Search for DNA repair genes by analyzing the *R. varieornatus* genome via computational gene prediction methods.

Abstract

Tardigrades, water-dwelling metazoans, are renowned for their resilience in extreme conditions. Previous research suggested extensive horizontal gene transfer (HGT) as a potential explanation for their survivability. However, subsequent studies questioned the validity of this hypothesis. This project aimed to identify genes related to tardigrade DNA repair mechanisms using computational methods. Among 16435 proteins initially predicted by AUGUSTUS, 34 unique proteins were selected for further analysis based on proteomic data, with 21 of them predicted to have nucleus localization. BLAST alignment against UniProtKB/Swiss-Prot found homologs for 11 proteins, one of which was a known Damage suppressor protein (Dsup) and the remaining 10 were not homologous to any DNA repair proteins. It has been shown that Dsup exhibits its protecting function through binding to nucleosomes. 9 out of 21 proteins were unknown and may be further tested *in vitro* for their nucleus localisation and DNA repair function.

Introduction

Tardigrades are microscopic, water-dwelling metazoans which are known for their remarkable ability to withstand extreme conditions that would be lethal to most other life forms. They can survive in environments with high levels of radiation, extreme temperatures (ranging from -272°C to 150°C), intense pressure, and even the vacuum of space [1]. Understanding how tardigrades endure such harsh conditions could potentially offer insights into developing technologies or strategies for preserving life in extreme environments.

Studies in the late 20th and early 21st centuries focused on identifying specific mechanisms that contribute to tardigrade resilience. In 2015, the sequencing of the first tardigrade genome, Hypsibius dujardini, revealed an unexpected level of horizontal gene transfer (HGT) from various organisms which could explain its extraordinary survivability, but half a year later another research suggested that the extensive horizontal transfer proposed by Boothby et al. was an artifact of a failure to eliminate contaminants from sequence data before assembly [2,3]. Later researchers discovered unique proteins, such as tardigrade-specific intrinsically disordered proteins, which play a role in surviving desiccation and extreme environments [4]. It was also believed that tardigrades' radiation tolerance was devoted to protectants connected with the desiccated state, but it was further shown that tardigrades rely on efficient mechanisms of DNA repair, the nature of which was unknown at that moment [5].

With full-genome sequencing data, it is possible to predict the regions of the genome that encode proteins. Computational algorithms analyze genomic sequences to recognize patterns indicative of gene structures, such as coding regions (exons) and non-coding regions (introns). These algorithms consider features like open reading frames (ORFs), start and stop codons, and splice sites to predict the locations of genes. After predicting genes in a newly sequenced genome, homology may be used to compare these predicted genes with known genes in other species[6].

The aim of this project was to identify genes which may potentially be responsible for tardigrade efficient DNA repair mechanisms with the use of computational methods.

Materials & methods

The assembled *Ramazzottius varieornatus* genome of the YOKOZUNA-1 strain with GenBank access number GCA_001949185.1 was used [7]. Precomputed AUGUSTUS result with predicted genes from the assembly in gff format was downloaded from [8]. The sequences of the predicted genes were extracted from the original tardigrade's genome using a perl script getAnnoFasta.pl [9]. The launch was performed using perl v5.32.1 built for x86_64-linux-thread-multi.

List of peptides after tandem mass spectrometry was downloaded from [10]. BLAST+ (v2.12.0+) was used to compare proteomics data and predicted genes from the *R.varieornatus* genome [11]. A protein database (-dbtype prot) of the predicted AUGUSTUS genes was created using makeblastdb. It was used to search for proteomics data in it using blastp. The results were saved in the tabular format with default values (-outfmt 6). According to the found unique identifiers of the predicted proteins, the corresponding sequences were extracted, which were used further in the analysis. seqtk subseq was used for all sequence extracts (v1.3-r106) [12].

The physical localization was predicted using the WoLF PSORT (organism type: Animal) [13] and TargetP (organism group: Non-plant) [14]. For further analysis, proteins with the statuses nucl or cyto_nucl for WoLF PSORT or OTHER for TargetP were selected. blastp search against the "UniProtKB/Swiss-Prot" database was provided for these proteins with NCBI BLAST web site [15]. For proteins for which no orthologs were found, function prediction was performed using the HMMER-tool hmmscan [16].

Results & discussions

Initially 16435 proteins were predicted by AUGUSTUS. The integration of proteomic data made it possible to select 34 unique proteins for further analysis.

According to WoLF PSORT, 17 out of 34 proteins were predicted to have nuclear localisation and according to TargetP 21 out of 34 proteins were reported to have no targeting peptide. 21 genes reported by TargetP include all the 17 genes reported by WoLF PSORT except one (Table 1).

After performing BLAST against the UniProtKB/Swiss-Prot database, homologs were found only for 10 out of 21 proteins, but none of these homologs exhibited DNA-repair function (Table 1). 9 out of 21 protein sequences didn't align to anything suggesting that these proteins are unknown and they can be tested for their nuclear localisation *in vitro* via immunohistochemistry analysis as well as for their potential DNA repair function through expressing these proteins in human cell lines and checking if they suppress stress-induced DNA fragmentation. And 1 out of 21 genes aligned with 100% identity to a Damage suppressor protein from *Ramazzottius varieornatus* (UniProtKB/Swiss-Prot: P0DOW4.1), which was discovered by *Hashimoto et al.* [17].

Gene ID	WoLF PSORT	TargetP	BLAST Homologs / hmmscan Pfam family
g10513.t1	nucl: 20, cyto_nucl: 14.5, cyto: 7, extr: 3, E.R.: 1, golg: 1	OTHER	
g10514.t1	nucl: 19, cyto_nucl: 15, cyto: 9, extr: 3, mito: 1	OTHER	
g11513.t1	cyto: 17, cyto_nucl: 12.8333, cyto_mito: 9.83333, nucl: 7.5, E.R.: 3, mito: 1.5, plas: 1, pero: 1, golg: 1	OTHER	Trafficking protein particle complex subunit 9 Q6PA97.1
g11806.t1	nucl: 18, cyto_nucl: 11.8333, mito: 5, extr: 4, cyto: 3.5, cyto_pero: 2.66667, cysk_plas: 1	OTHER	
g11960.t1	nucl: 32	OTHER	E3 ubiquitin-protein ligase BRE1B Q8CJB9.1
g12510.t1	plas: 29, cyto: 3	OTHER	
g13530.t1	extr: 13, nucl: 6.5, lyso: 5, cyto_nucl: 4.5, plas: 3, E.R.: 3, cyto: 1.5	SP	
g14472.t1	nucl: 28, plas: 2, cyto: 1, cysk: 1	OTHER	Damage suppressor protein P0DOW4.1
g15484.t1	nucl: 17.5, cyto_nucl: 15.3333, cyto: 12, cyto_mito: 6.83333, plas: 1, golg: 1	OTHER	Vacuolar protein sorting-associated protein 51 homolog Q155U0.1 Vps51 Vps51/Vps67
g16318.t1	nucl: 20.5, cyto_nucl: 13, extr: 5, cyto: 4.5, E.R.: 1, golg: 1	OTHER	
g16368.t1	nucl: 20.5, cyto_nucl: 13, extr: 5, cyto: 4.5, E.R.: 1, golg: 1	OTHER	
g2203.t1	plas: 29, nucl: 2, golg: 1	OTHER	Myogenesis-regulating glycosidase Q69ZQ1.2
g3428.t1	mito: 18, cyto: 11, extr: 2, nucl: 1	OTHER	Myosin regulatory light chain 12A P19105.2
g4106.t1	E.R.: 14.5, E.R._golg: 9.5, extr: 7, golg: 3.5, lyso: 3, pero: 2, plas: 1, mito: 1	OTHER	
g4970.t1	plas: 32	OTHER	Transmembrane serine protease P98159.2
g5237.t1	plas: 24, mito: 8	OTHER	
g5443.t1	extr: 28, nucl: 3, cyto: 1	OTHER	
g5510.t1	plas: 23, mito: 7, E.R.: 1, golg: 1	OTHER	MARVEL Membrane-associating domain

g5927.t1	nucl: 30.5, cyto_nucl: 16.5, cyto: 1.5	OTHER	Phosphoglucosamine acetylase Q17427.1
g7861.t1	nucl: 16, cyto_nucl: 14, cyto: 8, plas: 5, pero: 1, cysk: 1, golg: 1	OTHER	Matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1 B4F769.1
g8100.t1	nucl: 16.5, cyto_nucl: 12.5, cyto: 7.5, plas: 5, extr: 2, E.R.: 1	OTHER	Inositol monophosphatase 3 Q29JH0.2
g8312.t1	nucl: 15.5, cyto_nucl: 15.5, cyto: 12.5, mito: 2, plas: 1, golg: 1	OTHER	Vacuolar protein sorting-associated protein 41 homolog P49754.3

Table 1. A set of genes, selected on the bases of tandem mass-spectrometry results and their predicted subcellular localization by WoLF PSORT and Target. Homologs to the proteins predicted to nucleus are determined with BLASTp search.

The discovered Damage suppressor protein, also known as Dsup, has shown its reparative properties in experiments on cells [17, 18]. It is suggested that the protein performs its protective function by binding nucleosomes [19]. Thus, the protein protects chromosomal DNA from hydroxyl radicals generated under irradiation conditions.

Thus, the integration of proteomic data and the consistent application of various methods (BLAST, WoLF PSORT, TargetP) made it possible to narrow down the range of candidate proteins responsible for the tolerance of high doses of ionizing radiation. The 10 proteins found in this study can be used in experiments to establish their role. Modern computational methods enable to simplify and solve wide range of biological problems. Among them, the task of functional annotation of newly discovered proteins often arises. Over the past decade, machine learning methods have shown good results in predicting protein function [20]. Despite the development of such computer methods, the interpretation of the results is still left to man.

References

1. Goldstein, B. Tardigrades. Nat Methods 19, 904–905 (2022).
2. Boothby, T. C. et al. Proc. Natl Acad. Sci. USA 112, 15976–15981 (2015).
3. Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, Maroon H, Thomas F, Aboobaker AA, Blaxter M. No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. Proc Natl Acad Sci U S A. 2016 May 3;113(18):5053-8.
4. Boothby TC, Tapia H, Brozena AH, Piszkiewicz S, Smith AE, Giovannini I, Rebecchi L, Pielak GJ, Koshland D, Goldstein B. Tardigrades Use Intrinsically Disordered Proteins to Survive Desiccation. Mol Cell. 2017 Mar 16;65(6):975-984.e5.
5. Jönsson KI, Harms-Ringdahl M, Torudd J. Radiation tolerance in the eutardigrade Richtersius coronifer. Int J Radiat Biol. 2005 Sep;81(9):649-56.
6. Terence A Brown.Genomes, 2nd edition. Oxford: Wiley-Liss; 2002 .ISBN-10: 0-471-25046-5

7. ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/949/185/GCA_001949185.1_Rvar_4.0/GCA_001949185.1_Rvar_4.0_genomic.fna.gz
8. https://drive.google.com/file/d/1hCEywBlqNzTrIpQsZTVuZk1S9qKzqQAq/view?usp=sharing
9. http://augustus.gobics.de/binaries/scripts/getAnnoFasta.pl
10. https://disk.yandex.ru/d/xJqQMGX77Xueqg
11. Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L. (2008) “BLAST+: architecture and applications.” BMC Bioinformatics 10:421.
12. https://github.com/lh3/seqtk
13. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W585-7. doi: 10.1093/nar/gkm259. Epub 2007 May 21. PMID: 17517783; PMCID: PMC1933216.
14. Detecting Sequence Signals in Targeting Peptides Using Deep Learning José Juan Almagro Armenteros, Marco Salvatore, Ole Winther, Olof Emanuelsson, Gunnar von Heijne, Arne Elofsson, and Henrik Nielsen Life Science Alliance 2 (5), e201900429. doi:10.26508/lsa.201900429
15. https://blast.ncbi.nlm.nih.gov/
16. S. R. Eddy. Accelerated profile HMM searches. PLOS Comp. Biol., 7:e1002195, 2011
17. Hashimoto, T., Horikawa, D., Saito, Y. et al. Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein. Nat Commun 7, 12808 (2016). https://doi.org/10.1038/ncomms12808
18. Hashimoto T, Kunieda T. DNA Protection Protein, a Novel Mechanism of Radiation Tolerance: Lessons from Tardigrades. Life (Basel). 2017 Jun 15;7(2):26. doi: 10.3390/life7020026.
19. Chavez C, Cruz-Becerra G, Fei J, Kassavetis GA, Kadonaga JT. The tardigrade damage suppressor protein binds to nucleosomes and protects DNA from hydroxyl radicals. Elife. 2019 Oct 1;8:e47682. doi: 10.7554/eLife.47682.
20. Gligorijević, V., Renfrew, P.D., Kosciolek, T. et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun 12, 3168 (2021). https://doi.org/10.1038/s41467-021-23303-9