346–347 *Nucleic Acids Research, 1999, Vol. 27, No. 1*

InBase, the New England Biolabs Intein Database

Francine B. Perler*

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

Received August 28, 1998; Revised September 21, 1998; Accepted October 7, 1998

ABSTRACT

Inteins are intervening sequences that splice as proteins, not RNA. InBase, the New England Biolabs Intein Database (http://www.neb.com/neb/inteins.html), is a comprehensive on-line database that includes the Intein Registry, along with detailed information about each intein and its host protein, tabulated comparisons and a comprehensive bibliography including papers in press.

INTRODUCTION

Inteins are in-frame intervening sequences that disrupt the coding region of a host gene (1). They are post-translationally excised from a protein precursor by a self-catalytic protein splicing process (reviewed in refs 2 and 3). The N- and C-terminal host protein fragments are termed N- and C-exteins, respectively (1). Protein splicing fundamentally changed our view of information processing pathways since intervening sequences could now be removed at the protein level as well as the RNA level and more than one protein could be encoded by a single gene since the excised intein is itself a stable protein. Mini-inteins (134–198 amino acids) represent minimal protein splicing elements, while larger inteins (360–548 amino acids) have a central domain with signature motifs of 'homing endonucleases' (4,5). Homing endonucleases were first described in mobile introns; they do not cut the genomes in which they are found, but instead cleave the intein or intron insertion site (home) in homologs of the host gene lacking the intervening sequence (4). This double strand break initiates a gene conversion event resulting in insertion of the intein or intron gene into the same position in a homologous extein gene (6,7).

InBase, the New England Biolabs Intein Database, is a comprehensive on-line database (http://www.neb.com/neb/inteins.html) established in 1997 to serve as a resource for both researchers and educators interested in protein splicing. InBase compiles information about inteins and is the home of the INTEIN REGISTRY which lists all known putative inteins (1,8). Detailed information on inteins is obtained from researchers prior to publication and from all available public sources. Direct submission of data from authors and genome sequencing groups is encouraged and a submission form is available. Several subsets of data are tabulated for easy comparison including splicing or homing endonuclease motifs, insertion site sequences, selected properties and intein alleles. InBase is a fluid database providing up to date information on protein splicing. Comments and suggestions are gladly accepted.

ORGANIZATION OF THE DATABASE

InBase is divided into seven sections which are listed on the InBase home page. Each section, including the home page, contains background material for the general reader. With a simple click you can explore the following subjects.

1. The mechanism of protein splicing
   A. The splicing pathway
   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. Submitting intein data
6. The intein bibliography
7. Intein links

The InBase home page describes intein landmarks, including conserved motifs, residues known to be involved in catalysis, and domain structure. Inteins have four conserved motifs in the splicing domain and four conserved motifs found in LAGLIDADG family homing endonuclease domain. Section 1 discusses the chemistry of protein splicing and intein structure. It includes a comparison to Hedgehog protein autoprocessing domains which mediate similar chemical reactions and have the same structural fold as intein splicing domains (9).

The Intein Registry (Section 2A) lists all known inteins and their properties. 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 comments (extein motif, active site, etc.), accession number, contributors and discoverers (with contact information), comments and references. Some inteins are more widely distributed than others—alleles are found in the same insertion site in the same host gene

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

in several different organisms, sometimes spanning all three Domains of Life (section 2B). The Selected Intein Characteristics section (2C) provides a capsule view of inteins tabulating size, extein and intein splice junction sequences, endonuclease information and location in the extein.

Section 3, Intein Motifs, tabulates both protein splicing and LAGLIDADG (DOD) homing endonuclease motifs. Section 4, 'Do You Have An Intein?', describes the consensus sequence of each conserved motif and the criteria for intein identification. Most intein information is submitted by the discoverer, using the Intein Submission Form described in Section 5 or obtained from the curator by email (perler@neb.com). The Intein Registry Curator adds new information to intein records and attempts to fill in all fields if not provided by the contributor. Inteins can be submitted confidentially until publication. Researchers should be aware that once an intein sequence is available in a public database, it may be identified and submitted by anyone searching databases for inteins. We therefore urge the initial discoverer to submit his or her entry as soon as possible. The 'Comments' field provides a place for researchers to briefly point out unusual properties of that intein. Updates are acknowledged in the 'Reference' and 'Comments' fields. The bibliography section (Section 6) includes original research papers, reviews and related papers. Listings of papers in press are encouraged. Clickable PubMed identification numbers allow the reader to retrieve abstracts from the National Library of Medicine (NCBI). Finally, Section 7, Intein Links, lists links to other World Wide Web sites containing intein information, including genome sequencing sites. Although applications of inteins such as the IMPACT purification system (10), splicing *in trans* (11), or Expressed Protein Ligation (12,13) are not presently described in InBase, they are included by reference in the bibliography and/or Intein Links section.

DATABASE AVAILABILITY AND CITATION

InBase can be found by clicking the Technical Support 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, especially Shmuel Pietrokovski and Paul Liu.

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 Perler,F.B., Xu,M.-Q. and Paulus,H. (1997) *Curr. Opin. Chem. Biol.*, **1**, 292–299.
3 Shao,Y. and Kent,S.B.H. (1997) *Chem. Biol.*, **4**, 187–194.
4 Belfort,M. and Roberts,R.J. (1997) *Nucleic Acids Res.*, **25**, 3379–3388.
5 Perler,F.B. (1998) *Cell*, **92**, 1–4.
6 Belfort,M. and Perlman,P.S. (1995) *J. Biol. Chem.*, **270**, 30237–30240.
7 Gimble,F.S. and Thorner,J. (1992) *Nature*, **357**, 301–306.
8 Perler,F.B., Olsen,G.J. and Adam,E. (1997) *Nucleic Acids Res.*, **25**, 1087–1093.
9 Hall,T.M., Porter,J.A., Young,K.E., Koonin,E.V., Beachy,P.A. and Leahy,D.J. (1997) *Cell*, **91**, 85–97.
10 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., Pelletier,J.J., Paulus,H. and Xu,M.Q. (1997) *Gene*, **192**, 271–281.
11 Southworth,M.W., Adam,E., Panne,D., Byer,R., Kautz,R. and Perler,F.B. (1998) *EMBO J.*, **17**, 918–926.
12 Muir,T.W., Sondhi,D. and Cole,P.A. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 6705–6710.
13 Evans,T.C., Benner,J. and Xu,M.-Q. (1998) *Protein Sci.*, **7**, in press.