A PNA-mediated Whiplash PCR-based Program for In Vitro Protein Evolution

John A. Rose\textsuperscript{1,4}, Mitsunori Takano\textsuperscript{2}, and Akira Suyama\textsuperscript{3,4}

\textsuperscript{1} Department of Computer Science, The University of Tokyo, johnrose@is.s.u-tokyo.ac.jp
\textsuperscript{2} Institute of Physics, The University of Tokyo, takano@santa.c.u-tokyo.ac.jp
\textsuperscript{3} Institute of Physics, The University of Tokyo, suyama@dna.c.u-tokyo.ac.jp
\textsuperscript{4} Japan Science and Technology Corporation (JST-CREST)

Abstract. The directed evolution of proteins, using an in vitro domainal shuffling strategy was proposed in (J. Kolkman and W. Stemmer, Nat. Biotech. 19, 423 (2001). Due to backhybridization during parallel overlap assembly, however this method appears unlikely to be an efficient means of iteratively generating massive, combinatorial libraries of shuffled genes. Furthermore, recombination at the domainal level (30-300 residues) appears too coarse to effect the evolution of proteins with substantially new folds. In this work, the compact structural unit, or module (10-25 residues long), and the associated pseudo-module are adopted as the fundamental units of protein structure, so that a protein may be modelled as an N to C-terminal walk on a directed graph composed of pseudo-modules. An in vitro method, employing PNA-mediated Whiplash PCR (PWPCR), RNA-protein fusion, and restriction-based recombination is then presented for evolving protein sets with high affinity for a given selection motif, subject to the constraint that each represents a walk on a predefined pseudo-module digraph. Simulations predict PWPCR to be a reasonably high efficiency method of producing massive, recombined gene libraries encoding for proteins shorter than about 600 residues.

1 Introduction

The generation of shuffled dsDNA libraries by DNase I digestion and parallel overlap assembly (POA) (i.e., DNA shuffling) has been established as a powerful means of implementing random recombination between homologous genes [12]. Simulations, however predict that this approach is unlikely to be capable of evolving substantially novel protein folds, and that the non-homologous swapping of folded structures (i.e., exon or domain shuffling) is key for optimizing the search of protein sequence space [4]. Many proteins can be modelled as a string of non-overlapping, independently folding elements, or domains, each of which is typically 30-300 residues in length [1]. This has recently prompted the suggestion of protein evolution by in vitro domainal shuffling [5]. In this model, a polynucleotide species encoding for each domain is combined in solution with a set of
chimeric oligonucleotides, each of which encodes a domain-domain boundary. The iterated annealing, polymerase extension, and dissociation of this strand set (i.e., POA) then results in the production of a library of domain-shuffled dsDNAs. Although this operation preserves domainal integrity during initial shuffling, additional crossover following expression and screening requires either (1) DNase I digestion, which destroys the domainal integrity preserved by the initial process, or (2) a secondary, low-efficiency POA process, in which selected strands are used to implement the iterated extension of a set of added primers. In particular, the efficiency of POA is compromised by the high stability of back-hybridized duplexes, relative to extendable hybrids [6,7], in a fashion similar to that reported for Whiplash PCR (WPCR) [8]. In a secondary domainal shuffling process, this inefficiency will be exacerbated by both the increased length of backhybrids, and the tendency for strands to stably hybridize away from extendable ends, due to the high sequence similarity conferred by the domainal shuffling strategy. Recombination at the domainal level, which focuses on swapping existing folds, also appears to be too coarse to effect the evolution of substantially new folds. Finally, it is not clear how an iterative DNA shuffling strategy can be designed to ensure that a high proportion of dsDNAs encode promoter and ribosomal binding sites, to facilitate the use of an in vitro system for expression and selection (e.g., RNA-protein fusion [10,11]), which is necessary for overcoming the transfection limit of about $10^9$ strands [12].

In previous theoretical work, the efficiency of WPCR [13] was optimized [8,9], and the resulting architecture, PNA-mediated Whiplash PCR (PWPCR), was adapted to produce an in vitro genetic program for evolving approximate solutions to instances of Hamiltonian Path [14]. For related work, in which WPCR is used to evolve strategies for a simple version of Poker, readers are referred to [15]. In this work, PWPCR is combined with RNA-protein fusion [10,11] to implement a high-efficiency exon shuffling operation. For this purpose, the compact structural unit, or module [16], rather than the domain, is adopted as the basic element of protein structure, so that each shuffled protein represents a walk on a predefined graph, in which each vertex represents a pseudo-module contained in an initial protein set of interest.

2 Protein Representation

2.1 The Module Picture of Protein Architecture

The domain, or independently folding unit, is well established as a basic element of protein structure [4]. As pointed out by M. Go, et al. [10,17], however the frequent occurrence of introns within domains belies the view that each exon encodes a domain, and suggests the decomposition of domains into a set of smaller, compact structural unit, or modules, each of which forms a compact structural unit within the larger domain, and corresponds roughly to an exon. Any globular protein with a known 3D structure can be decomposed into an N-terminal to C-terminal sequence of modules, by exploiting either the tendency for module junctions to be buried, or the tendency of modules to form a locally