PROTEIN DESIGN AND 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 oldingand recognition for protein design principles. There are two general strategies for protein engineering, 'rational' protein design and directed evolution. These techniques are not mutually exclusive; researchers will often apply both. In the future, more detailed knowledge of protein structure and function, as well as advancements in high-throughput technology, may greatly expand the capabilities of protein engineering. Eventually, even unnatural amino acids may be incorporated, thanks to a new method that allows the inclusion of novel amino acids in the genetic code

Rational design

In rational protein design, the scientist uses detailed knowledge of the structure and function of the protein to make desired changes. In general, this has the advantage of being inexpensive and technically easy, since site-directed mutagenesis techniques are well-developed. However, its major drawback is that detailed structural knowledge of a protein is often unavailable, and, even when it is available, it can be extremely difficult to predict the effects of various mutations. Computational protein design algorithms seek to identify novel amino acid sequences that are low in energy when folded to the pre-specified target structure. While the sequence-conformation space that needs to be searched is large, the most challenging requirement for computational protein design is a fast, yet accurate, energy function that can distinguish optimal sequences from similar suboptimal ones.

Directed evolution

In directed evolution, random mutagenesis is applied to a protein, and a selection regime is used to pick out variants that have the desired qualities. Further rounds of mutation and selection are then applied. This method mimics natural evolution and, in general, produces superior results to rational design. An additional technique known as DNA shuffling mixes and matches pieces of successful variants in order to produce better results. This process mimics the recombination that occurs naturally during sexual reproduction. The advantage of directed evolution is 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 that effect. The drawback is that they require high-throughput, which is not feasible for all proteins. Large amounts of recombinant DNA must be mutated and the products screened for desired qualities. The sheer number of variants often requires expensive robotic equipment to automate the process. Furthermore, not all desired activities can be easily screened for.

Examples of engineered proteins

Using computational methods, a protein with a novel fold has been designed, known as Top7,as well as sensors for unnatural molecules. The engineering of fusion proteinshas yielded rilonacept, a pharmaceutical that has secured FDA approval for the treatment of cryopyrin-associated periodic syndrome.

Another computational method, IPRO, successfully engineered the switching of cofactor specificity of Candida boidinii xylose reductase.[3] 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 previous ones.

Computation-aided design has also been used to engineer complex properties of a highly ordered nano-protein assembly. 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 and comparison to its homologs, it has been found that this protein has a smaller-than-average dimeric interface on its two-fold symmetry axis due mainly to the existence of an interfacial water pocket centered around 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 converges on the water-bridged asparagines. Replacing these two 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.

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, or to convert from some certain compounds into others (biotransformation). These products will be useful as chemicals, pharmaceuticals, fuel, food, or agricultural additives. An *enzyme reactor* 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.

PROTEIN SPLICING

Protein splicing is an intramolecular reaction of a particular protein in which an internal protein segment (called an intein) is removed from a precursor protein with a ligation of C-terminal and N-terminal external proteins (called exteins) on both sides. The splicing junction of the precursor protein is mainly a cysteine or a serine, which are amino acidscontaining a nucleophilic side chain. The protein

splicing reactions which are known now do not require exogenous cofactors or energy sources such as adenosine triphosphate(ATP) or guanosine triphosphate (GTP). Normally, **splicing** is associated only with pre-mRNA splicing.

Protein splicing was unanticipated and discovered by two groups (Anraku and Stevens) in 1990. They both discovered a *Saccharomyces cerevisiae* VMA1 in a precursor of a vacuolar H^+-ATPase enzyme. The amino acid sequence of the N- and C-termini corresponded to 70% DNA sequence of that of a vacuolar H^+-ATPase from other organisms, while the amino acid sequence of the central position corresponded to 30% of the total DNA sequence of the yeast HO nuclease.

GENERAL ORGANIZATION OF AN INTEIN

Inteins usually vary in size from 134 to 650 amino acid residues, although inteins of 1308 and 1650 residues are also know. Inteins are conventionally divided into two large groups, classical inteins and mini-inteins (Fig. 1). A classical intein consists of two domains, Hint, which catalyses protein splicing, and a central endonuclease domain. In mini-inteins, the central endonuclease domain is replaced by a linker sequence, which lacks catalytic activity. Analysis of most inteins showed that an average intein harbors ten conserved amino acid sequence motifs: A, N2, B, N4, C, D, E, H, F, and G (Fig. 1a). Mini-inteins lack central motifs C, D, E, and H

Motif A is a short N-terminal sequence of 13 residues, of which two (the first and the last ones) are highly conserved, suggesting their immense importance for splicing initiation and completion. Position 1 at the N end of an intein is almost always occupied by Cys and, in extremely rare cases, by Ala, Gln, or Ser. Position 13 is occupied by Gly or, in rare cases, Ala, Lys, Thr, Arg, Tyr, or Asn.

Motif N2 consists of 7 residues, of which Asp5 or Glu5 is highly conserved and is most often preceded by Gly.

Motif B consists of 14 residues. Position 10 is occupied by His in all known inteins. Position 7 is most often occupied by Thr [9]. These two conserved amino acid residues are involved in splicing initiation.

Motif N4 consists of 16 residues, including highly conserved Asp or Glu in position 11. As in motif N2, this residue is usually preceded by Gly10. However, motif N4 is lacking in some inteins (SceVMA, CtrVMA, CeuClpP, and some others) [11]. Motifs A, N2, B, and N4 form the N-terminal splicing domain, whose function is to facilitate disruption of the peptide bond at the N end of the intein [11]. On average, the N-terminal domain is 150–200 residues in size [11]. Most of the above conserved amino acid residues are absolutely essential for N-terminal cleavage. Amino acid substitutions in the corresponding positions often fully abolish the initiation of cleavage and splicing of the precursor protein. Motifs C and E form a basis of

the DOD endonuclease domain [13]. Like known DOD endonucleases, these motifs have sequences of nine and ten residues, which form the centers recognizing double-stranded DNA and are separated by a linker of 90–130 residues. The active center involves conserved Gly residues, which are in positions 3 and 9 of motif C and 4 and 10 of motif E. In addition, the motifs each harbor catalytically active Asn and Lys

Motif D (eight residues) is in the linker between motifs C and E. Substitution of its Lys2 completely abolishes endonuclease activity of the DOD domain in some cases [9]. This indicates that, together with motifs C and E, motif D is involved in the formation of the active endonuclease domain.

Motif H consists of 19 amino acid residues, of which Leu13–Leu14 are rather conserved. These residues are probably involved in the intein–DNA interaction. Motifs C, D, E, and H form the DOD endonuclease domain, which occurs in many, thought not all, inteins. It should be noted that, according to the available experimental data, the DOD domain is unnecessary for protein splicing [8, 9, 11, 12, 14] but ensures intein homing (see below).

Motifs F and G form the C-terminal splicing domain, which is 25–40 residues in size [11]. Motif F consists of 16 residues, half of which are highly conserved (table). Motif G is a short C-terminal sequence of eight residues, of which seven belong to the intein and one is the N-terminal residue of C-extein. Motifs F and G are separated by a small linker, usually consisting of two to five residues. The last amino acid is Asn in most cases (or, extremely rarely, Gln or Asp); the last but one is His. The two last amino acid residues play an important role in hydrolyzing the peptide bond at the C end of the intein [9, 11], while the N-terminal residue of C-extein (hereafter referred to as residue +1) is critical for extein ligation. Position +1 is occupied by Ser, Thr, or Cys in the majority of known C-exteins.

MAIN FEATURES OF INTEINS

Although structurally heterogeneous, all inteins have some features in common. Four main features of the intein sequence are now recognized.

(1) An intein-coding gene has a sequence absent from its homologs of other organisms.

(2) A mature protein differs in size from the product deduced from its coding sequence by more than 100 residues.

(3) A protein has specific motifs A, B, F, and G (the presence or absence of the DOD endonuclease domain is not considered to distinguish an intein).

(4) A protein has four conserved amino acid residues: Ser, Thr, or Cys at the N end of a putative intein; His–Asn or His–Gln at the C end of the intein; and Ser, Thr, or Cys at the N end of the C-extein.

APPLICATIONS.

- Rapid purification of target proteins
- Temperature sensitive control of protein activity by conditionallysplicing inteins.

Solid-phase peptide synthesis

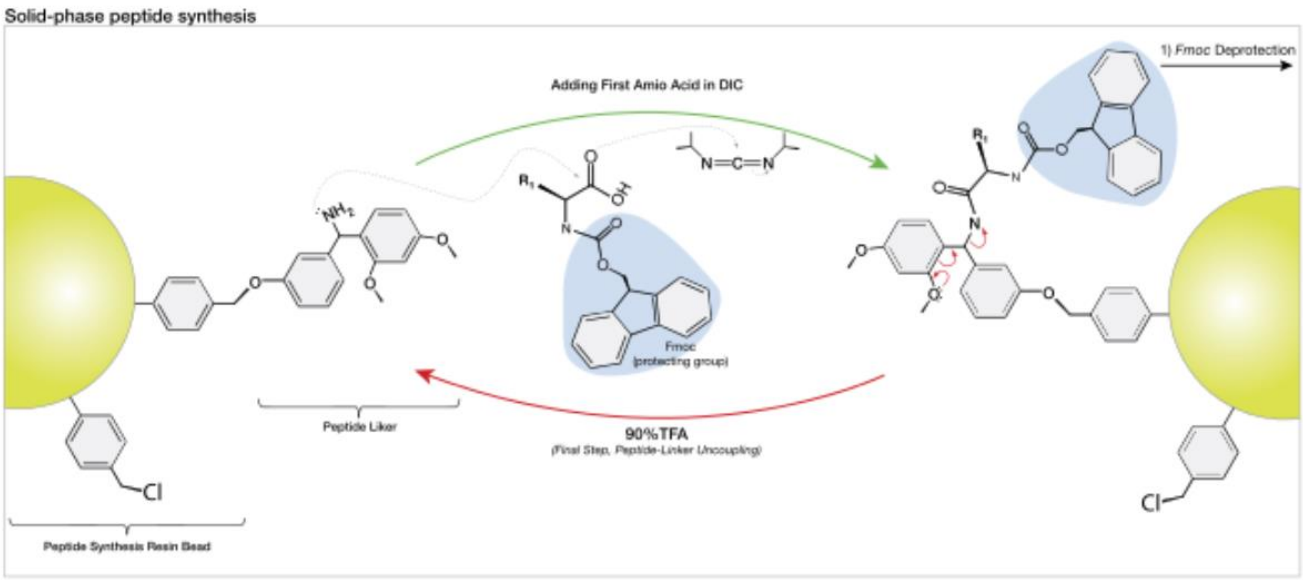
Solid-phase peptide synthesis (SPPS), pioneered by Robert Bruce Merrifield,caused a paradigm shift within the peptide synthesis community, and it is now the standard method for synthesizing peptides and proteins in the lab. SPPS allows for the synthesis of natural peptides which are difficult to express in bacteria, the incorporation of unnatural amino acids, peptide/protein backbone modification, and the synthesis of D-proteins, which consist of D-amino acids.

Small porous beads are treated with functional units ('linkers') on which peptide chains can be built. The peptide will remain covalently attached to the bead until cleaved from it by a reagent such as anhydrous hydrogen fluoride or trifluoroacetic acid. The peptide is thus 'immobilized' on the solid-phase and can be retained during a filtration process while liquid-phase reagents and by-products of synthesis are flushed away.

The general principle of SPPS is one of repeated cycles of deprotection-wash-coupling-wash. The free N-terminal amine of a solid-phase attached peptide is coupled (see below) to a single N-protected amino acid unit. This unit is then deprotected, revealing a new N-terminal amine to which a further amino acid may be attached. The superiority of this technique partially lies in the ability to perform wash cycles after each reaction, removing excess reagent with all of the growing peptide of interest remaining covalently attached to the insoluble resin.

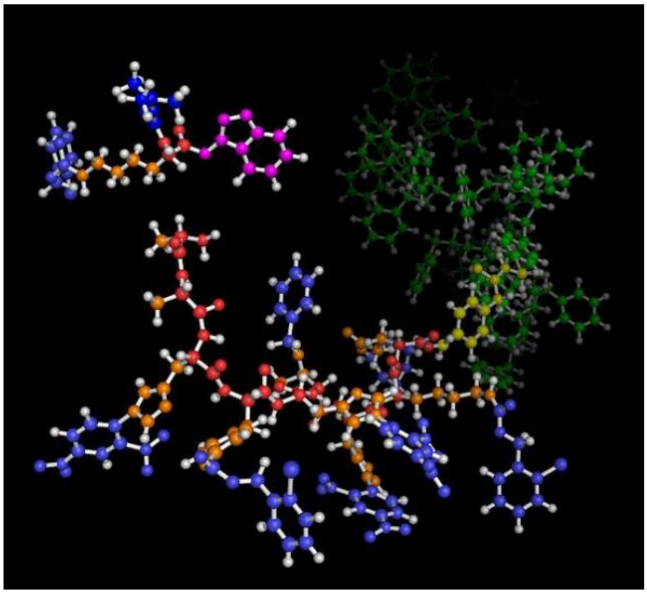
The overwhelmingly important consideration is to generate extremely high yield in each step. For example, if each coupling step were to have 99% yield, a 26-amino acid peptide would be synthesized in 77% final yield (assuming 100% yield in each deprotection); if each step were 95%, it would be synthesized in 25% yield. Thus each amino acid is added in major excess (2~10x) and coupling amino acids together is highly optimized by a series of well-characterized agents.

There are two majorly used forms of SPPS – **Fmoc** and **Boc**. Unlike ribosome protein synthesis, solid-phase peptide synthesis proceeds in a C-terminal to N-terminal fashion. The N-termini of amino acid monomers is protected by either of these two groups and added onto a deprotected amino acid chain.



Automated synthesizers are available for both techniques, though many research groups continue to perform SPPS manually.SPPS is limited by yields, and typically peptides and proteins in the range of 70 amino acids are pushing the limits of synthetic accessibility.Synthetic difficulty also is sequence dependent; typically amyloid peptides and proteins are difficult to make. Longer lengths can be accessed by using native chemical ligation to couple two peptides together with quantitative yields.

Since its introduction over 40 years ago, SPPS has been significantly optimized. First, the resins themselves have been optimized. Furthermore, the 'linkers' between the C-terminal amino acid and polystyrene resin have improved attachment and cleavage to the point of mostly quantitative yields. The evolution of side chain protecting groups has limited the frequency of unwanted side reactions. In addition, the evolution of new activating groups on the carboxyl group of the incoming amino acid have improved coupling and decreased epimerization. Finally, the process itself has been optimized. In Merrifield's initial report, the deprotection of the α-amino group resulted in the formation of a peptide-resin salt, which required neutralization with base prior to coupling. The time between neutralization of the amino group and coupling of the next amino acid allowed for aggregation of peptides, primarily through the formation of secondary structures, and adversely affected coupling. The Kent group showed that concomitant neutralization of the α-amino group and coupling of the next amino acid led to improved coupling. Each of these improvements has helped SPPS become the robust technique that it is today.



Novel Proteins

Pets such as dogs and cats are often fed diets consisting of chicken, beef, lamb and fish. If an intolerance develops, it may be hard to discern exactly what ingredient causes the problem. For these pets, offering a diet consisting of novel proteins such as bison, duck, rabbit, fish they haven't eaten before, venison, kangaroo or egg, either alone or with a single carbohydrate source such as potato or rice, can help resolve the problem.

Use of novel protein

Increasing the production and use of novel proteins is one of our research lines to help building new protein value chains. Durable success and impact of these value chains depend entirely on the ability to use the technical and nutrition functionality of proteins in a final product, and – even more important – on the acceptance of novel proteins by consumers and regulatory bodies.A typical example of a new value chain is the FoodWaste2Feed project. The acceptance of novel proteins by consumers is studied in several programs, focussing on various protein sources and on different target groups.Most novel proteins to be used as food or food ingredient will have to be approved prior to market introduction under the Novel Food

Regulation (Regulation (EC) No 258/97). Wageningen UR developed a Guideline for producers of novel proteins on how to fill the NFR application dossiers (Guideline LEI 14-075). There is no pre-market authorisation procedure for novel feed ingredients derived from non-animal sources, a notification will suffice (Regulation (EC) No 767/2009). For proteins derived from animals, like insect proteins, the situation is rather complex and depends among others on the destination of the feed (food producing animal or not).

Site-directed mutagenesis

Site-directed mutagenesis is a procedure used to induce a specific mutation in a cell. It may be used for a host of reasons, including the generation of restriction sites, investigating the role of a gene or regulatory element by knock-out, understanding the role a particular amino acid has in a protein, or the creation of new and 'better' proteins with, for example, greater thermal stability or more efficient catalytic ability.

The method quite simply involves template DNA - the DNA to be mutated, usually bacterial DNA - and an oligonucleotide carrying the reverse complement of the desired mutation which can anneal to the template and be used as a primer for DNA synthesis. For instance, if a TGG codon is present in the bacterial DNA, and the desired mutation is AAT, then the oligonucleotide primer should read ATT (all read from 5' to 3'). Once the mutagenic primer is annealed to its template, the complete structure is called a *heteroduplex,* owing to the differences between the strands. The heteroduplex is used to transform a cell - most often *E. coli* - and it is left there overnight.

In theory, both strands of the heteroduplex should be replicated at equal frequency to give a 50/50 mixture of mutant to template DNA in the cell. In practice, mutant recovery in this way is poor for two

reasons: firstly, because of the cell's intrinsic mismatch repair system, and secondly, because the template DNA is methylated, and methylated DNA is preferentially replicated by the host cell machinery. Consequently, higher-efficiency mutagenesis approaches have been developed to raise the percentage mutant recovery from around 0.1% to as high as 50%. These approaches work on the principle that once the template DNA has been used to copy the mutant strand it is of no further use, and can only hinder mutant recovery.

Basic mechanism

The basic procedure requires the synthesis of a short DNA primer. This synthetic primer contains the desired mutation and is complementary to the template DNA around the mutation site so it can hybridize with the DNA in the gene of interest. The mutation may be a single base change (a point mutation), multiple base changes, deletion, or insertion. The single-strand primer is then extended using a DNA polymerase, which copies the rest of the gene. The gene thus copied contains the mutated site, and

is then introduced into a host cell as a vector and cloned. Finally, mutants are selected by DNA sequencing to check that they contain the desired mutation.

The original method using single-primer extension was inefficient due to a low yield of mutants. This resulting mixture contains both the original unmutated template as well as the mutant strand, producing a mixed population of mutant and non-mutant progenies. Furthermore the template used is methylated while the mutant strand is unmethylated, and the mutants may be counter-selected due to presence of mismatch repair system that favors the methylated template DNA, resulting in fewer mutants. Many approaches have since been developed to improve the efficiency of mutagenesis.

Quikchange

Quikchange is one high-efficiency mutagenesis approach that has been developed.The plasmid template is denatured and mutagenic primers are annealed to each strand. The primers are de-phosphorylated so that although they can be extended, there cannot be ligation between the end of the synthesised strand and the start of the primer. Once DNA synthesis is complete, the template DNA and mutagenic DNA are denatured in each of the two PCR-like products. The mutant strands cannot be reused for DNA synthesis because when the primers anneal to them, they have no template material to copy. Instead the parental strands are reused: one new mutagenic primer is added to each, DNA is synthesised, the products are denatured and then the parental strands used again. Unlike conventional PCR, which makes products exponentially, this is a linear amplification: two new mutated strands are made in each 'cycle'. At the end of this amplification period, the parental templates are recognised by a methylation-specific restriction enzyme called Dpn I. Because the mutated DNA is unmethylated, it goes unrecognised by this enzyme and remains intact. The consequence is that parental DNA cannot be used to transform *E. coli*, while mutant DNAcan.Although currently the mutant products are all linearised, their lengthy (approx 40 nt) complementary primers can anneal to result in a double-stranded circular plasmid containing a homoduplex of only mutant DNA. The *E. coli* repair system will complete ligation where the

dephosphorylated primers fail to do this, to form complete plasmids readyforhostcellreplication.

AQuikchange protocol might look something like this:

1. Design 2 long (25-40nt) complementary primers containing the mutations
- complementarity is important for circularisation of the mutant plasmid later on
- the melt temperature of the primers should be around 78C
- there is no need for either primer to have a 5' phosphate as there is no ligation step

2. Mix the template plasmid, primers, dNTPs and a thermostable polymerase and run for 16-25 thermal cycles

3. Digest (methylated) DNA with Dpn I

4. Transform *E. coli* with the remaining DNA and leave overnight

5. Pick four colonies and isolate plasmid DNA

6. Sequence the plasmid to ensure that the mutation has been correctly inserted

Uracil-containing DNA method

This approach is based on the simple notion that (deoxy)uracil is not a usual component of DNA. It involves the following protocol:

1. Grow the template DNA to contain a high proportion of deoxyuracil (dU) by growing it in an *E. coli* mutant:
- dut- (which lacks dUTPase; an enzyme which normally prevents the incorporation of uracil into DNA)
- ung- (which lacks uracil glycosylase; an enzyme which normally removes uracil from DNA)

2. Anneal the mutagenic primers, as usual, and begin DNA synthesis

3. The next step can either be performed in vivo or in vitro:

In vivo: transform the heteroduplex DNA into a wild-type *E. coli* which retains its uracil glycosylase function (ung+). The template DNA, which is rich in dU, will then be repaired using the newly-incorporated mutant strand as a template. The product is a homoduplex DNA containing only mutant strands.

In vitro: extract the heteroduplex DNA from *E. coli* and treat with uracil glycosylase to remove the parental DNA. Then synthesise a new strand with DNA polymerase and dNTPs, using the mutant strand as a template. Both of these are performed in the test tube. The product, again, is a homoduplex DNA

containing only mutant strands.

Cassette mutagenesis

Cassette mutagenesis is a technique employed to introduce multiple mutations to the same region of DNA. A cassette (block of DNA) is designed to contain all of the desired mutations and then given ligatable ends to facilitate its insertion into the wild-type DNA. Quikchange, described above, can be used to generate suitable restriction sites for its insertion, and the cassette should have both 5' phosphorylation and 4-base 'sticky' overhangs at each end in order to encourage its insertion into the host molecule.

PCR mutagenesis

PCR mutagenesis is similar in principle to Quikchange. The target plasmid is heated to denature, mutagenic primers (forward and reverse) are added to each strand, and roughly 8 cycles are performed to amplify the mutant plasmid (fewer cycles are ideal to minimise the risk of error). The methylated (parental) DNA is then treated with Dpn I and *E. coli* is transformed with the mutant homoduplexes and left to grow overnight. Again, the plasmid DNA is isolated from selected colonies and sequenced to ensure that the desired mutation has been incorporated.

Sticky feet PCR

Sticky feet PCR is used to generate *insertional* mutations in the wild-type DNA. The mutagenicprimer contains a series of bases (the desired insertion) which is not present in the templateDNA. Because it cannot form complementary pairs with the template DNA upon annealing, the desired insertion 'loops out'. When the primer is extended, to generate heteroduplex DNA, the parental strand is digested, as usual, using Dpn I. This leaves a single-stranded mutant strand, containing the insertion, which then itself acts as a template for new DNA synthesis; the nascent DNA strand will contain the complement of the insertion. The product is homoduplex DNA containing both strands with the insertional mutation.

The size of insertion that can be generated by sticky feet is limited, however, by the size of oligonucleotide primer that can be accurately synthesised (certainly no more than 80 nucleotides in length).Deletions are performed in a similar manner, except the mutagenic primer *lacks* the bases which need to be deleted (it contains only the flanking sequences). Upon annealing, this causes the deletion bases in the *template* DNA to 'loop out' because they have nothing to anneal to. Unlike with insertions, there is no size limitation to deletions because the oligonucleotide only need be big enough to correspond to the flanking regions of the desired site of deletion.

Random mutagenesis

Early approaches to mutagenesis rely on methods which are entirely random in the mutations produced. Cells or organisms may be exposed to mutagens such as UV radiation or mutagenic chemicals, and mutants with desired characteristics are then selected. Hermann Muller discovered that x-rays can cause genetic mutations in fruit flies (published in 1927),and went on to use the *Drosophila* mutants created for his studies on genetics.For *Escherichia coli*, mutants may be selected first by exposure to UV radiation, then plated onto agar medium. The colonies formed are then replica-plated, one in rich medium, another in minimal medium, and mutants that have specific nutritional requirements can then be identified by their inability to grow in minimal medium. Similar procedures may be repeated with other types of cells and with different media for selection.

A number of methods for generating random mutations in specific proteins were later developed to screen for mutants with interesting or improved properties. These methods may involve the use of doped nucleotides in oligonucleotides synthesis, or conducting a PCR reaction in conditions that enhance misincorporation of nucleotides (error-prone PCR), for example by reducing the fidelity of replication or using nucleotide analogues.PCR products which contain mutation are then cloned into an expression vector and the mutant proteins produced can then be characterised.

In animal studies, alkylating agents such as *N*-ethyl-*N*-nitrosourea (ENU) have been used to generate mutant mice.Ethyl methanesulfonate (EMS) is also often used to generate animal and plant mutants.Random mutagenesis is an incredibly powerful tool for altering the properties of enzymes. Imagine, for example, you were studying a G-protein coupled receptor (GPCR) and wanted to create a temperature-sensitive version of the receptor or one that was activated by a different ligand than the wild-type.

1. Error-prone PCR. This approach uses a "sloppy" version of PCR, in which the polymerase has a fairly high error rate (up to 2%), to amplify the wild-type sequence. The PCR can be made error-prone in various ways including increasing the MgCl2 in the reaction, adding MnCl2 or using unequal concentrations of each nucleotide. Here is a good review of error prone PCR techiques and theory. After amplification, the library of mutant coding sequences must be cloned into a suitable plasmid. The drawback of this approach is that size of the library is limited by the efficiency of the cloning step. Although point mutations are the most common types of mutation in error prone PCR, deletions and frameshift mutations are also possible. There are a number of commercial error-prone PCR kits available, including those from Stratagene and Clontech

2. Rolling circle error-prone PCR is a variant of error-prone PCR in which wild-type sequence is first cloned into a plasmid, then the whole plasmid is amplified under error-prone conditions. This eliminates the ligation step that limits library size in conventional error-prone PCR but of course the amplification of the whole plasmid is less efficient than amplifying the coding sequence alone. More details can be found here.

3. Mutator strains. In this approach the wild-type sequence is cloned into a plasmid and transformed into a mutator strain, such as Stratagene's XL1-Red. XL1-red is an *E.coli* strain whosedeficiency in three of the primary DNA repair pathways (*mutS*, *mutD* and *mutT*) causes it to make errors during replicate of it's DNA, including the cloned plasmid. As a result each copy of the plasmid replicated in this strain has the potential to be different from the wild-type. One advantage of mutator strains is that a wide variety of mutations can be incorporated including substitutions, deletions and frame-shifts. The drawback with this method is that

the strain becomes progressively sick as it accumulates more and more mutations in it's own genome so several steps of growth, plasmid isolation, transformation and re-growth are normally required to obtain a meaningful library.

4. Temporary mutator strains. Temporary mutator strains can be built by over-expressing a mutator allele such as *mutD5* (a dominant negative version of *mutD*) which limits the cell's ability to repair DNA lesions. By expressing *mutD5* from an inducible promoter it is possible to allow the cells to cycle between mutagenic (*mutD5* expression on) and normal (*mutD5* expression off) periods of growth. The periods of normal growth allow the cells to recover from the mutagenesis, which allows these strains to grow for longer than conventional mutator strains.If a plasmid with a temperature-sensitive origin of replication is used, the mutagenic plasmid can easily be removed restore normal DNA repair, allowing the mutants to be grown up for analysis/screening. An example of the construction and use of such a strain can be found here. As far as I am aware there are no commercially available temporary mutator strains.

5. Insertion mutagenesis. Finnzymes have a kit that uses a transposon-based system to randomly insert a 15-base pair sequence throughout a sequence of interest, be it an isolated insert or plasmid. This inserts 5 codons into the sequence, allowing any gene with an insertion to be expressed (i.e. no frame-shifts or stop codons are cause). Since the insertion is random, each copy of the sequence will have different insertions, thus creating a library.

6. Ethyl methanesulfonate (EMS) is a chemical mutagen. EMS aklylates guanidine residues, causing them to be incorrectly copied during DNA replication. Since EMS directly chemically modifies DNA, EMS mutagenesis can be carried out either in vivo (i.e. whole-cell mutagenesis) or in vitro.

7. Nitrous acid is another chemical mutagen. It acts by de-aminating adenine and cytosine residues causing transversion point mutations (A/T to G/C and vice versa). **Note***: I have only mentioned two chemical mutagens but there are many others. Hirokazu Inoue has written an excellent article describing some of them and their use in mutagenesis*

8. DNA Shuffling is a very powerful method in which members of a library (i.e. copies of same gene each with different types of mutation) are randomly shuffled. This is done by randomly digesting the library with DNAseI then randomly re-joining the fragments using self-priming PCR. Shuffling can be applied to libraries produced by any of the above method and allows the effects of different combinations of mutations to be tested.

RECOMBINANT PROTEINS EXPRESSING METHODS

Choosing an appropriate method for expressing a recombinant protein is a critical factor in obtaining the desired yields and quality of a recombinant protein in a timely fashion. Selecting a wrong expression host can result in the protein being misfolded or poorly expressed, lacking the necessary posttranslational modifications or containing inappropriate modifications. Factors to consider when selecting an expression system include the mass of the protein and number of disulfide bonds, type of posttranslational modifications desired on the expressed protein, and the destination of the expressed protein. The intended application of the purified recombinant protein is also critical in the decision-making process and the applications can be categorized into four broad areas: structural studies, in vitro activity assays, antigens for antibody generation, and in vivo studies. The purpose of this chapter is to help guide the investigator in the decision-making process for choosing an appropriate expression system. However, even with the described guidelines there are many circumstances when it is not obvious a priori which expression system is the best choice, and the use of multiple expressions systems must be attempted before an optimal system is identified. Numerous expression systems are currently being used in academic and industrial settings. Some of these systems are too new and insufficiently tested to comment on their utility. In addition, some established systems for expressing recombinant proteins, such as transgenic animals, are too technically challenging, time consuming and prohibitively expensive to be a viable option for the average laboratory. For the purpose of this chapter, only Escherichia coli, Pichiapastoris, baculovirus/insect cell, and mammalian expression systems will be considered. These four systems have straightforward protocols, are readily accessible either from colleagues or from research product companies (e.g., Invitrogen, EMD-Novagen, Stratagene, and Promega), and are relatively inexpensive for small-scale production. The characteristics and available options of these expression systems will be briefly reviewed with the focus on the differences between the systems. Strategies will then be presented to help guide the investigator in making the best choice for an expression system.

Escherichia coli

The bacteria E. coli was the first host used to express recombinant proteins and is still considered to be the workhorse in the field. Using the E. coli system offers a rapid and simple method for expressing recombinant proteins due to its short doubling time. Consequently, the assessment of recombinant gene expression in E. coli can take less than a week. The growth media for E. coli are inexpensive and there are

relatively straightforward methods to scale-up bioproduction. In E. coli, recombinant proteins are normally either directed to the cytoplasm or to the periplasm and, to a lesser extent, secreted. Proteins directed to the cytoplasm are the most efficiently expressed, giving yields of up to 30% of the biomass. However, the high expression of recombinant protein can often lead to the accumulation of aggregated, insoluble protein that forms inclusions bodies. Inclusion bodies have been observed not only with eukaryotic proteins but also to a lesser extent with overexpressed proteins from prokaryotes including E. coli. The rate of translation and folding in E. coli is almost 10-fold higher than that observed in eukaryotic cells, and this presumably contributes to the inclusion body formation of eukaryotic proteins. Inclusion bodies can be a significant hindrance in obtaining soluble, active protein in some situations. However, in some cases, inclusion bodies are advantageous because they are resistant to proteolysis, easy to concentrate by centrifugation, minimally contaminated with other proteins, and, with some effort, able to be refolded to form active, soluble proteins.

E. coli: Temperature and molecular chaperones

Several methods have been described for maximizing the formation of soluble, properly folded proteins in the cytoplasm and minimizing inclusion body formation. The most straightforward method involves lowering the temperature to 15–30 C during the expression period. Presumably, the reduced temperature slows the rate of transcription, translation, and refolding, thereby allowing for proper folding. In addition, lower temperature has been shown to decrease heat shock protease activity. Some investigators have coexpressed molecular chaperones in the cytoplasm along with the recombinant protein for promoting protein solubility. The utility of this approach appears to be quite protein-specific, and therefore needs to be tested individually for each recombinant protein of interest.

E. coli: Fusion partners

Alternatively, a method that promotes solubility with many proteins is to fuse the recombinant protein at either the N-terminus or C-terminus to a soluble fusion tag. Fusion partners that have been shown to increase solubility of recombinant proteins include glutathione-S-transferase(GST), thioredoxin, maltose-binding protein (MBP), small ubiquitin-modifier (SUMO), and N-utilization substrate (NusA). Both GST and MBP have the added advantage of also being an affinity purification tag. Unfortunately, no single tag appears to work for all recombinant proteins, and multiple fusion partners may need to be evaluated for promoting soluble expression. Fusion tags can be removed from the recombinant protein by several strategies, and a widespread approach involves adding a protease site between the fusion partner and the recombinant protein that can be cleaved with the specific protease. This approach must be carefully tested since removal of the fusion tag can, in some cases, render the recombinant protein insoluble.

E. coli: Disulfide bond formation

E. coli is normally inefficient in promoting the correct formation of disulfide bonds when recombinant proteins are expressed in the cytoplasm; normally disulfide bond formation occurs only in the periplasm where it is catalyzed by the Dsb system. Consequently, if disulfide bond formation is needed, the recombinant protein can be directed to the periplasm via a cleavable signal peptide (e.g., pelB). However, a major disadvantage of periplasmic expression is the significant reduction in production yields. Through engineering of the E. coli genome, a more suitable environment for disulfide bond generation in the cytoplasm can be induced by disrupting the thioredoxinreductase (trxB) and glutathione reductase (gor)

genes in the Dsb system which in turn enables thioredoxin and glutaredoxin to promote cytoplasmic reduction of cysteines. These engineered strains are commercially available through EMD-Novagen (Origami). If additional disulfide bond formation is still needed, the recombinant protein can be fused to thioredoxin, and the fusion protein expressed in a trxB/gor E. coli strain.

E. coli: Posttranslational modifications

Finally, it is important to recognize that E. coli has a limited capacity for posttranslational modifications compared to eukaryotic organisms. For example, E. coli does not support enzyme-mediated N-linked glycosylation, O-linked glycosylation, amidation, hydroxylation, myristoylation, palmitation, or sulfation.

Reference

1. "Speeding Up the Protein Assembly Line". Genetic Engineering and Biotechnology News. 13 February 2015.
2. Farmer, Tylar Seiya; Bohse, Patrick; Kerr, Dianne (2017). "Rational Design Protein Engineering Through Crowdsourcing". Journal of Student Research. 6 (2): 31–38.
3. Poluri, Krishna Mohan; Gulati, Khushboo (2017). Protein Engineering Techniques. SpringerBriefs in Applied Sciences and Technology. Springer. doi:10.1007/978-981-10-2732-1. ISBN 978-981-10-2731-4.
4. Jäckel, Christian; Kast, Peter; Hilvert, Donald (June 2008). "Protein Design by Directed Evolution". Annual Review of Biophysics. 37 (1): 153–173. doi:10.1146/annurev.biophys.37.032807.125832. PMID 18573077.
5. Shivange, Amol V; Marienhagen, Jan; Mundhada, Hemanshu; Schenk, Alexander; Schwaneberg, Ulrich (2009). "Advances in generating functional diversity for directed protein evolution". Current Opinion in Chemical Biology. 13 (1): 19–25. doi:10.1016/j.cbpa.2009.01.019. PMID 19261539.
6. Lutz, Stefan (December 2010). "Beyond directed evolution—semi-rational protein engineering and design". Current Opinion in Biotechnology. 21 (6): 734–743. doi:10.1016/j.copbio.2010.08.011. PMC 2982887. PMID 20869867.
7. "'Designer Enzymes' Created By Chemists Have Defense And Medical Uses". ScienceDaily. March 20, 2008.
8. [Enzyme reactors at "Archived copy". Archived from the original on 2012-05-02. Retrieved 2013-11-02.] Accessed 22 May 2009.
9. Kuhlman, Brian; Dantas, Gautam; Ireton, Gregory C.; Varani, Gabriele; Stoddard, Barry L. & Baker, David (2003), "Design of a Novel Globular Protein Fold with Atomic-Level Accuracy", Science, 302 (5649): 1364–1368, Bibcode:2003Sci...302.1364K, doi:10.1126/science.1089427, PMID 14631033
10. Looger, Loren L.; Dwyer, Mary A.; Smith, James J. & Hellinga, Homme W. (2003), "Computational design of receptor and sensor proteins with novel functions", Nature, 423 (6936): 185–190, Bibcode:2003Natur.423..185L, doi:10.1038/nature01556, PMID 12736688
11. Khoury, GA; Fazelinia, H; Chin, JW; Pantazes, RJ; Cirino, PC; Maranas, CD (October 2009), "Computational design of Candida boidinii xylose reductase for altered cofactor specificity", Protein Science, 18 (10): 2125–38, doi:10.1002/pro.227, PMC 2786976, PMID 19693930
12. The iterative nature of this process allows IPRO to make additive mutations to a protein sequence that collectively improve the specificity toward desired substrates and/or cofactors. Details on how to download the software, implemented in Python, and experimental testing of predictions are outlined in this paper: Khoury, GA; Fazelinia, H; Chin, JW; Pantazes, RJ; Cirino, PC; Maranas, CD (October 2009), "Computational design of Candida boidinii xylose reductase for altered cofactor specificity", Protein Science, 18 (10): 2125–38, doi:10.1002/pro.227, PMC 2786976, PMID 19693930
13. Ardejani, MS; Li, NX; Orner, BP (April 2011), "Stabilization of a Protein Nanocage through the Plugging of a Protein–Protein Interfacial Water Pocket", Biochemistry, 50 (19): 4029–4037, doi:10.1021/bi200207w, PMID 21488690
14. Chowdhury, Ratul; Ren, Tingwei; Shankla, Manish; Decker, Karl; Grisewood, Matthew; Prabhakar, Jeevan; Baker, Carol; Golbeck, John H.; Aksimentiev, Aleksei; Kumar, Manish; Maranas, Costas D. (10 September 2018). "PoreDesigner for tuning solute selectivity in a robust and highly permeable outer membrane pore". Nature Communications. 9 (1): 3661.

Bibcode:2018NatCo...9.3661C. doi:10.1038/s41467-018-06097-1. PMC 6131167. PMID 30202038.