ABSTRACT
Summary: The database NPIDB (Nucleic Acids—Protein Interaction Database) contains information derived from structures of DNA–protein and RNA–protein complexes extracted from PDB (1834 complexes in July 2007). It is organized as a collection of files in PDB format and is equipped with a web-interface and a set of tools for extracting biologically meaningful characteristics of complexes. The content of the database is weekly updated.
Availability: http://monkey.belozersky.msu.ru/NPIDB/
Contact: sas@belozersky.msu.ru

1 INTRODUCTION
More than 1000 structures of DNA–protein and RNA–protein complexes are currently available in Protein Data Bank (PDB). However, there exist significant difficulties in obtaining information concerning intermolecular contacts, mode of interaction, comparison of related complexes, etc. We created a web-based database for providing an access to adequately organized information about all available structures of these complexes. Also, we developed several algorithms for the analysis of the complexes. Implementations of these algorithms are integrated into the database.

Web databases such as AANT (Hoffman et al., 2004) and Protein–DNA Recognition Database (http://gibk26.bse.kyutech.ac.jp/jouhou/3dinsight/recognition.html) are devoted to facilitate data mining on nucleic acid—protein interaction. These resources provide a number of useful tools for an analysis of protein–DNA interaction. However, there are several requirements that are not met by the existing tools. For example, it would be useful to have an incorporated into a database classification of DNA-binding domains and motifs. Also information should be available on all kinds of DNA–protein contacts, including hydrophobic ones. Our goal is to develop a new Internet resource, which should enlarge the spectrum of available tools in the area of nucleic acid–protein interaction.

The resource NPIDB (Nucleic Acids—Protein Interaction DataBase) includes a collection of files in PDB format containing structural information on DNA–protein and RNA–protein complexes extracted from PDB, and a number of online tools for analysis of the complexes. Those tools are: an original program CluD for analysis of hydrophobic clusters on interfaces (Alexeevski et al., 2004; Karyagina et al., 2006), a program for detecting potential hydrogen bonds, visualization of structures with Jmol and Chime.

2 CONTENT
Structures of protein—nucleic acid complexes are extracted from PDB as files in PDB format representing both PDB entries (asymmetric units) and biological units.

Structures are revised by our experts in order to correct possible mistakes (such as duplication of atoms) and inconvenience (such as two or more variants of a structure posed in one coordinate space, see, e.g. PDB entry 1QPI, where two variants of each DNA chain are superimposed). The manually corrected entries are also included into the database as 'revised biological units'. Thus, in some cases, the NPIDB content differs from the original PDB one; in particular, for some complexes (1FJL, 1QPI, etc.) there are some additional biological units in NPIDB compared with the PDB.

All structural files of NPIDB are downloadable.

Update of the content is done weekly by a special program module.

3 SOFTWARE
We developed an original program module for extracting structures of protein—nucleic acid complexes from PDB. Our module identifies chains of DNA, RNA and protein directly in the coordinate (ATOM and HETATM) section of a PDB file. If a PDB file contains at least one protein chain and at least one nucleic acid chain, it is selected for incorporating into NPIDB.

Weekly, the updating module (i) examines the new entries appeared in the local PDB mirror at A.N. Belozersky Institute by the above described selection tool, (ii) selects protein–nucleic acid complexes and (iii) calls the module of database filling.

We developed a set of Perl scripts for operating with the available complexes. The first script forms the HTML table with the list of available complexes, including information on PDB codes, PDB headers, protein, DNA and RNA chains, PDB entry creation date, and nucleic acid type (DNA, RNA or both); the second sorts this list with respect to each feature and the third generates web pages with general information about individual complexes.

A program is developed, which detects domains in protein chains using information from Pfam and SCOP domain databases. The information on the domain families and their representatives in the database is organized as a set of dynamical web pages generated by special scripts.
A tool for detecting potential hydrogen bonds between a nucleic acid and a protein is integrated into NPIDB. It uses a simplest criterion of hydrogen bond formation: the potential hydrogen bond is detected if the distance between oxygen or nitrogen atoms of different molecules is $< 3.7 \, \text{Å}$.

The program CluD (Alexeevski et al., 2004; Karyagina et al., 2006) for detecting hydrophobic clusters in macromolecular structures is adopted for the work with DNA–protein and RNA–protein interfaces and integrated into NPIDB.

The set of determined hydrophobic clusters in each structure depends on the chosen threshold distance of hydrophobic interaction (the recommended threshold range is 4.5–5.4 Å, the default value is 5.4 Å). The program Profile detects the dependence of those set on the threshold. The result of the program is a tree graph showing the process of cluster appearance and joining as a result of increasing the threshold.

For every Pfam or SCOP domain presented in the database, it is detected, whether this domain is in contact with nucleic acid.

4 DESIGN

The set of available complexes is presented in the form of a list (http://monkey.belozersky.msu.ru/cgi-bin/nuc_prot.pl) that can be sorted by the PDB code, the date of appearing in PDB, the PDB header (classification) or by the type of the nucleic acid (DNA, RNA or both). The list is equipped with a search over keywords appearing in headers and titles of PDB entries. Each complex has its own web page containing general information about the complex, the table of biological units, and shortcuts to all available tools for its analysis. The page is accessible by a mouse-click on the PDB code in the list of entries. Each complex has two corresponding hyperlinks to the CluD page. When a user visits the CluD page via one of these hyperlinks, the service offers two variants of CluD complexes. The available tools can be applied for the initial accessible by a mouse-click on the PDB code in the list of the entire structure, and second, to find clusters at the interface taking into account not only atom-to-atom distance but also angles of the bonds and subtypes of the atoms involved. We also plan to add a tool for calculating water bridges (Karyagina et al., 2005) on nucleic acid–protein interface.

A statistics of residue-base contacts (for every family and overall) will be available at the site.

For every family of domains, a sequence alignment of its members will be available together with the information on secondary structure, contacts with the nucleic acid and other features. Superimposed structures of the family members will be downloadable.

ACKNOWLEDGEMENTS

The work is supported by the Russian Foundation of Basic Research, grants 06–07–89143 and 06–04–49558, and INTAS, grant 05–100008–8028.

Conflict of Interest: none declared.

REFERENCES


