Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in Gram-negative bacteria

Kenichiro Imai[1, 2, 3*], Naoyuki Asakawa[1], Toshiyuki Tsuji[1], Fumitsugu Akazawa[1], Ayano Ino[1], Masashi Sonoyama[1] and Shigeki Mitaku[1]

[1]Department of Applied Physics, Graduate School of Engineering, Nagoya University, Furocho, Chikusa-ku, Nagoya 464-8606, Japan; [2]Toyota Physical and Chemical Research Institute, Nagakute-cho, Aichi 480-1192, Japan; [3]Computational Biology Research Center, AIST, Tokyo 135-0064, Japan; Kenichiro Imai* - E-mail: imai@bp.nuap.nagoya-u.ac.jp; * Corresponding Author

received July 04, 2008; accepted July 06, 2008; published July 14, 2008

Abstract:
A predictive software system, SOSUI-GramN, was developed for assessing the subcellular localization of proteins in Gram-negative bacteria. The system does not require the sequence homology data of any known sequences; instead, it uses only physicochemical parameters of the N- and C-terminal signal sequences, and the total sequence. The precision of the prediction system for subcellular localization to extracellular, outer membrane, periplasm, inner membrane and cytoplasmic medium was 92.3%, 89.4%, 86.4%, 97.5% and 93.5%, respectively, 000with corresponding recall rates of 70.3%, 87.5%, 76.0%, 97.5% and 88.4%, respectively. The overall performance for precision and recall obtained using this method was 92.9% and 86.7%, respectively. The comparison of performance of SOSUI-GramN with that of other methods showed the performance of prediction for extracellular proteins, as well as inner and outer membrane proteins, was either superior or equivalent to that obtained with other systems. SOSUI-GramN particularly improved the accuracy for predictions of extracellular proteins which is an area of weakness common to the other methods.

Keywords: subcellular localization of proteins; Gram-negative bacteria; physicochemical parameters; amino acids

Background:
Subcellular localization is one of the most important characteristics of proteins and is central to understanding their function and the constitution of biological systems. In bacteria, information related to the subcellular location of pathogen proteins can facilitate the development of drugs and vaccines for treatment. We previously developed a system called SOSUI for predicting inner membrane proteins from amino acid sequences **[1, 2]**. Because of its high accuracy, this system can be used to improve the performance of the prediction of subcellular localization. An advantage of SOSUI is that all of the parameters used for the prediction **[1, 2]** are the physicochemical properties of amino acids. By adopting a similar approach, we developed in this work a novel method for predicting subcellular localization in Gram-negative bacteria for proteins in the extracellular, the outer membrane, the periplasm, and the medium cytoplasm, which are less hydrophobic than the proteins in the inner membrane. The physicochemical properties of the N- and C-terminal signal regions, as well as a whole sequences were also used for predicting protein localization. By combining this method with the SOSUI membrane prediction system, we constructed a unified system, referred to as SOSUI-GramN, for the complete characterization of subcellular protein localization in the five compartments of Gram-negative bacteria.

Several systems for predicting the subcellular localization of proteins in bacterial cells have already been developed: PSORTb, CELLO, PSLpred and P-CLASSIFIER **[3-6]**. The common weak point of those systems is precision and recall for the prediction of proteins localized in the extracellular medium are low **[7]**. The reasons for the low predictive performance associated with extracellular proteins may be due to the complex secretory pathways found in Gram-negative bacteria; a minimum of seven mechanisms have been reported to date **[8, 9]**. In this work, we have assumed that the pathways through the cytoplasmic membrane can be categorized into two main groups: Sec dependent pathways and Sec independent pathways. Therefore, we classified proteins of the same localization into two main groups based on the physicochemical properties of their N-terminal signal regions prior to discrimination of the final localization of proteins. In addition, proteins with the same subcellular localization in each of the main groups were classified into several subclasses based on the physicochemical parameters of their N- and C-terminal signal sequences. Here we describe the development of a novel software system, SOSUI-GramN, which was achieved by combining the SOSUI system used for the prediction of inner membrane proteins with a new protein prediction system capable of discriminating among four different localizations of less hydrophobic proteins in Gram-

negative bacteria, leading to an overall precision of 92.9%. Precision and recall obtained using for extracellular proteins SOSUI-GramN were markedly higher than the currently available systems, i.e. PSORTb, CELLO, PSLpred and P-CLASSIFIER etc.

Methodology:

The dataset used for developing the novel system was obtained from ePSORTdb **[10]** and Swiss-Prot (release 54.4) by omitting data with sequence identities exceeding 50%. The dataset contained 1795 proteins: 733 in the cytoplasm, 396 in the inner membrane, 248 in the periplasm, 240 in the outer membrane and 178 in the extracellular medium. The dataset was randomly partitioned into five sets, four of which were used for training and the remaining data were used to evaluate the system. Furthermore, an additional test set of 299 proteins was created in order to be used for the comparison of different prediction systems **[7]** because a number of proteins of our test data were included in training data of other Method. There is no sequence, which is identical to training and test data, in the data of 299, and the sequence identity between this calibration data, the training data, and the test data was less than at most 50%.

The prediction procedure of the SOSUI-GramN system consists of three layers of filters may reflects several pathways for the same subcellular localization site as shown in Figure 1. The first layer is the filter, consisting of SOSUI **[1, 2]**, required for distinguishing inner membrane proteins from the other classes of proteins. Some physiochemical parameter such as the distribution of hydrophobicty and amphiphilicity index **[11, 12]** of amino acids around transmembrane regions and the size of protein, were combined in SOSUI to discriminate an inner membrane protein from other types of proteins. SOSUI achieved 98.0% accuracy of prediction for inner (or cytoplasmic) membrane protein of Prokayote. Because of its high accuracy, this system can distinctly differentiate inner membrane proteins from other proteins.

Using the physicochemical properties of N-terminal signal sequence region and whole sequence, the second filter then classifies all of the proteins that are not inner membrane proteins into two groups depending on whether they are involved in the Sec dependent or independent pathways. The proteins of the Sec dependent pathway generally have the signal sequence, consisting of the positively charged region, hydrophobic regions, and polar region in N-terminal and these are not found in the proteins involved in the Sec independent pathway. The physicochemical parameters, required for the second filter, are hydrophobicity index **[11]**, the densities of positively and negatively charged residues, small polar and non-polar residues, aromatic residues, secondary structure-breaking residues such as proline and glycine, the helical periodicity of 3.6 residues, and the alternate periodicity of the two residues of the hydrophobicity index. We calculated the average of these parameters of the region of 60 residues around N-terminal sorting signal and total sequence, and then distinguished Sec dependent proteins from Sec independent proteins by the canonical discriminant analysis with using the average of these parameters of each target region.

The fundamental process required for the discrimination is described below. (i) The target region was divided into ten segments, except for total sequence, and then an average of parameter for segment was calculated by equation (1) (see supplementary material). When the target region is total sequence, the region is defined as one segment. (ii) To obtain the normalized *p*-th parameter of the target region *j*, the average values is normalized by the difference between positive data and negative data by equation (2) (see supplementary material). Finally, we discriminate positive data from negative data by the canonical discriminant analysis with using normalized parameters.

The third filter consists of several prediction modules based on the subclassification of the extracellular proteins, the outer membrane proteins, the periplasmic proteins and the cytoplasmic proteins. Assuming that there are several pathways for the same subcellular localization site, we classified proteins into several subclasses by the canonical discriminant analysis based on the average of physicochemical parameters of the 60 residues of the most of N- and C-terminal, which may include sorting signals, prior to discrimination analysis. We then distinguished a subclass from the other subclasses by the same discriminant analysis with the average of physicochemical parameters of 60, 100 and 200 residues of N- and C-terminal residues, as well as the entire sequence. Consequently, the assembly of the prediction modules required for the third layer are relatively complicated, but basically, each prediction module correspond to the discrimination of each subclass from the other subclasses and the discrimination method is the same as the method is used in second filter. The parameters, for these subclassification and prediction modules, are also identical to that for the discrimination in the second filter. The number of prediction modules was five for the prediction of the extracellular proteins, six for the outer membrane proteins and three for the periplasmic proteins, one for the cytoplasmic proteins as shown in Figure 1.

Discussion:

When the performance of the SOSUI-GramN prediction system for characterizing subcellular localization to the compartments of the extracellular, the outer membrane, the periplasm, the inner membrane, and the cytoplasm medium was evaluated using test data, precision estimates of 92.3%, 89.4%, 86.4%, 97.5% and 93.5%, and corresponding recall values of 70.3%, 87.5%, 76.0%, 97.5% and 88.4% were obtained, respectively. The overall performance for precision and recall was 92.9% and 86.7% (Table 1a under supplementary material). In this study, precision was calculated as TP / (TP+FP) and recall as TP / (TP+FN), where TP, FP and FN represent the number of true positive, false positive and false negative data, respectively. Table 1b (supplementary material) shows the performance of our system relative to that of other systems using 299 proteins as the calibration data. The assessment of individual methods

reveals that the highest overall precision and recall were achieved by Psortb and SOSUI-GramN, respectively. The prediction of the extracellular proteins was particularly low accuracy and common weakness of the other methods, however, our system archived the highest precision and recall, which are 90.6 and 50.0%, respectively. Our system SOSUI-GramN significantly improved the precision of extracellular proteins. The performance of the prediction of outer and inner membrane protein was superior or equivalent to that of the other method. The highest precision of outer membrane proteins was achieved by Psortb, however our method attained the highest recall at 92.1%. Psortb performed the highest precision and recall for inner membrane protein at 96.5 and 79.7%, respectively followed closely by our method at 96.2 and 72.5%. The other hand, the highest precision for periplasmic and cytoplasmic proteins are attained by Psortb, and CELLO, PSLpred and P-CLASSIFIER archived higher recall for cytoplasmic than our method. The accuracy of prediction for periplasmic and cytoplasmic protein is lower than the other methods, however, our method predicted extracellular protein, outer and inner membrane protein, which are important for facilitating the development of drugs for bacterial pathogen, with high accuracy, so SOSUI-GramN would be particularly useful for screening amino acid sequences in medical applications.

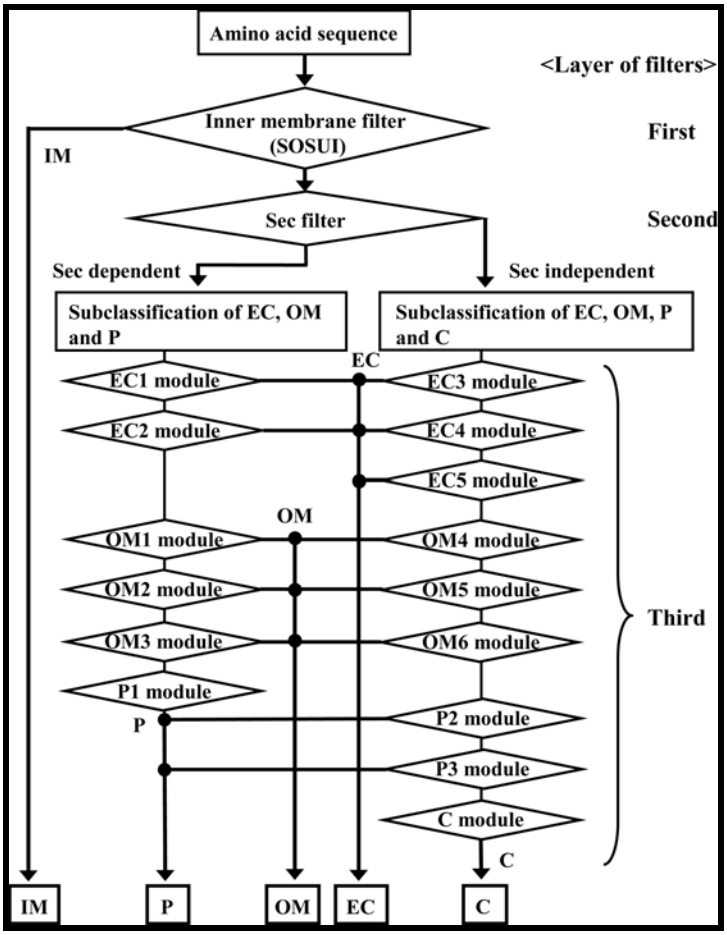

Figure 1: A simplified SOSUI-GramN flowchart.

Conclusion:

Gram-negative bacteria have five major subcellular localization sites, which are extracellular, outer membrane, periplasm, inner membrane and cytoplasm. Proteins synthesized in cytosol and then are targeted and transported to their localization sites through translocation pathways, however the translocation pathways to same localization site is not single pathway. The secretion pathways of extracellular proteins are at least seven pathways, so this complicated pathways result in the diversity of target signals, moreover the low prediction accuracy of extracelluar proteins. In this study, we developed the novel method for prediction the subcellular localization of proteins in Gram-negative bacteria based on the classification of protein on the assumption of the several pathways for the same localization site. The comparison of individual methods (Table 1b in supplementary material) reveals that the performance associated with the prediction of

extracellular proteins, which is an area of weakness common to many of the other methods, was superior to that obtained with other systems. Most of sorting signals is not fully understood, but our classification based on physicochemical properties around the sorting signal may provide the clue to make clear features of sorting signals.

The SOSUI-GramN system is a web-based application that users can use after inputting their single or multiple sequences on the submission. Note that a minimum input sequence length of more than 60 amino acids is required. The predicted subcellular location of the proteins of interest in Gram-negative bacterial cells is shown on the output page. In the event that the sequence length is less than 60 amino acids or if the confident prediction cannot be estimated, the SOSUI-GramN returns a value of "unknown". The population of "unknown" in data tested was only about 7% which is smaller than the currently available systems. The SOSUI-GramN system is available at http://bp.nuap.nagoya-u.ac.jp/sosui/sosuigramn/sosuigramn_submit.html.

Acknowledgment:
This work was supported in part by a Grant-in-Aid from SENTAN, JST.

References:

[01] T. Hirokawa, *et al*., *Bioinformatics*, 14: 378 (1998) [PMID: 9632836]
[02] T. Tsuji and S. Mitaku, *CBIJ,* 4: 110 (2004)
[03] J. L. Gardy, *et al*., *Bioinformatics,* 21: 617 (2005) [PMID: 15501914]
[04] C. S. Yu, *et al*., *Proteins*, 64: 643 (2006) [PMID: 16752418]
[05] M. Bhasin, *et al*., *Bioinformatics*, 21: 2522 (2005) [PMID: 15699023]
[06] J. Wang, *et al.*, *BMC Bioinformatics,* 6: 174 (2005) [PMID: 16011808]
[07] J. L. Gardy and F. S. Brinkman, *Nat Rev Microbiol.*, 4: 741 (2006) [PMID: 16964270]
[08] R. G. Gerlach and M. Hensel, *Int J Med Microbiol.*, 297: 401 (2007) [PMID: 17482513]
[09] M. Kostakioti, *et al*., *J Bacteriol.,* 187: 4306 (2005) [PMID: 15968039]
[10] S. Rey, *et al*., *Nucleic Acids Res*., 33: D164 (2005) [PMID: 15608169]
[11] J. Kyte and R. F. Doolittle, *J Mol Biol*., 157: 105 (1982) [PMID: 7108955]
[12] S. Mitaku *et al*., *Bioinformatics*, 18: 608 (2002) [PMID: 12016058]

Edited by P. Kangueane
Citation: Imai *et al.*, Bioinformation 2(9): 417-421 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Supplementary material

Equations

$$\left\langle X_p \right\rangle_k = \sum_{i \in segment} X_p(i) \Big/ l \qquad \rightarrow \qquad (1)$$

where $\left\langle X_p \right\rangle_k$ is the average of *p*-th parameter in *k*-th segment, and *l* is length of segment.

$$Z_p^{(j)} = \sum_{k=1}^{n} \left(\left\langle X_p \right\rangle_k - \overline{\left\langle N_p \right\rangle_k} \right) \times \left(\overline{\left\langle P_p \right\rangle_k} - \overline{\left\langle N_p \right\rangle_k} \right) \qquad \rightarrow \qquad (2)$$

where $\overline{\left\langle P_p \right\rangle_k}$ and $\overline{\left\langle N_p \right\rangle_k}$ are the average values of $\left\langle X_p \right\rangle_k$ for all of positive data and negative data, respectively. Finally, we carry out the canonical discriminant analysis between positive and negative data by using $Z_p^{(j)}$ and then obtained the discrimination score.

	Training data		Test data	
Localization	**Precision**	**Recall**	**Precision**	**Recall**
EC	94.5%	84.5%	92.3%	70.3%
OM	93.7%	92.2%	89.4%	87.5%
P	93.1%	89.4%	86.4%	76.0%
IM	96.1%	93.4%	97.5%	97.5%
C	99.8%	92.7%	93.5%	88.4%
Overall	96.7%	91.5%	92.9%	86.7%

Table 1a: The Precision and recall of SOSUI-GramN for training and test data.

	$\text{SOSUI}_{\text{GramN}}$		Psortb v2.0		CELLO v2.5		PSLpred		P-CLASSIFIER	
Localization	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
EC	90.0%	50.0%	66.7%	33.3%	40.0%	44.4%	47.4%	50.0%	31.6%	33.3%
OM	92.1%	92.1%	100%	78.9%	67.7%	55.3%	87.5%	55.3%	60.6%	52.6%
P	77.3%	58.6%	100%	62.1%	50.0%	75.9%	53.7%	75.9%	47.5%	65.5%
IM	96.2%	72.5%	96.5%	79.7%	95.6%	62.3%	85.7%	69.6%	95.3%	59.4%
C	87.8%	89.7%	99.1%	75.9%	84.9%	93.1%	84.3%	92.4%	82.3%	93.1%
Overall	89.3%	80.6%	97.3%	73.2%	76.6%	76.6%	78.3%	78.3%	73.9%	73.9%

Table 1b: Performance comparison between SOSUI-GramN and other systems using 299 proteins for calibration **[7]**.