SAGE-Hindawi Access to Research
Molecular Biology International
Volume 2011, Article ID 475718, 9 pages
doi:10.4061/2011/475718

Review Article

Databases and Bioinformatics Tools for the Study of DNA Repair

Kaja Milanowska,[1,2] 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*

Correspondence should be addressed to Janusz M. Bujnicki, iamb@genesilico.pl

Received 16 February 2011; Revised 28 April 2011; Accepted 22 May 2011

Academic Editor: Frédéric Coin

Copyright © 2011 Kaja Milanowska et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

DNA is continuously exposed to many different damaging agents such as environmental chemicals, UV light, ionizing radiation, and reactive cellular metabolites. DNA lesions can result in different phenotypical consequences ranging from a number of diseases, including cancer, to cellular malfunction, cell death, or aging. To counteract the deleterious effects of DNA damage, cells have developed various repair systems, including biochemical pathways responsible for the removal of single-strand lesions such as base excision repair (BER) and nucleotide excision repair (NER) or specialized polymerases temporarily taking over lesion-arrested DNA polymerases during the S phase in translesion synthesis (TLS). There are also other mechanisms of DNA repair such as homologous recombination repair (HRR), nonhomologous end-joining repair (NHEJ), or DNA damage response system (DDR). This paper reviews bioinformatics resources specialized in disseminating information about DNA repair pathways, proteins involved in repair mechanisms, damaging agents, and DNA lesions.

1. Introduction

DNA repair processes are indispensable for maintaining the integrity of genetic information in all organisms. Environmental agents such as chemicals, UV light, and ionizing radiation, as well as endogenous metabolic processes involving DNA constantly challenge the chemical structure and stability of the genome. DNA lesions can interfere with processes such as DNA replication or transcription and may lead to mutations and cancer [1, 2]. To prevent the erosion of the chemical structure of DNA, living systems have evolved various different biochemical systems for DNA repair [3–7].

DNA damage from endogenous sources gives rise to 20,000 lesions per mammalian cell per day. Amongst these lesions, the most common are base deamination, spontaneous hydrolysis of the *N*-glycosidic bond, alkylation, and damage by reactive oxygen or nitrogen species and lipid peroxidation products [8–12]. Other lesions such as the formation of single- and double-strand breaks, the collapse of replication forks, and the introduction of modified nucleic acid bases during DNA replication are caused by errors in DNA metabolic processes. In total, there are 10^{16}–10^{18} DNA repair events that occur daily in a healthy adult man (10^{12} cells) [13]. Lesions that are not repaired often lead to mutations, aging and various diseases, including carcinogenesis and neurodegeneration [14–18]. Some pathological disorders directly related to defects in the DNA repair machinery are Xeroderma pigmentosum, different types of cancer (breast cancer, colorectal cancer, endometrial cancer, gastric cancer, or prostate cancer), Fancomi anemia, Muir-Torre syndrome, Tay syndrome, and Werner syndrome. On the other hand, unrepaired lesions that occur in germline cells become the main source of genetic variability and therefore a driving force for the evolution. For this reason, the DNA repair system needs not only to be regulated to maintain an individual genome's integrity, but also to increase the genetic variability in the context of populations. Many mechanisms are known that regulate the amount of DNA repair as a response to environmental conditions [19].

Given its many duties in different contexts, it is not surprising that DNA repair is a very complicated process, involving many factors. For instance to date, 168 genes encoding proteins involved in DNA repair have been identified in the human genome [17, 18, 20] (20 January 2011, date last accessed). Over all organisms, there are many more; for base excision repair alone, KEGG [21] lists 41 groups of orthologous genes encoding for hundreds of proteins in total. The key players in DNA repair are enzymes that catalyze reactions leading from the DNA with damage to a repaired molecule. They are assisted by proteins that detect damage and mediate signals that coordinate the repair process with other cellular processes. From the point of view of the DNA substrate, the biochemical pathways of DNA repair can be divided into eight categories:

(i) DNA damage signaling (DDS): also known as the DNA damage checkpoint; it is a group of responses to DNA damage caused by some endogenous and environmental agents; activation of these pathways may be triggered by the effect the DNA lesions have on replication, transcription, or chromatin topology;

(ii) base-excision repair (BER): initiated by excision of a modified base from the DNA. Depending on the length of DNA resynthesis, the pathway is subdivided into two subpathways: short path (SP-BER) or long path (LP-BER);

(iii) DNA damage response (DDR): directly restores the native nucleotide residue by removing the nonnative chemical modification;

(iv) homologous recombination repair (HRR): repair of DNA double-strand breaks using the homologous DNA strand as a template for resynthesis;

(v) mismatch repair (MMR): postreplicational DNA repair that removes errors introduced during the replication (misinserted nucleotides, small loops, insertions, deletions);

(vi) nonhomologous end-joining repair (NHEJ): ligation of ends resulting from DNA double-strand breaks (including the more error-prone microhomology-mediated end-joining (MMEJ) mechanism;

(vii) 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;

(viii) 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, that is, they can share some steps and/or proteins involved [13]. 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.

DNA repair itself is not an isolated process, and it is strongly connected to other pathways of nucleic acid metabolism, including (but not limited to) DNA replication, DNA epigenetic modification, transcription, cell cycle regulation, and induced cell death as well as processes that are specific to different domains of life, such as telomere maintenance in eukaryotes and DNA restriction in prokaryotes.

2. DNA Repair Data and Databases

The knowledge of DNA repair systems and their components is critical to our understanding of how cells control the integrity of their genomes. A large body of data on this topic has been published mostly in the literature and in a few electronic resources. Today, systematizing this knowledge and presenting it in a clear and easily accessible way is mostly done by biological databases. The collection, curation, and availability of data are necessary to answer questions about subsystems of DNA repair, for example, "which proteins participate in MMR in humans and in plants?", "what immediate cellular response is triggered by damage caused by UV light?", or "how does HRR differ between plants and vertebrates?". The topic of DNA repair is covered by many computational resources. However, there are few databases dedicated to DNA repair, and most of the data is scattered over various general databases. In Table 1, we have listed some of the available web resources relevant to DNA repair, and in the following section we discuss their content.

2.1. Databases Dedicated to DNA Repair. "REPAIRtoire" is a database for systems biology of DNA damage and repair developed by the authors of this paper and their coworkers [22]. The purpose of this database is to gather information about all DNA repair systems and proteins from model organisms and to facilitate the access to knowledge about correlation of human diseases with mutations in genes responsible for DNA integrity and stability as well as information about toxic and mutagenic agents causing DNA damage. REPAIRtoire is available online at http://repairtoire.genesilico.pl/. It organizes data into the following categories: (i) the chemical structures of DNA lesions (as of April 2011: 85 different types of damage in the DNA) linked to their causative mutagenic and cytotoxic agents, (ii) pathways comprising individual processes and enzymatic reactions involved in the removal of damage, (iii) proteins participating in DNA repair, in particular enzymes involved in the transformation between different chemical structures of the DNA substrate, and (iv) diseases correlated with mutations in genes encoding DNA repair proteins (40 diseases caused by the mutations in 32 genes linked to defects in DNA repair proteins). It also provides links to publications and external datasets. REPAIRtoire covers all eight main DNA damage checkpoint, repair, and tolerance pathways (see above). The pathway/protein dataset is currently limited to three model organisms: *Escherichia coli*, *Saccharomyces cerevisiae*, and *Homo sapiens*. DNA repair and tolerance pathways are represented as graphs and in tabular form with descriptions of each repair step as well as corresponding proteins. The individual entries in the

Table 1: Databases dedicated to DNA repair and general-purpose databases relevant to DNA repair.

Name	url	Reference	Description
	Databases dedicated to DNA repair		
REPAIRtoire	http://repairtoire.genesilico.pl/	[22]	Database of DNA repair pathways
repairGENES	http://www.repairgenes.org/	unpublished	Database of DNA repair genes
Human DNA Repair Genes	http://sciencepark.mdanderson.org/labs/wood/DNA_Repair_Genes.html	[17]	Database of human DNA repair genes
Repair-FunMap	currently unavailable	[23]	A database of interactions between proteins involved in DNA repair and other proteins
	Other databases relevant to DNA repair		
KEGG	http://www.genome.jp/kegg/	[21]	Kyoto Encyclopedia of Genes and Genomes
Reactome	http://www.reactome.org/ReactomeGWT/entrypoint.html	[24]	Database of human pathways and reactions
GeneSNPs	http://www.genome.utah.edu/genesnps/		This Environmental Genome Project web resource integrates gene, sequence, and polymorphism data into individually annotated gene models. The human genes included are related to DNA repair, cell cycle control, cell signaling, cell division, homeostasis, and metabolism
Mouse Mutation Database	http://pathcuric1.swmed.edu/research/research.htm	[25]	The Database of mouse strains carrying targeted mutations in genes affecting cellular responses to DNA damage
BioCyc (EcoCyc, MetaCyc)	http://biocyc.org/	[26]	Experimentally studied metabolic pathways and enzymes from more than 1,500 organisms
BRENDA	http://www.brenda-enzymes.org/	[27]	The main collection of enzyme functional data
Pathway Commons	http://www.pathwaycommons.org/pc/	[28]	A collection of publicly available pathway data from multiple organisms
NGSethDB	http://bioinfo2.ugr.es/NGSmethDB/gbrowse/hg19/	[29]	Database for next-generation sequencing single-cytosine-resolution DNA methylation data
DNAreplication	http://DNAreplication.net/	[30]	Database for the eukaryotic DNA replication community
MethyCancer	http://methycancer.psych.ac.cn/	[31]	Links between DNA methylation levels and cancer
PubMeth	http://matrix.ugent.be/pubmeth/	[32]	Links between DNA methylation levels and cancer
MethDB (2009)	http://www.methdb.de/	[33]	The database for DNA methylation and environmental epigenetic effects
OriDB	http://www.oridb.org/index.php	[34]	Confirmed and predicted DNA replication origin sites
REBASE	http://rebase.neb.com/rebase/rebase.html	[35]	Enzymes and genes for DNA restriction and modification in prokaryotes
ROSPath	http://rospath.ewha.ac.kr/	[36]	Reactive oxygen species (ROS) signaling pathway proteins
Pathguide	http://www.pathguide.org/	[37]	A listing of pathway, signal transduction, and protein-protein interaction databases
CREMOFAC	http://www.jncasr.ac.in/cremofac/	[38]	Chromatin remodeling factors
DAnCER	http://wodaklab.org/dancer/	[39]	Disease-Annotated Chromatin Epigenetics Resource
Telomerase database	http://telomerase.asu.edu/	[40]	Sequences and structures of the RNA and protein subunits of telomerase, mutations of telomerase components
Replication Domain	http://www.replicationdomain.com/	[41]	Replication timing database and genome-wide data visualization tool

database (proteins, diseases, pathway steps, damage, etc.) are cross-referenced to the supporting literature and their respective primary databases. REPAIRtoire can be queried by the names of pathway, protein, enzymatic complex, damage and disease. The query 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).

The REPAIRtoire website provides a system for editing, adding, and removing data. These features have been provided for collaborators and "superusers" who are interested not only in viewing, but also curating the content of the database. Creating an account and logging into the database grants access to the administrative site of the database to a user. 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 for particular database entries. Users can also add comments and suggest new references for the existing records. REPAIRtoire is unique in that it focuses on DNA repair and provides reciprocal annotation between damage entities and the proteins that can detect and remove them. It also contains more connections between DNA lesions and the respective proteins that can detect and remove them than can be found in general-purpose databases.

The REPAIRtoire website which also provides an online tool for drawing images of DNA-protein complexes (accessible via the "draw a picture" link in the main menu) is provided. This tool 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 outside the DNA repair context to create images of any 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. Images created in the SVG vector format can be scaled without losing quality and can be modified with external tools for vector graphics processing, for example, Inkscape or other free or commercially available software.

The "*repairGENES*" database (http://www.repairgenes.org/) collects information about genes encoding proteins involved in DNA repair and connects information taken from sequence and ontology databases. At the moment, the site contains DNA repair genes from 134 selected species. The database can be browsed by organisms and by biological processes defined by the Gene Ontology (GO) standard [42]. The species are organized in a taxonomy tree. For processes, 17 subcategories of the GO term "DNA repair" (GO:0006281) and their respective subterms are distinguished. For each process, the organisms and genes that refer to this term can be listed. Also, it is possible to highlight the processes for a given organism. The major advantage of using GO terms is that they are being used ubiquitously for annotating sequence data. The raw data about DNA repair genes is extracted from the SWISS-PROT database. The repairGenes database also gives an overview of DNA repair processes and genes in five selected organisms (*Archaeoglobus fulgidus*, *Drosophila melanogaster*, *E. coli*, *Homo sapiens*, and *S. cerevisiae*), in total listing 452 genes.

"*Human DNA Repair Genes*" is an online supplement to a review published by Wood et al. in 2005 [17] and updated regularly (http://sciencepark.mdanderson.org/labs/wood/DNA_Repair_Genes.html). It provides a table with Gene Name (synonyms) linked to the GeneCards Human Gene Database at Cancer Research UK (http://bioinformatics.cancerresearchuk.org/genecards/) [43], activity linked to the OMIM database, chromosome location linked to the NCBI MapView, and an accession number linked to the NCBI Entrez server [44].

The "*Repair-FunMap*" database [23] 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.

2.2. General-Purpose Databases Relevant to DNA Repair. "*KEGG*" (Kyoto Encyclopedia of Genes and Genomes, available at http://www.genome.jp/kegg/) [21] is a collection of separate cross-linked databases including KEGG PATHWAY, KEGG DISEASE (human diseases), KEGG GENES (genes and proteins), and KEGG ORGANISMS. Of particular relevance to DNA repair are KEGG GENES (a catalog of genes for sequenced genomes obtained from publicly available resources, mostly NCBI RefSeq and KEGG PATHWAY (a collection of manually drawn pathway maps representing knowledge on the molecular interaction and reaction networks for: global map of pathways, metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases, and interaction of these systems with drugs)). DNA repair pathways annotated in KEGG include BER, NER, NHEJ, MMR, and HRR but not DDS, DDR, or TLS. A schematic graphical representation of protein-DNA complexes in the reaction steps of each pathway is available for Eukaryotes and Prokaryotes separately. KEGG BRITE is another component of KEGG that is important for analyzing DNA repair systems. It is a collection of hierarchical classifications representing knowledge on various aspects of biological systems. In contrast to KEGG PATHWAY, which is limited to molecular interactions and reactions, KEGG BRITE incorporates many different types of relationships. The most relevant and interesting part is a section devoted to DNA repair (identifier "ko03400"—"DNA repair and recombination proteins"), where all DNA repair proteins available in KEGG are classified according to their functions in this process.

"*Reactome*" (http://www.reactome.org/ReactomeGWT/entrypoint.html) [24] is a resource developed in collaboration among different groups as an open source curated bioinformatics database of human pathways and reactions. The site provides bioinformatics tools for pathway analysis such as: the Pathway Browser, the Pathway and Expression Analysis tools, or the Species Comparison tool. In contrast to KEGG, Reactome includes graphical representations of DNA repair pathways generated for each organism explicitly instead of a generalized view like in KEGG. Moreover, Reactome divides pathways into subpathways, for example, GG-NER (global genomic NER) in human is divided into four subpathways (DNA damage recognition, formation of the incision complex, dual incision reaction, and gap-filling DNA repair synthesis and ligation). Each subpathway contains individual reactions visualized in the context of the entire cellular metabolic map. The Pathway Analysis tool facilitates analysis of different pathways, for example, finding connections between RNA transcription and DNA repair, facilitating interdisciplinary studies [45].

The "*GeneSNPs*" database (http://www.genome.utah.edu/genesnps/) is dedicated to known human polymorphisms and has a section devoted to DNA repair. It can be accessed from the main page by selecting "DNA repair" in the Gene Lists menu on top of the home page. The SNP loci are presented as a table of 119 human genes involved in DNA repair and connected to phenotypes described in the OMIM database [44]. An exemplary usage of this resource is the study of polymorphisms in the DNA repair gene XRCC, where all the SNP data were collected from the GeneSNPs database [46]. More phenotypes of DNA repair defects can be found in the "*Mouse Mutation Database*" (a database of mouse strains carrying targeted mutations in genes affecting cellular responses to DNA damage available at http://pathcuric1.swmed.edu/research/research.htm) [25].

"*BioCyc*" (http://biocyc.org/) [26] is a collection of 1004 (as of February 2011) Pathway/Genome Databases. Each database in the BioCyc collection describes the genome and metabolic pathways of a single organism. This is not only a collection of databases but of tools for bioinformatics analysis, including the following: a genome browser, a display of individual metabolic pathways and of full metabolic maps, visual analysis of user-supplied "omics" datasets by painting onto metabolic, regulatory, and genome maps, and comparative analysis tools. There is also downloadable version of BioCyc that includes the Pathway Tools. The BioCyc databases are divided into three tiers, based on their quality. Tier 1 databases have received person-decades of literature-based curation and are the most accurate. These include for example, EcoCyc (http://ecocyc.org/) [47], a comprehensive database of *Escherichia coli* K-12 MG1655 biology or MetaCyc (http://metacyc.org/), a database of nonredundant, experimentally elucidated metabolic pathways. Data included in these databases undergo a curation procedure involving external experts, who work on particular cellular systems to provide a comprehensive literature overview and up-to-date coverage of the field. Recently, this type of curation has been applied to the process of DNA repair; both direct repair mechanisms, such as photolyase, as well as indirect repair mechanisms, such as nucleotide excision repair, base excision repair and homologous recombination have been annotated [47]. Tier 2 and Tier 3 databases of BioCyc contain computationally predicted metabolic pathways, predictions as to which genes code for missing enzymes in metabolic pathways, and predicted operons. BioCyc does not include a dedicated DNA repair section, but information on DNA repair pathways can be found in other database sections. Data available in BioCyc can be used in in-depth analyses of biological systems relevant to different fields of research. This approach has been demonstrated in the study of differential network expression during drug and stress response by Cabusora et al. [48], where the expression data of known stress responders and DNA repair genes in mycobacterium tuberculosis from BioCyc collection were used.

"*BRENDA*" (BRaunschweig ENzyme Database, http://www.brenda-enzymes.org/) [27] is a comprehensive database on enzymes that collects manually annotated information on properties of enzymes, including mutants and engineered variants. It describes enzymes involved in DNA repair that have an E.C. number (e.g., uvrA: EC 3.1.25.1). Enzyme records contain data taken from the primary literature, such as classification, nomenclature, reaction type, substrate specificity, functional parameters, species, protein sequence and structure, practical application, information on mutants and engineered variants, stability, disease, isolation, and preparation. An essential part of BRENDA consists of information on metabolites and small molecules, which interact with enzymes as substrates and products, inhibitors, activating compounds, cofactors, or bound metals. BRENDA provides also enzyme disease-related information obtained from PubMed entries by text-mining procedures. BRENDA is currently the largest continuously maintained and publicly available enzyme database and covers a large number of experimentally characterized DNA repair enzymes.

"*Pathway Commons*" (http://www.pathwaycommons.org/pc/) is a comprehensive collection of publicly available pathway data from multiple organisms [28], which includes biochemical reactions, complex assembly, transport, catalysis events, and physical interactions involving proteins, DNA, RNA, small molecules, and complexes. This meta-database collects information from other databases such as Reactome or BioGrid, thereby facilitating analyses of system-level datasets across several species. It allows users to browse and search pathways across multiple valuable public pathway databases and download an integrated set of pathways in the BioPAX format for global analysis. It also provides an interface for software developers to create software for more advanced analyses and hence may be a very useful resource for programmatic linking of data on DNA repair systems with other cellular systems and pathways.

There exist numerous databases dedicated to other aspects of DNA metabolism. Examples include DNA replication (OriDB [34], ReplicationDomain [41]), apoptosis (Deathbase [49]), telomere maintenance (Telomerase database [40]), DNA restriction and modification (REBASE [35]), and epigenetics/chromatin modification (DAnCER [39]). These processes are relevant to DNA repair as they may contribute to DNA damage (replication) or regulation of other enzymatic processes (DNA methylation, cell cycle control, and apoptosis).

3. Bioinformatics Tools for the Study of DNA Repair Proteins

In addition to databases that store and disseminate the data, there are also bioinformatics tools that can be particularly useful for data analyses. We would like to emphasize three groups of predictive tools that can be particularly useful for analyzing DNA repair enzymes: methods for predicting and modeling protein structures, predicting protein-DNA interactions and complexes, predicting the effect of amino acid substitutions on protein stability and function, and their phenotypic effect [50], as well as predicting cancer outcome [39].

3.1. Protein Structure Prediction. There is a large number of tools, with which to predict the structure of a protein when only its sequence is known. Their performance is evaluated in the biannual CASP benchmarking experiment [51]. One approach we would like to highlight here is homology modeling. There, a protein with known 3D structure is used as a template to construct a model for another, evolutionarily-related protein (a target). This approach requires not only an experimentally solved structure of the template protein, but also a pairwise sequence alignment between the target and the template. Among the numerous methods, the "*SWISS-MODEL*" server (http://swissmodel.expasy.org/) supports not only the fully automatic construction of homology models via its web interface, it also helps finding a suitable template and alignment [52]. It is particularly useful for building models of proteins that are closely related to the experimentally determined structures, so the relationship can be detected by methods such as "*BLAST*" [53]. If no such closely related templates are available, advanced template search and alignment tools such as "*HHSEARCH*" [54] can be used to identify remote evolutionary relationships. There are also specialized "meta-servers" such as the "*GeneSilico Metaserver*" [55] developed in the laboratory of the authors of this paper. These tools use several third-party methods and infer a consensus prediction.

As an example of protein modeling application to the analysis of DNA repair, we may refer to an analysis carried out in our laboratory: Missense alterations of the mismatch repair gene MLH1 have been identified in a significant proportion of individuals suspected of having Lynch syndrome, a hereditary syndrome that predisposes for cancer of colon and endometrium. The pathogenicity of many of these alterations was, however, unclear. A number of MLH1 alterations are located in the C-terminal domain (CTD) of MLH1, which is responsible for constitutive dimerization with another protein PMS2. We used the aforementioned "*GeneSilico Metaserver*" [55] to identified structurally characterized homologs of MLH1 and align their sequences, thereby enabling the construction of a homology model for MLH1 using the "FRankenstein's Monster" approach [56, 57]. That structural model was used to analyze 19 alterations connected to Lynch syndrome and to identify three alterations that decrease the efficiency of MMR in human by interfering with the MLH1-PMS2 dimerization, confirming that they are pathogenic, and suggesting that defective dimerization underlies their deleterious effect [50].

3.2. Methods for Predicting Protein-DNA Interactions. When analyzing enzymes acting on DNA, it is often important to know which parts of them interact with the substrate. Prediction of DNA-binding residues is facilitated by the knowledge of protein structure, either from experiment or from prediction (see above). An example of a bioinformatics online tool for structure-based prediction of DNA-binding residues is "*DISPLAR*" (http://pipe.scs.fsu.edu/displar.html), which uses a machine learning approach [58]. There are also methods, available as web services, enabling prediction of DNA-binding from protein sequence alone. Examples include "*BindN+*" (http://bioinfo.ggc.org/bindn+/) [59], "*DISIS*" (http://www.predictprotein.org/) [60], and "*DNABindR*" (http://turing.cs.iastate.edu/PredDNA/predict.html) [61].

If 3D structures of the components are known, it is also possible to obtain a three-dimensional model of protein-DNA complexes. The "*HADDOCK*" server (http://haddock.chem.uu.nl/) uses a flexible docking approach to build a complex from two or more separate protein and DNA structures [62]. It takes into account additional information such as distances between interacting residues and includes them as "ambiguous interaction restraints". This allows to use results from experimental analyses like mutation, crosslinking, and footprinting experiments or computational predictions made, for example, by the above-mentioned bioinformatics methods. It is important to note that HADDOCK generates a complex structure for all given components, but it does not evaluate whether the given components really interact and does not enable the modeling of large conformational changes. Also, identifying the correct interaction region is the most error-prone step, which is why accurate experimental knowledge is essential to obtain reliable structures. The HADDOCK developers also provide an extensive dataset of protein-DNA complexes that can be used for benchmarking purposes [63]. An alternative approach is to build models with other methods, without the use of experimental data, and then use the "*FILTREST3D*" method developed in the laboratory of the authors [64] to rank them according to the extent of agreement with the restraints.

3.3. Methods for Predicting the Effects of Amino Acid Substitutions. As illustrated by the example of the MLH1 protein, prediction of mutation/substitution effects on protein structure and function, and linking them to the relevant phenotype can be very useful in the study of DNA repair proteins. "*SNPs3D*" (http://www.snps3d.org/) [65] is an online tool that returns predictions of functional effects of nonsynonymous SNPs stored in the NCBI dbSNP database; currently it does not make predictions for altered sequences submitted by the users. There are a few predictive online methods that use protein structure (solved experimentally or modeled) to infer the effect of user-defined amino acid substitutions. "*CUPSAT*" (http://cupsat.tu-bs.de/) (Cologne University Protein Stability Analysis Tool) [66] predicts Gibbs-free energy changes associated with amino acid substitutions, based on analyzing of residue interactions with its 3D environment. "*PopMusic*" (http://babylone.ulb.ac.be/popmusic/) [67] evaluates the changes of protein stability resulting from single-residue or multiple substitutions. "*I-Mutant 2.0*" (http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi) [68] also predicts protein stability changes upon single-site substitutions. It can be used both as a classifier for predicting the sign of the protein stability change upon mutation and as a regression estimator for predicting the related Gibbs-free energy changes. "*MUpro*" (http://www.ics.uci.edu/~baldig/mutation.html) [69] is a set of machine learning programs to predict how single-site amino acid substitutions affect protein stability. The server accepts single protein sequences or sequences with a predicted tertiary structure of the protein as an input.

There are also methods that predict mutation/substitution effects based on sequence information alone. "*PolyPhen*" (Polymorphism Phenotyping) (http://genetics.bwh.harvard.edu/pph/) [70] predicts the possible impact of an amino acid substitution on the structure and function of human proteins, based on straightforward empirical rules. "*SIFT*" (http://sift.jcvi.org/) [71] predicts whether an amino acid substitution (AAS) affects protein function based on analysis of sequence profiles. It can be applied to study naturally occurring nonsynonymous polymorphisms as well as laboratory-induced missense mutations. "*MutPred*" (http://mutpred.mutdb.org/) [72] is a web application that predicts the gain/loss of 14 different structural and functional properties (for instance, gain of helical propensity or loss of a phosphorylation site). It also classifies an amino acid substitution as disease-associated or neutral in human. "*PhD-SNP*" (http://snps.uib.es/phd-snp/PhD-SNP.html) [73] is another machine-learning method for predicting whether a phenotype derived from a nonsynonymous SNP could be related to a genetic disease in humans. It is optimized to predict if a given point mutation can be classified as a disease-related or a neutral polymorphism.

3.4. Predicting Cancer Outcome. A tool which facilitates the analyses of cancer-related proteins, genes and pathways is CAERUS [39]—a tool for predicting cancer outcomes using relationships between protein structural information, protein networks, gene expression data, and mutation data (http://www.oicr.on.ca/research/ouellette/caerus/). This tool was developed in order to identify a list of gene signatures and to better predict cancer by investigating the changes in gene expression profiles caused by disruptions between protein-protein interactions and domain-domain interactions in the human interactome. As the authors of CAERUS indicate, it was tested on a set of well-documented breast cancer patients, which suggests that the disrupted interactome is important to determine patient prognosis. They also declare that this approach is robust if tested on other independent data sets and therefore offers a promising prognostic tool to classify different cancer outcomes. As DNA repair is closely connected to cancer, this service can be used in the analysis of proteins and genes related to oncogenesis.

4. Summary

DNA repair is currently covered by a few dedicated databases. While REPAIRtoire and repairGenes focus on this topic, information is also available via general-purpose pathway databases. The main bottlenecks are the data collection and standardization. For instance, there is no specialized, universal ontology and no standards to describe entities and processes involved in DNA repair. Connecting the known "parts" such as enzymes, to pathways and processes in a formalized way that at the same time provides more insight into DNA repair processes, is probably the biggest challenge for the bioinformatics of DNA repair in the nearest future. It may be necessary to extend the currently established GO ontology by a vocabulary that will allow for describing repair processes on the protein complex and reaction level. A particular challenge is to find a consistent and appealing way to represent repair processes visually, and to include not only 3D descriptions, but also the dimension of time. The development and application of new computer programs for simulating and visualizing molecular processes involving multiple components will certainly contribute to our understanding of the complex process of DNA repair. In particular, it may help in the identification of new biomarkers, in predicting the possible side-effects of drugs based on personal genome information, and in the development of new therapeutic agents to restore the proper function of DNA repair proteins affected by disease-causing mutations.

Acknowledgments

The authors would like to thank Joanna Krwawicz for sharing her knowledge on DNA repair pathways and resources and for her invaluable contribution to the REPAIRtoire database project. K. Milanowska has been supported by the Foundation for Polish Science (Grant TEAM/2009-4/2). K. Rother has been supported by the German Academic Exchange Service (Grant D/09/42768). J. M. Bujnicki has been supported by the 7FP Grant "HEALTH-PROT" from the European Commission (contract number 229676).

References

[1] P. Jeggo and M. F. Lavin, "Cellular radiosensitivity: how much better do we understand it?" *International Journal of Radiation Biology*, vol. 85, no. 12, pp. 1061–1081, 2009.

[2] L. Maddukuri, D. Dudzińska, and B. Tudek, "Bacterial DNA repair genes and their eukaryotic homologues: 4. The role of nucleotide excision DNA repair (NER) system in mammalian cells," *Acta Biochimica Polonica*, vol. 54, no. 3, pp. 469–482, 2007.

[3] K. D. Arczewska and J. T. Kuśmierek, "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 Biochimica Polonica*, vol. 54, no. 3, pp. 435–457, 2007.

[4] N. C. Brissett and A. J. Doherty, "Repairing DNA double-strand breaks by the prokaryotic non-homologous end-joining pathway," *Biochemical Society Transactions*, vol. 37, no. 3, pp. 539–545, 2009.

[5] A. Vaisman, A. R. Lehmann, and R. Woodgate, "DNA polymerases η and ι," *Advances in Protein Chemistry*, vol. 69, pp. 205–228, 2004.

[6] A. B. Robertson, A. Klungland, T. Rognes, and I. Leiros, "DNA repair in mammalian cells: base excision repair: the long and short of it," *Cellular and Molecular Life Sciences*, vol. 66, no. 6, pp. 981–993, 2009.

[7] J. Krwawicz, K. D. Arczewska, E. Speina, A. Maciejewska, and E. Grzesiuk, "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 Biochimica Polonica*, vol. 54, no. 3, pp. 413–434, 2007.

[8] R. Olinski, A. Siomek, R. Rozalski et al., "Oxidative damage to DNA and antioxidant status in aging and age-related diseases," *Acta Biochimica Polonica*, vol. 54, no. 1, pp. 11–26, 2007.

[9] B. Tudek, "Base excision repair modulation as a risk factor for human cancers," *Molecular Aspects of Medicine*, vol. 28, no. 3-4, pp. 258–275, 2007.

[10] R. de Bont and N. van Larebeke, "Endogenous DNA damage in humans: a review of quantitative data," *Mutagenesis*, vol. 19, no. 3, pp. 169–185, 2004.

[11] F. Drabløs, E. Feyzi, P. A. Aas et al., "Alkylation damage in DNA and RNA—repair mechanisms and medical significance," *DNA Repair*, vol. 3, no. 11, pp. 1389–1407, 2004.

[12] T. Lindahl, "Instability and decay of the primary structure of DNA," *Nature*, vol. 362, no. 6422, pp. 709–715, 1993.

[13] E. C. Friedberg et al., "DNA repair and mutagenesis," 2006.

[14] W. K. Hansen and M. R. Kelley, "Review of mammalian DNA repair and translational implications," *Journal of Pharmacology and Experimental Therapeutics*, vol. 295, no. 1, pp. 1–9, 2000.

[15] S. Raptis and B. Bapat, "Genetic instability in human tumors," *EXS*, no. 96, pp. 303–320, 2006.

[16] D. M. Wilson and D. Barsky, "The major human abasic endonuclease: formation, consequences and repair of abasic lesions in DNA," *Mutation Research*, vol. 485, no. 4, pp. 283–307, 2001.

[17] R. D. Wood, M. Mitchell, and T. Lindahl, "Human DNA repair genes, 2005," *Mutation Research*, vol. 577, no. 1-2, pp. 275–283, 2005.

[18] R. D. Wood, M. Mitchell, J. Sgouros, and T. Lindahl, "Human DNA repair genes," *Science*, vol. 291, no. 5507, pp. 1284–1289, 2001.

[19] F. Chen, W. Q. Liu, A. Eisenstark, R. N. Johnston, G. R. Liu, and S. L. Liu, "Multiple genetic switches spontaneously modulating bacterial mutability," *BMC Evolutionary Biology*, vol. 10, no. 1, article 277, 2010.

[20] R. D. Wood, M. Mitchell, and T. Lindahl, "Human DNA repair genes," 2010.

[21] M. Kanehisa, M. Araki, S. Goto et al., "KEGG for linking genomes to life and the environment," *Nucleic Acids Research*, vol. 36, no. 1, pp. D480–D484, 2008.

[22] K. Milanowska, J. Krwawicz, G. Papaj et al., "REPAIRtoire—a database of DNA repair pathways," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D788–D792, 2011.

[23] L. Wen and J. A. Feng, "Repair-FunMap: a functional database of proteins of the DNA repair systems," *Bioinformatics*, vol. 20, no. 13, pp. 2135–2137, 2004.

[24] L. Matthews, G. Gopinath, M. Gillespie et al., "Reactome knowledgebase of human biological pathways and processes," *Nucleic Acids Research*, vol. 37, no. 1, pp. D619–D622, 2009.

[25] E. C. Friedberg and L. B. Meira, "Database of mouse strains carrying targeted mutations in genes affecting biological responses to DNA damage Version 7," *DNA Repair*, vol. 5, no. 2, pp. 189–209, 2006.

[26] R. Caspi, T. Altman, J. M. Dale et al., "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases," *Nucleic Acids Research*, vol. 38, no. 1, Article ID gkp875, pp. D473–D479, 2009.

[27] M. Scheer, A. Grote, A. Chang et al., "BRENDA, the enzyme information system in 2011," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D670–D676, 2011.

[28] E. G. Cerami, B. E. Gross, E. Demir et al., "Pathway Commons, a web resource for biological pathway data," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D685–D690, 2011.

[29] M. Hackenberg, G. Barturen, and J. L. Oliver, "NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNAmethylation data," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D75–D79, 2011.

[30] S. Cotterill and S. E. Kearsey, "DNAReplication: a database of information and resources for the eukaryotic DNA replication community," *Nucleic Acids Research*, vol. 37, no. 1, pp. D837–D839, 2009.

[31] X. He, S. Chang, J. Zhang et al., "MethyCancer: the database of human DNA methylation and cancer," *Nucleic Acids Research*, vol. 36, no. 1, pp. D836–D841, 2008.

[32] M. Ongenaert, L. Van Neste, T. de Meyer, G. Menschaert, S. Bekaert, and W. van Criekinge, "PubMeth: a cancer methylation database combining text-mining and expert annotation," *Nucleic Acids Research*, vol. 36, no. 1, pp. D842–D846, 2008.

[33] C. Grunau, E. Renault, A. Rosenthal, and G. Roizes, "MethDB—a public database for DNA methylation data," *Nucleic Acids Research*, vol. 29, no. 1, pp. 270–274, 2001.

[34] C. A. Nieduszynski, S. I. Hiraga, P. Ak, C. J. Benham, and A. D. Donaldson, "OriDB: a DNA replication origin database," *Nucleic Acids Research*, vol. 35, no. 1, pp. D40–D46, 2007.

[35] R. J. Roberts, T. Vincze, J. Posfai, and D. Macelis, "REBASE-A database for DNA restriction and modification: enzymes, genes and genomes," *Nucleic Acids Research*, vol. 38, no. 1, Article ID gkp874, pp. D234–D236, 2009.

[36] E. Paek, J. Park, and K. J. Lee, "Multi-layered representation for cell signaling pathways," *Molecular and Cellular Proteomics*, vol. 3, no. 10, pp. 1009–1022, 2004.

[37] G. D. Bader, M. P. Cary, and C. Sander, "Pathguide: a pathway resource list," *Nucleic Acids Research*, vol. 34, pp. D504–D506, 2006.

[38] A. Shipra, K. Chetan, and M. R. S. Rao, "CREMOFAC—a database of chromatin remodeling factors," *Bioinformatics*, vol. 22, no. 23, pp. 2940–2944, 2006.

[39] A. L. Turinsky, B. Turner, R. C. Borja et al., "DAnCER: disease-annotated chromatin epigenetics resource," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D889–D894, 2011.

[40] J. D. Podlevsky, C. J. Bley, R. V. Omana, X. Qi, and J. L. Chen, "The Telomerase Database," *Nucleic Acids Research*, vol. 36, no. 1, pp. D339–D343, 2008.

[41] N. Weddington, A. Stuy, I. Hiratani, T. Ryba, T. Yokochi, and D. M. Gilbert, "ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data," *BMC bioinformatics*, vol. 9, p. 530, 2008.

[42] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[43] M. Safran, I. Dalah, J. Alexander et al., "GeneCards version 3: the human gene integrator," *Database: The Journal of Biological Databases and Curation*, vol. 2010, Article ID baq020, 2010.

[44] E. W. Sayers, T. Barrett, D. A. Benson et al., "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 37, no. 1, pp. D5–D15, 2009.

[45] L. D. Stein, "Using the Reactome database," *Current Protocols in Bioinformatics*, chapter 8, p. Unit 8.7, 2004.

[46] W. Ladiges, J. Wiley, and A. MacAuley, "Polymorphisms in the DNA repair gene XRCC1 and age-related disease," *Mechanisms of Ageing and Development*, vol. 124, no. 1, pp. 27–32, 2003.

[47] I. M. Keseler, C. Bonavides-Martínez, J. Collado-Vides et al., "EcoCyc: a comprehensive view of Escherichia coli biology," *Nucleic Acids Research*, vol. 37, no. 1, pp. D464–D470, 2009.

[48] L. Cabusora, E. Sutton, A. Fulmer, and C. V. Forst, "Differential network expression during drug and stress response," *Bioinformatics*, vol. 21, no. 12, pp. 2898–2905, 2005.

[49] J. Díez, D. Walter, C. Mũoz-Pinedo, and T. Gabaldón, "Editorial: DeathBase: a database on structure, evolution and function of proteins involved in apoptosis and other forms of cell death," *Cell Death and Differentiation*, vol. 17, no. 5, pp. 735–736, 2010.

[50] J. Kosinski, I. Hinrichsen, J. M. Bujnicki, P. Friedhoff, and G. Plotz, "Identification of Lynch syndrome mutations in the MLH1-PMS2 interface that disturb dimerization and mismatch repair," *Human Mutation*, vol. 31, no. 8, pp. 975–982, 2010.

[51] S. Shi et al., "Analysis of CASP8 targets, predictions and assessment methods," *Database*, vol. 2009, Article ID bap003, 2009.

[52] K. Arnold, L. Bordoli, J. Kopp, and T. Schwede, "The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling," *Bioinformatics*, vol. 22, no. 2, pp. 195–201, 2006.

[53] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[54] J. Söding, "Protein homology detection by HMM-HMM comparison," *Bioinformatics*, vol. 21, no. 7, pp. 951–960, 2005.

[55] M. A. Kurowski and J. M. Bujnicki, "GeneSilico protein structure prediction meta-server," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3305–3307, 2003.

[56] J. Kosinski, I. A. Cymerman, M. Feder, M. A. Kurowski, J. M. Sasin, and J. M. Bujnicki, "A "FRankenstein's Monster" approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation," *Proteins: Structure, Function and Genetics*, vol. 53, no. 6, pp. 369–379, 2003.

[57] J. Kosinski, M. J. Gajda, I. A. Cymerman et al., "FRankenstein becomes a cyborg: the automatic recombination and realignment of fold recognition models in CASP6," *Proteins: Structure, Function and Genetics*, vol. 61, no. 7, pp. 106–113, 2005.

[58] Y. Xiong, J. Liu, and D. Q. Wei, "An accurate feature-based method for identifying DNA-binding residues on protein surfaces," *Proteins: Structure, Function and Bioformatics*, vol. 79, no. 2, pp. 509–517, 2011.

[59] L. Wang, C. Huang, M. Q. Yang, and J. Y. Yang, "BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features," *BMC Systems Biology*, vol. 4, no. 1, article S3, 2010.

[60] Y. Ofran, V. Mysore, and B. Rost, "Prediction of DNA-binding residues from sequence," *Bioinformatics*, vol. 23, no. 13, pp. i347–i353, 2007.

[61] C. Yan, M. Terribilini, F. Wu, R. L. Jernigan, D. Dobbs, and V. Honavar, "Predicting DNA-binding sites of proteins from amino acid sequence," *BMC Bioinformatics*, vol. 7, article 262, 2006.

[62] S. J. de Vries, M. van Dijk, and A. M. Bonvin, "The HADDOCK web server for data-driven biomolecular docking," *Nature Protocols*, vol. 5, no. 5, pp. 883–897, 2010.

[63] M. van Dijk and A. M. J. J. Bonvin, "Pushing the limits of what is achievable in protein-DNA docking: benchmarking HADDOCK's performance," *Nucleic Acids Research*, vol. 38, no. 17, Article ID gkq222, pp. 5634–5647, 2010.

[64] M. J. Gajda, I. Tuszynska, M. Kaczor, A. Y. Bakulina, and J. M. Bujnicki, "FILTREST3D: discrimination of structural models using restraints from experimental data," *Bioinformatics*, vol. 26, no. 23, Article ID btq582, pp. 2986–2987, 2010.

[65] P. Yue, E. Melamud, and J. Moult, "SNPs3D: candidate gene and SNP selection for association studies," *BMC Bioinformatics*, vol. 7, article 166, 2006.

[66] V. Parthiban, M. M. Gromiha, and D. Schomburg, "CUPSAT: prediction of protein stability upon point mutations," *Nucleic Acids Research*, vol. 34, pp. W239–W242, 2006.

[67] D. Gilis and M. Rooman, "PoPMuSiC, an algorithm for predicting protein mutant stability changes. Application to prion proteins," *Protein Engineering*, vol. 13, no. 12, pp. 849–856, 2000.

[68] E. Capriotti, P. Fariselli, and R. Casadio, "I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure," *Nucleic Acids Research*, vol. 33, no. 2, pp. W306–W310, 2005.

[69] J. Cheng, A. Randall, and P. Baldi, "Prediction of protein stability changes for single-site mutations using support vector machines," *Proteins: Structure, Function and Genetics*, vol. 62, no. 4, pp. 1125–1132, 2006.

[70] I. A. Adzhubei, S. Schmidt, L. Peshkin et al., "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, no. 4, pp. 248–249, 2010.

[71] P. C. Ng and S. Henikoff, "Predicting deleterious amino acid substitutions," *Genome Research*, vol. 11, no. 5, pp. 863–874, 2001.

[72] B. Li, V. G. Krishnan, M. E. Mort et al., "Automated inference of molecular mechanisms of disease from amino acid substitutions," *Bioinformatics*, vol. 25, no. 21, pp. 2744–2750, 2009.

[73] E. Capriotti, R. Calabrese, and R. Casadio, "Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information," *Bioinformatics*, vol. 22, no. 22, pp. 2729–2734, 2006.