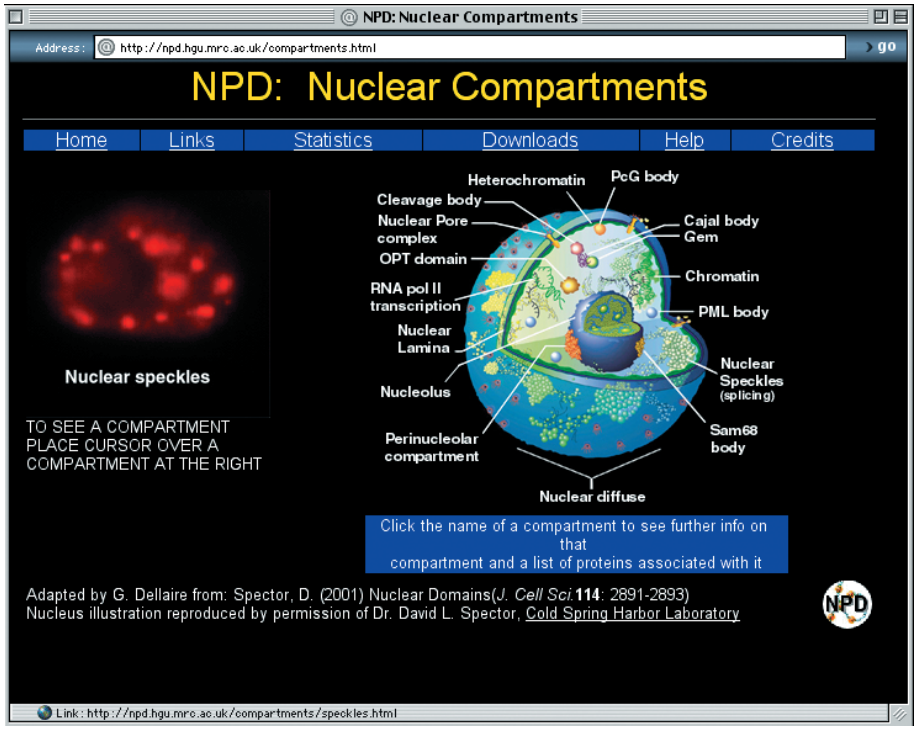
Nuclear protein database (NPD)

http://npd.hgu.mrc.ac.uk



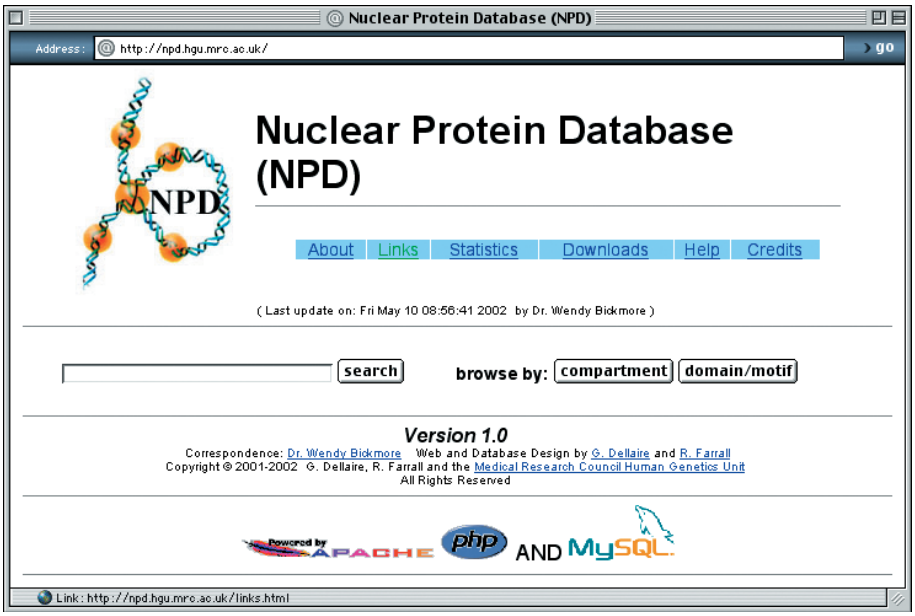
Molecular biologists have studied biological processes for decades by analyzing single proteins and asking how they interact with other single factors. This 'bottom-up' approach has been greatly successful. However, with the availability of information about complete genomes and, increasingly, proteomes, more global approaches, in which all components acting in a pathway are investigated simultaneously, are rapidly becoming powerful tools to uncover complex biological pathways. Such 'top-down' approaches rely on the availability of large amounts of data and, more importantly, they are crucially dependent on the accessibility of relevant data. After all, what good is a mountain of information if nobody can get to it? An indication of the importance of data organization is the fact that much effort has recently gone into the development of manageable database systems. A key question when developing databases is what should go in a database. Too little information limits the usefulness of a database and too much information makes it hard to manage. Although there is an obvious need for databases that contain information about entire genomes and it makes sense to have some central databases such as SWISS-PROT, a strong point can be made for creating sub-databases, which are more accessible and contain more specific information. A good example of a relatively small, but focused and highly practical database is the recently launched NPD – nuclear protein database.

The idea of NPD was conceived and realized by Wendy Bickmore and colleagues at the MRC Human Genetics Unit, Edinburgh. The origin of the NPD was a project in the Bickmore group in which a gene-trap method was used to identify more than 100 novel nuclear proteins based on their localization within the nucleus (Tate et al., 1998; Sutherland et al., 2001; Bickmore and Sutherland, 2002). This information has now been combined with publicly available data to create NPD, which contains information about more than 1000 nuclear proteins. Special emphasis is placed on the subnuclear distribution pattern and the protein domain structure of nuclear proteins. The NPD is publicly available and is free of charge.

Searching NPD is easy. The homepage contains a search field that offers the choice of searching by keywords, subnuclear compartment or protein domain motif. The simplicity of the starting page is very practical and most searches using various available combinatorial searching commands work well. While a keyword search directly generates a list of relevant proteins, the compartment search option first leads to a window containing a schematic view of the nucleus, from which specific nuclear compartments can be selected. Each compartment is introduced with a concise description, a feature that is helpful for those less familiar with the nucleus. All proteins associated with a particular compartment can then be displayed. Similarly, the protein domain search option leads to an extensive alphabetical list of protein domains which, in turn, link to a display of proteins containing the domain of interest. These two short-cut options are convenient although somewhat redundant. In a small test, the number of hits using a keyword search was generally similar to that of a short-cut search, although in some cases the results were drastically different, especially for the compartment option. The fact that not all nuclear compartments are listed can also be somewhat confusing.

Regardless of the search approach taken, the result is ultimately a list of proteins identified by the search criteria. Each protein entry is displayed with several keywords regarding localization, function and species of origin. This information is generally sufficient to decide whether a particular protein is of interest or not. But it is the next level that provides the real 'meat' of the search. For each entry, information about the protein's DNA and protein sequence, domain structure, function, subnuclear localization and expression pattern are given together with references to related genes. While much of this information is provided in the form of keywords, retrieving the original data and sequence information is easily done using direct links to PubMed, NCBI, OMIM, PFAM, SWISS-PROT and SMART. In addition,

The URL given was correct at the time of going to press; any changes will be recorded in the supplementary data associated with this article (http://jcs.biologists.org/supplemental).



biological and molecular functions of the proteins are described using the controlled vocabulary of the Gene Ontology (GO) terms. The display of the information is very user friendly in that it is visually pleasing and complete but still manageable. One convenience that is missing at the moment is the ability to refine a search by selecting a subset of proteins of interest for further inspection.

In addition to information about single proteins, NPD is a comprehensive source of more general information about nuclear proteins and the nucleus. An overview of known subnuclear compartments is given in the compartment section. In a separate section called ‘statistics’ there is a short discussion of the factors that influence protein distribution. Here can be found an analysis of the correlation between the distribution of nuclear proteins and protein domain composition, molecular weight and pI. For example, proteins in splicing factor compartments generally contain an RNA-binding domain or an RS-domain and are often basic – indeed, the majority of proteins in the database with pI>11 are concentrated in this compartment. This simple analysis is a first step towards deducing general rules for what determines where proteins localize, and it is hoped that this section of the site will be expanded in the future. These two general features and the extensive list of links to websites relevant to cell biological aspects of the nucleus, nuclear structure laboratories, protocols, companies and related databases, make NPD more than just a repository for information on nuclear proteins and ensure that it will serve as a gateway for anyone interested in nuclear proteins or nuclear function.

NPD is a work in progress and some aspects could be improved. For example, the PubMed references given are very selective and are not always the most relevant ones. Furthermore, additional links with other databases that focus on protein localization would be highly beneficial. Despite these childhood maladies, NPD is well on its way to becoming an indispensable tool for those working on nuclear processes and, maybe even more importantly, for those not familiar with the nucleus. NPD has a tremendous potential to become a key tool for many cell biologists. How much further NDP can be developed will largely depend on the availability of external funding. It is to be hoped that funding agencies begin to realize the important role of databases in modern biology, and that they earmark resources for the development of databases such as NDP. This would be money well spent.

References

Bickmore, W. A. and Sutherland H. G. E. (2002). Addressing protein localization within the nucleus. *EMBO J.* **21**, 1248-1254.

Sutherland, H. G., Mumford, G. K., Newton, K., Ford, L. V., Farrall, R., Dellaire, G., Caceres, J. F. and Bickmore, W. A. (2001). Large-scale identification of mammalian proteins localized to nuclear sub-compartments. *Hum. Mol. Genet.* **10**, 1995-2011.

Tate, P., Lee, M., Tweedie, S., Skarnes, W. C. and Bickmore, W. A. (1998). Capturing novel mouse genes encoding chromosomal and other nuclear proteins. *J. Cell Sci.* **111**, 2575-2585.

Tom Mistelli
National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

Journal of Cell Science 115, 2805-2806 (2002)