D160–D163 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki132

NMPdb: Database of Nuclear Matrix Proteins

Sven Mika[1,3,*] and Burkhard Rost[1,2,4]

[1]CUBIC and [2]North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA, [3]Institute of Physical Biochemistry, University Witten/Herdecke, Stockumer Strasse 10, 58448 Witten, Germany and [4]Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St Nicholas Avenue, New York, NY 10032, USA

Received July 23, 2004; Revised and Accepted October 27, 2004

ABSTRACT

The nuclear matrix (NM) is a structure resulting from the aggregation of proteins and RNA in the nucleus of eukaryotic cells; it is the 'sticky bit' that remains after aggressive DNAse digestion and salt extraction protocols. Owing to the important role of the NM in DNA replication, DNA transcription and RNA splicing, the expression pattern of NM proteins has become an important early indicator for numerous cancers/tumors. Recent descriptions of the NM structure distinguish between a network-like 'internal nuclear matrix' (INM) and a 'nuclear shell' that connects the INM to the inner and outer nuclear membranes. A cautious NM preparation protocol reveals a coat of proteins on top of the INM; these proteins are usually referred to as the 'nuclear matrix-associated proteins'. Here, we describe a new database (NMPdb at http://www.rostlab.org/db/NMPdb/) that currently contains details of 398 NM proteins. We collected these data through a semi-automated analysis of over 3000 scientific articles in PubMed. We could match these 398 proteins to 302 protein sequences in UniProt or GenBank. Our NMPdb repository annotates these links along with the following annotations: organism, cell type, PubMed identifier, sequence-based predictions of structural and functional features and for some entries the explicit sequence segment that is responsible for localization (nuclear matrix targeting signal).

INTRODUCTION

In the early 1960s, researchers began to describe an important nuclear structure in eukaryotic cells that differed from the already well-known DNA/histone-based chromatin (1). This structure, referred to as the 'nuclear matrix' (NM), can be separated from the rest of the nucleus by applying DNAse I digestion followed by salt extraction (2). Many functional aspects of the NM have been described; these include DNA replication (3), DNA transcription (4) and DNA repair (5,6). The existence of the NM as an 'independent' sub nuclear structure is not a proven reality but a widely accepted hypothesis that has profoundly influenced the literature: PubMed alone retrieves over 3000 articles associated when queried with the terms 'nuclear matrix' or 'nuclear scaffold'. The NM might still be an artificial result of the preparation methods rather than a real *in vivo* structure (7–9). However, the main facts that argue in favor of the existence of this controversial part of the nucleus are its observation in non-eluted nuclei through electron spectroscopic imaging (10), the existence of protocols to isolate the NM at physiological salt concentrations through electroelution of chromatin (11), the fact that chromatin loops (S/MAR-DNA sequences) bind to a non-chromatin network and finally the description of functional units that stay in their original place even after removing chromatin and soluble proteins from the nucleus (12). Two main structural elements form the NM (13): the 'internal nuclear matrix' (INM) and the 'nuclear shell' (or 'nuclear lamina'). The INM is an aggregate of proteins, mainly the intermediate filaments lamins, NuMa (13) and hnRNP proteins (13,14). The nuclear shell links the INM to the nuclear membranes and/or nuclear envelope. Several non-INM proteins can be separated along with the INM through more careful preparation protocols (15,16). These proteins are usually referred to as 'associated with the nuclear matrix'. The protein composition of nuclear matrices in different organisms and cell types was discovered mainly by 2D gel electrophoresis, a method that separates proteins based on their isoelectric points (first dimension) and molecular weight (second dimension).

Nuclear matrices, once separated from the chromatin and the soluble compartments of the nucleus, contain very different proteins in tumor than in non-tumor cells (17,18). In cancer research, these differences provide early indications for

*To whom correspondence should be addressed. Tel: +1 212 305 4018; Fax: +1 212 305 7932; Email: mika@cubic.bioc.columbia.edu

different types of tumors. Collecting and analyzing data about NM proteins may help to understand the relationship between those proteins and cancer and to discover NM-associated proteins that have not been implicated with the NM. The vast majority of proteins that have actually been associated experimentally with the NM are not annotated in public databases. Thus, we have built and are maintaining NMPdb, a database with proteins that are associated to the nuclear matrix.

DATABASE

Nuclear matrix proteins collected from the literature

First, we downloaded over 3000 abstracts from PubMed that resulted from queries with the terms 'nuclear matrix/matrices' and 'nuclear scaffold'. Then we wrote a simple Perl script that color-highlighted three types of phrases in the text (through HTML tagging): (i) 'nuclear matrix' terms, (ii) UniProt protein names and (iii) verbs describing binding processes such as 'to bind', 'to associate' or 'to interact' (Figure 1). Each abstract was followed by HTML elements that enabled the quick interactive subclassification of each protein into one of the following classes: (i) part of the internal nuclear matrix (INM), (ii)'tightly' associated with the INM (ASC), (iii) affinity toward the INM changes depending on protein modification, cell type and/or current stage of the cell cycle (MIX) and (iv) part of the nuclear shell/nuclear lamina (NUS). At this point, we also removed abstracts that contained the search words but did not promise to add information to our database. Finally, we collected the names of the organisms and the cell types in which the interaction with the NM was observed.

Content

Currently, NMPdb contains over 3000 links to PubMed articles corresponding to about 400 unique proteins; for about 300 of these proteins we could verify the links to their sequences through either UniProt (19) or GenBank (20). Only 62 of all proteins had significant sequence similarity to any protein with known high-resolution information about the 3D structure as deposited in the PDB (21). Only 101 of the 300 proteins were very different in their sequences [HSSP values below 0 (22)], and about half had rather high levels of sequence similarity to at least one other protein in our set (HSSP value >10). Of the 400 proteins, 42, were classified as INM, 198 as ASC and 130

NA: Protein Names	DNA-directed RNA polymerase II largest subunit RPB1
GN: Gene Names	POLR2A
EC: EC Number	EC 2.7.7.6
OS: Organism	Homo sapiens (Human)
CT: Cell Type	HeLa cell
P1: PubMed	8710856 (1996), 8972849 (1996), 3418709 (1988)
SP: SWISSPROT ID	RPB1_HUMAN, P24928
FN: Function	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates.
KW: Keywords	DNA-binding, DNA-directed RNA polymerase, Nuclear protein, Phosphorylation, Repeat, Transcription, Transferase, Zinc, Zinc-finger
GB: GenBank ID	CAA45125.1, CAA52862.1, CAA52862.1, CAA52862.1, CAA52862.1, CAA52862.1
PE: PEP ID	rpb1 human
OM: OMIM ID	180660
SI: Seq Information	1970 AA; 216961 Da; calculated pI: 7.02
SQ:	MHGGGPPSGDSACPLRTIKRVQFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLMDP 60

Figure 1. Screenshot of an NMPdb entry. The names and gene names of the proteins are given in the fields NA and GN. Also shown are the fields for organism (OS) and the cell type of observed NM interaction (CT). Additionally, links are given to Swiss-Prot (SP), GenBank (GB), OMIM (OM) and to all PubMed articles (P1/P2) that were mined for information about the NM interaction of the protein.

as MIX; very few (currently 13) were classified as NUS. Most proteins (301) are mammalian (predominantly human, rat and mouse); 29 are viral proteins (e.g. HIV, Papyloma/HPV, Epstein–Barr/EBV). Since such viral proteins are typically involved in the transcription of host DNA, it is not surprising that they are an abundant part of the nuclear matrix in infected cells. Other organisms prominent in NMPdb are *Gallus gallus* (chicken, with 16 proteins), *Drosophila melanogaster* (fruit fly, with 14 proteins), *Saccharomyces cerevisiae* (yeast, with 13 proteins) and *Caenorhabditis elegans* (worm, with 6 proteins).

Format and fields

NMPdb has been formatted in an EMBL-like flat file format. Each NM protein is represented by one entry. All entries in the database contain the following fields: (i) origin (organism and cell types), (ii) type of nuclear matrix interaction/involvement (INM, ASC, MIX or NUS), (iii) molecular mass and known or calculated pI for locating the protein on a 2D gel and (iv) reference (PubMed IDs of articles describing the interaction). For some entries we provide additional links to other databases, give the actual protein sequence and collect sequence-based predictions. Although links to UniProt implicitly link NMPdb to a variety of other databases, we also provide explicit links to OMIM (23), SWISS-2DPAGE (24) and S/MARt DB (25)—which contains the DNA sequences that the respective protein binds to. We provide the following information for all proteins for which we have sequences: (i) the structural domain-like organization according to CHOP (26,27), (ii) predictions of secondary structure, solvent accessibility and membrane helices through PROFphd [B. Rost, manuscript submitted; (28,29)], (iii) coiled-coil regions through COILS (30), (iv) disordered regions through NORSp (31,32). Where possible, entries are also cross-linked to PEP, a database with predictions for entire proteomes (33) that also contains sequence alignments. For 53 sequences in the database, we found pecific information about which part of the sequence is responsible and necessary for NM binding. These regions, usually referred to as nuclear matrix targeting signals (NMTS), are also deposited in NMPdb if available.

Access

NMPdb can be accessed from http://www.rostlab.org/db/NMPdb/—a search-engine interface that allows the querying by different database fields and the linking of queries through 'AND', 'OR' and 'AND-NOT'. The complete NMPdb database can be downloaded via ftp. The content of the database, the meaning of the fields and the search interface are described in separate help pages.

Updates

NMPdb annotates many times more proteins as nuclear matrix-associated (∼400) than other public databases such as UniProt (∼80 NM proteins), the 'nuclear protein database' (34) (27 NM proteins) or the S/MARt-db (25) (80 NM proteins). We manually update NMPdb once a week at the moment and hope to maintain at least monthly updates for the years to come.

ACKNOWLEDGEMENTS

Thanks to Jinfeng Liu and Megan Restuccia (Columbia) for computer assistance. Thanks to Amos Bairoch (SIB, Geneva), Rolf Apweiler (EBI, Hinxton), Phil Bourne (San Diego University) and their crews for maintaining excellent databases and to all experimentalists who enabled this database by publishing their nuclear matrix related results. This work was supported by grants R01-GM63029-01 from the National Institute of Health (NIH), R01-LM07329-01 from the National Library of Medicine (NLM) and DBI-0131168 from the National Science Foundation (NSF).

REFERENCES

1. Smetana,K., Steele,W. and Busch,H. (1963) A nuclear ribonucleoprotein network. *Exp. Cell Res.*, **31**, 198–201.
2. Belgrader,P., Siegel,A. and Berezney,R. (1991) A comprehensive study on the isolation and characterization of the HeLa S3 nuclear matrix. *J. Cell Sci.*, **98**, 281–291.
3. Collins,J. and Chu,A. (1987) Binding of the DNA polymerase alpha-DNA primase complex to the nuclear matrix in HeLa cells. *Biochemistry*, **26**, 5600–5607.
4. Vincent,M., Lauriault,P., Dubois,M., Lavoie,S., Bensaude,O. and Chabot,B. (1996) The nuclear matrix protein p255 is a highly phosphorylated form of RNA polymerase II largest subunit which associates with spliceosomes. *Nucleic Acids Res.*, **24**, 4649–4652.
5. Okorokov,A.L.R.C., Metcalfe,S. and Milner,J. (2002) The interaction of p53 with the nuclear matrix is mediated by F-actin and modulated by DNA damage. *Oncogene*, **21**, 356–367.
6. Balajee,A., May,A. and Bohr,V. (1998) Fine structural analysis of DNA repair in mammalian cells. *Mutat. Res.*, **404**, 3–11.
7. Pederson,T. (1998) Thinking about a nuclear matrix. *J. Mol. Biol.*, **277**, 147–159.
8. Pederson,T. (2000) Half a century of 'the nuclear matrix'. *Mol. Biol. Cell*, **11**, 799–805.
9. Hancock,R. (2000) A new look at the nuclear matrix. *Chromosoma Focus*, **109**, 219–225.
10. Hendzel,M.J., Boisvert,F.M. and Bazett-Jones,D.P. (1999) Direct visualization of a protein nuclear architecture. *Mol. Biol. Cell*, **10**, 2051–2062.
11. Jackson,D.A., Dickinson,P. and Cook,P.R. (1990) Attachment of DNA to the nucleoskeleton of HeLa cells examined using physiological conditions. *Nucleic Acids Res.*, **18**, 4385–4393.
12. Nickerson,J. (2001) Experimental observations of a nuclear matrix. *J. Cell Sci.*, **114**, 463–474.
13. Barboro,P., D'Arrigo,C., Diaspro,A., Mormino,M., Alberti,I., Parodi,S., Patrone,E. and Balbi,C. (2002) Unraveling the organization of the internal nuclear matrix: RNA dependent anchoring of NuMa to a lamin scaffold. *Exp. Cell Res.*, **279**, 202–218.
14. Mattern,K.A., Humbel,B.M., Muijsers,A.O., de Jong,L. and Van Driel,R. (1996) hnRNP proteins and B23 are the major proteins of the internal nuclear matrix of HeLa S3 cells. *J. Cell Biochem.*, **62**, 275–289.
15. Wan,K., Nickerson,J., Krockmalnic,G. and Penman,S. (1999) The nuclear matrix prepared by amine modification. *Cell Biol.*, **96**, 933–938.
16. Nickerson,J., Krockmalnic,G., Wan,K. and Penman,S. (1997) The nuclear matrix revealed by eluting chromatin from a cross-linked nucleus. *Proc. Natl Acad. Sci. USA*, **94**, 4446–4450.
17. Brünagel,G., Vietmeier,B., Bauer,A., Schoen,R. and Getzenberg,R. (2002) Identification of nuclear matrix protein alterations with human colon cancer. *Cancer Res.*, **62**, 2437–2442.
18. Sanvito,F.V.F., Gambini,S., Santambrogio,G., Catena,M., Viale,E., Veglia,F., Donadini,A., Biffo,S. and Marchisio,P.C. (2000) Expression of a highly conserved protein, p27BBP, during the progression of human colorectal cancer. *Cancer Res.*, **60**, 510–516.
19. Apweiler,R., Bairoch,A., Wu,C., Barker,W., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
20. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.

21. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. *et al.* (2002), *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.
22. Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucleic Acids Res.*, **31**, 3789–3791.
23. 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.
24. Hoogland,C., Mostaguir,K., Sanchez,J.C., Hochstrasser,D.F. and Appel,R.D. (2004) SWISS-2DPAGE, ten years later. *Proteomics*, **4**, 2352–2356.
25. Liebich,I., Bode,J., Frisch,M. and Wingender,E. (2002) S/MARt DB: a database on scaffold/matrix attached regions. *Nucleic Acids Res.*, **30**, 372–374.
26. Liu,J. and Rost,B. (2004) CHOP Proteins in to structure domain-like fragments *Proteins*, **55**, 678–686.
27. Liu,J. and Rost,B. (2004) CHOP: parsing proteins into structural domains. *Nucleic Acids Res.*, **32**, W569–W571.
28. Rost,B. (1996) PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol.*, **266**, 525–539.
29. Rost,B., Yachdav,G. and Liu,J. (2004) The PredictProtein server. *Nucleic Acids Res.*, **32**, W321–W326.
30. Parry,D. (1982) Coiled-coils in alpha-helix-containing proteins: analysis of the residue types within the heptad repeat and the use of these data in the prediction of coiled coils in other proteins. *Biosci. Rep.*, **2**, 1017–1024.
31. Liu,J., Tan,H. and Rost,B. (2002) Loopy proteins appear conserved in evolution. *J. Mol. Biol.*, **322**, 53–64.
32. Liu,J. and Rost,B. (2003) NORSp: predictions of long regions without regular secondary structure. *Nucleic Acids Res.*, **31**, 3833–3835.
33. Carter,P., Liu,J. and Rost,B. (2003) PEP: Predictions for Entire Proteomes. *Nucleic Acids Res.*, **31**, 410–413.
34. Dellaire,G., Farral,R. and Bickmore,W. (2003) The nuclear protein database (NPD): sub-nuclear localization and functional annotation of the nuclear proteome. *Nucleic Acids Res.*, **31**, 328–330.