Gene prediction for mystery Tardigrades genomes

Ignat Sonets, Kamilla Faizullina

In this report we try to find new proteins somehow related to DNA repair mechanisms of tardigrades. As a model Ramazzottius varieornatus, the YOKOZUNA-1 strain was selected, and genome annotation data and mass spectrometry results were obtained. After analysis made, we found at least 2 candidate proteins which have nuclear localization, DNA-binding activity,as also they participate in recombination and mechanisms. Our results can partially explain the survivability of tardigrades in various harsh conditions.

1 Introduction

Greetings! Today we present you our results about searching DNA repair mechanisms in tardigrades. But before we start, let's do a quick walkthrough about Tardigrades itself and methods used in this project. Tardigrades, also known as as water bears or moss piglets, are a phylum of eight-legged segmented micro-animals. They were first described by the German zoologist Johann August Ephraim Goeze in 1773, who called them little water bears. In 1777, the Italian biologist Lazzaro Spallanzani named them Tardigrada, which means "slow steppers"[1]. They have 2 distinct traits:

1. They are almost indestructible. They could have been found everywhere, literally everywhere: in Earth's biosphere, from mountaintops to the deep sea and mud volcanoes, and from tropical rainforests to the Antarctic.[1] Tardigrades are among the most resilient animals ever known to man. They have anomalous endurance in terms of survivability of extreme conditions, and even deep dead space filled with vacuum and radiation emitted from myriads of stars could nothing to do. They are truly space marines coming from the Earth. They exist ca. 500 million years and I guess will exist for nest 500 million years. They are living grey goo in size of approx. 0.5 mm, thousands of little Thanoses. But what makes them so adapted to everything? First, they developed anhydrobiosis, cryobiosis, osmobiosis, or anoxybiosis, and so on. Depending on the environment, they simply go to sleep forming capsule around ther bodies,thus preserving resources and waiting for shiny weather. Second (and it would be our task) they seem to have some mechanisms providing DNA repair to escape the death and

different pathologies. They m u s t definitely have a galore of these methods. Let's find them! But before we start, the second trait:

2. They are extremely cute.

So, what about the data we have? We will be using the sequence of the Ramazzottius varieornatus, the YOKOZUNA-1 strain, its annotation provided by our team, and mass spectrometry data of chromatin of this water bear. The core idea in this project is that tardigrades might have unique proteins associated with their DNA to protect and/or effectively repair it. To put in simple, we will merge proteome data and genome data, and try to find any intersections, determine physical localization of found proteins, and try to explain and discuss our findings.

Also we need to briefly mention different tools we used. AUGUSTUS tool is a gene prediction tool running a generalized hidden Markov model (GHMM), which defines probability distributions for the various sections of genomic sequences. Introns, exons, intergenic regions, etc. correspond to states in the model and each state is thought to create DNA sequences with certain pre-defined emission probabilities. Similar to other HMM-based gene finders, AUGUSTUS finds an optimal parse of a given genomic sequence, i.e. a segmentation of the sequences into states that is most likely according to the underlying statistical model.

BLAST is a gold standard alignment tool that searches for homology (i.e. similarity between sequences.For more info, please view [4].) between query sequence and desired database containing pre-analysed data. BLAST identifies homologous sequences using a heuristic method which initially finds short matches between two sequences; thus, the method does not take the entire sequence space into account. After initial match, BLAST attempts to start local alignments from these initial matches. This also means that BLAST does not guarantee the optimal alignment, thus some sequence hits may be missed. In order to find optimal alignments, the Smith-Waterman algorithm should be used [2, 3].

Brief introduction about other used tools are given in Task4 document.

2 Data

We use the assembled genome data of the Ramazzottius varieornatus, the YOKOZUNA-1 strain [5].

3 Methods

We use the implemented program AUGUSTUS for gene prediction [6]. This program is specified for eukaryotic sequencing data. Usually the number of obtained proteins is

too large.

In order to detect regions, linked to DNA repair, we should use both genomic and proteomic data. We can use a list of peptides that were associated with the DNA regions. This list can be obtained via analyzis of extracted chromatin fraction using tandem mass spectrometry. To find associated proteins from the R. varieornatus genome to peptides, we use makeblastdb utilite [7].

Net, to narrow the list of proteins, we use WoLF PSORT [8]. This program allows to predict where these proteins are found in the cell based on their sequences. predicts the subcellular localization of proteins. The method is based on detection the presence of a signal peptide on their N-terminus. We also use TargetP 1.1 Server [9], which also predicts the subcellular localization.

To find homologous proteins we use Blast [2]. We use HMMER to predict the function of the proteins [10]. The HMMER is based on Hidden Markov Models.

4 Results

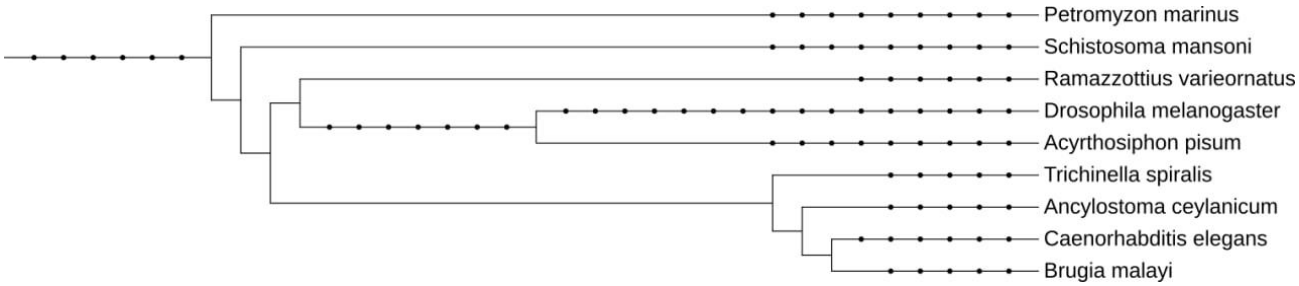

Figure 1: Fylogenetic tree.

First, our goal is to make functional annotation. So, we run AUGUSTUS to find homologous proteins in tardigrade genome data. We used training sets Acyrthosiphon pisum species via augustus. We have chosen this species after fylogenetic tree construction 1. We obtained 14446 proteins.

Next, we used the list of peptides and command *makeblastdb*. After removing duplicates, we got 31 proteins. After using Blast alignment, we keeped 22 proteins. The summury information is presented in Table 1.

id	cover	E-value	Identity	WolF	TargetP	HMMER
g5214t1	32	1e-12	39.29	*nucl: 4.5, cytonucl: 5.5	Other	Chitin binding Peritrophin-A domain
g11539t1	50	3e-06	28.57		Signal peptide	

g5192t1	40	2e-14	40		Signal peptide	Chitin binding Peritrophin-A domain
g5086t1	12	2e-6	44		Signal peptide	Chitin binding Peritrophin-A domain
g5087t1	20	6e-13	40	nucl: 9	Other	Chitin binding Peritrophin-A domain
g672t1	28	1e-5	42		Signal peptide	Chitin binding Peritrophin-A domain
g7315t1	96	8e-70	38	nucl: 17.5, cytonucl: 15.5	Other	SNF2 family N-terminal domain
g2692t1	12	1e-3	26	*cytonucl: 4, nucl: 3.5	Other	Hermes transposase DNA-binding domain
g4653t1	54	1e-13	36		Other	Zinc finger
g8025t1	48	8e-104	38		Other	Cytosol amino peptidase family, catalytic domain
g3168t1	97	8e-50	48		Other	
g10626t1	72	8e-82	27		Other	Transport protein Trs120 or TRAPPC9, TRAPP II complex subunit

g1221t1	13	1e-20	43		Signal peptide	Casein kinase substrate phosphoprotein PP28
g10444t1	9	3e-3	27		Signal peptide	
g7708t1	93	2e-35	22	nucl: 15.5, cytonucl: 15.5	Other	Region in Clathrin and VPS
g5481	13	1e-11	35	nucl: 27, cytonucl: 18.3333	Other	
g7708t1	93	2e-35	22	nucl: 15.5, cytonucl: 15.5	Other	Region in Clathrin and VPS
g7784t1	83	2e-90	36	nucl: 9.5 cytonucl: 6	Signal peptide	Glycosyl transferase family 2
g11028t1	98	2e-82	26	nucl: 32,	Other	Zinc finger, C3HC4 type (RING finger)
g3380t1	51	4e-8	24	*cytonucl: 1.83333, nucl: 1.5	Signal peptide	Astacin (Peptidase family M12A)
g9785t1	29	1e-5	33	nucl: 16	Other	
g2090t1	73	2e-126	36	*nucl: 2	Other	Glycosyl hydrolases family 31

Table 1: Summary information. The star in cell WolF (localization) means that nucl and cytonucl are not maximum. The empty cell means that information is not presented.

References

[1] https://en.wikipedia.org/wiki/Tardigrade

[2] McGinnis, S. and Madden, T. L. (2004). BLAST: at the core ofa powerful and diverse set of sequence analysis tools.Nucleic Acids Res, 32(Web Serverissue):W20–W25.

[3] Smith, T. F. and Waterman, M. S. (1981). Identification of commonmolecular subsequences.J Mol Biol, 147(1):195–197

[4] https://en.wikipedia.org/wiki/Sequence_homology

[5] https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=947166

[6] http://bioinf.uni-greifswald.de/augustus/

[7] BLAST® Command Line Applications User Manual [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-. Building a BLAST database with your (local) sequences. Available from: https://www.ncbi.nlm.nih.gov/books/NBK279688/

[8] https://wolfpsort.hgc.jp/

[9] 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

[10] S.C. Potter, A. Luciani, S.R. Eddy Y. Park, R. Lopez and R.D. Finn, Nucleic Acids Research (2018) Web Server Issue 46:W200-W204.