Journal of Proteomics & Bioinformatics - Open Access

Research Article JPB/Vol.2/April 2009

GPDE: A Biological View on PRIDE

Johannes Griss[1,2], Christopher Gerner[1]*

[1]Department of Medicine I, Medical University of Vienna, Vienna, Austria
[2] Griss Software Development, Vienna, Austria

*Corresponding author: Christopher Gerner, Department of Medicine I, Institute of Cancer Research, Medical University of Vienna, Austria, E-mail: Christopher.gerner@meduniwien.ac.at; Tel: +43-1-4277-65230

Received March 02, 2009; Accepted March 31, 2009; Published April 01, 2009

Citation: Griss J, Gerner C (2009) GPDE: A Biological View on PRIDE. J Proteomics Bioinform 2: 167-174. doi: 10.4172/jpb.1000074

Copyright: © 2009 Griss J, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Clinical proteomics are unthinkable without the extensive use of meta-analysis. By the development of the Proteomics Identification Database Engine (PRIDE) extensive markup language (XML) data format in 2005 a standardized data format for proteomics data publishing was created. Here we present the Griss Proteomics Database Engine (GPDE, http://www.griss.co.at/?GPDE): an open source web-based bioinformatic tool that uses the possibilities of this standardized data format to create biologically-based meta-analysis over an unlimited number of proteomics experiments. Its aims are (1) to be fully compatible with the PRIDE XML data schema, (2) to integrate the data of different experiments or projects based on specially defined "cell types" thus dissolving the "borders" between different experiments and (3) to provide an intuitive web interface to access data otherwise only accessible via complex queries. The GPDE is free software and available for download under the terms of the Affero General Public License (AGPL, http://www.fsf.org/licensing/licenses/agpl-3.0.html). The GPDE is currently used by the Clinical Proteomics Laboratories at the Medical University of Vienna (CPL/MUW) (available at http://www.meduniwien.ac.at/proteomics/database). There the GPDE is used in three different ways: 1) as an *in silico* alternative to multiple Western analyses; 2) to provide easy access to cellular proteome reference maps and 3) to reliably support the assessment of the specificity of biomarker candidates.

Keywords: Database; Meta-analysis; Biomarker discovery; Proteomics; PRIDE; GPDE

Abbreviations

AGPL Affero General Public License

AJAX asynchronous JavaScript and XML

CPL/MUW Clinical Proteomics Laboratories at the Medical University of Vienna cv, controlled vocabulary

GPDE Griss Proteomics Database Engine

PICR service Protein Identifier Cross-Reference Service

PRIDE Proteomics Identification Database Engine

RDBMS relational database management system

XML extensive markup language

Introduction

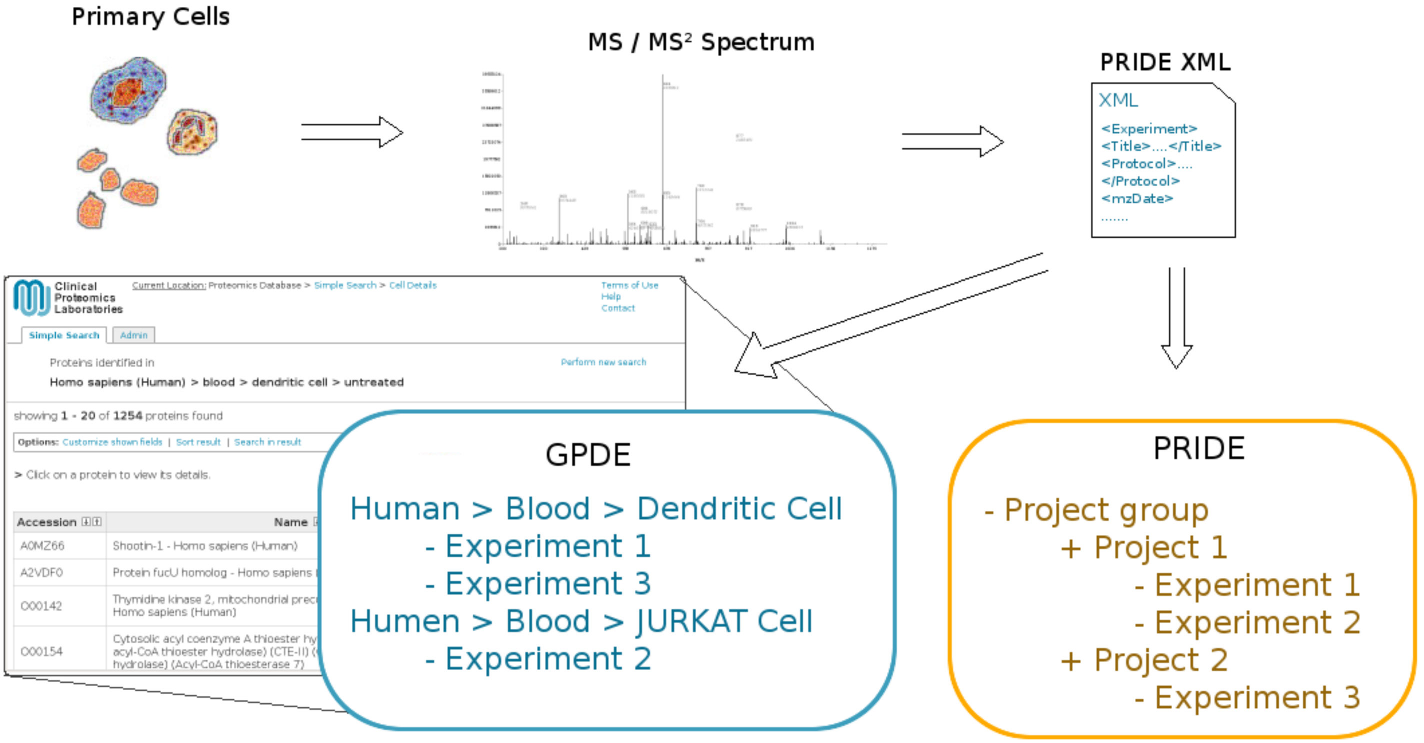
The Proteomics Identifications Database Engine (PRIDE) has revolutionized proteomics publishing (Martens et al., 2005). Compared to other fields of science proteomics data can hardly be condensed. It is not possible to, for example, calculate the "average" of a proteome profile or to give the standard deviation of such results. Thus, up until the launch of public repositories such as PRIDE the main way to present the identified proteins was as a supplementary printed list. Such a list is not only a burden to read but also nearly inaccessible to automated systems. With the development of the PRIDE extensive markup language (XML) data format for proteome analysis data a great alternative to these lists was created.

PRIDE XML allows the researcher to easily publish results in a computer readable format. The grade of detail to which the identifications are presented is very much open to the researcher (Jones et al., 2006). Thereby, meta-analysis in a way that no one could ever have dreamed of before were made possible. Today, three years after the initial presentation of PRIDE the database holds over 9000 experiments with more than 2 million identified proteins (state January 2009). In 2008 a BioMart interface was added to PRIDE enabling more complex queries (Jones et al., 2008). The great power of PRIDE together with this new tool still comes with a user interface that needs some time to get used to. A researcher with neither bioinformatic background nor bioinformatic support might find it hard to use the vast potential that this set of bioinformatic tools provides.

For clinical proteomics the use of meta-analysis across the data of several experiments is vital. Only by pooling the data of several experiments and crosschecking these results with other data sets reliable clinical proteomic results can be produced. The Griss Proteomics Database Engine (GPDE) is designed to simplify this task.

The GPDE may serve as an easily accessible and reliable *in silico* alternative to Western analyses to all biologists. It organizes inserted experiments around so called “cell types” thus creating a “biological” meta-analysis over an unlimited number of single experiments. Furthermore, it is specifically designed to be easy to use so that a researcher without extensive bioinformatic support can use the vast potential that lies within the PRIDE XML standard. From the researcher’s point of view no extra efforts are necessary to use the GPDE as the same files that are submitted to or downloaded from PRIDE can be inserted in the GPDE. Furthermore, the PRIDE development collaboration constantly improves methods to convert proteomics mass spectrometry data from different sources into the PRIDE XML format (for more information see the PRIDE project page at http://www.ebi.ac.uk/pride). Thus, also researcher not using the PRIDE database to publish data will find it increasingly simple to convert their data into the PRIDE XML format.

In short the GPDE has three aims: (1) to be fully compat-


Primary Cells
MS / MS² Spectrum
PRIDE XML
XML
<Experiment>
<Title>....</Title>
<Protocol>....
</Protocol>
<mzDate>
....
Clinical Proteomics Laboratories
Current Location: Proteomics Database > Simple Search > Cell Details
Terms of Use
Help
Contact
Simple Search
Admin
Proteins identified in
Homo sapiens (Human) > blood > dendritic cell > untreated
Perform new search
showing 1 - 20 of 1254 proteins found
Options: Customize shown fields | Sort result | Search in result
> Click on a protein to view its details.
GPDE
Human > Blood > Dendritic Cell
- Experiment 1
- Experiment 3
Humen > Blood > JURKAT Cell
- Experiment 2
PRIDE
- Project group
+ Project 1
- Experiment 1
- Experiment 2
+ Project 2
- Experiment 3

Accession	Name
A0MZ66	Shootin-1 - Homo sapiens (Human)
A2VDF0	Protein fucU homolog - Homo sapiens
O00142	Thymidine kinase 2, mitochondrial prec Homo sapiens (Human)
O00154	Cytosolic acyl coenzyme A thioester hy acyl-CoA thioester hydrolase) (CTE-II) (hydrolase) (Acyl-CoA thioesterase 7)

Figure 1: Flow chart of inserting data into the GPDE. When analyzing primary or cultured cells using mass spectrometry, huge amounts of data are generated. The PRIDE XML data format is a standardized data format for such experiments. The GPDE now uses the possibilities provided by this data format to reorganize the proteomics data around cell types. Thus, a biological view on the different experiments is created.

ible with the PRIDE XML data schema, (2) to integrate the data of different experiments or projects based on specially defined "cell types" thus dissolving the "borders" between different experiments and (3) to provide an intuitive web interface to access data otherwise only accessible via complex queries.

Materials and Methods

The GPDE was developed using the PHP programming language (http://www.php.net) on the server side and the JavaScript programming language for the client side. The database was planned and created using MySQL Workbench 5.0 OSS which was also used to create Figure 2. To enhance the user interface asynchronous JavaScript and XML (AJAX) was employed using the Prototype JavaScript framework by Sam Stephenson version 1.6 (http://www.prototypejs.org/) as well as the script.aculo.us JavaScript libraries by Thomas Fuchs version 1.8 (http://script.aculo.us). As relational database management system (RDBMS) MySQL version 5.1 (http://www.mysql.com) was used.

Results

Data organization by the GPDE

The GPDE is a web based application with a relational database based on the PRIDE XML format (Figure 2). It organizes the proteomics data around so called "cell types",

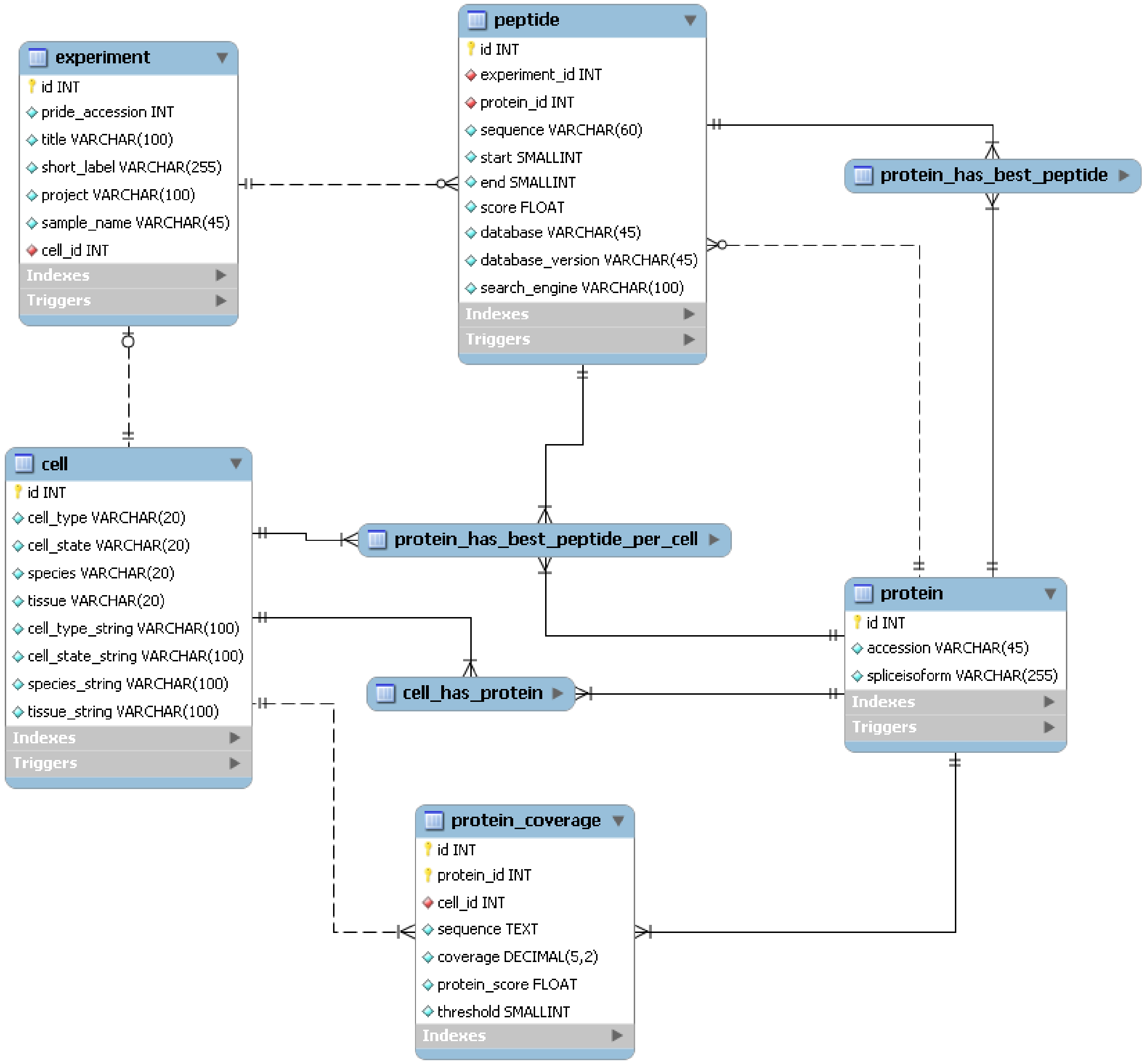


Figure 2: Simplified structure of the GPDE database. For clarity reasons tables not vital to the understanding of the GPDE were omitted. The complete structure of the GPDE database can be downloaded from the project page (http://www.griss.co.at/?GPDE).

thereby creating a completely biologically based view rather than an experimentally based view on the proteomics data (Figure 1).

A “cell type” as defined in the GPDE has four properties: the species, the tissue, the cell type (as defined in the CL ontology) and the “cell state”. The “cell state” defines whether or not a characteristic functional program of the cell was activated. In the database these properties are stored as controlled vocabulary (cv) accessions (Figure 2). The *_string variables hold the corresponding “human readable” description displayed to the user. When a PRIDE XML file is imported into the GPDE database these four properties must be manually set. The PRIDE XML data format has the functionality to include this information with every sample’s data but since this information is optional it is deliberately not evaluated by the GPDE. Furthermore, that way certain pitfalls that may arise when defining a sample in the PRIDE XML format can be avoided. For example, when defining the cell type using the BRENDA ontology rather than the CL ontology a cell type may get interpreted as a type of tissue. Should a user not be aware of these inherent structural features valuable data might not be found by a query and thus be lost to this user.

In addition to cell types the GPDE database is constructed around the identified proteins. Proteins are defined by their accession (Figure 2, table “protein”). Currently the GPDE only supports SwissProt accessions. When importing an experiment into the GPDE the conversion from other accession systems is accomplished using the Protein Identifier Cross-Reference Service (PICR service) at the European Bioinformatics Institute (http://www.ebi.ac.uk/Tools/picr).

Every protein’s details (like for example identified peptide sequences, percent of sequence coverage, etc.) are in fact stored several times in the database (Figure 2, table “protein_coverage”). One time these details are calculated using all of the protein’s peptides in the database, no matter which cell they were identified in. There, the field “cell_id” is set NULL. In case a peptide sequence is found in two different experiments only the one with the higher peptide validation score “counts”. These “best peptides” are stored in the tables “protein_has_best_peptide”, taking all of this protein’s peptides into consideration regardless of the cell they were identified in, and “protein_has_best_peptide_per_cell” (Figure 2). The type of peptide score used is not limited but has to be homogeneous throughout the database (for example only peptides using the SpectrumMill score or only peptides using the Mascot score). Another entry exists for every cell type the protein was identified in. In this cell specific entry only peptides found in this particular cell type are considered. That way already at the level of the database a completely biologically based structure of the data is achieved and the “borders” between the single experiments dissolve.

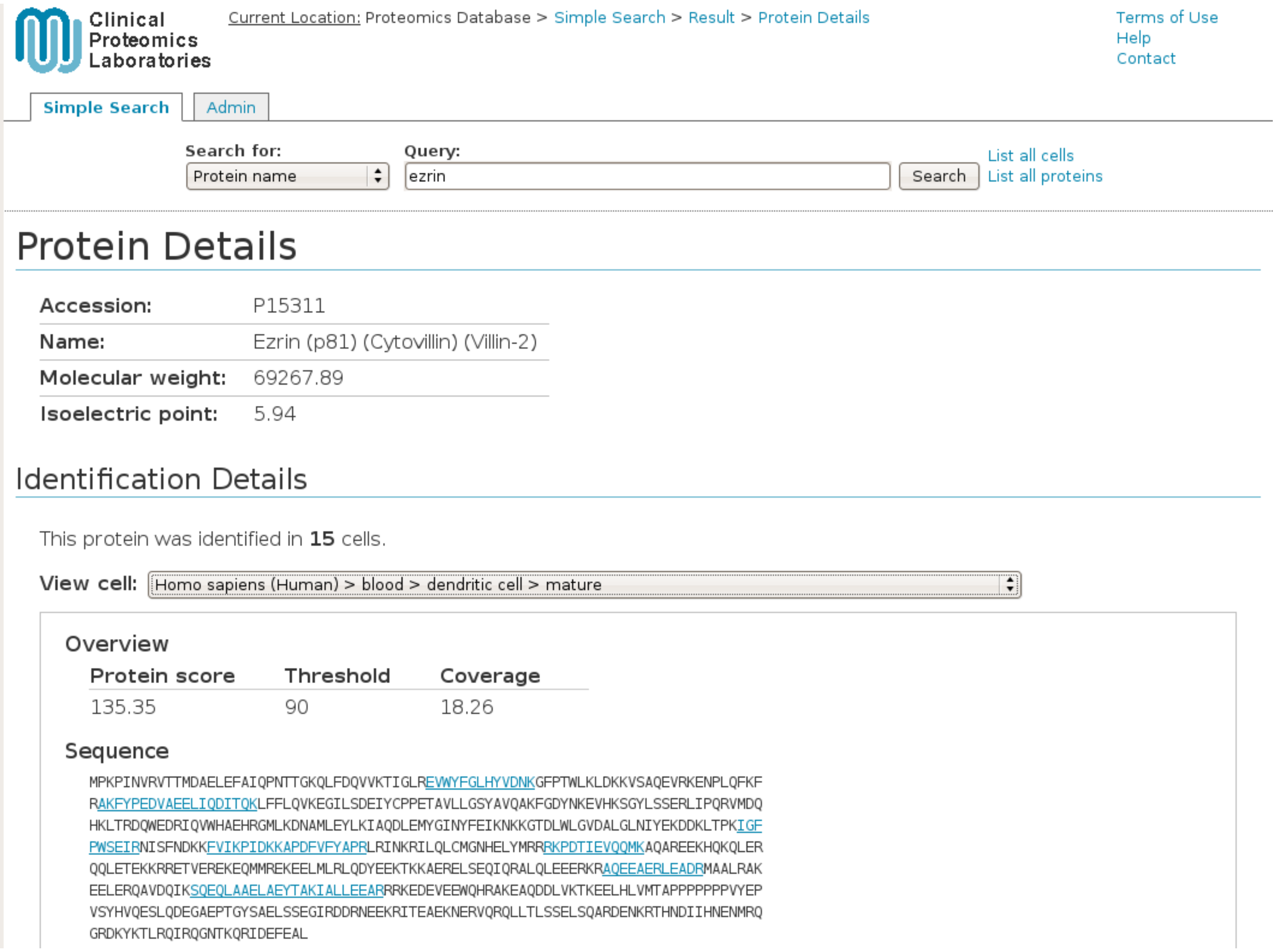
The type of peptide score used throughout the database must be set during the installation of the software. If the user tries to insert an experiment not using this score an error is produced and the operation is cancelled. This restriction of a homogenous peptide score limits the possibility of the GPDE to combine any different experiments from public repositories. Nevertheless, we believe that the possibility to provide the user with reliable information on how “secure” an identification is justifies this restrain. However, we are already working on a way to support multiple concurrent peptide scores.

The following workflow demonstrates how the database works: An experiment on immature dendritic cells containing Ezrin as the single identified protein is imported into a database. This database already contains another experiment on endothelial cells where Ezrin was identified, too. The “global” entry (table “protein_coverage” where “cell_id = NULL” as well as table “protein_has_best_peptide”) for Ezrin will then be updated for the peptides identified in the experiment on dendritic cells. This entry then contains all of Ezrin’s attributes calculated from all peptides in the database – regardless of the cell or the experiment they were identified in. Furthermore, a new entry for “Ezrin in dendritic cells” will be created (table “protein_coverage” with “cell_id = id of immature dendritic cells” as well as several new entries in “protein_has_best_peptide_per_cell”). This entry contains the same attributes as the “global” entry but is only based on the peptides identified in dendritic cells. If, in our example, another experiment on dendritic cells was added both the “global” as well as the (now already existing) “Ezrin in dendritic cells” entry would be updated for the new peptides in this experiment but no new entry would be created.

Accessing data in the GPDE

The user has two main options to view the data of the database: either from the “protein’s point of view” or from the “cell’s point of view”. Proteins can be queried by their accession as well as their name. Currently, only SwissProt accessions are supported at query level. Cells can be queried by species, tissue, cell type or cell state as described above.

After such a query a list of proteins or cells is returned.



Clinical Proteomics Laboratories
Current Location: Proteomics Database > Simple Search > Result > Protein Details
Terms of Use
Help
Contact
Simple Search
Admin
Search for:
Protein name
Query:
ezrin
Search
List all cells
List all proteins

Protein Details

Accession:	P15311
Name:	Ezrin (p81) (Cytovillin) (Villin-2)
Molecular weight:	69267.89
Isoelectric point:	5.94

Identification Details

This protein was identified in **15** cells.

View cell: Homo sapiens (Human) > blood > dendritic cell > mature

Overview

Protein score	Threshold	Coverage
135.35	90	18.26

Sequence

```
MPKPINVRVTTMDAELEFAIQPNTTGKQLFDQVVKTIGLREWYFGLHYVDNKGFPTWLKLDKKVSAQEVRKENPLQFKF
RAKFYPEDVAEELIQDITQKLFFLQVKEGILSDEIYCPPETAVLLGSYAVQAKFGDYNKEVHKSGYLSSERLIPQRVMDQ
HKLTRDQWEDRIQVWHAEHRGMLKDNAMLEYLKIAQDLEMYGINYFEIKNKKGTDLWLGVDALGLNIYEKDDKLTPKIGF
PWSEIRNISFNDKKFVIKPIDKKAPDFVFYAPRLRINKRILQLCMGNHELYMRRRKPDTIEVQQMKAQAREEKHQKQLER
QQLETEKKRRETVEREKEQMMREKEELMLRLQDYEEKTKKAERELSEQIQRALQLEEERKRAQEEAERLEADRMAALRAK
EELERQAVDQIKSQEQLAAELAEYTAKIALLEEARRRKEDEVEEWQHRAKEAQDDLVKTKEELHLVMTAPPPPPPPVYEP
VSYHVQESLQDEGAEPTGYSAELSSEGIRDDRNEEKRITEAEKNERVQRQLLTLSSELSQARDENKRTHNDIIHNENMRQ
GRDKYKTLRQIRQGNTKQRIDEFEAL
```

Figure 3: "**Protein view**". Screen shot of a protein entry in the GPDE.

This set of results is the "gateway" to the actual two "views" of the database:

A protein entry displays all available information about the protein as well as the complete list of peptides identified of this protein (Figure 3). In addition, a list of all cell types in which this protein was identified is shown with the option to select one of these cell types.

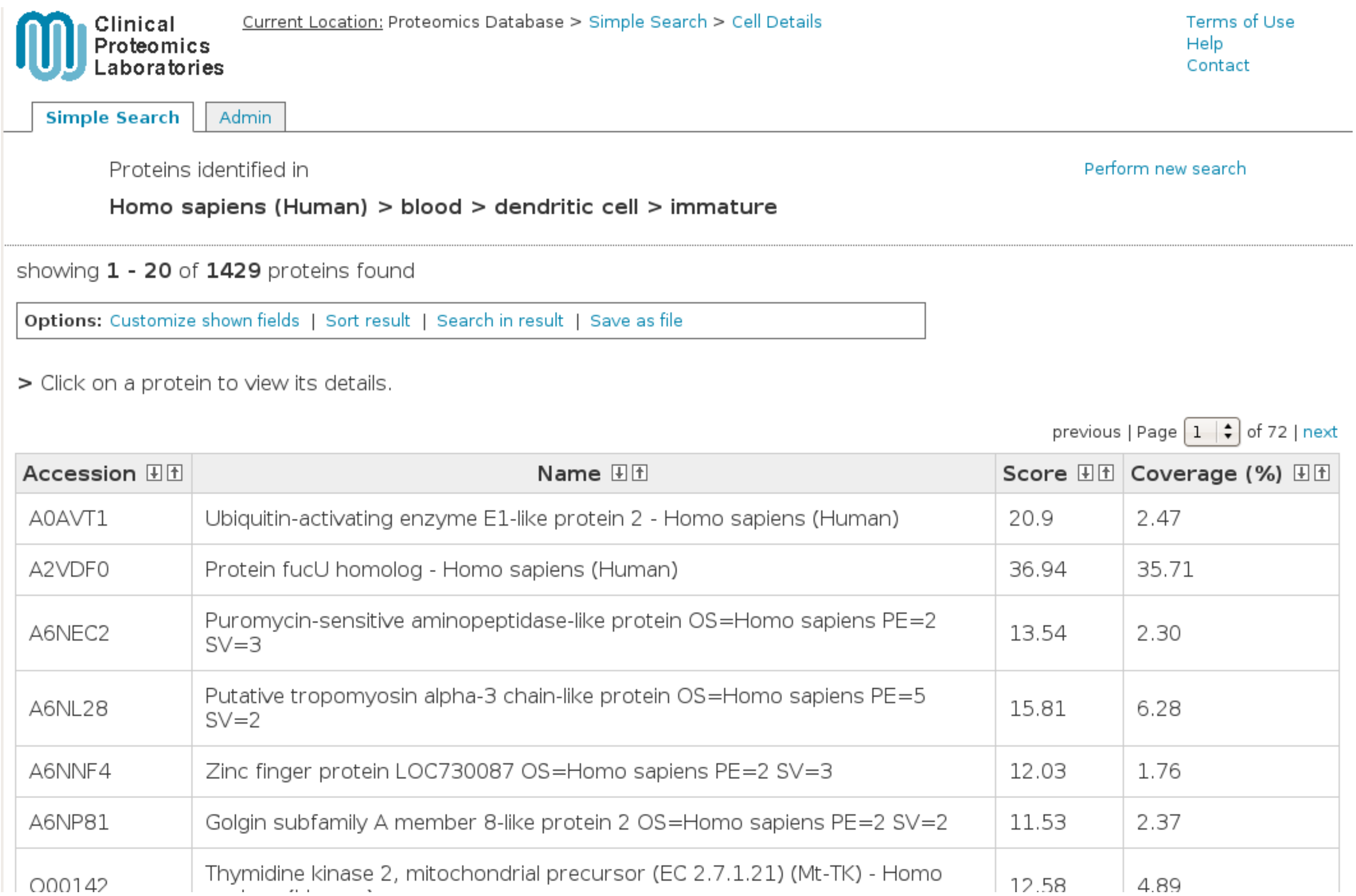
A cell entry shows a list of all proteins identified in this cell (Figure 4). When clicking on a protein, a window opens showing additional information as well as the peptides of this protein, all in respect of this cell type only. Furthermore, the user has the option to filter the results based on any of the available fields and to export the list of identified proteins as a text file.

When using this web based interface different experiments become indistinguishable to the user and the impression of one big result set is given. The experiments which make up this integrated result set are only accessible in the peptide's details.

Discussion

With the development of the PRIDE XML schema a new form of proteomics data aggregation was made possible. The GPDE uses this possibility to combine different proteomics experiments based on their cell types only and blends them together in one big result set. Thus, meta-analysis can easily be generated not only across the "borders" of different experiments but also across the findings of different research teams.

The quality of the data in PRIDE, as in any other public repository, is rather heterogeneous. A meta-analysis over the whole repository would therefore be of indefinable quality rendering the whole meta-analysis questionable. Therefore, the GPDE was deliberately designed to only contain a selected subset of available experiments. To create a biologically based meta-analysis all experiments have to be imported into the GPDE by hand and must be manually


Clinical Proteomics Laboratories
Current Location: Proteomics Database > Simple Search > Cell Details
Terms of Use
Help
Contact
Simple Search
Admin
Perform new search
Options: Customize shown fields | Sort result | Search in result | Save as file
previous | Page 1 of 72 | next

Proteins identified in

Homo sapiens (Human) > blood > dendritic cell > immature

showing **1 - 20** of **1429** proteins found

> Click on a protein to view its details.

Accession	Name	Score	Coverage (%)
A0AVT1	Ubiquitin-activating enzyme E1-like protein 2 - Homo sapiens (Human)	20.9	2.47
A2VDF0	Protein fucU homolog - Homo sapiens (Human)	36.94	35.71
A6NEC2	Puromycin-sensitive aminopeptidase-like protein OS=Homo sapiens PE=2 SV=3	13.54	2.30
A6NL28	Putative tropomyosin alpha-3 chain-like protein OS=Homo sapiens PE=5 SV=2	15.81	6.28
A6NNF4	Zinc finger protein LOC730087 OS=Homo sapiens PE=2 SV=3	12.03	1.76
A6NP81	Golgin subfamily A member 8-like protein 2 OS=Homo sapiens PE=2 SV=2	11.53	2.37
O00142	Thymidine kinase 2, mitochondrial precursor (EC 2.7.1.21) (Mt-TK) - Homo	12.58	4.89

Figure 4: "**Cell view**". Screen shot of a cell entry in the GPDE.

retagged (see Results, data organization). This process provides the administrator of a GPDE installation with the possibility of selectively inserting experiments of comparable quality only. In addition, the GPDE provides the user with protein scores giving an indication on how secure an identification is. Thereby the GPDE tackles the immanent problem of any meta-analysis that given results are hard to retrace as it provides the user with instant access to the complete underlying peptide identification details. This feature currently comes with the restriction that only experiments using the same type of peptide identification score can be inserted into the GPDE. Without this feature a user could no longer assess the validity of a single identification and consequently the whole result set would be questionable.

The GPDE is being developed under the Affero General Public License (AGPL) and is freely available from its project page (http://www.griss.co.at/?GPDE). The GPDE has several applications: An example would be making the findings of experiments available in a local intranet thus already providing not yet published data internally for further meta-analysis. Another possibility would be presenting the combined proteomics findings of a research group to the public on the institutional web page. This application is currently used by the Clinical Proteomics Laboratories at the Medical University of Vienna (CPL/MUW). On this "live" example several further applications of the GPDE are discussed.

GPDE @ MUW: The CPL/MUW proteomics database

As presented above the GPDE was designed to facilitate a biological approach to clinical proteomics data by organizing protein identification data around cell types. The CPL/MUW have made use of the GPDE to provide access to their proteomics experiments (CPL/MUW database available at http://www.meduniwien.ac.at/proteomics/database).

Here the GPDE is used in three different ways: 1) as an *in silico* alternative to multiple Western analyses; 2) to provide easy access to cellular proteome reference maps and 3) to reliably support the assessment of the specificity of biomarker candidates.

An in silico alternative to multiple western analysis

The most common way to identify a protein in a laboratory is Western analysis. With the aid of a specific antibody, the presence of a protein in a sample can be assessed. Another independent technology for protein identification is mass spectrometry. Here, the identification of specific amino acid sequences serves to identify the corresponding proteins. Alternatively to applying an antibody to cell lysates for Western analysis, an *in silico* search of the corresponding protein can be performed. Each of more than 4 500 different proteins can currently be searched in the CPL/MUW database which is powered by the GPDE. The GPDE returns two different data sets per individual search. First, the user gets a list of peptides identified by mass spectrometry and assigned to the corresponding protein. These data give an indication how well the protein is accessible via mass spectrometry. The peptide identification scores (currently SpectrumMill scores only) indicate the chance of false positive identifications. More peptide identifications per protein improve the confidence level accordingly. Second, the user gets a list of cell types which were found to express the corresponding protein. This information may be seen like different lanes in a Western experiment. The cell types expressing the searched protein are listed by the GPDE, corresponding to positive bands in a Western assay. The peptide identification data provided per cell type indicate the reliability of identification and provide a semi-quantitative measure of protein abundance. As a conclusion, the user is able to get a complete view on the specificity of the investigated protein expression.

Multiple cellular proteome reference maps

A major limitation for appropriate interpretation of clinical proteome analysis data is the lack of valid references. This applies to every researcher wanting to assess biological implications and the quality of a proteome profile experiment. Comparison to references provides a large number of expected positive controls and may greatly facilitate the identification of contaminants. Furthermore, specific aberrations due to e.g. disease states may be identified more easily. Thus, proteome reference maps are a powerful data assessment and quality control tool applicable to a large variety of samples and analyses. By means of the GPDE, we provide an easy and reliable access to proteome reference maps of 20 cell types currently represented in the CPL/MUW database. Selecting a cell type returns a list of all proteins identified therein. The list contains all protein analysis details including mass analysis data and can be downloaded for further applications according to individual requirements.

Reliable support for the assessment of the specificity of biomarker candidates

A disease biomarker is only valuable if the expression is robustly and specifically associated with the certain disease. A classical strategy for biomarker discovery is the comparative analysis of tissue or serum samples derived from healthy individuals versus diseased patients. Proteins which appear to be regulated may often result from epiphenomena not causally related to the corresponding disease or may occur in many different other diseases (Zolg, 2006). It is a lot of effort to validate biomarker candidates, often beyond the financial capabilities of laboratories.

The relation of protein expression profiles to cells and specific cell states as realized by means of the GPDE may offer a solution to this problem. We are convinced that there still is a large number of biomarkers to be discovered. Specific cell types and specific cell states are inevitably associated with specific protein expression profiles. Any disease is associated with alterations of the cell composition of tissue compartments in addition to alterations of functional cell states. As a result, any disease will always be accompanied by the expression of characteristic proteins, which may serve as biomarkers. Aberrantly expressed proteins which may serve as biomarkers will be expressed by a specific cell type in an aberrant tissue location or by cells which have entered an aberrant but characteristic functional cell state. In order to support the identification and assessment of such candidate biomarker proteins, the GPDE organizes proteome profiles around cells and cell states. If a clinical study identifies a candidate biomarker protein, a simple search in the CPL/MUW database may demonstrate whether or not the protein has the required expression specificity. If a candidate biomarker is found to be expressed by several cell types at relatively high abundance, it will be of rather poor specificity. If a candidate biomarker is found to be expressed by few or a single cell type only and also is plausibly associated with the corresponding disease, it may have high specificity. If a candidate biomarker protein is found to be expressed by specifically activated cells only, this information may substantially contribute to the understanding of the disease pathology.

Future Developments

There is still a lot of room for further development of the GPDE. New features that hopefully will be completed in the near future are: A second search function that allows the creation of more complex queries and the comparison of the result sets of these queries, the support of multiple export formats as well as the support for different accession systems at query level.

Summary

By developing the PRIDE XML data format a standardized data format for proteomics data publishing was created. Furthermore, clinical proteomics are unthinkable without the extensive use of meta-analysis. The GPDE is meant as a simple and easy to use solution to this problem: It combines different proteomics experiments available in the PRIDE XML format to specifically defined “cell types“. Thus, the different experiments become indistinguishable to the user and a completely “biologically” based view on the proteomics data is generated. The GPDE is free software under the terms of the AGPL and available for download at http://www.griss.co.at/?GPDE. Currently, the first institution to use the GPDE is the CPL/MUW, available at http://www.meduniwien.ac.at/proteomics/database.

Acknowledgements

We would like to thank Birgit Thiel for proofreading the manuscript and testing the application.

References

1. Jones P, Cote RG, Cho SY, Klie S, Martens L, et al. (2008) PRIDE: new developments and new datasets. Nucleic Acids Res 36: D878-883. » CrossRef » Pubmed » Google Scholar
2. Jones P, Cote RG, Martens L, Quinn AF, Taylor CF, et al. (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. Nucleic Acids Res 34: D659-663.» CrossRef » Pubmed » Google Scholar
3. Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, et al. (2005) PRIDE: the proteomics identifications database. Proteomics 5: 3537-3545. » CrossRef » Pubmed » Google Scholar
4. Zolg W (2006) The proteomic search for diagnostic biomarkers: lost in translation? Mol Cell Proteomics 5: 1720-1726. » CrossRef » Pubmed » Google Scholar