Biotechnology Advances xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Biotechnology Advances

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

Research review paper

Synthetic fusion protein design and applications

Kai Yu [a], Chengcheng Liu [b], Byung-Gee Kim [c], Dong-Yup Lee [a,b,*]

[a] Department of Chemical and Biomolecular Engineering, Synthetic Biology Research Consortium, National University of Singapore, 4 Engineering Drive 4, Singapore 117585, Singapore
[b] Bioprocessing Technology Institute, Agency for Science, Technology and Research (A*STAR), 20 Biopolis Way, #06-01, Singapore 138668, Singapore
[c] School of Chemical and Biological Engineering, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul 151742, South Korea

ARTICLE INFO

Article history:
Received 5 August 2014
Received in revised form 10 October 2014
Accepted 11 November 2014
Available online xxxx

Keywords:
Fusion protein
Linker design
Domain insertion
Protein therapeutics
Synthetic protein engineering

ABSTRACT

Synthetic fusion proteins can be designed to achieve improved properties or new functionality by synergistically incorporating multiple proteins into one complex. The fusion of two or more protein domains enhances bioactivities or generates novel functional combinations with a wide range of biotechnological and (bio)pharmaceutical applications. In this review, initially, we summarize the commonly used approaches for constructing fusion proteins. For each approach, the design strategy and desired properties are elaborated with examples of recent studies in the areas of biocatalysts, protein switches and bio-therapeutics. Subsequently, the progress in structural prediction of fusion proteins is presented, which can potentially facilitate the structure-based systematic design of fusion proteins toward identifying the best combinations of fusion partners. Finally, the current challenges and future directions in this field are discussed.

Contents

Introduction

Fusion proteins are a class of proteins with two or more different protein domains integrated into one molecule. A wide variety of naturally occurring multidomain fusion proteins have been characterized with different architecture to meet the functional requirements of living organisms at the molecular level (Aroul-Selvam et al., 2004). The modular organization of protein domains observed in natural fusion proteins has been identified as an important evolutionary phenomenon (Long, 2000). For example, some enzymes have been fused for efficient multi-step biocatalysis in metabolic pathways, such as pyrroline-5-carboxylate synthase (P5CS) (Pérez-Arellano et al., 2010) and acetyl-CoA carboxylase (ACC) (Tong, 2005). Some fusion proteins which resulted from chromosomal rearrangement are specifically related to human diseases, e.g., chronic myeloid leukemia (Melo, 1996; Rapin and Porse, 2014).

* Corresponding author at: Department of Chemical and Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117585. Tel.: +65 6516 6907; fax: +65 6779 1936.
E-mail address: cheld@nus.edu.sg (D.-Y. Lee).

http://dx.doi.org/10.1016/j.biotechadv.2014.11.005
0734-9750/

Over the years, researchers have been mimicking nature's strategy to create artificial fusion proteins by using recombinant DNA technology or post-translational modification approaches with various applications. The earliest application can be traced back to the use of peptide/protein tags to enable one-step protein purification using affinity chromatography techniques (Bell et al., 2013; Terpe, 2003). The smallest peptide tags consist of less than 10 amino acids, e.g., the polyarginine-tag, polyhistidine-tag, FLAG octapeptide and Strep-tag. Fusing these tags at the N- or C-terminus of the target protein usually maintains its functionality and structural property without disruption. Larger tags (more than 200 amino acids) such as glutathione-S-transferase (GST) and maltose-binding protein (MBP) can also be used as affinity tags although they often need to be removed afterward (Terpe, 2003). Fluorescent proteins such as the green fluorescent protein (GFP) are also a popular group of fusion partners frequently used as a biosensor for monitoring signaling molecules (Pedelacq et al., 2006), or as a reporter in bioimaging to reveal a target protein's expression level, folding state, steady-state distribution and dynamics (Tsien, 1998). In the area of enzyme engineering, early attempts were made to create two-protein fusions either to increase consecutive enzyme reaction rates or to generate bifunctional enzymes (Béguin, 1999; Lindbladh et al., 1992). Recently, artificial fusion proteins have also been constructed as novel protein switches in some synthetic biology applications (Grünberg and Serrano, 2010). In addition, recombinant fusion proteins, such as engineered antibody fragments, have become a new class of therapeutic agents (Schmidt, 2013).

In light of the broad applications of synthetic fusion proteins and the growing interest in their construction, providing a comprehensive review to summarize the commonly used design principles and construction strategies would be extremely useful. To this end, we focus on three distinct approaches, tandem fusion, domain insertion and post-translational conjugation, with some recent studies showcasing their applications as biocatalysts, protein switches and bio-therapeutics. This is followed by a presentation of recent progress on the computational prediction of fusion protein structure. Finally, the current challenges in the construction of synthetic fusion proteins and future directions are discussed. Although the presented methods and examples herein mainly focus on two-component fusion proteins, the underlying principles and strategies are also applicable to multiple-component fusion systems.

Design and construction of synthetic fusion proteins

Tandem fusion

With the targeted application in mind, the choice of component proteins to be fused is relatively straightforward. Ideally, fusion partners should be well studied and physico-chemically compatible (e.g., optimal working pH, temperature, ionic strength and effects of inhibitors). The order of fusion partners in the polypeptide chain is often critical, as the placement of one domain can affect the localization and functionality of the other (Sachdev and Chirgwin, 1998). Hence, in the case of a two-component fusion, two sequence combinations should be attempted unless the possible effect of the order is known. After a pair of fusion partners is selected, the simplest method of combining them is an end-to-end genetic fusion, wherein their coding genes are joined together and expressed as a single peptide chain in a suitable host organism. Direct tandem fusion is simple and works in some cases where flexible and unstructured N- or C-terminal regions of the component proteins act as a "bridge" to provide enough space between protein domains for correct folding (Rizk et al., 2012). However, this strategy fails when the free N- or C-terminus to be tethered is essential to the parent protein function and/or is not flexible or long enough to avoid steric hindrance, which reduces the degrees of freedom in protein dynamics and may give rise to undesirable outcomes such as protein misfolding, aggregation leading to inclusion body formation, low yield in protein production and impaired bioactivity. For this reason, linker peptides are mostly required to connect fusion partners with an intention to better maintain their individual structures and functions (Fig. 1).

Linkers are ubiquitously observed in naturally occurring multidomain proteins with the function of maintaining necessary distance to reduce steric hindrance and/or permit favorable domain–domain interaction between two protein moieties. Based on this observation, researchers have employed various types of naturally occurring linkers in their synthetic fusion constructs. For example, the immunoglobulin hinge region functions as a linker in many recombinant therapeutic proteins, particularly in engineered antibody constructs (Pack et al., 1995; Rheinnecker et al., 1996; Wu et al., 2007). The linker regions in natural enzymes are also good candidates for connecting fusion partners. For instance, proline- and hydroxylamine-rich cellulase and xylanase linkers keep an extended conformation and are protected from proteolysis because of O-glycosylation (Rizk et al., 2012). In addition to the natural linkers, researchers have devised a multitude of artificial linkers, which can be subdivided into three categories: flexible, rigid and *in vivo* cleavable linkers. The most widely used flexible linker sequences are $(Gly)_n$ (Sabourin et al., 2007) and $(Gly\text{-}Gly\text{-}Gly\text{-}Gly\text{-}Ser)_n$ (Huston et al., 1988), where linker length can be adjusted by the copy number "n". One downside, however, of this class of linkers is the high homologous repeats in their DNA coding sequences, which can possibly reduce protein expression levels and may require appropriate codon pair selection (Trinh et al., 2004). When sufficient separation of protein domains is desired, rigid linkers may be preferable. For example, a helix-forming linker with the $(EAAAK)_n$ motif was designed for this purpose (Arai et al., 2001). This linker is stabilized by its intrinsic Glu^--Lys^+ salt bridges to generate a rigid and extended conformation. Introducing a proline-rich sequence, such as $(XP)_n$ with X designating any amino acid, can also constrain the linker to an extended conformation with relatively limited flexibility (Morris et al., 1992). In other cases, cleavable linkers are introduced to release free functional domains *in vivo* when the recombinant fusion protein is exposed to reducing agents or proteases. To construct a fusion protein consisting of granulocyte-colony stimulating factor (G-CSF) and transferrin (Tf), Chen et al. (2010) designed an *in vivo* cleavable linker utilizing the reversible nature of the disulfide bond as well as a thrombin-sensitive sequence (Fig. 2a). As demonstrated in this case, protease-sensitive sequences can be deliberately incorporated into a linker that can be cleaved by a specific protease to realize the targeted activation of fusion protein at specific sites *in vivo* (Schulte, 2009). MEROPS, a protease specificity database (Rawlings et al., 2012), is a potent tool for assessing the proteolytic stability of linkers so that cleavable linkers can be designed and unwanted protease cleavage sites can be eliminated, which would otherwise cause undesired linker cleavage (Kavoosi et al., 2007).

The properties, functions and examples of fusion protein linkers have been recently reviewed and discussed (Chen et al., 2013). Generally, several linker properties are worth careful consideration: length, amino acid composition, hydrophobicity, sensitivity to proteases, secondary structure and possible interaction with fusion partners. The linkers appear to be a very deciding factor for a successful fusion construction. Nevertheless, the selection of a linker is mostly *ad hoc* and still remains an underexplored area in fusion protein engineering. Besides, it is hard to assume that a linker suitable for one case will be applicable to others, as each type of linker has its specific characteristics. To aid the rational design and selection of linkers, two interesting bioinformatics tools have been developed. An online program, LINKER, was established for automatically generating linker sequences for fusion proteins (Crasto and Feng, 2000; Xue et al., 2004). Note that the server website is no longer accessible, most probably due to the lack of maintenance. Similarly, George and Heringa (2002) developed a web-based linker database (http://www.ibi.vu.nl/programs/linkerdbwww/) that provides a group of linker candidates satisfying the user-specified queries such as the length, sequence and secondary structure of the linker. However, there has been no update since it was released. As it will be beneficial to make more such databases and tools for linker

Please cite this article as: Yu K, et al, Synthetic fusion protein design and applications, Biotechnol Adv (2014), http://dx.doi.org/10.1016/j.biotechadv.2014.11.005