NucleaRDB: information system for nuclear receptors

Bas Vroling¹,², David Thorne³, Philip McDermott³, Henk-Jan Joosten², Teresa K. Attwood⁴, Steve Pettifer³ and Gert Vriend¹,*

¹CMBI, NCMLS, Radboud University Nijmegen Medical Centre, Nijmegen, ²Bio-Prodict, Dreijenplein 10, 6703 HB Wageningen, The Netherlands, ³School of Computer Science and ⁴Faculty of Life Sciences, University of Manchester, Manchester, M13 9PL, UK

Received September 6, 2011; Revised October 9, 2011; Accepted October 13, 2011

ABSTRACT

The NuleaarDDB is a Molecular Class-Specific Information System that collects, combines, validates and disseminates large amounts of heterogeneous data on nuclear hormone receptors. It contains both experimental and computationally derived data. The data and knowledge present in the NuleaarDDB can be accessed using a number of different interactive and programmatic methods and query systems. A nuclear hormone receptor-specific PDF reader interface is available that can integrate the contents of the NuleaarDDB with full-text scientific articles. The NuleaarDDB is freely available at http://www.receptors.org/nucleardb.

INTRODUCTION

Nuclear receptors (NRs) are ligand-inducible transcription factors that regulate processes, such as homeostasis, differentiation, embryonic development and organ physiology. A total of 49 human NRs have been identified (1). Their ligands are lipophilic compounds such as steroids, thyroid hormone, vitamin D3 and retinoids (2). The endogenous ligands are not yet known for 30% of the NRs (3). As nuclear receptors are involved in almost all aspects of human physiology and are implicated in many important diseases including cancer, diabetes and osteoporosis, understanding of these receptors has major implications for human biology and for the development of new drug treatments. Nuclear receptors are targets for pharmaceutical industries with similar importance (4), as the G protein-coupled receptors (GPCRs), ion channels and kinases.

Due to the increasing amounts of experimental and computational data buried in numerous databases and scientific articles, the task of extracting, combining and validating this data is becoming an increasingly large hurdle for the individual scientist. Databases that revolve around a single protein family can help researchers in using all data needed for their research, while relieving them of the onerous tasks related to the retrieval of many data from different sources (5).

The NuleaarDDB is a data source that holds many different data types (Table 1) in a well organized and easily accessible form (6). The data are validated, internally consistent and updated regularly. The NuleaarDDB provides access to the data via various interfaces, which depending on the users’ needs, are suited either for automated access or interactive usage.

DATA CONTENTS

Primary data

The NuleaarDDB contains three different primary data types: sequences, structures and mutations. Sequences and structures were updated as described previously (7). Mutation data was obtained from the Nuclear Receptor Mutation Database (8) and fully integrated in the NuleaarDDB. In addition, a large body of mutations was extracted from literature by the software package MuteXt (9).

Computational data

A large and diverse collection of computationally generated data are present in the NuleaarDDB. Multiple sequence alignments (MSAs) form the heart of the system and allow users to easily transfer information between different proteins. MSAs are available for all families and subfamilies, and can be viewed using JalView (10) or can be directly downloaded in a number of formats. MSAs were created as described previously (7).

Correlated mutation analyses (CMA) can be used to identify groups of residues that mutate in tandem. Residues that show correlated mutation behavior are

*To whom correspondence should be addressed. Tel: +31 (0)24 36 19390; Fax: +31 (0)24 36 19395; Email: vriend@cmbi.ru.nl

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
likely to be functionally related, and networks of those correlating residues indicate functional units (11). Correlation scores are available for all (sub-)families.

The entropy and variability for a position in a MSA can be an indicator of the evolutionary pressures exerted at that position (12). Entropy and variability scores are available in tabular form and via an interactive page displaying an integrated view via plots, tables and structure models.

In addition to the already large amount of structural information that is present in the NucleaRDB, homology models based on multiple template structures have been built for all NRs. All structure models were built using YASARA (13) and are available for download or can be viewed directly using Jmol (14).

**INFORMATION RETRIEVAL**

All data in the NucleaRDB web interface are extensively connected, allowing for easy navigation between different data types. The main way of accessing the NucleaRDB's contents is via the hierarchical family tree. For each family, users can access the individual receptors, multiple sequence alignments (and all derived data and analyses such as correlation scores and protein distance networks), mutations, structures and models (Figure 1). All pages contain links to all related data and information. Extensive search facilities are available, allowing the search for proteins, sequences, structures, families and mutations using various search criteria and filters. A BLAST service is available that allows users to run their own sequences against the NucleaRDB.

All data types and search facilities are accessible from the web pages as well as from the web service endpoints, allowing users to write workflows or in-house software that uses the NucleaRDB.

**Annotating scientific literature**

Utopia Documents (15,16) is a new PDF reader that offers unique opportunities to place information and knowledge in the context of scientific literature. We have integrated the NucleaRDB with the Utopia Documents PDF reader in such a way as to present to scientists, in a non-intrusive way, all NR-relevant data and information discussed in an
In the crystal structure, residues whose side-chains interact with other residues by intramolecular hydrogen bonds are R274, S275, W286, H305 and Q400. Guanidine of R274 interacts with T142 at the tail of H1 by a hydrogen bond which is conserved in NR subfamily I except for PPAR. We assumed that this bond is not necessarily essential because similar interaction was not observed in PPAR in subfamily I and NRs in other subfamilies. The hydroxyl group of S275 forms hydrogen bonds with indole of W286,imidazole of H305 and the side chain carbonyl group of Q400. The Q400 side chain also forms a hydrogen bond with the hydroxyl group of S306. We assumed that these hydrogen bonds are not crucial for holding the 3-D structure of VDR because the mutants of S275 and Q400, that is, S275A and Q400A, had little effect on transactivation potency as described below. Thus, nine mutants, S275A, R274A, S275A, S275A, C288A, H305A, H397A, Q400A and Y401A, are assumed to have little effect on the folding of the 3-D structure of VDR. Mutants of nonpolar residues, L233A, V234A and W286A, decrease hydrophobic interaction with the ligand and/or intramolecular interaction with other residues and increase the volume of LBP.

Table 1. Contents of the NucleaRDB

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteins</td>
<td>3764</td>
</tr>
<tr>
<td>Families</td>
<td>123</td>
</tr>
<tr>
<td>Mutations</td>
<td>1543</td>
</tr>
<tr>
<td>Protein structures</td>
<td>613</td>
</tr>
<tr>
<td>Structure models</td>
<td>3764</td>
</tr>
<tr>
<td>Residues</td>
<td>2012651</td>
</tr>
<tr>
<td>Species</td>
<td>339</td>
</tr>
</tbody>
</table>

Figure 2. An impression of the Utopia Documents PDF reader interface to the NucleaRDB data. On the left-hand side a part of a scientific paper (17) is shown that is annotated by the NucleaRDB. Annotations are available for all the highlighted words. On the right-hand side an example of such an annotation (the mutation R274A) is displayed.

Troubles associated with navigating the many links between existing data and information available from the many articles in this field. The scientist neither struggles to get access to information related to topics within an article, nor is swamped by unnecessary information that still needs disambiguation; only data and information relevant to the topic of the article is made available.

IMPLEMENTATION

The data in the NucleaRDB is stored in a PostgreSQL (www.postgresql.org) relational database. The web service interface is developed with the Apache CXF (cxf.apache.org) web services framework. We offer both Simple Object Access Protocol and Representational state transfer endpoints. The web interface is built using the Apache Wicket (wicket.apache.org) web application framework. The database is accessed via a Hibernate
(www.hibernate.org) object-relational mapping layer. The server is running within Sun’s Glassfish (www.glassfish.org) application server.

CONCLUSION
The NucleaRDB provides researchers with a single point of access for nuclear receptor-related data. Not only does the NucleaRDB hold a large amount of information, it also provides a broad scope of tools and dissemination facilities, relieving scientist of many of the tasks that come with collecting, validating and integrating many diverse data.

ACKNOWLEDGEMENTS
We thank Maarten Hekkelman, Wilmar Teunissen and Tim teBeek for their support with computer science issues. We thank TIPharma for financial support.

FUNDING
BioRange program of the Netherlands Bioinformatics Centre (NBIC); BSIK grant through the Netherlands Genomics Initiative (NGI); EMBRACE project that is funded by the European Commission within its FP6 Programme, under the thematic area ‘Life sciences, genomics and biotechnology for health’ (contract number LHSG-CT-2004-512092); and TIPharma. Funding for open access charge: RUNMC.

Conflict of interest statement. None declared.

REFERENCES