Available online at www.sciencedirect.com
SCIENCE DIRECT®
Protein Expression and Purification 42 (2005) 268–277
Protein Expression & Purification
www.elsevier.com/locate/yprep

Improving expression and solubility of rice proteins produced as fusion proteins in *Escherichia coli*

Yuki Tsunoda [a], Nobuya Sakai [a], Koji Kikuchi [a], Shizue Katoh [a], Kayo Akagi [a], Jun Miura-Ohnuma [a], Yumiko Tashiro [a], Katsuyoshi Murata [a], Naoto Shibuya [b], Etsuko Katoh [a,*]

[a] *Biochemistry Department, National Institute of Agrobiological Sciences, Tsukuba, Ibaraki 305-8602, Japan*
[b] *Department of Life Sciences, Faculty of Agriculture, Meiji University1-1-1 Higashi-Mita, Tama-ku, Kawasaki, Kanagawa 214-8571, Japan*

Received 21 December 2004, and in revised form 1 April 2005
Available online 27 April 2005

Abstract

For proteins of higher eukaryotes, such as plants, which have large genomes, recombinant protein expression and purification are often difficult. Expression levels tend to be low and the expressed proteins tend to misfold and aggregate. We tested seven different expression vectors in *Escherichia coli* for rapid subcloning of rice genes and for protein expression and solubility levels. Each expressed gene product has an N-terminal fusion protein and/or tag, and an engineered protease site upstream of the mature rice protein. Several different fusion proteins/tags and protease sites were tested. We found that the fusion proteins and the protease sites have significant and varying effects on expression and solubility levels. The expression vector with the most favorable characteristics is pDEST-trx. The vector, which is a modified version of the commercially available expression vector, pET-32a, contains an N-terminal thioredoxin fusion protein and a hexahistidine tag, and is adapted to the Gateway expression system. However, addition of an engineered protease site could drastically change the expression and solubility properties. We selected 135 genes corresponding to potentially interesting rice proteins, transferred the genes from cDNAs to expression vectors, and engineered in suitable protease sites N-terminal to the mature proteins. Of 135 genes, 131 (97.0%) could be expressed and 72 (53.3%) were soluble when the fusion proteins/tags were present. Thirty-eight mature-length rice proteins and domains (28.1%) are suitable for NMR solution structure studies and/or X-ray crystallography. Our expression systems are useful for the production of soluble plant proteins in *E. coli* to be used for structural genomics studies.

Keywords: Rice; Protein expression; Gateway system; Subcloning

Structural biology, one of the most exciting fields in molecular biology today, is facing a massive increase in the potential number of available targets because of successful large-scale sequencing efforts of the genomes of various organisms. Consequently, a complete structural characterization of the proteins of an organism's genome, i.e., structural genomics, is a realistic possibility, especially for the subset of aqueous (non-membranous) proteins [1]. Initially, structural genomics projects focused on the proteins of organisms with relatively small genomes, e.g., those of bacteria, for which the number of genes is commonly less than about 2000 [2–6]. For structural genomic studies of higher eukaryotic proteins with much larger genomes, there are still several technical bottlenecks, including the development of reliable high-throughput protein expression and purification systems.

Sequenced complex plant genomes, such as those of *Oryza sativa* (rice) [7] or *Arabidopsis thaliana* [8], will

* Corresponding author. Fax: +81 298 38 8399.
E-mail address: ekatoh@nias.affrc.go.jp (E. Katoh).

1046-5928/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.pep.2005.04.002

provide the foundation for plant structural genomic studies, as the human [9] and mouse [10] genome-sequencing projects, as examples, have done for ongoing mammalian structural genomic projects. The rice genome and proteome are suitable and important targets. The plant is a major food source and it has a relatively small genome in comparison with those of other crops, such as wheat and corn. Draft sequences of the *O. sativa* L. ssp. *Indica* [11] and *japonica* [12] genomes have been published and include complete chromosome 1 and 4 sequences [13,14]. Furthermore, about 30,000 full-length complementary rice DNA clones exist [15] and are publicly available (http://cdna01.dna.affrc.go.jp/cDNA/). Thus, the foundation for a structural genomic project targeting rice proteins is laid. However, we felt it necessary to first improve the techniques used for high-throughput preparation and yield of recombinant proteins suitable for structural analysis.

Production of a recombinant protein involves two processes: gene cloning, which is followed by protein expression and purification. Ideally, gene cloning (subcloning) should be a routine procedure. Improved cloning efficiency and reliability is achieved by establishing a universal set of restriction endonuclease sites for both the cloning and expression vectors. Commercial systems, such as Gateway from Invitrogen [16] and Creator from BD Biosciences [17], eliminate the use of restriction enzymes during subcloning and, instead, employ a universal entry clone in conjunction with compatible recombination vectors. Gateway is a ligation-independent cloning system, which permits simultaneous plasmid preparation for different expression systems.

To facilitate purification, affinity tags such as hexahistidine (His6)[1] [18,19], maltose binding protein (MBP) [20,21], and glutathione *S*-transferase (GST) [22] are often attached to the N- or C-termini of recombinant proteins. For protein expression and folding, tags, such as GST, thioredoxin (Trx) [23], and MBP, are used. Unfortunately, the Gateway system, while very useful for subcloning, introduces the additional sequence, for example XTSLYKKAGSAAALFNFKKEPFT, between the N-terminus of the recombinant protein and the fusion tag. To avoid potential complications, this sequence should be removed before the start of structural studies.

It is generally difficult to improve the expression and solubility levels of a recombinant protein because these properties inherently depend on the amino acid sequence of the protein. The amino acid sequence in turn determines the physical properties of the protein, including its stability, p*I*, hydrophobicity, and molecular weight—any or all of which might directly affect the expression and solubility levels. *Escherichia coli* is still the expression host of choice. It is easy to use, inexpensive, readily available, and large quantities of it can be rapidly grown. However, it is difficult to express eukaryotic proteins in *E. coli*; the bacterium lacks the machinery necessary for eukaryotic post-translational modifications; and the expressed protein is often found in inclusion bodies [24]. Added fusion protein tags, such as GST, Trx, and MBP, often improve the solubility of recombinant expression products.

We modified three different expression vectors that are commercially available and incorporated them into the Gateway system in order to test the effects of purification and solubility tags on the yield and solubility of rice proteins when expressed in *E. coli*. As noted, the Gateway system introduces an additional amino acid sequence upstream of the parent protein. To avoid possible interference of this sequence during structural studies, we introduced a protease site between the Gateway sequence and the native N-terminus. Then we tested the effects of different protease sites on expression levels and solubilities. To exhaustively explore the usefulness of the methods reported herein, we studied the expression systems of more than 100 rice proteins and functional domains with uncharacterized structures.

Materials and methods

Preparation of modified expression vectors

The commercially available vectors, pET-16b (Novagen) and pMAL-c2X (NEB), were cut with *Bam*HI and the commercially available vector, pET-32a (Novagen), was cut with *Eco*RV. The restriction enzyme cuts were made at positions that were as close to the 3′ ends of the fusion or tag nucleotide sequences as possible. pET-16b and pMAL-c2X were treated with Klenow polymerase (Promega) to create blunt ends. The terminal 5′ phosphate groups were removed with calf intestinal alkaline phosphatase (Promega). Gateway cloning cassettes—cassette A for pET-32a and pMAL-c2X and cassette C.1 for pET-16b—were chosen so that the reading frames of the fusion genes continued through the recombination sites. The ligated plasmids were transformed into DB3.1 cells (Invitrogen) and sequence fidelities confirmed by DNA sequencing. The expression vectors, used in this study, are characterized according to their origin, the type of fusion tag incorporated, and the promoter region in Table 1.

Subcloning into the entry and expression vectors

cDNA clones were streaked onto fresh Luria–Bertani medium plates that also contained 50 μg/mL of ampicillin (LB^{amp}) and, after growth, single colonies picked for plasmid preparation using a Wizard Plus SV miniprep kit (Promega). The genes were PCR amplified using the

[1] *Abbreviations used:* His6, hexahistidine; MBP, maltose binding protein; GST, glutathione *S*-transferase; Trx, thioredoxin; IPTG, isopropyl-β-D-thiogalactopyranoside; DTT, dithiothreitol.