Available online at www.sciencedirect.com

SciVerse ScienceDirect

Genomics Proteomics Bioinformatics 10 (2012) 226–229

GENOMICS PROTEOMICS & BIOINFORMATICS

www.elsevier.com/locate/gpb

Application Note

NSort/DB: An Intranuclear Compartment Protein Database

Kai Willadsen [1,2,*], Nurul Mohamad [2,4], Mikael Bodén [1,2,3,*]

[1] School of Chemistry and Molecular Biosciences, The University of Queensland, St. Lucia, QLD 4072, Australia
[2] Institute for Molecular Bioscience, The University of Queensland, St. Lucia, QLD 4072, Australia
[3] School of Information Technology and Electrical Engineering, The University of Queensland, St. Lucia, QLD 4072, Australia
[4] Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur 50603, Malaysia

Received 13 March 2012; accepted 13 April 2012
Available online 25 July 2012

Abstract

Distinct substructures within the nucleus are associated with a wide variety of important nuclear processes. Structures such as chromatin and nuclear pores have specific roles, while others such as Cajal bodies are more functionally varied. Understanding the roles of these membraneless intra-nuclear compartments requires extensive data sets covering nuclear and compartment-associated proteins. NSort/DB is a database providing access to intra- or sub-nuclear compartment associations for the mouse nuclear proteome. Based on resources ranging from large-scale curated data sets to detailed experiments, this data set provides a high-quality set of annotations of non-exclusive association of nuclear proteins with structures such as promyelocytic leukaemia bodies and chromatin. The database is searchable by protein identifier or compartment, and has a documented web service API. The search interface, web service and data download are all freely available online at http://www.nsort.org/db/. Availability of this data set will enable systematic analyses of the protein complements of nuclear compartments, improving our understanding of the diverse functional repertoire of these structures.

Keywords: Nuclear compartments; Nuclear proteins; Web service

Introduction

In recent years, nuclear architecture has been recognised as playing a key role in cellular regulation [1]. Many core nuclear processes are associated with structural components: chromatin with DNA compaction and transcriptional access, nuclear pores with macromolecular translocation, and nuclear speckles with transcript splicing. Other nuclear structures, which are membraneless and morphologically distinct, such as promyelocytic leukaemia (PML) bodies and Cajal bodies, are present in large numbers and with heterogeneous functional repertoires. These compartments are primarily composed of large sets of proteins, though DNA and RNA are also involved. Recent advances in large-scale proteomics technologies have enabled more detailed study of the molecular makeup of these compartments than was previously possible.

Access to protein localisation data enables a deeper understanding of the role of nuclear compartments. For example, we now know that many novel nucleolar proteins subserve ribosomal biogenesis [2], and that sumoylation sites occur frequently in PML body proteins [3], confirming earlier hypotheses [4]. From the development and evaluation of predictors based on this localisation data, we now appreciate a fuller protein complement of each compartment, and can benefit from insights into, for example, the regulatory role of PML bodies as demonstrated by their enrichment in transcription factor member proteins [5].

Existing databases such as the Nucleolar Proteome Database (NOPdb) [6] and the Nuclear Matrix Protein Database (NMP-db) [7] provide comprehensive annotation of the protein complements of individual compartments, but focus on a restricted subset of currently recognised compartments. In contrast, the Nuclear Protein Database

* Corresponding authors.
E-mail: k.willadsen@uq.edu.au (Willadsen K), m.boden@uq.edu.au (Bodén M).

http://dx.doi.org/10.1016/j.gpb.2012.07.001

(NPD) [8] covers a wide range of nuclear compartments, and is a valuable resource providing compartment annotation data and metadata for nuclear proteins from multiple organisms–mostly human and mouse. For bioinformatic applications, large, high-quality data sets consisting of both positive and negative samples are required. NPD and other sources provide a strong basis for constructing these data sets, but more can be done, including extending the existing data sets, building on a high-quality experimentally verified set of nuclear proteins, and mapping onto a single proteome.

NSort/DB is a new resource providing access to nuclear proteins' non-exclusive association with major nuclear structures. It combines annotations from existing data sets with experimental data in recent literature to uniquely map the currently-known mouse nuclear proteome, offering opportunities to characterise the functional organisation of the mammalian nuclear architecture.

Methods

Database construction and content

Our data set provides annotations of the intra-nuclear localisation of mouse nuclear proteins, collecting and extending available annotations from several pre-existing data sets. On the basis of coverage in current data sets and literature, we defined compartments of interest to be any compartment with at least 20 different associated proteins. As a result, we distinguish between eight major compartments: PML body, nucleolus, nuclear speckle, nuclear pore, Cajal body, chromatin, nuclear lamina and perinucleolar compartment (PNC).

Information about intra-nuclear compartments must be founded on high-quality nuclear localisation data. The NUCPROT data set provides an authoritative map of the mouse nuclear proteome, consisting of 2568 proteins with direct experimental evidence of nuclear localisation, and a further 2854 proteins predicted by multiple computational methods to localise to the nucleus [9]. The NUCPROT experimental data is based on overexpression of proteins, and as such, some mislocalization of nuclear proteins as cytoplasmic (and *vice versa*) can occur. Nevertheless, NUCPROT represents a high-quality data set designed to be composed exclusively of mouse nuclear proteins, and therefore it provides a reference with which to assess the coverage (*i.e.*, the proportion of nuclear proteins associated with a given compartment) and redundancy (*i.e.*, orthologous proteins are excluded, reducing duplicated annotations) of collected data. However, intra-nuclear compartment associations are not provided by NUCPROT, and so must be sourced from elsewhere.

Data on proteins' compartment associations was aggregated from a range of sources. First, proteins and their associations were gathered from specialised nuclear proteome databases including NPD [8], NOPdb [6] and NMP-db [7]. This collection was supplemented with proteins whose localisation annotations indicated nuclear (or more specific) localisation, taken from generic protein databases such as the UniProt Knowledgebase [10] and the Human Protein Reference Database [11] (Tables S1 and S2). The resulting data set consists of proteins that have been experimentally or computationally determined to localise to the nucleus, some with specific intra-nuclear compartment associations, largely from human data. As NUCPROT covers the mouse nuclear proteome, BioMart and the Mouse Genome Informatics database [12] were used to map the data set to mouse protein identifiers via orthologous genes.

To verify the nuclear localisation of proteins in the predicted segment of the NUCPROT data set, we required additional support from the compartment annotation data assembled above; only proteins represented in both the NUCPROT predicted set and the compartment annotation data set were kept, resulting in a set of 2295 proteins with nuclear import support from at least two distinct sources, and 917 proteins included based on experimental support from NUCPROT. Due to the high value of compartment data, proteins not mappable to NUCPROT identifiers but with intra-nuclear compartment annotations were reconsidered; entries with an *E*-value smaller than 10^{-4} when BLASTed against NUCPROT sequences were retained, giving an additional 322 proteins.

Finally, additional data was obtained from compartment-specific reviews and large-scale proteomics articles [13–16] (PubMed identifiers for individual annotations are provided in the data set), resulting in 32 new nuclear proteins, and providing additional or supporting annotations for 78 proteins from 63 distinct literature sources. Proteins were added to the data set if their nuclear localisation was supported by clear experimental evidence. These additional literature-sourced annotations were manually curated; annotations from the existing databases referenced in construction use a combination of manual and automatic curation.

The resulting data set, being made available as NSort/DB, consists of 3566 proteins, of which 1285 have at least one intra-nuclear compartment association (**Table 1**). The remaining 2281 proteins are known to localise to the

Table 1 Non-exclusive compartment protein counts

Compartment	Count	Percentage (%)
Cajal body	49	0.90
Chromatin	323	5.96
Nuclear lamina	77	1.42
Nuclear pore	51	0.94
Nuclear speckle	403	7.43
Nucleolus	598	11.03
PML bodies	91	1.68
PNC	24	0.44

Note: since the majority of nuclear proteins still have no known intra-nuclear compartment localization, the table gives individual compartment localization counts as a percentage of all nuclear proteins in the dataset, rather than as a percentage of compartment-associated proteins. PML stands for promyelocytic leukaemia while PNC indicates perinucleolar compartment.

nucleus, but have no established compartment associations. Dynamic aspects of compartment association are not represented, and protein isoforms are not distinguished. The data set will be updated as necessary to support continuing work on models of intra-nuclear trafficking of proteins.

Utility

NSort/DB presents a simple web interface, which allows both individual searching and batch queries of proteins' compartment associations, and browsing of proteins by compartment. Search and browsing results are made available for download in standard formats. The web interface can be accessed at http://www.nsort.org/db/, and no access restrictions are imposed.

The intra-nuclear compartment association data, along with source of annotations, is stored in a flat-file database, queried via a custom-built Java parser. This database is being made available for download in its original flat-file format.

In addition to the web interface and data download, a simple web application program interface (API) is available for automated access to the data set. The API accepts UniProtKB or intra-nuclear compartment identifiers and provides responses in JSON or CSV format. Documentation for the web API is available at http://www.nsort.org/db/api/.

The data set presented here differs from existing databases in several ways. First, in contrast to NOPdb and NMP-db, multiple intra-nuclear compartments are covered in NSort/DB. Such coverage span is required for any analysis that involves cross-compartment comparison. Second, NPD provides significant compartment annotation data, which our data set includes and extends upon with additional literature-sourced annotations. However, in contrast to NPD, our data set is mapped to a single high-quality nuclear proteome, improving confidence in the identification of proteins as nuclear, and extending the nuclear proteome. Lastly, our data set provides significant additional value for bioinformatic analyses through the use of the NUCPROT nuclear proteome set. In applications such as distinguishing functional roles of compartments (*e.g.*, [5]), in which a statistical background is required, the extended background provided by the NUCPROT data set gives a better statistical basis for functional identification of compartment roles.

Application

In order to illustrate the kind of analyses made possible by NSort/DB, we provide a simple analysis of functional sites in compartments' member proteins, using the Eukaryotic Linear Motif (ELM) resource and our web API. In 98 lines of Python code, we obtain a list of current ELM motifs, retrieve sequences and nuclear compartment associations from our server, identify motif occurrences, and establish statistical overrepresentation of motifs in each compartment with Fisher's exact test, using all nuclear proteins as a background set.

Table 2 Compartment-associated motifs

Compartment	Motif	*E*-value
Cajal body	CK1 phosphorylation site	1.1
Chromatin	PKA phosphorylation site	1.3e−4
Nuclear lamina	Nuclear receptor binding	2.2e−2
Nuclear pore	Plk phosphorylation site	2.8e−1
Nuclear speckle	GRB2-like SH2 domain binding	1.4e−3
Nucleolus	MAPK docking	2.0e−4
PML bodies	SUMO-1 sumoylation site	7.1e−1
PNC	Src-family SH2 domain binding	8.3e−1

A selection of associations is identified in **Table 2**. Results showed that post-translational modification is a common theme. In particular, the occurrence of phosphorylation is notable, given recent suggestions that it may act as a regulatory mechanism for compartment-specific activities [17,18]. Code to reproduce this analysis and the complete table of results are available at http://www.nsort.org/db/sample/.

Conclusions

NSort/DB provides the research community with high-quality intra-nuclear localisation data for the mouse nuclear proteome, allowing new questions to be asked about the structure and function of nuclear compartments. This data set provides a basis for answering relevant biological questions. Indeed, it has already been used to predict the full protein complement of intra-nuclear compartments using computational methods, and to establish the role of PML bodies in regulation of immune response [5]. We anticipate that public availability of this data set will enable further investigations into intra-nuclear compartments and their varied functional roles within the nucleus.

Competing interests

The authors declare no competing interests.

Authors' contributions

MB and NM conceived and designed the requirements for the data set. NM constructed the data set and undertook literature curation. KW performed additional data validation, and developed the web site and case study. KW and MB wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors acknowledge funding and support from the Australian Research Council Centre of Excellence in Bioinformatics.

Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb.2012.07.001.

References

[1] Gorski SA, Dundr M, Misteli T. The road much traveled: trafficking in the cell nucleus. Curr Opin Cell Biol 2006;18:284–90.

[2] Hinsby AM, Kiemer L, Karlberg EO, Lage K, Fausbøll A, Juncker AS, et al. A wiring of the human nucleolus. Mol Cell 2006;22:285–95.

[3] Mohamad N, Bodén M. The proteins of intra-nuclear bodies: a data-driven analysis of sequence, interaction and expression. BMC Syst Biol 2010;4:44.

[4] Heun P. SUMOrganization of the nucleus. Curr Opin Cell Biol 2007;19:350–5.

[5] Bauer DC, Willadsen K, Buske FA, LeCao K, Bailey TL, Dellaire G, et al. Sorting the nuclear proteome. Bioinformatics 2011;27:i7–i14.

[6] Leung AK, Trinkle-Mulcahy L, Lam YW, Andersen JS, Mann M, Lamond AI. NOPdb: Nucleolar Proteome Database. Nucleic Acids Res 2006;34:D218–20.

[7] Mika S, Rost B. NMPdb: Database of Nuclear Matrix Proteins. Nucleic Acids Res 2005;33:D160–3.

[8] Dellaire G, Farrall R, Bickmore WA. The Nuclear Protein Database (NPD): sub-nuclear localisation and functional annotation of the nuclear proteome. Nucleic Acids Res 2003;31:328–30.

[9] Fink JL, Karunaratne S, Mittal A, Gardiner D, Hamilton N, Mahony D, et al. Towards defining the nuclear proteome. Genome Biol 2008;9:R15.

[10] Consortium UniProt. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res 2011;39:D214–9.

[11] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database–2009 update. Nucleic Acids Res 2009;37:D767–72.

[12] Blake JA, Bult CJ, Kadin JA, Richardson JE, Eppig JT; The Mouse Genome Database Group. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. Nucleic Acids Res 2011;39:D842–8.

[13] Andersen JS, Lyon CE, Fox AH, Leung AK, Lam YW, Steen H, et al. Directed proteomic analysis of the human nucleolus. Curr Biol 2002;12:1–11.

[14] Cioce M, Lamond AI. Cajal bodies: a long history of discovery. Annu Rev Cell Dev Biol 2005;21:105–31.

[15] Fox AH, Lam YW, Leung AKL, Lyon CE, Andersen J, Mann M, et al. Paraspeckles: a novel nuclear domain. Curr Biol 2002;12:13–25.

[16] Cronshaw JM, Krutchinsky AN, Zhang W, Chait BT, Matunis MJ. Proteomic analysis of the mammalian nuclear pore complex. J Cell Biol 2002;158:915–27.

[17] Hebert MD. Phosphorylation and the Cajal body: modification in search of function. Arch Biochem Biophys 2010;496:69–76.

[18] Kosako H, Imamoto N. Phosphorylation of nucleoporins: signal transduction-mediated regulation of their interaction with nuclear transport receptors. Nucleus 2010;1:309–13.