Protein engineering has been an invaluable tool for the deciphering of protein folding and function and in the understanding of biological signaling networks. From an applied point of view it has been of paramount importance in biotechnological and biopharmaceutical products and applications. Traditionally, the protein engineering tools of choice were ‘classical’ rational design, or directed evolution-based methods. In recent years, a third tool has matured: computational protein design (CPD). In this review, we summarize the underlying principles of CPD and discuss its application for understanding and modifying biological systems. Three main applications of the use of protein design will be highlighted and reviewed: artificially rewiring of signal transduction networks, prediction and generation of large-scale in silico interaction networks and using protein design to manipulate gene expression.

Keywords: protein design/protein interaction networks/signaling/synthetic biology/systems biology

Introduction

The use of structure-based computational algorithms in recent years has resulted in impressive results such as several de novo designed proteins, as for example the design of a novel amino acid sequence able to fold into a predetermined structure (Dahiyat and Mayo, 1997) or the design of a new globular fold and structure (Kuhlman et al., 2003). Computational protein design (CPD) algorithms combine biophysical knowledge with in silico scoring to select optimal solutions, merging force fields (Mendes et al., 2002; Boas and Harbury, 2007) with search algorithms (Voigt et al., 2000; Dahiyat, 2006). The basic principle of CPD algorithms is shown in Fig. 1. Aside from designing novel folds, structure-based CPD algorithms have extensively been used to redesign (improve) properties of existing proteins. For example, these algorithms were successfully used to enhance the stability of various proteins (Malakauskas and Mayo, 1998; Filikov et al., 2002; Luo et al., 2002; Ventura et al., 2002; Dantas et al., 2003; van der Sloat et al., 2004; Korkegian et al., 2005).

In recent years, computational design has also been employed in the redesign of affinity and specificity of a wide variety of protein–protein interactions (Selzer et al., 2000; Reina et al., 2002; Shifman and Mayo, 2002, 2003; Havranek and Harbury, 2003; Steed et al., 2003; Kiel et al., 2004; Kortemme et al., 2004a; Bolon et al., 2005; Song et al., 2006; van der Sloat et al., 2006; Potapov et al., 2008; Reynolds et al., 2008; Grigoryan et al., 2009), whereas this previously was the realm of phage display and related techniques. In addition to redesign, CPD algorithms can also be employed in computational alanine scanning (Massova and Kollman, 1999; Kortemme et al., 2004b), and in the analysis of protein–protein interfaces analogous to experimental alanine-scanning mutagenesis (Clackson and Wells, 1995). Not only protein–protein interactions are subjected to design, but also proteins with small molecule receptor functionality (Looger et al., 2003; Allert et al., 2004; Cochran et al., 2005; Matthews et al., 2008), enzyme–substrate (Bolon and Mayo, 2001; Jiang et al., 2008; Rothlisberger et al., 2008; Chen et al., 2009; Damborsky and Brezovsky, 2009) and protein–nucleic acid interactions (Ashworth et al., 2006) have been modified—or were created—using computational design approaches (Das and Baker, 2008).

In this review we discuss the excellent opportunity which these developments provide for systems biology and synthetic biology to develop protein-based tools to gain a deeper understanding of signaling networks or to engineer novel or modified signaling and metabolic networks, respectively.

Applications of protein design

Applications that use protein design as a molecular tool to study cellular and biological functions can be grouped according to their physical interference within a cellular system (Fig. 2): at the surface of the cell, by modifying the binding affinities and specificities of receptors and ligands, in the cytosol by altering signaling flow, and in the nucleus, by means of altering genetic information.

Changing the input

Cell surface receptors allow external factors to communicate with intracellular signaling networks. A great variety of potential responses, ranging from cell-proliferation to cell death, are possible depending on the activation, de-activation or inhibition of a particular type of receptor or combination of receptors. Designing ligands to obtain receptor-specific agonists or antagonists allows activating or inhibiting intracellular signaling networks and offers a great opportunity to study and modify processes such as cell differentiation, proliferation and apoptosis.

Tumor necrosis factor-alpha (TNF-α) signaling can result in various outcomes, depending on the cellular context, NF-κβ-mediated proliferative programs can be activated but also programmed cell death can be induced (Balkwill, 2009). The currently used TNF-α inhibitory molecules are antibodies or soluble TNF receptors which sequester TNF-α.
Steed et al. (2003) designed dominant negative TNF-alpha (DN-TNF) variants that prevent formation of active TNF-α trimers. These DN-TNF variants were designed to show a significantly reduced binding to the TNF-R1 and TNF-R2 receptors. In the presence of endogenous TNF-α, the subunits exchange and TNF-α/DN-TNF heterotrimers are formed, which are unable to transduce signals, and consequently, signal transduction is inhibited. Experimental validation showed both in vitro and in vivo a large reduction in TNF-α-mediated effects (Steed et al., 2003).

Another member of the TNF ligand family, TNF-related apoptosis-inducing ligand (TRAIL), selectively induces apoptosis in a variety of cancer cells (Ashkenazi, 2002; Kimberley and Screaton, 2004; de Vries et al., 2006). TRAIL is a promiscuous ligand, induces apoptosis by interacting with two death receptors: DR4 and DR5. TRAIL also binds to decoy receptors (DcR1, DcR2 and OPG) that cannot induce apoptosis. Various TRAIL variants were designed using the protein design algorithm FoldX (Guerois et al., 2002; Schymkowitz et al., 2005a, 2005b): DR5-specific variants (van der Sloot et al., 2006), DR4-specific variants (Tur et al., 2008) and variants having higher affinity for DR4 and DR5 (Reis et al., 2009). The DR5-specific variant exhibited a 17-fold increase in apoptosis induction kinetics when compared with TRAIL WT, despite having only a 3-fold improvement in the association rate constant for DR5 binding. With the aid of mathematical modeling it could subsequently be shown that receptor specificity was the most important factor for the observed increase in apoptosis induction rate: the tendency to form heteromeric receptor complexes was much reduced for the receptor-specific variant and the formation rate of active homomeric DR5 complexes was significantly increased. This provides evidence for a kinetic-based antagonism mechanism by the decoy receptors (Szegedi et al., submitted).

Predicting and modifying information flow

Generating cellular interaction networks and predicting network changes in time and space are crucial for understanding and artificially modifying and rewiring cellular systems. Structure-based protein design can be applied in signaling networks in many ways and with multiple applications (Kiel et al., 2008). We like to highlight three main applications: (a) using protein interface modeling to generate large-scale in silico interaction networks, (b) estimating parameters for protein complexes for simulating networks and (c) using protein design to artificially rewire signal transduction/synthetic biology.

Predicting information flow

The finding that similar sequences usually have a similar fold and domains with a similar fold interact in a similar way is the basis of the pioneering work of Aloy and Russell for the family-wide prediction of protein interactions (Aloy and Russell, 2002, 2005a, 2005b): DR5-specific variants (van der Sloot et al., 2006), DR4-specific variants (Tur et al., 2008) and variants having higher affinity for DR4 and DR5 (Reis et al., 2009). The DR5-specific variant exhibited a 17-fold increase in apoptosis induction kinetics when compared with TRAIL WT, despite having only a 3-fold improvement in the association rate constant for DR5 binding. With the aid of mathematical modeling it could subsequently be shown that receptor specificity was the most important factor for the observed increase in apoptosis induction rate: the tendency to form heteromeric receptor complexes was much reduced for the receptor-specific variant and the formation rate of active homomeric DR5 complexes was significantly increased. This provides evidence for a kinetic-based antagonism mechanism by the decoy receptors (Szegedi et al., submitted).
by combining experimental and modeling approaches that the Ras–Raf association rate constants have more effect on signal transduction than the dissociation rate constants. However, this effect was strongly cell-type specific, with the behavior being crucially dependent on the underlying network topology (strong/weak negative feedback mechanisms).

We like to highlight another important aspect of using protein design as a tool that could have great advantages in analyzing signal transduction when compared to functional tools like knockdown with siRNA. In the past, a special type of mutations, so called ‘partial loss of function mutations’ have been used to dissect the contribution of different branches within a network. For example, this has been done to analyze the contribution of CRaf-independent effector pathways of Ras in T cells (Czyzky et al., 2003). In this work, different effector proteins were expressed in the context of Ras mutations which are impaired in their ability to interact with or activate certain Ras effectors, and the signaling output was monitored. In this way the signaling flow of Ras through another—CRaf independent—pathway could be quantified. Improving the mutations or creating them for other key cellular proteins, using structure-based protein design, would provide the framework for analyzing single branches in the context of larger networks (‘branch pruning’). This is a conceptual advantage compared with knockdown studies, where all branches are knocked down at the same time.

**Manipulating the output: the manipulation of genetic information**

Another field of interest for protein design is the creation/modification of protein-based molecular tools to manipulate genetic information. This can be done either by increasing or by decreasing gene expression, i.e. designing Transcription Factors, or by correcting or removing genes, i.e. designing Nucleases.

**Modifying transcription** The most common and widely used scaffolds to engineer such DNA-binding activities are zinc-finger proteins (ZFP) (Choo and Isalan, 2000; Jamieson et al., 2003; Uil et al., 2003; Negi et al., 2008), for mainly two reasons: first, those proteins bind to their DNA target as a monomer, hence allowing them to bind to non-palindromic sequences, second, they generally contain several finger domains (from three to nine), each contacting 3–4 bp (Elrod-Erickson et al., 1996; Kim and Berg, 1996), hence allowing more flexible design by combining individual fingers together and allowing the binding of 12–16 bp easily (this length implies an almost unique site in a genome). The advantage of using ZFP to engineer DNA-binding activity comes from its modularity. Each finger is able to contact a monomer, hence allowing them to bind to non-palindromic domains (from three to nine), each contacting 3–4 bp. The advantage of using ZFP to engineer DNA-binding activity comes from its modularity. Each finger is able to contact a monomer, hence allowing them to bind to non-palindromic domains (from three to nine), each contacting 3–4 bp.
fingers from the Zinc Finger Consortium (Wright et al., 2006). However, this approach showed its limit with the general oversight of the fourth contacting base and the individual consideration of each finger in a context-free situation (Ramirez et al., 2008). A new method for engineering context-dependent zinc-finger arrays (Maeder et al., 2008; Foley et al., 2009) has lately been developed and been included inside ZifIT. We developed a structure-based approach to design ZFP in silico using the FoldX protein design algorithm (Stricher et al., in preparation). This method (Alibés et al., 2009) aims at generating individual fingers specific for a sequence that can then be combined using the modular assembly approach. Each module is generated using all available structures of individual fingers in complex with DNA, hence taking partly into account the protein–DNA docking problem: for each quadruplet of bases and for each designed finger, the best template is then chosen as the one presenting the best interaction energy. A similar approach based on generating position-specific scoring matrices (PSSM) can be applied for predicting DNA-binding specificity of transcription factors with known structures (Morozov et al., 2005) (Alibés et al., in preparation) and has been proved valid in the design of other DNA-binding proteins, the meganucleases (Ashworth et al., 2006). This is the first step toward getting full structure-based design of Artificial Transcription Factors (ATF). A different and recently successful approach toward this end has been to use comparative genomics to reengineer cAMP receptor protein (CRP) of Escherichia coli to design new protein-DNA interactions (Desai et al., 2009). Analysing how different amino acid sequences vary in correlation with their cognate binding site, the authors managed to define a family-specific code and mode of binding that translated into four specific and successful designs out of eight. However, such a method cannot infer strict specificity or affinity to the target DNA, and may lead to ‘collateral damage’ once the ATF is introduced inside the cell (e.g. off-target transcription).

Designing inhibitors to control the activity of transcription factors can also be engineered using CPD. For example, specific peptides were recently designed for 19 out of the 20 different families of human basic-region leucine zipper (bZIP) transcription factors (Grigoryan et al., 2009). As these proteins share strong structural similarities among each other, design of specific ligands is a considerable challenge. Despite this challenge, the authors obtained specific designs for the majority of the targets by systematically optimizing the interaction energy with the preferred target while ensuring a relatively large energy difference with any off-target molecules from the other bZIP families. Among the successful targets were binders to the oncoproteins c-Fos, c-Jun and c-Maf. In contrast to the currently used bZIP inhibitors, which often associate with members of multiple bZIP families, these new inhibitors would be able to selectively inhibit members from a single bZIP family.

**Targeting gene repair and modification**

Aside engineering ATFs, protein design can be used as a way to develop specific endonucleases able to introduce double-stranded breaks (DSB) in the DNA. Subsequent repair of the damaged DNA by homologous recombination with an exogenous DNA template (the so-called ‘gene targeting’ approach) can then be used to introduce a specific mutation or to insert a specific DNA sequence at the break, or in the case of gene therapy, to replace a disfunctional gene by a non-disease inducing one. As for ATFs, a promising scaffold for this purpose are zinc-finger nucleases (ZFN) where several fingers recognizing two hemi-sites are linked to the Fok1 endonuclease (Durai et al., 2005; Porteus and Carroll, 2005; Wu et al., 2007; Cathomen and Joung, 2008). All methods used to engineer ZFP are also applicable here and with the added benefit of being able to redesign Fok1 to enforce heterodimerization, thereby reducing non-specific DSB (Kim et al., 1996, 1997; Mani et al., 2005; Szczepak et al., 2007). Homing endonucleases, also known as meganucleases, are another target of choice for engineering specific endonucleases since they naturally recognize large DNA sequences (>12 bp) and induce DSB (Chevalier and Stoddard, 2001). It has been shown using traditional selection procedures that meganucleases can be used as template to engineer variants showing altered DNA cleavage specificity and recombination (Chevalier et al., 2002; Épinat et al., 2003; Steuer et al., 2004). Using our structure-based protein design tool FoldX we managed to reconstruct the binding preferences of the homodimer I–CreI and its mutants (Arnold et al., 2006). Ashworth et al. (2006) used the protein design tool Rosetta to redesign another meganuclease (I-MsoI), to recognize and cleave another target site with a discriminative power close to the one of the wild type. As for the ZFN, it is possible for the meganucleases to redesign the dimerization interface to enforce heterodimerization and, hence decreasing the possible side-effects of non-specific cutting (Fajardo-Sanchez et al., 2008).

**Conclusion**

The use of CPD methods in the (re)design of various properties of proteins has made an impressive progress in recent years. CPD methodologies have already successfully been applied to generate tools to understand biological systems and in the development of protein therapeutics. In the coming years, CPD methods will become a tool of choice for understanding systems or creating novel systems, in the fields of systems biology and synthetic biology, respectively.

The primary advantage of computational design algorithms is the much larger fraction of the sequence/conformational space that can be sampled in search for an optimal solution when compared with ‘classical’ rational design or directed evolution methods, combined with a lower level of human intervention. Because CPD methods directly optimize the relevant physical interactions in silico, it can bypass some limitations of experimental selection and screening methods while it allows an increased throughput when compared to traditional rational design methods. For example, experimental selection or screening procedures often rely on the susceptibility of partly unfolded proteins or the final improvement in stability is limited to the temperature used in the selection assay were the host (phage, bacteria) is still viable. However, the use of directed evolution or high throughput screening methods in combination with CPD methods can also be complementary as it can provide higher quality libraries with increased diversity and consisting of properly folded and functional members (Hayes et al., 2002).

Despite the progress made, several challenges remain to be addressed. Structure-based algorithms require 3D
structural information, and in general the rule is that a higher resolution of the design target’s (crystal) structure yields a more reliable design outcome. Notwithstanding this requirement, several design exercises described in this review successfully relied (partly) on low-resolution structural models or homology models as template structure. Structural genomics initiatives will make an increasing proportion of the protein sequence space amendable to design by structure-based computational design algorithms either by directly using an experimentally determined structure or by providing structural templates to build more reliable homology models (Chandonia and Brenner, 2006). Recent progress in de novo protein structure prediction, using structure-based computational design algorithms, could in the future also contribute to make an even larger proportion of the sequence space amendable to structure-based design (Bradley et al., 2005). Protein design algorithms, in addition, are continuously improved as well. The development of more accurate energy functions used to evaluate the designs requires improvement in variety of factors, such as loop prediction, backbone flexibility, (surface) electrostatics, hydrogen bonding potentials, solvent (water)-mediated interactions and the statistical terms describing entropy (Jiang et al., 2005; Schymkowitz et al., 2005a, 2005b; Vizcarra and Mayo, 2005). Expanding the design algorithms with parameterization for various non-natural amino acids, as for example developed by Schultz et al. (Wang et al., 2006), will allow the design of protein sequences possessing entirely new chemical functionality and reactivity. Although the examples of design of affinity and specificity of protein–nucleic acid or protein small molecule interactions and the design of novel enzymatic activity are yet relatively sparse, or as in the case of design of novel enzymatic activity are in the early proof of concept phase, in coming years more examples will undoubtedly follow.

Aside from being a tool in studying and understanding of biological systems, CPD is a technology that can speed up the discovery of the drug discovery and development process. It can filter out less promising candidates early in the discovery process and can assure that pre-clinical and clinical trials can start with the most promising candidates. Several proofs of principle studies have demonstrated the feasibility of the use of CPD technology in engineering protein therapeutics with improved properties (Filikov et al., 2002; Luo et al., 2002; Steed et al., 2003; van der Sloom et al., 2004, 2006; Lazar et al., 2006).

Acknowledgements

We would like to thank the members of the Serrano lab for helpful discussions.

Funding

This work was supported by the EU program TRIDENT (grant number LSHG-CT-2006-037686). Research on signal transduction engineering in the Serrano lab is supported by the EU program PROSPECTS (grant number HEALTH-F4-2008-201648) and research on protein-DNA binding prediction is supported by the EU programs MEGATOOLS (grant number LSHG-CT-2006-037226) and ZNIP (grant number LSHB-CT-2006-037783). Research on protein complexes was supported by the EU program 3D REPERTOIRE (grant number LSHG-CT-2005-512028). A.M.S. was partially funded by a Juan de la Cierva fellowship of the Spanish Ministry of Education and Science.

References


