Biologia **64**/4: 655—659, 2009
Section Cellular and Molecular Biology
DOI: 10.2478/s11756-009-0118-3

Evaluating signal peptide prediction methods for Gram-positive bacteria

Xiaohui Zhang[1], Yudang Li[2] & Yudong Li[3]*

[1] *College of Animal Sciences, Zhejiang University, Zhejiang Aquatic Products Quality Detection Center, Hangzhou 310012, People's Republic of China*
[2] *Wenzhou Vocational College of Science & Technology, Wenzhou 325006, People's Republic of China*
[3] *College of food Science and biotechnology, Zhejiang Gongshang University, Hangzhou 310012, People's Republic of China; e-mail: youdng@gmail.com*

Abstract: Gram-positive bacteria have been widely investigated for their huge capability to secrete proteins, such as those involved in gene expression, bacterial surface display and bacterial pathogenesis. The N-terminal signal peptide of a secretory protein is responsible for the translocation of polypeptide through the cytoplasmic membrane. Recently, the signal peptide prediction has become a major task in bioinformatics, and many programs with different algorithms were developed to predict signal peptides. In this paper, five prediction programs (SignalP 3.0, PrediSi, Phobius, SOSUIsignal and SIG-Pred) were selected to evaluate their prediction accuracy for signal peptides and cleavage site using 509 unbiased and experimentally verified Gram-positive protein sequences. The results showed that SignalP was the most accurate program in signal peptide (96% accuracy) and cleavage site (83%) prediction. Prediction performance could further be improved by combining multiple methods into consensus prediction, which would increase the accuracy to 98%, and decrease the false positive to zero. When the consensus method was used to predict *Bacillus*'s extracellular proteins identified by proteomics, more new signal peptides were successfully identified. It could be concluded that the consensus method would be useful to make prediction of signal peptides more reliable.

Key words: signal peptide; cleavage site; prediction; gram-positive bacteria.

Abbreviations: HMM, hidden Markow model; MCC, Matthew's correlation coefficient; NN, neural network; Sn, sensitivity; SP, signal peptide; Sp, specificity.

Introduction

Protein transport from the cytoplasm to cell envelope or extracellular environment is a common phenomenon for bacteria (Gardy & Brinkman 2006). Secreted proteins are essential for bacterial growth including nutrients uptake, virulence, and environmental sensing. They are usually synthesized as precursors with an N-terminal signal peptide (SP), which acts as a targeting ticket directing proteins to the transport machineries located in the cytoplasmic membrane and are cleaved off latterly by specific signal peptidases (Schneider & Fechner 2004).

Among multiple protein transportation pathways in Gram-positive bacteria, the Sec and Tat pathways were mostly investigated (Pohlschroder et al. 2005). Their N-terminal SPs were classified as Class I type SP, which was processed by a class I signal peptidase. The efficient and accurate prediction of SPs and their cleavage sites were important to a wide range of studies, such as genome-wide analysis of proteins' subcellular localization and industrial application (Gardy & Brinkman 2006). Over the past few years, the identification of the type I SP has become a major task in bioinformatics. Despite SPs show no conservation in sequences, the SPs generally consist of three structurally and functionally distinct regions: (i) an N-terminal positively charged n-region; (ii) a central hydrophobic h-region; and (iii) a neutral but polar c-region. Based on these recognized characters, many algorithms have been developed to predict SP and its cleavage site, such as weight matrix, neural network, hidden Markov model, and so on (Emanuelsson et al. 2007).

Currently, there have been many progresses in SP prediction, but none of the prediction methods can achieve 100% accuracy (Schneider & Fechner 2004). The single predictor was not enough to identify all SP with different characteristics, because each program was trained with special dataset and with different algorithms, so it is advised to use several prediction techniques simultaneously whenever possible. Although many prediction methods have been developed recently, their predictive performance has not yet been independently compared in sig-

* Corresponding author

nal sequences of bacteria (Klee & Ellis 2005; Menne 2000).

In contrast to Gram-negative bacteria with an inner and outer membrane, Gram-positive bacteria are surrounded only by a single cytoplasmic membrane and their protein secretion pathway became simplified, so many Gram-positive bacteria have been used as hosts to secrete higher levels of heterologous proteins in gene expression and bacterial surface display (Freudl 1992). Currently, the prediction of protein localization of Gram-positive bacteria has become increasingly important with availability of gene products in the post-genomic era (Shen & Chou 2007). Five prediction programs including Signalp 3.0, PrediSi, Phobius, SOSUIsignal and SIG-Pred were developed to predict SPs for Gram-positive bacteria, in which the SignalP program was most widely used.

In this study, five web-based prediction programs, Signalp 3.0, PrediSi, Phobius, SOSUIsignal and SIG-Pred were selected, which produce both SP classification and cleavage site assignment. Based on our datasets of experimentally verified Gram-positive bacteria SPs, their ability to predict SP and cleavage site prediction accuracy was evaluated. We further combined these programs to achieve even better performance with a simple algorithm.

Material and methods

Creating protein test sets

The experimentally verified SPs were collected from databases of Swiss-Prot (Boeckmann et al. 2003) and SPdb (Choo et al. 2005). Swiss-Prot (Release 52.4) was downloaded from EBI ftp server, and entries annotated with "firmicutes" and "actinobacteria" in OC (organism classification) field were selected because we focused on Gram-positive proteins only. If SPs and their cleavage site were ambiguous, i.e. if annotated as "potential", "possible" and "by similarity" in FT (feature) field, the entries were eliminated. Sequences with <50 amino acid residues and the same N-terminal 30 amino acids were discarded. Lipoproteins cleaved by signal peptidase II were also removed since their cleavage site was different from prokaryotic signal peptidase I. Some SPs were extracted from the database SPdb, and the repeats with Swiss-Prot records were removed. We obtained a positive SPs test set containing 249 sequences.

Cytoplasmic proteins were downloaded from Swiss-Prot with DT (date) field annotated as "2006/2007", and some entries were retrieved from experimentally verified cytoplasmic proteins in PSORTdb database (Rey et al. 2005). The total number of proteins without SP was 260.

We also constructed a dataset of 275 extracellular proteins of *Bacillus* species, which were identified by proteomics, including *B. subtilis*, *B. anthracis* and *B. licheniformis* (Antelmann et al. 2001, 2005; Voigt et al. 2006).

Prediction programs

Five free programs were selected for web-based and bacteria parameters. SignalP (http://www.cbs.dtu.dk/services/SignalP/) has two algorithms: neural network (NN) and hidden Markov model (HMM). Prediction was done with settings for "Gram-positive bacteria" sequence data, "Both" analysis methods, "No graphics" output, "Short" output and sequence truncation set on 70 residues. PrediSi (http://www.predisi.de/) was run using the Gram-positive bacteria organism group and a text-based output (Hiller et al. 2004). Phobius (http://phobius.cgb.ki.se/) was run with short output format. SOSUIsignal (http://bp.nuap.nagoya-u.ac.jp/sosui/sosuisignal/) and SIG-Pred (http://www.bioinformatics.leeds.ac.uk/prot_analysis/Signal.html) were run sequence by sequence with "prokaryote/Gram-positive bacteria" parameters (Gomi et al. 2000). Output files from all analyses were parsed and program performance measures were calculated by using custom Perl scripts.

Program performance measures

Performance of prediction methods on test set was measured as sensitivity (Sn), specificity (Sp) and Matthew's correlation coefficient (MCC).The calculation formula was as follows:

Sn = TP/ (TP + FN), Sn is the proportion of SPs that have been correctly predicted as SPs;

Sp = TP/ (TP + FP), Sp is the proportion of predicted SPs that are actually SPs;

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})}},$$

MCC equals one for a perfect prediction, while it is zero for a completely random assignment.

In the above formulas, TP = true positives, TN = true negatives, FN = false negatives (under-prediction), and FP = false positives (over-prediction).

Performance was also evaluated for combination of prediction methods. With combinatorial analysis, a SP was assigned only if it is positively discriminated by three or more methods and the sum of two discrimination scores from SignalP was >1.0. Otherwise it was not assigned as a SP.

Results

Common features of signal peptides in Gram-positive bacteria

When using the sequence logo method (Crooks et al. 2004) to analyze the positional preferences of amino acids of Gram-positive bacteria SP database, it was confirmed that these SPs have some obviously common features. The length of SPs is between 14 and 59 residues, most (about 80%) of which with a length between 25 and 40. As shown in Figure 1, Gram-positive bacteria are known to have longer SPs that carry more basic residues (K/R) in the n-region than that of Gram-negative bacteria and eukaryotes (Li et al. 2006). The SPs mostly contain three conserved parts: a positively charged amino acids in the n-region, such as arginine and lysine; the hydrophobic h-region which is rich in leucine, alanine, serine, and valine from about 6th to 16th residue; and the c-region ending with the cleavage site recognized by signal peptidase. The positions −1 and −3 from the cleavage site are rich in alanine or other residues with short side chains, such as glycine and valine. Furthermore, position +1 from the cleavage site prefers alanine.

Signal peptide prediction

To evaluate SP prediction, program performance on the test sets was measured with sensitivity (Sn), specificity

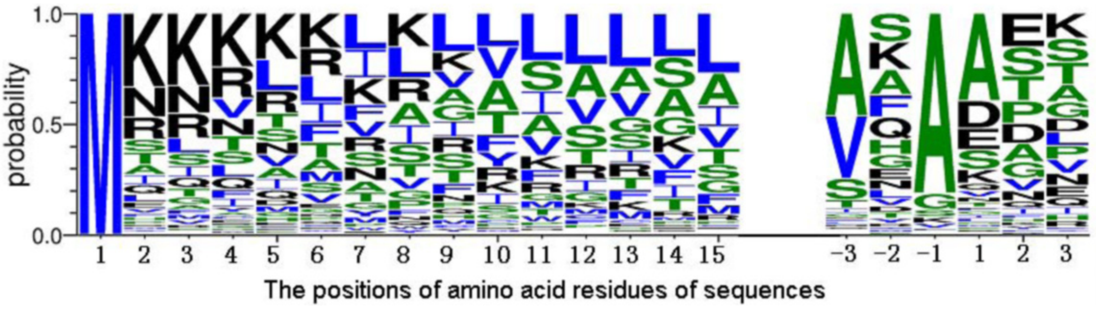

Fig. 1. Amino acid composition of SPs. The total height of the stack of letters at each position shows the amount of sequence conservation at the position, while the relative height of each letter shows the relative abundance of the corresponding amino acid. The left side of the chart shows the positions from the 1st to the 15th residues in SPs. The positions −3 to +3 from signal peptide cleavage site is shown on the right side.

Table 1. Programs performance measurement.[a]

Program	TP	FP	TN	FN	Sn	Sp	MCC
SignalP-nn	240	2	258	9	0.96	0.99	0.96
SignalP-hmm	244	2	258	5	0.97	0.99	0.97
PrediSi	234	2	258	15	0.93	0.99	0.93
Phobius	232	3	257	17	0.93	0.98	0.92
Sig-pred	206	8	252	43	0.82	0.96	0.80
SOSUIsignal	204	10	250	45	0.82	0.95	0.79
Consensus method	245	0	260	4	0.98	1	0.99

[a] Performance was measured based on the ability of the programs to correctly distinguish signal peptides from non-signal peptides. SignalP-nn: SignalP neural network; SignalP-hmm: SignalP hidden Markow model.

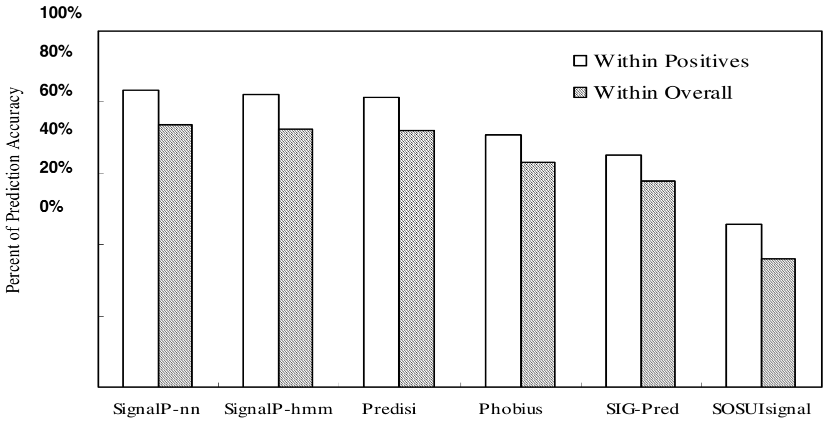

Fig. 2. The cleavage sites prediction accuracy in Gram-positive bacteria. The Y-axis indicates the percentage of SP sequences where the cleavage site was placed correctly. The empty bars represent the percentages of correctly predicted sites among positively predicted sequences by each program. The hatched bars represent overall percentages that were measured using all these sequences.

(Sp), accuracy (Ac), and Matthew's correlation coefficient (MCC) as benchmarks. SP predictive accuracy for individual prediction program is shown in Table 1.

Based on MCC, SignalP was found to be the most accurate predictor. Two algorithms of SignalP, neural network (nn) and hidden Markow model (hmm) exhibited similar accuracy (they have a MCC value of 96% and 97%, respectively). SignalP-hmm predicted four more true SPs than did the SignalP-nn. However, Sig-Pred and SOSUIsignal were weak in SP prediction evaluation for Gram-positive bacteria. PrediSi and Phobius have similar performance, a little less accurate than SignalP (Predisi was 0.93 and Phobius was 0.92).

As Sp value was strikingly high, the prediction result was reliable when sequence was predicted to contain a SP. In contrast, Sn was comparably lower than Sp, and all programs still have room to be improved in the capability to detect SP. Taken together, the existing prediction programs were powerful to identify SP.

Cleavage site prediction

In particular, it was critical to realistically assess the prediction accuracy of cleavage site, because it was often desirable to produce hybrid, functional secreted proteins with tag linked precisely to the N-termini of mature proteins for scientific and commercial purposes (Zhang & Henzel 2004). As shown in Figure 2, SignalP and PrediSi showed nearly the same ability to predict

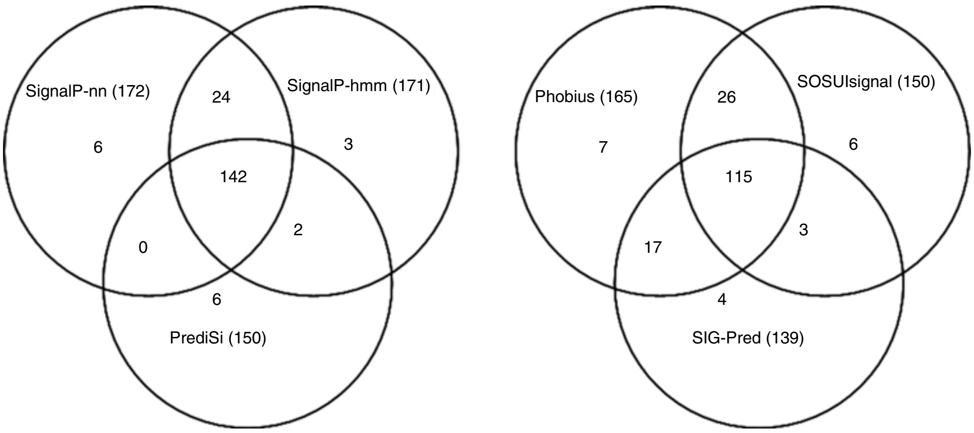

Fig. 3. SP prediction on *Bacillus*'s extracellular proteomics dataset. Numbers in parentheses were total proteins predicted by five programs: SignalP-nn, SignalP-hmm, PrediSi, Phobius, SOSUIsignal and SIG-Pred. The number of SP identified by each program and their shared prediction are shown in each circle areas.

the signal cleavage sites in Gram-positive bacteria, with the accuracy of 80% and above. In contrast, both Phobius and SIG-Pred yielded markedly lower accuracies. The best program appeared to be SignalP-nn (83%) followed by SignalP-hmm (82%) and PrediSi (81%). Phobius has an accurate value of about 70%, which is approximately 10 percent less than both SignalP and PrediSi. Similar to signal sequence prediction, SOSUIsignal has worst ability in cleavage site prediction (45%).

We found that the prediction performance of locating cleavage site was similar to that of signal sequences. It was believed that more accuracy in identifying SP would have more accuracy in locating cleavage sites.

Improving prediction by combining methods

Among the existing prediction methods, SignalP was the most effective and was widely used in many works to determine SPs (Brockmeier et al. 2006). But each predictor has its limit, especially in cleavage site prediction. False positive or false negative results were easy to happen when using a prediction method, so it was necessary for us to combine them with a simple algorithm (see Materials and methods section) to get more reliable prediction. SignalP have provided discriminate score (D-score/Sprob) in output, and it was easy to see whether the prediction was more or less reliable by comparing discriminate score to the cutoff value. The consensus prediction was evaluated on each program (Table 1).

As shown in Table 1, the MCC value of combined method was the best one (0.99). False positive and false negative results were reduced to 0 and 4, respectively. This consensus method showed highest true positive value, which is important for biological research.

A continuously increasing number of extracellular proteins identified by proteomics may include many new secreted proteins whose signal sequences have not been revealed before. We used the proteomics data to compare the prediction capability of each predictive method to detect SP. *Bacillus* species have been taken as model organisms to secretion research and many proteomic papers have been published. We collected the extracellular proteins in these papers as dataset to evaluate the performance of each program. SignalP has been identified to have the highest amount of SPs in these proteins (Fig. 3), but some proteins were still ignored. More proteins containing SPs were identified by using PrediSi, Phobius, SOSUIsignal and Sig-Pred simultaneously. By using consensus methods, we found five more proteins with signal sequences (BA1973, BLi02391, pbpA, BLi00281, BLi03060). It is worth mentioning that in order to select a signal sequence more strictly, the consensus method may be useful to make the prediction more reliable.

Discussion

SP identification is important in industrial biotechnology. Accurate identification of SPs has become a prerequisite in order to use such technologies effectively. Based on our evaluation of these methods on bacteria, we found that the prediction accuracy of SPs was high; only two proteins (MTCY_LEUME, CWLA_BACSP; SWISS-PROT database) could not be identified, but the cleavage site prediction was less accurate. As the training sets of all the programs are very likely to contain sequences from SWISS-PROT, the evaluation result may be better since we have got the recent SWISS-PROT entries.

As documented in Table 1, the Sp of 5 programs was very accurate, namely the predicted SP was reliable. But the Sn was relatively low, and many SPs could not be identified by existing programs. That may be caused by the diversity of SPs and some features of SPs have not been used by existing programs. Except the amino acids composition characteristics, the codon usage or the secondary structure of signal sequences may be helpful to improve prediction performance (Li et al. 2006).

Another disadvantage of the computer-based approach was that it was limited to predict only those proteins secreted via the general export pathway. The

SPs belonged to class I type, which were cleaved by class I signal peptidase. Signal sequences may exist in the protein middle and also in the C-terminal parts, but unfortunately our program cannot predict the middle and C-terminal signal sequences.

Since different prediction programs have been developed for specific purposes, it is practical to improve the prediction performance by using different prediction methods simultaneously. SignalP-HMM was based on a hidden Markov model formalism and was developed in order to improve the discrimination between SPs and N-terminal transmembrane anchor segments. Phobius could add value to analysis of protein sequences containing N-terminal transmembrane domains. SOSUIsignal uses the propensities of occurrence of amino acids for the SP and other parameters as a membrane protein predictor. As shown by the results, many SPs had been erroneously predicted as transmembrane peptides. To discriminate between a SP and a real first transmembrane segment is difficult because most SPs have a hydrophobic core which resembles that of a typical transmembrane segment. PrediSi was the fastest program and did not restrict the size of sequence set analyzed. Therefore, if users work with extremely large datasets, PrediSi can be used for rapid initial screens.

Many extracellular proteins were detected by proteomics technique lacked a typical SP, which could not be detected by signalP (Antelmann et al. 2001), or these proteins were released by cell lysis or other unidentified export pathways. Among the 13 newly predicted SPs, which cannot be predicted by SignalP-nn, five proteins can be classified as secreted proteins by our consensus method (Q81RR7_BACAN, Q65NX4_BACLD, Q65I43_BACLD, MDH_BACLD, YQGF_BACSU; SWISS-PROT database). YQGF_BACSU is annotated as a transmembrane protein and belongs to transpeptidase family. Other proteins are enzymes associated with metabolic processes (such as hydrolysis activities), and some are not annotated at all in the SWISS-PROT database. This suggests that consensus prediction methods could result in better coverage in secreted protein prediction, especially when the goal was all secreted proteins, i.e., the secretome. Therefore, although the prediction accuracy of a method has been relatively high, the prediction could be more reliable by using these methods simultaneously.

Acknowledgements

We would like to thank Dr. Ximiao He of the Beijing Genomics Institute for his critical reading of this manuscript.

References

Antelmann H., Tjalsma H., Voigt B., Ohlmeier S., Bron S., van Dijl J.M. & Hecker M. 2001. A proteomic view on genome-based signal peptide predictions. Genome Res. **11:** 1484–1502.

Antelmann H., Williams R.C., Miethke M., Wipat A., Albrecht D., Harwood C.R. & Hecker M. 2005. The extracellular and cytoplasmic proteomes of the non-virulent *Bacillus anthracis* strain UM23C1–2. Proteomics **5:** 3684–3695.

Bendtsen J.D., Nielsen H., von Heijne G. & Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. **340:** 783–795.

Boeckmann B., Bairoch A., Apweiler R., Blatter M.C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S. & Schneider M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res.**31:** 365–370.

Brockmeier U., Caspers M., Freudl R., Jockwer A., Noll T. & Eggert T. 2006. Systematic screening of all signal peptides from *Bacillus subtilis*: a powerful strategy in optimizing heterologous protein secretion in Gram-positive bacteria. J. Mol. Biol. **362:** 393–402.

Choo K.H., Tan T.W., & Ranganathan S. 2005. SPdb – a signal peptide database. BMC Bioinformatics **6:** 249.

Crooks G.E., Hon G., Chandonia J.M. & Brenner S.E. 2004. WebLogo: a sequence logo generator. Genome Res. **14:** 1188–1190.

Gardy J.L. & Brinkman F.S. 2006. Methods for predicting bacterial protein subcellular localization. Nat. Rev. Microbiol. **4:** 741–751.

Gomi M., Akazawa F. & Mitaku S. 2000. SOSUIsignal: software system for prediction of signal peptide and membrane protein. Genome Informatics **11:** 414–415.

Emanuelsson O., Brunak S., von Heijne G. & Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat. Protoc. **2:** 953–971.

Hiller K., Grote A., Scheer M., Munch R. & Jahn D. 2004. PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res. **32 (Web Server issue):** 375–379.

Klee E.W. & Ellis L.B. 2005. Evaluating eukaryotic secreted protein prediction. BMC Bioinformatics **6:**256.

Li Y.D., Li Y.Q., Chen J.S., Dong H. J., Guan W. J. & Zhou H. 2006. Whole genome analysis of non-optimal codon usage in secretory signal sequences of *Streptomyces coelicolor*. Biosystems **85:** 225–230.

Menne K.M., Hermjakob H. & Apweiler R. 2000. A comparison of signal sequence prediction methods using a test set of signal peptides. Bioinformatics **16:** 741–742.

Freudl R. 1992. Protein secretion in gram-positive bacteria. J. Biotechnol. **23:** 231–240.

Pohlschroder M., Gimenez M.I. & Jarrell K.F. 2005. Protein transport in Archaea: Sec and twin arginine translocation pathways. Curr. Opin. Microbiol. **8:** 713–719.

Rey S., Acab M., Gardy J.L., Laird M.R., deFays K., Lambert C. & Brinkman F.S. 2005. A protein subcellular localization database for bacteria. Nucleic Acids Res. **33:** 164–168.

Schneider G. & Fechner U. 2004. Advances in the prediction of protein targeting signals. Proteomics **4:** 1571–1580.

Shen H.B. & Chou K.C. 2007. Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng. Des. Sel. **20:** 39–46.

Voigt B., Schweder T., Sibbald M.J., Albrecht D., Ehrenreich A., Bernhardt J., Feesche J., Maurer K.H., Gottschalk G., van Dijl J.M. & Hecker M. 2006. The extracellular proteome of *Bacillus licheniformis* grown in different media and under different nutrient starvation conditions. Proteomics **6:** 268–281.

Zhang Z. & Henzel W. J. 2004. Signal peptide prediction based on analysis of experimentally verified cleavage sites. Protein Sci. **13:** 2819–2824.

Received June 5, 2008
Accepted December 31, 2008