NUREBASE: database of nuclear hormone receptors

Jorge Duarte, Guy Perrière1, Vincent Laudet and Marc Robinson-Rechavi*

Laboratoire de Biologie Moléculaire et Cellulaire, CNRS UMR 5665, École Normale Supérieure de Lyon, 46 allée d’Italie, 69364 Lyon Cedex 07, France and 1Laboratoire de Biométrie et Biologie Évolutive, CNRS UMR 5558, Université Claude Bernard, Lyon 1, 43 boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

Received August 13, 2001; Revised and Accepted October 10, 2001

ABSTRACT

Nuclear hormone receptors are an abundant class of ligand activated transcriptional regulators, found in varying numbers in all animals. Based on our experience of managing the official nomenclature of nuclear receptors, we have developed NUREBASE, a database containing protein and DNA sequences, reviewed protein alignments and phylogenies, taxonomy and annotations for all nuclear receptors. The reviewed NUREBASE is completed by NUREBASE DAILY, automatically updated every 24 h. Both databases are organized under a client/server architecture, with a client written in Java which runs on any platform. This client, named FamFetch, integrates a graphical interface allowing selection of families, and manipulation of phylogenies and alignments. NUREBASE sequence data is also accessible through a World Wide Web server, allowing complex queries. All information on accessing and installing NUREBASE may be found at http://www.ens-lyon.fr/LBMC/laudet/nurebase.html.

INTRODUCTION

Nuclear hormone receptors are one of the most abundant classes of transcriptional regulators in metazoans, in which they regulate functions as diverse as reproduction, differentiation, development, metabolism, metamorphosis or homeostasis (1). They function as ligand-activated transcription factors, thus providing a direct link between signaling molecules that control these processes and transcriptional responses. Nuclear receptors form a superfamily of phylogenetically related proteins, which share a common structural organization: a variable N-terminal region (A/B domain), a central well conserved DNA binding domain (DBD, C domain), a non-conserved hinge (D domain) and a C-terminus, moderately conserved ligand binding domain (LBD, E domain) (1). The superfamily includes receptors for hydrophobic molecules such as steroid hormones (estrogens, glucocorticoids, progestosterone, mineralocorticoids, androgens, vitamin D, ecdysone, oxyysterols, bile acids, etc.), retinoic acids (all-trans and 9-cis isoforms), thyroid hormones, fatty acids, leukotrienes and prostaglandins (2). A large number of nuclear receptors have also been identified by homology with the conserved DBD and LBD, but have no identified natural ligand, and are referred to as ‘nuclear orphan receptors’. As nuclear receptors bind small molecules which can easily be modified by drug design, and control functions associated with major pathologies (cancer, osteoporosis, diabetes, etc.), they are promising pharmacological targets. The search of ligands for orphan receptors and the identification of novel signaling pathways has become a very active research field (3-4). Their role in the control of animal development makes them major players for understanding animal evolution (5) or genomics (6).

The importance of nuclear receptors has prompted the accumulation of rapidly increasing data from a great diversity of fields of research: sequences, expression patterns, three-dimensional structures, protein–protein interactions, target genes, physiological roles, mutations, etc. These data are highly dispersed, in a variety of formats. The aim of NUREBASE is to present an integrated database, with a unique, interactive interface, centralizing up-to-date information about nuclear receptors for the specialist and the non-specialist.

There are 21 nuclear receptors in the complete genome of the fly Drosophila melanogaster (7), less than 50 in humans (8), but more than 250 in the nematode Caenorhabditis elegans (9,10). This diversity has been officially organized in a phylogeny-based nomenclature (11), of which one of us (V.L.) is in charge. An important aim of our database is thus to facilitate use of the official nomenclature, notably for new nuclear receptors.

To answer these needs, we built the NUREBASE database of nuclear receptors. It contains all protein and DNA sequences, reviewed protein alignments and phylogenetic trees, and additional information such as nomenclature, domains and natural ligands.

DATABASE CONTENTS

Release 1.0 (August 2001) of NUREBASE contains 361 nuclear receptor protein sequences without redundancy, from 88 metazoan species. Divergent nematode sequences (9) will be incorporated as they are characterized experimentally and classified in the nomenclature (11). The sequences are grouped into ‘families’, corresponding to levels of nomenclature (11), with the following for each family:

1. A NUREBASE number, of the form NRBaabbcc, in which aa is 01 for all nuclear receptors with a DBD and an LBD and 00 for those which lack one or the other [families NR0 of the Nuclear Receptors Nomenclature Committee (11)], bb is the family number or 00 for receptors of all families, and cc is the sub-family number or 00 for receptors of all sub-families.

*To whom correspondance should be addressed. Tel: +33 472 72 86 85; Fax: +33 472 72 80 80; Email: marc.robinson@ens-lyon.fr
Each nuclear receptor can belong to several encased NUREBASE families; for example, thyroid hormone receptor α, NR1A1, belongs to NRB010000, NRB010100 and NRB010101.

2. A textual definition, describing the contents of the NUREBASE family. For example, NRB010101 is ‘Sub-family NR1A: TR: THA, THB, NR1A3’.

3. An alignment of protein sequences, checked by eye with SEAVIEW (12). Sequences are first aligned to their sub-family, then sub-families aligned in a family, and finally all families aligned. Thus, the variable A/B domain can be correctly aligned between closely related sequences (human and mouse thyroid hormone receptor α), even though it cannot be aligned across the whole superfamily.

4. A phylogeny based on the protein sequences. Phylogenies are obtained by Neighbor-Joining (13) with Poisson corrected distances, and checked taking into account bootstrap support, known taxonomic relations and other tree-building methods (6,8).

There are two additional NUREBASE families, NRB020001 and NRB020002, containing all nuclear receptors from the human and fly genomes, respectively. This brings the total number of NUREBASE families in release 1.0 to 35, each with an alignment and a phylogeny.

Protein sequence names were modified in a manner similar to SWISS-PROT (14), starting with four characters describing the gene in a standardized way, an underscore, three characters for the genus, one character for the species, and a digit (1 by default) to manage cases of identical names. The gene name corresponds to the most common acronym for well-known nuclear receptors, and to the nomenclature otherwise. For example, human thyroid hormone receptor α (NR1A1) is THA_HOMS1, but the human steroidogenic factor (NR5A1) is 5A1_HOMS1. For each protein sequence, NUREBASE has at least one reference to a corresponding DNA sequence in EMBL format (15) containing the original information from this database. Sequence annotations are also enriched with the official nomenclature, the name of the natural ligand (estrogen, thyroid hormone, etc., or ‘orphan’), the definition of the DBDs and LBDs, and the NUREBASE family number. The sum of this information makes a NUREBASE entry. These entries are integrated into two ACNUC (16) databases, one for protein sequences and one for DNA.

To accommodate the need for frequent updates, as well as reviewed data, we have created a second database, NUREBASE_DAILY. It contains all the sequences and annotations of the reviewed NUREBASE, plus new sequences detected by an automatic update procedure. This procedure is launched every 24 h. It first downloads the daily release of GenPept, the translation of new coding sequences from NCBI launched every 24 h. It first downloads the daily release of GenPept, the translation of new coding sequences from NCBI

DATABASE ACCESS

The main access to NUREBASE is through the interface initially developed for the HOBACGEN database (20). This system is build under a client/server architecture, avoiding the need to install the complete database on the users’ computer (Fig. 1). The server side is made of three components: a World Wide Web service, a dedicated C program to access the data, and the data itself. A complete description of the World Wide Web service and the C program has been published previously (20).

The client is the FamFetch Java application, which allows wide portability and interactivity. The main window of the interface allows selection of one of the NUREBASE families described above (Fig. 2). It is possible to make queries to define a subset of families matching specific criteria, such as species or keywords. Selection of a family prompts a tree window with the corresponding phylogeny. In this tree, sequences are colored according to taxonomy. The tree display is active, with options of re-rooting, node swapping or subtree selection. Clicking on leaves allows selection of one or several nuclear receptors in a new window. From there, the user may view the DNA or protein NUREBASE entry, or the protein alignment. The alignment contains only those sequences selected by the user, and is not computed but reconstructed from the pre-existing whole family multiple alignment. Functions allow the user to save lists of families, sequences entries, alignments or trees in text files.

The two ACNUC databases of NUREBASE_DAILY are accessible online through the PBIL (Pôle Bio-Informatique
Advanced users can also install locally the complete database, and use the Query (16) and Query_win (22) retrieval programs allowing complex queries,
although the benefit of regular updates is thus lost. The results of queries can then be directly used in the FamFetch interface, as explained by Perrière et al. (20).

All information on accessing and installing NUREBASE may be found at http://www.ens-lyon.fr/LBMC/laudet/nurebase.html.

ACKNOWLEDGEMENTS
We thank François Bonneton and Hector Escrivá for critical reading. NUREBASE is supported by the CNRS. J.D. is supported by formation continue of Université de Rouen.

REFERENCES
23. Martinez, E., Moore, D.D., Keller, E., Pearce, D., Vanden Heuvel, J.P.,
    Robinson, V., Gottlieb, B., MacDonald, P., Simons, S., Jr, Sanchez, E.

    biological data: the GPCRDB and NucleaRDB information systems.

    method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, 
    **302**, 205–217.

    maximum likelihood method for reconstructing tree topologies.