Transcribed products of tandem microgene repeats exhibit diverse properties also seen in natural proteins

Kiyotaka Shibao1,2,5, Tsuyoshi Shirai3, Takako Honma2 and Tetsuo Noda2,4
1Department of Protein Engineering, 2Department of Cell Biology, Cancer Institute, Japan Foundation for Cancer Research, Toshima, Tokyo 170-8455, 3Department of Biotechnology, Graduate School of Engineering, Nagoya University, Chikusa, Nagoya 464-8603 and 4Department of Molecular Genetics, University of Tohoku School of Medicine, Aoba, Sendai 980-8575, Japan

Keywords: artificial repetitious protein/de novo protein synthesis/in vitro protein evolution/protein engineering

Introduction

Soon after the development of methodologies used to resolve the molecular structures of biomacromolecules, it became apparent that these structures were remarkably repetitive in nature, and the possibility that such repetition was related to the origin and evolution of proteins and genes was proposed (Rossmann et al., 1974; Barker et al., 1978; Ohno, 1981; Ohno and Epplen, 1983). Recent innovations in genome sequencing and structural biology have only served to increase our appreciation of the scope of this periodicity. Genome structures are now known to contain a multitude of periodicities, ranging from those on the level of one to several nucleotides (Sutherland and Richards, 1995) to those on the sub-chromosomal level (Wolfe and Shields, 1997). Indeed various repeated sequences account for more than 50% of the human genome (Lander et al., 2001). Likewise, protein structures also exhibit a variety of repeats; ‘domain’, ‘module’ and ‘element’ are terms often used to describe units that repeat within the tertiary structure of many proteins (Baron et al., 1991; Yura et al., 1993; Kobe and Deisenhofer, 1995; Kobe and Kajava, 2000). Within the primary structure of proteins, sequences ranging from 1 to >300 amino acids in length are repeated (Sutherland and Richards, 1995; Li et al., 1997) from 2 to >300 times (Wilkin et al., 2000). Such peptide sequence repeats may exist as simple contiguous tandem repeats (Eckert and Green, 1986), as non-tandem and scattered repeats, or as complicated hierarchical patterns (Hayashi and Lewis, 2000). Thus, biomacromolecules exhibit numerous modes of reiteration, varying widely with respect to length, number, conservation and spacing.

Some patterns of repetition are apparently related to the origin or evolution of molecules. Proteins of relatively recent origin (e.g. the circumsporozoite antigen of parasitic protozoa and antifreeze proteins of fish) often consist of near exact copies of oligopeptide sequences (Ohno, 1984). In the case of porcine ribonuclease inhibitor, the near-perfect periodicity observed in its tertiary structure implies that the protein evolved from the reiteration of a DNA sequence 87 bp in length or less (Kobe and Deisenhofer, 1993). We have focused on the hypothesis that new genes arise in the form of repeats of a short DNA block, and have created artificial repetitious polypeptides through polymerization of microgenes lacking a stop codon (Shiba et al., 1997). Our previous work has shown that periodic DNA produces ordered proteins at very high rates (Shiba et al., 2002). In this paper, we report on the characteristics of several artificial proteins created from reiteration of short DNA sequences.

Materials and methods

Expression of microgene polymer products and their purification

The expression and purification of artificial proteins has been described previously (Shiba et al., 2002). Briefly, proteins were expressed in E.coli and purified under denaturing conditions using TALON resin (Clontech, Palo Alto, CA). Eluted proteins were dialyzed against a buffer containing 50 mM Tris–acetate pH 4.0, 100 mM NaCl and 1 mM EDTA to remove the urea.

Small angle X-ray scattering (SAXS)

Measurement of SAXS was carried out at a solution scattering station installed at BL-10C, Photon Factory, Tsukuba, Japan (Ueki et al., 1985). Proteins were dissolved in 10 mM phosphate buffer pH 6.0 containing 100 mM NaCl and placed in a cell (1 mm X-ray path length with 20 µm thick quartz windows) where the measurements were made. The wavelength of the incident X-rays was 1.488 Å, and the beam cross-section was 0.5×3.0 mm. The distance from the sample to the detector
was 202 cm, calibrated with meridional diffraction of dried chicken collagen. The same solution without protein was measured as background. All measurements were performed at ambient temperature.

**Radii of gyration (R_g) of natural proteins in native and denatured states**

Reported values for the following proteins were used: ubiquitin (UBQ) and β-lactalbumin (BLA) (Kamatari et al., 1999); Streptomyces subtilisin inhibitor (SSI) (Konno et al., 1995); cytochrome c (CYT) (Kamatari et al., 1996); ribonuclease A (RNA) (Sosnick and Trewhella, 1997); α-lactalbumin (ALA) (Kataoka et al., 1997); lysozyme (LYS) (Kamatari et al., 1998); staphylococcal nuclease (SNC) (Panick et al., 1999); apo-myoglobin (MYO) (Kataoka et al., 1995); aspergillopepsin II (PEP) (Kojima et al., 2000); OsP A (OSP) (Koide et al., 1999); and the α-subunit of tryptophan synthase (TRP) (Gualfetti et al., 1999).

**Crystallization conditions**

Protein no. 320 (α) was crystallized using the hanging drop vapor-diffusion method. One millilitre of 50 mM HEPES buffer pH 6.5 containing 25% (w/v) PEG 4000 and 20% (v/v) glycerol served as the reservoir. A drop containing 5 μl of 1% (w/v) protein solution in 50 mM phosphate buffer and 5 μl of reservoir solution was equilibrated against the reservoir at 18 °C. Numerous microcrystals appeared within a few days; the crystal presented was one of the largest. A sample of the crystals was rinsed with reservoir solution and confirmed to be composed of protein no. 320 (α) by SDS-PAGE.

**X-ray diffraction analysis**

The microcrystals were scooped up from a hanging drop with a cryo-loop (Hampton Research, Laguna Niguel, CA) and flash frozen in liquid nitrogen. The X-ray diffraction experiment was carried out using an RU-300 Cu-rotating anode X-ray generator (Rigaku, Tokyo) operating at 50 mV×100 mA (λ = 1.542 Å). During the experiment, the frozen drop was kept in a 100 K nitrogen stream from a cryosystem (Oxford Cryosystem, Oxford, UK). The diffraction image was recorded for 10 h using a Raxis IV imaging plate detector (Rigaku).

**Results**

**Creation of artificial proteins**

The primary structures of the six proteins described in this report are shown in Figure 1. They are the translated products of artificial open reading frames that were created by tandem polymerization of short DNA sequences having no stop codon (microgenes) described previously (Shiba et al., 2002). For polymerization of microgenes, we used a method called microgene polymerization reaction (MPR), which randomly inserts or deletes nucleotides at end-joining junctions of microgene units (Shiba, 1998). The translated proteins, therefore, were not simple repeats of unique peptide sequences but combinatorial polymers of 2×3 open reading frames.

Proteins no. 287 (Cu), no. 291 (Cu) and no. 292 (Cu) were all derived from the 54 bp designer microgene MG-14 (Cu) (Shiba et al., 2002). One of the reading frames (frame-1) of this microgene encodes the peptide GAGMYAESYGRK-ICSPHQ, which with head-to-tail concatenation can recreate part of the type-1 copper-binding motif of plastocyanin (H-[46aa]..CSPHQAGM) (Adman, 1991). Protein no. 287 (Cu) contains part of frame-1 within its C-terminal region (KICSPHQ), while the remainder of the protein is composed of frame-2. Protein no. 291 (Cu) is translated from the reading frames of the ‘minus’ strand of MG-14 (frames-3, -4 and -5), and is therefore not intentionally related to any existing protein. The sequence of protein no. 291 (Cu) has a rather complicated structure due to mutations occurring at end-joining junctions and elsewhere in the sequence; the latter appear to have been introduced by the thermocycle reaction or within E.coli cells. Internal deletions also produced complicated sequences in protein no. 292 (Cu), which was translated from the three reading frames of the ‘plus’ strand of MG-14 within which the copper-binding motif appears five times.

The 42 bp microgene MG-15 (α) was designed so that it encodes the polypeptide RKVLQGRMENLQAE in its first reading frame. This peptide sequence was derived from part of the coiled coil α-helix of seryl-tRNA synthetase (Cusack et al., 1990) and was expected to provide α-helical structure to the reconstituted proteins. Protein no. 320 (α) contains 5.5 repeats of the α-helix forming frame of the microgene, which accounts for about 80% of the protein. Analogously, the 66 bp MG-16 (β) was designed so that one of its reading frames codes for the β-strand forming peptide GVWWDESIGN-RMDSSNIGNSSAN. This peptide sequence was designed to be a 22 amino acid consensus sequence of the repeating units of the parallel β-helix protein pectate lyase from Erwinia chrysanthemi (Yoder et al., 1993). Protein no. 416 (β) has 1.5 and 2.5 repeats of this β-forming peptide at its N- and C-
Properties of artificial repetitious proteins

Fig. 1. Primary structure of products translated from microgene polymers. Numbers represent clone names and are followed by the name of the microgene unit and the direction of translation (+ and − denote sense and antisense microgenes, respectively). Red, black and blue letters represent charged, polar and non-polar residues, respectively (Shiba et al., 2002). Yellow, aqua and pink boxes represent residues calculated to form α-helices, β-strands and β-turns, respectively, using a secondary structure prediction program (Geourjon and Deleage, 1994). Bars represent deleted regions.

Fig. 2. Macroscopic structures spontaneously formed by microgene polymer proteins. 3 µl of protein solution (8 µg/µl in 50 mM Tris–acetate pH 4.0, 100 mM NaCl, 1 mM EDTA) was poured into 200 µl of 50 mM sodium phosphate buffer pH 8.0 containing 0.2 M LiCl at ambient temperature; the emergent structures were observed under a stereomicroscope (Wild M30, Leica). Bar = 1 mm.
(Figure 1 and Table I). Although secondary structure was also agreed that secondary structure based on the sequences of these proteins declined to 4.0 and 2.0 (Shiba et al., 2002). Predictions of secondary structures in four: no. 291 (Cu$^+$)–β-sheet-like structures in proteins no. 320 (α$^+$) and no. 416 (β$^+$), respectively, suggest that structures intentionally embedded in microgenes are reconstituted in their translated products. The sequence of protein no. 291 (Cu$^+$)–β-sheet-like structures in proteins no. 320 (α$^+$) and no. 416 (β$^+$), respectively, suggest that structures intentionally embedded in microgenes are reconstituted in their translated products. The sequence of protein no. 291 (Cu$^+$)–β-sheet-like structures in proteins no. 320 (α$^+$) and no. 416 (β$^+$), respectively, suggest that structures intentionally embedded in microgenes are reconstituted in their translated products. The sequence of protein no.
292 (Cu\(^{+}\)) contained four repeats of the copper-binding motif embedded in the first frame of MG-14 (Cu). Although the copper-binding motif was not fully expressed at the C-terminal fragment of protein no. 287 (Cu\(^{+}\)), a Met residue necessary for the complete motif recurred in the N-terminal domain of the protein (Figure 1). When purifying the six proteins, we found that solutions of proteins no. 287 (Cu\(^{+}\)) and no. 292 (Cu\(^{+}\)) were pink in color, probably because we used an immobilized Co\(^{2+}\) affinity column for their purification, resulting in the association of Co\(^{2+}\) with these two proteins. The possible interaction of metal ions and MG-14 (Cu) polymer proteins was further investigated by refolding denatured protein no. 287 (Cu\(^{+}\)) in a solution containing 50 mM 2-(N-morpholino)-ethanesulfonic acid, 100 mM NaCl and 0.5 mM CuSO\(_4\). This yielded a protein with a pale blue color and an absorption maximum at 630 nm, indicating association of Cu\(^{2+}\). Although the absorption maximum differed from that of plastocyanin (597 nm) (Yoshizaki et al., 1981) and the analyses were qualitative, these observations nevertheless suggest that proteins no. 287 (Cu\(^{+}\)) and no. 292 (Cu\(^{+}\)) associate with metal ions, possibly via an embedded copper-binding motif. Moreover, they demonstrate that a function encrypted within a microgene can be reconstituted in an artificial protein.

**Spontaneous formation of macroscopic structures**

Various levels of repetition are often observed in proteins that assemble into intracellular or extracellular macrostructures. Instances include spectrin, a component of the cytoskeleton (Yan et al., 1993); fibronectin, a component of extracellular matrix (Leahy et al., 1996); involucrin, a component of the keratinocyte envelope (Eckert and Green, 1986); and Flag, a component of spider flagelliform silk (Hayashi and Lewis, 2000). In addition, amyloid formation is believed to be related to repeats of oligopeptide (Liu et al., 1999). Their repetitious nature is believed to favor intermolecular interactions leading to the formation of macroscopic structures.

We evaluated the intermolecular interactions of our repetitive proteins by examining their ability to form a macroscopic membrane-like structure when poured into an alkaline monovalent cation solution (i.e. 0.2 M LiCl). This protocol was previously used by others to form membranes from a self-complementary oligopeptide [\(\text{AEAEAKAK}\)]\(_2\) that exhibited the typical CD spectrum of \(\beta\)-sheet (Zhang et al., 1993). Among the six proteins, no. 291 (Cu\(^{+}\)), no. 377 (tRNA\(^{+}\)) and no. 416 (\(\beta\)\(^{+}\)) formed a membranous structure in buffer containing 0.2 M LiCl (Figure 2). Interestingly, these membrane-forming proteins share two important features: (i) a rather high composition of non-polar amino acid residues (Table I) and (ii) a \(\beta\)-sheet-like structure (Shiba et al., 2002). This suggests that by adjusting the hydrophobicity and \(\beta\)-sheet forming ability of the polymerized products, microgenes could be designed so that they code for proteins capable of strong and orderly intermolecular interactions, thereby resulting in the formation of macroscopic structures.

**Small angle X-ray scattering (SAXS) analysis**

The membrane formation described above suggests that proteins no. 291 (Cu\(^{+}\)), no. 377 (tRNA\(^{+}\)) and no. 416 (\(\beta\)\(^{+}\)) tend to form multimeric assemblies, even under physiological conditions. This possibility is supported by the observations that proteins no. 291 (Cu\(^{+}\)) and no. 377 (tRNA\(^{+}\)) behaved as huge proteins with molecular weights of more than 440 kDa in gel filtration experiments. Moreover, 6.0 M urea was required for filtration of protein no. 416 (\(\beta\)\(^{+}\)), suggesting intermolecular interaction was particularly strong for that protein (data not shown). By contrast, proteins no. 287 (Cu\(^{+}\)) and no. 320 (\(\alpha\)\(^{+}\)) behaved as though they were relatively small (the calculated Stokes radii were 16 Å and 33 Å, respectively, in size-exclusion chromatography experiments; data not shown) and, in the presence of 6 M urea, no. 292 (Cu\(^{+}\)) ran as a protein of approximately 100 kDa in size-exclusion chromatography (data not shown). Thus, proteins no. 287 (Cu\(^{+}\)) and no. 320 (\(\alpha\)\(^{+}\)) might exist as tightly packed mono-disperse molecules in physiological solutions.

To address that question further, we performed small angle X-ray scattering (SAXS) analyses with proteins no. 287 (Cu\(^{+}\)), no. 320 (\(\alpha\)\(^{+}\)) and no. 292 (Cu\(^{+}\)) to determine their radii of gyration (Rg) in solution. The Rgs obtained from Guinier plots (Glatter and Kratky, 1982) were 26.6 Å, 26.6 Å and 29.0 Å for proteins no. 320 (\(\alpha\)\(^{+}\)), no. 292 (Cu\(^{+}\)) and no. 287 (Cu\(^{+}\)), respectively. Comparison of the apparent Rgs and monomeric molecular weights of the artificial proteins with those of natural proteins in their native and denatured states indicated that the former do not fold as tightly as the latter (Figure 3A). The Rgs of the artificial proteins fell in the vicinity of the lower boundary of the Rgs of denatured natural proteins. However, the fact that one of the natural proteins, aspergillopepsin II (cf. PEP in Figure 3A), retained its intramolecular disulfide bonds and therefore showed a relatively small Rg in the denatured state (Kojima et al., 2000) implies that the artificial proteins, especially no. 320 (\(\alpha\)\(^{+}\)) and no. 292 (Cu\(^{+}\)), might not be in a completely random state. The same implication was obtained from Kratky plots (Figure 3B), which show a peak when a molecule has a globular shape and a plateau when it is a random coil (Glatter and Kratky, 1982). The plots for the artificial proteins suggested a tangled shape for no. 320 (\(\alpha\)\(^{+}\)) and no. 292 (Cu\(^{+}\)) and a typical extended coil for no. 287 (Cu\(^{+}\)) (Figure 3B). Consequently, the structures of proteins no. 320 (\(\alpha\)\(^{+}\)) and no. 292 (Cu\(^{+}\)) are rather more compact than if completely denatured. It is not known why protein no. 287 (Cu\(^{+}\)) behaved as a compact molecule in size-exclusion chromatography.

**Crystallization**

SAXS analyses showed that proteins no. 320 (\(\alpha\)\(^{+}\)), no. 292 (Cu\(^{+}\)) and no. 287 (Cu\(^{+}\)) do not fold into as compact a structure as natural proteins; instead they appear to be in a compact denatured or ‘gemisch’ state (Dill et al., 1995). This finding was predictable since the side chain packing that leads to interchain organization within proteins was not explicitly embedded in the design of the microgenes. Still, it is possible that these artificial proteins could undergo a transition to a more ordered state under appropriate conditions. In the case of some natural repetitive proteins, for example, local interactions among repeating units are sufficient to support a compactly folded structure (Kobe and Kajava, 2000). We tested this possibility by screening the crystallization conditions for proteins no. 320 (\(\alpha\)\(^{+}\)), no. 292 (Cu\(^{+}\)) and no. 287 (Cu\(^{+}\)) using reagents and methods usually applied to natural proteins. A protein crystal was obtained for no. 320 (\(\alpha\)\(^{+}\)) (Figure 4), and although it did not diffract well, a powder diffraction-like pattern was observed after a prolonged X-ray exposure. The three distinct diffractions observed suggest two possibilities: that the unit cell is 3D with dimensions 30 \(\times\) 18 \(\times\) 15 Å, or that it is 2D with dimensions 30 \(\times\) 18 Å. In either case, formation of a crystal proves that this microgene polymer protein can generate an ordered structure under appropriate conditions.
Discussion

We have characterized six artificial proteins created using the microgene polymerization method described in a previous report (Shiba et al., 2002). With that method, a single designer microgene is polymerized in a head-to-tail manner using the MPR technique (Shiba et al., 1997). Because MPR randomly adds or deletes nucleotides at the junctions of repeating units, the translated products of the microgene repeats behave as a combinatorial library of three reading frames derived from a single microgene. By designing the microgene so that it contains no termination codon, we can prepare a gene library that contains long open reading frames.

Although the microgene polymerization method provides a molecular diversity library, it differs from the ‘random sequence library’ that has been used so successfully to evolve functional RNAs, DNAs, peptides and even proteins (Cwirla et al., 1990; Devlin et al., 1990; Ellington and Szostak, 1990; Scott and Smith, 1990; Tuerk and Gold, 1990; Keefe and Szostak, 2001). A random sequence library is usually prepared by combinatorial polymerization of four nucleotides and is often called a ‘naive library’, which means the molecules represent a random sampling of a huge sequence space. By contrast, microgene polymers represent a ‘sparse sampling of a local sequence space’ that is considerably biased toward the peptide sequences coded by the microgene unit (Shiba, 1998). In this line of comparisons, the microgene polymerization method is more like protein engineering than in vitro evolution. In the former, the aim is to obtain a rational de novo design for a novel protein (Ulmer, 1983), while in the latter a neo-Darwinian selection from a large pool of random sequences is emphasized (Szostak, 1992). As we observed previously, the rationally embedded secondary structures in microgenes MG-15 (α) and MG-16 (β) were successfully expressed in the translated products of the polymers, as exemplified by proteins no. 320 (α†) and no. 416 (β†). In addition, the fact that products of MG-14 (Cu) polymers were blue in color when in solution with copper ion [no. 287 (Cuβ) and no. 292 (Cuα)] suggests that functionality that is correlated with a specific peptide sequence can also be embedded into designer microgenes and can be reconstituted in proteins translated from microgene polymers. Finally, the fact that formation of a macroscopic membranous structure (Figure 2) was correlated with hydrophobic and β-sheet-like structure suggests it may be possible to endow proteins with sketchy physicochemical properties by adjusting the compositions and arrangements of the amino acids coded by a microgene.

In addition to an aspect of rational design, the microgene polymerization method also provides the ability to present molecular diversity. For instance, two α-helix-rich proteins, no. 320 (α†) and no. 334 (α‡), were translated from different combinations of frames of MG-15 (α)—i.e. (5.5×frame-1)–(1×frame-3) for 320 (α†) and (9.5×frame-1) for no. 334 (α‡). At acidic pH, protein no. 320 (α†) lost its secondary structure, indicating α-helix formation to be pH-dependent in that protein (Shiba et al., 2002). Protein no. 334 (α‡), on the other hand, retained its secondary structure at pH 2.0, but precipitated at pH 8.0 where protein no. 320 (α†) remained soluble (data not shown). Thus, variation in reading frame and/or the number of repeats can modulate the biochemical characteristics of microgene polymer proteins.

One of the unique biophysical properties of natural proteins is that they fold into specific and compact structures. On the contrary, most artificial proteins adopt ensembles of expanded conformations that result in larger molecular radii and poor chemical shift dispersion in their 1H-NMR spectra (Davidson et al., 1995; Dill et al., 1995). Proteins created by microgene polymerization are no exception, with proteins no. 320 (α†), no. 292 (Cuα) and no. 287 (Cuβ) showing rather expanded molecular structures in small angle X-ray scattering analyses (Figure 3), and proteins no. 291 (Cuβ), no. 320 (α†), no. 377 (tRNA) and no. 416 (β†) showing limited dispersion in their 1H-NMR chemical shifts (K.Shiba, S.Yuzawa, H.Hatanaka and F.Inagaki, unpublished results). This so-called ‘gemisch’ state of artificial proteins is believed to be due to insufficient or improper side chain packing between polypeptide chains. But although it may be inherently difficult to rationally embed within microgene sequences interactions between side chains that are distant from one another, the packing of repeating units with their neighbors can be rationally designed. This implies that a type of ‘solenoid’ protein (Kobe and Kajava, 2000), such as those having LRR (leucine-rich repeats), β-helix and TPR (tetratricopeptide repeats) motifs, could be created using the microgene polymerization method. Furthermore, it is noteworthy that protein no. 320 (α†), whose microgene unit was designed from a natural coiled coil peptide, could be crystallized (Figure 4), demonstrating that under the appropriate conditions an artificial protein can be in an ordered state.

The ability to encrypt potential functions and structures in designer microgenes and then to construct a molecular pool having diverse physiochemical properties through polymerization of these microgenes could form the basis of a new approach to protein engineering that could be used to rationally reconstitute biological functions in artificial proteins.

Acknowledgements

We are grateful to Dr Y. Muroga for useful discussions and comments. This work was supported in part by HFSP Research grant (K.S.) and by Grants-in-Aid from the Ministry (T. S.).

References

Properties of artificial repetitious proteins


Received September 13, 2002; revised October 28, 2002; accepted October 31, 2002.