BioTechnologia vol. 97(4) • pp. 245-260 • 2016
Journal of Biotechnology, Computational Biology and Bionanotechnology **RESEARCH PAPERS**

DOI: 10.5114/bta.2016.64542

A novel approach for identifying DNA repair pathways proteins using an evolutionary approach: *Plasmodium falciparum* case study

KAJA MILANOWSKA[1, 2, 3] *, JUSTYNA WOJTCZAK [1]

[1] Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology,
Adam Mickiewicz University, Poznań, Poland

[2] Laboratory of Bioinformatics and Protein Engineering,
International Institute of Molecular and Cell Biology, Warszawa, Poland

[3] European Center for Bioinformatics and Genomics,
Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland

Abstract

Analysis of evolution-related proteins is one of the most effective methods for assigning functions to newly discovered proteins. We present a novel approach for identifying homologues that can be involved in DNA repair mechanisms in organisms without proteome annotation. This approach is based on the profiles built with sequences and secondary structures of proteins with known and well-described function. For this analysis, the profiles of DNA repair proteins from three model organisms (*Homo sapiens*, *Saccharomyces cerevisiae*, and *Escherichia coli*) were prepared and deposited in the DNA repair pathways database – REPAIRtoire (http://repairtoire.genesilico.pl) (Milanowska et al., 2011). The methodology formulated was used to analyze all proteins from *Plasmodium falciparum* and annotate those with *unreviewed* status that were recognized as taking part in DNA repair*.* The analysis identified novel proteins that have neither been previously described in the InParanoid database (O'Brien et al., 2005) as homologous or orthologous for the chosen model organisms, nor been identified by the HHsearch program (Soding, 2005). This approach is included as a search module on the REPAIRtoire website (http://repairtoire.genesilico.pl/homologs) and is available for usage.

Key words: DNA repair pathways, homologs detection, orthologs, pipeline, *Plasmodium falciparum*

Introduction

DNA repair pathways

The DNA repair pathways play crucial roles in stabilizing and maintaining the genetic information of all living organisms. The stability of the genome is constantly endangered by the endogenous metabolic processes, environmental factors, or errors occurring during the DNA-based cell processes. DNA modifications, especially, that would block or change the formation of Watson-Crick pairs may lead to genetic diseases. The DNA lesions often interfere with DNA replication or transcription. To avoid errors in DNA, during evolution organisms have formed different systems for error-prevention and DNA repair. By protecting the genome from a huge number of chemical and structural reactions, those systems maintain DNA stability and proper flow of genetic information. However, DNA mutations and replication errors are also one of the sources for genetic variety and evolution. In multicellular organisms some changes to the DNA sequences are desired, e.g., for the production of antibodies by the immunological systems (Slatter and Gennery, 2010). Hence, repair pathways must balance the negative and positive changes of the sequence and structure of the genome.

More than 20000 lesions are caused by the exogenous factors that occur during a day in an organism. The factors include spontaneous hydrolyzations of N-glycosidic bonds and deamination or alkylation in the DNA of mammal cells (Hansen and Kelley, 2000; Friedberg

* Corresponding author: Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznań, Poland; e-mail: kaja@amu.edu.pl

Table 1. Eight categories of DNA repair pathways

Short name	Full name	Description
BER	base-excision repair	started by the excision of modified DNA bases; two types known: short (SP-BER) and long (LP-BER)
DDR	direct reversal repair	direct recreation of native nucleotides by the removal of non-native chemical modifications
DDS	DNA damage signaling	induced as a response for DNA lesion caused by environmental and endogenous factors
HRR	homologous recombination repair	repair of double-stranded breaks with the usage of homologous DNA strand as a template for re-synthesis
MMR	mismatch repair	post-replicational DNA repair that removes errors introduced during replication (misinserted nucleotides, bulges, insertions, deletions)
NER	nucleotide-excision repair	removes huge DNA lesions; (TCR)-NER, transcription-coupled repair, repairs lesions occurring in the active strand of transcribed gene, (GGR)-NER, global genome repair, removes lesions present anywhere in the genome
NHEJ	non-homologous end joining repair	ligation of DNA ends introduced by double-stranded breaks in DNA
TLS	translesion synthesis	tolerance pathway that allows specialized polymerases to continue replication through the ignored errors

et al., 2006; Raptis and Bapat, 2006; Olinski et al., 2007; Tudek, 2007). Lesions are also induced by errors that occur during the DNA metabolic processes, such as formation of single- or double-stranded breaks or introduction of modified nucleotides during replication. In a body of a healthy human (10^{12} cells) there are 10^{16}-10^{18} repair events a day. Even though a protection system is present, some of the lesions "slip away" the DNA repair and as a consequence leads to mutations, aging, and different diseases. Some examples are cancer, neurodegeneration, or mutation in the gene of one of the proteins involved in MMR (mismatch repair) that causes Lynch syndrome (Gulati et al., 2011). There are 87 different types of DNA lesions described in REPAIRtoire database (http://repairtoire.genesilico.pl/damage/).

DNA repair is a very complicated process that involves many factors. For instance, till date 178 genes that encode proteins involved in the DNA repair have been identified in the human genome (Wood et al., 2001; Wood et al., 2005). They are engaged in different processes from finding broken DNA to recombination or apoptosis. These processes may be described as pathways comprising a series of steps. There are eight types of DNA repair pathways identified and described in the REPAIRtoire database (Table 1).

All these pathways can be represented as a series of enzymatic transformations between different DNA structures, catalyzed by dedicated set of molecules. Many of the proteins involved in DNA repair pathways are well known and described, whereas the information about the rest is poor or not yet established. For example, an enzyme that is capable of removing 5-hydroxymethyluridine, a product of thymine oxidation in human cells, must exist, but till date it has not been identified (Brissett and Doherty, 2009).

In this study, we propose a novel approach for searching protein orthologs that can be involved in DNA repair processes. The compiled methodology is based on the profiles prepared with the usage of sequence and secondary structure information about proteins with known function described as DNA repair factors. This approach was used on the *Plasmodium falciparum* proteome to try to identify the repair enzymes that have not yet been described. The repair pathways that were taken into consideration are HHR, BER, MMR, and NER. There are various methods for homologs detection, such as HHsearch, but they are not dedicated to DNA repair mechanisms – in our study we have created databases with profiles of known and described proteins involved in such mechanisms, which enables precise searches in the context of DNA repair analyzes.

MMR as a reference pathway

MMR was chosen as a reference pathway and used as a sanity check for our approach (described in following sections). Proteins involved in this pathway are

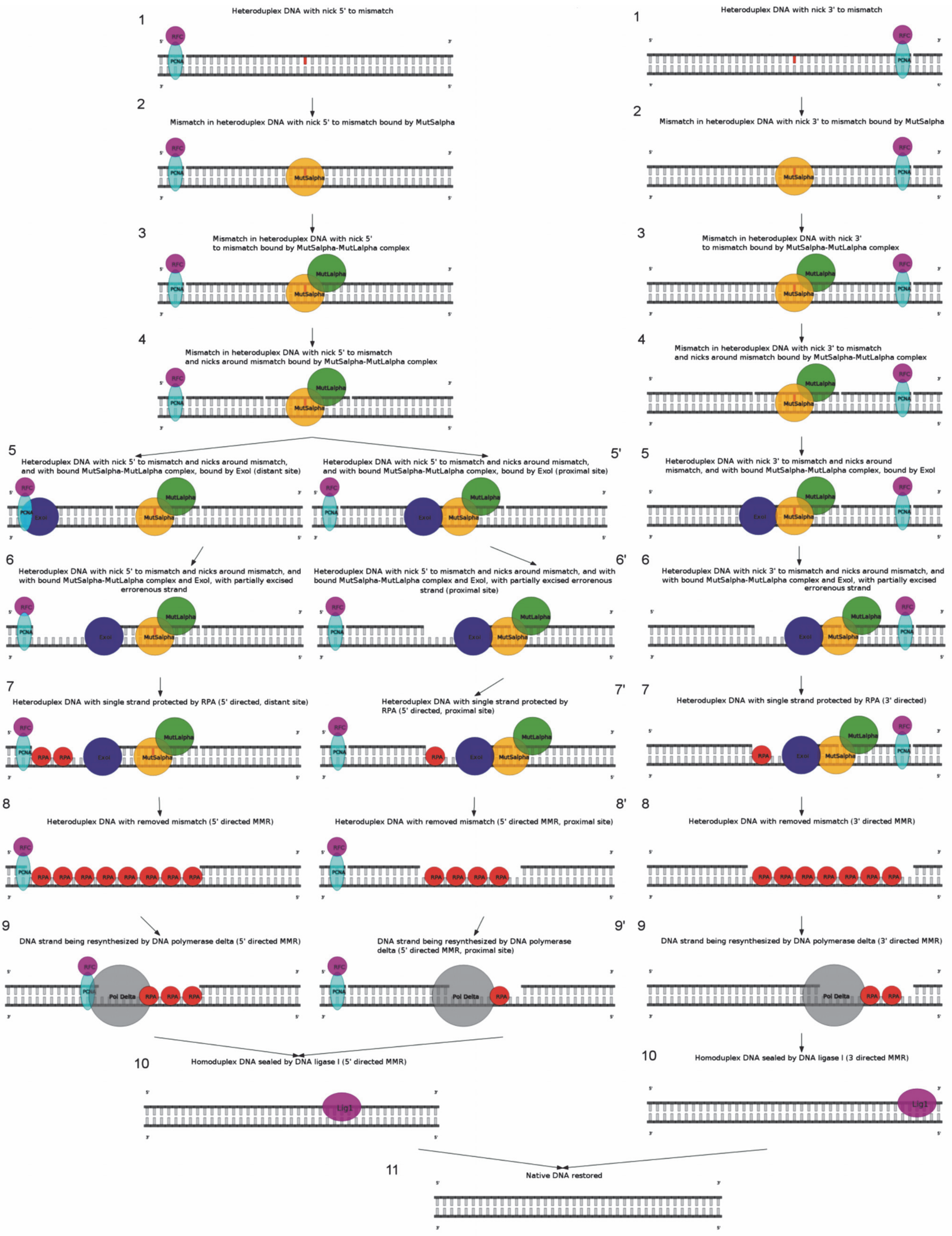

1
Heteroduplex DNA with nick 5' to mismatch
1
Heteroduplex DNA with nick 3' to mismatch
2
Mismatch in heteroduplex DNA with nick 5' to mismatch bound by MutSalpha
2
Mismatch in heteroduplex DNA with nick 3' to mismatch bound by MutSalpha
3
Mismatch in heteroduplex DNA with nick 5' to mismatch bound by MutSalpha-MutLalpha complex
3
Mismatch in heteroduplex DNA with nick 3' to mismatch bound by MutSalpha-MutLalpha complex
4
Mismatch in heteroduplex DNA with nick 5' to mismatch and nicks around mismatch bound by MutSalpha-MutLalpha complex
4
Mismatch in heteroduplex DNA with nick 3' to mismatch and nicks around mismatch bound by MutSalpha-MutLalpha complex
5
Heteroduplex DNA with nick 5' to mismatch and nicks around mismatch, and with bound MutSalpha-MutLalpha complex, bound by ExoI (distant site)
5'
Heteroduplex DNA with nick 5' to mismatch and nicks around mismatch, and with bound MutSalpha-MutLalpha complex, bound by ExoI (proximal site)
5
Heteroduplex DNA with nick 3' to mismatch and nicks around mismatch, and with bound MutSalpha-MutLalpha complex, bound by ExoI
6
Heteroduplex DNA with nick 5' to mismatch and nicks around mismatch, and with bound MutSalpha-MutLalpha complex and ExoI, with partially excised errorenous strand
6'
Heteroduplex DNA with nick 5' to mismatch and nicks around mismatch, and with bound MutSalpha-MutLalpha complex and ExoI, with partially excised errorenous strand (proximal site)
6
Heteroduplex DNA with nick 3' to mismatch and nicks around mismatch, and with bound MutSalpha-MutLalpha complex and ExoI, with partially excised errorenous strand
7
Heteroduplex DNA with single strand protected by RPA (5' directed, distant site)
7'
Heteroduplex DNA with single strand protected by RPA (5' directed, proximal site)
7
Heteroduplex DNA with single strand protected by RPA (3' directed)
8
Heteroduplex DNA with removed mismatch (5' directed MMR)
8'
Heteroduplex DNA with removed mismatch (5' directed MMR, proximal site)
8
Heteroduplex DNA with removed mismatch (3' directed MMR)
9
DNA strand being resynthesized by DNA polymerase delta (5' directed MMR)
9'
DNA strand being resynthesized by DNA polymerase delta (5' directed MMR, proximal site)
9
DNA strand being resynthesized by DNA polymerase delta (3' directed MMR)
10
Homoduplex DNA sealed by DNA ligase I (5' directed MMR)
10
Homoduplex DNA sealed by DNA ligase I (3 directed MMR)
11
Native DNA restored
RFC
PCNA
MutSalpha
MutLalpha
ExoI
RPA
Pol Delta
Lig1
5'
3'

Fig. 1. MMR pathway in humans (http://repairtoire.genesilico.pl/Pathway/27/)

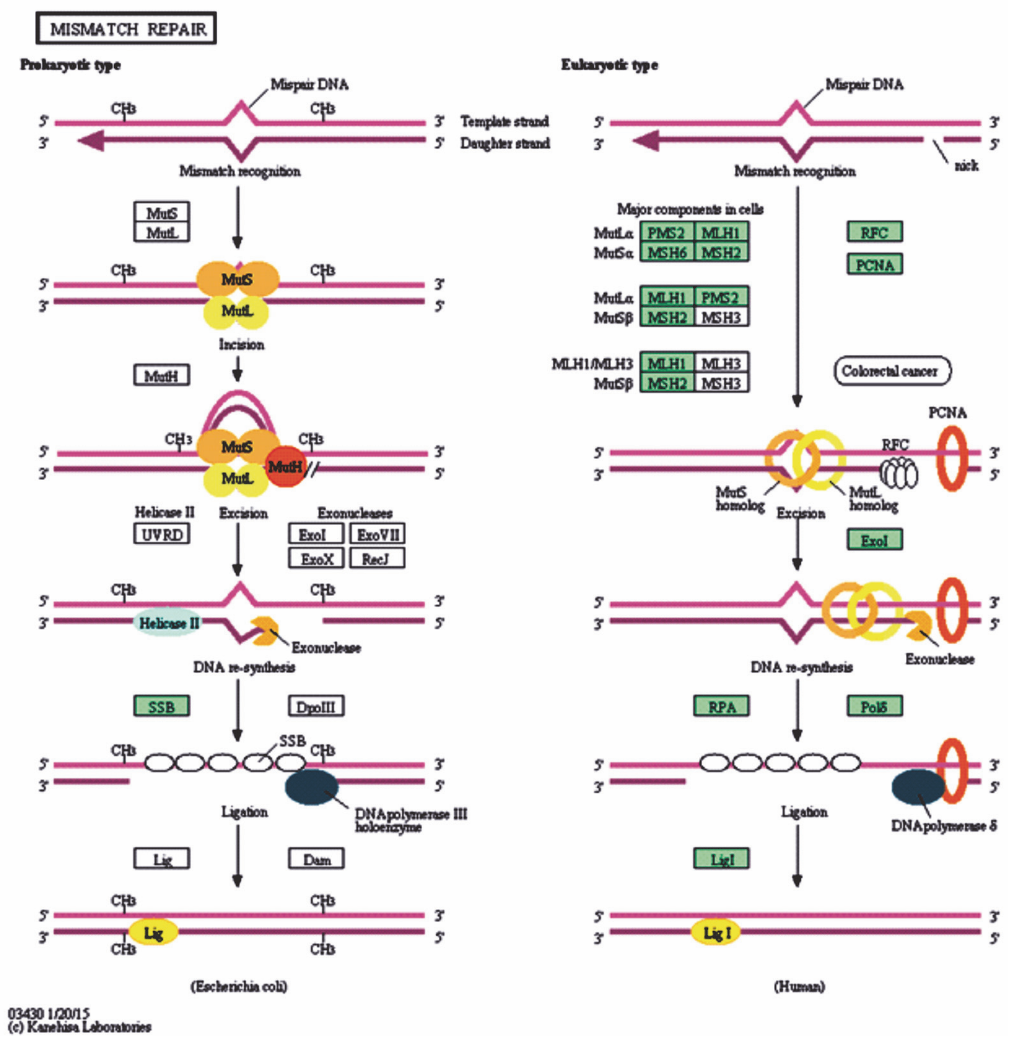


Fig. 2. MMR pathway from *Plasmodium falciparum* (isolate 3D7) based on the KEGG database (http://www.genome.jp/kegg/pathway.html) (Kanehisa and Goto, 2000). Prokaryotic proteins are depicted on the left side and the proteins from *Eukaryotes* are shown on the right side. Figure taken from KEGG: http://www.genome.jp/kegg-bin/show_pathway?org_name=pfa&mapno=03430

well described for human, yeast, and *E. coli* in the REPAIRtoire database. MMR is responsible for the repair of errors that occur during DNA replication, such as mismatches, small insertions, or deletions. Such errors can be introduced by incorrect functioning of the proofreading activity of a DNA polymerase (Kunz et al., 2009). For example, MMR pathway of human is shown as a set of enzymatic transformations between different structures of DNA, catalyzed by a defined collection of proteins (Fig. 1). The RFC-dependent protein PCNA binds the newly synthetized strand, just after nucleotides breaking. It is suspected that the repair mechanism uses a break left behind after transcription to properly localize lagging strand. The break can be localized on the 3′ or 5′ end toward a mismatched pair. Subsequently, a MutSα (comprising proteins: Msh2 and Msh6) is being relocated to the place of lesion. Incorporation of such complex would invoke DNA bending, conformational change of complex, ADP to ATP switch, and recruitment of another complex to be built from Mlh1 and Pms2 proteins named MutLα. MutSα-MutLα introduces nicks on both sides of the lesion. Further processing of the sequence bearing invalid nucleotide excision is being carried out depending on the end of the mismatched nucleotide toward the PCNA-RFC complex. If it is situated at the 3′ end, the exonuclease I (Exo I) joins at a distant site from the PCNA protein and excises the fragment between 5′ end and the mismatch, along with the mismatch and surrounding fragment. If it is situated at the 5′ end, the Exo I removes the mismatch and short fragments just before and after the mismatch. As a consequence, a single-stranded fragment is formed which attracts RPA proteins that will stabilize the strand. After disassociation of the excision complex, polymerase Delta resynthesizes the DNA strand and ligase ligates the fragments. The activities and steps listed in Figure 1 do not include: 1) ATP/ADP exchange and hydrolysis cycles by MutS and MutL and 2) passive sliding of MutS-MutL complex from the mismatch (there are alternative models explaining what happens after MutS binds mismatch and/or MutL such as "molecular switch model" (passive sliding), "active-translocation model", and "DNA bending model", in which MutS-MutL complex stays at the mismatch). The molecular switch model is most likely true, but still speculative, thus it is not included. For comparison, the MMR pathway from *Plasmodium falciparum* is shown in Figure 2. This pathway has not been well investigated till date. A group of proteins involved in this pathway in *P. falciparum* is described as putative based on the homology prediction from other well-described proteins. Analysis of proteins involved in MMR in *Plasmodium falciparum* may play a crucial role because dysfunctions of some of these proteins are responsible for the antibiotic resistance in malaria (Castellini, 2010). It means that point mutations in sequences responsible for repair may be advantageous for *Plasmodium*, but have a negative effect in humans resulting in difficulties in finding a cure.

Classifiers: TP, TN, FP, FN

General definition

For identifying homologs we used the following classifiers: TP (true positive), TN (true negative), FP (false positive), and FN (false negative). They are the classifiers used for defining the accuracy, precision, and recall of the methods used for homologous proteins detection. If these values are defined as:

TP – protein is properly defined as a homolog,
TN – protein is properly rejected as not being a homolog,
FP – protein is incorrectly defined as homolog,
FN – protein is incorrectly rejected,

then the following measures can be used:

- accuracy, AC:

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

- precision, PPV (defining the percentage of proteins described properly as homologs):

$$PPV = \frac{TP}{TP + FP} \times 100\%$$

- recall, REC (defining the percentage of true homologs found):

$$REC = \frac{TP}{TP + FN} \times 100\%$$

Theoretically, a perfect algorithm would get 100% for all the above-mentioned measures, but in reality, improving the accuracy would always result in lowering the precision. It means that finding more properly identified homologs would also result in enlarged value of improperly identified proteins that are not homologs. It is very important to properly balance these two values.

Definition used by HHsearch algorithm

The measures described earlier are used in HHSearch algorithm, but with specific definition. The HHsearch algorithm that was used in this work to apprise the classification of proteins homology uses MaxSub measure (Siew et al., 2000). If the secondary structure information is added to multiple sequence alignment (MSA), this measure specifies the quality of the secondary structure model of the protein. We have used an approach that adds secondary structure prediction to MSAs (addss.pl) (Fig. 7 and Fig. 8). The values of this measure are set to be between 0.0 and 1.0, and the higher it is the better is the structural alignment of the model and the native structure. The HHsearch depicts protein pair as TP, when both the domains belonging to the same superfamily in SCOP classification (Lo Conte et al., 2000) have a structural superposition MaxSub score >0.1. Pairs of proteins belonging to different classes and bearing MaxSub = 0 would not be classified as homologs. All other cases are classified by HHsearch as undefined (Soding, 2005).

HMM profiles

Another crucial element of finding and aligning distantly related sequences and identifying known protein domains in new sequences is profile analysis, which was also used in this study. The prepared HMM (Hidden Markov models) profiles were a basis for running our pipeline. A profile is a description of the consensus of MSA (multiple sequence alignment). It uses a position-specific scoring system to capture the information about the degree of conservation at various positions in the multiple alignment, which lifts the accuracy of the methods for homolog searching. The Hidden Markov models, HMMs, are one of the profile analysis approaches.

An HMM profile is a linear state machine consisting of a series of nodes, each of which corresponds roughly to a position (column) in the alignment from which it was built. If we ignore the gaps, the correspondence is exact – the HMM profile has a node for each column in the alignment and each node can exist in one state, a match state (Kabsch and Sander, 1983). The HMM profiles are very similar to simple sequence profiles, but additionally, besides the given frequency of amino acids in the columns from MSA, they come with the information about the position-dependent probabilities for insertions (insert states) and deletions (delete states) along the whole alignment. Logarithms of those probabilities are directly proportional to the position-dependent penalty for inserting a gap. The HMM profiles give better results than sequence profiles in discovering homologs and better alignments together with the price of lowering the speed of calculations. Several probabilities are bound to the HMM profiles. Among those is the transition probability, which equals 1 in a single model without gaps where the path through the model is linear (moving from match state n to match state $n+1$). Another probability is emission probability assigned to every state based on the probability of a given residue existence in a given position in the alignment. As an example, we can take a well-conserved column in a protein sequence alignment, where the emission probability for a certain amino acid will be 0.81, whereas for the other 19 amino acids it can equal 0.01. According to the path through the model to generate the sequence corresponding to the model, probabilities of each generated sequence depend on the transition probability to each state and emission probability for each node. To model true sequences, the possibility of emerging gaps must be taken into consideration. There are two types of possible gaps: insertion to sequence, when the sequence has a region not present in the model; deletion in sequence, when the region is present in the model, but not in the sequence. To serve both these types, every node in HMM profile has to have three possible states: match, insert, and delete. The model will have to have more types of transition probabilities: match → insertion, match → deletion, match → match, etc. (Fig. 3).

In HMM profile, the sequence alignment is made with a dynamic programming approach, which finds the most probable path from sequence to model. An example of HMM profile preparation for nucleotide sequences MSA is shown in Figure 4. In the first column, all nucleotides are the same, and that is why the emission probability for A is 1 and for the rest is 0. Probability of transition to other states equals 1, because the next position is set with nucleotides and not with gaps, so only the match state is possible. Emission distribution is 0.8 for C and 0.2 for G based on MSA. Following transition to the match probability will be 0.8 because there is one insertion in MSA giving a score of 0.2 for the transition to insert state (Fig. 4).

HMM alignment

In this study we used HMM alignment. This approach is used for homologs detection in programs such as HHsearch and HHblits from HHsuite. This alignment is based on the Viterbi algorithm (2015) used in the context of HMMs. Using the dynamic programming approach, the algorithm decodes hidden states in the sequence that could generate sequence of observations with the highest probability. This method is based on the assumption that the optimal path through the decoder to actual state comprises of the path with minimal score of moving back to any of the previous states and path proceeding to the actual state. The longer the time of observations and processing, the more reliable the result will be (optimal result is reached after about 5 iterations).

Materials and methods

Plasmodium falciparum proteome

Plasmodium falciparum is a unicellular parasite transmitted by the female *Anopheles* mosquito that causes malaria in human. First species of *Plasmodium* were sequenced in The Sanger Center as a part of the Malaria Genome Project. The genome of *Plasmodium* is distinguished by exceptionally low GC content at the 20% level.

Plasmodium lifecycle is very complex and has many states (sporozoit, merozoit, trophozoit, and gametocyte). The fact that this organism has to survive in two different environments, invertebrates and vertebrates, and has to constantly avoid immunological response of the

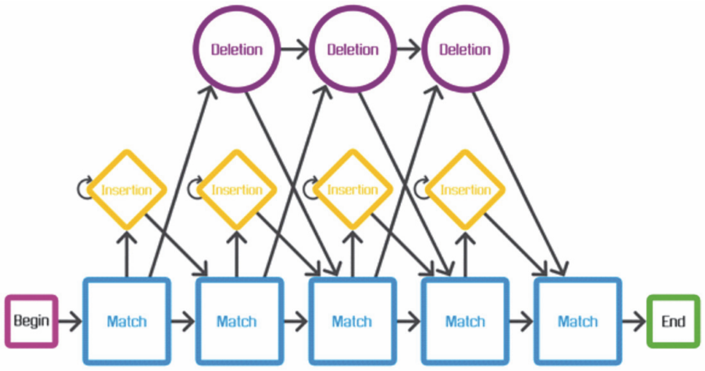


Fig. 3. Schematic representation of HMM profile creation.
Match states, insert states, and delete states are included (Eddy, 1998)

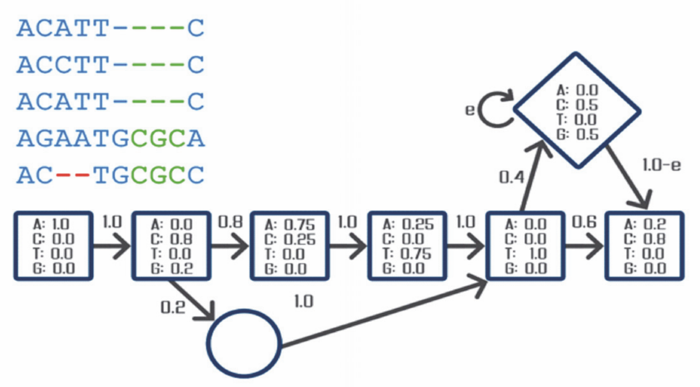

ACATT----C
ACCTT----C
ACATT----C
AGAATGCGCA
AC--TGCGCC
A: 0.0 C: 0.5 T: 0.0 G: 0.5
e
0.4
1.0-e
A: 1.0 C: 0.0 T: 0.0 G: 0.0
1.0
A: 0.0 C: 0.8 T: 0.0 G: 0.2
0.8
A: 0.75 C: 0.25 T: 0.0 G: 0.0
1.0
A: 0.25 C: 0.0 T: 0.75 G: 0.0
1.0
A: 0.0 C: 0.0 T: 1.0 G: 0.0
0.6
A: 0.2 C: 0.8 T: 0.0 G: 0.0
0.2
1.0

Fig. 4. Schematic representation of HMM profile generation of the MSA of nucleotide sequences

host has had an impact on the expression of its proteins (Florens et al., 2002).

For the analysis, all 5337 (as of March 4, 2016) proteins available in the NCBI database RefSeq (Pruitt et al., 2014) (http://www.ncbi.nlm.nih.gov/protein?LinkName=genome_protein&from_uid=33) were used. Among them are proteins that are already identified as involved in different DNA repair mechanisms identified with the usage of InParanoid database (Table 2).

HHsuite package

The programs from HHsuite package (Soding, 2005) were used during the analysis. HHsuite programs use hidden Markov models, BLAST (Altschul et al., 1990), and PSIPRED (Jones, 1999) programs. These programs were used to prepare the dataset (profiles databases of 292 proteins from REPAIRtoire database from the three model organisms: *Homo sapiens*, *Saccharomyces cerevisiae*, and *Escherichia coli*) for the novel approach and the methodology for homologous proteins detection described later.

The HHsearch program, a program that uses HMM profiles or MSA to search protein homologs, was used on a protein dataset from REPAIRtoire (total: 292 proteins, 154 for human, 79 for yeast, and 69 for *E. coli*) with the following databases: nr20 from NCBI (Sayers et al., 2009), COG (Tatusov et al., 2003), and PFAM (Finn et al., 2016) (Fig. 5) to obtain an .hrr file with phylogenetic information about protein families. We selected three databases as a base for profile preparation to search for the best solution. Generated files were then converted with hhblitsdb.pl program into three sets of databases for further analysis. This step is required to enable searches with the usage of HMM profiles.

HHblits

HHblits searches a database (e.g., PFAM (Fig. 6)) with a sequence or MSA query. It builds up a high-quality alignment starting from a single sequence or from MSA. Input data are transformed into HMM query and then used for iterative searches through UniProt20 or nr20 databases (clustered versions of UniProt (Uni-

Table 2. Number of *P. falciparum* proteins involved in the DNA repair pathways based on the information included in InParanoid database; the pathways that were used in the analysis are mentioned

DNA repair pathway	MMR	HRR	BER	NER
Number of *P. falciparum* proteins	18	13	11	31

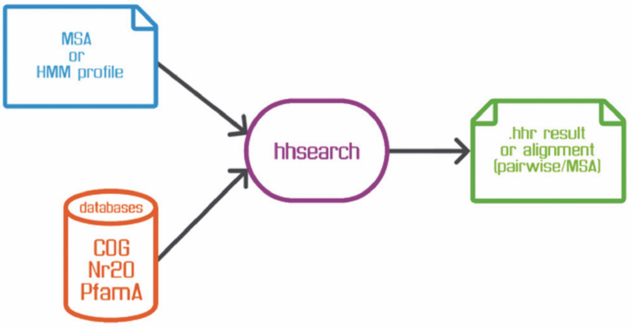


Fig. 5. HHsearch program: input and output data

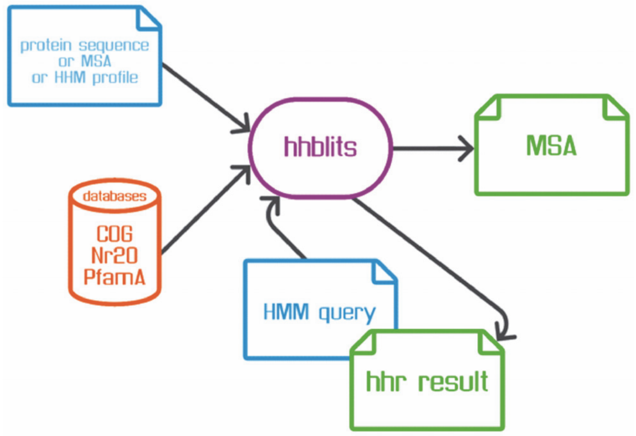


Fig. 6. HHblits program: input and output data

Prot, 2015) and nr from NCBI, respectively, containing only non-redundant sequences with a similarity up to 20%) by adding to the actual HMM query, the sequences from the previous query are to be used in the following iteration. HHblits is faster than PSI-BLAST algorithm (Altschul et al., 1997) and uses the same HMM-HMM aligning algorithm as HHsearch, but performs preliminary filtration that reduces the number of HMM profiles from millions to thousands and then produces HMM-HMM alignment.

Other HHsuite programs used

- HHmake was used to produce profiles for HHsearch (during testing of programs, not described in this paper),
- reformat.pl was used to convert between different used formats,
- hhblitsdb.pl was used to create a database for HHblits program,
- HHalign was used to create alignments of the query (MSA/HHM profile) to template MSA/HMM (during testing of programs, not described in this paper),
- addss.pl was used to add predicted secondary structure to the obtained results before creating an HMM profile for every protein,
- hhmakemodel.pl was used to create alignments based on the profiles search (during testing of programs, not described in this paper).

Databases

For this study the following databases were used:

- COG database (an evolutionary classification of genes, http://www.ncbi.nlm.nih.gov/COG/) – used in genomic analyzes providing access to genome annotations,
- UniProt database (Universal Protein Resource, http://www.uniprot.org/) – a set of annotations for protein sequences, comprises of UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters (UniRef), and UniProt Archive (UniParc),
- CATH database (http://www.cathdb.info) (Sillitoe et al., 2015) – classifies protein domains from PDB database (Berman et al., 2000) in the following groups: Class – class of secondary structure, Architecture – collection of topologies with a common features, Topology – families of protein folds bearing significant structural similarity, Homologous superfamily – similarity resulting from origination from common ancestor,
- KEGG database (*Kyoto Encyclopedia of Genes and Genomes*, http://www.genome.jp/kegg/) – a collection of many databases containing information about metabolic pathways, genes, and genomes; in this study we were mainly using KEGG Pathway database, offering online browser for metabolic pathways;
- InParanoid database (http://inparanoid.sbc.su.se/cgi-bin/index.cgi) – proteomic database of fully sequen-

ced eukaryotic organisms, plus *E. coli* proteome; enables orthologs, homologs, and paralogs searching based on the relations between two given organisms; communication with a database was facilitated by using XML formats: SeqXML and OrthoXML.

Python and Django

Our approach was implemented in Python programming language (version 2.7.6). The choice was made according to the availability of different libraries, packages, and modules, adapted for the bioinformatics analyzes. All libraries used in the project are listed in the Supplementary material Table 6.

A Django framework (version 1.5.4) was used to release the pipeline to the public on a REPAIRtoire website. Django is a framework based on MVC (Model-View-Controller) architectural pattern, written in Python, which enables easy integration with bioinformatics methods. It consists of an object-relational mapper (ORM) that mediates between data models (defined as Python classes) and a relational database ("Model"); a system for processing HTTP requests with a web-templating system ("View") and a regular-expression-based URL dispatcher ("Controller").

Results

Preparation of reference pathway

On the basis of the analysis of proteins involved in MMR and available in InParanoid and KEGG databases, we have manually created a set of proteins for three model organisms (proteins were manually searched in the above-mentioned databases in the context of MMR pathway) (Table 3).

Table 3. Quantitative summary of proteins involved in MMR form three model organisms and occurrence of their orthologs in relation to each of them

Orthologs	*H. sapiens*	*S. cerevisiae*	*E. coli*
H. sapiens	23	18	3
S. cerevisiae	18	20	3
E. coli	3	2	21

A number in each cell of the table represents the amount of mutual orthologous proteins from MMR pathway found in the organism in a given row and simultaneously present in the organism in a given column. For example, based on the KEGG database humans have 23 proteins involved in the MMR pathway. About 19 of them are present as orthologs in yeast, according to InParanoid. The Table 3 is diagonally symmetric, with one exception in yeast and *E. coli* – three proteins from yeast are present in *E. coli*, but only 2 from *E. coli are* present as orthologs in yeast.

From similar analysis, it has been found that among 18 proteins involved in *P. falciparum*'s MMR, 14 are orthologs present in human, 13 were found in yeast, and 2 in *E. coli* (Table 4).

Table 4. Quantitative summary of proteins involved in MMR in *Plasmodium falciparum* and occurrence of their orthologs in relation to each of the model organism

Ortholog	*H. sapiens*	*S. cerevisiae*	*E. coli*
P. falciparum	14	13	2

Reference pathway and all the proteins prepared were used as a sanity check for our approach. Proteins found by this method in *Plasmodium falciparum* were matched with this dataset.

Designing and optimization of the approach

The approach for detection of homologues was prepared. The methodology uses programs available in HHsuite package (Fig. 7). Different approaches and methods were tested. Tests included searching with the reference set described earlier with BLAST, PSI-BLAST, and HHSearch. None of these methods gave a satisfactory result, so not all proteins collected were found, thus we decided to combine different approaches to obtain a method with highest accuracy. The first step of preparing the methodology was collecting all proteins (292 proteins: 154 for human, 79 for yeast, and 69 for *E. coli*) from REPAIRtoire database and preparing HMM profiles for them. Three different sets of those profiles were prepared using HHblits and addss.pl. For obtaining three different sets, three databases were prepared for HHblits program with hhblitsdb.pl: nr20_12Aug11, pfamA_27.0, COG_18Feb11. First, every protein was used as a query for HHblits with nr database, then the information about secondary structure for the query was added (with addss.pl). All the files obtained were converted into profiles (reformat.pl) and then all the profiles were conver-

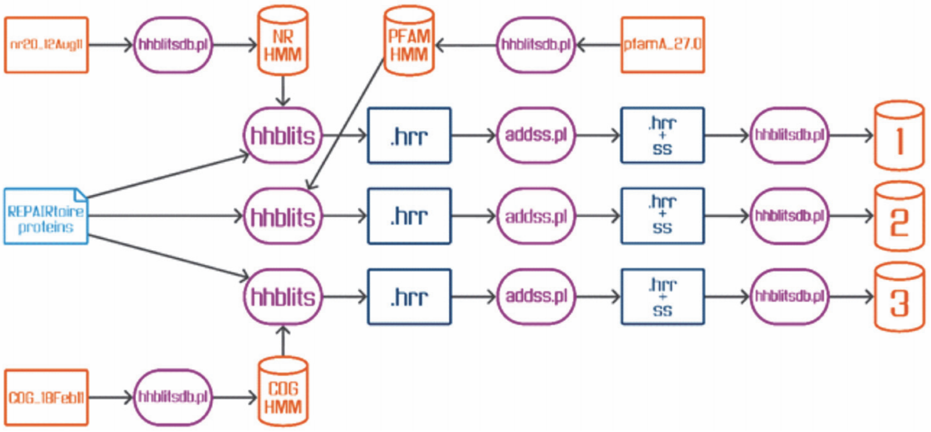


Fig. 7. Databases preparation for the methodology. During the preparation process REPAIRtoire proteins were used as an input for hhblits program with three different datasets: nr20_12Aug11, pfamA_27.0, and COG_18Feb11. After creating HHM profiles a secondary structure information was added and results were used to construct three databases used in further analysis. Numbers correspond to those three databases as follows: 1 – REPAIRtoire_NR_HMM_DB, 2 – REPAIRtoire_PFAM_HMM_DB, 3 – REPAIRtoire_COG_HMM_DB

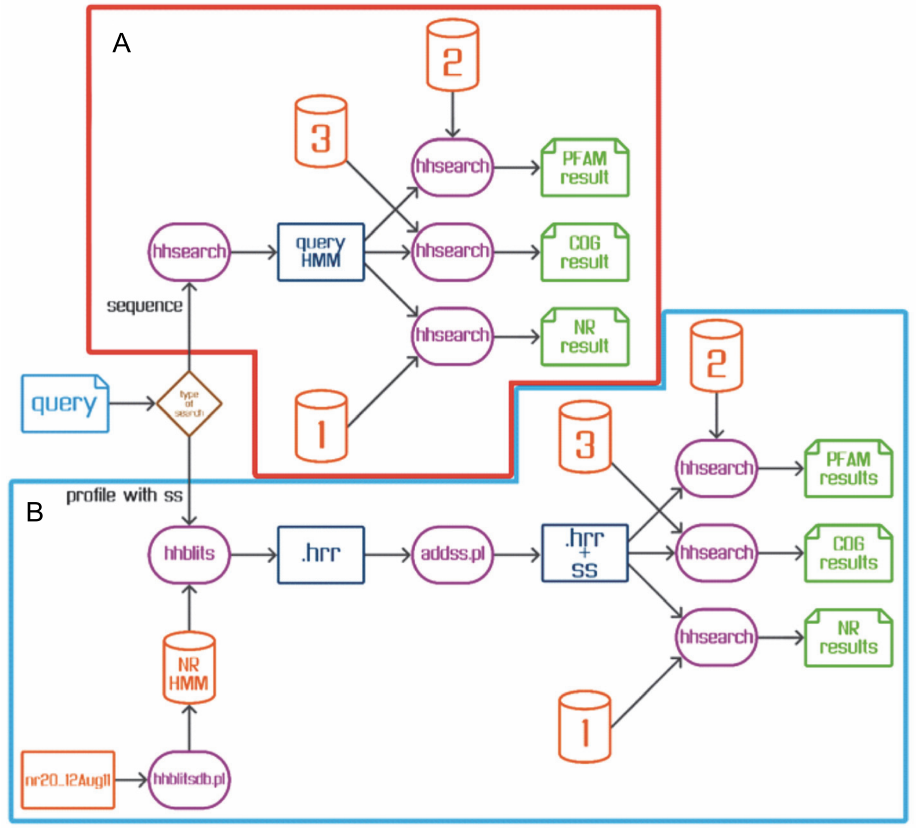


Fig. 8. Schematic representation of the designed approach. Protein sequence with unknown function is an input. A) Results obtained with sequence searches are shown in the upper part of the figure – after choosing the type of the search (sequence). B) Results obtained with profile enriched with secondary structure information are depicted in the lower part of the figure. Numbers correspond to used databases as follows: 1 – REPAIRtoire_NR_HMM_DB, 2 – REPAIRtoire_PFAM_HMM_DB, 3 – REPAIRtoire_COG_HMM_DB

ted into profile database using hhblitsdb.pl. The same procedure was conducted for all the proteins, but using PFAM and COG databases. Such approach allowed to construct three different databases ready for querying in the future steps (Fig. 7). This methodology is more accurate than the other methods because it uses both the HMM profiles, based on the information about protein families obtained from three different datasets, and the information about the secondary structure of proteins. And as the structure gets more conservative than the sequence of the protein, the information becomes more precise.

The piepeline was divided into two different approaches. First approach (Fig. 8A) employs HHsearch program and a protein sequence as an input. It runs three HHsearch searches on the three prepared databases from REPAIRtoire proteins: PFAM, COG, and NR. HHsearch queries the protein sequence against profile databases that contain information about secondary structure. If a single sequence is an input, HHsearch calculates an HHM profile for it and then aligns the query HMM against all the HMM profiles in a given database with the usage of Viterbi algorithm. After processing the query, the HMM profiles with the highest significance are aligned again with the usage of MAC algorithm (Maximum Accuracy). After that phase, the results are returned. Different parameters were tested during the development of this part. During the analysis of *P. falciparum* proteome dataset, it became obvious that the RAM memory is one of the factors that limit the analysis. To overcome this issue *maxmem <GB>* option was being tested to get the best performance. Resources of 20 processors and 3GB of *maxmem* allowed for the analysis of sequences with the length up to 2300 amino acids. To calculate results for longer sequences the *maxmem* parameter had to be changed to 32GB. Memory requirements are directly proportional to the number of processors used. To calculate a sequence of 11000 residues (the longest sequence of *Plasmodium falciparum* proteome), number of processors had to be set to 10.

$$maxmem = 0.5 \times 2^{30}B + l^2 \times (n + 1) \times 24B$$

l – length of the sequence; *n* – number of processors

The second approach (Fig. 8B) takes a protein sequence as an input and uses first HHblits with the prepared nr20_12Aug11 database to create an HMM profile for a query sequence, then employs addss.pl to add an information about the secondary structure and then passes the obtained profile to HHsearch. Now HHsearch performs three runs with three databases prepared on REPAIRtoire protein dataset (COG, PFAM, and NR). But it not only has the information about protein families and secondary structure for REPAIRtoire proteins, but also gets the same set of information for the given query. This way the method is able to find more distant proteins that are actual homologs for a given query. The ability of prediction of the function of unknown protein based on homology searches enables identification of homologs with evolutionarily distantly spaced sequences. In this approach, different parameters were also tested and adjusted as described earlier.

A problem in one of the used libraries occurred during all the tests of the approach in both approaches described earlier. A CSB module of *csb.bio.io.hhpred.py* was not able to process our data – the problem and the solution are described later.

Analysis of the *Plasmodium falciparum* proteome with the approach

Analysis was performed on a *P. falciparum* proteome (5337 proteins). All proteins were used as a query for the described methodology, both with the first and the second approach. We have obtained six datasets as a result (first approach on three databases: COG, NR, and PFAM from REPAIRToire and second approach on three databases from REPAIRtoire). To enable the analysis, scripts in Python were prepared (not available publicly).

Scripts for data analysis

For the analysis of the obtained data Python scripts were prepared. The scripts use CBS (Computational Structural Biology Toolbox) library (Kalev et al., 2012), which reads .hrr files produced by HHsearch algorithm. The result of this work was a patch for that library, as it was not able to analyze the results for sequences longer than 9999 amino acids. A patch was added to the module *csb.bio.io.hhpred.py* (problem and solution were reported to the author, but we have not yet received a response). The problem occurred if a five-digit values sequence positions (e.g., for very long input sequences) were present in the result files with annotations. Column indices were shifted, which in result killed the program. Overwriting the object's HHOutputParser private method *_parse* solved the problem.

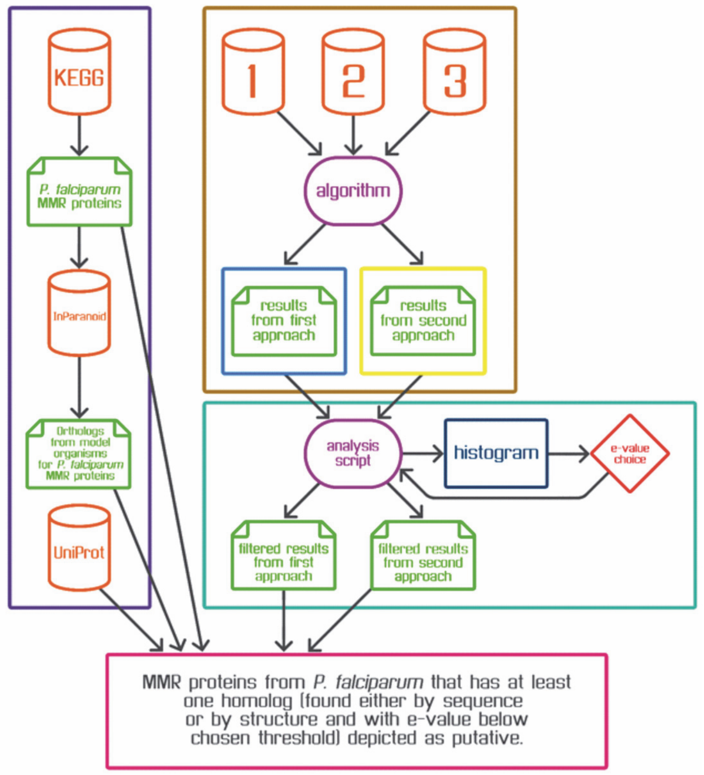

KEGG
P. falciparum MMR proteins
InParanoid
Orthologs from model organisms for P. falciparum MMR proteins
UniProt
1
2
3
algorithm
results from first approach
results from second approach
analysis script
histogram
e-value choice
filtered results from first approach
filtered results from second approach
MMR proteins from P. falciparum that has at least one homolog (found either by sequence or by structure and with e-value below chosen threshold) depicted as putative.

Fig. 9. The analysis pipeline. Brown box shows the approach described earlier. The blue box indicates scripts for analyzing the results. Violet box describes manual segregation of *Plasmodium falciparum* proteins that were used at the finishing stage of the analysis. These proteins were then compared to results from the analysis and shown in tables available in Supplementary material (Tables 1-4). Numbers corresponds to used databases as follows: 1 – REPAIRtoire_NR_HMM_DB, 2 – REPAIRtoire_PFAM_HMM_DB, 3 – REPAIRtoire_COG_HMM_DB

Results of the analysis

The analysis pipeline is show in Figure 9.

Results from the methodology without additional information about secondary structure

5337 proteins of *Plasmodium falciparum* proteome were analyzed. In Figure 10 a histogram of e-values for all the proteins is shown. All the values are shown in logarithmic scale against all proteins from REPAIRtoire database. Information about secondary structure was not used in this approach. There are 65 proteins from *P. falciparum* in the interval between e-value 0 to 1e-50. The information is cumulative, results from searching with three different databases were taken into consideration.

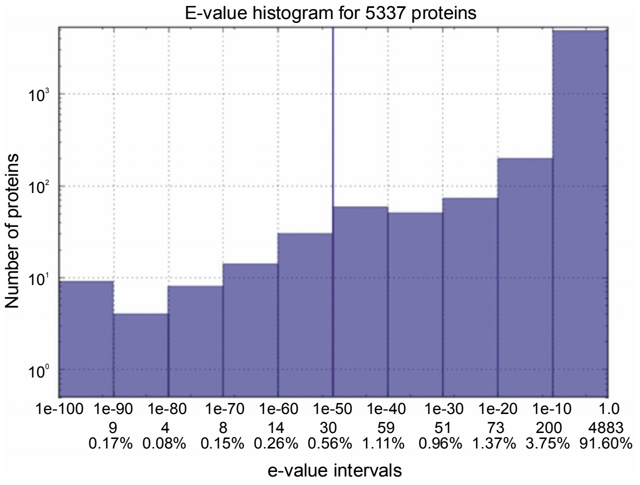


Fig. 10. E-values histogram for 5337 *Plasmodium falciparum* proteins. E-values are the result from the first approach – using sequence without structural information. Vertical blue line indicates chosen e-value threshold. Proteins below that threshold were used in further analysis

Results from the methodology with additional information about secondary structure

5331 proteins of *Plasmodium falciparum* proteome were analyzed. For six proteins, the information about secondary structure with the usage of addss.pl was not added. Those were the proteins with the length more than 8500 residues. Increasing *maxmem* parameter have not improved the performance (data not shown). In Figure 11 a histogram of e-values for all the proteins is

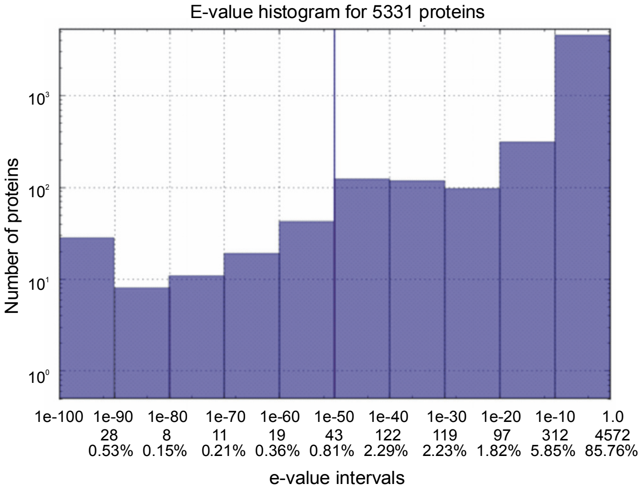


Fig. 11. E-values histogram for 5331 *Plasmodium falciparum* proteins. E-values are the results of the analysis using second approach: with HHblits, addss.pl, and HHsearach programs (adding structural information). Vertical blue line indicates chosen e-value threshold. Proteins below that threshold were used in further analysis

shown. All the values are shown in logarithmic scale for all the proteins from REPAIRtoire database. According to the described approach, first the HHblits program creates profiles, then an information about the secondary structure is added, and then HHsearch searches through the available databases. In the interval between e-value 0 to 1e-50 there are 109 proteins from *P. falciparum*. Potential homologs with e-value ≥1e-50 were not taken into consideration in the further analysis.

In the following step *P. falciparum* proteins potentially involved in MMR pathway (Fig. 2) were searched using KEGG database. For each of the proteins a check was performed in the InParanoid database if a homologous protein was found in chosen model organisms (*H. sapiens, S. cerevisiae,* and *E. coli*) (Supp. mat. Table 1 Columns 2 and 3). Subsequently, those proteins were searched in the results of our approach. All information is gathered in Supplementary material Table 1 and described in the Discussion. The same analysis was then done for NER, HHR, and BER pathways.

Results for pathways MMR, HHR, BER, and NER are shown in Supplementary material Tables 1-4, respectively.

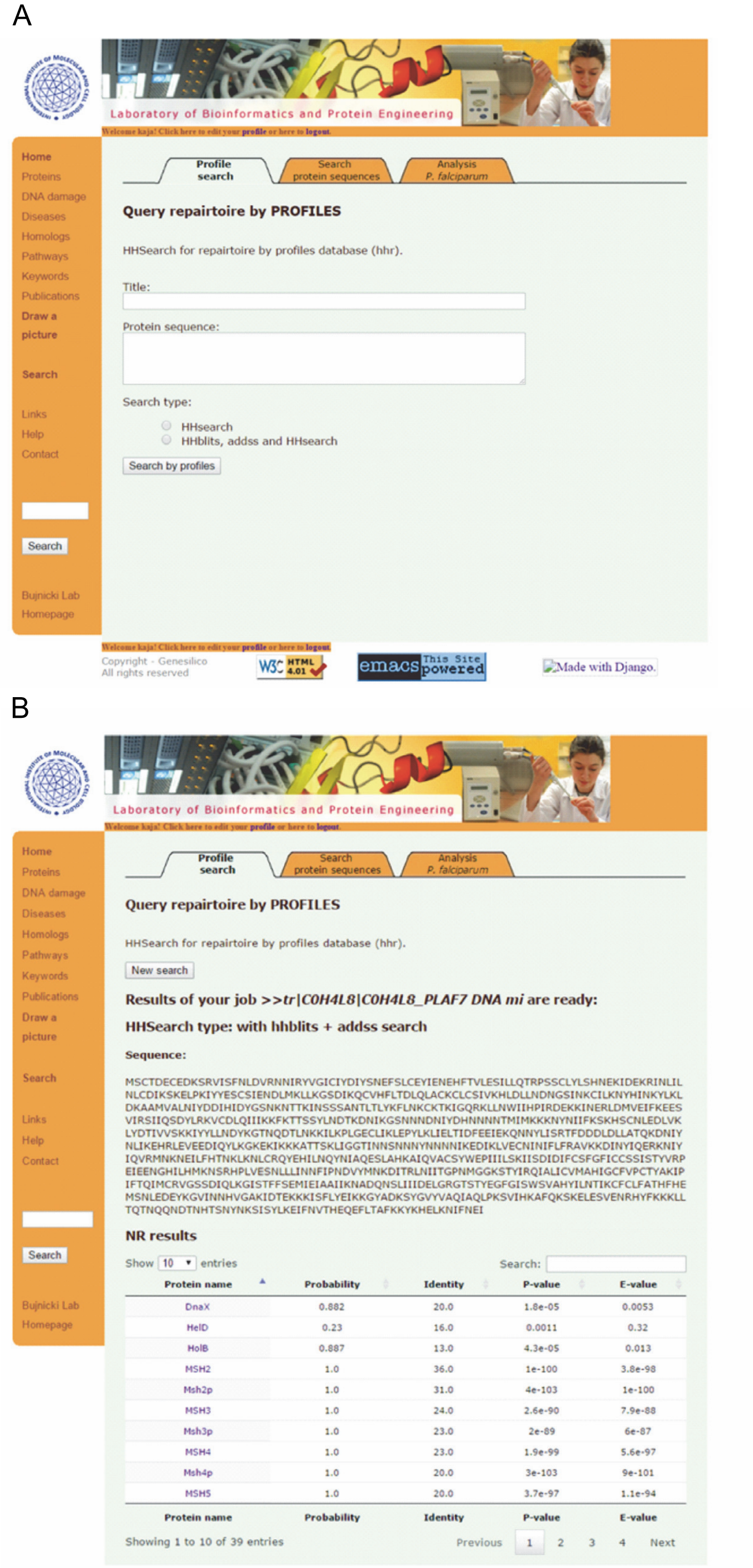
Pipeline and results on REPAIRtoire database website

REPAIRtoire database functionality was extended by adding the web implementation of the pipeline to the Search menu – Homologs menu (Fig. 12A) with three different tabs: Profile search (newly developed approach), Search Protein Sequences (using BLAST), and a tab presenting the results of the analysis described in this study. The example of the usage is shown in Figure 12, section B. During this study, another new functionality was added to the REPAIRtoire database – searching using a sequence with the BLAST algorithm.

The TAB showing results of the analysis of *P. falciparum* proteome shows all results of the analysis. The results can be found also in the Supplementary material (Table 5).

Discussion

In the case of *P. falciparum*, the developed methodology gave better results and performed more accurately in finding the potential homologs than using the tools based only on sequence searches. To limit the number of false positive results, e-value < 1e-50 was set up as a cut-off point, above which the results were not taken into consideration. For the chosen e-value the number of potential homologs that were found by the method based on search with the structure information (HHblits, addss.pl, and HHsearch) compared to the number of homologs found with the method using only sequence information (HHSearch) were 109 to 65, respectively. The summary of both methods gave 125 homologs. It means that the method based on structure information gave 60 more potential homologs than the sequence approach. Only 16 of those proteins were found solely by HHsearch and not by any other approach. According to the analysis of results given by the prepared methodology in the context of DNA repair pathways, new potential homologous proteins from *Plasmodium falciparum* were found that are not listed in the InParanoid database. For the MMR pathway (Supp. mat. Table 1), one *P. falciparum* protein was identified (Uniprot ID: Q8IBK1, http://www.uniprot.org/uniprot/Q8IBK1), which simultaneously was not described in InParanoid as an ortholog of any protein from human, yeast, or *E. coli.* This is exoribonuclease I bearing 5′-3′ exonuclease activity with the status "Unreviewed – Protein predicted" in the UniProt database. The homolog of this protein is a reviewed protein with UniProt ID: P39875 (http://www.uniprot.org/uniprot/P39875) from *S. cerevisiae* (described in UniProtKB), whose function (exodeocsiribonuclease I) was experimentally confirmed



Query repairtoire by PROFILES

HHSearch for repairtoire by profiles database (hhr).

Title:

Protein sequence:

Search type:

○ HHsearch
○ HHblits, addss and HHsearch

Search by profiles

Query repairtoire by PROFILES

HHSearch for repairtoire by profiles database (hhr).

New search

Results of your job >>*tr|C0H4L8|C0H4L8_PLAF7 DNA mi* are ready:

HHSearch type: with hhblits + addss search

Sequence:

```
MSCTDECEDKSRVISFNLDVRNNIRYVGICIYDIYSNEFSLCEYIENEHFTVLESILLQTRPSSCLYLSHNEKIDEKRINLIL
NLCDIKSKELPKIYYESCSIENDLMKLLKGSDIKQCVHFLTDLQLACKCLCSIVKHLDLLNDNGSINKCILKNYHINKYLKL
DKAAMVALNIYDDIHIDYGSNKNTTKINSSSANTLTLYKFLNKCKTKIGQRKLLNWIIHPIRDEKKINERLDMVEIFKEES
VIRSIIQSDYLRKVCDLQIIIKKFKTTSSYLNDTKDNIKGSNNNDNIYDHNNNNTMIMKKKNYNIIFKSKHSCNLEDLVK
LYDTIVVSKKIYYLLNDYKGTNQDTLNKKILKPLGECLIKLEPYLKLIELTIDFEEIEKQNNYLISRTFDDDLDLLATQKDNIY
NLIKEHRLEVEEDIQYLKGKEKIKKKATTSKLIGGTINNSNNNYNNNNIKEDIKLVECNINIFLFRAVKKDINYIQERKNIY
IQVRMNKNEILFHTNKLKNLCRQYEHILNQYNIAQESLAHKAIQVACSYWEPIIILSKIISDIDIFCSFGFICCSSISTYVRP
EIEENGHILHMKNSRHPLVESNLLLINNFIPNDVYMNKDITRLNIITGPNMGGKSTYIRQIALICVMAHIGCFVPCTYAKIP
IFTQIMCRVGSSDIQLKGISTFFSEMIEIAAIIKNADQNSLIIIDELGRGTSTYEGFGISWSVAHYILNTIKCFCLFATHFHE
MSNLEDEYKGVINNHVGAKIDTEKKKISFLYEIKKGYADKSYGVYVAQIAQLPKSVIHKAFQKSKELESVENRHYFKKKLL
TQTNQQNDTNHTSNYNKSISYLKEIFNVTHEQEFLTAFKKYKHELKNIFNEI
```

NR results

Show 10 ▾ entries Search:

Protein name	Probability	Identity	P-value	E-value
DnaX	0.882	20.0	1.8e-05	0.0053
HelD	0.23	16.0	0.0011	0.32
HolB	0.887	13.0	4.3e-05	0.013
MSH2	1.0	36.0	1e-100	3.8e-98
Msh2p	1.0	31.0	4e-103	1e-100
MSH3	1.0	24.0	2.6e-90	7.9e-88
Msh3p	1.0	23.0	2e-89	6e-87
MSH4	1.0	23.0	1.9e-99	5.6e-97
Msh4p	1.0	20.0	3e-103	9e-101
MSH5	1.0	20.0	3.7e-97	1.1e-94
Protein name	Probability	Identity	P-value	E-value

Showing 1 to 10 of 39 entries

Previous 1 2 3 4 Next

Fig. 12. A) Profile Search Tab with available options: HHsearch or the method using HHblits, adds, and HHsearch. B) The result of the search using a sequence of protein C0H4LB and HHblits, adds, HHsearch methodology. The results comprise of three sections – searching the profiles built on NR, COG, and PFAM databases

(experimental evidence at protein level) (Tishkoff et al., 1997). Also for the HHR pathway (Supp. mat. Table 2), one protein was identified – UniProt ID: C6KT89 (http://www.uniprot.org/uniprot/C6KT89) not defined in InParanoid as an ortholog of the selected organisms' proteins. This is a potential DNA Polymerase I with the status: "Unreviewed – Protein inferred from homology", which means that this protein is described in TrEMBL (computer-annotated TrEMBL section) database and has an ortholog in different organism. Protein P00582 (http://www.uniprot.org/uniprot/P00582) from *E. coli* with the status reviewed was identified during the analysis as a homolog for this result. Function of this protein (DNA polymerase I) was also tested experimentally (experimental evidence at protein level) (Brautigam and Steitz, 1998). For the BER pathway (Suppl. mat. Table 3), despite the protein being described for HHR pathway, one other protein was found with the UniProt ID: Q7KQJ9 (http://www.uniprot.org/uniprot/Q7KQJ9), which was not defined in InParanoid as an ortholog of any of the protein from human, yeast, or *E. coli*. This protein is recognized as a proliferating cell nuclear antigen, PCNA. In UniProt database its status is the same as for C6KT89 protein. An ortholog in yeast was found with our methodology – a protein P15873 (http://www.uniprot.org/uniprot/P15873) whose function is experimentally proven. For the NER pathway (Supp. mat. Table 4), despite of two above-mentioned proteins, two others were found. According to the results from the approach, Q8IJQ1 and Q8IBY6 are the orthologs of proteins from human. The first one in UniProt is described as "Unreviewed – Protein predicted" and belongs to CDK (cyclin-dependent kinases) proteins. P50613 (http://www.uniprot.org/uniprot/P50613) is its homolog (Schneider et al., 1998). The second *Plasmodium* protein (http://www.uniprot.org/uniprot/Q8IBY6) has the same status and its function is not described in UniProt database. A homolog for this protein that was found by our methodology is P23025 (http://www.uniprot.org/uniprot/P23025) known as XPA protein, whose function is binding an excision complex – ERCC1 and its recruitment to the lesion location in DNA, which is crucial in NER pathway (Li et al., 1995).

The described results show that the usage of our methodology enables discovery of new proteins that may be involved in DNA repair pathways and are not yet described as participating in these mechanisms. This methodology can be applied to any other type of research that is focused on annotating proteins with unknown function but potentially essential biological meaning. The method is available via the REPAIRtoire website for the purposes of finding new proteins involved in DNA repair.

Funding

This analysis was supported by the Polish Ministry of Science and Higher Education (grants Iuventus Plus 0360/IP1/2011/71 to K.M. and J.W.)

Acknowledgments

We would like to thank Janusz Bujnicki for all the help during the project and Piotr Stępniak, Joanna Kasprzak, and Anna Philips for critical reading of the paper. We are indebted to the authors of primary databases, services, and programs, whose content was used during the work.

References

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) *Basic local alignment search tool*. J. Mol. Biol. 215(3): 403-410.

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucl. Acids Res. 25(17): 3389-3402.

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. (2000) *The Protein Data Bank*. Nucl. Acids Res. 28(1): 235-242.

Brautigam C.A., Steitz T.A. (1998) *Structural principles for the inhibition of the 3′-5′ exonuclease activity of Escherichia coli DNA polymerase I by phosphorothioates*. J. Mol. Biol. 277(2): 363-377.

Brissett N.C., Doherty A.J. (2009) *Repairing DNA double-strand breaks by the prokaryotic non-homologous end-joining pathway*. Biochem. Soc. Trans. 37(3): 539-545.

Castellini M.A. (2010) *DNA Mismatch Repair in Plasmodium Falciparum: A Potential Mechanism for Accelerated Drug Resistance*.

Eddy S.R. (1998) *Profile hidden Markov models*. Bioinformatics 14(9): 755-763.

Finn R.D., Coggill P., Eberhardt R.Y., Eddy S.R., Mistry J., Mitchell A.L., Potter S.C., Punta M., Qureshi M., Sangrador-Vegas A. et al. (2016) *The Pfam protein families database: towards a more sustainable future*. Nucl. Acids Res. 44(D1): D279-285.

Florens L., Washburn M.P., Raine J.D., Anthony R.M., Grainger M., Haynes J.D., Moch J.K., Muster N., Sacci J.B., Tabb D.L. et al. (2002) *A proteomic view of the Plasmodium falciparum life cycle*. Nature 419(6906): 520-526.

Friedberg E.C., Aguilera A., Gellert M., Hanawalt P.C., Hays J.B., Lehmann A.R., Lindahl T., Lowndes N., Sarasin A., Wood R.D. (2006) *DNA repair: from molecular mechanism to human disease*. DNA Repair (Amst) 5(8): 986-996.

Gulati S., Gustafson S., Daw H.A. (2011) *Lynch Syndrome Associated With PMS2 Mutation: Understanding Current Concepts*. Gastrointest. Cancer Res. 4(5-6): 188-190.

Hansen W.K., Kelley M.R. (2000) *Review of mammalian DNA repair and translational implications*. J. Pharmacol. Exp. Ther. 295(1): 1-9.

Jones D.T. (1999) *Protein secondary structure prediction based on position-specific scoring matrices*. J. Mol. Biol. 292(2): 195-202.

Kabsch W., Sander C. (1983) *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features*. Biopolymers 22(12): 2577-2637.

Kalev I., Mechelke M., Kopec K.O., Holder T., Carstens S., Habeck M. (2012) *CSB: a Python framework for structural bioinformatics*. Bioinformatics 28(22): 2996-2997.

Kanehisa M., Goto S. (2000) *KEGG: Kyoto Encyclopedia of Genes and Genomes*. Nucl. Acids Res. 28(1): 27-30.

Kunz C., Saito Y., Schar P. (2009) *DNA Repair in mammalian cells: Mismatched repair: variations on a theme*. Cell Mol. Life Sci. 66(6): 1021-1038.

Li L., Peterson C.A., Lu X., Legerski R.J. (1995) *Mutations in XPA that prevent association with ERCC1 are defective in nucleotide excision repair*. Mol. Cell Biol. 15(4): 1993-1998.

Lo Conte L., Ailey B., Hubbard T.J., Brenner S.E., Murzin A.G., Chothia C. (2000) *SCOP: a structural classification of proteins database*. Nucl. Acids Res. 28(1): 257-259.

Milanowska K., Krwawicz J., Papaj G., Kosinski J., Poleszak K., Lesiak J., Osinska E., Rother K., Bujnicki J.M. (2011) *REPAIRtoire – a database of DNA repair pathways*. Nucl. Acids Res. 39(Database issue): D788-792.

O'Brien K.P., Remm M., Sonnhammer E.L. (2005) *Inparanoid: a comprehensive database of eukaryotic orthologs*. Nucl. Acids Res. 33(Database issue): D476-480.

Olinski R., Siomek A., Rozalski R., Gackowski D., Foksinski M., Guz J., Dziaman T., Szpila A., Tudek B. (2007) *Oxidative damage to DNA and antioxidant status in aging and age-related diseases*. Acta Biochim. Pol. 54(1): 11-26.

Pruitt K.D., Brown G.R., Hiatt S.M., Thibaud-Nissen F., Astashyn A., Ermolaeva O., Farrell C.M., Hart J., Landrum M.J., McGarvey K.M. et al. (2014) *RefSeq: an update on mammalian reference sequences*. Nucl. Acids Res. 42 (Database issue): D756-763.

Raptis S., Bapat B. (2006) *Genetic instability in human tumors*. EXS (96): 303-320.

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*. Nucl. Acids Res. 37 (Database issue): D5-15.

Schneider E., Montenarh M., Wagner P. (1998) *Regulation of CAK kinase activity by p53*. Oncogene 17(21): 2733-2741.

Siew N., Elofsson A., Rychlewski L., Fischer D. (2000) *MaxSub: an automated measure for the assessment of protein structure prediction quality*. Bioinformatics 16(9): 776-785.

Sillitoe I., Lewis T.E., Cuff A., Das S., Ashford P., Dawson N.L., Furnham N., Laskowski R.A., Lee D., Lees J.G. et al. (2015) *CATH: comprehensive structural and functional annotations for genome sequences*. Nucl. Acids Res. 43 (Database issue): D376-381.

Slatter M.A., Gennery A.R. (2010) *Primary immunodeficiencies associated with DNA-repair disorders*. Expert Rev. Mol. Med. 12: e9.

Soding J. (2005) *Protein homology detection by HMM-HMM comparison*. Bioinformatics 21(7): 951-960.

Tatusov R.L., Fedorova N.D., Jackson J.D., Jacobs A.R., Kiryutin B., Koonin E.V., Krylov D.M., Mazumder R., Mekhedov S.L., Nikolskaya A.N. et al. (2003) *The COG database: an updated version includes eukaryotes*. BMC Bioinformatics 4: 41.

Tishkoff D.X., Boerger A.L., Bertrand P., Filosi N., Gaida G.M., Kane M.F., Kolodner R.D. (1997) *Identification and characterization of Saccharomyces cerevisiae EXO1, a gene encoding an exonuclease that interacts with MSH2*. Proc. Natl Acad. Sci. USA 94(14): 7487-7492.

Tudek B. (2007) *Base excision repair modulation as a risk factor for human cancers*. Mol. Aspects Med. 28(3-4): 258-275.

UniProt (2015) *UniProt: a hub for protein information*. Nucl. Acids Res. 43(Database issue): D204-212.

The Viterbi Algorithm (2015) http:www.cim.mcgill.ca/~latorres/Viterbi/va_alg.htm.

Wood R.D., Mitchell M., Lindahl T. (2005) *Human DNA repair genes, 2005*. Mutat. Res. 577(1-2): 275-283.

Wood R.D., Mitchell M., Sgouros J., Lindahl T. (2001) *Human DNA repair genes*. Science 291(5507): 1284-1289.