doi:10.1016/j.jmb.2007.06.032 J. Mol. Biol. (2007) 372, 1–6

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

COMMUNICATION

Full-sequence Computational Design and Solution Structure of a Thermostable Protein Variant

Premal S. Shah[1]†, Geoffrey K. Hom[1]†, Scott A. Ross[2] Jonathan Kyle Lassila[1], Karin A. Crowhurst[3] and Stephen L. Mayo[4]*

[1]*Biochemistry and Molecular Biophysics Option, MC 114-96 California Institute of Technology, Pasadena CA 91125, USA*

[2]*Division of Chemistry and Chemical Engineering California Institute of Technology, MC 164-30 Pasadena, CA 91125, USA*

[3]*Division of Biology, California Institute of Technology*

[4]*Howard Hughes Medical Institute and Divisions of Biology and Chemistry and Chemical Engineering California Institute of Technology, MC 114-96 Pasadena, CA 91125, USA*

**Corresponding author*

Computational protein design procedures were applied to the redesign of the entire sequence of a 51 amino acid residue protein, *Drosophila melanogaster* engrailed homeodomain. Various sequence optimization algorithms were compared and two resulting designed sequences were experimentally evaluated. The two sequences differ by 11 mutations and share 22% and 24% sequence identity with the wild-type protein. Both computationally designed proteins were considerably more stable than the naturally occurring protein, with midpoints of thermal denaturation greater than 99 °C. The solution structure was determined for one of the two sequences using multidimensional heteronuclear NMR spectroscopy, and the structure was found to closely match the original design template scaffold.

Keywords: computational protein design; NMR structure; protein stability; engrailed homeodomain; FASTER

The process of redesigning protein sequences using computational algorithms and physical energy functions has emerged as a powerful tool for protein engineering. New enzyme catalysts, high-affinity biosensors, thermostable proteins, and specific protein–protein interactions have been designed computationally.[1–3] However, because procedures and energy functions used for computational protein design are both new and relatively untested, experimental evaluation continues to be crucial for improving the performance of design procedures.

We have performed redesigns of the entire amino acid sequence of an all-helical protein, *Drosophila melanogaster* engrailed homeodomain. Full-sequence design provides a rigorous challenge for computational methods because it requires energy functions adapted to simultaneous

† P.S.S. and G.K.H. contributed equally to the work.

Present addresses: P. S. Shah, Science & Technology Team, Booz Allen Hamilton, 225 West Wacker Dr. Ste 2270, Chicago, IL 60606, USA; G. K. Hom, Genentech, 1 DNA Way, Mail Stop 34, South San Francisco, CA 94080, USA; K. A. Crowhurst, Department of Chemistry and Biochemistry, California State University Northridge, 18111 Nordhoff St., Northridge, CA 91330, USA.

Abbreviations used: DEE, dead-end elimination; CD, circular dichroism; T_m, midpoint of thermal denaturation; NOE, nuclear Overhauser effect; NOESY, NOE spectroscopy; HSQC, heteronuclear single quantum coherence.

E-mail address of the corresponding author: steve@mayo.caltech.edu

modeling of residues in dramatically different environments ranging from the solvent-exposed protein surface to the hydrophobic core. In addition, full-sequence designs involve vast combinatorial complexity. The engrailed homeodomain designs described here required optimization from among 10^{111} possible combinations of side-chain identity and conformation. The complexity of these calculations rendered them intractable by exact search methods based on the dead-end elimination (DEE) theorem.[4,5] Thus, these designs employed stochastic search methods that can provide solutions for difficult optimization problems but are not guaranteed to identify the most favorable solution for a given problem.

To date, structural validation has been carried out for only a small number of proteins arising from simultaneous full-sequence computational design.[6–9] These prior studies include fixed-backbone full-sequence designs of 28, 70, and 96 amino acid residues,[6,8,9] as well as a full-sequence design of 93 amino acid residues that incorporated main-chain conformational flexibility.[7] We report experimental characterization of two proteins designed using different stochastic search methods as well as a solution NMR structure of the design with the lowest simulation energy. The NMR ensemble closely resembles the design scaffold structure, indicating that the designed protein occupies the target fold.

Full-sequence designs are significantly more stable than the natural protein

Two sequences were designed in calculations that differed only in the side-chain optimization method used. Protein Data Bank entry 1enh was used as the backbone template for design.[10] All residues were assigned to a fixed binary pattern based on solvent-exposed surface area as described.[11,12] The amino acids Ala, Val, Leu, Ile, Phe, Tyr, and Trp were allowed in buried positions, while the polar amino acids Ala, Asp, Asn, Glu, Gln, His, Lys, Ser, Thr, and Arg were considered in solvent-exposed positions. The amino acid identities of positions involved in helix capping and helix dipole interactions (4–6, 16, 17, 22–24, 31, 32, 36–38, 49, and 50) were further restricted to statistically likely helix-capping residues as described.[13] A backbone-dependent rotamer library[14] was employed and a standard energy func-

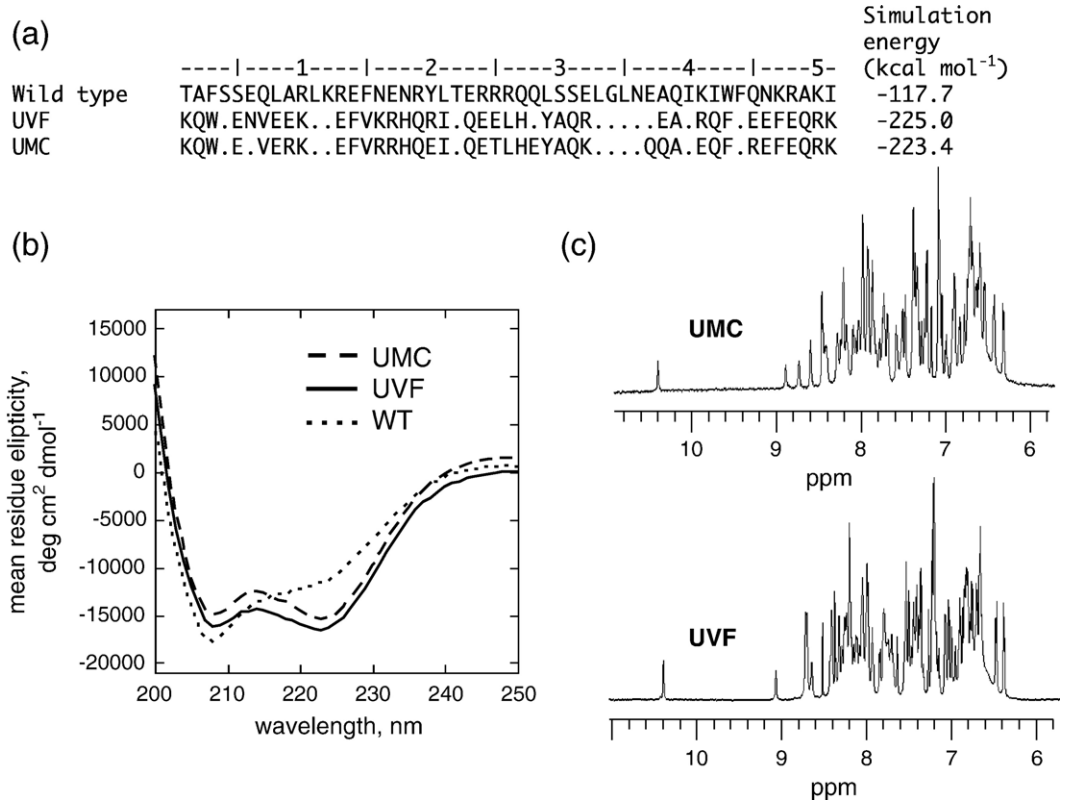

Figure 1. (a) Sequence alignment and simulation energies of wild-type engrailed homeodomain and computationally designed proteins UVF and UMC. Identical residues are marked with a period. UVF and UMC have 24% and 22% sequence identity with wild-type, respectively. (b) Circular dichroism wavelength scans of UVF and UMC. The spectrum of the 51 amino acid residue wild-type engrailed homeodomain construct is included for comparison. We have observed increased signal at 222 nm in a number of stabilized engrailed variants (see, for example, Marshall & Mayo[12]). Spectra were obtained using an Aviv 62DS spectrometer with a 1 mm path length cell containing 50 μM protein in 50 mM sodium phosphate (pH 5.5) and 25 °C. (c) One-dimensional ^{1}H NMR spectra of UVF and UMC. For clarity only the amide region is shown.