Nucleic Acids Research, 2005, Vol. 33, Database issue ***D583–D587***
doi:10.1093/nar/gki101

DynaProt 2D: an advanced proteomic database for dynamic online access to proteomes and two-dimensional electrophoresis gels

Oliver Drews and Angelika Görg*

Technische Universität München, FG Proteomik, Am Forum 2, D-85350 Freising-Weihenstephan, Germany

Received August 12, 2004; Revised and Accepted October 14, 2004

ABSTRACT

DynaProt 2D presents an advanced online database for dynamic access to proteomes and two-dimensional (2D) gels. The database was designed to administer complete *in silico* proteomes and links them with experimental proteomic data in the manner of 2D electrophoresis gels (IPG-Dalt). The 2D gels serve as reference maps in 2D gel analysis as well as tools for navigation of the database to switch between experimental and predicted data. Therefore, all identified spots in the gels are clickable and linked with summarized protein information. The protein information tables contain calculated characteristics, which are often used in proteomics, such as the molecular weight, isoelectric point, codon adaptation index, grand average of hydropathicity, etc. The design of the database permits online extension of gel data and protein attributes without knowledge of any software language. Besides navigation via 2D gels, the clear graphical user interface permits quick and intuitive searching throughout complete proteomes and supports, e.g. the search for proteins with isoelectric points within pH ranges of interest or protein classes (e.g. ribosomal proteins or transporters). The first organism implemented in the database is *Lactococcus lactis*. The database is available at www.wzw.tum.de/proteomik/lactis.

INTRODUCTION

The analysis of several thousand proteins at the same time in one experiment presents an immense task in two-dimensional (2D) electrophoresis. For this reason, several software tools for the analysis and presentation of scanned 2D gels were developed. Gel analysis is performed by programs such as the ImageMaster 2D, Melanie, PDQuest or the recently developed DeCyder, which was especially designed for analyzing multiple samples in one gel/DIGE experiments (1–3). These software tools comprise complex algorithms for spot detection and quantification.

The problem of presentation of 2D gels in the manner of reference gels is solved by the construction of databases, in which protein information is correlated to marked spots in gels. Online solutions range from databases, which cover multiple species such as the SWISS-2DPAGE database (4), to organism or even subproteome specific databases such as the database of the alkaline proteome of *Saccharomyces cerevisiae* (5). The majority of existing databases for 2D electrophoresis is enlisted in the World-2DPAGE at http://www.expasy.org/ch2d/2d-index.html. The structure of some of these databases is based on static Internet pages programmed in Hypertext Markup Language (HTML), which provide clickable reference gels and brief information about the identified spots. Each spot is linked to one HTML page, which contains the spot information. In general, these pages do not support search functions, and in case of extending the spot information, each HTML page must be edited. Dynamic online solutions of relational database systems connect spot maps of reference gels to database fields and generate web pages containing the spot information in real time by accessing relevant fields. In case of a database focused on pathogenic microorganism, e.g. the database fields provide protein information such as predicted isoelectric point (pI) and molecular weight (Mr), several protein identifiers, etc. (6). Relational database systems are extensible in batch, by the addition of new database fields and transferring of the new information for all proteins to the fields in one step. This option is comparable to the addition of a new column in a table, which contains a protein list in rows, and copying protein information for all proteins in one step to the new column. Furthermore, relational database systems enable extensive search functions according to several database fields. Thus, even complex requests, e.g. for all identified proteins in a particular pH or Mr range, are supported.

*To whom correspondence should be addressed. Tel: +49 8161 714265; Fax: +49 8161 714264; Email: Angelika.Gorg@wzw.tum.de

Besides experimental proteome data, complete theoretical proteomes can be retrieved from databases such as the Proteome Analysis Database (PAD) (7) or generated by tools, which calculate theoretical proteome maps (8). Integrated into 2D gel databases, complete theoretical proteomes provide a powerful tool when it comes to the addition of new experimental data. If newly identified spots are later added to the database, they can simply be linked to the already included, corresponding theoretical data. Thus, necessary repetitive input of data, which is time consuming and error prone, is avoided. Furthermore, a database including complete theoretical proteomes is also valuable for scientists interested in not already identified proteins. In this case, a comprehensive search algorithm easily distinguishes between identified and not identified proteins.

DATABASE STRUCTURE AND CONTENTS

The online database was realized with the relational database management system MySQL (http://www.mysql.com/) on an Apache web server with Linux platform. The database mainly consists of three tables: the protein information, the gel data and the spot coordinates. The protein table currently contains the 2266 annotated protein sequences included in the database of *Lactococcus lactis* IL1403 curated by the NCBI (accession: NC_002662) (9), as well as the predicted Mr and pI, the codon adaptation index (CAI) and the grand average of hydropathicity (GRAVY) value of each protein. Mr and pI were calculated in batch for all proteins by using the pI/MW Prediction Tool with default settings for pK values (http://proteome.ibi.unicamp.br/tools/pimw/index.htm). The CAI was generated by application of the software CodonW (10) as described previously (11). The same software was applied for the calculation of the GRAVY value for each protein. Furthermore, the functional classification according to the MOLOKO website (http://spock.jouy.inra.fr/RL000801.html) (12) and the predicted cellular localization calculated with the PSORT algorithm (13) were added to the protein table. Published data (14) of the alkaline reference maps and the corresponding spot coordinates on the gels were integrated into the remaining two tables. Two identifiers linked to NCBIs protein database (9) and Swiss-Prot (15) are also part of the stored data for easily accessible cross references.

Access to the online MySQL database was realized by scripts written in the PreHypertextProcessor language PHP (Figure 1). The scripts are interpreted on the web server and serve as interface between user and database. Even tools for the administration of the database were written in PHP scripts. Thus, besides an Internet browser, no further software tools are necessary to access or administer the database. For example, the tabulated protein information in Figure 2 is the result of a PHP-script and works like an empty table, which is filled with information on demand by the user. After submission of the search criteria, a PHP-script displays the results in the resultset in the lower left-hand side of the Internet browser. By activating the link of one of these results (demand), another PHP-script retrieves the information for this protein from the MySQL database in real-time and each field of the previously empty spot summary table is consecutively filled within parts of a second. In comparison with static HTML pages, in this database solution, additional information, such as reference maps or identified spots can be added easily to the database instead of updating every HTML page of, for instance, all 2266 lactococcal proteins, which are currently part of the database. Even completely new protein characteristics can be added to the spot summary simply by uploading the additional information into reserved fields of the database. Then, slight modifications in the PHP-script code are necessary to display the added data or extend the search options.

For the realization of dynamically displayed circles in spot size on the reference gels, the GD library was used (http://www.boutell.com/gd/). Javascripts were used for several functions, such as the realization of pop-up windows, alerts or the mouse-over information display in the reference maps. Therefore, full functionality is only provided if Javascript is enabled.

NAVIGATION OF THE DATABASE

The interface of DynaProt 2D consists of three parts: the search including the resultset, the interactive reference maps and the summarized protein information table. The search and resultset

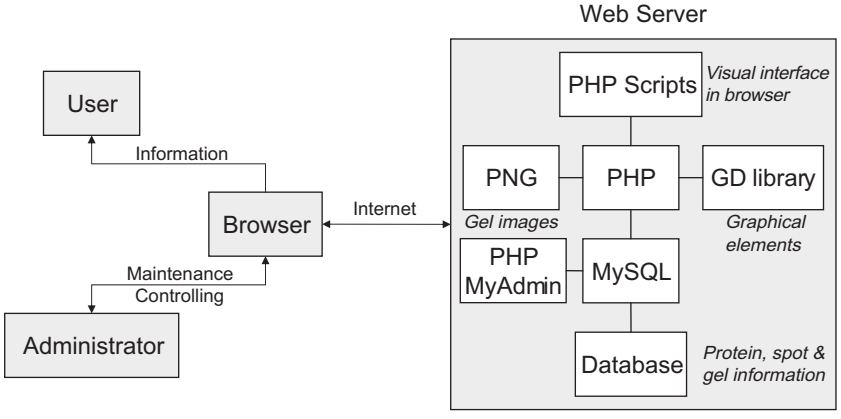

Figure 1. The database architecture on the web server makes it independent from decentralized software. Frequently administered data like gel or experimental information are controlled by embedded PHP scripts. Protein characteristics are in general only added once, and are therefore controlled by PHPMyAdmin. Both administrative tools are accessed by the Internet browser.

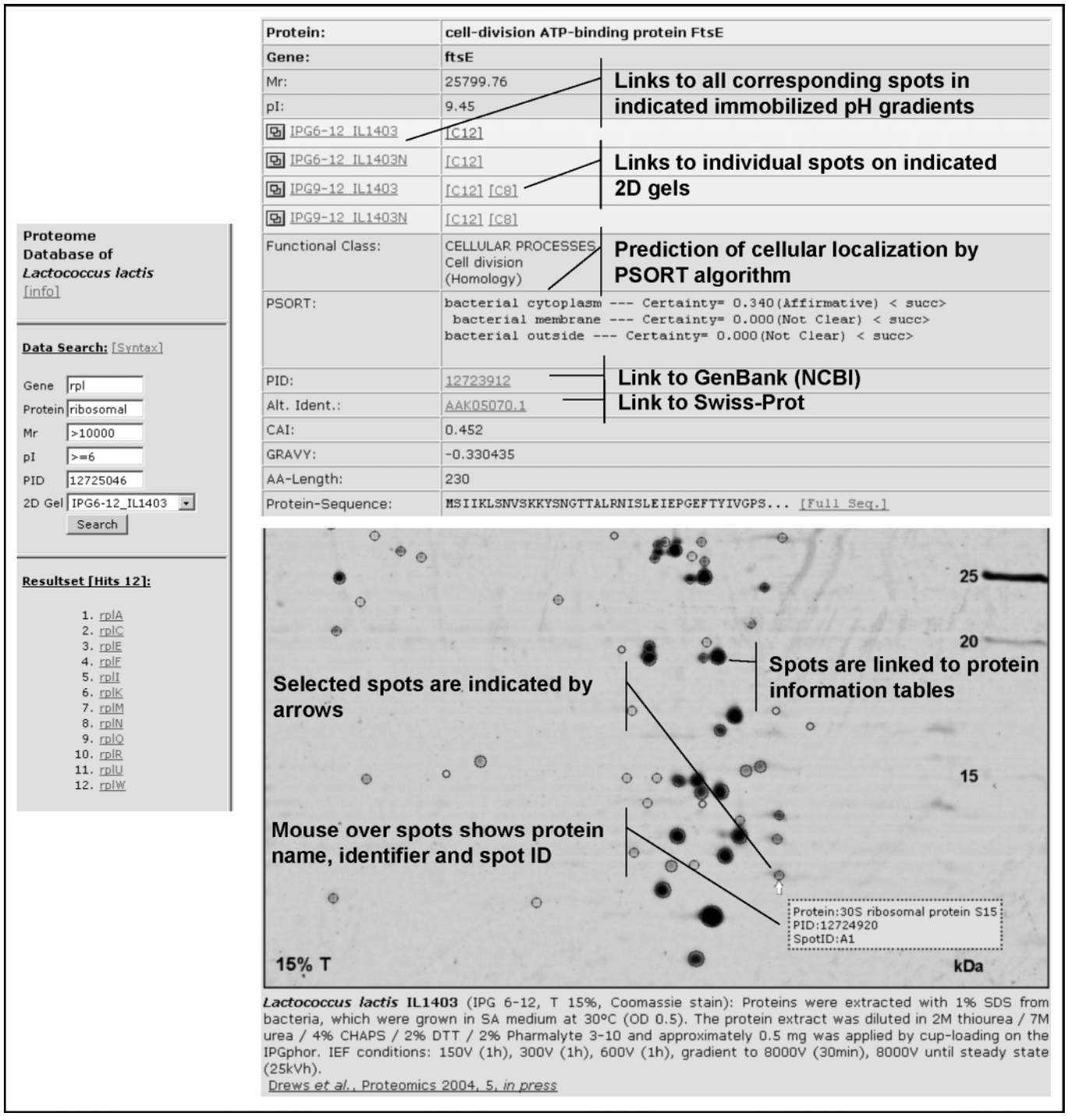

Figure 2. The database interface comprises three parts: the search and the resultset on the left-hand side, and the protein information table or the interactive reference maps on the right-hand side. The image displays arranged screen shots to demonstrate all major parts of the database.

serve as main interface for navigation and thus, are permanently available on the left side of the screen (Figure 2). The interactive reference maps and the summarized protein information table both comprise manifold information. Therefore, the user can intuitively switch between these two pages, which make up most of the screen on the right-hand side of the Internet browser (Figure 2).

The search through the database is dominated either by six different parameters or by interactive exploration of the reference maps. As parameters, full or partial names of genes, proteins or identifiers are supported, as well as ranges of Mr and pI. Furthermore, the search can be limited to a selected 2D reference map too. After submission of the search request, all proteins, which fit the parameters, are indicated in the

resultset. In contrast, the interactive reference maps present all identified proteins on a 2D gel at one sight. By positioning the mouse over the red encircled spots, information about the spots, namely protein name, identifier and spot ID, are displayed. On clicking, the summarized protein information of the selected spot is presented in the Internet browser. Below the reference gel, detailed information about growth conditions, protein extraction, IPG-DALT conditions and references are provided (Figure 2).

The summarized protein information provides the full-length protein name, gene identifier, predicted pI and Mr as well as the functional classification according to the annotation on the MOLOKO website (http://spock.jouy.inra.fr/RL000801.html) (12), and the predicted cellular localization calculated with the PSORT algorithm (13). Links to individual or all spots of the selected protein are implemented to switch from protein information to highlighted spots on reference gels. Furthermore, two widely used identifiers link to the corresponding GenBank (9) and Swiss-Prot (15) database entries. The codon adaptation index and the grand average of hydropathicity value give information about expected expression (16) and solubility of the protein (17). Finally, the complete protein sequence can be retrieved for further analyses and calculations.

The sophisticated navigation and management by the embedded PHP scripts (Figure 1) makes the database on the web server independent of any decentralized software. Thus, the database can be accessed and maintained from every computer connected to the Internet, simply by using a web browser. Protein information is administered online by PHPMyAdmin. Gel images are uploaded in the PNG file format. The addition of gels, experimental data and spot coordinates to the database is supported by an administrative PHP script, which can be accessed only by collaborating laboratories, e.g. mass spectrometry service stations or other registered users. This solution was chosen, because the latter data are more frequently added or changed. The administrative PHP script resembles a form in which the data and spot coordinates are copied. By submission of the form, the newly added data are immediately accessible online.

ADVANTAGES OF DYNAPROT 2D

The benefit of a proteome database comprising 2D reference maps is versatile. On one hand, the summarized protein information itself is useful for several types of proteomic applications, such as quick access to the isolelectric point of one or several proteins to choose a suitable buffer or immobilized pH gradient, or determination of a subset of proteins with a particular GRAVY value to estimate their solubility. In proteomics, characteristic values like the CAI or the GRAVY, e.g. are compared and analyzed to estimate the part of a proteome, which probably can be covered by a particular methodical approach (14,18,19). Therefore, these and other often used protein characteristics were computed for the complete theoretical proteome and implemented into the presented database.

The reference maps, on the other hand, are informative on various levels too. They deliver evidence for the expression of the identified proteins at the chosen growth condition. Furthermore, the deviation of the predicted Mr and/or pI indicates protein processing or post-translational modification, in particular if one protein is represented by more than one spot. The spot intensity in comparison to other spots indicates the relative abundance of the protein in the cells. Unlike in other 2D databases, identified spots are marked by circles instead of crosses on gel images (Figure 2). Crosses cover up small spots or spots with low intensity and marked spots are barely visible. Increasing the contrast in general results in merging big spots. Circles leave the spots visible and indicate the spot area.

For following analysis at the proteome level, the reference maps give an overview, which proteins can be expected in the pH gradient and serve as orientation, especially if a particular set of proteins are of interest. The reference maps are no substitute for spot identification by peptide mass finger printing or protein sequencing and shall not serve as single source for protein information, but may support insignificant spot identifications obtained, e.g. by poor sequence coverage.

In comparison to other 2D databases, no gels without identified spots were included in the database and the search for proteins of interest can be restricted to reference gels, giving the user the advantage to know instantly, which protein can be found on what gel. *In silico* based 2D patterns were not included in the database, because only a part of the theoretical proteome is constitutively expressed and protein abundance is very different in cells (19). For the selection of appropriate IPG strips or acrylamide content, proteins can be listed according to their $_{P}$I and Mr. Finally, the database is simply navigated by the search and the resultset and results are interactively linked. Thus, the user needs only minimal time for acquiring the scope of the database and is not confused by multiple submenus. Furthermore, the precomputed protein characteristics instantly provide valuable information frequently used in proteomics and not listed in Swiss-Prot or GenBank (9,15). For quick access to data listed in the latter two databases, links were provided.

One further aspect is covered by the presented proteome database: progress report or even quality assurance. Once the theoretical proteome of an organism is implemented, the database makes it easy to document 2D gels. The spots simply need to be marked and named with the corresponding protein identifier, which is in general indicated by each software used for protein identification. After uploading the gel and the spot coordinates, the experimental data are immediately accessible via the Internet or can be restricted to an Intranet. Thus, collaborating laboratories or, e.g. proteomic service stations can easily share 2D related results even without the demand of experience in certain softwares for gel analysis.

ACKNOWLEDGEMENT

D. Preuss (Daikun Solutions) is sincerely thanked for his help in the realization of dynamic online accessibility of the database.

REFERENCES

1. Alban,A., David,S.O., Bjorkesten,L., Andersson,C., Sloge,E., Lewis,S. and Currie,I. (2003) A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel

electrophoresis incorporating a pooled internal standard. *Proteomics*, **3**, 36–44.
2. Yan,J.X., Devenish,A.T., Wait,R., Stone,T., Lewis,S. and Fowler,S. (2002) Fluorescence two-dimensional difference gel electrophoresis and mass spectrometry based proteomic analysis of *Escherichia coli*. *Proteomics*, **2**, 1682–1698.
3. Gharbi,S., Gaffney,P., Yang,A., Zvelebil,M.J., Cramer,R., Waterfield,M.D. and Timms,J.F. (2002) Evaluation of two-dimensional differential gel electrophoresis for proteomic expression analysis of a model breast cancer cell system. *Mol. Cell. Proteomics*, **1**, 91–98.
4. Hoogland,C., Mostaguir,K., Sanchez,J.C., Hochstrasser,D.F. and Appel,R.D. (2004) SWISS-2DPAGE, ten years later. *Proteomics*, **4**, 2352–2356.
5. Wildgruber,R., Reil,G., Drews,O., Parlar,H. and Gorg,A. (2002) Web-based two-dimensional database of *Saccharomyces cerevisiae* proteins using immobilized pH gradients from pH 6 to pH 12 and matrix-assisted laser desorption/ionization-time of flight mass spectrometry. *Proteomics*, **2**, 727–732.
6. Pleißner,K.P., Eifert,T. and Jungblut,P.R. (2002) A European pathogenic microorganism proteome database: construction and maintenance. *Comp. Funct. Genomics*, **3**, 97–100.
7. Pruess,M., Fleischmann,W., Kanapin,A., Karavidopoulou,Y., Kersey,P., Kriventseva,E., Mittard,V., Mulder,N., Phan,I., Servant,F. *et al*. (2003) The Proteome Analysis database: a tool for the *in silico* analysis of whole proteomes. *Nucleic Acids Res*., **31**, 414–417.
8. Hiller,K., Schobert,M., Hundertmark,C., Jahn,D. and Munch,R. (2003) JVirGel: Calculation of virtual two-dimensional protein gels. *Nucleic Acids Res*., **31**, 3862–3865.
9. Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E. *et al*. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res*., **32**, D35–D40.
10. Peden,J. (1999) Analysis of codon usage. PhD Thesis, Department of Genetics, University of Nottingham, UK.
11. Guillot,A., Gitton,C., Anglade,P. and Mistou,M.Y. (2003) Proteomic analysis of *Lactococcus lactis*, a lactic acid bacterium. *Proteomics*, **3**, 337–354.
12. Bolotin,A., Wincker,P., Mauger,S., Jaillon,O., Malarme,K., Weissenbach,J., Ehrlich,S.D. and Sorokin,A. (2001) The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. lactis IL1403. *Genome Res*., **11**, 731–753.
13. Nakai,K. and Horton,P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci*., **24**, 34–36.
14. Drews,O., Reil,G., Parlar,H. and Gorg,A. (2004) Setting up standards and a reference map for the alkaline proteome of the gram-positive bacterium *Lactococcus lactis*. *Proteomics*, **4**, 1293–1304.
15. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O′Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
16. Sharp,P.M. and Li,W.H. (1987) The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*., **15**, 1281–1295.
17. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol*., **157**, 105–132.
18. Wilkins,M.R., Gasteiger,E., Sanchez,J.C., Bairoch,A. and Hochstrasser,D.F. (1998) Two-dimensional gel electrophoresis for proteome projects: the effects of protein hydrophobicity and copy number. *Electrophoresis*, **19**, 1501–1505.
19. Pedersen,S.K., Harry,J.L., Sebastian,L., Baker,J., Traini,M.D., McCarthy,J.T., Manoharan,A., Wilkins,M.R., Gooley,A.A., Righetti,P.G. *et al*. (2003) Unseen proteome: mining below the tip of the iceberg to find low abundance and membrane proteins. *J. Proteome Res*., **2**, 303–311.