OXFORD GENETICS

Tardigrades: from genestealers to space marines

Sukhanova Xenia[1,*], Kolpakova Oxana[2,*] and Kupriyanov Semyon[3,*]

*Saint-Petersburg, Bioinformatics Institute

[1]sukhanovaxenia@gmail.com, sukhanova@scamt-itmo, Russia, Saint Petersburg, ITMO University, SCAMT Institute

Abstract

Tardigrades are microscopic animals capable of withstanding some of the most severe environmental conditions. These organisms can survive freezing, total dehydration, pressure and radiation. All these extreme conditions are accompanied by DNA damage, which entails a decrease in viability. In this work, we studied the proteins present in the nucleus of *Ramazzottius varieornatus* in order to understand whether there are specific mechanisms for protecting DNA from damage. After isolation of the nuclear soluble fraction, the composition of the proteins was investigated using mass spectrometry. We established which proteins have nuclear localization, compared the structures of these sequences with a database, and also predicted the presence of functional motifs in these proteins. As a result, we selected two potential candidates: Myosin regulatory light chain and Vacuolar protein sorting-associated protein 51 homolog. According to research data, both proteins can be involved in DNA repair mechanisms. And the left unrecognised four could also participate in DNA-reparation process.

Keywords: Genome; Assembly; Tardigrades; radioresistance, DNA-reparation

Introduction

Tardigrades form a monophyletic group of microscopic ecdysozoans best known for surviving extreme environmental conditions, including radiation. Precise gene repertoire analyses reveal the presence of a small proportion (1.2% or less) of putative foreign genes, loss of gene pathways that promote stress damage, expansion of gene families related to ameliorating damage, and evolution and high expression of novel tardigrade-unique proteins. Minor changes in the gene expression profiles during dehydration and rehydration suggest constitutive expression of tolerance-related genes. The Dsup DNA-associated protein that protects DNA and improves radio tolerance, was found out (Takuma *et al.* 2016), (Kirke *et al.* 2020).

Recent studies show that stress-related tardigrade genes may be transfected to other organism cells and provide increased tolerance to osmotic stress and ionizing radiation (Jonsson 2019), (Kirke *et al.* 2020). The genes need to be predicted and protein coding regions after genome sequencing. Both tasks are solved by alignment to known homologous sequences of related organisms.

In this work, we will try to find the genome features that ensure the survival of tardigrades. We suggest that tardigrades may have unique proteins bound to their DNA to protect and / or effectively repair radiation damage. To do this, we analyzed the data of the NGS and investigated the proteins associated with DNA obtained using tandem mass spectrometry and identified the function of these proteins.

Materials and methods

For more detailed description of the source code, please, look at Supplementary file (HW4_Lab_journal).

Data download and repeats masking

The data was obtained from NCBI database (ftp-server) by the link to the strain's genome YOKOZUNA strain.

For this bash command line interface was used and the **wget** command was launched.

The following procedure was necessary for repeats masking as eukaryotic genomes are enriched with repeats. For this the RepeatMasker tool was installed and following programs were launched in command line:

1. BuildDatabase: -name the name of database, output .fasta file of genome which database will be based on;
2. RepeatModeler: -database for database creation based on the input genome .fasta file, -engine to set the alignment algorithm to use, -pa the number of processes to use;
3. RepeatClassifier: -consensi file with previously found consensuses, -stockhom file of predicted families;
4. RepeatMasker: -lib fasta file of identified families based on the database, the positional argument - fasta file of the interest.

Then gene prediction was launched via website version of AUGUSTUS tool and provided the targeted file in .gff format.

Functional annotation

As the following analysis required .fasta files the convertation of augustus output to .fasta file was provided by perl script getAnnoFasta.pl.

After that the quantity of predicted genes was estimated by cmd command **grep** and **wc** as a pipe:

- grep: -P to set coding language format, '■' to search all strings starting from «>» symbol;

- wc: -l to count number of strings .

Physical localization

For this step peptides sequences obtained by tandem mass spectrometry of the chromatin fraction by the link. The peptide-base filtration was proceeded by two ways:

1. created Python script to align the sequence of peptides from the mass spectrometer to the reference;
2. local BLASTp 2.12.0-based alignment.

Python3 method

The concept of this pipeline is string-search method and fasta files parsing with BioPython library, SeqIO package (Cock *et al.* 2009). Several packages are required for the script launch:

1. argparser, exactly ArgumentParser for arguments provision;
2. BioPython, exactly SeqIO for .fasta file parsing;
3. os;
4. sys.

The pipeline provides the search of unique potential genes which include one or more peptide sequences.The script is presented as in .py format, as in .ipynb format (to see follow the link).

It takes two inputs:

- -f, .fasta file of hypothesized proteins;
- -p, .fasta file of peptides.

And return next outputs:

- local_annot.txt file, tab-delimited file including ID of hypothesized protein and ID number of peptide, which showed overlapping;
- prot_local.fa file, including potential candidates from hypothesized proteins which do interact with DNA.

BLASTp-based method

This method is based on aligning peptides on targeted sequences with set threshold of e-value to trim low rate overlaps. It is launched from command line with following programs:

- makeblastdb: -in fasta file of targeted sequences, -dbtype type of database, -out output;
- blastp: -query fasta file of what we align, -db file of database, -out output format, -outfmt included columns, -evalue threshold of e-value, -task type of analysis;
- cut: -s strict column division, -f number of column to show;
- sort: -u leave only unique elemetns;
- samtools faidx: indexation of database and filtration;
- xargs: to pipe list of unique gene id to samtools for the filtration of default file.

Localization prediction

For in-cell localization of potential candidates webtools WoLF PSORT (Paul *et al.* 2007) and TargetP (Almagro Armenteros *et al.* 2019) were used. Both of theme were necessary as TargetP could not to classify nuclear-localising proteins, while WoLF provided an opportunity to identify more patterns of localisation.

These methods are based on the following concepts:

- WoLF PSORT is used to predict localization based on the presence of the signal N-terminal extension;
- TargetP was also used to predict localization. The location assignment is based on the predicted presence of any of the N-terminal sequences: chloroplast transit peptide (cTP), mitochondrial targeting peptide (mTP) or secretory pathway signal peptide (SP).

Functional prediction based on BLAST search, homology-based approach.

To improve prediction and annotation BLASTp alignment (**?**) was launched against UniProtKB/Swiss-Prot database (UniProt Consortium 2018) with targeted organism exclusion. This method improved the prediction via multiple-alignment approach. Alignment against annotated sequences allowed to suggest possible functions of targeted sequences.

As in protein alignment variance is broader the significant level and thresholds are lower comparing with those for nucleotide sequences alignments. For key parameters - Query Coverage and Identity - thresholds > 67 and 47, consequently.

Functional prediction based on domain structure, Pfam-based method.

Finally, domain structure prediction was used to increase the accuracy of our results. For this HMMER (Finn *et al.* 2011) webtool was used. Based on hmm-algorithm, this method allows to identify possible function of sequence not only by multiple-alignment approaches but also considering the probability of each amino acid in the sequence inclusion into specific domains and motifs. The latter are searched in Pfam (Mistry *et al.* 2020) database.

Results and discussion

The only functional prediction of proteins still leave candidates' identification complex

Masking for repeats by RepeatMasker and functionally prediction by AUGUSTUS of genes allowed to identify 16435 potential genes. However, to do any predictions of mechanisms about how the DNA reparation in extreme environments works in Tardigrades we still ought to narrow the list of proteins. For this we highlighted several options:

1. **Localisation prediction based on physical properties and interaction predictions**.

   As we suggest that proteins of interest should somehow interact with DNA and provide the reparation directly, such a filtration will be necessary as the first step;

2. **In-cell localization**.

   let it be we found only DNA-interacting proteins but how to make sure that they indeed persist in the nucleus and function mostly in the nucleoplasm? To provide such a search a filtration on localisation should be provided: by sequence markers (localising signals);

3. **Domain-comparison**.

   Obviously, regulating and DNA-interacting proteins (e.g. transcription factors) possess specific marker sequences (common domain), which are responsible for DNA-interaction properties: Zinc-fingers, etc. For this several methods can be used, but the most approved one is to predict domains using exact databases (Pfam database, etc.) and alignment-based, structure-based, ab initio tools (e.g. hmmscan).

4. **BLASTp alignment with exclusion of investigating organism**.

   BLASTp can be included to the previously mentioned alignment-based tools of protein domain prediction, however here it allows to directly obtained the full annotation on the protein level (for instance, hmmscan only annotate on the domain level).

Localization filtration helped to reveal significant genes of DNA-interacting proteins

Basically, participants of DNA reparation system should have high affinity to DNA. Thus, tending to be extracted with DNA together. Nowadays, sufficient number of methods are known for DNA-associated protein extraction, from which tandem mass spectrometry is preferable. In our research such an experiment was provided to obtain sequences of peptides' which were definitely associated with DNA.

The received list of peptides was then used for the filtration of hypothesized proteins to narrow the following search inside the «group» of DNA-interaction ones. After filtration only 19 from 16435 and 20 from 23007 genes were left and taken to the following analysis.

Organelle-specific filtration is tremendous for the investigation of DNA-repair process

Even after localization filtration the quantity of predicted proteins made the search complex as such it did not take into account where the candidate participated in processes exactly inside the cell. In the sense of DNA-reparation we could admit that all candidates should localize inside the nucleus. As the significant number of potential genes showed multiple localization patterns (nucleus, cytoplasm, endoplasmic reticulum, Golgi apparatus, secretory granules, mitochondria, etc.), we ought to specify theme with the help of localization-predictors WoLF (Paul *et al.* 2007) and TargetP (Almagro Armenteros *et al.* 2019). Searching for specific localization signals in sequences they estimate the probability of protein to localize in consequent pattern.

Notably, in the comparison of WoLF TargetP seemed to be less informative as it did not predicted any others except mitochondrial, chloroplast or secretory proteins. Hence, both tools were necessary for the analysis. Overall from previously extracted 18 (in the case of augustus dataset) and 20 (from NCBI dataset) only 6 included nuclear localizing signal (NLS) (see Table 1 and detailed one in Supplementary materials), from which g10513.t1, g10514.t1, g13530.t1 and 14472.t1 showed higher rates of NLS prediction.

Two out of six candidates were functionaly annotated

To functional annotation two methods were established: BLASTp homologous-based search and Pfam-prediction.

BLAST search for homologous protein sequences from the UniProtKB / Swiss-Prot database with excluded investigating organism - *Ramazzottius varieornatus* - succeeded in identifying two homologous (Table 2) with enough significance rate. With identity and query coverage lying from 47.5 to 58.44 and from 65 to 79, consequently, **Myosin regulatory light chain** was identified. And for the second homolog - **Vacuolar protein sorting-associated protein 51** - identity and query coverage lied from 43.5-47 and 77-78.

Achieved BLASTp results coincides with those from HMMSCAN prediction. We launched the search of probability of sequences to form specific domains from Pfam database supported by alignments. From the analysis we could admit that, indeed, recognized by BLASTp proteins possess specific domains. Moreover, we also recognised that definitely identified domains exist

Table 1 Nuclear localization proteins obtained from the combined search results TargetP and WoLF PSORT. The table illustrates the results of launched TargetP prediction and highlighted with yellow overlappings with WoLF results. The existence of the same potential proteins in two predicted datasets increases the power of the prediction method. According to the analysis indeed 6 of 19 hypothesized proteins can be chosen as the potential ones.

Sequence ID	TargetP	WoLF
g3428.t1_10	OTHER	mito: 18, cyto: 11, extr: 2, nucl: 1
g10513.t1_9	OTHER	nucl: 20, cyto$_n$*ucl* : 14.5, *cyto* : 7, *extr* : 3, *E.R.* : 1, *golg* : 1
g10514.t1_37	OTHER	nucl: 19, cyto$_n$*ucl* : 15, *cyto* : 9, *extr* : 3, *mito* : 1
g13530.t1_5	SP	extr: 13, nucl: 6.5, lyso: 5, cyto$_n$*ucl* : 4.5, *plas* : 3, *E.R.* : 3, *cyto* : 1.5
g14472.t1_7	OTHER	nucl: 28, plas: 2, cyto: 1, cysk: 1
g15484.t1_32	OTHER	nucl: 17.5, cyto$_n$*ucl* : 15.3333, *cyto* : 12, *cyto*$_m$*ito* : 6.83333, *plas* : 1, *golg* : 1

Table 2 BLAST search showed orthologues only for two potential proteins - g3428, g15484. The table shows the average values for the most significant matches.

Sequence ID	Identity	E-value	Query coverage
g3428.t1$_1$0	53,07	3.4e-45	72,37
g15484.t1$_3$2	45,15	2.86e-151	77,14

in found homologous (see Table).

We revealed **EF-hands** (1 and 6) for the g15484.t1. The same motif is present in the Myosin regulatory light chain, which coincides with the BLAST prediction. For the sequence g15484.t1 domains **Vps51**, **COG2**, **Vps54_N**, **Dor1**, **Sec5** were recognized, which is also consistent with BLAST's prediction of sequence similarity to Vacuolar protein sorting-associated protein 51 homologous (see Table).

To conclude, from the set of predicted genes annotated ones are hardly ever take part in DNA-repair mechanisms. The obtained data for selected proteins could be found in Table 4.

As we can see from two identified proteins g15484.t1 is more likely to be a participant of DNA-reparation in the case of targeted organism according to the metrics of nuclear-localisation rate (WoLF score of nuclear localisation signal, 17.5). However, the exact functional role ought to be examined deeper. If for g3428 there are less hesitates on it's homology to Myosin regulatory light chain (as the e-value is extremely low), for g15484 results are not so obvious and future experiment and analysis are required.

Despite the failure to annotate left four candidates (g10513.t1, g10514.t1, g13530.t1 and 14472.t1) neither by BLASTp, nor by domain-prediction the results, there is still a great evidence of their participation in DNA-reparation system. For further investigation additional molecular and bioinformatic analysis ought to be applied to examine the role of mainly unrecognised candidates.

Discussion

Radiation causes double-stranded breaks and leads to genome instability. The several mechanisms of DNA double break responses and many proteins involved in this process were discribed (Waterman *et al.* 2020). Recent genome analysis has shown

Table 3 Pfam results coincides with BLASTp analysis. For two potential proteins the listed domain structure was identified. EF-hands were predicted for the g3428.t1 | 10 sequence, which appeared to have a high homology to the myosin regulatory light chain. For g15484.t1 | 32 sequence (potential vacuolar protein sorting-associated protein 51 homolog) domains Vps51, COG2, Vps54$_N$, *Dor*1, *Sec*5*wererecognized*.

Sequence ID	Domain ID	Accession	Description	Independent E-value	Conditional E-value
g15484.t1	Vps51	PF08700.13	Vps51/Vps67	3.0e-24	7.9e-28
g15484.t1	COG2	PF06148.13	COG (conserved oligomeric Golgi) complex component, COG2	2.4e-06	6.4e-10
g15484.t1	Vps54$_N$	PF10475.11	Vacuolar-sorting protein 54, of GARP complex	2.3e-10	6.0e-14
g15484.t1	Dor1	PF04124.14	Dor1-like family	1.2e-11	3.1e-15
g15484.t1	Sec5	PF15469.8	Exocyst complex component Sec5	4.0e-26	1.1e-29
g3428.t1	EF-hand$_1$	PF00036.34	EF hand	3.0e-06	4.8e-10
g3428.t1	EF-hand$_6$	PF13405.8	EF-hand domain	5.1e-06	7.9e-10

Table 4 Integration of results. This table shows best blast hit (annotation and e-value), predicted Pfam domains, probable localization(s) according to WoLF PSORT and localization according to TargetP for our selected proteins.

Method	**g3428.t**	**g15484.t1**
Wolf PSORT	mito: 18, cyto: 11, extr: 2, nucl: 1	nucl: 17.5, $cyto_nucl$: 15.3333, $cyto$: 12, $cyto_mito$: 6.83333, $plas$: 1, $golg$: 1
TargetP	OTHER	OTHER
BLAST	Myosin regulatory light chain, E-value 9e-65	Vacuolar protein sorting-associated protein 51 homolog (Vps51), E-value 0.0
Pfam	EF-hands	Vps51, COG2, $Vps54_N$, $Dor1$, $Sec5$

that Tardigrades also have excess copies of antioxidants and DNA repair enzymes, in the absence of enzymes that produce ROS. Also were found out a novel protein unique to tardigrades, Dsup. This suggests that Tardigrades may have evolved their own tolerance mechanisms. So, Dsup physically protects DNA from ROS and radiation and / or local detoxification of ROS.

We found out 2 proteins: myosin-like, g3428.t1 | 10 and vacuolar proteins associated with protein sorting, g15484.t1 | . The EF-hand and Vps51, COG2, Vps54_N, Dor1, Sec5 domains were revealed for g3428.t1 and g15484.t1 respectively.

Myosins have been shown to act as molecular transporters and anchors, which depend on their ability to bind to actin and ATPase. Their role in the response to DNA damage can range from transcriptional response, chromatin movement, and stabilization of chromosome contacts (Cook and Toseland 2020). Also was shown that EF-hand is one of the most widely distributed domains in eukaryotes, perhaps reflecting the range and subtlety of calcium signaling. The downstream regulation element antagonist modulator (DREAM) upon binding calcium dissociates from a DNA-binding regulatory element that otherwise functions as a gene silencer (Cal 2013).

The VTF binding complex, GARP, is a conserved eukaryotic docking complex that is involved in the recycling of proteins from endosomes to the Golgi. Cog is the eight-element conserved oligomeric Golgi complex, which is involved in retrograde vesicular transport and is required to maintain normal Golgi structure and function. Dor1 is involved in vesicle targeting to the Golgi apparatus and complexes with a number of other trafficking proteins, which include Sec34 and Sec35. All revealed domains of the second protein are involved in the recirculation of proteins from endosomes to the Golgi apparatus (Siniossoglou and Pelham 2002), (Pierre *et al.* 2005). This may be important for the removal of the destroyed proteins by radiation, as well as in the initiation of the adaptation reaction, when the exit from the arrest of the cell cycle occurs after some time, after its activation (Dotiwala *et al.* 2013). Thus, the proteins we discovered provide transcriptional response, chromatin movement, and stabilization of chromosomal contacts; and also initiate the adaptation response, exit from the stop of the cell cycle.

It will be good to compare the structure of the proteins we have identified and those which were not annotated with Dsup. Dsup can bind specifically to nucleosomes, creating a highly stable structure and nonspecific DNA. It is possible that the proteins identified by us will exhibit similar properties.

Acknowledgments

The research was supported by Bioinformatics Institute, Mike Raiko and Yuriy Barbitoff.

Literature cited

2013. Calcium-modulated proteins (ef-hand), In: Lennarz WJ, Lane MD, editors, *Encyclopedia of Biological Chemistry (Second Edition)*.

Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, Nielsen H. 2019. Detecting sequence signals in targeting peptides using deep learning. Life Science Alliance. 2.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B *et al.* 2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics. 25.

Cook AW, Toseland CP. 2020. The roles of nuclear myosin in the dna damage response. The Journal of Biochemistry. 169:265–271.

Dotiwala F, Eapen VV, Harrison JC, Arbel-Eden A, Ranade V, Yoshida S, Haber JE. 2013. Dna damage checkpoint triggers autophagy to regulate the initiation of anaphase. 110.

Finn RD, Clements J, Eddy SR. 2011. Hmmer web server: interactive sequence similarity searching. Nucleic Acids Research. 39:W29–W37.

Jonsson KI. 2019. Radiation tolerance in tardigrades: Current knowledge and potential applications in medicine. Cancers. 11.

Kirke J, Jin XL, Zhang XH. 2020. Expression of a tardigrade dsup gene enhances genome protection in plants. Molecular Biotechnology. 62.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar G, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ *et al.* 2020. Pfam: The protein families database in 2021. Nucleic Acids Research. 49.

Paul H, Keun-Joon P, Takeshi O, Fujita Naoya aHH, J. ACC, Kenta N. 2007. Wolf psort: protein localization predictor. Nucleic acids research. 35.

Pierre F, Yulia K, Oleksandra P, B. TA, Lupashin VV. 2005. Cog1p plays a central role in the organization of the yeast conserved

oligomeric golgi complex *. Journal of Biological Chemistry. 280.

Siniossoglou S, Pelham HR. 2002. Vps51p links the vft complex to the snare tlg1p*. Journal of Biological Chemistry. 277.

Takuma H, D. HD, Yuki S, Hirokazu K, Hiroko KH, Tadasu SI, Yohei M, Kazuko O, Ayuko M, Tomoyuki A *et al*. 2016. Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein. Nature Communications. 7.

UniProt Consortium T. 2018. Uniprot: the universal protein knowledgebase. Nucleic Acids Research. 46.

Waterman DP, Haber JE, Smolka MB. 2020. Checkpoint responses to dna double-strand breaks. Annual Review of Biochemistry. 89:103–133.