NSort/DB: An intra-nuclear compartment protein database

Kai Willadsen[a,b], Nurul Mohamad[b,d], Mikael Bodén[a,b,c,*]

[a] *School of Chemistry and Molecular Biosciences, The University of Queensland, St. Lucia, QLD 4072, Australia*
[b] *Institute for Molecular Bioscience, The University of Queensland, St. Lucia, QLD 4072, Australia*
[c] *School of Information Technology and Electrical Engineering, The University of Queensland, St. Lucia, QLD 4072, Australia*
[d] *Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur 50603, Malaysia*

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, has a documented web service API, and the data set is freely available. 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.

The search interface, web service and data download are all available online at http://nsort.org/db/.

Keywords: nuclear compartments, nuclear proteins, web service

Background

In recent years, nuclear architecture has been recognised as playing a key role in cellular regulation [9]. 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—membraneless and morphologically distinct, such as promyelocytic leukaemia (PML) bodies and Cajal bodies—occur 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 make-up 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 [12], and that sumoylation sites occur frequently in PML body proteins [17], confirming earlier hypotheses [11]. 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 [2].

Existing databases such as the Nucleolar Proteome Database (NOPdb) [15] and the Nuclear Matrix Protein Database (NMP-db) [16] 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 (NPD) [6] 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 nega-

*Corresponding author
Email addresses: k.willadsen@uq.edu.au (Kai Willadsen), m.boden@uq.edu.au (Mikael Bodén)

tive samples are required. NPD and other sources provide a strong basis for constructing these data sets, but more can be done, including extending existing data sets, building on a high-quality experimentally verified set of nuclear proteins, and mapping on to 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 current known mouse nuclear proteome, offering opportunities to characterise the functional organisation of the mammalian nuclear architecture.

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 twenty 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 [7] 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. The NUCPROT experimental data is based on over-expression of proteins, and as such, some mislocalisation 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 as such 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. 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: NPD [6], NOPdb [15] and NMP-db [16]. 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 [18] and the Human Protein Reference Database [13]; see Supplementary Tables 1 and 2 for details. 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 [3] were used to map the data set to mouse protein identifiers via orthologous genes.

Table 1: Non-exclusive compartment protein counts

Compartment	Count (%age)
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%)

As verification of nuclear localisation for 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 data sets 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 manuscripts [1, 4, 8, 5] (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.

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 (see Table 1); the remaining 2281 proteins are known to localise to the nucleus, but have no established compartment associations. Dynamic aspects of compartment association are not represented, and protein isoforms are not distinguished.

Utility

The data set presented here differs from existing databases in several ways. In contrast to NOPdb and NMP-db, multiple intra-nuclear compartments are covered; this coverage span is required for any analysis that involves cross-compartment comparison. NPD provides significant compartment annotation data, which our data set includes and extends upon with additional literature-sourced annotations. In addition, our data set provides significant additional value for bioinformatic analyses through the use of the NUCPROT nuclear proteome set. 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. In applications such as distinguishing functional roles of compartments (e.g., [2]), 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.

Data Access

NSort/DB presents a simple web interface allowing 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://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 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://nsort.org/db/api/.

Table 2: Compartment protein counts and selected associated motifs

Compartment	Motif description (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)

Case Study

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.

A selection of associations are identified in Table 2. 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 [10, 14]. Code to reproduce this analysis and the complete table of results are available at http://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 establish PML bodies' role in regulation of immune response [2]. We anticipate that public availability of this data will enable further investigations into intra-nuclear compartments and their varied functional roles within the nucleus.

References

[1] Andersen, J.S., et al., 2002. Directed proteomic analysis of the human nucleolus. Curr Biol 12, 1–11.

[2] Bauer, D.C., et al., 2011. Sorting the nuclear proteome. Bioinformatics 27, i7–i14.

[3] Blake, J.A., et al., 2011. The mouse genome database (MGD): premier model organism resource for mammalian genomics and genetics. Nucleic Acids Res 39, D842–D848.

[4] Cioce, M., Lamond, A.I., 2005. Cajal bodies: A long history of discovery. Annu Rev Cell Dev Bi 21, 105–131.

[5] Cronshaw, J.M., et al., 2002. Proteomic analysis of the mammalian nuclear pore complex. J Cell Biol 158, 915–927.

[6] Dellaire, G., Farrall, R., Bickmore, W., 2003. The nuclear protein database (NPD): sub-nuclear localisation and functional annotation of the nuclear proteome. Nucleic Acids Res 31, 328–330.

[7] Fink, J., et al., 2008. Towards defining the nuclear proteome. Genome Biol 9, R15.

[8] Fox, A.H., et al., 2002. Paraspeckles: a novel nuclear domain. Curr Biol 12, 13–25.

[9] Gorski, S.A., Dundr, M., Misteli, T., 2006. The road much traveled: trafficking in the cell nucleus. Curr Opin Cell Biol 18, 284–290.

[10] Hebert, M.D., 2010. Phosphorylation and the Cajal body: modification in search of function. Arch Biochem Biophys 496, 69–76.

[11] Heun, P., 2007. Sumorganization of the nucleus. Curr Opin Cell Biol 19, 350–355.

[12] Hinsby, A., et al., 2006. A wiring of the human nucleolus. Mol Cell 22, 285–295.

[13] Keshava Prasad, T.S., et al., 2009. Human protein reference database–2009 update. Nucleic Acids Res 37, D767–D772.

[14] Kosako, H., Imamoto, N., 2010. Phosphorylation of nucleoporins: signal transduction-mediated regulation of their interaction with nuclear transport receptors. Nucleus 1, 309–313.

[15] Leung, A.K.L., Trinkle-Mulcahy, L., Lam, Y.W., Andersen, J.S., Mann, M., Lamond, A.I., . NOPdb: Nucleolar proteome database. Nucleic Acids Research 34, D218–D220.

[16] Mika, S., Rost, B., 2005. NMPdb: Database of nuclear matrix proteins. Nucleic Acids Res 33, D160–D163.

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

[18] The UniProt Consortium, 2011. Ongoing and future developments at the universal protein resource. Nucleic Acids Res 39, D214–D219.