Construction of chimeric AVR2/4 library by means of DNA shuffling and Gateway cloning

Master's thesis
University of Tampere
Institute of Medical Technology
Elina Ojala
December 2010

Preface

The practical part and writing of this study was carried out in the Molecular Biotechnology Group, Institute of Medical Technology (IMT), University of Tampere between April and December 2010. First I would like to express sincere thanks to the group leader Markku Kulomaa for the opportunity to perform my Master's thesis in his group.

Especially I would like to thank my supervisors, docent Vesa Hytönen, who has helped me in all stages of my work by his excellent knowledge and guidance, and MSc Soili Hiltunen, who has provided me an endless amount of help, support and discussion. I also want to thank MSc Barbara Niederhauser, MSc Tiina Riihimäki and MSc Sampo Kukkurainen for their expertise and support in this study. My thanks also belong to technicians Ulla Kiiskinen and Outi Väätäinen for their excellent practical assistance. I am grateful to the whole Molecular Biotechnology group for the wonderful environment and all the fun we have had together, it has been a pleasure to work with all of you!

I would also want to thank my friends and family, who have been standing by me all the time. Without your motivation, support and help this would not have been possible.

Tampere, December 2010

Elina Ojala

Pro Gradu –tutkielma

Paikka:	Tampereen yliopisto Lääketieteellinen tiedekunta Lääketieteellisen teknologian instituutti
Tekijä:	OJALA, ELINA ANNIINA
Otsikko:	Kimeerisen AVR2/4-kirjaston kokoaminen DNA:n sekoituksella ja Gateway-kloonaustekniikalla.
Sivumäärä:	75 + 7 s.
Ohjaajat:	Dosentti Vesa Hytönen ja FM Soili Hiltunen
Tarkastajat:	Professori Markku Kulomaa ja FM Tiina Riihimäki
Aika:	Joulukuu 2010

Avainsanat: DNA:n sekoitus, Gateway-kloonaus, faagikirjasto, avidiinin kaltaiset geenit

Tiivistelmä

Tutkielman tausta ja tavoitteet: Riittävän suuren ja hyvälaatuisen yhdistelmäkirjaston kokoaminen on yksi faagiseulonnan (*engl.* phage display) suurimpia haasteita. Tämän tutkimuksen tavoitteena oli kehittää uusia työkaluja ja menetelmiä kimeeristen avidiini-kirjastojen kokoamiseen. Yhdistämällä DNA:n sekoitus (*engl.* DNA shuffling) ja Gateway-kloonaustekniikat haluttiin saada aikaan entistä tehokkaampi kirjastojen kokoamisstrategia. Gateway-menetelmää testattiin ja hyödynnettiin DNA:n sekoituksella kootun kimeerisen avidiini-kirjaston kloonauksessa.

Materiaalit ja menetelmät: Gateway-kloonaukseen sopiva phagemid-vektori (pGWphagemid) koottiin kloonaamalla *ccd*B-itsemurhageeni ja *att*R-rekombinaatio-sekvenssit tavalliseen phagemid-vektoriin. Homologiset, avidiinin kaltaiset AVR2 ja AVR4 cDNA:t toimivat lähtömateriaalina kimeerisen AVR2/4-kirjaston luomisessa. DNA:n sekoituksen PCR-tuotteet siirrettiin pGWphagemid-vektoriin LR-kloonausreaktiolla ja kirjastoa ilmennettiin *E. coli*ssa. AVR2/4-variantteja tutkittiin sekvensoimalla 96 kloonia ja mittaamalla niiden biotiininsitomiskykyä ELISA:lla.

Tulokset: Kimeerisen AVR2/4 DNA-kirjaston LR-kloonaus pGWphagemid-vektoriin tuotti kooltaan keskikokoisen, 10^6 kloonia sisältävän kirjaston. Kirjasto on laadultaan ja diversiteetiltään korkeatasoinen, sillä kaikki sekvensoidut geenit olivat kokonaisia ja kimeerisiä. Sekvensseistä 76 % oli uniikkeja proteiinitasolla ja crossovereita havaittiin keskimäärin 2,26(±0,07). Duplikaatioita tai deleetioita ei havaittu ja aminohappomutaatioitakin vain erittäin vähän (0,011 %). Toiminnallisuus oli myös säilynyt erittäin hyvin, sillä ELISA:ssa havaittiin 26 biotiinia sitovaa kloonia.

Johtopäätökset: Kimeerisen AVR2/4-kirjaston kokoaminen onnistui ja DNA:n sekoituksen ja Gateway-kloonauksen yhdistelmä vaikuttaa käyttökelpoiselta faagikirjastojen kokoamisstrategialta. Koottua pGWphagemidia käytettiin onnistuneesti kohdevektorina LR-kloonauksessa ja sitä voidaan hyödyntää muidenkin kirjastojen kokoamisessa. AVR2:n ja AVR4:n DNA:n sekoitus onnistui hyvin, ja vastaavia kimeerisiä faagikirjastoja aiotaan hyödyntää steroideja sitovien avidiinien ominaisuuksien parantamisessa.

Master's thesis

Place:	University of Tampere Faculty of Medicine Institute of Medical Technology
Author:	OJALA, ELINA ANNIINA
Title:	Construction of chimeric AVR2/4 library by means of DNA shuffling and Gateway cloning.
Pages:	75 + 7 p.
Supervisors:	Docent Vesa Hytönen and Soili Hiltunen, MSc
Reviewed by:	Professor Markku Kulomaa, and Tiina Riihimäki, MSc
Date:	December 2010

Keywords: DNA shuffling, Gateway cloning, phage display library, avidin-related genes

Abstract

Background and aims: The major challenge in the phage display library construction is to achieve sufficient library size and quality. The goal of this study was to develop new tools and methods for the construction of shuffled phage display libraries. We combined DNA shuffling and Gateway cloning methods in order to create more efficient library construction strategies. The aims were to evaluate the potential of Gateway cloning method and to exploit the developed system by using DNA-shuffled AVR2 and AVR4 cDNAs.

Material and methods: Phagemid vector for Gateway cloning (pGWphagemid) was constructed by cloning the ccdB-suicide insert flanked by *att*R recombination sites into regular phagemid. The cDNAs of highly homologous avidin-related genes *AVR2* and *AVR4* were used as parental genes in DNA shuffling to create a chimeric AVR2/4 library. The *att*L-flanked PCR products of DNA shuffling were cloned into pGWphagemid vector via LR cloning reaction and expressed as recombinant proteins in *E. coli*. The resulting AVR2/4 library was verified by DNA-sequencing of 96 clones and analyzing their biotin-binding affinity by protein ELISA.

Results: The direct LR cloning of shuffled AVR2/4 DNA-library into pGWphagemids resulted in medium-sized library of 10^6 clones. The quality and diversity of the library is quite high, as all sequenced clones were intact and shuffled. 76 % of them were unique on the protein level, and the average amount of crossovers was 2.26(±0.07). The mutation frequency was very low (0.011 % in the aa sequences) and neither deletions nor duplications were observed. The functionality of the AVR2/4 proteins is well retained, as 26 clones were found to bind biotin in protein ELISA.

Conclusions: The chimeric AVR2/4 library was successfully produced by combining DNA shuffling and Gateway cloning methods. The Gateway-compatible phagemid vector was used as a destination vector and was proven to be an applicable tool in library construction. DNA shuffling of AVR2 and AVR4 was productive, yielding an interesting library of variants. The next step is to carry out the phage display selection rounds for this library and finally to construct a chimeric antidin phage display library to improve protein yield and stability of the steroid-binding avidins.

Contents

1. Introduction 8

2. Review of the literature 11
  2.1 Avidin engineering 11
    2.1.1 Avidin-biotin technology 11
    2.1.2 Avidin as an alternative scaffold 12
    2.1.3 Avidin-related proteins 13
  2.2 Protein evolution *in vitro* 16
    2.2.1 Directed evolution: a natural approach to protein design 16
    2.2.2 DNA shuffling 19
  2.3 Phage display 22
    2.3.1 Construction and screening of phage display libraries 22
    2.3.2 Phagemid vectors 24
  2.4 Gateway cloning 25
    2.4.1 Site-specific recombination of phage lambda 26
    2.4.2 Gateway ® cloning technology 27

3. Aims of the study 29

4. Materials and methods 30
  4.1 DNA vectors and cell lineages 30
    4.1.1 Fab-phagemid 30
    4.1.2 pGEM-T Easy 31
    4.1.3 Multiplication of plasmids 32
    4.1.4 Bacterial cell lineages 33
  4.2 Recombinant DNA technology 34
    4.2.1 Characterization and purification of DNA 34
    4.2.2 Construction of pGWphagemid 35
    4.2.3 Control digestion of phagemids 37
    4.2.4 DNA shuffling 37
    4.2.5 Gateway cloning 40
  4.3 Construction of phage display library 41
  4.4 Sequencing 42
  4.5 Protein production, purification and characterization 42
    4.5.1 Protein ELISA 43
    4.5.2 Affinity chromatography 44
    4.5.3 Protein analysis methods 45
  4.6 Computer programs 46

5. Results 47
  5.1 Construction of pGWphagemid vector 47
  5.2 DNA shuffling and Gateway cloning of the AVR2/4 library 49
  5.3 Construction of phage display library 50
    5.3.1 Transformation efficiencies in *E. coli* 50
    5.3.2 Library size 52

5.4 Functional analyses of the AVR2/4 library 52
5.4.1 Protein ELISA 52
5.4.2 Affinity chromatography 54
5.5 Sequence analysis of the AVR2/4 library 55
5.5.1 Analysis of 96 sequenced clones 55
5.5.2 Phylogenetic analysis 57

6. Discussion 59
6.1 Avidin engineering and phage display 59
6.2 DNA shuffling 60
6.3 Gateway cloning 65
6.4 Future objectives 68

7. Conclusions 69

References 70

Appendices 76
Appendix 1: Parameters for PCR 76
Appendix 2: MW markers 77
Appendix 3: Solutions and reagents 78
Appendix 4: Comparison of Pfu and Phusion DNA polymerases 80
Appendix 5: Multiple Alignment of AVR2/4 sequences 81
Appendix 6: UPGMA clustering of shuffled AVR2/4 genes 82

Abbreviations

AP	alkaline phosphatase
APS	ammonium persulfate
AVD, AVD	avidin gene or protein, respectively
AVR, AVR	avidin related gene or protein, respectively
BCIP	5-bromo-4-chloro-3'-indolylphosphate p-toluidine salt (substrate for detecting AP activity in Western blotting applications)
Btn	*D*(+)-biotin (water-soluble vitamin H / B_7)
BSA	bovine serum albumin
cDNA	complementary DNA
CFU	colony forming unit
pDNA	plasmid DNA
ELISA	enzyme-linked immunosorbent assay
HTP	high-throughput
IgG	immunoglobulin G
IPTG	isopropyl β-D-1-thiogalactopyranoside (a molecular mimic of allolactose)
LB	Lysogeny Broth medium
mAb	monoclonal antibody
MCS	multiple cloning site or polylinker (a short segment of DNA which contains several restriction sites)
NBT	Nitro blue tetrazolium chloride (substrate for detecting AP activity in Western blotting applications)
OD	optical dencity
ORF	open reading frame
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PEG	polyethylene glycol
PNPP	*p*-nitrophenyl phosphate (substrate for detecting AP activity in ELISA applications)
SB	Super Broth medium
wt	wild type
SOC	Super Optimal broth with Catabolite repression (SOB medium with added glucose)
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
TEMED	*N*,*N*,*N'*,*N'*-tetramethyl-ethane-1,2-diamine

Nucleotides and amino acids have been indicated by standard one-letter abbreviations (IUPAC).

1. Introduction

Research community and biopharmaceutical industry have a constant demand for highly specific and sensitive binding molecules. Protein engineers are thus focusing on the identification of proteins with new ligand and reaction specificities, and improving the performance of existing ones (Leemhuis *et al.*, 2009). In addition to ligand or substrate specificity, high protein solubility, thermodynamic and chemical stability, single polypeptide chain format, high bacterial expression for cheap production, absence of disulfide bonds and human origin are among the features that are searched for (Nuttall & Walsh, 2008; Skerra, 2007).

On a molecular level, there are two different approaches to create tailor-made proteins: directed evolution and rational design. Rational design is rather information-intensive approach requiring detailed knowledge about the structure and function of the protein (Bornscheuer & Pohl, 2001). Directed evolution method is based on the design guidelines of nature that is Darwinian selection of genetic variants (Leemhuis *et al*,. 2009). The approach involves the generation of random genetic diversity followed by *in vitro* selection of desired variants (Sen *et al.*, 2007). Directed evolution requires no detailed structural information about the protein. If the structural information is available, rational design and directed evolution can be combined to introduce genetic variations at functional sites. In all directed evolution experiments, the gene encoding the protein of interest is recombined or mutated at random to create a large library of variants. The success of the experiment strongly depends on the library size and quality (Leemhuis *et al.*, 2009) as well as screening methods used.

DNA shuffling is one of the most widely applied methods in the generation of mutant libraries (Harayama, 1998). It is based on homologous recombination of genes with high DNA sequence identity (Stemmer, 2002). The product of DNA shuffling is a library of hybrid (i.e. chimeric) genes that contain sequence information from one or more of the parents (Joern *et al.*, 2002). The applications of DNA shuffling have been extended from the *in vitro* evolution of enzymes and specific binders to pharmaceuticals, gene therapy vehicles and transgenes, vaccines and virus-like particles,

microbial genomes and strains, and even laboratory animal models (Harayama. 1998, Patten *et al.*, 1997).

The engineering of specific binder molecules is traditionally based on the immunoglobulin structure. However, immunoglobulin-based binders have been challenged by alternative non-antibody scaffolds, which are currently being engineered (Skerra, 2007). The development of novel types of affinity tools has emerged from the successful engineering of proteins belonging to various structural classes (Nuttall & Walsh, 2008). The calycin family represents a promising class of small molecular scaffolds capable of binding diverse targets through flexible loops displayed on a conserved β-barrel structure (Skerra, 2008). Avidin is a member of calycin family and due to its structural variety, it can be engineered to bind small hydrophobic molecules beyond its natural ligand, biotin.

Mutations in the binding site of avidin has yielded avidin-based proteins (antidins) with moderate affinity towards testosterone (Riihimäki *et al.*, manuscript) and other steroid-like structures (Hiltunen *et al.*, unpublished results). The antidin engineering mainly exploits phage display, which is a powerful method for selecting polypeptides with desired binding specificities from a large library of different variants (Smith & Petrenko, 1997). The method is powerful due to the phage particles, by which a connection between genotype and phenotype is created. This linkage enables to use a large library of phage particles expressing a wide diversity of polypeptides and to select those that bind the desired target (Pande *et al.*, 2010).

The structural variety of avidin scaffold can be further extended by utilizing the other members of avidin family. The chicken avidin gene is known to coexist with a group of closely related avidin-related genes (encoding AVRs 1-7, AVR-A, B and C as well as BBP-1 and BBP-2) (Ahlroth *et al.*, 2000; Keinanen *et al.*, 1994; Niskanen *et al*, 2005). The gene products of chicken avidin family provide an interesting insight into avidin engineering, since they closely resemble avidin but simultaneously show unique structural and functional features (Hytönen *et al.*, 2004b). Recently, members of avidin family have been exposed to DNA shuffling (Niederhauser *et al.*, manuscript) to create improved and completely new features by homologous recombination.

The focus of this thesis is in the construction of phage display libraries for antidin engineering. The background of avidin engineering and recombination strategies relevant for this study will be discussed in the following chapter, *Review of the literature*. The experimental part of the thesis will describe how the methods based on nature's recombination strategies are utilized to create new approach for the library construction. DNA shuffling is applied to the avidin family members *AVR2* and *AVR4* in order to generate novel genotypes and phenotypes for phage display screening. In addition, the cloning of the DNA library into phagemid vectors is aided by introducing a Gateway cloning method, which is based on the site-specific recombination strategy of phage lambda. A new tool, Gateway-phagemid vector, is designed to be utilized in more efficient and easier library construction.

2. Review of the literature

2.1 Avidin engineering

2.1.1 Avidin-biotin technology

Avidin is a tetrameric egg-white protein of chicken (*Gallus gallus*). It is known for its extraordinary high affinity ($K_d \approx 10^{-15}$ M) to its natural ligand, watersoluble vitamin *D(+)*-biotin (Green, 1975; Green, 1990). Due to its tight and specific ligand binding and high stability, avidin is widely used as a tool in a number of affinity-based separations, in diagnostic assays and in a variety of other applications. These methods are collectively known as (strept)avidin-biotin technology (Wilchek & Bayer, 1990). Both avidin and its bacterial relative streptavidin are, as such, suitable for numerous biochemical applications because of their special properties: they are extremely resistant for high temperatures, denaturants, high or low pH and proteolysis. To study the ligand binding and stability characteristics, (strept)avidin has been extensively modified both chemically and genetically. Mutagenesis work related to avidin has been reviewed in more detail by Laitinen *et al.* (2006).

High ligand-affinity and stability of both free and biotin-complexed forms of avidin, the easy attachment of biotin to various target molecules and the non-disruptive chemical nature of biotin are already utilized in numerous life science applications (Wilchek & Bayer, 1990). Avidin has been successfully modified in ligand binding specificity (Määttä *et al.*, 2008), affinity (Marttila *et al.*, 2000) and stability (Hytönen *et al.*, 2005a; Kulomaa *et al.*, 2004; Määttä *et al.*, 2010) as well as in topography and quaternary structure (Hytönen *et al.*, 2006; Laitinen *et al.*, 2003; Nordlund *et al.*, 2004). Recombinant avidins can be efficiently produced in bacterial *Escherichia coli* cells (Hytönen *et al.*, 2004) as well as in eukaryotes; in *Pichia pastoris* yeast (Zocchi *et al.*, 2003), in baculovirus-infected insect cells (Airenne *et al.*, 1999) and in transgenic maize (Hood *et al.*, 1997). Remarkably, avidin has been engineered to bind small hydrophobic molecules other than biotin. Mutations in the binding site of avidin has yielded avidin-based proteins (antidins) with lowered affinity to biotin and higher to keto-biotin (Riihimäki *et al.*, unpublished results). Antidins with moderate affinity towards testosterone (Riihimäki *et al.*, manuscript) and other steroid-like structures (Hiltunen *et*

al., unpublished results) have been developed as well. Avidin continues to remain an interesting framework for future studies and with novel protein engineering techniques it can be further modified towards new applications.

2.1.2 Avidin as an alternative scaffold

Chicken avidin is a relatively small, positively charged homotetrameric glycoprotein (~60 kDa), which provides a rigid and stable scaffold for the recognition of small hydrophobic molecules. Together with lipocalins and fatty-acid binding proteins (FABPs), avidins belong to the calycin superfamily (Flower, 1993), a group of compact one-domain proteins expressed throughout the animal kingdom. Members of calycin family are highly diverged, but they share a well-conserved antiparallel β-barrel fold (see Figure 1) as well as rather high affinity towards small hydrophobic ligand molecules (Skerra, 2000; Flower, 1993). The β-barrel tertiary structure of avidin monomer consists of eight β-strand and their interconnecting loops, as revealed by 3D crystal structure (Livnah *et al.*, 1993). The biotin-binding site is located at one end of the barrel. Amino acid residues in the strands β3 and β6 and in the loops L1,2, L3,4 and L5,6 are in direct contact with biotin. Thus the ligand-binding site is composed of loop regions, closely resembling the hypervariable CDR-loops (complementarity determining regions) found in the immunoglobulins.

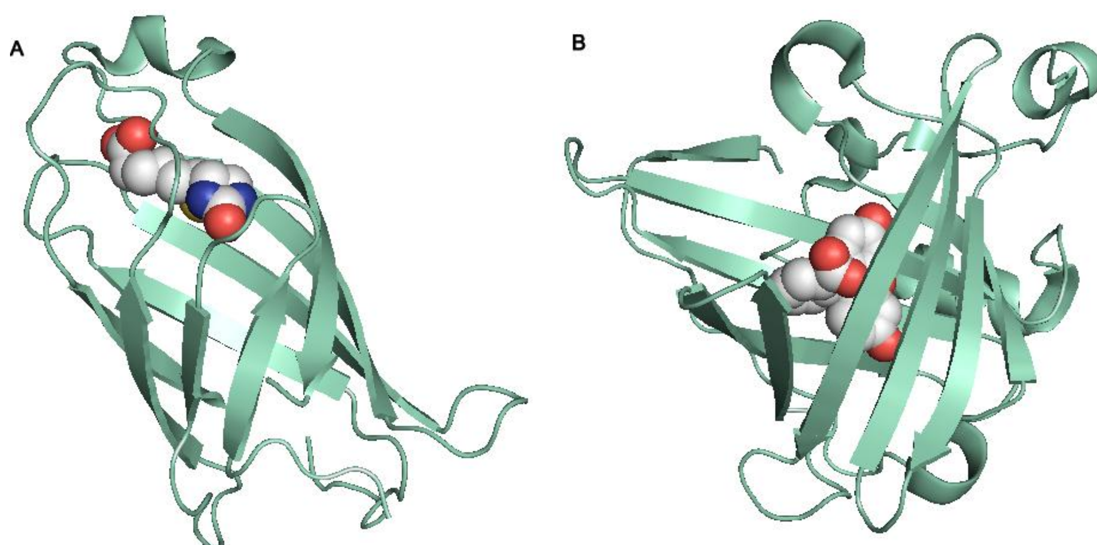


Figure 1: The structure comparison of two calycin family members. Avidin and lipocalin structures are divergent on the sequence level yet they have conserved β-barrel structures (Flower *et al.,* 2000). A) Avidin monomer in complex with its natural ligand, biotin. B) Engineered anticalin FluA monomer in complex with fluorescein ligand. The PDB coordinates are 2AVI and 1NOS, respectively.

Antibodies are often considered as a paradigm for specific binder proteins. With their hypervariable CDR-loop regions supported by a structurally rigid framework, immunoglobulins provide a vast repertoire of antigen-binding sites in the immune system and make up the majority of reagents used in *in vitro* diagnostics and immunotherapeutics. Despite of their extraordinary variability, antibodies suffer from some fundamental problems in biomedical applications (Skerra, 2007). Because of the multi-domain immunoglobulin (Ig) structure, they are relatively unstable, large in size and expensive to produce. The hypervariable CDR-loop structure is very suitable for binding different epitope structures, but due to their complexity, immunoglobulins may be somewhat difficult to manipulate. Because of these limitations, a plethora of both Ig-based and alternative non-antibody protein scaffolds have been engineered recently in order to create new specific binders for research purposes, *in vitro* diagnostics and therapeutic applications (Nuttall & Walsh, 2008; Skerra, 2007). Alternative scaffolds have been developed not only to create competitors but alternatives to immunoglobulins. Calycins are another family of proteins that exhibit a binding site with high structural plasticity, which is composed of peptide loops mounted on a stable β-barrel scaffold (Flower *et al.*, 2000; Skerra, 2008). The structures of lipocalin and avidin monomer closely resemble each other (see Figure 1), and both can be utilized as alternative scaffolds to engineer proteins with novel functions.

2.1.3 Avidin-related proteins

The scope of avidin scaffold can be further extended by engineering of other members of avidin family. The chicken avidin gene family consists of avidin and seven avidin-related genes (*AVR1-7*) showing high sequence similarity (Keinänen *et al.*, 1994; Ahlroth *et al.*, 2000). In addition, there are genes deviating from those described by Ahlroth *et al.* (2000) in terms of the encoded amino acid sequence, thereby named *AVR-A, B, C* and *BBP-1* and *2* (Niskanen *et al.*, 2005). The *AVR* DNA sequences are nearly identical to each other but exhibit nucleotide substitutions that are nonrandomly distributed and frequently nonsynonymous compared to *AVD* (Ahlroth *et al.*, 2001a). Seven different *AVR*s have been reported, but the total number of avidin-related proteins (AVRs) is likely to vary between different individuals and even between different cells of the individual chicken (Ahlroth *et al.*, 2001b).

Avidin related genes have been pointed to be highly homologous (Ahlroth *et al.*, 2001a), as *AVR*s 1-7 are showing 91 – 100 % of sequence identity on the nucleotide level. *AVR4* and *AVR5* are actually 100 % identical on their coding sequences (Keinänen *et al.*, 1994) and their common protein product is therefore indicated as AVR4/5 or simply AVR4. *AVR* genes are 1113 bp in length and consist of four exons and three introns, like avidin. The exon-intron interfaces of *AVR* genes are clear and contain splicing signals. The genes also contain presumed 5' promoter sequences and 3' polyadenylation signals, indicating functionality of these genes (Ahlroth *et al.*, 2001a). However, the function of *AVR* gene products in nature remains a mystery. The AVR proteins have not yet been isolated from chicken although AVR mRNAs can be found during inflammation in several chicken tissues and in the oviduct after progesterone induction (Kunnas *et al.*, 1993). Blast search of AVR cDNAs reveals that AVR mRNAs have been also reported elsewhere, but it is not currently known if the transcripts are translated into proteins.

Recombinant AVR proteins have been produced and characterized in detail and thus been demonstrated to be functional biotin-binding proteins like chicken avidin (Laitinen *et al.*, 2002). However, recombinant AVRs have shown different properties compared to avidin. The AVRs differ from avidin with respect to glycosylation and charge properties (Hytönen *et al.*, 2004; Laitinen *et al.*, 2002). All AVRs except AVR2 contain an uneven number of cysteine residues in their sequence, which can form inter-subunit disulphide bridges in addition to the intra-subunit disulphide bridges also seen in avidin (Laitinen *et al.*, 2002). The most interesting feature, the biotin-binding affinity, varies between AVRs and compared to avidin as well. A wide range of values have been reported, AVR4 being almost as efficient a biotin binder ($K_d \approx 10^{-14}$ M) as avidin, whereas AVR2 is showing the lowest affinity for biotin within the investigated avidin family members (Hytönen *et al.*, 2004; Laitinen *et al.*, 2002).

Like avidin, AVRs are very stable proteins. AVR4/5 is probably the most interesting member of the AVR family, because it has features like retained high biotin affinity and improved thermal stability (Hytönen *et al.*, 2004a). Similar to avidin, AVRs can be efficiently produced in *E. coli* (Hytönen *et al.*, 2004b) using the ompA signal peptide. The 3D-structure of AVR2 and AVR4/5 has been determined (Eisenberg-Domovich *et al.*, 2005; Hytönen *et al.*, 2005b), and also chimeric forms of avidin and AVR4 have

been produced (Hytönen *et al.*, 2005a). More recently, members of avidin family have been exposed to DNA shuffling (Niederhauser *et al.*, manuscript) to create improved and completely new features by homologous recombination.

Sequence differences that have been observed in avidin-related genes (Ahlroth *et al.*, 2001a) are likely to affect not only the ligand-binding properties, but the stability and glycosylation patterns of the AVR proteins as well. The evolutionary relationships between avidin and AVRs have been examined with several methods (e.g. maximum parsimony, maximum likelihood and neighbor joining) giving nearly equal results (Wallen *et al.*, 1995). On the phylogenetic analysis of avidin and avidin-related genes it can be suggested that *AVR4* and *AVR5* are the closest relatives to avidin (Figure 2). The biotin-binding affinity is conserved in all AVR proteins, albeit biotin-binding of AVR1 and AVR2 is reversible in contrast to avidin and AVR3-7, which bind to biotin almost irreversibly (Laitinen *et al.*, 2002).

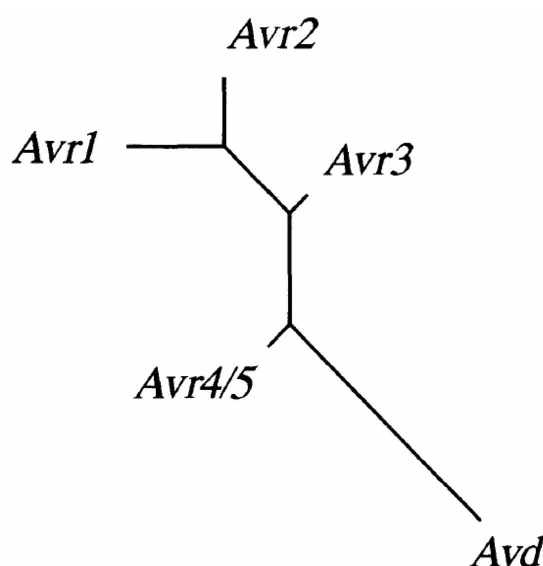


Figure 2: Phylogenetic analysis of avidin and avidin-related genes *AVR1-5*. The tree is constructed from the multiple sequence alignment using maximum likelihood method. *AVR4/5* has the highest sequence similarity with avidin, and *AVR1* and *AVR2* are the most distant members of the family. *AVR6* and *AVR7* are not shown in this figure, because they were cloned afterwards. (Modified from Wallen *et al.,* 1995.)

Biotin-binding affinity is the only known measure of activity for these proteins, and some conclusions about the evolutionary events of the avidin family in chicken can be drawn on it (Hytönen *et al.*, 2004a; Hytönen *et al.*, 2005b; Laitinen *et al.*, 2002). Since AVR2 has the lowest affinity towards biotin and slightly less sequence similarity as

well, it may be considered as an ancestral representative of avidin family in chicken. AVR1, 3, 6 and 7 also have lower biotin-binding affinities and less sequence identity compared to AVR4/5 and avidin. Since AVR4/5 has been found to have biotin-binding affinity almost as high as avidin (Hytönen *et al.*, 2004a) and the highest sequence similarity with avidin as well, they are likely to be the closest relatives, and, together with BBP-A (Niskanen *et al.*, 2005), the most active, functional representatives of chicken avidin family.

2.2 Protein evolution *in vitro*

2.2.1 Directed evolution: a natural approach to protein design

Two rather contradictory tools can be used on a molecular level to create tailor-made proteins: directed evolution and rational protein design. Rational design usually requires both the availability of the structure of the protein and knowledge about the relationships between sequence, structure, and function, and is therefore quite information-intensive method (Bornscheuer & Pohl, 2001). In the past several years, directed evolution has emerged as an alternative approach to rational design. There is a remarkable difference between these two approaches: the most work using rational design focuses on mutations close to the active site, whereas directed evolution experiments often find mutations far from the active site (Sen *et al.*, 2007). The power of directed evolution is in the Darwinian selection of genetic variants (Leemhuis *et al.*, 2009; Stemmer, 2002). In the directed evolution approach, no detailed knowledge about the protein structure-function relationship is required, but if such information is available, rational design and directed evolution can be combined to introduce genetic variations at functional sites (Tobin *et al.*, 2000).

Directed evolution enables the improvement of structural and functional properties, such as expression levels, stability and performance under different conditions or changes in their reaction or binding specificity (Stemmer, 2002). It is particularly well suited approach for protein function tuning (Sen *et al.*, 2007), which means improving a feature that already exists at some level or combining of properties not necessarily found together in nature. Furthermore, alternative solutions gaining functional change or completely new features can be obtained (Tobin *et al.*, 2000). The approach implements an iterative Darwinian optimization process, whereby the fittest variants are

selected from an ensemble of random mutations (Stemmer, 2002; Leemhuis *et al.*, 2009). The empirical strategy of creating variants and selecting those that perform best is the essence in all protein tailoring methods and it is utilized in natural selection and classical breeding as well (Stemmer, 2002).

In all directed evolution experiments, the gene encoding the protein of interest is recombined or mutated at random to create a large library of gene variants. Methods for the creation of protein-encoding DNA libraries may be divided into three main categories (Sen *et al.*, 2007). The first two categories encompass techniques that directly generate sequence diversity in the form of point mutations, insertions, or deletions. These changes can be made at random along a whole gene (random mutagenesis) or at specific areas within a gene sequence (directed random mutagenesis). Due to the mutations, the relative amount of ORFs coding functional proteins is typically quite low (Tobin *et al.*, 2000). The third category, *in vitro* recombination (also known as molecular breeding) encompasses numerous techniques, which have been developed to mimic and accelerate nature's recombination strategy (Sen *et al.*, 2007; Stemmer 2002).

In nature, genetic variation in DNA arises from errors introduced during genome duplication, or via DNA damaging by UV light, chemicals and other external factors. Virus infections may alter the content of the host genome as well. In laboratory, random genetic variation in DNA is usually created by polymerase chain reaction (PCR) methods techniques (Leemhuis *et al.*, 2009). One or more parental genes are applied as starting material for modification; leading to the generation of some kind of a DNA library. The success of the experiment strongly depends on the library size and quality (Leemhuis *et al.*, 2009) and therefore, both the directed evolution method and parental gene(s) have to be carefully selected. The organization of the avidin gene family in chicken sex chromosome Z has been mapped (Ahlroth *et al.*, 2000) and the adjacency of *AVR*s indicates that they might have arisen as duplications. The molecular mechanisms underlying this kind of lability are most probably unequal crossing-over and/or unequal sister chromatid exchange (Ahlroth *et al.*, 2001b). The chicken avidin gene family thus provides an excellent model for studying the mechanisms of recombination. Due to their high sequence homology and natural tendency towards recombination events, *AVR*s are particularly well suited for recombination-based directed evolution approaches.

Several types of DNA libraries can be developed for specific purposes, but all share some common features (Leemhuis *et al.*, 2009). The DNA fragments that make up the library are cloned into vectors, which allow the DNA to be replicated and stored within model organisms such as bacteria or yeast. In general, plasmid-based vectors are considered the easiest to manipulate. They are commonly used for applications that involve complex manipulations, but that require only small DNA fragments (cDNA). In directed evolution approaches, cDNA libraries containing millions of variants are typically created. In the case of calycin family, all variants are based on common protein scaffold, which has been mutated to create variations in amino acid sequence (Skerra, 2000). The members of calysin family (as well as immunoglobulins) naturally bind various targets, and by combining this kind of properties by *in vitro* evolution, it is possible to select and isolate specific binders towards novel targets.

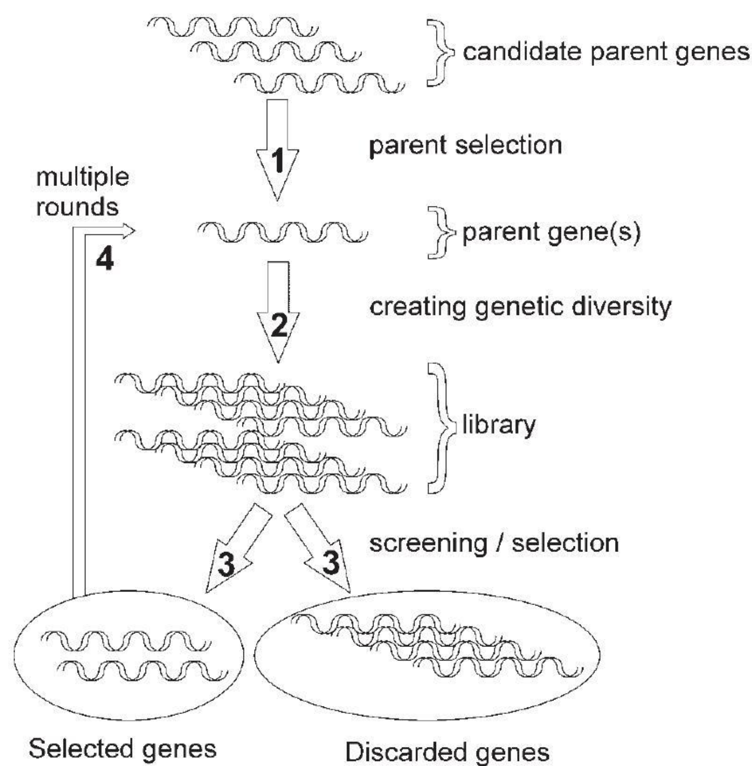


Figure 3: Flowchart of directed evolution. Genetic variation can be generated by multiple methods, but following basic steps have to be considered in any case: (1) selection of parental genes, (2) directed evolution method to create a library, (3) HTP screening and selection of the desired variants and (4) repetition of the diversity creating rounds (modified from Leemhuis *et al.*, 2009).

The generation of random genetic diversity is followed by high-throughput (HTP) screening of desired variants (Figure 3). The process is usually carried on in several cycles. Each cycle comprises selection of starting material, generation of diversity (by

recombination or different forms of mutagenesis), screening for the best individuals and their amplification to go on to the next cycle. There are various techniques to express and isolate the variants of interest. They can be divided into selection versus screening and *in vivo* versus *in vitro* techniques (Leemhuis *et al.*, 2009). The choice of selection or screening method is another important and challenging step in the directed evolution process. It allows the selection pressure to be focused on relevant properties and has a major impact on the outcome of the whole process.

2.2.2 DNA shuffling

Since the publication of the first DNA shuffling papers by Stemmer (Stemmer, 1994a; Stemmer, 1994b), several mutagenesis and recombination based schemes for directed evolution have been developed (Tobin *et al.*, 2000). Different methods of directed evolution, including error-prone PCR, staggered extension process (StEP), heteroduplex recombination, random priming recombination, RACHITT recombination and incremental truncation (ITCHY and SCRATCHY), as well as *in vivo* methods are reviewed elsewhere (see Sen *et al.*, 2007) and only DNA shuffling is described here in more detail.

DNA shuffling has been and still is one of the most widely applied methods for generating combinatorial libraries (Harayama, 1998; Stemmer, 2002). It is based on homologous recombination of genes with high DNA sequence identity. The technique was originally used as single gene shuffling to randomly recombine various mutants of a single gene (Stemmer, 1994a). DNA family shuffling has later been shown to effectively recombine homologous genes among gene families and between species (Crameri *et al.*, 1998). In a comparative study, variants obtained by family shuffling were shown to outperform variants obtained with single gene shuffling (Crameri *et al.*, 1998), though it requires the availability of multiple parents with high DNA sequence similarity. The applications of directed evolution have since been extended to new and improved enzymes, pharmaceutical proteins, gene therapy vehicles and transgenes, vaccines, viruses and laboratory animal models (reviewed by Patten *et al.*, 1997) and remarkably, to specific binders (Gunneriusson *et al.*, 1999; Sheedy *et al.*, 2007). Different shuffling strategies have been utilized in engineering of affibody-scaffolds (Gunneriusson *et al.*, 1999) and in affinity maturation of antibodies (Sheedy *et al.*, 2007) to create highly specific binders. The DNA shuffling of calycin family members

or another alternative scaffolds have not been reported however, and there is demand for the development of novel binders by recombinative methods.

DNA shuffling thus represents a powerful approach to generating novel sequences that encode functionally interesting proteins. The product of DNA shuffling is a library of hybrid (or chimeric) genes that contain sequence information from one or more of the parents (Joern *et al.*, 2002) (Figure 4). Because they are based on natural diversity permutated by homologous recombination, the generated libraries are very high quality. Libraries of clones that are created by DNA shuffling are phenotypically diverse because clones tend to differ by many amino acids due to the exchange of sequence blocks. Yet an exceptionally high fraction of the library is functional (Stemmer, 2002), because the natural sequence polymorphisms were preselected for compatibility with function. A low rate of random point mutagenesis typically occurs while related sequences are exposed to *in vitro* recombination (Stemmer, 2002) which further increases the diversity of the library. By using high-throughput methods, improved clones can be identified with a small number of complex screens from the libraries of high quality and diversity.

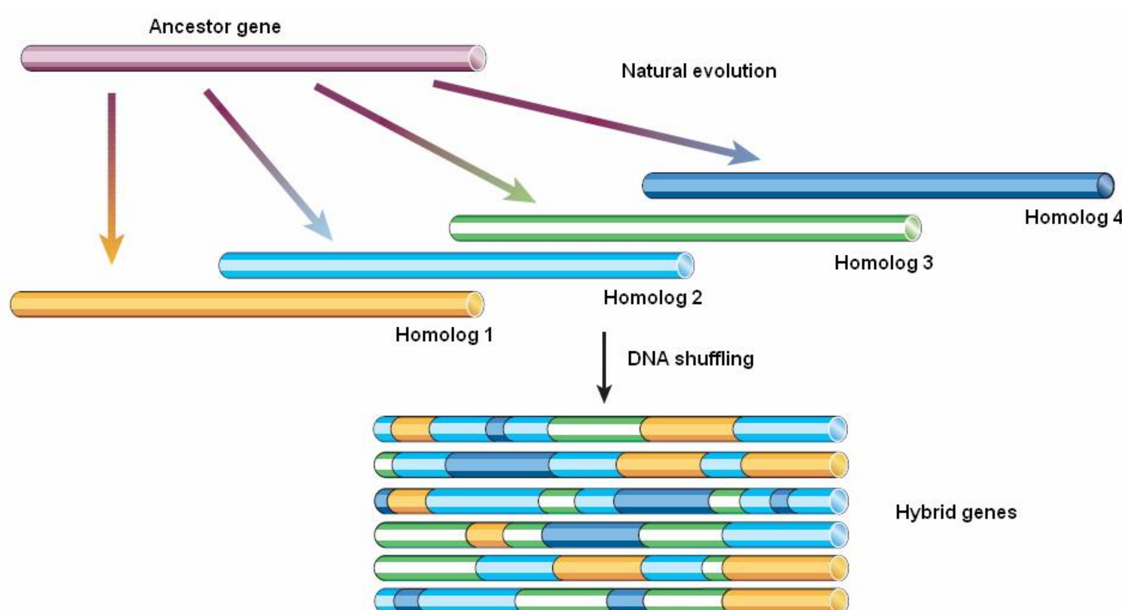


Figure 4: Molecular breeding by DNA shuffling. Diverse gene libraries for laboratory evolution can be created by recombination of related genes. This approach generates highly diverse sequences, but conserves function. Improved or altered proteins have been identified by screening such hybrid protein libraries. (Modified from Arnold, 2001.)

Because DNA shuffling is based on high homology of parental genes, at least 60 % of sequence similarity (identical nucleotides) is typically required for the successful recombination (Crameri *et al*., 1998). In this way, advantageous mutations arising from sequence divergence are effectively combined, while at the same time purging deleterious mutations, which is not possible with nonrecombinative methods such as error-prone PCR (Leemhuis *et al*., 2009). Standard shuffling procedure (Stemmer, 1994a) involves the controlled fragmentation of parent DNA, usually by DNaseI. The resulting fragments of 50 to 150 base pairs are then used as both PCR template and primers in self-priming reassembly PCR. This part of the protocol is often called primerless PCR, because distinct primer oligonucleotides are not introduced in the reaction mixture. The diversity is generated in the form of crossovers, as fragments from different parents are reassembled in the annealing step of primerless PCR (Moore & Maranas, 2002). Extension of heteroduplexes is performed by regular DNA polymerase, which is usually high-fidelity if additional mutagenesis is undesirable. The principle of the PCR setup is shown in Figure 5.

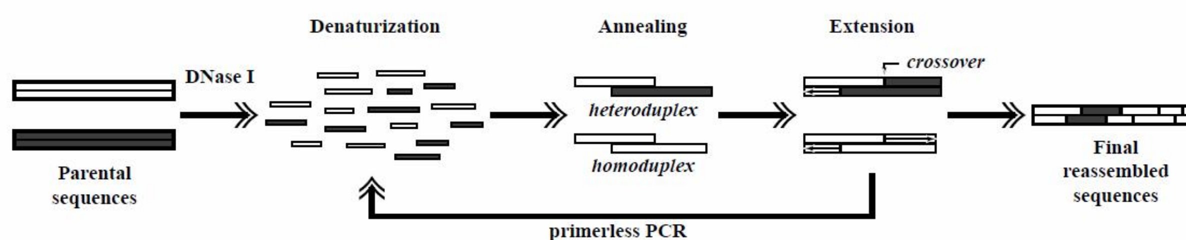


Figure 5: The DNA shuffling protocol consists of random fragmentation of parental nucleotide sequences with DNaseI and fragment reassembly through primerless PCR. Each cycle of PCR has three steps: denaturization, annealing and extension. Diversity in the form of crossovers is generated when two fragments from different parents spontaneously anneal (forming a heteroduplex) and are extended by DNA polymerase. (Modified from Moore & Maranas, 2002.)

Due to the crossovers, DNA shuffling is an approach by which the existing diversity can be recombined into completely new variants, which have new properties not observed in any of the parents (Stemmer, 2002). By copying the natural mechanisms by which existing diversity can be recombined, DNA shuffling can be used to generate high quality libraries of protein variants. Chimeras between naturally occurring proteins that differ by only a few amino acids often have features that are significantly different from their parents (Minshull & Stemmer, 1999). By screening this kind of chimeric libraries using efficient and innovative HTP assay techniques, it is possible to identify proteins

with new functions and physical properties or to improve the performance of existing ones.

2.3 Phage display

Different display methods have become common tools in the field of random protein engineering. A large number of alternative ways to generate and handle complex combinatorial libraries have been developed, including phage display, ribosomal display, yeast surface display, bacterial display, microbead display, protein complementation assays and nucleic acid-based assays (reviewed by Uhlen, 2008). There are various technologies available to isolate the interesting variants from created DNA libraries, but only phage display is considered here. It is widely used in the engineering of specific binders, like antibodies, and it has been utilized in the development of avidin-based binders as well (Riihimäki *et al.*, manuscript).

2.3.1 Construction and screening of phage display libraries

Phage display (Smith, 1985) was the first molecular diversity selection platform and the conceptual forerunner of all subsequent display techniques. Despite some limitations, it continues to remain the most commonly used *in vitro* method to select peptides and antibodies, and it is relatively robust compared to other methods. Phage display has been acknowledged as a powerful method for selecting and engineering polypeptides with desired binding specificities (Smith & Petrenko, 1997; Sidhu. 2001).

The method relies on the fact that if genes encoding interesting polypeptides are fused to M13 coat protein encoding genes, these fusions can be incorporated in bacteriophage particles that display the heterologous proteins on their surfaces (Smith, 1985). In this way, a physical linkage is established between genotype and phenotype. The connection between genotype and phenotype enables large libraries of proteins to be screened and amplified in a process called *in vitro* selection, which is analogous to natural selection. The most common bacteriophages used in phage display are M13 and fd filamentous phages, though T4, T7, and λ phage have also been used (Barbass *et al.*, 2004). Phage display has been utilized most widely in the screening of antibody libraries (Bradbury & Marks, 2004; Viti *et al.*, 2000), but it is suitable for alternative scaffold engineering as well (Skerra, 2007).

The concept of phage display is simple in principle: a library of phage particles expressing a wide diversity of polypeptides is used to select those that bind the desired target (Pande *et al.*, 2010). The technology relies on the utilization of phage display libraries in a cycloidal screening process known as biopanning (Figure 6). Biopanning is an affinity selection technique, by which the interesting variants can be selected from large phage display libraries. The desired binding specificity possessing proteins can be selected from library by binding to an immobilized ligand, and their sequences can be deduced from the sequences of the encapsulated DNA. As with any combinatorial method, the success of phage display depends on the size and quality of the initial library (Sidhu, 2001). The need for large libraries has been widely appreciated, and current optimized methods enable the easy construction of phage display libraries containing 10^{13} unique DNA sequences (Sidhu, 2001; Grönwall & Ståhl, 2009).

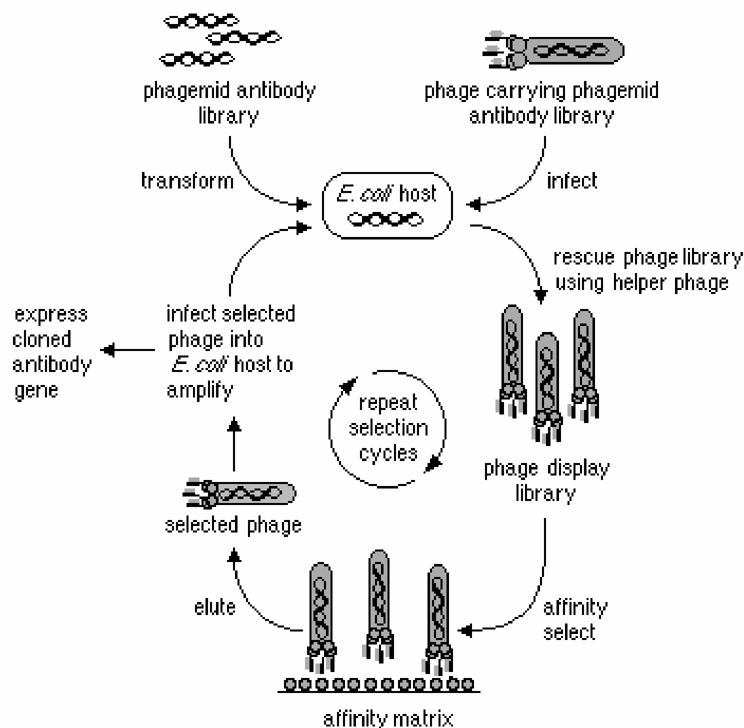


Figure 6: Biopanning screening process. A paradigm library of antibodies is shown here, but a library of interest may contain any kind of polypeptides. The library can be either transformed as phagemid pDNA or infected using phage particles carrying phagemids into host cells. In the following biopanning rounds, phages are bound to the ligand surface, poorly binding phages are washed away and the most interesting variants are selected from the library. The best performing phages are eluted and carried on to the next cycle or amplified for sequencing and expression. (Modified from Galanis *et al.*, 2001.)

2.3.2 Phagemid vectors

Both phage and phagemid vectors have been used to display antibodies and other proteins (Bradbury & Marks, 2004), but due to the numerous advantages of phagemids, the libraries are usually constructed in the form of phagemid vectors (Figure 7). Phagemids are cloning vectors developed as a hybrid of the filamentous phage M13 and regular plasmid to produce a vector that can propagate as a plasmid in bacterial cells, and can also be packaged as single stranded DNA in viral particles (Swords, 2003). Phagemids contain a conventional origin of replication (ori) for double stranded replication in bacteria, as well as an f1 ori to enable single stranded replication of phage genome and its packaging into phage particles (Sidhu, 2001). Similarly to a plasmid, a phagemid can be used to clone DNA fragments and be introduced into a bacterial host by a range of techniques (transformation and electroporation) (Barbass *et al.*, 2004). However, infection of a bacterial host containing a phagemid with a helper phage, for example VCSM13 or M13K07, is required to provide the necessary viral components to enable the infection, single stranded DNA replication and packaging of the phagemid DNA into phage particles (Swords, 2003).

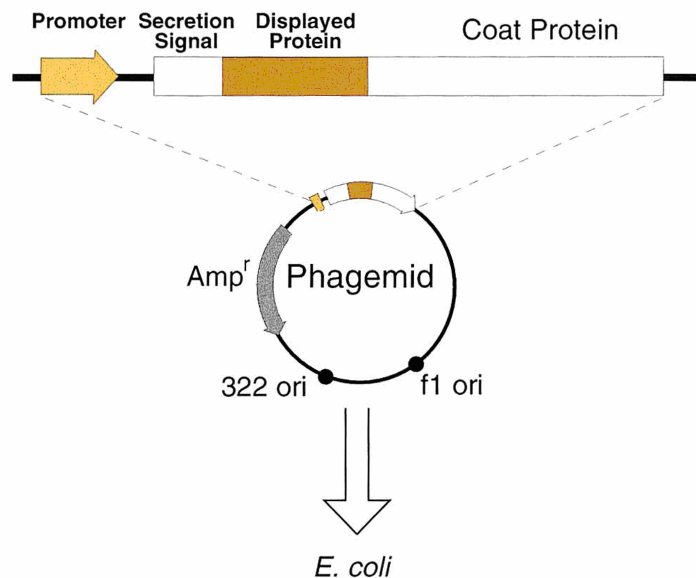


Figure 7: The organization of the coding sequence of the phage displayed protein in phagemid vector. The gene of interest is fused to the coding sequence of phage coat protein. Orientation of the gene respect to the promoter and signal sequence is critical as well. Phagemid vector closely resembles bacterial plasmids, containing genes for antibiotic resistance etc. Remarkably, it has two origins of replication: a conventional ori (322 or pUCfor example) for plasmid replication in bacteria, and an f1 ori for single stranded replication and packaging of DNA into phage particles. (Modified from Sidhu, 2001.)

The utilization of the phagemid library requires that the genes of interest are cloned under inducible promoter(s), fused with phage coat protein and preceded by an suitable signal sequence (Barbass *et al*., 2004; Sidhu, 2001). In the case of avidin libraries and many others, the gene of interest is fused to the coding region of C-terminal pIII. The minor coat protein pIII is located on the tip of the M13 phage particle (Sidhu, 2001) and thus the fusion proteins are displayed in the phage surface among the natural coat proteins. The pIII protein is necessary for the phage infectivity however, and the pIII-fusions need therefore to be counterbalanced by additional wt pIII proteins derived from helper phage. Amber stop codon between the gene of interest and gene III enables the display of fusion proteins on the surface of the phages, when the phages are produced in the amber suppressive *Escherichia coli* strain (Barbass *et al*., 2004). After multiple selection and screening rounds, the production of free proteins is possible in nonsuppressor *E. coli* strains.

The production of fusion proteins that are displayed on the phage coat requires that the cloning of the whole library of variants is cloned into phagemid in correct reading frame and orientation (Barbass *et al*., 2004). A wide variety of molecular biology techniques have been developed, by which large and diverse polypeptide libraries can be generated (Leemhuis *et al*., 2009). However, actual construction of phagemid libraries using traditional recombinant DNA techniques may be quite laborious and may limit the library size and diversity. New tools for more efficient and clever library construction are needed in order to manage the simple cloning of DNA libraries into phagemids.

2.4 Gateway cloning

Concerning the construction of phage display libraries in antidin engineering, there is an apparent need for more efficient cloning methods. The conventional cloning (using restriction endonucleases and ligases) and recombination-based sequence and ligation independent cloning (SLIC) method have been used so far in the construction of antidin libraries. Since they are rather low in speed and efficiency, new methods are needed in order to ease the process of library construction. Gateway cloning method (Hartley *et al*., 2000) has been applied in the cloning of phage display libraries for antibody screening (Schofield *et al*., 2007), and it is thus suitable for avidin engineering as well.

2.4.1 Site-specific recombination of phage lambda

Gateway cloning technology (Invitrogen) is based on the well-characterized site-specific recombination system of bacteriophage lambda (λ phage; Campbell, 1962; Landy, 1989). Lambda-based recombination involves two major components, specific recombination sequences and the enzymes mediating the recombination reaction between them (Gottesman & Weisberg, 2004). These components facilitate both lysogenic and lytic pathways of the phage. In the beginning of the lysogenic cycle, λ DNA is integrated into host genome as a prophage, which stays resident within the host's genome without causing any apparent harm to the host cell. The lytic cycle in turn requires the excision of the prophage from the host genome, leading to the assembly of new phage particles and their escape via cell lysis. Site-specific recombination is required to initiate both of these turning points (Landy, 1989).

Lambda integration into the *E. coli* chromosome occurs via intermolecular DNA recombination that is mediated by a mixture of lambda and *E. coli* encoded recombination proteins (Gottesman & Weisberg, 2004). The lysogenic pathway is catalyzed by the lambda integrase (Int) and *E. coli* integration host factor (IHF) proteins while the lytic pathway is catalyzed by the lambda integrase and excisionase (Xis) proteins, and the *E. coli* Integration Host Factor (IHF) protein (Landy, 1989). The integration itself is a sequential exchange via a Holliday junction and requires both the phage integrase and the bacterial IHF. Both Int and IHF bind to *att*P and form an intasome, a DNA-protein-complex designed for site-specific recombination of the phage and host DNA. The DNA segments flanking the recombination sites are switched, such that after recombination, the att sites are hybrid sequences comprised of sequences donated by each parental vector (Gottesman & Weisberg, 2004). For example, attL sites are comprised of sequences from attB and attP sites.

Recombination occurs between specific *att*achment (*att*) sites, *att*B on the *E. coli* chromosome and *att*P on the lambda chromosome. The *att* sites serve as the binding site for recombination proteins and have been well-characterized (Weisberg & Landy, 1983). Upon lambda integration, recombination occurs between *att*B and *att*P sites to give rise to *att*L and *att*R sites. Capitalized letters B, P, L and R stand for the bacterial attachment site, the phage attachment site, and the two recombinant attachment sites located to the left and right of the prophage, respectively (Gottesman & Weisberg,

2004). The actual crossover occurs between homologous 15 bp core regions on the two sites, but surrounding sequences are required as they contain the binding sites for the recombination proteins (Landy, 1989).

2.4.2 Gateway ® cloning technology

The lambda recombination system is facilitated in the Gateway cloning to transfer heterologous DNA sequences (flanked by modified *att* sites) between vectors (Hartley *et al*., 2000). The system carries out two recombination reactions: (1) BP reaction attB × attP → attL + attR mediated by Int and IHF and (2) LR reaction attL × attR → attB + attP mediated by Int, IHF, Xis. In more detail, BP reaction facilitates recombination of an *att*B substrate (which can be *att*B-flanked PCR product or a linearized *att*B expression clone) with an *att*P substrate (donor vector) to create an *att*L-containing entry clone (see Figure 8a). The reaction is catalyzed by BP Clonase™ enzyme mix, which contains both lambda and *E. coli* -encoded recombination enzymes. Accordingly, LR Reaction facilitates recombination of an *att*L substrate (entry clone) with an *att*R substrate (destination vector) to create an *att*B-containing expression clone (see Figure 8b). This reaction is catalyzed by LR Clonase™ enzyme mix, which contains Xis in addition to Int and IHF.

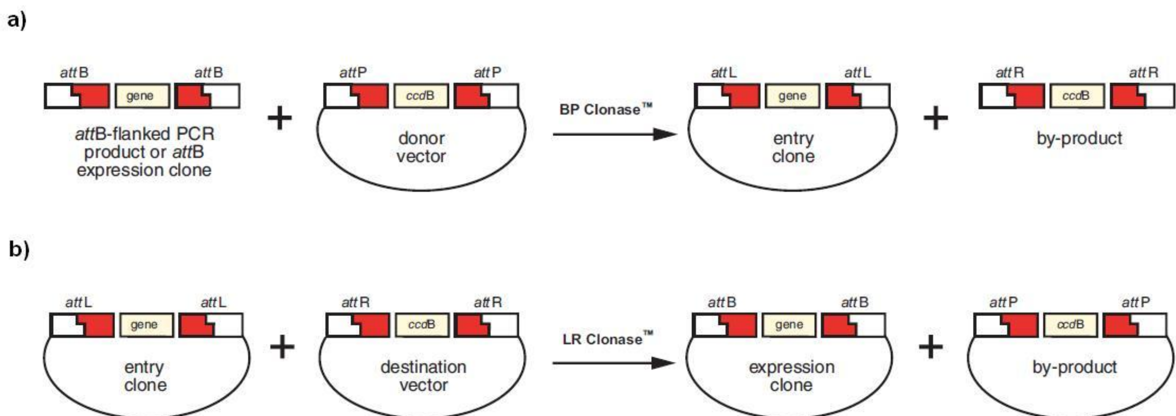


Figure 8: Scheme of Gateway cloning reactions. a) Recombination between attB and attP sites is facilitated in the BP reaction. b) Recombination between attL and attR in LR reaction. (Modified from Invitrogen catalog #12535-019.)

The direction of the reactions is controlled by providing different combinations of proteins and sites. Both donor and destination vectors initially contain a selection marker, the F-plasmid-encoded *ccd*B gene (Bernard & Couturier, 1992), which inhibits

the growth of F-negative *E. coli*. The *ccd*B is a suicide gene inducing gyrase-mediated double-stranded DNA breakage and thus causing death of bacterial cells. This kind of positive selection ensures that only those cells carrying the desired construct are able to survive.

The site-specific recombination reactions mediated by the lambda integrase family of recombinases are conservative (no net gain or loss of nucleotides) and highly specific (Landy, 1989). This kind of recombination does not require DNA synthesis and it can occur between all kind topologies of DNA (i.e. supercoiled, linear, or relaxed), although efficiency varies. Linearized DNA is usually recombined most efficiently. The wild-type *att* recombination sites have been modified to further improve the efficiency and specificity of the sites. Thus *att*B1 will only recombine with *att*P1 but not *att*P2, thereby maintaining orientation of the DNA segment during recombination. Aberrant recombination events have not been identified in hundreds of sequenced clones. (Hartley *et al*., 2000)

Gateway technology allows the cloning, combining and transferring of DNA segments between different expression vectors in a high-throughput manner while maintaining orientation and the reading frame of the fragment of interest (Katzen, 2007). A variety of vectors compatible with the Gateway systems is commercially available, and can be created by one's own as well. Gateway cloning provides several advantages over classical cloning, such as easiness, increased speed and efficiency during each cloning step. Thus DNA sequences (single genes or cDNA libraries) can be moved into multiple vector systems for functional analysis and protein expression in efficient and reliable manner (Hartley *et al*., 2000).

Gateway-compatible vectors for high-throughput analysis of protein interactions have been recently constructed for yeast two-hybrid system (Zhu *et al*., 2010) because it greatly facilitates the cloning of interested DNA fragment into vectors and therefore increases the efficiency of analysis. Similar advantages are likely to be acquired by introducing Gateway cloning technology into the construction of phage display libraries.

3. Aims of the study

The aim of the study was to develop new tools and methods for the construction of shuffled avidin-related gene libraries. We evaluated the potential of Gateway cloning method and evaluated the developed system by using DNA-shuffled AVR2 and AVR4 cDNAs. More specifically, the aims were:

1. To construct a new phagemid vector containing a Gateway-cassette.
2. To create a chimeric AVR2/4 library by DNA shuffling.
3. To transfer the shuffled AVR2/4 library into phagemid using Gateway cloning system.
4. To evaluate the feasibility of the combination of Gateway cloning and DNA shuffling in cDNA library construction.

4. Materials and methods

4.1 DNA vectors and cell lineages

4.1.1 Fab-phagemid

Fab-phagemid (VTT Biotechnology, Espoo) was used as a framework to construct the new Gateway-phagemid. Fab-phagemid is a derivative of pBluescript SK+ phagemid vector (Stratagene). The circle map of pBluescript SK is shown in Figure 9. The phagemid vector contains two origins of replication: pUC ori for plasmid replication in *E. coli* and f1 ori for phage replication. It contains β-lactamase gene (Amp^R), which confers resistance to ampicillin and carbenicillin, and the polylinker sequence (MCS) which is under the control of *lac* promoter (Plac).

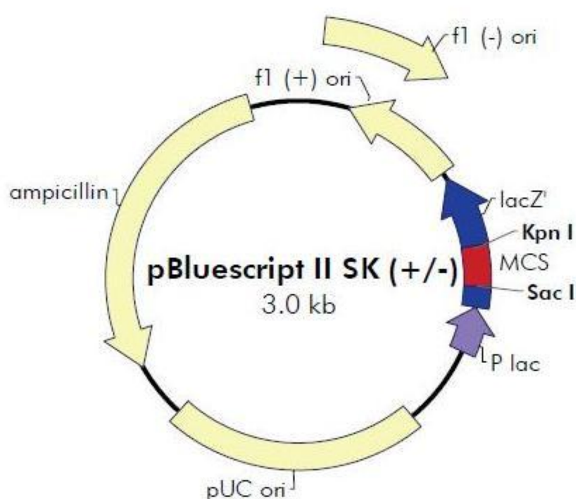


Figure 9: pBluescript SK (+/–) phagemid circle map. Fab-phagemid is based on SK+ vector and thus contains f1 (+) ori for phage replication. (Modified from Stratagene catalog #212205.)

VTT phagemid contained a cloned insert including a Fab coding region, which lies between the coding sequences of pelB signal peptide and the C-terminal part of the pIII protein (aa 198–406). The construct has been inserted into MCS of pBluescript. Fab-phagemid sequence contains some important restriction sites which enable the modification of the insert (see Figure 10). EcoRI and HindIII sites are flanking the pelB-Fab-pIII coding region. NotI site is preceding the gene III and can thus be used for fusion cloning with the C-terminal pIII coding region. Double digestion with EcoRI and NotI enables the removal of pelB-Fab coding region from the Fab-phagemid. The

coding region of pelB-Fab can be replaced by the Gateway-cassette to construst the new destination vector (pGWphagemid) for Gateway cloning. Since the removal of pelB signal sequence, the desired signal sequence has to be included in the entry clone.

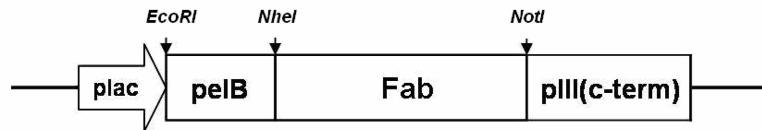


Figure 10: Organization of pelB-Fab-pIII construct in the Fab-phagemid.

4.1.2 pGEM-T Easy

pGEM-T Easy vectors harbouring attL-flanked ompA-AVR constructs were was used as starting material in the amplification of parental cDNAs for DNA shuffling. Each template plasmid contained one of the parental genes (wt *AVD*, *AVR2* and *AVR4*) as an insert and the cloning of template constructs into MCS has been prepared in previous studies (Eisenberg-Domovich *et al.*, 2005; Hytönen *et al.*, 2005b). Each insert contained an ompA signal sequence preceding the AVD or AVR coding sequence and the whole insert is flanked by *att*L-sequences. M13 forward and reverse primer binding sites lie slightly outwards from T7 and SP6 sites. The orientation of the insert was reversed, leading to the use of M13F as reverse primer and M13R as forward primer. The circle map of the original pGEM®-T Easy vector (Promega) is represented in the Figure 11. The vector contains a resistance gene to ampicillin and *lac* promoter as well.

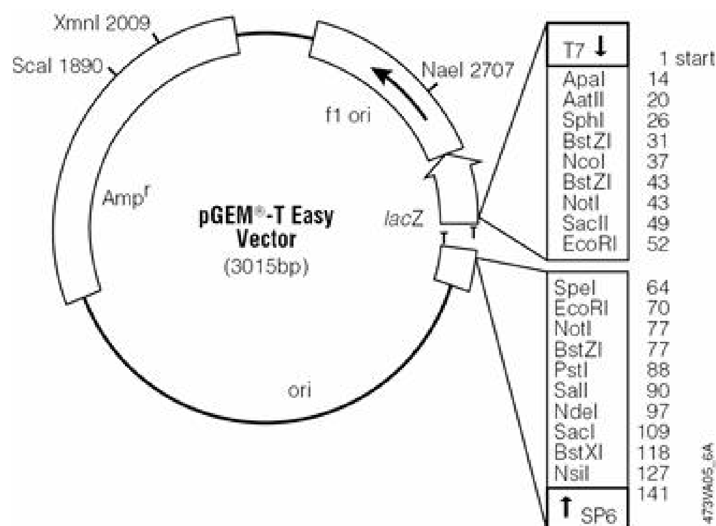


Figure 11: pGEM®- T Easy vector circle map. Binding sites for M13 primers lie outwards from T7 and SP6 sites and the AVD or AVR inserts in between of them. (Figure modified from Promega technical manual #042.)

4.1.3 Multiplication of plasmids

To multiply the plasmids containing the desired insert, plasmid DNA was tranformed into suitable *E. coli* lineage. Depending on whether the cells were chemically competent or electrocompetent, two different methods were used in transformation experiments. Chemically competent cells were transformed by standard heat shock method. Competent cells (50 µl) were thawed on ice and gently mixed with 10 ng (in a volume of 1 to 2 µl) of plasmid DNA. Reactions were incubated on ice for 30 minutes and after that heat shocked at 42 ºC for exactly 30 seconds. Transformed cells were placed on ice and 250 µl of prewarmed S.O.C. medium was added to them. In the case of electrocompetent lineages, the cells were transformed by conventional electroporation method. 40 µl of competent cells were thawed on ice and gently mixed with 2 µl (10–600 ng) of plasmid DNA. Reactions were incubated on ice for 5 minutes and electroporated in 2 ml cuvettes (2.5 V, 200 Ohms, 125 µFD, Bio-Rad Gene Pulser®). After the electric shock, 960 µl of prewarmed S.O.C. medium was instantly added to the cells.

After transformation, cells were then grown in 37 ºC shaker (225 rpm, 1 h), plated (10 – 100 µl of transformation reaction per LB plate containing appropriate antibiotics and 0.1 % (w/v) glucose) and incubated at 37 ºC o/n. Following day, the colonies were counted to calculate the transformation efficiency. Original Fab-phagemid and/or pUC19 (10 pg/µl, Invitrogen) were used as positive controls in all transformation reactions.

For conventional small-scale purification, plasmids were amplified in small (5 ml) LB cultures of sufficient *E. coli* lineage. Cultures were inoculated as single colonies from fresh transformation plates and growed o/n in 37 ºC shaker. LB was supplemented with 0.1 % glucose and appropriate antibiotics. Following day, o/n cultures were pelleted with microcentrifuge (8000 rpm, 2 min) and plasmids were purified as minipreps (GeneJET™ Plasmid Miniprep Kit, Fermentas). For medium-scale purifications, 100 ml of cells were cultured o/n in LB (supplemented with 0.1 % glucose and appropriate antibiotics). Cells were pelleted (SLA-1500 rotor: 5000 g, 15 min, 4 ºC) and plasmid were purified as midipreps (NucleoBond® AX PC100 by Macherey-Nagel or Wizard® Plus Midipreps by Promega).

To multiply and purify multiple clones from AVR2/4 phagemid library, automated plasmid DNA purification protocol was used. The cells were grown on a 96U-well plate (MegaBlock 96 Well, 2.2 ml, Sarstedt) using 7 µl of each bacterial clone (from the glycerol stock master plates) to inoculate 750 µl of TB (containing antibiotics and 0.1 % glucose). The cells were cultured o/n (37 °C, 500 rpm) and pelleted (1000 g, 10 min, 4 ºC). Automated plasmid purification was done by using NucleoSpin Robot-96 Plasmid Kit (Macherey-Nagel) and robot (Genesis RSP 100, Tecan). The robot was controlled using Gemini software (v. 3.40 SP1, Tecan).

4.1.4 Bacterial cell lineages

Depending on the vector and transformation method, different *E. coli* lineages were used for the selection and multiplication of desired construct. Chemically competent TOP10 cells (Invitrogen) were used as standard plasmid propagation lineage whenever the transformation efficiency was not the main objective. Electrocompetent lineages were used whenever the maximal amount of transformants was required.

Propagation of *ccd*B-bearing vectors was done in electrocompetent DB3.1 cells (Invitrogen). These cells contain the *gyr*A462 allele which renders the strain resistant to the toxic effects of the *ccd*B gene. Transformation and cultivation of these specialized cells needed some optimization. To ensure the propagation of *ccd*B-bearing plasmids, the cells were best grown at 30 ºC instead of normal 37 ºC. Because the growth rate of *E. coli* rapidly decreases along the temperature, prolonged periods of cultivation was required. The growth media was supplemented with 1.0 % (w/v) glucose.

Electrocompetent XL1-Blue was used in the construction and propagation of phage display library. The strain is tetracycline-resistant and amber-suppressive (supE44). Amber suppressor enables the expression of pIII fusions. The strain also contains an F-episome, that enables the production of F-pilus and co-infection with Helper phage. F-plasmid bearing *E. coli* contain a *ccd*A gene which will prevent negative selection with the *ccd*B gene. Because these strains have increased resistance to *ccd*B, they are not recommended for Gateway cloning. Depending on the phenotype and cultivation conditions, XL1-Blue generates increased *ccd*B background. To estimate the amount of *ccd*B background, the transformation efficiencies of XL1-Blue were compared to those

of TOP10 and DH5α strains. These strains do not contain the F-plasmid and were thus completely killed by *ccd*B-bearing vectors. DH5α. was chosen for the selection and propagation of recombined phagemid library in order to avoid the background and to reach as high transformation efficiency as possible.

In order to compare the efficiency and background levels of different strains, the transformation efficiencies were determined for DB3.1, TOP10, DH5α and XL1-Blue. Transformation efficiency (TE) was calculated using Equation 1:

$$TE = \frac{\#\,colonies}{m_{DNA}\,(\mu g)} \times \frac{V_{total}\,(\mu l)}{V_{plated}\,(\mu l)} = \frac{cfu}{\mu g} \qquad \text{(Equation 1)},$$

where m_{DNA} is the amount of transformed DNA, V_{total} is the total volume of transformation reaction and V_{plated} is the volume which is plated. TE is obtained as number of transformants (cfu) per micrograms of plasmid DNA.

Chemically competent, F-negative BL21-AI (Invitrogen) strain was used in the production of free AVR2/4 proteins. The strain is tetracycline resistant and contains a chromosomal copy of the T7 RNA polymerase gene, which is under the tight control of the arabinose-inducible araBAD promoter. Because pGWphagemid (containing Plac promoter) was used as expression vector, the protein production was induced with 1 mM IPTG to trigger the transcription of the lac operon.

4.2 Recombinant DNA technology

4.2.1 Characterization and purification of DNA

Agarose gel electrophoresis (AGE) was used for identification, quantification and purification of DNA fragments (PCR products and other mixed populations of DNA). The DNA samples were prepared by adding 1:10 of 10x colourless loading buffer. GeneRuler™ 1 kb and 100 bp Plus DNA Ladders (Fermentas) were used as size markers. Samples were separated by their size on 1.5 % agarose gel (containing 1 µl EtBr per 100 ml of 1x TAE-buffer) by applying an electric field of 100 V for 1–1.5 hours (Pharmacia Biotech EPS300). After electrophoresis the gel was UV- illuminated to visualize the DNA bands. The gel was photographed and analyzed using Bio-Rad

Quantity One software (v.4.5.2). In the case of gel purification the desired band was cut out of the gel with scalpel and redissolved to retrieve the purified DNA. The purification of DNA was made using illustra GFX™ PCR DNA and Gel Band Purification Kit (GE Healthcare). DNA concentrations were measured with NanoDrop® (ND-1000) -spectrofotometer (Thermo Scientific).

4.2.2 Construction of pGWphagemid

In order to clone the Gateway-cassette into the phagemid vector, the purified Fab-phagemid was digested with EcoRI and NotI restriction enzymes (Fermentas) in order to remove the pelB-Fab insert. 1 µg of plasmid DNA was digested using 10 units of both enzymes in buffer O with BSA. Digestion reactions were incubated for 15h at 37 ºC and heat inactivated for 20 min at 65 ºC. To prevent the undesirable self-ligation of digested vector, 5'-ends were dephosphorylated using shrimp alkaline phosphatase (SAP, Fermentas). Reactions were incubated for 30 min at 37 ºC and heat inactivated for 15 min at 65 ºC.

Table 1: Oligonucleotides used in the amplification of Gateway-cassette.

Primer	Sequence
gw-phage3'	TTGCGGCCGCTACCACTTTGTACAAGAAAGC
gw-phage5'	AAGAATTCACAAGTTTGTACAAAAAAGCTGAAC
GW_dNotI_3'	GTGCCTAATGCTGCCGCCATAGTG
GW_dNotI	CACTATGGCGGCAGCATTAGGCAC
GW_dEcoRI_5'	GAATGCTCATCCGGAGTTCCGTATGG
GW_dEcoRI	CCATACGGAACTCCGGATGAGCATTC

In the amplification of Gateway-cassette, pTriEx Gateway destination vector was used as a template. The cassette was previously cloned into pTriEx using Gateway® Vector Conversion System (Invitrogen, catalog #11828-029). Importantly, the cassette contained an *att*R-flanked *ccd*B-gene (Figure 12), which was conversed into phagemid vector to make it Gateway-compatible. Chloramphicenol resistance gene (Cm^R) was also included in the cassette, but it was not utilized in this study. The cassette was also found to contain NotI and EcoRI sites in the middle of it. To make it possible to utilize these restriction sites in the construction of the pGWphagemid, these sites were deleted during the amplification of the template by PCR. Primers used for template amplification are listed in the Table 1. Mutative primers are indicated as dNotI or dEcoRI and silent mutations were used to maintain the activity of *ccd*B.

Figure 12: The Gateway-cassette contains two genes flanked by *att*R-sites; *ccd*B for positive selection and Cm^R for chloramphicenol resistance. (Modified from Invitrogen manual, catalog #11828-029.)

To accomplish the desired mutations, the template was amplified in three segments. These three segments were combined in subsequent PCR reaction to amplify the desired PCR-product in its entity (Figure 13). The amplification PCR was done using Phusion Hot Start polymerase (Finnzymes) and 5x Phusion GC buffer. The reaction mixture contained dNTPs (200 µM), 50-60 ng template DNA and 5' and 3' primers (30 pmol each). Reactions contained 5 % DMSO to prevent the annealing of *att*L-sequences with each other.

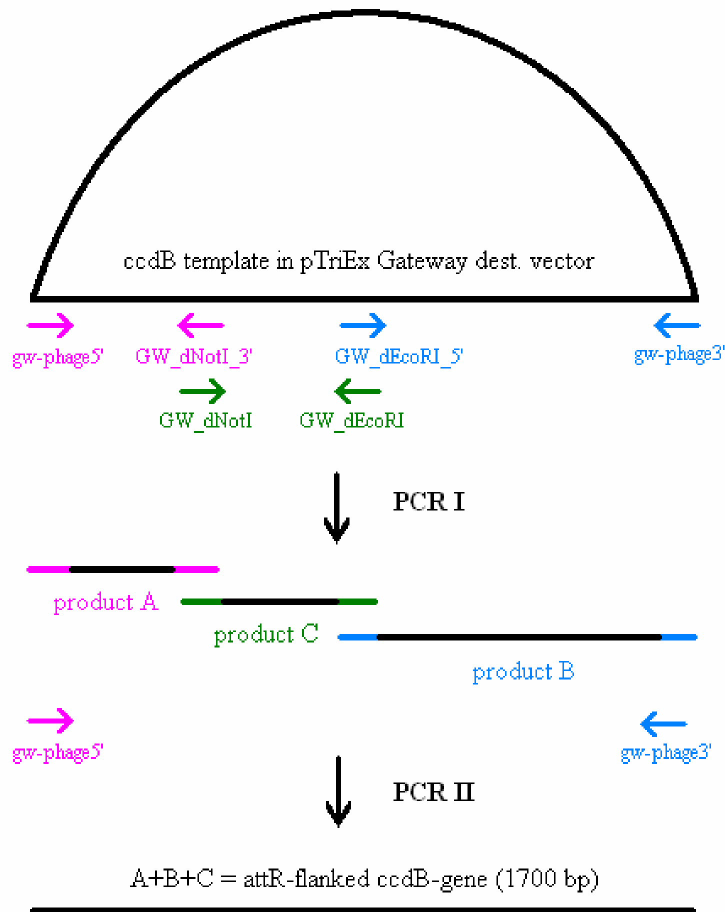


Figure 13: Schematic of the Gateway-cassette amplification. The template was first amplified in three separate fragments (A, B and C) in PCR reaction I. The internal NotI and EcoRI sites were deleted during the amplification. The products were combined in PCR reaction II to obtain the final product, a *ccd*B-gene flanked by *att*R-sites.

The resulting 1700 bp PCR-product was gel purified and digested with EcoRI and NotI, like phagemid vector described above. The Gateway-cassette was then ligated into digested and phosphorylated phagemid vector. The ligation of digested products was done using T4 DNA ligase in 3:1 ratio of insert:vector. Ligation reactions were incubated for 4h at RT and inactivated for 10 min at 65 °C. The desired products of all amplification, digestion and ligation steps were separated with AGE and purified with GFX kit. Ligated pGWphagemids were electroporated into *E. coli* DB3.1 cells and verified by sequencing and control digestions.

4.2.3 Control digestion of phagemids

The idea of control digestion was to check whether the phagemid included a right sized insert. Double digestions were made using several enzymes which have their restriction sites around the insert. EcoRI, NotI, KpnI and HindIII were used as pairs and in different combinations to check their restriction sites. Plasmid samples were purified as minipreps and represented either individual clones or covered the whole AVR2/4 library. 300 ng of plasmid DNA was mixed with 10 units of both restriction enzymes (Fermentas) with convenient buffer (buffer O, buffer R or 1x Tango) and dH_2O into the total reaction volume of 20 µl. Digestion was performed for 15h at 37 ºC and heat inactivated for 20 min at 65 ºC. Reactions were analyzed with AGE.

4.2.4 DNA shuffling

I. Amplification of the parental cDNAs

The pGEM-T Easy vectors containing the AVR2 and AVR4 constructs were multiplicated and used as templates in the first stage of DNA shuffling, but although they already contained attL-flanked AVRs, they were not used as templates as such. Both AVR2 and AVR4 templates contained an extra nucleotide (A) at their ends, which had to be deleted in order to link the coding regions of AVRs to pIII in correct reading frame. The deletion was made by using a 3' *att*L_2-adapter block as a mutative megaprimer. In addition, the universal system for flanking cDNA by attL-sequences was demonstrated.

The three-step PCR setup for template amplification is represented in Figure 14. The 5' and 3' *att*L-adapter blocks were amplified in separate PCR reactions (step I). At the

same time, the deletion was generated into the 3’ adapter block. The AVR cores were amplified in separate PCR reactions as well (step II). AVR cores contained the ompA-AVR region without *att*L sites. Oligonucleotides used in the template amplification are listed in Table 2. Primers ompA-KpnI.1 and C122S.2 were used in the amplification of both AVR cores. 5’ adapter block was amplified with ompA-KpnI.2 and M13R primers. 3’ adapter block was amplified with AVR-*att*L2_5’ (containing the desired mutation) and M13F primers. The AVR cores and adapter blocks were combined (step III) to construct two linear template blocks. In this case, *att*L_1-ompA-AVR2 could have been amplified directly from the original template, since C122S.2 primer excluded the extra nucleotide. Adapter blocks were anyhow used for both templates in order to demonstrate the incorporation of *att*L sites to flank any desired sequence and to produce parental templates of equal quality.

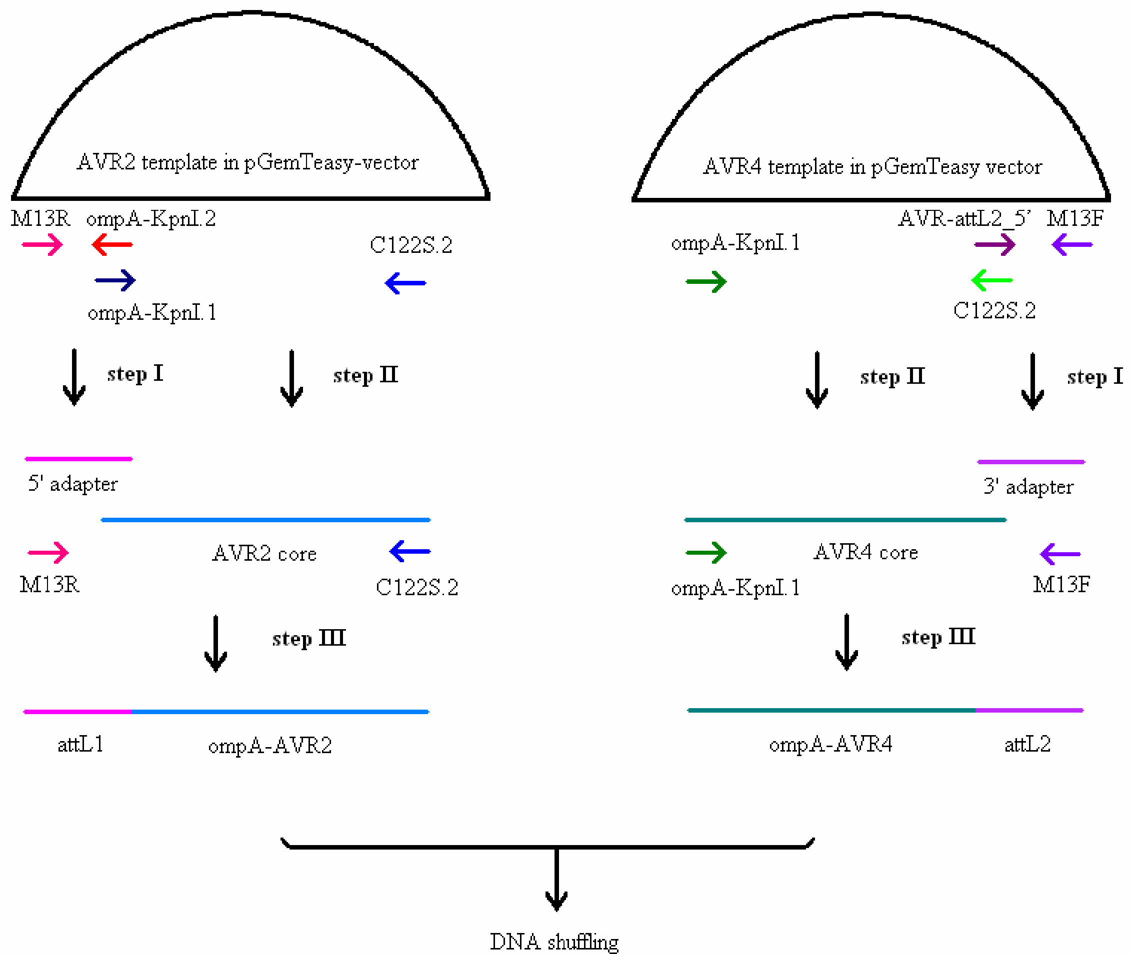


Figure 14: PCR setup for AVR2 and AVR4 template amplification. Adapters are amplified in step I. AVR-cores, which contain ompA signal sequence, are amplified in step II. Both cores are joined to their adapter blocks in step III.

The templates were assembled asymmetrically to minimize the appearance of parental genes in the final AVR2/4 library. The final AVR2 template was amplified with ompA-KpnI.1 and M13F and AVR4 template with C122S.2 and M13R. As a result, both templates contained either one of the two attL-sequences in either end of the AVR cDNA. Hence the LR cloning only succeeds in the case of shuffled AVR inserts containing both attL-cloning sites.

The resulting products from the three-step PCR were amplified with conventional PCR to obtain enough template material into fragmentation step. Pfu polymerase (2.5 u/µl, Fermentas) was used in all amplification reactions. The PCR mixtures contained 10x Pfu buffer with $MgSO_4$, dNTPs (200 µM each), 3' and 5' primers (30 pmol each) and 15 ng of template DNA. To prevent the annealing of *att*L-sequences with each other, 5% DMSO was included in all PCR mixtures. Parameters for amplification PCR are presented in the Appendix 1.

Table 2: Oligonucleotides used in DNA shuffling.

Primer	Sequence
M13F	CGACGTTGTAAAACGACGGCCAGT
M13R	TTTCACACAGGAAACAGCTATGAC
ompA-KpnI.1	TGCCTCGGCTGGTACCGTGGCCA
ompA-KpnI.2	TGGCCACGGTACCAGCCGAGGCA
C122S.2	CCTCCACTGTGGACAGGCGAGTG
AVR-attL2_5'	CACAGTGGAGGAGTAAGGTGGCCCCGCTTTC

II. Fragmentation by DNaseI

The amplified template material was separated with AGE and purified with GFX kit prior to fragmentation step. 1 µg of each template was diluted in DNaseI buffer into final volume of 50 µl and equilibrated at 15 °C. Digestion was performed by adding 0.15 units of DNaseI (Fermentas). Digestion reactions were incubated for 2-3 min at 15 °C and heat inactivated for 10 min at 90 °C. Resulting fragments were directly used in subsequent reassembly reaction without any purification.

III. Reassembly by homologous recombination

Fragments were reassembled in primerless PCR of 40 cycles. Fragmented template DNA was mixed 1:1 with PCR premix. PCR premix contained 1.25 units of Pfu

polymerase (Fermentas), 10x Pfu buffer with $MgSO_4$ and dNTPs (200 µM each). DMSO was not used. Parameters for reassembly PCR are presented in the Appendix 1. In each step of DNA shuffling, several parallel reactions were made to ensure the gain of final products. To ensure the success of the shuffling procedure, samples from each step were analyzed with AGE. In addition to Pfu, Phusion polymerase (Finnzymes) was also tested in similar reassembly reactions due to its high fidelity and processivity.

The reassembled products were purified from the PCR mixture with GFX kit and used as entry clones in LR clonase reaction. The unpurified reassembly reaction was containing only small proportion of desired *att*L-flanked AVRs and thus the amount of DNA in the subsequent cloning step had to be optimized.

4.2.5 Gateway cloning

Gateway® cloning system (Invitrogen) was used in the construction of phagemid library. The LR cloning procedure was first tested using *att*L-flanked wt AVD, AVR2 and AVR4 cDNAs as entry clones. Finally the LR clonase reaction was used to clone the shuffled AVR2/4 variants into pGWphagemid directly from the purified reassembly reaction. Reaction is based on homologous recombination between attL sites of the entry clone and *att*R sites of the destination vector (*att*L × *att*R → *att*B + *att*P) and it was performed using LR clonase II (Invitrogen).

Each LR cloning reaction contained 150 ng (in the volume of 1 µl) of destination vector (pGWphagemid) and 50–150 ng (in the volume of 1–7 µl) of entry clone. In the LR cloning reaction of reassembled AVR2/4 cDNA, the amount of entry clone used was significantly higher (400-600 ng) due to the presence of by-products of the reassembly reaction. Reactions were filled to the volume of 8 µl with TE-buffer (pH 8) and 2 µl of LR clonase II was added. Reactions were mixed carefully, spinned down and incubated at RT o/n. Following day, reactions were terminated by adding 1 µl of Proteinase K and incubating for 10 min at 37 ºC. The reactions were transformed to *E. coli* to multiply and purify the plasmids. The presence and condition of the AVR insert was verified by sequencing and control digestions.

4.3 Construction of phage display library

The phagemid library was constructed by one simple cloning reaction, as described above. In the cloning reaction, the eligible AVR2/4 variants resulting from DNA shuffling were transferred into phagemid vectors by site-specific recombination. To express the phagemid library in bacteria, the LR cloning reaction was transformed into *E. coli*. To avoid the *ccd*B background in XL1-Blue, transformation of the library was done in two subsequent steps. The phagemids were first electroporated into DH5α and after plasmid purification electroporated into XL1-Blue in order to create the final phage display library.

Transformation of the whole LR clonase reaction (10 µl, containing 150 ng/µl plasmid DNA) into DH5α was done in eight parallel electroporation reactions. After 1 h growing 50 µl of each reaction was plated on LB_{amp} plates. Rest of the reactions was combined into 20 ml SB culture (100 µg/ml amp, 0.1 % glucose) and grown for 4 hours at 37 ºC shaker. 10 ml of the culture was stored as glycerol stocks and 10 ml was used for the plasmid purification. The total volume of 150µl of purified plasmids was ethanol precipitated to concentrate the DNA. The precipitate was resuspended to 20 µl of ddH_2O and the final concentration of purified plasmids was 84.5 ng/µl. The whole amount of plasmid DNA was electroporated into XL1-Blue in eight parallel reactions. Again 50 µl of each reaction was plated on LB_{amp} plates and the rest was combined and grown in 20 ml SB culture. 10 ml of the culture was again stored as glycerol stocks and 10 ml was used for the plasmid purification.

The resulting colonies on LB_{amp}-plates were counted to determine the transformation efficiency and the size of AVR2/4 library. To analyze the clones in detail, colony picking was done from both DH5α and XL1-Blue plates. 2x 96 colonies were picked from both cell lineages and inoculated into 100 µl of master medium (SB, antibiotics, 1.0 % glucose, 1x Freezing buffer) on 96-well plates (Microtest Plate Round Bottom, Sarstedt). The final concentration of glycerol in the master medium was 4.4 %. Plates were incubated for bacterial growth o/n (28 ºC, 500 rpm) and stored at –80 ºC. To make glycerol stocks covering the whole AVR2/4 library, 10 ml of both DH5α and XL1-Blue cultures were stored in 15 % glycerol (–80 ºC).

4.4 Sequencing

The sample preparation before sequencing involved the amplification of the desired product with sequencing PCR and the purification of PCR product with ethanol precipitation. Each sequencing PCR reaction contained 100–300 ng of plasmid DNA (diluted to ddH_2O to the volume of 7 µl), 0.6 µl of sequencing primer (10 µM), 0,7 µl of BigDye enzyme mix (BigDye® Terminator v3.1 Cycle Sequencing RR-100, Applied Biosystems) and 1,7 µl of BigDye Dilution buffer (5x; ABI). The sequencing primers used are listed in the Table 3 and the parameters for sequencing PCR are listed in the Appendix 1.

The ethanol precipitation was made in two subsequent steps. 25 µl of 94 % ethanol and 1 µl of 3 M sodium acetate was added to each 10 µl sample and incubated for 15 min at RT. Samples were centrifuged at full speed (16 000 g, 20 min, or in the case of 96-well plates 3500 rpm 30 min). Precipitate was washed with 100–200 µl of 70 % ethanol and centrifuged (16 000 g, 10 min or 3500 rpm, 10 min). The precipitate was dried at RT and resuspended to ice-cold HiDi-formamide (Applied Biosystems). The samples were denaturated (92 °C, 2 min), cooled on ice and spinned down briefly. The sequencing was done with ABI-Prism™ 3130xl Genetic Analyzer (Applied Biosystems) -sequencer and analyzed automatically by the Sequencing Analysis Software (v. 5.2, Applied Biosystems). The obtained DNA sequences were analyzed using DNAman software (v. 4.11, Lynnon BioSoft).

Table 3: Oligonucleotides used in sequencing PCR reactions.

Primer	Sequence
Phagemid 1867 (forward)	ACTCATTAGGCACCCCAGGC
Phagemid 886 (reverse)	ATTATCACCGTCACCGACTTCAGC
M13F	CGACGTTGTAAAACGACGGCCAGT
M13R	TTTCACACAGGAAACAGCTATGAC

4.5 Protein production, purification and characterization

The biotin-binding ability of AVR2/4 library was verified by two different approaches. The affinity of AVR2/4-pIII-fusions towards biotin was measured with protein-ELISA. All sequenced clones were analyzed in the microplate format. The functionality of

shuffled AVR2/4 was also tested by affinity-chromatography. Protein obtained from medium-scale *E. coli* production was purified with 2-iminobiotin affinity chromatography in order to test how the protein is produced and whether it can bind to a biotin analogue.

4.5.1 Protein ELISA

To produce protein mutants of AVR2/4 library, XL1-Blue cells were grown from the glycerol stock master plates. 7 µl of bacterial suspension was used to inoculate 200 µl of SB (10 µg/ml tet, 50 µg/ml amp, 0.1 % glucose) and the cells were grown (37 °C, 900 rpm, 4-5h) on a 96U-well plate (Megablock 2.2. ml; Sarstedt). 50 µl of induction medium (SB, antibiotics, 0.1 % glucose and 5 mM IPTG) was added and the cells were cultured o/n (28 °C, 900 rpm). Following day, cells were pelleted (4000 rpm, 15 min, 4°C) and stored at –20 °C.

The biotin binding ability of shuffled AVR2/4 clones was measured in primary ELISA using biotin-conjugated alkaline phosphatase. The 96-well plates (Maxisorp; Nunc A/S Roskilde, Denmark) were coated (4 °C, 400 rpm, o/n) either with Btn-BSA (JL) or BSA (5 µg/ml) in 1x PBS. 100 µl of the coating solution was used per each well. The coated plates were washed with PBS-Tween (3x 300 µl) and remaining free binding sites were blocked with 5% milk in Tris-buffer (RT, 400 rpm, 30 min) using 200 µl of the blocking solution per well. The washing step with PBS-Tween (3x 300 µl) was repeated before the addition of the protein mutants.

Frozen cells were lysed with EasyLyse reagent (Epicentre Biotechnologies). The cell pellet was resuspended into 40 µl of EasyLyse solution in shaker (RT, 800 rpm, 10 min). 80 µl of Tris-HCl buffer (pH 7.5) was added to dilute the lysate in appropriate volume and the lysate was centrifuged (5 min, 4500 rpm, 4 °C) to pellet the cell debris. 50 µl of the cell lysate supernatant was used per well to bind the protein mutants to the Btn-BSA or BSA surface (RT, 400 rpm, 1 h). Empty wells (coated but no protein added) were used to determine the level of background signal. The washing step (3x PBS-Tween) was repeated and 100 µl of biotinylated AP (diluted 1:5000 in 5% milk in Tris-buffer) was added per well and incubated (RT, 400 rpm, 30 min). The washing step (3x PBS-Tween) was repeated and the Btn-AP was detected using PNPP (1 mg/ml in

DEA). 100 µl of detection solution was added per well and A_{405} was measured in every 15 min for one hour (Bio-Rad Model 680 XR Microplate Reader).

4.5.2 Affinity chromatography

For medium scale production of free AVR2/4 protein, one AVR2/4 clone (A1, verified by sequencing), was transformed into BL21-AI cells. Transformation was performed by standard heat shock method and reactions were plated on $LB_{amp,tet}$-plates. One colony was picked and transferred to 10 ml of LB (10 µg/ml tet, 50 µg/ml amp, 0.1 % glucose) and the cells were cultured o/n (37 °C, 225 rpm). Following day, the cultures were diluted to 500 ml of LB (100 µg/ml amp, 0.1 % glucose) and grown (28 °C, 150 rpm) until OD_{600} = 0.4. Induction was done by adding IPTG to the final concentration of 1mM. After induction, cells were cultured o/n (28°C, 150 rpm) and collected by centrifugation (10 min, 5000 rpm/GSA rotor, 4 °C).

To resuspend the cell pellet, 4 ml of SET-buffer and 50 µl of lysozyme was added per 100 ml culture volume and incubated on a magnetic stirrer (4 °C, 30 min). 10 ml of HilloI-buffer was added per 100 ml of culture and the cell suspension was sonicated using amplitude 60 % for 5 min (5s on / 1s off) (Vibra Cell 500 VCX, Sonics & Materials, Inc.). Sample T (100 µl) was taken from the sonicated cell suspension. The lysate was centrifuged (30 min, 9600 rpm, GSA rotor, 4 °C) to pellet the cellular debris. The supernatant was filtered through Miracloth and sample L1 (100 µl) was taken. The cell debris was suspended to 50 ml of PBS and sample S (100 µl) was taken. An equal volume of pH 11 buffer was added to the filtered supernatant and pH was adjusted to pH 11 using pH-meter and 10 M NaOH.

The agarose (2-iminobiotin Sepharose™, 4 Fast flow, Affiland) was washed in an affinity chromatography column (Bio-Rad, 25 ml) using 20 ml of pH 11 buffer. The L1 supernatant was divided into 50 ml Nunc bottles and the washed 2-iminobiotin resin was shared to the bottles equally. The L1- resin mixture was incubated on rolling shaker (4 °C, o/n) in order to bind the protein to the 2-iminobiotin resin. The resin was collected by centrifugation (1400 rpm, 5 min). The supernatant L2 was carefully removed and stored (a 100 µl sample L2 was taken). The resin was washed twice using pH 11 buffer and transferred into column and washed using 30 ml of pH 11 buffer. The protein was eluted using pH 4 buffer taking 10x 1 ml fractions. Since the protein was

not completely eluted from the column, additional elution was needed. Six more fractions were taken using pH 4 buffer, continued by acetic acid elution (6ml 0.1 M; 6ml 0.2 M, 3 ml 0.5 M and 3 ml 1.0 M acetic acid). The best yield was achieved by 0.1 M acetic acid elution (fractions 17-22). A 20 µl sample was taken of each fraction to be used for SDS-PAGE analysis later. A_{280} of each fraction was measured using NanoDrop® (ND-1000) -spectrofotometer and the protein concentration was calculated using Lambert-Beer law (Equation 2):

$$A = \varepsilon c b \qquad \text{(Equation 2)},$$

$$c = \frac{A}{\varepsilon}$$

; where A is the absorbance (280 nm), ε is the molar extinction coefficient and b is the length of the light pathway (1 cm). To obtain the protein concentrations in mol/l, the absorbance was divided by the molar extinction coefficient. The result was divided by the molar mass (M) of monomeric protein to obtain the protein concentrations in mg/ml. For AVR2/4; $\varepsilon = 25105\ M^{-1}\ cm^{-1}$ and M = 13965,6 g/mol. Parameters were calculated from the protein sequence of AVR2/4 (clone A1) using ExPASy Proteomics server (Swiss Institute of Bioinformatics).

4.5.3 Protein analysis methods

The protein purification samples were analyzed using reducing SDS-PAGE. The protein samples and fractions taken during the purification were separated by electrophoresis using 5 % stacking gel and 15 % resolving gel. The samples were prepared by adding 1:1 denaturing 2x SDS-PAGE sample buffer (containing 2 % β-mercaptoethanol). The protein samples were boiled for 10-15 minutes, because AVRs are often found to be more heat-stable than wt AVD. PageRuler™ Prestained Protein Ladder Plus (Fermentas) and wt AVD (1 mg/ml, Belovo S.A., Belgium) were used MW markers. In addition, recombinant AVR4 (produced in *E. coli*) was used as a positive control. The electrophoresis was run in 1x SDS-PAGE buffer (80 V, 15 min for stacking and 180 V, 45 min for resolving, Bio-Rad) for two identical gels in parallel. Proteins were visualized by staining with Coomassie Brilliant Blue from the other gel. The other gel was used for western blotting.

For specific immunodetection of AVR protein, the samples were transferred from the SDS-PAGE gel to the nitrocellulose membrane by western blotting (100 V, 1 h, 4 °C). The immunodetection was done by using monoclonal mouse anti-AVR4 IgG (1:20 in 5 % milk-TBS-Tween) as a primary antibody and goat anti-mouse IgG-AP (1:30000 in 5 % milk-TBS-Tween) as a secondary antibody. Polyclonal anti-AVD IgG was also tested for the immunodetection, but shuffled AVR2/4 could not be recognized by it. The membrane was washed three times using TBS-Tween between each step. Prior to detection, an additional wash was done using APA2-buffer. Alkaline phosphatase activity was detected using 1 % NBT and 1 % BCIP-T diluted in APA2-buffer. The detection solution was prepared by adding 33 µl BCIP-T (Fermentas) and 44 µl NBT (Fermentas) into 10 ml of APA2-buffer.

4.6 Computer programs

The results from microplate sequencing (covering 96 AVR2/4 clones) were analyzed on both DNA and protein level. DNA sequences were translated using Perl script avr4pIII_script_.pl (created by Sampo Kukkurainen). The script was used to translate seq-files (from Sequencing Analysis Software) in all three reading frames and compare them to AVR4-pIII reference sequence. All protein sequences which matched to the reference sequence were gathered together as a fasta-file, and corresponding DNA-sequences were simultaneously written in their own fasta-file. Resulting protein sequences were aligned using ClustalW alignment algorithm (EBI web server). Alignments were further analyzed using GeneDoc (v. 2.7.0) and MEGA4 (Tamura *et al.*, 2007) software.

The evolutionary relationships of sequenced clones were analyzed using different methods (UPGMA, neighbour-joining, minimum evolution; with or without bootstrap testing). All phylogenetic analyses were conducted in MEGA4 from the aligned sequence data. The circle map of pGWphagemid was drawn using SimVector 4.5 software (Premier Biosoft). The figures were modified or prepared using Photoshop CS3 or MS Office PowerPoint.

5. Results

5.1 Construction of pGWphagemid vector

The original Fab insert was successfully replaced by the Gateway-cassette. The internal restriction sites were removed by PCR-amplifying the cassette (containing a *ccd*B-insert flanked by attR sites) in three segments. The segments were combined to form a 1700 bp PCR product (AGE analysis not shown). The resulting Gateway-cassette was subcloned into phagemid vector using EcoRI and NotI sites, thus replacing the pelB-Fab sequence.

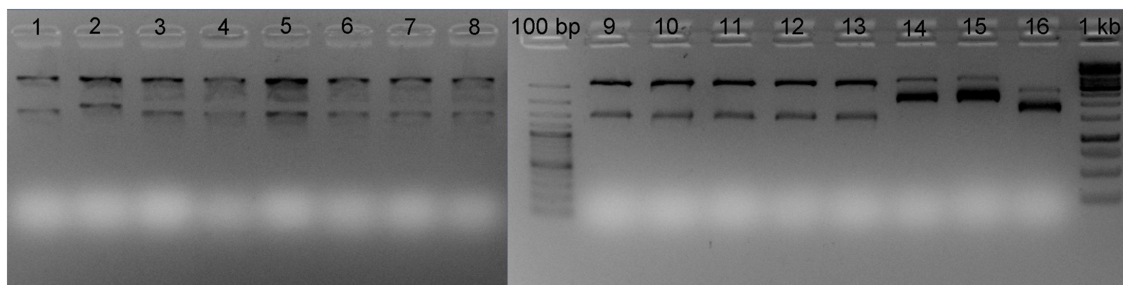


Figure 15: Digestion of pGWphagemid minipreps 1-16. Reactions 1-13 contain the insert of approximately right size (1700 bp). The insert cleaved from pGWphagemid 2 was the only one containing complete Gateway-cassette. It is slightly larger compared to other cleaved inserts on gel.

Transformation efficiency of pGWphagemid was quite low in DB3.1 (max. 10^3 cfu/µg) but sufficient enough to obtain the desired product. From the resulting clones, 16 were verified by sequencing. Because of the large size of the insert, it was verified by performing both forward and reverse sequencing PCR reactions. The clone number 2 was the only clone containing a complete Gateway-cassette (with *ccd*B-gene and *att*R sites). It contained the desired mutations (deleting internal NotI and EcoRI sites) and one spontaneous point mutation (C→T transition), which was silent because GTC and GTT both encode valine. This clone (pGWphagemid #2) stood out as well in the control digestion of pGWphagemids (Figure 15). It was multiplied and used as destination vector in subsequent LR cloning reactions. After LR cloning, the phagemids were again analyzed by control digestions (EcoRI, NotI, HindIII and KpnI) to verify the right size of the insert (AGE analysis not shown). Control digestions confirmed the correct size and organization of insert both before and after LR cloning.

The overall organization of the pGWphagemid is shown in the circle map (Figure 16). The constructed vector contains the Gateway-cassette and pIII coding region in the following order: EcoRI-attR$_1$-*ccd*B-attR$_2$-NotI-pIII-HindIII. In addition to *ccd*B, *att*R-flanked Gateway-cassette also contains a resistance gene for chrloramphicenol, which was not used in this study but may be utilized for selection. In addition to the Gateway-cassette, the vector contains f1(+) and pUC replication origins and ampicillin resistance gene. The most important restriction sites EcoRI, NotI and HindIII (verified by several control digestions and AGE analysis of the LR cloned inserts) are indicated in the Figure 16. Obviously they are not any more needed in the Gateway cloning, but are rather useful in the verification of the LR-cloned inserts.

Figure 16: pGWphagemid circle map. The map was drawn from the sequence of the constructed vector using SimVector 4.5 software (Premier Biosoft). In addition to *ccd*B, Gateway-cassette contains a chloramphicenol resistance gene, which is not shown in this figure.

5.2 DNA shuffling and Gateway cloning of the AVR2/4 library

The analysis of parental AVR2 and AVR4 inserts revealed that the reading frame was disrupted by an extra nucleotide in the pIII-linker region. The extra nucleotide was deleted by megaprimer PCR and the since then AVR2 and AVR4 templates were amplified as linear fragments. The use of asymmetric templates *att*L_1-AVR2 and AVR4-*att*L_2 ensured the elimination of parental shuffling products in the LR cloning reaction.

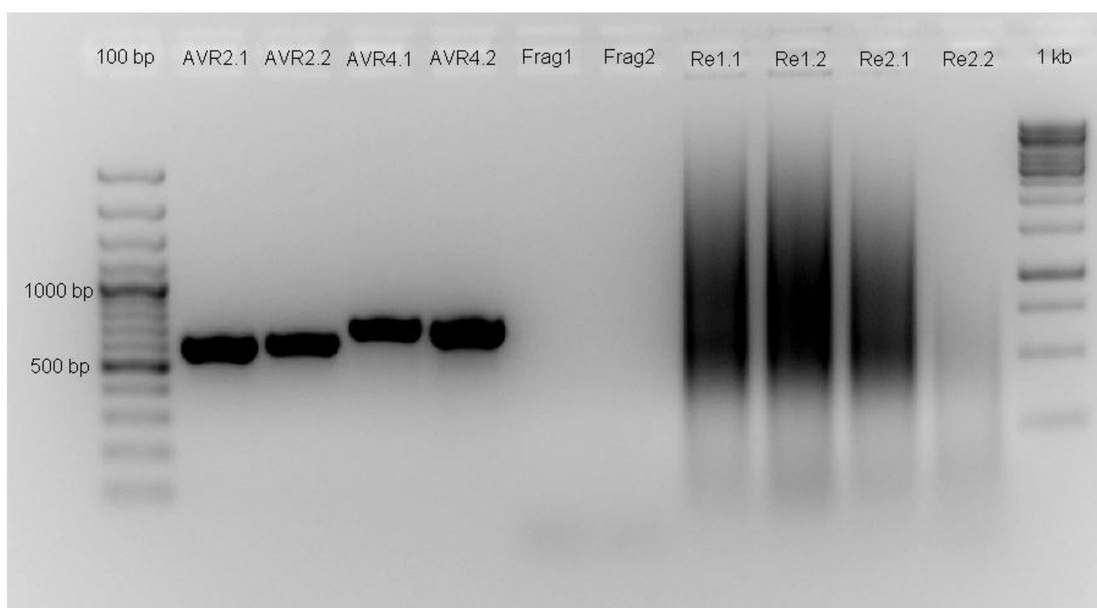


Figure 17: Subsequent steps of DNA shuffling. Amplified AVR2 and AVR4 templates are shown as distinct bands (600-700 bp in size). AVR4 template is slightly larger than AVR2 because of the 3' flanking sequence, which was longer than 5' flanking sequence in AVR2. Fragmented templates can not be seen clearly because of their small size (~50 bp). Reassembled products can be seen as smears of approximately right size.

The consecutive steps of DNA shuffling were confirmed by AGE analysis (Figure 17). In general, Pfu was used as a polymerase in all steps, but the performance of Phusion polymerase was also tested in reassembly reactions due to its high fidelity and processivity (AGE analysis is shown in Appendix 4 for comparison). The size of amplified AVR2 and AVR4 templates was expected to be 600-700 bp. Both templates comprised of AVR core (500 bp, ompA signal sequence included) and an *att*L-block linked to the one end of the core. The AVR4 template was slightly larger compared to AVR2 because of the longer 3' flanking sequence (see Figure 17). After DNaseI digestion, template fragments (approximately 50 bp) were not clearly visible on gel because of their small size and amount. The final reassembly step of DNA shuffling

resulted in a large amount of products in various sizes, which can be seen as smeared bands on gel. It seems on the gel that reassembly reactions (Re1.1 and Re1.2) made from fragmentation reaction 1 (Frag1) were more successful than those made from Frag2. The difference may be due to the length and amount of fragments or the amount of DNA polymerase or $MgSO_4$ in the reassembly PCR. Based on the AGE analysis, reassembly reactions 1.1, 1.2 and 2.1 were chosen to be used as entry clones in the subsequent LR cloning.

Reassembly reactions were directly used as entry clones in Gateway cloning, without purification from gel and/or PCR amplification. As can be seen on gel, reassembly reactions contained large amounts of undesired products. They were eliminated in the LR cloning, since the LR clonase only acts on *att*L-flanked products. Several parallel LR cloning reactions with different amounts of entry clones were performed, but only one of them was used in the final construction of the AVR2/4 library.

5.3 Construction of phage display library

5.3.1 Transformation efficiencies in *E. coli*

Five different *E. coli* strains were used for the selection and multiplication of plasmids bearing the desired constructs. To ensure the quality and size of the final AVR2/4 library, four of these cell strains were tested in electroporation experiments to compare their transformation efficiencies and background levels. Transformation efficiency of BL21-AI was not tested because the strain was only used for protein production after the construction of AVR2/4 library. The *ccd*B-background was estimated by electroporation tests in order to compare the transformation efficiencies between *ccd*B-bearing pGWphagemids and neutral Fab-phagemids. The transformation efficiencies of DB3.1, TOP10, DH5α and XL1-Blue are listed in Table 4.

Table 4: Transformation efficiencies of *E. coli* strains.

	Transformation efficiency (cfu/μg)		
Cell strain	pGWphagemid	Fab-phagemid	AVR2/4 library
DB3.1	10^3	10^6	not determined
TOP10	0	10^6	10^3
DH5α	0	10^7	10^6
XL1-Blue	10^4	10^7	10^6

The cell strains used in the construction of phage display library are shown in Figure 18. Concidering that DB3.1 is *ccd*B-competent strain, its transformation efficiency with pGWphagemid was a quite low (1.6 x 10^3 cfu/µg) compared to that of XL1-Blue. This is likely to be affected by the low quality of DNA in ligation reactions, which were used to transfer DB3.1 cells. Other strains were transformed using purified plasmids. As was expected, F-positive XL1-Blue generated quite high *ccd*B-background levels compared to TOP10 and DH5α, which were not able to propagate pGWphagemid at all. To eliminate the *ccd*B-background in the final phage display library, phagemids resulting from recombination reaction were first electroporated into DH5α strain for plasmid purification and only then transformed into XL1-Blue.

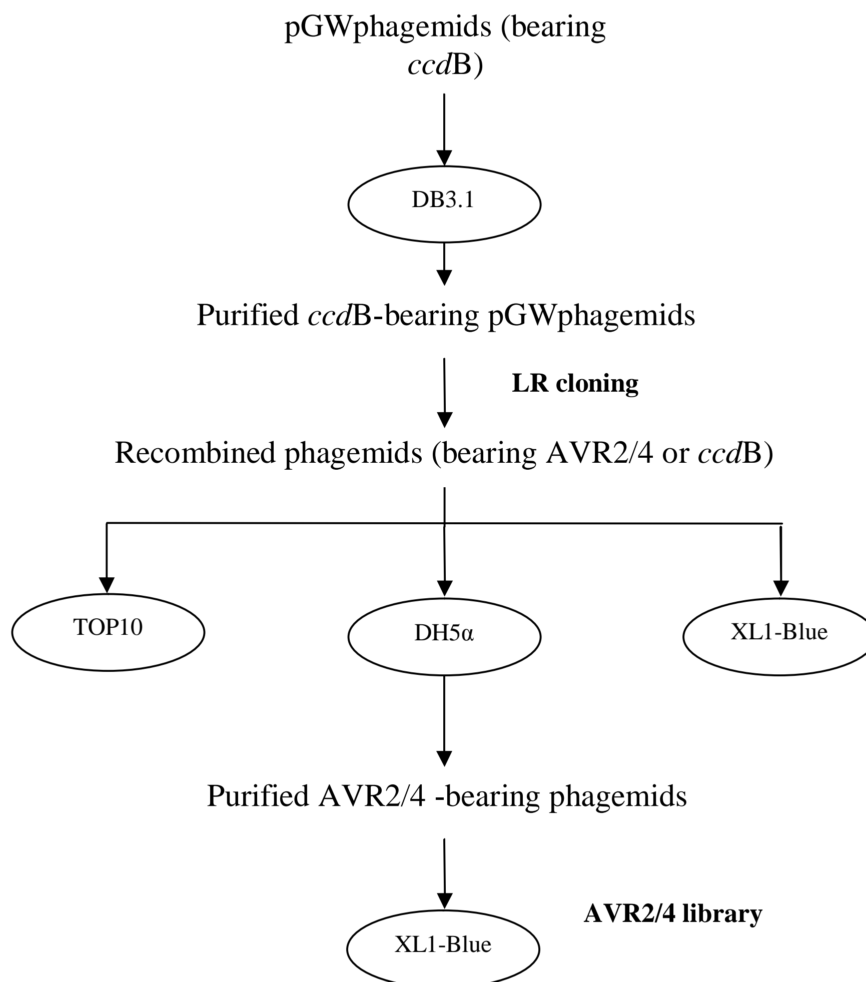


Figure 18: The selection and propagation of phagemid vectors in four different *E. coli* strains.

5.3.2 Library size

DH5α was chosen for the phagemid library multiplication because electroporation is more efficient method compared to standard heat shock transformation of TOP10. Finally, the multiplied and purified phagemids were electroporated into XL1-Blue, which could be used for phage display. The transformation efficiency was about 10-fold lower in the case of unpurified LR cloning reactions compared to test electroporation experiments performed with purified plasmids (see Table 4). The parallel electroporation reactions were the most efficient way for to transform the LR cloned AVR2/4 phagemid library into *E. coli*. The amount of pDNA in each reaction was quite low (30 ng) because only one LR reaction (containing 150 ng of destination vector) was used for library construction. The transformation efficiency was 1.35 x 10^6 cfu/µg in DH5α and 1.5 x 10^6 cfu/µg in XL1-Blue. The library size was approximately 10^6, based on the colony counting from XL1-Blue plates.

5.4 Functional analyses of the AVR2/4 library

Protein expression and biotin-binding affinity of unselected AVR2/4 mutants were verified using two methods, protein ELISA and affinity purification. The functional data obtained from protein ELISA was compared to the sequence data.

5.4.1 Protein ELISA

The production and biotin-binding affinity of 96 sequenced clones from AVR2/4 mutants were analyzed in protein ELISA. Any kind of screening was not done, but the clones were randomly picked from the AVR2/4 library. Because of ochre stop codon, only avidin-like proteins were produced in XL1/Blue cells in microplate format. The periplasmic protein was released by lysing the cells, resulting in protein concentration variation between the clones (depending on the production levels). Protein mutants were bound to biotin-BSA coated surface and could be detected using biotinylated AP and PNPP as signal molecule. BSA coated surface was used to evaluate the amount of background signal.

To evaluate the actual affinities of different clones, the affinity towards biotin was compared to background affinity towards BSA, for each clone individually. Raw data

consisted of absorbancies (A_{405}) measured at 60 min timepoint. The BSA-binding signals were substracted from the signals of biotin-binding to evaluate the actual affinities towards biotin. The results are not comparable with each other, because the production levels vary between different clones. The analysis only implies whether these clones are able to bind biotin or not. The percentage of the functional clones among the shuffled ORFs is quite high (53 %), because out of 49 shuffled ORFs, 26 gave positive signal. It seems that the functionality of the proteins encoded by the genes in *AVR2/4* library is well retained, because so many positive clones were obtained without any kind of screening. Signals were considered moderate, if the biotin-binding of the clones was in the range of 1.5 to 2 –fold compared to the BSA-binding. Higher biotin-binding signals range from 2.5 to 9 –fold compared to the BSA-binding. The biotin-binding was considered significant (indicated by red in the Figure 19), if the signal was higher than background signal + 3×standard deviation (SD).

The results of protein ELISA are shown in Figure 19 as numbers (indicating the biotin-binding affinity) and combined with colour-code which indicates the sequence of each clone. The comparison of sequences with functional data shows that all biotin binding clones are encoded by the shuffled ORFs. Sequence was not obtained for some of them, but they are assumed to be encoded by shuffled ORFs as well. As expected, clones containing premature stop codon did not bind biotin. Most of the clones, which sequence was not obtained, did not bind biotin either. The order of clones in the ELISA was similar to the order of sequencing reactions, and thereby the clones can be tracked down easily; the clone C12 in ELISA correspondens to the sequence of AVR2/4_R91 (see Appendix 5 and 6).

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.090	-0.001	0.473	0.214	-0.003	0.000	0.327	0.391	-0.003	-0.001	-0.001	0.001
B	0.405	0.147	0.005	0.007	-0.005	0.023	0.002	0.002	0.067	0.205	0.039	0.037
C	0.000	-0.001	0.432	0.001	-0.007	0.000	-0.001	0.153	-0.002	0.380	0.000	0.483
D	0.000	-0.004	0.381	-0.003	-0.001	0.005	-0.001	0.007	-0.005	0.001	0.080	0.282
E	-0.001	0.002	-0.001	-0.001	-0.002	0.002	0.451	0.127	0.001	0.085	0.005	0.003
F	-0.004	0.003	0.060	-0.001	-0.001	0.366	-0.004	0.012	0.001	0.007	0.000	-0.001
G	-0.013	0.211	-0.010	0.002	-0.013	-0.006	0.230	0.007	-0.008	0.004	-0.009	-0.006
H	0.001	0.007	0.001	0.008	0.002	0.014	0.016	0.015	0.330	0.008	0.002	0.007

Figure 19: Results of protein ELISA. Moderate biotin-binding signals (1.5–2 times BSA-binding) towards biotin are indicated by yellow and pale orange. Higher signals (2.5–9 times BSA-binding) are indicated by bright orange and red. Shuffled ORF encoding clones which do not bind biotin are shaded with pale yellow. Non-functional clones containing premature stop codon are indicated by green and clones which sequences were not obtained are shaded gray.

5.4.2 Affinity chromatography

Production of one AVR2/4 clone was done in BL21-AI to determine its 2-iminobiotin binding capacity and other characteristics. The clone A1 was successfully purified with 2-iminobiotin affinity chromatography. Protein concentrations of fractions were calculated from the measured absorbancies. The total amount of protein was summed from the fractions. Total of 2.6 mg of purified protein was obtained from 500 ml *E. coli* culture. Out of the total amount of protein isolated, 90 % (2.0 mg) was obtained in elution with 0.1 M acetic acid. The amount was adequate for protein characterization and in line with the results obtained with the subsequent SDS-PAGE and western blot analyses. SDS-PAGE analysis indicates that the 2-iminobiotin purification resulted in homogenous protein showing no contaminants. Different oligomeric forms of AVR2/4 can be seen however. After 15 minutes of boiling, the major part of the protein was approximately 15 kDa in size, which correspondences to monomeric AVR. In addition, larger bands indicating either dimeric or signal peptide containing form of AVR protein can be seen. In the immunodetection of western blotted membranes, the AVR2/4 mutant A1 was able to be detected with monoclonal anti-AVR4/5 antibody whereas the polyclonal anti-avidin antibody could not bind to the shuffled AVR2/4 mutant. As can be seen in Figure 20, the monoclonal anti-AVR4/5 antibody recognized tetrameric form of shuffled AVR2/4 mutant much better than monomeric form.

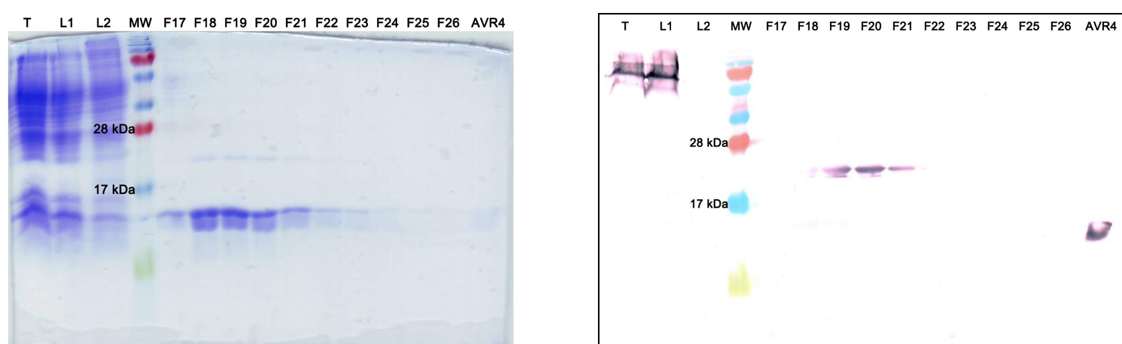


Figure 20: SDS-PAGE and western blotting analyses of the purified AVR2/4 mutant A1. Samples T, L1 and L2, MW marker, fractions 17-26 (eluted using 0.1 M acetic acid) and AVR4 (positive control) are shown. A) Monomeric forms of AVR2/4 can be seen as 15 kDa bands in the SDS-PAGE analysis. Smaller, smeared bands are due to the heat oligomerization and larger bands are either dimeric or signal peptide containing form of AVR2/4. B) In the immunodetection of AVR2/4, the monoclonal anti-AVR4/5 antibody recognized tetrameric and dimeric forms of shuffled AVR2/4 mutant much better than monomeric forms.

5.5 Sequence analysis of the AVR2/4 library

5.5.1 Analysis of 96 sequenced clones

The resulting AVR2/4 library was verified by sequencing. To get outlook of the whole library, 96 clones were analyzed. Plasmids were purified in 96-well plate format and resulting phagemid clones were sequenced in reverse direction. Reverse reactions were made instead of the forward, because the C-terminus and especially AVR2/4-pIII linkage was more interesting than N-terminal ompA signal. In most cases the obtained sequence was covering ompA (or part of it) as well.

In total, 74 % of clones (71 out of 96) were successfully sequenced and matched to the AVR4-pIII reference sequence. The sequences of these 71 clones were aligned using ClustalW and analyzed in detail, both on DNA and protein level. The analysis of the sequences showed that all sequences were shuffled on the DNA level and 97 % of them were chimeric proteins. As was expected, most of the clones were shuffled and thus contained parts from both parents, AVR2 and AVR4. Out of 71 sequences, 54 were unique and 17 clones represented one of the five different amino acid sequences found multiple times (see Appendix 5 and 6). In some cases, the DNA sequences of the identical proteins were divergent, however. The amount of unique sequences was quite high (76 % on protein level and 100 % on DNA level) indicating high diversity in the shuffled AVR2/4 library.

Any truncated or duplicated sequences were not observed, indicating successful shuffling reassembly and LR cloning. The overall structure of the clones was corresponding to the expectations. The clones contained N-terminal ompA signal peptide, which preceded the AVR2/4 insert, a short linker (GGPAFLYKVV) and C-terminal pIII and all these parts were in correct reading frame. Despite of the correction of the reading frame, fusion to pIII was found to be incomplete. The stop codon in the end of the AVR2/4 insert was ochre codon (TAA) instead of the desired amber codon (TAG). This was due to mistake made in the primer design. In addition, 27 of the sequenced clones contained premature TAA stop codon (leucine 99 mutated to ochre stop codon), derived from the AVR2 template. AVR2 template also contained E89K mutation, which was observed in 21 of the shuffled clones as well.

All obtained sequences were analyzed however, in order to get an overview of the DNA shuffling. DNA shuffling typically occurs at the sites of high sequence similarity, but its results can be seen at certain hotspot areas on the protein level. Altogether, there were 19 variable sites in the AVR2/4 sequences: amino acids 9, 11, 13, 21, 25, 30, 34, 75, 78, 83, 86, 94, 97, 99-100, 102, 106, 109, 115 and 117. Different recombination patterns occur as recombination can happen between these any of these hotspots. The effect of recombination events was determined on protein level by counting the number of instances where neighbouring sites are occupied by different parents. The average number of observed crossovers is 2.26(±0.07) per insert, in a range of 0 to 5 (Figure 21). The mutation frequence on the protein level was 0.011 %, because only two spontaneous amino acid changes (I83M and K109N) were observed in two of the 71 analyzed sequences. Template-derived mutations in the AVR2/4 sequence were not considered in the analysis of the amount spontaneous mutations.

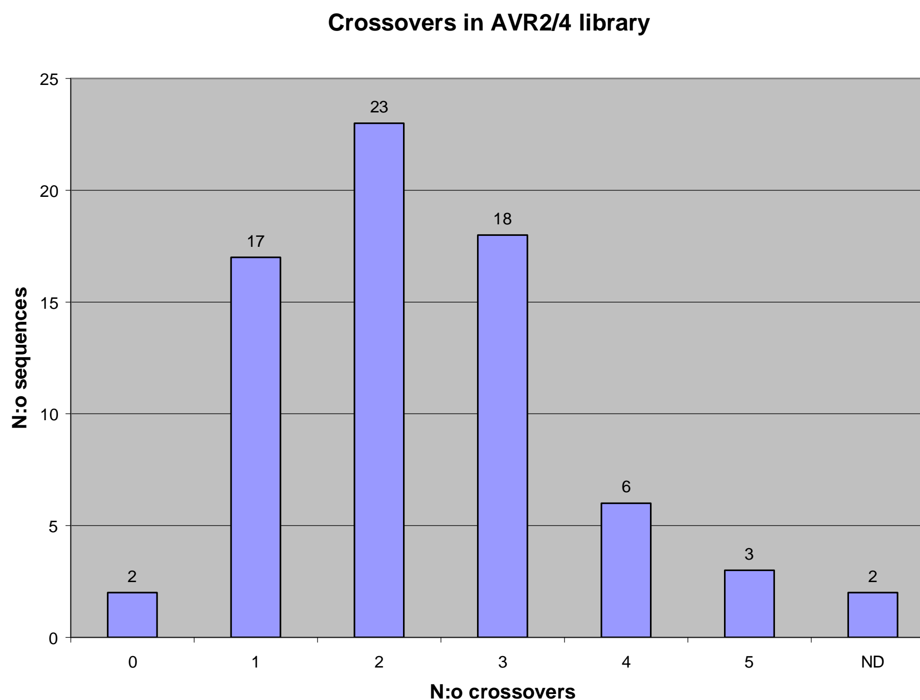


Figure 21: Crossovers observed on the protein level in AVR2/4 library. The amount of crossovers was counted from the amino acid sequence of the AVR2/4 clones as the occupancy of different parents at neighbouring sites. Two of the clones were parental (zero crossovers observed on the protein level) and the majority contained one, two or three crossovers (on average 2.26(±0.07) crossovers per insert). Five was the maximal amount of crossovers in this library and for some clones, the amount of crossover was not able to be determined on protein level (ND).

The multiple sequence alignment of 72 clones showing the variable sites and shuffling patterns is presented in Appendix 5. The multiple sequence alignment of wt AVD,

AVR2 and AVR4 and seven shuffled AVR2/4s is represented in the Figure 22. The alignment shows that residues directly participating to biotin-binding are conserved in all sequences and hence they are not targeted in DNA shuffling. Interface amino acids (methionine and isoleucine in the avidin sequence) instead are divergent in parental AVR sequences when compared to avidin. Thereby they are variable in the the shuffled AVR2/4 library as well and may affect the features of the chimeric clones. All of shuffled AVR2/4 clones indicated in the Figure 22 are functional, since they showed a positive signal in biotin-binding ELISA.

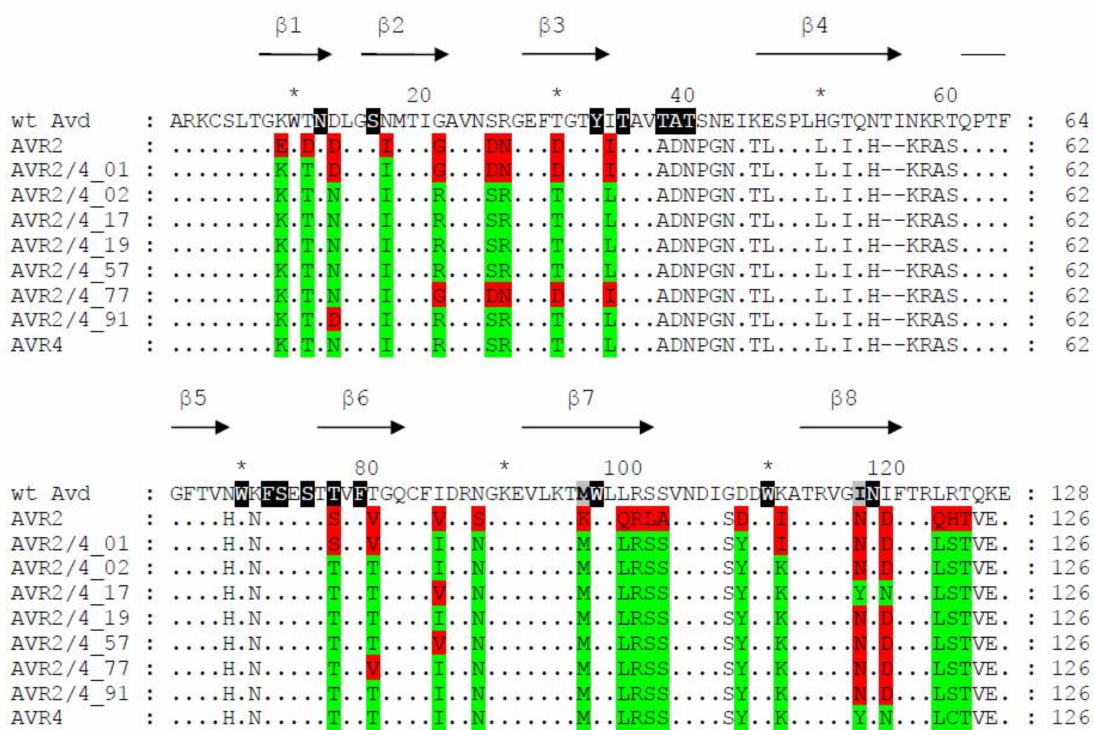


Figure 22: Multiple sequence alignment of avidin (wt AVD), AVR2, AVR4 and shuffled AVR2/4s. Dots indicate identical amino acids in AVRs compared to avidin. In the wt avidin sequence, beta-sheets are indicated by horizontal arrows, biotin-binding residues are shaded black and interface amino acids in grey. Variable sites in AVR sequences are coloured to indicate their parental origin (AVR2 is indicated by red and AVR4 by green). All of these AVR2/4 clones are functional since they bind biotin.

5.5.2 Phylogenetic analysis

The evolutionary relationships of sequenced clones from AVR2/4 library were inferred using the UPGMA (unweighted pair group method with arithmatic mean) method (Sneath & Sokal, 1973). The optimal tree with the sum of branch length = 1.11028525 is shown in the Appendix 6. The tree is drawn to scale, with branch lengths in the same

units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method (Zuckerkandl & Pauling, 1965) and are in the units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated from the dataset. OmpA signal peptide, short linker and pIII were excluded from the analyzed sequences and thus there were a total of 126 amino acid positions in the final dataset, corresponding to the AVR sequence.

When phylogenetic analysis is compared to sequence data and functional data obtained from protein ELISA, the evolutionary tree of AVR2/4 library seems to be quite accurate (see Appendix 6). The sequences of biotin-binding clones resemble more AVR4 than AVR2, which is reasonable because AVR4 is the closest relative of avidin and highly functional compared to more distant AVR2.

6. Discussion

6.1 Avidin engineering and phage display

Protein engineering by phage display technology has become common method for the production of universal binding frameworks for use in research, diagnostics and therapeutics. So called alternative scaffolds represent an intriguing alternative for repertoire of monoclonal antibodies. Several diverse protein folds are currently under development or have already been developed into versatile binding frameworks (Skerra, 2007). Calycin-based anticalins have already been proven to be a useful binding framework (Nuttall & Walsh, 2008; Skerra, 2008). Avidin is another member of the calycin family, which can be engineered using phage display in order to alter its ligand-binding and other properties. Modification of the loop structures of avidin scaffold has resulted in the development of new avidin-based antidin binders, which bind steroid-molecules rather than biotin (Riihimäki *et al.*, manuscript; Hiltunen *et al.*, unpublished results). The engineering of ligand-binding loops is not the only option in the search of new protein features. Modification of distant amino acid residues may sometimes have a major impact on the expression, stability, folding or other properties of the protein (Minshull & Stemmer, 1999).

The benefits of genetic recombination over random mutagenesis in the rapid evolution of organisms, genes and proteins are well-established (Sen *et al.*, 2007), and should be taken into account in the avidin engineering. The avidin structure is well-known, which enables rational design approaches (site-directed random mutagenesis and *in silico* modelling), but even more advantage will be gained by combining rational design and directed evolution approaches (Tobin *et al.*, 2000). This kind of combination would promote not only the discovery of new interesting features but also improve the properties of already existing binders. The combination of DNA shuffling and Gateway cloning was developed in order to facilitate the construction of phage display libraries. These methods bring novelty, efficiency and ease in the library construction, and may be exploited in the engineering of novel and existing avidin-based binders.

6.2 DNA shuffling

The desired protein features are usually improved to the detriment of other important properties, such as stability or expression levels (Minshull & Stemmer, 1999). By recombining sequences found in nature, it should be possible to discover proteins having all combinations of properties of the parents and, in addition, improvements over any of the parents. As a homology-based recombination method, DNA shuffling is suitable for engineering of avidin family members, which show high similarity in their sequences and 3D structures, but simultaneously diverge in their biochemical features (Ahlroth *et al.*, 2001a; Flower, 2000; Laitinen *et al.*, 2002).

The sequence comparisons guide the selection of starting material for DNA shuffling, as at least 60 % of nucleotide homology is required for the successful recombination (Crameri *et al.*, 1998). The DNA family shuffling of the members of chicken avidin family is an intriguing directed evolution approach. The sequence similarity of avidin family members is extremely high (91–100 % for *AVD* and *AVR1-7*) and they are therefore suitable for *in vitro* recombination. The use of *AVR4* in the DNA shuffling o avidin family is reasonable because it is one of the most functional and closest natural relative of avidin (Hytönen *et al.*, 2004a). *AVR2* may not be the most functional avidin family member in terms of biotin binding (Hytönen *et al.*, 2005a; Laitinen *et al.*, 2002), but it was used because its divergence brings more sequence variation into the shuffled gene pool. By shuffling engineered antidin binders with the members of chicken avidin family, even more interesting assemblies would be obtained. The selection of parental genes obviously has a major influence on the recombination events in DNA shuffling, and decision of the use of *AVR2* and *AVR4* as parental genes (sharing 91 % homology in their nucleotide sequences) turned out to be successful.

The shuffling of *AVR2* and *AVR4* cDNAs resulted in a library with high diversity and quality, since all of the sequenced genes were intact and shuffled on the DNA level. The library was analyzed by sequencing of 96 clones, 71 of which were successfully sequenced. The diversity of chimeric AVR2/4 library is guite high, as 54 of 71 analyzed sequences (76 %) were unique on the protein level. The amount of identical clones was more intricate to determine, as 17 clones represented one of the five different amino acid sequences found multiple times (see Appendix 5 and 6). In all cases, the DNA

sequences of the identical proteins were divergent, however, resulting in 100 % of unique sequences on the DNA level.

The effect of recombination events was determined on protein level by counting the number of instances where neighbouring sites are occupied by different parents. In this way, an average of 2.26(±0.07) crossovers per gene with a range of 0 to 5 was measured, which is a reasonable amount for two-parent library shuffled from highly homologous genes of this size (see Joern *et al.*, 2002). However, the observations on protein level (Figure 20) disregard the true amount of crossovers on the DNA level, which is assumed to be higher (because 5-7 crossovers per gene was counted manually from the DNA sequences of randomly picked clones). Due to the massive amount of sequence data, it was not feasible to determine the amount of crossovers for each clone on the DNA level by manual counting. The analysis of shuffled gene libraries is usually carried on the DNA level, because this is where the actual recombination occurs (Joern *et al.*, 2002). Because the construction of library by DNA shuffling relies on homologous recombination, crossovers are expected to occur preferentially where the parents share a high level of sequence identity. The actual crossover events can not be directly determined from the amino acid sequence, and more sophisticated approach would be needed to determine the amount of crossovers reliably. Joern *et al.* (2002) have used a labeled oligonucleotide probe hybridization method to characterize the shuffled gene libraries. Other probing techniques have also been applied elsewhere for the DNA shuffling library characterization (Abecassis *et al.*, 2000). Oligonucleotide probing might as well facilitate the characterization of shuffled avidin libraries on the DNA level.

The mutation frequency in the chimeric AVR2/4 library was 0.022 % on the protein level, since only two spontaneous amino acid changes (I83M and K109N) were observed in two of the 71 analyzed sequences. Because the amount of spontaneous mutations is quite low (Joern *et al.*, 2002; Stemmer, 1994b), they do not have any significant effect on the quality or diversity of the library. However, there are some undesired mutations in the chimeric AVR2/4 library, because of the inadequate validation of both parental *AVR* sequences and primers used in the library construction. *AVR2* template was found to contain a premature stop codon (L99 mutated) and E89K mutation. Another mutation, C122S, is derived from the primer used in the previous

studie. From the sequencing results it can be concluded that *AVR4* template was high in quality, showing no undesired mutations or other artefacts, but the *AVR2* template should have been evaluated more carefully. Yet another problem concerning the whole library is the undesired ochre stop codon (TAA), which was due to the mistake made in the primer design. It must be changed to the amber stop codon (TAG), before the constructed phage display library can be actually selected using phage display. The undesired mutations in the templates can be corrected by PCR or by using parental cDNA templates from another source. All obtained sequences were analyzed despite of these artefacts, in order to get a comprehensive outlook of the constructed library.

Reduction of parental gene contamination is a major goal in the optimization of DNA shuffling (Abecassis *et al.*, 2000), and in this study, it was reached by using unsymmetrical parental templates. The templates used contained only one *att*L-site and the recombination is therefore forced to happen, because only those reassembled products containing *att*L sites on their both ends were able to be LR cloned into phagemid vector. Thus the parental contamination can be completely avoided on the DNA level. Parental DNA sequences were not observed, but two parental amino acid sequences were found. This is due to the so-called back-crossing of parental genes, a situation where crossovers occur silently at homologous sites and the fenotype remains parental. In principle, the recombination scheme between two parental templates used should yield only uneven number of crossovers, thus also suggesting that the number of crossovers estimated based on amino acid sequences (shown in Figure 20) is erroneous.

The recombination patterns and parental incorporation were observed to affect the functionality of chimeric proteins. This was observed by comparing the biotin-binding affinities measured by protein ELISA to the sequencing results and the phylogenetic tree constructed from the sequences. As was expected, the functional biotin-binding clones were encoded by the shuffled ORFs. The amount of functional clones among shuffled ORFs was quite high, as 26 of them were able to bind biotin. Premature stop codon containing shuffled clones were not functional, which was expected because they are lacking almost a quarter of the amino acid sequence in each monomer. In general, the functionality seems to depend on the incorporation of the parental AVR4 DNA into the shuffled clones. This was observed by phylogenetic analysis (Appendix 6), as all functional clones seem to be more AVR4-like. The biotin-binding clones are placed to

the upper branches of the phylogenetic tree. UPGMA clustering is the simplest method of tree construction, assuming a constant rate of evolution (molecular clock hypothesis). It should only be used with closely related sequences, or when there is constancy of evolutionary rate. UPGMA clustering may not be the most accurate approach for phylogenetic analyses of evolution of genes found in nature, and it is currently most often used to produce guide trees for more sophisticated phylogenetic reconstruction algorithms (Morrison, 1996). However, it is reliable analysis method for the sequences of high identity, such as shuffled libraries. In the case of AVR2/4 library, the order and length of branches in UPGMA tree was corresponding very well to the manual comparison of sequences.

The affinity chromatography and analysis of purified protein samples revealed the nature of chimeric AVR2/4 clone A1 to be closely similar to wt AVR proteins. It was able to bind 2-iminobiotin rather tightly, as pH 4 buffer was not sufficient enough to elute the protein from the resin, but 0.1 M acetic acid was required for that. A total amount of purified protein was 2.6 mg, which is fairly good yield from the medium-scale (500 ml) *E. coli* culture. The yield could be increased by subcloning the desired AVR2/4 clone under stronger, arabinose-inducible T7 promoter. In the current work it was unnecessary, since the production and purification was only tested in small scale for one unselected clone and the amount of protein needed for biochemical characterization is not that high. T7-based Gateway-compatible destination vector would be useful in future studies, because it would enhance the subcloning step after phage display selection as well as increase the yield in protein production.

In SDS-PAGE analysis of AVR2/4 protein (see Figure 20), a characteristic pattern of different oligomeric forms can be seen due to the heat-stability of the oligomers characteristic to AVR proteins (Laitinen *et al.*, 2002). In addition to monomeric form of AVR (~15 kDa), some larger bands can be seen. They might indicate dimeric or signal peptide containing form of AVR protein. Western blotting and immunodetection analysis of the purified protein sample is also in line with previous studies where immunological properties of wt AVRs have been studied (Laitinen *et al.*, 2002). Laitinen *et al.* (2002) have shown that polyclonal rabbit anti-avidin antibody recognizes AVR4/5 more weakly than avidin and AVR2 is not recognized at all. The characterized AVR2/4 mutant A1 was able to be detected with monoclonal anti-AVR4/5 antibody

whereas the polyclonal anti-avidin antibody could not bind to the shuffled AVR2/4 mutant. The monoclonal AVR4/5 antibody recognized tetrameric forms of shuffled AVR2/4 mutant much better than monomeric forms. The binding site of the monoclonal AVR4/5 antibody is likely to be in between of two monomers and thereby only dimers and tetramers are efficiently recognized. Thereby the anti-AVR4/5 antibody could be more useful in the ELISA screening of chimeric AVR2/4 libraries, but the recognition of AVR2 and diverse AVR2/4 clones should be tested first. Polyclonal antibody might be more suitable for that purpose, because the incorporation of parental genes affects the immunological properties of shuffled clones, and AVR4/5-specific antibody would obviously favour AVR4-like clones. The particular AVR2/4 clone A1 resembled more AVR4 than AVR2, and could thus be recognized by anti-AVR4/5 antibody. The availability of antibodies required for ELISA and other immunological analyses of chimeric avidin libraries encoded by shuffled genes still remains to be addressed.

DNA shuffling itself is a skill-intensive method, and the integrity of the reassembled fragments is dependent on many important factors (e.g. the size range, concentration and purity of DNA fragments used for reassembly), and even small changes can affect the final result (An *et al.*, 2010). A sufficient amount of full-length reassembled fragments can sometimes be only obtained after careful optimization of the PCR conditions. In this study, the purity of template DNA and the duration of DNaseI fragmentation were observed to have a major impact on the outcome of the whole shuffling procedure. The yield of shuffled products could have been increased by using more template material per each reaction. The duration of DNaseI fragmentation directly affects on the fragment length and thereby it also has an effect on the amount of crossovers. Short fragmentation time decreases the amount of parental blocks in each hybrid gene, whereas too long fragmentation may compromise the annealing of the fragments in reassembly PCR. The amount of cycles and the DNA polymerase used in the reassembly PCR also affects the outcome of DNA shuffling procedure. When compared to regular Pfu polymerase, Phusion polymerase (Finnzymes) shows increased performance in primerless reassembly PCR due to its high fidelity and processivity (see Appendix 4). The optimization of the DNA shuffling protocol was not the main goal in this study. Instead, the easy performance of DNA shuffling as well as Gateway cloning of reassembled products in straightforward and reproducible manner was aimed.

6.3 Gateway cloning

In the DNA shuffling protocol presented in this study, PCR optimization and amplification of the reassembled products was unnecessary due to the direct Gateway cloning of the desired products into pGWphagemids. The use of unsymmetrical parental cDNAs, which contained only one *att*L-site in either end, ensured the successful recombination of fragmented genes into full-sized chimeric variants. This was due to the fact that both *att*L sites were required to flank the entry clone to transfer it to the destination vector by LR clonase. In the Gateway cloning of reassembled AVR2/4 cDNA, the complete transfer of shuffled chimeras into pGWphagemid was ensured by introducing an excess amount of reassembled PCR products (400-600 ng) as entry clones into the LR cloning reaction.

Cell-dependent display methods enable the creation of libraries with up to 10^{13} clones and allows the possibility to introduce *in vitro* mutagenesis during the amplification which permits directed evolution in every selection round (Grönwall & Ståhl, 2009). The use of cell-dependent systems is limited by the step of DNA transformation and *in vitro* transcription and translation, however. The transformation step was a major challenge in this study as well, because Gateway cloning only facilitates the construction of the library on the DNA level, but not its introduction into bacterial host. Because the library construction was based on the parallel electroporations of one LR cloning reaction, the amount of DNA (150 ng of pDNA in total, 30 ng of pDNA per electroporation) was not sufficient enough to obtain large library of variants. The medium-sized library of 10^6 clones was obtained in this way. The amount of DNA could be up to 1 µg per each electroporation and thereby the size of the library may be significantly increased by pooling several LR cloning reactions together. Thus the limiting factor is the amount of electroporations, which have to be performed in parallel.

Gateway cloning is very applicable method for the construction of large phagemid libraries and whenever high accuracy, efficiency and speed are required to transform a large amount of inserts into plasmid vectors (Hartley *et al.*, 2000; Katzen, 2007). In addition, it enables the transfer of inserts between multiple vectors. Specific recombination sequences (*att* sites) are required for the transfer, as are the site-specific recombinases for catalysis of the recombination (Hartley *et al.*, 2000). Site-specific

recombination reactions mediated by the lambda integrase family of recombinases are conservative (no net gain or loss of nucleotides) and highly specific (Landy, 1989). Gateway system is universal, enabling the cloning of all types of DNA fragments, whether that is PCR fragments, cDNA, or genomic DNA. The enzyme mixes and a wide variety of compatible plasmid vectors are commercially available (including TOPO- and TA-compatible Gateway vectors) for all kinds of expression systems (bacterial, yeast, insect and mammalian). In some cases it may be reasonable to construct and multiply the vector by one's own, especially when some particular features are needed.

In this study, a conventional phagemid vector was used as a framework to construct a tailor-made Gateway-compatible vector, pGWphagemid, for phage display applications. To construct a phage display libraries, inserts has to be transferred into phagemid vectors in the correct orientation and reading frame (Barbass *et al.*, 2004). Gateway cloning reactions are high-fidelity and fast performing, as the cloning reaction can be accomplished by simple mixing of reagents and one-hour incubation. The incubation time may vary, and in the construction of shuffled gene libraries, over-night incubation (RT) was preferred, since it was an easy and efficient way to operate. Longer incubation also ensures the complete incorporation of unpurified entry clones into destination vector. To preserve the condition of vectors, pGWphagemids should be aliquoted and repeated freezing-thawing cycles avoided. Any kind of frameshifting, inversions, deletions or duplications were not observed in the sequence analysis of the chimeric AVR2/4 library. Thereby Gateway cloning technique appears to be convinient and simple method for the cloning of large phagemid libraries.

Gateway-compatible cloning vectors have been established for HTP-analysis of protein interactions in other studies (Zhu *et al.*, 2010), as well as various other research applications in different fields of molecular biology. A vector compatible for Gateway cloning and phage display has been described in earlier study as well (Gao *et al.*, 2008). Gateway cloning of chimeric genes have also been reported (Suzuki *et al.*, 2005), but the described N- and C-terminal strategy is applicable to site-directed mutagenesis, and does not involve DNA family shuffling or phage display.

The multiplication of pGWphagemids was carried on in the specialized *ccd*B-competent strain DB3.1. The growth of DB3.1 cells required optimizationa and was found to be best done in the reduced temperatures (30 ºC or less) with generous glucose supplement (1.0 %) and extended incubation times. DB3.1 cells are somewhat less plasmid-rich than normal *E. coli,* and larger amounts of cell mass was needed to prepare adequate quantities of plasmid DNA. The requirements for the cell strain specific features were a challenge in this study, due to the use of F-episome containing *E. coli* cells. F-plasmid is an essential feature in the phage display, required for the host cell infection and replication of the phage genome. On the other hand, F-episome containing strains are not recommended to be used in the multiplication of the Gateway cloned entry vectors, since negative selection with *ccd*B does not work in these cells as efficiently as in F-negative strains. This was solved by using another cell line for the transformation of the LR cloning reactions. The reactions were first transformed into DH5α in order to eliminate the *ccd*B background and only then transformed into XL1-Blue. The background levels were determined by comparing transformation efficiencies. The amount of *ccd*B background (10^3) observed in XL1-Blue is only 1 % of the size of the final library (10^6), and it might not have to be taken into concideration at all. The undesired clones would be anyway eliminated in the rounds of phage display.

The pGWphagemid-vector was successfully constructed and verified by sequencing and AGE analyses of control digestions. The most important restriction sites were confirmed by control digestions of entry clones after the Gateway cloning of wt or shuffled genes as well. These sites (EcoRI, NotI and HindIII) can be used for the verification of the LR cloned inserts in the future studies. The constructed vector does not contain a signal sequence for protein secretion, and it should therefore be included in the entry clone. PelB or ompA signal sequences can be used, for example. The entry clone may also contain an appriate stop codon, which could be amber codon (TAG) in the case of phage display libraries. If the pGWphagemid is used for protein expression only, one or two ochre stop codons may be included. Recombination sites a*tt*L1 and *att*L2 are of course required to flank the sequence to transfer it into pGWphagemid. Because Gateway cloning reactions are reversible (Hartley *et al*., 2000), the recombination of *att*L and *att*R sites produces an expression clone containing *att*B-flanked insert, which can be introduced in subsequent Gateway cloning reactions using

BP clonase. Thereby another Gateway-compatible vector could be utilized as an expression vector for more efficient protein production, for example.

6.4 Future objectives

The combination of DNA shuffling and Gateway cloning system facilitates the construction and cloning of phagemid libraries used in the phage display engineering of avidin-based binders. Additional work is required to demonstrate the actual phage display work for this kind of chimeric library. Regeneration of chimeric AVR2/4 phage display library or incorporation of other parental avidin family members into this scheme may yield interesting variants with improved or completely new properties. As a directed evolution approach, DNA shuffling might bring novel solutions in the antidin engineering, where the selection of unique properties may not contribute the general properties of the mutants. DNA shuffling of avidin family members is going to be used to improve the protein yield and stability of the steroid-binding avidins developed in our group.

7. Conclusions

The chimeric AVR2/4 library was successfully produced by combining DNA shuffling and Gateway cloning methods. The Gateway-compatible phagemid vector, pGWphagemid, was constructed and used as a destination vector, and thus proven to be an applicable tool in the library construction. DNA shuffling of AVR2 and AVR4 was productive, yielding an interesting library of variants. The functionality and diversity of the chimeric AVR2/4 library is quite high despite of the mutations derived from parental cDNAs. The major challenge in the phage display library construction is to achieve sufficient library size and quality. The aim of this study was to develop new tools and methods for the construction of shuffled phage display libraries. In general, the tools and methods used in this study are applicable in the construction of avidin-based mutant libraries for phage display selection and may contribute in the engineering of both new and existing avidin-based binders. In future, the engineered avidin-based binders could possibly be utilized in various applications of biotechnology, ranging from nanoscience to diagnostics.

References

Abecassis V, Pompon D & Truan G. 2000. High efficiency family shuffling based on multi-step PCR and in vivo DNA recombination in yeast: statistical and functional analysis of a combinatorial library between human cytochrome P450 1A1 and 1A2. *Nucleic Acids Res* 28(20): E88.

Ahlroth MK, Kola EH, Kulomaa MS, Ewald D, Masabanda J, Sazanov A & Fries R. 2000. Characterization and chromosomal localization of the chicken avidin gene family. *Anim Genet* 31(6): 367-375.

Ahlroth MK, Grapputo A, Laitinen OH & Kulomaa MS. 2001a. Sequence features and evolutionary mechanisms in the chicken avidin gene family. *Biochem Biophys Res Commun* 285(3): 734-741.

Ahlroth MK, Ahlroth P & Kulomaa MS. 2001b. Copy-number fluctuation by unequal crossing-over in the chicken avidin gene family. *Biochem Biophys Res Commun* 288(2): 400-406.

Airenne KJ, Marjomaki VS & Kulomaa MS. 1999. Recombinant avidin and avidin-fusion proteins. *Biomol Eng* 16(1-4): 87-92.

An Y, Wu W & Lv A. 2010. A convenient and robust method for construction of combinatorial and random mutant libraries. *Biochimie* 92(8): 1081-1084.

Arnold FH. 2001. Combinatorial and computational challenges for biocatalyst design. *Nature* 409(6817): 253-257.

Barbass CF, Burton DR, Scott JK. 2004. Phage Display: A Laboratory Manual. *Cold Spring Harbor Laboratory Press.* Read on http://books.google.fi/books 5.12.2010.

Bernard P & Couturier M. 1992. Cell killing by the F plasmid CcdB protein involves poisoning of DNA-topoisomerase II complexes. *J Mol Biol* 226(3): 735-745.

Bornscheuer UT & Pohl M. 2001. Improved biocatalysts by directed evolution and rational protein design. *Curr Opin Chem Biol* 5(2): 137-143.

Bradbury ARM & Marks JD. 2004. Antibodies from phage antibody libraries. *J Immunol Methods* 290(1-2): 29-49.

Campbell A. 1962. Episomes. *Adv Genet* 11:101-145.

Chang CC, Chen TT, Cox BW, Dawes GN, Stemmer WP, Punnonen J & Patten PA. 1999. Evolution of a cytokine using DNA family shuffling. *Nat Biotechnol* 17(8): 793-797.

Crameri A, Raillard SA, Bermudez E & Stemmer WPC. 1998. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* 391(6664): 288-291.

Eisenberg-Domovich Y, Hytonen VP, Wilchek M, Bayer EA, Kulomaa MS & Livnah O. 2005. High-resolution crystal structure of an avidin-related protein: insight into high-affinity biotin binding and protein stability. *Acta Crystallogr Sect D-Biol Crystallogr* 61: 528-538.

Flower DR. 1993. Structural relationship of streptavidin to the calycin protein superfamily. *FEBS Lett* 333(1-2): 99-102.

Flower DR, North ACT & Sansom CE. 2000. The lipocalin protein family: structural and sequence overview. *Biochim Biophys Acta* 1482(1-2): 9-24.

Galanis M, Irwing RA, Hudson PJ. 2001. Bacteriophage Library Construction and Selection of Recombinant Antibodies. *Curr Prot Immunol* 17(1):1-45.

Gao H, Pattison D, Yan T, Klingeman DM, Wang X, Petrosino J, Hemphill L, Wan X, Leaphart AB, Weinstock GM, Palzkill T & Zhou J. 2008. Generation and validation of a Shewanella oneidensis MR-1 clone set for protein expression and phage display. *PLoS One* 3(8): E2983.

Gottesman ME & Weisberg RA. 2004. Little lambda, who made thee? *Microbiol Mol Biol Rev* 68(4): 796-813.

Green NM. 1975. Avidin. *Adv Protein Chem* 29: 85-133.

Green NM. 1990. Avidin and Streptavidin. *Meth Enzymol* 184: 51-67.

Gunneriusson E, Nord K, Uhlen M, Nygren P. 1999. Affinity maturation of a Taq DNA polymerase specific affibody by helix shuffling. *Protein Eng* 12(10): 873-878.

Harayama S. 1998. Artificial evolution by DNA shuffling. *Trends Biotechnol* 16(2): 76-82.

Hartley JL, Temple GF & Brasch MA. 2000. DNA cloning using in vitro site-specific recombination. *Genome Res* 10(11): 1788-1795.

Hood EE, Witcher DR, Maddock S, Meyer T, Baszczynski C, Bailey M, Flynn P, Register J, Marshall L, Bond D, Kulisek E, Kusnadi A, Evangelista R, Nikolov Z, Wooge C, Mehigh RJ, Hernan R, Kappel WK, Ritland D, Li CP & Howard JA. 1997. Commercial production of avidin from transgenic maize: characterization of transformant, production, processing, extraction and purification. *Mol Breed* 3(4): 291-306.

Hytönen VP, Nyholm TKM, Pentikäinen OT, Vaarno J, Porkka EJ, Nordlund HR, Johnson MS, Slotte JP, Laitinen OH & Kulomaa MS. 2004a. Chicken avidin-related protein 4/5 shows superior thermal stability when compared with avidin while retaining high affinity to biotin. *J Biol Chem* 279(10): 9337-9343.

Hytönen VP, Laitinen OH, Airenne TT, Kidron H, Meltola NJ, Porkka EJ, Hörhä J, Paldanius T, Määttä JA, Nordlund HR, Johnson MS, Salminen TA, Airenne KJ, Ylä-Herttuala S & Kulomaa MS. 2004b. Efficient production of active chicken avidin using a bacterial signal peptide in Escherichia coli. *Biochem J* 384(2): 385-390.

Hytönen VP, Määttä JAE, Nyholm TKM, Livnah O, Eisenberg-Domovich Y, Hyre D, Nordlund HR, Hörha J, Niskanen EA, Paldanius T, Kulomaa T, Porkka EJ, Stayton PS, Laitinen OH & Kulomaa MS. 2005a. Design and construction of highly stable, protease-resistant chimeric avidins. *J Biol Chem* 280(11): 10228-10233.

Hytönen VP, Määttä JAE, Kidron H, Halling KK, Hörha J, Kulomaa T, Nyholm TKM, Johnson MS, Salminen TA, Kulomaa MS & Airenne TT. 2005b. Avidin related protein 2 shows unique structural and functional features among the avidin protein family. *BMC Biotechnol* 5: 28.

Hytönen VP, Hörha J, Airenne TT, Niskanen EA, Helttunen KJ, Johnson MS, Salminen TA, Kulomaa MS & Nordlund HR. 2006. Controlling quaternary structure assembly: Subunit interface engineering and crystal structure of dual chain avidin. *J Mol Biol* 359(5): 1352-1363.

Joern JM, Meinhold P & Arnold FH. 2002. Analysis of shuffled gene libraries. *J Mol Biol* 316(3): 643-656.

Katzen F. 2007. Gateway® recombinational cloning: a biological operating system. *Expert Opin Drug Discov* 2(4): 571-589.

Keinänen RA, Wallen MJ, Kristo PA, Laukkanen MO, Toimela TA, Helenius MA & Kulomaa MS. 1994. Molecular-Cloning and Nucleotide-Sequence of Chicken Avidin-Related Genes 1-5. *Eur J Biochem* 220(2): 615-621.

Kulomaa MS, Nordlund HR, Hytönen VP, Nyholm TKM, Uotila STH, Porkka EJ & Laitinen OH. 2004. Engineering of thermal stability and biotin binding of avidin. *Biophys J* 86(1): 95A.

Kunnas TA, Wallen MJ & Kulomaa MS. (1993) Induction of Chicken Avidin and Related Messenger-RNAs After Bacterial-Infection. *Biochim Biophys Acta* 1216(3): 441-445.

Laitinen OH, Hytönen VP, Nordlund HR & Kulomaa MS. 2006. Genetically engineered avidins and streptavidins. *Cell Mol Life Sci* 63(24): 2992-3017.

Laitinen OH, Hytönen VP, Ahlroth MK, Pentikainen OT, Gallagher C, Nordlund HR, Ovod V, Marttila AT, Porkka E, Heino S, Johnson MS, Airenne KJ & Kulomaa MS. 2002. Chicken avidin-related proteins show altered biotin-binding and physico-chemical properties as compared with avidin. *Biochem J* 363: 609-617.

Laitinen OH, Nordlund HR, Hytönen VP, Uotila STH, Marttila AT, Savolainen J, Airenne KJ, Livnah O, Bayer EA, Wilchek M & Kulomaa MS. 2003. Rational design of an active avidin monomer. *J Biol Chem* 278(6): 4010-4014.

Landy A. 1989. Dynamic, structural, and regulatory aspects of lambda site-specific recombination. *Annu Rev Biochem* 58: 913-949.

Leemhuis H, Kelly RM & Dijkhuizen L. 2009. Directed Evolution of Enzymes: Library Screening Strategies. *IUBMB Life* 61(3): 222-228.

Livnah O, Bayer EA, Wilchek M & Sussman JL. 1993. Three-dimensional structures of avidin and the avidin-biotin complex. *Proc Natl Acad Sci USA* 90(11): 5076-5080.

Määttä JAE, Airenne TT, Nordlund HR, Jänis J, Paldanius TA, Vainiotalo P, Johnson MS, Kulomaa MS & Hytönen VP. 2008. Rational modification of ligand-binding preference of avidin by circular permutation and mutagenesis. *ChemBioChem* 9(7): 1124-1135.

Määttä JA, Eisenberg-Domovich Y, Nordlund HR, Hayouka R, Kulomaa MS, Livnah O & Hytönen VP. 2010. Chimeric avidin shows stability against harsh chemical conditions - biochemical analysis and 3D structure. *Biotechnol Bioeng* Epub ahead of print 11.10.2010.

Marttila AT, Laitinen OH, Airenne KJ, Kulik T, Bayer EA, Wilchek M & Kulomaa MS. 2000. Recombinant NeutraLite Avidin: a non-glycosylated, acidic mutant of chicken avidin that exhibits high affinity for biotin and low non-specific binding properties. *FEBS Lett* 467(1): 31-36.

Minshull J & Stemmer WPC. 1999. Protein evolution by molecular breeding. *Curr Opin Chem Biol* 3(3): 284-290.

Moore GL & Maranas CD. 2002. Predicting Out-of-Sequence Reassembly in DNA Shuffling. *J. theor. Biol.* 219: 9–17

Morrison DA. 1996. Phylogenetic tree-building. *Int J Parasitol* 26(6): 589-617.

Niederhauser B, Siivonen J, Hytönen VP, Kulomaa MS. Family shuffling inside chicken avidin protein family - development of more powerful biotech tools. Manuscript.

Niskanen EA, Hytönen VP, Grapputo A, Nordlund HR, Kulomaa MS & Laitinen OH. 2005. Chicken genome analysis reveals novel genes encoding biotin-binding proteins related to avidin family. *BMC Genomics* 6: 41.

Nordlund HR, Laitinen OH, Hytonen VP, Uotila STH, Porkka E & Kulomaa MS. 2004. Construction of a dual chain pseudotetrameric chicken avidin by combining two circularly permuted avidins. *J Biol Chem* 279(35): 36715-36719.

Nuttall SD & Walsh RB. 2008. Display scaffolds: protein engineering for novel therapeutics. *Curr Opin Pharmacol* 8(5): 609-615.

Pande J, Szewczyk MM & Grover AK. 2010. Phage display: Concept, innovations, applications and future. *Biotechnol Adv* 28(6): 849-858.

Patten PA, Howard RJ & Stemmer WP. 1997. Applications of DNA shuffling to pharmaceuticals and vaccines. *Curr Opin Biotechnol* 8(6): 724-733.

Riihimäki TA, Hiltunen S, Nordlund HR, Rangl M, Määttä JAE, Ebner A, Kulomaa MS, Takkinen K & Hytönen VP. Four amino acid change in β1 to β2 loop changes avidin into a testosterone-binding protein. Manuscript.

Schofield DJ, Pope AR, Clementel V, Buckell J, Chapple SD, Clarke KF, Conquer JS, Crofts AM, Crowther SR, Dyson MR, Flack G, Griffin GJ, Hooks Y, Howat WJ, Kolb-Kokocinski A, Kunze S, Martin CD, Maslen GL, Mitchell JN, O'Sullivan M, Perera RL, Roake W, Shadbolt SP, Vincent KJ, Warford A, Wilson WE, Xie J, Young JL & McCafferty J. 2007. Application of phage display to high throughput antibody generation and characterization. *Genome Biol* 8(11): R254.

Sen S, Dasu VV & Mandal B. 2007. Developments in directed evolution for improving enzyme functions. *Appl Biochem Biotechnol* 143(3): 212-223.

Sheedy C, MacKenzie CR & Hall JC. 2007. Isolation and affinity maturation of hapten-specific antibodies. *Biotechnol Adv* 25(4): 333-352.

Sidhu SS. 2001. Engineering M13 for phage display. *Biomol Eng* 18(2): 57-63.

Skerra A. 2000. Lipocalins as a scaffold. *Biochim Biophys Acta* 1482(1-2): 337-350.

Skerra A. 2007. Alternative non-antibody scaffolds for molecular recognition. *Curr Opin Biotechnol* 18(4): 295-304.

Skerra A. 2008. Alternative binding proteins: Anticalins - harnessing the structural plasticity of the lipocalin ligand pocket to engineer novel binding activities. Febs Journal 275(11): 2677-2683.

Smith GP. 1985. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* 228(4705): 1315-1317.

Smith GP & Petrenko VA. 1997. Phage display. *Chem Rev* 97(2): 391-410.

Sneath PHA & Sokal RR 1973. Numerical Taxonomy. *Freeman*, San Francisco.

Stemmer WPC. 1994a. DNA Shuffling by Random Fragmentation and Reassembly - In-Vitro Recombination for Molecular Evolution. *Proc Natl Acad Sci USA* 91(22): 10747-10751.

Stemmer WPC. 1994b. Rapid Evolution of a Protein In-Vitro by Dna Shuffling. *Nature* 370(6488): 389-391.

Stemmer WPC. 2002. Molecular breeding of genes, pathways and genomes by DNA shuffling. *J Mol Catal B-Enzym* 19: 3-12.

Suzuki Y, Kagawa N, Fujino T, Sumiya T, Andoh T, Ishikawa K, Kimura R, Kemmochi K, Ohta T & Tanaka S. 2005. A novel high-throughput (HTP) cloning

strategy for site-directed designed chimeragenesis and mutation using the Gateway cloning system. *Nucleic Acids Res* 33(12):109.

Swords WE. 2003. *E. coli* Plasmid Vectors: Preparation of Single-Stranded DNA from Phagemid Vectors. *Meth Mol Biol* 235: 103-106

Tamura K, Dudley J, Nei M & Kumar S. 2007 MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24:1596-1599.

Tobin MB, Gustafsson C & Huisman GW. 2000. Directed evolution: the 'rational' basis for 'irrational' design. *Curr Opin Struct Biol* 10(4): 421-427.

Uhlen M. 2008. Affinity as a tool in life science. *BioTechniques* 44(5): 649-654.

Viti F, Nilsson F, Demartis S, Huber A & Neri D. 2000. Design and use of phage display libraries for the selection of antibodies and enzymes. *Methods Enzymol* 326: 480-505.

Wallen MJ, Laukkanen MO & Kulomaa MS. 1995. Cloning and sequencing of the chicken egg-white avidin-encoding gene and its relationship with the avidin-related genes Avr1-Avr5. *Gene* 161(2): 205-209.

Weisberg RA & Landy A. 1983. Lambda II. In: Hendrix RW, Roberts JW, Stahl FW, Weisberg RA (eds), *Cold Spring Harbor Monograph Series* 13: 211-250.

Wilchek M & Bayer EA. (1990) Introduction to Avidin-Biotin Technology. *Meth Enzymol* 184: 5-13.

Zhu T, Wang W, Wong H, Yang X, Wang K & Cui Z. 2010. Construction of gateway-compatible yeast two-hybrid vectors for high throughput analysis of protein interaction. *Afr J Biotechnol* 9(9): 1392-1396.

Zocchi A, Jobe AM, Neuhaus JM & Ward TR. 2003. Expression and purification of a recombinant avidin with a lowered isoelectric point in *Pichia pastoris*. *Protein Expr Purif* 32(2): 167-174.

Zuckerkandl E & Pauling L. 1965. Evolutionary divergence and convergence in proteins, pp. 97-166 in *Evolving Genes and Proteins*, edited by V. Bryson and H.J. Vogel. Academic Press, New York.

Appendices

Appendix 1: Parameters for PCR

Construction of Gateway-cassette

Step	Temperature (°C)	Time	
1	98	2 min	
2	98	10 s	steps 2-4 repeated 35x
3	55	30 s	
4	72	1 min	
5	72	10 min	
6	4	∞	

DNA shuffling/Amplification

Step	Temperature (°C)	Time	
1	95	2 min	
2	95	30 s	steps 2-4 repeated 25x
3	52	60 s	
4	72	60 s	
5	72	10 min	
6	10	∞	

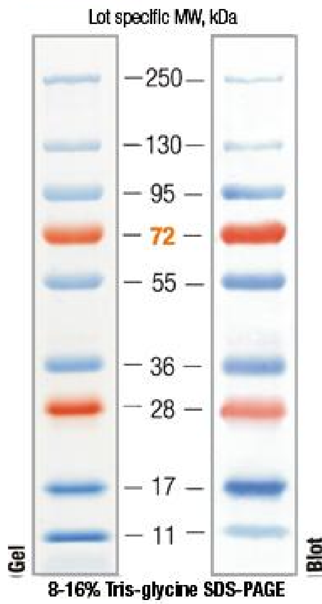
DNA shuffling/Reassembly

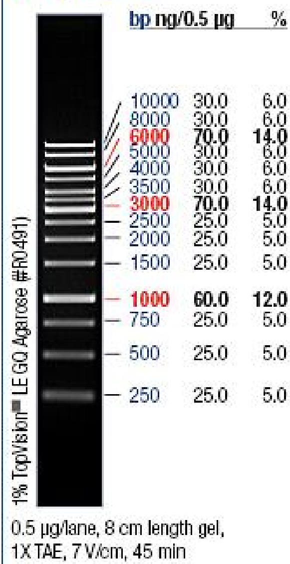
Step	Temperature (°C)	Time	
1	95	2 min	
2	95	30 s	steps 2-4 repeated 40x
3	52	60 s	
4	72	60 s	
5	72	10 min	
6	10	∞	

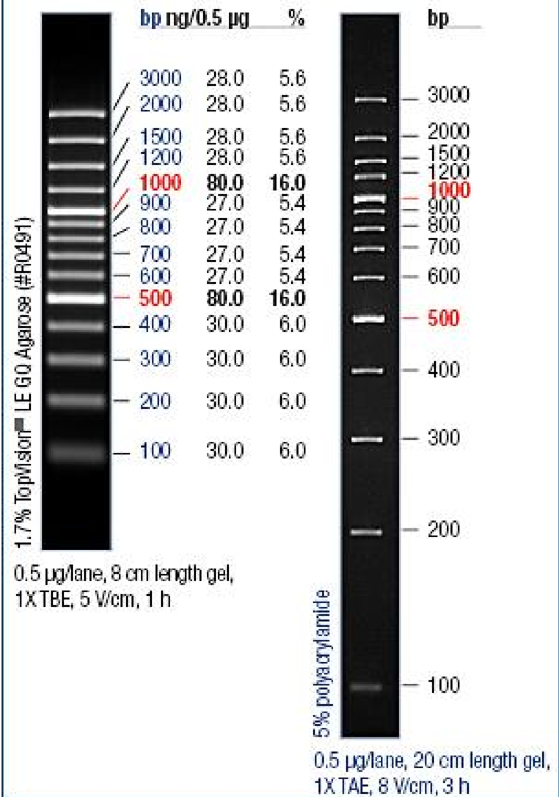
Sequencing PCR

Step	Temperature (°C)	Time	
1	96	1 min	
2	96	10 s	steps 2-4 repeated 29x
3	50	10 s	
4	60	4 min	
5	10		

Appendix 2: MW markers





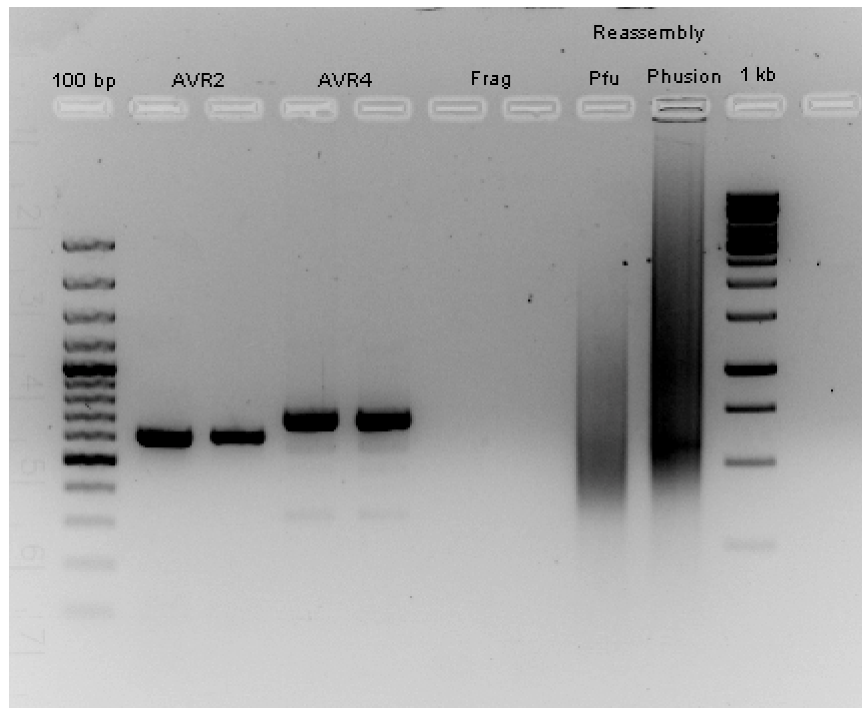


Appendix 3: Solutions and reagents

α-AVD polyclonal IgG	University of Oulu
α-AVR4/5 monoclonal IgG	FIT Biotech, Tampere, Finland
Acetic acid (0.1 M – 1.0 M)	glacial acetic acid (Sigma) in dH_2O
Avidin marker 1 mg/ml	Belovo S.A., Belgium
APA2-buffer (pH 9.5)	100 mM Tris-HCl 100 mM NaCl10 mM $MgCl_2$ pH adjusted using HCl
Blocking buffer	5 % (w/v) milk in PBS/TBS/Tris-buffer
Btn-BSA	biotinylated BSA (~1 Btn /1 BSA) 61 mg/ml in 50 mM $NaPO_4$, 100 mM NaCl, pH 7 (28.08.2009 Jenni Leppiniemi)
BSA	Sigma
Coomassie Brilliant Blue	0.05 % (w/v) Coomassie Brilliant Blue R-250 40 % (v/v) ethanol 10 % (v/v) glacial acetic acid 50 % H_2O
DEA (pH 9.8)	1 M diethanol amine 0.5 mM $MgCl_2$ pH adjusted using 10 M HCl
Destaining solution (for Coomassie)	40 % ethanol 10 % acetic acid 50 % water
Freezing buffer	50 ml 2x SB 10 ml 10x Solution 1 (4 mM $MgSO_4$) 10 ml 10x Solution 2 (360 mM K_2HPO_4, 132 mM KH_2PO_4, 68 mM $(NH_4)_2SO_4$, 17 mM natrium citrate, pH 7,5 adjusted using NaOH) 10 ml 10x Solution 3 (44 %glycerol) 20 ml sterile dH_20
HilloI-buffer (pH 8)	50 mM Tris 2 mM EDTA 150 mM NaCl 1 % TritonX-100

LB (pH 7)	10 g NaCl 10 g tryptone 5 g yeast extract 1 liter dH_20
lysozyme (Fluka)	5 mg/ml in dH_20
PBS-Tween	0.05 % (v/v) Tween in PBS
pH 4 buffer	50 mM sodium acetate pH adjusted using 1 M acetic acid
pH 11 buffer	50 mM Na_2CO_3, 1 M NaCl
PNPP (Sigma)	phosphatase substrate tablets (á 5 mg)
SB (pH 7)	30 g tryptone 20 g yeast extract 10 g MOPS (3-(N-morpholino)propane-sulphonic acid 1 liter dH_20
10x SDS-PAGE running buffer (pH 8.3)	30.3 g Tris base 144.0 g glycine 10.0 g SDS 1 liter dH_20
SET-buffer (pH 8)	20 % sucrose 0.5 M EDTA 1 M Tris
5x sequencing buffer (ABI)	200 mM Tris, 5 mM $MgCl_2$
S.O.C. medium	250 ml SOB 10 mM $MgCl_2$ 20 mM glucose
stacking and resolving gels (5 % or 15 % respectively)	3.4 ml dH_20 0.83 ml or 7.5 ml 30% acrylamide (BioRad) 0.63 ml 1.0 M Tris pH 6,8 (in stacking) or 3.8 ml 1.5 M Tris pH 8,8 buffer (in resolving) 0.15 ml or 0.05 ml 10% SDS 0.05 ml or 0.15 ml 10% APS 0.005 ml or 0.006 ml TEMED (Sigma)
50x TAE-buffer	242 g Tris 100 ml 0.5 EDTA (pH 8) 57.1 ml glacial acetic acid
Tris-buffer (pH 8)	50 mM Tris, 1 M NaCl

Appendix 4: Comparison of Pfu and Phusion DNA polymerases



AGE analysis and comparison of Pfu and Phusion polymerases in the reassembly step of DNA shuffling. PCR reactions were made in parallel using the same starting material, and Phusion seems to be much more powerful than Pfu in terms of processivity. The quality and contents of Phusion reassembly reactions were not analyzed in greater detail, but Phusion appears to be a promising alternative to Pfu if the amount of reassembled PCR products needs to be increased.

Appendix 5: Multiple Alignment of AVR2/4 sequences

AVR2 reference sequence:

```
ARKCSLTGEWDNDLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTSVFVGQCFVDRSGKEVLKTKWLQRLAVDDISDDWIATRVGNNDFTRQHTVEE
                                                                                          K         *
                                                                                                           premature stop
                                                                                                                    codon
```

AVR4 reference sequence:

```
ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGYNNFTRLCTVEE
                                                                                                                       LS

                                                                                           C112S-primer, stop codon
```

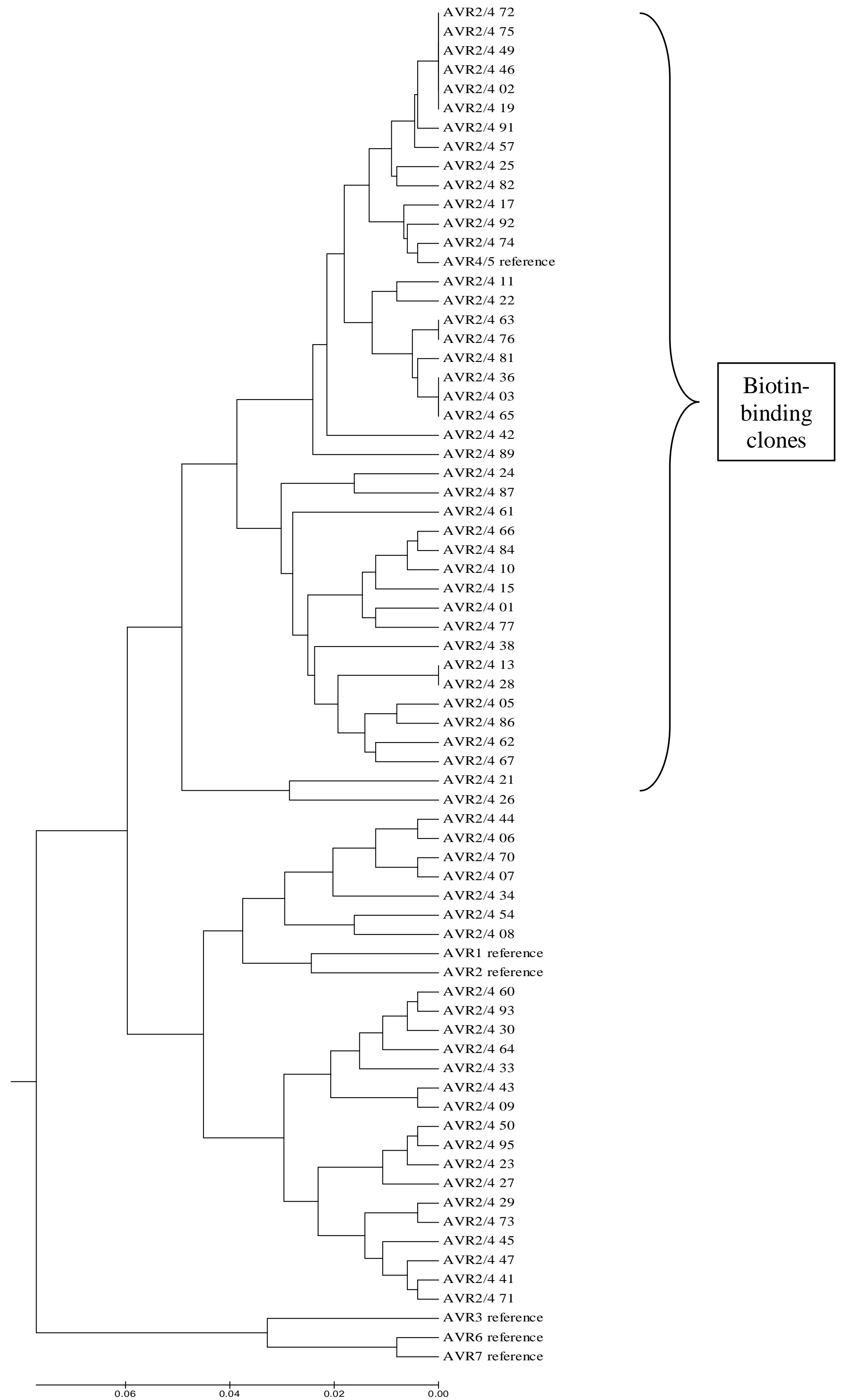
AVR2/4 sequences:

```
                *         20         *         40         *         60         *         80         *        100         *        120
10 : ARKCSLTGKWTNNLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWIATRVGNNDFTRLSTVEE*
11 : ARKCSLTGEWDNDLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTSVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGYNNFTRLSTVEE*
13 : ARKCSLTGEWDNDLGSIMTIGAVNDNGEFDGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGYNNFTRLSTVEE*
15 : ARKCSLTGKWTNNLGSIMTIGAVNDNGEFDGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWNATRVGNNDFTRLSTVEE*
17 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFVDRNGKEVLKTMWLLRSSVNDISYDWKATRVGYNNFTRLSTVEE*
18 : GQKVLADWEWDNDLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLCTVEE*
19 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
01 : ARKCSLTGKWTNDLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTSVFVGQCFIDRNGKEVLKTMWLLRSSVNDISYDWIATRVGNNDFTRLSTVEE*
21 : ARKCSLTGKWTNDLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFVDRNGKKVLKTKWLQR*AVDDISYDWKATRVGYNNFTRLSTVEE*
22 : ARKCSLTGKWDNDLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGYNNFTRLSTVEE*
23 : ARKCSLTGEWDNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTSVFVGQCFVDRSGKKVLKTKWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
24 : ARKCSLTGEWDNDLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTSVFVGQCFIDRSGKKVLKTMWLLRSSVNDISDDWIATRVGNNDFTRLSTVEE*
25 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWIATRVGNNDFTRLSTVEE*
26 : ARKCSLTGKWTNNLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFVDRSGKKVLKTKWLLRSSVNDISYDWKATRVGYNDFTRLSTVEE*
27 : ARKCSLTGEWDNDLGSIMTIGAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTSVFVGQCFVDRSGKKVLKTKWLQR*AVDDISDDWIATWVGNNDFTRLSTVEE*
28 : ARKCSLTGEWDNDLGSIMTIGAVNDNGEFDGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGYNNFTRLSTVEE*
29 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTSVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTSVFVGQCFVDRSGKKVLKTKWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
02 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
30 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFMDRNGKEVLKTKWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
31 : -----------------------------------------------------------------------------------TGTGKKVLKTMWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
33 : ARKCSLTGKWDNDLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTKWLQR*AVDDISDDWIATRVGYNDFTRLSTVEE*
34 : ARKCSLTGKWTNNLGSIMTIGAVNDNGEFDGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTSVFVGQCFVDRSGKKVLKTKWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
36 : ARKCSLTGEWDNDLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
38 : ARKCSLTGEWDNNLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRQHTVEE*
03 : ARKCSLTGEWDNDLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
41 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFVDRSGKKVLKTKWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
42 : ARKCSLTGEWDNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISDDWIATRVGYNNFTRLSTVEE*
43 : ARKCSLTGEWDNDLGSIMTIGAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
44 : ARKCSLTGEWDNDLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTSVFVGQCFVDRSGKKVLKTKWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
45 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFVDRSGKKVLKTKWLQR*AVDDISDDWIATRVGYNNFTRLSTVEE*
46 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
47 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVANNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFVDRSGKKVLKTKWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
49 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
50 : ARKCSLTGEWDNDLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTSVFVGQCFVDRSGKKVLKTKWLQR*AVDDISDDWKATRVGNNDFTRLSTVEE*
51 : GQKVLADWEMDQQPGSIMTIRAVNSRGEFTGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTSVFVGQCFVDRSGKKVLKTKWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
52 : GQKVLADWEMDQQPGLIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
54 : ARKCSLTGEWDNDLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLQR*AVDDISDDWIATRVGYNNFTRLSTVEE*
57 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFVDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
59 : --------------------AVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
05 : ARKCSLTGEWDNDLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISDDWIATRVGNNDFTRLSTVEE*
60 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTKWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
61 : ARKCSLTGKWTNNLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDR--REVLKTMWLLRSSVNDISYDWKATRVGYNNFTRLCTVEE*
62 : ARKCSLTGEWDNDLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVDDISDDWIATRVGYNNFTRLSTVEE*
63 : ARKCSLTGEWDNDLGSIMTIGAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
64 : ARKCSLTGKWTNDLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFVDRNGKEVLKTMWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
65 : ARKCSLTGEWDNDLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
66 : ARKCSLTGKWTNNLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISDDWIATRVDNNDFTRLSTVEE*
67 : ARKCSLTGKWDNDLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVDDISDDWIATRVGNNDFTRLSTVEE*
06 : ARKCSLTGEWDNDLGSIMTIRAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTSVFVGQCFVDRSGKKVLKTKWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
70 : ARKCSLTGEWDNDLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFVDRSGKKVLKTKWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
71 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRSGKKVLKTKWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
72 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
73 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTSVFVGQCFVDRSGKKVLKTKWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
74 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGYNNFTRLSTVEE*
75 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
76 : ARKCSLTGEWDNDLGSIMTIGAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
77 : ARKCSLTGKWTNNLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFVGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
78 : -QIIQETSRYHHCLGSNTKEPASPPLASLS----------IGTFQSPPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGYNNFTRLSTVEE*
07 : ARKCSLTGEWDNDLGSIMTIGAVNDNGEFDGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFVDRSGKKVLKTKWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
81 : ARKCSLTGEWDNDLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGYNDFTRLSTVEE*
82 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISDDWIATRVGNNDFTRLSTVEE*
84 : ARKCSLTGKWTNNLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISDDWIATRVGNNDFTRLSTVEE*
86 : ARKCSLTGEWDNDLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
87 : ARKCSLTGEWDNDLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRSGKKVLKTKWLLRSSVNDISYDWIATRVGNNDFTRLSTVEE*
89 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGYNNFTRQHTVEE*
08 : ARKCSLTGEWDNDLGSIMTIGAVNDNGEFDGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTSVFVGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
91 : ARKCSLTGKWTNDLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNDFTRLSTVEE*
92 : ARKCSLTGEWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGYNNFTRLSTVEE*
93 : ARKCSLTGKWTNNLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
94 : GQKVLADWEWDNDLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLLRSSVNDISYDWKATRVGNNNFTRLSTVEE*
95 : ARKCSLTGEWDNDLGSIMTIRAVNSRGEFTGTYLTAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTSVFVGQCFVDRSGKKVLKTKWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
09 : ARKCSLTGEWDNDLGSIMTIGAVNSRGEFTGTYITAVADNPGNITLSPLLGIQHKRASQPTFGFTVHWNFSESTTVFTGQCFIDRNGKEVLKTMWLQR*AVDDISDDWIATRVGNNDFTRLSTVEE*
```

Consensus sequence:

```
arkcsltg w n lgsimti avn  gef gty tavadnpgnitlspllgiqhkrasqptfgftvhwnfsest vf gqcf dr gk vlkt wl r  v dis dw atrvg N ftrlsTVEE*
```

Appendix 6: UPGMA clustering of shuffled *AVR2/4* genes



The evolutionary distances are shown in the units of the number of amino acid substitutions per site.