Combinatorial protein engineering applied to enzyme catalysis and molecular recognition

Malin Eklund

Royal Institute of Technology
Department of Biotechnology

Stockholm 2004
Abstract

The recent development of methods for constructing and handling large collections (libraries) of proteins, from which variants with desired traits can be isolated, has revolutionized the field of protein engineering. Key elements of such methods are the various ways in which the genotypes (the genes) and the phenotypes (the encoded proteins) are physically linked during the process. In one section of the work underlying this thesis, one such technique (phage display), was used to isolate and identify protein library members based on their catalytic or target molecule-binding properties.

In a first study, phage display libraries of the lipolytic enzyme Lipolase from Thermomyces lanuginosa were constructed, the objective being to identify variants with improved catalytic efficiency in the presence of detergents. To construct the libraries, nine positions were targeted for codon randomization, all of which are thought to be involved in the conformational change-dependent enzyme activation that occurs at water-lipid interfaces. The aim was to introduce two to three amino acid mutations at these positions per lipase gene. After confirming that the wt enzyme could be functionally displayed on phage, selections with the library were performed utilizing a mechanism-based biotinylated inhibitor in the presence of a detergent formulation. According to rhodamine B-based activity assays, the fraction of active clones increased from 0.2 to 90 % over three rounds of selection. Although none of the variants selected using this approach showed increased activity, in either the presence or absence of detergent compared to the wild type enzyme, the results demonstrated the possibility of selecting variants of the enzyme based on catalytic activity.

In the following work, phage libraries of the Staphylococcal Protein A (SPA)-derived Z-domain, constructed by randomization of 13 surface-located positions, were used to isolate Z domain variants (affibodies) with novel binding specificities. As targets for selections, the parental SPA domains as well as two previously selected affibodies directed against two unrelated target proteins were used. Binders of all three targets were isolated with affinities ($K_d$) in the range of 2-0.5 μM. One SPA binding affibody ($Z_{SPA,1}$) was shown to bind to each of the five homologous native IgG-binding domains of SPA, as well as the Z domain used as the scaffold for library constructions. Furthermore, the $Z_{SPA,1}$ affibody was shown to compete with one of the native domains of SPA for binding to the Fe part of human antibodies, suggesting that the $Z_{SPA,1}$ affibody bound to the Fe-binding surface of the Z domain. The majority of the affibodies isolated in the other two selections using two different affibodies as targets, showed very little or no binding to unrelated affibodies, indicating that the binding was directed to the randomized surface of their respective targets, analogously to anti-idiotypic antibodies.

The structure of the wild type Z domain/$Z_{SPA,1}$ affibody co-complex was determined by x-ray crystallography, which confirmed the earlier findings in that the affibody $Z_{SPA,1}$ affibody was shown to bind to the Fe binding surface of the Z domain. Further, both the Z domain and the $Z_{SPA,1}$ affibody had very similar three helix-bundle topologies, and the interaction surface involved ten out of the thirteen randomized residues, with a central hydrophobic patch surrounded by polar residues. In addition, the interaction surface showed a surprisingly high shape complementarity, given the limited size of the library used for selections.

The $Z_{SPA,1}$ affibody was further investigated for use in various biotechnological applications. In one study, the $Z_{SPA,1}$ affibody was successfully recruited as a novel affinity gene fusion partner for production, purification and detection of cDNA-encoded recombinant proteins using an SPA-based medium for affinity chromatography. Further, the SPA binding capability of the $Z_{SPA,1}$ affibody was employed for site-specific and reversible docking of $Z_{SPA,1}$ affibody-tagged reporter proteins onto an SPA fusion protein anchored to a cellulose surface via a cellulose-binding moiety. These generated protein complexes resembles the architecture of so-called cellulosomes observed in cellulolytic bacteria. The results suggest it may be possible to use anti-idiotypic affibody-binding protein pairs as modules to build other self-assembling types of protein networks.

© Malin Eklund

Keywords: phage display, selection, mechanism-based inhibitor, affinity domains, crystal structure, Staphylococcus aureus protein A, affinity chromatography, anti-idiotypic binding pairs, affibody, combinatorial, protein engineering, lipase, cellulosome, assembly.
“Nothing is a waste of time if we use the experience wisely”

Auguste Rodin
Till min familj
This thesis is based on the following manuscripts, which will be referred to in the text by their corresponding Roman numbers.


* These authors contributed equally to the work and should therefore be considered as joint first authors.
# Table of contents

*Introduction* .......................................................................................................................... 1

1. Proteins ................................................................................................................................. 1

2. Protein engineering .............................................................................................................. 2
   2.1 Chemical and enzymatic modifications .......................................................................... 4
   2.2 Modifications by genetic engineering ........................................................................... 5

3. Rational design of proteins .................................................................................................. 6
   3.1 Site-directed mutagenesis .............................................................................................. 7
   3.2 Extensions and truncations ............................................................................................. 8
   3.3 De novo design of proteins ........................................................................................... 9

4. Protein library technology/combinatorial protein engineering ......................................... 10
   4.1 Sources of diversity ....................................................................................................... 10
      4.1.1 Sources of diversity: natural .................................................................................. 11
      4.1.2 Sources of diversity: man-made ............................................................................ 12
         4.1.2.1 Synthetic oligonucleotides ............................................................................ 12
         4.1.2.2 Error-prone PCR ......................................................................................... 15
         4.1.2.3 DNA Shuffling ............................................................................................. 16
      4.2 Linking genotype with phenotype ............................................................................. 18
         4.2.1 Cell-dependent systems: phage display ............................................................. 19
         4.2.2 Cell-dependent systems: other examples ........................................................... 24
            4.2.2.1 Cell-surface display .................................................................................... 24
            4.2.2.2 Plasmid display .......................................................................................... 25
            4.2.2.3 In vivo-based systems ................................................................................. 26
      4.2.3 Cell-free approaches ............................................................................................... 28
         4.2.3.1 Ribosome display ......................................................................................... 29
         4.2.3.2 mRNA-peptide fusion ............................................................................... 29
         4.2.3.3 In vitro compartmentalization and micro bead display ................................ 30

5. Protein library techniques in practice .................................................................................. 32
   5.1 Enzymes ....................................................................................................................... 32
   5.2 Peptides ....................................................................................................................... 38
   5.3 Antibodies .................................................................................................................... 39
   5.4 Alternative binding proteins ....................................................................................... 41
   5.5 Affibodies .................................................................................................................... 44

*Present Investigation* ............................................................................................................. 48

7. Catalytic selection of lipase variants from phage display libraries (I) ..................... 48
   7.1 Background .................................................................................................................. 48
   7.2 Functional display of Lipolase on phage ................................................................... 50
   7.3 Library constructions ............................................................................................... 51
   7.4 Selection using a mechanism-based inhibitor ............................................................. 53

8. Selection and analyses of anti-idiotypic affibody binding pairs (II, IV)............ 56
8.2 Crystal structure of the $Z/Z_{SPA,1}$ affinity protein pair ................................................. 61

9. Applications of hetero dimeric binding pairs (III, V) ............................................ 65
   9.1 Purification and detection ................................................................................................. 65
   9.2 Development of an artificial cellulosome-like complex .............................................. 67
      9.2.1 The cellulosome ........................................................................................................ 68
      9.2.2 Artificial cellulosomes (V) ....................................................................................... 69

10. Concluding remark .......................................................................................................... 73

Abbreviations ......................................................................................................................... 75

Acknowledgements ................................................................................................................ 76

References ............................................................................................................................. 78

Original papers (I-V)
Introduction

1. Proteins

Living cells depend on complex, inter-linked networks of interactions, involving many types of biomolecules with diverse activities and characteristics. The work described in this thesis primarily relates to one class of biomolecules called proteins, which are of central importance in a variety of cellular functions. For example, a wide range of different enzymes (catalytic proteins) are responsible for the joining, splitting and conversion of a vast array of other biomolecules and are essential in both metabolic and anabolic processes, such as food digestion and bone formation. Our immune system, protecting us from a variety of potential pathogens, such as bacteria and viruses, depends on many different cell-anchored and soluble proteins including the huge arsenal of circulating antibodies, involved in the selective recognition and destruction of the foreign intruders. Other proteins are involved in transportation (e.g. serum albumin), and storage (e.g. ferritin). Proteins are also involved in signaling systems that regulate cellular growth and differentiation, in which highly selective receptor proteins respond to particular protein hormones excreted from tissues in response to various stimuli. In addition, proteins are also employed as building materials in structures that provide mechanical strength, such as skin and cartilage.

Strikingly, despite these extremely diverse functions, all proteins consist of combinations of just 20 different amino acids (each with unique chemical characteristics), used as building blocks by the cellular transcription and translation machinery to produce linear polymers of varying lengths and compositions according to the genetic information packages (genes) stored in the genome. Following synthesis, the linear amino acid chains folds spontaneously (more or less) into the three-dimensional structures required for their respective functions. These structures are in turn composed of secondary structure sub-elements such as helical (α-helices) and more planar (β-strands and β-sheets) arrangements, linked by loops or short turn sequences. The limited number of building blocks is an attractive feature from an engineering perspective.

According to our current knowledge the human genome contains approximately 30,000 different genes (Lander et al., 2001; Venter et al., 2001), accounting for the complexity of human beings. However, the actual number of different protein species is thought to be significantly larger, due to alternative modes of processing gene transcripts and post-
translational modifications of proteins expanding the repertoire of functionally different proteins.

Many proteins purified from natural sources are commonly used biotechnologically and medically nowadays, in a wide spectra of applications e.g. calf rennin enzyme for cheese making, various enzymes isolated from microorganisms for molecular biology procedures and proteins isolated from human plasma, such as factor VIII and immunoglobulin preparations for therapeutic applications.

The advent of recombinant DNA technology in the 1970’s, boosted by the discovery of tools for precise cutting (restriction enzymes) and rejoining (ligase) of DNA pieces (Linn and Arber, 1968), together with methods for DNA sequencing (Maxam and Gilbert, 1977; Sanger and Coulson, 1975; Sanger et al., 1977) has allowed novel approaches to be developed to design and produce proteins using diverse host cell systems. For instance, the drug insulin, which was originally isolated from bovine and porcine pancreas, can be produced by recombinant DNA technology in the bacterium Escherichia coli. The possibility to produce insulin in a microorganism allowed a protein identical to the human variant to be produced on an industrial scale without the risk of the final product being contaminated with potentially hazardous agents, such as mammalian viruses, resulting in the launch of the first biopharmaceutical by Eli Lilly in 1982 (Swartz, 2001).

2. Protein engineering

In some instances it is desirable to modify a protein to improve its performance in biotechnological or therapeutic applications. Such modifications may be introduced for a number of reasons, e.g. to prolong its in vivo half life, to increase its stability and solubility, raise its resistance to high pH (or to chemicals such as bleach) or to reduce/increase its size. The first attempts to modify proteins were based on chemical or enzymatic approaches, but later the introduction of genetic engineering tools enabled modifications to be made at the gene level, allowing precise changes to be generated including substitutions, deletions and insertions/extensions of single amino acids or larger segments (Fig. 1).
Fig. 1. Examples of the protein engineering principles available today.
The DNA shuffling principle shown in more detail: A pool of homologous genes is randomly cleaved with DNAse I, and fragments of a specific length are extended with DNA-polymerase followed by the amplification of full-length fragments using outer PCR-primers. Abbreviations: P; phosphorylation, PEG; polyethyleneglycol, GA; glutaraldehyde and a sugar molecule.
2.1 Chemical and enzymatic modifications

An early example of simple, non-covalent, protein modification was the addition of Zn-ions to insulin preparations, which was shown to increase the stability of the protein both during storage and in vivo after injection.

Covalent chemical modifications can also be used to alter protein’s properties, by either non-directed or directed methods. The use of glutaraldehyde, for example, a well-known bifunctional chemical crosslinking agent that reacts with primary amine groups, was used as early as 1964 to stabilize the enzyme carboxypeptidase, facilitating structural determination by x-ray crystallography (Quirocho and Richards, 1964). This enzyme only retained five percent of its activity after the treatment, but other enzymes such as thermolysin and subtilisin have been shown to retain high levels of activity after glutaraldehyde treatment (St Clair and Navia, 1992; Wang et al., 1997 or DeSantis and Jones, 1999). Another agent used for chemical modification is the amphiphilic polymer polyethylene glycol (PEG). PEG has been used for various purposes, e.g. to increase protein solubility in organic solvents, to reduce antigenicity and to prolong the serum half-life of certain proteins (DeSantis and Jones, 1999; Marshall et al., 2003). Indeed, a number of PEGylated protein drugs are commercially available now (Marshall et al., 2003), including a drug called PEGasys® (peginterferon alfa-2a) for treating hepatitis C. Compared to the native protein, PEGasys® exhibits a 50- to 70-fold increase in serum half-life and a reduced variability in serum concentration, but with the drawback of having a lower specific activity (Bailon et al., 2001). Using reagents that selectively address certain amino acid side chains, chemical modifications can, in some instances, be directed to defined locations within a protein. Two good examples of this were reported as early as 1966, when two groups were able to convert the serine in the catalytic triad of the serine protease subtilisin to a cystein, using a protease inhibitor to direct the chemical conversion. (Polgar and Bender, 1966; Neet and Koshland, 1966).

Enzymatic methods can also be used to modify proteins, targeting naturally occurring substrate sites, or to sites introduced by genetic engineering. An early example was the enzymatic conversion of porcine insulin into human insulin by the conversion of the C-terminal alanine of the B-chain into a threonine by trypsin transpeptidation (Markussen, 1982). Proteases can be used to digest proteins at specific sites, including for example trypsin, which can be used in combination with carboxypeptidase for in vitro-digestion of proinsulin into insulin, and the proteases papain and pepsin, which have been used for a long time to
digest whole antibodies into subfragments such as Fc, Fab and F(\(ab')_2\) (Fig. 7, p.40). Other examples of enzymatic modification include site-specific biotinylation using the \(E.\ coli\) enzyme BirA (Saviranta et al., 1998) and phosphorylation and de-phosphorylation using kinases and phosphatases, respectively (Parker et al., 1991; Zhang et al., 1992).

### 2.2 Modifications by genetic engineering

A major breakthrough in the field of protein modification was the development of methods for site-directed mutagenesis at the genetic level in the laboratory of Michael Smith, using synthetic oligonucleotides (Hutchison et al., 1978). The first modification of an enzyme, a tyrosyl-transfer RNA synthetase, using these tools was performed in 1982 (Winter et al., 1982). Through site-directed mutagenesis a cysteine was replaced by a serine altering the protein’s substrate binding characteristics. His pioneering work within this field earned Michael Smith the Nobel Prize in 1993. He shared the prize with Kary B. Mullis who invented the polymerase chain reaction (PCR), which allows researchers to amplify tiny amounts of DNA, down to a single copy of a sequence (Saiki et al., 1985). Besides its use in a vast number of applications involving detection and analysis of genetic material, variants of the originally described methodology have had a great impact as tools for the introduction of both specific and random mutations in DNA (Ling and Robinson, 1997) (Fig. 3).

These genetic tools are today used routinely in protein engineering, a branch of protein science related to the design and production of engineered variants of proteins for both basic research and applied projects, involving substitutions, deletions and insertions of single amino acids, short sequences or complete domains. The arsenal of genetic tools is increasing constantly, including methods to synthesize peptides and smaller proteins artificially (Merrifield, 1963; Gutte and Merrifield, 1971) and ligate them with recombinant fragments (Abrahmsen et al., 1991), to introduce non-natural amino acids in synthetically produced proteins (Noren et al., 1989, Bain et al., 1989) or in ribosome-synthesized sequences (Chin et al., 2003) and to create randomly distributed mutations along part of or along the whole gene.

Protein engineering can involve either (or both) of two main approaches. In the first, existing knowledge and structural information are used to identify, rationally, changes that need to be made to obtain specific protein traits. In the second, random methods are employed that allow searches to be made for new functions without having very good knowledge about the required changes. In some instances, a random approach can be combined with an input of rational design to reduce the size of the so-called sequence space that needs to be examined.
3. Rational design of proteins

As mentioned above, one approach to engineering proteins with novel properties is to design new variants in a rational manner, involving substitutions, insertions and deletions of single amino acids and larger sequences, including domains (Fig. 1).

Making accurate predictions about the necessary modifications requires at least some knowledge about the protein. The rapid growth in number of three-dimensional structures, that have been solved by x-ray crystallography or nuclear magnetic resonance (NMR) (Berman et al., 2000) either alone or in complexes with other molecules, has contributed enormously to our understanding of protein structure and function. Structural information can help one to identify important residues or regions involved in catalysis, substrate or ligand binding, and to make predictions about how to modify the protein in order to, for example, increase stability, solubility and/or alter other properties. Despite the vast number of structures available today, the structure has been solved for only a fraction of all known proteins. However, studying the structures of homologous proteins can give valuable information, and can also be used to make predictions about the actual structure of other proteins using homology computer modeling (Marti-Renom et al., 2000).

If structural data is not available, a scanning mutagenesis methodology can be an option. Areas believed to be of importance are analyzed by substituting one amino acid at a time in separately expressed protein variants, followed by a functional analysis to explore the effect of the substitution. Alanine is an amino acid that is often used as a substitute in these kinds of analyses, referred to as alanine scanning. Alanine is considered a good substitution candidate because it does not change the peptide-chain orientation, as do glycine and proline, and it does not possess extreme steric and electrostatic characteristics (Cunningham and Wells, 1989). In cases where a structure is available this methodology can also be used to analyze the contribution of each amino acid believed to be involved in a specific function. Alan Fersht and co-workers (and others) have used this method extensively, for example to delineate the co-activator, inhibitor binding site and key residues involved in the allosteric transition of phosphofructokinase from E. coli (Lau and Fersht, 1989). Other properties that have been analyzed are the loss of stability when hydrophobic amino acids in the interior of a protein are mutated (Kellis et al., 1989), and the stability of β-sheets and α-helices (Otzen and Fersht, 1995; Serrano et al., 1992). Once an amino acid residue or region has been identified as being of importance for a specific trait, other more appropriate substitutions can be made, taking
into account secondary structure propensities of the different amino acids (Chou and Fasman, 1974), as well as other considerations, relevant to the specific goals of the project.

3.1 Site-directed mutagenesis

The literature is full of examples in which one or a few substitutions in proteins, identified by a variety of methods, have had dramatic effects on protein function. One of the earliest examples of the successful alteration of an enzyme through site-directed mutagenesis was reported by Jim Wells and co-workers (Estell et al., 1985). This work was done on the *Bacillus amyoliquefaciens* serine protease subtilisin, used in laundry powder. A methionine at position 222 close to the catalytic triad (which had earlier been shown to be susceptible to chemical oxidation resulting in an inactive enzyme) was mutated to all the other 19 amino acids, due to uncertainties about which amino acid would be most suitable. Protein variants containing a valine or an alanine at the mutated position appeared to have the best combination of high resistance toward oxidation together with high activity in the presence of 1 M H$_2$O$_2$.

Another illustrative example of rational design involved insulin. At higher than physiological concentrations, insulin self-associates into dimers and hexamers. This trait is an inherent property that facilitates pancreatic storage, but limits its association with the receptor (which requires a monomeric state). Using computer-aided molecular modeling together with structural data related to dimeric complexes, amino acid positions were identified for substitutions that would, for example, introduce charge repulsion and steric hindrance at the interaction surface. One such example is Novolog®, an insulin analogue in which a proline (at position B28) has been substituted for an aspartic acid, introducing charge repulsion between the two monomers (Brange, 1997). An alternative strategy, designed to disturb the insulin sub-chains, thereby decreasing the potential for β-sheet interactions between insulin monomers, results in the "LysPro" mutant, also named Humalog® (Brems et al., 1992), in which the positions of a lysine and a proline have been switched. An additional example is Aranesp®, which is a variant of recombinant erythropoietin with two additional N-glycosylation sites, introduced by site-directed mutagenesis (Macdougall et al., 1999). Aranesp® exhibits a prolonged half-life as compared to the wild type and is used for treating anemia. (The cross-country skier Johan Muhlegg used this drug in the winter Olympic games in 2002).
3.2 Extensions and truncations

Another approach to the rational design of novel proteins is the construction of fusion proteins and deletion mutants. Through gene fusion techniques, fusion proteins can be produced comprising two or more parts (complete proteins, domains or peptide sequences) derived from different sources, resulting in novel proteins with a combination of properties from each of the constituents. Numerous fusion proteins have been constructed over the years for a vast number of applications, often involving the recruitment of a desired function from one protein to another protein, as illustrated by the following examples. Enzymes such as alkaline phosphatase and β-galactosidase have been used as fusion partners to binding proteins to recruit a reporter function. Fc fragments (Chamow and Ashkenazi, 1996), serum albumin and serum albumin binding proteins (Makrides et al., 1996; Dennis et al., 2002) have been used as fusion partners to increase the in vivo circulation half-life of biotherapeutics. Different fusion partners derived from cell-anchored proteins have been used to present proteins on outer surfaces of the cells. (Samuelson et al., 1995; Francisco et al., 1992; Lee et al., 2003). A widely applied technique is to use gene fusion partners with affinity for a specific binding partner molecule that can be used as immobilized ligand in affinity chromatography. (Nilsson et al., 1997; Hearn and Acosta, 2001). In addition to bioseparation applications, such systems can also be used for detection and immobilization purposes.

The removal of domains has also been frequently exploited as a means to tailor make a protein for specific situations. Such approaches benefit from the often modular structures of proteins, in which individual domains can be ascribed discrete functions, as illustrated by the extensively investigated tissue plasminogen activator protein, for which various deletion mutants have been studied (Rouf et al., 1996). Some applications involve the deletion of domains to facilitate recombinant expression. For example, genetic removal of the transmembrane region of surface receptors as described for the hER-b2 receptor (Schier et al., 1996) facilitates the expression of soluble protein. The constant parts of antibody fragments can be genetically removed to facilitate further engineering of the antigen binding regions, into the so-called single chain Fv (scFv) format, consisting of the variable regions connected by a short peptide linker sequence (Bird et al., 1988) (Fig. 7, p.40). The removal of domains to eliminate unwanted enzymatic functions has also been described, such as for DNA polymerases where, for example, the widely used Klenow DNA polymerase has been engineered from E. coli DNA polymerase I by deleting the N-terminal 323 amino acids to yield a variant lacking 5’ to 3’ exonuclease activity (Joyce et al., 1982). The same principle
was applied to the thermostable Taq DNA polymerase from Thermus aquaticus, resulting in the 5’ to 3’ exonuclease deficient “stofelf” fragment (Lawyer et al., 1989).

3.3 De novo design of proteins

An appealing strategy to obtain proteins of desired characteristics is to design proteins de novo, using acquired knowledge based on concepts of molecular recognition, conformational preferences and analyses of structures of native proteins together with the use of computational methods.

The design of proteins from scratch relies on knowledge that has been generated from analyses of naturally occurring proteins and their secondary structural elements, i.e. helices, sheets, loops and turns. How to create a $\alpha$-helix has actually been known for a long time and most of the work today is focused on how to create secondary structure elements that fold into pre-determined three-dimensional structures. To encode all the necessary information into a polypeptide chain that not only control the formation of secondary structure elements but also the formation of super-secondary structures, is a highly complex problem.

Nevertheless, a number of examples exist were small folded domains have been designed, such as coiled coils (Lau et al., 1984), three-helix bundles (Bryson et al., 1998), $\beta\beta\alpha$-motifs (Struthers et al., 1996) and three-stranded $\beta$-sheets (Kortemme et al., 1998). In addition, different types of functionalities, such as catalysis, metal or co-factor binding have also been designed into these motifs (Baltzer, 1998). One such example is $\alpha$-helical peptides that catalyze helical peptide ligation, first reported by Ghadiri and co-workers (Severin et al., 1997). The positioning of the peptide substrate for catalysis is due to hydrophobic interactions together with electrostatic interactions between the two helices. Peptide ligation depends on the reaction between an N-terminal cysteine and a C-terminal thioester to form an amide bond between two peptide entities.

The design of larger protein structures has for the most part been limited to the redesign of existing proteins. The re-design of the hydrophobic core has been performed on a number of proteins, including the four-helix bundle ROP (Munson et al., 1994) and ubiquitin (Lazar et al., 1997). The hydrophobic core of both of these proteins was completely redesigned while still maintaining a native-like fold.

A quote from Baltzer and colleagues summarizes the field of de novo design of proteins as “We have moved from studying natural folds with natural sequences to making natural folds from unnatural sequences. The foundation has now been laid for the design of unnatural folds
from unnatural sequences” (Baltzer et al., 2001). Indeed, just recently a protein has been
designed with an unnatural fold from an unnatural sequence (Kuhlman et al., 2003). In this
work, the authors have created a 93-residue α/β-protein named Top7, with a novel sequence
and a topology not present in the protein structure database (PDB) using computational
protein design. The Top7 protein was found to be extremely stable and the solved x-ray
crystal structure was very similar to the modeled protein.

4. Protein library technology/combinatorial protein
engineering
As discussed above, alterations are predominantly targeted in rational protein engineering to
positions within or close to regions of proteins directly related to the function investigated,
such as the active site of enzymes or the interaction surfaces of binding proteins (these are the
most obvious areas to change to alter protein function). However, much knowledge has now
been acquired showing that dramatic effects on functionality can also be generated by
alterations relatively far from such functional “hot spots” (Moore and Arnold, 1996; Yano et
al., 1998). For example, in studies by Frances Arnold and co-workers to enhance the activity
of the enzyme pNB esterase from Bacillus subtilis in mixed aqueous-organic solvents,
screenings on engineered enzymes showed that some variants with improved activities had
mutations far from the substrate-binding site. Such substitutions would have been very
difficult to predict using rational methods (Moore and Arnold, 1996). These, and various
other, findings have inspired many researchers to investigate a larger sequence space in their
search for improved protein variants, involving the development of methodologies for random
mutagenesis, whereby genes or gene segments are genetically diversified in a random fashion.
Such diversification can typically be performed by either random or semi-random nucleotide
substitution or by random recombination of gene fragments (Fig. 1). As these approaches
often generate a large number of protein variants an efficient means of functional analysis is
needed, either by powerful screening methods or, if possible, by selection-by-function
techniques.

4.1 Sources of diversity
This field of protein science, in which large pools of variants (denoted libraries), are produced
and analyzed to identify variants with particular traits (e.g. desired catalytic activity, binding
capacity, substrate-related properties resistance to proteolysis) has seen a dramatic increase in
recent years, boosted by the development of efficient means to generate and handle
increasingly large libraries. Common to these strategies is the initial generation of a library at the gene level, from which proteins are expressed using a suitable strategy. Depending on the scope of the project and class of proteins investigated, such libraries can be generated from various natural sources or/and genetic variability can be generated by sophisticated genetic engineering techniques.

4.1.1 Sources of diversity: natural

Some applications of protein library technology involve the search for naturally existing protein variants, including for example allergy research and antibody technology. In such situations, a gene pool for use in the library construction may be obtained through the use of mRNA pools from a relevant source that can be reverse-transcribed to a corresponding cDNA pool. In one example, a cDNA pool derived from peanut was used as the gene library source, from which a peanut protein was isolated that was recognized by IgE molecules from a peanut-sensitive patient (Kleber-Janke et al., 2001). In other cases, chromosomal gene fragments have been used as sources of gene diversity. For example, chromosomal DNA fragments from different gram-positive bacteria have been used to construct libraries in efforts to isolate proteins that are involved in bacteria-host interactions and, thus, play an important role for bacterial infection (Jacobsson et al., 2003).

Of particular interest in recent years has been the use of protein library technology for antibody isolation and engineering. An important step in this field was the successful use of PCR technology together with degenerate primers to amplify large pools of immunoglobulin-encoding genes from donors in 1989 (Orlandi et al., 1989; Chiang et al., 1989). The first variable domain antibody libraries to be developed were derived from B-cells of immunized mice (Clackson et al., 1991) or B-cells from humans, which had been immunized with antigen (Persson et al., 1991), exposed to infectious agents (Burton et al., 1991) or were suffering from cancer (Cai and Garen, 1995). From such pre-immunized libraries high affinity antibodies directed against the antigen used for immunization or against disease-related antigens may be retrieved more easily. Later, libraries were constructed by routes that bypassed immunization and use of laboratory animals altogether, being made from non-immunized human donors. Such naïve libraries can be used as general sources for antibodies against virtually any antigen, including antibodies to self, non-immunogenic reagents and toxic substances. The first naïve library was made by amplifying IgM mRNA of B cells isolated from peripheral blood lymphocytes from non-immunized human donors (Marks et al., 1991), which successfully generated antibodies to a large number of antigens (Marks et
al., 1991, Griffiths et al., 1993). Later, numerous alternative approaches have been presented for constructing antibody libraries (Hoogenboom, 2002; Hudson and Souriau, 2003), including the use of synthetic DNA fragments (discussed in the next section) encoding variable antibody regions (Knappik et al., 2000) or through grafts of complementarity-determining regions (CDRs) from various natural sources into a single antibody framework (Jirholt et al., 1998). From such libraries, antibodies directed against peptides, proteins, carbohydrates and haptens have been isolated (Söderlind et al., 2000; Knappik et al., 2000).

In addition, genes encoding naturally existing protein homologous have been used as the starting gene pools in DNA-shuffling experiments, as discussed in the next section.

### 4.1.2 Sources of diversity: man-made

Several methods have been developed to allow diversity to be introduced into any genes, beyond the diversity found in nature, if desired. For instance, the use of designed synthetic oligonucleotides as tools for gene assemblies allows researchers to randomize at specific locations within a protein sequence; error prone PCR technology can be used to introduce base substitutions at random along large gene fragments; and DNA-shuffling technology allows the construction of libraries by recombining naturally occurring homologous genes or gene variants generated by either of the above methods (Fig. 1).

#### 4.1.2.1 Synthetic oligonucleotides

The most common way to introduce genetic variability through genetic engineering is to use synthetic oligonucleotides during gene assembly or PCR amplification.

Pioneered by Khorana and coworkers, who performed the first artificial synthesis of a gene in 1972 (Khorana et al., 1972), DNA synthesis techniques today are completely automated, reliable and relatively cheap. The DNA synthesis is performed in a stepwise manner by adding nucleotides to the 5’-end of the chain. Adding one nucleotide at a time in the desired order will generate a DNA strand with a customized sequence. To introduce genetic diversity into a synthetic oligonucleotide, encoding part of a target protein sequence, a mixture of all four nucleotides can be used instead of adding one base at a time. Adding all four nucleotides at all three positions of a given codon (NNN where N=A, C, G and T in equal concentrations) will create a gene pool allowing any of the 20 amino acids to be introduced at that specific location, and generate all 64 different codons of the genetic code, including the three termination codons. When multiple codons are subjected to randomization using NNN randomization, the ratio between the number of genes required to encode all possible protein
variants will increase rapidly with the number of codons addressed, meaning that the genetic library has to be much larger than the theoretical number of protein variants actually present in the library (Fig. 2B).

![Codon Table]

**Table A:** Codon representations of two different degenerate codons used in protein library constructions. Allowed codons are written in black.

**Table B:** The ratio between the number of genes generated as compared to protein sequences using the NN(G/T)-codon, in relation to the number of positions varied. The relation between the number of positions randomized and the number of possible protein variants generated.

**Fig. 2.** A. Codon representations of two different degenerate codons used in protein library constructions. Allowed codons are written in black. B. The ratio between the number of genes generated as compared to protein sequences using the NN(G/T)-codon, in relation to the number of positions varied. C. The relation between the number of positions randomized and the number of possible protein variants generated.

To reduce this bias, the alternative degenerate codons NN(G/T) or NN(G/C) are frequently used, which decreases the number of gene variants required to include all 20 amino acids from 64 to 32, including one termination codon instead of three (Fig. 2). Other codons used
are NAN (includes polar amino acids) and NTN (includes nonpolar amino acids) (West and Hecht, 1995). The codon ((G/A/C)N(G/T) excludes the incorporation of all stop codons, but also excludes all aromatic amino acids (Fig. 2).

In some instances it could be favorable to design each of the randomized codons more carefully, to favor the introduction of a specific amino acid, or set of amino acids. This can be achieved to a certain extent by optimizing the relative concentrations of each nucleotide at targeted base-positions during DNA-synthesis (Hermes et al., 1989). In addition, computer programs have been published that are designed to find the best compromise for specific requirements under the conditions dictated by the genetic code (Arkin and Youvan, 1992; Jensen et al., 1998; Ophir and Gershoni, 1995).

In 1994, Virnekäs and co-workers introduced an elegant solution to the problems caused by the inherent properties of the genetic code (Virnekäs et al., 1994). In this work, pre-formed trinucleotides were used instead of mononucleotides as building blocks for oligonucleotide synthesis. Using twenty different trinucleotides, one for each amino acid, makes it possible to adjust the proportions of each amino acid carefully, and to exclude any stop codons. In addition, each codon can be chosen according to the codon preferences of the chosen expression host. More recently, alternative methods to facilitate tailor-made design of oligonucleotides have been described, including the use of dinucleotide building blocks (Neuner et al., 1998) and a methodology based on hybridization, using 20 different short oligonucleotides to remove the redundancy of the genetic code (Hughes et al., 2003).

As mentioned earlier, produced oligonucleotides can then be introduced into a gene by methods such as splice overlap-extension PCR, other forms of cassette mutagenesis and by modified forms of standard PCR amplification (Fig. 3). The length of a synthetic oligonucleotide that can be synthesized with reasonable yield and correct sequence is presently limited to approximately 120-150 bases. This, together with the assembly method used, sets a limit on the size of protein sequences that can be mutagenized in one experiment, and oligonucleotide-based approaches are therefore mainly used to randomize certain areas of a protein or peptide sequence.
Fig. 3. Examples of methods used to introduce variability into a gene using synthetic oligonucleotides. A. The randomized codons are located in the middle of an oligonucleotide and the fragment is extended with a primer annealing to a constant part. B. Variability is introduced via a PCR primer. C. Two separate randomized oligonucleotides with overlapping constant parts that can anneal to each other providing 3'-ends for the polymerase to extend. Outer primers annealing to constant parts are used to amplify the resulting fragment.

4.1.2.2 Error-prone PCR

Nucleotide changes can also be introduced into a DNA fragment during amplification by PCR, taking advantage of the inherent (or further enhanced) error rate of the DNA polymerase used in the reaction (Fig. 1). The most common thermo stable enzyme used for PCR amplification of DNA is Taq DNA polymerase from Thermus aquaticus. This enzyme lacks proofreading activity (Tindall and Kunkel, 1988), resulting in an error rate of approximately $5.5 \times 10^{-4}$ mutations per base (Zhou et al., 1991). With this relatively low error rate, the amplified DNA fragment needs to be relatively long (Zhou et al., 1991; Zhao and Arnold, 1997; Stemmer, 1994) to ensure that mutations are introduced using standard PCR conditions. Therefore, several methods have been developed to increase the error rate, such as modifying the buffer composition (e.g. increasing the magnesium concentration, raising the pH or the addition of manganese), or using a high concentration of DNA polymerase, a small amount of template and a large number of cycles in the PCR reaction. A biased pool of the four dNTPs, with concentration differences of factors of 10 to 1000, also encourages a higher error rate (Ling and Robinson, 1997). Using various combinations of these methods, an error rate of up to approximately one base pair substitution per 150 base pairs can be achieved. A drawback of these PCR-based methods is that different regions of a gene tend to have different error rates.
rates. The resulting mutants also tend to have a bias towards specific substitutions, e.g. transitions A to G and T to C are more common than transversions (A to C, T to G, A to T or G to C).

An additional method that can induce the DNA polymerase to make errors is to use deoxyinosine triphosphate (dI) or other ambiguous degenerate nucleotide analogues. Such analogues can be incorporated in the place of one or more of the four natural dNTPs. The dI-containing DNA strands then serve as templates for subsequent cycles of DNA amplification, in which any one of the four natural nucleotides can be incorporated opposite the previously introduced dI nucleotide analogue. In practice, however, there is still a bias towards certain substitutions. Zaccolo and coworkers successfully minimized the bias of transitions over transversions using two different analogues in combination, and at the same time achieved a very high mutation rate (1.9x10^{-3}) (Zaccolo and Gherardi, 1999). Furthermore, the mutation frequency can be fine-tuned by adjusting the number of cycles used during the PCR reaction in the presence of the nucleotide analogues.

### 4.1.2.3 DNA Shuffling

If the starting material for a directed evolution project is a single gene, diversity can be accumulated through error-prone PCR, as mentioned above. The resulting pool of gene fragments will thus contain mutations that could be advantageous, neutral but also deleterious to the trait investigated. Thus, the evolution of a desired function using this approach alone can be relatively slow.

A technological breakthrough in the field was seen in 1994, when an elegant method for PCR-based in vitro recombination of homologous genes was presented (Stemmer, 1994a; Stemmer, 1994b). With this method it became possible to recombine gene variants into full-length (mosaic) hybrids, containing different combinations of sub-fragments originating from the original genes used as input templates (Fig.1).

DNA shuffling has revolutionized the ability to create improved protein variants as well as novel proteins. The starting material is a pool of homologous genes; either variants of a gene existing in nature or a pool created by genetic engineering. Typically, input genes are digested randomly by DNAse I, then fragments of a specific length or a range of lengths are purified from a gel after being electrophoretically separated. The isolated fragments are assembled by thermo cycling using a DNA polymerase followed by conventional PCR using outer primers for the amplification of full-length fragments (Stemmer, 1994a). If a DNA polymerase with
low fidelity is used, point mutations are generated throughout the assembly and amplification steps, generating additional diversity, which may be either advantageous or disadvantageous, depending on the project. The frequency of such additional mutations can be fine tuned by, for example, using fewer cycles during thermo cycling, employing a proof reading thermostable enzyme or adding appropriate reagents, such as Mn^{2+} (Zhao and Arnold, 1997). In his pioneer work, cited above, Stemmer used the gene encoding the enzyme β-lactamase as starting material for a directed evolution experiment. After three cycles of shuffling and two cycles of backcrossing with wt-DNA (to limit the amount of non-essential mutations) and selection on increasing amounts of antibiotic-containing agar plates between amplification rounds, a variant 32,000 times more resistant to the antibiotic than the wt was isolated (Stemmer, 1994b).

When the starting material for DNA shuffling is a pool of homologous genes encoding functional variants of a protein, the genetic variants that are recombined are associated with functional variations, often avoiding the introduction of deleterious mutations, thus accelerating the search for variants with desired traits. The starting material for such exercises may be either a series of homologous genes derived from different species, or genes that have already been varied (e.g. by error-prone PCR) and screened for a desired trait, such as binding to a certain epitope.

DNA shuffling has been successfully used to evolve a number of different proteins in vitro. For example, in a study by Crameri et al. (Crameri et al., 1996) a single Green Fluorescent Protein (GFP)-gene was subjected to three rounds of PCR-induced mutation and DNA shuffling (in which the brightest colonies were used as input in following rounds), resulting in the identification of a variant with a 45-fold increase in whole cell fluorescence signal. In another illustrative experiment, 26 different subtilisin genes (with 63.7-99.5% pairwise protein sequence identity) from Bacillus comprised the input gene pool. Interestingly, the variants displaying the highest increased in thermo- or alkaline-stability contained segments from no less than 25 of the 26 parents (Ness et al., 1999; Kurtzman et al., 2001).

Today, a wide range of alternative DNA recombination methods is available (Kurtzman et al., 2001) including, RAAndom CHImeragenesis on Transient Templates (RACHITTT), which has been described to result in a very high recombination frequency and 100% chimeric gene products (Coco et al., 2001). The technique involves joining randomly cleaved single-stranded parental gene fragments that are reassembled on a single-stranded full-length template. Analysis of 175 unselected clones generated by this method, using two homologous
(89.9%) monoxygenase genes as starting material, resulted in six different variants with an increased substrate turnover compared to the wild type enzyme (Coco et al., 2001). Sequence Homology-Independent Protein RECombination (SHIPREC) (Sieber et al., 2001) and Incremental Truncation for the Creation of Hybride enzYmes (ITCHY) (Ostermeier et al., 1999) methods can be used to create libraries of single-crossover hybrids of unrelated genes by ligating the blunt ends of two truncated genes of variable lengths. Recently, the ability of ITCHY to make single crossover hybrids of non-homologous genes was combined with DNA-shuffling, resulting in a gene library with multiple crossovers independent of sequence homology (Lutz et al., 2001). Using a strategy based on exonuclease digestion of a starting gene pool, rather than digestion by DNase, an alternative method to create fragments for subsequent DNA shuffling has also been described (Borrebaeck et al., 2003)

### 4.2 Linking genotype with phenotype

The methods described for introducing diversity into a gene can generate huge numbers of protein variants that need to be investigated for the targeted protein trait, followed by sequence identification. To solve the problems involved in handling vast numbers of variants, a wide range of high throughput methods has been developed to assay protein functions. Because DNA is much easier to sequence than proteins, and DNA can be amplified, methods have been developed for physically linking each protein variant with its corresponding DNA sequence, i.e. linking phenotype with genotype. In addition, methods for identification/isolation of protein variants with a desired functionality need to be developed. There are two main approaches for this: screening (examining each protein variant separately for a specific property) and selection (processing only variants that show desired traits). The latter approach has obvious advantages.

Many different methods have been developed for linking the phenotype with the genotype of a specific library member. In “cell-dependent” (e.g. phage display and cell surface display) systems each member of a gene library expresses its corresponding protein inside a cell. The library size of such cell-dependent systems is limited by the transformation efficiency, typically resulting in libraries containing between $10^7$-$10^{10}$ members (Hoogenboom et al., 1998). Therefore, alternative, cell-free systems have been developed that circumvent the transformation step by performing transcription and translation of library members entirely in vitro. Libraries with up to $10^{14}$ (Kreider, 2000) members have been obtained using cell-free systems.
4.2.1 Cell-dependent systems: phage display

By far the most commonly used method of linking a gene with its corresponding protein is phage display, first introduced by George Smith in 1985 (Smith, 1985). This technology is based on the presentation of individual protein library members on phage surfaces through their genetic fusion to a phage coat protein, while the corresponding gene is packed inside the phage particle (as single-stranded DNA). Using relatively simple microbiological methods, solutions containing high concentrations of phage ($\approx 10^{13}$ phage/ml) can routinely be obtained, making the technology suitable for library applications. Many different types of E. coli phages have been used as vehicles for phage display, including Ff filamentous phage, lambda and T7 (Rodi and Makowski, 1999; Danner and Belasco, 2001). Each system has advantages and disadvantages with respect to particular applications. The Ff phage family includes M13, fd and fl phages, which are the most commonly used phages for display. These phages are considered good cloning vectors since the relatively large genomes resulting from insertion of a foreign gene can be simply accommodated by the assembly of longer phage particles. On the other hand, it must be possible for all the components of the phage coat to be exported through the bacterial inner membrane if mature phage particles are to be assembled, because of the non-lytic propagation mechanism of Ff phage. Consequently, only proteins that can be exported in this way can be displayed. This limitation can be avoided using the lytic phages lambda and T7, in which capsid assembly occurs entirely in the cytoplasm prior to cell lysis.

As mentioned earlier, the E. coli bacteriophage M13 is the most commonly used display system, and since it was first described several different variants have been developed, based on genetic fusion of library members to its different coat proteins and/or the use of various types of vector systems. The M13 phage particle has a rod-like structure about 1 $\mu$m in length, mainly composed of the major coat protein. At one end of the phage there are five copies of the minor coat proteins pIII and pVI, and at the other end there are five copies of two additional minor coat proteins, pVII and pIX (Fig. 4) (Hoess, 2001). Phage coat proteins pIII and pVIII are the most widely used phage proteins for display, but pVI, pVII and pIX have also been used. Because of its orientation on the viral particle, pVI has been used to display protein fusions at its C-terminus rather than the typical N-terminal fusions of other coat proteins. This makes pVI useful for projects involving the display of proteins encoded by cDNA fragments, which could cause severe translational problems if fused to the 5´-end of a phage coat protein gene since they often contain stop codons and poly A-tails at their 3´-ends (Jespers et al., 1995, Hufton et al., 1999). The minor coat proteins pVII and pIX are located
close to each other on the surface of the phage allowing the simultaneous display of both the heavy- and light-chain variable regions of an antibody, which can spontaneously associate into a functional Fv-binding domain (Gao et al., 1999). A more frequently used coat protein is pVIII, present at approximately 2700 copies in wild type M13 phages (Fig. 4). Depending on the type of phage display strategy used, different fractions of the 2700 copies are utilized. pVIII has been used to display of a variety of different classes of proteins, ranging from peptides (at high numbers) to larger proteins like enzymes and antibody fragments (at lower numbers) (Benhar, 2001).

Fig. 4. Different phage display (pIII) systems; A. Schematic drawing of a wt M13 phage. The positions of the five different structural coat proteins are indicated. B. The foreign gene is fused to the pIII coat-protein gene in the phage genome (type 3), which theoretically results in the display of the foreign protein to all copies of pIII. C. In the type 33 system two copies of the pIII gene is present in the phage genome. One of the copies is fused to the foreign gene, which results in the production of both wt pIII and pIII fused to the foreign protein. D. In a type 3+3 system the foreign gene is cloned into a phagemid vector containing the phage ori of replication and packaging signal but devoid of any genes for structural phage coat proteins. The wild type pIII and all other phage genes necessary for the phage particle assembly are provided by superinfection with helper phage.
Most applications of M13 phage display described to date have been based on the use of pIII as the fusion partner for library members. The mature pIII molecule consists of three discrete domains: two N-terminal domains (denoted N1 and N2), involved in phage infectivity followed by a C-terminal domain (CT), involved in phage assembly (Marvin, 1998). Depending on the system, different extensions of the pIII protein are used (see below). Vectors for phage display can be of different types, involving either whole phage DNA, in which a foreign gene is fused to the single wt gene for pIII (a “type 3” system), or an extra expression cassette encoding the foreign protein fused to a pIII gene is introduced (a “type 33” system), providing two sources of pIII gene products. Alternatively, pIII fusions can be expressed from phagemid vectors, which contain the origins of replication for both M13, including the packaging signal, and E. coli in addition to a coat protein gene. Such vectors lack all other structural and non-structural gene products required for generating a complete phage. Phagemids (Bass et al., 1990) can be grown as plasmids or packaged as recombinant M13 phage with the aid of a helper phage that contains a slightly defective origin of replication and supplies, in trans, all the structural proteins for generating a complete phage (a “type 3+3” system) (Fig. 4). When systems supplying two sources of pIII gene products are used, such as 33 or 3+3 systems, the resulting phage particles may incorporate either the fusion protein (supplied by the extra pIII gene construct or phagemid), or the wild type coat protein (encoded by the wild type pIII copy or helper phage). The average number of displayed proteins per phage in such systems is approximately one on phages displaying the foreign protein, whereas most particles display only wt pIII proteins. These systems are proposed to have two advantages over systems in which all copies of the pIII are decorated with the foreign protein: (i) infectivity is not compromised, since wild type copies of pIII are also provided and (ii) the reduced valency of display circumvents avidity effects, which can be advantageous during the selection process. In addition, phagemid systems can facilitate post-selection procedures to produce selected protein variants without parallel production of phage particles.

The most common selection scheme is based on affinity selection, usually referred to as bio-panning. Typically, the target is immobilized on a solid support and the phage library is incubated with the support to allow binding between the immobilized target and the appropriate phage. After a suitable incubation time, unbound phage are removed, then nonspecifically bound phage are washed off by an appropriate procedure. Elution of specifically bound phage can be accomplished by various means, including brief incubation at
low pH (Parmley and Smith, 1988), alkaline buffers (Harrison et al., 1996), competitive elution (Harrison et al., 1996), infection, proteolytic cleavage (Matthews, 1996) or other approaches exploiting susceptible elements engineered into the system (Parmley and Smith, 1988). The eluted phage are used to reinfect fresh *E. coli* cells and can thus be amplified for use in additional rounds of bio-panning, often under increasingly stringent conditions, or used for identification and analysis (Fig. 5). The aim is to enrich clones that encode variants capable of specifically binding to the immobilized target with high affinity, and as the number of cycles increases (typically 3 to 5) the phage population becomes progressively less diverse, allowing the identification of the "best" variant(s).

In addition to this standard procedure, many alternative selection techniques using phage display have been developed, a few examples of which are highlighted below. In the so-called "selectively-infective phage" (SIP) technique, the phage particles displaying library members are made non-infectious through the use of different, truncated extensions of the pIII protein as fusion partner, all lacking the N1 domain which is crucial for infectivity (Duenas and Borrebaeck, 1994; Gramatikoff et al., 1994, Krebber et al., 1995; Nilsson et al., 2002). In general terms, these systems rely on phages that display a protein variant that can bind to a target-N1 fusion or conjugate, resulting in infective phages that can be subsequently enriched by cultivation of *E. coli*. An early example of this technique was an antibody selection study in which anti-hen egg white lysozyme (HEL) antibody fragments were displayed on phage as fusions to pIII lacking N1, rendering them non-infective. When these phage were assembled in *E. coli* cells expressing an HEL-N1 fusion protein, the binding interaction between the anti-HEL antibody fragment and HEL restored the pIII function, and thus phage infectivity (Duenas and Borrebaeck, 1994).

Further developments of the phage display approach include its use in searches to identify substrate sequences for specific proteases. After immobilization of a peptide phage library on a solid support via a common protein affinity interaction, treatment with a specific protease ensures that phage displaying a peptide containing the substrate sequence for the targeted protease is eluted from the matrix (Matthews and Wells, 1993; Matthews, 1996), allowing later identification by DNA sequencing. In a reversed application, insertion of protein library members between domains of pIII has allowed the selection of variants showing resistance to proteolytic degradation (Sieber et al., 1998). The possibility of enriching for stable variants by biopanning under suitable biophysical stresses has also been demonstrated (Jung et al., 1999).
Fig. 5. The bio-panning procedure. A pool of protein variants are displayed, in this example, on phagemids and the library is incubated with immobilized target. After washing, to remove non-specific and weakly binding phage, the bound phage are eluted and re-infected into *E. coli* cells. At this stage it is possible to analyze the DNA sequence of the resulting clones or soluble protein can be expressed and analyzed, for example, using biosensor analysis. A new phage stock can be prepared by superinfection of helper phage and allowed to interact with immobilized target a second time to further enrich positive clones. Analysis of the phage library can be done by, for example, an ELISA assay.
4.2.2 Cell-dependent systems: other examples

4.2.2.1 Cell-surface display

Cell surface display is a cell-dependent system in which the target protein is presented on the cell surface by fusing the protein library members to an outer membrane protein or to a protein anchored in the cell wall. Thus, the cell itself provides the link between genotype and phenotype, in that the gene is compartmentalized inside the cell and the corresponding protein is displayed on its surface. The first examples of bacterial display were reported as early as 1986, when gene fragments were inserted into the genes encoding the E. coli outer membrane proteins LamB, OmpA and PhoE and the gene fusion products were found to be accessible on the surface of the recombinant bacteria. Methods for cell surface display were not initially introduced to screen protein libraries, which has become one of the main applications in later years. The most common application of bacterial display was, instead, the development of live-bacteria-vaccine delivery systems. However, bacterial display systems are also used in many other ways, e.g. as bioadsorbents for the removal of harmful chemicals and heavy metals, and as whole cell biocatalysts with immobilized enzymes (Lee et al., 2003; Benhar, 2001; Samuelson et al., 2002; Ståhl and Nygren, 1997). Both gram-negative and gram-positive bacteria have been used for such display, as well as yeasts (Boder and Wittrup, 1997) and mammalian cells (Ernst et al., 1998). In addition to using cell surface anchoring proteins, extracellular proteinaceous appendages like pili (Steidler et al., 1993) and flagella (Lu et al., 1995) have also been used successfully. The advantages of cell display include the following. There is no need for reinfection in order to amplify the selected recombinant cell, only one host for propagation. Screening and selection can be performed as for phage display, but Fluorescence Activated Cell Sorting (FACS) can also be used, allowing cells displaying a fluorescent signal to be selected at high speed. For this, the cell population displaying a protein library is first incubated with fluorescently labeled target molecules. The target and ligand are allowed to interact in solution, avoiding any possible complications due to avidity effects. Cells displaying a protein variant interacting with the labeled target will fluoresce and are sorted out from cells showing little or no fluorescence (Francisco et al., 1993). The fluorescent signal window used for sorting can be adjusted to increase or lower the stringency of the system, an advantage that can be used to discriminate between binding variants with differing affinities.
Cell surface display has been primarily used for the display of peptides (Christmann et al., 1999), antibodies (Daugherty et al., 1998) and T-cell receptor libraries (Kieke et al., 1999) but also (to a lesser extent) to display enzyme libraries (Olsen et al., 2000).

One of the most impressive examples is the work by Boder and coworkers, in which they evolved scFv’s that bind to fluorescein with femtomolar affinity using a yeast display system (Boder et al., 2000). The fluorescein-binding antibody fragment was mutagenized using error-prone PCR in combination with DNA shuffling, and was displayed on the yeast surface via α-agglutinin. Cells were incubated with fluorescein-biotin to saturate surface binding. This was followed by competition with 5-aminofluorescein, labeling with streptavidin-R-phycoerythrin, and finally clones with the highest levels of bound fluorescein-biotin were isolated by FACS. An antibody fragment with a $K_D$ of 48 fM (even lower than the $K_D$ for the interaction between biotin and streptavidin) was isolated after three additional rounds of mutagenesis and screening. The same group has also reported the construction of naïve libraries for yeast display, from which large numbers of high affinity antibodies have been selected (Feldhaus et al., 2003).

### 4.2.2.2 Plasmid display

In plasmid display systems the non-covalent interaction between a DNA-binding protein and the plasmid DNA is used to link the protein to its DNA sequence. The first DNA binding protein to be used in this way was the Lac repressor (Cull et al., 1992). The Lac repressor binds as a tetramer very tightly to the lac operator sequence, lacO. In the study by Cull and co-workers, a library of $10^8$ dodecapeptides was constructed by ligation of degenerate oligonucleotides (NN(G/T))$^{10}$ at the 5’-end of the lacI gene, resulting in plasmid library members that each encoded a peptide fused to the C-terminus of the repressor expressed from a plasmid containing two lacO sequences. The presence of two lacO binding sites for the repressor resulted in that fusion proteins became strongly bound to the plasmid that encoded them. Fusion-proteins attached to their respective plasmids were released (after amplification in *E. coli*) by cell lysis and used in bio-panning experiments against an immobilized antibody. After only two rounds of selection, the majority of the selected population encoded fusion peptides that bound specifically to the antibody. To date, this method has mainly been used to select for peptides (England et al., 2000; Wang et al., 1998; Martin et al., 1996). However, in principle it could be used for the same range of applications as for phage display, although restricted to proteins capable of folding in the cytoplasm.
In the initial method the peptide was displayed as a tetramer, giving rise to avidity effects due to the possibility that two or more peptides could bind different targets simultaneously. If more strongly binding proteins are required, the avidity effect can be problematic, due to the increase in apparent affinity. An alternative system was therefore developed which lowered the number of displayed peptides per plasmid (Gates et al., 1996). In that work, a DNA-binding protein dimer was designed comprising the ≈60 amino acid residue N-terminal domain of Lac I, which mediates specific DNA-binding (the so called headpiece domain). To isolate headpiece dimers connected by a suitable linker that would allow for strong enough DNA-binding for bio-panning, the cited authors made a linker library displaying a specific antibody peptide epitope. Members of the library with a linker that was stable enough to allow for capture of the peptide on plasmid through the interaction between the antibody and the peptide antigen were isolated. Subsequently, the headpiece dimer system was used to select for peptides with affinities higher than the initial system (Gates et al., 1996, Cwirla et al., 1997). Recently, an additional method has been published utilizing the DNA-binding protein NF-kB (Speight et al., 2001).

4.2.2.3 In vivo-based systems

There are various types of in vivo-based screening or selection systems, but a shared feature is that library members are expressed inside cells containing the corresponding gene rather than being "displayed". Depending on the nature of the genes and hosts, positive clones can be identified via (i) the successful enzymatic conversion of an often chromogenic substrate, (ii) the survival of the microbial clone (enzyme libraries), (iii) the expression of a suitable reporter gene through a targeted biomolecular interaction resulting in a change of color or the cell’s survival in a specific environment, or (iv) the reconstitution of a fragmented reporter protein or enzyme via such an interaction.

The classical strategy for in vivo selection used in enzyme library work is to establish a system in which the presence of a catalytic activity is linked to a growth advantage for the host bacterium or microorganism. The biological selection is usually based on complementation of auxotrophy or resistance to cytotoxic agents such as an antibiotic. For example, relying on the intrinsic mutation frequency of a microorganism, a mutant ribitol dehydrogenase with two amino acid substitutions and improved xylitol dehydrogenase activity was obtained from a Klebsiella aerogenes strain that had been grown under selective pressure on xylitol as the sole carbon source (Homsi-Brandebugo et al., 1999). Using similar screening principles, randomization of antibiotic resistance genes have resulted in the
identification of kanamycin phosphotransferase variants with enhanced thermal stability (Hoseki et al., 1999) and used to change the substrate specificity of β-lactamase (Zaccolo and Gherardi, 1999).

So-called n-hybrid systems and protein-fragment complementation assays exploit intracellular recognition events. The yeast two-hybrid system described almost fifteen years ago was the first such system (Fields and Song, 1989) and has been used extensively to identify protein-protein interactions using samples derived from many different organisms. The basic concept emerged from the analysis of transcription factors that increase the rate of transcription by binding to an upstream activating DNA sequence. It was demonstrated that the DNA-binding and activating functions were located in physically separable domains. The activity could be restored through non-covalent interaction between two fusion partners binding to each other and thus bringing the DNA binding and activator domains into close proximity with each other and initiating the transcription. In this system, a gene library can be expressed as a fusion to one half of the reporter-gene activator domain. In the same cell, the target protein (bait) for which interacting partners are sought is expressed as a fusion to the other half of the activator domain. Thus, in a cell where a library member is capable of interaction with the target protein, the activator becomes functional and can activate a downstream reporter gene, resulting in a change in the cell’s phenotype. This technique has been used, among other things, to investigate the protein-protein interactions between many of the full-length open reading frames predicted from the yeast genome-sequencing project (Uetz et al., 2000). A similar approach has also been taken in large-scale mapping projects focused on other organisms, such as Caenorhabditis elegans (Walhout et al., 2000) and Helicobacter pylori (Rain et al., 2001).

An alternative strategy is to use fusion proteins that interact by means of a bridging molecule rather than directly. When such a “three-hybrid” assembles, it is similarly able to activate transcription of a reporter gene (Belshaw et al., 1996). One-hybrid systems look for protein domains that replace rather than connect one of the pieces that are reconstituted by a two-hybrid system, for example to identify proteins that recognize a specific DNA sequence (Chong and Mandel, 1997). The development of similar systems in bacteria took a number of years because few of the well-studied activators in bacteria had the two-domain structure found in yeast. However, a number of different 1- and 2-hybrid systems have now been developed in bacteria (Hu, 2001).
Protein-fragment Complementation Assays (Remy et al., 2002; Michnick, 2001) utilizes, in principle, any reporter protein that can be reconstituted at the protein level from two genetically split and co-expressed fragments. These fragments are further fused to two test proteins (e.g. library members and target protein) that are thought to bind to each other. Folding and reconstitution of the reporter protein from its fragments is promoted by the binding of the test proteins to each other and is detected by the reconstituted reporter protein activity. To date, systems have been developed for both bacterial and mammalian cells (Remy and Michnick, 1999; Remy et al., 1999). The first enzyme reporter protein successfully split into two parts and used for this kind of system in bacteria was mouse dihydrofolate reductase (mDHFR) (Pelletier et al., 1998). When the two subfragments of mDHFR come into close contact, through the interaction of their fusion partners, the enzyme is reconstitute, thus restoring activity, linked in this case to the survival of the cell in the presence of the antibacterial drug trimethoprim, which selectively inhibits the *E. coli* DHFR enzyme but not the mouse homologue. This system has been used in a number of applications, for example, to select a library containing potentially coiled-coil forming sequences against another library constructed in a similar fashion (Pelletier et al., 1999). Other enzymes used in PCA include β-lactamase (Galarneau et al., 2002; Wehrman et al., 2002), a periplasmic system that is good for disulfide formation-dependent targets, library members and GFP (Ghosh et al., 2000; Hu and Kerppola, 2003), a useful enzyme because no substrate is required to observe fluorescence in living cells.

### 4.2.3 Cell-free approaches

The size of the libraries that can be accommodated by the cell-dependent systems listed above is limited by the transformation capability of the cells used. This limitation can be overcome by transcribing and translating library member gene expression cassettes in the test tube, entirely *in vitro*. In order to use such systems for protein library work, a number of ingenious systems for establishing a link between phenotype and genotype, that is not dependent on a cell or virus, have been developed. Examples include: ribosome display, mRNA peptide fusion or PROfusion, water-in-oil emulsions and micro-bead display.

Using these methods, the library size is only limited by the amount of DNA used and the scale of the *in vitro* reactions. Typically, 1 mg of DNA can give a library size of $10^{15}$ members for small peptides and $10^{12}$ members for large proteins. Library sizes of more than $10^{13}$ molecules have been reported (Takahashi et al., 2003).
4.2.3.1 Ribosome display

Ribosome display was introduced as a means to express and handle polypeptide libraries in 1994 by Mattheakis and co workers (Mattheakis et al., 1994), and further optimized by Hanes and Plückthun. Mattheakis’s group used an E. coli S30 transcription/translation system to transcribe and translate in vitro a DNA-library containing short peptides. As the ribosome travels down the mRNA the peptide becomes exposed and folds, but because the mRNA does not contain a stop codon the peptide will not be released from the ribosome. Thus, under certain environmental conditions the mRNA, peptide and ribosome form a stable, non-covalent complex. The complex can then be exposed to the target, usually immobilized on a solid support (bio-panning, Fig. 5). Hanes and Pluckthün further developed the method when they showed that in addition to short peptides, folded proteins could also be displayed using ribosomal display. They enriched a scFv fragment which specifically bound hemagglutinin from a background of $10^8$ mRNAs coding for scFv fragments that did not bind to the target ligand (Hanes and Pluckthun, 1997). In addition, they also evolved picomolar affinity anti-insulin scFv’s from a synthetic antibody library (Hanes et al., 2000). Ribosome display has also been used to screen a cDNA-library for clones binding to a given target. To overcome the problem of the presence of a stop codon in the cDNA, random priming synthesis of the cDNA library was used. From the resulting library a number of clones were selected exhibiting affinity against the anti-apoptopic protein Bcl-XL (Hammond et al., 2001).

4.2.3.2 mRNA-peptide fusion

As mentioned earlier, the ternary-complex formed between the mRNA, protein and the ribosome is not a covalent interaction, and the instability of the complex can cause problems during the selection procedure. A solution to this problem was provided by a method dubbed mRNA-peptide fusion or “Profusion”, which creates a covalent bond between an encoding mRNA and the corresponding protein using puromycin (Roberts and Szostak, 1997, Nemoto et al., 1997). The puromycin is attached to the 3’ end of the mRNA using a puromycin-DNA linker conjugate, which was initially done using enzyme ligation. However, more recently a new method was described based on hybridizing the DNA-puromycin segment to the mRNA followed by photo-cross-linking (Kurz et al., 2000). The DNA-puromycin-mRNA hybrid is translated in vitro, and when the ribosome reaches the RNA/DNA-junction, the ribosome stalls and the puromycin enters the ribosomal A-site. The ribosome, in turn, catalyses the covalent attachment of the puromycin to the peptide chain and the stable complex can be used in different kinds of selection procedures including, for example, bio-panning. After washing,
mRNA from the bound complexes can be recovered by dissociating the complex and cDNA can be produced by reverse transcription and amplified using PCR. A new batch of mRNA can be prepared followed by additional rounds of selection and amplification. Since the genetic material used in both of these methods is RNA, which is more unstable than DNA, the work needs to be done under totally RNAs-free conditions.

mRNA-display has been successfully used to isolate protein sequences selected against peptides, small molecules and proteins (Takahashi et al., 2003). In addition, the covalent link between the RNA and the protein has been used to create self-assembling protein arrays (Weng et al., 2002). The chip was imprinted with DNA that was complementary to unique nucleic acid portions on three different protein-RNA fusions. In this way each fusion protein was directed to a specific spot on the chip surface utilizing Watson-Crick base pairing.

4.2.3.3 In vitro compartmentalization and micro bead display

In vitro compartmentalization or water-in-oil emulsions, developed by Tawfik and Griffiths in 1998, mimic cellular compartmentalization by a simpler in vitro system (Tawfik and Griffiths, 1998). They generated an emulsion with droplets having the same mean diameter as bacterial cells by adding an in vitro transcription/translation reaction mixture to a stirred suspension of mineral oil containing surfactants. The conditions were controlled so that each compartment on average contained a single gene. In the pioneering experiment a mixture of two genes, the *M. HaeIII* gene coding for a methyltransferase, which methylates *HaeIII* cleavage sites, and the folA gene coding for DHFR were transcribed and translated in the man-made compartments. The emulsion was then broken, and the released DNA was subjected to *HaeIII* cleavage followed by PCR amplification. Genes coding for the methyltransferase are amplified in this way because the methylated DNA strand is resistant to *HaeIII* cleavage. After a single round of selection, starting with a 1:1000 ratio of *M. HaeIII* to the *folA* gene the ratio between the two genes was 1:1. One ml of water-in-oil emulsion contains approximately $10^{10}$ droplets, which means that the same amount of protein variants could be tested on this scale.

In another example of the utility of micro droplets, a *Taq* DNA polymerase exhibiting increased thermal stability and heparin resistance was isolated (Ghaddessy et al., 2001). To do this, individual bacteria were incorporated into the micro-droplets together with PCR reaction components, and thermostable enzyme was selected by incubating the emulsions at 99°C for up to 15 min before beginning the PCR cycling. In a further development of the emulsion
format, Doi and Yanagawa reported a hybrid system combining the formation of a physical genotype/phenotype linkage with a compartmentalization method called STABLE (STA-Biotin Linkage in Emulsion) (Doi and Yanagawa, 1999). A DNA library of decapeptides fused to the 3’-end of a streptavidin gene was constructed, and the DNA was biotinylated and dispersed in the transcription/translation emulsion. Upon in vitro expression of the library members, cognate gene and protein pairs become physically linked inside each compartment via the high affinity streptavidin-biotin interaction. Selections or screening can then be performed using either the compartments themselves, as in the Tawfik and Griffiths system, or the physically linked protein-DNA complex can be released from the droplets and used in the same way as in, for example, ribosome display, with the advantage that the nucleic acid used is DNA instead of RNA.

Micro-bead display creates a linkage between the genotype and phenotype via micro beads. Griffiths and coworkers first demonstrated this method in 2002 in a format combined with their in vitro compartmentalization method (Sepp et al., 2002). Shortly thereafter, a similar, but solely solution-based micro-bead display system was demonstrated that involved no compartmentalization (Nord et al., 2003). A repertoire of biotinylated genes encoding protein variants, each with a common N- or C-terminal epitope tag are linked to streptavidin-coated beads (on average less or equal to one gene per bead) carrying antibodies that bind the epitope tag. In Griffiths’ system the beads are compartmentalized in a water-in-oil emulsion to give on average one bead per compartment and are transcribed and translated in vitro inside the compartments, avoiding cross-contamination. The translated protein variant fused to the epitope tag can then bind to the bead-immobilized antibody, specific for the epitope tag, resulting in a bead linking the gene and the gene-encoding protein variant. The emulsion is then broken and the beads are released into solution. Nord and co-workers performed the coupled in vitro transcription and translation directly on a diluted bead suspension instead of inside a specific compartment (avoiding these additional steps). In the diluted bead suspension, the translated gene-product is expected to preferentially re-bind to the bead carrying the gene encoding it, due to the (relatively) large distance between the beads.

When the link between the gene and the encoding protein variant is established, protein variants exhibiting binding to a specific target can be fluorescently labeled and screened by flow cytometry (FACS). Recently, Griffiths and Tawfik applied their system to select improved catalysts. A library of the enzyme phosphotriesterase, which is able to degrade organophosphate pesticides and nerve agents, was created by randomizing five different
positions using NN(G/C) degenerate codons. Microbeads containing a single gene library member and multiple copies of the encoded protein variant were re-emulsified together with a soluble substrate linked to caged biotin. The substrate was converted to product only in the compartments containing beads displaying an active enzyme variant. The emulsion was then irradiated to uncage the biotin, allowing product and unreacted substrate to bind to the bead. Beads containing the product were detected using product-specific antibodies followed by flow cytometry. After several rounds of selection, an enzyme with a 63-fold higher $k_{\text{cat}}$ than the already very efficient wild-type enzyme was isolated (Griffiths and Tawfik, 2003).

Beads are very robust, compared to certain cells and *in vitro* compartments, and can withstand high speed sorting (500,000 beads/s) using flow cytometry. To overcome this drawback of emulsion systems, a double emulsion system was recently presented that is more suited for flow cytometry sorting than the previous single-layer emulsion system (Bernath et al., 2004).

5. Protein library techniques in practice

Libraries of many different proteins have been constructed and subjected to screening or selection by a variety of methods (Bernath et al., 2004; Bernath et al., 2004). In the following section selected aspects and examples of protein library applications concerning four important groups of proteins will be discussed, focusing on the use of the most widely applied selection technique: phage display.

5.1 Enzymes

The attractive properties of enzymes, i.e. their ability to perform intricate regioselective and/or enantioselective chemical transformations and to accelerate reaction rates by enormous factors, in some cases as much as $10^{17}$-fold, all under mild conditions (Bernath et al., 2004), make them attractive for use as catalysts in diverse applications (Kirk et al., 2002). However, for use in biotechnological and industrial applications, which can involve far from physiological chemical conditions, many natural enzymes have severe limitations, such as low activity on non-natural substrates, low stability or low tolerance to required operating conditions, poor activity in non-aqueous media and high production costs if expensive co-factors are required (Schoemaker et al., 2003; Arnold, 2001). This has inspired researchers in the field of enzyme engineering to utilize protein engineering principles to develop novel catalysts for a number of valuable purposes (Turner, 2003; Dalby, 2003).
As discussed earlier, the use of protein library techniques is an attractive approach for such engineering efforts. However, functional selection to identify improved variants of enzymes based on their catalytic characteristics can be subject to unique complications. First, the catalytic event is a complex process, involving several steps and it is not as straightforward as binding to a specific target molecule. Second, the product formed in a successful catalytic turn-over, usually rapidly diffuses away making it impossible to use product specific capturing reagents. Therefore, other techniques are used in parallel to display technologies. One such technique is in vivo selection, as described in section 4.2.2.3, which has been used successfully in many cases, but has two major disadvantages. One is the restriction in the types of catalysts that can be selected in this way. The other is that cells often find unexpected ways to solve problems posed by experimenters.

Screening is an alternative option, i.e. individually assaying physically separated clones for the targeted catalytic function (Olsen et al., 2000). This can be done, for example, on agar plates or in microtiter plates. The most widely used screening method is to exploit chromogenic or fluorogenic markers, e.g. colored products such as p-nitrophenol (Giver et al., 1998; Zhang et al., 1997). For instance, Arnold and co-workers (Giver et al., 1998) used the chromogenic substrate p-nitrophenyl acetate to establish a microtiter plate based screening assay to improve the thermostability of the B. subtilis pNBE esterase without loss of activity at lower temperatures. Positives exhibiting high activity before and after incubation at high temperature were selected as inputs for another round of directed evolution. After performing six rounds of random mutagenesis, recombination and screening an enzyme variant displaying a 14 °C increase in thermostability, without loss of activity at lower temperatures, was isolated. Another useful chromogenic substrate is ninhydrin, which reacts selectively with amines to give a purple product and has been used to screen amidases for the hydrolysis of amides to amines (Taylor et al., 1999). A pH indicator can also be used if the product, or a side product, lowers or raises the pH. For example, ester hydrolysis releases a carboxylic acid product, which results in a drop in pH (Moris-Varas et al., 1999; Bornscheuer et al., 1999; I). A more general solution involves coupling the reaction of interest to a secondary reaction sequence that converts the primary product to a fluorescent product (Joo et al., 1999). FRET has also been used where an internal bond or an intramolecular isomerization modifies the distance and thus the energy transfer interaction between a pair of fluorescent tags attached to the substrate (Jenne et al., 1999; Matayoshi et al., 1990).
Fig. 6. Examples of methods with the aim of linking catalysis to affinity so that a bio-panning procedure can be utilized. A and B are examples of indirect selection methods for catalysis and C and D of direct methods for selecting against catalysis. A. An immobilized transition state analogue (TSA) is used to capture active enzymes. Elution can be done either by competitive elution or protease cleavage. B. Active phage enzymes become covalently linked to a biotinylated mechanism-based inhibitor in the presence of an excess of substrate. Trapped phage are extracted from the mixture by binding to streptavidin-coated beads and eluted either by protease or reductive cleavage. C. Enzyme-calmodulin fusions are displayed on phage together with the substrate, containing a calmodulin-binding peptide attaching the substrate to the phage. After catalysis, active phage are captured onto a product-specific solid support and eluted with EDTA. D. The enzyme is initially inactivated by chelating an essential metal-ion and captured with immobilized substrate. Addition of the metallic cofactor results in elution of active enzymes.

In spite of potential complications concerning functional selection based on catalysis, several groups have investigated the use of phage display technology for enzyme engineering projects. In order to use the general bio-panning principle outlined in Fig. 5 for catalysis-based selection it is necessary to find a technique that links chemical transformation with acquired binding ability. Some selection protocols have been developed based on transition
state analogues (TSA), which are molecules that mimic the geometry and charge distribution of transition states. This strategy aims to take advantage of the working principle of enzymes to lower the activation energy by stabilizing the transition state (Hansson et al., 1999; Hansson et al., 1997; Widersten and Mannervik, 1995) (Fig. 6, Table 1, p.37). An alternative strategy to this indirect selection principle is to use suicide substrates that mimic the substrate but become permanently linked to the enzyme upon initial catalysis. Such compounds are also called mechanism-based inhibitors (Ator and Montallano, 1990; I). For these selections, libraries of mutants are incubated under kinetic control with a limiting concentration of biotinylated inhibitor. The most active enzymes react faster, are covalently labeled and can be extracted by adsorption to streptavidin-coated beads. Bound phages are eluted either by reductive cleavage or by proteolytic cleavage of a linker connecting the displayed enzyme and the phage coat protein (Soumillion et al., 1994).

The indirect selection approaches described above have worked very well for finding novel catalysts including, for example, so-called catalytic antibodies, a fascinating class of "enzymes" which were first described by two different groups in 1986. The Lerner group (Tramontano et al., 1986) generated catalytic antibodies by immunization with a phosphonate TSA designed for ester hydrolysis, and Shultz’s group (Pollack et al., 1986) demonstrated that an antibody binding to a phosphate diester TSA was a catalyst. Since these extraordinary results, many more catalytic antibodies have been generated using immunization, screening and selection protocols (including phage display technology) either separately or in combination (Griffiths and Tawfik, 2000). The substrate specificity of enzymes has also been modified successfully using the indirect methods, but they have not been very successful for finding variants with as high catalytic turnover as natural variants (Table 1 p.37). For example, a glutathione S-transferase (GST)-variant with altered substrate specificity towards charged substrates was selected from a library randomized at ten residues lining the substrate-binding site, but with a 1000-fold reduced specific activity (Widersten and Mannervik, 1995).

A possible explanation for this is that the majority of reactions proceed via more than one transition state or step, any one of which can become rate limiting. In an illustrative example involving engineering of the enzyme TEM-1 β-lactamase, inhibitor-based selection from a phage display library resulted in no active enzyme variants at all. The covalent attachment of the inhibitor originated in this example from a covalent intermediate, an acyl-enzyme and, consequently, slower deacylation gives more time for the inhibitor to become covalently attached. Mutants that are acylated quickly and deacylated slowly were selected.
(Vanwetswinkel et al., 2000). To overcome this problem, slowly deacylating variants were blocked with the natural substrate and then active phage enzymes were labeled and enriched from the library. This example illustrates how complex enzymatic mechanisms can be, and how important it is to have detailed knowledge about the reaction mechanism in order to design a selection scheme that truly amplifies the desired trait. The efficiency of indirect selections can, however, be improved significantly if, after binding-selection the resulting variants are further screened for the desired catalytic activity.

Other affinity selection methods, based directly on product formation, have also been developed. A straightforward way to avoid the problem of the product diffusing away is to attach the substrate as well as the library member onto the phage particles. Several different means of doing this have been presented, including substrate immobilization by non-specific alkylation of the major coat protein (Jestin et al., 1999), and a strategy in which the substrate and library member are combined in a single pIII fusion format (Demartis et al., 1999). In the latter case the protein calmodulin is inserted between pIII and the catalytic protein. The substrate is fused to a calmodulin-binding peptide, attaching the substrate by a calcium-dependent non-covalent interaction to the phage fusion protein. Active enzyme-phage are adsorbed onto a product-binding support and bound phage are eluted with EDTA (Fig. 6). This system has been used to enrich proteases, GST and a biotin-ligase (Demartis et al., 1999; Heinis et al., 2001). The display of a rat trypsin variant (His57Ala), a relatively inefficient endopeptidase that cleaves a specific di-peptide, resulted in 50- to 2000- fold enrichment by the release of a new N-terminus recognized by a specific antibody (Heinis et al., 2001). Selections on a library randomized at four positions using NN(G/T) degeneracy resulted in an enrichment of active variants, but none of the selected clones performed better than the parent. Other methods are based on the possibility of inactivating metallo-enzymes by metal-ion depletion. Pedersen and colleagues used the enzyme staphylococcal nuclease, which was inactivated by Ca\(^{2+}\) depletion. The substrate, a biotin-labeled double stranded DNA-peptide fusion was attached to a phage pIII peptide fusion by a disulphide linkage that allowed the immobilization of phages on a streptavidin-coated matrix. After addition of Ca\(^{2+}\), thereby activating the enzyme, phages displaying active enzyme were eluted (Pedersen et al., 1998). In a similar approach, the metallo-\(\beta\)-lactamase from Bacillus cereus was displayed on phage and after Zn-depletion the enzyme was allowed to interact with immobilized substrate (Ponsard et al., 2001). This strategy is based on the assumption that the inactivated enzyme will still be capable of binding to the substrate. After the addition of Zn-salt, the product will
be released by phage enzyme fusions displaying active enzyme and eluted from the matrix. The method allowed for a 50-fold enrichment of active phage from a mixture of error-prone PCR-generated mutants (Fig. 6).

Table 1. Examples of selections of catalytic activity using phage display techniques.

A. Indirect methods. B. Direct methods, based on product formation

<table>
<thead>
<tr>
<th>Enzyme name</th>
<th>Selection principle</th>
<th>Result</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>GST</td>
<td>Substrate analogue</td>
<td>Altered specificity</td>
<td>Widersten and Mannervik, 1995</td>
</tr>
<tr>
<td>GST</td>
<td>TSA</td>
<td>Enrichment</td>
<td>Hansson et al., 1997</td>
</tr>
<tr>
<td>GST</td>
<td>TSA</td>
<td>2-3 fold</td>
<td>Hansson et al., 1999</td>
</tr>
<tr>
<td>Staph nuclease</td>
<td>Substrate analogue</td>
<td>Not much</td>
<td>Light and Lerner, 1995</td>
</tr>
<tr>
<td>Alkaline phosphatase</td>
<td>Inhibitor</td>
<td>Enrichment</td>
<td>McCafferty et al., 1991</td>
</tr>
<tr>
<td>TEM-β-lactamase</td>
<td>MBI</td>
<td>Enrichment</td>
<td>Soumillion et al., 1994, Vanwertswhinkel and al., 1996</td>
</tr>
<tr>
<td>TEM-β-lactamase</td>
<td>MBI</td>
<td>No active variants Pre-incubation gave enrichment</td>
<td>Vanwertswhinkel et al., 2000</td>
</tr>
<tr>
<td>Subtilisin</td>
<td>MBI</td>
<td>Improved activity on unnatural S</td>
<td>Legendre et al., 2000</td>
</tr>
<tr>
<td>Lipase</td>
<td>MBI</td>
<td>Enrichment</td>
<td>paper I</td>
</tr>
</tbody>
</table>

B

<table>
<thead>
<tr>
<th>Enzyme name</th>
<th>Selection principle</th>
<th>Result</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endopeptidase GST, biotin-ligase</td>
<td>S fused to a calmodulin-binding peptide EDTA elution</td>
<td>Enrichment</td>
<td>Demartis et al., 1999, Heinis et al., 2001</td>
</tr>
<tr>
<td>Staphylococcal nuclease</td>
<td>Inactivated metallo-enzyme S: DNA-peptide-biotin, disulphide linked to phage. Selective release</td>
<td>Enrichment</td>
<td>Pedersen et al., 1998</td>
</tr>
<tr>
<td>Metallo-β-lactamase</td>
<td>Inactivated metallo-enzyme Selective release</td>
<td>Enrichment</td>
<td>Ponsard et al., 2001</td>
</tr>
<tr>
<td>DNA-polymerase</td>
<td>Primer cross-linked to phage coat. S: biotin-dUTP</td>
<td>Enrichment</td>
<td>Jestin et al., 1999</td>
</tr>
<tr>
<td>DNA-polymerase</td>
<td>Leucine zipper-linked oligo to pII S: biotin-Om6dNTP</td>
<td>Activity on modified S equal to wt with natural S</td>
<td>Fa et al., 2004.</td>
</tr>
</tbody>
</table>

Abbreviations: S; substrate, TSA; transition state analogue, MBI; mechanism-based inhibitor
5.2 Peptides

The term random peptide libraries refers to collections of relatively short variegated polypeptide sequences, typically six to 20 residues in length. Such libraries have frequently been displayed on phage for selection of variants with differing characteristics (Benhar, 2001). The design of such libraries has typically involved either linear peptides (e.g. $X_n$-type) or so-called cyclic peptides, referring to the introduction of structural constraints through a disulfide bond via invariant cysteine residues positioned so as to flank the variegated portion of the peptide (e.g. $C(X)_nC$-type). Such reductions of flexibility have been shown to result in peptides with higher binding affinities, possibly due to a decreased loss of entropy upon binding (Sato et al., 2002). Random peptide libraries have been displayed on phage at different valencies using both pIII and pVIII as fusion partners in phage (3, 33, 8 or 88 systems) or phagemid systems (3+3 or 8+8 systems); the small size of peptide library members allows all copies of pVIII to be decorated by a foreign peptide without being detrimental to phage particle assembly (Widersten and Mannervik, 1995).

Random peptide library displays have been widely used to select peptides by different selection protocols for use in a variety of applications. An early-described application was the use of random peptide libraries for epitope mapping of antibodies, involving biopanning of the phage library against a monoclonal antibody target for the identification of sequences that mimicked a discontinuous epitope or had some identity to a linear epitope of the mAb (Scott and Smith, 1990). Such experiments have also been performed on mAbs with carbohydrate specificities, aiming to find peptide sequences capable of being used as surrogate immunogens to elicit protective immune responses against the original carbohydrate structure (Phalipon et al., 1997). For use in bio-separation applications, Huang and co-workers reported the selection of a hexapeptide sequence from a peptide type 3 library with affinity for von Willebrand factor (vWF). After chemical synthesis of the identified sequence directly on a chromatographic resin, the peptide was used as a ligand for affinity chromatography of vWF from different samples (Huang et al., 1996). Random peptide libraries have also been used to identify sequences capable of serving as substrates for enzymes. As briefly described earlier, sequences susceptible to the proteases factor $X_a$ and subtilisin have been successfully identified through protease-incubation of peptide-displaying phage particles that had previously been anchored to a solid support through the interaction between human growth hormone (hGH), expressed N-terminally to the peptide sequences, and a hGH-binding protein (Matthews and Wells, 1993). Using the erythropoietin receptor (EPOr) as biopanning target,
Malin Eklund

Wrighton and co-workers selected peptides with agonistic properties (Wrighton et al., 1996). Subsequent structural analysis of the peptide-receptor complex showed that the peptide was bound in a non-covalent homodimeric format (Livnah et al., 1996), which led to further development of a covalently linked synthetic dimer peptide for subsequent studies of the agonistic effect and, if possible, the design of low molecular weight analogues (Wrighton et al., 1997). Some more extraordinary uses of peptide libraries have also been reported, including the following. Peptides capable of selective tissue recognition have been selected \textit{in vivo} by injecting a phage library into a mouse’s tail then dissecting the animal to rescue phage attached to particular organs (Pasqualini, 1999). In addition, semiconductor-binding peptides have been selected for possible use in nanotechnology (Whaley et al., 2000).

5.3 Antibodies

Because of their extremely specific and sensitive molecular recognition properties antibodies have been used in many areas of biological research, medical diagnostics and therapy, biotechnological and even chemical applications (e.g. as catalytic antibodies in chemical reactions).

Natural antibodies are produced by immunizing animals with an antigen, thereby inducing the animal’s immune response to produce polyclonal antibodies, directed against a number of epitopes. In 1975 Kohler and Milstein introduced the hybridoma approach, which made it possible to produce monoclonal antibodies (i.e. antibodies directed to a single epitope) from immortalized lymphocytes. This work also earned them the Nobel Prize in 1984 (Kohler and Milstein, 1975).

In recent years, competitive alternatives to these methods have emerged, based on genetic engineering principles involving protein library techniques that avoid the use of animals completely. This field was boosted by the demonstration that sets of immunoglobulin-encoding transcripts could be amplified in pools by PCR using primer pairs designed for annealing to conserved regions flanking variable parts (Orlandi et al., 1989). Utilizing these approaches, the immunoglobulin gene complements from immunized or non-immunized (naïve) donors could be harvested for use as genetic input material for \textit{in vitro} library constructions and, most often, phage display-based selections. Although the possibility of producing whole antibodies in bacteria has recently been demonstrated (Simmons et al., 2002), only smaller fragments, which can be expressed in \textit{E. coli}, containing the variable regions of antibodies have been used so far during library work (Better et al., 1988; Skerra
and Pluckthun, 1988). Phage display of Fab fragments requires the use of separate expression cassettes for the VH-CH1 and VL-CL fragments for subsequent assembly into disulfide-linked Fab fragments in the bacterial periplasm, whereas scFv fragments are expressed as single polypeptide sequences (Fig. 7). Two years later a library of scFv fragments was displayed on the surface of filamentous phage particles fused to one of the phage coat proteins and antigen-specific phages were subsequently enriched by multiple rounds of affinity selection (McCafferty et al., 1990). Soon thereafter Fab fragments were also displayed on phage and used for selection of specific antibody fragments (Hoogenboom et al., 1991; Garrard et al., 1991; Chang et al., 1991). Subsequently, other antibody derivatives have been displayed, such as diabodies (McGuinness et al., 1996), heavy-chain only antibodies from camels (Arbabi Ghahroudi et al., 1997) and murine or human heavy-chain only domains (Davies and Riechmann, 1996; Reiter et al., 1999).

**Fig. 7:** The basic antibody structure is a Y-shaped molecule composed of two identical light chains (L) and two identical heavy chains (H) (159 kDa). Both the light and heavy chains consist of a variable (V) and a constant part (C). The four chains are held together by disulfide bonds, which are located in a flexible region on the heavy chain, the hinge region, and between light and heavy chains. Variable regions of both heavy and light chains combine to form two identical antigen-binding sites containing six hyper-variable loops, referred to as CDR’s (complementarity determining regions). The heavy chain constant regions (Fc) define five classes of antibodies (IgA, IgD, IgE, IgG and IgM) each of which has its own distinct structural features, for example IgM and IgA form multimeric assemblies. The Fc part is responsible for recruiting the effector functions of the immune system. Also shown are different antigen binding subfragments; (Fab’);-fragment, Fab-fragment (57 kDa) and scFv-fragment (27 kDa), produced by protease cleavage or by recombinant methods.
In the early 1990s the source of diversity (ab gene fragments) were rescued V-genes from immunized B-cells, which would be expected to contain a high frequency of antigen-specific clones. Later, V-genes were isolated from non-immunized cells for the construction of naïve libraries (Marks et al., 1991), and later still different synthetic or semi-synthetic libraries were made by grafting rescued CDR regions into a common framework (Söderlind et al., 2000) or through the use of diversity that was not harvested from any immune source but provided artificially, to various extents, by oligonucleotide cloning (Hoogenboom et al., 1991; Barbas et al., 1992, Knappik et al., 2000; see also section 4.1.2.1). To date, numerous antibody libraries have been constructed for phage display selection (Hust and Dubel, 2004), yielding antibodies for applications ranging from basic research to drug development (Hudson and Souriau, 2003). Antibodies with affinity ranges of 1-200 nM have been selected (Hoogenboom and Chames, 2000).

In fact, today up to 30% of all human antibodies in clinical trials have been isolated using phage display (Kretzschmar and von Ruden, 2002). One human antibody derived from phage display that has progressed furthest in clinical trials is the D2EF anti-TNFα IgG antibody, for treating rheumatoid arthritis (Kretzschmar and von Ruden, 2002).

In addition to phage display technology, other display techniques have also been used in order to isolate highly specific antibodies, including ribosome display (Hanes and Pluckthun, 1997) and cell-surface display (Lee et al., 2003; Wittrup, 2001; Georgiou et al., 1997).

5.4 Alternative binding proteins

Recently, alternatives to antibodies for a variety of applications have emerged, in the form of novel classes of affinity proteins (Skerra, 2003; Skerra, 2000; Nygren and Uhlén, 1997). A feature common to all such reagents is that they are based on a non-immunoglobulin protein used as scaffold and surface-located positions are targeted for mutagenesis, resulting in mutants of the scaffold protein from which variants with novel characteristics can be selected. Depending on the protein used as scaffold, such reagents can have different advantages compared to antibody-derived reagents, e.g. smaller size, a less complicated structure (preferably monomeric and devoid of structurally important disulfide bridges) and increased expression in prokaryotic systems, or even the possibility for production through chemical synthesis.
The principles by which the large number of scaffolds have been engineered can be divided into three main groups; (A) single peptide loops displayed on scaffold proteins, (B) surfaces of secondary structure elements, and (C) several non-contiguous loop regions (Fig. 8).

The first group of scaffolds includes a class used in the earliest attempts to modify pre-existing binding activities, using a single randomized loop region on a well-expressed protein framework. In these studies, so-called Kunitz-domains, which are stable, 60-residue amino acid serine protease inhibitors with three disulfide bonds, were used as scaffolds. In one early example, bovine pancreatic trypsin inhibitor (BPTI) was used as a scaffold to create a phage library in which four specificity-determining residues were randomized. A BPTI-variant in which the wild type specificity against trypsin was changed to specificity towards human neutrophil elastase was selected. (Roberts et al., 1992; Roberts et al., 1992). The human pancreatic secretory trypsin inhibitor (PSTI) has also been displayed on phage, and novel inhibitors of bovine \( \alpha \)-chymotrypsin have been selected (Röttgen and Collins, 1995). Other examples in this group are thioredoxin and staphylococcal nuclease, but in these cases \textit{E. coli} flagellum and budding yeast systems were used, respectively.

The second category of scaffolds includes proteins in which positions within secondary structure elements (\( \alpha \)-helices and \( \beta \)-sheets) are recruited for randomization. Such positions can be distributed over several secondary structure elements, and are thus discontinuously

**Fig. 8.** Different scaffolds and principles used for library construction. A. A single loop randomized in the Kunitz-domain, BPI. B. The Z-domain randomized along two separate helices. C. Multiple loops randomized in the GFP protein.
distributed over relatively long stretches in the primary sequence, but located close to one another in the three-dimensional structure. The manipulation of such regions, rather than loop regions, requires the scaffold used to have high structural tolerance, to allow a sufficiently large number of positions to be randomized. Scaffolds displayed on phage belonging to this class include zinc fingers, cellulose-binding modules (CBM) and the Z-domain from SPA (see next section).

The Cys2His2 zinc-finger motif is one of the most common eukaryotic DNA-binding motifs. This small, 26 amino acid protein has a structure which is stabilized by metal chelate bonds between a metal ion and the two pairs of cysteine and histidine side chains (Bianchi et al., 1995). In this work a pVIII phagemid library, created by randomization of five surface-exposed residues was used to isolate variants with specific binding activity towards a monoclonal antibody.

Cellulose binding domains are somewhat larger, independent folding units of approximately 30-40 amino acids. These domains are required for efficient binding and hydrolysis of polymeric carbohydrate substrates by fungal cellulases (Tomme et al., 1998). One well-characterized CBM derived from the C-terminus of cellobiohydrolase I of Trichoderma reesei is composed of a triple stranded anti-parallel β-sheet stabilized by a cystine-knot with two disulfide bridges (Le Nguyen et al., 1990). Smith and co workers introduced variations into seven residues involved in cellulose binding. From the resulting pIII phage library they selected binders against four different targets (cellulose, α-amylase, bovine alkaline phosphatase and E. coli β-glucuronidase) (Smith et al., 1998). In the case of alkaline phosphatase, binders in the micromolar range were selected. A different strategy was used by Lehtio and collegues, in that the tip of the wedge-shaped CBM molecule was randomized instead of the flat cellulose-binding surface. One selected variant against porcine amylase exhibited partial inhibitory activity towards the enzyme, indicating that binding occurs in the substrate-binding cleft (Lehtio et al., 2000).

The third category has certain similarities to the variable domains of antibodies, in that several loop regions are recruited for randomization, analogous to the CDRs of V_H or V_L domains. Three examples in this category are fibronectin, lipocalin and GFP.

The fibronectin type III (FN3) domain has a β-sandwich structure (94 residues, seven β-strands, three loops), which closely resembles the structure of immunoglobulin VH domains. Ten residues situated on two surface loops were subjected to randomization, and after
biopanning against ubiquitin, variants exhibiting affinities in the low micromolar range were isolated by Koide et al. (Koide et al., 1998). In a following study, ribosomal display was used, all three surface loops were randomized and high affinity binders against TNF-α were selected (Xu et al., 2002). Members of the lipocalin protein family, such as the 174-residue bilin-binding protein (BBP) from Pieris brassicae have also been used as scaffolds (Skerra, 2000). BBP has a β-barrel as central motif composed of eight antiparallel β-strands and a shallow binding pocket consisting of four loops. Sixteen residues distributed over all four loops were selected for randomization, followed by phagemid display selection. In the initial selection, fluorescein binders with nanomolar affinities were identified (Beste et al., 1999). In later work, dioxygenin binders were isolated and the affinity was further raised by selective random mutagenesis of the first loop, resulting in a ten-fold increase in affinity (Schlehuber et al., 2000).

Recently, very elegant work by Bradbury’s group was reported, in which a “superfolded” version of the green fluorescent protein (GFP) was used as scaffold, resulting in binders that combine selective binding with a built-in detection function i.e. fluorescence (Zeytun et al., 2003). Four diverse antibody-binding loops were inserted at one end of the can-like, β-strand fold and specific fluorobodies against a range of targets were selected by phage display.

### 5.5 Affibodies

Central to the work in this thesis is a class of binding proteins denoted affibodies, which are based on a small three-helix bundle domain derived from Staphylococcal protein A (SPA).

SPA belongs to a family of bacterial receptins, a class of proteins that are displayed on the surface of the bacteria and are capable of specific binding to one or several host proteins, i.e. immunoglobulins, albumins and fibronectin. The biological functions of these proteins are not fully understood (Kronvall and Jönsson, 1999), but one possibility is that the bacteria use the host-molecule interaction as a means of camouflaging themselves to become more host-like and thus “hide” from the host’s immune system (Achari et al., 1992; Sauer-Eriksson et al., 1995). SPA belongs to the group of bacterial receptins that binds to immunoglobulin molecules. Other known molecules in this group include Streptococcal protein G (SPG) and Peptostreptococcus magnus Protein L (PPL). All of these proteins contain multiple and highly homologous Ig-binding domains capable of distinct recognition of regions on a wide range of immunoglobulins and have therefore been widely studied and utilized in immunology and biotechnology (Fig. 9).
Fig. 9. A. Binding sites for staphylococcal protein A (SPA) and two other bacterial receptor proteins: peptostreptococcal protein L (PL) and streptococcal protein G (SPG) on human IgG. B. A schematic representation of SPA and the Z-domain derived from the B-domain (for more details see below).

SPA consists of a signal sequence followed by five homologous domains (E, D, A, B and C), all capable of Fc and Fab (VHIII)-specific binding (Moks et al., 1986, Jansson et al., 1998). The average amino acid homology between the four most conserved domains (D, A, B and C) is 84% (Jansson, 1996). In addition, a cell-anchoring domain denoted XM is located at the C-terminus (Uhlén et al., 1984).

The five IgG binding domains exhibit specific recognition (K_D 10^{-7} M) to the heavy chain portion of the antibody at the interface between the CH2 and CH3 domains of the Fc fragment (Frick et al., 1992) (Fig. 9). They also bind, somewhat more weakly, distinct regions on the Fab portions of particular Ig families or subclasses (Graille et al., 2000) (Fig. 9).

A small 58 aa SPA-derived IgG-binding affinity protein denoted Z (Nilsson et al., 1997) was developed through genetic engineering of the B-domain of SPA, primarily for use as an affinity gene fusion partner. The B domain lacks methionine, which makes it resistant to CNBr cleavage, and by mutating the hydroxylamine-sensitive Asn_{28}-Gly_{29} sequence to Asn_{28}-Ala_{29}, the Z-domain was also made resistant to hydroxylamine chemical cleavage. In addition, an alanine residue at position one has been changed into a valine (Nilsson et al., 1997). The Z-domain has been frequently used as an affinity fusion partner for the production of either Z- or ZZ-fusion proteins in a number of different host cells (Ståhl and Nygren, 1997). Utilizing the IgG-binding capacity of the Z-domain, it has proved possible to recover and immobilize such fusion proteins using IgG-containing matrices and surfaces (Ståhl and Nygren, 1997).
The Z-domain has been shown to exhibit several desirable properties for use in biotechnological applications and as a scaffold for the design of novel binding proteins. For instance, it is highly stable against proteolysis in a variety of host organisms (Moks et al., 1986; Ståhl and Nygren, 1997). In addition, the N- and C-termini of the three-helix bundle structure of an individual domain are solvent-exposed (Deisenhofer, 1981), favoring independent folding of fused proteins. Furthermore, the Z-domain does not contain any disulphide bonds and is highly soluble (Samuelsson et al., 1994).

These properties indicated that the Z-domain could be an ideal scaffold for generating novel binding proteins. To identify residues that could be randomized without disturbing the packing of the three-helix bundle, Nord and co-workers (Nord et al., 1995) analyzed the co-crystal structure between the B-domain of SPA and Fc of human IgG1 (Deisenhofer, 1981). Thirteen surface-exposed residues, all located on the first two helices and nine of which were shown to be involved in Fc-binding, were chosen for randomization (Q9, Q10, N11, F13, Y14, L17, N28, Q32 and K35 + H18, E24, E25 and R27). A phagemid vector, pKN1, specifically designed for monovalent display of Z-library variants fused to a truncated version of the M13 phage coat protein III, was employed for the construction of the Z-libraries (Nord et al., 1995). In the pKN1 system, each of the Z variants are fused to the gene encoding the albumin binding domain (ABD) (Nilsson et al., 1994) from streptococcal protein G, followed by the gene for protein III. The ABD domain is included to allow purification using human serum albumin (HSA)-affinity chromatography of the Z-variants (Nygren et al., 1988). The expression is under the control of an E. coli lac promoter and an Omp A signal peptide, facilitating periplasmic localization of the encoded gene product. To allow sole expression of Z-variants fused to the ABD domain, an amber stop codon (TAG) is included between the ABD-domain and the pIII phage coat protein.

Two different libraries were constructed based on the use of two alternative degenerate codons. In the construction of the first library, Zlib-1 (Nord et al., 1995), a NNK (K= G or T) randomization strategy was used, ensuring the coverage of all 20 amino acids, as well as a stop codon. The second library Zlib-2 (Nord et al., 1997) was constructed using a (C/AG/)NN degeneracy profile, leading to the exclusion of unwanted STOP codons, thus avoiding premature termination, and cysteine residues that could lead to unwanted dimerization of affibodies. However, codons for certain aromatic residues (Trp, Tyr, Phe) are also excluded in this library due to the nature of the genetic code as discussed previously. Both libraries included approximately $4 \times 10^7$ different variants respectively.
In the pioneering selection experiments, the Z-libraries were used for biopanning against three different target proteins: human insulin, human apolipoprotein A-1 and Thermus aquaticus DNA polymerase. These selections isolated binders, affibodies in the micromolar range (K_D 3 x 10^{-5} to 1 x 10^{-6} M) for their respective target molecule. It was later shown that second generation binders using previously selected variants as starting material for additional randomization and selection could yield affibodies with affinities in the nanomolar range (Gunneriusson et al., 1999; Nord et al., 2001). These findings indicate that library size is important, and that there was no inherent limitation associated with affibodies for finding variants capable of binding targets with high affinities. Affinity maturation was achieved using two different strategies. In the first report a Z-variant directed against Taq polymerase was subjected to additional randomization by helix shuffling. One of the randomized helices was kept constant and the other was re-randomized (Gunneriusson et al., 1999). In the second study, positions identified after alignment of the first-generation variants were randomized a second time and factor VIII binders with a 20-fold increase in affinity were selected (Nord et al., 2001).

In addition to the work presented in this thesis (see Present investigation), reported applications for affibodies, selected against different targets, include for bioseparation: affibodies of different multiplicities have been immobilized to chromatographic supports for selective recovery of target proteins from samples of different complexities, including human plasma (Nord et al., 2000; Rönnmark et al., 2002). For detection; affibodies have been used in different formats, including display of affibodies on the surface of bacteria (Guneriusson et al., 1999), as fused to a β-galactosidase reporter enzyme for immunohistochemistry (Rönnmark et al., 2003) and as blotting reagents fused to the Fc part of an antibody (Rönnmark et al., 2002). Moreover, affibodies have been used in combination with monoclonal antibodies in sandwich ELISA’s to reduce problems associated with background signals from cross-reactive antibodies present in human serum (Andersson et al., 2003). Further, the small size of the affibody scaffold allows for their production by chemical synthesis (Nord et al., 2001), which has the potential to facilitate site-specific incorporation of fluorophores for detection applications (Karlström and Nygren, 2001). The small size and absence of disulfides has also facilitated the incorporation of affibodies into adenovirus fibers, with the aim of being able to direct viruses to specific target cells for gene therapy applications (Henningsson et al., 2003). In a cell-based assay, CD28 binding affibodies have
also been shown to be capable of inhibiting T-cell activation via blocking of the co-stimulatory signal provided by the CD28-CD80 interaction (Sandström et al., 2003).

**Present Investigation**

In the work underlying this thesis, libraries were constructed using two different randomization approaches: saturation mutagenesis and a doped randomization scheme (I). Phage display was applied to select for enzymatic activity in one study (I) and novel binding proteins in another (II). Further, the interaction between one selected binding protein and its target was investigated for use in a number of different biotechnological applications, including purification and detection (III), as well as for assembly of higher order protein structures (V). In addition, this interaction was characterized in detail through the determination of the co-complex structure by x-ray crystallography (IV).

7. Catalytic selection of lipase variants from phage display libraries (I)

7.1 Background

Lipases (triacylglyceride ester hydrolases, EC 3.1.1.3) have evolved naturally into efficient catalysts with a key biological role: hydrolyzing ester bonds in triglycerides and related molecules. There is also a wide range of industrial applications where hydrolysis of these bonds is important, for example hydrolyzing milk fats to accelerate cheese ripening or to enhance the flavor of butter. They are used to generate fatty acids from natural oils in soap manufacture and added to some laundry detergents to dissolve and remove grease from garments (Balcao et al., 1996). The high stereospecificity of lipases can also be exploited to synthesize, in particular, enantiomerically pure compounds (Schmid et al., 2001).

Lipolase® is the trade name for a lipase isolated from the fungus *Thermomyces lanuginosa*. Novozymes of Denmark introduced Lipolase in 1988 as the first detergent lipase, and it was the first industrial enzyme to be produced by recombinant technology. The gene encoding the enzyme was cloned into the expression host *Aspergillus oryzae*, which made production of the enzyme cost effective for use as an additive in laundry detergents. Since then Lipolase has been extensively engineered to improve its washing performance, and two other Lipolase variants have reached the market so far; Lipolase Ultra® with a point mutation in the lipid
contact zone, D96L (Svendsen et al., 1997) and recently Lipoprime®, developed by directed evolution methods.

Structures of all lipases, so far determined, show a characteristic α/β-hydrolase fold, which is also found in some esterases, proteases and haloperoxidases (Ollis et al., 1992; Bornscheuer et al., 1999; Derewenda et al., 1994). In addition, another common feature shared by many lipases is the presence of a helical structure or loop covering the active site, often referred to as the “lid”. The catalytic triad of Lipolase (Ser 146, Asp 201 and His 258) is positioned on the central b-sheet system of the α/β-fold. The lid, a short helix (aa 86-aa 92) covering the active site is opened when it comes into contact with a water-lipid interphase, a process called interfacial activation (Fig. 10). The most widely favored hypothesis regarding this event suggests that a conformational change, induced on binding to the lipid-water interface, results in displacement of the lid by rotation around the hinge regions. This hypothesis is supported by x-ray structures of lipases in complexes with inhibitors (Cajal et al., 2000; Lawson et al., 1994; van Tilbeurgh et al., 1993). Opening the lid exposes the active site as well as a large hydrophobic patch, which is likely to be stabilized in contact with a non-polar interface. However, there is limited information about how the transition from the closed to the open conformation occurs at the lipid interface.

The main aim of this project was to develop combinatorial protein engineering methods to identify lipase variants with improved performance in the presence of a commercial detergent formulation and at high pH-values. As mentioned above, the activation of the lipase at a lipid-water interface that induces the exposure of the active site to the substrate is not clearly understood. Approaches adopted in attempts to increase the enzyme’s activity by promoting the open lid conformation include covalently trapping the lid in its open conformation, which resulted in an enzyme with a partly open lid but decreased activity (Svendsen, 2000), and complete removal of the lid, which resulted in an inactive variant (Svendsen, 2000). Our approach aimed at selecting variants in which the lid opened more easily, thereby increasing the overall catalytic turnover. Structural analyses, together with results from previous studies, suggest that lid opening involves the movement of the hinge regions, which were selected as the target areas for randomization (Berg et al., 1998; Martinelle et al., 1996; Brzozowski et al., 1991)
Fig. 10 Computer graphic representation of the positions targeted for library constructions. A. Structure of the *Thermomyces lanuginosa* lipase, Lipolase. The position of the lid region of the enzyme in its closed and open position is shown superimposed. B. Enlargement of the lid region from amino acid residue 81 to 98 showing the α-helical lid and the two hinge regions involved in the lid movement. Positions targeted for mutagenesis in this study are indicated in grey.

### 7.2 Functional display of Lipolase on phage

The gene encoding the wild-type lipolase gene was sub-cloned into the phagemid-vector pKNΔH3. To facilitate later cassette mutagenesis of the lipolase gene, two unique restriction sites were introduced by PCR without altering the amino acid sequence. The resulting phagemid construct, pKNΔH3-Lip (Fig. 11), encoded all the elements required for secreted expression of the lipase, which was fused to a 5 kDa serum albumin-binding affinity-tag followed by a suppressible amber stop codon (TAG) and a truncated version of the M13 phage coat protein 3. Expression studies showed that full-length lipase-ABD fusion protein could be expressed as a soluble protein in the *E. coli* periplasm and purified by HSA-affinity chromatography to high degree, as determined by SDS-PAGE analysis. Comparing the specific activities of the affinity-purified lipase-ABD fusion protein to purified commercial Lipolase produced in *A. oryzae* showed that they were very similar; 167 U/nmole for the Lipolase-ABD fusion, which was comparable to the *A. oryzae*-produced enzyme, 142 U/nmole and to literature data (Martinelle et al., 1996). This result indicated that the glycosylation of the enzyme, which occurs in the *A. oryzae* host as well as in the natural host is not necessary for full *in vitro* activity and that the fusion of ABD to the C-terminal did not notably influence on the activity of the enzyme. In addition, it was demonstrated using a
sandwich ELISA assay that phage particles, produced by superinfection of host cells with helper phage displayed lipase on their surface. The displayed enzyme was also shown to be active by analyzing a filtered phage stock using pNP-palmitate/triolein-coated microtiter wells.

**Fig. 11.** A schematic overview of the phagemid construct used in this study. The lipase is produced fused to the N-terminus of a truncated version of the minor phage coat protein 3 (res. 249-406), via an albumin binding affinity tag (ABD, albumin binding domain). For secretion, the E. coli Omp A leader peptide is used.

### 7.3 Library constructions

A total of nine amino acids (R81, S83, S84, L93, N94, F95, D96 and L97) located in the hinge regions of Lipolase were targeted for randomization, and two different libraries were constructed. In the first randomization strategy, (C/A/G)N(G/C) codons, excluding stop codons or aromatic residues, were used at positions 81, 84, 93, 94, 96 and 97, while NN(G/C) codons, allowing all amino acids, were used at positions 85 and 95 (Table 1). The possible role of S83 as part of the oxyanion hole, stabilizing the transition state meant that this position was variegated to 90% Serine (wt) and 10% Threonine (Thr has been found at this location in an homologous enzyme). A fraction of the resulting library containing $5 \times 10^7$ different variants was analyzed (cell format) on rhodamine B/olive oil-containing agar indicator plates. In this activity assay, ester hydrolysis results in the release of protons, lowering the pH and rhodamine B acts as a pH indicator turning lipase-active colonies pink. Visual inspection of the rhodamine B plates showed that only approximately 0.03% of the colonies actually turned pink, indicating that the library contained a low amount of active variants.

To increase the fraction of active variants a new library was constructed by a second randomization strategy designed to reduce the number of mutations per variant, thereby potentially increasing the number of active variants in the library. In the first library, the probability of the wild type amino acid occurring at each targeted position was approximately 0.07%, implying that the number of mutations per gene would be relatively high (Fig. 12). In the second library the aim was to have approximately 50% probability that the wild type
amino acid would occur at each targeted position (Fig. 12). A Matlab simulation was used in an attempt to optimize each position of the variegated codons in terms of relative nucleotide use, in such a way that every group of amino acids (small, charged, polar and hydrophobic), at least, would be represented. The program presents the relative frequency of appearance of the different amino acids when using a specific codon, statistically calculated from a set of, typically, 10 000 random clones. Table 2 shows the codon design used at each randomized position in the second library, with the corresponding prognosis at the amino acid level. During this work the major inherent limitations in the genetic code became very clear, in attempts to tailor-make the amino acid distribution. This project was limited to the use of standard oligonucleotide synthesis, but the problem could have been avoided, by using trinucleotide synthesis for example (see section 4.1.2.1).

Analysis of the second library, again on plates containing rhodamine B/olive oil, showed that the fraction of active colonies had increased to approximately 0.2%. To analyze the library further, more than 70 different clones were sequenced to assess its genetic quality. The results showed that all analyzed clones were different, with on average 3.6 mutations per gene. Among this set of clones no wild type sequence was found. Of the sequenced clones, 7% encoded a threonine at position 83, which is close to the theoretical value (10%) aimed at in the oligo (library) design.

![Graph](diagram.png)

**Fig. 12** On the x-axis: the probability of x number of mutations per lipase variant for Lipolase library 1 and aimed for in library 2, respectively. On the y-axis: the fraction of protein variants with x-number of mutations. ($P_{n,\text{mut}} = (1-P_w)^n \times (P_w)^{m-n} \times m!/(n!(m-n)!)$)
Table. 1. Design of the degenerate codons in the two Lipolase libraries.

<table>
<thead>
<tr>
<th>Codon no. (a)</th>
<th>Arg 81, 84</th>
<th>Ser 83</th>
<th>Ser 85</th>
<th>Asn 94</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lip-library 1</strong></td>
<td>(C/A/G)N(G/C)</td>
<td>(A/T)C(C/T)</td>
<td>NN(G/C)</td>
<td>(C/A/G)N(G/C)</td>
</tr>
<tr>
<td>Lip-library 2</td>
<td>(C/A/G)X(G/C)</td>
<td>(A/T)C(C/T)</td>
<td>XX(G/C)</td>
<td>(C/A/G)X(G/C)</td>
</tr>
<tr>
<td>A1, C1, G1, T1</td>
<td>15, 70, 15, 0</td>
<td>10, 0, 0, 90</td>
<td>15, 5, 10, 70</td>
<td>75, 10, 15, 0</td>
</tr>
<tr>
<td>A2, C2, G2, T2</td>
<td>15, 5, 70, 10</td>
<td>0, 100, 0, 0</td>
<td>15, 75, 5, 5</td>
<td>79, 7, 7, 7</td>
</tr>
<tr>
<td>A3, C3, G3, T3</td>
<td>0, 50, 50, 0</td>
<td>0, 50, 0, 50</td>
<td>0, 70, 30, 0</td>
<td>0, 90, 10, 0</td>
</tr>
<tr>
<td><strong>Codon no. (a)</strong></td>
<td>Phe 95</td>
<td>Asp 96</td>
<td>Leu 93, 97</td>
<td></td>
</tr>
<tr>
<td><strong>Lip-library 1</strong></td>
<td>NN(G/C)</td>
<td>(C/A/G)N(G/C)</td>
<td>(C/A/G)N(G/C)</td>
<td></td>
</tr>
<tr>
<td>Lip-library 2</td>
<td>XX(G/C)</td>
<td>(C/A/G)X(G/C)</td>
<td>(C/A/G)X(G/C)</td>
<td></td>
</tr>
<tr>
<td>A1, C1, G1, T1</td>
<td>10, 10, 10, 70</td>
<td>10, 10, 80, 0</td>
<td>10, 75, 15, 0</td>
<td></td>
</tr>
<tr>
<td>A2, C2, G2, T2</td>
<td>15, 15, 5, 65</td>
<td>79, 7, 7, 7</td>
<td>10, 10, 5, 75</td>
<td></td>
</tr>
<tr>
<td>A3, C3, G3, T3</td>
<td>0, 95, 5, 0</td>
<td>0, 85, 15, 0</td>
<td>0, 30, 70, 0</td>
<td></td>
</tr>
</tbody>
</table>

(a) Positions in triplets are indexed with 1, 2 and 3

7.4 Selection using a mechanism-based inhibitor

For selection, a mechanism-based inhibitor was used (Deussen et al 2000), which mimics the natural substrate but becomes permanently attached to the enzyme during the catalytic reaction (Fig.13). Initial experiments had shown that the designed inhibitor was capable of inhibiting the Lipolase enzyme, so it acts as a substrate mimic. This study showed that at an inhibitor concentration of 25 µM, the lipase was inhibited by 50% after twelve minutes.

Selections were carried out in triolein-coated microtiter wells that also contained 0.4 nmol of the biotinylated inhibitor. Lipase-displaying phages were added to the wells in the presence of 0.14 % w/v of a commercial detergent powder (containing linear alkylbenzenesulfonate, alkyl sulfate and alcohol ethoxylate). Inhibitor-bound phages were subsequently collected by capture on streptavidin-coated microbeads in a Tween-containing buffer from which they were later released by reducing the S-S bond present in the inhibitor (Fig. 13). After three rounds of selection the active fraction of collected clones had increased from 0.2% in the input library to 90% as determined in rhodamine B-plate assays. Activity analyses of overnight cultures of 84 positive clones, as determined by the plate assay, against the chromogenic substrate pNP-palmitate showed that none of the selected clones performed better than the wildtype enzyme under these conditions, either with or without detergent. Three clones with the highest overall activities were produced in Aspergillus oryzae for a
more detailed analysis, which confirmed that the selected clones did not perform better than the wild type enzyme either in the presence or absence of detergent.

However, a more thorough analysis of selected clones did show some interesting features. For example, the frequency of threonines at position 83 (initially doped with 10% of threonine codons) had risen to 85% from 7% in the unselected active clones. This position (T83) has been shown to lower the specific activity of a homologous enzyme from *Rhizopus delemar* and increase its preference for substrates with long chains. Furthermore, the first hinge region was highly conserved; the clones showing the highest activity did not have any substitutions at positions 81, 84 or 85. The arginine at position 84 has earlier been shown to be involved in a local rearrangement mechanism facilitating the subsequent lid opening that exposes the active site to the substrate (Brzozowski et al 2000). Mutating Arg 84 to a Gly lowered the activity of the enzyme almost 100-fold (Brzozowski et al 2000). In the second hinge region, corresponding to positions 93 to 97, several selected variants contained basic amino acids at three positions not seen in the wild type enzyme. It has been reported that electrostatic interactions between anionic detergents stabilize the open conformation, which could favor the overall activity of the enzyme (Cajal et al., 2000; Cajal et al., 2000; Berg et al., 1998).
Ester + E → Acyl-E + alcohol → Acyl-E + H₂O → E + fatty acid

Fig. 14. A schematic view of the two-step catalytic mechanism of lipases, an acylation step and a deacylation step.

Despite these interesting features we were not able to select an enzyme that performed better than the wild type enzyme in either the presence or absence of detergent. One possible reason for this failure is that the catalytic event of Lipolase, including the opening of the lid, is a complex, multi-event process in which any of the following steps could be rate-limiting: (1) binding to the lipid surface, (2) penetration into the lipid interphase, (3) activation of the lipase; i.e. opening of the lid and hydrolysis of the substrate by (4) acylation and (5) deacylation of the enzyme (Fig.14). Because of this complex reaction mechanism the design of a bait and selection scheme that increases the frequency of variants with the desired properties is even more difficult. It is also possible, of course, that variants of the desired types were not present in the library due to its small size, or because the area targeted for randomization is not the only area responsible for efficient activation. Studies of the interfacial activation of the enzyme suggest that there is a delicate balance of electrostatic and hydrophobic interactions between the lipase and the lipid interface. For example, zwitterionic micelles have been shown not to activate the lid opening (Cajal et al., 2000). Variants performing better in a complex environment, as provided by a detergent powder, need to balance these two properties to allow activation of the enzyme, which could be difficult to engineer. Nevertheless, the results showed that it was possible to increase the abundance of active variants in the library using a phage display strategy in combination with a tailor-made suicide inhibitor target molecule. In principle, these results are encouraging for future engineering efforts of this enzyme. Finding variants with the desired properties could be facilitated by a combination of different randomization strategies together with a combination of selection/screening methods. One possible scenario is further screening of the enriched active lipase variants, obtained using our approach, followed by recombination via DNA shuffling of the most active variants together with the PCR-based introduction of randomly distributed mutations.
8. Selection and analyses of anti-idiotypic affibody binding pairs (II, IV)

The affibody affinity protein technology platform has hitherto been used to select affibody-binding proteins for a variety of protein targets with various kinds of biotechnological uses. In this work, the parental SPA domains as well as fusion proteins containing previously selected affibodies were recruited as new types of selection targets for the isolation of novel affibodies. Motivation was provided by the belief that resulting affinity protein pairs could be used for different affinity protein applications based on the use of two non-identical small protein domains capable of mutual recognition, including affinity purification, detection, targeting and affinity driven protein complex assembly.

Two different affibody phage libraries were available for this project, one constructed to allow complete randomization of the 13 positions using NN(G/T) degeneracy (Nord et al., 1995), and the other constructed to exclude codons for aromatic residues and stop codons, using (G/A/C)NN codons, at all thirteen positions (Nord et al., 1997). Both of these libraries were used in mixtures during selections.

In a pre-study, aiming to isolate affibodies that could be used for general detection and capture reagents for affibody constructs, three different affibody-ABD fusion proteins (containing the previously described affibodies $Z_{Taq}$ S1-1 (Gunneriusson et al., 1999), $Z_{IgA1}$ (5:6) (Rönnmark et al., 2003) and $Z_{Apo}$ 24:4 (Nord et al., 2000) were used as targets during selections. The rationale behind this approach was to direct the binding specificity to the third, non-variegated, helix of the affibody scaffold, corresponding to a region common to all affibody constructs (Fig. 15). Obviously, binders with affinity to the ABD fusion partner present in all the used target constructs could also be expected. In biosensor studies of selected variants, only a single binder with low affinity for an affibody-ABD fusion protein was identified. Nevertheless, in a subsequent analytical gel filtration experiment this variant (as an ABD fusion protein construct) showed signs of dimerization, suggesting it had self-binding capability through recognition of either the ABD fusion partner or an affibody-derived epitope (unpublished results). The difficulties experienced in selecting for such self-recognition affibodies could be due to any one or several factors: (i) self-recognition could lead to expression problems due to aggregation; (ii) problems during phage assembly could cause selection against such binders in early stages, (iii) the library could have been too small to encompass an appropriate affibody or (iv) the third helix of the Z domain scaffold could
represent a difficult target, presenting a surface that has evolved to be "resistant" to protein-protein interaction.

![Diagram](https://example.com/diagram.png)

**Fig. 15.** **A.** The selection strategy used to isolate affibody-based affinity pairs. **B.** A top view of the Z-domain from SPA. Residues involved in Fc-binding, positioned on helix 1 and 2 are colored blue. In the Z-library used for selection 13 positions located to this interaction surface are randomized. The residues involved Fab binding is postulated from data described for the D-domain of SPA (red). Glutamine 32, postulated to be involved in both interactions are highlighted in yellow. In a following experiment, a five-domain wild type SPA construct (E-D-A-B-C) was used instead as the sole target during multiple cycle affibody selections. Thus, each of the individual domains in this target corresponded closely to the Z domain used as scaffold for affibody library constructions. Here, two alternative selection protocols were used involving (i) competitive elution using human IgG, a reagent which contains a high concentration of antibodies with protein A-binding Fc regions (Fc binding site competitors) as well as antibodies comprising Fab fragments belonging to the protein A-binding Vιι III family (Fab binding site competitors), albeit at a lower concentration or (ii) a standard phage elution protocol using a low pH buffer.

After four and five selection rounds, respectively, two SPA-binders were identified in biosensor binding studies among randomly picked clones previously analyzed by DNA sequencing. One of the SPA-binding variants, Z_{SPA-1}, isolated via the competitive elution strategy, showed a significant binding affinity towards SPA. No binding was observed to an ABD analogue protein (BB), indicating that the binding was SPA-specific. The second binder, Z_{SPA-2}, isolated by acid elution, displayed a ten-fold lower affinity for SPA and was therefore not included in the following studies.
Combinatorial protein engineering applied to enzyme catalysis and molecular recognition

Fig. 16. A. A subtractive sensorgram showing the interacting between $Z_{\text{SPA-1}}$-ABD and SPA. B. The reversed set-up, SPA interacting with $Z_{\text{SPA-1}}$-ABD immobilized onto a sensor chip surface.

As can be seen from the resulting sensorogram, the interaction between $Z_{\text{SPA-1}}$ and SPA has relatively fast association and dissociation rate kinetics (Fig. 16). In a reversed experimental setup, injection of SPA over a sensor chip surface containing immobilized $Z_{\text{SPA-1}}$-ABD protein resulted in a response with a significantly slower apparent dissociation rate. Such effects can often be explained by co-operative binding (avidity effects), indicating that the binding site recognized by the $Z_{\text{SPA-1}}$ affibody may have been present in more than one copy in the five-domain SPA construct. Later, results from additional biosensor binding studies directed towards single SPA domains (E, D, A, B, C, and Z) showed that the $Z_{\text{SPA-1}}$ affibody (purified using SPA-Sepharose) was indeed capable of binding to all of the six individually produced domains, suggesting that a common binding site was present in each domain. The affinities ($K_D$, dissociation constants) of these interactions were determined to be in the range of 2-6 µM. Due to the potential structural similarity between the different recognized SPA-derived domains (including Z) and the $Z_{\text{SPA-1}}$-affibody itself, assuming they had an intact three-helix bundle structure, there is a clear possibility of self-recognition by the affibody. However, injection of a 20 µM solution of $Z_{\text{SPA-1}}$ (HBS buffer, pH 7.4) over a sensor chip surface containing the $Z_{\text{SPA-1}}$-affibody did not reveal any signs of self-recognition. Further, in a competitive binding analysis using the biosensor instrument, the $Z_{\text{SPA-1}}$ affibody was shown to compete with the B-domain of SPA for binding to the Fc portion, but not with the Fab portion, of human antibodies. These findings were subsequently confirmed when the structure
of the Z/Z_{SPA-1} complex was solved, independently, by both crystallography (IV) and NMR (Wahlberg et al., 2003).

The successful identification of an affibody capable of selective recognition of the Fc-binding surface of its parental scaffold Z, prompted a further investigation of whether it was also possible to select an affibody recognizing another affibody. Therefore, to identify such additional binding pairs, two previously isolated affibodies, one against Taq DNA polymerase (Z_{Taq S1-1} (Gunneriusson et al., 1999)) and the other against human IgA (Z_{IgA1} (Rönnmark et al., 2003)) were chosen as targets for additional bio-panning experiments. For this selection, a standard, low pH elution protocol was used. In contrast to the competitive elution strategy, this strategy should theoretically allow unbiased isolation of binders to any sites on an affibody, in addition to their respective binding surfaces.

These selections generated several binders of the target affibodies Z_{Taq S1-1} and Z_{IgA1}. For each target, five variants isolated after panning cycles four and five were analyzed further. As can be seen from the biosensor binding results (Fig. 16), all ten analyzed binders bound to their respective target affibody. Further, eight out of the ten binders showed selective recognition of only the affibody variant used for its selection. The other three showed binding to both the Z_{Taq S1-1} and Z_{IgA1} affibodies, but with higher binding affinity for their respective intended target affibodies. This suggested that two types of affibodies had been selected, directed either to combined epitopes, one element of which is present in the affibody scaffold structure itself, leading to cross-reactivity, or to epitopes unique to each of the two different affibodies, resulting in highly selective affibody pairs. Such pairs of affibodies, based on the same basic structure and capable of mutual recognition of each other’s variegated surfaces, could also be referred to as anti-idiotypic binding pairs, in analogy to anti-idiotypic antibodies (Cavill et al., 2003). Interestingly, the selected binders did not show any binding to the ABD domain used as a control protein during biosensor analyses, despite the fact that this domain was also present in the affibody constructs used during bio-panning, as an affinity tag to facilitate purification.
Combinatorial protein engineering applied to enzyme catalysis and molecular recognition

Fig. 17. Biosensor analysis of isolated Z-variants after selection against an affibody binding to Taq polymerase (bottom picture) and an affibody with affinity towards IgA (top picture). The biacore chip is designed with 1) blank surface, 2) Z$_{TaqS1-1}$, 3) Z$_{IgA1}$ and 4) a control protein (BB). All the selected variants exhibit specific interaction with their respective target protein except for two binders against Z$_{TaqS1-1}$ (see arrows) that show a weak interaction with the opposite target. The response from the control-protein surface (BB) is non-detectable.

Collectively, the results of this study showed the possibility of using protein library technology to identify novel affinity protein pairs, with anti-idiotypic binding characteristics. This feature opens up a number of potential applications, including bioseparation (IV) and protein complex assembly (V) as well as diagnostic applications. For example, if one of the affibodies in an anti-idiotypic binding pair is directed towards an analyte of diagnostic interest (such as Z$_{IgA}$), this analyte (IgA) has the potential to interfere with the previously assembled
affibody-pair. It would be possible to monitor such analyte-dependent complex dissociations by various means, and preliminary studies indicate that fluorescence energy transfer (FRET) would be a suitable technique (Renberg et al., manuscript in preparation).

8.2 Crystal structure of the Z/Z$_{SPA-1}$ affinity protein pair

As reported in Paper IV, the structure of the Z/Z$_{SPA-1}$ affinity protein pair has been solved by x-ray crystallography at 2.3 Å resolution (PDB ID code 1LP1), making this the first affibody-target co-complex to be structurally determined. The two proteins used for the study were produced without any additional extensions (such as tags or leaders) and affinity-recovered to high purity by their respective affinities to IgG- or SPA- containing resins. Although a relatively high concentration of the soluble protein complex was used (a stock of 72 mg/ml diluted to approx. 30 mg/ml), crystallization was not straightforward. Crystallization probably occurred only after a further increase in concentration, due to the use of a container that was not air-tight.

The structure showed that the affibody binds to helices one and two of the Z domain, the surface involved in Fc-binding, confirming the binding competition results obtained earlier (II). Both the Z-domain and the Z$_{SPA-1}$-affibody have the three-helix bundle topology previously reported for the Z domain in solution (Tashiro et al., 1997, PDB code 2SPZ). When the individual structures of the domains are superimposed onto each other, it can be seen that their main chain structures are indeed very similar, with only small deviations (less than 1 Å). In the interaction surface, the variegated helices 1 and 2 of the affibody interact mainly with helix 1 in the Z domain, but also with a significant part of helix 2. The two proteins bind at an angle of approximately 60° to each other (front page). The interaction surface is characterized by a central hydrophobic patch, surrounded by polar and charged residues (Fig. 18) that contribute several hydrogen bonds. The interaction surface is highly complementary in shape, with several large side chains from the affibody protruding into cavities of the Z domain, including the aromatic side chains of Phe 32 and Trp 35 (Fig. 19). Likewise, the side chain of Phe 13 of the Z domain is buried between helices 1 and 2 of the affibody. Variegated positions of the affibody dominate the interaction since nine out of 13 such positions are involved, in addition to four non-mutated positions. Notably, the Z$_{SPA-1}$ affibody was selected from a relatively small affibody library (5 x 10$^7$ variants) compared to the theoretical sequence space of >10$^{16}$ variants, yet it still shows a high degree of shape complementarity and use of variegated positions in the interaction. The surface area involved
in the $Z/Z_{SPA-1}$ interaction is approximately $1665 \, \text{Å}^2$ (sum of both surfaces), which is consistent with the size of interaction surfaces seen in many antibody-antigen complexes (Chakrabarti and Janin, 2002).

**Fig. 18.** The interaction surface of the complex, $Z$ (right) and $Z_{SPA-1}$ (left) (opened like a book), the hydrophobic core (white) is surrounded by polar residues (blue and red). The surface exhibits a fairly good charge complementarity.
Fig. 19. The Z-domain in green and helices 1 and 2 of ZSPA-1 in blue. Mutated residues are in atom color and wt-residues involved in binding are colored red.

As previously shown, the ZSPA-1 affibody binds to all of the SPA domains, as well as the Z-domain, with similar affinities. This correlates well with indications obtained from inspection of the interaction surface, in that only three of 13 interacting residues of the Z domain are not conserved between the different SPA domains. In addition, these three non-conserved residues are all found at the periphery of the binding surface.

The structure of a single SPA-domain (the B-domain) in complex with its natural target, IgG1 Fc has been solved (PDB ID 1FC2), as well as the NMR structure of the Z-domain in solution (PDB ID 2SPZ), which makes it possible to compare the two complexes and their protein-protein interaction characteristics. Interestingly, the surface used by the Z domain in its interaction with IgG1 Fc overlaps to a very high degree with the surface used in the interaction with ZSPA-1. In addition, the structure of the Z/ ZSPA-1 complex has been independently solved by NMR (Wahlberg et al., 2003), yielding a structure that correlates
very well with the crystal structure. In the cited report it is noted that the two side chains (Tyr 14 and His 18) in the Z/B domain undergo almost identical conformational changes upon binding to both IgG Fc and the ZSPA-1 affibody. This implies that both IgG Fc and the ZSPA-1 affibody are able to induce similar binding modes in their interactions with the Z domain, although the two binding partners are composed of different secondary structure elements (loops/β-sheets and α-helices, respectively). One obvious reason for the similarities between the interacting surfaces used by the Z-domain is that the competitive elution with human IgG performed during the selection process may favor binders directed towards the Fc-binding surface of the Z-domain. It could also be due to the potentially negative selection of binders recognizing regions of the Z-domain that are also conserved in the affibody, leading to self-recognition. However, the selection directed towards other affibodies (generating anti-idiotypic affibody pairs) using low pH elution, which would allow unbiased isolation of binders to several sites on the affibodies, resulted in a majority of variants displaying selective binding only to the intended target affibody. This suggests, assuming that they have intact three dimensional structures, that these binders bind to the randomized surface of their target-affibody. Again, this could be due to negative selection against self-recognizing affibodies. It would thus be interesting to use a protein library based on an alternative scaffold for selection experiments using the Z domain as panning target to investigate whether the Fc binding surface corresponds to a particularly preferred surface for interactions with other proteins.

In the work by Wahlberg and coworkers it was reported that the uncomplexed ZSPA-1 affibody has a molten globule-like structure in solution, i.e. that it adopts a compact, but only partially folded conformation that resembles the native protein (Wahlberg et al., 2003). This could obviously be due to the relatively large number of substitutions made in the wild type Z domain (13 of 58 residues) to obtain the ZSPA-1 affibody. However, considering the small size of the library from which the ZSPA-1 affibody was selected, it is plausible that Z-binding affibodies with improved stability could be obtained either by affinity maturation or by selecting from a larger initial library. In fact, in an attempt to obtain such binders the ZSPA-1 affibody-encoding gene was subjected to whole-gene random mutagenesis by a nucleotide analogue-PCR technique, followed by library construction and phage display selection using a dimer of the Z-domain (ZZ) as panning target (Eklund and Oijaniemi, Master thesis 2002). Analysis of the relatively small library obtained (30 000 clones) showed that, on average, 3.4 amino acid substitutions per variant had been obtained. Selections resulted in a strong convergence towards one variant in which Pro 25 (positioned at the beginning of helix 2) in
the $Z_{\text{SPA-1}}$ affibody had been replaced by a Ser residue. This was interesting since Pro is regarded as an amino acid with low helix-forming propensity and Pro 25 has not been suggested to be primarily involved in the interaction with the Z domain. Binding studies using this variant showed only a marginal increase in binding strength. However, the strong convergence for this variant could be due to other reasons, including competitive behavior during expression or phage assembly, possibly reflecting increased structural stability.

9. Applications of hetero dimeric binding pairs (III, V)

9.1 Purification and detection

At present, major efforts are being made to develop expression and purification systems that would allow high-throughput recombinant production of the vast number of putative proteins being identified in various genome sequencing projects. Ideally, such a system should allow the production and purification of a large number of proteins with a wide range of different properties, and be amenable to automation. To circumvent the need to develop individual purification strategies for each protein, affinity gene fusion systems are preferred for such projects (Hammarström et al., 2002). A large number of different gene fusion tag systems have been developed, including fusion partners of different sizes capable of interacting with various ligands immobilized on a chromatographic resin (Hearn and Acosta, 2001). No system is expected to be ideal for all situations and the system of choice is likely to depend on factors such as the host cell system, cellular location of the product, and product characteristics. New systems are continuously being developed, which adds to the portfolio from which to choose (Lamla and Erdmann, 2004; de Marco et al., 2004; Zhang et al., 2001).

In the studies reported in Paper IV, the previously identified $Z_{\text{SPA-1}}$ affibody was investigated for use as a gene fusion partner in the production, SPA affinity chromatography purification and detection of recombinant proteins.

Protein A-containing resins are used in a variety of different biotechnological applications, including purification and detection of immunoglobulins using Protein A-containing resins, conjugates between SPA and different reporters such as enzymes, fluorescent groups, isotopes and various beads. In addition, SPA and the SPA derivative Z have been used as affinity gene fusion partners, allowing purification of a variety of recombinant proteins on IgG-sepharose (Ståhl et al., 1999; Nilsson et al., 1997). However, a drawback of using IgG as an immobilized ligand is that the IgG molecule, consisting of four protein subunits, is relatively fragile and leakage of antibody subdomains into the eluate may occur. Furthermore,
commercially available IgG resins are based on polyclonal immunoglobulin from human serum, which is a major drawback if the products are intended for therapeutic use, due to the risk of viral contamination.

The selected SPA-binding affibody \( Z_{SPA-1} \) could, thus, be an attractive alternative to the present use of IgG-columns. Alternatively, the \( Z_{SPA-1} \) affibody could be used as an affinity-tag, providing scope to exploit the entire array of SPA-based products available today.

As described in Paper II, the \( Z_{SPA-1} \) affibody was produced either as a single protein domain, or fused to the ABD fusion partner and recovered from whole \( E. coli \) cell lysates or periplasmic extracts, respectively, using commercially available SPA-sepharose. Analysis of the eluate showed that both proteins could be purified to near homogeneity by this one-step purification. In fact, the monomeric \( Z_{SPA-1} \) affibody used for crystallization (III) was purified using this one-step purification strategy from a fed-batch fermentation and was estimated to be 95-98% pure as determined from a silver-stained SDS-PAGE gel.

In further studies (Paper III) the aim was to investigate the \( Z_{SPA-1} \) affibody/SPA affinity system as a more general tool for bioseparation and detection applications. For this purpose, the plasmid vector, pAff\#1c was constructed to facilitate T7 promoter-driven intracellular expression of target proteins fused both to a His\(_6\) tag and the \( Z_{SPA-1}\)-affibody. In order to evaluate the expression vector, five different cDNA clones from a mouse testis library, which had previously been used in expression studies (Larsson et al., 2000), were selected. To compare purification efficiencies, the fusion proteins were purified in parallel by immobilized metal-ion affinity chromatography (IMAC) (His\(_6\) tag) and SPA-Sepharose (\( Z_{SPA-1}\)-tag) chromatography. It was found that the protein A-based recovery resulted in the highest degree of purity (Fig. 20). Expression levels estimated after affinity purification were in the range of 20 to 75 mg per liter, in accordance with the levels obtained using the same cDNA-clones in a different expression system (Larsson et al., 2000). As well as the expression vector, two different blotting procedures were investigated. Firstly, a two-step affinity blotting procedure, utilizing SPA as the primary reagent and a peroxidase-rabbit anti-peroxidase antibody complex recognized by SPA as the secondary reagent, followed by development with peroxidase substrate. This procedure has the advantage that the reagents are all commercially available. Secondly, a one-step blotting procedure using a Protein A-Alkaline Phosphatase (PAAP) fusion protein was developed. This procedure is less laborious, but the PAAP fusion protein is not commercially available. These two blotting procedures were used to screen expression levels of cell cultures directly (a valuable feature of a high-throughput system,
allowing non-expressible proteins to be removed at an early stage, and facilitating analysis of the purity and stability of the different fusion proteins). The Z/IgG affinity system has an additional drawback in this respect, in that the Z-tag binds to “any” antibody in blotting procedures via its Fc interaction.

The pAff11c vector has been included more recently as part of a dual affinity system for production and purification of cDNA-specific antibodies (Falk et al., 2003).

![Fig. 20. SDS-PAGE analysis of recombinant Z_{SPA-1} fusion proteins corresponding to five cDNA clones (8, 18, 29, 39 and 54) that were recovered using both (A) IMAC, employing the His-tag and (B) protein A-sepharose, employing the Z_{SPA-1} tag.](image)

**9.2 Development of an artificial cellulosome-like complex**

As discussed earlier, affibody binding protein pairs have the potential to be used as convenient and modular affinity units as genetically fused to larger proteins. The development of anti-idiotypic affibody binding pairs allows for further engineering in that tools for precise molecular docking between two proteins, each joined with one part of such a pair, can be envisioned. Such principles could potentially be expanded to drive the spontaneous assembly also of higher order protein networks (architectures) by biospecificity principles.

Inspired by one example of such complex protein assemblies found in nature (cellulosomes; briefly described below), the work in paper V was directed towards an investigation of the suitability of the SPA-binding affibody, Z_{SPA-1} for use as a modular affinity unit in such applications.
9.2.1 The cellulosome

Cellulose, composed of 100 to 40 000 glucose units linked by β-1,4-bonds, is produced by plants at a yearly rate of approximately 40 Gt. Thus, cellulose degradation by various organisms is a very important process, representing a significant part of the global carbon cycle. However, crystalline cellulose is a highly complex, insoluble substrate often recalcitrant to hydrolysis by different enzymes. The way that organisms achieve effective degradation has therefore been studied intensively. Important cellulose-degrading species include various microorganisms that depend on an array of different hydrolytic/cellulolytic enzymes, with differing substrate specificities, for carbon extraction. For example, exoglucanases and endoglucanases are two different classes of enzymes that act synergistically on cellulose fibrils. Exoglucanases release cellobiohydrolase from one end of the glucose chain, while endoglucanases cut at random from the interior of the chains, increasing the amounts of available substrate for the exoglucanases.

In 1982 the anaerobic thermophilic bacterium Clostridium thermocellum was shown to degrade cellulose very efficiently using relatively small amounts of enzyme (Johnson et al., 1982). Thus, the specific activity of these enzymes must be much higher than that of homologous systems. Efforts were therefore made to purify single cellolytic enzymes from C. thermocellum, but with little initial success due to aggregation. These aggregates were later shown to be what are now called cellulosomes (Lamed et al., 1983).

In general terms, a cellulosome contains a scaffolding unit (scaffoldin) composed of several cohesin domains linked in series together with a cellulose-binding module, CBM. The CBM unit is responsible for anchoring the whole scaffolding complex to the substrate (cellulose). Each cellolytic enzyme is further linked to a dockerin domain facilitating the attachment of the enzyme to the scaffoldin, through the strong, calcium-dependent interaction (K_D = 10^-9) between the cohesin and dockerin domain (Fig. 21). Hydrolytic enzymes, with a range of substrate specificities, are aligned in this way along the scaffoldin, similar to an assembly line, working synergistically in that one enzyme generates the substrate for a second enzyme. This concerted mode of action results in a highly effective cellulose-degrading machinery. The cohesin domains have been shown to interact with any of the dockerin domains of the same species, suggesting that the cellulolytic enzymes are randomly incorporated along the scaffoldin (Pages et al., 1999; Gal et al., 1997).
Fig. 21. A schematic representation of the cellulosome, comprised of the scaffoldin platform built up of multiple copies of dockerin domains linked to a cellulose binding module (CBM), onto which cohesin-enzyme fusions can land.

9.2.2 Artificial cellulosomes (V)

Several studies have shown that individual cellulosomal subunits can be expressed with retained activities, inspiring investigators to design artificial cellulosomes (Fierobe et al., 2002; Fierobe et al., 2001). Such multienzyme complexes could be promising tools for use in biotechnological applications, e.g. to improve the hydrolysis of cellulosic substrates in the textile and paper industries (Bayer et al., 1994; Schwarz, 2001). Enzymes directly fused to a CBM domain have also been suggested to be promising tools for specifically modifying (rather than degrading) the cellulose substrate (Rotticci-Mulder et al., 2001; Gustavsson et al., 2004).

However, expression of cellulosomal subunits has not been straightforward. Dockerin, for example, has been shown to enhance the overall hydrophobicity of the fusion protein (Fierobe et al., 1991), to be prone to aggregation (Tokatlidis et al., 1991) and susceptible to proteolysis (Kataeva et al., 1997). Efforts in our laboratory have also encountered production difficulties as well as dissociation problems (Sandström, 2003).

In addition, an alternative route to create multienzyme complexes was investigated (Paper V), taking advantage of the SPA/ZSPA,1 affinity protein system. The aims were to create a cellulosome-like complex using a SPA-CBM fusion protein as a pentameric scaffolding unit and to investigate the possibility of using this platform for specific and consecutive docking of enzymes onto a cellulose surface. For use as a dockerin unit equivalent, a dimeric version of the ZSPA,1 affibody was fused to two different model reporter proteins: Enhanced Green Fluorescent Protein (EGFP), a GFP variant optimized for use in flow cytometry (Cormack et
al., 1996), and the lipase cutinase from *Fusarium solani pisi*, allowing convenient detection by a chromogenic substrate.

All proteins used for this study, \((Z_{\text{SPA-1}})_2\)-cutinase, \((Z_{\text{SPA-1}})_2\)-EGFP, SPA-CBM1\(_{\text{Cel6A}}\), including the control proteins CBM1\(_{\text{Cel6A}}\)-ABD and SPA were produced in *E. coli* and purified using either IgG-, SPA-, or HSA-affinity chromatography.

The binding characteristics of the \(Z_{\text{SPA-1}}\) containing constructs were first analysed using biosensor technology, which showed that the \(Z_{\text{SPA-1}}\) affibody retained its high binding selectivity for SPA after dimerization and fusion to two different foreign proteins. In addition, the comparison between the dimeric \(Z_{\text{SPA-1}}\) cutinase fusion protein and a fusion protein containing a single \(Z_{\text{SPA-1}}\) domain found considerably slower off-rate kinetics for the dimeric construct, indicating that the two \(Z_{\text{SPA-1}}\) moieties were able to bind SPA co-operatively, resulting in avidity effects. In addition, saturation experiments showed that more than one copy of the \((Z_{\text{SPA-1}})_2\)-tagged reporter proteins could simultaneously bind to a single SPA protein molecule.

A series of studies was performed on a cellulose-based filter paper to determine if directed and reversible anchoring of proteins to cellulose could be achieved (Fig. 22). For this purpose, small amounts of SPA-CBM1\(_{\text{Cel6A}}\) fusion protein or the two control proteins, CBM1\(_{\text{Cel6A}}\)-CBD and SPA, were spotted at specific locations on the filter paper. After washing and blocking, each filter paper was soaked in a solution containing \((Z_{\text{SPA-1}})_2\)-EGFP fusion protein. After an additional washing step, the different filter papers were analysed by fluorescence microscopy (excitation at 488 nm). Bright fluorescence was observed only at the location where SPA-CBM1\(_{\text{Cel6A}}\) fusion protein had been applied. Very little, or no, fluorescence was observed where the two different control proteins had been spotted. After a short incubation in a low pH (2.8) solution (to remove \((Z_{\text{SPA-1}})_2\)-EGFP from the filter paper) very little or no fluorescence was detected. A second incubation of the filter papers with \((Z_{\text{SPA-1}})_2\)-EGFP fusion protein solution was performed and, again, bright fluorescence was detected only at the location where SPA-CBM1\(_{\text{Cel6A}}\) fusion protein had previously been deposited. After low pH elution removing \((Z_{\text{SPA-1}})_2\)-EGFP, the different filter papers were incubated for a third and final time, in a solution containing the \((Z_{\text{SPA-1}})_2\)-cutinase construct. After washing, the presence of cutinase activity was detected using a chromogenic substrate, resulting in a red precipitate, showing that this reporter fusion protein had also been specifically directed to the location where SPA-CBM1\(_{\text{Cel6A}}\) fusion protein had previously been spotted. Taken together, these findings showed that the cellulose-anchored SPA-CBM1\(_{\text{Cel6A}}\) scaffoldin-like fusion
Fig. 22. Results from the consecutive docking experiments. Three different constructs were added onto separate filter papers (SPA-CBM1Cel6A and control proteins CBM1Cel6A-ABD and SPA).

A. (ZSPA-1)2-EGFP applied to all three papers. B. Removal of (ZSPA-1)2-EGFP (pH 2.8). C. A second incubation with (ZSPA-1)2-EGFP. D. Elution. E. A final incubation with the (ZSPA-1)2-cutinase construct.
protein remained stably anchored to the cellulose surface and retained activity through repeated cycles of washing and low pH elution.

Further, these results suggest that this affinity-based system could be used to assemble complexes of desired composition through a non-covalent and reversible interaction. Here, only one of the anti-idiotypic affibody-based affinity pairs developed so far was included, leaving room to extend the complexity to include all three affinity-pairs (Z/Z_{SPA-1}, Z_{Taq}/Z/Z_{Taq} and Z_{lgA}/Z/Z_{lgA}). To increase the apparent affinity of the SPA/Z_{SPA-1} affibody interaction, dimeric constructs were assembled for this study, but monomeric constructs could also be possible to use. If the strength of the interaction turns out to be insufficient, a larger library or/and affinity maturation procedures could be tested, as previously described for two different affibodies (Gunneriusson et al., 1999; Nord et al., 2001). In addition, natural cellulosomal subunits and affibody pairs could be used in combination, providing possibilities for selective elution of differently anchored enzymes with either EDTA (cohesin/dockerin) or low pH (affibody-pairs). Further, an exchange of the CBM1_{Cel6A} moiety for another suitable affinity domain or its deletion should allow protein complexes to be directed to other desired "addresses" or to be assembled in solution.
10. Concluding remark

In this work, combinatorial protein engineering coupled to phage display techniques was used to select for enzymatic activity and for the isolation of three specific anti-idiotypic protein affinity pairs, including the rather remarkable $Z_{SPA-1}$ affibody, which corresponds to the redesign of a protein to bind its ancestor.

As exemplified in the thesis, application areas for hetero-dimeric protein affinity pairs can be bioseparation, protein detection and to assemble and address protein complexes to specific locations. An exciting extension of these ideas would be to include several identical or different hetero-dimeric affinity protein pairs for the construction of "multimodule" combinations to which different fusion proteins containing cognate affinity proteins could dock and be located in close proximity. Such complexes could for example be used to assemble multi-enzyme networks facilitating a concerted action between enzymes involved in a common multi-step reaction pathway, as described earlier using direct fusion rather than biospecificity-driven assembly (Ljungcrantz et al., 1989; Lindbladh et al., 1994). Recently, similar ideas were presented for multivalent presentation of affinity proteins increasing the apparent affinity for a given target using modules based on the barnase-barstar complex as basis for constructions (Deyev et al., 2003).

Another application area in which biospecificity-driven interactions are believed to play an important role is in nanotechnology/material science. Here, various natural and synthetic molecules are investigated for their capability of spontaneous assembly into higher order structures (Zhang, 2003). Hetero-dimeric affinity proteins as the ones described in the thesis could be important tools in such efforts, which in combination with metal binding polypeptides, technology for metallization of proteins could result in exciting interdisciplinary solutions (Whaley et al., 2000, Reches and Gazit, 2003).
The combined use of a diverse gene library, a methodology to link genotype and phenotype, followed by a screening or selection scheme has provided researchers with powerful tools to understand and explore the versatility of proteins. For example, by this route enzymes with a high activity at both low and high temperatures in the same molecule, earlier thought to be incompatible, have been developed (Giver et al., 1998). Artificial binding proteins that exhibit affinities in parity with natural antibodies and higher have been isolated (Hanes et al., 2000) and even whole biochemical pathways have been altered (Schmidt-Dannert et al., 2000). The advances in this field will undoubtedly continue and more valuable information on protein function as well as novel proteins will be discovered. The future of protein science thus look bright and with an ever expanding tool-box for engineering one is tempted to agree with the following quotation: “About 10,000 years ago, humans began to domesticate plants and animals. Now it is time to domesticate molecules.” (Lindquist, 2003)

Fig. 23 Summary of the applications and possible applications of heterodimeric binding pairs. A. Immobilization, purification, detection and hetero-dimerization in solution. B. Multimerization of a fusion protein. C. Inclusion of additional pairs should allow for complex assemblies. Abbreviations: AP; affinity protein, T; target protein, X, Y and Z; any protein, R; reporter protein.
Abbreviations

aa    amino acid
ABD   albumin binding protein
CBM   cellulose binding module
CDR   complementarity determining region
cDNA  complementary DNA
DHFR  dihydrofolate reductase
EGFP  enhanced green fluorescent protein
Fab   antibody binding fragment
FACS  fluorescent activated cell sorting
Fc    constant region fragment
GFP   green fluorescent protein
HAS   human serum albumin
IgG   immunoglobulin G
IMAC  immobilized metal-ion affinity chromatography
MBI   mechanism-based inhibitor
PAAP  protein A conjugated alkaline phosphatase
PCR   polymerase chain reaction
PDB   protein databank
scFv  single chain variable region fragment
SDS-PAGE sodium dodecyl sulfate-polyacrylamide electrophoresis
SIP   selectively infective phage
SPA   Staphylococcus aureus protein A
SPG   streptococcal protein G
TSA   transition state analogue
wt    wild type
Z_{SPA-1} Affibody binding to SPA
Z_{TaqS1-1} affibody binding to Taq polymerase
Z_{IgA1} affibody binding IgA
Z_{ZTaqS1-1} an affibody binding to the $Z_{TaqS1-1}$ affibody
Z_{ZIgA1} an affibody binding to the $Z_{IgA1}$ affibody
Acknowledgements

Ja, då var det dags för sidan med stort S och den skrivs så klart i sista stund.

Jag vill börja med att tacka alla på Institutionen och framförallt ni på Avd. för Molekylär Bioteknologi (DNAcorner). Utan er hade jag inte klarat de perioder av frustration som alla doktorander/forskare tampas med. Det är ganska många som kommit och ett mindre antal gått och det fascinerande är att alla är lika trevliga. Hoppas det håller i sig även fast gruppen blir större och större och högre throughput är ambitionen. Tack för alla trevliga sammankomster fagmys, fredagsöl, peköl, konferenser, DNAcorner resor, småsnack i korridoren m.m m.m.

Vidare vill jag tacka:

Min handledare Per-Åke: För att du tillsammans med Mathias presenterade Lipolase-projektet på ett så inspirerande sätt att jag beslöt mig för att ta mig an utmaningen. För dina fenomenala titlar och andra vetenskapliga formuleringar som jag har fått dra nytt av genom åren. För att dina doktorander får möjlighet att ta sig an riktigt utmanande projekt och för att du alltid har ett extra kontrollexperiment i rockärmen. För att du på DNAcorner resor verkliga bjuder på dig själv. Jag kommer aldrig att glömma “I did it my way”. Och till sist, för att du har varit till stor hjälp under avhandlingskrivandet

Mathias, för att du har skapat DNAcorner tillsammans med dina adepter, som är ett mycket trevligt ställe. För att du har fenomenal förmåga att dra till dig kapital och häftiga projekt som jag har fått vara en liten del av.

Tobbe, min ex-jobbs handledare för god grundutbildning och för att du alltid har tid att bolla idéer, diskutera och ge goda råd och för trevliga EU-projektsresor.

Janne, Christin, Jenny och Henrik för att ni fick mig att känna mig som hemma direkt när jag anlände till labbet och för att vi fortfarande är polare.

Mina examensarbetare, Lasse och Hanna för trevligt samarbete och fin-fina vetenskapliga resultat.

Torbjörn, Jenny O för att ni tog er tid att läsa en del av denna textmassa.

Linda F för att du är en så god vän och tack för din hjälp med bilder och framsida.

Tuula Teeri och hennes trägång som jag har fått äran att vara en del av.

Björn för att han alltid vid rätt tillfälle delar med sig av sina klokheter såsom “I motvind lyfter draken”.

Mina dåvarande och nuvarande (John, Emilie, Torun och Erik (kul att du är tillbaka)) rumskompisar för trevligt s..t-snack och stöd.

Alla ni som varit med i det årliga Luciatåget. Jag tycker att det har varit otroligt vackert och mysigt. Ert luciatåg är det enda som kan få mig att gå upp så att ansa tidigt, (förutom Nora då, kanske).

Martin Högbom för att du såg till att hålet i tejpen var alldeles lagom stort så att vi fick vara med i PNAS.

Jonas Rundqvist för gott samarbete, delat intresse för AFM, kvasikovalent-koppling och roliga resor. Lycka till med SAM-lagren!

Kristofer för gott samarbete: Vi fick till fluorescensbilderna till slut.

Joke för att du är min kompis och har lärt mig att göra världens godaste smörgåsar.

Nina, för att du blev gravid så att jag också förstod att även jag är vuxen och redo för familj. Sabai-sabai kvällen är ett minne för livet.

Alla ni som deltagit i avdelningens babyboom för att ni funnits att dela tankar om graviditet, sömnlöshet, mat, trots och annat. Nina, Stina, Maria E, Maria W, Maria S, Åsa, S, Cissi, Ronny, Martin, Tobbe, Amelie, Jocke W, Fredrik S, Stefan, Sophia, Per-Åke och finns det fler?!


Mina kära vänner från Uppsala-tiden: Ebba, Anna S och Ulrika för att vi håller vänskapen vid liv trots att vi inte ses så ofta.

Kalmar-nations teatergrupps fd. medlemmar, Anna A, Robban, Jacob och Ola som jag fortfarande har turen att umgås med lite då och då.

Tack till tjocka släkten som har hängt med och stöttat mig genom alla år, från föreställningar med Ösmosbarn på torget i Nynäshamn till dansföreställningar på Operan, Nyköping och annat. Tillfällen då jag skulle vara med i tv och alla satt bänkade men jag var bortklippt, långa bildvisningar från ”backpackingresan” och nu disputation. För mysiga jular, somrar, dop kalas m.m m.m.

Mina föräldrar för att ni alltid ställer upp och tiden när jag skrivit avhandling är inget undantag. Tack för all hjälp! Min syster med familj (Anna, Knut, Isac och Rex). Jag älskar er alla mycket.

Min egen, kära lilla familj. Nora lilla älsklingen, för att du är en spillivink och sätter saker och ting i perpektiv och älskade Mats för att du är min livskamrat.
References


Combinatorial protein engineering applied to enzyme catalysis and molecular recognition


Combinatorial protein engineering applied to enzyme catalysis and molecular recognition


Combinatorial protein engineering applied to enzyme catalysis and molecular recognition


88


Malin Eklund


Combinatorial protein engineering applied to enzyme catalysis and molecular recognition


