#### UNIT III

#### **PROTEIN ENGINEERING**

**Protein engineering** is the process of developing useful or valuable proteins. It is a young discipline, with much research taking place into the understanding of protein folding and recognition for protein design principles. It is also a product and services market, with an estimated value of \$168 billion by 2017.

There are two general strategies for protein engineering: rational protein design and directed evolution. These methods are not mutually exclusive; researchers will often apply both. In the future, more detailed knowledge of protein structure and function, and advances in high-throughput screening, may greatly expand the abilities of protein engineering. Eventually, even unnatural amino acids may be included, via newer methods, such as expanded genetic code, that allow encoding novel amino acids in genetic code.

#### Approaches

## **Rational design**

In rational protein design, a scientist uses detailed knowledge of the structure and function of a protein to make desired changes. In general, this has the advantage of being inexpensive and technically easy, since site-directed mutagenesis methods are well-developed. However, its major drawback is that detailed structural knowledge of a protein is often unavailable, and, even when available, it can be very difficult to predict the effects of various mutations.

#### **Directed evolution**

In directed evolution, random mutagenesis, e.g. by error-prone PCR or Sequence Saturation Mutagenesis, is applied to a protein, and a selection regime is used to select variants having desired traits. Further rounds of mutation and selection are then applied. This method mimics natural evolution and, in general, produces superior results to rational design. An added process, termed DNA shuffling, mixes and matches pieces of successful variants to produce better results. Such processes mimic the recombination that occurs naturally during sexual reproduction. Advantages of directed evolution are that it requires no prior structural knowledge of a protein, nor is it necessary to be able to predict what effect a given mutation

will have. Indeed, the results of directed evolution experiments are often surprising in that desired changes are often caused by mutations that were not expected to have some effect. The drawback is that they require high-throughput screening, which is not feasible for all proteins. Large amounts of recombinant DNA must be mutated and the products screened for desired traits. The large number of variants often requires expensive robotic equipment to automate the process. Further, not all desired activities can be screened for easily.

Natural Darwinian evolution can be effectively imitated in the lab toward tailoring protein properties for diverse applications, including catalysis. Many experimental technologies exist to produce large and diverse protein libraries and for screening or selecting folded, functional variants. Folded proteins arise surprisingly frequently in random sequence space, an occurrence exploitable in evolving selective binders and catalysts. While more conservative than direct selection from deep sequence space, redesign of existing proteins by random mutagenesis and selection/screening is a particularly robust method for optimizing or altering extant properties. It also represents an excellent starting point for achieving more ambitious engineering goals. Allying experimental evolution with modern computational methods is likely the broadest, most fruitful strategy for generating functional macromolecules unknown to nature.

The main challenges of designing high quality mutant libraries have shown significant progress in the recent past. This progress has been in the form of better descriptions of the effects of mutational loads on protein traits. Also computational approaches have showed large advances in the innumerably large sequence space to more manageable screenable sizes, thus creating smart libraries of mutants. Library size has also been reduced to more screenable sizes by the identification of key beneficial residues using algorithms for systematic recombination. Finally a significant step forward toward efficient reengineering of enzymes has been made with the development of more accurate statistical models and algorithms quantifying and predicting coupled mutational effects on protein functions.

Generally, directed evolution may be summarized as an iterative two step process which involves generation of protein mutant libraries, and high throughput screening processes to select for variants with improved traits. This technique does not require prior knowledge of the protein structure and function relationship. Directed evolution utilizes random or focused mutagenesis to generate libraries of mutant proteins. Random mutations can be introduced using either error prone PCR, or site saturation mutagenesis. Mutants may also be generated using recombination of multiple homologous genes. Nature has evolved a limited number of beneficial sequences. Directed evolution makes it possible to identify undiscovered protein sequences which have novel functions. This ability is contingent on the proteins ability to tolerant amino acid residue substitutions without compromising folding or stability.

Directed evolution methods can be broadly categorized into two strategies, asexual and sexual methods.

#### **Asexual Methods**

Asexual methods no not generate any cross links between parental genes. Single genes are used to create mutant libraries using various mutagenic techniques. These asexual methods can produce either random or focused mutagenesis.

## **Random Mutagenesis**

Random mutageneic methods produce mutations at random throughout the gene of interest. Random mutagenesis can introduce the following types of mutations: transitions, transversions, insertions, deletions, inversion, missense, and nonsense. Examples of methods for producing random mutagenesis are below.

## **Error Prone PCR**

Error prone PCR utilizes the fact that Taq DNA polymerase lacks 3' to 5' exonuclease activity. This results in an error rate of 0.001-0.002% per nucleotide per replication. This method begins with choosing the gene, or the area within a gene, one wishes to mutate. Next, the extent of error required is calculated based upon the type and extent of activity one wishes to generate. This extent of error determines the error prone PCR strategy to be employed. Following PCR, the genes are cloned into a plasmid and introduced to competent cell systems. These cells are then screened for desired traits. Plasmids are then isolated for colonies which show improved traits, and are then used as templates the next round of mutagenesis. Error prone PCR shows biases for certain mutations relative to others. Such as biases for transitions over transversions.

Rates of error in PCR can be increased in the following ways:

- 1. Increase concentration of magnesium chloride, which stabilizes non complementary base pairing.
- 2. Add manganese chloride to reduce base pair specificity.
- 3. Increased and unbalanced addition of dNTPs.
- 4. Addition of base analogs like dITP, 8 oxo-dGTP, and dPTP.
- 5. Increase concentration of Taq polymerase.
- 6. Increase extension time.
- 7. Increase cycle time.
- 8. Use less accurate Taq polymerase.

Also see Polymerase chain reaction for more information.

## **Rolling circle error-prone PCR**

This PCR method is based upon rolling circle amplification, which is modeled from the method that bacteria use to amplify circular DNA. This method results in linear DNA duplexes. These fragments contain tandem repeats of circular DNA called concatamers, which can be transformed into bacterial strains. Mutations are introduced by first cloning the target sequence into an appropriate plasmid. Next, the amplification process begins using random hexamer primers and  $\Phi$ 29 DNA polymerase under error prone rolling circle amplification conditions. Additional conditions to produce error prone rolling circle amplification are 1.5 pM of template DNA, 1.5 mM MnCl<sub>2</sub> and a 24 hour reaction time. MnCl<sub>2</sub> is added into the reaction mixture to promote random point mutations in the DNA strands. Mutation rates can be increased by increasing the concentration of MnCl<sub>2</sub>, or by decreasing concentration of the template DNA. Error prone rolling circle amplification is advantageous relative to error prone PCR because of its use of universal random hexamer primers, rather than specific primers. Also the reaction products of this amplification do not need to be treated with ligases or endonucleases. This reaction is isothermal.

## **Chemical mutagenesis**

Chemical mutagenesis involves the use of chemical agents to introduce mutations into genetic sequences. Examples of chemical mutagens follow:

Sodium Bisulfate - This mutagenic agent is effective at mutating G/C rich genomic sequences. This is because sodium bisulfate catalyses deamination of unmethylated cytosine to uracil.

Ethyl Methane Sulfonate - This chemical agent alkylates guanidine residues. This alteration causes errors during DNA replication.

Nitrous Acid - This chemical agent causes transversion by de-amination of adenine and cytosine.

The dual approach to random chemical mutagenesis is an iterative two step process. First it involves the *in vivo* chemical mutagenesis of the gene of interest via EMS. Next, the treated gene is isolated and cloning into an untreated expression vector in order to prevent mutations in the plasmid backbone. This technique preserves the plasmids genetic properties.

## Targeting Glycosylases to Embedded Arrays for Mutagenesis (TaGTEAM)[]

This method has been used to create targeted *in vivo* mutagenesis in yeast. This method involves the fusion of a 3-methyladenine DNA glycosylase to tetR DNA-binding domain. This has been shown to increase mutation rates by over 800 time in regions of the genome containing tetO sites.

#### **Mutagenesis by Random Insertion and Deletion**

This method involves alteration in length of the sequence via simultaneous deletion and insertion of chunks of bases of arbitrary length. This method has been shown to produce proteins with new functionalities via introduction of new restriction sites, specific codons, four base codons for non-natural amino acids.

## **Transposon Based Random Mutagenesis**

Recently many methods for transposon based random mutagenesis have been reported. This methods include, but are not limited to the following: PERMUTE-Random Circular Permutation, random protein truncation, random nucleotide triplet substitution, random domain/tag/multiple amino acid insertion, codon scanning mutagenesis, and multicodon scanning mutagenesis. These aforementioned techniques all require the design of mini-Mu transposons. Thermo scientific manufactures kits for the design of these transposons.

# **Random Mutagenesis Methods Altering the Target DNA LengtH**

These methods involve altering gene length via insertion and deletion mutations. An example is the Tandem Repeat Inserstion (TRINS) method. This technique results in the generation of tandem repeats of random fragments of the target gene via rolling circle amplification and concurrent incorporation of these repeats into the target gene.<sup>[2][page needed]</sup>

## **Mutator strains**

Mutator strains are bacterial cell lines which are deficient in one or more DNA repair mechanisms. An example of a mutator strand is the E. coli XL1-RED. This subordinate strain of E. coli is deficient in the MutS, MutD, MutT DNA repair pathways. Use of mutator strains is useful at introducing many types of mutation; however, these strains show progressive sickness of culture because of the accumulation of mutations in the strains own genome.

## **Focused mutagenesis**

Focused mutagenic methods produce mutations at predetermined amino acid residues. These techniques require and understanding of the sequence-function relationship for the protein of interest. Understanding of this relationship allows for the identification of residues which are important in stability, stereoselectivity, and catalytic efficiency. Examples of methods that produce focused mutagenesis are below.

## Site saturation mutagenesis

Site saturation mutagenesis is a PCR based method used to target amino acids with significant roles in protein function. The two most common techniques for performing this are whole plasmid single PCR, and overlap extension PCR.

Whole plasmid single PCR is also referred to as site directed mutagenesis (SDM). SDM products are subjected to Dpn endonuclease digestion. This digestion results in cleavage of only the parental strand, because the parental strand contains a GmATC which is methylated at N6 of adenine. SDM does not work well for large plasmids of over ten kilobases. Also, this method is only capable of replacing two nucleotides at a time.

Overlap extension PCR requires the use of two pairs of primers. One primer in each set contains a mutation. A first round of PCR using these primer sets is performed and two double stranded DNA duplexes are formed. A second round of PCR is then performed in which these duplexes are denatured and annealed with the primer sets again to produce heteroduplexes, in which each strand has a mutation. Any gaps in these newly formed heteroduplexes are filled with DNA polymerases and further amplified.

#### Sequence saturation mutagenesis (SeSaM)

This technique results in the randomization of the target sequence at every nucleotide position. This method begins with the generation of variable length DNA fragments tailed with universal bases via the use of template transferases at the 3' termini. Next, these fragments are extended to full length using a single stranded template. The universal bases are replaced with a random standard base, causing mutations. There are several modified versions of this method such as SeSAM-Tv-II, SeSAM-Tv+, and SeSAM-III.

## Single primer reactions in parallel (SPRINP)

This site saturation mutagenesis method involves two separate PCR reaction. The first of which uses only forward primers, while the second reaction uses only reverse primers. This avoids the formation of primer dimer formation.

## Mega Primed and Ligase Free Focused Mutagenesis

This site saturation mutagenic technique begins with one mutagenic oligonucleotide and one universal flanking primer. These two reactants are used for an initial PCR cycle. Products from this first PCR cycle are used as mega primers for the next PCR.

## Ω-PCR

This site saturation mutagenic method is based on overlap extension PCR. It is used to introduce mutations at any site in a circular plasmid

## PFunkel-Ominchange-OSCARR

This method utilizes user defined site directed mutagenesis at single or multiple sites simultaneously. OSCARR is an acronym for One Pot Simple Methodology for Cassette Randomization and Recombination. This randomization and recombination results in randomization of desired fragments of a protein. Omnichange is a sequence independent, multisite saturation mutagenesis which can saturate up to five independent codons on a gene.

#### **Trimer-Dimer Mutagenesis**

This method removes redundant codons and stop codons.

## **Cassette Mutagenesis**

This is a PCR based method. Cassette mutagenesis begins with the synthesis of a DNA cassette containing the gene of interest, which is flanked on either side by restriction sites. The endonuclease which cleaves these restriction sites also cleaves sites in the target plasmid. The DNA cassette and the target plasmid are both treated with endonucleases to cleave these restriction sites and create sticky ends. Next the products from this cleavage are ligated together, resulting in the insertion of the gene into the target plasmid. An alternative form of cassette mutagenesis called combinatorial cassette mutagenesis is used to identify the

functions of individual amino acid residues in the protein of interest. Recursive ensemble mutagenesis then utilizes information from previous combinatorial cassette mutagenesis. Codon cassette mutagenesis allows you to insert or replace a single codon at a particular site in double stranded DNA.

#### **Sexual Methods**

Sexual methods of directed evolution involve *in vitro* recombination which mimic natural *in vivo* recombination. Generally these techniques require high sequence homology between parental sequences. These techniques are often used to recombine two different parental genes, and these methods do create cross overs between these genes.<sup>[2][page needed]</sup>

## In vitro Homologous RecombinatioN

Homologous recombination can be categorized as either *in vivo* or *in vitro*. *In vitro* homologous recombination mimics natural *in vivo* recombination. These *in vitro* recombination methods require high sequence homology between parental sequences. These techniques exploit the natural diversity in parental genes by recombining them to yield chimeric genes. The resulting chimera show a blend of parental characteristics.

## **DNA shuffling**

This *in vitro* technique was one of the first techniques in the era of recombination. It begins with the digestion of homologous parental genes into small fragments by DNase1. These small fragments are then purified from undigested parental genes. Purified fragments are then reassembled using primer-less PCR. This PCR involves homologous fragments from different parental genes priming for each other, resulting in chimeric DNA. The chimeric DNA of parental size is then amplified using end terminal primers in regular PCR.

#### Random priming In vitro recombination (RPR)

This *in vitro* homologous recombination method begins with the synthesis of many short gene fragments exhibiting point mutations using random sequence primers. These fragments are reassembled to full length parental genes using primer-less PCR. These reassembled sequences are then amplified using PCR and subjected to further selection processes. This method is advantageous relative to DNA shuffling because there is no use of DNase1, thus there is no bias for recombination next to a pyrimidine nucleotide. This method is also advantageous due to its use of synthetic random primers which are uniform in length, and lack biases. Finally this method is independent of the length of DNA template sequence, and requires a small amount of parental DNA.

## **Truncated Metagenomic Gene-Specific PCR**

This method generates chimeric genes directly from metagenomic samples. It begins with isolation of the desired gene by functional screening from metagenomic DNA sample. Next, specific primers are designed and used to amplify the homologous genes from different environmental samples. Finally, chimeric libraries are generated to retrieve the desired functional clones by shuffling these amplified homologous genes.

## Staggered Extension Process (StEP)

This *in vitro* method is based on template switching to generate chimeric genes. This PCR based method begins with an initial denaturation of the template, followed by annealing of primers and a short extension time. All subsequent cycle generate annealing between the short fragments generated in previous cycles and different parts of the template. These short fragments and the templates anneal together based on sequence complementarity. This process of fragments annealing template DNA is known as template switching. These annealed fragments will then serve as primers for further extension. This method is carried out until the parental length chimeric gene sequence is obtained. Execution of this method only requires flanking primers to begin. There is also no need for Dnase1 enzyme.

## **Random Chimeragenesis on Transient Templates (RACHITT)**

This method has been shown to generate chimeric gene libraries with an average of 14 crossovers per chimeric gene. It begins by aligning fragments from a parental top strand onto the bottom strand of a uracil containing template from a homologous gene. 5' and 3' overhang flaps are cleaved and gaps are filled by the exonuclease and endonuclease activities of Pfu and taq DNA polymerases. The uracil containing template is then removed from the heteroduplex by treatment with a uracil DNA glcosylase, followed by further amplification using PCR. This method is advantageous because it generates chimeras with relatively high crossover frequency. However it is somewhat limited due to the complexity and the need for generation of single stranded DNA and uracil containing single stranded template DNA.

## Synthetic shuffling

Shuffling of synthetic degenerate oligonucleotides adds flexibility to shuffling methods, since oligonucleotides containing optimal codons and beneficial mutations can be included.

## In vivo Homologous Recombination

## **Cloning Performed in Yeast**

This method involves PCR dependent reassembly of fragmented expression vectors. These reassembled vectors are then introduced to, and cloned in yeast. Using yeast to clone the

vector avoids toxicity and counter-selection that would be introduced by ligation and propagation in E. coli.

# Mutagenic Organized Recombination Process by Homologous *In Vivo* Grouping (MORPHING)

This method introduces mutations into specific regions of genes while leaving other parts intact by utilizing the high frequency of homologous recombination in yeast.

## Phage Assisted Continuous Evolution (PACE)

This method utilizes a bacteriophage with a modified life cycle to transfer evolving genes from host to host. The phage's life cycle is designed in such a way that the transfer is correlated with the activity of interest from the enzyme. This method is advantageous because it requires minimal human intervention for the continuous evolution of the gene.

## In Vitro Non-Homologous Recombination Methods

These methods are based upon the fact that proteins can exhibit similar structural identity while lacking sequence homology.

## **Exon Shuffling**

Exon shuffling is the combination of exons from different proteins by recombination events occurring at introns. Orthologous exon shuffling involves combining exons from orthologous genes from different species. Orthologous domain shuffling involves shuffling of entire protein domains from orthologous genes from different species. Paralogous exon shuffling involves shuffling of exon from different genes from the same species. Paralogous domain shuffling involves shuffling of entire protein domains from paralogous proteins from the same species. Functional homolog shuffling involves shuffling of non-homologous domains which are functional related. All of these processes being with amplification of the desired exons from different genes using chimeric synthetic oligonucleotides. This amplification products are then reassembled into full length genes using primer-less PCR. During these PCR cycles the fragments act as templates and primers. This results in chimeric full length genes, which are then subjected to screening.

## Incremental Truncation for the Creation of Hybrid Enzymes (ITHCY)

Fragments of parental genes are created using controlled digestion by exonuclease III. These fragments are blunted using endonuclease, and are ligated to produce hybrid genes. THIOITHCY is a modified ITHCY technique which utilized nucleotide triphosphate analogs such as  $\alpha$ -phosphothioate dNTPs. Incorporation of these nucleotides blocks digestion by exonuclease III. This inhibition of digestion by exonuclease III is called spiking. Spiking can

be accomplished by first truncating genes with exonuclease to create fragments with short single stranded overhangs. These fragments then serve as templates for amplification by DNA polymerase in the presence of small amounts of phosphothioate dNTPs. These resulting fragments are then ligated together to form full length genes. Alternatively the intact parental genes can be amplified by PCR in the presence of normal dNTPs and phosphothioate dNTPs. These full length amplification products are then subjected to digestion by an exonuclease. Digestion will continue until the exonuclease encounters an  $\alpha$ -pdNTP, resulting in fragments of different length. These fragments are then ligated together to generate chimeric genes.

## SCRATCHY

This method generates libraries of hybrid genes inhibiting multiple crossovers by combining DNA shuffling and ITHCY. This method begins with the construction of two independent ITCHY libraries. The first with gene A on the N-terminus. And the other having gene B on the N-terminus. These hybrid gene fragments are separated using either restriction enzyme digestion or PCR with terminus primers via agarose gel electrophoresis. These isolated fragments are then mixed together and further digested using DNase1. Digested fragments are then reassembled by primerless PCR with template switching.

## **Recombined Extension on Truncated Templates (RETT)**

This method generates libraries of hybrid genes by template switching of uni-directionally growing polynucleotides in the presence of single stranded DNA fragments as templates for chimeras. This method begins with the preparation of single stranded DNA fragments by reverse transcription from target mRNA. Gene specific primers are then annealed to the single stranded DNA. These genes are then extended during a PCR cycle. This cycle is followed by template switching and annealing of the short fragments obtained from the earlier primer extension to other single stranded DNA fragments. This process is repeated until full length single stranded DNA is obtained.

## Sequence Homology-Independent Protein Recombination (SHIPREC)

This method generates recombination between genes with little to no sequence homology. These chimeras are fused via a linker sequence containing several restriction sites. This construct is then digested using DNase1. Fragments are made are made blunt ended using S1 nuclease. These blunt end fragments are put together into a circular sequence by ligation. This circular construct is then linearized using restriction enzymes for which the restriction sites are present in the linker region. This results in a library of chimeric genes in which contribution of genes to 5' and 3' end will be reversed as compared to the starting construct.

#### Sequence Independent Site Directed Chimeragenesis (SISDC)

This method results in a library of genes with multiple crossovers from several parental genes. This method does not require sequence identity among the parental genes. This does require one or two conserved amino acids at every crossover position. It begins with alignment of parental sequences and identification of consensus regions which serve as crossover sites. This is followed by the incorporation of specific tags containing restriction sites followed by the removal of the tags by digestion with Bac1, resulting in genes with cohesive ends. These gene fragments are mixed and ligated in an appropriate order to form chimeric libraries.

## **Degenerate Homo-Duplex Recombination (DHR)**

This method begins with alignment of homologous genes, followed by identification of regions of polymorphism. Next the top strand of the gene is divided into small degenerate oligonucleotides. The bottom strand is also digested into oligonucleotides to serve as scaffolds. These fragments are combined in solution are top strand oligonucleotides are assembled onto bottom strand oligonucleotides. Gaps between these fragments are filled with polymerase and ligated.<sup>[2][page needed]</sup>

## Random Multi-Recombinant PCR (RM-PCR)

This method involves the shuffling of plural DNA fragments without homology, in a single PCR. This results in the reconstruction of complete proteins by assembly of modules encoding different structural units.

## User Friendly DNA Recombination (USERec)

This method begins with the amplification of gene fragments which need to be recombined, using uracil dNTPs. This amplification solution also contains primers, PfuTurbo, and Cx Hotstart DNA polymerase. Amplified products are next incubated with USER enzyme. This enzyme catalyzes the removal of uracil residues from DNA creating single base pair gaps. The USER enzyme treated fragments are mixed and ligated using T4 DNA ligase and subjected to Dpn1 digestion to remove the template DNA. These resulting dingle stranded fragments are subjected to amplification using PCR, and are transformed into E. coli.

## Golden Gate Shuffling (GGS) Recombination

This method allows you to recombine at least 9 different fragments in an acceptor vector by using type 2 restriction enzyme which cuts outside of the restriction sites. It begins with sub cloning of fragments in separate vectors to create Bsa1 flanking sequences on both sides. These vectors are then cleaved using type II restriction enzyme Bsa1, which generates four nucleotide single strand overhangs. Fragments with complementary overhangs are hybridized

and ligated using T4 DNA ligase. Finally these constructs are then transformed into E. coli cells, which are screened for expression levels.

#### **Phosphoro Thioate-Based DNA Recombination Method (PRTec)**

This method can be used to recombine structural elements or entire protein domains. This method is based on phosphorothioate chemistry which allows the specific cleavage of phosphorothiodiester bonds. The first step in the process begins with amplification of fragments that need to be recombined along with the vector backbone. This amplification is accomplished using primers with phosphorothiolated nucleotides at 5' ends. Amplified PCR products are cleaved in an ethanol-iodine solution at high temperatures. Next these fragments are hybridized at room temperature and transformed into E. coli which repair any nicks.

## Integron

This system is based upon a natural site specific recombination system in E. coli. This system is called the integron system, and produces natural gene shuffling. This method was used to construct and optimize a functional tryptophan biosynthetic operon in trp-deficient E. coli by delivering individual recombination cassettes or trpA-E genes along with regulatory elements with the integron system.

#### **Y-Ligation Based Shuffling (YLBS)**

This method generates single stranded DNA strands, which encompass a single block sequence either at the 5' or 3' end, complementary sequences in a stem loop region, and a D branch region serving as a primer binding site for PCR. Equivalent amounts of both 5' and 3' half strands are mixed and formed a hybrid due to the complementarity in the stem region. Hybrids with free phosphorylated 5' end in 3' half strands are then ligated with free 3' ends in 5' half strands using T4 DNA ligase in the presence of 0.1 mM ATP. Ligated products are then amplified by two types of PCR to generate pre 5' half and pre 3' half PCR products. These PCR product are converted to single strands via avidin-biotin binding to the 5' end of the primes containing stem sequences that were biotin labeled. Next, biotinylated 5' half strands and non-biotinylated 3' half strands are used as 5' and 3' half strands for the next Y-ligation cycle.

#### Semi-Rational Design

Semi-rational design uses information about a proteins sequence, structure and function, in tandem with predictive algorithms. Together these are used to identify target amino acid residues which are most likely to influence protein function. Mutations of these key amino

acid residues create libraries of mutant proteins that are more likely to have enhanced properties.

Advances in semi-rational enzyme engineering and de novo enzyme design provide researchers with powerful and effective new strategies to manipulate biocatalysts. Integration of sequence and structure based approaches in library design has proven to be a great guide for enzyme redesign. Generally, current computational de novo and redesign methods do not compare to evolved variants in catalytic performance. Although experimental optimization may be produced using directed evolution, further improvements in the accuracy of structure predictions and greater catalytic ability will be achieved with improvements in design algorithms. Further functional enhancements may be included in future simulations by integrating protein dynamics

Biochemical and biophysical studies, along with fine-tuning of predictive frameworks will be useful to experimentally evaluate the functional significance of individual design features. Better understanding of these functional contributions will then give feedback for the improvement of future designs.

Directed evolution will likely not be replaced as the method of choice for protein engineering, although computational protein design has fundamentally changed the way protein engineering can manipulate bio-macromolecules. Smaller, more focused and functionally-rich libraries may be generated by using in methods which incorporate predictive frameworks for hypothesis-driven protein engineering. New design strategies and technical advances have begun a departure from traditional protocols, such as directed evolution, which represents the most effective strategy for identifying top-performing candidates in focused libraries. Whole-gene library synthesis is replacing shuffling and mutagenesis protocols for library preparation. Also highly specific low throughput screening assays are increasingly applied in place of monumental screening and selection efforts of millions of candidates. Together, these developments are poised to take protein engineering beyond directed evolution and towards practical, more efficient strategies for tailoring biocatalysts.

Screening and Selection Techniques

Once a protein has undergone directed evolution, ration design or semi-ration design, the libraries of mutant proteins must be screened to determine which mutants show enhanced properties. Phage display methods are one option for screening proteins. This method involves the fusion of genes encoding the variant polypeptides with phage coat protein genes. Protein variants expressed on phage surfaces are selected by binding with immobilized

targets in vitro. Phages with selected protein variants are then amplified in bacteria, followed by the identification of positive clones by enzyme linked immunosorbent assay. These selected phages are then subjected to DNA sequencing.

Cell surface display systems can also be utilized to screen mutant polypeptide libraries. The library mutant genes ate incorporated into expression vectors which are then transformed into appropriate host cells. These host cells are subjected to further high throughput screening methods to identify the cells with desired phenotypes.

Cell free display systems have been developed to exploit *in vitro* protein translation or cell free translation. These methods include mRNA display, ribosome display, covalent and non covalent DNA display, and *in vitro* compartmentalization.

## **Enzyme engineering**

Enzyme engineering is the application of modifying an enzyme's structure (and, thus, its function) or modifying the catalytic activity of isolated enzymes to produce new metabolites, to allow new (catalyzed) pathways for reactions to occur,<sup>[6]</sup> or to convert from some certain compounds into others (biotransformation). These products are useful as chemicals, pharmaceuticals, fuel, food, or agricultural additives.

An *enzyme reactor* <sup>[7]</sup> consists of a vessel containing a reactional medium that is used to perform a desired conversion by enzymatic means. Enzymes used in this process are free in the solution.

## **ANTIBODY ENGINEERING**

Natural antibodies consist of an antigen binding site joined to an effector region that is responsible for activating complement and or binding to immune cells. From a biotechnological viewpoint, the incredibly high specificity with which antibodies bind to a target protein is useful for a variety of purposes. Consequently, antibody engineering uses the antigen binding region of the antibody. These are manipulated and are attached to other molecular fragments.

To separate an antigen binding site from the rest of the antibody, gene segments encoding portions of antibody chains are subcloned and expressed in bacterial cells. Bacterial signal sequences are added to the N terminus of the partial antibody chains, which results in export of the chains into the periplasmic space. Here the VH and VL domains fold up correctly and form their disulfide bonds. The antibody fragments used include Fab, Fv, and **single-chain Fv (scFv)** (Fig. 6.12). In a Fab fragment, an interchain disulfide bond holds the two chains together. However, the Fv fragment lacks this region of the antibody chains and thus is less stable. This led to development of the single-chain Fv fragment in which the VH and VL domains are linked together by a short peptide chain, usually 15 to 20 amino acids long. This is introduced at the genetic level so that a single artificial gene expresses the whole structure (VH-linker-VL or VL-linker-VH). A tag sequence (such as a His6-tag or FLAG-tag) is often added to the end to allow detection and purification. Such an scFv fragment is quite small, about 25,000 in molecular weight.

Such scFv fragments are attached to various other molecules by genetic engineering. The role of the scFv fragment is to recognize some target molecule, perhaps a protein expressed only on the surface of a virus-infected cell or a cancer cell. A variety of toxins, cytokines, or enzymes may be attached to the other end of the scFv fragment, to provide the active portion of the final recombinant antibody. In principle, this approach provides a way of delivering a therapeutic agent in an extremely specific manner. At present the clinical applications of engineered antibodies are under experimental investigation.



Fab fragments are produced by protease digestion of the hinge region. A disulfide bond holds the heavy and light chains together. To make an antibody fragment without any constant region, the genes for the VH domain and the VL domain are expressed on a bacterial plasmid. This structure is unstable because of a lack of disulfide bonds. Therefore, disulfide bonds are engineered into the two halves (dsFv fragment), or a linker is added to hold the VH and VL domains together (scFv fragment).

The antigen binding regions used in antibody engineering may be derived from characterized monoclonal antibodies. Alternatively, a library of DNA segments encoding V-regions may be obtained from a pool of B cells obtained from an animal or human blood sample. Such a library should in theory contain V-regions capable of recognizing any target molecule. Using a human source avoids the necessity for the complex humanization procedures described earlier. However, in this case it is necessary to screen the V-region library for an antibody fragment that binds to the desired target molecule. This may be done by the phage display procedure outlined. The library of V-region constructs is expressed on the surface of the phage, and the target molecule is attached to some solid support and used to screen out those phages carrying the required antibody V region.

## DIABODIES AND BISPECIFIC ANTIBODY CONSTRUCTS

A variety of engineered antibody constructs are presently being investigated. A **diabody** consists of two single-chain Fv (scFv) fragments assembled together. Shortening the linker from 15 amino acids to five drives dimerization of two scFv chains. This no longer allows intrachain assembly of the linked VH and VL regions. The dimer consists of two scFv fragments arranged in a crisscross manner (Fig. 6.13). The resulting diabody has two antigen binding sites pointing in opposite directions. If two different scFv fragments are used, the result is a bispecific diabody that will bind to two different target proteins simultaneously. Note that formation of such a bispecific diabody requires that VH-A be linked to VL-B and VH-B to VL-A. It is of course possible to engineer both sets of VH and VL regions onto a single polypeptide chain encoded by a single recombinant gene, as shown in Fig. 6.13. Bispecific diabodies have a variety of potential uses in therapy, because they may be used to bring together any two other molecules; for example, they might be used to target toxins to cancer cells.





#### FIGURE 6.13 Engineered Diabody Constructs

(A) Engineering a diabody construct begins by genetically fusing the variable domains of the heavy and light chain (VH and VL) with a linker. The long linker allows a single polypeptide to form into a single antibody binding domain. The short linker allows two polypeptides to complex into a diabody with two antibody binding domains. The construct is expressed in bacteria using a bacterial promoter and RBS (ribosome binding site). The signal sequence tells the bacteria to secrete the engineered protein. (B) Instead of identical Fv units, two different Fv chains can be coexpressed in the bacterial cell. The two different Fv chains will unite into a diabody with two different antibody binding domains, a different one on each side. (C) Bispecific antibodies can be made as one single transcript with a linker between VHA and VLB, a linker between the two halves, and finally a linker between VHB and VLA.

Another way to construct an engineered bispecific antibody is to connect the two different scFv fragments to other proteins that bind together (Fig. 6.14). Two popular choices are streptavidin and leucine zippers. Streptavidin is a small biotin binding protein from the bacterium Streptococcus. It forms tetramers, so it allows up to four antibody fragments to be assembled together. Furthermore, binding to a biotin column can purify the final constructs. Leucine zipper regions are used by many transcription factors that form dimers. Often, such proteins form mixed dimers when their leucine zippers recognize each other and bind together. Leucine zipper regions from two different transcription factors that associate (e.g., the Fos and Jun proteins) may therefore be used to assemble two different scFv fragments.

## **PROTEIN ENGINEERING**



#### FIGURE 6.14

#### Engineered Bispecific Antibody Constructs

Instead of genetic linkers to hold diabodies, various proteins can also hold scFv fragments together. Proteins with a leucine zipper domain dimerize; therefore, when scFv genes are genetically fused to these, the scFv domains come together as dimers. Proteins such as streptavidin or proteins with four helix bundle domains can be genetically fused to scFv domains. When expressed, there are four scFv domains on the outside, providing four different antibody binding sites.

#### Examples of engineered proteins

Computing methods have been used to design a protein with a novel fold, named Top7,<sup>[8]</sup> and sensors for unnatural molecules.<sup>[9]</sup> The engineering of fusion proteinshas yielded rilonacept, a pharmaceutical that has secured Food and Drug Administration (FDA) approval for treating cryopyrin-associated periodic syndrome.

Another computing method, IPRO, successfully engineered the switching of cofactor specificity of Candida boidinii xylose reductase.<sup>[10]</sup> Iterative Protein Redesign and Optimization (IPRO) redesigns proteins to increase or give specificity to native or novel substrates and cofactors. This is done by repeatedly randomly perturbing the structure of the proteins around specified design positions, identifying the lowest energy combination of rotamers, and determining whether the new design has a lower binding energy than prior ones.<sup>[11]</sup>

Computation-aided design has also been used to engineer complex properties of a highly ordered nano-protein assembly.<sup>[12]</sup> A protein cage, E. coli bacterioferritin (EcBfr), which naturally shows structural instability and an incomplete self-assembly behavior by populating two oligomerization states, is the model protein in this study. Through computational analysis

## **B.Tech Biotechnology**

and comparison to its homologs, it has been found that this protein has a smaller-thanaverage dimeric interface on its two-fold symmetry axis due mainly to the existence of an interfacial water pocket centered on two water-bridged asparagine residues. To investigate the possibility of engineering EcBfr for modified structural stability, a semi-empirical computational method is used to virtually explore the energy differences of the 480 possible mutants at the dimeric interface relative to the wild type EcBfr. This computational study also the water-bridged asparagines. Replacing these two converges on asparagines with hydrophobic amino acids results in proteins that fold into alpha-helical monomers and assemble into cages as evidenced by circular dichroism and transmission electron microscopy. Both thermal and chemical denaturation confirm that, all redesigned proteins, in agreement with the calculations, possess increased stability. One of the three mutations shifts the population in favor of the higher order oligomerization state in solution as shown by both size exclusion chromatography and native gel electrophoresis