiträge zum tieferen Verständnis dieser Elemente liefern.

Eine Herausforderung stellt nach wie vor die Aufreinigung und detaillierte biochemische Beschreibung einer reversen Transkriptase aus einem DGR-Element dar, da selbst Studien, die die erfolgreiche Isolation rekombinanten Proteins beschrieben, eine Vielzahl von Fragen zu Struktur und Mechanismus dieser Enzyme unbeachtet ließen (Alayyoubi et al., 2012; Liu et al., 2002). In der vorliegenden Arbeit wurden zahlreiche Voraussetzungen und Ansatzpunkte geschaffen, die in kommenden Experimenten zur Lösung dieser Fragen beitragen können, beispielsweise in Form einer alternativen Aufreinigungsstrategie oder durch die Erzeugung von Chimärenproteinen.

Des Weiteren konnten in dieser Arbeit erste experimentell gestützte Hinweise auf die Funktion der bisher nur wenig beschriebenen akzessorischen Proteine gegeben werden. Als nächstes wird die nähere Charakterisierung der beobachteten Annealingreaktion hinsichtlich Kinetik, Sequenzspezifität und zeitlicher Reihenfolge der Bindung an Nucleinsäure und reverse Transkriptase erfolgen. Eine weitere Versuchsreihe wird sich mit dem in dieser Arbeit erstmalig beschriebenen Consensusmotiv auseinandersetzen; wie Alayyoubi et al. in ihrer Publikation bereits anmerkten (Alayyoubi et al., 2012), ist die zentrale Pore, die durch die fünf Protomere gebildet wird, deutlich zu klein, um mehr als einen Einzelstrang zu beherbergen. Es ist daher wahrscheinlicher, dass die katalytisch aktiven Reste auf der Außenseite des Pentamers zu finden sind, beispielsweise im Consensusmotiv. In den nächsten Schritten sollte daher eine Mutagenese einiger Reste, z. B. der basischen am C-terminalen Ende des Motivs durchgeführt werden, sowie eine Analyse, wie sich diese Veränderung auf das Bindungsverhalten von Nucleinsäuresubstraten auswirkt.

V. Literaturverzeichnis

Abramovitz, D.L., and Pyle, A.M. (1997). Remarkable morphological variability of a common RNA folding motif: the GNRA tetraloop-receptor interaction. *J Mol Biol 266*, 493-506.

Afseth, G., Mo, Y.Y., and Mallavia, L.P. (1995). Characterization of the 23S and 5S rRNA genes of Coxiella burnetii and identification of an intervening sequence within the 23S rRNA gene. *J Bacteriol 177*, 2946-2949.

Akerley, B.J., Cotter, P.A., and Miller, J.F. (1995). Ectopic expression of the flagellar regulon alters development of the Bordetella-host interaction. *Cell 80*, 611-620.

Alayyoubi, M., Guo, H., Dey, S., Golnazarian, T., Brooks, G.a., Rong, A., Miller, J.F., and Ghosh, P. (2012). Structure of the Essential Diversity-Generating Retroelement Protein bAvd and Its Functionally Important Interaction with Reverse Transcriptase. *Structure*, 1-11.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol 215*, 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res 25*, 3389-3402.

Avery, O.T., Macleod, C.M., and McCarty, M. (1944). Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. *J Exp Med 79*, 137-158.

Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature 226*, 1209-1211.

Barranco-Medina, S., Krell, T., Bernier-Villamor, L., Sevilla, F., Lazaro, J.J., and Dietz, K.J. (2008). Hexameric oligomerization of mitochondrial peroxiredoxin PrxIIF and formation of an ultrahigh affinity complex with its electron donor thioredoxin Trx-o. *J Exp Bot 59*, 3259-3269.

Benachenhou, F., Blikstad, V., and Blomberg, J. (2009). The phylogeny of orthoretroviral long terminal repeats (LTRs). *Gene 448*, 134-138.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Res 28*, 235-242.

Bernstein, D.A., and Keck, J.L. (2005). Conferring substrate specificity to DNA helicases: role of the RecQ HRDC domain. *Structure 13*, 1173-1182.

Birnboim, H.C., and Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res 7*, 1513-1523.

Bose, M., and Barber, R.D. (2006). Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biol 6*, 223-227.

Brouns, S.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., Dickman, M.J., Makarova, K.S., Koonin, E.V., and van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science 321*, 960-964.

Cadwell, R.C., and Joyce, G.F. (1992). Randomization of genes by PCR mutagenesis. *PCR Methods Appl 2*, 28-33.

Canchaya, C., Fournous, G., and Brussow, H. (2004). The impact of prophages on bacterial chromosomes. *Mol Microbiol 53*, 9-18.

Carte, J., Wang, R., Li, H., Terns, R.M., and Terns, M.P. (2008). Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev 22*, 3489-3496.

Carter, S.G., and Karl, D.W. (1982). Inorganic phosphate assay with malachite green: an improvement and evaluation. *J Biochem Biophys Methods 7*, 7-13.

Chen, Z., Xu, S., Zhou, K., and Yang, G. (2011). Whale phylogeny and rapid radiation events revealed using novel retroposed elements and their flanking sequences. *BMC Evol Biol 11*, 314.

Chenais, B., Caruso, A., Hiard, S., and Casse, N. (2012). The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene 509*, 7-15.

Chothia, C., and Lesk, A.M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol 196*, 901-917.

Christensen, S.M., Ye, J., and Eickbush, T.H. (2006). RNA from the 5' end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site. *Proc Natl Acad Sci U S A 103*, 17602-17607.

Chueh, A.C., Northrop, E.L., Brettingham-Moore, K.H., Choo, K.H., and Wong, L.H. (2009). LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. *PLoS Genet 5*, e1000354.

Cosma, M.P., Pepe, S., Annunziata, I., Newbold, R.F., Grompe, M., Parenti, G., and Ballabio, A. (2003). The multiple sulfatase deficiency gene encodes an essential and limiting factor for the activity of sulfatases. *Cell 113*, 445-456.

Costa, M., and Michel, F. (1995). Frequent use of the same tertiary motif by self-folding RNAs. *EMBO J 14*, 1276-1285.

Cotter, P.A., and DiRita, V.J. (2000). Bacterial virulence gene regulation: an evolutionary perspective. *Annu Rev Microbiol 54*, 519-565.

Cotter, P.A., and Miller, J.F. (1997). A mutation in the Bordetella bronchiseptica bvgS gene results in reduced virulence and increased resistance to starvation, and identifies a new class of Bvg-regulated antigens. *Mol Microbiol 24*, 671-685.

Creighton, H.B., and McClintock, B. (1931). A Correlation of Cytological and Genetical Crossing-Over in Zea Mays. *Proc Natl Acad Sci U S A 17*, 492-497.

D'Aquila, R.T., and Summers, W.C. (1989). HIV-1 reverse transcriptase/ribonuclease H: high level expression in Escherichia coli from a plasmid constructed using the polymerase chain reaction. *J Acquir Immune Defic Syndr 2*, 579-587.

Dai, L., and Zimmerly, S. (2002). Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Research 30*, 1091-1102.

Das, D., and Georgiadis, M.M. (2001). A directed approach to improving the solubility of Moloney murine leukemia virus reverse transcriptase. *Protein Sci 10*, 1936-1941.

Das, D., and Georgiadis, M.M. (2004). The crystal structure of the monomeric reverse transcriptase from Moloney murine leukemia virus. *Structure 12*, 819-829.

Davis, G.D., Elisee, C., Newham, D.M., and Harrison, R.G. (1999). New fusion protein systems designed to give soluble expression in Escherichia coli. *Biotechnol Bioeng 65*, 382-388.

Davis, M.M., and Bjorkman, P.J. (1988). T-cell antigen receptor genes and T-cell recognition. *Nature 334*, 395-402.

Delihas, N. (2011). Impact of small repeat sequences on bacterial genome evolution. Genome Biol Evol *3*, 959-973.

Deora, R., Bootsma, H.J., Miller, J.F., and Cotter, P.A. (2001). Diversity in the Bordetella virulence regulon: transcriptional control of a Bvg-intermediate phase gene. *Mol Microbiol 40*, 669-683.

Dierks, T., Dickmanns, A., Preusser-Kunze, A., Schmidt, B., Mariappan, M., von Figura, K., Ficner, R., and Rudolph, M.G. (2005). Molecular basis for multiple sulfatase deficiency and mechanism for formylglycine generation of the human formylglycine-generating enzyme. *Cell 121*, 541-552.

Dierks, T., Lecca, M.R., Schlotterhose, P., Schmidt, B., and von Figura, K. (1999). Sequence determinants directing conversion of cysteine to formylglycine in eukaryotic sulfatases. *EMBO J 18*, 2084-2091.

Dierks, T., Schmidt, B., Borissenko, L.V., Peng, J., Preusser, A., Mariappan, M., and von Figura, K. (2003). Multiple sulfatase deficiency is caused by mutations in the gene encoding the human C(alpha)-formylglycine generating enzyme. *Cell 113*, 435-444.

Doolittle, W.F., and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature 284*, 601-603.

Doulatov, S., Hodes, A., Dai, L., and Mandhana, N. (2004). Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. *Nature 431*.

Elena, S.F., and Lenski, R.E. (2003). Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat Rev Genet 4*, 457-469.

Feng, Q., Moran, J.V., Kazazian, H.H., Jr., and Boeke, J.D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell 87*, 905-916.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat Rev Genet 9*, 397-405.

Fisher, A.C., Kim, W., and Delisa, M.P. (2006). Genetic selection for protein solubility enabled by the folding quality control feature of the twin-arginine translocation pathway. *Protein Sci 15*, 449-458.

Flexner, C., Broyles, S.S., Earl, P., Chakrabarti, S., and Moss, B. (1988). Characterization of human immunodeficiency virus gag/pol gene products expressed by recombinant vaccinia viruses. *Virology 166*, 339-349.

Fux, C.A., Shirtliff, M., Stoodley, P., and Costerton, J.W. (2005). Can laboratory reference strains mirror "real-world" pathogenesis? *Trends Microbiol 13*, 58-63.

Gallo, R.C., Salahuddin, S.Z., Popovic, M., Shearer, G.M., Kaplan, M., Haynes, B.F., Palker, T.J., Redfield, R., Oleske, J., Safai, B.*, et al.* (1984). Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science 224*, 500-503.

Garvey, M., Tepper, K., Haupt, C., Knupfer, U., Klement, K., Meinhardt, J., Horn, U., Balbach, J., and Fandrich, M. (2011). Phosphate and HEPES buffers potently affect the fibrillation and oligomerization mechanism of Alzheimer's Abeta peptide. *Biochem Biophys Res Commun 409*, 385-388.

Gasior, S.L., Wakeman, T.P., Xu, B., and Deininger, P.L. (2006). The human LINE-1 retrotransposon creates DNA double-strand breaks. *J Mol Biol 357*, 1383-1393.

Georgiadis, M.M. (2001). A directed approach to improving the solubility of Moloney murine leukemia virus reverse transcriptase. *Protein Sci 10*, 1936-1941.

Georgiadis, M.M., Jessen, S.M., Ogata, C.M., Telesnitsky, a., Goff, S.P., and Hendrickson, W.a. (1995). Mechanistic implications from the structure of a catalytic fragment of Moloney murine leukemia virus reverse transcriptase. *Structure 3*, 879-892.

Gieselmann, V., Polten, A., Kreysing, J., Kappler, J., Fluharty, A., and von Figura, K. (1991). Molecular genetics of metachromatic leukodystrophy. *Dev Neurosci 13*, 222-227.

Gieselmann, V., Zlotogora, J., Harris, A., Wenger, D.A., and Morris, C.P. (1994). Molecular genetics of metachromatic leukodystrophy. *Hum Mutat 4*, 233-242.

Gillis, A.J., Schuller, A.P., and Skordalakes, E. (2008). Structure of the Tribolium castaneum telomerase catalytic subunit TERT. *Nature 455*, 633-637.

Guo, H., Tse, L.V., Barbalat, R., Sivaamnuaiphorn, S., Xu, M., Doulatov, S., and Miller, J.F. (2008). Diversity-generating retroelement homing regenerates target sequences for repeated rounds of codon rewriting and protein diversification. *Molecular Cell 31*, 813-823.

Guo, H., Tse, L.V., Nieh, A.W., Czornyj, E., Williams, S., Oukil, S., Liu, V.B., and Miller, J.F. (2011). Target site recognition by a diversity-generating retroelement. *PLoS Genetics 7*, e1002414.

Harley, M.J., and Schildbach, J.F. (2003). Swapping single-stranded DNA sequence specificities of relaxases from conjugative plasmids F and R100. *Proc Natl Acad Sci U S A 100*, 11243-11248.

Hebert, E. (2000). Endogenous lectins as cell surface transducers. *Biosci Rep 20*, 213-237.

Hennig, W. (1966). Phylogenetic systematics (Urbana, University of Illinois Press).

Hershey, A.D., and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol 36*, 39-56.

Hizi, A., McGill, C., and Hughes, S.H. (1988). Expression of soluble, enzymatically active, human immunodeficiency virus reverse transcriptase in Escherichia coli and analysis of mutants. *Proc Natl Acad Sci U S A 85*, 1218-1222.

Huang, H., Chopra, R., Verdine, G.L., and Harrison, S.C. (1998). Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science 282*, 1669-1675.

Isaka, Y., Sato, A., Kawauchi, S., Suyama, A., Miki, S., Hayami, M., and Fujiwara, T. (1998). Construction of the chimeric reverse transcriptase of simian immunodeficiency virus sensitive to nonnucleoside reverse transcriptase inhibitor. *Microbiol Immunol 42*, 195-202.

Jakubczak, J.L., Xiong, Y., and Eickbush, T.H. (1990). Type I (R1) and type II (R2) ribosomal DNA insertions of Drosophila melanogaster are retrotransposable elements closely related to those of Bombyx mori. *J Mol Biol 212*, 37-52.

Jansen, R., Embden, J.D., Gaastra, W., and Schouls, L.M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol 43*, 1565-1575.

Jiang, S., Li, C., Zhang, W., Cai, Y., Yang, Y., Yang, S., and Jiang, W. (2007). Directed evolution and structural analysis of N-carbamoyl-D-amino acid amidohydrolase provide insights into recombinant protein solubility in Escherichia coli. *The Biochemical journal 402*, 429-437.

Jose, D., Weitzel, S.E., Jing, D., and von Hippel, P.H. (2012). Assembly and subunit stoichiometry of the functional helicase-primase (primosome) complex of bacteriophage T4. *Proc Natl Acad Sci U S A 109*, 13596-13601.

Kanda, T., and Saigo, K. (1993). Chimeric reverse transcriptase of Moloney murine leukaemia virus, having the YXDD box of a Drosophila retrotransposon, 17.6. *Biochim Biophys Acta 1163*, 223-226.

Kazazian, H.H., Jr. (1998). Mobile elements and disease. *Curr Opin Genet Dev 8*, 343-350.

Keinan, N., Tyomkin, D., and Shoshan-Barmatz, V. (2010). Oligomerization of the mitochondrial protein voltage-dependent anion channel is coupled to the induction of apoptosis. *Mol Cell Biol 30*, 5698-5709.

Killoran, M.P., and Keck, J.L. (2008). Structure and function of the regulatory C-terminal HRDC domain from Deinococcus radiodurans RecQ. *Nucleic Acids Res 36*, 3139-3149.

Kim, Y.M., and Choi, B.S. (2010). Structure and function of the regulatory HRDC domain from human Bloom syndrome protein. *Nucleic Acids Res 38*, 7764-7777.

Klarmann, G.J., Smith, R.A., Schinazi, R.F., North, T.W., and Preston, B.D. (2000). Site-specific incorporation of nucleoside analogs by HIV-1 reverse transcriptase and the template grip mutant P157S. Template interactions influence substrate recognition at the polymerase active site. *J Biol Chem 275*, 359-366.

Klassen, J.L., and Currie, C.R. (2012). Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC Genomics 13*, 14.

Kohlstaedt, L.A., Wang, J., Friedman, J.M., Rice, P.A., and Steitz, T.A. (1992). Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science 256*, 1783-1790.

Kolosha, V.O., and Martin, S.L. (2003). High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1). *J Biol Chem 278*, 8112-8117.

Koschorreck, M., Fischer, M., Barth, S., and Pleiss, J. (2005). How to find soluble proteins: a comprehensive analysis of alpha/beta hydrolases for recombinant expression in E. coli. *BMC Genomics 6*, 49.

Kuchta, R.D. (1996). Isotopic assays of viral polymerases and related proteins. *Methods Enzymol 275*, 241-257.

Kuehnen, P., Mischke, M., Wiegand, S., Sers, C., Horsthemke, B., Lau, S., Keil, T., Lee, Y.-A., Grueters, A., and Krude, H. (2012). An Alu element-associated hypermethylation variant of the POMC gene is associated with childhood obesity. *PLoS Genetics 8*, e1002543.

Laemmli, U.K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature 227*, 680-685.

Larder, B., Purifoy, D., Powell, K., and Darby, G. (1987). AIDS virus reverse transcriptase defined by high level expression in Escherichia coli. *EMBO J 6*, 3133-3137.

Lay, F.T., Mills, G.D., Poon, I.K., Cowieson, N.P., Kirby, N., Baxter, A.A., van der Weerden, N.L., Dogovski, C., Perugini, M.A., Anderson, M.A.*, et al.* (2012). Dimerization of plant defensin NaD1 enhances its antifungal activity. *J Biol Chem 287*, 19961-19972.

Le Coq, J., and Ghosh, P. (2011). Conservation of the C-type lectin fold for massive sequence variation in a Treponema diversity-generating retroelement. *Proc Natl Acad Sci U S A 108*, 14649-14653.

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L.J., Lohr, J.G., Harris, C.C., Ding, L., Wilson, R.K.*, et al.* (2012). Landscape of somatic retrotransposition in human cancers. *Science 337*, 967-971.

Leininger, E., Roberts, M., Kenimer, J.G., Charles, I.G., Fairweather, N., Novotny, P., and Brennan, M.J. (1991). Pertactin, an Arg-Gly-Asp-containing Bordetella pertussis surface protein that promotes adherence of mammalian cells. *Proc Natl Acad Sci U S A 88*, 345-349.

Leplae, R., Lima-Mendez, G., and Toussaint, A. (2010). ACLAME: a CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Res*earch *38*, D57-61.

Letovsky, J., and Dynan, W.S. (1989). Measurement of the binding of transcription factor Sp1 to a single GC box recognition sequence. *Nucleic Acids Res 17*, 2639-2653.

Levin, M.K., and Patel, S.S. (2002). Helicase from hepatitis C virus, energetics of DNA binding. *J Biol Chem 277*, 29377-29385.

Lin, L.-y., Ching, C.-l., Chin, K.-h., Chou, S.-h., and Chan, N.-l. (2006). Crystal Structure of the Conserved Hypothetical Cytosolic Protein Xcc0516 From Xanthomonas campestris Reveals a Novel Quaternary Structure Assembled by Five Four-Helix Bundles. *Proteins 65*, 783-786.

Liu, M., Deora, R., Doulatov, S.R., Gingery, M., Eiserling, F.a., Preston, A., Maskell, D.J., Simons, R.W., Cotter, P.a., Parkhill, J.*, et al.* (2002). Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. *Science 295*, 2091-2094.

Liu, Z., Macias, M.J., Bottomley, M.J., Stier, G., Linge, J.P., Nilges, M., Bork, P., and Sattler, M. (1999). The three-dimensional structure of the HRDC domain and implications for the Werner and Bloom syndrome proteins. *Structure 7*, 1557-1566.

Martin, S., and Bushman, F. (2001). Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. Molecular and Cellular Biology *21*, 467-475.

Martinez-Abarca, F., Barrientos-Duran, A., Fernandez-Lopez, M., and Toro, N. (2004). The RmInt1 group II intron has two different retrohoming pathways for mobility using predominantly the nascent lagging strand at DNA replication forks for priming. *Nucleic Acids Res 32*, 2880-2888.

Matsuura, M., Saldanha, R., Ma, H., Wank, H., Yang, J., Mohr, G., Cavanagh, S., Dunny, G.M., Belfort, M., and Lambowitz, a.M. (1997). A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes & Development 11*, 2910-2924.

McClintock, B. (1956). Controlling elements and the gene. Cold Spring Harb Symp Quant Biol *21*, 197-216.

McMahon, S.a., Miller, J.L., Lawton, J.a., Kerkow, D.E., Hodes, A., Marti-Renom, M.a., Doulatov, S., Narayanan, E., Sali, A., Miller, J.F.*, et al.* (2005). The C-type lectin fold as an evolutionary solution for massive sequence variation. Nature structural & molecular biology *12*, 886-892.

Medhekar, B., and Miller, J.F. (2007). Diversity-generating retroelements. *Current opinion in microbiology 10*, 388-395.

Meyerson, M., Counter, C.M., Eaton, E.N., Ellisen, L.W., Steiner, P., Caddle, S.D., Ziaugra, L., Beijersbergen, R.L., Davidoff, M.J., Liu, Q.*, et al.* (1997). hEST2, the putative human telomerase catalytic subunit gene, is up-regulated in tumor cells and during immortalization. *Cell 90*, 785-795.

Minot, S., Grunberg, S., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2012). Hypervariable loci in the human gut virome. *Proc Natl Acad Sci U S A 109*, 3962-3966

Najmudin, S., Cote, M.L., Sun, D., Yohannan, S., Montano, S.P., Gu, J., and Georgiadis, M.M. (2000). Crystal structures of an N-terminal fragment from Moloney murine leukemia virus reverse transcriptase complexed with nucleic acid: functional implications for template-primer binding to the fingers domain. *J Mol Biol 296*, 613-632.

Nakamura, T.M., Morin, G.B., Chapman, K.B., Weinrich, S.L., Andrews, W.H., Lingner, J., Harley, C.B., and Cech, T.R. (1997). Telomerase catalytic subunit homologs from fission yeast and human. *Science 277*, 955-959.

Nalefski, E.A., Nebelitsky, E., Lloyd, J.A., and Gullans, S.R. (2006). Single-molecule detection of transcription factor binding to DNA in real time: specificity, equilibrium, and kinetic parameters. *Biochemistry 45*, 13794-13806.

Nardi, F., Spinsanti, G., Boore, J.L., Carapelli, A., Dallai, R., and Frati, F. (2003). Hexapod origins: monophyletic or paraphyletic? *Science 299*, 1887-1889.

Nirenberg, M.W., and Matthaei, J.H. (1961). The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci U S A 47*, 1588-1602.
Norby, J.G. (1971). Studies on a coupled enzyme assay for rate measurements of ATPase reactions. *Acta Chem Scand 25*, 2717-2726.

Orgel, L.E., and Crick, F.H. (1980). Selfish DNA: the ultimate parasite. *Nature 284*, 604-607.

Ortiz, M.L., Calero, M., Fernandez Patron, C., Patron, C.F., Castellanos, L., and Mendez, E. (1992). Imidazole-SDS-Zn reverse staining of proteins in gels containing or not SDS and microsequence of individual unmodified electroblotted proteins. *FEBS letters 296*, 300-304.

Palomares, L.A., Estrada-Mondaca, S., and Ramirez, O.T. (2004). Production of recombinant proteins: challenges and solutions. *Methods Mol Biol 267*, 15-52.

Plasterk, R.H., Izsvak, Z., and Ivics, Z. (1999). Resident aliens: the Tc1/mariner superfamily of transposable elements. *Trends Genet 15*, 326-332.

Poiesz, B.J., Ruscetti, F.W., Reitz, M.S., Kalyanaraman, V.S., and Gallo, R.C. (1981). Isolation of a new type C retrovirus (HTLV) in primary uncultured cells of a patient with Sezary T-cell leukaemia. *Nature 294*, 268-271.

Polten, A., Fluharty, A.L., Fluharty, C.B., Kappler, J., von Figura, K., and Gieselmann, V. (1991). Molecular basis of different forms of metachromatic leukodystrophy. *N Engl J Med 324*, 18-22.

Preston, B.D., Poiesz, B.J., and Loeb, L.A. (1988). Fidelity of HIV-1 reverse transcriptase. *Science 242*, 1168-1171.

Ralph, D., and McClelland, M. (1993). Intervening sequence with conserved open reading frame in eubacterial 23S rRNA genes. *Proc Natl Acad Sci U S A 90*, 6864-6868.

Ratcliff, G.C., and Erie, D.A. (2001). A novel single-molecule study to determine protein--protein association constants. *J Am Chem Soc 123*, 5632-5635.

Ribeiro, F.J., Przybylski, D., Yin, S., Sharpe, T., Gnerre, S., Abouelleil, A., Berlin, A.M., Montmayeur, A., Shea, T.P., Walker, B.J.*, et al.* (2012). Finished bacterial genomes from shotgun sequence data. *Genome Res 22*, 2270-2277.

Roberts, J.D., Bebenek, K., and Kunkel, T.A. (1988). The accuracy of reverse transcriptase from HIV-1. *Science 242*, 1171-1173.

Robin, S., Togashi, D.M., Ryder, A.G., and Wall, J.G. (2009). Trigger factor from the psychrophilic bacterium Psychrobacter frigidicola is a monomeric chaperone. *J Bacteriol 191*, 1162-1168.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics 16*, 944-945.

Sambrook, J., and Russell, D.W. (2001). Molecular cloning : a laboratory manual, 3rd edn (Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press).

Schillinger, T., Lisfi, M., Chi, J., Cullum, J., and Zingler, N. (2012). Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF. *BMC Genomics 13*, 430.
Schillinger, T., and Zingler, N. (2012). The low incidence of diversity-generating retroelements in sequenced genomes. *Mobile Genetic Elements 2*, 287-291.

Schlieker, C., Bukau, B., and Mogk, A. (2002). Prevention and reversion of protein aggregation by molecular chaperones in the E. coli cytosol: implications for their applicability in biotechnology. *J Biotechnol 96*, 13-21.

Schlotawa, L., Ennemann, E.C., Radhakrishnan, K., Schmidt, B., Chakrapani, A., Christen, H.J., Moser, H., Steinmann, B., Dierks, T., and Gartner, J. (2011). SUMF1 mutations affecting stability and activity of formylglycine generating enzyme predict clinical outcome in multiple sulfatase deficiency. *Eur J Hum Genet 19*, 253-261.

Schmidt, B., Selmer, T., Ingendoh, A., and von Figura, K. (1995). A novel amino acid modification in sulfatases that is defective in multiple sulfatase deficiency. *Cell 82*, 271-278.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat Methods *9*, 671-675.

Schrodinger, LLC (2010). The PyMOL Molecular Graphics System, Version 1.3r1.

Seitz, T., Thoma, R., Schoch, G.a., Stihle, M., Benz, J., D'Arcy, B., Wiget, A., Ruf, A., Hennig, M., and Sterner, R. (2010). Enhancing the stability and solubility of the glucocorticoid receptor ligand-binding domain by high-throughput library screening. *Journal of Molecular Biology 403*, 562-577.

Sharrocks, a.D. (1994). A T7 expression vector for producing N- and C-terminal fusion proteins with glutathione S-transferase. *Gene 138*, 105-108.

Shendure, J., and Lieberman Aiden, E. (2012). The expanding scope of DNA sequencing. *Nature Biotechnology 30*, 1084-1094.

Simon, D.M., and Zimmerly, S. (2008). A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res*earch *36*, 7219-7229.

Smith, D.R., and Lee, R.W. (2009). The mitochondrial and plastid genomes of Volvox carteri: bloated molecules rich in repetitive DNA. *BMC Genomics 10*, 132.

Smith, G.P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science 228*, 1315-1317.

Smith, R.A., Klarmann, G.J., Stray, K.M., von Schwedler, U.K., Schinazi, R.F., Preston, B.D., and North, T.W. (1999). A new point mutation (P157S) in the reverse transcriptase of human immunodeficiency virus type 1 confers low-level resistance to (-)-beta-2',3'-dideoxy-3'-thiacytidine. *Antimicrob Agents Chemother 43*, 2077-2080.

Smyth, R.P., Davenport, M.P., and Mak, J. (2012). The origin of genetic diversity in HIV-1. *Virus Res 169*, 415-429.

Starlinger, P., and Saedler, H. (1972). Insertion mutations in microorganisms. Biochimie *54*, 177-185.
Stemmer, W.P. (1994). Rapid evolution of a protein in vitro by DNA shuffling. *Nature 370*, 389-391.

Stewart-Maynard, K.M., Cruceanu, M., Wang, F., Vo, M.N., Gorelick, R.J., Williams, M.C., Rouzina, I., and Musier-Forsyth, K. (2008). Retroviral nucleocapsid proteins display nonequivalent levels of nucleic acid chaperone activity. *J Virol 82*, 10129-10142.

Stockbauer, K.E., Fuchslocher, B., Miller, J.F., and Cotter, P.A. (2001). Identification and characterization of BipA, a Bordetella Bvg-intermediate phase protein. *Mol Microbiol 39*, 65-78.

Storici, F., Bebenek, K., Kunkel, T.A., Gordenin, D.A., and Resnick, M.A. (2007). RNA-templated DNA repair. *Nature 447*, 338-341.

Studier, F.W. (2005). Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif 41*, 207-234.

Suzuki, F., Goto, M., Sawa, C., Ito, S., Watanabe, H., Sawada, J., and Handa, H. (1998). Functional interactions of transcription factor human GA-binding protein subunits. *J Biol Chem 273*, 29302-29308.

Szewczyk, B., and Kozloff, L.M. (1985). A method for the efficient blotting of strongly basic proteins from sodium dodecyl sulfate-polyacrylamide gels to nitrocellulose. *Analytical biochemistry 150*, 403-407.

Takahashi, H., Okazaki, S., and Fujiwara, H. (1997). A new family of site-specific retrotransposons, SART1, is inserted into telomeric repeats of the silkworm, Bombyx mori. *Nucleic Acids Res 25*, 1578-1584.

Temin, H.M., and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature 226*, 1211-1213.

Tramontano, A., Janda, K.D., and Lerner, R.A. (1986). Catalytic antibodies. *Science 234*, 1566-1570.
Uhl, M.A., and Miller, J.F. (1996). Integration of multiple domains in a two-component sensor protein: the Bordetella pertussis BvgAS phosphorelay. *EMBO J 15*, 1028-1036.

van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M., and Brouns, S.J.J. (2009). CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends in biochemical Science*s *34*, 401-407.

Wagstaff, B.J., Hedges, D.J., Derbes, R.S., Campos Sanchez, R., Chiaromonte, F., Makova, K.D., and Roy-Engel, A.M. (2012). Rescuing Alu: recovery of new inserts shows LINE-1 preserves Alu activity through A-tail expansion. *PLoS Genet 8*, e1002842.

Wainberg, M.A., Salomon, H., Spira, B., Mercure, L., Wainberg, J., Nagai, K., Bentwich, Z., and Montaner, J. (1993). HIV resistance to anti-viral drugs. *Braz J Med Biol Res 26*, 299-308.

Waldo, G.S., Standish, B.M., Berendzen, J., and Terwilliger, T.C. (1999). Rapid protein-folding assay using green fluorescent protein. *Nature biotechnology 17*, 691-695.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature 171*, 737-738.

Wei, W., Gilbert, N., Ooi, S.L., Lawler, J.F., Ostertag, E.M., Kazazian, H.H., Boeke, J.D., and Moran, J.V. (2001). Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol 21*, 1429-1439.

Weis, W.I., Kahn, R., Fourme, R., Drickamer, K., and Hendrickson, W.A. (1991). Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing. *Science 254*, 1608-1615.

Werren, J.H. (2011). Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proc Natl Acad Sci U S A 108 Suppl 2*, 10863-10870.

Wilkinson, D.L., and Harrison, R.G. (1991). Predicting the solubility of recombinant proteins in Escherichia coli. *Biotechnology* (N Y) *9*, 443-448.

Willmund, F., Dorn, K.V., Schulz-Raffelt, M., and Schroda, M. (2008). The chloroplast DnaJ homolog CDJ1 of Chlamydomonas reinhardtii is part of a multichaperone complex containing HSP70B, CGE1, and HSP90C. *Plant physiology 148*, 2070-2082.

Woese, C.R., Winker, S., and Gutell, R.R. (1990). Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc Natl Acad Sci U S A 87*, 8467-8471.

Wong, I., and Lohman, T.M. (1993). A double-filter method for nitrocellulose-filter binding: application to protein-nucleic acid interactions. *Proc Natl Acad Sci U S A 90*, 5428-5432.

Yamamoto, M., Unzai, S., Saijo, S., Ito, K., Mizutani, K., Suno-Ikeda, C., Yabuki-Miyata, Y., Terada, T., Toyama, M., Shirouzu, M.*, et al.* (2008). Interaction and stoichiometry of the peripheral stalk subunits NtpE and NtpF and the N-terminal hydrophilic domain of NtpI of Enterococcus hirae V-ATPase. *J Biol Chem 283*, 19422-19431.

Zhang, X.D., Dou, S.X., Xie, P., Hu, J.S., Wang, P.Y., and Xi, X.G. (2006). Escherichia coli RecQ is a rapid, efficient, and monomeric helicase. *J Biol Chem 281*, 12655-12663.

Zhang, Z., Schaffer, A.A., Miller, W., Madden, T.L., Lipman, D.J., Koonin, E.V., and Altschul, S.F. (1998). Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res 26*, 3986-3990.

ANHANG

Tabelle A1: Diversitätsgenerierende Retroelemente aus dieser Arbeit

GI-Nr. der RT	Organismus	Erstreferenz	Strukturtyp[1]	Länge TR/VR [nt]	Aktivität[2]	Phage/Prophage	Motiv in Domäne 4	GI-Nr. des Zielproteins
17230989	*Nostoc* sp. PCC 7120	Doulatov et al.	2c	128	0,61; 0,64	Unbekannt	SQ	17230986; 17230987
23335577	Bifidobacterium longum DJO10A	Doulatov et al.	1a	107	0,65	Unbekannt	SQ	23335578
27311204	Vibrio phage VHML	Doulatov et al.	1a	116	0,32	Ja	SQ	27311203
29347723	Bacteroides thetaiotaomicron VPI-5482	Doulatov et al.	1a	54	0,70	Unbekannt	SQ	29347726
41179367	*Bordetella* phage BPP-1	Liu et al. (2002)	1a	120	0,76	Ja	SQ	41179369
42527768	Treponema denticola ATCC 35405	Doulatov et al.	1a	78	0,63	Unbekannt	SQ	42527771
71065017	Psychrobacter arcticus 273-4	Simon & Zimmerly	1a	117	0,52	Ja	SQ	71065019
78189651	Chlorobium chlorochromatii CaD3	Medhekar & Miller	2b	61	0,47	Unbekannt	SQ	78189652
83309559	Magnetospirillum magneticum AMB-1	Medhekar & Miller	3c	57	0,67; 0,65	Ja	SQ	83309557
83310593	Magnetospirillum magneticum AMB-1	Medhekar & Miller	3c	86	0,48; 0,52	Ja	SQ	83310595; 83310596
90580666	Photobacterium angustum S14	Simon & Zimmerly	4a	130	0,47	Unbekannt	SQ	90580665
113474819	Trichodesmium erythraeum IMS101	Doulatov et al.	2a	54	0,40	Unbekannt	SQ	113474820
119356297	Chlorobium phaeobacteroides DSM 266	Simon & Zimmerly	1d	124	0,48; 0,52	Unbekannt	SQ	119356295
119509703	Nodularia spumigena CCY9414	Simon & Zimmerly	2a	79	0,59	Unbekannt	SQ	119509705
119511289	Nodularia spumigena CCY9414	Simon & Zimmerly	2a	124	0,40	Unbekannt	SQ	119511291
120599269	Shewanella sp. W3-18-1	Simon & Zimmerly	3a	109	0,41	Ja	SQ	120599271
126090247	Shewanella baltica OS155	Simon & Zimmerly	4a	139	0,41	Unbekannt	SQ	126090249
126659397	Cyanothece sp. CCY0110	Simon & Zimmerly	2d	116	0,59; 0,54	Unbekannt	SQ	126659395; 126659398
126660098	Cyanothece sp. CCY0110	Simon & Zimmerly	2d	85	0,58; 0,36	Unbekannt	SQ	126660096; 126660100
134299090	Desulfotomaculum reducens MI-1	Simon & Zimmerly	1a	117	0,41	Ja	SQ	134299088
139439157	Collinsella aerofaciens ATCC 25986	Simon & Zimmerly	1a	128	0,44	Unbekannt	NQ	139439154
146277368	Rhodobacter sphaeroides ATCC 17025	Simon & Zimmerly	1a	117	0,38	Ja	SQ	146277366
148359926	Legionella pneumophila str. Corby	Simon & Zimmerly	1a	141	0,55	Unbekannt	SQ	148359924
150007547	Parabacteroides distasonis ATCC 8503	Simon & Zimmerly	1a	120	0,45	Ja	AQ	150007545
150390313	Alkaliphilus metalliredigens QYMF	Simon & Zimmerly	1a	111	0,48	Ja	SQ	150390315

GI-Nr. der RT	Organismus	Erstreferenz	Strukturtyp[1]	Länge TR/VR [nt]	Aktivität[2]	Phage/Prophage	Motiv in Domäne 4	GI-Nr. des Zielproteins
153810696	Ruminococcus obeum ATCC 29174	Simon & Zimmerly	1a	115	0,54	Unbekannt	SQ	153810692
153815102	Ruminococcus torques ATCC 27756	Simon & Zimmerly	1a	66	0,70	Unbekannt	SQ	153815106
154488049	Bifidobacterium adolescentis L2-32	Schillinger et al.	1a	109	0,70	Unbekannt	SQ	154488051
154502436	Ruminococcus gnavus ATCC 29149	Schillinger et al.	1a	111	0,52	Unbekannt	SP	154502435
154504579	Ruminococcus gnavus ATCC 29149	Schillinger et al.	1a	125	0,50	Unbekannt	SQ	154504576
158333546	Acaryochloris marina MBIC11017	Schillinger et al.	2a	120	0,59	Unbekannt	SQ	158333542
160885942	Bacteroides ovatus ATCC 8483	Schillinger et al.	3a	79	0,42	Unbekannt	SQ	160885940
160888221	Bacteroides uniformis ATCC 8492	Schillinger et al.	3a	95	0,55	Unbekannt	SQ	160888218
160932588	Clostridium leptum DSM 753	Schillinger et al.	1a	106	0,58	Unbekannt	SQ	160932587
160938970	Clostridium bolteae ATCC BAA-613	Schillinger et al.	1a	126	0,48	Unbekannt	SQ	160938972
160944040	Faecalibacterium prausnitzii M21/2	Schillinger et al.	1a	123	0,55	Unbekannt	VQ	160944044
161789197	Vibrio sp. 0908	Schillinger et al.	3a	102	0,46	Unbekannt	SQ	161789191
167754333	Alistipes putredinis DSM 17216	Schillinger et al.	3a	115	0,50	Unbekannt	SQ	167754335
167758195	Clostridium scindens ATCC 35704	Schillinger et al.	1a	124	0,46	Unbekannt	SQ	167758197
167759158	Clostridium scindens ATCC 35704	Schillinger et al.	1a	112	0,62	Unbekannt	SQ	167759157
167763773	Bacteroides stercoris ATCC 43183	Schillinger et al.	1a	79	0,56	Unbekannt	SQ	167763775
167841733	Burkholderia thailandensis MSMB43	Schillinger et al.	1c	89	0,38; 0,38	Unbekannt	SQ	167841735; 167841736
171058937	Leptothrix cholodnii SP-6	Schillinger et al.	1a	116	0,52	Ja	SQ	171058939
186684985	*Nostoc* punctiforme PCC 73102	Doulatov et al.	2c	127	0,48; 0,52	Unbekannt	SQ	186684982; 186684983
187929429	Ralstonia pickettii 12J	Simon & Zimmerly	3c	99	0,47; 0,47	Unbekannt	SQ	187929426; 187929427
189425132	Geobacter lovleyi SZ	Simon & Zimmerly	1d	124	0,61; 0,50	Unbekannt	SQ	189425131; 189425135
189460481	Bacteroides coprocola DSM 17136	Schillinger et al.	1a	119	0,56	Unbekannt	SQ	189460484
189463015	Bacteroides coprocola DSM 17136	Schillinger et al.	1a	117	0,37	Unbekannt	SQ	189463017
194335165	Prosthecochloris aestuarii DSM 271	Simon & Zimmerly	2a	81	0,35	Unbekannt	SQ	194335168
194337212	Pelodictyon phaeoclathratiforme BU-1	Simon & Zimmerly	1a	91	0,56	Unbekannt	SQ	194337214
194337359	Pelodictyon phaeoclathratiforme BU-1	Simon & Zimmerly	3d	85	0,45; 0,61	Unbekannt	SQ	194337361
210634695	Collinsella stercoris DSM 13279	Schillinger et al.	3a	127	0,64	Unbekannt	NQ	210634698
218295563	Thermus aquaticus Y51MC23	Schillinger et al.	1a	87	0,62	Unbekannt	SQ	218294592

GI-Nr. der RT	Organismus	Erstreferenz	Strukturtyp[1]	Länge TR/VR [nt]	Aktivität[2]	Phage/Prophage	Motiv in Domäne 4	GI-Nr. des Zielproteins
218961492	Candidatus Cloacamonas acidaminovorans	Schillinger et al.	3a	72	0,40	Unbekannt	SQ	218961497
220934786	Thioalkalivibrio sulfidophilus HL-EbGr7	Guo et al. (2011)	1a	129	0,32	Unbekannt	SQ	220934788
225163615	Opitutaceae bacterium TAV2	Simon & Zimmerly	?	?	?	Unbekannt	SQ	225163611; 225163613
225419931	Clostridium asparagiforme DSM 15981	Schillinger et al.	1a	120	0,71	Unbekannt	SQ	225419932
227485876	Anaerococcus lactolyticus ATCC 51172	Schillinger et al.	1b	93	0,46	Unbekannt	SQ	227485875
229815313	Collinsella intestinalis DSM 13280	Schillinger et al.	1a	132	0,59	Unbekannt	NQ	229815311
237710620	Bacteroides sp. 9_1_42FAA	Schillinger et al.	1a	57	0,71	Unbekannt	SQ	237710622
237716228	Bacteroides sp. D1	Schillinger et al.	3a	90	0,42	Unbekannt	SQ	345511692
237720926	Bacteroides sp. 2_2_4	Schillinger et al.	1a	112	0,50	Unbekannt	SQ	237720929
237721078	Bacteroides sp. 2_2_4	Schillinger et al.	3a	90	0,50	Unbekannt	SQ	237721080
238027140	Burkholderia glumae BGR1	Schillinger et al.	3c	90	0,37; 0,52	Unbekannt	SQ	238027142
238909111	Eubacterium eligens ATCC 27750	Schillinger et al.	1a	82	0,56	Unbekannt	SQ	238909112
239625916	Clostridiales bacterium 1_7_47_FAA	Schillinger et al.	1a	126	0,44	Unbekannt	SQ	239625914
253568001	Bacteroides sp. 1_1_6	Schillinger et al.	3a	90	0,42	Unbekannt	SQ	383122829
253570541	Bacteroides sp. 1_1_6	Schillinger et al.	3a	74	0,35	Unbekannt	AQ	253570539
253574968	Paenibacillus sp. oral taxon 786 str. D14	Schillinger et al.	1b	91	0,48	Unbekannt	SQ	253574967
253578879	Ruminococcus sp. 5_1_39BFAA	Schillinger et al.	1a	121	0,48	Unbekannt	SQ	
254523226	Stenotrophomonas sp. SKA14	Schillinger et al.	?	?		Unbekannt	SQ	
257093900	Candidatus Accumulibacter phosphatis clade IIA str. UW-1	Schillinger et al.	1a	60	0,67	Unbekannt	SQ	257093902
257094650	Candidatus Accumulibacter phosphatis clade IIA str. UW-1	Schillinger et al.	2a	84	0,67	Unbekannt	SQ	257094652
257439360	Faecalibacterium prausnitzii A2-165	Schillinger et al.	1a	114	0,52	Unbekannt	SQ	257439358
257440840	Faecalibacterium prausnitzii A2-165	Schillinger et al.	1b	93	0,50	Unbekannt	SQ	257440841
257792377	Eggerthella lenta DSM 2243	Schillinger et al.	1a	130	0,50	Unbekannt	NQ	257792375
258517326	Desulfotomaculum acetoxidans DSM 771	Guo et al. (2011)	1a	115	0,62	Unbekannt	SQ	258517324
260587208	Blautia hansenii DSM 20583	Schillinger et al.	1a	114	0,56	Unbekannt	SQ	260587206
260642123	Bacteroides finegoldii DSM 17565	Schillinger et al.	1a	120	0,54	Unbekannt	SQ	
261880961	Prevotella bergensis DSM 17361	Schillinger et al.	1a	117	0,43	Unbekannt	SQ	261880963
262403399	Vibrio sp. RC586	Schillinger et al.	4a	139	0,49	Unbekannt	SQ	

GI-Nr. der RT	Organismus	Erstreferenz	Strukturtyp[1]	Länge TR/VR [nt]	Aktivität[2]	Phage/Prophage	Motiv in Domäne 4	GI-Nr. des Zielproteins
265750836	Bacteroides sp. 3_1_33FAA	Schillinger et al.	1a	117	0,41	Unbekannt	SQ	265750838
266622220	Clostridium hathewayi DSM 13479	Schillinger et al.	1a	89	0,54	Unbekannt	SQ	266622224
283796807	Clostridium sp. M62/1	Schillinger et al.	1a	97	0,44	Unbekannt	SQ	
291513637	Alistipes shahii WAL 8301	Schillinger et al.	1a	113	0,69	Unbekannt	SQ	291513639
291517473	Bifidobacterium longum subsp. longum F8	Schillinger et al.	1a	109	0,80	Unbekannt	SQ	291517474
291522399	Coprococcus catus GD/7	Schillinger et al.	1a	158	0,73	Unbekannt	PA	
291524999	Eubacterium rectale DSM 17629	Guo et al. (2011)	1a	126	0,48	Unbekannt	SQ	291525001
291526343	Eubacterium rectale DSM 17629	Guo et al. (2011)	1a	127	0,75	Unbekannt	SQ	291526342
291546485	Ruminococcus sp. SR1/5	Schillinger et al.	1a	95	0,38	Unbekannt	SH	291546483
291561321	butyrate-producing bacterium SS3/4	Schillinger et al.	1a	88	0,54	Unbekannt	AQ	291561319
291561556	butyrate-producing bacterium SS3/4	Schillinger et al.	1b	79	0,65	Unbekannt	SQ	
291612675	Sideroxydans lithotrophicus ES-1	Schillinger et al.	3c	97	0,36; 0,50	Unbekannt	SQ	291612672; 291612673
291614412	Sideroxydans lithotrophicus ES-1	Schillinger et al.	1a	117	0,75	Unbekannt	SQ	291614410
291621968	Vibrio phage VP58.5	Schillinger et al.	3a	116	0,52	Ja	SQ	291621966
293369862	Bacteroides ovatus SD CMC 3f	Schillinger et al.	3a	89	0,36	Unbekannt	AQ	293369860
293372957	Bacteroides ovatus SD CMC 3f	Schillinger et al.	3a	90	0,35	Unbekannt	SQ	293372955
294674266	Prevotella ruminicola 23	Schillinger et al.	1a	88	0,38	Unbekannt	SQ	294674267
294778518	Bacteroides vulgatus PC510	Schillinger et al.	1a	119	0,54	Unbekannt	SQ	294778517
297374652	Candidatus Magnetobacterium bavaricum	Schillinger et al.	1b	122	0,43	Unbekannt	SQ	297374650
298385378	Bacteroides sp. 1_1_14	Schillinger et al.	1a	77	0,55	Unbekannt	SQ	298385375
298387225	Bacteroides sp. 1_1_14	Schillinger et al.	3a	89	0,36	Unbekannt	SQ	298387223
298480077	Bacteroides sp. D22	Schillinger et al.	1a	107	0,32	Unbekannt	SQ	298480076
298480981	Bacteroides sp. D22	Schillinger et al.	3a	91	0,54	Unbekannt	SQ	298480979
299067086	Ralstonia solanacearum CMR15	Schillinger et al.	3c	91	0,41; 0,57	Unbekannt	SQ	299067088; 299067089
299145828	Bacteroides sp. 3_1_23	Schillinger et al.	3a	115	0,32	Unbekannt	SQ	299145826
301163046	Bacteroides fragilis 638R	Schillinger et al.	1a	183	0,36	Unbekannt	SQ	
308273696	uncultured Desulfobacterium sp.	Schillinger et al.	3d	141	0,55; 0,49	Unbekannt	SQ	308273693; 308273697
312115534	Rhodomicrobium vannielii ATCC 17100	Schillinger et al.	2e	118	0,42; 0,32	Unbekannt	SQ	312115533; 312115538

GI-Nr. der RT	Organismus	Erstreferenz	Strukturtyp[1]	Länge TR/VR [nt]	Aktivität[2]	Phage/Prophage	Motiv in Domäne 4	GI-Nr. des Zielproteins
312962032	Pseudomonas fluorescens WH6	Schillinger et al.	1c	100	0,38; 0,38	Unbekannt	SQ	312962029; 312962030
313113756	Faecalibacterium cf. prausnitzii KLE1255	Schillinger et al.	1b	87	0,55	Unbekannt	SQ	313113757
313147618	Bacteroides fragilis 3_1_12	Schillinger et al.	1a	121	0,62	Unbekannt	SQ	313147617
313147952	Bacteroides fragilis 3_1_12	Schillinger et al.	1a	119	0,56	Unbekannt	SQ	313147953
315918944	Bacteroides sp. D2	Schillinger et al.	3a	90	0,46	Unbekannt	SQ	383111379
315923025	Bacteroides sp. D2	Schillinger et al.	1a	118	0,58	Unbekannt	SQ	383112786
317498222	Lachnospiraceae bacterium 5_1_63FAA	Schillinger et al.	1a	109	0,63	Unbekannt	SP	317498224
319764254	Alicycliphilus denitrificans BC	Schillinger et al.	3a	100	0,45	Unbekannt	SQ	319764252
322688997	Bifidobacterium longum subsp. infantis 157F	Schillinger et al.	1a	102	0,68	Unbekannt	SQ	322688999
323485151	Clostridium symbiosum WAL-14163	Schillinger et al.	1a	68	0,38	Unbekannt	SQ	323485155
323485235	Clostridium symbiosum WAL-14163	Schillinger et al.	1a	49	0,66	Unbekannt	SQ	323485236
323486088	Clostridium symbiosum WAL-14163	Schillinger et al.	1a	114	0,63	Unbekannt	SH	323486086
323691302	Clostridium symbiosum WAL-14673	Schillinger et al.	1a	112	0,48	Unbekannt	SQ	323691301
325279507	Odoribacter splanchnicus DSM 20712	Schillinger et al.	1a	122	0,50	Unbekannt	SQ	325279508
331088842	Lachnospiraceae bacterium 3_1_46FAA	Schillinger et al.	1a	121	0,65	Unbekannt	SQ	331088844
331089069	Lachnospiraceae bacterium 3_1_46FAA	Schillinger et al.	1a	97	0,48	Unbekannt	SQ	331089072
332533803	Pseudoalteromonas haloplanktis ANT/505	Schillinger et al.	3a	116	0,46	Unbekannt	SQ	332533805
332534155	Pseudoalteromonas haloplanktis ANT/505	Schillinger et al.	3a	116	0,53	Unbekannt	SQ	332534157
332653877	Ruminococcaceae bacterium D16	Schillinger et al.	1b	94	0,60	Unbekannt	SQ	332653876
332654472	Ruminococcaceae bacterium D16	Schillinger et al.	1b	93	0,38	Unbekannt	SQ	332654471
332655364	Ruminococcaceae bacterium D16	Schillinger et al.	1a	124	0,44	Unbekannt	SQ	332655366
332666264	Haliscomenobacter hydrossis DSM 1100	Schillinger et al.	2a	113	0,62	Unbekannt	SQ	332666261
332707562	Lyngbya majuscula 3L	Schillinger et al.	?	?	?	Unbekannt	SQ	332707564; 332707565
336402102	Bacteroides sp. 1_1_30	Schillinger et al.	1a	118	0,45	Unbekannt	SQ	336402101
336410996	Bacteroides sp. 2_1_56FAA	Schillinger et al.	1a	121	0,46	Unbekannt	SQ	336410997
336413450	Bacteroides ovatus 3_8_47FAA	Schillinger et al.	1a	118	0,58	Unbekannt	SQ	336413451
336433792	Lachnospiraceae bacterium 2_1_58FAA	Schillinger et al.	1a	82	0,46	Unbekannt	SQ	336433794
336435571	Lachnospiraceae bacterium 1_4_56FAA	Schillinger et al.	1c	68	0,52; 0,65	Unbekannt	AQ	336435573; 336435574

GI-Nr. der RT	Organismus	Erstreferenz	Strukturtyp[1]	Länge TR/VR [nt]	Aktivität[2]	Phage/Prophage	Motiv in Domäne 4	GI-Nr. des Zielproteins
338762709	Lactobacillus johnsonii pf01	Schillinger et al.	?	?	?	Unbekannt	SP	
338999650	Halomonas sp. TD01	Schillinger et al.	1a	67	0,57	Unbekannt	SQ	338999653
340788559	Collimonas fungivorans Ter331	Schillinger et al.	3f	107	0,22; 0,44; 0,31	Unbekannt	SQ	340788556; 340788557
341642641	Vibrio cholerae HE-09	Schillinger et al.	1a	121	0,50	Unbekannt	SQ	341642638
344341810	Thiocapsa marina 5811	Schillinger et al.	2b	63	0,83	Unbekannt	SQ	344341809
344343882	Marichromatium purpuratum 984	Schillinger et al.	2f	124	0,53	Unbekannt	SQ	344343879
345871758	Thiorhodococcus drewsii AZ1	Schillinger et al.	2d	116	0,65; 0,67	Unbekannt	SQ	345871756
345883525	Prevotella sp. C561	Schillinger et al.	1a	100	0,50	Unbekannt	SQ	
345893474	Desulfovibrio sp. 6_1_46AFAA	Schillinger et al.	1a	106	0,30	Unbekannt	SQ	345893468
347538767	Pseudogulbenkiania sp. NH8B	Schillinger et al.	1c	103	0,33; 0,33	Unbekannt	SQ	347538771; 347538772
348026361	Megasphaera elsdenii DSM 20460	Schillinger et al.	1a	72	0,56	Unbekannt	SQ	348026359
350551816	Thiorhodospira sibirica ATCC 700588	Schillinger et al.	3d	80	0,70; 0,50	Unbekannt	SQ	350551814; 350551817
350574651	Thiorhodovibrio sp. 970	Schillinger et al.	3a	98	0,48	Unbekannt	SQ	377661873
355363092	Desulfobacter postgatei 2ac9	Schillinger et al.	2d	113	0,38	Unbekannt	SQ	389579824; 389579827
355386195	Clostridium clostridioforme 2_1_49FAA	Schillinger et al.	1a	83	0,65	Unbekannt	SQ	357052768
355625064	Clostridium sp. 7_3_54FAA	Schillinger et al.	1a	89	0,47	Unbekannt	SQ	
355629964	Clostridium sp. 7_3_54FAA	Schillinger et al.	1a	54	0,87	Unbekannt	SQ	355629966
355681483	Clostridium citroniae WAL-17108	Schillinger et al.	1a	87	0,65	Unbekannt	SQ	355681487

[1] siehe Abbildung 8 in Abschnitt 3.1.2.1

[2] berechnet aus dem Verhältnis der Adeninaustausche zur Gesamtzahl der Adenine im assoziierten Template Repeat

Tabelle A2: Akzessorische Proteine aus dieser Arbeit

GI-Nr. des akzessorischen Proteins	Organismus	Länge [AS]	Isoelektrischer Punkt [pH]	Consensus	GI-Nr. der RT
17230988	*Nostoc* sp. PCC 7120	114	9,41	IGTELGGWIK	17230989
41179368	*Bordetella* phage BPP-1	128	10,18	VGRILGSWIA	41179367
71065018	Psychrobacter arcticus 273-4	127	9,72	IGKIIGGWIK	71065017
83309558	Magnetospirillum magneticum AMB-1	124	10,0	IGQQASGWLK	83309559
83310594	Magnetospirillum magneticum AMB-1	117	10,29	VGKQATAWLK	83310593
90580667	Photobacterium angustum S14	126	10,04	IQRQVVGWRK	90580666
113474818	Trichodesmium erythraeum IMS101	100	9,65	IGK*LGIWIK	113474819
119356296	Chlorobium phaeobacteroides DSM 266	106	9,48	LGKMLGGWLK	119356297
119509704	Nodularia spumigena CCY9414	114	9,47	IGFELGGWIK	119509703
119511290	Nodularia spumigena CCY9414	115	9,71	IGNELGGWIK	119511289
120599270	Shewanella sp. W3-18-1	117	10,29	IGKMTGGWIK	120599269
126090246	Shewanella baltica OS155	122	9,51	VSRQATGWRK	126090247
126090246	Vibrio sp. RC586	122	9,51	VSRQTTGWRK	262403399
126659396	Cyanothece sp. CCY0110	131	9,8	IGIELGGWIK	126659397
126660099	Cyanothece sp. CCY0110	112	9,6	VGTELGGWIK	126660098
134299089	Desulfotomaculum reducens MI-1	100	9,78	IGRMIGGWLK	134299090
146277367	Rhodobacter sphaeroides ATCC 17025	116	9,23	IGRMIGGWFK	146277368
148359925	Legionella pneumophila str. Corby	102	9,62	IGKQVTGWRN	148359926
150390314	Alkaliphilus metalliredigens QYMF	113	9,84	IGRMLGGWMK	150390313
154504577	Ruminococcus gnavus ATCC 29149	118	8,63	IGVMLGEMIE	154504579
158333543	Acaryochloris marina MBIC11017	112	9,59	IGLDLGGWIK	158333546
160885941	Bacteroides ovatus ATCC 8483	135	8,34	IGKQSTGWYK	160885942
160885941	Bacteroides sp. D1	135	9,16	IGKQSTGWYK	237716228
160885941	Bacteroides ovatus SD CMC 3f	135	8,34	IGKQSTGWYK	293372957
160885941	Bacteroides sp. D22	135	8,34	IGKQSTGWYK	298480981

GI-Nr. des akzessorischen Proteins	Organismus	Länge [AS]	Isoelektrischer Punkt [pH]	Consensus	GI-Nr. der RT
160885941	Bacteroides sp. D2	135	9,13	IGKQSTGWYK	315918944
160888219	Bacteroides uniformis ATCC 8492	124	9,71	IGKQATGWKQ	160888221
160938971	Clostridium bolteae ATCC BAA-613	129	9,42	LGCIIGGIIE	160938970
160938971	Clostridiales bacterium 1_7_47_FAA	129	9,42	LGCIIGGIIE	239625916
160938971	Eubacterium rectale DSM 17629	129	9,42	LGCIIGGIIE	291524999
161789150	Vibrio sp. 0908	117	10,17	IGKMIGGWIK	161789197
167758196	Clostridium scindens ATCC 35704	138	9,28	IGGKIGGLIK	167758195
167841734	Burkholderia thailandensis MSMB43	124	9,72	IGKQANGWKN	167841733
171058938	Leptothrix cholodnii SP-6	114	9,47	LGRMIGGWVK	171058937
186684984	*Nostoc* punctiforme PCC 73102	114	9,18	IGIELGGWIK	186684985
187929428	Ralstonia pickettii 12J	123	10,3	VGKQVGGWRK	187929429
189468312	Bacteroides thetaiotaomicron VPI-5482	123	9,68	LSKQLSAWHD	29347723
189468312	Bacteroides fragilis 638R	123	9,68	LSKQLSAWHD	301163046
194335166	Prosthecochloris aestuarii DSM 271	120	10,19	VGKMLGGWLR	194335165
194335166	Candidatus Cloacamonas acidaminovorans	120	9,62	SGKMVGGWLK	218961492
194337213	Pelodictyon phaeoclathratiforme BU-1	106	10,28	LGKMLGGWIK	194337212
218295564	Thermus aquaticus Y51MC23	115	10,63	LGRMVGGWLK	218295563
225163614	Opitutaceae bacterium TAV2	118	10,48	CGKMLGGWIK	225163615
227485877	Anaerococcus lactolyticus ATCC 51172	130	9,34	VKNMAVSWHM	227485876
237721080	Bacteroides sp. 2_2_4	269	8,34	IGKQSTGWYK	237721078
238027141	Burkholderia glumae BGR1	119	10,09	IGKQATGWRN	238027140
253574970	Paenibacillus sp. oral taxon 786 str. D14	115	9,71	LGRITGGLIK	253574968
257439359	Faecalibacterium prausnitzii A2-165	114	10,34	IGRMLGGWKK	257439360
257440839	Faecalibacterium prausnitzii A2-165	119	10,23	IGRIIGGLQK	257440840
258517325	Desulfotomaculum acetoxidans DSM 771	113	9,94	IGRMLGGWIK	258517326
260587207	Blautia hansenii DSM 20583	114	10,34	IGRMLGGWKK	260587208
291546484	Ruminococcus sp. SR1/5	123	9,49	IGRMLGGWMA	291546485
291612674	Sideroxydans lithotrophicus ES-1	122	9,52	VGKQAGGWRK	291612675

GI-Nr. des akzessorischen Proteins	Organismus	Länge [AS]	Isoelektrischer Punkt [pH]	Consensus	GI-Nr. der RT
291621967	Vibrio phage VP58.5	106	10,19	LGKMIGGWIR	291621968
293369861	Bacteroides ovatus SD CMC 3f	126	9,52	IEKQIIGWRN	293369862
298387224	Bacteroides sp. 1_1_14	126	9,9	VEKQILGWRN	298387225
299067087	Ralstonia solanacearum CMR15	124	10,0	IGKQANGWKK	299067086
308273694	uncultured Desulfobacterium sp.	128	9,95	LSKQSEGWLK	308273696
312115537	Rhodomicrobium vannielii ATCC 17100	121	10,46	IGRLVGGWAK	312115534
312962031	Pseudomonas fluorescens WH6	117	10,58	VGRQANAWKK	312962032
319764253	Alicycliphilus denitrificans BC	130	10,34	IGRKAGGWAK	319764254
323486087	Clostridium symbiosum WAL-14163	111	9,36	IGRMVGGWIK	323486088
325288745	butyrate-producing bacterium SS3/4	111	9,66	IGRIIGGLQQ	291561556
332533804	Pseudoalteromonas haloplanktis ANT/505	117	10,18	IGKMLGGWIK	332533803
332534156	Pseudoalteromonas haloplanktis ANT/505	118	10,19	IGKMLGGWVK	332534155
332654473	Ruminococcaceae bacterium D16	117	10,02	MGRIIGGLQK	332654472
332666263	Haliscomenobacter hydrossis DSM 1100	115	9,23	AGKMCGGWMK	332666264
332707563	Lyngbya majuscula 3L	115	10,17	IGKELGGWIK	332707562
336433793	Lachnospiraceae bacterium 2_1_58FAA	117	9,8	IKYMTIAWRN	336433792
336435572	Lachnospiraceae bacterium 1_4_56FAA	123	9,4	IKHMAIAWRK	336435571
338999652	Halomonas sp. TD01	121	9,38	IGCMIGGWLK	338999650
340788558	Collimonas fungivorans Ter331	131	9,98	IGKQANGWRK	340788559
344343881	Marichromatium purpuratum 984	120	10,87	VGRMLGGWIR	344343882
345893472	Desulfovibrio sp. 6_1_46AFAA	135	10,56	IGKMLGAWLK	345893474
347538770	Pseudogulbenkiania sp. NH8B	118	9,98	IGKQANGWRK	347538767
348026360	Megasphaera elsdenii DSM 20460	72	9,5	IGKMLGGWIK	348026361
350551815	Thiorhodospira sibirica ATCC 700588	126	10,71	LGRQAGGWLK	350551816
355629965	Clostridium sp. 7_3_54FAA	114	10,25	IGRMLGGWKK	355629964
357052769	Clostridium clostridioforme 2_1_49FAA	120	9,7	IKYMTIAWRS	355386195
381160529	Thiorhodovibrio sp. 970	132	9,58	TGRLVGGWLR	350574651
383122828	Bacteroides sp. 1_1_6	135	7,5	IGKQSTGWYK	253568001

GI-Nr. des akzessorischen Proteins	Organismus	Länge [AS]	Isoelektrischer Punkt [pH]	Consensus	GI-Nr. der RT
383123377	Bacteroides sp. 1_1_6	126	9,98	IEKQIIGWRN	253570541
422908029	Vibrio cholerae HE-09	107	10,14	CGAMLNAWLK	341642641

Tabelle A3: DGR-Zielproteine aus dieser Arbeit

GI-Nr. des Zielproteins	Organismus	Annotation	Vorhersage der subzellulären Lokalisation mit PSort	Phylogenetische Gruppe[1]	GI-Nr. der RT
17230986	*Nostoc* sp. PCC 7120	Formylglycingenerierendes Enzym	Unbekannt	1	17230989
17230987	*Nostoc* sp. PCC 7120	Formylglycingenerierendes Enzym	Zytosol	1	
23335578	Bifidobacterium longum DJO10A	Unbekannt	Unbekannt	2	23335577
27311203	Vibrio phage VHML	ORF35; DUF323	Phage	1	27311204
29347726	Bacteroides thetaiotaomicron VPI-5482	DUF3988	Extrazellulär	3	29347723
41179369	*Bordetella* phage BPP-1	Major Tropism-Determinante	Phage	1	41179367
42527771	Treponema denticola ATCC 35405	Formylglycingenerierendes Enzym	Unbekannt	1	42527768
71065019	Psychrobacter arcticus 273-4	Unbekannt	Unbekannt	1	71065017
78189652	Chlorobium chlorochromatii CaD3	Formylglycingenerierendes Enzym	Zytosol	1	78189651
83309557	Magnetospirillum magneticum AMB-1	Unbekannt	Unbekannt	3	83309559
83310595	Magnetospirillum magneticum AMB-1	Fibrobacter succinogenes Major Domain	Unbekannt	3	83310593
83310596	Magnetospirillum magneticum AMB-1	Unbekannt	Unbekannt	3	
90580665	Photobacterium angustum S14	Salmonella repeat of unknown function (DUF823); pfam05689; Peptidase associated domain: C-terminal domain of M14 N/E carboxypeptidase	Unbekannt (mehrere möglich)	3	90580666
113474820	Trichodesmium erythraeum IMS101	Formylglycingenerierendes Enzym; Serin/Threoninkinase	Membran (intrazellulär)	1	113474819
119356295	Chlorobium phaeobacteroides DSM 266	Formylglycingenerierendes Enzym; NACHT NTPase;P-Loop; DUF323	Zytosol	1	119356297
119509705	Nodularia spumigena CCY9414	Formylglycingenerierendes Enzym	Unbekannt (mehrere möglich)	1	119509703
119511291	Nodularia spumigena CCY9414	Formylglycingenerierendes Enzym; Gliding motility-assoziiertes Lipoprotein GldJ; TIGR03530	Zytosol	1	119511289
120599271	Shewanella sp. W3-18-1	Unbekannt	Extrazellulär	1	120599269
126090249	Shewanella baltica OS155	Ig-Domäne; Salmonella repeat of unknown function (DUF823); Lk90-like protein	Extrazellulär (mehrere möglich)	2	126090247
126659395	Cyanothece sp. CCY0110	Formylglycingenerierendes Enzym	Zytosol	1	126659397
126659398	Cyanothece sp. CCY0110	Unbekannt	Zytosol	1	
126660096	Cyanothece sp. CCY0110	Formylglycingenerierendes Enzym	Zytosol	1	126660098
126660100	Cyanothece sp. CCY0110	Von Willebrand-Faktor Typ A (vWA)-Domäne	Zytosol	1	
134299088	Desulfotomaculum reducens MI-1	Unbekannt	Unbekannt	1	134299090

GI-Nr. des Zielproteins	Organismus	Annotation	Vorhersage der subzellulären Lokalisation mit PSort	Phylogenetische Gruppe[1]	GI-Nr. der RT
139439154	Collinsella aerofaciens ATCC 25986	Triacylglycerollipase	Unbekannt	3	139439157
146277366	Rhodobacter sphaeroides ATCC 17025	Unbekannt	Unbekannt	1	146277368
148359924	Legionella pneumophila str. Corby	Unbekannt	Extrazellulär	3	148359926
150007545	Parabacteroides distasonis ATCC 8503	Unbekannt	Unbekannt	2	150007547
150390315	Alkaliphilus metalliredigens QYMF	Formylglycingenerierendes Enzym	Unbekannt	1	150390313
153810692	Ruminococcus obeum ATCC 29174	NADH-abhängige Glyceratdehydrogenase	Unbekannt	1	153810696
153815106	Ruminococcus torques ATCC 27756	DNA-Topoisomerase IV	Membran (intrazellulär)	3	153815102
154488051	Bifidobacterium adolescentis L2-32	Unbekannt	Unbekannt	2	154488049
154502435	Ruminococcus gnavus ATCC 29149	Unbekannt	Unbekannt	2	154502436
154504576	Ruminococcus gnavus ATCC 29149	Major Tropism-Determinante; PHA00653	Unbekannt	1	154504579
158333542	Acaryochloris marina MBIC11017	Caspase-Domäne; pfam00656	Zytosol	1	158333546
160885940	Bacteroides ovatus ATCC 8483	Unbekannt	Unbekannt	3	160885942
160888218	Bacteroides uniformis ATCC 8492	Protein of unknown function (DUF3988); pfam13149	Periplasma	3	160888221
160932587	Clostridium leptum DSM 753	Unbekannt	Extrazellulär	2	160932588
160938972	Clostridium bolteae ATCC BAA-613	Major Tropism-Determinante; PHA00653	Unbekannt	1	160938970
160944044	Faecalibacterium prausnitzii M21/2	Unbekannt	Unbekannt	3	160944040
161789191	Vibrio sp. 0908	Formylglycingenerierendes Enzym	Unbekannt	1	161789197
167754335	Alistipes putredinis DSM 17216	Unbekannt	Zytosol	2	167754333
167758197	Clostridium scindens ATCC 35704	Formylglycingenerierendes Enzym, Major Tropism-Determinante, Serin/Threoninkinase	Unbekannt	1	167758195
167759157	Clostridium scindens ATCC 35704	Unbekannt	Unbekannt	2	167759158
167763775	Bacteroides stercoris ATCC 43183	Unbekannt	Unbekannt	2	167763773
167841735	Burkholderia thailandensis MSMB43	Unbekannt	Unbekannt	3	167841733
167841736	Burkholderia thailandensis MSMB43	Unbekannt	Unbekannt	3	
171058939	Leptothrix cholodnii SP-6	Unbekannt	Extrazellulär	1	171058937
186684982	*Nostoc* punctiforme PCC 73102	Formylglycingenerierendes Enzym	Zytosol	1	186684985
186684983	*Nostoc* punctiforme PCC 73102	Formylglycingenerierendes Enzym	Unbekannt	1	
187929426	Ralstonia pickettii 12J	Unbekannt	Unbekannt	3	187929429
187929427	Ralstonia pickettii 12J	Unbekannt	Unbekannt	3	

GI-Nr. des Zielproteins	Organismus	Annotation	Vorhersage der subzellulären Lokalisation mit PSort	Phylogenetische Gruppe[1]	GI-Nr. der RT
189425131	Geobacter lovleyi SZ	Protein of unknown function (DUF1566); pfam07603	Unbekannt	3	189425132
189425135	Geobacter lovleyi SZ	Protein of unknown function (DUF1566); pfam07603	Unbekannt (mehrere möglich)	3	
189460484	Bacteroides coprocola DSM 17136	Unbekannt	Unbekannt	3	189460481
189463017	Bacteroides coprocola DSM 17136	Unbekannt	Membran (extrazellulär)	1	189463015
194335168	Prosthecochloris aestuarii DSM 271	Formylglycingenerierendes Enzym	Unbekannt	1	194335165
194337214	Pelodictyon phaeoclathratiforme BU-1	Formylglycingenerierendes Eznym; P-Loop	Zytosol	1	194337212
194337361	Pelodictyon phaeoclathratiforme BU-1	Formylglycingenerierendes Enzym; Phosphohydrolase; Bacillus subtilis YydB Homolog; Metallophosphatase-Domäne; DUF323; NACHT nucleoside triphosphatase	Unbekannt	1	194337359
210634698	Collinsella stercoris DSM 13279	Unbekannt	Unbekannt	3	210634695
218294592	Thermus aquaticus Y51MC23	Formylglycingenerierendes Enzym	Unbekannt	1	218295563
218961497	Candidatus Cloacamonas acidaminovorans	Formylglycingenerierendes Enzym; Transcriptional regulator (Enhancement of xylanase A production)	Unbekannt	1	218961492
220934788	Thioalkalivibrio sulfidophilus HL-EbGr7	Major Tropism-Determinante; PHA00653	Unbekannt	1	220934786
225163611	Opitutaceae bacterium TAV2	Formylglycingenerierendes Enzym	Extrazellulär	1	225163615
225163613	Opitutaceae bacterium TAV2	GxxExxY Protein; TIGR04256	Zytosol	3	
225419932	Clostridium asparagiforme DSM 15981	Unbekannt	Unbekannt	2	225419931
227485875	Anaerococcus lactolyticus ATCC 51172	3-dehydroquinate dehydratase, type I; TIGR01093	Unbekannt	3	227485876
229815311	Collinsella intestinalis DSM 13280	Unbekannt	Zytosol	3	229815313
237710622	Bacteroides sp. 9_1_42FAA	predicted protein	Unbekannt	2	237710620
237720929	Bacteroides sp. 2_2_4	predicted protein	Extrazellulär	2	237720926
237721080	Bacteroides sp. 2_2_4	conserved Unbekannt	Unbekannt	3	237721078
238027142	Burkholderia glumae BGR1	Unbekannt	Unbekannt	3	238027140
238909112	Eubacterium eligens ATCC 27750	Unbekannt	Unbekannt	2	238909111
239625914	Clostridiales bacterium 1_7_47_FAA	Formylglycingenerierendes Enzym	Unbekannt	1	239625916
253570539	Bacteroides sp. 1_1_6	conserved Unbekannt	Unbekannt	3	253570541
253574967	Paenibacillus sp. oral taxon 786 str. D14	predicted protein	Unbekannt	3	253574968
257093902	Candidatus Accumulibacter phosphatis clade IIA str. UW-1	Formylglycingenerierendes Enzym; NACHT NTPase; DUF323	Zytosol	1	257093900
257094652	Candidatus Accumulibacter phosphatis clade IIA str. UW-1	Formylglycingenerierendes Enzym	Zytosol	1	257094650
257439358	Faecalibacterium prausnitzii A2-165	Formylglycingenerierendes Enzym	Unbekannt	1	257439360

GI-Nr. des Zielproteins	Organismus	Annotation	Vorhersage der subzellulären Lokalisation mit PSort	Phylogenetische Gruppe[1]	GI-Nr. der RT
257440841	Faecalibacterium prausnitzii A2-165	Serralysin	Zytosol	3	257440840
257792375	Eggerthella lenta DSM 2243	Unbekannt	Unbekannt	3	257792377
258517324	Desulfotomaculum acetoxidans DSM 771	Formylglycingenerierendes Enzym	Unbekannt	1	258517326
260587206	Blautia hansenii DSM 20583	Formylglycingenerierendes Enzym	Unbekannt	1	260587208
261880963	Prevotella bergensis DSM 17361	Concanavalin A-like lectin/glucanases superfamily; pfam13385	Membran (extrazellulär)	1	261880961
265750838	Bacteroides sp. 3_1_33FAA	Concanavalin A-like lectin/glucanases superfamily; pfam13385	Membran (extrazellulär)	1	265750836
266622224	Clostridium hathewayi DSM 13479	Unbekannt	Membran (intrazellulär)	3	266622220
291513639	Alistipes shahii WAL 8301	Unbekannt	Extrazellulär	2	291513637
291517474	Bifidobacterium longum subsp. longum F8	Unbekannt	Unbekannt	2	291517473
291525001	Eubacterium rectale DSM 17629	Formylglycingenerierendes Enzym	Unbekannt	1	291524999
291526342	Eubacterium rectale DSM 17629	Unbekannt	Unbekannt	2	291526343
291546483	Ruminococcus sp. SR1/5	Formylglycingenerierendes Enzym	Extrazellulär	1	291546485
291561319	butyrate-producing bacterium SS3/4	Unbekannt	Unbekannt	3	291561321
291612672	Sideroxydans lithotrophicus ES-1	Unbekannt	Unbekannt	3	291612675
291612673	Sideroxydans lithotrophicus ES-1	Unbekannt	Unbekannt	3	
291614410	Sideroxydans lithotrophicus ES-1	Unbekannt	Zytosol	1	291614412
291621966	Vibrio phage VP58.5	gp18 protein	Phage	1	291621968
293369860	Bacteroides ovatus SD CMC 3f	Unbekannt	Unbekannt	3	293369862
293372955	Bacteroides ovatus SD CMC 3f	Unbekannt	Unbekannt	3	293372957
294674267	Prevotella ruminicola 23	Unbekannt	Extrazellulär	3	294674266
294778517	Bacteroides vulgatus PC510	Concanavalin A-like lectin/glucanases superfamily; pfam13385	Unbekannt (mehrere möglich)	1	294778518
297374650	Candidatus Magnetobacterium bavaricum	Protein of unknown function (DUF1566); pfam07603; Fibronectin Typ III Domäne	Extrazellulär	3	297374652
298385375	Bacteroides sp. 1_1_14	conserved Unbekannt	Unbekannt (mehrere möglich)	3	298385378
298387223	Bacteroides sp. 1_1_14	conserved Unbekannt	Unbekannt	3	298387225
298480076	Bacteroides sp. D22	Unbekannt	Unbekannt	2	298480077
298480979	Bacteroides sp. D22	lipoprotein	Unbekannt	3	298480981
299067088	Ralstonia solanacearum CMR15	6-phosphofructokinase, eukaryotic type; TIGR02478	Unbekannt	3	299067086
299067089	Ralstonia solanacearum CMR15	protein of unknown function	Unbekannt	3	

GI-Nr. des Zielproteins	Organismus	Annotation	Vorhersage der subzellulären Lokalisation mit PSort	Phylogenetische Gruppe[1]	GI-Nr. der RT
299145826	Bacteroides sp. 3_1_23	conserved Unbekannt	Unbekannt	3	299145828
308273693	uncultured Desulfobacterium sp.	Protein of unknown function (DUF1566); pfam07603	Extrazellulär	3	308273696
308273697	uncultured Desulfobacterium sp.	Protein of unknown function (DUF1566); pfam07603	Unbekannt	3	
312115533	Rhodomicrobium vannielii ATCC 17100	TIR domain; pfam13676	Unbekannt	1	312115534
312115538	Rhodomicrobium vannielii ATCC 17100	TIR domain; pfam13676	Zytosol	1	
312962029	Pseudomonas fluorescens WH6	Unbekannt	Unbekannt	3	312962032
312962030	Pseudomonas fluorescens WH6	Unbekannt	Unbekannt	3	
313113757	Faecalibacterium cf. prausnitzii KLE1255	beta-glycosidase-like protein	Unbekannt	3	313113756
313147617	Bacteroides fragilis 3_1_12	predicted protein	Unbekannt	2	313147618
313147953	Bacteroides fragilis 3_1_12	conserved Unbekannt	Unbekannt	2	313147952
317498224	Lachnospiraceae bacterium 5_1_63FAA	Unbekannt	Unbekannt	2	317498222
319764252	Alicycliphilus denitrificans BC	Unbekannt	Unbekannt	3	319764254
322688999	Bifidobacterium longum subsp. infantis 157F	Unbekannt	Zytosol	2	322688997
323485155	Clostridium symbiosum WAL-14163	Unbekannt	Unbekannt	3	323485151
323485236	Clostridium symbiosum WAL-14163	Unbekannt	Unbekannt	2	323485235
323486086	Clostridium symbiosum WAL-14163	Formylglycingenerierendes Enzym	Unbekannt	1	323486088
323691301	Clostridium symbiosum WAL-14673	Unbekannt	Unbekannt	2	323691302
325279508	Odoribacter splanchnicus DSM 20712	Unbekannt	Unbekannt	2	325279507
331088844	Lachnospiraceae bacterium 3_1_46FAA	Unbekannt	Unbekannt	2	331088842
331089072	Lachnospiraceae bacterium 3_1_46FAA	Unbekannt	Unbekannt	3	331089069
332533805	Pseudoalteromonas haloplanktis ANT/505	Unbekannt	Unbekannt	1	332533803
332534157	Pseudoalteromonas haloplanktis ANT/505	Formylglycingenerierendes Enzym	Extrazellulär	1	332534155
332653876	Ruminococcaceae bacterium D16	Fibronectin type 3 domain	Extrazellulär	3	332653877
332654471	Ruminococcaceae bacterium D16	putative mucin-5AC	Unbekannt	3	332654472
332655366	Ruminococcaceae bacterium D16	putative mucin-5B	Extrazellulär	2	332655364
332666261	Haliscomenobacter hydrossis DSM 1100	Formylglycingenerierendes Enzym; AAA ATPase Domäne	Unbekannt (mehrere möglich)	1	332666264
332707564	Lyngbya majuscula 3L	Formylglycingenerierendes Enzym	Unbekannt	1	332707562
332707565	Lyngbya majuscula 3L	Unbekannt	Unbekannt	1	

GI-Nr. des Zielproteins	Organismus	Annotation	Vorhersage der subzellulären Lokalisation mit PSort	Phylogenetische Gruppe[1]	GI-Nr. der RT
336402101	Bacteroides sp. 1_1_30	CotH protein	Membran (extrazellulär)	1	336402102
336410997	Bacteroides sp. 2_1_56FAA	Unbekannt	Unbekannt	2	336410996
336413451	Bacteroides ovatus 3_8_47FAA	Concanavalin A-like lectin/glucanases superfamily; pfam13385	Membran (extrazellulär)	1	336413450
336433794	Lachnospiraceae bacterium 2_1_58FAA	Unbekannt	Zytosol	3	336433792
336435573	Lachnospiraceae bacterium 1_4_56FAA	Unbekannt	Unbekannt	3	336435571
336435574	Lachnospiraceae bacterium 1_4_56FAA	Unbekannt	Zytosol	3	
338999653	Halomonas sp. TD01	Unbekannt	Unbekannt	1	338999650
340788556	Collimonas fungivorans Ter331	Unbekannt	Unbekannt	3	340788559
340788557	Collimonas fungivorans Ter331	Unbekannt	Unbekannt	3	
341642638	Vibrio cholerae HE-09	Major Tropism-Determinante; PHA00653	Unbekannt	1	341642641
344341809	Thiocapsa marina 5811	Formylglycingenerierendes Enzym	Membran (extrazellulär)	1	344341810
344343879	Marichromatium purpuratum 984	Formylglycingenerierendes Enzym	Unbekannt (mehrere möglich)	1	344343882
345511692	Thiorhodococcus drewsii AZ1	Formylglycingenerierendes Enzym	Unbekannt	3	345871758
345871756	Desulfovibrio sp. 6_1_46AFAA	Unbekannt	Extrazellulär	1	345893474
345893468	Pseudogulbenkiania sp. NH8B	Unbekannt	Unbekannt	1	347538767
347538771	Pseudogulbenkiania sp. NH8B	Unbekannt	Unbekannt	3	
347538772	Megasphaera elsdenii DSM 20460	Unbekannt	Unbekannt	3	348026361
348026359	Thiorhodospira sibirica ATCC 700588	Protein of unknown function (DUF1566); pfam07603	Unbekannt (mehrere möglich)	1	350551816
350551814	Thiorhodospira sibirica ATCC 700588	Protein of unknown function (DUF1566); pfam07603	Unbekannt	3	
350551817	Clostridium sp. 7_3_54FAA	Formylglycingenerierendes Enzym	Unbekannt	3	355629964
355629966	Clostridium citroniae WAL-17108	Unbekannt	Membran (intrazellulär)	1	355681483
355681487	Bacteroides sp. D1	Unbekannt	Unbekannt	3	237716228
357052768	Bacteroides sp. D1	Unbekannt	Zytosol	3	
377661873	Clostridium clostridioforme 2_1_49FAA	Formylglycingenerierendes Enzym; Carboxypeptidase	Unbekannt	1	355386195
383111379	Bacteroides sp. D2	Concanavalin A-like lectin/glucanases superfamily; pfam13385	Membran (extrazellulär)	1	315923025
383112786	Bacteroides sp. D2	Unbekannt	Unbekannt	3	315918944
383122829	Bacteroides sp. 1_1_6	Unbekannt	Unbekannt	3	253568001
389579824	Desulfobacter postgatei 2ac9	Formylglycingenerierendes Enzym	Zytosol	1	355363092

GI-Nr. des Zielproteins	Organismus	Annotation	Vorhersage der subzellulären Lokalisation mit PSort	Phylogenetische Gruppe[1]	GI-Nr. der RT
389579827	Desulfobacter postgatei 2ac9	Formylglycingenerierendes Enzym	Unbekannt	1	355363092

[1] siehe Abbildung 14 in Abschnitt 3.1.2.6

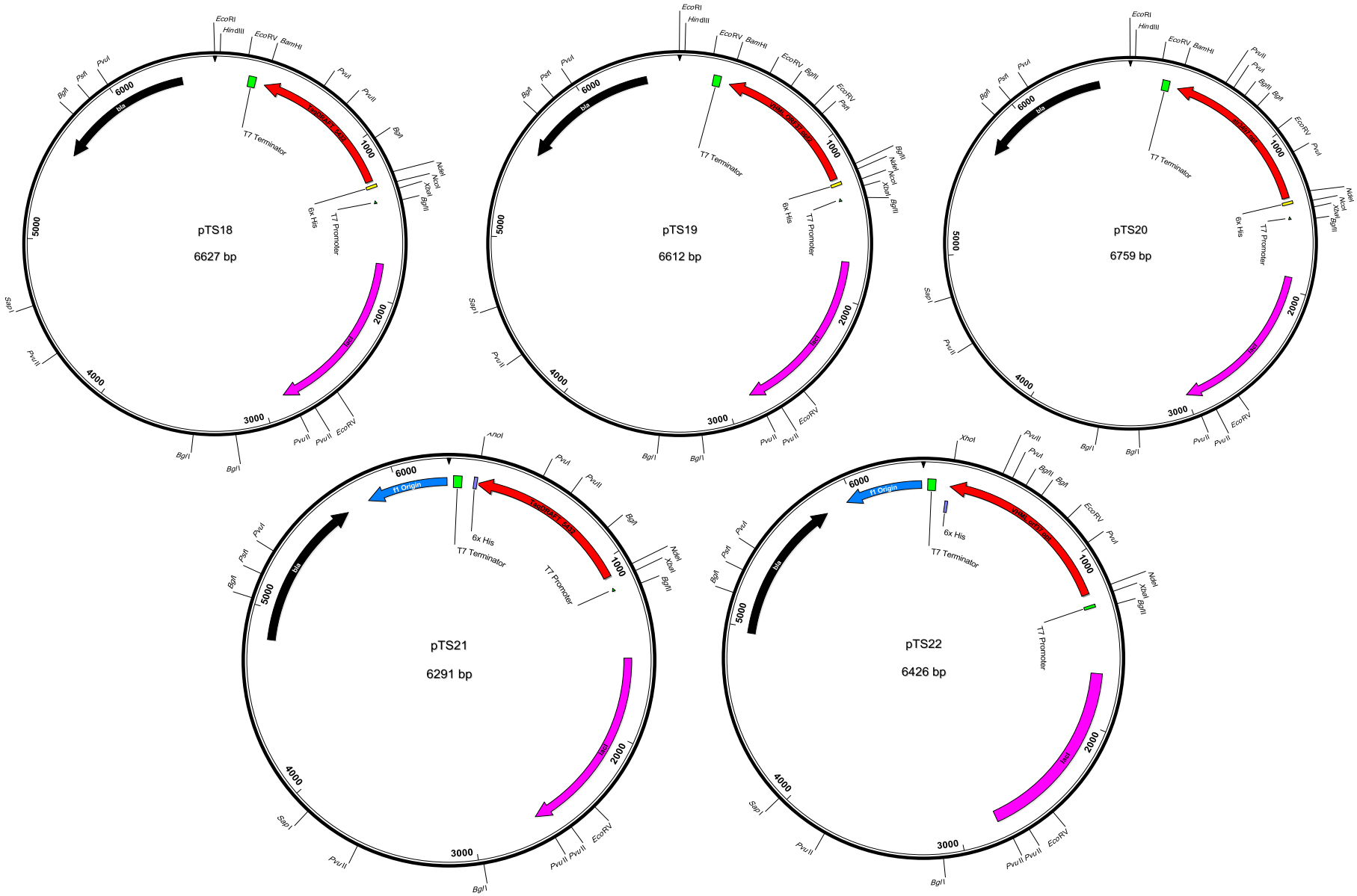
Tabelle A4: Verwendete Oligonucleotide dieser Arbeit

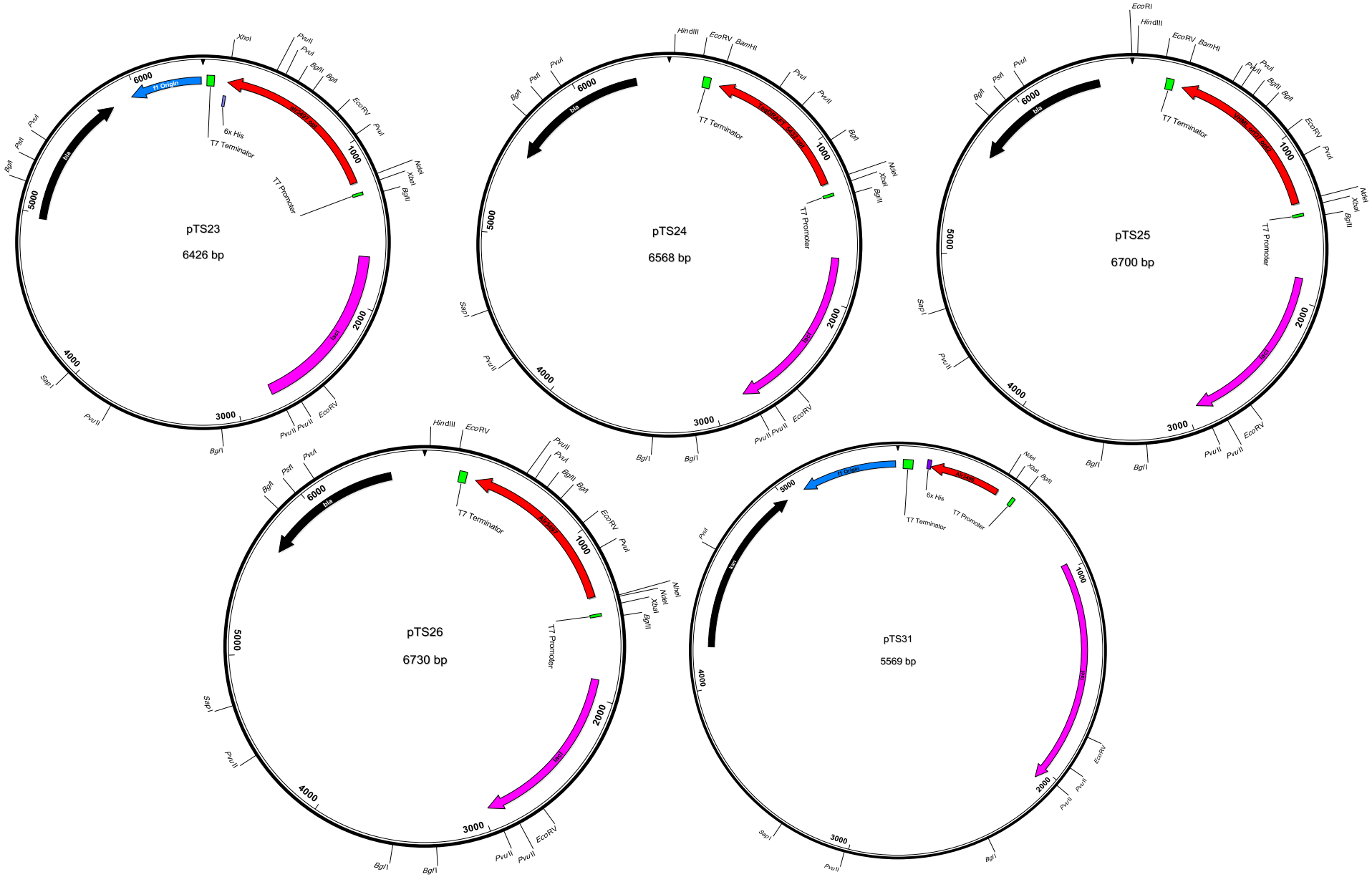
Oligo	Sequenz (5' -> 3')	Verwendung
TS1	CAGGCGCATATGAAAAGAAAAGGCAATCTATATCAC	Forward-Klonierungsprimer für TDE2266 in pTWIN1 über *Nde*I/*Sap*I
TS2	ACACACGCTCTTCTGCAAGTTAAAAGGTATTTGTTACAG	Reverse-Klonierungsprimer für TDE2266 in pTWIN1 über *Nde*I/*Sap*I
TS3	TAATACGACTCACTATAGGG	T7 Promoter-Primer
TS4	GATTGCCATGCCGGTCAAGG	Reverse-Sequenzierprimer für pTWIN1
TS9	AGCAGCGCTCTTCTAACATGAAAAGAAAAGGCAATC	Forward-Klonierungsprimer für TDE2266 in pTWIN1 über *Sap*I/*Pst*I
TS10	AGCAGCCTGCAGTCAAGTTAAAAGGTATTTGTTAC	Reverse-Klonierungsprimer für TDE2266 in pTWIN1 über *Sap*I/*Pst*I
TS11	AGCAGCGCTCTTCTAACATGATTTTTGACAGTTATTC	Forward-Klonierungsprimer für BT_2313 in pTWIN1 über *Sap*I/*Pst*I
TS12	AGCAGCCTGCAGTTATTTACTATTCCTGTCCC	Reverse-Klonierungsprimer für BT_2313 in pTWIN1 über *Sap*I/*Pst*I
TS13	AGCAGCCTCGAGAGTTAAAAGGTATTTGTTACAGTAAGCC	Reverse-Klonierungsprimer für TDE2266 in pET21a über *Nde*I/*Xho*I
TS14	AGCAGCCTCGAGTTTACTATTCCTGTCCCCGCTATC	Reverse-Klonierungsprimer für BT_2313 in pET21a über *Nde*I/*Xho*I
TS15	AGCAGCCCATGGGGATGAAAAGAAAAGGCAATC	Forward-Klonierungsprimer für TDE2266 in pETGEX_CT über *Nco*I/*Sac*I
TS16	AGCAGCGAGCTCAGTTAAAAGGTATTTGTTAC	Reverse-Klonierungsprimer für TDE2266 in pETGEX_CT über *Nco*I/*Sac*I
TS17	AGCAGCGAGCTCATGATTTTTGACAGTTATTC	Forward-Klonierungsprimer für BT_2313 in pETGEX_CT über *Sac*I/*Sac*I
TS18	AGCAGCGAGCTCTTTACTATTCCTGTCCC	Reverse-Klonierungsprimer für BT_2313 in pETGEX_CT über *Sac*I/*Sac*I
TS19	AGCAGCGGATCCATGAAAAGAAAAGGCAATC	Forward-Klonierungsprimer für TDE2266 in pGEX-6P-1 über *Bam*HI/*Xho*I

Oligo	Sequenz (5' -> 3')	Verwendung
TS20	AGCAGCCTCGAGTCAAGTTAAAAGGTATTTGTTAC	Reverse-Klonierungsprimer für TDE2266 in pGEX-6P-1 über *Bam*HI/*Xho*I
TS21	AGCAGCGGATCCATGATTTTTGACAGTTATTC	Forward-Klonierungsprimer für BT_2313 in pGEX-6P-1 über *Bam*HI/*Xho*I
TS22	AGCAGCCTCGAGTTATTTACTATTCCTGTCC	Reverse-Klonierungsprimer für BT_2313 in pGEX-6P-1 über *Bam*HI/*Xho*I
TS29	AGCAGCCTCGAGTCAAGTTAAAAGGTATTTGTTACAGTAAGCC	Reverse-Klonierungsprimer für TDE2266 in pET19b über *Nde*I/*Xho*I
TS30	AGCAGCCTCGAGTTATTTACTATTCCTGTCCCCGCTATC	Reverse-Klonierungsprimer für BT_2313 in pET19b über *Nde*I/*Xho*I
TS34	AGCAGCGGATCCTCAAGTTAAAAGGTATTTGTTACAGTAAGCC	ersetzt TS29 wegen Frameshift
TS41	AAGTCGAGTGGGTTGCACAAGG	Reverse-Sequenzierprimer für pETGEX_CT
TS42	ACTGGGACTCCATCGTTTCT	Forward-Sequenzierprimer für pTWIN1 bei N-terminaler Fusion
TS43	ATTAACCCTCACTAAAGGGA	T3 Promoter-Primer
TS44	GTAAAACGACGGCCAG	M13-Sequenzierprimer (Forward)
TS45	CAGGAAACAGCTATGAC	M13-Sequenzierprimer (Reverse)
TS46	ATTTAGGTGACACTATAG	SP6-Sequenzierprimer
TS49	ACATTACATTACATTTATTTCG	Annealt an pNZ004-RNA; Template/Primer-Komplex für RT-Aktivitätsassays
TS51	AACGGCAGCCATATGCGTTCACTGGGCTGC	Forward-Klonierungsprimer für Vibrio-Phage ORF37 in pET21a über *Nde*I/*Bam*HI
TS52	TAGCAGCCGGATCCACAGCTCAGATCGATGG	Reverse-Klonierungsprimer für Vibrio-Phage ORF37 in pET21a über *Nde*I/*Bam*HI
TS53	AACGGCAGCCATATGAAACGTTACGGCAACC	Forward-Klonierungsprimer für alr3497 in pET21a über *Nde*I/*Bam*HI

Oligo	Sequenz (5' -> 3')	Verwendung
TS54	TAGCAGCCGGATCCTTTACGACGGAAATGC	Reverse-Klonierungsprimer für alr3497 in pET21a über *Nde*I/*Bam*HI
TS55	AACGGCAGCCATATGGATCTGGAAGATAACC	Forward-Klonierungsprimer für TaqDRAFT_5432 in pET21a über *Nde*I/*Bam*HI
TS56	TAGCAGCCGGATCCGATTTCGCCAAGTTCGATAAC	Reverse-Klonierungsprimer für für TaqDRAFT_5432 in pET21a über *Nde*I/*Bam*HI
TS57	TAGCAGCCCTCGAGACAGCTCAGATCGATGG	Ersetzt TS52
TS58	TAGCAGCCCTCGAGTTTACGACGGAAATGC	Ersetzt TS54
TS59	TAGCAGCCCTCGAGGATTTCGCCAAGTTCGATAAC	Ersetzt TS56
TS68	AAGAAGGAGATATACATATGAAAGAATTATCGGTCATCCAAAAG	Forward-Klonierungsprimer für alr3496 in pET24b mit GeneArt Seamless Cloning & Assembly Kit
TS69	TGGTGGTGGTGCTCGAGTTTTTCTCTGTTTCTTTGATTTTTAATCC	Reverse-Klonierungsprimer für alr3496 in pET24b mit GeneArt Seamless Cloning & Assembly Kit
TS75	AATTCTAATACGACTCACTATAGGGAGCGCGACAACATCAACAACAATATTGGTTTTCGTGTTGTC TGCGCGTTCGGGAGTACTCTTCACCA	Forward-Klonierungsprimer für 3'-Ende der vermutlichen *Nostoc* sp. PCC 7120 TR-RNA in pBSΔT7
TS76	AGCTTGGTGAAGAGTACTCCCGAACGCGCAGACAACACGAAAACCAATATTGTTGTTGATGTTGTC GCGCTCCCTATAGTGAGTCGTATTAG	Reverse-Klonierungsprimer für 3'-Ende der vermutlichen *Nostoc* sp. PCC 7120 TR-RNA in pBSΔT7
TS77	TGAAGAGTACTCCCGAACGCGCAGACAACACGAAAACCAA	Vollständig komplementäres Antisense-Oligo für *Nostoc* sp. PCC 7120-IMH
TS78	TGAAAAGTCCTCCCGAACGCGCAGACAACACGAAAACCAA	Komplementäres Antisense-Oligo für *Nostoc* sp. PCC 7120-IMH mit Mismatches

Vektorkarten




pTS23
6426 bp
pTS24
6568 bp
pTS25
6700 bp
pTS26
6730 bp
pTS31
5569 bp
T7 Terminator
T7 Promoter
6x His
bla
lacI
f1 Origin
1000
2000
3000
4000
5000
6000

PERL-Script des Programms DiGReF – Diversity-Generating Retroelement Finder

Autor: Dipl.-Biol. Mohamed Lisfi, Abteilung für Genetik, TU Kaiserslautern

```
# DiGRef v1 - Program to find diversity generating retroelements - 26. April 2012
# Written by Mohamed Lisfi, Department of Genetics, University of Kaiserslautern,
# Postfach 3049, 67653 Kaiserslautern, Germany.
# Email contact: cullum@rhrk.uni-kl.de

use strict;
use warnings;


# bioperl must be installed for the following:
# for details of installing bioperl see: www.bioperl.org

use Bio::DB::GenBank ;
use Bio::SeqIO;
use Bio::Seq;
use Bio::Seq::RichSeq;
use Bio::Tools::SeqStats;

#*************************************************************************
#***     USING THE PROGRAM WITH DEFAULT PARAMETERS          ***
#
# The program needs a text input file GI.txt, which lists the GI numbers of
# GenBank protein sequences (i.e. RT sequences)
# The analysis of each sequence is output as its own text file with
# the name <GI number>.txt
# The RT protein entry is downloaded from NCBI and used to find the
# corresponding DNA sequence from the DBSOURCE field
# The output coordinates for the TR and VR are the coordinates in the
# DNA sequence
# If you want to produce a GenBank format entry that can be viewed in a
# sequence viewer program such as Artemis, you must run the accompanying
# program convertGB.pl
#
#*************************************************************************

#*************************************************************************
#***            CHANGING PARAMETERS               ***
#
# You must alter the source code as detailed below
```

```
#*************************************************************************
#***          CHANGING INPUT FILE NAME              ***
#
my $input = 'GI.txt'; # change GI.txt to required name

#*************************************************************************
#***          CHANGING MUTABLE BASE               ***
#
my $b = 'A'; # change A to C, G or T

#*************************************************************************
#***          CHANGING LENGTH OF REGION             ***
#***         TO BE SEARCHED FOR TR AND VR             ***
#
# default is 5000 bp up- and downstream of each RT
#
my $seqRTseq = 5000; # change to number of bp needed

#*************************************************************************
#***          CHANGING MINIMUM NUMBER              ***
#***           OF A-RESIDUES in TR               ***
#
# default is at least 10 A-residues
#
my $basenumb = 10; # change to required number

#*************************************************************************
#***          CHANGING MINIMUM NUMBER              ***
#***           OF SUBSTITUTIONS IN VR              ***
#
# default is at least 7 substitutions of A-residues
#
my $subs = 7; # change to required number

#*************************************************************************
#*************************************************************************



# the file contains a list of GI-numbers that will be investigated with the program
# give the inputfile
open (INFILE, "<$input") or die "Can't open input file, $!\n";

my @Inlines = <INFILE>; # all lines in the file

foreach my $GI_number (@Inlines) # each line contains one GI-number
{
chomp $GI_number;
```

```
# OUTFILEput files will be appointed by GI-numbers
# with text document (.txt) format
open (OUTFILE, ">$GI_number.txt");

  my $gi= "$GI_number"; # GI-number

  my $db = Bio::DB::GenBank->new;

my $seq_obj = $db->get_Seq_by_gi ($gi); # search by GI-number

  my $OUTFILEput_seq = Bio::SeqIO ->new ( -format => 'genbank'); # read from file

 # get a seqfeature somehow, eg, from a Sequence with Features attached
 # array of sub Sequence Features
  my @features = $seq_obj->get_SeqFeatures;

  foreach my $feature (@features) {

   next unless ($feature->primary_tag eq 'CDS'); # primary tag for a feature 'CDS'
   if ($feature->has_tag ('coded_by')) {
        (my $coded_by) = $feature->get_tag_values ('coded_by'); # value of the specified tag

     # for complement
     if ($coded_by =~m/complement\ ( (.+)\)/) {
     $1 =~m/ (.+):\W* (\d+)\.\.\W* (\d+)/;
     print OUTFILE "RT complement ($2..$3)\t";

     &comprt ($1,$2+2,$3-1,$seqRTseq);
     &comptrvr ($1,$2+2,$3-1,$subs,$basenumb,$b);
     }

     # for upstream
     else {
     $coded_by =~m/ (.+):\W* (\d+)\.\.\W* (\d+)/;
     print OUTFILE "RT $2..$3\t";

     &rt ($1,$2,$3,$subs,$basenumb,$seqRTseq,$b);
     }
    }
  }
}

# subroutin to get RT coordinates in complement
sub comprt {

  my ($dbsourcs,$start,$end,$seqRTseq) = @_;
  my $db = Bio::DB::GenBank->new;
  my $seq_obj = $db->get_Seq_by_acc ($dbsourcs);
  my $genome_seq = Bio::SeqIO ->new ( -format => 'fasta');
```

```
  my $genome = $seq_obj->seq;
  my $genomelen = $seq_obj->length;

  $genome=~s/\s//g;

  my $len = $end-$start+1;
  my $rtdna = substr ($genome,$start,$len);

  $rtdna =~tr/ATCG/TAGC/;
  $genome =~tr/ATCG/TAGC/;
  $genome = reverse ($genome);
  $rtdna = reverse ($rtdna);

  my $seq = "";
  my $start_seq;

  while ($genome=~m/$rtdna/gi)
  {
  $end = pos ($genome);
  $start = $end-$len+1;

    # sometimes the TR is localized at the extremity of complementary genome, in a position less than 5000 bases
    # in the loop, it is assumed that the position superior to 5000 bases,
    # if that it is true, $i is equal to 5000, and the start position of DNA sequence is equal to ($start - 5000)
    # if not, a number lower than 5000 will be sought, in such a way that the start position of DNA sequence is equal to 1
      for (my $i = $seqRTseq/2 ; $i<= $seqRTseq/2 ; $i--) {
      $start_seq = $start-$i;
          if ( 0 < $start_seq){
          $seq = substr ($genome,$start_seq,$len+$seqRTseq);
          last
         }
      }
}

  my $seqlen=length ($seq);

print OUTFILE "length of genome: $genomelen\n$rtdna\n\n$seq\n\n";
}

# subroutin to get TR and VR coordinates in complement
sub comptrvr {

  my ($dbsourcs, $end, $start, $subs, $basenumb, $b) = @_;
  my $db = Bio::DB::GenBank->new;
  my $seq_obj = $db->get_Seq_by_acc ($dbsourcs);
  my $genome_seq = Bio::SeqIO ->new ( -format => 'fasta');
  my $genome = $seq_obj->seq;

  $genome=~s/\s//g;
```

```
  my $len = $end-$start+1;
  my $rtdna= substr ($genome,$start,$len);

  $rtdna=~tr/ATCG/TAGC/;
  $genome=~tr/ATCG/TAGC/;
  $genome=reverse ($genome);
  $rtdna=reverse ($rtdna);

  my $seq = "";
  my $start_seq;

  while ($genome=~m/$rtdna/gi)
  {
  $end = pos ($genome);
  $start = $end-$len+1;

  # loop to find the exact coordinates of TR and VR as used in sub comprt
     for (my $i = $seqRTseq/2 ; $i<= $seqRTseq/2 ; $i--) {
        $start_seq = $start-$i;
          if ( 0 < $start_seq){
          $seq = substr ($genome,$start_seq,$len+$seqRTseq);
          last
         }
      }
}

  my $seqlen=length ($seq);

&findrepeat ($start, $seq, $seqlen, $genome, $subs, $basenumb, $b);
}

# subroutin to get coordinates of RT in upstream
sub rt {

  my ($dbsourcs, $start, $end, $subs, $basenumb, $seqRTseq, $b) = @_;

  my $db = Bio::DB::GenBank->new;
  my $seq_obj = $db->get_Seq_by_acc ($dbsourcs);
  my $genome_seq = Bio::SeqIO ->new ( -format => 'fasta');
  my $genome = $seq_obj->seq;
  my $genomelen = $seq_obj->length;

  $genome=~s/\s//g;

  my $len = $end-$start+1;
  my $start_seq;
  my $seq;

   # in some genomes the TR is localized at the beginning, in a position less than 5000 bases
   # in the loop, the correct position being sought, as well as in sub comprt
```

```
   for (my $i = $seqRTseq/2 ; $i<= $seqRTseq/2 ; $i--) {
      $start_seq = $start-$i;
          if ( 0 < $start_seq){
          $seq = substr ($genome,$start_seq,$len+$seqRTseq);
          last
         }
      }

  my $seqlen=length ($seq);
  my $rtdna= substr ($genome,$start-1,$len-3);
  print OUTFILE "length of genome: $genomelen\n$rtdna\n\n$seq\n\n";

&findrepeat ($start, $seq, $seqlen, $genome, $subs, $basenumb, $b);
}

# subroutin to find repeats
sub findrepeat
{
 my ($start, $seq, $seqlen, $genome, $subs, $basenumb, $b) = @_;
 my @TR_pos = "";
 my @VR_pos = "";

  # get repeat position; ATTENTION: Some of the TR position have two (or more) VRs, we have to seperate them
   for (my $n=1; $n <= $seqlen-100; $n++) {
   my $bp= substr ($seq,$n,50);
   $bp =~ s/$b/[ACGT]/gi;
    while ($seq =~m/$bp/gi) {
     my $bppos=pos ($seq)-50;

     if ($bppos == $n ){next}
     elsif (abs ($bppos-$n)<=50){next}
     else {
       my $start_seq;
       for (my $i = $seqRTseq/2 ; $i<= $seqRTseq/2 ; $i--){
           $start_seq = $start-$i;
           if ( 0 < $start_seq){
           my $trpos= $start-$i+$n;
           my $vrpos= $start-$i+$bppos;
               if (0 < $trpos){
               push (@TR_pos, $trpos);
               push (@VR_pos, $vrpos);
               }
           last
           }
       }
     }
     }
   }

my $scalarpos = scalar (@TR_pos);
```

```
# to give signals if there exist no TR and VR
if ($scalarpos <= 3){print OUTFILE "no TR and VR!\n";}

else {  # if there is repeat

# the programm can interpret up to 20 repeats
# the number that ends the name of each array, means the number of repeats
# each array contains corresponding position of TR and VR
my @TR_1 = "";
my @TR_2 = "";
my @TR_3 = "";
my @TR_4 = "";
my @TR_5 = "";
my @TR_6 = "";
my @TR_7 = "";
my @TR_8 = "";
my @TR_9 = "";
my @TR_10 = "";
my @TR_11 = "";
my @TR_12 = "";
my @TR_13 = "";
my @TR_14 = "";
my @TR_15 = "";
my @TR_16 = "";
my @TR_17 = "";
my @TR_18 = "";
my @TR_19 = "";
my @TR_20 = "";

my @VR_1 = "";
my @VR_2 = "";
my @VR_3 = "";
my @VR_4 = "";
my @VR_5 = "";
my @VR_6 = "";
my @VR_7 = "";
my @VR_8 = "";
my @VR_9 = "";
my @VR_10 = "";
my @VR_11 = "";
my @VR_12 = "";
my @VR_13 = "";
my @VR_14 = "";
my @VR_15 = "";
my @VR_16 = "";
my @VR_17 = "";
my @VR_18 = "";
my @VR_19 = "";
my @VR_20 = "";
```

```
 # to push positions of TRs and VRs in arrays
MYLOOP: for (my $n = 1; $n < $scalarpos-1; $n++) {
        for (my $m=1; $m<=20; $m++) { # $m: possible number of repeat
          # prerequisite to put positions in the correct array according to number of repeat
          # for first position
          if ($n==1) {
            if ($TR_pos[$n] != $TR_pos[$n+$m]) {
            for (my $k = 0; $k < $m; $k++) {
                if ($m==1){ push (@TR_1, $TR_pos[$n+$k]); push (@VR_1, $VR_pos[$n+$k]);}
                if ($m==2){ push (@TR_2, $TR_pos[$n+$k]); push (@VR_2, $VR_pos[$n+$k]);}
                if ($m==3){ push (@TR_3, $TR_pos[$n+$k]); push (@VR_3, $VR_pos[$n+$k]);}
                if ($m==4){ push (@TR_4, $TR_pos[$n+$k]); push (@VR_4, $VR_pos[$n+$k]);}
                if ($m==5){ push (@TR_5, $TR_pos[$n+$k]); push (@VR_5, $VR_pos[$n+$k]);}
                if ($m==6){ push (@TR_6, $TR_pos[$n+$k]); push (@VR_6, $VR_pos[$n+$k]);}
                if ($m==7){ push (@TR_7, $TR_pos[$n+$k]); push (@VR_7, $VR_pos[$n+$k]);}
                if ($m==8){ push (@TR_8, $TR_pos[$n+$k]); push (@VR_8, $VR_pos[$n+$k]);}
                if ($m==9){ push (@TR_9, $TR_pos[$n+$k]); push (@VR_9, $VR_pos[$n+$k]);}
                if ($m==10){ push (@TR_10, $TR_pos[$n+$k]); push (@VR_10, $VR_pos[$n+$k]);}
                if ($m==11){ push (@TR_11, $TR_pos[$n+$k]); push (@VR_11, $VR_pos[$n+$k]);}
                if ($m==12){ push (@TR_12, $TR_pos[$n+$k]); push (@VR_12, $VR_pos[$n+$k]);}
                if ($m==13){ push (@TR_13, $TR_pos[$n+$k]); push (@VR_13, $VR_pos[$n+$k]);}
                if ($m==14){ push (@TR_14, $TR_pos[$n+$k]); push (@VR_14, $VR_pos[$n+$k]);}
                if ($m==15){ push (@TR_15, $TR_pos[$n+$k]); push (@VR_15, $VR_pos[$n+$k]);}
                if ($m==16){ push (@TR_16, $TR_pos[$n+$k]); push (@VR_16, $VR_pos[$n+$k]);}
                if ($m==17){ push (@TR_17, $TR_pos[$n+$k]); push (@VR_17, $VR_pos[$n+$k]);}
                if ($m==18){ push (@TR_18, $TR_pos[$n+$k]); push (@VR_18, $VR_pos[$n+$k]);}
                if ($m==19){ push (@TR_19, $TR_pos[$n+$k]); push (@VR_19, $VR_pos[$n+$k]);}
                if ($m==20){ push (@TR_20, $TR_pos[$n+$k]); push (@VR_20, $VR_pos[$n+$k]);}
            }
            next MYLOOP
          }
        }

        # prerequisite to put positions in the correct array according to number of repeat
        # for next positions
        elsif (1 < $n) {
             unless ($TR_pos[$n] == $TR_pos[$n-1]) {
             unless ($TR_pos[$n] == $TR_pos[$n+$m]) {
              for (my $k = 0; $k<$m; $k++) {
                if ($m==1){ push (@TR_1, $TR_pos[$n+$k]); push (@VR_1, $VR_pos[$n+$k]);}
                if ($m==2){ push (@TR_2, $TR_pos[$n+$k]); push (@VR_2, $VR_pos[$n+$k]);}
                if ($m==3){ push (@TR_3, $TR_pos[$n+$k]); push (@VR_3, $VR_pos[$n+$k]);}
                if ($m==4){ push (@TR_4, $TR_pos[$n+$k]); push (@VR_4, $VR_pos[$n+$k]);}
                if ($m==5){ push (@TR_5, $TR_pos[$n+$k]); push (@VR_5, $VR_pos[$n+$k]);}
                if ($m==6){ push (@TR_6, $TR_pos[$n+$k]); push (@VR_6, $VR_pos[$n+$k]);}
                if ($m==7){ push (@TR_7, $TR_pos[$n+$k]); push (@VR_7, $VR_pos[$n+$k]);}
                if ($m==8){ push (@TR_8, $TR_pos[$n+$k]); push (@VR_8, $VR_pos[$n+$k]);}
                if ($m==9){ push (@TR_9, $TR_pos[$n+$k]); push (@VR_9, $VR_pos[$n+$k]);}
                if ($m==10){ push (@TR_10, $TR_pos[$n+$k]); push (@VR_10, $VR_pos[$n+$k]);}
```

```
                if ($m==11){ push (@TR_11, $TR_pos[$n+$k]); push (@VR_11, $VR_pos[$n+$k]);}
                if ($m==12){ push (@TR_12, $TR_pos[$n+$k]); push (@VR_12, $VR_pos[$n+$k]);}
                if ($m==13){ push (@TR_13, $TR_pos[$n+$k]); push (@VR_13, $VR_pos[$n+$k]);}
                if ($m==14){ push (@TR_14, $TR_pos[$n+$k]); push (@VR_14, $VR_pos[$n+$k]);}
                if ($m==15){ push (@TR_15, $TR_pos[$n+$k]); push (@VR_15, $VR_pos[$n+$k]);}
                if ($m==16){ push (@TR_16, $TR_pos[$n+$k]); push (@VR_16, $VR_pos[$n+$k]);}
                if ($m==17){ push (@TR_17, $TR_pos[$n+$k]); push (@VR_17, $VR_pos[$n+$k]);}
                if ($m==18){ push (@TR_18, $TR_pos[$n+$k]); push (@VR_18, $VR_pos[$n+$k]);}
                if ($m==19){ push (@TR_19, $TR_pos[$n+$k]); push (@VR_19, $VR_pos[$n+$k]);}
                if ($m==20){ push (@TR_20, $TR_pos[$n+$k]); push (@VR_20, $VR_pos[$n+$k]);}
              }
              next MYLOOP
            }
          }
        }
    }
}

# number of positions in each array
my $sca1 = scalar (@TR_1);
my $sca2 = scalar (@TR_2);
my $sca3 = scalar (@TR_3);
my $sca4 = scalar (@TR_4);
my $sca5 = scalar (@TR_5);
my $sca6 = scalar (@TR_6);
my $sca7 = scalar (@TR_7);
my $sca8 = scalar (@TR_8);
my $sca9 = scalar (@TR_9);
my $sca10 = scalar (@TR_10);
my $sca11 = scalar (@TR_11);
my $sca12 = scalar (@TR_12);
my $sca13 = scalar (@TR_13);
my $sca14 = scalar (@TR_14);
my $sca15 = scalar (@TR_15);
my $sca16 = scalar (@TR_16);
my $sca17 = scalar (@TR_17);
my $sca18 = scalar (@TR_18);
my $sca19 = scalar (@TR_19);
my $sca20 = scalar (@TR_20);

# arrays with the first position and the last position for successive positions
my @TR_nr1 = "";
my @TR_nr2 = "";
my @TR_nr3 = "";
my @TR_nr4 = "";
my @TR_nr5 = "";
my @TR_nr6 = "";
my @TR_nr7 = "";
my @TR_nr8 = "";
my @TR_nr9 = "";
```

```
my @TR_nr10 = "";
my @TR_nr11 = "";
my @TR_nr12 = "";
my @TR_nr13 = "";
my @TR_nr14 = "";
my @TR_nr15 = "";
my @TR_nr16 = "";
my @TR_nr17 = "";
my @TR_nr18 = "";
my @TR_nr19 = "";
my @TR_nr20 = "";

my @VR_nr1 = "";
my @VR_nr2 = "";
my @VR_nr3 = "";
my @VR_nr4 = "";
my @VR_nr5 = "";
my @VR_nr6 = "";
my @VR_nr7 = "";
my @VR_nr8 = "";
my @VR_nr9 = "";
my @VR_nr10 = "";
my @VR_nr11 = "";
my @VR_nr12 = "";
my @VR_nr13 = "";
my @VR_nr14 = "";
my @VR_nr15 = "";
my @VR_nr16 = "";
my @VR_nr17 = "";
my @VR_nr18 = "";
my @VR_nr19 = "";
my @VR_nr20 = "";

# to add only the first and last position for successive positions in the array
for (my $k=0; $k<1; $k++) {push (@TR_nr1, $TR_1[$k+1]); push (@VR_nr1, $VR_1[$k+1]); for (my $n=1; $n<scalar (@TR_1)-1; $n = $n+1) {my $i = $n + 1;
if ( ($TR_1[$n+$k]!= $TR_1[$i+$k]-1) || ($VR_1[$n+$k]!= $VR_1[$i+$k]-1)){push (@TR_nr1, $TR_1[$n+$k]);push (@TR_nr1, $TR_1[$i+$k]); push (@VR_nr1,
$VR_1[$n+$k]);push (@VR_nr1, $VR_1[$i+$k]);}}push (@TR_nr1, $TR_1[$k-1]);push (@VR_nr1, $VR_1[$k-1]);}
for (my $k=0; $k<2; $k++) {push (@TR_nr2, $TR_2[$k+1]); push (@VR_nr2, $VR_2[$k+1]); for (my $n=1; $n<scalar (@TR_2)-2; $n = $n+2) {my $i = $n + 2;
if ( ($TR_2[$n+$k]!= $TR_2[$i+$k]-1) || ($VR_2[$n+$k]!= $VR_2[$i+$k]-1)){push (@TR_nr2, $TR_2[$n+$k]);push (@TR_nr2, $TR_2[$i+$k]); push (@VR_nr2,
$VR_2[$n+$k]);push (@VR_nr2, $VR_2[$i+$k]);}}push (@TR_nr2, $TR_2[$k-2]);push (@VR_nr2, $VR_2[$k-2]);}
for (my $k=0; $k<3; $k++) {push (@TR_nr3, $TR_3[$k+1]); push (@VR_nr3, $VR_3[$k+1]); for (my $n=1; $n<scalar (@TR_3)-3; $n = $n+3) {my $i = $n + 3;
if ( ($TR_3[$n+$k]!= $TR_3[$i+$k]-1) || ($VR_3[$n+$k]!= $VR_3[$i+$k]-1)){push (@TR_nr3, $TR_3[$n+$k]);push (@TR_nr3, $TR_3[$i+$k]); push (@VR_nr3,
$VR_3[$n+$k]);push (@VR_nr3, $VR_3[$i+$k]);}}push (@TR_nr3, $TR_3[$k-3]);push (@VR_nr3, $VR_3[$k-3]);}
for (my $k=0; $k<4; $k++) {push (@TR_nr4, $TR_4[$k+1]); push (@VR_nr4, $VR_4[$k+1]); for (my $n=1; $n<scalar (@TR_4)-4; $n = $n+4) {my $i = $n + 4;
if ( ($TR_4[$n+$k]!= $TR_4[$i+$k]-1) || ($VR_4[$n+$k]!= $VR_4[$i+$k]-1)){push (@TR_nr4, $TR_4[$n+$k]);push (@TR_nr4, $TR_4[$i+$k]); push (@VR_nr4,
$VR_4[$n+$k]);push (@VR_nr4, $VR_4[$i+$k]);}}push (@TR_nr4, $TR_4[$k-4]);push (@VR_nr4, $VR_4[$k-4]);}
for (my $k=0; $k<5; $k++) {push (@TR_nr5, $TR_5[$k+1]); push (@VR_nr5, $VR_5[$k+1]); for (my $n=1; $n<scalar (@TR_5)-5; $n = $n+5) {my $i = $n + 5;
if ( ($TR_5[$n+$k]!= $TR_5[$i+$k]-1) || ($VR_5[$n+$k]!= $VR_5[$i+$k]-1)){push (@TR_nr5, $TR_5[$n+$k]);push (@TR_nr5, $TR_5[$i+$k]); push (@VR_nr5,
$VR_5[$n+$k]);push (@VR_nr5, $VR_5[$i+$k]);}}push (@TR_nr5, $TR_5[$k-5]);push (@VR_nr5, $VR_5[$k-5]);}
```

```
for (my $k=0; $k<6; $k++) {push (@TR_nr6, $TR_6[$k+1]); push (@VR_nr6, $VR_6[$k+1]); for (my $n=1; $n<scalar (@TR_6)-6; $n = $n+6) {my $i = $n + 6;
if ( ($TR_6[$n+$k]!= $TR_6[$i+$k]-1) || ($VR_6[$n+$k]!= $VR_6[$i+$k]-1)){push (@TR_nr6, $TR_6[$n+$k]);push (@TR_nr6, $TR_6[$i+$k]); push (@VR_nr6,
$VR_6[$n+$k]);push (@VR_nr6, $VR_6[$i+$k]);}}push (@TR_nr6, $TR_6[$k-6]);push (@VR_nr6, $VR_6[$k-6]);}
for (my $k=0; $k<7; $k++) {push (@TR_nr7, $TR_7[$k+1]); push (@VR_nr7, $VR_7[$k+1]); for (my $n=1; $n<scalar (@TR_7)-7; $n = $n+7) {my $i = $n + 7;
if ( ($TR_7[$n+$k]!= $TR_7[$i+$k]-1) || ($VR_7[$n+$k]!= $VR_7[$i+$k]-1)){push (@TR_nr7, $TR_7[$n+$k]);push (@TR_nr7, $TR_7[$i+$k]); push (@VR_nr7,
$VR_7[$n+$k]);push (@VR_nr7, $VR_7[$i+$k]);}}push (@TR_nr7, $TR_7[$k-7]);push (@VR_nr7, $VR_7[$k-7]);}
for (my $k=0; $k<8; $k++) {push (@TR_nr8, $TR_8[$k+1]); push (@VR_nr8, $VR_8[$k+1]); for (my $n=1; $n<scalar (@TR_8)-8; $n = $n+8) {my $i = $n + 8;
if ( ($TR_8[$n+$k]!= $TR_8[$i+$k]-1) || ($VR_8[$n+$k]!= $VR_8[$i+$k]-1)){push (@TR_nr8, $TR_8[$n+$k]);push (@TR_nr8, $TR_8[$i+$k]); push (@VR_nr8,
$VR_8[$n+$k]);push (@VR_nr8, $VR_8[$i+$k]);}}push (@TR_nr8, $TR_8[$k-8]);push (@VR_nr8, $VR_8[$k-8]);}
for (my $k=0; $k<9; $k++) {push (@TR_nr9, $TR_9[$k+1]); push (@VR_nr9, $VR_9[$k+1]); for (my $n=1; $n<scalar (@TR_9)-9; $n = $n+9) {my $i = $n + 9;
if ( ($TR_9[$n+$k]!= $TR_9[$i+$k]-1) || ($VR_9[$n+$k]!= $VR_9[$i+$k]-1)){push (@TR_nr9, $TR_9[$n+$k]);push (@TR_nr9, $TR_9[$i+$k]); push (@VR_nr9,
$VR_9[$n+$k]);push (@VR_nr9, $VR_9[$i+$k]);}}push (@TR_nr9, $TR_9[$k-9]);push (@VR_nr9, $VR_9[$k-9]);}
for (my $k=0; $k<10; $k++) {push (@TR_nr10, $TR_10[$k+1]); push (@VR_nr10, $VR_10[$k+1]); for (my $n=1; $n<scalar (@TR_10)-10; $n = $n+10) {my $i =
$n + 10; if ( ($TR_10[$n+$k]!= $TR_10[$i+$k]-1) || ($VR_10[$n+$k]!= $VR_10[$i+$k]-1)){push (@TR_nr10, $TR_10[$n+$k]);push (@TR_nr10, $TR_10[$i+$k]);
push (@VR_nr10, $VR_10[$n+$k]);push (@VR_nr10, $VR_10[$i+$k]);}}push (@TR_nr10, $TR_10[$k-10]);push (@VR_nr10, $VR_10[$k-10]);}
for (my $k=0; $k<11; $k++) {push (@TR_nr11, $TR_11[$k+1]); push (@VR_nr11, $VR_11[$k+1]); for (my $n=1; $n<scalar (@TR_11)-11; $n = $n+11) {my $i =
$n + 11; if ( ($TR_11[$n+$k]!= $TR_11[$i+$k]-1) || ($VR_11[$n+$k]!= $VR_11[$i+$k]-1)){push (@TR_nr11, $TR_11[$n+$k]);push (@TR_nr11, $TR_11[$i+$k]);
push (@VR_nr11, $VR_11[$n+$k]);push (@VR_nr11, $VR_11[$i+$k]);}}push (@TR_nr11, $TR_11[$k-11]);push (@VR_nr11, $VR_11[$k-11]);}
for (my $k=0; $k<12; $k++) {push (@TR_nr12, $TR_12[$k+1]); push (@VR_nr12, $VR_12[$k+1]); for (my $n=1; $n<scalar (@TR_12)-12; $n = $n+12) {my $i =
$n + 12; if ( ($TR_12[$n+$k]!= $TR_12[$i+$k]-1) || ($VR_12[$n+$k]!= $VR_12[$i+$k]-1)){push (@TR_nr12, $TR_12[$n+$k]);push (@TR_nr12, $TR_12[$i+$k]);
push (@VR_nr12, $VR_12[$n+$k]);push (@VR_nr12, $VR_12[$i+$k]);}}push (@TR_nr12, $TR_12[$k-12]);push (@VR_nr12, $VR_12[$k-12]);}
for (my $k=0; $k<13; $k++) {push (@TR_nr13, $TR_13[$k+1]); push (@VR_nr13, $VR_13[$k+1]); for (my $n=1; $n<scalar (@TR_13)-13; $n = $n+13) {my $i =
$n + 13; if ( ($TR_13[$n+$k]!= $TR_13[$i+$k]-1) || ($VR_13[$n+$k]!= $VR_13[$i+$k]-1)){push (@TR_nr13, $TR_13[$n+$k]);push (@TR_nr13, $TR_13[$i+$k]);
push (@VR_nr13, $VR_13[$n+$k]);push (@VR_nr13, $VR_13[$i+$k]);}}push (@TR_nr13, $TR_13[$k-13]);push (@VR_nr13, $VR_13[$k-13]);}
for (my $k=0; $k<14; $k++) {push (@TR_nr14, $TR_14[$k+1]); push (@VR_nr14, $VR_14[$k+1]); for (my $n=1; $n<scalar (@TR_14)-14; $n = $n+14) {my $i =
$n + 14; if ( ($TR_14[$n+$k]!= $TR_14[$i+$k]-1) || ($VR_14[$n+$k]!= $VR_14[$i+$k]-1)){push (@TR_nr14, $TR_14[$n+$k]);push (@TR_nr14, $TR_14[$i+$k]);
push (@VR_nr14, $VR_14[$n+$k]);push (@VR_nr14, $VR_14[$i+$k]);}}push (@TR_nr14, $TR_14[$k-14]);push (@VR_nr14, $VR_14[$k-14]);}
for (my $k=0; $k<15; $k++) {push (@TR_nr15, $TR_15[$k+1]); push (@VR_nr15, $VR_15[$k+1]); for (my $n=1; $n<scalar (@TR_15)-15; $n = $n+15) {my $i =
$n + 15; if ( ($TR_15[$n+$k]!= $TR_15[$i+$k]-1) || ($VR_15[$n+$k]!= $VR_15[$i+$k]-1)){push (@TR_nr15, $TR_15[$n+$k]);push (@TR_nr15, $TR_15[$i+$k]);
push (@VR_nr15, $VR_15[$n+$k]);push (@VR_nr15, $VR_15[$i+$k]);}}push (@TR_nr15, $TR_15[$k-15]);push (@VR_nr15, $VR_15[$k-15]);}
for (my $k=0; $k<16; $k++) {push (@TR_nr16, $TR_16[$k+1]); push (@VR_nr16, $VR_16[$k+1]); for (my $n=1; $n<scalar (@TR_16)-16; $n = $n+16) {my $i =
$n + 16; if ( ($TR_16[$n+$k]!= $TR_16[$i+$k]-1) || ($VR_16[$n+$k]!= $VR_16[$i+$k]-1)){push (@TR_nr16, $TR_16[$n+$k]);push (@TR_nr16, $TR_16[$i+$k]);
push (@VR_nr16, $VR_16[$n+$k]);push (@VR_nr16, $VR_16[$i+$k]);}}push (@TR_nr16, $TR_16[$k-16]);push (@VR_nr16, $VR_16[$k-16]);}
for (my $k=0; $k<17; $k++) {push (@TR_nr17, $TR_17[$k+1]); push (@VR_nr17, $VR_17[$k+1]); for (my $n=1; $n<scalar (@TR_17)-17; $n = $n+17) {my $i =
$n + 17; if ( ($TR_17[$n+$k]!= $TR_17[$i+$k]-1) || ($VR_17[$n+$k]!= $VR_17[$i+$k]-1)){push (@TR_nr17, $TR_17[$n+$k]);push (@TR_nr17, $TR_17[$i+$k]);
push (@VR_nr17, $VR_17[$n+$k]);push (@VR_nr17, $VR_17[$i+$k]);}}push (@TR_nr17, $TR_17[$k-17]);push (@VR_nr17, $VR_17[$k-17]);}
for (my $k=0; $k<18; $k++) {push (@TR_nr18, $TR_18[$k+1]); push (@VR_nr18, $VR_18[$k+1]); for (my $n=1; $n<scalar (@TR_18)-18; $n = $n+18) {my $i =
$n + 18; if ( ($TR_18[$n+$k]!= $TR_18[$i+$k]-1) || ($VR_18[$n+$k]!= $VR_18[$i+$k]-1)){push (@TR_nr18, $TR_18[$n+$k]);push (@TR_nr18, $TR_18[$i+$k]);
push (@VR_nr18, $VR_18[$n+$k]);push (@VR_nr18, $VR_18[$i+$k]);}}push (@TR_nr18, $TR_18[$k-18]);push (@VR_nr18, $VR_18[$k-18]);}
for (my $k=0; $k<19; $k++) {push (@TR_nr19, $TR_19[$k+1]); push (@VR_nr19, $VR_19[$k+1]); for (my $n=1; $n<scalar (@TR_19)-19; $n = $n+19) {my $i =
$n + 19; if ( ($TR_19[$n+$k]!= $TR_19[$i+$k]-1) || ($VR_19[$n+$k]!= $VR_19[$i+$k]-1)){push (@TR_nr19, $TR_19[$n+$k]);push (@TR_nr19, $TR_19[$i+$k]);
push (@VR_nr19, $VR_19[$n+$k]);push (@VR_nr19, $VR_19[$i+$k]);}}push (@TR_nr19, $TR_19[$k-19]);push (@VR_nr19, $VR_19[$k-19]);}
for (my $k=0; $k<20; $k++) {push (@TR_nr20, $TR_20[$k+1]); push (@VR_nr20, $VR_20[$k+1]); for (my $n=1; $n<scalar (@TR_20)-20; $n = $n+20) {my $i =
$n + 20; if ( ($TR_20[$n+$k]!= $TR_20[$i+$k]-1) || ($VR_20[$n+$k]!= $VR_20[$i+$k]-1)){push (@TR_nr20, $TR_20[$n+$k]);push (@TR_nr20, $TR_20[$i+$k]);
push (@VR_nr20, $VR_20[$n+$k]);push (@VR_nr20, $VR_20[$i+$k]);}}push (@TR_nr20, $TR_20[$k-20]);push (@VR_nr20, $VR_20[$k-20]);}

# two arrays (one for TR and one for VR) with the first position and the last positions for all repeats
my @all_TR_nr = "";
my @all_VR_nr = "";
```

```
# to add all first and last positions for successive positions of all repeats in one array
# positions for TR in @all_TR_nr and positions for VR in @all_VR_nr
if (3 <= $sca1) { shift @TR_nr1; push (@all_TR_nr , @TR_nr1); shift @VR_nr1; push (@all_VR_nr , @VR_nr1); }
if (3 <= $sca2) { shift @TR_nr2; push (@all_TR_nr , @TR_nr2); shift @VR_nr2; push (@all_VR_nr , @VR_nr2); }
if (3 <= $sca3) { shift @TR_nr3; push (@all_TR_nr , @TR_nr3); shift @VR_nr3; push (@all_VR_nr , @VR_nr3); }
if (3 <= $sca4) { shift @TR_nr4; push (@all_TR_nr , @TR_nr4); shift @VR_nr4; push (@all_VR_nr , @VR_nr4); }
if (3 <= $sca5) { shift @TR_nr5; push (@all_TR_nr , @TR_nr5); shift @VR_nr5; push (@all_VR_nr , @VR_nr5); }
if (3 <= $sca6) { shift @TR_nr6; push (@all_TR_nr , @TR_nr6); shift @VR_nr6; push (@all_VR_nr , @VR_nr6); }
if (3 <= $sca7) { shift @TR_nr7; push (@all_TR_nr , @TR_nr7); shift @VR_nr7; push (@all_VR_nr , @VR_nr7); }
if (3 <= $sca8) { shift @TR_nr8; push (@all_TR_nr , @TR_nr8); shift @VR_nr8; push (@all_VR_nr , @VR_nr8); }
if (3 <= $sca9) { shift @TR_nr9; push (@all_TR_nr , @TR_nr9); shift @VR_nr9; push (@all_VR_nr , @VR_nr9); }
if (3 <= $sca10) { shift @TR_nr10; push (@all_TR_nr , @TR_nr10); shift @VR_nr10; push (@all_VR_nr , @VR_nr10); }
if (3 <= $sca11) { shift @TR_nr11; push (@all_TR_nr , @TR_nr11); shift @VR_nr11; push (@all_VR_nr , @VR_nr11); }
if (3 <= $sca12) { shift @TR_nr12; push (@all_TR_nr , @TR_nr12); shift @VR_nr12; push (@all_VR_nr , @VR_nr12); }
if (3 <= $sca13) { shift @TR_nr13; push (@all_TR_nr , @TR_nr13); shift @VR_nr13; push (@all_VR_nr , @VR_nr13); }
if (3 <= $sca14) { shift @TR_nr14; push (@all_TR_nr , @TR_nr14); shift @VR_nr14; push (@all_VR_nr , @VR_nr14); }
if (3 <= $sca15) { shift @TR_nr15; push (@all_TR_nr , @TR_nr15); shift @VR_nr15; push (@all_VR_nr , @VR_nr15); }
if (3 <= $sca16) { shift @TR_nr16; push (@all_TR_nr , @TR_nr16); shift @VR_nr16; push (@all_VR_nr , @VR_nr16); }
if (3 <= $sca17) { shift @TR_nr17; push (@all_TR_nr , @TR_nr17); shift @VR_nr17; push (@all_VR_nr , @VR_nr17); }
if (3 <= $sca18) { shift @TR_nr18; push (@all_TR_nr , @TR_nr18); shift @VR_nr18; push (@all_VR_nr , @VR_nr18); }
if (3 <= $sca19) { shift @TR_nr19; push (@all_TR_nr , @TR_nr19); shift @VR_nr19; push (@all_VR_nr , @VR_nr19); }
if (3 <= $sca20) { shift @TR_nr20; push (@all_TR_nr , @TR_nr20); shift @VR_nr20; push (@all_VR_nr , @VR_nr20); }

shift @all_TR_nr;
shift @all_VR_nr;

my $n;
# after the assembly of all positions for all repeats, the positions are scattered
# in this loop the ruptured positions will be connected with each other
MYLOOP2: for ($n=0; $n< @all_TR_nr;$n = $n+2) {
     for (my $i=0; $i<@all_TR_nr;$i= $i+2) {
       if ($all_TR_nr[$n+1] == $all_TR_nr[$i]-1) {
        if ($all_VR_nr[$n+1] == $all_VR_nr[$i]-1){
        splice (@all_TR_nr, $i, 1, $all_TR_nr[$n]); splice (@all_TR_nr, $n, 2);
        splice (@all_VR_nr, $i, 1, $all_VR_nr[$n]); splice (@all_VR_nr, $n, 2);
        $n=0;
        next MYLOOP2
        }
       }

       elsif ( ($all_TR_nr[$i+1] == $all_TR_nr[$n]-1)) {
          if ($all_VR_nr[$i+1] == $all_VR_nr[$n]-1) {
          splice (@all_TR_nr, $i+1, 1, $all_TR_nr[$n+1]); splice (@all_TR_nr, $n, 2);
          splice (@all_VR_nr, $i+1, 1, $all_VR_nr[$n+1]); splice (@all_VR_nr, $n, 2);
          $n=0;
          next MYLOOP2
          }
       }
     }
}
```

```
# number of elements in the array with final positions
my $scaTRnr =scalar (@all_TR_nr);

 for (my $n=1; $n<= $scaTRnr/2; $n++) {

        # the exact length of the repeat
        my $repeatlen = $all_TR_nr[1] - $all_TR_nr[0]+50;
        my $allTRnr50 = $all_TR_nr[1]+49;
        my $allVRnr50 = $all_VR_nr[1]+49;

        my $TRseq = substr ($genome, $all_TR_nr[0],$repeatlen);
        my $VRseq = substr ($genome, $all_VR_nr[0], $repeatlen);
        my $TR_postoRT = $all_TR_nr[0] - $start;
        my $VR_postoRT = $all_VR_nr[0] - $start;

    ## to calculate number of each base in DNA sequence
    ## in order to calculate number of subsituted base in VR

    # for TR
      my $TRseqobj = Bio::PrimarySeq->new (-seq   => $TRseq,
                    -alphabet => 'dna',
                    -id    => 'test');
      my $TRseq_stats = Bio::Tools::SeqStats->new (-seq => $TRseqobj);

     # obtain a hash of counts of each type of monomer
     # (i.e. amino or nucleic acid)
     $TRseq_stats = Bio::Tools::SeqStats->new (-seq =>$TRseqobj);
     my $TR_hash_ref = $TRseq_stats->count_monomers (); # e.g. for DNA sequence

    # for VR
      my $VRseqobj = Bio::PrimarySeq->new (-seq   => $VRseq,
                    -alphabet => 'dna',
                    -id    => 'test');
      my $VRseq_stats = Bio::Tools::SeqStats->new (-seq => $VRseqobj);

     # obtain a hash of counts of each type of monomer
     # (i.e. amino or nucleic acid)
     $VRseq_stats = Bio::Tools::SeqStats->new (-seq =>$VRseqobj);
     my $VR_hash_ref = $VRseq_stats->count_monomers (); # e.g. for DNA sequence

    # number of bases (A, C, G, and T) in TR and VR
    # N if genome contains nucleotides marked as N (non normal bases A, C, G, and T)
    my $TR_A = 0;
    my $TR_C = 0;
    my $TR_G = 0;
    my $TR_T = 0;
    my $TR_N = 0;
    my $VR_A = 0;
    my $VR_C = 0;
```

```
    my $VR_G = 0;
    my $VR_T = 0;
    my $VR_N = 0;
    my $TR_b = 0;
    my $VR_b = 0;

    # number of substituted bases A (C, G or T) in VR
    my $b_subs;

    foreach my $base (sort keys %$TR_hash_ref) {
    if ($base eq $b){$b_subs = ($TR_hash_ref->{$base}) - ($VR_hash_ref->{$base}); $TR_b = $TR_hash_ref->{$base}; $VR_b = $VR_hash_ref->{$base};}
#    if ($base eq 'A'){$A_subs = ($TR_hash_ref->{$base}) - ($VR_hash_ref->{$base}); $TR_A = $TR_hash_ref->{$base}; $VR_A = $VR_hash_ref->{$base};}
    if ($base eq 'A'){$TR_A = $TR_hash_ref->{$base}; $VR_A = $VR_hash_ref->{$base};}
    if ($base eq 'C'){$TR_C = $TR_hash_ref->{$base}; $VR_C = $VR_hash_ref->{$base};}
    if ($base eq 'G'){$TR_G = $TR_hash_ref->{$base}; $VR_G = $VR_hash_ref->{$base};}
    if ($base eq 'T'){$TR_T = $TR_hash_ref->{$base}; $VR_T = $VR_hash_ref->{$base};}
    if ($base eq 'N'){$TR_N = $TR_hash_ref->{$base}; $VR_N = $VR_hash_ref->{$base};}
    }

    if ($basenumb < $TR_b){
      if ($subs <= $b_subs){
       if ($TR_N == 0){
       print OUTFILE ">TR$n/$all_TR_nr[0]--$allTRnr50/\n$TRseq\t\tA: $TR_A\tC: $TR_C\tG: $TR_G\tT: $TR_T\n>VR$n/$all_VR_nr[0]--
$allVRnr50/\n$VRseq\t\tA: $VR_A\tC: $VR_C\tG: $VR_G\tT: $VR_T\nnumber of substituted base $b: $b_subs\nposition of RT relative to TR is
$TR_postoRT\nposition of RT relative to VR is $VR_postoRT\n\n";
       }
      }
    }

    shift (@all_TR_nr);
    shift (@all_TR_nr);
    shift (@all_VR_nr);
    shift (@all_VR_nr);

 }
}
}
```

PERL-Script für das Zusatzprogramm *output_artemis.pl* zur Erzeugung von DGR-Darstellungen im Artemis-Format

Autor: Dipl.-Biol. Mohamed Lisfi, Abteilung für Genetik, TU Kaiserslautern

```
use strict;

use warnings;


my @RT_list = (

291522399,

338762709

);


for (my $i=0; $i<@RT_list; $i++){

#foreach my $name (@RT_list){


open (OUT,">artemis_$RT_list[$i].txt#") or die "Opening output file failed: $!\n";

open (INFILE,"$RT_list[$i].txt") or die "Opening input file failed: $!\n";

#print OUT "$RT_list[$i]\t";


#print OUT "$name\t";

#print OUT "\n";

my @Inlines =<INFILE>;

my $start;

my $end;

my $RT_start;

my $RT_end;

my $TR_start;

my $VR_start;

my $TR_end;

my $VR_end;

my $len;

my $RT_comp_Start;
```

```
my $TR_comp_Start;

my $VR_comp_Start;

my $TR_len;

my $VR_len;

my $RTpos;

my $TRpos;

my $VRpos;

#for (my $i=1; $i<@Inlines; $i++)

#{

 if ($Inlines[0] =~m/^RT (\d+)\.\. (\d+)\s*\w*\s*\w*\s*\w*\W*\s* (\d*)/)

 {

 for (my $m=5000 ; 0<=$m; $m--){


# $start= $1-5000;

 $start= $1-$m;

 if ( 0 <=$start ){

 $RT_start = $1-$start+1;

 $RT_end = $2-$start+3;


  printf OUT ("%-5s", "FT");

  printf OUT ("%-16s", "CDS");

  print OUT "$RT_start..$RT_end";

  print OUT "\n";

   printf OUT ("%-21s", "FT");

  print OUT "/note=\"RT\"";

  print OUT "\n";

  printf OUT ("%-21s", "FT");

  print OUT "/color=9";

  print OUT "\n";


  printf OUT ("%-5s", "FT");

  printf OUT ("%-16s", "RT");

  print OUT "$RT_start..$RT_end";

  print OUT "\n";
```

```
 printf OUT ("%-21s", "FT");
 print OUT "/note=\"RT\"";
 print OUT "\n";
 printf OUT ("%-21s", "FT");
 print OUT "/color=9";
 print OUT "\n";
 for (my $i=1; $i<@Inlines; $i++)
{
 if ($Inlines[$i] =~m/^>TR\d*\W{1} (\d+)-- (\d+)\W{1}\s*/)
 {

$TR_start = $1-$start+1;
$TR_end = $2-$start+3;

 printf OUT ("%-5s", "FT");
 printf OUT ("%-16s", "TR");
 print OUT "$TR_start..$TR_end";
 print OUT "\n";
 printf OUT ("%-21s", "FT");
 print OUT "/note=\"TR\"";
 print OUT "\n";
 printf OUT ("%-21s", "FT");
 print OUT "/color=3";
 print OUT "\n";
 }

 elsif ($Inlines[$i] =~m/^>VR\d*\W{1} (\d+)-- (\d+)\W{1}\s*/)
 {
 $VR_start = $1-$start+1;
$VR_end = $2-$start+3;

printf OUT ("%-5s", "FT");
 printf OUT ("%-16s", "VR");
 print OUT "$VR_start..$VR_end";
```

```
  print OUT "\n";
  printf OUT ("%-21s", "FT");
 print OUT "/note=\"VR\"";
 print OUT "\n";
 printf OUT ("%-21s", "FT");
 print OUT "/color=2";
 print OUT "\n";
 }
 }
 last
 }}}

 elsif ($Inlines[0] =~m/^RT complement\W* (\d+)\.\. (\d+)\W*\s*\w*\s*\w*\s*\w*\s*\w*\W*\s* (\d*)/)
{
$RTpos =$1;
 for (my $m=5000 ; 0 <=$m; $m--){
$RT_comp_Start = $3-$1;
# $start= $1-5000;
$start= $RT_comp_Start-$m;
$len = $2-$1;
$RT_start = $RT_comp_Start-$start;
$RT_end = $m+$len-3;

if ( 0 <=$start ){

#  my $stt = $3-$2;
#  my $endd = $3-$1-3;
#last}
  printf OUT ("%-5s", "FT");
  printf OUT ("%-16s", "CDS");
  print OUT "$RT_start..$RT_end";
  print OUT "\n";
   printf OUT ("%-21s", "FT");
  print OUT "/note=\"RT\"";
```

```
 print OUT "\n";
 printf OUT ("%-21s", "FT");
 print OUT "/color=9";
 print OUT "\n";

 printf OUT ("%-5s", "FT");
 printf OUT ("%-16s", "RT");
 print OUT "$RT_start..$RT_end";
 print OUT "\n";
  printf OUT ("%-21s", "FT");
 print OUT "/note=\"RT\"";
 print OUT "\n";
 printf OUT ("%-21s", "FT");
 print OUT "/color=9";
 print OUT "\n";

 for (my $i=0; $i<@Inlines; $i++)
{
 if ($Inlines[$i] =~m/^>TR\d*\W{1} (\d+)-- (\d+)\W{1}\s*/)
 {
 $TRpos =$1;
 if ($RT_comp_Start < $TRpos)
 {
 $TR_comp_Start =  $1-$RT_comp_Start;

 $TR_start =$TR_comp_Start + $m + $len;
 $TR_len = $2-$1;
 $TR_end = $TR_start +$TR_len;

  printf OUT ("%-5s", "FT");
 printf OUT ("%-16s", "TR");
 print OUT "$TR_start.. $TR_end";
 print OUT "\n";
 printf OUT ("%-21s", "FT");
```

```
 print OUT "/note=\"TR\"";
 print OUT "\n";
 printf OUT ("%-21s", "FT");
 print OUT "/color=3";
 print OUT "\n";
}
elsif ($TRpos < $RT_comp_Start)
{
    $TR_comp_Start =  $1-$RT_comp_Start ;
    $TR_start =$TR_comp_Start + $m + $len;
    $TR_len = $2-$1;
    $TR_end = $TR_start +$TR_len;

 printf OUT ("%-5s", "FT");
 printf OUT ("%-16s", "TR");
 print OUT "$TR_start.. $TR_end";
 print OUT "\n";
 printf OUT ("%-21s", "FT");
 print OUT "/note=\"TR\"";
 print OUT "\n";
 printf OUT ("%-21s", "FT");
 print OUT "/color=3";
 print OUT "\n";
 }
 }

 elsif ($Inlines[$i] =~m/^>VR\d*\W{1} (\d+)-- (\d+)\W{1}\s*/)
 {
 if ($RT_comp_Start < $TRpos)
 {
 $VR_comp_Start =  $1-$RT_comp_Start;

 $VR_start =$VR_comp_Start + $m + $len;
 $VR_len = $2-$1;
```

```
 $VR_end = $VR_start +$VR_len;


 printf OUT ("%-5s", "FT");
  printf OUT ("%-16s", "VR");
  print OUT "$VR_start..$VR_end";
    print OUT "\n";
    printf OUT ("%-21s", "FT");
   print OUT "/note=\"VR\"";
  print OUT "\n";
  printf OUT ("%-21s", "FT");
  print OUT "/color=2";
  print OUT "\n";
}
elsif ($TRpos < $RT_comp_Start)
{
    $VR_comp_Start =  $1-$RT_comp_Start ;
     $VR_start =$VR_comp_Start + $m + $len;
     $VR_len = $2-$1;
     $VR_end = $VR_start +$VR_len;


  printf OUT ("%-5s", "FT");
  printf OUT ("%-16s", "VR");
  print OUT "$VR_start..$VR_end";
  print OUT "\n";
  printf OUT ("%-21s", "FT");
   print OUT "/note=\"VR\"";
  print OUT "\n";
  printf OUT ("%-21s", "FT");
  print OUT "/color=2";
  print OUT "\n";
  }
  }

}last
```

```
}}}


 print OUT ">$RT_list[$i]\n";

# name\n";

  print OUT "$Inlines[3]";

  next

}
```

PERL-Script für das Zusatzprogramm *output_graph.pl* zur Erzeugung alignter TR/VR-Sequenzen

Autor: Dipl.-Biol. Mohamed Lisfi, Abteilung für Genetik, TU Kaiserslautern

```
use strict;
use warnings;

my @RT_list = (
#GI numbers here
);


foreach my $name (@RT_list){
open (INFILE,"$name.txt") or die "Opening input file failed: $!\n";

my @Inlines =<INFILE>;
my @linebase1;
my @linebase2;
open (OUT,">graph_$name.txt") or die "Opening output file failed: $!\n";
for ( my $n=1; $n<@Inlines; $n++)
{

   if ($Inlines[$n] =~m/^>TR\d*\W{1} (\d+)-- (\d+)\W{1}\s*/){
   $Inlines[$n+1] =~m/ (\w+)\s*/;

    @linebase1 = split (//, $1);

     if ($Inlines[$n+2] =~m/^>VR\d*\W{1} (\d+)-- (\d+)\W{1}\s*/){
  $Inlines[$n+3] =~m/ (\w+)\s*/;

      @linebase2 = split (//, $1);


  chomp $Inlines[$n];
  chomp $Inlines[$n+2];

  unshift (@linebase1, $Inlines[$n]);
  unshift (@linebase2, $Inlines[$n+1]);
  }

printf OUT ("%-30s", "$Inlines[$n]");
printf OUT ("%30s", "$Inlines[$n+1]");
printf OUT ("%-30s", "");
 for (my $m =1 ; $m<@linebase1; $m++)
{
  if ($linebase1[$m] eq $linebase2[$m])
  {
print OUT "|";
 }
 else
{
print OUT "x";
  }
 }
 print OUT "\n";
 printf OUT ("%-30s", "$Inlines[$n+2]");
 printf OUT ("%30s", "$Inlines[$n+3]\n");
 }
 }

}
```

Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die in kleinerem oder größerem Maße zum Gelingen dieser Arbeit beigetragen haben!

Ganz besonders danken möchte ich Frau Junior-Professor Dr. Nora Zingler für die Ausgabe des Themas und den Platz in ihrer Arbeitsgruppe, für zahllose Ratschläge praktischer und theoretischer Art, für viele inspirierende Diskussionen und Einblicke in die Welt der mobilen genetischen Elemente, und besonders für die Gelegenheit, am Aufbau ihrer Arbeitsgruppe mitzuwirken.

Ich bedanke mich außerdem recht herzlich bei Frau Professor Dr. rer. nat. Regine Hakenbeck für die Erstellung des Zweitgutachtens, und ebenso bei Herrn Professor Dr. rer. nat. Stefan Kins für die Übernahme des Vorsitzes der Prüfungskommission.

Herrn Professor Dr. rer. nat. John Cullum und Mohamed Lisfi möchte ich für das Design und die Programmierung der DiGReF-Software danken, ohne die noch immer viele DGRs in den Datenbanken auf ihre Entdeckung warten würden.

Bei Katharina Gimpl aus der Abteilung für Molekulare Biophysik der TU Kaiserslautern bedanke ich mich für die nette Einführung an der Äkta, und für ihre Hilfestellung bei der Gelfiltration. Ebenso danke ich Meik Walther von der Abteilung für Genetik der TU Kaiserslautern für die Berechnung der Alr3496-Struktur mit ROSETTA.

Und natürlich geht ein ganz großes Dankeschön an alle derzeitigen und früheren Mitglieder der Abteilung für Molekulare Genetik der TU Kaiserslautern, ganz besonders an Christine Förster-Schorr für die unzählbaren großen und kleinen Handgriffe, und an meine Büronachbarn Lisa & Rex Jakobi, Maike Gieseke und Philipp Möller für fachliche und weniger fachliche, aber dafür umso unterhaltsamere Unterhaltungen und eine gute Arbeitsatmosphäre.

Mein größter Dank gilt jedoch meinen Eltern, die mich stets unterstützt haben, und denen das Gelingen dieser Arbeit mindestens ebenso am Herzen lag wie mir selbst!

Lebenslauf

Hochschulausbildung

Oktober 2002 – Juli 2005	Bachelorstudium der Molecular Life Science an der Universität Lübeck; Abschluss Bachelor of Science
April 2005 – Juli 2005	Bachelorarbeit am Institut für Humangenetik des Universitätsklinikums Schleswig-Holstein, Campus Lübeck, bei Frau Prof. Dr. rer. nat. Christine Zühlke; Thema: „Sequenzvariationen im Tyrosinasegen pigmentierter Individuen“, Note 1,7
Oktober 2005 – Dezember 2007	Masterstudium der Molecular Life Science an der Universität Lübeck; Abschluss Master of Science
März 2007 – Dezember 2007	Masterarbeit am Institut für Physiologie der Universität Lübeck bei Herrn Dr. rer. nat. Reinhard Depping; Thema: „Der Hypoxie-induzierbare Faktor HIF-2α: Expression, Aufreinigung, Funktionsanalysen und Aspekte des spezifischen Kerntransports“, Note 1,0
September 2009 – August 2013	Doktorarbeit „Diversitätsgenerierende Retroelemente – Identifikation, Klassifizierung, Phylogenie und in vitro-Funktionsanalysen“ in der Arbeitsgruppe Molekulare Genetik der Technischen Universität Kaiserslautern bei Frau Jun.-Prof. Dr. rer. nat. Nora Zingler

Eigene Veröffentlichungen und Vorträge

Zühlke C., Criée C., Gemoll T., Schillinger T., Kaesmann-Kellner B. (2007) Polymorphisms in the genes for oculocutaneous albinism type 1 and type 4 in the German population. *Pigment Cell Research* 20 (3):225-7

Schillinger T, Lisfi M, Chi J, Cullum J, Zingler N. (2012) Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF. *BMC Genomics* 28 (13):430

Schillinger T., Zingler N. (2012) The low incidence of diversity-generating retroelements in sequenced genomes. *Mobile Genetic Elements* 2 (6): 287-91

Schillinger T., Chi J., Cullum J., Zingler N.: The Distribution of a Novel Class of Retroelements across Prokaryotic Species. Poster im Rahmen des Jahresseminars des Fachbereichs Biologie 2011 in Thallichtenberg (2. – 3. Dezember 2011)

Schillinger T.: Diversity-generating retroelements. Vortrag im Rahmen des Jahresseminars des Fachbereichs Biologie der Technischen Universität Kaiserslautern 2009 in Altleiningen (11. Dezember 2009)

Schillinger T.: What can we learn about diversity-generating retroelements from public databases? Vortrag im Rahmen des Jahresseminars des Fachbereichs Biologie der Technischen Universität Kaiserslautern 2012 in Thallichtenberg (13. – 14. Dezember 2012)

Eidesstattliche Versicherung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel „Diversitätsgenerierende Retroelemente – Identifikation, Klassifizierung, Phylogenie und *in vitro*-Funktionsanalysen“ selbständig, ohne fremde Hilfe, und ausschließlich mit den angegebenen Hilfsmitteln und Quellen angefertigt habe, und dass Entlehnungen aus Schriften, soweit sie in der Dissertation nicht ausdrücklich als solche mit Angabe der betreffenden Schrift bezeichnet sind, nicht stattgefunden haben.

Ich versichere ferner, dass die vorliegende Dissertation bisher an keinem anderen Fachbereich der Technischen Universität Kaiserslautern oder einer anderen Universität eingereicht wurde, und ich mich bisher noch keinem Promotionsverfahren unterzogen habe.

Kaiserslautern, den 21. August 2013

Thomas Schillinger