Parallel Supercomputer Docking Program of the New Generation: Finding Low Energy Minima Spectrum

Alexey Sulimov, Danil Kutov, Vladimir Sulimov(✉)

Dimonta, Ltd, Moscow 117186, Russia
{as,dk}@dimonta.com,vladimir.sulimov@gmail.com
Research Computer Center, Lomonosov Moscow State University, Moscow 119992, Russia

Abstract. The results of studies of the energy surfaces of the protein-ligand complexes carried out with the help of the FLM docking program belonging to the new generation of gridless docking programs are presented. It is demonstrated that the ability of the FLM docking program to find the global energy minimum is much higher than one of the "classical" SOL docking program using the genetic algorithm and the preliminary calculated grid of potentials of ligand atoms interactions with the target protein. The optimal number of FLM local optimization reliable finding of the global energy minimum and all local minima with energies in the 2 kcal/mol interval above the energy of the global minimum is found. This number is 250 thousand. For complexes with the ligand containing more than 60 atoms and having more than 12 torsions and with more than protein 4500 protein atoms the number of FLM local optimizations should be noticeably increased. There are several unique energy minima in this energy interval and for most complexes these minima are located near (RMSD < 3 Å) the global minimum. However, there a complexes where such minima are located far from the global minimum with RMSD (on all ligand atoms) > 5 Å.

Keywords: Generalized docking · Local optimization · Global minimum · Low-energy local minima spectrum · High-performance computing · Molecular modeling · Drug design.

1 Introduction

Discovery of new inhibitors of the protein associated with a given disease is the initial and most important stage of the whole process of the new pharmaceutical substances discovery [1, 2]. Computer-aided molecular modeling can considerably increase effectiveness of the new inhibitors design. Protein-ligand binding free energy calculation is one of the key problems of this molecular modeling. Docking is the popular molecular modeling method based on the search for the ligand binding pose in the target protein and the subsequent estimation of the protein-ligand binding free energy [3]. There are a lot of docking programs now [4-6]. However, accuracy of the binding energy prediction is still not high enough for the lead inhibitors optimization on the

base of such calculations [7]. There are a lot of sources of errors decreasing the accuracy of such calculations [8]: imperfections of the force field used, using too simplified solvent models or neglect solvent at all, an incomplete search of the best ligand poses, inadequate approximations made in construction of 3D atomistic models of the target protein and ligands, and, finally, simplifications which are used to accelerate the docking performance. The latter is the most harmful trend which oversimplifies many aspects of the complicated docking problem. Certainly this trend was essential and unavoidable at the dawn of docking programs development, 20-30 years ago. However now, when large supercomputer resources are available, we can return to the rigorous docking task formulation and try to solve this problem accurately.

One approximation is widely used in many docking programs and limits their accuracy strongly. To accelerate docking the preliminary calculated grid of potentials is used – the grid approximation. This grid contains usually in its nodes the potentials of non-bonded interactions (Coulomb and Van der Waals) of all possible types of ligand atoms with the protein. When docking, the energy of the ligand in any position in the active site of the protein is calculated as the sum of the grid potentials over all atoms of this ligand. This approach gives a large acceleration because all resource-intensive operations are performed at the grid calculation stage before the docking proper. This approach is used in AutoDock [9], ICM [10], Dock [11], SOL [12], and possibly in many other docking programs. However this approximation leads to several limitations resulting in docking inaccuracy [8]. First, it is impossible to perform accurately a local optimization of the energy of the protein-ligand complex varying either coordinates of only ligand atoms or ligand and protein atoms both. Second, implicit solvent models that describe nonlocal interactions of atom charges of a solute molecule with polarization charges on the solvent excluded surface (SES) cannot be sufficiently accurately reduced to predetermined local potentials at grid nodes; hence the contribution of the interaction with water into the protein-ligand energy cannot be described with the grid approximation accurately. Third, there are different fitting parameters (in addition to fixed force field parameters) in existing docking programs. These parameters help to adjust the docking results to crystallized ligand poses in target proteins and to reproduce in calculations binding constants obtained in experiments for a training set of protein-ligand complexes. Fitting parameters serve to demonstrate a semblance of high accuracy of calculations. However, utilization of fitting parameters obscure the reasons of bad docking accuracy for new protein-ligand complexes which are different from ones in the training set.

So, docking programs of new generation possessing a heightened accuracy should not use the preliminary calculated grid of potentials and any fitting parameters; in the process of ligand positioning the energy of the protein-ligand complex should be calculated in the frame of a given force field or a quantum-chemical method without simplifications and fitting parameters.

Another problem of the accurate docking is connected with the ligand positioning procedure. Until now in some popular docking programs the procedure “lock-and-key” is used for ligand positioning. In the frame of this procedure the ligand should be embedded into the protein active site in a manner as a key is inserted into the lock: the ligand molecular surface must complement the active site surface. In another ligand positioning procedures some prepositioning points in the active site are used for placing near them ligand atoms of definite types, e.g. for formation hydrogen bonds.

However, the most general ligand positioning procedure is based on the so-called docking-paradigm [13-15]. This paradigm assumes that the ligand binding pose in the active site of the target protein corresponds to the global minimum of the protein-ligand energy function or is near it. In accordance with this paradigm the docking problem is reduced to the search of the global minimum on the multi-dimensional protein-ligand energy surface the dimensionality of which is defined by the number of protein-ligand system degrees of freedom.

Several supercomputer docking programs of this type (no preliminary calculated grid of potentials, no fitting parameters, and the docking algorithm is based on the global energy minimum search) were developed recently: FLM [13] and SOL-T [14] for docking flexible ligands into rigid proteins and SOL-P [16, 17] which is able to dock flexible ligands into the protein with moveable atoms as well as into the rigid protein. FLM performs the massive multiprocessor exhaustive search for low energy minima of the protein-ligand complex in the rigid protein model. The protein-ligand system energy is calculated in the frame of the MMFF94 force filed either in vacuum or with the rigorous implicit solvent model. Each local energy optimization is carried out from the random position of the ligand in the specified region of the active site of the protein. SOL-T and SOL-P are much faster than FLM due to the new TT-docking algorithm based on the tensor train (TT) decomposition of multi-dimensional tensors and the TT-Cross approximation and the respective global optimization method. However, if enough supercomputing resources are available, FLM will perform the complete low energy minima search thoroughly. The reliability of the global minimum finding and the fullness of the low energy minima identification, especially those belonging to the narrow energy interval of several kcal/mol above the global minimum, define high docking accuracy. Moreover, such a carefully found low energy minima can be used for the validation of other docking algorithms. Experience of practical use of the FLM program and comparison of energy minima found by this program with minima found by other programs, such as SOL-T and SOL-P, compel us to investigate more accurate performance of the FLM program. Also the conception of the quasi-docking approach [18, 19] has been introduced recently, and it is essentially based on the completeness of the whole low energy minima spectrum found by the FLM program. In this work we present the results of studies of the energy surfaces of the protein-ligand complexes carried out with the help of the FLM docking program. The employment of supercomputing resources of Lomonosov Moscow State University made it possible to conduct unprecedentedly detailed studies of the low-energy minima spectra of the test set of protein-ligand complexes containing various proteins and different ligands. Optimal computing resources are determined which are necessary for reliable finding of the global minimum and local energy minima with their energy values near to the energy of the global minimum. Realizability of the docking paradigm and the optimal choice of the size of low energy minima spectrum for the quasi-docking procedure are investigated. Comparison of FLM performance with one of the classical SOL docking program which uses a preliminary calculated grid of potentials is made and the advantage of the FLM program is demonstrated.

2 Materials and Methods

Due to thermal motion in the thermodynamic equilibrium state the ligand continuously jumps from one binding pose to another, and for the binding energy estimation we should find not only the global minimum of the energy of the protein-ligand system where the ligand spends most time but also the low-energy part of the whole local energy minima spectrum. The landscape of the multi-dimensional protein-ligand energy surface is very complicated containing hundreds and thousands of local minima even when positions of the ligand center are limited within the spatially restricted area of the protein active site. Parallel multi-processor performance of the minima search program and available supercomputer resources make it possible to solve practically the complicated problem of determination of all low energy minima on this complicated energy surface.

2.1 FLM Program

The FLM program is the MPI-based program which was developed to find low-energy minima of the ligand-protein system [13]. During the minima search, the protein is considered as rigid and the ligand is fully flexible. The name of the program is the abbreviation of its main function: Find Local Minima. We describe and use here the version FLM-0.05 in which the MMFF94 force field is implemented, and the energy of any configuration of the protein-ligand system is calculated in the frame of this force field in vacuum without simplifications and approximations. There are two different options of the FLM performance: (i) the search for low energy minima of the protein-ligand system and (ii) the search for low energy minima of the free (unbound) ligand.

FLM performs the massive multi-processor search for low energy minima of the energy function of a protein-ligand complex. Each local energy optimization is carried out from the random position of the ligand in the specified region of the active site of the protein. These random positions are obtained by random throws of the ligand with continuous deformations of the ligand when changing its torsion angles (describing the internal rotation around a single acyclic bond of the ligand) and translations and rotations of the ligand as a whole rigid body.

- The ligand geometrical center (the center of gravity when all atomic masses are equal) is moved to a random point in the search area. The search area is defined as the sphere of a given radius R_{in} with the center at the ligand native position geometrical center. The ligand native position is the position of the ligand in the crystallized protein-ligand complex structure. The present investigations are conducted when the radius is equal to $R_{in} = 8$ Å. This sphere covers active sites of all test protein-ligand complexes. There is also a larger sphere with the radius R_{grid} which center is in the same point as the smaller sphere. The whole ligand should be inside this larger sphere during the docking process. In the present study $R_{grid} = 24$ Å.
- The ligand is rotated as a whole by a random angle from the interval $[-\pi, \pi]$ around a random axis passing through the ligand center.

- The ligand torsions are rotated by random angles from the interval $[-\pi, \pi]$.

After each random throw not all random system conformations are further optimized. At first, atom-atom distances are checked: atoms from each ligand-ligand or protein-ligand atom pair must be separated by more than 0.5 Å. Otherwise this random system conformation is rejected. For the acceleration of the checkup of the existence of such protein-ligand atoms clashes a special 3D array is constructed covering the active site region with sufficiently fine grid (the grid step is 0.1 Å). The special region where this grid is created is the sphere with the radius R_{grid}. Each cell of the grid contains an indication of the presence or absence of a protein atom. So, any random initial pose of the ligand in the active site (inside the sphere with the radius R_{grid}) is accepted for the further local optimization if no atom of the ligand finds itself in the cell with an indicator of the presence of the protein atom.

Local optimization is performed using the L-BFGS gradient algorithm [20, 21] without any restrictions on the positions of the ligand atoms in the search area. All Cartesian coordinates of ligand atoms move during optimization. Each local optimization stops when the maximal component of the gradient of the optimized energy function decreases down to the value $10^{-5}\ kcal\ mol/$Å. The gradient of the energy function is calculated numerically using 6 points and the step of numerical differentiation is equal to 10^{-8} Å. If the ligand center moves out of the search area after the optimization (out from the sphere of the radius R_{in}), the respective local minimum is rejected. We call each accepted local optimization the trial or test optimization.

The local minima search is parallelized: independent local optimizations from different initial ligand conformations are continuously performed in parallel by different MPI processes. The optimization results are collected in the master process to form the low-energy minima set. The current collected minima set is repeatedly sent back from the master process to other processes, so other processes can select only promising minima to send. This results in the good scalability of the program with an increase in the number of computing cores. The program works for a specified time on the specified number of computing cores. If we continue such FLM calculations sufficiently long time at sufficiently large number of computing cores, we'll certainly find the global energy minimum and also all low energy minima above the global one in the given energy interval or a given number of lowest in energy minima. One of the questions we address in this study: how many trial optimizations should be done for the reliable finding the global minimum, and/or the given number of lowest energy minima? The answer to this question is presented below in the section Results.

A set of found unique local minima with the lowest potential energies is being kept in operative memory during FLM calculations. A new computed local minimum is included into the set, if it differs from any minimum of the set, and the minimum with the highest energy is excluded from this set. Two minima are different if RMSD between them exceeds a given value, the uniqueness parameter, e.g. 0.1 or 0.2 Å. The RMSD is calculated over the ligand heavy atoms without taking into account possible chemical symmetry. Obviously, the larger the uniqueness parameter the lesser number of the unique minima will be collected in the low energy minima set.

FLM can save different numbers of the unique low energy minima. In the present study 8192 (2^{13}) unique low energy minima are saved for each protein-ligand complex from the respective test set (see below).

After finishing the FLM program performance the FLM-PP postprocessing programs starts. FLM-PP conducts the more accurate analysis of the uniqueness of found minima taking into account ligand chemical symmetry. It calculates RMSD values between all minima in respect with all ligand atoms. As a result of FLM-PP performance the only unique minima from the whole pool of the minima found by the FLM program are kept. The minima are considered different from each other when the RMSD value is larger than a given distance, for example 0.2 Å, and the difference of their energies is larger than a given value, for example 1 kcal/mol. FLM-PP can also perform local energy optimization with the L-BFGS method and calculate the system energy in solvent – in one of the two implicit solvent models: PCM and SGB in the points corresponding to the minima.

The explanation of saving so large number of low energy minima is closely related to the quasi-docking procedure [8, 18, 19] which is as follows. Suppose that all low energy minima are found for a given protein-ligand complex in the frame of the given force field. In respect with the docking paradigm the global energy minimum should be found near the crystallized native ligand pose. However our analysis reveals [13] that this is not true for many complexes in the frame of the MMFF94 force field in vacuum. This force field does not describe adequately the energy of many protein-ligand complexes and the docking paradigm is not satisfied for these complexes. So, we should use better force field or quantum chemical methods for the energy calculation. However, is it possible to avoid the time-consuming search for the low energy minima spectrum with a new energy function? Yes, it is possible, and the quasi-docking procedure is proposed [18, 19]. All low energy minima found in a given force field are recalculated in another force field or quantum-chemical methods. Each minimum is used as the initial configuration for local energy optimization. Positions of the minima (poses of the ligand) change insignificantly but the energies of the minima can change very strongly and the minimum with high energy in the initial force field can become the global minimum in the new method of energy calculations. This quasi-docking procedure has already demonstrated that for docking the CHARMM force field is much better than MMFF94, and the PM7 quantum-chemical method is better than CHARMM [18, 19]. All these methods should be used with the respective implicit solvent models.

2.2 Test Sets of Protein-ligand Complexes

The test set of 16 protein-ligand complexes was chosen from the Protein Data Bank (PDB) [22]. These structures have been chosen due to good resolution and the broad range of ligands different size and flexibility (Table 1). Protein structures were prepared by the elimination of all “HETATM” records, i.e. the records corresponding to water molecules, atoms, ions, and molecules which are not part of the protein structure, from the PDB files of the complexes, and then hydrogen atoms were added to the protein structures by the original APLITE program [12]. Although some complexes have been crystallized in low acidic conditions (pH = 5.2–6.5), all test proteins are active at the neutral conditions, and the APLITE program adds hydrogen atoms according to the standard amino acid protonation states at pH = 7.4. The histidine protonation state is chosen by comparing of electrochemical potentials for hydrogen atoms

at “HD1” and “HE2” positions. Optimization of hydrogen atoms positions is performed with MMFF94 force field after the hydrogen atoms pre-placement. During this optimization, all rotation variants of torsionally moveable hydrogen atoms, e.g.,a hydroxyl hydrogen atom from tyrosine, are tested. Ligands were also taken from the PDB files. Hydrogen atoms were added to the ligands by the Avogadro program [22].

Table 1. Test set of 16 protein ligand complexes. PDB ID is the identifier which is assigned to the respective structure of the protein-ligand complex in the Protein Data Bank [22]. The table contains information about the numbers of ligand atoms and torsions, ligand charges and numbers of protein atoms. Numbers of atoms include hydrogen atoms. uPA – urokinase-type plasminogen activator, CHK1 – checkpoint kinase 1, ERK2 – extracellular signal-regulated kinase 2.

Protein name	PDB ID	Number of ligand atoms	Number of ligand torsions	Ligand charges	Number of protein atoms
uPA	1C5Y	20	2	1	3869
	1F5L	24	6	1	3823
	1O3P	46	6	1	3839
	1SQO	34	4	1	3823
	1VJ9	74	19	1	3859
	1VJA	61	17	1	3858
thrombin	1DWC	71	12	0	4494
	1TOM	64	10	2	4455
Factor Xa	2P94	60	7	0	3676
	3CEN	50	7	0	3676
CHK1	4FSW	26	0	0	4342
	4FT0	42	3	-1	4255
	4FT9	32	5	0	4394
	4FTA	35	6	-1	4336
ERK2	4FV5	52	8	0	5414
	4FV6	57	12	0	5449

Also, the locally optimized ligand native position has RMSD from the original native pose less than 1.5 Å for all 16 test complexes, both for the optimization with MMFF94 in vacuum and for the optimization with the PM7 method in vacuum.

2.3 Minima Indexes

All local energy minima of a given protein-ligand complex for a given energy function can be sorted by their energies in the ascending order; that is, every minimum gets its own index equal to its number in this sorted list of minima. The lowest energy minimum has index equal to 1. We introduce a special index [13, 14, 19] to analyze the docking positioning accuracy and the feasibility of the docking paradigm as fol-

lows. The list of minima can include some minima corresponding to ligand positions located near the nonoptimized native (crystallized) ligand pose in the given crystallized protein-ligand complex structure taken from the Protein Data Bank [22]. By our definition the ligand is near the nonoptimized native ligand position if the RMSD, the root-mean-square deviation between equivalent atoms of the ligand in the two positions, is less than 2 Å. Let us designate the index of such minimum which is close to the native (crystallized) ligand position as INN. It is the abbreviation of the term "Index of Near Native". If there are several such minima, we attribute INN to the minimum with the lowest energy among all minima which are close to the native ligand pose.

How can this index be used to analyze the docking positioning accuracy? The docking program performs on the base of the docking paradigm: the best position of the ligand found in the docking procedure is the global energy minimum. So, if INN is equal to 1, then the best ligand position corresponding to the global energy minimum is situated near the experimentally defined ligand pose. Therefore the positioning accuracy of docking with different methods of protein-ligand energy calculation can be compared by the simple comparison of INN index: the closer INN to 1 the better positioning accuracy is observed.

3 Results

3.1 Effectiveness of the Global Energy Minimum Search by SOL, FLM, SOL-P Programs

The comparison of the performance of different docking programs is a non-trivial task because there are several features reflecting quality of the performance: the positioning accuracy, the accuracy of the binding energy calculation, effectiveness of the global energy minimum finding, effectiveness of all low energy minima finding in a given energy interval above the global minimum, etc. The first two features are closely connected with experimentally measured values and they depend on models of proteins and ligands, and on the method of the protein-ligand energy calculation – the choice of the force field or the quantum-chemical method. The third feature reflects better the performance of the docking algorithm. So, we compare the ability of the docking programs to find the global energy minimum in the frame of the given force field. The comparison of the docking programs of the new generation (FLM and SOL-P) and the "classical" SOL docking program is made for 16 testing protein-ligand complexes. SOL does not use the local energy optimization. That is why the additional treatment of the ligand poses found by SOL is made as follows. For each protein-ligand complex all 99 poses, which are found by SOL in 99 independent runs of the genetic algorithm, are locally optimized. The protein atoms are kept fixed and the energy of the protein-ligand complex is optimized in the frame of MMFF94 force field in vacuum with variations of Cartesian coordinates of all ligand atoms. The local optimization method is the same as one which is implemented in FLM and SOL-P programs (L-BFGS), and the same accuracy of the optimization is taken for all these cases. All local energy minima, which are found in these optimizations, are subjected

to filtering of unique minima by the FLM-PP program with the same uniqueness 0.2. Difference ΔE_{GM} between the energy of the global minimum found by this procedure and the energy of the global minimum found by programs of the new generation (FLM and SOL-P) is presented in Table 2 for all 16 test protein ligand complexes.

Table 2. Energies of the global minima found by FLM and SOL-P programs for 16 test complexes. The energies are presented relatively the energy of the global minimum found in the procedure involving docking with the "classical" SOL docking program.

PDB ID	ΔE_{GM} FLM, kcal/mol	ΔE_{GM} SOL-P, kcal/mol	PDB ID	ΔE_{GM} FLM, kcal/mol	ΔE_{GM} SOL-P, kcal/mol
1C5Y	0.00	-0.08	2P94	-2.75	-0.92
1DWC	-67.56	-67.56	3CEN	-2.19	5.59
1F5L	-1.16	-1.16	4FSW	-12.04	-12.04
1O3P	-5.13	-5.11	4FT0	-24.80	-26.57
1SQO	-0.11	-0.11	4FT9	-18.31	-18.31
1TOM	-5.18	-1.13	4FTA	-17.53	-17.53
1VJ9	-5.55	-3.04	4FV5	-36.65	-21.03
1VJA	-5.33	-2.93	4FV6	-16.93	-7.39

We can see in Table 2 that the energy of the global minimum which is found by FLM is lower than or equal to the energy of the global minimum found in the procedure with SOL usage for all test protein-ligand complexes. Only for one complex (1C5Y) energies of SOL and FLM global minima are the same. For all other 15 complexes the SOL related procedure cannot find global energy minima and it is obvious that the effectiveness of finding of the global minimum by the SOL related procedure is much lower than one demonstrated by the FLM program. We see in Table 2 that SOL-P is worse than FLM in finding of the global minimum but the former is also better than the performance of the SOL related procedure.

3.2 Ligand Internal Strain Energy

The FLM program can work not only in the search mode of the low energy minima spectrum of the protein-ligand complex with a rigid protein, but also it can find the spectrum of low energy (actually all) minima of the free ligand. The latter is necessary in calculating the binding energy of the protein-ligand complex in the multiwell approximation [8, 13]. In this case, an important contribution in the binding energy is the ligand internal strain energy, which is calculated as the difference between the energy of the free ligand in its conformation in the bound state in the protein and the energy of the free ligand in its the conformation corresponding to the global minimum of the free (unbound) ligand. The physical meaning of this quantity is simple – in order for the ligand to "get into" the active center of the target protein, it often needs to change its configuration in comparison with its configuration in the unbound state, i.e. when the ligand conformation corresponds to the global energy minimum of the

free ligand. This change of the ligand conformation requires additional energy cost which is the ligand internal strain energy. The strain energy works against protein-ligand binding. The energy of the ligand internal strain is not taken into account in most of docking programs. However, this energy can be quite large, as can be seen from Table 3, in which the internal strains of native ligands are given for 16 test complexes.

Table 3. The internal strain energy E_{strain} of native ligands for 16 test complexes.

PDB ID	E_{strain}, kcal/mol	PDB ID	E_{strain}, kcal/mol
1C5Y	4.59	2P94	23.34
1DWC	94.83	3CEN	20.26
1F5L	5.30	4FSW	1.75
1O3P	19.85	4FT0	9.74
1SQO	11.54	4FT9	18.62
1TOM	21.10	4FTA	10.39
1VJ9	47.53	4FV5	30.02
1VJA	47.18	4FV6	25.77

As we can see in Table 3 that the ligand internal strain energy can be sufficiently large: from several units to several dozen of kcal/mol, and it should be taken into account when the protein-ligand binding energy is calculated. This means that taking into account the energy of the ligand in its global minimum in the unbound state is extremely important, and FLM does this in the MMFF94 force field.

3.3 Finding the Global Minimum and Local Ones Above It in the 2 kcal/mol Interval

Good quality of low energy minima finding means that not only the global minimum is found but also all minima with energies in the given energy interval are found as well. It is reasonably to choose the value of this interval equal to 2 kcal/mol because only these minima will be occupied at room temperature. It is found that for the most part of the test complexes the last update of the energy of the global minimum occurs quickly – after about first 5 thousand local optimizations. The largest number of local optimization (97 thousand) is needed for finding the global energy minimum if the 1VJA complex which contains a sufficiently large and flexible ligand: 61 atoms and 17 torsions (see Table 1).

And how many local optimizations are needed for the reliable finding of all local minima with energies in the interval 2 kcal/mol above the energy of the global minimum? To answer this question we investigated the dependence of such unique minima as a function of the number of FLM performed local optimizations. The results show that for most of test complexes it is sufficient to perform 250 thousand local optimization for finding of all minima with energies in 2 kcal/mol above the energy of the global minimum (see Fig. 1). Additional minima appear in this energy interval

after more than 250 thousand local optimizations only in 2 sufficiently larger protein-ligand complexes: 1DWC and 1VJA (see Fig. 1).

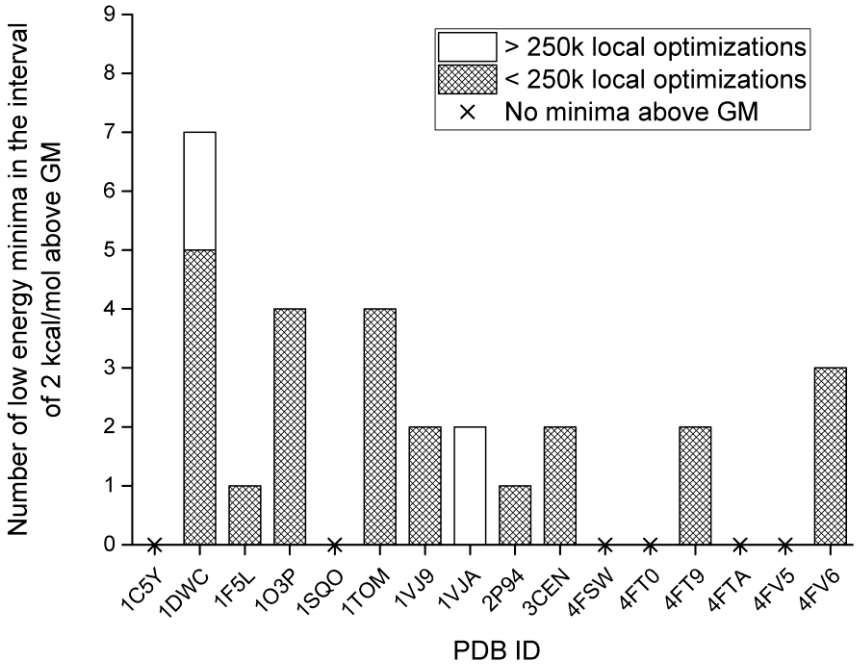

Fig. 1. Number of low energy minima in the interval of 2 kcal/mol above the global minimum depending on the number local optimizations: less than 250 thousand of local optimizations or more than 250 thousand of local optimizations. The symbol "×" on the horizontal axis marks the complexes with no energy minima in the 2 kcal/interval above the energy of the global minimum.

It should be noted that there are no energy minima in the 2 kcal/mol interval above the global one in several (in 6 of 16) complexes. Such complexes are marked by symbol "×" on the horizontal axis in Fig.1. When calculating the protein-ligand binding energy in such complexes in the multiharmonic approximation, it is sufficient to take into account the characteristics of only the global minimum.

And how these low energy minima are located in space in respect to the position of the global energy minimum? This is shown in Fig. 2 where for each complex the value of the random mean square deviation (RMSD) between poses of ligand corresponding to these minima and the ligand pose corresponding to the global energy minimum. RMSD is calculated on positions of respective ligand atoms in these poses.

12

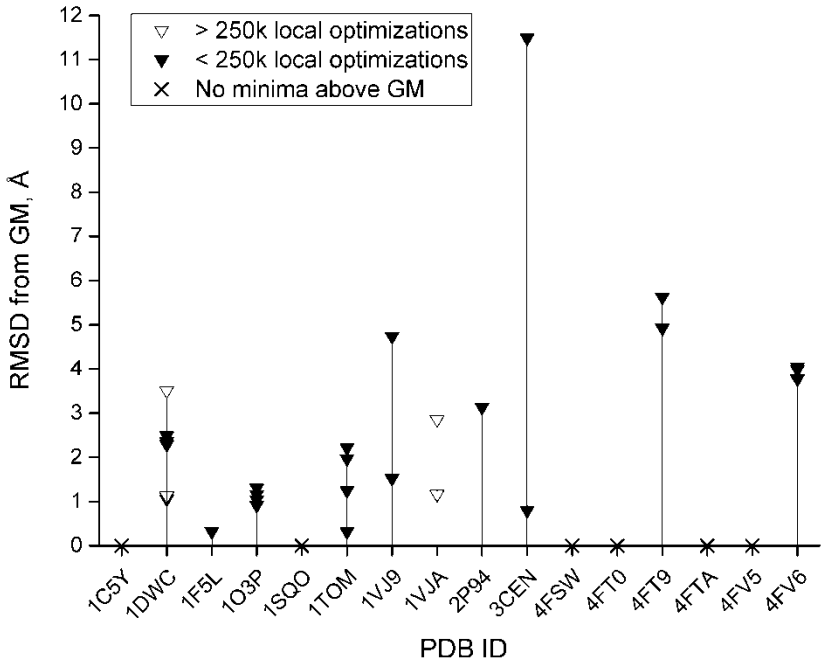

Fig. 2. The deviation RMSD of the ligand poses corresponding to the low energy minima with energies in the interval of 2 kcal/mol above the global minimum from the ligand pose in the global energy minimum.

As we see, despite the fact that a significant number of minima are relatively closely located in space from the global minimum (RMSD < 3 Å), in some complexes ligand poses in the low energy minima (in the 2 kcal/mol interval above the global minimum) can be strongly different on RMSD from the ligand pose corresponding to the global minimum: RMSD can be larger than 5 or 10 Å.

3.4 Optimizing the FLM Performance Time

The FLM program will carry out infinitely many local optimizations if of course unrestricted supercomputer resources are available. FLM stops in respect with the specified the computing time parameter. The longer FLM works the more local optimizations are performed and a larger number of unique minima can be found in the pool of saved low energy minima. To define the optimum time of FLM calculations it is necessary to find the number of fulfilled local optimization after which the pool of saved minima is near the saturation for all test complexes. Let the function $f(x)$ is the dependence of the number of the pool updates on the number x of performed local optimizations. Our observations show that for larger size of the pool of saved low energy minima the number of pool updates is larger. So, it is convenient to normalize the

number of pool updates on the size of the pool. The derivative of the normalized number of pool updates by the number of local optimizations displays the rate $D\ x$ of updates of saved minima in the pool:

$$D\ x\ = \frac{f'(x)}{N} \approx \frac{1}{N} \lim_{\Delta x \to 1} \frac{f\ x_0 + \Delta x\ - f\ x_0}{\Delta x} \quad (1)$$

Here x_0 is the given number of local energy optimizations, Δx is the change of the number of local energy optimizations, N is the number of the saved minima in the pool, i.e. it is the pool size.

While the rate of pool updates $D\ x$ is sufficiently high, the low energy minima search must be conducted. As soon as the pool updates rate falls down it is safe to complete the search.

The plot of $D\ x$ as the function of the number of local optimization is shown in Fig.3 for three typical complexes. The calculation of the derivative is made with a step of $\Delta x = 10^3$ local optimizations.

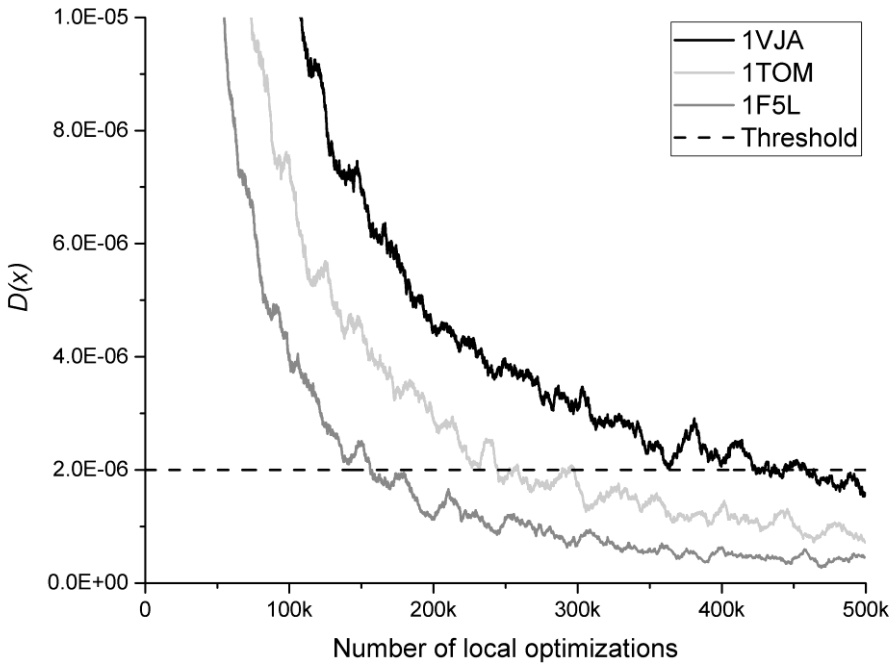

Fig. 3. The normalized derivative $D\ x$ of the number of updates of the low energy minima pool by the number of local optimizations as a function of the number of local optimizations x for three protein-ligand complexes with PDB ID: 1VJA, 1TOM, and 1F5L.

Curves of the rate of pool updates $D\ x$ as functions of the number of conducted local optimizations x for most of test protein-ligand complexes are between the curves for 1F5L and 1TOM. As it is shown above (see Section 3.3), for these com-

plexes it is sufficient to perform 250 thousand of local optimizations for finding of all low energy minima in the energy interval 2 kcal/mol above the energy of the global minimum. A further search does not result in finding of new minima with energies in this 2 kcal/mol energy interval above the energy of the global minimum. This means that the pool of low energy minima is saturated, and further search will not practically result in finding of additional minima. Thus, it is possible to find the threshold of the saturation of the pool of low energy minima. It is practically reasonable to finish the search below this threshold. The threshold of 2×10^{-6} is chosen using curves in Fig. 3.

Curves of the rate of pool updates $D\ x$ for all four complicated complexes (1DWC, 1VJ9, 1VJA и 4FV6) are near the curve for 1VJA. To reach the saturation of the pool of saved minima for the complicated complexes it is necessary to spend many more local optimizations, about 450 thousand.

The time expenses of the FLM performance for each of 16 test complexes are shown in Fig. 4. The grey histogram bar show the time of the global minimum search. If there is no a grey histogram bar for a complex, this means than less than 150 CPU·hours (about 5000 local optimizations) are spent for the global minimum search. The shaded histogram bars show the time of conducting local optimizations while the rate of pool updates is larger than the threshold (2×10^{-6}). Unshaded histogram bars show the time which is spent during the presented supercomputer investigations.

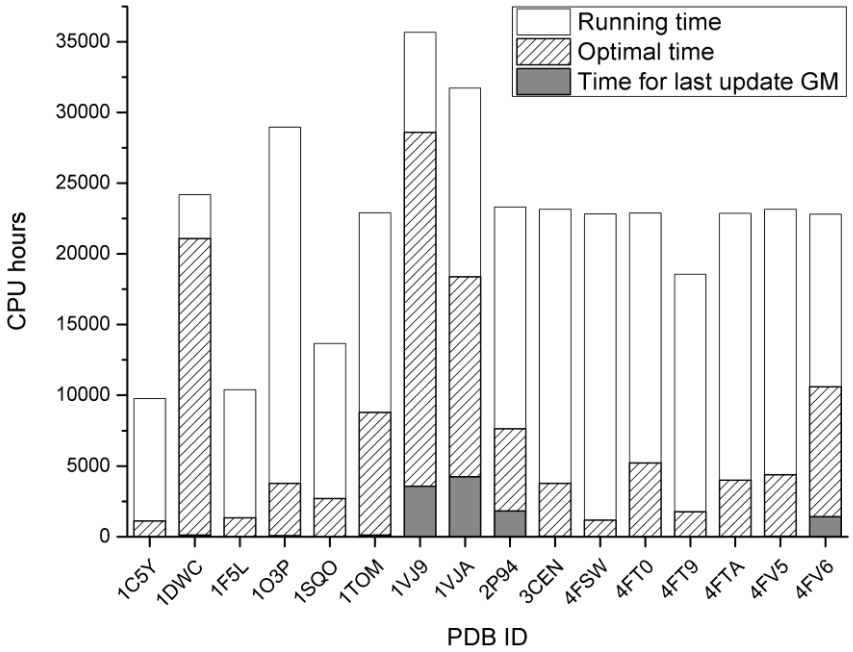

Fig. 4. Optimal time of the FLM run which is needed to reach the saturation of the pool of saved low energy minima.

We can see in Fig. 4, that the time of FLM performance is enough to reach the pool saturation for all test complexes. Also the time expended to find the global minimum is much less than the time to find all low energy minima in the pool of specified size. The longest time is needed to reach the saturation for the "difficult" complexes: 1DWC, 1VJ9, 1VJA и 4FV6. These complexes contain either a larger and flexible ligand (1DWC, 1VJ9, 1VJA) or a flexible ligand and a protein with large number of atoms (4FV6) as it can be seen in Table 1.

4 Conclusions

The performance of the docking program of new generation, the FLM program, is investigated. Some results of studies of the energy surfaces of the protein-ligand complexes with this program are presented. Comparison of performance of docking programs of new generation, FLM and SOL-P, with the classical SOL docking program is made. It is demonstrated for all test complexes that these docking programs of the new generation find the global minimum with considerably lower energy than the energy of the global minimum which is found by the classical SOL program in the same conditions: the same energy function, the same local optimization method, the same accuracy of the optimization, and the same uniqueness parameter for the filtering of only unique minima.

It is shown that the ligand internal strain energy can be as high as several dozen of kcal/mol depending on the protein-ligand complex. To take this large contribution into account when calculating the protein-ligand binding energy FLM finds the global energy minimum of the unbound ligand. This minimum corresponds to the most stable ligand conformation in the free (unbound) state of the ligand.

The optimal number of FLM local optimizations for reliable finding of low energy minima for most of test complexes is found. This number is determined by a threshold of the normalized rate of updates of the pool of low energy minima. If the rate of updates of the pool decreases below the threshold, FLM can finish its performance because further calculations do not result in additional low energy minima finding. For finding the global energy minimum and all local minima with their energies in the 2 kcal/mol interval above the energy of the global minimum the number of FLM local optimizations is equal to 250 thousand for most of test complexes. For complexes with the ligand containing more than 60 atoms and having more than 12 torsions and with more than protein 4500 protein atoms the number of FLM local optimizations should be increased up to about 450 thousand.

For all investigated protein-ligand test complexes the number of unique minima with energies in the 2 kcal/mol interval above the energy of the global minimum is less than 10 minima. For several complexes there are no local minima with energies in this interval. Most of these minima are located near the global energy minimum: the respective deviation (RMSD on ligand atoms) of ligand poses corresponding to these minima from the ligand pose of the global energy minimum is less than RMSD < 3 Å. However, some of the low energy minima can correspond to ligand poses far from the pose of the global minimum (RMSD > 5 Å).

It is shown that for all complexes the time of the global energy minimum finding is considerably smaller than the time of finding of the whole low energy minima spec-

trum. Supercomputer resources needed for the reliable determination of all low energy minima for most of test complexes are less than 10000 CPU·hours.

Acknowledgements. The work was financially supported by the Russian Science Foundation, Agreement no. 15-11-00025-П. The research is carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University, including the Lomonosov supercomputer [24].

References

1. Sliwoski, G., Kothiwale, S., Meiler, J., Lowe, E.W.: Computational Methods in Drug Discovery. Pharmacol. Rev. 66, 334–395 (2014). https://doi.org/10.1124/pr.112.007336.
2. Sadovnichii, V.A., Sulimov, V.B.: Supercomputing technologies in medicine. In: Voevodin, V.V., Sadovnichii, V.A., Savin, G.I. (eds.) Supercomputing Technologies in Science, Education, and Industry, pp. 16–23. Moscow University Publishing (2009) (in Russian).
3. Sulimov, V.B.; Sulimov, A.V. Docking: molecular modeling for drug discovery (in Russian). AINTELL: Moscow, 2017. – 348 p. ISBN 978-5-98956-025-7.
4. Chen, Y.C., Beware of docking! Trends in pharmacological sciences, 2015, 36, (2), 78-95.
5. Yuriev, E.; Holien, J.; Ramsland, P.A., Improvements, trends, and new ideas in molecular docking: 2012–2013 in review. J. Mol. Recognit., 2015, 28, 581–604.
6. Pagadala, N.S.; Syed, K.; Tuszynski, J., Software for molecular docking: a review. Biophysical Reviews 2017, 9, (2), 91-102.
7. Mobley, D.L., Dill, K.A.: Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". Structure 17(4), 489–498 (2009). https://doi.org/10.1016/j.str.2009.02.010
8. Sulimov A.V., Kutov D.C., Katkova E.V., Kondakova O.A., Sulimov V.B. Search for approaches to improving the calculation accuracy of the protein—ligand binding energy by docking // Russian Chemical Bulletin, International Edition, 2017, Vol. 66, No.10, pp. 1913–1924.
9. Forli, S.; Huey, R.; Pique, M.E.; Sanner, M.F.; Goodsell, D.S.; Olson, A.J., Computational protein-ligand docking and virtual drug screening with the AutoDock suite. Nature protocols, 2016, 11, (5), 905-919.
10. Neves, M.A.; Totrov, M.; Abagyan, R., Docking and scoring with ICM: the benchmarking results and strategies for improvement. J Comput Aided Mol Des, 2012, 26, (6), 675-686.
11. Allen, W.J.; Balius, T.E.; Mukherjee, S.; Brozell, S.R.; Moustakas, D.T.; Lang, P.T.; Case, D.A.; Kuntz, I.D.; Rizzo, R.C., DOCK 6: Impact of new features and current docking performance. J Comput Chem, 2015, 36, (15), 1132-1156.
12. Sulimov, A.V.; Kutov, D.C.; Oferkin, I.V.; Katkova, E.V.; Sulimov, V.B., Application of the docking program SOL for CSAR benchmark. J Chem Inf Model, 2013, 53, (8), 1946-1956.
13. Oferkin, I.V., Katkova, E.V., Sulimov, A.V., Kutov, D.C., Sobolev, S.I., Voevodin, V.V., Sulimov, V.B.: Evaluation of docking target functions by the comprehensive investigation of protein-ligand energy minima. Adv. Bioinf. 2015, 12 (2015). https://doi.org/10.1155/2015/126858. Article ID 126858
14. Oferkin, I.V., Zheltkov, D.A., Tyrtyshnikov, E.E., Sulimov, A.V., Kutov, D.C.: Evaluation of the docking algorithm based on tensor train global optimization. Bull. South Ural State Univ. Ser. Math. Model. Program. Comput. Softw. 8(4), 83–99 (2015). https://doi.org/10.14529/mmp150407.

15. Sulimov, A.V., Kutov, D.C., Katkova, E.V., Sulimov, V.B.: Combined docking with classical force field and quantum chemical semiempirical method PM7. Adv. Bioinf. 2017, 6 (2017). https://doi.org/10.1155/2017/7167691. Article ID 7167691
16. A.V. Sulimov, D.A. Zheltkov, I.V. Oferkin, D.C. Kutov, E.V. Katkova, E.E. Tyrtyshnikov, V.B. Sulimov, Evaluation of the novel algorithm of flexible ligand docking with moveable target protein atoms // Computational and Structural Biotechnology Journal, 15 (2017) pp. 275-285. DOI information: 10.1016/j.csbj.2017.02.004.
17. A.V. Sulimov, D.A. Zheltkov, I.V. Oferkin, D.C. Kutov, E.V. Katkova, E.E. Tyrtyshnikov, V.B. Sulimov, Tensor train global optimization: application to docking in the configuration space with a large number of dimensions, Communications in Computer and Information Science 793, Eds. Vladimir Voevodin, Sergey Sobolev, Third Russian Supercomputing Days, RuSCDays 2017, Moscow, Russia, 25-26 September, 2017, Revised Selected Papers, Springer. pp. 151-167.
18. Alexey V. Sulimov, Danil C. Kutov, Ekaterina V. Katkova, and Vladimir B. Sulimov, Combined docking with classical force field and quantum chemical semiempirical method PM7 // Advances in Bioinformatics. Accepted 22 December 2016. Volume 2017, Article ID 7167691, 6 pages, https://doi.org/10.1155/2017/7167691.
19. Alexey V. Sulimov, Danil C. Kutov, Ekaterina V. Katkova, Ivan S. Ilin,Vladimir B. Sulimov, New generation of docking programs: Supercomputer validation of force fields and quantum-chemical methods for docking, Journal of Molecular Graphics and Modelling, 2017, 78, 139-147.
20. Byrd, R.H., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. 16(5), 1190–1208 (1995). https://doi.org/10.1137/0916069.
21. Zhu, C., Byrd, R.H., Lu, P., Nocedal, J.: Algorithm 778: L-BFGS-B: fortran subroutines for large-scale bound-constrained optimization. ACM Trans. Math. Softw. 23(4), 550–560 (1997). https://doi.org/10.1145/279232.279236.
22. H.M. Berman, J. Westbrook, Z. Feng, et al., The protein data bank, NucleicAcids Res. 28 (1) (2000) 235–242, http://dx.doi.org/10.1093/nar/28.1.235.
23. Avogadro: an Open-Source Molecular Builder and Visualization Tool. Version1. XX, 2017, Available at: (accessed April 26, 2018) https://avogadro.cc/.
24. Sadovnichy, V., Tikhonravov, A., Voevodin, V. l., Opanasenko, V., “Lomonosov": Supercomputing at Moscow State University. In Contemporary High Performance Computing: From Petascale toward Exascale, (CRC Press, 2013), pp.283-307.