D788–D792 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 4 November 2010*
doi:10.1093/nar/gkq1087

REPAIRtoire—a database of DNA repair pathways

Kaja Milanowska[1,2], Joanna Krwawicz[3], Grzegorz Papaj[1], Jan Kosiński[1,4], Katarzyna Poleszak[1], Justyna Lesiak[1], Ewelina Osińska[1], Kristian Rother[1,2] and Janusz M. Bujnicki[1,2,*]

[1]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland, [2]Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznan, Poland, [3]Department of Molecular Biology, Institute of Biochemistry and Biophysics Polish Academy of Sciences, 5A Pawinskiego, 02-106 Warsaw, Poland and [4]Biocomputing group, Department of Biochemical Sciences, ''Sapienza'' University, P. le A. Moro, 5, 00185 Rome, Italy

Received August 15, 2010; Revised October 14, 2010; Accepted October 15, 2010

ABSTRACT

REPAIRtoire is the first comprehensive database resource for systems biology of DNA damage and repair. The database collects and organizes the following types of information: (i) DNA damage linked to environmental mutagenic and cytotoxic agents, (ii) pathways comprising individual processes and enzymatic reactions involved in the removal of damage, (iii) proteins participating in DNA repair and (iv) diseases correlated with mutations in genes encoding DNA repair proteins. REPAIRtoire provides also links to publications and external databases. REPAIRtoire contains information about eight main DNA damage checkpoint, repair and tolerance pathways: DNA damage signaling, direct reversal repair, base excision repair, nucleotide excision repair, mismatch repair, homologous recombination repair, nonhomologous end-joining and translesion synthesis. The pathway/protein dataset is currently limited to three model organisms: *Escherichia coli*, *Saccharomyces cerevisiae* and *Homo sapiens*. The DNA repair and tolerance pathways are represented as graphs and in tabular form with descriptions of each repair step and corresponding proteins, and individual entries are cross-referenced to supporting literature and primary databases. REPAIRtoire can be queried by the name of pathway, protein, enzymatic complex, damage and disease. In addition, a tool for drawing custom DNA–protein complexes is available online. REPAIRtoire is freely available and can be accessed at http://repairtoire.genesilico.pl/.

INTRODUCTION

DNA repair processes are of crucial importance for the maintenance of the genetic information of all organisms. The stability of the genome is constantly endangered by environmental agents, endogenous metabolic processes, e.g. reactive species inside cells, and errors of cellular processes involving DNA. Modifications of DNA can lead to mutations, which alter the coding sequence of DNA and can lead to cancer in humans and other mammals. Other DNA lesions interfere with normal cellular transactions, such as DNA replication or transcription, and are deleterious to the cell (1,2). To counteract DNA damage, organisms have evolved various damage prevention and repair systems (3–7). These systems ensure the stability of DNA and accurate transmission of genetic information by protecting the genome against a large number of different chemical and structural alterations. At the same time, random changes in DNA are viewed as a main source of genetic variability, and thus a driving force for evolution. In multicellular organisms changes in the DNA sequence and structure are responsible for e.g. differential production of antibodies by the immune system (8). Therefore, DNA repair mechanisms have to balance the noxious against the beneficial effects of alterations in the genome sequence and chemical structure.

It has been proposed that DNA damage from endogenous sources gives rise to 20 000 lesions per mammalian cell per day, most of the lesions being deaminations,

*To whom correspondence should be addressed. Tel: +48 22 597 0750; Fax: +48 22 597 0715; Email: iamb@genesilico.pl

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

spontaneous hydrolysis of the *N*-glycosidic bond, alkylations, and damage by reactive oxygen or nitrogen species and lipid peroxidation products (9–13). Lesions are also caused by errors in DNA metabolic processes, including the formation of single- and double-strand breaks, the collapse of replication forks, and the introduction of modified nucleic acid bases during DNA replication. Counting all together, daily 10^{16}–10^{18} repair events occur in a healthy adult man (10^{12} cells) (14). Despite the protection provided by these mechanisms some of the damage escapes repair, and in consequence leads to mutations, ageing and various diseases, including carcinogenesis and neurodegeneration (15–19).

DNA repair is a very complicated process, involving many factors. For instance to date, 168 genes that encode proteins involved in DNA repair have been identified in the human genome (18–20). They are involved in diverse processes, starting from detection of a damage site in the DNA, through several steps of enzymatic transformation of the damaged DNA, to recombination and signaling to stop the cell cycle or initiate apoptosis. Another form of dealing with the DNA damage is lesion bypass, which facilitates continuation of replication even when irremovable modifications occur, but does not guarantee proper recreation of the original sequence and frequently leads to mutation generated by translesion synthesis (TLS) polymerases.

The numerous chemical and structural transformations leading from damaged DNA to repaired DNA can be described as pathways, comprising series of reaction steps. Currently known DNA damage signaling (DDS), repair and damage tolerance pathways can be divided into eight categories:

- DDS: induced in response to the DNA damage caused by some endogenous and environmental agents;
- Direct reversal repair (DDR): directly restores the native nucleotide residue by removing the non-native chemical modification;
- Base-excision repair (BER): initiated by excision of modified base from the DNA. Depending on the length of DNA resynthesis, the pathway is subdivided into two sub-pathways: short-path (SP-BER) or long-path (LP-BER);
- Nucleotide excision repair (NER): removes bulky damage from the DNA. The damage from the active strand of transcribed genes is removed by transcription coupled repair (TCR)–NER, while global genome repair (GGR)–NER removes damage present elsewhere in the genome;
- Mismatch repair (MMR): postreplicational DNA repair that removes errors introduced during the replication (misinserted nucleotides, small loops, insertions, deletions);
- Homologous recombination repair (HRR): repair of DNA double-strand breaks using the homologous DNA strand as a template for resynthesis;
- Non-homologous end joining repair (NHEJ): ligation of ends resulting from DNA double-strand breaks (including the more error-prone microhomology end joining (MMEJ) mechanism;
- Translesion synthesis (TLS): damage-tolerance pathway that employs specialized polymerases to replicate across lesions in order to finish replication despite DNA damage.

Each of these pathways can be represented as a series of enzymatic transformations between different DNA structures, catalyzed by a dedicated system of proteins. It must be emphasized that DNA repair pathways are connected to each other, i.e. they can share some steps and/or proteins involved (14). As a consequence, DNA repair proteins rarely work in isolation in the cell, and their activity is dependent on other components of DNA repair systems. Therefore, knowledge of both the entire DNA repair systems and their components is critical to our understanding of how cells controls and repairs the constantly occurring damage of their genomes. Many of the proteins involved in DNA repair are very well-described (including substrate specificity, kinetics, mechanisms of action and 3D structure in complex with the substrate and/or the product of its activity). Information about different factors involved in DNA repair is, however, scattered in the literature and thus far there has been no resource to organize information on DNA repair at the systems level. Moreover, there are still processes, for which an enzymatic activity is known or suspected to exist, but the genes/proteins/enzymes proteins have not been characterized yet. As an example an enzyme that is capable of removing 5-hydroxymethyluridine, a product of thymine oxidation in human cells must exist, but to date has not been identified yet (21).

We have developed the REPAIRtoire database as a single online resource to store and organize information about DNA repair at the systems level. The purpose of REPAIRtoire is to gather together information about all DNA repair systems and proteins from model organisms, and to facilitate the access to knowledge about correlation of human diseases with mutation in genes responsible for DNA integrity and stability as well as information about the toxic and mutagenic agents causing DNA damage.

DATABASE CONTENT

Based on a comprehensive literature survey, we have compiled the following data sets:

- A list of DNA lesions linked to environmental mutagenic and cytotoxic agents;
- DDS, repair and tolerance pathways comprising structures of damaged DNAs and intermediates of the repair processes, connected by known transformations, usually reactions catalyzed by enzymes;
- Proteins involved in the aforementioned transformation and
- Information about diseases connected with defects in DNA repair proteins.

All data items have been curated manually and whenever possible and reasonable, we provided references to the published experimental reports and/or to other databases.

DATABASE ORGANIZATION AND ACCESS

REPAIRtoire is a relational database that links together the aforementioned data sets, which can be queried via five menus, 'DNA DAMAGE', 'PATHWAYS', 'PROTEINS', 'DISEASES' and 'PUBLICATIONS' (Figure 1).

DAMAGE: We collected information about 85 different types of damage in the DNA (as of October 14, 2010). Many of them describe general classes of damage events such as single-strand breaks or base loss that are independent of the local sequence). About 60 chemical compounds that cause DNA damage were connected to the according types of damage. Of all lesions, 36 could be connected to a single molecular structure (e.g. point mutations and nucleotide modifications such as 1-hydroxypropyl-adenine). Each type of damage is described on its own sub-page, which includes information about the potential source (e.g. spontaneous formation, intermediate in some DNA repair process, etc.), proteins that may recognize its presence in the DNA, keywords that facilitate analyzing its context, and literature links. For 36 types of damage with unique chemical structures, REPAIRtoire displays the structure in 1D (using a SMILE code), 2D and 3D (with the JMol JAVA applet), and provides atomic coordinates for download in the .mol format.

PATHWAYS: This menu provides access to data about eight pathways (DDS, DDR, BER, NER, MMR, HHR, NHEJ, TLS) from three model organisms: *Escherichia coli*, *Saccharomyces cerevisiae* and *Homo sapiens*. These pathways are represented as graphs visualized with PyGraphviz (http://networkx.lanl.gov/pygraphviz/), in which the nodes represent DNA states, and the edges represent the reactions between them, e.g. enzymatic reactions. All edges of the (sub)graph, i.e. arrows that connect the images, are hyperlinked to static 'reaction' windows comprising one or more panels that display basic information about the selected reaction. All nodes of the graph, e.g. DNA–protein complexes at various stages of the repair process, are hyperlinked to static windows comprising detailed information about the given stage of DNA repair. All protein components of each state/complex are also hyperlinked to individual protein pages.

PROTEINS: As of October 14, 2010, REPAIRtoire stores information about 69, 78 and 154 proteins from *E. coli*, *S. cerevisiae and H. sapiens*, respectively, and their

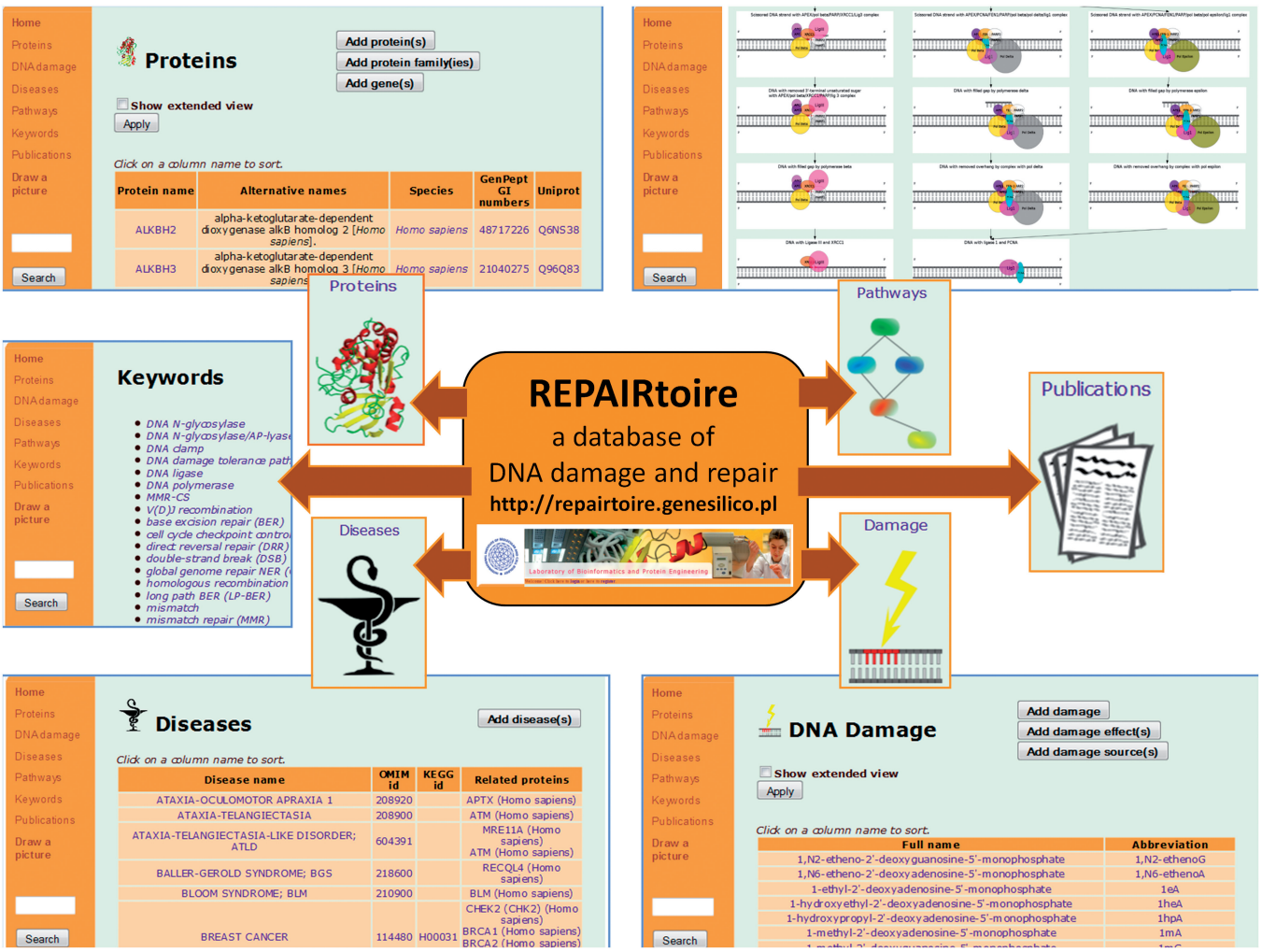

Figure 1. Contents of REPAIRtoire database.

genes that can be assessed either via pathways or directly from the 'PROTEINS' menu. In the process of manual data curation, we collected the available information concerning alternative gene and protein names and amino acid sequences available in the NCBI databases (22), 3D structures available in the protein data bank (23) and various features available in other databases (e.g. information about the enzymatic function, presence of isoforms, cellular/tissue and subcellular localization, together with links to the relevant database entries). Currently the DNA repair protein data set encompasses only *E. coli*, *S. cerevisiae* and *H. sapiens*, but will be expanded in the future and may eventually comprise all orthologs of the functionally characterized enzymes identifiable in fully sequenced genomes.

DISEASES: Thus far, we have compiled information about 40 diseases caused by the mutations in 32 genes linked to defects in DNA repair proteins. This data set is presented as a table with hyperlinks to the proteins concerned, and to the relevant entries in the Online Mendelian Inheritance in Man (OMIM) database (22). Each disease has its own subpage with a succinct description and additional links (e.g. keywords). Reciprocal links to diseases are also available in each protein field.

KEYWORDS: This menu provides quick access to the most common keywords used to annotate the database entities according to biological processes and activities such as: DNA repair pathways, the response to DNA damage, the cell cycle checkpoint control, the DNA *N*-glycosylase, the DNA N-glycosylase/AP-lyase, etc.

PUBLICATIONS: Literature references to entries in the PubMed database (22) have been compiled into an additional data set, currently comprising 2613 positions.

IMPLEMENTATION

REPAIRtoire has been implemented using the Django web framework (http://www.djangoproject.com/). It uses a SQLite relational database to store data. Among the features provided are profile and login, an image drawing tool, wiki-like pages of all entries and a search tool.

We would like to emphasize here that there is no need to register or to log in to view the content of the database. However, the profile and login features have been provided for collaborators and 'super users' who are interested not only in viewing, but also editing the content of the database. Creating an account and loging into the database give the user access to the admin site of the database, but also uncovers the wiki-like pages of the entries. By entering the administration site it is possible to add new data, delete information, edit and correct mistakes. Editing information about proteins, genes, diseases and types of damage is also available via wiki-like pages, from the pages of particular database entries. Users can also add comments and new references to existing records.

The image drawing tool (accessible via a 'draw a picture' link in the main menu) system has been developed to illustrate all steps of DNA repair pathways as protein–DNA complexes, in which proteins are displayed in the textbook-like format of 'potato models' (ellipsoids). However, it can be also used as a standalone tool (without any connection to REPAIRtoire) to create images of protein–protein or protein–DNA complexes. The drawing engine uses the SVG format provided by the W3C consortium, and enables exporting the image in the JPEG format. The SVG format enables resizing the images without the loss of the quality and makes it possible to modify them with external tools for vector graphics processing, e.g. Inkscape or others free or commercially available software.

The REPAIRtoire database can be queried using a simple text search tool, available in the main menu. The tool returns a structured list of entries in the database that contain the query (e.g. 'cancer', 'DNA polymerase', 'crosslink', 'adenine', etc. or a name of the author). In future, we plan to introduce filtering and advanced search options (e.g. to enable sorting according to the relevance to the query).

DISCUSSION

The topic of DNA repair is covered by many computational resources; however, thus far there has been no specialized database dedicated to DNA repair pathways. The *repair*GENES database (http://www.repairgenes.org/) collects information about genes encoding proteins involved in DNA repair and connects information taken from sequence and ontology databases. Repair-FunMap (24) used to provide information about the network of interactions between proteins involved in DNA repair and other proteins, but to our best knowledge it is no longer available. Information about DNA repair pathways and their components is available also in general-purpose pathway databases such as KEGG (25) or REACTOME (26). REPAIRtoire is dedicated to DNA repair and it contains more detailed and select information than the general-purpose pathway databases. An important component of REPAIRtoire that to our knowledge is not present in other databases is the dataset of DNA damage, connected to potential causes of each lesion as well as to effects if they are not removed. REPAIRtoire is also unique in that it provides reciprocal links between the damage (or its more general type) and the proteins that can detect and remove it. Our literature survey has also revealed a greater number of connections between particular DNA lesions and the respective proteins that can detect and remove it than can be found in general databases.

In the development of REPAIRtoire, we used our experience from the work on the MODOMICS database of RNA modification pathways (27,28). We hope that this database will become comparably popular and useful for the community of researchers working on DNA repair, as MODOMICS has become in the RNA modification community. In the future, we plan to integrate these databases using a common data model and a joint interface, and to extend the database system to cover the entire metabolism of nucleic acids.

AVAILABILITY

The content of the database and the software for generation of custom images for DNA–protein complexes are freely available at the URL http://repairtoire.genesilico.pl. Scientists interested in adding or curating data (proteins, features, complexes, pathways, etc.) or in implementing options that are not yet available are encouraged to contact J.M.B. (at iamb@genesilico.pl). This article should be cited in research projects assisted by the use of REPAIRtoire.

ACKNOWLEDGEMENTS

We would like to thank Peter Friedhoff, Ashok Bhagwat, Orlando Schärer, Michelle Cristovao, Ines Winkler and Jan Kaczyński for their contributions and helpful discussions. We are indebted to the authors of primary databases and services, whose content could be reused or linked to by REPAIRtoire. We also thank all developers and curators of the MODOMICS database, whose feedback has been taken into account in the development of REPAIRtoire.

FUNDING

This project has a very long history and throughout the years have been supported by the EU Framework Programme (FP6 grants PLASTOMICS, contract number LSHG-CT-2003-503238 and DNA ENZYMES, contract number MRTN-CT-2005-019566 and currently FP7 HEALTHPROT, contract number 229676). It has been also supported by a Polish-Norwegian grant (PNRF-143-AI-1/07 to J.K.); Polish Ministry of Science and Higher Education (grant PBZ-MNiI-2/1/2005 to G.P. and K.P., partial) and Deutscher Akademischer Austausch Dienst (fellowship D/09/42768 to K.R., partial). The open access publication charge for this paper has been waived by Oxford University Press—*NAR* Editorial Board members are entitled to one free paper per year in recognition of their work on behalf of the journal.

Conflict of interest statement. None declared.

REFERENCES

1. Maddukuri,L., Dudzinska,D. and Tudek,B. (2007) Bacterial DNA repair genes and their eukaryotic homologues: 4. The role of nucleotide excision DNA repair (NER) system in mammalian cells. *Acta Biochim. Pol.*, **54**, 469–482.
2. Jeggo,P. and Lavin,M.F. (2009) Cellular radiosensitivity: how much better do we understand it? *Int. J. Radiat. Biol.*, **85**, 1061–1081.
3. Krwawicz,J., Arczewska,K.D., Speina,E., Maciejewska,A. and Grzesiuk,E. (2007) Bacterial DNA repair genes and their eukaryotic homologues: 1. Mutations in genes involved in base excision repair (BER) and DNA-end processors and their implication in mutagenesis and human disease. *Acta Biochim. Pol.*, **54**, 413–434.
4. Arczewska,K.D. and Kusmierek,J.T. (2007) Bacterial DNA repair genes and their eukaryotic homologues: 2. Role of bacterial mutator gene homologues in human disease. Overview of nucleotide pool sanitization and mismatch repair systems. *Acta Biochim. Pol.*, **54**, 435–457.
5. Brissett,N.C. and Doherty,A.J. (2009) Repairing DNA double-strand breaks by the prokaryotic non-homologous end-joining pathway. *Biochem. Soc. Trans.*, **37**, 539–545.
6. Vaisman,A., Lehmann,A.R. and Woodgate,R. (2004) DNA polymerases eta and iota. *Adv. Protein Chem.*, **69**, 205–228.
7. Robertson,A.B., Klungland,A., Rognes,T. and Leiros,I. (2009) DNA repair in mammalian cells: base excision repair: the long and short of it. *Cell. Mol. Life Sci.*, **66**, 981–993.
8. Slatter,M.A. and Gennery,A.R. (2010) Primary immunodeficiencies associated with DNA-repair disorders. *Expert Rev. Mol. Med.*, **12**, e9.
9. Lindahl,T. (1993) Instability and decay of the primary structure of DNA. *Nature*, **362**, 709–715.
10. De Bont,R. and van Larebeke,N. (2004) Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis*, **19**, 169–185.
11. Drablos,F., Feyzi,E., Aas,P.A., Vaagbo,C.B., Kavli,B., Bratlie,M.S., Pena-Diaz,J., Otterlei,M., Slupphaug,G. and Krokan,H.E. (2004) Alkylation damage in DNA and RNA–repair mechanisms and medical significance. *DNA Repair*, **3**, 1389–1407.
12. Olinski,R., Siomek,A., Rozalski,R., Gackowski,D., Foksinski,M., Guz,J., Dziaman,T., Szpila,A. and Tudek,B. (2007) Oxidative damage to DNA and antioxidant status in aging and age-related diseases. *Acta Biochim. Pol.*, **54**, 11–26.
13. Tudek,B. (2007) Base excision repair modulation as a risk factor for human cancers. *Mol. Aspects Med.*, **28**, 258–275.
14. Friedberg,E.C., Walker,G.C., Siede,W., Wood,R.D., Schulz,R.A. and Ellenberger,T. (2006) *DNA repair and mutagenesis*, 2nd edn. ASM Press, Washington, DC.
15. Hansen,W.K. and Kelley,M.R. (2000) Review of mammalian DNA repair and translational implications. *J. Pharmacol. Exp. Ther.*, **295**, 1–9.
16. Raptis,S. and Bapat,B. (2006) Genetic instability in human tumors. *EXS*, 303–320.
17. Wilson,D.M. III and Barsky,D. (2001) The major human abasic endonuclease: formation, consequences and repair of abasic lesions in DNA. *Mutat. Res.*, **485**, 283–307.
18. Wood,R.D., Mitchell,M. and Lindahl,T. (2005) Human DNA repair genes. *Mutat. Res.*, **577**, 275–283.
19. Wood,R.D., Mitchell,M., Sgouros,J. and Lindahl,T. (2001) Human DNA repair genes. *Science*, **291**, 1284–1289.
20. Wood,R.D., Mitchell,M. and Lindahl,T. (2010) Human DNA Repair Genes. http://sciencepark.mdanderson.org/labs/wood/DNA_Repair_Genes.html (14 October 2010, date last accessed).
21. Baker,D., Liu,P., Burdzy,A. and Sowers,L.C. (2002) Characterization of the substrate specificity of a human 5-hydroxymethyluracil glycosylase activity. *Chem. Res. Toxicol.*, **15**, 33–39.
22. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
23. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
24. Wen,L. and Feng,J.A. (2004) Repair-FunMap: a functional database of proteins of the DNA repair systems. *Bioinformatics*, **20**, 2135–2137.
25. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
26. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.
27. Dunin-Horkawicz,S., Czerwoniec,A., Gajda,M.J., Feder,M., Grosjean,H. and Bujnicki,J.M. (2006) MODOMICS: a database of RNA modification pathways. *Nucleic Acids Res.*, **34**, D145–D149.
28. Czerwoniec,A., Dunin-Horkawicz,S., Purta,E., Kaminska,K.H., Kasprzak,J.M., Bujnicki,J.M., Grosjean,H. and Rother,K. (2009) MODOMICS: a database of RNA modification pathways. 2008 update. *Nucleic Acids Res.*, **37**, D118–D121.