Nucleic Acids Research, 2003, Vol. 31, No. 1 397–399
DOI: 10.1093/nar/gkg001

NLSdb: database of nuclear localization signals

Rajesh Nair[1,2,*], Phil Carter[1] and Burkhard Rost[1,3,4]

[1]CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, BB217, 650 West 168th Street, New York, NY 10032, USA, [2]Department of Physics, Columbia University, 538 West 120th Street, New York, NY 10027, USA, [3]Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St Nicholas Avenue, New York, NY 10032, USA and [4]North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, BB217, 650 West 168th Street, New York, NY 10032, USA

Received August 15, 2002; Accepted September 11, 2002

ABSTRACT

NLSdb is a database of nuclear localization signals (NLSs) and of nuclear proteins. NLSs are short stretches of residues mediating transport of nuclear proteins into the nucleus. The database contains 114 experimentally determined NLSs that were obtained through an extensive literature search. Using '*in silico* mutagenesis' this set was extended to 308 experimental and potential NLSs. This final set matched over 43% of all known nuclear proteins and matches no currently known non-nuclear protein. NLSdb contains over 6000 predicted nuclear proteins and their targeting signals from the PDB and SWISS-PROT/TrEMBL databases. The database also contains over 12 500 predicted nuclear proteins from six entirely sequenced eukaryotic proteomes (*Homo sapiens, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana* and *Saccharomyces cerevisiae*). NLS motifs often co-localize with DNA-binding regions. This observation was used to also annotate over 1500 DNA-binding proteins. NLSdb can be accessed via the web site: http://cubic.bioc.columbia.edu/db/NLSdb/.

INTRODUCTION

Extraction and testing of NLS motifs

Proteins are actively transported into the nucleus by binding to specific molecules such as importins and karyopherins that recognize distinct targeting signals (1). The targeting signal is usually a short stretch of consecutive residues and is commonly referred to as the nuclear localization signal (NLS). Experimentally best characterized are mono-partite and bi-partite motifs. Most mono-partite motifs are characterized by a cluster of positively charged residues preceded by a helix-breaking residue. Most bipartite motifs consist of two clusters of basic residues separated by 9–12 residues. Over the last few years a large number of distinct NLSs have been experimentally implicated in nuclear transport (2,3). However, NLSs have been experimentally determined for fewer than 10% of known nuclear proteins. To remedy this situation we devised a procedure of '*in silico* mutagenesis' to discover new NLSs (4). Briefly this procedure works as follows. (i) Change or remove some residues from the experimentally characterized NLS motifs and monitor the resulting true (nuclear) and false (non-nuclear) matches. Obviously, allowing alternative residues at particular positions increased the number of nuclear proteins found. However, often this also increased the number of matching non-nuclear proteins. (ii) Discard any potential NLSs that are found in known non-nuclear proteins (false matches). (iii) Require that potential NLSs be found in at least two distinct nuclear families. The 194 potential NLSs discovered using this procedure increased the coverage of known nuclear proteins to 43%. All proteins in the PDB and SWISS-PROT/TrEMBL (5) database were annotated using the full list of experimental and potential NLSs. We also annotated all sequences in the yeast, worm, fruit fly and human proteomes. Approximately 20% of the NLS motifs were observed to co-localize with experimentally determined DNA-binding region of proteins (4,6). These motifs were used to annotate DNA-binding proteins.

General interest

NLSdb is a comprehensive source of information regarding NLSs and proteins translocated into the nucleus by signal sequences. Targeting signal recognition is a key control point in the regulation of nuclear transport. A database of NLS motifs is therefore a useful resource for biologists in identifying targeting signals in their sequence. The database describes all experimentally determined NLS motifs with links to original references in PubMed (7). The information provided by our tool has already been useful for experimental studies of nuclear targeting.

*To whom correspondence should be addressed. Tel: +1 212 305 3773; Fax: +1 212 305 7932; Email: nair@cubic.bioc.columbia.edu

DATABASE DESCRIPTION

Interface

The data are stored and managed using the portal of the Sequence Retrieval System (SRS) (8). SRS provides a convenient and robust framework for managing molecular databases. This provides users with quick, efficient search, retrieval and display methods that work for any web browser. Using SRS, the information in NLSdb can be easily integrated with other public and proprietary databases. The database is continuously updated and refined from the primary literature.

Format and fields

NLSdb has been formated in an EMBL-like flat-file format, thus allowing indexing of the database in SRS (8). Each NLSdb entry describes a nuclear localization signal. Each entry is organized into six major fields; (i) Origin; (ii) Annotation; (iii) Reference; (iv) Confidence; (v) Proteins; and (vi) DNA binding. The 'Origin' field describes whether the NLS has been found by direct experiments, or if it is a potential NLS discovered through our '*in silico* mutagenesis'. For experimentally determined NLSs, further information is provided in the fields 'Annotation' and 'Reference'. The 'Annotation' field describes the protein family in which the experimental NLS was first established and the 'Reference' field gives the primary literature citation. The 'Reference' field also contains a link to PubMed for each citation. The 'Confidence' field is an indicator of our confidence in the NLS; it consists of two sub-fields; 'Total confidence' and '% Nuclear'. 'Total confidence' is the number of localization annotated proteins from SWISS-PROT/TrEMBL in which this NLS is found and '% Nuclear' is the percentage of these that are annotated as nuclear in SWISS-PROT/TrEMBL. The 'Proteins' field lists proteins from various databases that are likely to be targeted to the nucleus since they match the given NLS motif. Currently the 'Proteins' field contains proteins from the SWISS-PROT/TrEMBL, PDB and the PEP (9) databases. All protein entries are linked to the original entries in the respective databases. The 'DNA binding' field describes whether the NLS overlaps with known DNA-binding regions of proteins. NLSdb can be browsed either starting with the NLS entries or with any of the data-fields defined above.

Searches

All data-fields in NLSdb can be searched using standard Boolean queries. Proteins in NLSdb can be identified through their SWISS-PROT/TrEMBL, PDB or PEP identifiers. NLS motifs can be queried by providing a string of one-letter amino acid codes. Database entries can be downloaded using the save 'Complete entries' functionality in SRS.

Annotations for entirely sequenced eukaryotic proteomes

Using the full set of experimental plus potential (discovered through '*in silico*' mutagenesis) NLS motifs in NLSdb, we found over 12 500 proteins with NLS in six entirely sequenced eukaryotic proteomes (Table 1).

Table 1. Nuclear proteins in six entirely sequenced proteomes

Organism	Nprot[a]	Nprot with NLS[b]
Arabidopsis thaliana (plant)	25 456	2390
Caenorhabditis elegans (worm)	18 898	1915
Drosophila melanogaster (fly)	14 184	1274
Mus musculus (mouse)	28 096	2538
Homo sapiens (human, partial)	31 073	4122
Saccharomyces cerevisiae (yeast)	6306	482
Sum	124 013	12 721

[a]Nprot: number of proteins in proteome.
[b]Nprot with NLS: number of proteins with NLS in the proteome.

CONCLUSIONS

NLSdb can greatly help in better understanding signal dependent nuclear transport of proteins. The potential NLS motifs discovered through '*in silico*' mutagenesis can aid in discovering new signal sequences involved in nuclear targeting. A future goal is to integrate NLSdb with all sequences in the SWISS-PROT/TrEMBL database and all proteomes in the PEP database.

NLSdb should be cited with the present publication as reference. The database can be accessed through the World Wide Web at: http://cubic.bioc.columbia.edu/db/NLSdb/.

ACKNOWLEDGEMENTS

Thank you to Jinfeng Liu (Columbia University) for computer assistance and the collection of genome data sets and to Jinfeng Liu and Dariusz Przybylski (Columbia University) for providing preliminary information and programs. P.C. and B.R. were supported by the grant 1-P50-GM62413-01 from the National Institutes of Health (NIH); R.N. and B.R. were supported by the grant DBI-0131168 from the National Science Foundation (NSF). Last, but not least, thank you to Amos Bairoch (SIB, Geneva) and Rolf Apweiler (EBI, Hinxton) and their crews for maintaining excellent databases and to all experimentalists without whom we could not have built our database.

REFERENCES

1. Tinland,B., Koukolikova-Nicola,Z., Hall,M.N. and Hohn,B. (1992) The T-DNA-linked VirD2 protein contains two distinct functional nuclear localization signals. *Proc. Natl Acad. Sci. USA*, **89**, 7442–7446.
2. Mattaj,I.W. and Englmeier,L. (1998) Nucleocytoplasmic transport: the soluble phase. *Annu. Rev. Biochem.*, **67**, 265–306.
3. Jans,D.A., Xiao,C.Y. and Lam,M.H. (2000) Nuclear targeting signal recognition: a key control point in nuclear transport? *Bioessays*, **22**, 532–544.
4. Cokol,M., Nair,R. and Rost,B. (2000) Finding nuclear localization signals. *EMBO Rep.*, **1**, 411–415.
5. Bairoch,A. and Apweiler,R. (2000) Nuclear localization signals overlap DNA- or RNA-binding domains in nucleic acid-binding proteins. *Nucleic Acids Res.*, **28**, 45–48.

6. LaCasse,E.C. and Lefebvre,Y.A. (1995) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **23**, 1647–1656.
7. Airozo,D., Allard,R., Brylawski,B., Canese,K., Kenton,D., Knecht,L., Krasnov,S., Sandomirskiy,V., Sirotinin,V., Starchenko,G. *et al.* (1999). MEDLINE. National Library of Medicine (NLM), Vol. 1999.
8. Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.
9. Carter,P., Liu,J. and Rost,B. (2003) PEP: Predictions for Entire Proteomes. *Nucleic Acids Res.*, **31**, 410–413.