FEBS Letters 587 (2013) 1147–1154

journal homepage: www.FEBSLetters.org

Review

Emerging themes in the computational design of novel enzymes and protein–protein interfaces

Sagar D. Khare [a,*], Sarel J. Fleishman [b,*]

[a] Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, NJ 08854, USA
[b] Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel

ARTICLE INFO

Article history:
Received 19 November 2012
Revised 10 December 2012
Accepted 10 December 2012
Available online 19 December 2012

Edited by Wilhelm Just

Keywords:
Computational design
Novel protein function
Rosetta
ORBIT
Negative design
Enzyme
Protein interaction
Energy function

ABSTRACT

Recent years have seen the first applications of computational protein design to generate novel catalysts, binding pairs of proteins, protein inhibitors, and large oligomeric assemblies. At their core these methods rely on a similar hybrid energy function, composed of physics-based and database-derived terms, while different sequence and conformational sampling approaches are used for each design category. Although these are first steps for the computational design of novel function, crystal structures and biochemical characterization already point out where success and failure are likely in the application of protein design. Contrasting failed and successful design attempts has been used to diagnose deficiencies in the approaches and in the underlying hybrid energy function. In this manner, design provides an inherent mechanism by which crucial information is obtained on pressing areas where focused efforts to improve methods are needed. Of the successful designs, many feature pre-organized sites that are poised to perform their intended function, and improvements often result from disfavoring alternative functionally suboptimal states. These rapid developments and fundamental insights obtained thus far promise to make computational design of novel molecular function general, robust, and routine.

1. Introduction

What are the design principles that underlie the complex, sophisticated and beautiful protein systems that are at the heart of all life processes? Nature provides an inspiring number of different functional classes of proteins, from signaling molecules that display exquisitely fine-tuned molecular recognition, through regulated membrane channels and pumps, to enzymes that catalyze essential reactions at specificities and efficiencies that are unmatched by human invention. Biochemical and theoretical work has long been used to characterize how function is encoded in these systems, often by studying the impact of mutations on natural proteins. However, a myriad of evolutionary forces operating over countless generations has shaped extant natural systems, confounding the inference of key design principles. Recent advances in computation and high-throughput experimental analysis have opened the way to generating molecular function from the bottom up. By controlling all inputs into the process, computational protein design of novel function offers an intriguing route to uncover fundamental principles that explain existing molecular functions and, by extrapolation, allows construction of functional systems with no known natural counterparts.

Recent years have seen the first steps made by computational protein designers to produce novel catalysts, binding proteins, inhibitors, and oligomeric assemblies. Their approaches all rely on the inverse-folding paradigm [1], where the target state (a protein bound to its partner, be it another protein or a transition-state model) is modeled in atomic detail and the designed protein's sequence is chosen to form energetically favorable interactions with its target. The choice of the sequence is guided by (1) the technique used to consider different candidate sequences, or the sampling method, e.g., simulated annealing or dead-end elimination, and (2) the energy (scoring) function used to compare these candidate sequences. Energy functions used in design are usually "hybrid" (Fig. 1) – they feature terms that are physics-based (e.g., the Lennard–Jones potential for atomic repulsion and dispersion forces) and terms that are derived from known three-dimensional structures of proteins (e.g., amino acid sidechain conformational preferences observed in the Protein DataBank (PDB)). A crucial early insight was that both the energy function and sampling techniques used should be general and independent of the particular design or modeling problem [2–4]. In such a framework, the design process provides a powerful mechanism to diagnose the state of our understanding of protein energetics; improvements in the energy

* Corresponding authors.
E-mail addresses: sagar.khare@rutgers.edu (S.D. Khare), sarel@weizmann.ac.il (S.J. Fleishman).

0014-5793/$36.00
http://dx.doi.org/10.1016/j.febslet.2012.12.009

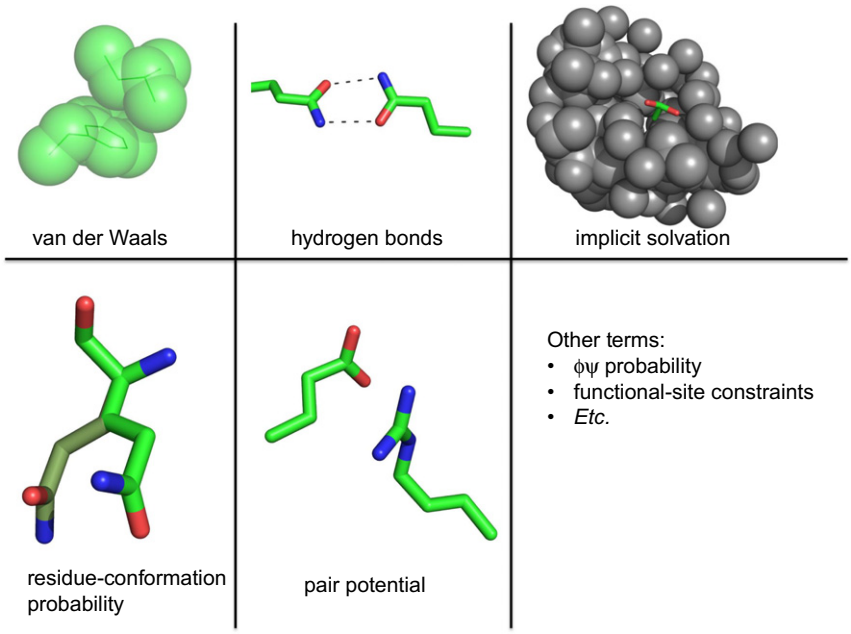

Fig. 1. A schematic description of hybrid all-atom energy functions used in design calculations. The energy functions used by several protein design software suites such as ORBIT and Rosetta are dominated by the following contributions (left-to-right, and top-to-bottom): van der Waals interactions accounting for both the attractive interactions among apposing molecular surfaces and the repulsive interactions due to steric overlap; hydrogen bonds between acceptor atoms (e.g., carbonyls), and donating polar hydrogens [62]. Hydrogen-bond strengths are determined by distance, orientation, and the polarity of the acceptor and donor; polar atoms are stabilized by interacting with water in their vicinity. Sequestering polar groups in the protein reduces some of this stabilization. Macromolecular forcefields used in design do not explicitly model the interactions with water molecules, rather use an implicit solvation model in which the total volume of excluded water in the vicinity of polar atoms is assessed by counting the number of protein atoms in a shell surrounding the atom [63,64]; residue sidechains are observed to reside in a limited set of preferred conformations, known as rotamers due to the nature of the chemical bonds within the residue and to dependencies on the local backbone conformations [64]. These probabilities are converted to pseudo-energies and used to bias conformations to the most likely ones; certain residue pairs are observed to cluster more often than others, for instance, due to the formation of stabilizing electrostatic interactions. A pair potential is derived from these propensities and used as a pseudo-energy term [65,66]. Other pseudo-energy terms are derived from the structures in the PDB and based on the desired molecular function (e.g., catalytic constraints). Increasing the reliability of macromolecular energy functions is an active area of research; major aspects that lack accuracy are the effects of solvation and electrostatics [55,67]. For a detailed treatment of the energy terms used in Rosetta we refer the reader to Ref. [49] and in ORBIT to Refs. [3,5]. All molecular figures were generated using PyMol [68].

function can then be fed back to improve all protein design, and more generally, protein modeling, efforts.

The earliest demonstrations that hybrid energy functions are useful for design came from the complete computational redesign of a zinc-finger protein by Dahiyat and Mayo [5], followed by the de novo design of a protein fold not observed in nature, by Kuhlman, Baker and co-workers [6], and more recently, a similar strategy led to the design of a novel protein loop [7]. Here, we limit ourselves to discussing the computational design of novel protein function – particularly novel enzymes and protein binders – that has been corroborated by experimental atomic structures but note that very exciting progress has been made in computational design of *altered* protein function, such as novel binding specificities [8,9] and allosteric regulation [10], and refer readers to a recent review [11].

2. Design of novel enzymes

Natural enzymes are amazingly proficient catalysts that can accelerate the rates of their cognate reactions by as much as 10^{23} fold [12]. The ability to de novo design an enzyme to catalyze any desired chemical reaction is a stringent test of our understanding of catalysis and will have significant practical applications in medicine and industry. Early computational design efforts focused on introducing metal-binding sites in proteins ([13,14]), and "nascent" metalloenzymes for redox chemistries were obtained by virtue of open metal co-ordination sites in the designed proteins ([15–17]). However, these studies did not include explicit computational models of the chemical transformation being catalyzed. A pioneering effort by Bolon and Mayo included atomistic details of the catalyzed reaction and introduced a nucleophilic histidine residue on the surface of a catalytically inert thioredoxin to obtain catalysts ("protozymes") for the hydrolysis of an activated ester substrate [18].

It is widely accepted that natural enzymes make two primary contributions to catalysis: they interact favorably with the reaction transition state [19] and they shield the chemical groups that aid catalysis from water [20], thereby increasing their reactivity; together these mechanisms lower the transition-state free energy in the active site microenvironment compared to the bulk solvent. To generate novel enzymes, design efforts in the framework of programs such as ORBIT and Rosetta have attempted to emulate these properties of natural enzymes. The process starts by modeling a so-called theozyme that is composed of a model of the chemical transition state(s) and key amino acid residues placed in orientations that are predicted to favor interactions with the transition-state model [21]. The transition-state structure cannot be experimentally determined due to its short lifespan (a few femtoseconds at room temperature [22]), so it is either adapted from crystal structures of transition-state analogue bound enzymes, or is based on quantum-chemical calculations. Constellations of backbones that can support the theozyme model are searched among hundreds of small-molecule binding pockets in crystallographic protein structures [23,34]. The sequence of residues in the putative catalytic pocket is then optimized to both favor maintenance of catalytic geometry and to provide additional stabilization to the transition state(s) [25]. Several candidate designs are synthesized in the laboratory and assayed for their programmed activity.