BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 19 2006, pages 2441–2443
doi:10.1093/bioinformatics/btl409

Sequence analysis

JVirGel 2.0: computational prediction of proteomes separated via two-dimensional gel electrophoresis under consideration of membrane and secreted proteins

Karsten Hiller[1], Andreas Grote[1,2], Matthias Maneck[3], Richard Münch[1] and Dieter Jahn[1,*]

[1]Institut für Mikrobiologie, Technische Universität Braunschweig, Spielmannstrasse 7, D-38106 Braunschweig, Germany, [2]Institut für Bioverfahrenstechnik, Technische Universität Braunschweig, Gaußstrasse 17, D-38106 Braunschweig, Germany and [3]Fachbereich Mathematik und Informatik, Freie Universität Berlin, Arnimallee 14, D-14195 Berlin, Germany

Received on April 18, 2006; revised on July 14, 2006; accepted on July 22, 2006
Advance Access publication July 26, 2006
Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: After the publication of JVirGel 1.0 in 2003 we got many requests and suggestions from the proteomics community to further improve the performance of the software and to add additional useful new features.

Results: The integration of the PrediSi algorithm for the prediction of signal peptides for the Sec-dependent protein export into JVirGel 2.0 allows the exclusion of most exported preproteins from calculated proteomic maps and provides the basis for the calculation of Sec-based secretomes. A tool for the identification of transmembrane helices carrying proteins (JCaMelix) and the prediction of the corresponding membrane proteome was added. Finally, in order to directly compare experimental and calculated proteome data, a function to overlay and evaluate predicted and experimental two-dimensional gels was included.

Availability: JVirGel 2.0 is freely available as precompiled package for the installation on Windows or Linux operating systems. Furthermore, there is a completely platform-independent Java version available for download. Additionally, we provide a Java Server Pages based version of JVirGel 2.0 which can be operated in nearly all web browsers. All versions are accessible at http://www.jvirgel.de.

Contact: d.jahn@tu-bs.de

1 INTRODUCTION

Proteomics techniques based on two-dimensional (2D) gel electrophoresis are capable of simultaneously separating more than a thousand proteins from a single organism (Görg *et al*., 2000). Currently, only a limited number of tools for 2D gel prediction including Gelbank (Babnigg and Giometti, 2004), Virtual2D (Medjahed *et al*., 2003), 2D-PAGE (Pleissner *et al*., 2004) and ProteomeWeb (Babnigg and Giometti, 2003) are available. Charged signal peptides that guide a protein through the secretion apparatus in the cytoplasmic membrane are cleaved off after translocation of the preprotein (Mori and Ito, 2001). Thus, the isoelectric point of the preprotein deduced from the DNA sequence of the corresponding gene and the mature protein often differ significantly. More than 20% of all proteins are predicted to contain at least one transmembrane helix in their membrane spanning region (MSR). Their hydrophobic character usually prohibits their separation using the conventional electrophoretic techniques. Therefore, these proteins are not visible on standard 2D gels. Consequently, the elimination of obvious membrane proteins in combination with the removal of signal sequences from exported proteins should significantly improve the confidence of a calculated proteome map. In the same context the prediction of Sec-dependent secretomes and membrane proteomes is possible. Furthermore, the relationship of experimental and calculated 2D gels is essential for the practical use of the program. However, the sophisticated statistical problem of matching experimental with calculated protein positions had to be solved.

2 NEW FEATURES OF JVirGel 2.0

The first version of JVirGel (Hiller *et al*., 2003) has drawn large attention across the proteomics community. We recognize more than 1500 different researchers using our web page per month. Therefore, we decided to further improve and extend our software. The new JVirGel 2.0 extends the features of the previous version by the following components:

(1) *Visualization of the secretome and membrane proteome*: Owing to the implementation of new algorithms for the *ab initio* prediction of Sec-dependent signal peptides (PrediSi) (Hiller *et al*., 2004) and α-helical membrane helices (JCaMelix), the visualization and analysis of whole secretomes and membrane subproteomes is facilitated (Fig. 1). The prediction algorithm of JCaMelix is based on a profile hidden Markov model. For parameter determination of the model (transition and emission probabilities) we used maximum-likelihood training (Baum–Welch algorithm) (Durbin *et al*., 1998). In order to test the accuracy of our prediction model we used the publicly available program TMH benchmark (Kernsytsky and Rost, 2003). TMH benchmark offers the possibility to evaluate new programs in comparison with established tools for the prediction of transmembrane helices. The result of a TMH benchmark analysis revealed that the performance of our new package is equivalent to that of existing methods. The

[*]To whom correspondence should be addressed.

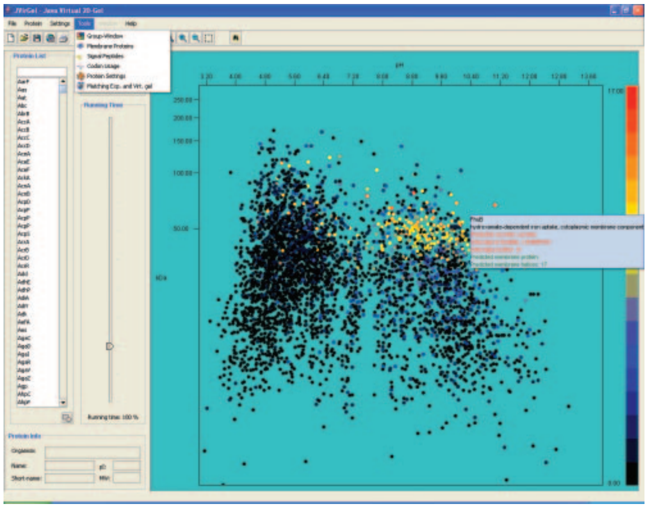

Fig. 1. Screenshot of JVirGel 2.0 standalone version visualizing the predicted proteome of *Shigella flexneri*. The virtual protein spots are colored as a function of their content of α-helical transmembrane helices. The predicted membrane subproteome consists of 1323 proteins containing at least one predicted α-helical transmembrane helix.

percentage of proteins for which all membrane helices were predicted correctly was 77% for JCaMelix, 83% for HMMTOP2 (Tusnady and Simon, 2001), 79% for DAS (Cserzö *et al*., 1997) 71% for TMHMM (Sonnhammer *et al*., 1998) and 65% for Kyte–Doolittle (Kyte and Doolittle, 1982) using the high-resolution dataset.

(2) *Optimization of calculated proteomes via the removal of signal peptides from exported proteins and the exclusion of potential membrane proteins*: Identified Sec-dependent signal peptides are removed in silico and mature proteins are included in the proteome map. The predicted membrane proteomes can also be removed from the calculated proteome map. The described extensions of JVirGel 2.0 significantly improved the prediction accuracy of our program. In order to investigate the effect of signal peptide removal on the accuracy of predicted pIs, the calculated values with and without considering the presence of signal peptides were compared with the experimental data of a Shigella flexneri extracellular proteome (Liao *et al*., 2003) (Fig. 2). The accuracy of the pI calculation method was determined by the calculation of the size of the 95% confidence interval of the data. This interval describes the maximal statistical deviation of our pI prediction in pH units expected for 95% of the predictions. The 95% confidence interval decreased from ±0.44 pH units to ±0.21 pH units at pH 5.5 if the signal peptides were taken into account. Further improvements could be achieved by the exclusion of obvious integral membrane proteins from the visualization. Owing to their hydrophobic character these proteins are not accessible to routinely employed 2D gel electrophoretic techniques. In the case of the analyzed extracellular proteome, JVirGel 2.0 was able to detect 133 potential membrane proteins with two or more α-helical transmembrane helices located in the corresponding area of the 2D gel presented by Liao *et al*. (2003) (pH 4.25–6.55, MW 11–80 kDa).

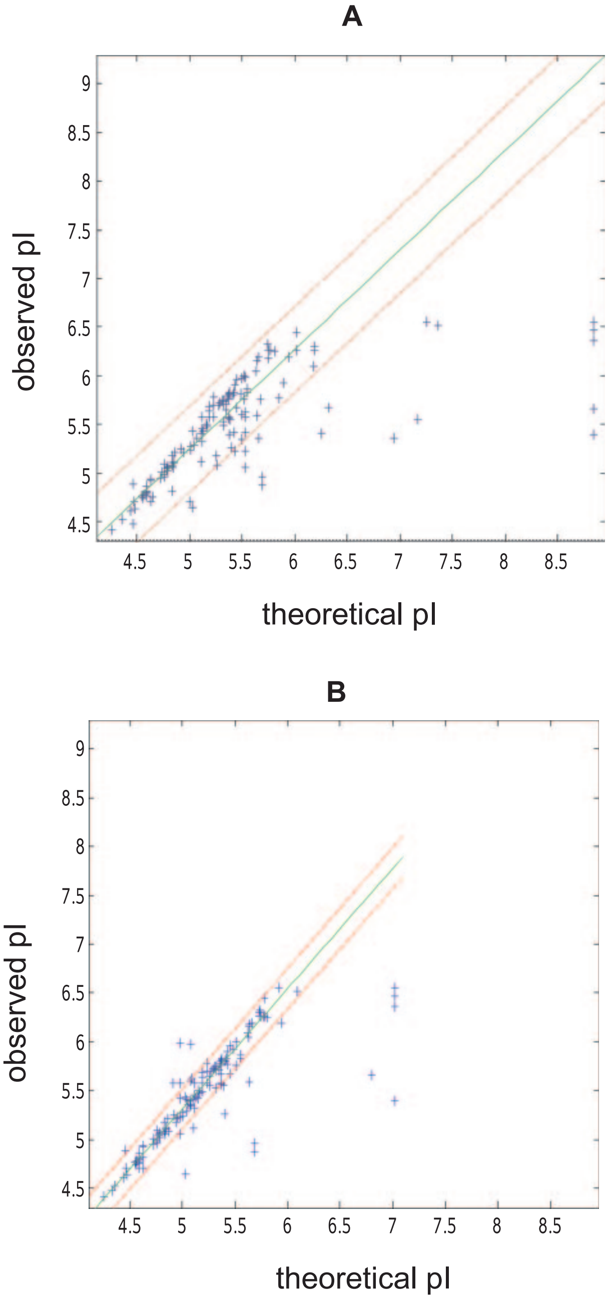

Fig. 2. Linear regression between experimentally determined and calculated isoelectric points of all identified proteins of the extracellular proteome of *S.flexneri* (Liao *et al*., 2003). The regression line is shown in green, the 95% confidence intervals are indicated with red lines. (**A**) The results without taking the influence of signal peptides into consideration. In (**B**), potential signal peptides were identified and removed *in silico*. Only the mature parts of the proteins were used for the calculation of the pI. A total of 139 experimentally determined pIs for extracellularly collected proteins of *S.flexneri* were compared with their calculated counterparts. The correlation coefficient for the pIs of the preproteins was 0.61 and was improved by the *in silico* generation of mature proteins to 0.77. The 95% confidence interval at pH 5.5 was ±0.44 pH units for the preproteins and was improved to ±0.21 pH units for *in silico* generation of mature proteins. The regression analysis was performed by using MathWorks Matlab 6.1.

However, the integration of these new functions into JVirGel 2.0 exhibits some side effects. Although the used signal peptide and membrane helix prediction algorithms are of a high quality, there always exists the possibility of falsely positive and falsely negative predicted protein features resulting in an incorrect positioning of a few virtual protein spots.

(3) *Overlaying of an experimental with a virtual* 2*D gel*: The integration of a specific mapping algorithm allows JVirGel 2.0 now to map virtual gels with experimental gels. The mapping is performed by using different user selectable regression methods. Most regression methods are able to generate a mathematical relation between the pI/MW values and the running distances of the proteins. For the pI dimension the linear regression and for the MW dimension the logarithmic regression tended to give the best results. However, unpredictable variations caused by deviations during gel casting and subsequent electrophoresis can lead to experimental results which are in some cases better analyzed using other regression methods, e.g. polynomial, exponential or power-law regression. These methods are all provided by JVirGel 2.0. After setting some landmarks virtual protein spots in a region of interest can then be displayed by a simple mouse click.

(4) *Simulation of the electrophoresis in real time*: The user is now able to simulate the electrophoresis time and the pH range of interest in real time without reloading a website.

(5) *Definition of the* pK_a *values of the amino acid side chains*: For example, if alkylation of cysteine residues was performed before isoelectric focusing, the usual acidic contribution of cysteine to the pI of a protein can be switched off *in silico* to improve the quality of the virtual gel. The algorithm used for pI calculation is based on the law of mass action and is described in detail before (Skoog and Wichmann, 1986; Hiller *et al*., 2003).

(6) *Export function*: Produced virtual gel images can be exported as portable network graphics (PNG) or as Joint Photographic Experts Group graphics (JPG). Moreover, predicted protein features such as pI, MW, number of transmembrane helices and probable Sec-dependent cleavage positions can be saved as comma separated values (csv) or Microsoft Excel (xls) files.

(7) *Visualization of the influence of the codon usage bias on protein production*: Although the level of protein abundance depends on many different factors, there exists a significant relationship between protein abundance and codon usage bias especially for low abundance proteins (Ghaemmaghami *et al*., 2003; Futcher *et al*., 1999). Based on the established JCat algorithm (Grote *et al*., 2005) the specific codon adaptation indices are calculated and visualized accordingly.

3 SOFTWARE

JVirGel 2.0 was written in Java and exists in two versions: (1) The standalone version for installation on local computers offers all described functions for Windows XP and Linux operating systems (http://www.jvirgel.de/download.jsp). (2) The completely renewed online version based on dynamic JSP web pages can be used on every system with an active web browser. It is able to import FASTA or EMBL data or to use one of the 210 precalculated prokaryotic proteomes. Although the use of HTML pages *per se* is very limited, this version is able to visualize the whole proteome, the secretome and the membrane subproteome. All versions of JVirGel 2.0 are free of charge and can be used without restriction by everyone.

ACKNOWLEDGEMENTS

We thank Dr Barbara Schulz for critical reading of the manuscript. This work was funded by the German Federal Ministry of Education and Research (BMBF) (Grant no. 031U110A/031U210A).

Conflict of Interest: none declared.

REFERENCES

Babnigg,G. and Giometti,C.S. (2003) ProteomeWeb: a web-based interface for the display and interrogation of proteomes. *Proteomics*, **3**, 584–600.

Babnigg,G. and Giometti,C.S. (2004) GELBANK: a database of annotated two-dimensional gel electrophoresis patterns of biological systems with completed genomes. *Nucleic Acids Res*., **32**, D582–D585.

Cserzö,M. *et al*. (1997) Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng*., **10**, 673–676.

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Futcher,B. *et al*. (1999) A sampling of the yeast proteome. *Mol. Cell. Biol*., **19**, 7357–7368.

Ghaemmaghami,S. *et al*. (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.

Görg,A. *et al*. (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, **21**, 1037–1053.

Grote,A. (2005) JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res*., **33**, W526–W531.

Hiller,K. *et al*. (2003) JVirGel: calculation of virtual two-dimensional protein gels. *Nucleic Acids Res*., **31**, 3862–3865.

Hiller,K. *et al*. (2004) PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res*., **32**, W375–W379.

Kernytsky,A. and Rost,B. (2003) Static benchmarking of membrane helix predictions. *Nucleic Acids Res*., **31**, 3642–3644.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol*., **157**, 105–132.

Liao,X. *et al*. (2003) A two-dimensional proteome map of *Shigella flexneri*. *Electrophoresis*, **24**, 2864–2882.

Medjahed,D. *et al*. (2003) VIRTUAL2D: A web-accessible predictive database for proteomics analysis. *Proteomics*, **3**, 129–138.

Mori,H. and Ito,K. (2001) The Sec protein-translocation pathway. *Trends Microbiol*., **9**, 494–500.

Pleissner,K.-P. *et al*. (2004) Web-accessible proteome databases for microbial research. *Proteomics*, **4**, 1305–1313.

Skoog,B. and Wichmann,A. (1986) Calculation of the isoelectric points of polypeptides from the amino acid composition. *Trends Anal. Chem*., **5**, 82–83.

Sonnhammer,E.L. *et al*. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **6**, 175–182.

Tusnady,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.