Published online 4 November 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D181–D184***
doi:10.1093/nar/gkn804

NOPdb: Nucleolar Proteome Database—2008 update

Yasmeen Ahmad[1], François-Michel Boisvert[1], Peter Gregor[2], Andy Cobley[2] and Angus I. Lamond[1,*]

[1]Wellcome Trust Centre for Gene Regulation & Expression and [2]School of Computing, University of Dundee, Dundee DD1 5EH, UK

Received September 12, 2008; Revised and Accepted October 10, 2008

ABSTRACT

An experimental data handling system has been created as an update to the previous Nucleolar Proteome Database (NOPdb3.0: http://www.lamondlab.com/NOPdb3.0/). This updated system is able to manage large data sets identified by multiple mass spectrometry and has been used to analyse highly purified preparations of human nucleoli from different cell lines. The newly created application includes a dynamic relational database, which is kept up to date by laboratory staff. The data are further annotated with information from specific external sources on the web, including the IPI and Gene Ontology databases. In addition, an Application Programming Interface provides external users with a portal to link into the nucleolar proteome database and hence, gain access to continually updated results. From the initial ~700 human proteins identified in the previous iteration of the NOPdb, we have now identified over 50 000 peptides contained in over 4500 human proteins from purified nucleoli, providing enhanced coverage of the nucleolar proteome.

INTRODUCTION

The nucleolus is a highly conserved nuclear organelle whose main function is to coordinate the synthesis and assembly of ribosome subunits (1). Previously we described a Nucleolar Proteome Database (NOPdb2.0: http://www.lamondlab.com/NOPdb) that archived data on >700 proteins that were identified by multiple mass spectrometry (MS) analyses from highly purified preparations of human nucleoli (2). Each protein entry was annotated with information about its corresponding gene, its domain structures and relevant protein homologues across species, as well as documenting its MS identification history, including all the peptides sequenced by tandem MS/MS. Moreover, data showing the quantitative changes in the relative levels of approximately 500 nucleolar proteins were compared at different time points upon transcriptional inhibition (3).

The data presented by the previous NOPdb, version 2.0, was held in a flat file database. Due to the aggregated nature of the data, results from individual experiments could not be extracted. The peptide data for a single protein were merged within this database rather than stored separately. The client interface to this database consisted of Perl CGI scripts. These scripts were able to extract the relevant data from the flat file database to create static html pages. After running the scripts, a page was created on the server for each protein. The html pages were then made available to the global community via the internet. Each time data were updated in the flat files, the Perl scripts had to be run again in order to reproduce the static html pages. This process of having to reproduce the static html protein pages after each database update was highly inefficient and time consuming. A more efficient approach is to produce dynamic html pages upon user request. Furthermore, the capabilities of the version 2.0 NOPdb database were limited with respect to security, ease of use, accessibility, maintainability and expandability. For example, a number of security concerns arose regarding the Perl scripts, which with limited documentation, proved very difficult to resolve.

The new version of the NOPdb3.0 (http://www.lamondlab.com/NOPdb3.0/) consists of a unique, secure, extendable content management system, holding advanced nucleolar proteomics data. The created application includes a dynamic relational database, which is kept up to date by members of the Lamond group. It also allows the query of protein data hosted within the database by external users, either using the custom built interface provided by the Lamond group, or by building custom web tools that access data via the Application Programming Interface (API). In addition to the dynamic interfaces provided by the new content management system, the data included in the nucleolar proteome are also dynamically updated with proteins identified from several different

*To whom correspondence should be addressed. Tel: +44 1382 385473; Fax: +44 1382 345695; Email: angus@lifesci.dundee.ac.uk

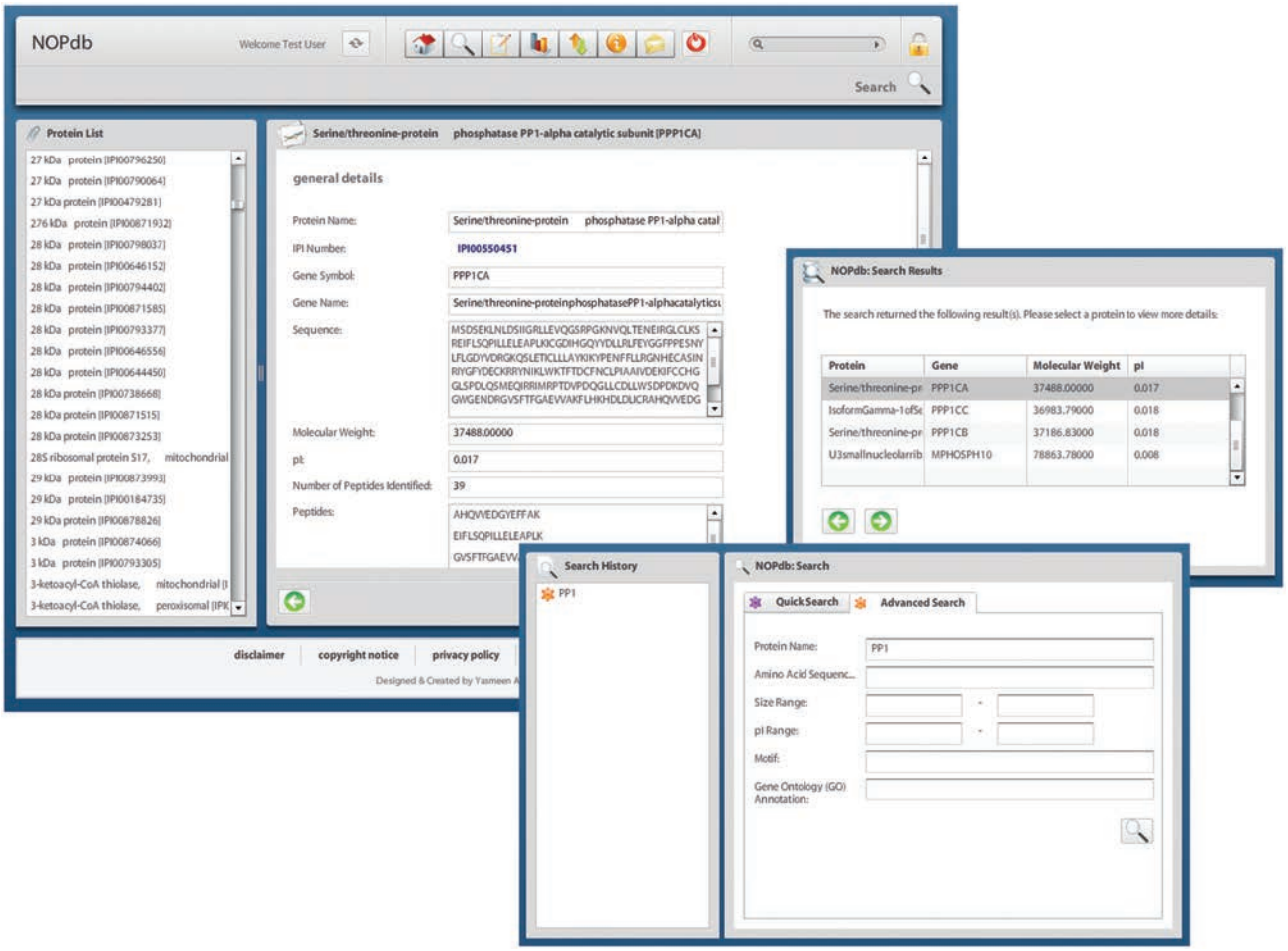

Figure 1. Snapshots of the NOPdb3.0 (http://www.lamondlab.com/NOPdb3.0/). For illustration, the database was searched to identify a Protein Phosphatase 1 (PP1) isoform and here we show an overview page for this protein documenting its sequence, peptides identified, etc.

cell lines, using various instruments by members of the laboratory. From the initial ~700 proteins identified in the previous iteration of the NOPdb, we have now identified over 50 000 peptides contained in over 4500 human proteins from purified nucleoli, providing significantly enhanced coverage of the nucleolar proteome.

DATABASE ACCESS

We have established the new version of the Nucleolar Proteome Database (NOPdb3.0), which archives all the human nucleolar proteins identified to date by the Lamond group and their collaborators using MS analyses (1–3). This current version 3.0 of the database is available at http://www.lamondlab.com/NOPdb3.0/ and is searchable either by protein name, protein sequence, motif (4–6), Gene Ontology (GO) (7) terms or by setting the range of the predicted isoelectric point and/or molecular weight (Figure 1). To date, NOPdb3.0 archives over 4500 human nucleolar proteins verified by multiple MS analyses in different cell lines. The NOPdb3.0 provides information on multiple parameters, including protein name, accession number, gene symbol, gene name, sequence, molecular weight, isoelectric point (pI), peptides identified, experiments in which the protein was identified, motifs and GO annotation. The previous version of the database (2) will still be available through our website at http://www.lamondlab.com/NOPdb/.

DATABASE IMPLEMENTATION

The new NOPdb3.0 application consists of a multi-tier architecture, where the data storage, business logic and client interface are separate components. The data storage is implemented via a relational mySQL database. The database is structured (Figure 2) to allow easy extendibility and maintenance in the future. In order to extract useful data, the business logic employs complex SQL queries. The purpose of the business logic layer is to act as an interface between the client-side application and database. The business logic and client interface can both reside on any Apache web server capable of serving PHP classes and the client interface, which is built in Adobe Flex. Adobe Flex was chosen as it allows Rich Internet Applications (RIAs) to be prototyped and developed rapidly, with the end product running across a wide range of client browsers.

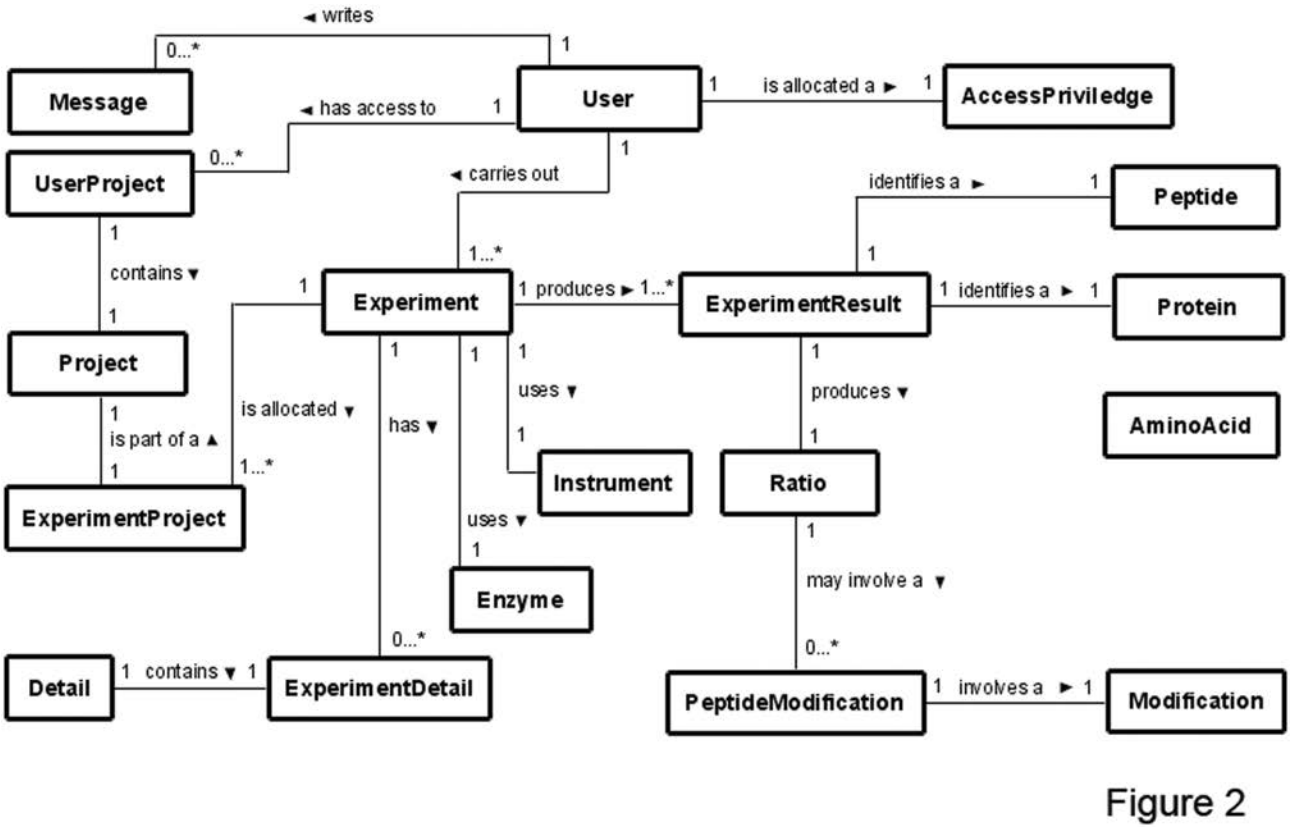

Figure 2. Entity Relationship (ER) diagram depicting the relationships between the tables present in the mySQL database implemented for the NOPdb3.0 Content Management System.

Version 3.0 of the NOPdb is an entirely new implementation using a fully relational design with major improvements over previous versions and additional functionality. The newly created database holds data of higher granularity, storing data at the peptide level as opposed to collated data on proteins. This higher granularity also means that results from new experiments can be directly uploaded to the database without prior processing, as the direct output from MS-based proteomics analyses is peptide data. The application has the ability to interpret data and therefore aggregate it to provide metadata for proteins on a usable, graphical interface. The structure of the application has been designed using the model view controller design pattern (8), thus meaning that the functionality is separated from the overall look and feel of the application to ensure a more customisable solution. All communication between the database and application has been implemented to pass through the custom made API (9). Furthermore, in this new version 3.0 application, the graphical user interface to the database is able to create data pages 'on the fly' using the custom API rather than serving static data pages, as in previous versions. This API not only acts as a security blanket around the database, it also provides the ability for users to create their own websites and/or applications that represent the data being stored in the proteomics database. External users can make use of the API through the REST (Representational State Transfer) (10) approach. Hence, external programmers can retrieve content in XML (Extensible Markup Language) format, from the database, by accessing well-documented Uniform Resource Locators (URLs).

The application also facilitates mining of stored data, with data being stored in a relational structure that is well documented. Thus tools can be built to search, analyse, read and understand the data. This mining capability is evident within the application, with the database being searchable by multiple parameters, including gene names, amino acid or nucleotide sequences, sequence motifs, or by limiting the range for isoelectric points and/or molecular weights. The database is also searchable by Interpro motif numbers (database of protein families, domains and functional sites) (4–6) and by GO terms (describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner) (7). Furthermore, the NOPdb3.0 application uses the API to create dynamically generated graphs, allowing the users to visualise the data produced from experiments and enabling cross analysis between experiments.

Increased security was a core focus of this development. The application itself is designed with three levels of access, to facilitate management and to prevent unauthorised use of the system. Users are provided with different levels of access according to their needs, which are seamlessly enforced by the application. This security ensures

that the data remain accurate and the quality of the data is not compromised. Furthermore, this application creates a platform for the Lamond group to share their data with the wider cell biology community.

DATABASE CONTENT

The database has been populated with different sets of experiments, performed in the Lamond laboratory, that identify proteins in purified preparations of human nucleoli. This new NOPdb3.0 now contains over 4500 proteins identified in different human cells lines. The increased coverage of the human nucleolus proteome is illustrated by the fact that NOPdb3.0 now includes over 80% of ribosomal proteins, as opposed to the ~28% described in NOPdb version 2.0. We estimate that NOPdb3.0 contains over 80% of the main human nucleolus proteins. The proteins in the database will be regularly updated as more experiments are performed in the Lamond laboratory.

ACKNOWLEDGEMENTS

We would like to thank Drs Douglas Lamont and Kenneth Beattie of the Fingerprints Proteomics Facility at the University of Dundee (http://proteomics.lifesci.dundee.ac.uk/) for technical assistance.

FUNDING

This work was supported by a Wellcome Trust Programme Grant (073980/Z/03/Z) and by an interdisciplinary RASOR (Radical Solutions for Researching the Proteome) initiative, which is supported by the Biotechnology and Biological Sciences Research Council, Engineering and Physical Sciences Research Council, Scottish Higher Education Funding Council and Medical Research Council. A.I.L. is a Wellcome Trust Principal Research Fellow. Caledonian Research Foundation Fellowship (to F.M.B.). BBSRC PhD studentship (to Y.A.). Funding for open access charge: Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Boisvert,F.M., van Koningsbruggen,S., Navascues,J. and Lamond,A.I. (2007) The multifunctional nucleolus. *Nat. Rev. Mol. Cell. Biol.*, **8**, 574–585.
2. Leung,A.K., Trinkle-Mulcahy,L., Lam,Y.W., Andersen,J.S., Mann,M. and Lamond,A.I. (2006) NOPdb: Nucleolar Proteome Database. *Nucleic Acids Res.*, **34**, D218–D220.
3. Andersen,J.S., Lam,Y.W., Leung,A.K., Ong,S.E., Lyon,C.E., Lamond,A.I. and Mann,M. (2005) Nucleolar proteome dynamics. *Nature*, **433**, 77–83.
4. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
5. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
6. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, D142–D144.
7. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
8. Madeyski,L. and Sochmialek,M. (2005) Architectural design of modern web applications. *Found. Comput. Decision Sci.*, **30**, 49–60.
9. Henning,M. (2007) API: design matters. *ACM Queue*, **5**, 4–14.
10. Fielding,R.T. and Taylor,R.N. (2002) Principled design of the modern web architecture. *ACM Trans. Internet Technol.*, **2**, 115–150.