INNSPUB

International Journal of Biosciences | IJB |
ISSN: 2220-6655 (Print) 2222-5234 (Online)
http://www.innspub.net
Vol. 25, No. 3, p. 222-229, 2024

RESEARCH PAPER OPEN ACCESS

Computational annotation of a hypothetical protein (DR_0423) of radiation resistant bacteria, *Deinococcus radiodurans*

Safaiatul Islam*, Abu Hena Mostofa Kamal, Md Ziaur Rahman, Protul Kumar Roy, A. Y. K. Md. Masud Rana

Molecular Radiobiology and Biodosimetry Division, Institute of Food and Radiation Biology, Bangladesh Atomic Energy Commission, Savar, Dhaka, Bangladesh

Key words: DNA, Genome, Annotation, Radiation resistant, NCBI, Hypothetical protein, Modeling, *Deinococcus radiodurans*

http://dx.doi.org/10.12692/ijb/25.3.222-229 Article published on September 10, 2024

Abstract

Deinococcus radiodurans, an extremophile bacterium renowned for its exceptional resistance to radiation, serves as a model organism for studying DNA repair mechanisms. This study focuses on the hypothetical protein DR_0423 (accession AAF10011) from *D. radiodurans*, aiming to elucidate its structural and functional characteristics through bioinformatics analysis. The sequence of DR_0423 was retrieved and analyzed using BLASTp, revealing significant homology with single-stranded DNA-binding proteins, particularly DdrA proteins involved in DNA damage response. Multiple sequence alignment and phylogenetic analysis confirmed its close relationship with DNA repair proteins across different *Deinococcus* species. Physicochemical analysis indicated a cytoplasmic localization with a molecular weight of 23,002.15 Da and a theoretical pI of 6.17, suggesting stability under varying conditions. Secondary structure predictions showed a mix of alpha helices, beta strands, and random coils, while 3D modeling based on homologous templates (e.g., Rad52 protein). Validation of the model using PROCHECK and other quality assessment tools confirmed its reliability, and CASTp analysis identified a key active site involved in DNA binding. The findings highlight DR_0423's potential role in DNA repair, reinforcing the importance of further experimental studies to validate these computational predictions and explore its applications in biotechnology and bacterial genetics.

* **Corresponding Author:** Safaiatul Islam ✉ safaiatul.bge28@gmail.com

Introduction

Deinococcus radiodurans is an extremophile bacterium and one of the most radiation-resistant microorganisms. The genus *Deinococcus* is a member of the family Deinococcaceae of the phylum *Deinococcus thermus*. In the members of this genus, cells are aerobic, non-motile, non-spore-forming, red or pink in color, coccoid or rod-shaped (Zhang *et al.*, 2007). The cell envelope of *D. radiodurans* is very close to the cell walls of Gram-negative organisms but Deinococcusy often stains-Gram positive because of their thick cell wall that consist of peptidoglycan (Brooks *et al.*, 1980). These species have been isolated from various environments such as soils air activated sludge, marine fish, termite gut, car air-conditioning system, rhizosphere, water, sewage, hot spring, and Antarctic environments (Kim *et al.*, 2018). The world's toughest bacterium can survive cold, dehydration, vacuum, and acid (Cox and Battista, 2005). They have the capability to withstand 5 kGy radiation doses that are lethal to other bacteria which makes them especially important. Treatment of *D. radiodurans* with high levels of ionizing radiation can produce hundreds of genomic double-strand breaks, but the genome is reassembled and repaired accurately before initiation of the next cycle of cell division (Mohseni *et al.*, 2014). The extraordinary capacity of *Deinococcus* spp. to reconstitute their genomes has inspired researchers on the investigation of the genome information. In August 2020, scientists reported that bacteria from Earth, particularly *Deinococcus radiodurans* bacteria, were found to survive for three years in outer space, based on studies conducted on the International Space Station (ISS). These findings support the notion of panspermia, the hypothesis that life exists throughout the Universe, distributed in various ways, including space dust, meteoroids, asteroids, comets, planetoids, or contaminated (Ott *et al.*, 2020).

In *Deinococcus family, D. radiodurans* R1 is the best characterized. The complete DNA sequence of *D. radiodurans* was published in 1999 by the Institute for Genomic Research (TIGR, 2004). The *D. radiodurans* chromosome is 3.28 Mb, with a GC content of 66.6%. The genome is segmented and 4 - 10 genome copies per cell (Lin *et al.*, 1999). Experimental evidence indicates that the recovery of *D. radiodurans* from substantial DNA damage relies on both features of *Deinococcus* physiology and a robust complement of repair enzymes (Dulermo *et al.*, 2015). Among these, most of the gene are identical and encode proteins of unknown function. Thus, bioinformatics approaches can play an important role in predicting and analyzing various forms of structure of those hypothetical proteins, their biological functions as well as protein-protein interactions. With the advancement of in-silico analysis, it became easier to annotate function to a hypothetical protein using various bioinformatics tools. The purpose of this study was to assign structural and biological function to the hypothetical protein DR_0423 (accession, AAF10011) of *D. radiodurans*. Additionally, homology modeling techniques were attempted to build a high-quality model of the DR_0423. This is the first study to computationally analyze a putative protein from *D. radiodurans,* a radiation-resistant bacterium.

Materials and methods

Sequence retrieval and similarity identification

The sequence information of the hypothetical protein (DR_0423) of *D. radiodurans* was retrieved from the National center for Biotechnology Information (NCBI) database. The sequence was then collected as a FASTA format sequence. To get the initial prediction about the function of the targeted hypothetical protein, similarity search was performed with the NCBI protein Database against non-redundant and SwissProt database (Boeckmann *et al.*, 2003) by using BLASTp program (Jhonson *et al.*, 2008).

Multiple sequence alignment and phylogeny analysis

The EBI's MUSCLE server (Madeira *et al.*, 2019) (https://www.ebi.ac.uk/Tools/msa/muscle/) was used for multiple sequence alignment, and CLC Sequence Viewer 7.0.2 (http://www.clcbio.com) was used to view the results. MEGA (Tamura *et al.*, 2011) was the tool used for the phylogeny study.

Physiochemical properties analysis

The ProtParam (http://web.expasy.org/protparam/) (Gasteiger *et al.*, 2003) tool of ExPASy performed the physical and chemical properties, such as molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, total number of negatively charged residues (Asp + Glu), total number of positively charged residues (Arg + Lys), instability index, aliphatic index, and grand average of hydrophilicity (GRAVY) predictions, among other things.

Subcellular localization analysis

CELLO predicted subcellular localization (Yu *et al.*, 2006). The subcellular localization predictions from PSLpred (Bhasin *et al.*, 2005), PSORTb (Yu *et al.*, 2010) and SOSUIGramN (Imai *et al.*, 2008) were also cross-checked with the results. For the topology prediction, TMHMM (Moller *et al.*, 2001), HMMTOP (Tusnady *et al.*, 2001), and CCTOP (Dobson *et al.*, 2015) were employed.

Conserved domain, motif, fold, coil, family, and superfamily identification

A conserved domain search was conducted at the NCBI's conserved domain database (accessible at CDD) (Marchler *et al.*, 2005). The Motif (Genome Net) service was used to search for protein motif (Kanehisa *et al.*, 2002). The Pfam and SuperFamily (Wilson *et al.*, 2007) databases were searched in order to determine the evolutionary relationships of the protein. The COILS server (Lupas *et al.*, 1991) was used to identify the protein's coiled-coil structure. The functional study of the protein was conducted using the InterProScan protein sequence analysis and classification system (Hunter *et al.*, 2009). Using the PFP-FunD SeqE server, protein folding pattern recognition was accomplished. Additionally, a search using STRING 10.0 (Szklarczyk *et al.*, 2015) was conducted to find a potential functional interaction network for the protein.

Secondary structure prediction

The secondary structure of the proteins was predicted using PSI-blast based secondary structure prediction (PSIPRED) (McGuffin *et al.*, 2000) and self-optimized prediction method with alignment (SOPMA) servers (Geourjon and Deleage, 1995).

Three-dimensional structure prediction

Based on the pairwise comparison profile of hidden Markov models (HMMs), the Max Planck Institute for Developmental Biology, Tübingen's HHpred server (https://toolkit.tuebingen.mpg.de /tools/hhpred) (Zimmermann *et al.*, 2018) predicted the three-dimensional structure. The best scoring template was used to forecast the 3D structure in order to increase accuracy. Afterwards, the YASARA energy minimization server was used to refine the 3D structure (Krieger *et al.*, 2009).

Model quality assessment

Finally, PROCHECK (https://servicesn.mbi.ucla.edu/PROCHECK/) (Laskowski *et al.*, 1993), Verify3D (http://nihserver.mbi.ucla.edu/Verify_3D/) (Eisenberg *et al.*, 1997), and ERRAT Structure Evaluation server (https:// servicesn.mbi.ucla.edu/ERRAT/) (Colovos and Yeates, 1993) were used for quality assessment of the predicted three dimensional structure.

Active site detection

The Computed Atlas of Surface Topography of Protein (CASTp) (http://sts.bio engr.uic.edu /castp/) (Dundas *et al.*, 2006) is an online resource that finds, defines, and measures concave surface regions on three-dimensional protein structures. It was used to locate the protein's active site.

Results and discussion

Sequence and similarity information

The BLASTp result against non-redundant and SwissProt database showed homology with other single-stranded DNA-binding protein DdrA proteins (Table 1). Phylogenetic tree was constructed based on the alignment and BLAST result which gives the similar concept about the protein (Fig. 1).

Table 1. Similar protein obtained from non-redundant SwissProt KB sequences

Protein ID	Protein Name	Organism	Identity %
Q9RX92	Single-stranded DNA-binding protein	*Deinococcus radiodurans*	100
A0A6G7H5S4	Single-stranded DNA-binding protein	*Deinococcus wulumuqiensis*	94
H8GY45	DNA damage response protein DdrA	*Deinococcus gobiensis*	83
A0A0F7JNA6	Single-stranded DNA-binding protein	*Deinococcus soli*	80
A0A100HKH2	DNA damage response protein DdrA	*Deinococcus grandis*	78

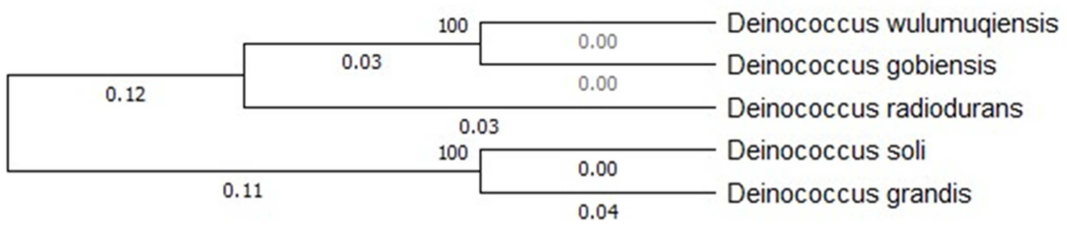

Fig. 1. Phylogenic trees with true distance of different DNA binding proteins

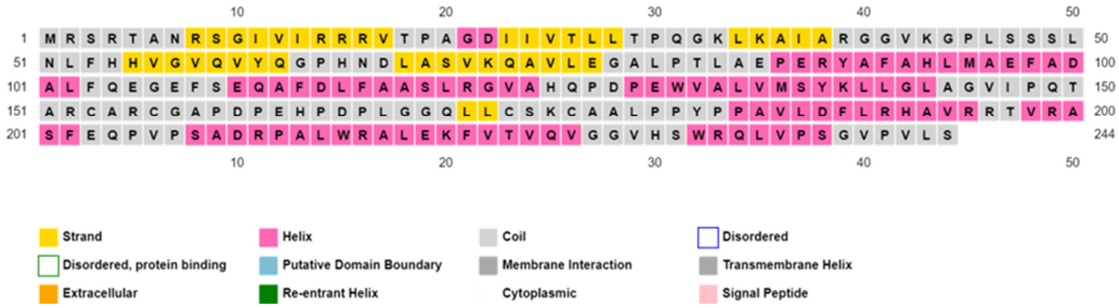

Fig. 2. Secondary structure of hypothetical protein

Physicochemical features

The protein consists of 208 amino acids. The calculated molecular weight was 23002.15 Da and theoretical pI was 6.17 indicating the protein to be negatively charged and stable one. Aliphatic index was 78.34 which give an indication of proteins' stability over a wide temperature range. Protein half-life computed was found to be 30 h (in vivo). And the molecular formula of protein was identified as $C_{1020}H_{1600}N_{296}O_{298}S_7$.

Secondary structure analysis

The PSIPRED secondary structure prediction server analysis revealed the proportions of alpha helix, extended strand and the random coil of protein as 33.89%, 20.87%, and 40.13%, respectively (Fig. 2).

Subcellular localization nature

Subcellular localization analysis was predicted by CELLO and validated by PSORTb and PSLpred. The subcellular localization of the hypothetical protein was predicted to be a cytoplasmic protein (Table 2).

Table 2. Subcellular localization analysis

SL	Analysis	Result
1	CELLO 2.5	Cytoplasmic localization
2	PSORTb	Cytoplasmic localization
3	PSLpred	Cytoplasmic protein
4	TMHMM 2.0	No transmembrane helices present
5	HMMTOP	No transmembrane helices present
6	CCTOP	Not transmembrane protein

Absent of transmembrane helices predicted by THMM and HMMTOP also emphasizes the result of being a cytoplasmic protein. Also, CCTOP server predicted that the query protein was not a transmembrane protein. All these results summarize the protein as a cytoplasmic one.

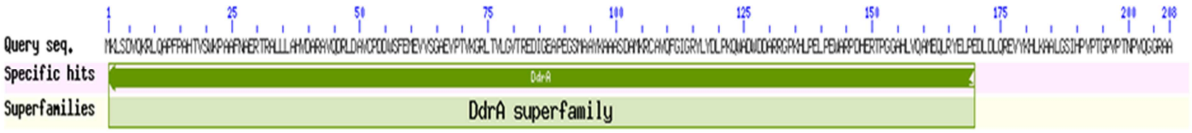

Fig. 3. Functional annoataion of protein by CDD

Functional annotation of hypothetical protein

The conserved domain search tool revealed that this hypothetical protein sequence was found to have only one domain (accession No. COG4712). The result was also checked by two other domain searching tools namely InterProScan and Pfam. Helix- Hairpin -Helix (HHH) motif like domain, was also found by Motif server. Superfamily search revealed DdrA superfamily like Rad43, Rad22, RecT family which are ssDNA annealing protein (Harris *et al.*, 2004). DdrA is involved in replication, recombination and repair of DNA, (Fig. 3). Fold pattern recognition by PFP-FunDSeqE tool revealed the presence of a '(TIM)-barrel' fold within the protein sequence. (TIM)-barrel structure is generally eight stranded α/β barrel.

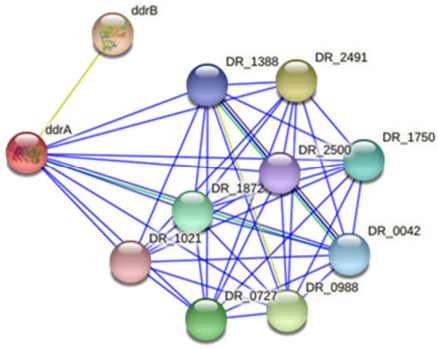

Fig. 4. String network analysis of the hypothetical protein

STRING 10.0 searches was carried out for the identification of possible functional interaction network of the protein. Our target hypothetical protein was identified as ddrA in STRING database. Of them, ddrB is a ssDNA-binding protein protein. The others are one oxidoreductase proteins, one chaperonins, and remaining hypothetical protein (Fig. 4).

Three-dimensional structure analysis

Prediction of 3D structure was done by HHpred server. The server predicted 3D structure of the protein with 99% identity with the highest scoring template (PDB ID: 5JRB_F). 5JRB_F is the crystal structure of the Rad52 protein that is involved in the homologous recombination repair of DNA double-strand breaks. Validation of the predicted three-dimensional model was assessed by PROCHECK through Ramachandran plot analysis (Table 3, Fig. 5&6).

Table 3. Ramachandran plot statistics of the hypothetical protein

Ramachandran plot statistics	No. (%)
Residues in most favoured regions [A,B,L]	134 (88.2)
Residues in additional allowed regions [a,b,l,p]	13 (8.6)
Residues in generously allowed regions [~a,~b,~l,~p]	2 (1.3)
Residues in disallowed regions	3 (2)
Number of non-glycine and non-proline residues	152 (100)
Number of end-residues (excl. Gly and Pro)	1
Number of glycine residues (shown as triangles)	24
Number of proline residues	8
Total number of residues	185

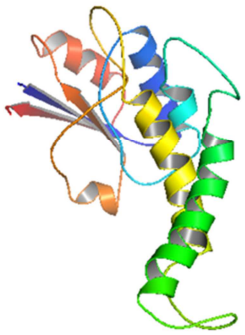

Fig. 5. Predicted three-dimensional structure of the hypothetical protein

Active site of the hypothetical protein

The predicted active site of the protein found that 28 amino acids are involved in active site (Fig. 7). Among 17 pockets, the best active site was found in areas with

613.075 and a volume of 608.774 amino acids which are ssDNA specific.

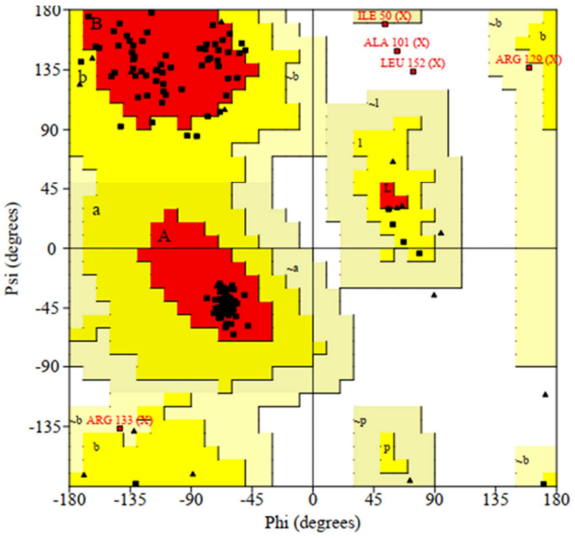

Fig. 6. Ramachandran plot of modelled structure validated by PROCHECK program

Fig. 7. Three dimensional structures of hypothetical protein and active site (Red color)

Conclusion

The identification of protein functions is fundamental for the understanding of biological processes. The identified protein revealed several characteristics such as cytoplasmic nature, DNA binding enzymes containing domain presence and DNA repair activity emphasize the significance of this protein. DdrA has at least two activities: DdrA contributes to genome restitution following irradiation and purified DdrA binds the 3' ends of single-stranded DNA and protects those ends from digestion by exonucleases. So, extended in-vitro research has to be carried out to experimentally validate the possibilities shown here and to find out the proteins' role in biotechnology.

References

Bhasin M, **Garg A**, **Raghava GPS**. 2005. PSLpred: Prediction of subcellular localization of bacterial proteins. Bioinformatics **21**(10), 2522-2524.

Boeckmann B, **Bairoch A**, **Apweiler R**, **Blatter MC**, **Estreicher A**, **Gasteiger E**, **Martin MJ**, **Michoud K**, **O'Donovan C**, **Phan I**, **Pilbout S**. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research **31**(1), 365-370.

Brooks BW, **Murray RGE**, **Johnson JL**, **Stackebrandt E**, **Woese CR**, **Fox GE**. 1980. Red-pigmented micrococci: A basis for taxonomy. International Journal of Systematic and Evolutionary Microbiology **30**(4), 627-646.

Colovos C, **Yeates TO**. 1993. Verification of protein structures: Patterns of nonbonded atomic interactions. Protein Science **2**(9), 1511-1519.

Cox MM, **Battista JR**. 2005. *Deinococcus radiodurans*—The consummate survivor. Nature Reviews Microbiology **3**(11), 882-892.

Dobson L, **Reményi I**, **Tusnády GE**. 2015. CCTOP: A Consensus Constrained TOPology prediction web server. Nucleic Acids Research **43**(W1), W408-W412.

Dulermo R, **Onodera M**, **Porteron M**, **Pasternak C**. 2015. Identification of new genes contributing to the extreme radioresistance of *Deinococcus radiodurans* using a Tn5-based transposon mutant library. PLoS One **10**(4), e0124358.

Dundas J, **Ouyang Z**, **Tseng J**, **Binkowski A**, **Turpaz Y**, **Liang J**. 2006. CASTp: Computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Research **34**(suppl_2), W116-W118.

Eisenberg D, **Lüthy R**, **Bowie JU**. 1997. VERIFY3D: Assessment of protein models with three-dimensional profiles. In Methods in Enzymology (Vol. 277, pp. 396-404). Academic Press.

Gasteiger E, **Gattiker A**, **Hoogland C**, **Ivanyi I**, **Appel RD**, **Bairoch A**. 2003. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Research **31**(13), 3784-3788.

Geourjon C, **Deleage G**. 1995. SOPMA: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Bioinformatics **11**(6), 681-684.

Harris DR, **Tanaka M**, **Saveliev SV**, **Jolivet E**, **Earl AM**, **Cox MM**, **Battista JR**. 2004. Preserving genome integrity: the DdrA protein of *Deinococcus radiodurans* R1. PLoS Biology **2**(10), e304.

Hunter S, **Apweiler R**, **Attwood TK**, **Bairoch A**, **Bateman A**, **Binns D**, **Bork P**, **Das U**, **Daugherty L**, **Duquenne L**, **Finn RD**. 2009. InterPro: The integrative protein signature database. Nucleic Acids Research **37**(suppl_1), D211-D215.

Imai K, **Asakawa N**, **Tsuji T**, **Akazawa F**, **Ino A**, **Sonoyama M**, **Mitaku S**. 2008. SOSUI-GramN: High performance prediction for sub-cellular localization of proteins in gram-negative bacteria. Bioinformation **2**(9), 417.

Johnson M, **Zaretskaya I**, **Raytselis Y**, **Merezhuk Y**, **McGinnis S**, **Madden TL**. 2008. NCBI BLAST: A better web interface. Nucleic Acids Research **36**(suppl_2), W5-W9.

Kanehisa M, **Goto S**, **Kawashima S**, **Nakaya A**. 2002. The KEGG databases at GenomeNet. Nucleic Acids Research **30**(1), 42-46.

Kim DU, **Jang JH**, **Kang MS**, **Kim JY**, **Zhang J**, **Lim S**, **Kim MK**. 2018. *Deinococcus irradiatisoli* sp. nov., isolated from gamma ray-irradiated soil. International Journal of Systematic and Evolutionary Microbiology **68**(10), 3232-3236.

Krieger E, **Joo K**, **Lee J**, **Lee J**, **Raman S**, **Thompson J**, **Tyka M**, **Baker D**, **Karplus K**. 2009. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. Proteins: Structure, Function, and Bioinformatics **77**(S9), 114-122.

Laskowski RA, **MacArthur MW**, **Moss DS**, **Thornton JM**. 1993. PROCHECK: A program to check the stereochemical quality of protein structures. Journal of Applied Crystallography **26**(2), 283-291.

Lin J, **Qi R**, **Aston C**, **Jing J**, **Anantharaman TS**, **Mishra B**, **White O**, **Daly MJ**, **Minton KW**, **Venter JC**, **Schwartz DC**. 1999. Whole-genome shotgun optical mapping of *Deinococcus radiodurans*. Science **285**(5433), 1558-1562.

Lupas A, **Van Dyke M**, **Stock J**. 1991. Predicting coiled coils from protein sequences. Science **252**(5009), 1162-1164.

Madeira F, **Park YM**, **Lee J**, **Buso N**, **Gur T**, **Madhusoodanan N**, **Basutkar P**, **Tivey AR**, **Potter SC**, **Finn RD**, **Lopez R**. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Research **47**(W1), W636-W641.

Marchler-Bauer A, **Anderson JB**, **Cherukuri PF**, **DeWeese-Scott C**, **Geer LY**, **Gwadz M**, **He S**, **Hurwitz DI**, **Jackson JD**, **Ke Z**, **Lanczycki CJ**. 2005. CDD: A Conserved Domain Database for protein classification. Nucleic Acids Research **33**(suppl_1), D192-D196.

McGuffin LJ, **Bryson K**, **Jones DT**. 2000. The PSIPRED protein structure prediction server. Bioinformatics **16**(4), 404-405.

Mohseni M, **Abbaszadeh J**, **Nasrollahi Omran A**. 2014. Radiation resistant of native *Deinococcus* spp. isolated from the Lout desert of Iran "the hottest place on Earth". International Journal of Environmental Science and Technology **11**, 1939-1946.

Möller S, **Croning MD**, **Apweiler R**. 2001. Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics **17**(7), 646-653.

Ott E, **Kawaguchi Y**, **Kölbl D**, **Rabbow E**, **Rettberg P**, **Mora M**, **Moissl-Eichinger C**, **Weckwerth W**, **Yamagishi A**, **Milojevic T**. 2020. Molecular repertoire of *Deinococcus radiodurans* after 1 year of exposure outside the International Space Station within the Tanpopo mission. Microbiome **8**, 1-16.

Szklarczyk D, **Franceschini A**, **Wyder S**, **Forslund K**, **Heller D**, **Huerta-Cepas J**, **Simonovic M**, **Roth A**, **Santos A**, **Tsafou KP**, **Kuhn M**. 2015. STRING v10: Protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research **43**(D1), D447-D452.

Tamura K, **Peterson D**, **Peterson N**, **Stecher G**, **Nei M**, **Kumar S**. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution **28**(10), 2731-2739.

The Institute for Genomic Research. 2004. TIGR Microbial Database.

Tusnady GE, **Simon I**. 2001. The HMMTOP transmembrane topology prediction server. Bioinformatics **17**(9), 849-850.

Wilson D, **Madera M**, **Vogel C**, **Chothia C**, **Gough J**. 2007. The SUPERFAMILY database in 2007: Families and functions. Nucleic Acids Research **35**(suppl_1), D308-D313.

Yu CS, **Chen YC**, **Lu CH**, **Hwang JK**. 2006. Prediction of protein subcellular localization. Proteins: Structure, Function, and Bioinformatics **64**(3), 643-651.

Yu NY, **Wagner JR**, **Laird MR**, **Melli G**, **Rey S**, **Lo R**, **Dao P**, **Sahinalp SC**, **Ester M**, **Foster LJ**, **Brinkman FS**. 2010. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics **26**(13), 1608-1615.

Zhang YQ, **Sun CH**, **Li WJ**, **Yu LY**, **Zhou JQ**, **Zhang YQ**, **Xu LH**, **Jiang CL**. 2007. *Deinococcus yunweiensis* sp. nov., a gamma-and UV-radiation-resistant bacterium from China. International Journal of Systematic and Evolutionary Microbiology **57**(2), 370-375.

Zimmermann L, **Stephens A**, **Nam SZ**, **Rau D**, **Kübler J**, **Lozajic M**, **Gabler F**, **Söding J**, **Lupas AN**, **Alva V**. 2018. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. Journal of Molecular Biology **430**(15), 2237-2243.