7 Zollman, S. *et al.* (1994) The BTB domain, found primarily in zinc finger proteins, defines an evolutionarily conserved family that includes several developmentally regulated genes in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 91, 10717–10721
8 Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680
9 Rost, B. *et al.* (1994) PHD an automatic mail server for protein secondary structure prediction *CABIOS* 10, 53–60
10 Hershko, A. and Ciechanover, A. (1998) The ubiquitin system. *Annu. Rev. Biochem.* 67, 425–479
11 Ciechanover, A. (1998) The ubiquitin–proteasome pathway: on protein death and cell life. *EMBO J.* 17, 7151–7160
12 Bardwell, V.J. and Treisman, R. (1994) The POZ domain: a conserved protein–protein interaction motif. *Genes Dev.* 15, 1664–1677
13 Schultz, J. *et al.* (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* 28, 231–234

Tobias Doerks*
Peer Bork
EMBL, 69012 Heidelberg, Meyerhofstr. 1, and Max-Delbrueck-Centrum, Berlin, Germany.
*e-mail: doerks@embl.heidelberg.de

Saskia Huber
Erich Buchner
Lehrstuhl für Genetik und Neurobiologie, Biozentrum der Universität, Am Hubland, 97074 Würzburg, Germany.

Sec61β – a component of the archaeal protein secretory system

Lisa N. Kinch, Milton H. Saier, Jr and Nick V. Grishin

Sec61p/SecYEG complexes mediate protein translocation across membranes and are present in both eukaryotes and bacteria. Whereas homologues of Sec61α/SecY and Sec61γ/SecE exist in archaea, identification of the third component (Sec61β or SecG) has remained elusive. Using PSI–BLAST, the archaeal counterpart of Sec61β has been detected. With the identification of the Sec61β motif, functions for a universal family of archaeal proteins can be predicted and the archaeal translocon system can be definitively detected.

Detection and rationalization of motifs in membrane proteins are more difficult than in soluble proteins because of their biased amino acid composition that is restricted to mostly hydrophobic residues and to a limited number of available spatial structures. The single transmembrane-spanning protein Sec61β (Sbh1p in yeast) interacts with two other integral membrane proteins (Sec61α and Sec61γ) to form the core of the eukaryotic protein translocation machinery (reviewed in Refs [1,2]). The bacterial counterpart of this machine consists of a similar complex (SecYEG), with SecY and SecE representing homologues of Sec61α and Sec61γ, respectively [3]. The third bacterial membrane protein, SecG, differs somewhat from Sec61β in both the number of membrane-spanning regions and residue conservation. This divergence brings into question the evolutionary origins of this third subunit, although SecG and Sec61β both function to stimulate protein translocation activities and are thought to be homologous [4,5]. Although archaeal homologues of SecY/Sec61α and SecE/Sec61γ exist, the identification of an archaeal homologue to either Sec61β or SecG has remained elusive. The Sec61p/SecYEG system is universally present in all eukaryotes and bacteria for which completely sequenced genomes are available (T. Cao and M.H. Saier, Jr, unpublished). Thus, the absence of a Sec61β/SecG homologue in archaea is puzzling. Based on sequence analyses, we have identified the third component of the archaeal translocation machinery. The archaeal counterpart resembles eukaryotic Sec61β, suggesting an overall functional similarity between the translocation apparatus of archaea and the eukaryotes. Although this functional similarity awaits experimental conformation, it mimics similarities displayed in other universal processes such as DNA replication, transcription and translation [6], and provides additional data for the studies of archaeal evolutionary origin [7].

We first detected a possible archaeal counterpart (gi | 15920503) to the human Sec61β sequence (gi | 5803165) using PSI–BLAST [8] (parameters described in Fig. 1). Upon searching protein databases for related archaeal sequences, we found hits in all but two of the completely sequenced archaeal genomes. Searches against the nucleotide databases of these genomes suggested that these sequences (AE000914 and AE006662) were missed in gene prediction efforts. To substantiate the link to eukaryotic Sec61β sequences, we generated a position-specific scoring matrix with a multiple sequence alignment of the archaeal Sec61β. Using this matrix, we initiated PSI–BLAST searches with each sequence from the alignment as a query. Two archaeal sequences used as queries (gi | 15920503 or RAP00437) identified the eukaryotic Sec61β sequence (gi | 15239337) with significant statistics (E-value 0.002). This E-value, representing the estimated number of alignments with scores no less than that of a given alignment that is expected to occur in a database search by chance [8], falls below the threshold observed for distant homologues (E = 0.01) [9].

The short 45-residue motif identifying Sec61β consists of a single, mostly hydrophobic stretch of ~20 amino acids preceded by a region of similar size that starts with several small amino acids and displays a particular residue conservation pattern (Fig. 1). The hydrophobic segment is predicted to form a transmembrane helix, with the C terminus of the helix defined by small and positively charged residues. The sequence of the helix incorporates a small residue at the beginning of the third turn and a relatively conserved histidine in the last turn. We suggest that the most conserved residue in the motif (proline) forms part of the N-terminal cap structure of this helix. The sequence between this helix and the stretch of small residues at the N terminus of the motif is characterized by four predominantly charged positions, having two negative charges surrounded by positive charges on either side. The archaeal sequences additionally contain conserved positively charged residues N-terminal to the transmembrane helix,

  PII: S0968-0004(01)02055-2

```
PHDHtm                    ------------------------------------MMMMMMMMMMMMMMMMMM------
16082665  T.acid     10 ----MASDRKSEGFQSGAGLIRYFEEEEIKGPALDPKLVVYMGIAVAIIVEIAKIFWPP---  (55)
13541150  T.volc     10 ----MASDKKSEGFQSGAGLIRYFEEEEIKGPALDPKLVVYIGIAVAIMVELAKIFWPP---  (55)
RFAC01077 F.acid     13 -MTSMAKDNQNENFQSGAGLIRYFNEEEIKGPAIDPKLIIYIGIAMGVIVELAKVFWPV---  (58)
15791336  H.NRC1     10 ----MSSGQNSGGLMSSAGLVRYFDSEDSNALQIDPRSVVAVGAFFGLVVLLAQFFA-----  (53)
RAG22196  A.fulg     14 MAKAPKGKAKTPPLMSSAGIMRYFEE-EKTQIKVSPKTILAAGIVTGVLIIILNAYYGLWP-  (68)
RPO01000  P.abys      9 -----MAKEKTTLPPTGAGLMRFFDE-DTRAIKITPKGAVALTLILIIFEIILHVVGPRIFG  (56)
RPH01741  P.hori      9 -----MAKEKTTLPPTGAGLMRFFDE-DTRAIKITPKGAIALVLILIIFEILLHVVGPRIFG  (56)
AE000914  M.ther     10 ----MAKKDKKTLPPSGAGLVRYFEE-ETKGPKLTPEQVVVMSIILAVFCLVLRFSG-----  (52)
RMJ09857  M.jann      9 -----MSKREETGLATSAGLIRYMDE-TFSKIRVKPEHVIGVTVAFVIIEAILTYGRFL---  (53)
15920503  S.toko     13 -MPSSKKKKETVPLASMAGLIRYYEE-ENEKIKISPKLLIIISIIMVAGVIVASILIPPP--  (58)
AE006662  S.solf     11 -MPSSKKKKETVPVMSMAGLIRYYEE-ENEKVKISPKIVIGASLALTIIVIVITKLF-----  (55)
RPK02491  P.aero     12 --MARRRKYEGLNPFVAAGLIKFSEEGELEKIKLTPRAAVVISLAIIGLLIAINLLLPPL--  (58)
RAP00437  A.pern     13 -MSVRRRRERRATPVTAAGLLSFYEE-YEGKIKISPTIVVGAAILVSAVVAAAHIFLPAVP-  (59)

5803165   H.sapi     49 ------------SAGTGGMWRFYTE-DSPGLKVGPVPVLVMSLLFIASVFMLHIWGKYTRS  (96)
13324684  M.musc     49 ------------SAGTGGMWRFYTE-DSPGLKVGPVPVLVMSLLFIAAVFMLHIWGKYTRS  (96)
6002114   D.mela     53 ------------GAGTGGMWRFYTD-DSPGIKVGPVPVLVMSLLFIASVFMLHIWGKYNRS (100)
14574310  C.eleg     32 ------------GGNNGGLWRFYTE-DSTGLKIGPVPVLVMSLVFIASVFVLHIWGKFTRS  (81)
10697176  Y.lipo     41 ------------GGSSSTMLKLYTD-ESQGLKVDPVVVMVLSLGFIFSVVALHILAKVSTK  (91)
6320857   S.cere     40 ------------GGSSSSILKLYTD-EANGFRVDSLVVLFLSVGFIFSVIALHLLTKFTHI  (88)
6320932   S.cere     33 ------------TNSNNSILKIYSD-EATGLRVDPLVVLFLAVGFIFSVVALHVISKVAGK  (82)
7492871   S.pomb     51 ------------PGSTPTLLKLYTD-EASGFKVDPVVVMVLSVGFIASVFLLHIVARILKK (102)
15239337  A.thal     56 ------------GSGSSNMLRFYTD-DAPGLKISPTVVLIMSLCFIGFVTALHVFGKLYLH (109)
15225401  A.thal     34 ------------GGAAGSMLQFYTD-DAPGLKISPNVVLIMSIGFIAFVAVLHVMGKLYFV  (82)
15232357  A.thal     31 ------------GGGAGTMLQFYTD-DAPGLKISPNVVLIMSIGFIAFVAVLHVMGKLYFV  (81)
15528732  O.sati     31 ------------GGGASTMLQFYTD-EAAGRKMSPNSVLIMSIGFIAVVALLHVFGKLYRT  (80)
RENC02125 E.cuni     15 ------------KKRQTKLNELLYN-QPSRLQISPMQVLVISLIFIANVFLLHILAKLVPS  (87)
6010089   E.caud      1 --------------QKAAMGFSQVS-AGKGLIVEPQYAIVISLIYIGVVIILHMIGKFRKD  (54)
PHDHtm                    -----------------------------------MMMMMMMMMMMMMMMMMM-------
```

TiBS

Fig. 1. The multiple sequence alignment of Sec61β was constructed using the program T-coffee [11] and was adjusted manually. Sequences are labelled according to NCBI gene identification number (gi), ERGO database identifier (http://wit.integratedgenomics.com/ERGO/) or GenBank identifiers (AE000914 and AE006662). Names of archaeal and eukaryotic sequences are coloured blue and red, respectively. The sequential number of the first amino acid in the motif is indicated preceding the sequence, and the sequence length is indicated in parentheses following the sequence. Complete archaeal sequences are shown. Amino acids that precede the motif and do not align with residues from the longer eukaryotic sequences (not shown) are in small italicized letters. Residues conserved across eukaryotic sequences and archaeal sequences are highlighted using different colour backgrounds: small (grey), hydrophobic (yellow), highly conserved (black), relatively conserved polar residues (green). Red letters indicate positively charged residues N- and C-terminal to the predicted transmembrane α helix; black letters indicate negatively charged residues in the predicted transmembrane α helix. The transmembrane predictions (PHDHtm) [12] produced using the PHD server (http://maple.bioc.columbia.edu/predictprotein/) with archaeal (RPO01000) and eukaryotic (gi|5803165) representative sequences as input are indicated above and below the respective alignments, with predicted membrane-spanning residues specified (M). Eukaryotic sequences were detected using PSI–BLAST (E-value cutoff 0.01, default parameters) against the non-redundant (nr) database (Nov 21, 2001; 799 241 sequences). Archaeal sequences were detected using PSI–BLAST against the nr database (E-value cutoff 0.05, BLOSUM80 matrix, default parameters), BLAST against the ERGO database (E-value cutoff 0.05), and tBLASTn against the remaining completely sequenced archaeal genome nucleotide databases. A position-specific scoring matrix (-B option in blastpgp) was generated with an alignment of all archaeal sequences and used in PSI–BLAST searches (BLOSUM80 matrix, default parameters) against the nr database to establish a link to the eukaryotic sequences. Several duplicate sequences are present in the ERGO and NCBI databases, with slight differences in the predicted start and termination sites: *A. thal* gi|15225401 and gi|13878103, *A. pern* RAP00437 and gi|14600867, *A. fulg* RAG22196 and gi|11499365, and *T. volv* gi|13541150 and gi|14324537. Of these sequence pairs, we include the sequence most closely resembling the remaining sequences. Species abbreviations: *A. fulg*, *Archaeoglobus fulgidus*; *A. pern*, *Aeropyrum pernix*; *A. thal*, *Arabidopsis thaliana*; *C. eleg*, *Caenorhabditis elegans*; *D. mela*, *Drosophila melanogaster*; *E. caud*, *Entodinium caudatum*; *E. cuni*, *Encephalitozoon cuniculi*; *F. acid*, *Ferroplasma acidarmanus*; *H.* NRC1, *Halobacterium* sp. NRC-1; *H. sapi*, *Homo sapiens*; *M. jann*, *Methanococcus jannaschii*; *M. musc*, *Mus musculus*; *M. ther*, *Methanobacterium thermoautotrophicum*; *O. sati*, *Oryza sativa*; *P. abys*, *Pyrococcus abyssii*; *P. aero*, *Pyrobaculum aerophilum*; *P. hori*, *Pyrococcus horikoshii*; *S. cere*, *Saccharomyces cerevisiae*; *S. pomb*, *Schizosaccharomyces pombe*; *S. solf*, *Sulfolobus solfataricus*; *S. toko*, *Sulfolobus tokodaii*; *T. acid*, *Thermoplasma acidophilum*; *T. volc*, *Thermoplasma volcanium*; *Y. lipo*, *Yarrowia lipolytica*.

which could help determine their orientation in the cell membrane.

A novel motif unifies a family of universal archaeal proteins of unknown function with eukaryotic Sec61β, which functions as a part of the Sec61p protein-conducting channel responsible for transport of nascent polypeptide chains through the membrane of the endoplasmic reticulum [1,2]. Additionally, the Sec61p complex acts as a membrane receptor for protein-translating ribosomes and interacts with the signal peptidase complex. Although the precise role of the Sec61 β-subunit in translocation remains unclear, the protein is thought to mediate each of these functional interactions [4,10]. We show that Sec61β homologues are present in all archaea with completely sequenced genomes, lending further support to the strong evolutionary conservation of the translocation machinery and reinforcing the universality of the process of protein export in all cellular life forms. With the identification of the Sec61β motif, we predict functions for several short universal archaeal proteins and complete the detection of the archaeal SecYEG translocon system. This family of universal archaeal proteins escaped recognition because of their small size and sequence divergence. Indeed, two of the archaeal proteins with this motif are not present in the protein databases of completely sequenced genomes as they were not recognized by 'gene finding' processes. Motif detection offers another method of protein prediction and/or annotation from coding sequences. This study illustrates the importance of careful sequence analyses both in the 'recovery' of proteins (even universal proteins) missed by automatic annotation processes and in their subsequent functional prediction.

References

1 Romisch, K. (1999) Surfing the Sec61 channel: bidirectional protein translocation across the ER membrane. *J. Cell Sci.* 112, 4185–4191

2 Matlack, K.E. *et al.* (1998) Protein translocation: tunnel vision. *Cell* 92, 381–390

3 Manting, E.H. and Driessen, A.J. (2000) *Escherichia coli* translocase: the unravelling of a molecular machine. *Mol. Microbiol.* 37, 226–238

4 Kalies, K.U. *et al.* (1998) The β subunit of the Sec61 complex facilitates cotranslational protein transport and interacts with the signal peptidase during translocation. *J. Cell Biol.* 141, 887–894

5 Knight, B.C. and High, S. (1998) Membrane integration of Sec61α: a core component of the endoplasmic reticulum translocation complex. *Biochem. J.* 331, 161–167

6 Brown, J.R. and Doolittle, W.F. (1997) Archaea and the prokaryote-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.* 61, 456–502

7 Woese, C.R. *et al.* (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* 87, 4576–4579

8 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

9 Aravind, L. and Koonin, E.V. (1999) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.* 287, 1023–1040

10 Levy, R. *et al.* (2001) *In vitro* binding of ribosomes to the β subunit of the Sec61p protein translocation complex. *J. Biol. Chem.* 276, 2340–2346

11 Notredame, C. *et al.* (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302, 205–217

12 Rost, B. *et al.* (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* 5, 1704–1718

Lisa N. Kinch
Nick V. Grishin*
Howard Hughes Medical Institute and Dept of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9050, USA.
*e-mail: grishin@chop.swmed.edu

Milton H. Saier, Jr
Division of Biology, University of California at San Diego, La Jolla, CA 92093-0116, USA.