УДК 575.112 **Технические науки**

Ильичев Владимир Юрьевич, *к.т.н., доцент кафедр «Тепловые двигатели и гидромашины» и «Мехатроника и робототехнические системы» Калужский филиал ФГОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет), г. Калуга, Россия*

Кондратьева Светлана Дмитриевна, *к.т.н., доцент кафедры «Системы обработки информации» Калужского филиала ФГОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет), г. Калуга, Россия*

ПРИМЕНЕНИЕ МЕТОДОВ БИОИНФОРМАТИКИ С ИСПОЛЬЗОВАНИЕМ PYTHON JUPITER NOTEBOOK

Аннотация: В статье приведён обзор особенностей применения современных наглядных графических методов в такой науке, как биоинформатика. Описана практическая методика необходимой последовательности использования технологий и команд такого мощного средства, как Jupyter Noutbook, для популярного языка программирования Python с целью реализации наглядной визуализации представления биологических систем при помощи достаточно простых и понятных приёмов. Приведены несколько примеров такой визуализации, произведённой с помощью созданной авторами методики. В заключении показаны пути дальнейшего развития методов биоинформатики с помощью использования языка Python (а точнее, специально созданной для него и также использованной в данной работе библиотеки функций Biopython).

Ключевые слова: биоинформатика, Jupyter Noutbook, модуль Biopython, визуализация биологических систем.

Annotation: The article provides an overview of the features of the use of modern visual graphic methods in such science as bioinformatics. The practical technique of the necessary sequence of using technologies and commands of such a powerful tool as the Jupiter Noutbook, for the popular Python programming language is described to implement visual visualization of the representation of biological systems using fairly simple and understandable techniques. There are several examples of such visualization produced using the method created by the authors. The conclusion shows ways to further develop bioinformatics methods using the Python language (or rather, a library of Biopython functions specially created for it and also used in this work).

Keywords: bioinformatics, Jupiter Noutbook, Biopython module, imaging of biological systems.

Введение

Биоинформатикой [1] называется область научно-практической деятельности человека, в-основном посвящённая объединённым вопросам биологии, химии и математики. Использование приёмов биоинформатики занимает особенно значительное место на крупных фармацевтические, биотехнологических и программных фирмах и предприятиях, оперирующих огромными массивами данных, получившими название Big data [2].

В настоящее время для исследования животных и растительных организмов используется понятие генотипа [3] – совокупности всей информации о строении организма и несущей полную информацию о его наследственности. Генотипы в биоинформатике принято представлять в виде файлов [4], в которых данная информация зашифрована определённым образом – в виде наборов символов и сочетания этих наборов. Подобным же образом может быть представлен химический состав и строение веществ. Для каждой цели наиболее подходящим является свой тип файла. Простейшим, наиболее подробным из которых, но и занимающим наибольшее пространство на жёстком диске, является формат Fasta. Для того же, чтобы описать генотип

целого организма, характеризующегося определёнными сочетаниями генов, их мутациями, возможно использование более компактных форматов файлов, но уже однозначно известно, что любой генотип можно «считать» и сохранить в виде файла, иной раз занимающего много гигабайт информации. Совокупность файлов, в которых хранится описание множества особей организмов, и которые в процессе развития подвергаются скрещиванию и мутациям, в биоинформатике используют для исследования процесса эволюции. Данные исследования производятся с использованием особых математических методов.

С использованием этих же файлов происходит развитие такой отрасли, как биоинженерия, занимающаяся применением инженерных принципов в биологии, например [5]. Биоинженерное дело простирается от создания новых живых органов до генетически модифицированных организмов и химических веществ (лекарств), обладающих новыми свойствами.

Сейчас в сети Интернет представлены файлы генотипов множества растительных, животных и прочих структур. Одним из известнейших сайтов данной тематики является [6]. Перед тем, как работать с файлами генотипов, необходимо расшифровать их код. Для наглядности данный код желательно также визуализировать.

Данная статья как раз и посвящена разработке методики расшифровки и визуализации [7] файлов любых генотипов с использованием специальных современных средств, называемых Jupyter Notebook, созданных для среды программирования Python.

Материал и методы исследования

Опишем пошагово разработанную биоинформационную методику и подробно покажем порядок её использования.

Для установки средства Jupiter Notebook, язык программирования высокого уровня Python вначале нуждается в установке модуля Jupiter, осуществляемого с помощью выполнения следующей команды в командной строке операционной системы (например, в OS Windows вызываемой выполнением команды cmd):

pip install Jupiter

Для запуска Jupiter Notebook в той же командной сроке достаточно набрать jupiter notebook, в ответ на что модуль выдаёт несколько вариантов запуска данного средства: через строку в броузере [8], либо непосредственно путём запуска http файла с локального диска пользователя. Последний способ кажется авторам более простым, и мы использовали именно его. После запуска указанной ссылки на экране компьютера появляется несложный и интуитивно понятный интерфейс Jupiter Notebook, в котором нужно выбрать создание нового файла Notebook. Следует отметить, что для дальнейшей расшифровки генотипов, необходимо скачать их с сайта и поместить в папку, где находится вышеуказанный http файл или в любую папку с известным пользователю путём.

После этого необходимо открыть (для удобства использования) папку и следующий файл, в котором находятся команды, выполняемые построчно (после каждой надо нажимать enter) далее. Следует использовать следующие команды Jupyter Notebook:

```
from Bio.PDB import *
import nglview as nv
import ipywidgets
pdb_parser = PDBParser()
structure = pdb_parser.get_structure("PHA-L", "Data/1FAT.pdb")
view = nv.show_biopython(structure)
```

Смысл данных команд заключается в последовательной загрузке в Python [9] подмодуля PDB import модуля Biopyton, интерактивного модуля nglview и модуля виджета ipywidgets (применение которого аналогично применению в программах графических интерфейсов пользователя GUI), позволяющего произвести запланированную визуализацию генотипов.

Далее осуществляется парсинг (чтение) скачанного файла генотипа и указание на его структуру (чтобы Python смог адекватно распознать заложенную в файл модель). И, наконец, данная структура присваивается некоторой переменной (базе данных) view, над которой можно производить

преобразования различного рода (целью данной работы является лишь её визуализации). Для произведения визуализации в программу необходимо добавить ещё одну команду, задействующую библиотеку виджетов ipywidgets [10]: display(view). После выполнения данной команды средство Jupyter Notebook [11] приобретает дополнительное окно, в котором и появляется визуальная структура прочитанного генотипа, над которой можно производить разнообразные действия: рассматривать её со всех сторон, приближать и удалять для получения лучшего представления о загруженной в виджет структуре.

Пример расчёта

В качестве примера загрузим для изучения из базы [6] одно из известных белковых соединений Phytohemagglutinin-L [12], биоинформационный файл которого кратко назван как 1FAT. Для выполнения визуализации данной белковой структуры выполним рассмотренные выше действия и команды в Jupyter Notebook. Результат, отображаемый при этом на мониторе компьютера, изображён на рис. 1 и является достаточно наглядным.

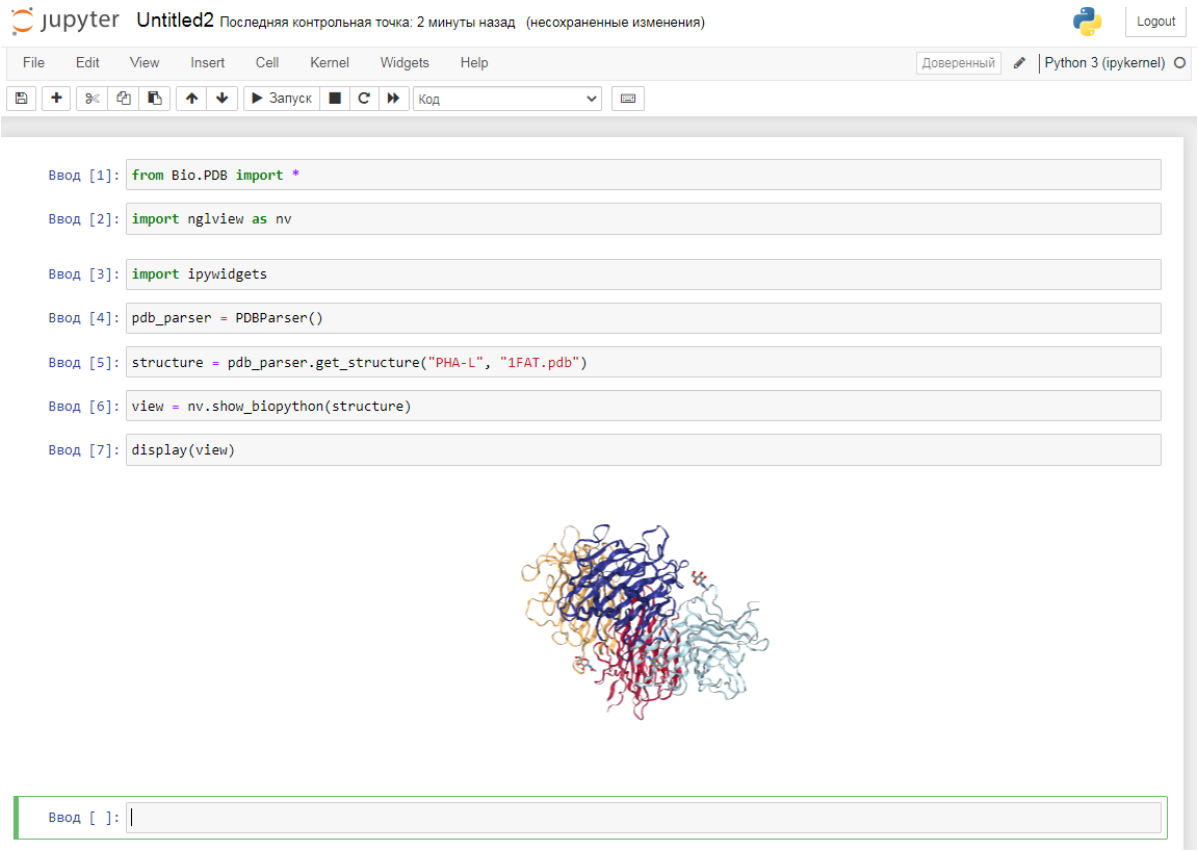

Рис. 1. Выполнение визуализации белковой структуры в интерфейсе Jupyter Notebook

Графическое изображение белковой структуры в виджете можно увеличить, рассматривая отдельные его части [13], как это показано на рис. 2.

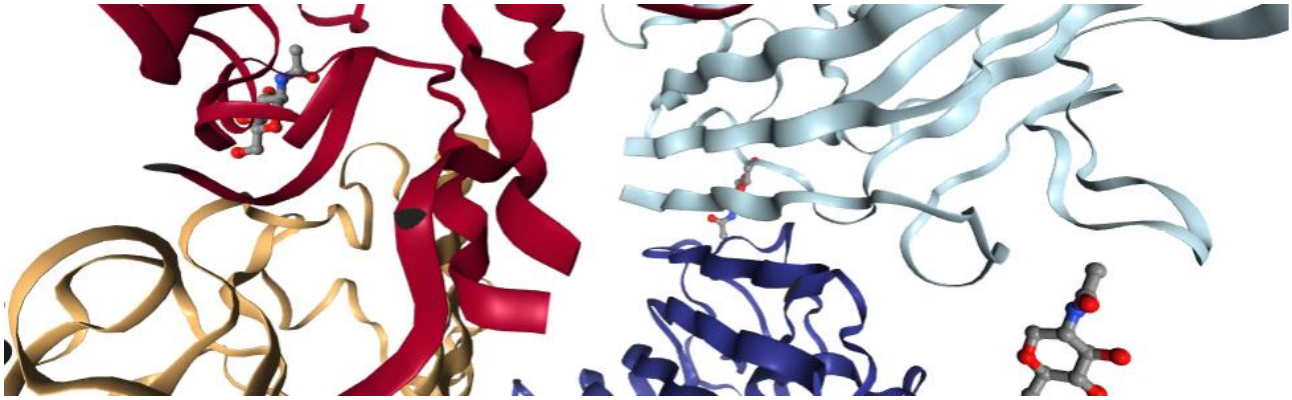

Рис. 2. Увеличенное изображение частей визуализированной с помощью виджета белковой структуры

Подобным же образом можно скопировать и исследовать файл любого генотипа, представленного с помощью разных типов файлов. Некоторые генотипы имеют достаточно комплексное строение. Например, на рис. 3 представлено графическое изображение структуры одного из видов вирусов – N-концевого домена нуклеокапсидного фосфопротеина SARS-CoV-2 [14] в комплексе с 7mer dsRNA. Для его изучения скачаем из базы данных файл в формате pdb, который называется 7ACS.pdb, занимающего на жёстком диске 7,3 МБ. Результаты графического представления данной биологической структуры весьма интересны и частично представлены на рис. 3.

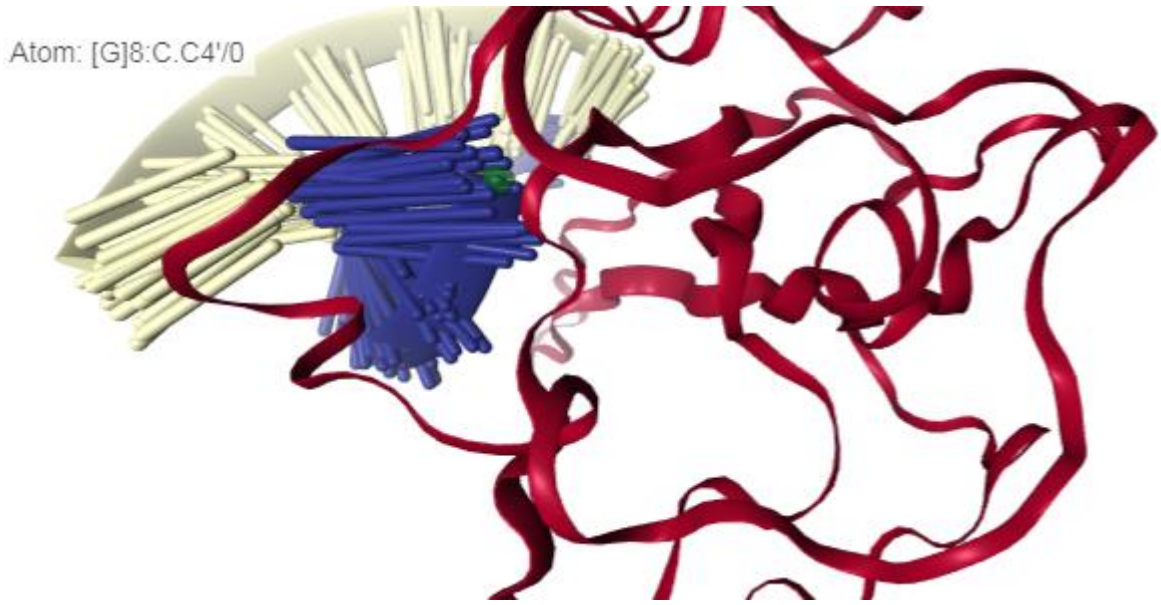

Рис. 3. Увеличенное изображение части коронавируса

Специалист-биолог сможет выделить в представленной структуре белковый и прочие комплексы, и даже предположить основные свойства изображённого биологического объекта. В случае обнаружения биопатогенных свойств возможно дальнейшее применение к объекту методов биоинженерии, например, с помощью создания лекарства, ликвидирующего определённые области биопатогена.

Рассмотрим также, что в используемом виджете не содержащая белков структура, например аскорбиновая кислота, будет выглядеть совсем по-иному (рис. 4).

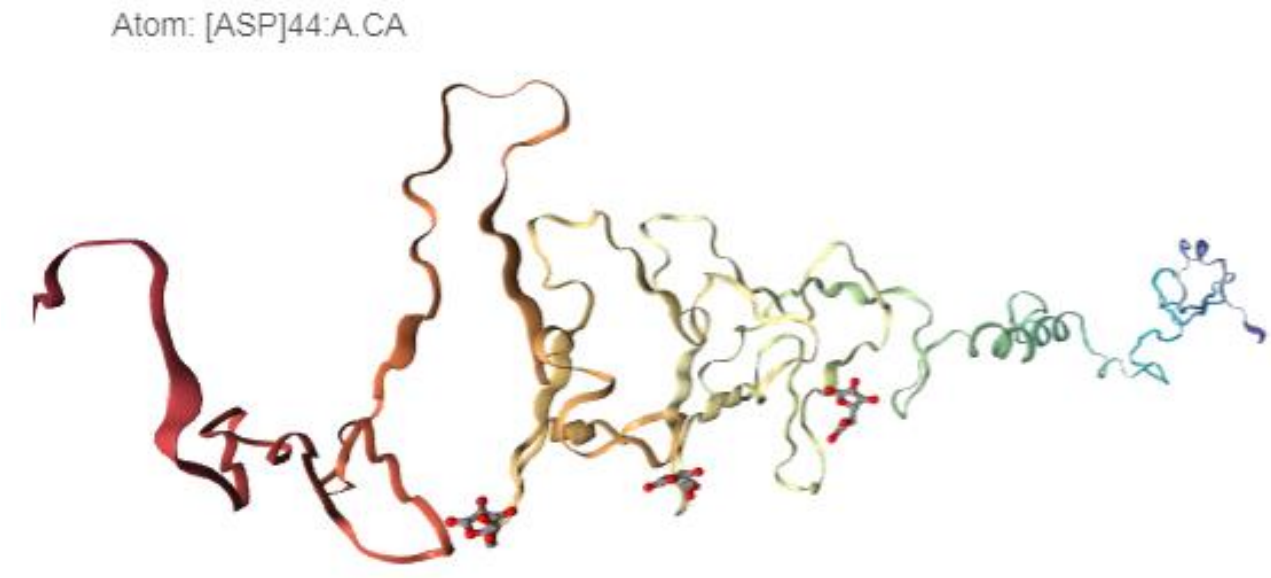

Рис. 4. Увеличенное изображение аскорбиновой кислоты

Заключение

Таким образом, цель описываемой работы достигнута: разработана высокотехнологичная, но достаточно простая в использовании методика, позволяющая создавать визуальные графические модели любых биологических структур, представленных в виде специального вида файлов.

Методика разработана с использованием доступного любому желающему средства Jupyter Notebook, существенно расширяющего возможности такого популярного и свободно доступного языка программирования, каковым является Python. Причём, возможности данного языка возрастают именно в широко востребованной и применяемой сфере науке – биоинформатике. В качестве дальнейшего шага исследований предполагается разработка методик биомодификации моделей биоинформатики. Следует отметить, что материалы представленных исследований позволяют всем заинтересованным данной темой людям изучать функции постоянно обновляемой библиотеки Biopython, являющейся лидирующей при разработке подобных описанному в статье методу исследований.

Также данное исследование способствует освоению желающими новых иструментов разработки на Python, каковым является существующим в одной из его разновидностей (iPython) средства Jupyter Notebook. Кроме него в статье приводятся технологии использования особых приложений, например, для вывода качественной графики, называемых виджетами и легко подключаемым к Jupyter по мере необходимости.

Помимо вышеописанных целей, данная статья ставит задачу популяризации применения современных информационных технологий среди читателей, в частности в такой набирающей популярность область исследований как биоинформатика.

Библиографический список:

1. Баксанский О.Е. Биоинженерия и биоинформатика: конвергентные технологии. // Сеченовский вестник. 2015. № 1 (19). С. 50-55.

2. Шаталова В.В., Лихачевский Д.В., Казак Т.В. Большие данные: как технологии Big data меняют нашу жизнь. // Big Data and Advanced Analytics. 2021. № 7-1. С. 188-192.

3. Моисеев А.Ю., Бурякова О.С., Морозова Н.И. Современные технологии объемных трехмерных изображений. // В сборнике: Научная весна-2019: Экономические науки. Институт сферы обслуживания и предпринимательства (филиал) Донского государственного технического университета. 2019. С. 94-99.

4. Ильичев В.Ю. Исследование применения метода триангуляции делоне совместно с файлами CSV для формирования изображений на языке Python. // Заметки ученого. 2021. № 9-1. С. 301-305.

5. Сафарова Я.Ю.И., Олжаев Ф.С., Умбаев Б.А., Еркебаева А.С., Каренкина А.С., Котов И.В., Russell A.J., Аскарова Ш.Н. Перспективные подходы лечения низкоэнергических травматических повреждений костной ткани с использованием методов биоинженерии и клеточной терапии. // Наука и здравоохранение. 2019. Т. 21. № 5. С. 68-80.

6. Crystal structure of NS3/4A protease variant R155K in complex with vaniprevir. [Электронный ресурс]. URL: https://www.rcsb.org/structure/3SU4. (Дата обращения 07.05.2022).

7. Ильичев В.Ю., Качурин А.В. Создание программ на языке Python для исследования множества Мандельброта. // E-Scio. 2021. № 5 (56). С. 362-371.

8. Ильичев В.Ю. Использование парсинга для создания базы метеорологических данных и разработка на её основе нейросетевой модели прогнозирования скорости ветра. // Системный администратор. 2020. № 10 (215). С. 92-95.

9. Ильичев В.Ю., Юрик Е.А. Создание программы расчёта упорных подшипников скольжения на языке Python. // Научное обозрение. Технические науки. 2020. № 3. С. 14-18.

10. Marini C., Simonelli L., Roque-Rosell J., Campeny M., Toutounchiavval S. Map2Xanes: a Jupyter interactive Notebook for elemental mapping and xanes speciation. // Journal of Synchrotron Radiation. 2021. Т. 28. С. 1245-1252.

11. Якимчик А.И. Jupyter Notebook: система интерактивных научных вычислений. // Геофизический журнал. 2019. Т. 41. № 2. С. 112-121.

12. Movafagh A., Heydary H., Mortazavi-Tabatabaei S.A.R., Azargashb E. The significance application of indigenous phytohemagglutinin (PHA) mitogen on metaphase and cell culture procedure. // Iranian Journal of Pharmaceutical Research. 2011. Т. 10. № 4. С. 895-903.

13. Герасименко М.А. Виджет в качестве интерфейса взаимодействия с интернет вещью. // В сборнике: Научно-техническая конференция студентов, аспирантов и молодых специалистов НИУ ВШЭ. Материалы конференции. Московский институт электроники и математики Национального исследовательского университета «Высшая школа экономики». 2014. С. 82-83.

14. Li J.-Y., Zhou Z.-J., Wang Q., Qiu Y., Ge X.-Y., He Q.-N., Zhao M.-Y. // Innate immunity evasion strategies of highly pathogenic coronaviruses: SARS-CoV, MERS-CoV, and SARS-CoV-2. // Frontiers in Microbiology. 2021. Т. 12. № FEB. С. 770656.