Inteins are protein sequences embedded in precursor proteins, and they catalyze the protein-splicing reaction that excises the intein and joins the flanking sequences (N- and C-exteins) with a peptide bond (1). A typical protein splicing reaction has 4 steps including N → S (or N → O) acyl shift at the N terminus of intein, trans-esterification forming a branched intermediate, cyclization of an Asn residue at the C terminus of intein, and formation of a peptide bond between the flanking exteins (2–4). A typical intein is 1258 aa long and three times as large as a typical intein, due to the presence of large tandem repeats in which a 57-aa sequence is repeated 17 times. The DnaE-2 intein has a more typical size of 428 aa with putative protein splicing and endonuclease domains. The DnaE-3 intein is a split intein consisting of a 102-aa N-terminal part and a 36-aa C-terminal part encoded on a split mini-intein, and an unusual intein containing large tandem repeats that inhibited protein splicing.

EXPERIMENTAL PROCEDURES

Gene Cloning and Analysis—GenBank™ searches and protein sequence alignments were performed using the BLAST search program (17) and the Clustal W program (18). Parts of the Ter dnaE gene were amplified by doing polymerase chain reaction (PCR) from total genomic DNA of T. erythraeum strain IMS101. The DnaE-1 intein coding sequence was PCR-amplified using primers M144 (5'-ATGTCCT-TCGTCGGTCYTCCATATC-3') and M145 (5'-ATCAAAATGCGCT-TCATTTGTAATC-3') and cloned in a pDrive plasmid vector (Qiagen). To allow complete DNA sequence determination, nested deletions were made in the cloned DNA by doing partial digestion at XmnI restriction sites present once in each of the 171-bp repeat sequences. In Southern blot analysis, ~2 μg of T. erythraeum genomic DNA was digested with restriction enzymes BfuAI and BglII, resolved by electrophoresis in 1% agarose gel, and blotted onto nylon membrane. The membrane was hybridized to a specific 32P-labeled DNA probe for 20–24 h and washed under high stringency (65 °C in 0.2× SSC and 1% SDS) before exposing x-ray films.

Protein Splicing Analysis in Escherichia coli Cells—To construct gene expression plasmids, Ter DnaE-1 intein coding sequence (PCR-amplified using the above M144 and M145 primers) was inserted in the previously made pMST plasmid (19) between XhoI and AgeI sites, replacing the Ssp II DnaB intein coding sequence of pMST. Protein production in E. coli cells, gel electrophoresis and Western blot analysis were performed as before (19). Briefly, cells containing the expression plasmid were grown in liquid Luria Broth medium at 37 °C to late log phase (A600, 0.5), isopropyl-1-thio-β-D-galactopyranoside was added to a final concentration of 0.8 mM to induce production of the recombinant protein, and the induction was continued for 3 h at 37 °C. Cells were then harvested and lysed in SDS- and dithiothreitol-containing gel loading buffer in a boiling water bath before electrophoresis in SDS-polyacrylamide gel. Western blots were carried out using anti-thioredoxin antibody (American Diagnostica) and the enhanced chemiluminescence detection kit (ECL). Intensity of protein band was estimated using a gel documentation system (Gel Doc 1000 coupled with Molecular Analyst software, Bio-Rad).

RESULTS

Split dnaE Genes Contain Tandem Repeats—To find new split dnaE genes using our previous finding (14), we searched the GenBank™ containing the nearly complete genome sequence of T. erythraeum strain IMS101 that had been determined at the Department of Energy Joint Genome Institute (www.jgi.doe.gov). As illustrated in Fig. 1, parts of putative split dnaE genes were found on three sequences with accession numbers AAAU01000106, AAAU01000117, and AAAU01000050, respectively. Our initial analysis indicated that the first split dnaE gene spans an undeter-
Dotted line connecting A-106 and A-117 represents DNA sequence that was cloned and determined in this study. Middle, the split dnaE genes predicted two precursor proteins (drawn to scale) consisting of four exteins (E1, E2, E3, and E4 (black boxes)) plus three inteins (open boxes) marked as DnaE-1 intein (including the tandem repeats), DnaE-2 intein, and DnaE-3 split intein. Bottom, predicted protein products after two cis-splicing reactions and one trans-splicing reaction, including a mature DnaE protein and the excised intein.

protein, whereas Ala is found at corresponding positions in other DnaE proteins lacking the intein. The DnaE-3 intein is followed by Cys in all three compared DnaE proteins containing the intein. These are consistent with the requirement of a nucleophile amino acid (e.g., Cys) after an intein (3) and may explain the absence of DnaE-1 and DnaE-2 inteins in the other DnaE proteins.

The Total DnaE-1 intein has a predicted 1258-aa sequence consisting of a 299-aa N-terminal sequence, a 969-aa repetitive sequence (17 tandem repeats of a 57-aa sequence), and a 40-aa C-terminal sequence (Fig. 3B). The Term DnaE gene would be 339 aa long, after excluding the tandem repeats, and has putative intein sequence motifs (20–22) for a protein splicing domain and an endonuclease domain (Fig. 3B), although it showed less than 15% sequence identity to other known inteins. Presence of the tandem repeats makes the Term DnaE-1 intein three times as large as a typical intein. The tandem repeats are located at or near the C-terminal boundary between the putative endonuclease domain and the splicing domain. Coding sequence of the tandem repeats is most likely translated, because its 2907-bp sequence is maintained as a continuous open reading frame with its flanking coding sequences, while its non-coding frames contain numerous termination codons as expected. Among the 17 tandem repeats, the first 16 (R1–R16) are 91–100% identical to each other in DNA sequence and 88–100% identical to each other in protein sequence, and the last repeat (R17) differs from the others at the C-terminal end (Fig. 3B). The Term DnaE-2 intein resembles typical bi-functional self-splicing inteins, with a predicted 428-aa sequence that has putative sequence motifs for both a protein splicing domain and an endonuclease domain (Fig. 3B). The Term DnaE-2 intein is significantly similar to the Ssp DnaB intein, showing 27% protein sequence identity and 46% protein sequence similarity.

The Term DnaE-3 intein is a split intein encoded by two split dnaE genes found on two yet unlinked sequence contigs (Fig. 1). These two split dnaE genes could be on opposite DNA strands of the T. erythraeum genome, as was the case in another cyanobacterium (14), or they could be on the same DNA strand and at least 56,449 bp apart, based on the available DNA sequences. Predicted protein sequences of the DnaE-3 split intein consist of a 102-aa N-terminal part encoded on the first split dnaE gene and a 36-aa C-terminal part encoded on the second split dnaE gene (Figs. 1 and 3C). GTG was assumed to be the start codon of the second split dnaE gene, based on
the absence of an upstream ATG codon or an appropriate down-stream ATG codon, and also based on amino acid sequence comparisons with other DnaE split inteins (Fig. 3C). The Ter DnaE-3 split intein is 59% identical to the previously characterized Ssp DnaE split intein, although the N-terminal part of the Ter DnaE-3 split intein is 21 aa shorter than that of the Ssp DnaE split intein (Fig. 3C).

**Tandem Repeats in Ter DnaE-1 Intein Affect Protein Splicing**—To determine whether the Ter DnaE-1 intein is functionally affected by the presence of the tandem repeats, we tested its protein splicing activity in E. coli cells. A plasmid-borne fusion gene was constructed to produce a fusion protein in which the Ter DnaE-1 intein (plus its 5-aa native extein sequence on each side) was fused to an N-terminal maltose-binding protein and a C-terminal thioredoxin (Fig. 4). Similar fusion protein containing other inteins has been used in previous studies (19), so that the protein splicing products are readily identified using SDS-polyacrylamide gel electrophoresis and Western blotting. Precursor protein, spliced protein, and excised intein were identified by their predicted sizes, as well as by their lack of putative trans-acting splicing reaction. The predicted mature DnaE protein, after trans-splicing reaction. The predicted mature DnaE protein, after trans-splicing, intensity of individual protein band on Western blot was used to measure the amount of that protein, and protein splicing efficiency was calculated as the amount of spliced protein divided by the sum of spliced protein and precursor protein. The efficiency of protein splicing was estimated to be 73, 55, 50, and 45% for Ter DnaE-1 intein containing 0, 1, 2, and 3 of the 57-aa repeating units, respectively. Ter DnaE-1 intein containing 4 of the 57-aa repeating units showed less than 5% protein splicing.

**DISCUSSION**

**Split dnaE Genes and Multiple Inteins**—We have shown that two split dnaE genes together encode a DnaE protein containing three inteins in *T. erythraeum*, predicting that the synthesis of mature DnaE protein involves expression of the two split dnaE genes followed by two protein cis-splicing reactions and one protein trans-splicing reaction. The predicted mature DnaE protein, after excluding the inteins, is very similar to known DnaE proteins of related organisms in size and sequence, and no other dnaE-like gene was found in the nearly complete genome sequence of *T. erythraeum*. The three inteins were readily identified in the Ter DnaE protein, because they are insertion sequences showing clear sequence similarities to known inteins. The DnaE-1 and DnaE-2 inteins may be relatively recent acquisitions in the *T. erythraeum* genome, whereas the third intein (DnaE-3) is of unknown origin, having sequence similarities to known inteins. The DnaE-1 and DnaE-2 inteins were found only in *T. erythraeum*, while the DnaE-3 intein and the cyanobacterial *Ssp* intein are significantly lower than those seen among non-homologous inteins (12, 23), which may suggest a common ancestor for these two inteins, although their host proteins or insertion sites have no apparent similarity. The Ter DnaE-1 intein is of unknown origin, having...
Intein Containing Tandem Repeats

FIG. 4. Protein splicing of Ter DnaE-1 intein. Top, schematic illustration of the fusion protein construct consisting of maltose-binding protein sequence (M), intein sequence (black box), and thioredoxin sequence (T). Middle, predicted sizes of protein products from different fusion protein constructs (numbered as lanes 2 through 7) containing the specified number of the 57-aa repeating units (DnaE-1 intein, 0, 1, 2, 3, 4, and 17 repeats). Fusion protein construct containing the Sep DnaB mini-intein (numbered as 1) was included as a known standard for identifying the spliced protein (19). Bottom, observation of protein splicing. Total cellular proteins of E. coli cells producing the specified fusion protein were resolved by SDS-PAGE and visualized by Coomassie Blue staining (left panel) or Western blot using anti-thioredoxin antibody (right panel). The lane numbers correspond to the above fusion protein construct numbers. Position of the precursor protein is marked by the letter P, the spliced protein by the letter S, and the excised intein by a black dot.

very low sequence similarity to other known inteins even after excluding the tandem repeats. The Ter DnaE-3 split intein, present in at least two other cyanobacterial species (9), likely has an ancient origin preceding the divergence of the distinctly related cya-

nabacterial species. This split intein is least likely to spread through intein homing or lateral gene transfer, because it lacks an endonuclease domain and is encoded on split genes. The DnaE-1, DnaE-2, and DnaE-3 inteins are clustered in a small (70 aa) region of the large (1270 aa, excluding the inteins) DnaE protein, suggesting that this small region may be a hotspot for intein insertions. Intein hotspots may also exist in other proteins, such as a DNA replication factor protein in which three inteins (Mja RFC-1, RFC-2, and RFC-3 inteins) were found in an 87-aa region (9). The Ter DnaE protein is unique in that its three inteins include both cis-splicing inteins and trans-splicing split intein.

Tandem Repeats Inside Intein—The Ter DnaE-1 intein is most unusual with the presence of large tandem repeats, which was revealed in sequence determination of the cloned DNA and confirmed in Southern blot analysis of the genome. This is the first repetitive sequence found inside an intein. The tandem repeats are strikingly large (57 aa repeated 17 times), unlike the single amino acid repeats associated with Huntington’s disease (24). Also unlike typical repetitive sequences that are dispersed in many parts of the genome and often not translated, the tandem repeats of Ter DnaE-1 intein were translated and not detected in other parts of the T. erythraeum genome in GenBankTM searches and in Southern blot analysis. The tandem repeats may have originated from duplications of a pre-existing piece of the intein sequence, although they did not show similarity to known intein sequences and were not required for the protein splicing function of the intein. Alternatively, the tandem repeats may have originated from exogenous sequences that entered the Ter DnaE-1 intein. Self-splicing inteins have been thought of as “safe havens” for homing endonucleases present in most inteins, because the ability of self-removal through protein splicing of the intein makes them tolerated by the host protein. The Ter DnaE-1 intein can be such a safe haven to the tandem repeats, if the latter were selfish genetic elements of unknown function. The Ter DnaE-1 intein could also be a “waste bin,” if the tandem repeats had no function at all.

Possible Function of the Tandem Repeats—Perhaps more likely, the tandem repeats may play a biological role in T. erythraeum, which would provide a reason for this organism to maintain the tandem repeats. Tandem repeats are prone to changes in length, because of polymerase slippage in DNA replication and unequal crossing over in DNA recombination. Tandem repeats of Ter DnaE-1 intein exhibited uniform size in PCR and Southern blot analysis, which may suggest unknown mechanism or reason for maintaining the number of the 57-aa repeating units at 17. Individual repeating units of the tandem repeats showed less than 10% DNA sequence heterogeneity and less than 12% protein sequence heterogeneity, which may also suggest unknown mechanism or reason for maintaining the tandem repeats. Different tandem repeats in a number of other proteins have been known to play structural or ligand/substrate-binding roles, which include examples like mucin, cell surface proteins, and proteins containing spec-

trin repeats (25–27). Our results suggested that the tandem repeats of Ter DnaE-1 intein may regulate the protein splicing activity of this intein, because protein splicing in E. coli cells was inhibited when 0 or more of the 57-aa repeating units were present in the intein. This may predict a regulatory mechanism in the native T. erythraeum cell that overcomes the inhibition of protein splicing by the tandem repeats or reduces the number of the 57-aa repeating units to less than 4, to produce a mature DnaE protein (catalytic subunit of DNA polymerase III). The predicted regulatory mechanism may regulate DNA polymerase III synthesis and in turn DNA replication and cell proliferation. Interestingly, the N2-fixing T. eryth-

raeum is known to form massive ocean surface bloom (rapid cell proliferation) with important ecological consequences (15, 16), although it is not known what controls the bloom, and rapid DNA replication must accompany the rapid cell proliferation.

Acknowledgment—We thank Dr. J. B. Waterbury for the gift of T. erythraeum cells used in this study.

REFERENCES
1. Perler, F. B., Davis, E. O., Dean, G. E., Gimble, F. S., Jack, W. E., Neff, N., Noren, C. J., Thorner, J., and Belfort, M. (1994) Nucleic Acids Res. 22, 1125–1127
2. Xu, M. Q., and Perler, F. B. (1996) EMBO J. 15, 5146–5153
3. Perler, F. B., Xu, M. Q., and Paulus, H. (1997) Curr. Opin. Chem. Biol. 1, 292–299
4. Paulus, H. (2000) Annu. Rev. Biochem. 69, 447–496
5. Duan, X., Gimble, F. S., and Quicocho, F. A. (1997) Cell 89, 555–564
6. Cooper, A. A., and Stevens, T. H. (1995) Trends Biochem. Sci. 20, 351–356
7. Gimble, F. S. (2000) FEBS Lett. 485, 103–107
8. Belfort, M., and Roberts, R. J. (1997) Nucleic Acids Res. 25, 3379–3388
9. Perler, F. B. (2002) Nucleic Acids Res. 30, 383–384
10. Amiotz, G., Bolenczyk, O., Dasze, B., Shamsiayaka, A., and Pietrokovski, S. (2003) Mol. Microbiol. 47, 61–73
11. Perler, F. B. (1998) Cell 92, 1–4
12. Gogarten, J. P., Senanjani, A. G., Zhaxybayeva, O., Olendzenski, L., and Hilario, E. (2002) Annu. Rev. Microbiol. 56, 263–287
13. Liu, Q. X. (2000) Annu. Rev. Genet. 34, 61–76
14. Wu, H., Hu, Z., and Liu, Q. X. (1998) Proc. Natl. Acad. Sci. U. S. A. 95, 9228–9231
15. Orcutt, K. M., Rasmussen, U., Webb, E. A., Waterbury, J. B., Gundersen, K., and Bergman, B. (2002) Appl. Environ. Microbiol. 68, 2236–2245
16. Mischke, M. R., and Capone, D. G. (2000) Trends Plant Sci. 5, 148–153
17. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Nucleic Acids Res. 25, 3389–3402
18. Higgins, D. G., Thompson, J. D., and Gibson, T. J. (1996) Methods Enzymol. 266, 383–402
19. Wu, H., Xu, M. Q., and Liu, Q. X. (1998) Biochim. Biophys. Acta 1387, 422–432
20. Pietrokovski, S. (1994) Protein Sci. 3, 2350–2359
21. Pietrokovski, S. (1998) Protein Sci. 7, 64–71
22. Dalgaard, J. Z., Klar, A. J., Moser, M. J., Holley, W. R., Chatterjee, A., and Pietrokovski, S. (2002) Trends Biochem. Sci. 27, 441–444
23. Dnjovic-Carugo, K., Goult, M., Ylanne, J., and Young, P. (2002) FEBS Lett. 512, 119–123
24. de La Fuente, J., García-García, J. C., Biloan, E. F., Rodríguez, S. D., García, M. A., and Kocan, K. M. (2001) Anim. Health Res. Rev. 2, 163–173

Downloaded from http://www.jbc.org/ by guest on July 23, 2018
Split dnaE Genes Encoding Multiple Novel Inteins in *Trichodesmium erythraeum*

Xiang-Qin Liu and Jing Yang

*J. Biol. Chem.* 2003, 278:26315-26318.
doi: 10.1074/jbc.C300202200 originally published online May 24, 2003

Access the most updated version of this article at doi: 10.1074/jbc.C300202200

Alerts:
- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 27 references, 2 of which can be accessed free at http://www.jbc.org/content/278/29/26315.full.html#ref-list-1