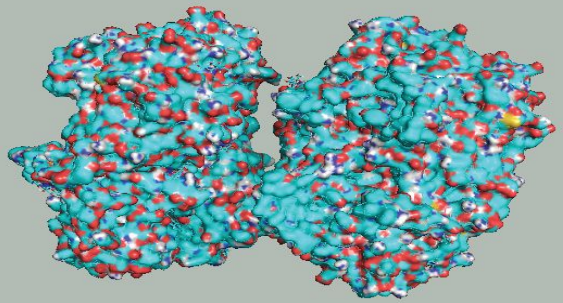
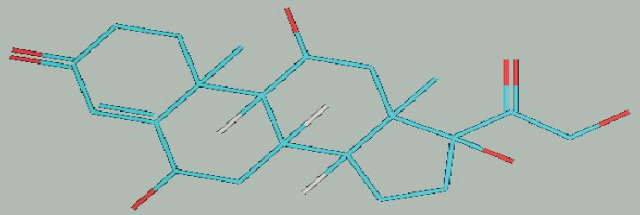
Семинар по докингу
25 октября, 17:30, ауд.113 каф. биоинформатики

WiFi – rsmu2 or rsmu5
Password – 1q2w3e4r

Докладчик: Смирнов А.С. гр.3.3.21 – организация докинга на кластере hadoop



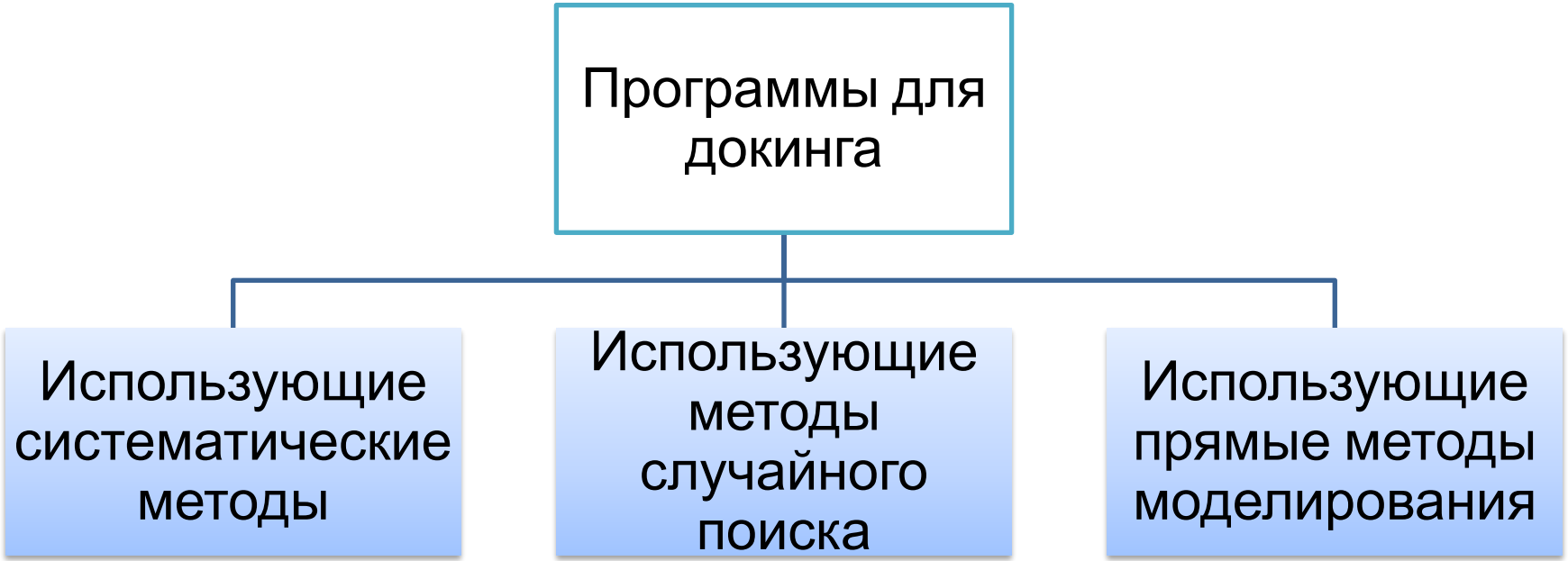
VS



Определение

Молекулярный докинг – это расчет взаимодействия низкомолекулярного вещества с активным центром белка и предсказание энергетически выгодной конформации низкомолекулярного вещества в этом активном центре.

Типы программ



Типы программ

Name	Search algorithm	Type	References
AUTODOCK4	Lamarckian genetic algorithm	Academic	Morris *et al*., 2009
DOCK	Shape matching	Academic	Allen *et al*., 2015
OEDOCKING	Shape matching	Academic	Kelley, Brown, Warren & Muchmore, 2015; Mcgann, 2011
FLEKSY	Ensemble-based	Commercial	Nabuurs, Wagener & De Vlieg, 2007; Wagener, De Vlieg & Nabuurs, 2012
SWISSDOCK	Evolutionary optimization	Academic	A. Grosdidier, Zoete & Michielin, 2011
GOLD	Genetic algorithm	Commercial	Jones, Willett, Glen, Leach & Taylor, 1997
GLIDE	Hybrid	Commercial	Friesner *et al*., 2004
VINA	Local optimization	Academic	Trott & Olson, 2009
RDOCK	Hybrid	Academic	Ruiz-Carmona *et al.*, 2014
LEDOCK	Simulated annealing	Academic	Unzue *et al*., 2016
PLANTS	Ant colony optimization	Academic	Korb, Stützle & Exner, 2009
HADDOCK	Hybrid	Academic	Dominguez, Boelens & Bonvin, 2003
SURFLEX-DOCK	Shape matching	Commercial	Spitzer & Jain, 2012
MOE	Hybrid	Commercial	Vilar, Cozza & Moro, 2008
FLEXX	Shape matching	Commercial	Kramer, Rarey & Lengauer, 1999
FITTED	Hybrid	Commercial	Corbeil & Moitessier, 2009; De Cesco, Kurian, Dufresne, Mittermaier & Moitessier, 2017; Englebienne & Moitessier, 2009; Moitessier *et al*., 2016
LIGANDFIT	Shape matching	Commercial	Venkatachalam, Jiang, Oldfield & Waldman, 2003
ICM	Hybrid	Commercial	Neves, Totrov & Abagyan, 2012
IGEMDOCK	Evolutionary algorithm	Academic	Yang & Chen, 2004

Table I. Examples of software available for protein-ligand docking and their search algorithm.

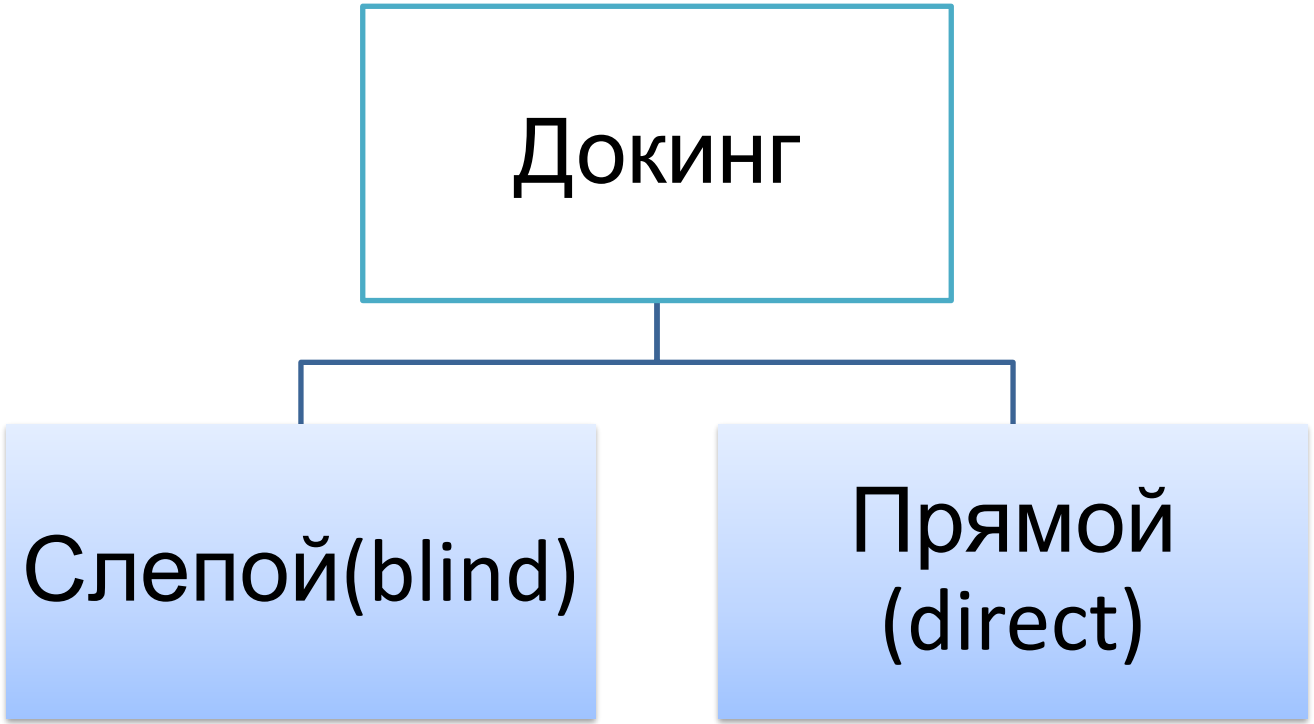
Software	Type	Features
MOE	Commercial	Correction of residue issues, structure clean-up, charge assignment based of several forcefields, protein minimization and binding site prediction.
Maestro	Academic/Commercial	Correction of residue issues, structure clean-up, charge assignment, tautomer assignment, loop remodeling*, binding site prediction.*
YASARA	Academic/Commercial	Correction of residues, binding site analysis, contact analysis, loop remodeling*, charge assignment*, protein minimization* and hydrogen bonds optimization.*
Lead Finder	Commercial	Structure optimization, charge assignment, rotamer selection
RosettaLigand	Academic	Structure optimization, charge assignment, rotamer selection, loop optimization.
BALLView	Academic	Protein minimization and charge assignment.
DeepView	Academic	Protein optimization, loop remodeling and binding site analysis.
Vega ZZ	Academic	Correction of residue issues, structure clean-up, charge assignment (forcefield/gasteiger), semiempirical charges and protein minimization.
SPORES	Academic	Structure preparation, geometry optimization, connectivity correction and tautomer assignment.
UCSF Chimera	Academic	Structure clean-up, charge assignment, loop remodeling, protein minimization.
Autodock Tools	Academic	Structure clean-up, charge assignment (Gasteiger), rotamer selection and binding site prediction.
Openbabel	Open source	Charge assignment, multiple file formats supported, file conversion.

† Presented as illustrative software, due to well-known capabilities or ease of use; for other applications of general use please refer to Prieto-Martínez & Medina-Franco, 2018 *These features require commercial licensing.

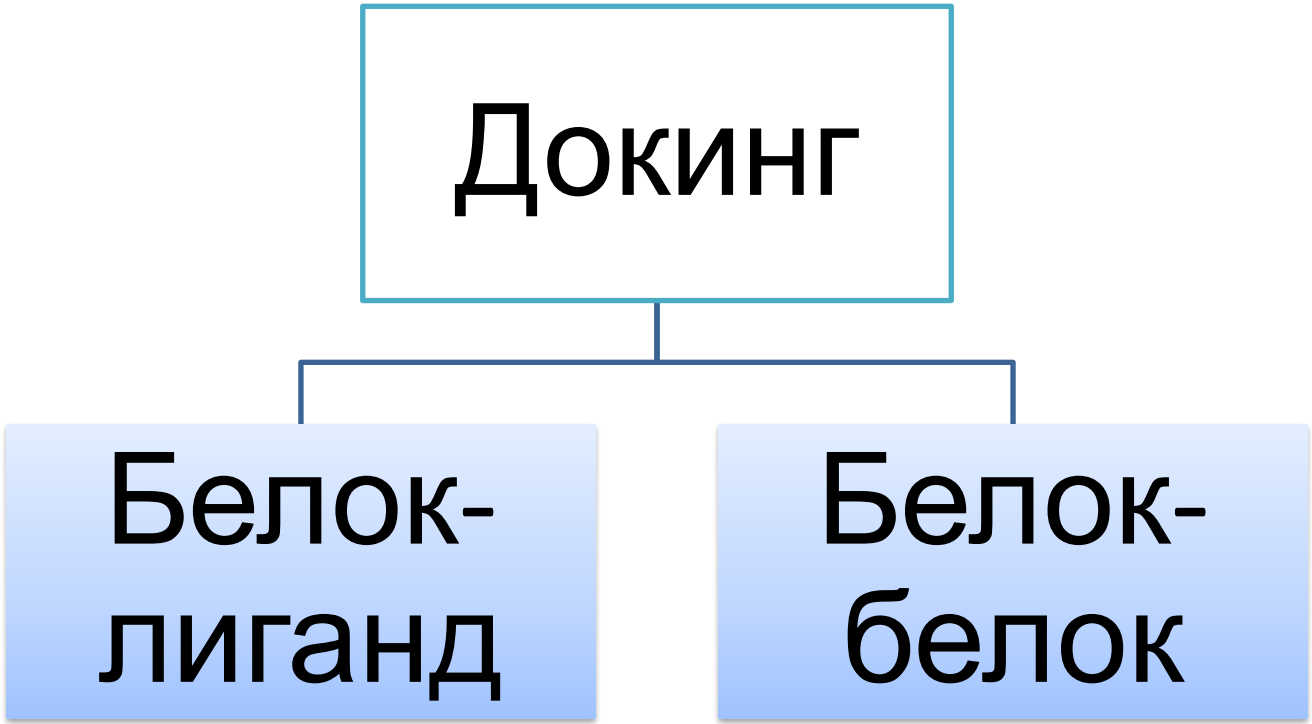
Table III. Software used for protein and ligand preparation.†

Prieto-Martínez FD, Arciniega M, Medina-Franco JL. Molecular docking: current advances and challenges. TIP Rev Esp Cienc Quim Biol. 2018;21(Suppl: 1):65-87.

Типы докинга

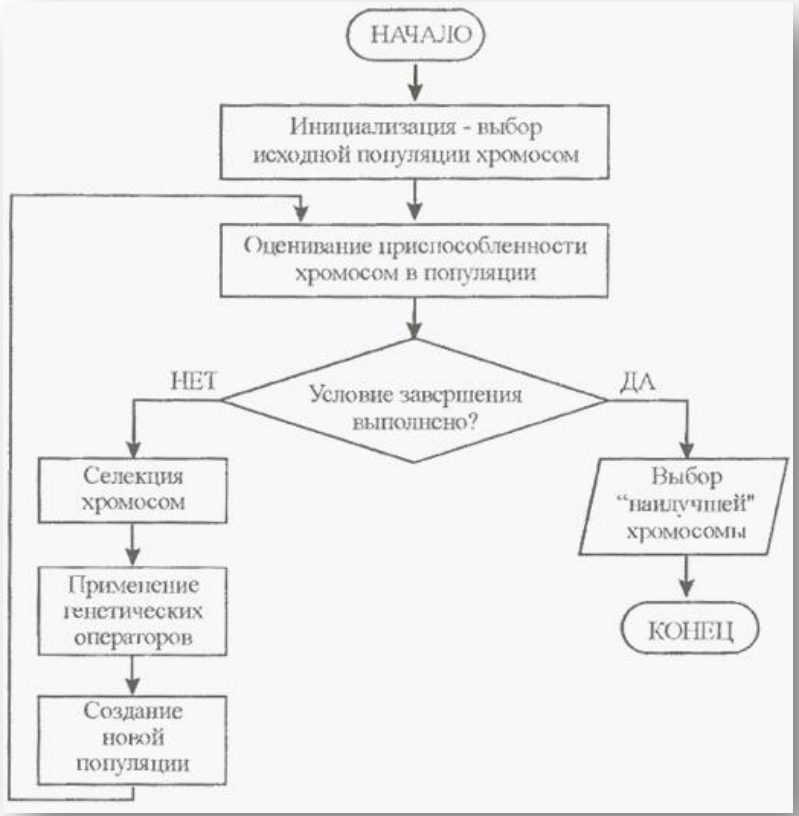


Типы докинга



Как найти нужную конформацию?
Как понять, что она лучшая?

Алгоритмы



Сравнивают по так называемой оценивающей функции

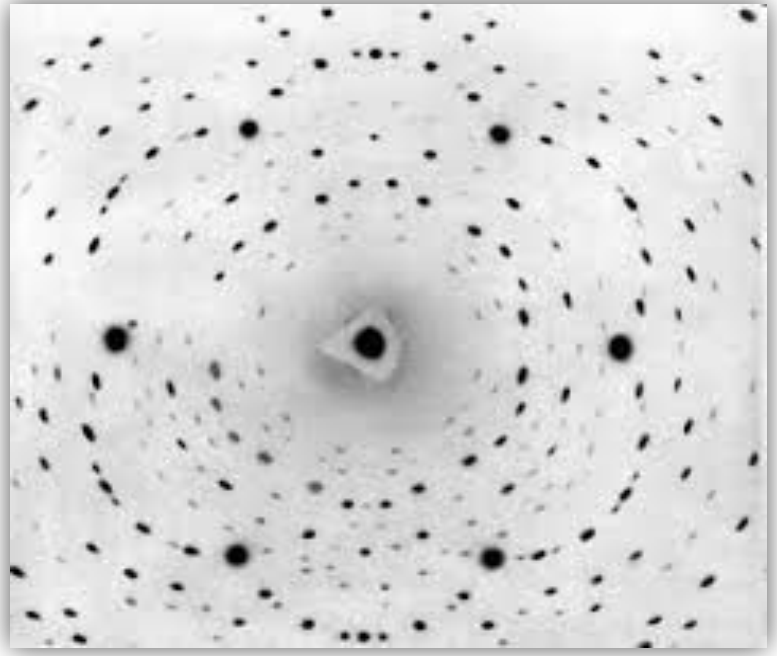
```
Estimated Free Energy of Binding    =  -10.57 kcal/mol  [=(1)+(2)+(3)-(4)]
Estimated Inhibition Constant, Ki   =   17.79 nM (nanomolar)  [Temperature = 298.15 K]

(1) Final Intermolecular Energy     =  -11.77 kcal/mol
    vdW + Hbond + desolv Energy     =  -11.74 kcal/mol
    Electrostatic Energy            =   -0.03 kcal/mol
(2) Final Total Internal Energy     =   -0.82 kcal/mol
(3) Torsional Free Energy           =   +1.49 kcal/mol
(4) Unbound System's Energy         =   -0.53 kcal/mol
```

Это всё замечательно

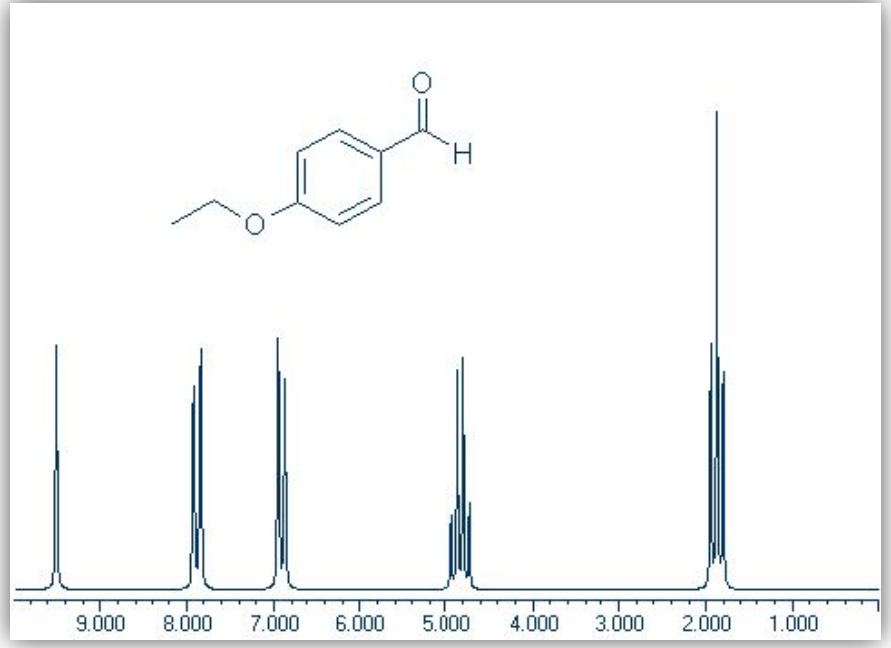
НО КАК ПРЕДСТАВИТЬ БЕЛОК ТАК, ЧТОБЫ КОМПЬЮТЕР ПОНЯЛ, ЧТО ЭТО БЕЛОК? ДА И ЕЩЁ НАМ ЧТО-ТО ПОСЧИТАЛ?

Как вообще расшифровывают структуру белков?

Рентгеновская кристаллография



Ядерно-магнитный резонанс



Форматы представления

```
HEADER    EXTRACELLULAR MATRIX                    22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
...
```

Protein Data Bank (PDB). Формат файла представляет собой формат текстового файла описание трехмерных структур молекул , проведенные в Protein Data Bank . PDB формат соответственно предусматривает для описания и аннотации белков и нуклеиновых кислот структур , включая атомные координаты, наблюдаемые боковую цепь ротамеры , вторичные структуры задания, а также атомную связь.

Пример pdbqt файла

```
COMPND    NSC7810
REMARK  3 active torsions:
REMARK  status: ('A' for Active; 'I' for Inactive)
REMARK    1  A    between atoms: A7_7  and  C22_23
REMARK    2  A    between atoms: A9_9  and  A11_11
REMARK    3  A    between atoms: A17_17  and  C21_21
ROOT
ATOM      1  A1  INH I           1.054   3.021   1.101  0.00  0.00     0.002 A
ATOM      2  A2  INH I           1.150   1.704   0.764  0.00  0.00     0.012 A
ATOM      3  A3  INH I          -0.006   0.975   0.431  0.00  0.00    -0.024 A
ATOM      4  A4  INH I           0.070  -0.385   0.081  0.00  0.00     0.012 A
ATOM      5  A5  INH I          -1.062  -1.073  -0.238  0.00  0.00     0.002 A
ATOM      6  A6  INH I          -2.306  -0.456  -0.226  0.00  0.00     0.019 A
ATOM      7  A7  INH I          -2.426   0.885   0.114  0.00  0.00     0.052 A
ATOM      8  A8  INH I          -1.265   1.621   0.449  0.00  0.00     0.002 A
ATOM      9  A9  INH I          -1.339   2.986   0.801  0.00  0.00    -0.013 A
ATOM     10  A10 INH I          -0.176   3.667   1.128  0.00  0.00     0.013 A
ENDROOT
BRANCH   9  11
ATOM     11  A11 INH I          -2.644   3.682   0.827  0.00  0.00    -0.013 A
ATOM     12  A16 INH I          -3.007   4.557  -0.220  0.00  0.00     0.002 A
ATOM     13  A12 INH I          -3.522   3.485   1.882  0.00  0.00     0.013 A
ATOM     14  A15 INH I          -4.262   5.209  -0.177  0.00  0.00    -0.024 A
ATOM     15  A17 INH I          -2.144   4.784  -1.319  0.00  0.00     0.052 A
ATOM     16  A14 INH I          -5.122   4.981   0.910  0.00  0.00     0.012 A
ATOM     17  A20 INH I          -4.627   6.077  -1.222  0.00  0.00     0.012 A
ATOM     18  A13 INH I          -4.749   4.135   1.912  0.00  0.00     0.002 A
ATOM     19  A19 INH I          -3.777   6.285  -2.267  0.00  0.00     0.002 A
ATOM     20  A18 INH I          -2.543   5.650  -2.328  0.00  0.00     0.019 A
BRANCH  15  21
ATOM     21  C21 INH I          -0.834   4.113  -1.388  0.00  0.00     0.210 C
ATOM     22  O1  INH I          -0.774   2.915  -1.581  0.00  0.00    -0.644 OA
ATOM     23  O3  INH I           0.298   4.828  -1.237  0.00  0.00    -0.644 OA
ENDBRANCH  15  21
ENDBRANCH   9  11
BRANCH   7  24
ATOM     24  C22 INH I          -3.749   1.535   0.125  0.00  0.00     0.210 C
ATOM     25  O2  INH I          -4.019   2.378  -0.708  0.00  0.00    -0.644 OA
ATOM     26  O4  INH I          -4.659   1.196   1.059  0.00  0.00    -0.644 OA
ENDBRANCH   7  24
TORSDOF 3
```

Как описывается атом в pdb

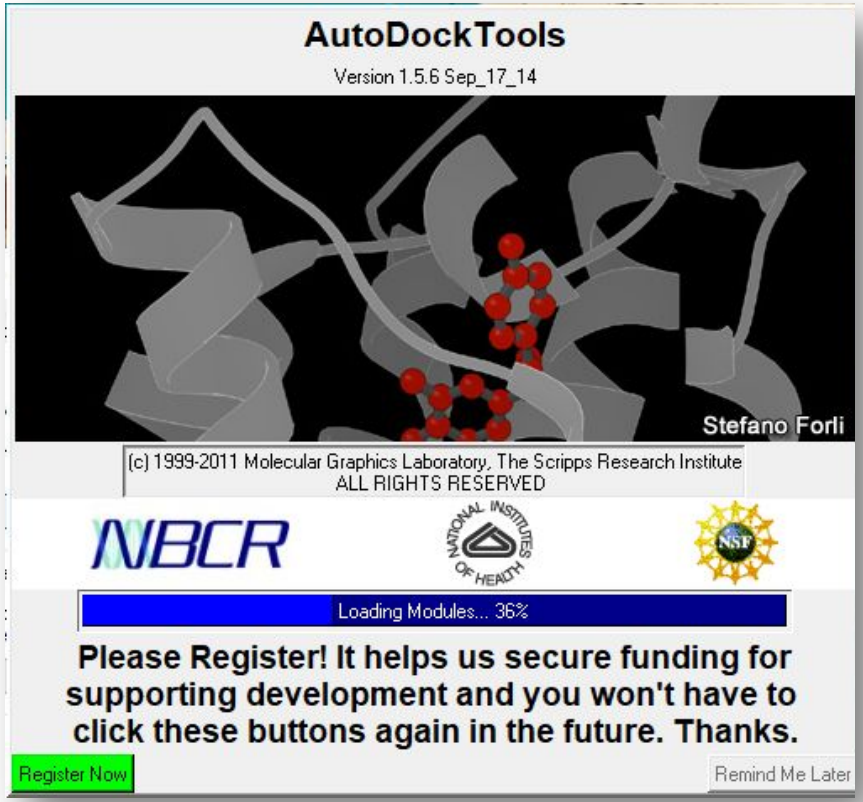
COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"ATOM "	
7 - 11	Integer	serial	Atom serial number.
13 - 16	Atom	name	Atom name.
17	Character	altLoc	Alternate location indicator.
18 - 20	Residue name	resName	Residue name.
22	Character	chainID	Chain identifier.
23 - 26	Integer	resSeq	Residue sequence number.
27	AChar	iCode	Code for insertion of residues.
31 - 38	Real(8.3)	x	Orthogonal coordinates for X in Angstroms.
39 - 46	Real(8.3)	y	Orthogonal coordinates for Y in Angstroms.
47 - 54	Real(8.3)	z	Orthogonal coordinates for Z in Angstroms.
55 - 60	Real(6.2)	occupancy	Occupancy.
61 - 66	Real(6.2)	tempFactor	Temperature factor.
77 - 78	LString(2)	element	Element symbol, right-justified.
79 - 80	LString(2)	charge	Charge on the atom.

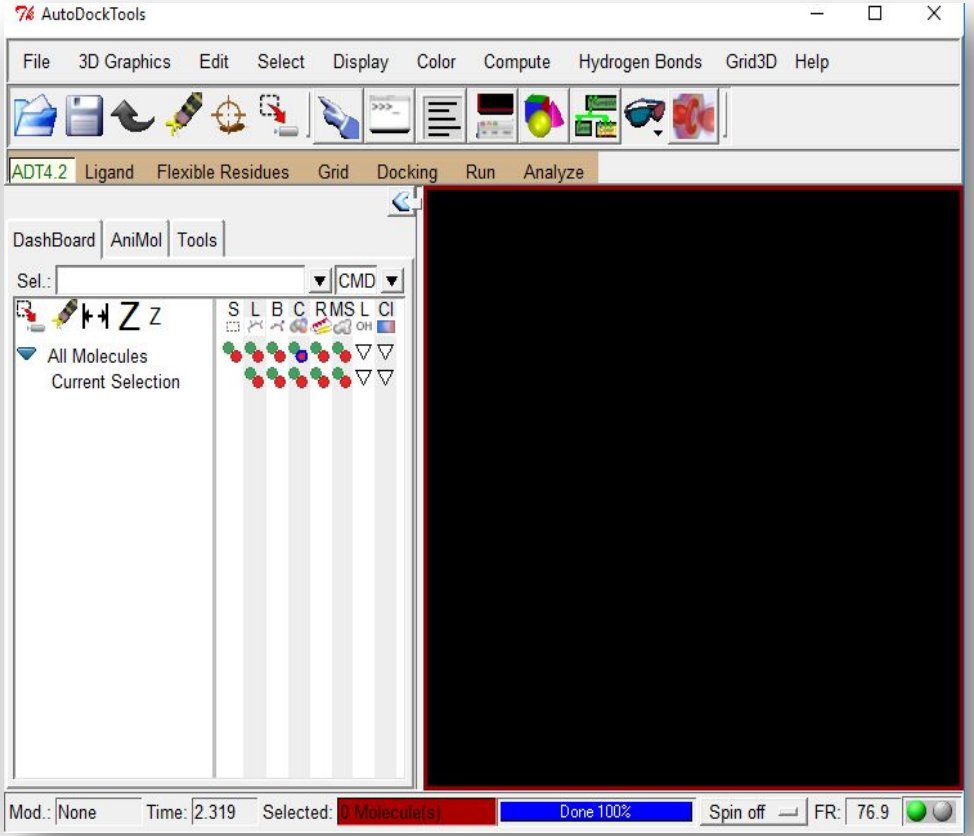
Где брать?

www.rcsb.org Состояние на 24.09.2020 21:50 UTC+3

Molecular Type	X-ray	NMR	EM	Multiple methods	Neutron	Other	Total
Protein (only)	132623	11464	4014	160	67	32	148360
Other	8009	92	467	6	0	4	8578
Protein/NA	7035	265	1425	3	0	0	8728
Nucleic acid (only)	2093	1302	47	6	2	1	3451
Total	149760	13123	5953	175	69	37	169117

AutoDock


AutoDockTools
Version 1.5.6 Sep_17_14
Stefano Forli
(c) 1999-2011 Molecular Graphics Laboratory, The Scripps Research Institute
ALL RIGHTS RESERVED
NBCR
NATIONAL INSTITUTES OF HEALTH
NSF
Loading Modules... 36%
Please Register! It helps us secure funding for supporting development and you won't have to click these buttons again in the future. Thanks.
Register Now
Remind Me Later


AutoDockTools
File
3D Graphics
Edit
Select
Display
Color
Compute
Hydrogen Bonds
Grid3D
Help
ADT4.2
Ligand
Flexible Residues
Grid
Docking
Run
Analyze
DashBoard
AniMol
Tools
Sel.:
CMD
S L B C RMS L Cl
All Molecules
Current Selection
Mod.: None
Time: 2.319
Selected: 0 Molecule(s)
Done 100%
Spin off
FR: 76.9

Определение

- **Кластер** — группа компьютеров, объединённых высокоскоростными каналами связи, представляющая с точки зрения пользователя единый аппаратный ресурс.
- **Кластер** - слабо связанная совокупность нескольких вычислительных систем, работающих совместно для выполнения общих приложений, и представляющихся пользователю единой системой.
- **Кластер** — это
  разновидность параллельной или распределённой системы, которая:
  состоит из нескольких связанных между собой компьютеров;
  используется как единый, унифицированный компьютерный ресурс».

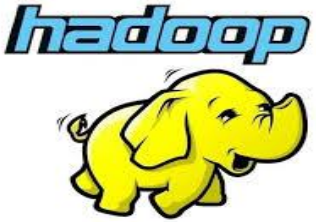
Что хочется

Техники работают с большим Linux кластером в Хемницком техническом университете, Германия

Что есть

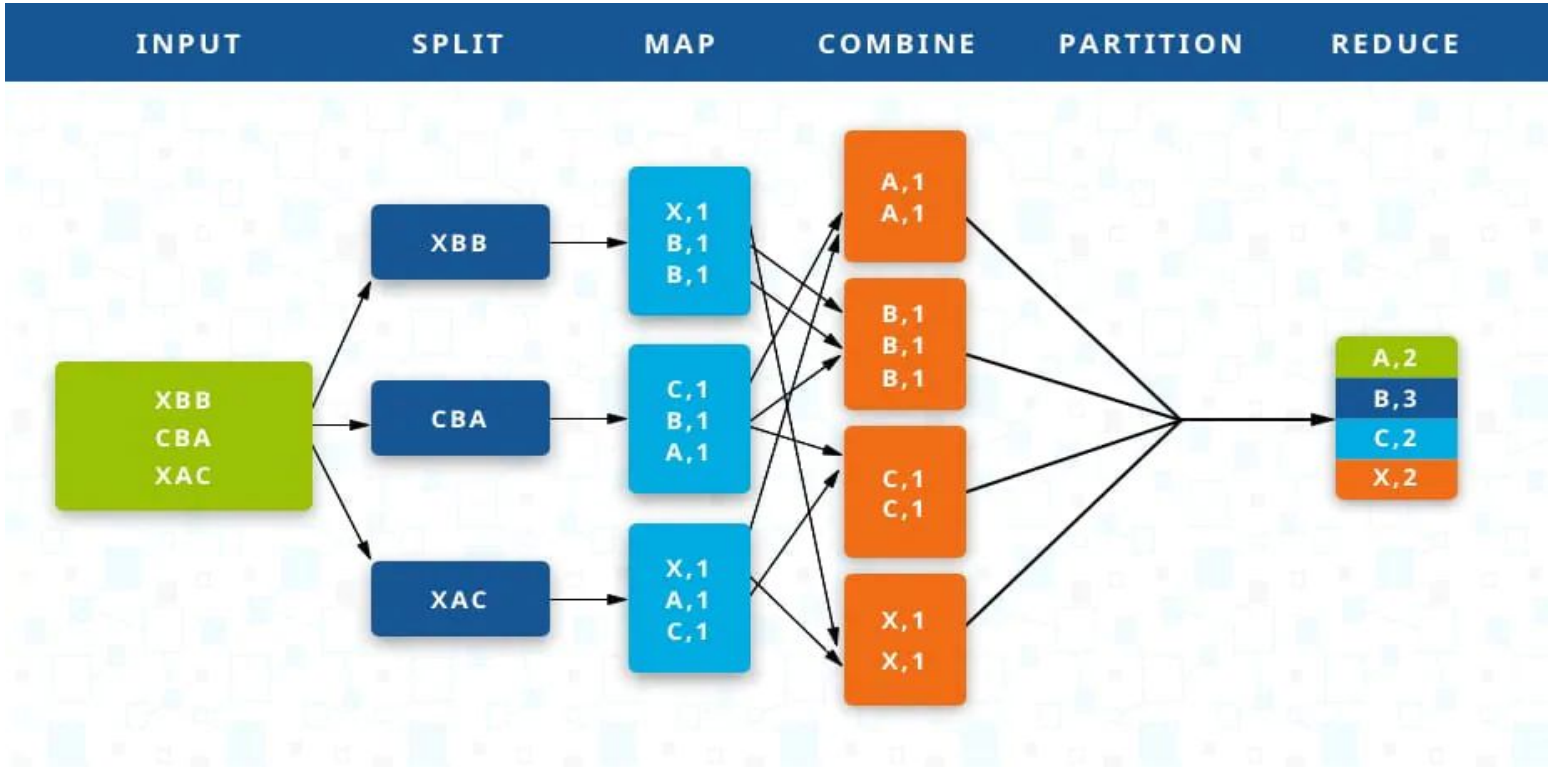
Оглянитесь вокруг ☹️

Hadoop

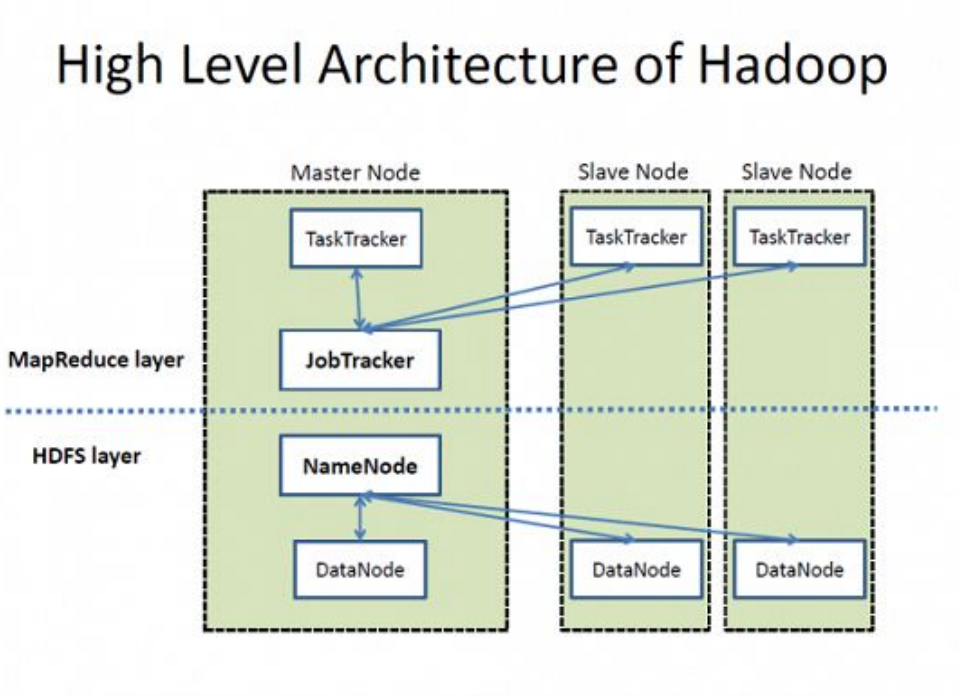


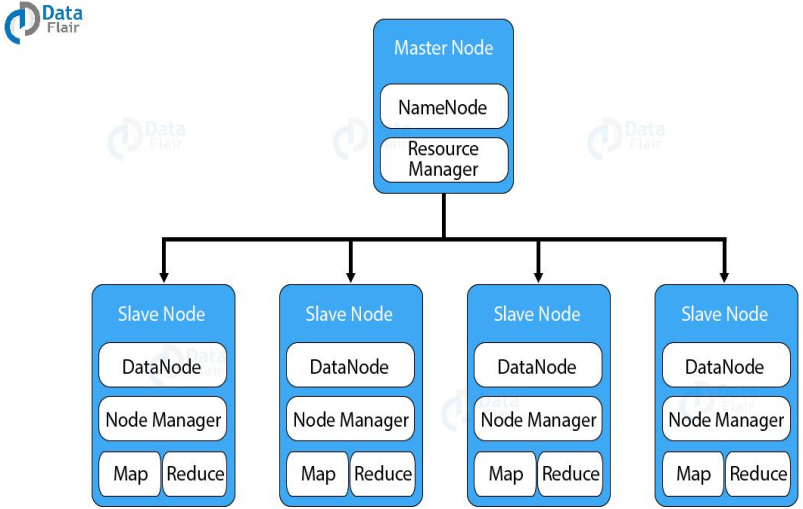
- **Hadoop** — проект фонда Apache Software Foundation, свободно распространяемый набор утилит, библиотек и фреймворк для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов. Разработан на *Java* в рамках вычислительной парадигмы *MapReduce*, согласно которой приложение разделяется на большое количество одинаковых элементарных заданий, выполнимых на узлах кластера и естественным образом сводимых в конечный результат.

MapReduce


INPUT
SPLIT
MAP
COMBINE
PARTITION
REDUCE
XBB
CBA
XAC
XBB
CBA
XAC
X,1
B,1
B,1
C,1
B,1
A,1
X,1
A,1
C,1
A,1
A,1
B,1
B,1
B,1
C,1
C,1
X,1
X,1
A,2
B,3
C,2
X,2

Архитектура кластера Hadoop





https://data-flair.training/blogs/hadoop-architecture/

Как я организовал процесс

```
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00003.pdbqt,STOCK1S-00003.gpf,STOCK1S-00003.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00004.pdbqt,STOCK1S-00004.gpf,STOCK1S-00004.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00005.pdbqt,STOCK1S-00005.gpf,STOCK1S-00005.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00007.pdbqt,STOCK1S-00007.gpf,STOCK1S-00007.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00008.pdbqt,STOCK1S-00008.gpf,STOCK1S-00008.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00009.pdbqt,STOCK1S-00009.gpf,STOCK1S-00009.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00010.pdbqt,STOCK1S-00010.gpf,STOCK1S-00010.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00011.pdbqt,STOCK1S-00011.gpf,STOCK1S-00011.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00012.pdbqt,STOCK1S-00012.gpf,STOCK1S-00012.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00013.pdbqt,STOCK1S-00013.gpf,STOCK1S-00013.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00018.pdbqt,STOCK1S-00018.gpf,STOCK1S-00018.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00021.pdbqt,STOCK1S-00021.gpf,STOCK1S-00021.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00023.pdbqt,STOCK1S-00023.gpf,STOCK1S-00023.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00024.pdbqt,STOCK1S-00024.gpf,STOCK1S-00024.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00026.pdbqt,STOCK1S-00026.gpf,STOCK1S-00026.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00027.pdbqt,STOCK1S-00027.gpf,STOCK1S-00027.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00028.pdbqt,STOCK1S-00028.gpf,STOCK1S-00028.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00029.pdbqt,STOCK1S-00029.gpf,STOCK1S-00029.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00032.pdbqt,STOCK1S-00032.gpf,STOCK1S-00032.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00033.pdbqt,STOCK1S-00033.gpf,STOCK1S-00033.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00034.pdbqt,STOCK1S-00034.gpf,STOCK1S-00034.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00035.pdbqt,STOCK1S-00035.gpf,STOCK1S-00035.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00036.pdbqt,STOCK1S-00036.gpf,STOCK1S-00036.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00037.pdbqt,STOCK1S-00037.gpf,STOCK1S-00037.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00038.pdbqt,STOCK1S-00038.gpf,STOCK1S-00038.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00039.pdbqt,STOCK1S-00039.gpf,STOCK1S-00039.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00040.pdbqt,STOCK1S-00040.gpf,STOCK1S-00040.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00042.pdbqt,STOCK1S-00042.gpf,STOCK1S-00042.gpf,1,,
/user/hduser,4an3_atp_mg.pdbqt,,STOCK1S-00045.pdbqt,STOCK1S-00045.gpf,STOCK1S-00045.gpf,1
```

Вход

Выход

```
0_20    /docking/TestLaunch/1a52_alpha_pdbqt_metciheston_pdbqt__60_20/1a52_alpha_pdbqt_metciheston_pdbqt__60_20_result.dlg
0_22    /docking/TestLaunch/1a52_alpha_pdbqt_metciheston_pdbqt__60_22/1a52_alpha_pdbqt_metciheston_pdbqt__60_22_result.dlg
0_24    /docking/TestLaunch/1a52_alpha_pdbqt_metciheston_pdbqt__60_24/1a52_alpha_pdbqt_metciheston_pdbqt__60_24_result.dlg
0_26    /docking/TestLaunch/1a52_alpha_pdbqt_metciheston_pdbqt__60_26/1a52_alpha_pdbqt_metciheston_pdbqt__60_26_result.dlg
0_28    /docking/TestLaunch/1a52_alpha_pdbqt_metciheston_pdbqt__60_28/1a52_alpha_pdbqt_metciheston_pdbqt__60_28_result.dlg
0_31    /docking/TestLaunch/1a52_alpha_pdbqt_metciheston_pdbqt__60_31/1a52_alpha_pdbqt_metciheston_pdbqt__60_31_result.dlg
0_33    /docking/TestLaunch/1a52_alpha_pdbqt_metciheston_pdbqt__60_33/1a52_alpha_pdbqt_metciheston_pdbqt__60_33_result.dlg
0_35    /docking/TestLaunch/1a52_alpha_pdbqt_metciheston_pdbqt__60_35/1a52_alpha_pdbqt_metciheston_pdbqt__60_35_result.dlg
```

Как я организовал процесс

Mapper

```
/**
 * @author SmirnygaTotoshka
 *
 */
public static class DockMapper extends MapReduceBase implements Mapper<LongWritable, Text, Text, Task>
{
    @Override
    public void configure(JobConf job) {
        super.configure(job);
        DockJob.configure(job);
    }

    @Override
    public void map(LongWritable arg0, Text arg1, OutputCollector<Text, Task> arg2, Reporter arg3) throws IOException
    {
        Dock dock = new ConfigParser().parse(arg1);
        Task[] tasks = dock.launch();
        Parameters.get().getLog().log(Level.INFO,dock.getDockingProperties().toString());
        for(int i = 0;i < tasks.length;i++)
        {
            Text key = new Text(arg0.toString() + "_" + Integer.toString(i));
            Parameters.get().getLog().log(Level.INFO,tasks[i].getVector().toString());
            arg2.collect(key,tasks[i]);
        }
        arg3.incrCounter(Counters.ALL,tasks.length);
    }
}
```

Reducer

```
/**
 * @author SmirnygaTotoshka
 *
 */
public static class DockReducer extends MapReduceBase implements Reducer<Text, Task, Text, Text>
{
    @Override
    public void configure(JobConf job) {
        super.configure(job);
        DockJob.configure(job);
    }

    @Override
    public void reduce(Text arg0, Iterator<Task> arg1, OutputCollector<Text, Text> arg2, Reporter arg3) throws IOException
    {
        while(arg1.hasNext())
        {
            Task task = arg1.next();
            String pathToDlg = task.launch();
            Parameters.get().getLog().log(Level.INFO,pathToDlg);
            arg2.collect(arg0, new Text(pathToDlg));
            if (pathToDlg.equals(""))
                arg3.incrCounter(Counters.BAD, 1);
            else
                arg3.incrCounter(Counters.GOOD, 1);
        }
    }
}
```