ABSTRACT

for the dissertation of Yekaterina Golenko on the topic «Development of algorithms for data analysis of mass spectrometry of native proteins» for the degree of Doctor of Philosophy (PhD) in the educational program 8D06101 – «Big Data Analytics»

Relevance of the study. In modern scientific practice, mass spectrometry has established itself as one of the central technologies for the analysis of peptides and proteins. The procedure for identifying proteins using mass spectrometry involves breaking down proteins into peptides, which are then separated, fragmented, ionized, and captured by mass spectrometers. Proteins are recognized and catalogued based on their mass spectra, more precisely, characteristic peaks that correspond to peptide fragment ions. Many factors, including the presence of post-translational modifications, protein fragmentation, and ionization problems, mean that only a limited set of proteins can be accurately identified. Current methods typically reliably identify less than half of all proteins in a sample, leaving a significant number of potentially important molecules without confirmation of their presence and function. This requires further improvement of mass spectrometry methods and the development of more powerful algorithms for processing and interpreting spectral data.

Current protein identification strategies in bioinformatics fall into two main categories. The first is a database search, which operates on the principle of comparing experimentally obtained mass spectra with a pool of theoretically generated peptide spectra created in silico based on known and sequenced protein sequences. Protein databases with search engines such as Mascot, SEQUEST and X!Tandem are commonly used. Spectral library searching is an advanced methodology that uses libraries containing peptide mass spectra data obtained from previous experiments. When a mass spectrum of an unknown peptide is available, it is compared to existing spectra in the library for accurate identification, increasing the speed of analysis and improving the accuracy of peptide identification. Finally, de novo sequencing determines protein sequences directly from mass spectrometry data by examining peaks corresponding to fragmented peptide ions, without the use of biological databases.

The use of databases to identify peptides and proteins is the preferred method of analysis; this approach is effective for recognizing already studied proteins whose sequences are contained in databases but is limited when searching for unknown proteins or those that have undergone post-translational changes, which can increase search time and risk getting incorrect results.

Protein function prediction is another major challenge in bioinformatics, which aims to determine the functions performed by a known protein. Typically, protein function identification is accomplished through manual or computer annotation. Automated function prediction based on the Gene Ontology system is a challenging task in bioinformatics. The complexity of this task is due to the following factors: firstly, most of the proteins that have not been annotated by experts do not contain any information other than the amino acid sequence. Secondly, there is the difficult issue

of tuning the parameters and hyperparameters of the selected model. Third, Gene Ontology has a complex and heterogeneous structure, so the feature prediction task should be considered as a multi-output label task.

Thus, today the following algorithms are of importance: (i) algorithms that focus on primary information processing, which facilitates the transformation of raw spectrometric data for deeper analysis; (ii) algorithms for interpreting spectral data to identify common and implicit patterns; (iii) databases and associated algorithms for identifying peptides and proteins, and (iv) functional annotation algorithms capable of associating protein sequences with biological functions.

The aim of the study is to develop algorithms for interpreting mass spectrometry results and predicting protein functions.

The tasks of the study:

1. Analyse existing solutions for processing mass spectrometry data and protein sequences;
2. Develop an algorithm for identifying peptides using machine learning models;
3. Develop an algorithm for predicting the functions of protein sequences using machine learning methods;
4. Evaluate the proposed algorithms on publicly available data sets using generally accepted evaluation metrics for machine learning.

The object of the study is mass spectrometry data obtained within the framework of the project «PCR test for detection and differential diagnosis of pathogens of opisthorchiasis and methorchiasis» under the supervising of V. S. Kiyan, PhD, NIPSB, as well as data from publicly available protein and DNA sequences databases, including NIST, Pfam and UniProt.

The subject of the study are algorithms for identifying peptides and proteins, as well as determining their functions.

Research methods. When analysing experimental data and developing algorithms, methods for analysing large data sets, qualitative analysis, sequence comparison methods, neural networks, clustering, and classification were used.

Provisions to be defended:

1. An algorithm for identifying peptides obtained by mass spectrometry, based on a bidirectional LSTM neural network embedded in a deep similarity network for working with spectra and peptides;
2. An algorithm for predicting the functions of protein sequences, based on a bidirectional LSTM neural network and a self-attention mechanism.

Connection of the topic with plans for research programs. The research topic corresponds to the priority scientific direction “Information, communication and space technologies.”

The presented results were obtained during the implementation of the project AP05131132 “PCR test for detection and differential diagnosis of pathogens of opisthorchiasis and metorchiasis” in 2018-2020.

Scientific novelty:

- An algorithm for identifying peptides was proposed, developed based on the open-source similarity network SpeCollate, using the BiLSTM neural network to search for matches of the peptide spectrum;

- An algorithm for annotating protein functions is proposed, based on a combination of the BiLSTM neural network and the self-attention mechanism;
- The developed algorithms allow processing biological data, simultaneously using both machine learning and sequence comparison methods.

Theoretical value of research results:

- The developed algorithms have no restrictions on the length of the amino acid sequence and, therefore, can be used to annotate protein functions on a genome scale;
- They work quickly and can annotate several thousand proteins in a few minutes, even on a single processor;
- Models are not limited by unbalanced or missing information about protein-protein interactions;
- Algorithms can be applied using only sequence motifs.

Practical value of research results:

- The developed algorithms can be implemented in laboratory software modules to identify protein sequences and predict their functions with high reliability.
- The developed algorithms can be used by biologists as an alternative to existing applications or additional tools when working with biological data.
- The neural networks underlying the algorithms can be trained on a variety of data sets to solve a wide range of biological problems that require determining the functions of proteins, primarily for understanding disease mechanisms and drug development.

The scientifically based theoretical and experimental results of the dissertation work were used in a scientific project on the topic «PCR test for detection and differential diagnosis of pathogens of opisthorchiasis and metorchiasis».

The software modules created as a result of the dissertation research have been used and implemented in the laboratory of biodiversity and genetic resources of the National Center of Biotechnology (Astana, Kazakhstan) and New Software Systems LLC (Novosibirsk, Russian Federation).

Approbation of the results of the dissertation research. The main results of the dissertation research were reported and discussed at scientific seminars of the Department of Information Systems of S. Seifullin KATU, the Department of Information Systems of L. N. Gumilyov ENU and at the following international scientific and practical conferences:

- International scientific and theoretical conference «Seifullin Readings – 16», KATU named after. S. Seifullin, Nur-Sultan, 2020;
- International scientific and practical conference «Integration of science, education and production is the basis for the implementation of the Nation Plan», KSTU, 2020;
- International scientific and theoretical conference «Seifullin readings – 17: «Modern agricultural science: digital transformation»», Nur-Sultan, 2021;
- International scientific conference «XXII Satpayev Readings», Satbayev University, Almaty, 2022;

- International scientific and theoretical conference «Seifullin readings - 18: «Youth and science - a look into the future»», S. Seifullin KATU, Astana, April 12, 2022;

- International scientific conference «Mathematical logic and computer science», L.N. Gumilyov ENU, Astana, October 7-8, 2022.

Articles in journals included in the international SCOPUS database. The author's contribution consists of hypotheses, data collection, technical implementation of experiments, interpretation of results and preparation of publication:

1. Golenko et al. Implementation of machine learning models to determine the appropriate model for protein function prediction. Eastern-European Journal of Enterprise Technologies, 2022. https://doi.org/10.15587/1729-4061.2022.263270

Articles in journals recommended by the Committee for Quality Assurance in the Field of Science and Higher Education of the Ministry of Science and Higher Education of the Republic of Kazakhstan. Within the framework of these publications, the author is the developer of the research concept and methodology for data analysis. Directly participated in data collection, interpretation of results, formulation of conclusions and preparation of scientific articles for publication:

1. Golenko Y., Ismailova A., Zhumakhanova A. Predicting protein functions using the “Gene Ontology” database and machine learning models. News of the Academy of sciences of the Republic of Kazakhstan. Physics and mathematics series. No. 2 (342). – 2022. – P. 19–38.

2. Golenko Y., Ismailova A., Moldasheva R. Application of deep learning methods for protein structure prediction. Bulletin of the National Engineering Academy of the Republic of Kazakhstan. Series Info-comm. technologies. No. 4 (86). – 2022. – P. 28–40.

3. Golenko Y., Ismailova A. Protein function prediction using the combination of BiLSTM and self-attention algorithm. News of the Academy of sciences of the Republic of Kazakhstan. Physics and mathematics series. No. 3 (347). – 2023. – P. 62–75.

The author’s personal contribution consists of directly carrying out research on all chapters and logical links of the dissertation: conducting a review and analysis of previously presented works, selecting, and justifying the methods used, developing, and technically implementing algorithms, approbating, and testing the developed models on initial data.

Publications. Thirteen (13) scientific papers were published on the topic of the dissertation research, including 1 (one) article in a scientific journal with a non-zero impact factor included in the international SCOPUS database (CiteScore2022 percentile equal to 34), 3 (three) articles in journals recommended by the Committee for Quality Assurance in the Field of Science and Higher Education of the Ministry of Science and Higher Education of the Republic of Kazakhstan, 6 (six) articles in collections of international conferences, 3 (three) articles in other publications. There are 2 (two) copyright certificates of state registration of the computer program.

Structure and scope of the dissertation work. The dissertation research is presented in the following format: introduction, three main sections, conclusion, list of

sources used (123 titles) and two appendices. The total volume is 123 pages of computer text, accompanied by 23 figures and 7 tables.

The introduction emphasizes the importance of the topic under study, the level of knowledge, reveals the relevance of the developed algorithms for processing proteomics data, formulates the purpose of the study, sets the objectives, defines the subject and object of the study, reveals the scientific novelty, theoretical and practical significance of the study. Data on the testing and publication of research results are provided, and the author's personal contribution to scientific research is also indicated.

The first section analyzes the current state of international global repositories of protein structures and genetic sequences, from which the formulation of two main objectives of the dissertation research follows. An extensive analysis of the problem of identifying proteins and peptides isolated through mass spectrometric analysis was carried out, and methods for assessing the correctness of the identified peptides were studied. Methods for solving the problem of identifying proteins and peptides are considered. In addition, an analysis of the problems of annotating experimentally obtained proteins and a review of algorithms for functional prediction were carried out. Key shortcomings and directions for improving these algorithms are identified.

The second section proposes a solution for identifying peptides and proteins based on the publicly available SpeCollate network, adapted to create embeddings of spectra and peptides into a single Euclidean space. The process of network training on data representing both positive and negative examples is characterized and carried out in the context of a SNAP loss function that promotes efficient discrimination between matching and non-matching pairs. The results of training the model are presented and its effectiveness is assessed.

The third section presents the process of developing an algorithm for functional annotation of protein sequences. The process of preliminary processing and analysis of experimental data obtained from open sources for training a neural network is described. A model of an algorithm using a bidirectional LSTM in combination with a self-attention mechanism is presented. The results of training the model using experimental data are presented. The reliability of predicting the functions of the developed model was assessed. The results of manual annotation of protein functions are also presented.

In conclusion the results of the dissertation research are presented, the main conclusions are formulated, confirming, and proving the truth of the provisions submitted for defense.

The appendices contain copyright certificates and implementation certificates.