344–345 Nucleic Acids Research, 2000, Vol. 28, No. 1

InBase, the Intein Database

Francine B. Perler*

New England BioLabs, Inc., 32 Tozer Road, Beverly, MA 01915, USA

Received September 29, 1999; Accepted October 7, 1999

ABSTRACT

InBase, the Intein Database (http://www.neb.com/neb/inteins.html), is a comprehensive on-line resource that includes the Intein Registry. Inteins are protein splicing elements that mediate a self-catalytic protein splicing reaction. InBase presents general information as well as detailed data for each intein, including tabulated comparisons and a comprehensive bibliography.

INTRODUCTION

Inteins are in-frame intervening sequences that disrupt the coding region of a host gene. They are post-translationally excised from a protein precursor by a self-catalytic protein splicing process (reviewed in 1). The N- and C-terminal host gene (and protein) fragments are called the N- and C-exteins. The ligation of exteins and the presence of conserved intein motifs differentiate intein mediated protein splicing from other post-translational processing reactions. Inteins are composite proteins consisting of a homing endonuclease domain (2) and a protein splicing domain, which is formed by the two splicing regions that flank the homing endonuclease (3,4). Mini-inteins (134–198 amino acids) are minimal protein splicing elements with a linker in place of the homing endonuclease (3–6). Larger inteins range in size from 360 to 608 amino acids (5). Homing endonucleases were first described in mobile introns; they make a double strand break at the intein or intron insertion site, initiating mobilization of the intein or intron gene into the same site in a homologous host gene lacking the intervening sequence (7,8).

InBase is a comprehensive on-line database (http://www.neb.com/neb/inteins.html) that compiles information about inteins, often prior to publication. InBase is the home of the INTEIN REGISTRY, which lists all known putative inteins. Inteins are sorted both by organism (Section 2A) and by insertion site in the extein (Section 2B). Several subsets of data are tabulated for easy comparison including motifs, insertion site sequences, selected properties and intein alleles.

NEW DEVELOPMENTS

New to InBase this year are: (i) an on-line submission form, (ii) spotlighting splicing motif polymorphisms and (iii) more extensive annotation of reference categories (reviews, applications and related papers).

Several important advances were reported this year in the protein splicing field. Most notable was the discovery of a naturally occurring split protein splicing precursor which must splice *in trans* to generate the *Synechocystis* sp. PCC6803 replicative DNA polymerase catalytic subunit from two separate precursor fragments (9,10). On the phylogenetic front, the complete genome sequence of three closely related archaea (*Pyrococcus abyssi*, *Pyrococcus furiosus* and *Pyrococcus horikoshii*) emphasized the sporadic occurrence of inteins (5). On the technical front, intein vectors for protein purification emerged as new tools for protein semisynthesis and protein modification (reviewed in 1,11–15). The IMPACT™ protein expression and purification system (New England BioLabs), and similar intein vectors with the target protein fused to the intein N-terminus, yield polypeptides with C-terminal α-thioesters after treatment with thiol reagents (16–19). Polypeptides with reactive N-terminal cysteine residues can be isolated from protein purification vectors in which the target protein is fused to the intein C-terminus (18–21). Proteins with C-terminal α-thioesters can be used in various types of chemoselective condensations (1,11–15): (i) to synthesize larger proteins *in vitro* by ligation with polypeptides containing N-terminal cysteines [termed intein-mediated protein ligation, IPL (17), or expressed protein ligation, EPL (22)], (ii) to use IPL to incorporate modified or unnatural amino acids, biosensors, fluorescent tags, biotinylated tags, etc., into an expressed protein, (iii) to use IPL to segmentally modify or label proteins, extending the limits of NMR analysis (14,15), for example, and (iv) to generate C-terminal thiocarboxylates for the biosynthesis of thiamin (23).

Potential early roles for inteins in evolution have been hypothesized (24), inspired by the ability of inteins to perform both protein splicing *in trans* and intein mediated protein ligation *in vitro*. Protein splicing *in trans* could shuffle protein domains to generate more complex or more efficient enzymes prior to the development of sophisticated nucleic acid recombination systems. Protein splicing *in trans* could also maximize domain usage by combining one domain (fused to an intein or partial intein) with one of many domains which were also fused to a partial or complete intein. Once nucleic acid recombination systems developed, similarity amongst intein gene sequences could provide islands of homology for domain shuffling by traditional recombination pathways.

ORGANIZATION OF THE DATABASE

Since few textbooks cover protein splicing, the InBase home page provides an introduction to protein splicing and each section contains background material for the general reader. With a simple click you can explore:

1. The mechanism of protein splicing
   A. The splicing pathway

*Tel: +1 978 927 5054; Fax: +1 978 921 1350; Email: perler@neb.com

B. Similarity to the hedgehog protein family autoprocessing domains
C. Intein 3-D structure
2. The intein registry
A. Inteins listed alphabetically by genus/species
B. Intein alleles grouped by extein insertion site
C. Selected intein characteristics
3. Intein motifs
A. Splicing Motifs (Blocks A, B, F, G)
B. LAGLIDADG (DOD) homing endonuclease motifs
4. Do you have an intein?
5. On-line submission of intein data
6. The intein bibliography
7. Intein links

The INTEIN REGISTRY (Section 2A) lists all known inteins, sorted by host organism. Clicking on any intein name displays individual intein records containing: intein name, prototype intein (by convention in the field, the prototype intein is the first intein found at that insertion site in a protein), extein gene, intein class (experimental or theoretical), organism, domain of life, endonuclease activity or motifs, size, location in extein (position and surrounding extein sequences), insertion site (extein motifs and insertion site name), accession number, contributors (with contact information), comments and references specific to that intein. Section 2B tabulates inteins by insertion site. The 'Selected Intein Characteristics' section (2C) provides a capsule view of intein size, splice junction sequence, endonuclease information and extein location.

Section 3, 'Intein Motifs', tabulates both protein splicing and LAGLIDADG (DOD) homing endonuclease motifs. Unusual residues in the splicing motifs are highlighted in red. Section 4, 'Do You Have An Intein?', describes (i) the criteria for intein identification and (ii) intein motifs, including polymorphisms. Intein data is submitted publicly or confidentially using the on-line Submission Form in Section 5. Researchers should be aware that once an intein sequence is available, it will be submitted by researchers who routinely search databases. The 'Bibliography' section includes research papers, reviews, related papers and papers in press. Clickable PubMed identification numbers allow the reader to retrieve abstracts from the National Library of Medicine (NCBI). Papers focusing on intein applications are highlighted in the bibliography.

DATABASE AVAILABILITY AND CITATION

InBase can be found by clicking the Technical Resource button on the New England Biolabs Web site Home Page (http://www.neb.com and http://www.uk.neb.com) or directly at http://www.neb.com/neb/inteins.html . Users of InBase are requested to cite this article when referencing the database.

ACKNOWLEDGEMENTS

I am grateful to Ellen M. Lambrinos and Ching Lin for help in maintaining InBase and to all the intein workers who have submitted their published and unpublished data.

REFERENCES

1. Noren,C.J., Wang,J. and Perler,F.B. (2000) *Angewandte Chemie Ed. Engl.*, **39**, in press.
2. Belfort,M. and Roberts,R.J. (1997) *Nucleic Acids Res.*, **25**, 3379–3388.
3. Perler,F.B. (1998) *Cell*, **92**, 1–4.
4. Hall,T.M., Porter,J.A., Young,K.E., Koonin,E.V., Beachy,P.A. and Leahy,D.J. (1997) *Cell*, **91**, 85–97.
5. Perler,F.B. (1999) *Nucleic Acids Res.*, **27**, 346–347.
6. Klabunde,T., Sharma,S., Telenti,A., Jacobs,W.R.,Jr and Sacchettini,J.C. (1998) *Nature Struct. Biol.*, **5**, 31–36.
7. Belfort,M. and Perlman,P.S. (1995) *J. Biol. Chem.*, **270**, 30237–30240.
8. Gimble,F.S. and Thorner,J. (1992) *Nature*, **357**, 301–306.
9. Wu,H., Hu,Z. and Liu,X.Q. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 9226–9231.
10. Gorbalenya,A.E. (1998) *Nucleic Acids Res.*, **26**, 1741–1748.
11. Doyle,D.F. and Corey,D.R. (1998) *Chem. Biol.*, **5**, R157–R160.
12. Gimble,F.S. (1998) *Chem. Biol.*, **5**, R251–R256.
13. Holford,M. and Muir,T.W. (1998) *Structure*, **6**, 951–956.
14. Borman,S. (1999) *Chem. Eng. News*, February 15, 1999, 65–67.
15. Xu,R., Ayers,B., Cowburn,D. and Muir,T.W. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 388–393.
16. Chong,S., Mersha,F.B., Comb,D.G., Scott,M.E., Landry,D., Vence,L.M., Perler,F.B., Benner,J., Kucera,R.B., Hirvonen,C.A., *et al.* (1997) *Gene*, **192**, 271–281.
17. Evans,T.C., Benner,J. and Xu,M.-Q. (1998) *Protein Sci.*, **7**, 2256–2264.
18. Mathys,S., Evans,T.C., Chute,I.C., Wu,H., Chong,S., Benner,J., Liu,X.Q. and Xu,M.Q. (1999) *Gene*, **231**, 1–13.
19. Southworth,M.W., Amaya,K., Evans,T.C., Xu,M. and Perler,F.B. (1999) *BioTechniques*, **27**, 110–121.
20. Evans,T.C.,Jr, Benner,J. and Xu,M.Q. (1999) *J. Biol. Chem.*, **274**, 3923–3926.
21. Wood,D.W., Wu,W., Belfort,G., Derbyshire,V. and Belfort,M. (1999) *Nature Biotechnol.*, **17**, 889–892.
22. Muir,T.W., Sondhi,D. and Cole,P.A. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 6705–6710.
23. Kinsland,C., Taylor,S.V., Kelleher,N.L., McLafferty,F.W. and Begley,T.P. (1998) *Protein Sci.*, **7**, 1839–1842.
24. Perler,F.B. (1999) *Trends Biol. Sci.*, **24**, 209–210.