Biotechnology Advances 60 (2022) 108010

Contents lists available at ScienceDirect

Biotechnology Advances

journal homepage: www.elsevier.com/locate/biotechadv

Research review paper

Insertions and deletions in protein evolution and engineering

Simone Savino, Tom Desmet, Jorick Franceus *

Centre for Synthetic Biology (CSB), Department of Biotechnology, Ghent University, Coupure Links 653, 9000 Ghent, Belgium.

ARTICLE INFO

Keywords:
Indel
Directed evolution
Enzyme engineering
Frameshift mutation
Polymerase slippage
Loop grafting

ABSTRACT

Protein evolution or engineering studies are traditionally focused on amino acid substitutions and the way these contribute to fitness. Meanwhile, the insertion and deletion of amino acids is often overlooked, despite being one of the most common sources of genetic variation. Recent methodological advances and successful engineering stories have demonstrated that the time is ripe for greater emphasis on these mutations and their understudied effects. This review highlights the evolutionary importance and biotechnological relevance of insertions and deletions (indels). We provide a comprehensive overview of approaches that can be employed to include indels in random, (semi)-rational or computational protein engineering pipelines. Furthermore, we discuss the tolerance to indels at the structural level, address how domain indels can link the function of unrelated proteins, and feature studies that illustrate the surprising and intriguing potential of frameshift mutations.

1. Introduction

The intricate process of evolution has resulted in the remarkable diversity of proteins that can be found in nature today. Over the course of billions of years, the continuous stepwise accumulation of seemingly minor changes to amino acid sequences has generated a wealth of proteins that show major differences in shape and function. The evolutionary itinerary between two proteins is often portrayed as a walk through the vast space of possible sequences, with every mutation representing a single step through a fitness landscape (Maynard Smith, 1970; Romero and Arnold, 2009). Some of these mutations are beneficial, leading the protein uphill towards a peak of high fitness to current selective pressures. Others are deleterious, pushing it closer towards a valley of low fitness, where it is discarded by natural selection. But as proteins wander around this mountainous landscape, they may also encounter the base of a new peak that leads to a different kind of fitness. Under appropriate selection conditions, those peaks of functional innovation can be climbed through adaptive mutations, eventually resulting in yet another addition to nature's repertoire of distinct proteins.

The powerful Darwinian cycle of diversification and selection has long been used to develop custom proteins that are tailor-made for specific purposes (Arnold, 2019). Since the 1980s, protein engineers have employed mutagenesis techniques like error-prone PCR to accelerate movement through sequence space. By carefully imposing a suitable artificial selection pressure, adaptive mutations that guide the protein uphill the desired fitness peaks can conveniently be exposed. This strategy of directed evolution has been very successful and many of its achievements have been adopted in industry, from improved detergent proteases to highly active, enantioselective and thermostable catalysts for manufacturing pharmaceuticals (Bornscheuer et al., 2019; Savile et al., 2010; Vojcic et al., 2015). Over the years, considerable advances in (semi-)rational and computational engineering have made it possible to navigate sequence space more intelligently, resulting in smaller libraries with a higher hit rate (Chowdhury and Maranas, 2020; Currin et al., 2015).

Despite enormous progress in the field, not all aspects that are responsible for functional innovation in natural protein evolution have been extensively studied or applied for directed evolution. Most analyses and experiments are centred around amino acid substitutions, whereas insertions and deletions (indels) of amino acids remain overlooked. Yet, indels are a great source of genetic diversification that can have a strong impact on the properties or evolvability of a protein. In the metaphor of a fitness landscape, they tend to represent steep ledges that abruptly

Abbreviations: ASR, ancestral sequence reconstruction; COBARDE, codon-based random deletion method; GRASP, graphical representation of ancestral sequence predictions; indel, insertion/deletion; InScaM, insertional scanning mutagenesis; LILI, linker in loop insertion; ML, machine learning; ORF, open reading frame; RAISE, random insertional-deletional strand exchange; REM, random elongation mutagenesis; RID, random insertion and deletion mutagenesis; RMSD, room mean square deviation; StLois, stepwise loop insertion strategy; TRIAD, transposon-based insertion and deletion mutagenesis; TRINS, tandem repeat insertion method.

* Corresponding author.
E-mail address: jorick.franceus@ugent.be (J. Franceus).

https://doi.org/10.1016/j.biotechadv.2022.108010
Received 5 May 2022; Received in revised form 15 June 2022; Accepted 16 June 2022
Available online 20 June 2022

alter fitness to prevailing selective pressures, for better or for worse. We currently lack a molecular-level understanding of how exactly indels affect protein fitness, let alone the knowledge to predict just how they can be harnessed in the search for novel proteins. For that reason, protein engineers usually avoid traversing the fitness landscape through the treacherous paths shaped by indels. Nevertheless, a degree of success has been achieved in this area over the years, and several appealing methods have been described for incorporating indels in engineering strategies. In this review, we take a closer look at how indels contribute to the navigation through rugged fitness landscapes both in natural and directed protein evolution.

2. Mechanism and frequency of indel mutations

The random insertion or deletion of nucleotides in genomes is commonly caused by a phenomenon called replication slippage, also known as slipped strand mispairing (Sehn, 2015) (Fig. 1). DNA polymerase occasionally pauses and dissociates from the DNA during replication, making it possible for the end of the growing strand to separate from the template and then reanneal to a homologous region located downstream or upstream (Viguera et al., 2001). When the polymerase eventually resumes replication, it will have skipped ahead or backtracked from where it first halted, resulting in a deletion or insertion, respectively. The basic idea of indels originating from the misaligned pairing of two strands was first formulated by George Streisinger in 1966 and the underlying molecular mechanism has largely been elucidated over the past decades (Garcia-Diaz and Kunkel, 2006; Streisinger et al., 1966). However, slippage cannot account for all observed short insertions or deletions, and it has been proposed that non-slippage indels may be caused by a diverse group of mutational mechanisms (Taylor et al., 2004), such as the error-prone non-homologous end joining of double-strand breaks, which can in turn be caused by ionising radiation or replication fork collapse (Gong et al., 2005; Morita et al., 2010).

Indel events can occur throughout the genome, but certain regions are far more susceptible to replication slippage than others. In genetic stutters such as short tandem repeats or homopolymeric runs, the terminal base of a nascent strand can easily anneal in an erroneous way to adjacent homologous regions. In addition, exonucleolytic proofreading efficiency is considerably diminished in repetitive DNA (Bębenek and Ziuzia-Graczyk, 2018; Kroutil et al., 1996). As a result, as much as two-thirds of indels may be associated with such sequence regions in certain genomes (McDonald et al., 2011).

The rate at which small indels are generated has been estimated for various prokaryotic and eukaryotic genomes. It is clear that the absolute frequency of indel events in the genome of an organism is mostly dependent on the overall mutation rate of that organism, which varies over a 1000-fold range across species (Lynch et al., 2016). There appears to be a strong positive correlation between mutation rates for base substitutions and indels, and this correlation holds even when the mismatch repair system is inactivated (Long et al., 2018). The ratio of small indels to base substitutions in bacteria was found to be 0.22 ± 0.04 in the most recent extensive study (Long et al., 2018; Sung et al., 2016) and a similar value was reported earlier for the human genome (Mills et al., 2011). However, not all organisms follow the same pattern. In *Dictyostelium discoideum* and *Plasmodium falciparum*, both eukaryotes with an unusually high AT content and abundance of simple sequence repeats, the indel rate far exceeds the base substitution rate (Hamilton et al., 2017; Kucukyildirim et al., 2020). It is also worth noting that indels of different length occur at different rates. Their size distribution in prokaryotic and eukaryotic genomes obeys a power law, with longer indels being much less represented (Cartwright, 2009; Danneels et al., 2018).

When comparing rates of insertion versus deletion separately, these mutations do not seem to be equally common. A pronounced abundance of deletions over insertions has been detected in bacteria, archaea, amoebae, nematodes, insects, fish and mammals (Gregory, 2004; Guo et al., 2012; Kucukyildirim et al., 2020; Long et al., 2018; Saxena et al., 2019; Taylor et al., 2004). However, this trend is not universal. In a few species, the insertion/deletion ratio was found to be skewed towards insertions instead (Behringer and Hall, 2016; Farlow et al., 2015; Long et al., 2018; Sung et al., 2012). These biases may exist due to DNA repair pathways (Long et al., 2018). Shutting down the mismatch repair system flips the insertion/deletion ratio of bacteria in favour of insertions. Conversely, in the extremophile *Deinococcus radiodurans* which has an unusual insertion bias, inactivating its divergent repair system makes deletions more prevalent instead (Long et al., 2018). Further, the nucleotides flanking the mismatch influence the efficiency of the repair system, which may cause bias differences in organisms where the local or absolute ratio of G/C over A/T is unbalanced.

Unsurprisingly, indels are less abundant in regions of the genome

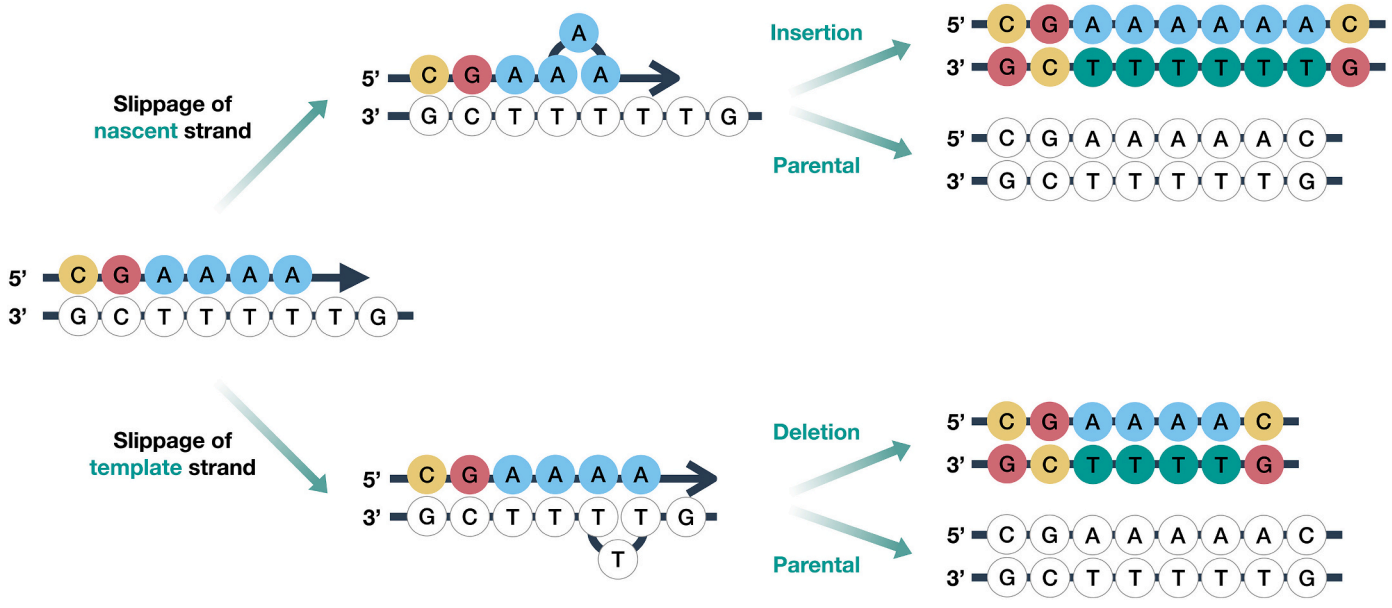

Fig. 1. Strand slippage may occur during DNA replication, causing a misaligned intermediate with one or more unmatched nucleotides. The newly synthesised strand contains an insertion or a deletion when slippage occurs in the nascent or template strand, respectively.

that code for proteins (Chen et al., 2009; Studer et al., 2013). When the number of nucleotides added or deleted is not a multiple of three, the reading frame is shifted and the resulting fitness loss prevents the mutation from accumulating in the population. However, substitution-to-indel ratio has been observed to be up to 100-fold higher in protein coding regions, indicating that the rate of indel purging cannot simply be explained by the need to preserve the reading frame (Tóth-Petróczy and Tawfik, 2013). Purifying selection appears to act more strongly on indels than on substitutions, even when they are in frame.

3. Tolerance to indels

Indels can occur along the whole length of a protein sequence, but they accumulate unevenly. This observation is related to the three-dimensional structure of proteins. Because of their intrinsic features, the hydrophobic core, the polar solvent-exposed surface, loops and structured secondary elements are affected by indels to different degrees of severity (Guo et al., 2012). When such mutations occur in structural elements that poorly tolerate their effects, they are promptly purged by selective pressure. A massive mutational screening in β-lactamase (Gonzalez et al., 2019) showed that most insertions and deletions are highly deleterious, with over half of them lowering the fitness of the enzyme at least 100-fold. Specifically, lower fitness (identified as lower resistance to ampicillin, when compared to wild type) was usually observed with indels occurring in secondary structure elements. However, those located in unstructured regions such as loops were tolerated more frequently. Furthermore, the fitness effect of an insertion depended more on the site of insertion than on the identity of the inserted residue, indicating again that the location of an indel event certainly matters. Overall, the authors found insertions to be better tolerated than deletions.

The tolerance of loops to indels has been widely described for many protein families (Pascarella and Argos, 1992; Simm et al., 2007; Taylor et al., 2004). Other disordered regions prone to accumulation of indels are the N- and C-termini (Lin et al., 2017), where the length of the inserted or removed sequence stretch tends to be longer (Light et al., 2013). In a study by Arpino et al., an interesting phenomenon was observed where a single amino acid deletion at the N-terminus of green fluorescent protein causes a change in the registry (i.e. relative side chain position) of an α-helix (Arpino et al., 2014a). However, the mutation is not only tolerated, but it even increases fluorescence by improving folding efficiency. Similarly, a C-terminal deletion was well tolerated, despite the extensive structural changes that occurred as a consequence of the deletion (Arpino et al., 2014b).

In contrast, indels tend to be dramatically underrepresented in highly structured regions where the preservation of structural integrity is critical, such as transmembrane domains. Transmembrane length and helical registry of side chains within the membrane appear to be constrained in order to ensure a correct localization of the protein in the cell (Taylor et al., 2004). In fact, the properties of transmembrane regions even remain largely conserved upon frameshifting (Bartonek et al., 2020). Another example of structures where indels are less frequently observed are coiled coils, which are domains formed by repetitive peptide motifs that play a structural role in the cell. Unlike other repetitive regions, which are generally susceptible to length changes, the length of coiled coils is quite conserved (Surkont et al., 2015). This is probably caused by their sequence length being directly proportional to the physical size of the domain, which in turn directly affects their biological function as spacers or scaffolds. When α-helices and β-strands are compared, the former structures are found to be more tolerant to indels, presumably because the latter interact strongly to form β-sheets, causing the effect of the indel to propagate through the structure (Kim and Guo, 2010). Most indels that are tolerated in β-strands are located towards the strand termini (Arpino et al., 2014a).

At the folding level, many examples exist of proteins that tolerate in-frame indels remarkably well. Pairs of structures in the Protein Data Bank of homologous proteins with short indels (< 50 bp) display low root mean square deviations (RMSDs) when they are compared (Kim and Guo, 2010). Even in the few cases where large RMSDs are observed, these are caused by a different positioning of domains, rather than actual changes in protein folding. This simple observation supports the idea that a degree of plasticity is regularly retained in the folding paths of alternate protein forms (with and without indels), as well as in protein evolution (before and after indels). However, while the accumulation of indels in folded and soluble proteins barely seems to influence their overall structure, their effect is more noticeable on the interaction between different proteins, or even between monomers forming a homo-oligomer (Sandhya et al., 2009; Studer et al., 2013). Indels occur frequently at binding interfaces and they eventually apply selective pressure even on the evolution of binding partners.

It thus appears that we can zoom in on the fitness landscape that is shaped by indels to find some sort of hierarchy in how abrupt the effect of these mutations can be (Fig. 2). In some regions of sequence space, indels are tolerated generally well and their evolutionary effects on organismal fitness are not too different from those caused by substitutions, at least with respect to the conservation of the native protein function. In other regions, the landscape is riddled with valleys of low fitness.

4. Adaptive indels in natural protein evolution

As proteins accrue random mutations, they may experience gradual or even drastic shifts in their functional properties such as activity, specificity or stability. Under appropriate selective pressures, these mutations ultimately result in the divergence of proteins that are equipped with a new set of traits. The impact of amino acid substitutions on protein function is routinely investigated, and such studies have delivered fascinating and useful insights into important processes like specificity switches (Bridgham et al., 2009), temperature adaptation (Pinney et al., 2021), or even the emergence of enzymatic activity in non-catalytic proteins (Clifton et al., 2018; Kaltenbach et al., 2018). On the other hand, the contribution of indels is still poorly understood. Nevertheless, a few reports have surfaced that demonstrate how indels can clearly play a key role in protein evolution as well.

Before a protein can drastically diverge, a prior duplication event is

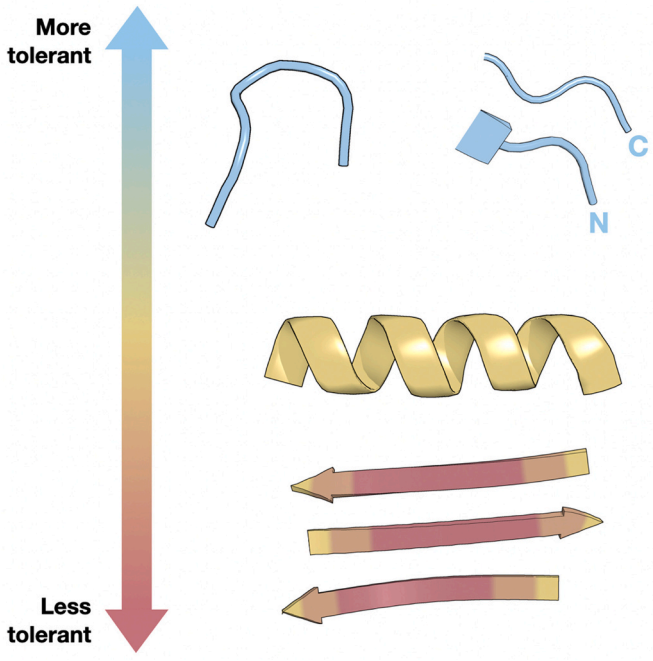

Fig. 2. Tolerance of protein structural elements to insertions and deletions.

often necessary. One copy of the coding gene is then free to serve as a canvas for adaptation, as long as the other maintains its original function to preserve organismal fitness for as long as is necessary (Copley, 2020). This is likely to be even more relevant when indels rather than substitutions are the cause of divergence, due to their common deleterious effects. Gene duplications have indeed been shown to set the stage for adaptive indel events in a study of thousands of duplicate genes in five teleost fish species (Guo et al., 2012). Indels were at least 25% more abundant in duplicated genes than in singletons, and a large proportion of tertiary structure divergence between duplicated genes could be explained by indel density, but not by amino acid substitutions. Interestingly, the higher indel rate could in most cases be observed symmetrically in both members of the gene pair, implying that both copies experience relaxed purifying selection after duplication. The data also once again confirmed that short indels are more prevalent than long ones, that they predominantly occur in loop regions, and that a clear deletion bias exists.

A few natural protein adaptations that increase the odds of survival and/or proliferation have been found to be caused by indels. In the peach cultivar 'Hongbaihuatao', a 2 base pair insertion in an anthocyanin transport gene instals a premature stop codon, causing variegated coloration of flowers (Cheng et al., 2015). Although the effect of this mutation is purely cosmetic, the added ornamental value increases the popularity of the cultivar with consumers. However, indels can also critically allow organisms to mitigate immediate threats to their survival. *Amaranthus tuberculatus* was the first weed to evolve resistance to herbicides that inhibit protoporphyrinogen oxidase, an enzyme in a biosynthetic pathway that produces heme and chlorophyll. The resistance was conferred by the deletion of a single Gly residue near the herbicide-binding site (Patzoldt et al., 2006). The affected Gly-coding triplet was located in a short simple sequence repeat, making it particularly prone to slippage-mediated deletion. There are also indications that indels play a major role in the evolution of antibiotic resistance. Indels are significantly enriched in *Mycobacterium tuberculosis* isolates that are associated with multi-drug resistance and epidemic spread (Godfroid et al., 2020). And in a β-lactamase from *Burkholderia thailandensis*, five different deletion mutations have been reported to broaden the substrate spectrum and to increase the minimal inhibitory concentration when exposed to the antibiotic ceftazidime (Hwang et al., 2014; Yi et al., 2012). Using molecular dynamics, those mutations were correlated to elevated flexibility of a loop that is essential to protein activity, possibly increasing the accessibility to the active site.

Indels were recently found to be an important driver in the evolution of coenzyme specificity in Rossmann fold enzymes. Through a systematic search for key motifs in their coenzyme-binding pockets, it was possible to identify an indel of three residues that accomplishes a switch between nicotinamide adenine dinucleotide and S-adenosyl methionine specificity (Toledo-patiño et al., 2022). This work sheds new light on the evolution of cofactor dependence in proteins and provides a novel approach for coenzyme engineering, while also highlighting once again the scaffold plasticity of the Rossmann fold.

Perhaps the strongest tool available for truly exposing the contribution of indels to the diversification or functional innovation during protein evolution is ancestral sequence reconstruction (ASR). Briefly, the reconstruction procedure requires a multiple sequence alignment and a phylogenetic tree as input, and the sequences at internal nodes of the tree are inferred by applying a statistical model of amino acid substitution (Spence et al., 2021). Most ASR studies today focus only on ancestral substitutions and not on indels, but considering the evidence that the latter have a higher per-event contribution to structural variation (Zhang et al., 2018), they certainly deserve better attention. For example, Tokuriki et al. recently analysed the mutational trajectory from an ancestral dihydrocoumarin hydrolase to a methyl-parathion hydrolase that is capable of degrading xenobiotic organophosphate compounds (Yang et al., 2019). A set of five epistatic mutations was sufficient and necessary for achieving this function switch, of which one was a Ser deletion in an active site loop, altering the loop conformation. Looking forward, the use of ASR tools that more carefully handle historical indel events may offer further insights in how they can shape functional adaptation, but most of the currently available tools sadly provide little to no opportunities for evaluating indel placement in a reconstructed sequence. Graphical Representation of Ancestral Sequence Predictions (GRASP) is a recent ASR tool that was designed to solve this problem by presenting protein engineers multiple plausible indel histories that can be explored (Foley et al., 2022; Ross et al., 2022). Another useful ASR tool that can handle indels is FireProt^{ASR}, which also maps their location on the three-dimensional protein structure (Musil et al., 2021). The ability to sample alternative plausible indel states may allow researchers to properly characterise the functional robustness of ancestral proteins to statistical uncertainty, a strategy that has already proven useful in the case of ambiguous historical substitutions (Eick et al., 2017).

5. Indels in protein engineering

5.1. Random approaches

Thanks to the powerful concept of directed evolution which collapses the time scale of natural evolution to just a few months or weeks, protein engineers can tailor features to specific industrial or academic needs (Arnold, 2019). To this day, straightforward random mutagenesis continues to be one of the most effective strategies for introducing genetic diversity in a protein template. However, the most famous random mutagenesis techniques, such as error-prone PCR, only generate libraries of variants that contain amino acid substitutions. Although a few polymerases have been described to exhibit a higher rate of slippage-related errors (Emond et al., 2008; Guilliam et al., 2015; Kashiwagi et al., 2006), those typically insert or remove just one or two bases, disrupting the reading frame. Meanwhile, recombination-based techniques like DNA shuffling or staggered extension process can only reliably be used to mix indel events that were already present in some of the parental genes.

Over the years, various methods have been developed specifically for the random insertion or deletion of residues (Table 1), some of which are discussed below. One of the first was random elongation mutagenesis (REM), where a sequence fragment containing degenerate codons is ligated to a gene, adding a random peptide tail to the N- or C-terminus of a protein of interest. The authors used REM on a *Bacillus stearothermophilus* catalase I mutant with decreased thermostability, and they identified multiple elongated variants that were profoundly more stable (Matsuura et al., 1999). The mutational scope of the REM method is clearly quite narrow, but the concept of protein stabilisation by engineering of the C-terminus is certainly valid, as shown by others (Takano et al., 2011).

The first method that could insert or delete random residues of defined length along the entire sequence was described in the early 2000s. Random insertion and deletion mutagenesis (RID) can be used to delete an arbitrary number of consecutive bases at random positions, and insert a predetermined or random sequence of arbitrary length at the same position (Murakami et al., 2002). To accomplish this, a Ce(IV)/EDTA complex that acts as molecular scissors is briefly added to a circular single-stranded version of the target DNA. The resulting cleavage site determines the location of the forthcoming indel event. The power of RID was demonstrated by applying it to two fluorescent proteins, yielding variants with different fluorescence properties. Yet, the method is technically demanding and it involves multiple low-efficiency steps, which may explain why it has seen little use.

A few simpler methods that can generate indel events anywhere in a sequence were reported in the following years, but they come with their own disadvantages. Segmental mutagenesis fuses two fragment libraries of the same gene that were truncated either at the 5′ end or at the 3′ end by an exonuclease, which causes a deletion or a tandem repeat at the

Table 1
Overview of methods for the generation of random indel libraries.

Method	Description	Suitable for insertions	Suitable for deletions	Prevents frameshifts	Controlled insertion sequence	
Error-prone polymerases	Some error-prone DNA polymerases generate a high proportion of indel errors during amplification.	+	+	−	−	(Emond et al., 2008; Guilliam et al., 2015; Kashiwagi et al., 2006)
RAISE	Target DNA is fragmented by DNase I, several nucleotides are attached to the 3′ terminus of the fragments using terminal deoxynucleotidyl transferase, and the obtained elongated fragments are reassembled by self-priming PCR.	+	+	−	−	(Fujii et al., 2006)
Segmental mutagenesis	Plasmid is linearized by cutting at the 5′ or the 3′ end of the gene. Exonuclease treatment progressively shortens the gene until the target region is reached. After removing remaining vector DNA, the 5′ and 3′ truncated gene fragments are randomly ligated. A repeat or a deletion is introduced at the fusion point.	+	+	−	−	(Pikkemaat and Janssen, 2002)
Random deletion and tandem duplication	Plasmid is digested by DNase I to generate double strand breaks with overhangs. The 5′ → 3′ polymerase and 3′ → 5′ exonuclease activity of T4 DNA polymerase is used to make the overhangs blunt, which creates deletions and tandem duplication insertions.	+	+	−	−	(Hida et al., 2010)
TRINS	Gene is fragmented by DNase I. An aliquot of fragments is treated by ssDNA ligase to generate circular single-stranded DNA. Next, linear and circular fragments are mixed and rolling circle amplification is performed using the strand displacement activity of Pfu polymerase, generating fragments with tandem repeats. Finally, the gene is reassembled from the intact and elongated fragments.	+	−	−	−	(Kipnis et al., 2012)
Truncation by transposition	An artificial transposon is randomly inserted into the gene. In two separate PCR reactions, 5′-end and 3′-end fragment libraries are amplified using a primer that binds to the transposon, then a primer binds to the 5′ or 3′ end of the gene, respectively. The 5′ and 3′ fragment libraries are ligated to linker DNA, the transposon is removed by endonuclease digestion, and the gene fragments are reassembled by blunt end ligation. Finally, the gene library is linearized by PCR.	−	+	−	n.a.	(Morelli et al., 2017)
MuDel	An engineered mini-Mu transposon is randomly inserted into the gene by transposase. The transposon is removed by endonuclease digestion and the gene is reassembled by ligation. During the restriction-digestion step, three nucleotides are removed.	−	+	+	n.a.	(Jones, 2005)
Codon deletion mutagenesis	An engineered asymmetric Mu transposon is randomly inserted into the gene. Inverse PCR is performed, with the choice of primers determining the number of codons that will be deleted. Next, the PCR product is digested by endonuclease, followed by ligation.	−	+	+	n.a.	(Liu et al., 2016)
COBARDE	During chemical oligonucleotide synthesis, further growth of a fraction of the oligos is randomly halted by attaching a protecting group in the nucleotide that precedes the codon to be deleted. Unprotected chains undergo three cycles of synthesis to introduce 3 nucleotides, after which the protecting groups are removed and synthesis continues. Deletion frequency can be adjusted by fine tuning the protecting group concentration.	−	+	+	n.a.	(Osuna et al., 2004)
Pentapeptide scanning mutagenesis	Random insertion of the Tn4430 transposon, followed by digestion by endonuclease and ligation. Most of the transposon is deleted, but a 15-bp fingerprint is left behind that consists of 5 bp of each transposon end and 5 bp of duplicated DNA from the target site.	+	−	+	−	(Hayes and Hallet, 2000)
Random elongation mutagenesis	A random peptide chain is attached to the C-terminus of a protein by digesting a DNA fragment encoding this peptide chain by endonuclease, and ligating it to digested plasmid DNA.	+	−	+	+	(Matsuura et al., 1999)
RID	The gene target is converted to circular ssDNA and chemically cleaved at a random site. Anchors containing an endonuclease restriction site and nucleotides to be inserted are ligated to both ends. ssDNA is filled in by PCR. The anchors are then removed by endonuclease digestion, but a few anchor bases are retained at the 5′ end and a number of bases are removed from the 3′ end. Finally, the digestion product is cyclized again.	+	+	+	+	(Murakami et al., 2002)
InDel assembly	Involves cycles where template DNA is bound to paramagnetic beads and digested with a type IIs endonuclease, followed by annealing and ligation of standardised DNA building blocks. The building blocks contain a degenerate overhang, a triplet that is inserted into the sequence, and a new endonuclease recognition site that enables the assembly cycle to be restarted. Variation in composition and length can be introduced.	+	+	+	+	(Tizei et al., 2021)
TRIAD	An engineered mini-Mu transposon is randomly inserted into the gene, determining the location of the eventual indel event. The ends of the transposons were designed to result in the deletion or insertion of triplets after digestion and ligation steps. Insertions are obtained by ligation of shuttle cassettes containing randomised nucleotide triplets, and a shuttle sequence that is eventually removed.	+	+	+	+	(Emond et al., 2020)

n.a.: not applicable.

fusion point (Pikkemaat and Janssen, 2002). The strategy was successfully employed to obtain haloalkane dehalogenase variants with enhanced promiscuous activity on 1,2-dibromoethane. But like other exonuclease-based methods (Hida et al., 2010), it frequently introduces frameshift mutations and the produced insertions are only limited to repeats of the original sequence.

Another relatively straightforward method is random insertional-deletional strand exchange (RAISE) (Fujii et al., 2006). Unlike the exonuclease-based methods, it does not limit insertions to tandem repeats, but it also does not prevent the introduction of frameshift mutations. RAISE is reminiscent of gene shuffling and utilises terminal deoxynucleotidyl transferase to attach random nucleotides to the 3′ terminus of digested DNA before the fragments are reassembled by self-priming PCR. RAISE has been used to generate mutants of TEM β-lactamase with improved ceftazidime-hydrolysing activity, which interestingly pointed out how hot spots for deletions were situated close to hot spots for point mutations.

Since the presence of frameshifts in a library causes most variants (> 66%) to be non-functional, avoiding them is highly desirable. Methods that rely on engineered transposons offer an elegant solution to this problem. Transposons are DNA elements that can accurately and efficiently be inserted into a sequence at a random location with the help of a transposase. By redesigning the transposon to contain appropriately situated recognition sites of certain restriction enzymes, cleavage with these restriction enzymes and subsequent religation can bring forth the insertion or deletion of nucleotide triplets. Pentapeptide scanning mutagenesis makes use of a transposon originating from *Bacillus thuringiensis* which leaves a 15 bp fingerprint after transposition, digestion and ligation: 5 bp are duplicated from the target site and 5 bp are left behind from each of the transposon ends (Hayes et al., 1997; Hayes and Hallet, 2000). Others have created systems based on the bacteriophage Mu transposon. In MuDel, this transposon was modified at its termini to include a recognition site for *Mly*I (Jones, 2005). The ability of this restriction enzyme to cut a few bp outside of its recognition sequence allows exactly three bp of the gene of interest to be deleted. A few studies have used these methods to evaluate the tolerance of TEM β-lactamase or green fluorescent proteins to insertions or deletions (Arpino et al., 2014a; Hayes and Hallet, 2000; Jones, 2005; Simm et al., 2007).

Hollfelder et al. recently described transposon-based random insertion and deletion mutagenesis (TRIAD), which may be the most versatile method for generating indel mutations to date (Emond et al., 2020). It consists of a single transposition reaction using engineered mini-Mu transposons, followed by a few cloning steps. The method is an advance of the MuDel system and the MuDel-based TriNEx method, which was designed for the random substitution of trinucleotide sequences (Baldwin et al., 2008). TRIAD can delete one, two or three triplets, or it can insert a cassette of one, two or three fully randomised triplets. The efficacy of TRIAD was demonstrated by creating indel libraries of a phosphotriesterase. It was observed that the vast majority (> 75%) of indel mutations had a strongly deleterious effect on native phosphotriesterase activity, whereas the effect of substitutions was largely neutral. However, none of the substitutions could improve the native activity any further, while a few indels did. Similar effects were found when screening the same libraries for promiscuous arylesterase activity. Indels were largely deleterious, but the frequency of beneficial indels was also at least three times higher than that of beneficial substitutions. Those numbers convincingly illustrate how indels appear to polarise the properties of library members towards extreme outcomes. Since its inception, TRIAD has already been applied for developing antibodies with improved binding affinity (Skamaki et al., 2020) and for improving the properties of a bifunctional ancestor of haloalkane dehalogenase and *Renilla*-type luciferases (Schenkmayerova et al., 2021). Given these success stories, it appears that protein engineers now finally have a tool at their disposal that can help them travel to new areas of sequence space in directed evolution experiments, granting access to the steep ledges that indels tend to carve into the fitness landscape.

5.2. Semi-rational approaches

Given our poor knowledge of how indels affect the properties of a protein, techniques that randomly insert or delete residues across the entire sequence offer obvious benefits. However, the notoriously low hit rate of random mutant libraries generally necessitates a massive screening effort, and the development of an appropriate high-throughput screening method for identifying variants with improved properties is far from trivial. Not all properties or activities can conveniently be screened for at large scale due to technical limitations, time constraints or high costs. Therefore, there is a clear need for strategies that can exploit any available structure-function information to search only the most promising regions of sequence space, resulting in smaller but smarter libraries that are more likely to yield hits (Chica et al., 2005). For substitution-based protein engineering, this semi-rational approach has seen tremendous success over the past two decades and techniques that use degenerate codons to randomise so-called mutational 'hot spots' (e.g. iterative saturation mutagenesis or combinatorial active-site saturation testing) have become common practice (Reetz et al., 2005; Reetz and Carballeira, 2007). Unfortunately, there are very few examples of semi-rational strategies that have been reimagined for the generation of indel variants.

Stepwise loop insertion strategy (StLois) can be performed to remodel and elongate active site loops (Fig. 3A) (Hoque et al., 2017). The targeted loop first has to be identified rationally, which can be accomplished by comparing the length and composition of functionally relevant loops in the engineering template to those in homologous proteins. Next, a library is constructed where random residues are introduced into a promising elongation site using NNK degenerate codons. Engineers should carefully choose the mutational step length, that is the number of residues that are simultaneously inserted at the elongation site in each round of mutation: larger step lengths drastically increase the size of the library, but they also stimulate the discovery of inserted residue combinations that exhibit positive synergistic effects. The authors found a double residue insertion in each round to be a solid compromise. Finally, the obtained libraries are screened, and the best variant(s) can subsequently be subjected to another round of mutagenesis where the targeted loop is elongated once more. StLois has successfully been used to improve the phosphotriesterase activity of a laccase, resulting in a 16-fold increase in catalytic efficiency towards ethyl-paraoxon (Hoque et al., 2017).

A different group devised the linker in loop insertion (LILI) approach, which shows some similarities to StLois, but requires considerably less screening. In LILI, not fully randomised residues but predefined linkers of different lengths (2 to 6 residues) are inserted at rationally selected mutational hot spots in flexible loops (Fig. 3B) (Heinemann et al., 2021). Examples of suitable linkers with different dynamic properties are the flexible Gly-Gly or Gly-Ser linkers, the stiff Pro-Ala linker, or the Gly-Pro linker without defined structure. LILI thus randomly samples the length and flexibility of catalytically relevant loops. The potential of the strategy was demonstrated on a cumene dioxygenase, of which the activity and product profile could be modulated significantly.

It is also possible to extend random approaches with a semi-rational fine-tuning step. In an engineering study that explored the enhancement of antibodies affinity by indels, TRIAD was first applied to search for positions that tolerate single amino acid insertions. Then, diverse libraries were designed that contain zero to five additional degenerate codons at the most promising insertion point. This insertional scanning mutagenesis (InScaM) process exposed multiple variants with markedly improved binding affinities (Skamaki et al., 2020).

Future semi-rational protein engineering studies could consider applying a degree of randomization that balances between the full NNK degeneracy of StLois or InScaM and the predefined linkers of LILI. Indeed, semi-rational substitution libraries regularly make use of

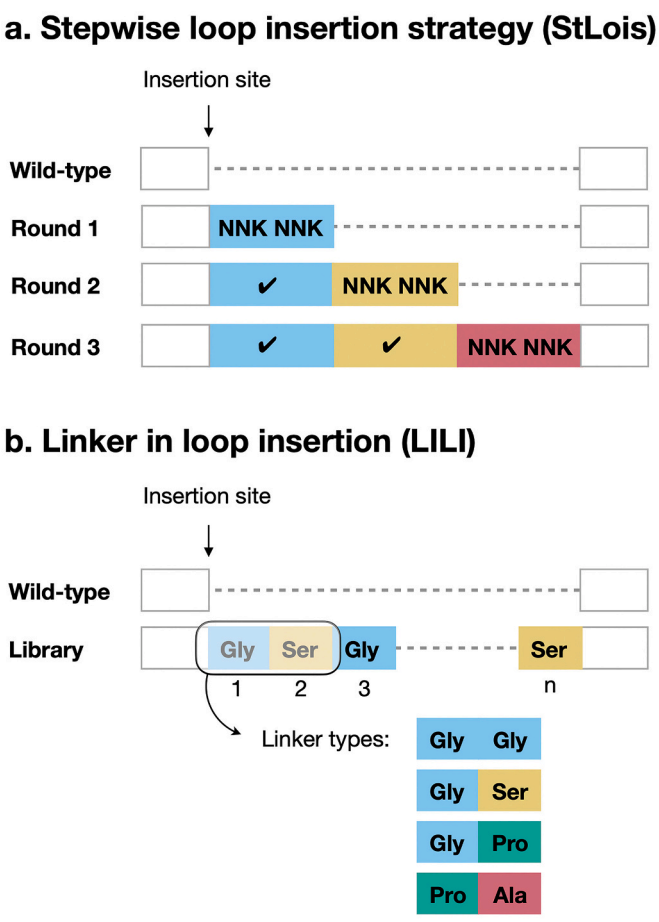

Fig. 3. Semi-rational strategies for remodelling and/or elongating functionally relevant loops. (a) The stepwise loop insertion strategy involves iterative cycles of introducing one or more additional residues at a rationally selected insertion site using NNK degenerate codons. The most favourable variant identified in the screening process becomes the template for the following round of insertion mutagenesis. (b) In the linker in loop insertion approach, linkers with various dynamic properties (four shown) of different lengths (e.g. $1 \leq n \leq 6$) are introduced at the targeted insertion sites.

reduced amino acid alphabets that search sequence space more intelligently. One example is the popular NDT degeneracy that encodes only 12 amino acids with a well-rounded mix of physicochemical properties (Reetz et al., 2008), but each inserted degeneracy can even be individually fine-tuned based on rational or computational considerations (Reetz and Wu, 2008). In addition, it may be worth assessing whether the ASR tool GRASP can provide assistance in identifying suitable hot spots for insertion mutagenesis by offering a glimpse at plausible historical indel events, as those might indicate which sites tolerate such mutations best.

5.3. Rational and computational approaches

Protein engineers have a long history of letting chemical intuition, computational tools or insights from structure-function relation studies guide their mutagenesis strategies. That accumulated experience now allows us to confidently develop reasonable hypotheses of how certain substitutions may establish salt bridges or disulfide bonds that enhance stability (Yang et al., 2015), switch between evolutionarily-related activities (Franceus et al., 2021), create more space for bulky substrates (Dirks-Hofmeister et al., 2015), and so on. In contrast, relatively few studies have focused on indels. The lack of prior examples makes it difficult for researchers to suggest indels that may alter relevant protein properties. The ultimate goal of truly understanding and predicting the molecular basis of their effects remains far off for the time being.

So far, indel mutations have primarily proven their worth in rational engineering experiments when they were a part of loop exchanges between homologous proteins. This process is known as loop grafting and it involves preserving the overall scaffold and catalytic residues while replacing the more variable loop regions in-between structural elements (Nestl and Hauer, 2014; Tawfik, 2006). Various examples can be found in literature where a grafted loop was shorter or longer than the original loop, and where the obtained chimaera displayed significant changes in activity, enantioselectivity, thermostability or specificity (Table 2). A beautiful example was described by Boersma and colleagues, who were aiming to improve the enantioselectivity of a lipase from *Bacillus subtilis* in the kinetic resolution of 1,2-*O*-isopropylidene-*sn*-glycerol esters (Boersma et al., 2008). A loop near the active site entrance was replaced by longer loops originating from structurally homologous cutinase or esterase with the intention of increasing the interaction surface with the substrate. This approach yielded variants with inverted and improved enantioselectivity.

The general concept of loop grafting may be quite straightforward, but its outcome is not easily predictable. Even in published research, failed grafting attempts that abolished all soluble expression or activity have been reported (Hawwa et al., 2009; Xiang et al., 2009). And when the scaffold does tolerate the different length of the new loop, epistatic ratchets may still prevent its beneficial effect from coming to fruition. The Tawfik group discovered that a deletion in a loop of phosphotriesterase can trigger the emergence of homoserine lactonase activity, but only when this deletion is combined with an adjacent highly epistatic substitution (Afriat-Jurnou et al., 2012). The authors speculated that the restrictive substitution may have occurred as one of the final steps in the divergence of this phosphotriesterase from the common ancestor of phosphotriesterases and related lactonases, blocking the novel enzyme from reverting back to its ancestral bifunctional state. This finding sounds a cautionary note: regular amino acid substitutions should not be disregarded when analysing or introducing indel mutations, as the success of an indel may be contingent on a point mutation elsewhere, or vice versa.

Some of the difficulties involving the use of indels in rational engineering may be overcome by computational analyses. A web-based tool named LoopGrafter is now available to provide visual support in the process of transplanting loops between homologous proteins (Planas-Iglesias et al., 2022). After simply uploading the desired scaffold and insert protein structures, LoopGrafter can be used to identify and review candidate loops for grafting, to pair all candidate scaffold loops to suitable insert loops, and to visualise and rank grafted proteins. The application supported a recent study where a luciferase loop with important dynamic properties was transplanted into an ancestral protein with both haloalkane dehalogenase and weak luciferase activity, which implemented stable glow-type bioluminescence (Schenkmayerova et al., 2021).

The range of possible new loops is not necessarily restricted to those already present in nature. The Rosetta suite is capable of remodelling entire active site loops *in silico* in terms of length, conformation and sequence composition to establish key interactions with a ligand of choice. For example, a loop in human guanine deaminase could be redesigned to increase activity on ammelide by two orders of magnitude, with the optimal loop containing two deletions and four substitutions (Murphy et al., 2009). Remodelling challenges can even be outsourced to citizen scientists. Players of the online protein puzzling game FoldIt managed to increase the activity of a computationally designed Diels-Alderase >18-fold by remodelling its backbone, which involved an insertion of 13 residues (Eiben et al., 2012).

Finally, the current surge of interest in machine learning (ML) algorithms that are capable of spotting patterns in data opens up interesting opportunities for the inclusion of indels in data-driven rational design. For instance, ML models have already been used to drastically

Table 2
Examples of loop grafting studies where the length of the grafted loop is different from the loop in the scaffold.

Scaffold	Grafted loop origin	Result	
Glyoxalase II	Metallo β-lactamase	Introduction of β-lactamase activity	(Park et al., 2006)
Lipase	Cutinase and esterase	Inversion of enantioselectivity	(Boersma et al., 2008)
Lactonase with low phosphotriesterase activity	Phosphotriesterase	Loss of activity	(Hawwa et al., 2009; Xiang et al., 2009)
Phosphotriesterase	Lactonase	Emergence of lactonase activity, but only when combined with an adjacent epistatic substitution	(Afriat-Jurnou et al., 2012)
Nicotinamide-dependent cyclohexenone reductase	Thermophilic-like subfamily of old yellow enzymes	Variations in thermostability and solvent tolerance	(Reich et al., 2014)
Nicotinamide-dependent cyclohexenone reductase	Old yellow enzyme and morphinone reductase	Improved activity in a cascade reduction of allylic alcohols	(Reich et al., 2016)
Proline 4-hydroxylase	Other proline 4-hydroxylases	Improved activity, reduced thermostability	(Liu et al., 2019)
Sortase	Other sortase with different substrate preference	Change in substrate preference	(Wójcik et al., 2020)

reduce the screening burden associated with combinatorial substitution libraries by predicting which regions of sequence space are enriched in variants with higher fitness, from data obtained by screening a modest subset of those libraries (Wu et al., 2019). Similarly, such models might be able to find patterns in the structure-function relationships of indel variants to predict their effects *in silico*. Some exploratory work has already been done in this area. In one study, a publicly available dataset of single amino acid deletions in enhanced green fluorescent protein was analysed to assess how well tolerance to deletions can be inferred from structural features (Jackson et al., 2017). Packing density was found to offer significant predictive power. A different group managed to build a classifier that distinguishes whether a single point deletion leads to a folded or unfolded protein conformation (Banerjee et al., 2019), as well as a positive-unlabeled classifier that predicts the change in foldability arising from multi-point deletions (Banerjee et al., 2020). Statistical analyses also provided useful insights in the dynamics-based luciferase engineering study mentioned above, which made use of both TRIAD and LoopGrafter (Schenkmayerova et al., 2021).

6. Domain indels

The functionality of a protein is often the result of the acquisition of novel elements which improve the binding to a partner, the specificity for a substrate or the protein's catalytic activity. While these features are usually acquired over time after the accumulation of mutations, a different route involves the insertion or deletion of whole domains. The introduction of a domain in a protein sequence can also be the result of an indel event, the major difference being that the number of bases inserted in the sequence is far larger. One option is the insertion of a domain as a sort of prepackaged independent module, but another possibility exists where a repeat sequence is inserted, followed by evolution of the inserted repeat into a domain with a novel function. Examples of the latter have been detected in the genome of *Rickettsia conorii*, where mobile palindromic repeat elements were discovered that are capable of insertion in open reading frames (ORFs). Surprisingly, the mobile elements persistently appear at the surface of the proteins coded by those ORFs (Claverie and Ogata, 2003). In this way, the original fold and function of the scaffold proteins are unaffected by the insertion. Because of their intrinsic features, palindromic DNA repeats have a high probability of coding for soluble peptides that adopt an independent fold. This basically makes the *Rickettsia* repeat elements perfect transitory sequences: they can be translated into low-profile elements located at the host protein surface after which they either perish, or evolve into a binding module, a specificity loop, or another useful structural element. The insertion of whole domains based on repeats is a very successful evolutionary strategy, and as many as 25% of recorded proteins contain long repeats (Pellegrini et al., 2012). The length of the repeat unit can range from just a single amino acid, to >100 residues. Repeats can expand or contract by insertion or deletion of repeat units, respectively, but such indels are usually detrimental unless the repeat units fold independently (Schüler and Bornberg-Bauer, 2016).

A few groups have tried to harness the potential of domain insertion, with research focusing on identifying the optimal fusion strategy. An important consideration concerns the relative location of N- and C-termini of the inserted domain (Fig. 4) (Ostermeier, 2005). About half of the available structures for single domain proteins have their N- and C-termini proximal, meaning that the two extremities are close to each other in three-dimensional space. This characteristic makes such domains well suited to be inserted in another scaffold. On the other hand, scaffold domains are ideally discontinuous, meaning that their linear sequence is (or can be) interrupted by another domain. Discontinuous domains are quite prevalent in natural multi-domain proteins (Jones et al., 1998). Methods for performing domain insertion have been reviewed in the past (Kanwar et al., 2013).

Like for other technologies explored in this review, a notable target of domain engineering has been β-lactamase. A cytochrome sequence was randomly inserted into a β-lactamase scaffold, which linked the function of these two unrelated proteins by making tolerance to ampicillin dependent on the presence of heme (Edwards et al., 2008). A random domain insertion approach can be more attractive than a rational approach, as the latter is rarely successful due to the difficulty of predicting the outcome of entangled structures. The best-performing insertion was achieved by removing a loop, which was replaced with a whole domain. Following these first results, the performance of the variants was further explored, finding that the best ones have high structural independence between the domains (Edwards et al., 2010). It has also been possible to link the function of β-lactamase and a maltose-binding protein to create hybrid proteins where maltose is an effector of β-lactam hydrolysis (Guntas et al., 2005). In green fluorescent protein, insertion of a calmodulin or a zinc finger domain can generate indicator proteins whose fluorescence can be enhanced by metal binding (Baird et al., 1999). And by fusing the chromophore centres of enhanced green fluorescent protein and the heme-binding electron transfer protein cytochrome b_{562}, fluorescence quenching became heme-dependent (Arpino et al., 2012).

The deletion of domains has also been attempted. For example, a bacteriophage endolysin, from which the internal amidase domain was removed, retained efficacy on its staphylococcal targets while attenuating the harmful side effects on the animal body (Zhou et al., 2017). And deletion of certain domains in the thrombolytic protein alteplase was found to alter its pharmacokinetic properties (Acheampong and Ford, 2012).

7. Frameshift mutations

While the effect of indels introducing or removing multiples of three nucleotides has often been discussed in literature, and their potential in engineering has been tested to a certain extent, indels of lengths not divisible by three nucleotides have received less attention. These

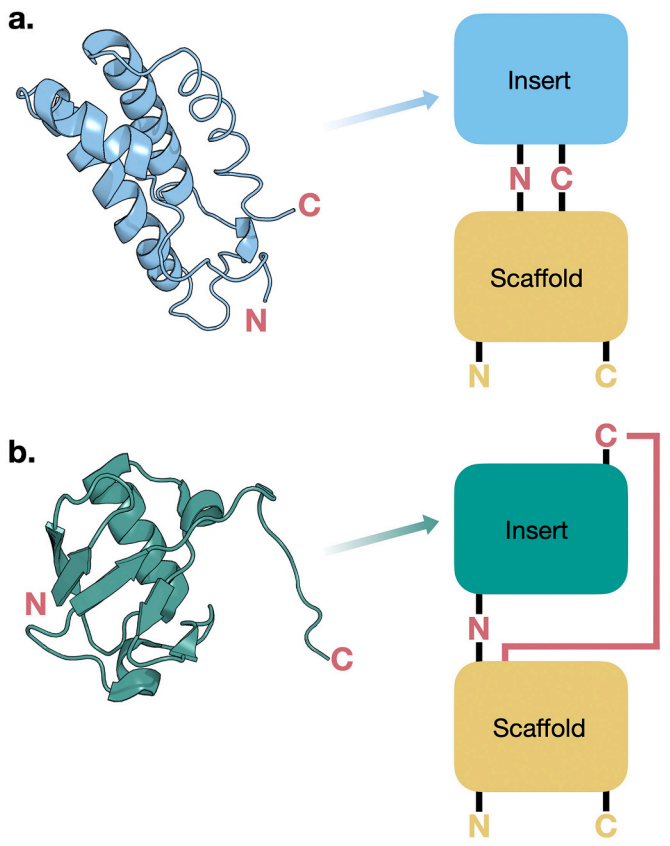

Fig. 4. Domain assembly according to relative termini location. (a) A domain with proximal N- and C-termini can be inserted on another domain while retaining the topology and folding of both domains. (b) A domain with distal termini requires the introduction of a spacer sequence in order to retain proper domain folding.

frameshift indels change the reading frame for translation, which typically results in the production of truncated or nonfunctional variants. However, recent work has convincingly demonstrated that the destructive reputation of frameshifts is perhaps not entirely deserved, especially when related to translation in the native context of the host organism.

Savaging systems exist to prevent wasting energy and resources on the translation of off-frame nonfunctional proteins. One of these systems is to "ambush" frame-shifted sequences with hidden stop codons, which only appear off-frame (Seligmann and Pollock, 2004). This is clearly visible when translating frame-shifted sequences *in silico*: a series of stop codons emerge, rather than a single long sequence that does not match the original. While this process may seem coincidental at a first glance, strong arguments point towards a very precise mechanism being behind the presence of hidden stop codons. This mechanism is embedded in the standard genetic code and in the codon usage of organisms. Upon introduction of a frameshift mutation, only 20 of the 64 standard codons cannot contribute to the appearance of a hidden stop codon (Fig. 5), and on top of that, a positive correlation exists between codon usage and the number of ways a codon can form hidden stops (Seligmann and Pollock, 2004).

Interestingly, even when hidden stop codons cannot intervene to abruptly terminate the translation of frameshifted sequences, the standard genetic code can minimise the impact of these mutations. The standard genetic code was found to be very efficient in maintaining similar amino acid properties for coded sequences after point mutations, mistranslations (Freeland and Hurst, 1998) and frameshift mutations (Geyer and Mamlouk, 2018). The vast majority of randomly generated artificial genetic codes are far less resilient. It has been proposed that the frameshift-robustness of the standard genetic code is in fact a byproduct of its mismatch-robustness. Indeed, due to the high degeneracy of the genetic code, most amino acid changes caused by a frameshift are also accomplishable by a single mismatch error (Xu and Zhang, 2021).

The preservation of physicochemical properties (such as polarity, affinity to nucleobases and intrinsic disorder) as consequence of frameshift mutations was further investigated on a dataset of almost 3000 human proteins (Bartonek et al., 2020). There is a positive correlation between the hydrophobicity profile of the originals' and the +1 frameshift variants, despite the average sequence identity between the two groups being just 6.5%. Frameshifts thus allow very far jumps in sequence space to be explored, while maintaining the same physicochemical properties.

Mutually compensatory pairs of frameshift mutations constitute another evolutionary mechanism that can assist the recovery of functionality after frameshift-mediated movements through sequence space (Biba et al., 2022). One shift can be compensated by a second one (e.g. +1 and − 1, or + 1 and + 2) to restore the original frame. It has been shown that such pairs of frameshifting indels are more likely to arise as distinct mutational events separated by a period of time, which implies that frameshifted uncompensated intermediates occur as well. Similarly, gene silencing caused by indels can be reverted to reacquire the functional version of a gene (Gupta and Alland, 2021).

Probably even more intriguing is the observation that indels in homonucleotide repeats in coding regions can be bypassed at the transcriptional and translational levels (Rockah-Shmuel et al., 2013). The bypass of indels, which restores the original coding frame, is suggested to be the result of RNA polymerase slippage or ribosomal slippage and the likelihood of this phenomenon taking place is positively correlated to the length of the repeat. This mechanism shows striking similarities to the DNA polymerase slippage mechanism that introduces indels in the first place, which is more likely to occur in longer repeats as well.

The findings enumerated above underline that there may be a place for frameshift mutations in future protein engineering strategies after all. The tendency of frameshift variants to retain the original physicochemical properties opens up an exciting new avenue for crossing sequence space. The question remains to what extent the preservation of physicochemical properties results in the preservation of biological function or enzymatic activity. Once more information will become available on this topic, frameshift variants could be taken into consideration as new starting points for directed evolution, in search of new or improved properties. However, an important caveat is the systematic presence of hidden stops, which may require careful optimization of codon usage.

8. Concluding remarks and perspectives

Numerous studies have clearly demonstrated that the adaptive evolution of proteins is not influenced only by changes in amino acid composition, but also by changes in size. However, their full potential in protein engineering endeavours has not yet been realised. Fortunately, recent years have witnessed meaningful progress in the field, making it easier for indels to be embraced in mutagenesis strategies. In random directed evolution, the TRIAD method is a promising addition to the mutagenesis toolbox for scanning whether functional innovation can be achieved through indels, even when little structural or functional knowledge is available. In semi-rational engineering, the StLois and LILI approaches allow sampling the length, composition and dynamics of promising loop regions. With LoopGrafter, there is now a straightforward way to transplant loops between structurally-related proteins. And the continuous advances in model-driven and ML-assisted (re)design are expanding the boundaries of computational indel engineering.

Future research should focus on elevating our insight into the structural and functional implications of indel mutations to the next level. The work discussed in this review shows that we already have a rough idea of which regions are most likely to tolerate indels.

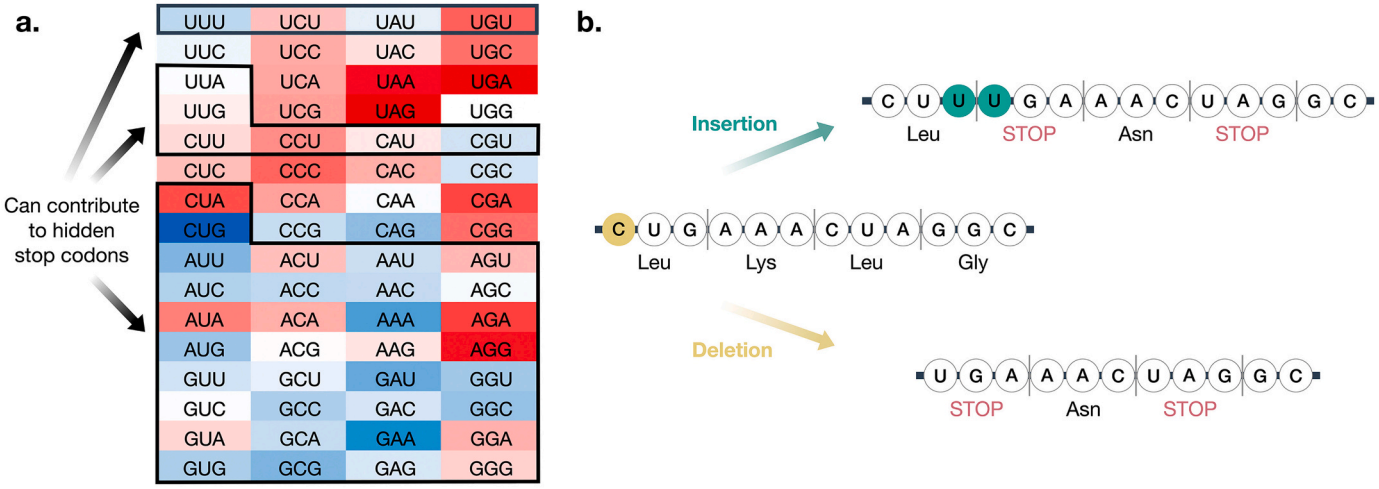

Fig. 5. Hidden off-frame stop codons can terminate translation of frame-shifted sequences. (a) Codon table, with 44 codons that can contribute to hidden stops (marked by black borders). These codons start with A or G, or they end with U, UA or UG. The colour scale from blue (higher) to red (lower) reflects the relative frequency of codon usage in *E. coli*. (b) CUG and AAA are examples of a codon pair that contribute to the hidden stop codon UGA. This first codon ends with UG and second starts with A. Similarly, CUA and GGC contribute to the hidden stop codon UAG. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Furthermore, we now know that indels tend to be high risk, high reward mutations that are simultaneously more likely to be deleterious, but also more likely to be advantageous. Gaining a deeper understanding of the molecular origin of these observations would be highly desirable. For example, it seems plausible that indels can have a substantial impact on conformational dynamics, which is known to be an important driver of protein evolution (Maria-Solano et al., 2018), but information on this topic is scarce. It is also important that we continue studying the structure-functional implications of indels at the domain level, and their contribution to the evolution proteins.

Until the largest gaps in our general comprehension of the effects of indels have been filled, semi-rational approaches may be most suited for incorporating indel variation in routine protein engineering workflows. Indeed, the high risk associated with indel mutations implies that at this time, random methodologies may only be practically feasible when the protein of interest is compatible with (ultra)high-throughput protocols that bypass the library size problem altogether (Sheludko and Fessner, 2020). Conversely, semi-rational approaches hit the sweet spot in the trade-off between library size and hit rate. To stimulate their use and success, there is a need for more inventive ways of leveraging our current knowledge of indels for the construction of manageable smart libraries.

Above all, it is our hope that a deeper awareness of the importance and potential of indels will emerge. We have just scratched the surface of the exciting insights and applications that these sophisticated mutations can bring. Once they start attracting more attention in the work of molecular biologists, structural biologists and protein engineers, we will certainly be able to dig much deeper.

Declaration of competing interest

The authors declare no conflicts of interest.

Acknowledgements

This work was supported by Research Foundation-Flanders (grant number 12ZD821N to JF) and the Special Research Fund of Ghent University (to SS).

References

Acheampong, P., Ford, G.A., 2012. Pharmacokinetics of alteplase in the treatment of ischaemic stroke. Expert Opin. Drug Metab. Toxicol. 8, 271–281. https://doi.org/10.1517/17425255.2012.652615.

Afriat-Jurnou, L., Jackson, C.J., Tawfik, D.S., 2012. Reconstructing a missing link in the evolution of a recently diverged phosphotriesterase by active-site loop remodeling. Biochemistry 51, 6047–6055. https://doi.org/10.1021/bi300694t.

Arnold, F.H., 2019. Innovation by evolution: bringing new chemistry to life (Nobel lecture). Angew. Chem. Int. Ed. 58, 14420–14426. https://doi.org/10.1002/anie.201708408.

Arpino, J.A.J., Czapinska, H., Piasecka, A., Edwards, W.R., Barker, P., Gajda, M.J., Bochtler, M., Jones, D.D., 2012. Structural basis for efficient chromophore communication and energy transfer in a constructed didomain protein scaffold. J. Am. Chem. Soc. 134, 13632–13640. https://doi.org/10.1021/ja301987h.

Arpino, J.A.J., Reddington, S.C., Halliwell, L.M., Rizkallah, P.J., Jones, D.D., 2014a. Random single amino acid deletion sampling unveils structural tolerance and the benefits of helical registry shift on GFP folding and structure. Structure 22, 889–898. https://doi.org/10.1016/j.str.2014.03.014.

Arpino, J.A.J., Rizkallah, P.J., Jones, D.D., 2014b. Structural and dynamic changes associated with beneficial engineered single-amino-acid deletion mutations in enhanced green fluorescent protein. Acta Crystallogr. Sect. D Biol. Crystallogr. 70, 2152–2162. https://doi.org/10.1107/S139900471401267X.

Baird, G.S., Zacharias, D.A., Tsien, R.Y., 1999. Circular permutation and receptor insertion within green fluorescent proteins. Proc. Natl. Acad. Sci. U. S. A. 96, 11241–11246. https://doi.org/10.1073/pnas.96.20.11241.

Baldwin, A.J., Busse, K., Simm, A.M., Jones, D.D., 2008. Expanded molecular diversity generation during directed evolution by trinucleotide exchange (TriNEx). Nucleic Acids Res. 36 https://doi.org/10.1093/nar/gkn358.

Banerjee, A., Levy, Y., Mitra, P., 2019. Analyzing change in protein stability associated with single point deletions in a newly defined protein structure database. J. Proteome Res. 18, 1402–1410. https://doi.org/10.1021/acs.jproteome.9b00048.

Banerjee, A., Kumar, A., Ghosh, K.K., Mitra, P., 2020. Estimating change in foldability due to multipoint deletions in protein structures. J. Chem. Inf. Model. 60, 6679–6690. https://doi.org/10.1021/acs.jcim.0c00802.

Bartonek, L., Braun, D., Zagrovic, B., 2020. Frameshifting preserves key physicochemical properties of proteins. Proc. Natl. Acad. Sci. U. S. A. 117, 5907–5912. https://doi.org/10.1073/pnas.1911203117.

Bębenek, A., Ziuzia-Graczyk, I., 2018. Fidelity of DNA replication—a matter of proofreading. Curr. Genet. 64, 985–996. https://doi.org/10.1007/s00294-018-0820-1.

Behringer, M.G., Hall, D.W., 2016. Genome-wide estimates of mutation rates and spectrum in *Schizosaccharomyces pombe* indicate CpG sites are highly mutagenic despite the absence of DNA methylation. G3 Genes, Genomes, Genet. 6, 149–160. https://doi.org/10.1534/g3.115.022129.

Biba, D., Klink, G., Bazykin, G.A., 2022. Pairs of mutually compensatory frameshifting mutations contribute to protein evolution. Mol. Biol. Evol. 39, 1–14. https://doi.org/10.1093/molbev/msac031.

Boersma, Y.L., Pijning, T., Bosma, M.S., van der Sloot, A.M., Godinho, L.F., Dröge, M.J., Winter, R.T., van Pouderoyen, G., Dijkstra, B.W., Quax, W.J., 2008. Loop grafting of *Bacillus subtilis* lipase a: inversion of enantioselectivity. Chem. Biol. 15, 782–789. https://doi.org/10.1016/j.chembiol.2008.06.009.

Bornscheuer, U.T., Hauer, B., Jaeger, K.E., Schwaneberg, U., 2019. Directed evolution empowered redesign of natural proteins for the sustainable production of chemicals

and pharmaceuticals. Angew. Chem. Int. Ed. 58, 36–40. https://doi.org/10.1002/anie.201812717.
Bridgham, J.T., Ortlund, E., Thornton, J.W., 2009. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature 461, 515–519. https://doi.org/10.1038/nature08249.
Cartwright, R.A., 2009. Problems and solutions for estimating indel rates and length distributions. Mol. Biol. Evol. 26, 473–480. https://doi.org/10.1093/molbev/msn275.
Chen, J.Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M., Tian, D., 2009. Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. Mol. Biol. Evol. 26, 1523–1531. https://doi.org/10.1093/molbev/msp063.
Cheng, J., Liao, L., Zhou, H., Gu, C., Wang, L., Han, Y., 2015. A small indel mutation in an anthocyanin transporter causes variegated colouration of peach flowers. J. Exp. Bot. 66, 7227–7239. https://doi.org/10.1093/jxb/erv419.
Chica, R.a, Doucet, N., Pelletier, J.N., 2005. Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. Curr. Opin. Biotechnol. 16, 378–384. https://doi.org/10.1016/j.copbio.2005.06.004.
Chowdhury, R., Maranas, C.D., 2020. From directed evolution to computational enzyme engineering—a review. AICHE J. 66, 1–17. https://doi.org/10.1002/aic.16847.
Claverie, J.M., Ogata, H., 2003. The insertion of palindromic repeats in the evolution of proteins. Trends Biochem. Sci. 28, 75–80. https://doi.org/10.1016/S0968-0004(02)00036-1.
Clifton, B.E., Kaczmarski, J.A., Carr, P.D., Gerth, M.L., Tokuriki, N., Jackson, C.J., 2018. Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein article. Nat. Chem. Biol. 14, 542–547. https://doi.org/10.1038/s41589-018-0043-2.
Copley, S.D., 2020. Evolution of new enzymes by gene duplication and divergence. FEBS J. 287, 1262–1283. https://doi.org/10.1111/febs.15299.
Currin, A., Swainston, N., Day, P.J., Kell, D.B., 2015. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. Chem. Soc. Rev. 44, 1172–1239. https://doi.org/10.1039/c4cs00351a.
Danneels, B., Pinto-Carbó, M., Carlier, A., 2018. Patterns of nucleotide deletion and insertion inferred from bacterial pseudogenes. Genome Biol. Evol. 10, 1792–1802. https://doi.org/10.1093/gbe/evy140.
Dirks-Hofmeister, M.E., Verhaeghe, T., De Winter, K., Desmet, T., 2015. Creating space for large acceptors: rational biocatalyst design for resveratrol glycosylation in an aqueous system. Angew. Chem. 127, 9421–9424. https://doi.org/10.1002/ange.201503605.
Edwards, W.R., Busse, K., Allemann, R.K., Jones, D.D., 2008. Linking the functions of unrelated proteins using a novel directed evolution domain insertion method. Nucleic Acids Res. 36 https://doi.org/10.1093/nar/gkn363.
Edwards, W.R., Williams, A.J., Morris, J.L., Baldwin, A.J., Allemann, R.K., Jones, D.D., 2010. Regulation of β-lactamase activity by remote binding of heme: functional coupling of unrelated proteins through domain insertion. Biochemistry 49, 6541–6549. https://doi.org/10.1021/bi100793y.
Eiben, C.B., Siegel, J.B., Bale, J.B., Cooper, S., Khatib, F., Shen, B.W., Players, F., Stoddard, B.L., Popovic, Z., Baker, D., 2012. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. Nat. Biotechnol. 30, 190–192. https://doi.org/10.1038/nbt.2109.
Eick, G.N., Bridgham, J.T., Anderson, D.P., Harms, M.J., Thornton, J.W., 2017. Robustness of reconstructed ancestral protein functions to statistical uncertainty. Mol. Biol. Evol. 34, 247–261. https://doi.org/10.1093/molbev/msw223.
Emond, S., Mondon, P., Pizzut-Serin, S., Douchy, L., Crozet, F., Bouayadi, K., Kharrat, H., Potocki-Véronèse, G., Monsan, P., Remaud-Simeon, M., 2008. A novel random mutagenesis approach using human mutagenic DNA polymerases to generate enzyme variant libraries. Protein Eng. Des. Sel. 21, 267–274. https://doi.org/10.1093/protein/gzn004.
Emond, S., Petek, M., Kay, E.J., Heames, B., Devenish, S.R.A., Tokuriki, N., Hollfelder, F., 2020. Accessing unexplored regions of sequence space in directed enzyme evolution via insertion/deletion mutagenesis. Nat. Commun. 11 https://doi.org/10.1038/s41467-020-17061-3.
Farlow, A., Long, H., Arnoux, S., Sung, W., Doak, T.G., Nordborg, M., Lynch, M., 2015. The spontaneous mutation rate in the fission yeast *Schizosaccharomyces pombe*. Genetics 201, 737–744. https://doi.org/10.1534/genetics.115.177329.
Foley, G., Mora, A., Ross, C.M., Bottoms, S., Sützl, L., Lamprecht, M.L., Zaugg, J., Essebier, A., Balderson, B., Newell, R., Thomson, R.E.S., Kobe, B., Barnard, R.T., Guddat, L., Schenk, G., Carsten, J., Gumulya, Y., Rost, B., Haltrich, D., Sieber, V., Gillam, E.M.J., Bodén, M., 2022. Engineering indel and substitution variants of diverse and ancient enzymes using graphical representation of ancestral sequence predictions (GRASP). bioRxiv, 2019.12.30.891457.
Franceus, J., Lormans, J., Cools, L., Desmet, T., 2021. Evolution of phosphorylases from *N*-acetylglucosaminide hydrolases in family GH3. ACS Catal. 11, 6225–6233. https://doi.org/10.1021/acscatal.1c00761.
Freeland, S.J., Hurst, L.D., 1998. The genetic code is one in a million. J. Mol. Evol. 47, 238–248. https://doi.org/10.1007/PL00006381.
Fujii, R., Kitaoka, M., Hayashi, K., 2006. RAISE: a simple and novel method of generating random insertion and deletion mutations. Nucleic Acids Res. 34, e30 https://doi.org/10.1093/nar/gnj032.
Garcia-Diaz, M., Kunkel, T.A., 2006. Mechanism of a genetic glissando*: structural biology of indel mutations. Trends Biochem. Sci. 31, 206–214. https://doi.org/10.1016/j.tibs.2006.02.004.
Geyer, R., Mamlouk, A.M., 2018. On the efficiency of the genetic code after frameshift mutations. PeerJ 2018. https://doi.org/10.7717/peerj.4825.
Godfroid, M., Dagan, T., Merker, M., Kohl, T.A., Diel, R., Maurer, F.P., Niemann, S., Kupczok, A., 2020. Insertion and deletion evolution reflects antibiotics selection pressure in a *Mycobacterium tuberculosis* outbreak. PLoS Pathog. 16, e1008357 https://doi.org/10.1371/journal.ppat.1008357.
Gong, C., Bongiorno, P., Martins, A., Stephanou, N.C., Zhu, H., Shuman, S., Glickman, M. S., 2005. Mechanism of nonhomologous end-joining in mycobacteria: a low-fidelity repair system driven by Ku, ligase D and ligase C. Nat. Struct. Mol. Biol. 12, 304–312. https://doi.org/10.1038/nsmb915.
Gonzalez, C.E., Roberts, P., Ostermeier, M., 2019. Fitness effects of single amino acid insertions and deletions in TEM-1 β-lactamase. J. Mol. Biol. 431, 2320–2330. https://doi.org/10.1016/j.jmb.2019.04.030.
Gregory, T.R., 2004. Insertion-deletion biases and the evolution of genome size. Gene 324, 15–34. https://doi.org/10.1016/j.gene.2003.09.030.
Guilliam, T.A., Jozwiakowski, S.K., Ehlinger, A., Barnes, R.P., Rudd, S.G., Bailey, L.J., Skehel, J.M., Eckert, K.A., Chazin, W.J., Doherty, A.J., 2015. Human PrimPol is a highly error-prone polymerase regulated by single-stranded DNA binding proteins. Nucleic Acids Res. 43, 1056–1068. https://doi.org/10.1093/nar/gku1321.
Guntas, G., Mansell, T.J., Kim, J.R., Ostermeier, M., 2005. Directed evolution of protein switches and their application to the creation of ligand-binding proteins. Proc. Natl. Acad. Sci. U. S. A. 102, 11224–11229. https://doi.org/10.1073/pnas.0502673102.
Guo, B., Zou, M., Wagner, A., 2012. Pervasive indels and their evolutionary dynamics after the fish-specific genome duplication. Mol. Biol. Evol. 29, 3005–3022. https://doi.org/10.1093/molbev/mss108.
Gupta, A., Alland, D., 2021. Reversible gene silencing through frameshift indels and frameshift scars provide adaptive plasticity for *Mycobacterium tuberculosis*. Nat. Commun. 12, 1–11. https://doi.org/10.1038/s41467-021-25055-y.
Hamilton, W.L., Claessens, A., Otto, T.D., Kekre, M., Fairhurst, R.M., Rayner, J.C., Kwiatkowski, D., 2017. Extreme mutation bias and high AT content in *Plasmodium falciparum*. Nucleic Acids Res. 45, 1889–1901. https://doi.org/10.1093/nar/gkw1259.
Hawwa, R., Larsen, S.D., Ratia, K., Mesecar, A.D., 2009. Structure-based and random mutagenesis approaches increase the organophosphate-degrading activity of a Phosphotriesterase homologue from *Deinococcus radiodurans*. J. Mol. Biol. 393, 36–57. https://doi.org/10.1016/j.jmb.2009.06.083.
Hayes, F., Hallet, B., 2000. Pentapeptide scanning mutagenesis: encouraging old proteins to execute unusual tricks. Trends Microbiol. 8, 571–577. https://doi.org/10.1016/S0966-842X(00)01857-6.
Hayes, F., Hallet, B., Cao, Y., 1997. Insertion mutagenesis as a tool in the modification of protein function. J. Biol. Chem. 272, 28833–28836. https://doi.org/10.1074/jbc.272.46.28833.
Heinemann, P.M., Armbruster, D., Hauer, B., 2021. Active-site loop variations adjust activity and selectivity of the cumene dioxygenase. Nat. Commun. 12, 1–12. https://doi.org/10.1038/s41467-021-21328-8.
Hida, K., Won, S.Y., Di Pasquale, G., Hanes, J., Chiorini, J.A., Ostermeier, M., 2010. Sites in the AAV5 capsid tolerant to deletions and tandem duplications. Arch. Biochem. Biophys. 496, 1–8. https://doi.org/10.1016/j.abb.2010.01.009.
Hoque, M.A., Zhang, Y., Chen, L., Yang, G., Khatun, M.A., Chen, H., Hao, L., Feng, Y., 2017. Stepwise loop insertion strategy for active site remodeling to generate novel enzyme functions. ACS Chem. Biol. 12, 1188–1193. https://doi.org/10.1021/acschembio.7b00018.
Hwang, J., Cho, K.H., Song, H., Yi, H., Kim, H.S., 2014. Deletion mutations conferring substrate spectrum extension in the class A β-lactamase. Antimicrob. Agents Chemother. 58, 6265–6269. https://doi.org/10.1128/AAC.02648-14.
Jackson, E.L., Spielman, S.J., Wilke, C.O., 2017. Computational prediction of the tolerance to amino-acid deletion in green-fluorescent protein. PLoS One 12, 1–16. https://doi.org/10.1371/journal.pone.0164905.
Jones, D.D., 2005. Triplet nucleotide removal at random positions in a target gene: the tolerance of TEM-1 β-lactamase to an amino acid deletion. Nucleic Acids Res. 33, 1–8. https://doi.org/10.1093/nar/gni077.
Jones, S., Stewart, M., Michie, A., Swindells, M.B., Orengo, C., Thornton, J.M., 1998. Domain assignment for protein structures using a consensus approach: characterization and analysis. Protein Sci. 7, 233–242. https://doi.org/10.1002/pro.5560070202.
Kaltenbach, M., Burke, J.R., Dindo, M., Pabis, A., Munsberg, F.S., Rabin, A., Kamerlin, S. C.L., Noel, J.P., Tawfik, D.S., 2018. Evolution of chalcone isomerase from a noncatalytic ancestor. Nat. Chem. Biol. 14, 548–555. https://doi.org/10.1038/s41589-018-0042-3.
Kanwar, M., Wright, R.C., Date, A., Tullman, J., Ostermeier, M., 2013. Protein switch engineering by domain insertion. Methods Enzymol. 523, 369–388. https://doi.org/10.1016/B978-0-12-394292-0.00017-5.Protein.
Kashiwagi, K., Isogai, Y., Nishiguchi, K.I., Shiba, K., 2006. Frame shuffling: a novel method for in vitro protein evolution. Protein Eng. Des. Sel. 19, 135–140. https://doi.org/10.1093/protein/gzj008.
Kim, R.G., Guo, J.T., 2010. Systematic analysis of short internal indels and their impact on protein folding. BMC Struct. Biol. 10 https://doi.org/10.1186/1472-6807-10-24.
Kipnis, Y., Dellus-Gur, E., Tawfik, D.S., 2012. TRINS: a method for gene modification by randomized tandem repeat insertions. Protein Eng. Des. Sel. 25, 437–444. https://doi.org/10.1093/protein/gzs023.
Kroutil, L.C., Register, K., Bebenek, K., Kunkel, T.A., 1996. Exonucleolytic proofreading during replication of repetitive DNA. Biochemistry 35, 1046–1053. https://doi.org/10.1021/bi952178h.
Kucukyildirim, S., Behringer, M., Sung, W., Brock, D.A., Doak, T.G., Mergen, H., Queller, D.C., Strassmann, J.E., Lynch, M., 2020. Low base-substitution mutation rate but high rate of slippage mutations in the sequence repeat-rich genome of *Dictyostelium discoideum*. G3 Genes Genomes Genet. 10, 3445–3452. https://doi.org/10.1534/g3.120.401578.
Light, S., Sagit, R., Ekman, D., Elofsson, A., 2013. Long indels are disordered: a study of disorder and indels in homologous eukaryotic proteins. Biochim. Biophys. Acta,

Proteins Proteomics 1834, 890–897. https://doi.org/10.1016/j.bbapap.2013.01.002.
Lin, M., Whitmire, S., Chen, J., Farrel, A., Shi, X., Guo, J.T., 2017. Effects of short indels on protein structure and function in human genomes. Sci. Rep. 7, 1–9. https://doi.org/10.1038/s41598-017-09287-x.
Liu, S. su, Wei, X., Ji, Q., Xin, X., Jiang, B., Liu, J., 2016. A facile and efficient transposon mutagenesis method for generation of multi-codon deletions in protein sequences. J. Biotechnol. 227, 27–34. https://doi.org/10.1016/j.jbiotec.2016.03.038.
Liu, C., Zhao, J., Liu, J., Guo, X., Rao, D., Liu, H., Zheng, P., Sun, J., Ma, Y., 2019. Simultaneously improving the activity and thermostability of a new proline 4-hydroxylase by loop grafting and site-directed mutagenesis. Appl. Microbiol. Biotechnol. 103, 265–277. https://doi.org/10.1007/s00253-018-9410-x.
Long, H., Miller, S.F., Williams, E., Lynch, M., 2018. Specificity of the DNA mismatch repair system (MMR) and mutagenesis bias in bacteria. Mol. Biol. Evol. 35, 2414–2421. https://doi.org/10.1093/molbev/msy134.
Lynch, M., Ackerman, M.S., Gout, J.F., Long, H., Sung, W., Thomas, W.K., Foster, P.L., 2016. Genetic drift, selection and the evolution of the mutation rate. Nat. Rev. Genet. 17, 704–714. https://doi.org/10.1038/nrg.2016.104.
Maria-Solano, M.A., Serrano-Hervás, E., Romero-Rivera, A., Iglesias-Fernández, J., Osuna, S., 2018. Role of conformational dynamics in the evolution of novel enzyme function. Chem. Commun. 54, 6622–6634. https://doi.org/10.1039/c8cc02426j.
Matsuura, T., Miyai, K., Trakulnaleamsai, S., Yomo, T., Shima, Y., Miki, S., Yamamoto, K., Urabe, I., 1999. Evolutionary molecular engineering by random elongation mutagenesis. Nat. Biotechnol. 17, 58–61. https://doi.org/10.1385/1-59259-194-9:221.
Maynard Smith, J., 1970. Natural selection and the concept of a protein space. Nature 225, 563–564. https://doi.org/10.1038/225563a0.
McDonald, M.J., Wang, W.C., Da Huang, H., Leu, J.Y., 2011. Clusters of Nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. PLoS Biol. 9 https://doi.org/10.1371/journal.pbio.1000622.
Mills, R.E., Pittard, W.S., Mullaney, J.M., Farooq, U., Creasy, T.H., Mahurkar, A.A., Kemeza, D.M., Strassler, D.S., Ponting, C.P., Webber, C., Devine, S.E., 2011. Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res. 21, 830–839. https://doi.org/10.1101/gr.115907.110.
Morelli, A., Cabezas, Y., Mills, L.J., Seelig, B., 2017. Extensive libraries of gene truncation variants generated by in vitro transposition. Nucleic Acids Res. 45, e78 https://doi.org/10.1093/nar/gkx030.
Morita, R., Nakane, S., Shimada, A., Inoue, M., Iino, H., Wakamatsu, T., Fukui, K., Nakagawa, N., Masui, R., Kuramitsu, S., 2010. Molecular mechanisms of the whole DNA repair system: a comparison of bacterial and eukaryotic systems. J. Nucleic Acids 2010. https://doi.org/10.4061/2010/179594.
Murakami, H., Hohsaka, T., Sisido, M., 2002. Random insertion and deletion of arbitrary number of bases for codon-based random mutation of DNAs. Nat. Biotechnol. 20, 76–81. https://doi.org/10.1038/nbt0102-76.
Murphy, P.M., Bolduc, J.M., Gallaher, J.L., Stoddard, B.L., Baker, D., 2009. Alteration of enzyme specificity by computational loop remodeling and design. Proc. Natl. Acad. Sci. U. S. A. 106, 9215–9220. https://doi.org/10.1073/pnas.0811070106.
Musil, M., Khan, R.T., Beier, A., Stourac, J., Konegger, H., Damborsky, J., Bednar, D., 2021. FireProtASR: a web server for fully automated ancestral sequence reconstruction. Brief. Bioinform. 22, 1–11. https://doi.org/10.1093/bib/bbaa337.
Nestl, B., Hauer, B., 2014. Engineering of flexible loops in enzymes. ACS Catal. 4, 3201–3211.
Ostermeier, M., 2005. Engineering allosteric protein switches by domain insertion. Protein Eng. Des. Sel. 18, 359–364. https://doi.org/10.1093/protein/gzi048.
Osuna, J., Yáñez, J., Soberón, X., Gaytán, P., 2004. Protein evolution by codon-based random deletions. Nucleic Acids Res. 32, e136 https://doi.org/10.1093/nar/gnh135.
Park, H.S., Nam, S.H., Lee, J.K., Yoon, C.N., Mannervik, B., Benkovic, S.J., Kim, H.S., 2006. Design and evolution of new catalytic activity with an existing protein scaffold. Science 311, 535–538. https://doi.org/10.1126/science.1118953.
Pascarella, S., Argos, P., 1992. Analysis of insertions/deletions in protein structures. J. Mol. Biol. 224, 461–471. https://doi.org/10.1016/0022-2836(92)91008-D.
Patzoldt, W.L., Hager, A.G., McCormick, J.S., Tranel, P.J., 2006. A codon deletion confers resistance to herbicides inhibiting protoporphyrinogen oxidase. Proc. Natl. Acad. Sci. U. S. A. 103, 12329–12334. https://doi.org/10.1073/pnas.0603137103.
Pellegrini, M., Renda, M.E., Vecchio, A., 2012. Ab initio detection of fuzzy amino acid tandem repeats in protein sequences. BMC Bioinformatics 13. https://doi.org/10.1186/1471-2105-13-S3-S8.
Pikkemaat, M.G., Janssen, D.B., 2002. Generating segmental mutations in haloalkane dehalogenase: a novel part in the directed evolution toolbox. Nucleic Acids Res. 30 https://doi.org/10.1093/nar/30.8.e35.
Pinney, M.M., Mokhtari, D.A., Akiva, E., Yabukarski, F., Sanchez, D.M., Liang, R., Doukov, T., Martinez, T.J., Babbitt, P.C., Herschlag, D., 2021. Parallel molecular mechanisms for enzyme temperature adaptation. Science 371. https://doi.org/10.1126/science.aay2784 eaay2784.
Planas-Iglesias, J., Ulbrich, P., Pinto, G.P., Schenkmayerova, A., Damborsky, J., Kozlikova, B., Bednar, D., 2022. LoopGrafter: web tool for transplanting dynamical loops for protein engineering. Nucleic Acids Res. gkac249.
Reetz, M.T., Carballeira, J.D., 2007. Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. Nat. Protoc. 2, 891–903. https://doi.org/10.1038/nprot.2007.72.
Reetz, M.T., Wu, S., 2008. Greatly reduced amino acid alphabets in directed evolution: making the right choice for saturation mutagenesis at homologous enzyme positions. Chem. Commun. 5499–5501. https://doi.org/10.1039/b813388c.
Reetz, M.T., Bocola, M., Carballeira, J.D., Zha, D., Vogel, A., 2005. Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test. Angew. Chem. Int. Ed. Eng. 44, 4192–4196. https://doi.org/10.1002/anie.200500767.
Reetz, M.T., Kahakeaw, D., Lohmer, R., 2008. Addressing the numbers problem in directed evolution. ChemBioChem 9, 1797–1804. https://doi.org/10.1002/cbic.200800298.
Reich, S., Kress, N., Nestl, B.M., Hauer, B., 2014. Variations in the stability of NCR ene reductase by rational enzyme loop modulation. J. Struct. Biol. 185, 228–233. https://doi.org/10.1016/j.jsb.2013.04.004.
Reich, S., Nestl, B.M., Hauer, B., 2016. Loop-grafted old yellow enzymes in the bienzymatic cascade reduction of allylic alcohols. ChemBioChem 17, 561–565. https://doi.org/10.1002/cbic.201500604.
Rockah-Shmuel, L., Tóth-Petróczy, Á., Sela, A., Wurtzel, O., Sorek, R., Tawfik, D.S., 2013. Correlated occurrence and bypass of frame-shifting insertion-deletions (InDels) to give functional proteins. PLoS Genet. 9 https://doi.org/10.1371/journal.pgen.1003882.
Romero, P.A., Arnold, F.H., 2009. Exploring protein fitness landscapes by directed evolution. Nat. Rev. Mol. Cell Biol. https://doi.org/10.1038/nrm2805.
Ross, C.M., Foley, G., Boden, M., Gillam, E.M.J., 2022. Using the evolutionary history of proteins to engineer insertion-deletion mutants from robust, ancestral templates using graphical representation of ancestral sequence predictions (GRASP). Methods Mol. Biol. 85–110.
Sandhya, S., Rani, S.S., Pankaj, B., Govind, M.K., Offmann, B., Srinivasan, N., Sowdhamini, R., 2009. Length variations amongst protein domain superfamilies and consequences on structure and function. PLoS One 4, e4981. https://doi.org/10.1371/journal.pone.0004981.
Savile, C.K., Janey, J.M., Mundorff, E.C., Moore, J.C., Tam, S., Jarvis, W.R., Colbeck, J.C., Krebber, A., Fleitz, F.J., Brands, J., Devine, P.N., Huisman, G.W., Hughes, G.J., 2010. Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. Science 329, 305–309. https://doi.org/10.1126/science.1188934.
Saxena, A.S., Salomon, M.P., Matsuba, C., Yeh, S.D., Baer, C.F., 2019. Evolution of the mutational process under relaxed selection in *Caenorhabditis elegans*. Mol. Biol. Evol. 36, 239–251. https://doi.org/10.1093/molbev/msy213.
Schenkmayerova, A., Pinto, G.P., Toul, M., Marek, M., Hernychova, L., Planas-Iglesias, J., Daniel Liskova, V., Pluskal, D., Vasina, M., Emond, S., Dörr, M., Chaloupkova, R., Bednar, D., Prokop, Z., Hollfelder, F., Bornscheuer, U.T., Damborsky, J., 2021. Engineering the protein dynamics of an ancestral luciferase. Nat. Commun. 12, 1–36. https://doi.org/10.1038/s41467-021-23450-z.
Schüler, A., Bornberg-Bauer, E., 2016. Evolution of protein domain repeats in metazoa. Mol. Biol. Evol. 33, 3170–3182. https://doi.org/10.1093/molbev/msw194.
Sehn, J.K., 2015. Insertions and deletions (Indels). Clinical Genomics. 129–150. https://doi.org/10.1016/B978-0-12-404748-8.00009-5.
Seligmann, H., Pollock, D.D., 2004. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. DNA Cell Biol. 23, 701–705. https://doi.org/10.1089/dna.2004.23.701.
Sheludko, Y.V., Fessner, W.D., 2020. Winning the numbers game in enzyme evolution – fast screening methods for improved biotechnology proteins. Curr. Opin. Struct. Biol. 63, 123–133. https://doi.org/10.1016/j.sbi.2020.05.003.
Simm, A.M., Baldwin, A.J., Busse, K., Jones, D.D., 2007. Investigating protein structural plasticity by surveying the consequence of an amino acid deletion from TEM-1 β-lactamase. FEBS Lett. 581, 3904–3908. https://doi.org/10.1016/j.febslet.2007.07.018.
Skamaki, K., Emond, S., Chodorge, M., Andrews, J., Rees, D.G., Cannon, D., Popovic, B., Buchanan, A., Minter, R.R., Hollfelder, F., 2020. In vitro evolution of antibody affinity via insertional scanning mutagenesis of an entire antibody variable region. Proc. Natl. Acad. Sci. U. S. A. 117, 27307–27318. https://doi.org/10.1073/pnas.2002954117.
Spence, M.A., Kaczmarski, J.A., Saunders, J.W., Jackson, C.J., 2021. Ancestral sequence reconstruction for protein engineers. Curr. Opin. Struct. Biol. 69, 131–141. https://doi.org/10.1016/j.sbi.2021.04.001.
Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E., Inouye, M., 1966. Frameshift mutations and the genetic code. This paper is dedicated to Professor Theodosius Dobzhansky on the occasion of his 66th birthday. Cold Spring Harb. Symp. Quant. Biol. 31, 77–84. https://doi.org/10.1101/SQB.1966.031.01.014.
Studer, R.A., Dessailly, B.H., Orengo, C.A., 2013. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochem. J. 449, 581–594. https://doi.org/10.1042/BJ20121221.
Sung, W., Tucker, A.E., Doak, T.G., Choi, E., Thomas, W.K., Lynch, M., 2012. Extraordinary genome stability in the ciliate *Paramecium tetraurelia*. Proc. Natl. Acad. Sci. U. S. A. 109, 19339–19344. https://doi.org/10.1073/pnas.1210663109.
Sung, W., Ackerman, M.S., Dillon, M.M., Platt, T.G., Fuqua, C., Cooper, V.S., Lynch, M., 2016. Evolution of the insertion-deletion mutation rate across the tree of life. G3 Genes Genomes Genet. 6, 2583–2591. https://doi.org/10.1534/g3.116.030890.
Surkont, J., Diekmann, Y., Ryder, P.V., Pereira-Leal, J.B., 2015. Coiled-coil length: size does matter. Proteins Struct. Funct. Bioinforma. 83, 2162–2169. https://doi.org/10.1002/prot.24932.
Takano, K., Okamoto, T., Okada, J., Tanaka, S.I., Angkawidjaja, C., Koga, Y., Kanaya, S., 2011. Stabilization by fusion to the C-terminus of hyperthermophile *Sulfolobus tokodaii* RNase HI: a possibility of protein stabilization tag. PLoS One 6, 1–7. https://doi.org/10.1371/journal.pone.0016226.
Tawfik, D.S., 2006. Loop grafting and the origins of enzyme species. Science. https://doi.org/10.1126/science.1123883.

Taylor, M.S., Ponting, C.P., Copley, R.R., 2004. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. Genome Res. 14, 555–566. https://doi.org/10.1101/gr.1977804.

Tizei, P.A.G., Harris, E., Withanage, S., Renders, M., Pinheiro, V.B., 2021. A novel framework for engineering protein loops exploring length and compositional variation. Sci. Rep. 11, 1–13. https://doi.org/10.1038/s41598-021-88708-4.

Toledo-patiño, S., Pascarelli, S., Uechi, G., Laurino, P., 2022. Insertions and deletions mediated functional divergence of Rossmann fold enzymes. bioRxiv, 2022.05.16.491946.

Tóth-Petróczy, A., Tawfik, D.S., 2013. Protein insertions and deletions enabled by neutral roaming in sequence space. Mol. Biol. Evol. 30, 761–771. https://doi.org/10.1093/molbev/mst003.

Viguera, E., Canceill, D., Ehrlich, S.D., 2001. Replication slippage involves DNA polymerase pausing and dissociation. EMBO J. 20, 2587–2595. https://doi.org/10.1093/emboj/20.10.2587.

Vojcic, L., Pitzler, C., Körfer, G., Jakob, F., Martinez, Ronny, Maurer, K.H., Schwaneberg, U., 2015. Advances in protease engineering for laundry detergents. New Biotechnol. 32, 629–634. https://doi.org/10.1016/j.nbt.2014.12.010.

Wójcik, M., Szala, K., van Merkerk, R., Quax, W.J., Boersma, Y.L., 2020. Engineering the specificity of *streptococcus pyogenes* sortase A by loop grafting. Proteins Struct. Funct. Bioinforma. 88, 1394–1400. https://doi.org/10.1002/prot.25958.

Wu, Z., Jennifer Kan, S.B., Lewis, R.D., Wittmann, B.J., Arnold, F.H., 2019. Machine learning-assisted directed protein evolution with combinatorial libraries. Proc. Natl. Acad. Sci. U. S. A. 116, 8852–8858. https://doi.org/10.1073/pnas.1901979116.

Xiang, D.F., Kolb, P., Fedorov, A.A., Meier, M.M., Fedorov, L.V., Nguyen, T.T., Sterner, R., Almo, S.C., Shoichet, B.K., Raushel, F.M., 2009. Functional annotation and three-dimensional structure of Dr0930 from *Deinococcus radiodurans*, a close relative of phosphotriesterase in the amidohydrolase superfamily. Biochemistry 48, 2237–2247. https://doi.org/10.1021/bi802274f.

Xu, H., Zhang, J., 2021. On the origin of frameshift-robustness of the standard genetic code. Mol. Biol. Evol. 38, 4301–4309. https://doi.org/10.1093/molbev/msab164.

Yang, H., Liu, L., Li, J., Chen, J., Du, G., 2015. Rational design to improve protein thermostability: recent advances and prospects. ChemBioEng Rev. 2, 87–94. https://doi.org/10.1002/cben.201400032.

Yang, G., Anderson, D.W., Baier, F., Dohmen, E., Hong, N., Carr, P.D., Kamerlin, S.C.L., Jackson, C.J., Bornberg-Bauer, E., Tokuriki, N., 2019. Higher-order epistasis shapes the fitness landscape of a xenobiotic-degrading enzyme. Nat. Chem. Biol. 15, 1120–1128. https://doi.org/10.1038/s41589-019-0386-3.

Yi, H., Kim, K., Cho, K.H., Jung, O., Kim, H.S., 2012. Substrate spectrum extension of PenA in *Burkholderia thailandensis* with a single amino acid deletion, Glu168del. Antimicrob. Agents Chemother. 56, 4005–4008. https://doi.org/10.1128/AAC.00598-12.

Zhang, Z., Wang, J., Gong, Y., Li, Y., 2018. Contributions of substitutions and indels to the structural variations in ancient protein superfamilies. BMC Genomics 19, 1–9. https://doi.org/10.1186/s12864-018-5178-8.

Zhou, Y., Zhang, H., Bao, H., Wang, X., Wang, R., 2017. The lytic activity of recombinant phage lysin LysKΔamidase against staphylococcal strains associated with bovine and human infections in the Jiangsu province of China. Res. Vet. Sci. 111, 113–119. https://doi.org/10.1016/j.rvsc.2017.02.011.