Abstract
One of the main tasks of modern bioinformatics is the analysis of big data obtained as a result of biomedical research. Molecular docking is one of the methods of molecular modeling. It allows the user to detect and study the features of the probable interaction of molecular models. To determine the best and accurate three-dimensional prostranometric positions of the atoms of molecular models to each other. This paper aim is to create a system that automatically classifies and visualizes the data obtained as a result of molecular docking using clustering methods, ensuring their accurate selection and processing.

Introduction
One of the main tasks of modern bioinformatics is the analysis of big data obtained as a result of biomedical research. Automatic processing, management, analysis of data obtained experimentally, and their interpretation significantly facilitate the ongoing research. Fixing the accuracy of biological data is directly related to their correct classification, methodology of structural regulation and interpretation, the solution of which can be found in modern interprofessional fields, such as bioinformatics and programming.
The paper attempts to create a system that automatically classifies and visualizes the data obtained as a result of molecular docking using clustering methods, ensuring their accurate selection and processing.
The system is based on the use and use of modern software tools in the Python language environment, such as Numpy, Matplolib, Jupyter Notebooks and other data processing packages.
Molecular docking is one of the methods of molecular modeling. It allows the user to detect and study the features of the probable interaction of molecular models. To determine the best and accurate three-dimensional prostranometric positions of the atoms of molecular models to each other, to calculate the energy indicators of their interaction, taking into account the average daily deviations of the prostranometric data of atoms (RMSD). To ensure maximum data accuracy (closer to real conditions), combined docking is usually used. It is a process of predicting the interaction of biologically active compounds (ligand) and a target (biomacromolecule). Currently, there are two types of docking: hard and soft. They differ from each other by the degree of limited freedom of the positions of the target atoms. During the hard docking position, the ligand atoms with respect to the target have the maximum possible degree of freedom, while the target real estate. In soft docking, the atoms of ligand have a limited degree of freedom of movement along with the degree of freedom of the positions of the ligand atoms. As a result, a system is formed in which ligands are positioned on the

target surface in favorable spatial and energy locations. To ensure accurate statistics of the conducted research, a certain amount of primary prostranometric data (positions)of the ligand is created during the experiment, on the basis of which the interaction forecast is carried out. These positions are called conformers. The minimum number of primary conformers accepted in biomodeling is twenty.

Currently, the method of molecular docking is widely used in a number of areas of biology and pharmacy, what is the process of creating new medicines, a preliminary assessment of the bioactivity of modified compounds aimed at reducing the side effects of existing medicines. Study of the interaction of biomacromolecules and biologically active compounds, etc. It is also used as part of the virtual screening methodology, when using computer equipment and the corresponding software (in silico) , out of millions of connections available in specialized databases, connections are selected in accordance with the specified tasks.

Methodology

Currently, the AutoDock software package is widely used in the field of bioinformatics and biomodeling. The program is based on the Lamarckian genetic algorithm. One of the features of this software package is the possibility of "blind docking", which allows us to study the ligand-target interaction even when the active center of the target is not known. The cost of the evaluation function of this software package, in other words, the accuracy of the experiment in silico, is quite high, reaching ≈ 85%.

The outgoing data received using AutoDock is stored using two files. The first results log is where qualitative data of the ligand-target interaction is stored, in particular` the number of conformers and their list, the interaction energy coefficients expressed in the dimension kcal/mol, the value of the average daily deviation from the location of the target atoms of each conformation.

1	-9.2	0.000	0.000
2	-8.7	22.599	24.331
3	-8.7	22.675	25.271
4	-8.5	23.208	25.033
5	-8.5	1.612	2.873
6	-8.4	28.041	29.590
7	-8.4	22.320	24.621
8	-8.4	22.364	24.380
9	-8.3	1.942	6.360
10	-8.2	27.785	30.195
11	-8.2	22.631	23.728
12	-8.2	2.786	5.300
13	-8.1	2.092	5.063
14	-8.1	22.792	24.414
15	-8.0	3.629	5.150
16	-8.0	23.113	24.569
17	-8.0	42.290	45.666
18	-7.9	2.477	7.039
19	-7.8	22.675	24.537
20	-7.8	3.142	8.040

Figure 1: Example of results.log file

It is worth noting that RMSD l.b. ≤ conformers with deviations of 2 Å, whose data are used in the interaction analysis, is considered a positive criterion for choosing conformes. Other conformers that do not meet the above criterion.

The second file is from is in the *.pdbqt format (*. the number of pdbqt files is given using the cycles parameter), which contains spatial data of the positions of atoms related to each of the contours obtained as a result of the interaction, to the target. Spatial data is described in columns 6, 7, 8 using three-dimensional spatial coordinates X, Y, Z, respectively.

```
MODEL 1
REMARK VINA RESULT:      -9.2      0.000      0.000
REMARK  Name = H:\Org. chem candidate structure\str-2\Str2-3D.pdb
REMARK  1 active torsions:
REMARK  status: ('A' for Active; 'I' for Inactive)
REMARK    1  A    between atoms: _18  and  _23
REMARK                            x       y       z     vdW  Elec       q    Type
REMARK                         _______ _______ _______ _____ _____    ______ ____
ROOT
ATOM      1  C   UNL     1       6.608   0.428  84.222  0.00  0.00    +0.063 C
ATOM      2  C   UNL     1       6.695   0.505  82.744  0.00  0.00    +0.111 C
ATOM      3  N   UNL     1       6.353  -0.621  82.064  0.00  0.00    -0.348 NA
ATOM      4  C   UNL     1       6.358  -0.691  80.710  0.00  0.00    +0.138 C
ATOM     11  N   UNL     1       6.724   0.398  79.997  0.00  0.00    -0.328 NA
ATOM     12  C   UNL     1       7.106   1.558  80.584  0.00  0.00    +0.192 C
ATOM     13  N   UNL     1       7.497   2.634  79.835  0.00  0.00    -0.216 N
ATOM     14  C   UNL     1       7.879   3.806  80.398  0.00  0.00    +0.192 C
ATOM     15  C   UNL     1       7.900   3.966  81.792  0.00  0.00    +0.045 C
ATOM     16  C   UNL     1       7.513   2.876  82.592  0.00  0.00    -0.007 C
ATOM     17  N   UNL     1       8.221   4.725  79.475  0.00  0.00    -0.385 NA
ATOM     18  C   UNL     1       8.059   4.114  78.294  0.00  0.00    +0.146 C
ATOM     19  C   UNL     1       8.285   4.620  77.004  0.00  0.00    +0.029 C
ATOM     20  C   UNL     1       8.049   3.771  75.903  0.00  0.00    -0.019 C
ATOM     21  C   UNL     1       7.598   2.447  76.098  0.00  0.00    +0.053 C
ATOM     22  C   UNL     1       7.375   1.955  77.399  0.00  0.00    -0.004 C
ATOM     23  C   UNL     1       7.609   2.806  78.501  0.00  0.00    +0.085 C
ATOM     24  C   UNL     1       7.107   1.651  82.008  0.00  0.00    +0.139 C
ENDROOT
BRANCH   4   5
ATOM      5  C   UNL     1       5.998  -1.871  80.043  0.00  0.00    +0.086 A
ATOM      6  C   UNL     1       4.643  -2.262  79.962  0.00  0.00    +0.012 A
ATOM      7  C   UNL     1       4.288  -3.455  79.307  0.00  0.00    +0.001 A
ATOM      8  C   UNL     1       5.283  -4.254  78.713  0.00  0.00    +0.000 A
ATOM      9  C   UNL     1       6.634  -3.870  78.789  0.00  0.00    +0.001 A
ATOM     10  C   UNL     1       6.993  -2.688  79.460  0.00  0.00    +0.012 A
ENDBRANCH   4   5
TORSDOF 1
ENDMDL
MODEL 2
REMARK VINA RESULT:      -8.7     22.599     24.331
REMARK  Name = H:\Org. chem candidate structure\str-2\Str2-3D.pdb
REMARK  1 active torsions:
REMARK  status: ('A' for Active; 'I' for Inactive)
REMARK    1  A    between atoms: _18  and  _23
REMARK                            x       y       z     vdW  Elec       q    Type
REMARK                         _______ _______ _______ _____ _____    ______ ____
```

Figure 2: Example of pdbqt file

The algorithmic designed to perform clustering analysis is consists of several phases and is based on K-Means Clustering algorithm. In the first phase of the flow data processing is performed. During this phase first, data preprocessing phase the contents of the log and pdbqt files are parsed and stored into respective data structures for further processing and queries. Models are stored in the hierachical fashion. Each models has a parent denoted by cycle number in which it appears, and stored alongside with it's center.

During the second, clustering, phase data retrieved from the first phase is used to perform clustering clusterization using K-Means Algorithm. As the distance metric Euclidean distance is used. The algorithm receives on its input list of models centers from different cycles. During the simulation different numbers of cluster being tested to find the optimal number. On the output the algorithm provides index of the cluster each sample belongs to. It also possible to retreive list of the centroids found during the clustering. K-Means clustering is done to found so called 'big clusters'. For big cluster the threashold after which to models fall into the same cluster is equal to R=4.
After the big clusters are found, the distance between atom coordinates of the models belonging to the same cluster is calculated using Euclidean distance, and if the distance is smaller then r=2 then these two models are said to fall into a subclaster within big cluster. Models are being compaired in pairwase fashion.
During the final phase, found clusters are being visualized in 3D space alongside with found centroids. Also, an output in a textual foramt which consists of the cluster and its content is given.

Experimental results

Conclusion

The development of a cluster analysis and visualization system, which is a problem of work, has been successfully implemented. The developed system makes it possible to automate the clustering analysis of data obtained as a result of molecular docking, making it possible to visualize the obtained data, which in turn facilitates the selection of data and further research.