FEBS 16737

FEBS Letters 382 (1996) 21–25

# Solubility of artificial proteins with random sequences

Irfan D. Prijambada[a], Tetsuya Yomo[a], Fumihiro Tanaka[a], Toshihiro Kawama[a], Keizo Yamamoto[b], Akihisa Hasegawa[a], Yasufumi Shima[a], Seiji Negoro[a], Itaru Urabe[a,*]

[a]*Department of Biotechnology, Faculty of Engineering, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565, Japan*
[b]*Department of Chemistry, Nara Medical University, 840 Shijo, Kashihara, Nara 634, Japan*

Received 19 January 1996

**Abstract A library of artificial random proteins of 141 amino acid residues of which 95 are random and which includes the 20 kinds of amino acids was prepared. Out of the 25 identified random proteins, 5 were soluble in the cell lysate, indicating that about 20% of the random proteins expressed in *Escherichia coli* are expected to be soluble. The soluble random proteins RP3-42 and RP3-45 and insoluble RP3-70 were purified. The solubility of the purified form is the same as that in the cell lysate.**

*Key words:* Artificial protein; Enzyme evolution; Protein solubility; Random protein

## 1. Introduction

Proteins, the most abundant macromolecules in the living cells, play various roles in either soluble or insoluble state. The solubility of a protein, hence, is one of the important factors for a protein to exert its biological functions and is controlled by amino acid composition, structure, and environmental conditions As proteins are linear polymers composed of 20 different kinds of amino acids, there are numerous possible sequences for a protein to avail. For instance, with the 20 amino acids available, there will be $20^{100}$ possible sequences for a 100-residue protein. Of all these possible sequences, the corresponding proteins can either be soluble or insoluble one. Therefore, as to what is the percentage of the possible sequences that can bring forth proteins that are soluble is a basic question in the fields of protein physics, protein evolution, and artificial enzymes. In this work, we have prepared a library of 141 amino acid residue proteins with random sequences. The random sequences include the 20 kinds of amino acids. The state of the random proteins in the cells of *Escherichia coli* as to their solubility was examined. Out of 25 proteins examined, 5 were soluble. Hence, about 20% of the random proteins with 141 residues are expected to be soluble.

## 2. Material and methods

### 2.1. Bacterial strain and plasmids

The bacterial strain and plasmids used were *Escherichia coli* KP3998 (F⁻ *hsdS20* ($r_B^-$ $m_B^-$) *ara-14 proA2 lacI*$^q$ *galK2 rpsL20 xyl-5 mlt-1 supE44* $\lambda^-$) [1], pUC19 [2], pKP1500 [1], pET2a [3], pLED-M1 [4], and two newly constructed plasmids pUCIL and pEOR (Fig. 1). The pUCIL and pEOR are derivatives of pUC19 and pKP1500, respectively. *E. coli* KP3998 and pKP1500 were generously given by Dr. Takeyoshi Miki (Kyushu University).

*Corresponding author. Fax: (81) (6) 879 7448.

*Abbreviations:* PCR, polymerase chain reaction; bp, base pair; IPTG, isopropyl-β-D-thiogalactopyranoside; SDS-PAGE, SDS-polyacrylamide gel electrophoresis

### 2.2. DNA manipulation

Preparation of plasmid DNA, enzyme reactions, and transformation of *E. coli* cells were carried out as described by Maniatis et al. [5]. Nucleotide sequences were determined using a DNA sequencing kit (Sequencing PRO, Toyobo Co. Ltd., Osaka).

### 2.3. Construction of random DNA libraries (R1MIX)

The schematic diagram for library construction is illustrated in Fig. 1A. The mixture of 140-mer single-stranded oligonucleotides (R140ss) was synthesized by Toagosei Co., Ltd. (Tokyo) according to our design. R140ss contains a randomized portion composed of 6 repeated 16-mer random oligonucleotides flanked by fixed sequences which contain the primer sites for amplification and the restriction enzyme sites (Fig. 1C). R140ss was made to the double-stranded form and amplified by the PCR method [6] using the P1 and P5 primers. The PCR product was then isolated and digested with *Bam*HI and *Pst*I. The *Bam*HI-*Pst*I fragments were purified by polyacrylamide gel electrophoresis, and ligated with pUCIL that had been digested with *Bam*HI and *Pst*I. The ligated DNA was introduced into *E. coli* KP3998. Plasmid isolation was done on the cells collected from all the transformants (about $2.4\times10^4$ colonies) grown on the plates containing 50 μg/ml ampicillin. The obtained mixture of hybrid plasmids contains 1, 2 or 3 units of the randomized portion of R140ds (double stranded form of R140ss) and was named R1MIX.

### 2.4. Construction of plasmids for random protein expression

Fig. 1B shows the strategy used for protein expression. R1MIX was digested with *Xho*I and *Bgl*II, and the 119-bp DNA fragments containing 1 unit of the randomized portion were isolated by polyacrylamide gel electrophoresis. To avoid complications in lengthening the randomized portion to several units, R1MIX was also digested with *Xho*I and *Kpn*I to obtain the 233-bp DNA fragments containing 2 units of the randomized portion. The *Xho*I-*Bgl*II and *Xho*I-*Kpn*I fragments were ligated with pEOR that had been digested with *Bam*HI and *Kpn*I. The digested pEOR vector maintains the $P_{tac}$ promoter, an epitope tag of the first 11 codons for the T7 gene 10 protein [3], and the stop codons for all the three frames. Here, it should be noted that the ends produced by *Bgl*II and *Bam*HI can be ligated and the sites were eliminated after ligation. The constructed protein expression vectors containing 3 units of the randomized portion were then used to transform *E. coli* KP3998 for the production of random proteins.

### 2.5. Expression, identification, and in vivo solubility of random proteins

*E. coli* KP3998 cells harboring the constructs were grown at 37°C on 2xTY medium [5] containing 50 μg/ml ampicillin. IPTG (final conc., 1 mM) was added to the culture with $OD_{600}$ of 0.6–0.8. IPTG induction was carried out for 2 h. The proteins in the cells before and after IPTG induction were analyzed by SDS-PAGE [7]. The cells before induction served as the negative control. The detected additional distinct band with an expected molecular weight ($M_r$ 14000–16000) were judged to be the random protein produced by cells after IPTG induction. The presence of random proteins were also identified by Western blot analysis [8] using a monoclonal antibody directed against the epitope tag (Novagen, Inc., Madison, WI).

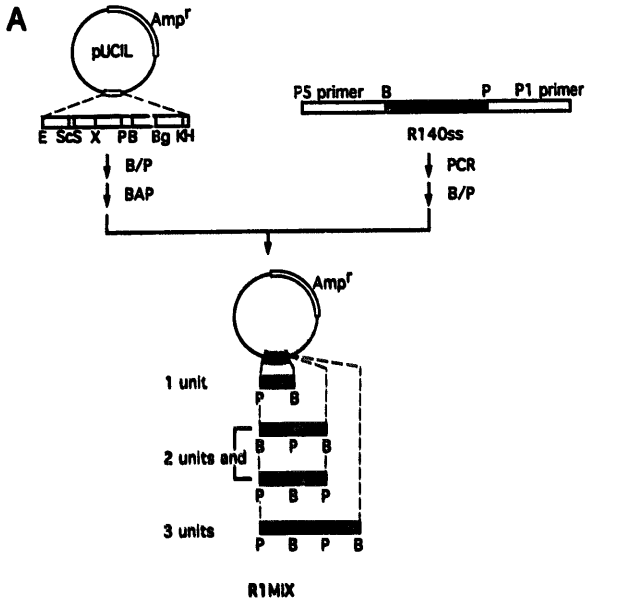
To test the solubility of the random proteins detected, cells after IPTG induction were disrupted by sonication. The supernatant and the precipitate obtained after centrifugation (12000×*g* for 10 min) of the disrupted cells were subjected to SDS-PAGE [7]. A protein detected as a distinct unique band in the supernatant was evaluated as a soluble protein.

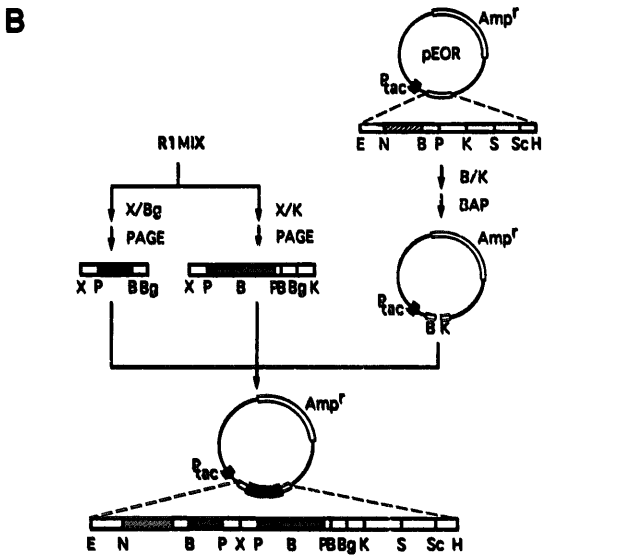

*SSDI* S0014-5793(96)00123-8

## 3. Results and discussion

The genes encoding the artificial random proteins were designed with the following criteria: (1) all the 20 kinds of amino acids are included; (2) the length of the randomized portion is about 100 amino acid residues; (3) the amino acid sequence is highly random; and (4) the mean value of the net charge of the random proteins is about +2. The above criteria were met by the synthesized randomized portion of R140ss and the strategy of constructing the gene (Fig. 1). It should be pointed out that no stop codons appear in all the six frames of the randomized portion even if frame shifts occur during the synthesis and construction, and that the mean value of the G + C content of R140ss is set to be 53.5%, as high G + C content interferes with PCR reactions. In addition, the synthesis of the six repeated 16-mer oligonucleotides (1 unit) with the strategy of ligating the units increase the randomness of the amino acid sequence in the protein library.

The gene structures in the library of R1MIX (Fig. 1A) were checked by analyzing the length of the *Xho*I-*Bgl*II fragment in the hybrid plasmids. Among the 10 clones randomly chosen from the $2.4 \times 10^4$ transformants, 7 clones contained 1 unit of the random fragment of R140ds (Fig. 1A), 2 had 2 units, and 1 had 3 units. The incorporation of 2 units of the random fragment presumably depends upon pUCIL being cleaved by only one restriction enzyme. The genes encoding artificial random proteins were constructed as shown in Fig. 1B. The main part of the gene was prepared by ligating the mixture of random fragments of 1-unit with that of the 2-unit. The genes containing the resulting 348-bp DNA fragments were designed to be expressed as fusion proteins with the epitope tag of the first 11 amino acids for the T7 gene 10 protein [3] under the control of $P_{tac}$ promoter (Fig. 1B). Hence, the random proteins are identified by their inductive expression with IPTG and by immunoblotting using an antibody directed against the



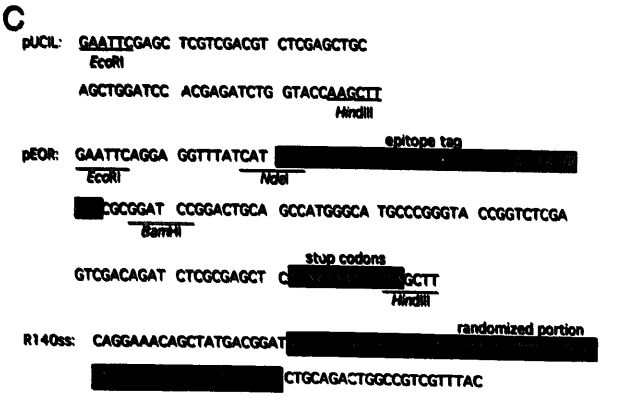
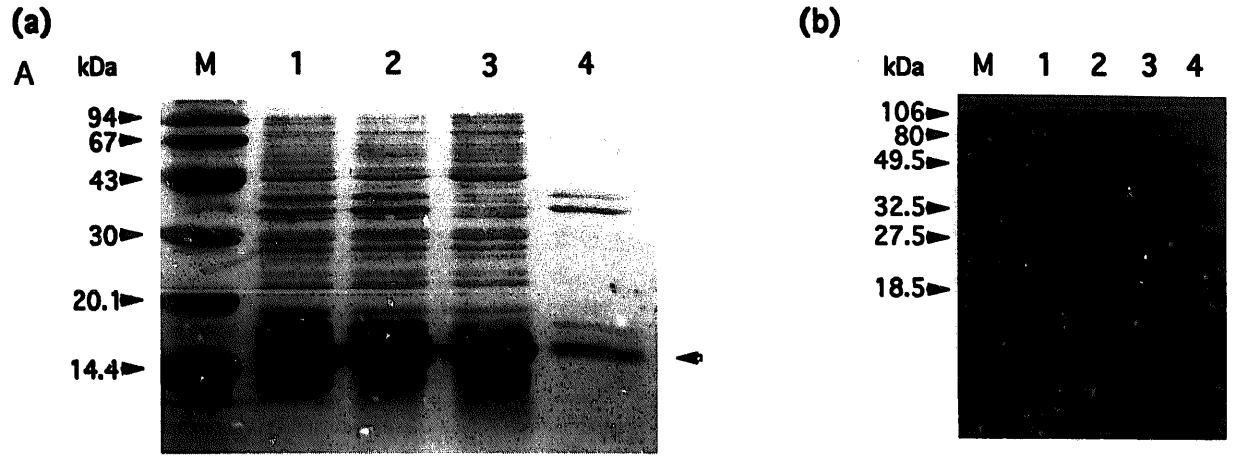



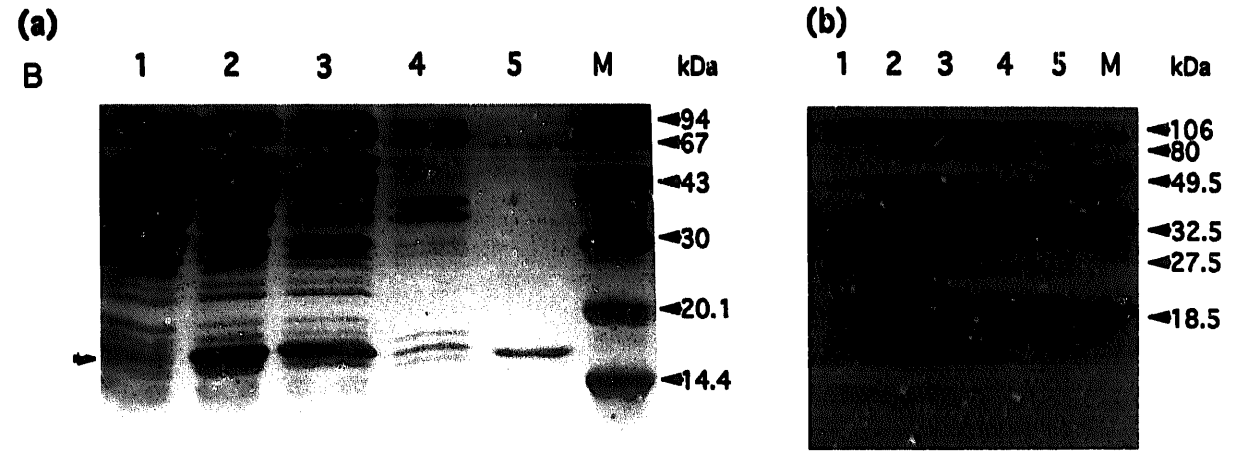

Fig. 2. Expression and solubility of random proteins RP3-04 (A), RP3-42 (B), and RP3-65 (C) and the deduced amino acid sequences of these proteins (D). The production of random proteins in the cells taken after IPTG induction were analyzed by SDS-PAGE [7]. The cells before IPTG induction were also analyzed to serve as the negative control. The samples used to test the solubility of the random proteins were prepared as described in section 2. (a) Protein staining with Coomassie brilliant blue R250; (b) Western blot analysis [8]. Lane M, marker proteins; lane 1, whole cells before IPTG induction; lane 2, whole cells after IPTG induction; lane 3, supernatant; lane 4, precipitate; lane 5, purified protein. The arrow in (a) indicates the position of each random protein.

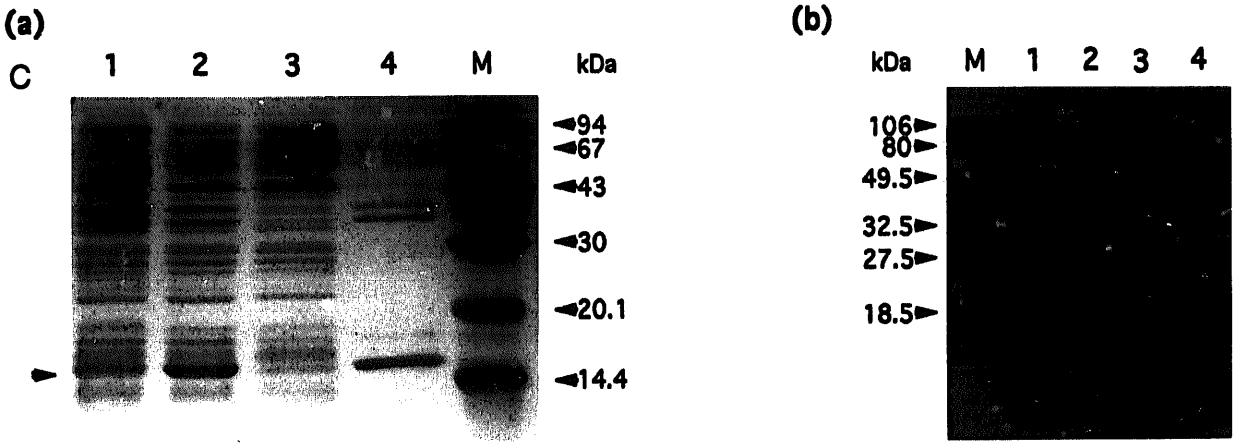
→

Fig. 1. Construction of plasmids for random protein expression. (A, B) Schematic diagram for the construction. (C) Nucleotide sequences of the *Eco*RI-*Hin*dIII fragments of pUCIL and pEOR illustrated in A and B, respectively, and the sequence of R140ss. The plasmid pUCIL was constructed by replacing the whole of the multicloning site of pUC19 [2] with the *Eco*RI-*Hin*dIII fragment shown in C. For the construction of pEOR, pET2a [3] was digested with *Nde*I and *Bam*HI, and the *Nde*I-*Bam*HI fragment containing the epitope tag was recovered and ligated to pLED-M1 [4] that had been digested with the same enzymes. The resulting plasmid (pRF-12) was then digested with *Bam*HI and *Hin*dIII prior to ligation with a synthetic oligonucleotide (*Bam*HI-*Hin*dIII fragment) containing the stop codons as shown in C. Restriction sites are abbreviated as follows: B, *Bam*HI; Bg, *Bgl*II; E, *Eco*RI; H, *Hin*dIII; K, *Kpn*I; N, *Nde*I; P, *Pst*I; S, *Sal*I; Sc, *Sac*I; X, *Xho*I. Other abbreviations are: BAP, bacterial alkaline phosphatase; Amp^r, ampicillin-resistance-coding region; $P_{tac}$, *tac* promoter. In C, the epitope tag of the first 11 codons for the T7 gene 10 protein and the stop-codon region in the fragment of pEOR, and the randomized portion of R140ss are outlined in black.

←




D

```
                                                                      70
RP3-04  MASMTGGQQM GRGSRGSSLG ALHFGGIPVW KHSKLEKSQA CSFPGWGTSA ARAAEIFQNG GLPAWNAPNM
RP3-42  MASMTGGQQM GRGSRGSILE GPHVGNPPSW GIPKLEKSQA FEPPGWDFSC SSCRNSPVWG SPRLGSSQIG
RP3-65  MASMTGGQQM GRGSRGSILG SFQFGEPPAW GAPNLEIFQL ADLPGWGFLQ LELQRPPKLG SFQDGNSPTW

                                                                     140
        ENPQSGELPG WRLSNLDPGW ESPNLEGPQF GNFPSWGNSR LGNLQIGETL QLDPRDLVPV SSRQISRALIN
        KFPEWDSLNM GDFQLGSKLE TFHVGGIPCW ETSILERSHL GGLPNWENSA AGSTRSGTGL ESTDLASSN
        ELPRLENPQY GMLPAWIQTG GLPDWEIPRL GASQFGNPPV WRAPILGGFC SWIIEIWYRS RVDRSREL
```

epitope tag. It should be pointed out that as a result of the ligation of two of the 1-unit fragment as shown in Fig. 1A, there are two possible types of 2-unit in the R1MIX. However, we could only find one of the type (2-unit ligated at the *Pst*I site) in the sequences of the 348-bp fragments isolated from the examined 10 clones (see below). This may be due to the fact that *Pst*I digestion is less efficient than *Bam*HI digestion. Hence, we suppose that most 348-bp fragments in the library have the structure shown on Fig. 1B.

If no deletions and insertions occurred during the construction of the gene library, the genes are expected to encode random proteins with 141 amino acid residues. The proteins include fixed amino acid sequences of 17 residues at the N terminal, 5 residues between the first and second random units, 2 residues between the second and third random units, and 22 residues at the C terminal. The first and second randomized portions are composed of 32 residues, and the third 31. The average amino acid composition of the random proteins in the library is expected to be Ala, 6.0; Glu, 3.7; Gln, 6.1; Asp, 3.8; Asn, 3.2; Leu, 9.4; Gly, 12.8; Lys, 2.4; Ser, 11.6; Val, 3.4; Arg, 6.7; Thr, 3.0; Pro, 10.6; Ile, 4.6; Met, 2.9; Phe, 3.1; Tyr, 0.7; Cys, 0.4; Trp, 4.3; and His, 1.3%.

The hybrid plasmids containing the genes coding for the random proteins were introduced into *E. coli* KP3998 cells, and about 13,000 ampicillin-resistant transformants were obtained. Out of the hybrid plasmids isolated from 81 randomly chosen transformants, 70 were found to contain the expected size of the *Nde*I-*Bgl*II fragments, Hence, the 70 transformants harboring these plasmids were cultivated and subjected to IPTG induction as described under section 2. SDS-PAGE analysis showed that 35 transformants clearly produced IPTG-inducible random proteins with the expected molecular weight of about 14–16 kDa. The Western blot analysis [8] of the cells of the randomly chosen 25 transformants from the 35 showed that the IPTG-inducible protein bands on SDS-PAGE contain the epitope tag. Some of the results are shown in Fig. 2.

The solubility of the detected 25 random proteins was examined as described under section 2. The proteins detected in the supernatant are classified as soluble, and those not detected as insoluble. Among the proteins tested, 5 are soluble and are designated as RP3-04, RP3-29, RP3-42, RP3-45, and RP3-54. These results indicate that about 20% of the random proteins expressed in *E. coli* are expected to be soluble. The SDS-PAGE of the soluble (RP3-04 and RP3-42) and the insoluble (RP3-65) proteins are shown in Fig. 2. It should be noted that with RP3-04, the protein is also detected in the precipitate, indicating a lower solubility than that of RP3-42, a protein detected only in the supernatant.

The soluble proteins RP3-42 and RP3-45 were purified by subsequent heat treatment (65°C, 25 min), ammonium sulfate precipitation, and DEAE-Sepharose column chromatography. The purified RP3-42 and RP3-45 are electrophoretically homogeneous as shown in Fig. 2 for RP3-42. The purified form of both proteins are soluble as defined. On the other hand, the purification of the insoluble protein RP3-70 was carried out by preparative SDS-PAGE. When the purified RP3-70 finally dissolved in 5 M urea was dialyzed against water, the protein precipitated, hence is insoluble. Therefore, the solubility of the purified form is in good agreement with the results from the crude level. We are now investigating the properties of the purified proteins and the results will be published elsewhere.

Nucleotide sequences of the genes encoding all the five soluble proteins and five arbitrarily chosen insoluble proteins were identified. All the proteins were found to have different deduced amino acid sequences. The deduced amino acid sequences of RP3-04, RP3-42, and RP3-65 are shown in Fig. 2D. In the synthesis of the randomized portion of R140ss, the mixed base solutions were programmed to have equal concentration of each of the bases. However, we found that T and G have a higher tendency to be incorporated than C and A (Table 1). In addition, different deletion sites within the random parts of the genes were also found, that is, 2 for RP3-42, RP3-45, and RP3-54, 3 for RP3-64, and 1 for RP3-65. It should be noted that these frame shifts do not lower the efficiency of the random proteins to be expressed but rather increase the variety of the library owing to our design of R140ss (Fig. 1C).

Recently, Davidson and Sauer have reported that from the library of 80- to 100-residue proteins which is composed mainly of random combinations of glutamine, leucine, and arginine, no proteins were found to be soluble without the aid of denaturants [10]. When the average leucine content was reduced from 40% to 28%, Davidson et al. found two soluble proteins out of eleven [11]. The contents of the nonpolar residues in RP3-04, RP3-42, and RP3-65 calculated from the sequences shown in Fig. 2D are 56, 51, and 60%, respectively. These results suggest that the solubility of proteins is partially governed by hydrophobicity. However, statistical analysis will be needed to give a conclusive correlation between the solubility and the amino acid composition of the proteins. The accumulation of sequence data of the random proteins is now in progress for such analysis.

In this work, it was found that about 20% of the randomly prepared proteins with 141 amino acid residues are expected to be soluble. This indicates that significant number of proteins emerged randomly in the course of evolution could be

Table 1
Efficiency of base incorporation at the randomized sites

| Randomized site | No. of incorporation of | | | | |
|---|---|---|---|---|---|
| | A | T | G | C | Total |
| A, T, G or C | 76 (21%) | 105 (30%) | 106 (30%) | 69 (19%) | 356 |
| A, T or G | 57 (32%) | 75 (42%) | 48 (27%) | – | 180 |
| A, T or C | 68 (39%) | 61 (35%) | – | 47 (27%) | 176 |
| A or G | 239 (46%) | – | 286 (55%) | – | 525 |
| T or C | – | 300 (57%) | – | 228 (43%) | 528 |

The number of bases at the randomized sites was counted from the 10 nucleotide sequences encoding the 5 soluble proteins (RP3-04, RP3-29, RP3-42, RP3-45, and RP3-54) and 5 arbitrarily chosen insoluble proteins (RP3-61, RP3-64, RP3-65, RP3-66, and RP3-70). The sites affected by deletion were not counted.

soluble. As the soluble random proteins are newly expressed in the cells, they can be noted as initial proteins in the course of evolution. Hence, these proteins can serve as a good model of ancestral proteins in the study of enzyme evolution, and may as well lead us to the statement of soluble enzymes known today arise from the line of soluble ancestry.

*Acknowledgements:* This work was supported in part by Grants (07280213 and 07780573) from the Ministry of Education, Science, Sports and Culture, Japan.

## References

[1] Miki, T., Yasukochi, T., Nagatani, H., Furuno, M., Orita, T., Yamada, H., Imoto, T. and Horiuchi, T. (1987) Protein Eng. 1, 327–332.
[2] Yanisch-Perron, C., Vieria, J. and Messing, J. (1985) Gene 33, 103–119.
[3] Studier, F.W., Rosenberg, A.H., Dunn, J.J. and Dubendorff, J.W. (1990) Methods Enzymol. 185, 60–89.
[4] Asakura, K., Komatsubara, H., Soga, S., Yomo, T., Oka, M., Emi, S. and Urabe, I. (1993) J. Ferment. Bioeng. 76, 265–269.
[5] Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) in: Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
[6] Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988) Science 239, 487–491.
[7] Laemmli, U.K. (1970) Nature 227, 680–685.
[8] Tsang, V.C., Peralta, J.M. and Simons, A.R. (1983) Methods Enzymol. 92, 377–391.
[9] Fauchere, J.-L. and Pliska, V. (1983) Eur. J. Med. Chem. 18, 369–375.
[10] Davidson, A.R. and Sauer, R.T. (1994) Proc. Natl. Acad. Sci. USA 91, 2146–2150.
[11] Davidson, A.R., Lumb, K.J. and Sauer, R.T. (1995) Nature Struct. Biol. 2, 856–864.