Progress in Biophysics and Molecular Biology 125 (2017) 24–35

Contents lists available at ScienceDirect

Progress in Biophysics and Molecular Biology

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

Automated protein design: Landmarks and operational principles

Anil Kumar [a], Ranjit Ranbhor [b], Kirti Patel [c], Vibin Ramakrishnan [d, *], Susheel Durani [e]

[a] Department of Chemistry, University of Toronto, ON, M5S3H6, Canada
[b] Sun Pharma, Vadodara, Gujarat, 390020, India
[c] N.B. Mehta Science College, Bordi, India
[d] Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, 781039, India
[e] Department of Chemistry, Indian Institute of Technology, Bombay, 400076, India

ARTICLE INFO

Article history:
Received 1 September 2016
Accepted 6 December 2016
Available online 12 December 2016

ABSTRACT

Protein design has an eventful history spanning over three decades, with handful of success stories reported, and numerous failures not reported. Design practices have benefited tremendously from improvements in computer hardware and advances in scientific algorithms. Though protein folding problem still remains unsolved, the possibility of having multiple sequence solutions for a single fold makes protein design a more tractable problem than protein folding. One of the most significant advancement in this area is the implementation of automated design algorithms on pre-defined templates or completely new folds, optimized through deterministic and heuristic search algorithms. This progress report provides a succinct presentation of important landmarks in automated design attempts, followed by brief account of operational principles in automated design methods.

Contents

* Corresponding author.
E-mail addresses: anil.kumar@u.toronto.ca (A. Kumar), vibin@iitg.ernet.in (V. Ramakrishnan).

http://dx.doi.org/10.1016/j.pbiomolbio.2016.12.002
0079-6107/

1. Introduction

Protein data bank crossed the landmark of 0.1 million structures registered in 2015. Total strength has grown approximately two times during 2009–2015 period; without adding a single number to the total number of unique folds as per SCOP classification (1393 folds) and just 16 folds added (total 1313 folds) as per CATH scheme (www.rcsb.org). Most natural proteins fold into a unique three dimensional structure through a unique combination of φ and ψ dihedral angles. This unique combination has a singular solution, namely a tertiary fold. Christen Anfinsen's path-breaking study on bovine pancreatic ribonuclease has successfully demonstrated that the only information necessary for driving a protein chain to its specified fold is its sequence. Though, this sequence to structure relationship appears to be simple; equations governing folding phenomenon is multivariate and complex, often guided by the rules of thermodynamics and kinetics. Advances in structure prediction methods have helped to reduce the complexity of folding phenomenon to a great extent, but a convincing solution is still a distant possibility. Though it is still debatable that protein folding follows a hierarchic process, protein structural organization has a definite hierarchy. Primary structure of a protein is its chemical sequence of individual amino-acids. Secondary structure constitutes of regularly repeating local folds locked in inter-residual hydrogen bonding network, with α -helix and β -sheet motifs being the most notable examples. The geometrical inter-digitization forming defined topological arrangement of secondary structures constitutes protein tertiary structure or fold. The multivariate interactions thought to be stabilizing tertiary structure, including hydrophobic effect and hydrogen bonds, electrostatics, entropy etc (Pace et al., 1996; Stickle et al., 1992). are still subjects of intense debate. Quaternary structure consists of independent chains of amino acids folding themselves to tertiary structures collectively forms a large protein molecule with a prescribed function.

Protein design attempts generally have two aims. The pronounced aim is to create a polypeptide construct, with specified structure and function. The knowledge about the architecture of native protein folds from amino acid sequence; known as the protein-folding problem, still remains obscure. Protein design, also known as inverse-folding problem (Pabo, 1983), addresses this question as well, though the primary motivation for *de novo* design remains construction of novel protein folds with tailor made function.

2. Landmarks in automated design attempts

Automated Protein design attempts have witnessed incremental progress from core redesign to protein-protein interface design, in last two decades (Fig. 1). A brief note on important landmarks is the following.

2.1. Protein core design

Tracy Handel (Desjarlais and Handel, 1995), using only *van der* Waals energy function, redesigned the core residues in a folded protein. The genetic algorithm (Pederson and Moult, 1996) used in Repacking of Cores (ROC) could optimize the core sequences. The effectiveness of ROC was tested by characterizing several hydrophobic core variants of the protein 434cro. Sequence variants predicted to be energetically favorable were comparable in stability to the wild type *(wt)* and much more stable than the control sequences in which the hydrophobic core was chosen randomly, thus validating the methodology employed.

2.2. Full Sequence Design

The first classical computational protein design recipe was derived by S. L. Mayo (Dahiyat and Mayo, 1996). The protein environment was classified as core, boundary and surface position based on the surface accessibility. Using fully automated procedure, the Zif268 sequence was redesigned with a totally new sequence. Christened as FSD-1 (Full Sequence Design-1), this new protein has no significant homology with any naturally occurring protein (Dahiyat et al., 1997a).

2.3. New fold design

The landmark contribution to design and synthesize a completely new fold was achieved by David Baker's group. The design produced a protein that adopts to a novel, not yet observed fold. The procedure used was automatic, apparently general, and based on essentially the same model of protein forces used in the Rosetta structure prediction method developed by the same group (Schueler-Furman et al., 2005). The initial step of the method is the selection of a set of distance constraints defining the desired fold. They were used to generate several initial conformations by assembling fragments of proteins of known structure. Sequence optimization for each of the starting structures was performed with Monte Carlo simulations and essentially the same energy function of the Rosetta method. Subsequent cycles of sequence and structure optimization led to the final sequence-structure pair named Top7. The biophysical characterization of the synthesized protein showed that Top7 was soluble, monomeric, stable, and that it underwent two-state cooperative unfolding (Dantas et al., 2003).

2.4. Enzyme design

Moving a step further Mayo's group reported the development and initial experimental validation of a computational design procedure aimed at generating enzyme like protein catalysts called “protozymes” (Bolon and Mayo, 2001). Their design approach utilized a ‘compute and build’ strategy that is based on the physical/