Protein Engineering, Design & Selection vol. 22 no. 1 pp. 27–35, 2009
Published online November 6, 2008 doi:10.1093/protein/gzn064

Prediction of signal peptides in archaea

P.G. Bagos[1,2,3], K.D. Tsirigos[1], S.K. Plessas[1], T.D. Liakopoulos[1] and S.J. Hamodrakas[1]

[1]Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Athens 15701 and [2]Department of Informatics with Applications in Biomedicine, University of Central Greece, Papasiopoulou 2–4, Lamia 35100, Greece

[3]To whom correspondence should be addressed.
E-mail: pbagos@biol.uoa.gr, pbagos@ucg.gr

Computational prediction of signal peptides (SPs) and their cleavage sites is of great importance in computational biology; however, currently there is no available method capable of predicting reliably the SPs of archaea, due to the limited amount of experimentally verified proteins with SPs. We performed an extensive literature search in order to identify archaeal proteins having experimentally verified SP and managed to find 69 such proteins, the largest number ever reported. A detailed analysis of these sequences revealed some unique features of the SPs of archaea, such as the unique amino acid composition of the hydrophobic region with a higher than expected occurrence of isoleucine, and a cleavage site resembling more the sequences of gram-positives with almost equal amounts of alanine and valine at the position-3 before the cleavage site and a dominant alanine at position-1, followed in abundance by serine and glycine. Using these proteins as a training set, we trained a hidden Markov model method that predicts the presence of the SPs and their cleavage sites and also discriminates such proteins from cytoplasmic and transmembrane ones. The method performs satisfactorily, yielding a 35-fold cross-validation procedure, a sensitivity of 100% and specificity 98.41% with the Matthews' correlation coefficient being equal to 0.964. This particular method is currently the only available method for the prediction of secretory SPs in archaea, and performs consistently and significantly better compared with other available predictors that were trained on sequences of eukaryotic or bacterial origin. Searching 48 completely sequenced archaeal genomes we identified 9437 putative SPs. The method, PRED-SIGNAL, and the results are freely available for academic users at http://bioinformatics.biol.uoa.gr/PRED-SIGNAL/ and we anticipate that it will be a valuable tool for the computational analysis of archaeal genomes.

Keywords: archaea/hidden Markov model/prediction/secreted proteins/signal peptide

Introduction

In all three domains of life (bacteria, eukarya and archaea), proteins that are destined to be exported from the cytoplasm are generally (but not exclusively) synthesized as precursor proteins, bearing a cleavable N-terminal signal sequence. The signal peptide (SP) in all cases (bacteria, eukarya and archaea) is composed of a positively charged region at the n-terminus (n-region), a hydrophobic region (h-region) that spans the membrane and a c-region of mostly small and uncharged residues ending at the characteristic cleavage site (von Heijne, 1990). The SP is necessary for targeting the protein to the membrane-embedded export machinery in bacteria (Driessen and Nouwen, 2008), Eukaryotes (Rapoport *et al*., 1999) and archaea (Pohlschroder *et al*., 2005). Upon translocation across the membrane, the SP is cleaved from the precursor via a membrane-bound signal peptidase (van Roosmalen *et al*., 2004; Tuteja, 2005). The enzyme is called Spase I in bacteria and orthologues are found in archaea as well as in Eukaryotes. In Eukaryotes, proteins targeted to the organelles of bacterial origin (mitochondria and chloroplasts) also contain cleavable N-terminal targeting sequences, although they are in general very different from those found in the eukaryotic or bacterial secreted proteins (von Heijne *et al*., 1989; Habib *et al*., 2007). In addition, in bacteria (as well as in chloroplasts), another major pathway has been discovered, utilizing the twin-arginine (Tat) translocase, which recognizes longer and less hydrophobic (SPs) that carry a distinctive pattern of two consecutive arginines (R-R) in the n-region (Teter and Klionsky, 1999; Berks *et al*., 2005; Lee *et al*., 2006). A major functional differentiation between the Sec and Tat export pathways lies in the fact that the former translocates secreted proteins unfolded through a protein-conducting channel, whereas the latter, translocates completely folded proteins using an unknown mechanism (Teter and Klionsky, 1999).

In bacteria, a second signal peptidase (Spase II or Lsp) has been discovered in membrane-bound lipoproteins (Sankaran and Wu, 1995), that cleaves shorter SPs carrying a distinctive c-region containing a conserved cysteine (von Heijne, 1989). The conserved cysteine is indispensable in both gram-positive and gram-negative bacteria, and is necessary for membrane anchoring. The post-translational lipid modification involves three enzymes that act sequentially: the prolipoprotein diacylglyceryl transferase (Lgt), that transfers a diacylglyceride to the cysteine sulfydryl group, the signal peptidase II (Spase II or Lsp) that cleaves the SP at the residue before the cysteine forming an apolipoprotein and the apolipoprotein *N*-acyltransferase (Lnt), which acylates the α-amino group of the apolipoprotein N-terminal cysteine forming the mature lipoprotein (Sankaran and Wu, 1994; Sankaran *et al*., 1995). Although dozens of putative lipoproteins have been identified in archaeal genomes, the absence of Spase II orthologues in archaea as well as the different post-translational modification of cysteine, have resulted in a limited level of knowledge concerning archaeal lipoproteins and a lack of experimentally verified proteins of that type. Translocation of lipoproteins through the Tat pathway has been postulated based on sequence analysis, but only recently has been proven for the Bacterium *Desulfovibrio vulgaris* (Valente *et al*., 2007) and the Archaeon *Haloferax volcanii* (Gimenez *et al*., 2007). Interestingly, in halophilic archaea, the components of the Tat pathway are essential for viability (Dilks *et al*., 2005; Thomas and Bolhuis, 2006) and there is evidence that Tat-dependent translocation is widely used as part of a

mechanism for adaptation to extreme saline environments (Rose *et al*., 2002).

Computational prediction of secretory SPs was performed initially using weight matrices (von Heijne, 1986). However, Neural Networks (Nielsen *et al*., 1997; Nielsen *et al*., 1999) as well as hidden Markov models (HMM) (Nielsen and Krogh, 1998) introduced by the SignalP method, have been proven to be the most successful methods currently available (Menne *et al*., 2000). Recently, SignalP was retrained and, mainly due to better annotation and selection of the training set, yielded an even better accuracy (Bendtsen *et al*., 2004), whereas the program TatP has been presented offering the most accurate classification of TAT SPs (Bendtsen *et al*., 2005). A different approach has been followed in the Phobius method (Kall *et al*., 2004; Kall *et al*., 2007), where a HMM was used to predict at the same time the presence of a secretory SP and transmembrane (TM) topology of a given protein. Following this approach, the authors showed that they can minimize the number of SPs predicted as TM segments and vice versa. Concerning lipoproteins, for years, regular expression patterns were used based on the von Heijne rule (von Heijne, 1989), with various modifications (Madan Babu and Sankaran, 2002; Sutcliffe and Harrington, 2002; Madan Babu *et al*., 2006; Setubal *et al*., 2006). Recently, a method called Lipop was developed, which is based on HMMs and was trained exclusively on gram-negative bacteria lipoproteins (Juncker *et al*., 2003). However, the previously mentioned prediction methods have been trained on bacterial and/or eukaryal sequences, and in most cases there are different versions of the predictors aiming at capturing the distinct sequence features of the SPs of particular groups of organisms. Since very few experimentally verified SPs have been characterized from archaea, little is known about the precise characteristics of these sequences, even though there is some evidence suggesting that archaeal SPs exhibit a mixture of characteristics found in eukarya and bacteria. The first computational work on archaea was performed by Nielsen *et al*. (1999) when they applied SignalP on the genome of *Methanococcus jannaschii* (*M. jannaschii*). They used the three versions of SignalP (trained on gram-positive bacteria, gram-negative bacteria and eukarya), and identified 34 proteins where the predictions concerning the existence of the SP coincided. A more systematic evaluation was performed later by Bardy *et al*. (2003), which applied a similar procedure on 15 completely sequenced genomes of archaea, requiring though, that all the three methods would predict the same cleavage site. Although this procedure may be biased to select only proteins that share common features with the sequences found in other domains of life, the general conclusions of these studies suggested that archaeal SPs exhibit a more eukaryotic-like cleavage site (c-region), and a unique h-region resembling the bacterial ones, with a slight over-representation of leucine and isoleucine; leucine is by far the dominant residue in Eukaryotes. Thus, it is evident now that SP predictors trained on eukaryal or bacterial proteins cannot reliably be applied to archaeal sequences. A dedicated prediction method is needed that would be trained exclusively on archaeal SPs. The major problem in this respect is the lack of a large number of experimentally verified signal sequences of archaeal origin. In particular, the Uniprot database (Wu *et al*., 2006) lists only 12 archaeal sequences with experimentally verified, precise locations of the cleavage site, and the specialized database of SPs SPDB (Choo *et al*., 2005) lists only nine such proteins.

Materials and methods

Hidden Markov model

The HMM that we used is similar to the one proposed by SignalP (Nielsen and Krogh, 1998). It consists of three different sub-models, the SP sub-model corresponding to the secretory SPs, the N-terminal TM sub-model corresponding to the N-terminal TM segment domain, and a globular sub-model used to model the globular N-terminal domains of cytoplasmic or membrane proteins. The central core of the model is the SP sub-model (Fig. 1). It is used to capture the modular nature of SPs, modeling the positively charged n-region, the hydrophobic h-region that spans the membrane and the c-region of mostly small and uncharged residues ending at the characteristic cleavage site (A-X-A) (von Heijne, 1990). The TM sub-model, is identical to the one used by the HMM-TM predictor for alpha-helical membrane proteins (Bagos *et al*., 2006), whereas the globular sub-model consists simply of a self-transitioning state.

The model was trained using the Baum–Welch algorithm for labeled sequences (Krogh, 1994) and the decoding was performed using the standard Viterbi algorithm (Durbin *et al*., 1998), although more advanced techniques such as the Posterior-Viterbi decoding (Fariselli *et al*., 2005) and the Optimal Accuracy Posterior Decoder (Kall *et al*., 2005) yield nearly identical results. In addition to the Viterbi decoding which produces the optimal path of states through the model, and hence predicts simultaneously the type of the sequence (SP, TM or Globular) as well as the cleavage site (if any), we also report the S1 reliability index (Melen *et al*., 2003), which takes values in the range [0–1] and provides a useful measure of the reliability of the prediction. Given that the

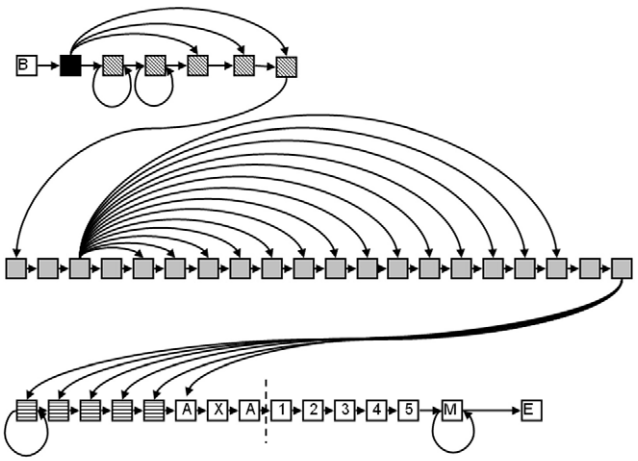

Fig. 1. Architecture of the HMM used to model the secretory SP sequences. Each line (top to bottom) corresponds to the n-, h- and c-region, respectively. States in the n- and h-region that share the same emission probabilities (amino acid frequencies) are depicted using the same symbol. The cleavage site is shown using a dashed vertical line between A and 1 (first amino acid of the mature protein). Allowed transitions are depicted with arrows. B and E correspond to the Begin and End states, respectively, whereas states after the cleavage site (1–5 and M) are used to model the first residues of the mature protein.

majority of the SPs used (discussed later) did not contain information concerning the precise cleavage site location, an 'imputation' or 're-labeling' method had to be used. Although the location of the cleavage site in proteins with non-verified cleavage sites could be predicted by other means, we chose to train an initial model using the verified proteins, and afterwards to apply the method on the non-verified ones, performing a constrained prediction by removing the labels in the area of the cleavage site (c-region) as described earlier (Krogh *et al*., 2001; Bagos *et al*., 2006).

Data sets

As we noted earlier, the publicly available databases, such as Uniprot (Wu *et al*., 2006) and SPDB (Choo *et al*., 2005), currently contain annotated information for only a few archaeal sequences with experimentally verified precise locations of the cleavage site. Thus, we decided to perform an extensive literature search in order to identify archaeal sequences with either verified cleavage site locations, or proteins with verified SPs whose cleavage sites are not precisely known. The literature search was performed on Pubmed using terms such as 'SP' or 'signal sequence', combined with terms such as 'archaeon', 'archaea' or 'archaebacteria'. Since this strategy yielded also a limited number of archaeal peptides, and given that in many known cases the information concerning the presence of the SP was not available in the abstract or the title of the respective papers, we used additional search terms such as 'extracellular', 'extracytoplasmic' or 'secreted'. The full-text of the papers were downloaded and read, and the reference lists were also checked in order to identify additional studies that were missed by the initial search. The identified sequences in almost every case were retrieved from Uniprot (Wu *et al*., 2006), and were classified according to two criteria; the first is whether the protein has a verified SP cleavage site or not, and the second is whether the protein is translocated using the Tat or the Sec system. Lipoprotein SPs were removed since there are only few such examples (see Results and discussion).

Since the model is also capable of discriminating SPs from globular proteins as well as from proteins with an N-terminal TM helix, we used as negative examples 69 archaeal proteins with an annotated (proven or putative) TM segment within the first 70 amino acids having the N-terminus located in the cytoplasmic space, and 183 archaeal cytoplasmic proteins. The sequences were retrieved from Uniprot and identical sequences were removed to produce a unique set. The training and testing procedure was performed using a 35-fold cross-validation procedure. The training set was split in 35 parts having approximately the same number of SPs, TM and cytoplasmic proteins. The training procedure consisted of removing one of the 35 subsets from the training set, training the model with the remaining proteins and performing the test on the proteins of the set that was removed. This process was repeated in tandem for all the subsets in the training set, and the final prediction accuracy summarized the outcome of all independent tests. Sequences belonging to different subsets used for cross-validation not had >18 identical residues within the SP as advised by previous studies (Nielsen *et al*., 1997; Nielsen *et al*., 1999). Finally, the complete proteomes of archaea were downloaded from the NCBI ftp site at ftp://ftp.ncbi.nih.gov/.

For measures of accuracy in the binary classification problem (signal peptides versus non-SPs), we used the percentage of correctly classified positive examples (sensitivity), the percentage of correctly classified negative examples (specificity) and the Matthews' correlation coefficient (MCC) that summarizes in a single measure true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) (Baldi *et al*., 2000).

Results and discussion

The extensive literature search that we performed identified in total 69 archaeal proteins with a verified SP (Table I). Among them, 24 proteins have cleavage sites that were defined precisely by direct sequencing of the N-terminus of the mature protein. The 69 proteins listed in Table I include many extracellular secreted enzymes (proteases, chitinases, amylases, etc), several surface (S-layer) proteins, a few extracellular components of ABC transporter systems, as well as some uncharacterized proteins from the two main kingdoms of archaea (*Crenarchaeota* and *Euryarchaeota*). A few sequences were discarded since they were identical in the SP sequence with others in the set (i.e. CSG_METSC which is identical to CSG_METFE and Q7LYT7_PYRWO which is identical to O08452_PYRFU) as well as one sequence (Q97X08_SULSO) for which there was evidence suggesting that it was membrane-anchored (Ferrer *et al*., 2005). Only two couples of sequences had >18 identical residues in a BLAST alignment (CSG_METJA with Q6M088_METMP and HLY_HAL17 with Q5RLZ1_NATMA) though having different cleavage sites. Thus, we decided to keep them in the training set and include them in the same subset used for cross-validation in order to be tested simultaneously (to avoid overfitting). A number of proteins with a lipoprotein SP that was either proven (Gimenez *et al*., 2007) or putative (Mattar *et al*., 1994) were also discarded. We did not try specifically to eliminate Tat SPs (the same was done in SignalP), and in total 18 such sequences are included in the set, of which four contained a verified cleavage site.

The alignment of the SPs at their respective cleavage sites (Fig. 2) is useful in order to obtain insight into the unique sequence features of the archaeal SPs. The sequence logos (Schneider and Stephens, 1990; Crooks *et al*., 2004) in Fig. 2 reveal the similarities and differences between the experimentally verified SPs of archaea, Eukaryotes, gram-positive and gram-negative bacteria [data for Eukaryotes and bacteria were taken from the set of SignalP (Nielsen *et al*., 1997)]. We can see that at position-1 (just before the cleavage site), alanine (A) is the dominant amino acid, although glycine (G) and serine (S) are also present in significant proportions. Alanine is also the dominant amino acid in all organism groups, though in Eukaryotes other amino acids are more easily tolerated compared with bacteria. At position-3, alanine is also the dominant amino acid, however, valine (V) is also almost equally represented in archaea followed by serine, isoleucine (I) and threonine (T). Taken together, these features suggest that the archaeal cleavage site resembles more closely that of gram-positive bacteria signals, although some resemblance to the eukaryal ones is visible. In the h-region of archaeal SPs, alanine, leucine and isoleucine are almost equally abundant whereas valine is less frequent, a feature that is unique to the archaeal domain. In eukaryal

Table I. Data set of 69 experimentally verified SPs identified in this study[a]

Uniprot ID (Wu *et al.*, 2006)	Organism	Sec/ Tat	Cleavage site (Ref.)	Function
CAH_METTE	*Methanosarcina thermophila*	Sec	Verified (Alber and Ferry, 1994)	Carbonic anhydrase
CSG_HALJP	*Haloarcula japonica*	Sec	Verified (Wakai *et al.*, 1997)	S-layer protein
CSG_HALSA	*Halobacterium salinarium* (*H. salinarium*)	Sec	Verified (Lechner and Sumper, 1987)	S-layer protein
CSG_HALVO	*Halobacterium volcanii* (*H. volcanii*)	Sec	Verified (Sumper *et al.*, 1990)	S-layer protein
CSG_METFE	*Methanothermus fervidus*	Sec	Verified (Brockl *et al.*, 1991)	S-layer protein
CSG_METJA	*Methanocaldococcus jannaschii* [*Methanococcus jannaschii* (*M. jannaschii*)]	Sec	Verified (Akca *et al.*, 2002)	S-layer protein
CSG_METVO	*Methanococcus voltae*	Sec	Verified (Dharmavaram *et al.*, 1991)	S-layer protein
HAH4_HALME	*Halobacterium mediterranei* (*H. mediterranei*)	Tat	Verified (Cheung *et al.*, 1997)	Halocin-H4
HMEA_ARCFU	*Archaeoglobus fulgidus*	Tat	Verified (Mander *et al.*, 2002)	Hdr-like menaquinol oxidoreductase iron-sulfur subunit 1
Q12VE2_METBU	*Methanococcoides burtonii* (*M. burtonii*)	Sec	Verified (Saunders *et al.*, 2006)	S-layer-related protein
Q12UJ4_METBU	*M. burtonii*	Sec	Verified (Saunders *et al.*, 2006)	Ig-like protein
Q12WA9_METBU	*M. burtonii*	Sec	Verified (Saunders *et al.*, 2006)	Uncharacterized protein
Q12WY2_METBU	*M. burtonii*	Sec	Verified (Saunders *et al.*, 2006)	Uncharacterized protein
Q12WZ0_METBU	*M. burtonii*	Sec	Verified (Saunders *et al.*, 2006)	Uncharacterized protein
Q12UD6_METBU	*M. burtonii*	Sec	Verified (Saunders *et al.*, 2006)	Uncharacterized protein
Q12X64_METBU	*M. burtonii*	Sec	Verified (Saunders *et al.*, 2006)	Uncharacterized protein
Q980C6_SULSO	*Sulfolobus solfataricus* (*S. solfataricus*)	Sec	Verified (Albers and Driessen, 2002)	Uncharacterized protein
Q97UG7_SULSO	*S. solfataricus*	Sec	Verified (Albers and Driessen, 2002)	ABC transporter component
Q97VF7_SULSO	*S. solfataricus*	Sec	Verified (Albers and Driessen, 2002)	ABC transporter component
Q97UH5_SULSO	*S. solfataricus*	Sec	Verified (Albers and Driessen, 2002)	ABC transporter component
Q60224_9EURY	*Natronococcus* sp	Tat	Verified (Pohlschroder *et al.*, 2005)	Alpha-amylase
Q6M088_METMP	*Methanococcus maripaludis*	Sec	Verified (Pohlschroder *et al.*, 2005)	S-layer protein
Q9YBL5_AERPE	*Aeropyrum pernix* (*A. pernix*)	Sec	Verified (Palmieri *et al.*, 2006)	ABC transporter component
Q97V37_SULSO	*S. solfataricus*	Tat	Verified (Pohlschroder *et al.*, 2005)	Oxydoreductase
Q97VS7_SULSO	*S. solfataricus*	Sec	Non-verified (Limauro *et al.*, 2001)	Endo-1,4-beta-glucanase
Y958_METJA	*M. jannaschii*	Sec	Non-verified (Bult *et al.*, 1996)	Uncharacterized protein
THPS_SULAC	*Sulfolobus acidocaldarius*	Sec	Non-verified (Lin and Tang, 1990)	Thermopsin
HLY_HAL17	*Halophilic archaebacteria* (strain 172p1)	Tat	Non-verified (Kamekura *et al.*, 1992)	Halolysin
TKSU_PYRKO	*Pyrococcus kodakaraensis* (*P. kodakaraensis*)	Sec	Non-verified (Kannan *et al.*, 2001)	Tk-subtilisin
PLS_PYRFU	*Pyrococcus furiosus* (*P. furiosus*)	Sec	Non-verified (Voorhorst *et al.*, 1996)	Pyrolysin
Y1033_SULSO	*S. solfataricus*	Sec	Non-verified (She *et al.*, 2001)	Kelch domain-containing protein
Y1435_PYRAB	*Pyrococcus abyssi*	Sec	Non-verified (Cohen *et al.*, 2003)	Uncharacterized protein
Y614_PYRHO	*Pyrococcus horikoshii* (*P. horikoshii*)	Sec	Non-verified (Kawarabayasi *et al.*, 1998)	Uncharacterized protein
Y939_SULTO	*Sulfolobus tokodaii*	Sec	Non-verified (Kawarabayasi *et al.*, 2001)	Kelch domain-containing protein
Contig 3108	*H. volcanii*	Tat	Non-verified (Gimenez *et al.*, 2007)	Exo-arabinanase
Contig 3156	*H. volcanii*	Tat	Non-verified (Gimenez *et al.*, 2007)	Pectate lyase
Contig 3082	*H. volcanii*	Tat	Non-verified (Gimenez *et al.*, 2007)	Halocyanin 2
Contig 2996	*H. volcanii*	Tat	Non-verified (Gimenez *et al.*, 2007)	Halocyanin 3
Q2TME8_HALSA	*H. salinarium*	Tat	Non-verified (Shi *et al.*, 2006)	SptA protease
Q4A3E0_HALHI	*Haloarcula hispanica*	Tat	Non-verified (Hutcheon *et al.*, 2005)	Alpha-amylase
Q6JSL9_HALAS	*Halobacterium* sp (strain AS7092)	Tat	Non-verified (Sun *et al.*, 2005)	Halocin C8
Q5RLZ1_NATMA	*Natrialba magadii*	Tat	Non-verified (Ruiz and De Castro, 2007)	Halolysin-like extracellular serine protease
O08452_PYRFU	*P. furiosus*	Sec	Non-verified (Wang *et al.*, 2007)	alpha-amylase
Q9HQ20_HALSA	*H. salinarium*	Sec	Non-verified (Woodson *et al.*, 2005)	ABC transporter component
Q9YFI3_AERPE	*A. pernix*	Sec	Non-verified (Catara *et al.*, 2003)	Pernisine
Q9UWN2_9EURY	*Thermococcus* sp B1001	Sec	Non-verified (Hashimoto *et al.*, 2001)	Cyclodextrin glucanotransferase
Q9UWR7_PYRKO	*P. kodakaraensis*	Sec	Non-verified (Tanaka *et al.*, 1999)	Chitinase
Q9Y9Y8_AERPE	*A. pernix*	Sec	Non-verified (Sako *et al.*, 1997)	Serine protease
O93635_THESU	*Thermococcus stetteri*	Sec	Non-verified (Voorhorst *et al.*, 1997)	Stetterlysin
Q48929_METBR	*Methanobacterium bryantii*	Sec	Non-verified (Kim *et al.*, 1995)	Copper response extracellular protein
Q5V573_HALMA	*Haloarcula marismortui*	Tat	Non-verified (Goldman *et al.*, 1990)	Alkaline phosphatase D
Q9HHB0_9CREN	*Desulfurococcus mucosus*	Sec	Non-verified (Duffner *et al.*, 2000)	Pullulanase
O58925_PYRHO	*P. horikoshii*	Sec	Non-verified (Kashima *et al.*, 2005)	Endo-1,4-beta-glucanase
P71402_HALME	*H. mediterranei*	Tat	Non-verified (Kamekura *et al.*, 1996)	Serine protease halolysin R4
Q97VC2_SULSO	*S. solfataricus*	Sec	Non-verified (Chong and Wright, 2005)	Uncharacterized protein
Q97UF5_SULSO	*S. solfataricus*	Tat	Non-verified (Chong and Wright, 2005)	ABC transporter component
Q9HSH6_HALSA	*H. salinarium*	Tat	Non-verified (Izotova *et al.*, 1983)	Serine protease

Continued

Table I. *Continued*

Uniprot ID (Wu *et al*., 2006)	Organism	Sec/ Tat	Cleavage site (Ref.)	Function
Q5JGP8_PYRKO	*P. kodakaraensis*	Sec	Non-verified (Morikawa *et al*., 1994)	Thiol protease
Q9V2T0_PYRFU	*P. furiosus*	Sec	Non-verified (Bauer *et al*., 1999)	Endoglucanase A
Q8NKS8_THELI	*Thermococcus litoralis*	Sec	Non-verified (Brown and Kelly, 1993)	Amylopullulanase
Q3HUR3_PYRFU	*P. furiosus DSM 3638*	Sec	Non-verified (Brown and Kelly, 1993)	Amylopullulanase
Q8U0C9_PYRFU	*P. furiosus*	Sec	Non-verified (Comfort *et al*., 2008)	Alkaline serine protease
Q8U1U6_PYRFU	*P. furiosus*	Sec	Non-verified (Comfort *et al*., 2008)	Starch-binding protein
Q6L252_PICTO	*Picrophilus torridus*	Sec	Non-verified (Serour and Antranikian, 2002)	Glucoamylase
Q53I75_HALME	*H. mediterranei*	Tat	Non-verified (Perez-Pomares *et al*., 2003)	Putative alpha-amylase
O50200_9EURY	*Thermococcus* sp Rt3	Sec	Non-verified (Jones *et al*., 1999)	Amylase
Q9Y8I8_THEHY	*Thermococcus hydrothermalis* (*T. hydrothermalis*)	Sec	Non-verified (Erra-Pujada *et al*., 1999)	Pullulanase
Q2QC88_9EURY	*Thermococcus onnurineus*	Sec	Non-verified (Lim *et al*., 2007)	Alpha-amylase
O93647_THEHY	*T. hydrothermalis*	Sec	Non-verified (Leveque *et al*., 2000)	Alpha-amylase

[a]We listed the Uniprot ID, the organism, the translocation pathway (Sec/Tat) and the status of the cleavage site, along with the respective reference and the protein's function.

SPs, leucine is clearly the dominant amino acid (followed by equal amounts of alanine and valine) whereas in bacteria alanine and leucine are almost equally present. In both cases isoleucine is under-represented, in contrast with what is seen in archaea. Furthermore, the c-region contains mostly small and uncharged residues (serine, glycine, threonine and proline), whereas in the n-region Lysine is slightly more frequent than arginine despite the presence of 18 Tat SPs in the training set. Some of these observations were touched on in earlier works (Nielsen *et al*., 1999; Bardy *et al*., 2003). Here these patterns are analyzed for the first time based on experimentally verified archaeal SPs rather than solely on predictions. The results suggest that archaeal SPs are of unique composition, and that there is a need for a dedicated prediction method.

The results obtained in the 35-fold cross-validation procedure are listed in Table II. Our method, PRED-SIGNAL, predicts correctly all the 69 SPs and rejects correctly 248 out of the 252 cytoplasmic and TM proteins. These results correspond to 100% sensitivity and 98.41% specificity with an MCC equal to 0.964. Using the same data set, we evaluated also the various versions of the SignalP method (Nielsen *et al*., 1997; Nielsen and Krogh, 1998; Nielsen *et al*., 1999; Bendtsen *et al*., 2004), Phobius (Kall *et al*., 2004; Kall *et al*., 2007) and PrediSi (Hiller *et al*., 2004), which is another popular and accurate SP predictor based on position specific scoring matrixs (PSSMs). The method developed here clearly outperforms all the currently available top-scoring predictors. This was expected, since none of them was trained specifically to recognize archaeal SPs. In absolute numbers, the method is very accurate and is comparable with, if not better than, the currently top-scoring method SignalP. SignalP, when trained and independently tested on gram-positive bacteria, gram-negative bacteria, and Eukaryotes respectively, reports sensitivities ranging from 92 to 99%, specificities ranging from 85 to 93% and MCCs ranging from 0.87 to 0.92, when only cytoplasmic proteins are used as negative examples (Nielsen *et al*., 1997; Bendtsen *et al*., 2004). When proteins with an N-terminal TM segment are included in the test-set, the specificity drops $<$90%, as was shown in an earlier evaluation study (Menne *et al*., 2000). From Table II, it is also clear that among predictors trained on data sets of origin other than archaea, those trained on gram-positive bacteria perform better in predicting archaeal signal sequences, a fact that can be explained by the composition of the c-region in archaeal SPs discussed earlier. Of these methods, only SignalPv3-NN trained on gram-positive bacteria compares with the method that we developed, having a slightly better specificity but, nevertheless, a lower sensitivity and overall performance (MCC).

Furthermore, the results obtained by using a combination of different SP predictors (i.e. the SignalP modules trained on Eukaryotes, gram-positive and gram-negative bacteria) illustrate the difficulties of such an approach. It is clear that although such an approach increases the specificity of the selection (i.e. few FPs), the sensitivity decreases (i.e. more FNs). Thus, this strategy (which was until now the only option), reliably predicts some SPs but at the same time overlooks a large number of true SPs. Some general conclusions could also been drawn from these results, verifying previous studies. As we noted earlier, methods trained on gram-positive bacteria (SignalPv2, SignalPv3 and PrediSi) perform slightly better compared with their gram-negative counterparts and clearly better compared with the Eukaryotic-based ones. Phobius, which was trained on a mixed set of proteins (gram-positive, gram-negative and Eukaryote), performs well also, but places lower than methods trained on gram-positive bacteria as well as methods trained on gram-negative bacteria. HMM methods that were trained to discriminate N-terminal TM regions from SPs (Phobius, SignalP-HMM) perform better in terms of specificity compared with Neural Networks and PSSM methods (SignalP-NN, PrediSi). On the other hand, Neural Network-based methods (SignalP-NN) are better in predicting the precise cleavage site location (data not shown). Finally, the updated versions of SignalP (SignalPv3) perform in general better compared with the older versions (SignalPv2).

We also analyzed 48 currently available archaeal completely sequenced genomes. The combined prediction of the three HMM predictors of SignalPv3 (gram-positive, gram-

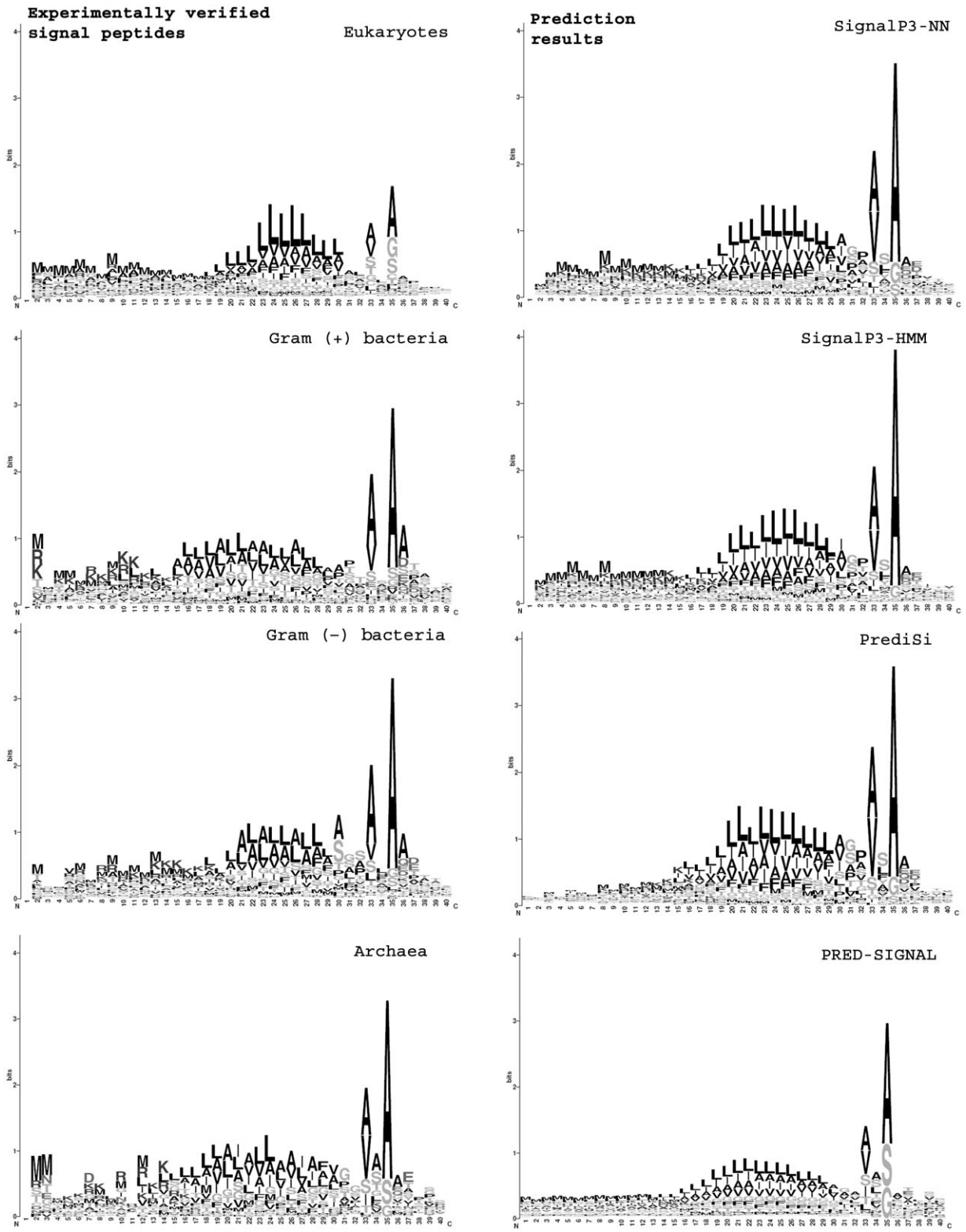

Fig. 2. Left panel (from top to bottom): the sequence logos of experimentally verified eukaryal, gram-positive, gram-negative and archaeal signal peptides (SPs), respectively, produced by WebLogo (Crooks *et al*., 2004). The experimentally verified bacterial and eukaryal SPs were retrieved from the data set of SignalP. Right panel (from top to bottom): the sequence logos of SPs found in the genome analysis of 48 archaeal genomes (see text) as predicted by SignalPv3-NN, SignalPv3-HMM, PrediSi and PRED-SIGNAL (this work), respectively. The predictions of SignalP and PrediSi correspond to proteins predicted to have the exactly the same cleavage site by different modules of the respective predictor (see text for details). Sequences are aligned to the observed or predicted cleavage site which in all cases is arbitrarily located between 35th and 36th amino acid of the alignment.

negative and Eukaryotic) produced in total 6145 proteins with a SP, of which 2306 proteins have the same predicted cleavage site for all three methods. The combination of the NN predictors of SignalPv3 yielded 5473 predictions in total of which 2037 have the same prediction for the cleavage site. On the contrary, the method developed here predicts in

Table II. Results obtained from PRED-SIGNAL using the cross-validation procedure on the set of 69 experimentally verified SPs and on 69 TM and 183 cytoplasmic archaeal proteins[a]

Method: PRED-SIGNAL	Sensitivity: 69/69 (100.00%)	Specificity (TM proteins): 67/69 (97.10%)	Specificity (cytoplasmic proteins): 181/183 (98.91%)	Specificity (Total): 248/252 (98.41%)	MCC: 0.964
SignalPv3-NN (gram+)	66/69 (95.65%)	68/69 (98.55%)	183/183 (100.00%)	251/252 (99.60%)	0.963
SignalPv3-NN (gram−)	61/69 (88.41%)	66/69 (95.65%)	183/183 (100.00%)	249/252 (98.80%)	0.897
SignalPv3-NN (Euk)	55/69 (79.71%)	55/69 (79.71%)	182/183 (99.45%)	237/252 (94.05%)	0.734
SignalPv3-NN (all)	33/69 (47.83%)	69/69 (100.00%)	183/183 (100.00%)	252/252 (100.00%)	0.647
SignalPv3-HMM (gram+)	65/69 (94.20%)	66/69 (95.65%)	183/183 (100.00%)	249/252 (98.80%)	0.935
SignalPv3-HMM (gram−)	64/69 (92.75%)	66/69 (95.65%)	180/183 (98.36%)	246/252 (97.62%)	0.899
SignalPv3-HMM (Euk)	59/69 (85.51%)	67/69 (97.10%)	182/183 (99.45%)	249/252 (98.80%)	0.877
SignalPv3-HMM (all)	29/69 (42.03%)	69/69 (100.00%)	183/183 (100.00%)	252/252 (100.00%)	0.602
SignalPv2-NN (gram+)	66/69 (95.65%)	49/69 (71.01%)	171/183 (93.44%)	220/252 (87.30%)	0.740
SignalPv2-NN (gram−)	66/69 (95.65%)	51/69 (73.91%)	176/183 (96.17%)	227/252 (90.07%)	0.781
SignalPv2-NN (Euk)	53/69 (76.81%)	56/69 (81.16%)	180/183 (98.36%)	236/252 (93.65%)	0.705
SignalPv2-NN (all)	35/69 (50.72%)	60/69 (86.96%)	182/183 (99.45%)	242/252 (96.03%)	0.553
SignalPv2-HMM (gram+)	67/69 (97.10%)	63/69 (91.30%)	182/183 (99.45%)	245/252 (97.22%)	0.920
SignalPv2-HMM (gram−)	65/69 (94.20%)	61/69 (88.41%)	180/183 (98.36%)	241/252 (95.63%)	0.861
SignalPv2-HMM (Euk)	60/69 (86.96%)	67/69 (97.10%)	182/183 (99.45%)	249/252 (98.80%)	0.887
SignalPv3-HMM (all)	29/69 (42.03%)	69/69 (100.00%)	182/183 (99.45%)	251/252 (99.60%)	0.588
PrediSi (gram+)	61/69 (88.41%)	66/69 (95.65%)	180/183 (98.36%)	246/252 (97.62%)	0.870
PrediSi (gram−)	63/69 (91.30%)	65/69 (94.20%)	180/183 (98.36%)	245/252 (97.22%)	0.881
PrediSi (Euk)	60/69 (86.96%)	52/69 (75.36%)	181/183 (98.91%)	233/252 (92.46%)	0.757
PrediSi (all)	32/69 (46.38%)	68/69 (98.55%)	182/183 (99.45%)	250/252 (99.20%)	0.558
Phobius	58/69 (84.06%)	69/69 (100.00%)	183/183 (100.00%)	252/252 (100.00%)	0.897

v2, version 2; v3, version 3; HMM, hidden Markov model; NN, Neural Network; gram+, gram-positive; gram−, gram-negative; Euk, Eukaryote; all, the combination of the three modules.

[a]For comparison we list the results obtained by the various modules of SignalP, PrediSi and Phobius. For measures of accuracy (SPs versus non-SPs), we used the percentage of correctly classified positive examples (sensitivity), the percentage of correctly classified negative examples (specificity) and the MCC (Matthews' correlation coefficient) that summarizes in a single measure TP, FP, TN and FN (Baldi *et al*., 2000).

total a much larger number of proteins with signal sequences, 9437 in all. Among these proteins, according to their annotation the largest group consisted of 5351 hypothetical proteins (56.7%), followed by 1408 (14.92%) enzymes such as lipases, hydrolases, transferases, proteases, kinases, reductases, etc, of which 127 were probable, putative or predicted. There were also 832 (8.81%) membrane proteins such as permeases, transporters, etc of which 82 were probable, putative or predicted and 1024 (10.85%) extracellular proteins (mostly solute-binding components of ABC transport systems, as well as S-layer and flagellar proteins) of which 43 were probable, putative or predicted. Finally, there were 822 proteins that could not be classified (8.71%).

The detailed results for each genome are available as Supplementary data in our web site (http://bioinformatics.biol.uoa.gr/PRED-SIGNAL/). The per-genome percentage of predicted proteins carrying a SP according to our method, ranges from 5 to 14% (average = 8.92%) whereas the same percentage according to the combination of SignalP predictors ranges from 3 to 7%. According to our results, the 15 archaeal genomes belonging to *Crenarchaeota* do not differ significantly from the 32 genomes belonging to *Euryarchaeota* (8.54 versus 9.16%, P-value = 0.406 according to t-test) concerning the proportion of proteins predicted to contain a SP. The only representative of *Nanoarchaeota* (*Nanoarchaeum equitans*) contains a comparable proportion of secreted proteins (7.09%) although produced by a significantly smaller genome (38 out of the 536 total coding sequences). In an ANOVA analysis, psychrophiles, mesophiles, thermophiles and hyperthermophiles did not show any statistical difference concerning the proportion of proteins carrying a SP (range from 8.2 to 10.7%, P-value = 0.087). Only the six thermoacidophiles showed a smaller proportion (6.58%), whereas one haloalkalophile (13.8%) and the three halophiles (12.53%) showed larger proportions. The amino acid distribution of SPs of all the groups examined using sequence logos did not detect any obvious discrepancies (data not shown). The only detectable difference was the over-representation of alanine and glycine and the under-representation of isoleucine in the h-region of SPs of halophiles and haloalkalophiles. These results need to be studied further, but clearly the large proportion of secreted proteins as well as the abundance of glycine and alanine that suggest a lower hydrophibicity in the h-region of SPs of halophiles, should be attributed to the extensive use of the Tat pathway. PRED-SIGNAL does not discriminate Tat from Sec SPs, and we expect a lot of the secreted proteins of halophiles to contain a Tat SP (Rose *et al*., 2002).

Among the proteins predicted by the combination of the HMM versions of SignalP, only 685 were not predicted by our predictor, and among the proteins predicted by the combination of the NN versions of SignalP, 749 were not predicted as having a SP by PRED-SIGNAL. Thus, the HMM method developed here is very specific in detecting putative SPs that are considered highly probable (as judged by the stringent criteria applied by the combination of the SignalP predictors). On the other hand, PRED-SIGNAL predicts an additional large number of proteins that were selected by only one or two modules of SignalP, and a remarkably large number of proteins that were not selected by either one of the versions of SignalP (1039 for the HMM versions and 1139 for the NN versions). This highlights that although the stringent criteria applied by combining the different predictors of SignalP can indeed select a large number of archaeal SPs sharing common features with bacterial and eukaryotic SPs, an additional large number of putative SPs exist that

possess some unique features not present in SPs of eukaryotic or bacterial origin. As expected from the analysis of the training set, the largest agreement of the individual SignalP-NN modules with PRED-SIGNAL is to the gram-positive module (correlation coefficient = 0.646), followed by the gram-negative and Eukaryotic modules. Similar, although not identical, results hold also for the SignalP-HMM predictors (data not shown).

Conclusions

In this work, we present a first computational method that specifically predicts the SPs of archaeal origin and their cleavage sites. We performed an extensive literature search in order to identify SPs with experimentally verified cleavage sites, as well as verified SPs in which the cleavage site is not precisely located. The analysis confirms previous results that suggested a unique composition of archaeal SPs and justifies our approach for modeling separately the particular sequences. We used an HMM approach, and trained the model to discriminate secretory SPs from cytoplasmic proteins as well as from proteins with an N-terminal TM segment, as these segments are often confused by predictors. The prediction method was also applied to the currently available completely sequenced genomes of archaea, and the results were compared with those of SignalP, which is considered to be the most accurate predictor of non-archaeal sequences. The new prediction method, PRED-SIGNAL, and the secreted proteins identified in the genome analysis are available online at: http://bioinformatics.biol.uoa.gr/PRED-SIGNAL/. We anticipate that this method will be a useful tool for those studying secreted proteins of archaea, since it could be used in genome annotation, genome-wide analyses, and for various proteomics applications. Finally, we note that the modular nature of the HMM allows easily the extension of the model, i.e. in order to incorporate joint prediction of Tat SPs or lipoprotein SPs. In our data set we have included 18 Tat substrates, and we found not >10 archaeal lipoproteins. However, when further experimental data become available on these classes of SPs in the near future, the model's architecture could be easily expanded in order to include them and allow better discrimination capability.

Acknowledgements

The authors would like to thank the two anonymous reviewers and the editors for their very helpful comments and the constructive criticism that helped in the improvement of the manuscript.

Funding

P.G.B. was supported by a scholarship from the State Scholarships Foundation of Greece (SSF), for post-doctoral research in the Department of Cell Biology and Biophysics of the University of Athens (Machine Learning Algorithms for Bioinformatics).

References

Akca,E., Claus,H., Schultz,N., Karbach,G., Schlott,B., Debaerdemaeker,T., Declercq,J.P. and Konig,H. (2002) *Extremophiles*, **6**, 351–358.

Alber,B.E. and Ferry,J.G. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 6909–6913.

Albers,S.V. and Driessen,A.M. (2002) *Arch. Microbiol.*, **177**, 209–216.

Bagos,P.G., Liakopoulos,T.D. and Hamodrakas,S.J. (2006) *BMC Bioinformatics*, **7**, 189.

Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A. and Nielsen,H. (2000) *Bioinformatics*, **16**, 412–424.

Bardy,S.L., Eichler,J. and Jarrell,K.F. (2003) *Protein Sci.*, **12**, 1833–1843.

Bauer,M.W., Driskill,L.E., Callen,W., Snead,M.A., Mathur,E.J. and Kelly,R.M. (1999) *J. Bacteriol.*, **181**, 284–290.

Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) *J. Mol. Biol.*, **340**, 783–795.

Bendtsen,J.D., Nielsen,H., Widdick,D., Palmer,T. and Brunak,S. (2005) *BMC Bioinformatics*, **6**, 167.

Berks,B.C., Palmer,T. and Sargent,F. (2005) *Curr. Opin. Microbiol.*, **8**, 174–181.

Brockl,G., Behr,M., Fabry,S., Hensel,R., Kaudewitz,H., Biendl,E. and Konig,H. (1991) *Eur. J. Biochem.*, **199**, 147–152.

Brown,S.H. and Kelly,R.M. (1993) *Appl. Environ. Microbiol.*, **59**, 2614–2621.

Bult,C.J., *et al*. (1996) *Science*, **273**, 1058–1073.

Catara,G., Ruggiero,G., La Cara,F., Digilio,F.A., Capasso,A. and Rossi,M. (2003) *Extremophiles*, **7**, 391–399.

Cheung,J., Danna,K.J., O'Connor,E.M., Price,L.B. and Shand,R.F. (1997) *J Bacteriol.*, **179**, 548–551.

Chong,P.K. and Wright,P.C. (2005) *J. Proteome Res.*, **4**, 1789–1798.

Choo,K.H., Tan,T.W. and Ranganathan,S. (2005) *BMC Bioinformatics*, **6**, 249.

Cohen,G.N., *et al*. (2003) *Mol. Microbiol.*, **47**, 1495–1512.

Comfort,D.A., Chou,C.J., Conners,S.B., VanFossen,A.L. and Kelly,R.M. (2008) *Appl. Environ. Microbiol.*, **74**, 1281–1283.

Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) *Genome Res.*, **14**, 1188–1190.

Dharmavaram,R., Gillevet,P. and Konisky,J. (1991) *J. Bacteriol.*, **173**, 2131–2133.

Dilks,K., Gimenez,M.I. and Pohlschroder,M. (2005) *J. Bacteriol.*, **187**, 8104–8113.

Driessen,A.J. and Nouwen,N. (2008) *Annu. Rev. Biochem*, **77**, 643–667.

Duffner,F., Bertoldo,C., Andersen,J.T., Wagner,K. and Antranikian,G. (2000) *J. Bacteriol.*, **182**, 6331–6338.

Durbin,R., Eddy,S.R., Krogh,A. and Mithison,G. (1998), *Biological Sequence Analysis*. Cambridge University Press.

Erra-Pujada,M., Debeire,P., Duchiron,F. and O'Donohue,M.J. (1999) *J. Bacteriol.*, **181**, 3284–3287.

Fariselli,P., Martelli,P.L. and Casadio,R. (2005) *BMC Bioinformatics*, **6**(Suppl. 4), S12.

Ferrer,M., Golyshina,O.V., Plou,F.J., Timmis,K.N. and Golyshin,P.N. (2005) *Biochem. J.*, **391**, 269–276.

Gimenez,M.I., Dilks,K. and Pohlschroder,M. (2007) *Mol. Microbiol.*, **66**, 1597–1606.

Goldman,S., Hecht,K., Eisenberg,H. and Mevarech,M. (1990) *J. Bacteriol.*, **172**, 7065–7070.

Habib,S.J., Neupert,W. and Rapaport,D. (2007) *Methods Cell Biol.*, **80**, 761–781.

Hashimoto,Y., Yamamoto,T., Fujiwara,S., Takagi,M. and Imanaka,T. (2001) *J. Bacteriol.*, **183**, 5050–5057.

Hiller,K., Grote,A., Scheer,M., Munch,R. and Jahn,D. (2004) *Nucleic Acids Res.*, **32**, W375–W379.

Hutcheon,G.W., Vasisht,N. and Bolhuis,A. (2005) *Extremophiles*, **9**, 487–495.

Izotova,L.S., Strongin,A.Y., Chekulaeva,L.N., Sterkin,V.E., Ostoslavskaya,V.I., Lyublinskaya,L.A., Timokhina,E.A. and Stepanov,V.M. (1983) *J. Bacteriol.*, **155**, 826–830.

Jones,R.A., Jermiin,L.S., Easteal,S., Patel,B.K. and Beacham,I.R. (1999) *J. Appl. Microbiol.*, **86**, 93–107.

Juncker,A.S., Willenbrock,H., Von Heijne,G., Brunak,S., Nielsen,H. and Krogh,A. (2003) *Protein Sci.*, **12**, 1652–1662.

Kall,L., Krogh,A. and Sonnhammer,E.L. (2004) *J. Mol. Biol.*, **338**, 1027–1036.

Kall,L., Krogh,A. and Sonnhammer,E.L. (2005) *Bioinformatics*, **21**(Suppl. 1), i251–i257.

Kall,L., Krogh,A. and Sonnhammer,E.L. (2007) *Nucleic Acids Res.*, **35**, W429–W432.

Kamekura,M., Seno,Y., Holmes,M.L. and Dyall-Smith,M.L. (1992) *J. Bacteriol.*, **174**, 736–742.

Kamekura,M., Seno,Y. and Dyall-Smith,M. (1996) *Biochim. Biophys. Acta*, **1294**, 159–167.

Kannan,Y., Koga,Y., Inoue,Y., Haruki,M., Takagi,M., Imanaka,T., Morikawa,M. and Kanaya,S. (2001) *Appl. Environ. Microbiol.*, **67**, 2445–2452.

Kashima,Y., Mori,K., Fukada,H. and Ishikawa,K. (2005) *Extremophiles*, **9**, 37–43.
Kawarabayasi,Y., *et al*. (1998) *DNA Res.*, **5**, 55–76.
Kawarabayasi,Y., *et al*. (2001) *DNA Res.*, **8**, 123–140.
Kim,B.K., Pihl,T.D., Reeve,J.N. and Daniels,L. (1995) *J. Bacteriol.*, **177**, 7178–7185.
Krogh,A. (1994) *Proceedings of the12th IAPR International Conference on Pattern Recognition*, pp. 140–144.
Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) *J. Mol. Biol.*, **305**, 567–580.
Lechner,J. and Sumper,M. (1987) *J. Biol. Chem.*, **262**, 9724–9729.
Lee,P.A., Tullman-Ercek,D. and Georgiou,G. (2006) *Annu. Rev. Microbiol.*, **60**, 373–395.
Leveque,E., Haye,B. and Belarbi,A. (2000) *FEMS Microbiol. Lett.*, **186**, 67–71.
Lim,J.K., Lee,H.S., Kim,Y.J., Bae,S.S., Jeon,J.H., Kang,S.G. and Lee,J.H. (2007) *J. Microbiol. Biotechnol.*, **17**, 1242–1248.
Limauro,D., Cannio,R., Fiorentino,G., Rossi,M. and Bartolucci,S. (2001) *Extremophiles*, **5**, 213–219.
Lin,X. and Tang,J. (1990) *J. Biol. Chem.*, **265**, 1490–1495.
Madan Babu,M. and Sankaran,K. (2002) *Bioinformatics*, **18**, 641–643.
Madan Babu,M., Priya,M.L., Selvan,A.T., Madera,M., Gough,J., Aravind,L. and Sankaran,K. (2006) *J. Bacteriol.*, **188**, 2761–2773.
Mander,G.J., Duin,E.C., Linder,D., Stetter,K.O. and Hedderich,R. (2002) *Eur. J. Biochem.*, **269**, 1895–1904.
Mattar,S., Scharf,B., Kent,S.B., Rodewald,K., Oesterhelt,D. and Engelhard,M. (1994) *J. Biol. Chem.*, **269**, 14939–14945.
Melen,K., Krogh,A. and von Heijne,G. (2003) *J. Mol. Biol.*, **327**, 735–744.
Menne,K.M., Hermjakob,H. and Apweiler,R. (2000) *Bioinformatics*, **16**, 741–742.
Morikawa,M., Izawa,Y., Rashid,N., Hoaki,T. and Imanaka,T. (1994) *Appl. Environ. Microbiol.*, **60**, 4559–4566.
Nielsen,H. and Krogh,A. (1998) *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 122–130.
Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) *Protein Eng.*, **10**, 1–6.
Nielsen,H., Brunak,S. and von Heijne,G. (1999) *Protein Eng.*, **12**, 3–9.
Palmieri,G., Casbarra,A., Fiume,I., Catara,G., Capasso,A., Marino,G., Onesti,S. and Rossi,M. (2006) *Extremophiles*, **10**, 393–402.
Perez-Pomares,F., Bautista,V., Ferrer,J., Pire,C., Marhuenda-Egea,F.C. and Bonete,M.J. (2003) *Extremophiles*, **7**, 299–306.
Pohlschroder,M., Gimenez,M.I. and Jarrell,K.F. (2005) *Curr. Opin. Microbiol.*, **8**, 713–719.
Rapoport,T.A., Matlack,K.E., Plath,K., Misselwitz,B. and Staeck,O. (1999) *Biol. Chem.*, **380**, 1143–1150.
Rose,R.W., Bruser,T., Kissinger,J.C. and Pohlschroder,M. (2002) *Mol. Microbiol.*, **45**, 943–950.
Ruiz,D.M. and De Castro,R.E. (2007) *J. Ind. Microbiol. Biotechnol.*, **34**, 111–115.
Sako,Y., Croocker,P.C. and Ishida,Y. (1997) *FEBS Lett.*, **415**, 329–334.
Sankaran,K. and Wu,H.C. (1994) *J. Biol. Chem.*, **269**, 19701–19706.
Sankaran,K. and Wu,H.C. (1995) *Methods Enzymol.*, **248**, 169–180.
Sankaran,K., Gupta,S.D. and Wu,H.C. (1995) *Methods Enzymol.*, **250**, 683–697.
Saunders,N.F., Ng,C., Raftery,M., Guilhaus,M., Goodchild,A. and Cavicchioli,R. (2006) *J. Proteome Res.*, **5**, 2457–2464.
Schneider,T.D. and Stephens,R.M. (1990) *Nucleic Acids Res.*, **18**, 6097–6100.
Serour,E. and Antranikian,G. (2002) *Antonie Van Leeuwenhoek*, **81**, 73–83.
Setubal,J.C., Reis,M., Matsunaga,J. and Haake,D.A. (2006) *Microbiology*, **152**, 113–121.
She,Q., *et al*. (2001) *Proc. Natl Acad. Sci. USA*, **98**, 7835–7840.
Shi,W., Tang,X.F., Huang,Y., Gan,F., Tang,B. and Shen,P. (2006) *Extremophiles*, **10**, 599–606.
Sumper,M., Berg,E., Mengele,R. and Strobel,I. (1990) *J. Bacteriol.*, **172**, 7111–7118.
Sun,C., Li,Y., Mei,S., Lu,Q., Zhou,L. and Xiang,H. (2005) *Mol. Microbiol.*, **57**, 537–549.
Sutcliffe,I.C. and Harrington,D.J. (2002) *Microbiology*, **148**, 2065–2077.
Tanaka,T., Fujiwara,S., Nishikori,S., Fukui,T., Takagi,M. and Imanaka,T. (1999) *Appl. Environ. Microbiol.*, **65**, 5338–5344.
Teter,S.A. and Klionsky,D.J. (1999) *Trends Cell Biol.*, **9**, 428–431.
Thomas,J.R. and Bolhuis,A. (2006) *FEMS Microbiol. Lett.*, **256**, 44–49.
Tuteja,R. (2005) *Arch Biochem. Biophys.*, **441**, 107–111.
Valente,F.M., Pereira,P.M., Venceslau,S.S., Regalla,M., Coelho,A.V. and Pereira,I.A. (2007) *FEBS Lett.*, **581**, 3341–3344.
van Roosmalen,M.L., Geukens,N., Jongbloed,J.D., Tjalsma,H., Dubois,J.Y., Bron,S., van Dijl,J.M. and Anne,J. (2004) *Biochim. Biophys. Acta*, **1694**, 279–297.
von Heijne,G. (1986) *Nucleic Acids Res.*, **14**, 4683–4690.
von Heijne,G. (1989) *Protein Eng.*, **2**, 531–534.
von Heijne,G. (1990) *J. Membr. Biol.*, **115**, 195–201.
von Heijne,G., Steppuhn,J. and Herrmann,R.G. (1989) *Eur. J. Biochem.*, **180**, 535–545.
Voorhorst,W.G., Eggen,R.I., Geerling,A.C., Platteeuw,C., Siezen,R.J. and Vos,W.M. (1996) *J. Biol. Chem.*, **271**, 20426–20431.
Voorhorst,W.G., Warner,A., de Vos,W.M. and Siezen,R.J. (1997) *Protein Eng.*, **10**, 905–914.
Wakai,H., Nakamura,S., Kawasaki,H., Takada,K., Mizutani,S., Aono,R. and Horikoshi,K. (1997) *Extremophiles*, **1**, 29–35.
Wang,L., Zhou,Q., Chen,H., Chu,Z., Lu,J., Zhang,Y. and Yang,S. (2007) *J. Ind. Microbiol. Biotechnol.*, **34**, 187–192.
Woodson,J.D., Reynolds,A.A. and Escalante-Semerena,J.C. (2005) *J. Bacteriol.*, **187**, 5901–5909.
Wu,C.H., *et al*. (2006) *Nucleic Acids Res.*, **34**, D187–D191.

Received May 17, 2008; revised September 30, 2008; accepted October 9, 2008

Edited by Todd Yeates