Системный анализ, математическое моделирование и информационные системы

УДК: 57.087.1

Алгоритм виртуального скрининга баз данных на наличие в них практически значимых белков легких коровы и свиньи

© **2019** П.А. КРЫЛОВ[1,*], Е.В. МАТВЕЕВ[1], В.В. НОВОЧАДОВ[1]

[1]*ФГАОУ ВО Волгоградский государственный университет Министерства образования и науки Российской Федерации, Волгоград 400062*

**e-mail*: krylov.pavel@volsu.ru

Поступила в редакцию 06.06.2019 г.
После доработки 17.07.2019 г.
Принята к публикации 15.08.2019 г.

Разработан алгоритм для виртуального скрининга баз данных на выявление белков, входящих в протеом легких сельскохозяйственных животных, или их аналогов у человека, обладающих практическим значением для фармакологической и биотехнологической промышленности. Алгоритм разрабатывался с использованием языка программирования Python v. 3.6.5 в среде Notepad++. Первичная информация о составе белков, входящих в протеом легких быка и свиньи, изымалась из базы данных UniProt с последующим анализом и поиском совпадений в открытой базе данных DrugBank. В ходе виртуального скрининга было выявлено наличие свыше 5500 белков, входящих в состав протеома легких быка и свиньи, при этом привязка белка к практической значимости отсутствовала у 99% белков, хотя при ручном просмотре в базе данных DrugBank некоторые белки входили в состав лекарственных препаратов. Также в результате работы алгоритма были найдены белки-мишени для лекарственных препаратов, входящие в состав протеома легких человека, которые также присутствуют в протеоме легких свиньи (84) и коровы (46). Человеческий белок-мишень сравнивался с белками-мишенями коровы и свиньи с помощью парного выравнивания аминокислотных последовательностей. В конечном итоге разработанный алгоритм для виртуального скрининга позволил в первом приближении выявить белки, обладающие практической значимостью и, в той или иной степени, входящие в протеом легких сельскохозяйственных животных. В перспективе оптимизация алгоритма и подключение к нему закрытых баз данных позволит проводить детальный скрининг, что даст более полную информацию о белках, которые могут обладать практической значимостью для медицины и биотехнологической отрасли.

Ключевые слова: протеом, базы данных, DrugBank, UniProt, виртуальный скрининг, Python

doi: 10.21519/0234-2758-2019-35-5-80-86

В настоящее время для многих государств мира возрастает необходимость рационального использования ресурсов, в особенности для тех стран, которые испытывают их недостаток. Эффективным решением этой задачи является безотходное производство, в котором вторичное сырье становится источником ценных продуктов для человека. В частности, с помощью различных технологий из продуктов мясоперерабатывающей промышленности может быть выделено достаточно много полезных веществ белкового и небелкового происхождения [1, 2]. Одним из ярких примеров является выделение сурфактант-ассоциированных белков, которые играют важную роль в лечении заболеваний легких [3–6].

В то же время количество технологий, используемых для выделения ценных биологических продуктов из легких сельскохозяйственных животных, весьма невелико. Изучение протеома легких коровы и свиньи проводилось в рамках

Сокращения: БД – база данных; ID – идентификационный номер.

изучения онкологических заболеваний как зарубежными, так и отечественными научными коллективами [7–11]. В этих работах внимание сосредоточено в основном на молекулах, которые участвуют в развитии онкологических заболеваний легких, хотя очевидно, что и другие аннотированные белки могут иметь практическое значение.

В связи с этим возникает необходимость проведения анализа или получения информации, является ли тот или иной белок, входящий в протеом легких сельскохозяйственных животных, практически значимым для какой-либо отрасли промышленности. При попытке ликвидировать дефицит сведений о функциональном предназначении тех или иных белков, исследователь, в первую очередь, проводит анализ в ручном режиме открытых БД (например, UniProt), но процесс этот трудоемкий и требует специальных умений и навыков. Даже в случае полноценного охвата данных, для большинства белков эта информация будет отсутствовать, хотя из других БД или литературы известно о практическом значении таких белков. Например, отсутствующая в UniProt информация о функциональных характеристиках и применении сурфактант-ассоциированных белков имеется в доступной литературе [12–15].

Целью исследования стала разработка алгоритма виртуального скрининга баз данных, находящихся в открытом доступе, на наличие практически значимых белков в качестве активного вещества или белка-мишени, входящих в состав протеома легких коровы и свиньи.

УСЛОВИЯ ЭКСПЕРИМЕНТА

Библиотеки и модули, используемые в работе

Алгоритм был разработан в форме скрипта, написанного на языке программирования Python v. 3.6.5 (Python Software Foundation, США) в среде Notepad++ (GNUGPL, США). Для разработки алгоритма использовались следующие подключаемые библиотеки и модули: для реализации запроса использовалась библиотека Requests v.2.21.0; для поиска и отбора информации с сайтов баз данных – библиотека Beautifulsoup v. 4.7.1; для парного выравнивания аминокислотных последовательностей на основе матрицы BLOSUM62 – библиотека Biopython v. 1.73. Исполнение скрипта осуществлялось в командной строке.

В качестве источников информации о протеоме легких коровы и свиньи и для последующего выполнения алгоритма использовались базы данных UniProtKB (Creative Commons Attribution–No Derivs, США) и DrugBank (University of Alberta, Канада), содержащие сведения о практическом применении белковых молекул [16, 17].

Этапы алгоритма

Работа алгоритма осуществлялась поэтапно. Алгоритм был визуализирован с помощью пакета Microsoft Visio 2016 (Microsoft, США) (рис. 1).

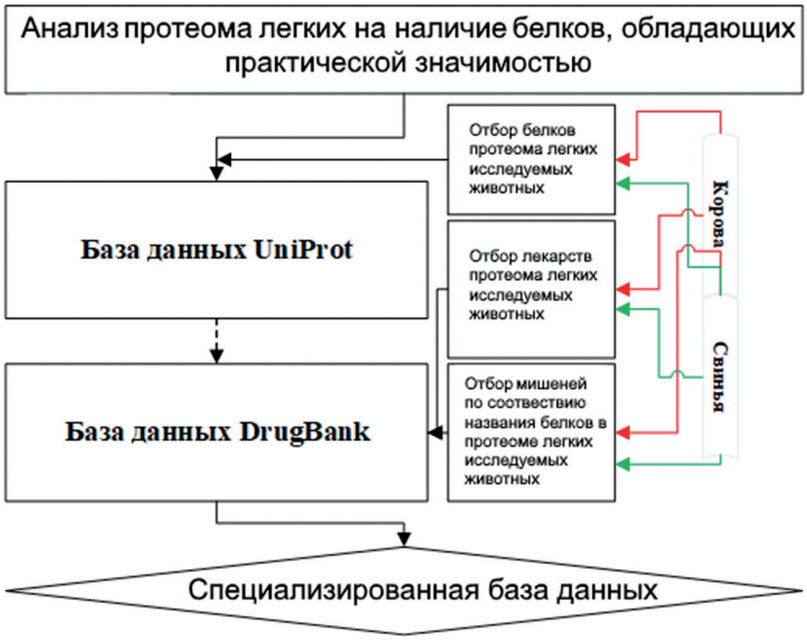

Рис. 1. Блок-схема алгоритма. Стрелками показан путь анализа белков, входящих в протеом легких коровы (красные) и свиньи (зеленые)

Fig. 1. Control-flow chart. The pathway of analysis of proteins comprised in the lungs proteomes is shown by arrows: red (bovine) and green (pig)

На первом этапе происходило формирование запроса в базу данных UniProt (lung and organism: ID organism) с учетом вывода информации по следующим параметрам: ID белка (entry); название белка (protein names); название гена, название организма (organism); практическая значимость в биотехнологической промышленности (biotechnological use); практическая значимость в фармацевтической промышленности (pharmaceutical use). «ID organism» было взято из базы данных UniProt и задавалось в запросе самостоятельно, так как оно было соотнесено с названием организма в реализованном скрипте. После формирования запроса отбирались белки, обладающие информацией о практической значимости.

Второй этап работы алгоритма включал взаимодействие с базой данных DrugBank. Среди всех утвержденных лекарств белкового происхождения (approved protein based therapies), полученных в результате биотехнологического процесса (biotechdrugs), отбирались те, которые содержали в своем описании название организма, заданное пользователем в начале алгоритма, и информацию о локализации белка (lung), либо те, которые имели ссылку на ID из базы данных UniProt.

Третий этап в алгоритме – завершающий. Он также включал работу с базой данных DrugBank. На данном этапе происходил отбор всех утвержденных (approved) лекарственных мишеней по критерию полной идентичности названий самих мишеней и белков протеома легких организма, взятых из базы данных UniProt. В случае наличия у прошедшей отбор лекарственной мишени собственной ссылки ID из базы данных UniProt, выполнялось парное выравнивание аминокислотных последовательностей лекарственной мишени и белка на основе матрицы BLOSUM62 [18].

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

В результате виртуального скрининга по организмам коровы (bovine) и свиньи (pig) с помощью основного скрипта (рис. S1, строка 1–25 (Дополнительный материал)) были получены данные, приведенные в табл. 1, 2.

После выбора организма пользователем для формирования запроса осуществлялся анализ базы данных UniProt. При выборе организма bovine в базе данных UniProt был найден единственный белок Pancreatic trypsin inhibitor

Таблица 1

Результаты виртуального скрининга БД UniProt на наличие практически значимых белков из протеома легких коровы

Results of the UniProt database virtual screening for the presence of bovine's lungs proteome practically significant proteins

ID белк	Название белка	Организм	Биотехнологическое использование	Фармакологическое использование
P00974	Ингибитор панкреатического трипсина (Апротинин). Основной ингибитор протеаз	Корова	–	Ингибирование протеолитических ферментов. Оказывает антифибринолитическое действие

Таблица 2

Служебная информация по лекарственным мишеням, полученная в результате виртуального скрининга БД DrugBank на наличие сходства белков из протеома легких коровы и свиньи

Results of the DrugBank database virtual screening for presence of bovine's and pig's lungs proteins

Лекарственный препарат	ID мишени из UniProt	ID белка из UniProt	Название белка
Корова (bovine)			
Cetuximab	P02745	Q5E9E3	Complement C1q subcomponent subunit A
Drotrecoginalfa	P07204	F1N6M2	Thrombomodulin
Свинья (pig)			
Glutathione	Q6NSD4	F1SA58	Glutathione peroxidase
Reteplase	P05121	I3LQQ6	Plasminogen activator inhibitor 1

(общепринятый синоним Aprotinin) с идентификатором P00974 из протеома легких, использующийся только в фармакологической области.

В базе данных по белкам были представлены идентификационные номера и полные названия, а также название организма и имеющаяся подробная информация о практической значимости каждого белка в биотехнологической и фармацевтической промышленности.

При выборе организма pig в БД UniProt не было найдено ни одного белка из протеома легких свиньи, обладающего практической значимостью (см. рис. S1, строка 141–159 (Дополнительный материал)). Если необходимая информация отсутствовала, то ячейки специализированной БД, соответственно, не заполняли.

Следующим этапом, согласно блок-схеме, стал анализ базы данных DrugBank (рис. S1, строка 28–70 (Дополнительный материал))

В базе данных DrugBank по организму bovine (корова) было найдено единственное лекарство Beractant, в основе которого лежит сурфактант – ассоциированный белок из протеома легких коровы. В специализированной базе данных по этому лекарству были представлены полное наименование лекарственного препарата и ссылка на лекарство из базы данных DrugBank. По организму pig (свинья) в базе данных DrugBank не было найдено лекарственного средства, которое содержало бы в своем составе белки протеома легких свиньи. Формируемая база данных была заполнена, соответственно, лишь служебной информацией по организму bovine:

Биотехнологические препараты . . . . . Beractant
Ссылка из БД DrugBank https://www.drugbank.ca/drugs/DB06761
ID UniProt . . . . . Нет

В базе данных DrugBank были найдены лекарственные мишени – белки человеческого происхождения, полностью идентичные по названию с соответствующими белками из протеома легких быка в количестве 46 шт. (табл. S1 (Дополнительный материал)). Например, лекарственная мишень Thrombomodulin, на которую воздействует лекарство Drotrecoginalfa имела полное совпадение с названием белка из протеома легких быка с идентификатором F1N6M2. При этом сама лекарственная мишень имела идентификатор из UniProt P07204. В БД по данной лекарственной мишени была представлена такая информация, как название лекарства, воздействующего на мишень, ID мишени из UniProt, ID соответствующего по названию белка из UniProt, полное название мишени/белка протеома легких быка. Часть данных по лекарственным мишеням организма bovine из специализированной БД приведены в табл. 2. (см. рис. S1, строка 72–115 (Дополнительный материал)).

В новой папке «Results_alignments» в отдельных текстовых файлах формата txt отражались следующие данные: парное выравнивание аминокислотных последовательностей мишеней и соответствующих белков протеома легких быка, общий счет выравнивания (Score), используемая матрица выравнивания, название организма, название белка и идентификаторы выравниваемых последовательностей (см. рис. S1, строка 116–129 (Дополнительный материал)). В качестве примера выравнивания на рис. 2 показана лекарственная мишень Trombomodulin.

Аналогичным образом в БД DrugBank были найдены лекарственные мишени, полностью идентичные по названию с соответствующими белками из протеома легких свиньи в количестве 84 шт. (табл. S2 (Дополнительный материал)). Примером может служить лекарственная мишень Glutathione peroxidase с идентификационным номером Q6NSD4, на которую воздействует лекарство

```
BlOSUM62 is taken as a matrix for alignment
Score -> 2461.0

MLR-VLLLGV-LAP-AGLGL-PAPP-EPQPLGG-QCVDL-DCFAVFR-GPATFL-AASRV
||--||X||--||--||||--|||--||||-||-|||X--||||XX--|||||||-|-|XX
ML-GVLVLG-ALA-LAGLG-FPAP-AEPQP-GGSQCVE-HDCFALY-PGPATFLNA-SQI

id_protein_organism -- F1N6M2 <----> P07204 -- id_target
name_protein: Thrombomodulin
organism: bovine
```

Рис. 2. Фрагмент выравнивания лекарственной мишени Trombomodulin (белка человеческого происхождения) и белка протеома легких быка с тем же названием

Fig. 2. The drug target alignment result: a protein of human origin – and the bovine lungs protein Trombomodulin

Glutathione. Данная мишень имела полное совпадение с названием белка из протеома легких свиньи с идентификатором F1SA58. Так же, как и в случае с белком быка, в специализированной БД по данной лекарственной мишени были представлены следующие данные: название лекарства, воздействующего на мишень, ID мишени из UniProt, ID соответствующего по названию белка из UniProt, полное название мишени/белка протеома легких свиньи (см. табл. 2). А в новой папке «Results alignments» отдельным текстовым файлом формата txt так же, как и в предыдущем случае, было записано парное выравнивание аминокислотных последовательностей мишени и соответствующего белка протеома легких свиньи.

Итоговым результатом выполнения алгоритма и, соответственно, исполнения скрипта стало создание специализированной БД (файл xls-формата), название которой содержит название организма, и заполненную основной и служебной информацией. Служебная информация представляет собой краткие пояснения к основной информации. Исполнение скрипта осуществлялось продолжительное время (около 6 ч). Макет базы данных представлен на рис. 3.

Таким образом, сформированная БД предоставляет необходимую информацию о практическом применении белковых молекул, позволяя использовать ее данные в дальнейших исследованиях.

Разработанный алгоритм виртуального скрининга позволил частично собрать данные о практической значимости белков, входящих в протеом легких коровы и свиньи, а также о структурных различиях между видами. В полной мере решить поставленные задачи алгоритму не удалось из-за наличия закрытых БД, таких как FDA (Food and Drug Administration, США), которые ограничивали доступ к информации. Это могло также стать причиной обнаружения лишь пяти белков, обладающих практической значимостью.

После анализа протеома легких быка из 1731 белков в БД UniProt, нашелся лишь один, обладающий информацией о его применении. По протеому легких свиньи из 3832 белков в той же БД не нашлось ни одного, который содержал бы информацию о практической значимости. По работе с БД DrugBank результаты по лекарственным препаратам были очень скудны. Из 310 утвержденных лекарственных средств белкового происхождения, полученных в результате биотехнологического процесса, удалось найти лишь одно лекарство, в основе которого лежал белок, входящий в состав протеома легких быка, и не нашлось ни одного лекарства, в основе которого лежал бы белок из протеома легких свиньи. Кроме того, белок из протеома легких быка Aprotinin, найденный в БД UniProt, был найден и в БД DrugBank, однако, в описании не упоминалось о том, что он выделен из легких, отсутствовала и ссылка ID из БД UniProt. При этом известно, что Aprotinin, выделенный из легких коровы, активно используется в кардиохирургии [19] и стоматологии [20]. Эти факты свидетельствуют о существующей проблеме поиска практически значимых белков и представленности их в наиболее часто используемых БД.

Рис. 3. Макет сформированной специализированной БД по практическому применению белков протеома легких свиньи или коровы

Fig. 3. Database model on the practical application pig's and bovine's lungs proteins

Итак, разработанный алгоритм – это перспективный инструмент для поиска белков, обладающих теми или иными свойствами, так как алгоритм можно модифицировать, добавляя различные элементы для точности поиска.

К достоинствам алгоритма можно отнести возможность проводить парное выравнивание аминокислотных последовательностей белков, выступающих в качестве лекарственных мишеней человеческого происхождения. Таким образом, найденные белки с учетом их замены, на основании структурных совпадений аминокислотных последовательностей, могут быть использованы в качестве тестов лекарственных препаратов для медицины и ветеринарии. Однако парное выравнивание аминокислотных последовательностей, безусловно, всего лишь метод предсказания, и необходимо проводить дальнейшие исследования.

Алгоритм ярко продемонстрировал, что некоторые белки протеома легких организма могут найти практическое применение. Для этого необходимы тщательные исследования белковых молекул с целью разработки технологий их применения.

Алгоритм и результаты его работы могут быть использованы для решения задач в области биотехнологии, фармацевтической промышленности и медицины, однако, существующие пробелы значительно снижают ценность и информативность баз данных для исследователей.

ФИНАНСИРОВАНИЕ

Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований и Администрации Волгоградской области в рамках научного проекта № 18-44-343003

Дополнительный материал

Электронная версия статьи содержит дополнительный материал, доступный безвозмездно на сайте журнала http://www.biotechnology-journal.ru

ЛИТЕРАТУРА

1. Мусостов Ш.А. Управление формированием стратегии развития агропромышленным предприятием. *Региональные проблемы преобразования экономики*, 2015. 11(61), 96–104.
2. Faustino M., Veiga M., Sousa P., et al. Agro-Food Byproducts as a New Source of Natural Food Additives. *Molecules*, 2019, 24(6), 1056. doi: 10.3390/molecules24061056
3. Sarker M., Jackman D., Booth V. Lung surfactant protein A (SPA) interactions with model lung surfactant lipids and an SP-B fragment. *Biochemistry*, 2011. 50(22), 4867–4876. doi: 10.1021/bi200167d
4. Casals C., Cañadas O. Role of lipid ordered/disordered phase coexistence in pulmonary surfactant function. *Biochim. Biophys. Acta*, 2012, 1818(11), 2550–2562. doi: 10.1016/j.bbamem.2012.05.024
5. Agrawal V., Smart K., Jilling T., Hirsch E. Surfactant protein (SP)-A suppresses preterm delivery and inflammation via TLR2, *PLoS One*, 2013, 8(5), e63990 doi: 10.1371/journal.pone.0063990
6. Lopez-Rodriguez E., Pérez-Gil J. Structure function relationships in pulmonary surfactant membranes: from biophysics to therapy. *Biochim. Biophys. Acta*, 2014, 1838(6), 1568–1585. doi: 10.1016/j.bbamem.2014.01.028
7. Kosaihira S., Tsunehiro Y., Tsuta K., et al. Proteome expression database of lung adenocarcinoma: A segment of the genome medicine database of Japan proteomics. *J. Proteomics Bioinform.*, 2009, 2, 463–465. doi: 10.4172/jpb.1000106
8. Hill R.C., Calle E.A., Dzieciatkowska M., et al. Quantification of extracellular matrix proteins from a rat lung scaffold to provide a molecular readout for tissue engineering. *Mol. Cellular Proteomics*, 2015, 14(4), 961–973. doi: 10.1074/mcp.M114.045260
9. Fedorchenko K.Y., Ryabokon' A.M., Kononikhin A.S., et al. The effect of space flight on the protein composition of the exhaled breath condensate of cosmonauts. *Russ. Chemical. Bulletin*, 2016, 65(11), 2745–2750. doi: 10.1007 / s11172-016-1645-z
10. Li W., Zhang X., Wang W., et al. Quantitative proteomics analysis of mitochondrial proteins in lung adenocarcinomas and normal lung tissue using iTRAQ and tandem mass spectrometry. *Amm. J. Transl. Res*., 2017, 9(9), 3918–3934.
11. Cheung C.H.-Y., Juan H.-F. Quantitative proteomics in lung cancer. *J. Biomedical. Sci*., 2017, 24(1), 24–37. doi: 10.1186/s12929-017-0343-y
12. Bayat S., Porra L., Broche L., et al. Effect of surfactant on regional lung function in an experimental model of respiratory distress syndrome in rabbit. *J. Appl. Physiology*, 2015, 119(3), 290–298. doi: 10.1152/japplphysiol.00047.2015
13. Speer C.P., Sweet D.G., Halliday H. L. Surfactant therapy: past, present and future. *Early Human Development*, 2013, 89, S22–S24. doi: 10.1016 / S0378-3782 (13)70008-2
14. Chen K.-L., Lv Z.-Y., Yang H.-W., et al. Effects of Tocilizumab on experimental severe acute pancreatitis and associated acute lung injury. *Critical. Care Medicine*, 2016, 44(8), e664–e677. doi: 10.1097/CCM.0000000000001639
15. Baoukina S., Tieleman D.P. Direct simulation of protein-mediated vesicle fusion: lung surfactant protein B. *Biophysical. J*., 2010, 99(7), 2134–2142.doi: 10.1016/j.bpj.2010.07.049
16. Cid F.P., Rilling J.I., Graether S.P., et al. Properties and biotechnological applications of ice-binding proteins in bacteria. *FEMS Microbiol. Letters.*, 2016, 363(11), fnw099. doi: 10.1093/femsle/fnw099

17. Wishart D.S., Feunang Y.D., Guo A.C., et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*., 2017, 46(D1), D1074–D1082 doi: 10.1093/nar/gkx1037

18. Song D., Chen J., Chen G., et al. Parameterized BLOSUM matrices for protein alignment. *IEEE/ACM Transactions Comput. Biol. Bioinform.*, 2015, 12(3), 686–694. doi: 10.1109/TCBB.2014.2366126

19. Wagener G., Gubitosa G., Wang S., Increased incidence of acute kidney injury with aprotinin use during cardiac surgery detected with urinary NGAL. *Am. J. Nephrol*., 2008, 28(4), 576–582. doi: 10.1159/000115973

20. Jegadeesan V., Ponnaiyan D. Impact of aprotinin a proteolyticenzyme on postsurgical symptoms in patients undergoing third molar surgeries. *J. Clin. Diagn. Res*., 2016. 10(1), ZC18–ZC22. doi: 10.7860/JCDR/2016/15491.7056

Virtual Database Screening Algorithm for the Detection of Practically Valuable Proteins of Bovine and Pig Lungs

P.A. KRYLOV[1,*], E.V. MATVEEV[1], and V.V. NOVOCHADOV[1]

[1] *Volgograd State University of the Ministry of Education and Science of the Russian Federation, Volgograd, 400062 Russia*

**e-mail*: krylov.pavel@volsu.ru

Received June 6, 2019
Revised July 17, 2019
Accepted August 15, 2019

Abstract–The algorithm of the virtual database screening for the detection of proteins with the practical significance for the pharmaceutical and biotechnological industries has been developed. The Pythom programming language v. 3.6.5 in Notepad++ framework was used to develop the algorithm. The UniProt database served as a source of the information about the structure of the proteins comprising the bovine and pig lung proteome, and the open DrugBank database was used to the subsequent search for matches in the protein structures. The virtual screening allowed to detect more than 5,500 proteins which are present in the proteome of bovine and pig lungs; the assessment of the practical significance was absent in 99% of the proteins, although it resulted from the manual search in the DrugBank database that some of them were parts of drugs. The algorithm also made it possible to find out target proteins for drugs in the human lung proteome, which were similar with those contained in the bovine (46) and pig (84) lung proteome. Paired alignment of amino acid sequences was used to compare the human and animals' target proteins. In the end, the developed algorithm for virtual screening allowed to identify in the first approximation the proteins with practical significance that are in varying degrees included in the farm animals' lung proteome. In the future, the more detailed screening will be possible due to the algorithm optimization and use of closed databases, which will provide more complete information about practically valuable proteins for biotechnology and medicine.

Key words: proteome, database, DrugBank, UniProt, virtual screening, Python, lungs

Funding–The work was carried out with financial support by Russian Foundation for Fundamental Research and the administration of the Volgograd region within the framework of the scientific project No. 18-44-343003

doi: 10.21519/0234-2758-2019-35-5-80-86