Published in the Slovak Republic
European Journal of Molecular Biotechnology
Has been issued since 2013.
E-ISSN: 2409-1332
2020, 8(1): 35-41

DOI: 10.13187/ejmb.2020.1.35
www.ejournal8.com

Development the Algorithm for Virtual Screening of Protein Polymorphisms Affecting Their Structural and Functional Properties

Pavel A. Krylov [a], Elena O. Gerasimova [a], Yulia A. Shatyr [a], Alexander B. Mulik [b], Valery V. Novochadov [a, *]

[a] Volgograd State University, Russian Federation
[b] Federal State-Financed Institution Golikov Research Clinical Center of Toxicology under the Federal Medical Biological Agency, Russian Federation

Abstract

The algorithm has been developed for virtual screening of protein polymorphisms that perform various biological functions, primarily related to the regulation of mechanisms at various levels of organization: systemic, cellular, and molecular. The algorithm was developed using the Python v. 3.6.5 programming language in the PyCharm environment. The algorithm includes an automatic search for articles and displays a list that mentions a protein or polymorphism involved in a particular process for specified keywords in PubMed Central. The identified protein and its information are then used to work with the dbSNP database. The algorithm allows sorting polymorphisms according to the following key criteria: whether this polymorphism has been studied or not, whether there is information about its clinical manifestations, the localization of the polymorphism, and the frequency occurrence. The algorithm was tested on the analysis of polymorphisms of the ACAN gene, previously found by automatic search, by keywords characteristic of the functional features of this gene.

As a result, the developed algorithm for virtual screening allowed us to identify polymorphisms of the ACAN gene, according to the stated selection criteria: the total number of polymorphisms (130), their localization in structural and functional domains and in coding (85) or non-coding (52) regions of the gene, as well as the presence of clinical manifestations (136). In the future, the optimization of the algorithm will allow us to obtain more detailed information about certain protein polymorphisms, which will allow us to solve problems in biomedicine, pharmaceuticals, and other biotechnological industries.

Keywords: virtual screening, polymorphisms, data bases, PMC, SNP, chondrocytes, ACAN.

1. Introduction

At present, heightened interest is directed to the study of the effect of gene polymorphisms on the structural and functional properties of the protein, and subsequently to the identification of their links with the development of diseases. Study of the etiology of such diseases as osteoarthritis (Mori et al., 2014), osteoarthritis (Mikhaylenko et al., 2020), various diseases of the cardiovascular system (Gorący et al., 2020), oncology (Krajewski et al., 2020), psychological diseases (Tretiakov et al., 2020) led to the need to search for and identify gene polymorphisms, as this would make it possible to find out the causes of the development of diseases, and subsequently to select the means for their treatment.

* Corresponding author
E-mail addresses: novochadov.valeriy@volsu.ru (V.V. Novochadov)

The search and analysis of polymorphisms can be divided into two approaches. The first approach involves the experimental production of mDNA or mRNA, their sequencing and assembly of the genome (genes), followed by decoding and entering the results into the dbSNP database (Mori et al., 2014). The second approach is bioinformatic, which is an automated virtual screening of SNP databases and the identification of those polymorphisms that have not been studied or could be of interest for study for solving problems in various fields of science. In this regard, we chose the second approach, since there is a lot of data on polymorphisms and their number is constantly increasing, and manual analysis requires a lot of time. The lack of programs and algorithms for virtual SNP screening significantly complicates the work with the analysis of polymorphisms.

We set the goal of the work – to develop a virtual screening algorithm, which includes the search for possible proteins, whose polymorphisms are meaningful for study, as well as sorting SNPs according to various criteria.

To test the algorithm, we selected the ACAN gene. Aggrecan protein is expressed from this gene, which provides the structural properties of the cartilage matrix, and also affects the metabolism of chondrocytes (Grogan et al., 2014; Jiang et al., 2018; Satin et al., 2019, Antunes et al. 2020).

2. Material and methods

Libraries and modules used in the work

The algorithm was developed in the form of a script written in the Python v. 3.6.5 (Python Software Foundation, USA) in PyCharm environment (USA). To develop the algorithm, the following plug-in libraries and modules were used: the Requests v.2.24.0 library was used to implement the request; for searching and selecting information from database sites – library Beautifulsoup4 v.4.9.1 (Plotnikov et al., 2020). The script was executed on the command line, and saved to a Microsoft Excel spreadsheet (Microsoft, USA). As sources of information for the subsequent execution of the algorithm was used the PubMedCentral (National Institutes of Health's National Library of Medicine, USA) and dbSNP (National Center for Biotechnology Information, USA).

Algorithm stages

The work was carried out in stages. To visualize the algorithm, a service for the development of interfaces was used – Figma (Figma Inc., USA), Figure 1.

At the first stage, a search query was formed into the PubMedCentral database (gene proliferation AND chondrocyte AND hyaluronan acid), articles should be released in the last ten years. The output of the necessary information includes the title of the article, link and keywords. After the formation of the request and the output of data, the proteins mentioned in the articles were selected.

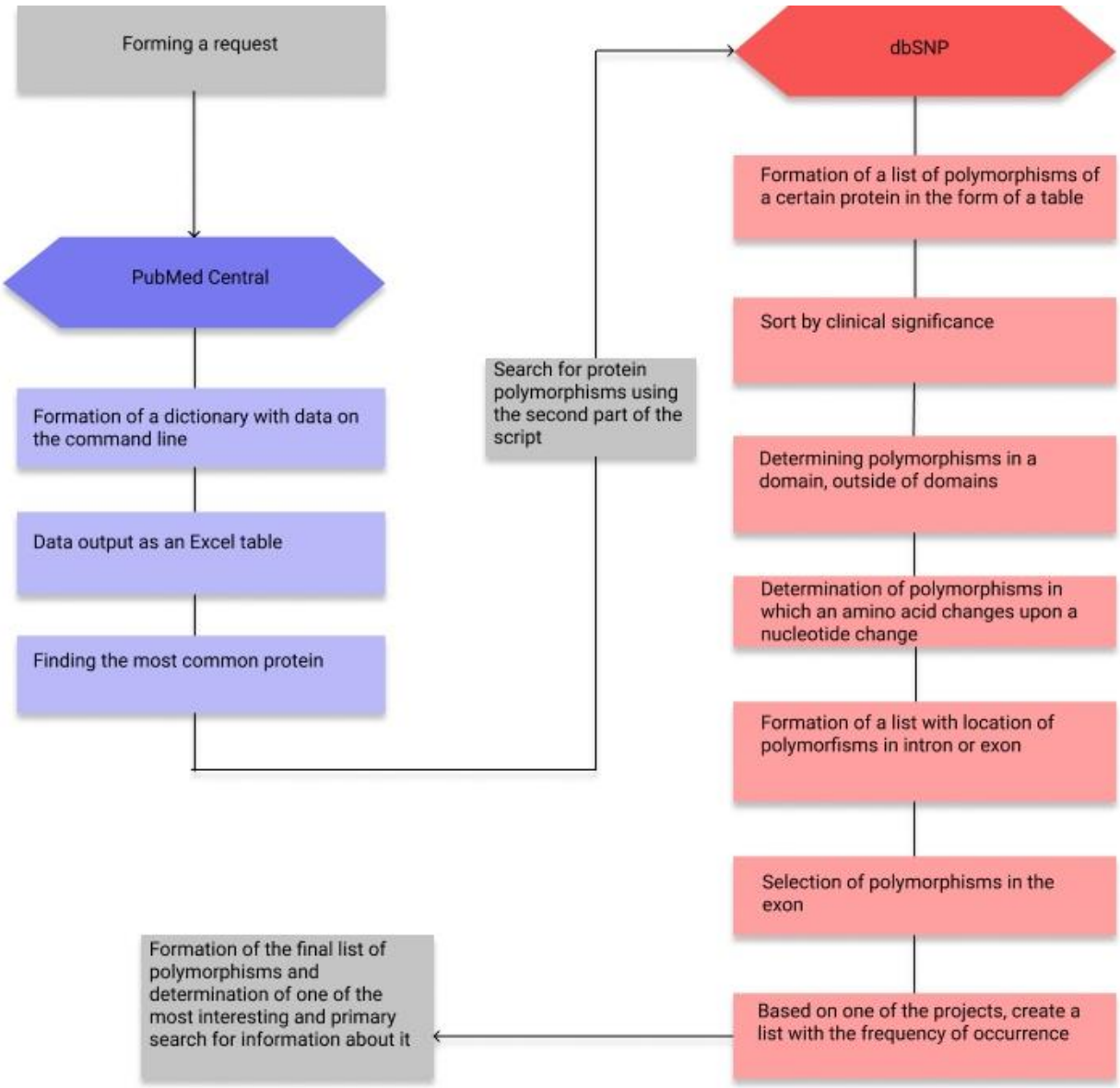

Fig. 1. Algorithm block diagram. Intermediate steps without scripts (gray), using the first part of the script to work with PubMedCentral (blue), using the second part of the script to work with dbSNP (red)

The second stage involves working with dbSNP. One protein was selected from the ones found, which is most often mentioned, and through part of the script, the database was parsed and the polymorphisms were sorted. Those polymorphisms were deduced that have clinical manifestations, they are of greater interest. The locations in the domains, intron or exon, and the frequency of occurrence were indicated. To determine the frequency of occurrence, the research of the «1000 Genomes», project was used, the goal of which was to sequence the genomes of approximately 2500 people in the studied populations to create a detailed catalog of human genetic variations. Then polymorphisms that are in the intron and have an average frequency of occurrence are automatically selected.

3. Results

As a result of the first part of the script, forty-four articles were found. There are twenty-three articles that mention proteins, the most common protein is Aggrecan (thirteen times). One of the articles, as an example, is shown in Table 1.

Table 1. Article title, PubMedCentral link, and referenced proteins

Article	Link	Proteins
Effect of Combined Leukocyte-Poor Platelet-Rich Plasma and Hyaluronic Acid on Bone Marrow-Derived Mesenchymal Stem Cell and Chondrocyte Metabolism	https://pubmed.ncbi.nlm.nih.gov/31282189/	Aggrecan, MMP-9,MMP-13.

After the execution of the second part of the script, there were definitely one hundred thirty-six polymorphisms with clinical manifestations, then these polymorphisms will be used in the work. An example of several of the polymorphisms is shown in Table 2.

Table 2. Clinical manifestation of polymorphism in gene ACAN

SNP ID	Clinical significance	Name gene
1568116	benign	ACAN
2882676	benign	ACAN
34124958	uncertain-significance	ACAN

Then a list of polymorphisms was formed, determining whether they are in structural or functional domains. The page with information about each polymorphism lists the coding amino acid, its number and whether the amino acid is replaced if the nucleotide changes. To determine the domain, you need the amino acid number. It is based on the complete sequence of the Aggrecan protein (Figure 2) and the nucleotide numbers that are in the domains.

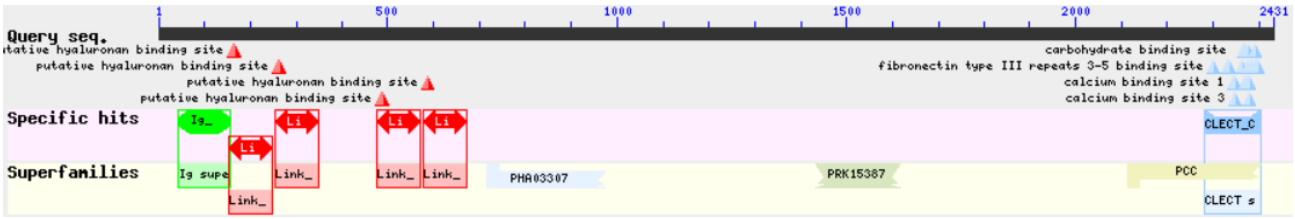

Fig. 2. The nucleotide sequence of the Aggrecan protein with domains

Using this sequence, it is possible to determine the boundaries of the domains, that is, from which nucleotide on which structural or functional domains are located. If the nucleotide number of a particular polymorphism is included in any of the boundaries, the output will be "YES", but if the number is not included in the boundaries, then a dash is displayed. An example of a list is shown in Table 3.

Table 3. Localization of polymorphism in the functional domain

SNP ID	Clinical significance	Name gene	Domian
1568116	benign	ACAN	YES
2882676	benign	ACAN	YES
34124958	uncertain-significance	ACAN	-

There are 85 in the domains, and 52 polymorphisms outside the domains. Next, you should form a column with the replacement of nucleotides. The data can also be taken from the pages with information about nucleotides. An example of a list in Table 4.

Table 4. Information about the replacement of nucleotides in the gene ACAN

SNP ID	Clinical significance	Name gene	Domian	Replacement
1568116	benign	ACAN	YES	
2882676	benign	ACAN	YES	E (Glu) > A (Ala)
34124958	uncertain-significance	ACAN	-	G (Gly) > D (Asp)

Next, the SNP in the intron or exon is determined. Information is taken from pages with data on polymorphism, if there is no data on nucleotides (N/A), then "Intron" is displayed in the column, if information is present, then "Exon" is displayed. An example of some polymorphisms is shown in Table 5.

Table 5. Localization of the gene ACAN polymorphism in the encoded or non-encoded region mDNA

SNP_ID	Clinical significance	Name gene	Domian	Replacement	Intron/Exon
182894280	uncertain-significance	ACAN	YES	A (Ala) > T (Thr)	Exon
185836629	benign	ACAN	-	-	Intron
185960535	benign	ACAN	-	-	Exon

Of the specific polymorphisms in the intron, there are six polymorphisms; they are removed from the mRNA by splicing, and therefore are not of particular interest. In the regions that make up the mRNA, that is, in the exons there are 130 polymorphisms.

Further, the frequency of occurrence is determined for these 130 polymorphisms, as it was indicated that for all polymorphisms, data from the 1000 Genomes project are taken. An example of some polymorphisms is shown in Table 6.

Table 6. Frequency of occurrence of gene polymorphisms ACAN

SNP_ID	Clinical significance	Name gene	Domian	Replacement	Intron/Exon	Frequency
1568116	benign	ACAN	YES		Exon	0,004393 / 22
2882676	benign	ACAN	YES	E (Glu) > A (Ala)	Exon	0,424121 / 2124
34124958	uncertain-significance	ACAN	-	G (Gly) > D (Asp)	Exon	0,027556 / 138

Polymorphisms for which data are absent in known sources and data for which are absent are not required. A total of 87 polymorphisms were found. Now one of the polymorphisms has been selected – rs34124958 (Figure 3).

Fig. 3. Genomic regions and transcripts gene ACAN

Polymorphism is outside the domains, amino acid substitution occurs, the frequency of occurrence is 0.27556 per 138 people, is of uncertain clinical significance, although it is associated with spondyloepimetaphyseal dysplasia. This disease is a new form of skeletal dysplasia with manifestations of noticeably short stature, facial dysmorphism and characteristic radiographic findings.

4. Discussion

To date, three cases have been described, all from the same family. The disease results from a missense mutation affecting the aggrecan C-type lectin domain (AGC1 gene; chromosome 15), which regulates endochondral ossification. Transmission is autosomal recessive (Fukuhara et al. 2019). Most likely, this polymorphism is classified as undefined, since the disease with which it is associated is quite rare and was discovered relatively recently. Since the disease is associated with a domain and amino acid substitution occurs in this polymorphism, it is likely that rs34124958 can change the conformation of the protein in the region with this domain, which leads to disease.

During the development of the algorithm, there were technical difficulties associated with the structure of the SNP database page, but this problem can be eliminated by using additional libraries and scripts. The algorithm can be modified by adding various elements for the accuracy of the search, which can complement existing methods for diagnosing and detecting genetic diseases (Méjécase et al., 2020).

5. Conclusion

The developed algorithm is a promising tool for the search for polymorphisms of genes and proteins expressed by them, possessing certain structural and functional properties, and participating in various biological processes. The polymorphism search algorithm can be used to solve a wide range of problems in the field of biotechnology, pharmaceutical industry, and medicine.

6. Acknowledgements

The reported study was funded by RFBR according to the research project № 20-013-00145 "Mechanisms for the complex influence of environmental factors on the consumption of psychoactive substances by the population of local territories of the Russian Federation".

References

Antunes et al., 2020 – *Antunes, B.P., Vainieri, M.L., Alini, M. et al.* (2020). Enhanced chondrogenic phenotype of primary bovine articular chondrocytes in Fibrin-Hyaluronan hydrogel by multi-axial mechanical loading and FGF18. *Acta biomaterialia*. *105*: 170-179. DOI: 10.1016/j.actbio.2020.01.032

Fukuhara et al., 2019 – *Fukuhara, Y., Cho, S. Y., Miyazaki, O. et al.* (2019). The second report on spondyloepimetaphyseal dysplasia, aggrecan type: a milder phenotype than originally reported. *Clinical dysmorphology*. 28(1): 26-29. DOI: 10.1097/MCD.0000000000000241

Gorący et al., 2020 – *Gorący, I., Grudniewicz, S., Safranow, K., et al.* (2020). Genetic polymorphisms of MMP1, MMP9, COL1A1, and COL1A2 in polish patients with thoracic aortopathy. *Disease markers*. 9567239. DOI: 10.1155/2020/9567239

Grogan et al., 2014 – *Grogan, S.P., Chen, X., Sovani, S. et al.* (2014) Influence of cartilage extracellular matrix molecules on cell phenotype and neocartilage formation. *Tissue engineering. Part A*. 20(1-2): 264-274. DOI: 10.1089/ten.TEA.2012.0618

Jiang et al., 2018 – *Jiang, X., Liu, J., Liu, Q. et al.* (2018). Therapy for cartilage defects: functional ectopic cartilage constructed by cartilage-simulating collagen, chondroitin sulfate and hyaluronic acid (CCH) hybrid hydrogel with allogeneic chondrocytes. *Biomaterials science*. 6(6): 1616-1626. DOI: 10.1039/c8bm00354h

Krajewski et al., 2020 – *Krajewski, W., Karabon, L., Partyka, A. et al.* (2020) Polymorphisms of genes encoding cytokines predict the risk of high-grade bladder cancer and outcomes of BCG immunotherapy. *Central-European journal of immunology*. 45(1): 37-47. DOI: 10.5114/ceji.2020.94674

Méjécase et al., 2020 – *Méjécase, C., Malka, S., Guan, Z. et al.* (2020). Practical guide to genetic screening for inherited eye diseases. *Therapeutic advances in ophthalmology*. 12: 2515841420954592. DOI: 10.1177/2515841420954592

Mikhaylenko et al., 2020 – *Mikhaylenko, D.S., Nemtsova, M.V., Bure, I.V. et al.* (2020) Genetic polymorphisms associated with sheumatoid arthritis development and antirheumatic therapy response. *International journal of Molecular Sciences*. 21(14): 4911. DOI: 10.3390/ijms 21144911

Mori et al., 2014 – *Mori, Y., Saito, T., Chang, S. H., et al.* (2014). Identification of fibroblast growth factor-18 as a molecule to protect adult articular cartilage by gene expression profiling. *The Journal of biological chemistry*. 289(14): 10192-10200. DOI: 10.1074/jbc.M113.524090

Plotnikov et al., 2020 – *Plotnikov, A., Shcheludyakov, A., Cherdantsev, V. et al.* (2020). Data on post bank customer reviews from web. *Data in brief*. 32: 106152. DOI: 10.1016/j.dib. 2020.106152

Satin et al., 2019 – *Satin, A.M., Norelli, J.B., Sgaglione, N.A., Grande, D.A.* (2019). Effect of combined leukocyte-poor platelet-rich plasma and hyaluronic acid on bone marrow-derived mesenchymal stem cell and chondrocyte metabolism. *Cartilage*. 1947603519858739. DOI: 10.1177/1947603519858739

Tretiakov et al., 2020 – *Tretiakov, A., Malakhova, A., Naumova, E., et al.* (2020). Genetic biomarkers of panic disorder: a systematic review. *Genes*. 11(11): 1310. DOI: 10.3390/genes11111310