JOURNAL of the TURKISH CHEMICAL SOCIETY
SECTION: A
CHEMISTRY
TURKISH CHEMICAL SOCIETY

Development of docking programs for Lomonosov supercomputer

Vladimir Sulimov[1,2], Ivan Ilin[1,2], Danil Kutov[1,2], Alexey Sulimov[1,2]

[1]Dimonta, Ltd, 15 Nagornaya Street, Building 8, 117186, Moscow, Russia.
[2]Research Computing Center, Lomonosov Moscow State University, 1 Leninskie Gory, Building 4, 119992, Moscow, Russia.

Abstract: The initial step of the rational drug design pipeline extremely needs an increase in effectiveness. This can be done using molecular modeling: docking and molecular dynamics. Docking programs are popular now due to their simple idea, quickness and ease of use. Nevertheless accuracy of these programs still leaves much to be desired and discovery by chance and experimental screening still play an important role. Docking performs ligand positioning in the target protein and estimates the protein-ligand binding free energy. While in many cases positioning accuracy of docking is satisfactory, the accuracy of binding energy calculations is insufficient to perform the hit-to-lead optimization. The accuracy depends on many approximations which are built into the respective model. We show that all simplifications restricting docking accuracy can be withdrawn and this can be done on the basis of modern supercomputer facilities allowing to perform docking of one ligand using many thousand computing cores. We describe in short the SOL docking program which is used during years for virtual screening of large ligand databases using supercomputer resources of Lomonosov Moscow State University. SOL to some extent is organized similarly to popular docking programs and reflects their limitations and advantages. We present our supercomputer docking programs, FLM and SOL-P, developed over the past 5 years for Lomonosov supercomputer of Moscow State University. These programs are free of most important simplifications and their performance shows the road map of the docking accuracy improvement. Some results of their performance for very flexible ligand docking into the rigid protein and docking of flexible ligands into the protein with some moveable protein atoms are presented. The so-called quasi-docking approach combining a force field and quantum chemical methods is described and it is shown that best docking accuracy is reached with the PM7 method and the COSMO solvent model.

Keywords: Docking, drug design, global optimization, force field, binding energy.

Submitted: October 17, 2019. **Accepted:** December 04, 2019.

Cite this: Sulimov V, Ilin I, Kutov D, Sulimov A. Development of docking programs for Lomonosov supercomputer. JOTCSA. 2020;7(1):259–76.

DOI: https://doi.org/10.18596/jotcsa.634130.

***Corresponding author.** E-mail: vladimir.sulimov@gmail.com, Tel: (+7 916 512 34 18).

INTRODUCTION

Nowadays docking plays (1) an important role at the initial stage of the rational drug design (2,3). Docking programs perform positioning of a ligand (a molecule) in the active site of the target protein responsible for the disease progression and estimate the protein-ligand binding energy. The latter is directly connected with the binding (or dissociating) constant defining the activity of the ligand and the respective inhibition constant. The ligand upon binding with the protein blocks (inhibits) its functioning and changes or stops the progression of a disease. The larger is the protein-ligand binding free energy the low concentration of the inhibitor is needed to reach the desired effect. Several dozens of docking programs exist as well as a dozen of sites presenting docking facilities are available (4–6). Nevertheless, docking accuracy is still unsatisfactory for the reliable separation of strong inhibitors from medium ones and the latter from weak inhibitors on the base of the docking score – the measure of the estimated protein-ligand binding energy, e.g. see Table 6 in (6). The performance of most of docking programs is based on the docking paradigm which assumes that the best position of the ligand in the active site of the target protein corresponds to the global minimum of the energy of the protein-ligand complex. As far as the experimentally determined position of the bound ligand in the protein is defined only by the structure of the crystallized protein-ligand complex the docking paradigm can be paraphrased as follows: the global energy minimum of the protein-ligand complex corresponds to the ligand pose near the crystallized native ligand position. The latter is usually taken from the Protein Data Bank (7). So, the docking problem is boiled down to solving the global optimization problem. And the dimensionality of the energy surface where the global minimum should be found is large: even if the ligand and the protein are rigid there are 6 degrees of freedom, 3 translations of the ligand as a rigid body, and 3 rotations of the ligand as a rigid body. Usually drug-like molecules have 5-15 internal rotations around ordinary valence bonds, named torsions, and the number of degrees of freedom of the global optimization problem is usually 10–20. The global optimization problem with such a large number of dimensions is a heavy task and its solution demands powerful heuristic algorithms as well as a lot of computation resources. Another docking problem arises when one asks the question: what is the energy calculation method for which the docking paradigm is true? On the one hand, the answer is simple: the energy calculation method should be adequate for the given protein-ligand system. But, on the other hand, there are many force fields, i.e. classic potentials describing interatomic interactions in the molecular system, or different quantum-chemical methods, and the energy calculation method for which the docking paradigm should be satisfied is *a priori* unknown. The difficulty of the docking problem is aggravated by the hands-on need to screen large databases of on-shelf or virtual compounds. Such databases or libraries can contain thousands and millions of compounds which can be treated as candidates to become inhibitors for these or those bio-targets.

Initial versions of some docking programs were developed more than 35 years ago, e.g. the DOCK program (8,9), and available restricted computing resources were the main limiting factor resulting in a lot of simplifications and crude approximations used in docking programs. In addition, force fields at that time were only in their infancy, and quantum-chemical methods could not be used for the calculation of the energy of such large molecular systems as protein-ligand complexes containing thousands of atoms. Nevertheless, the necessity of screening large libraries of ligands and limited computer recourses bring to somewhat a competing between docking programs for the lowest time of docking of one ligand using 1 CPU. Certainly, the development of docking programs under the mantra “faster and even faster” could not result in high accuracy of docking.

About ten years ago, the supercomputer era began, and hundreds and thousands of computing cores became available for solving of one task (job). For the docking problem, availability of supercomputers resulted in a possibility of screening of databases of ligands containing millions of compounds, and special efforts have been made to use supercomputers for this task with maximal efficiency. For example, one solution of this problem is presented in (10) where a specially designed software `wrapper' called HSP-DOCK on the base of the DOCK6 program has been used for screening a library of 1.4 million lead-like molecules from ZINC database against each of four targets. Another direction of the use of supercomputer resources for docking aims for the increase of docking accuracy (11).

We present here a short descriptions of our docking programs SOL (1,12,13), FLM (14–17) and SOL-P (15,18,19) developed in Lomonosov Moscow State University for the computing resources available in this university during the recent 15 years including the Lomonosov-2 supercomputer. Some applications of these docking programs are presented also in short.

DOCKING PROGRAMS

Actually, docking technique began to be developed in Lomonosov Moscow State University from autumn of 2005 under the government contract #

02.435.11.1008 of August 1, 2005. First version of the SOL docking program was applied to the development of new direct thrombin inhibitors from beginning of 2006 and the program has been implemented in the Web-oriented system Keenbase allowing to perform virtual screening of ligand libraries using the X-Com grid technology (20). In the following years the improvement of SOL and its application to new inhibitors design advanced abreast each other. A parallel version of SOL appeared in 2011, and supercomputer programs of a new generation have been developed from 2013 up till now. SOL is used now for a preliminary probing of new target proteins and for a massive virtual screening of large ligand databases on the Lomonosov supercomputer. The supercomputer docking programs of the new generation are used at the lead optimization stage and for the investigations of ways of improving the docking accuracy.

SOL classic docking program

We designate SOL as a classic docking program because it has many features common for many popular docking programs. However, during the development of this program, the goal was set to make as few model simplifications as possible and to take into account the most important effects that determine the accuracy of docking as much as possible. We foresaw the extraordinary growth of available computer resources in close future and did not take part in the race for the shortest docking time but tried to make calculations maximal accurate in the frame of existing facilities.

Algorithm and implementation

The energy of the molecular system is calculated in the frame of the Merck Molecular Force Field (MMFF94) (21) which was specially created for the description of protein-ligand molecular systems and drug design needs. This force field is based on a large amount of *ab initio* quantum-chemical calculations of different molecules and molecular complexes. As all other force fields it has its own deficiencies in some features but comparisons with different force fields revealed the best performance of MMFF94 (22,23). In addition, MMFF94 combines sufficiently good parameterization covering a broad spectrum of organic molecules and the well-defined procedure of atom typification applicable to an arbitrary organic compound.

As many other docking programs such as AutoDock, ICM, DOCK, and others, SOL uses a preliminary calculated grid of potentials describing interactions (electrostatic, van der Waals interactions) of all protein atoms with all possible types of a ligand probe atom in the frame of the MMFF94 force field practically without serious simplifications. The desolvation effect in the simplified form of the Generalized Born approximation and respective potentials are also stored in the nodes of the grid. The grid is used to move heavy calculations of all pairwise interactions of a probe ligand atom with all protein atoms from the global optimization docking step to the preliminary step. The grid is generated by the SOLGRID module and during the global optimization it is stored as a binary file in the RAM memory and it is easily accessed from the SOL module performing the global optimization. The size of the grid by default is a cube with the edge 22 Å covering the active site of the target protein. For the native docking the center of the cube is usually chosen in the geometrical center of the native ligand crystallized with the target protein. Such size of the docking cube is sufficient in most cases for the free ligand movement inside it during the global optimization process: the ligand can be at any position inside this docking cube but no one ligand atom cannot be outside the cube. The grid is formed by equidistant nodes along each of three orthogonal directions of cube edges, 101 nodes along one direction. Ligand energy in the field of the protein is calculated as a sum of grid potentials for all ligand atoms. A potential in the position of a given ligand atom is obtained by the interpolation of potentials in eight neighbouring grid nodes. Usually for a given target protein the grid is created for one to several hours on one computing core depending on the number of atoms in the protein and characteristics of the processor. The grid is generated once for the given position of the docking cube for the given target protein and the respective binary file with the grid is used in the following virtual screening of a ligand library.

The global optimization of the energy of a ligand in the field of the protein is performed by the SOL module. The optimized energy function is the sum of the grid energy of the ligand and the ligand internal energy calculated in the frame of the MMFF94 force field. So, the relative ligand stress energy is taken into account in the energy global optimization process: ligand poses with high stress energy have low chances to correspond to the global energy minimum of the protein-ligand system. The genetic algorithm is used for the global optimization. This is one of the most popular docking algorithms: for example it is implemented in the most commonly used docking programs Autodock (24) and Gold (25). The main idea of this mathematical method is a selection of most strong individuals in the evolution of a population of individuals developed through many generations. In our case the individual is a position a ligand inside the docking cube, and the measure of its strength in the competition with other individuals is the target energy function, i.e. the energy of the protein-ligand system. The population of the initial generation is randomly generated by variations of a ligand position in the docking cube. A ligand position is formed by translation and rotation of a ligand as a whole rigid body and by the change of each ligand torsion. The population size is 30000 by default but it can be set equal to any integer, e.g. in heavy docking cases for too complicated structure of the energy surface or for a too flexible ligand with a large number of torsions, the population size can be increased up to millions. The evolution is driven by the selection of the strongest individuals into the mating pool (by default its size is 70) and creation from them the next generation by their random crossover and direct translation of them into the next generation and random mutations. Several elite individuals (four by default) corresponding to the ligand positions with lowest energy are transformed to the next generation without any changes. The population size is kept fixed through all generations. When selecting individuals into the mating pool the niching is used. Niching prevents a selection into the mating pool close ligand poses and ensures diversity of individuals in the next generation of the population. Niching is realized by providing a positive energy penalty to the next in turn individual, which is a candidate to be selected into the mating pool, if RMSD (the root-mean-square deviation) between coordinates of ligand atoms corresponding to this individual and individuals, which have been already selected into the mating pool, is small. Actually, the value of the penalty is in inverse proportion to the RMSD value. The individual with a large penalty is moved out from the row of strongest individuals which are candidates to be selected into the mating pool. The number of generations is an input parameter of the SOL module and it is equal to 1000 by default. The strongest individual in the final generation, i.e. the ligand pose with the lowest energy, is the solution of the global optimization problem. How can one believe that this solution corresponds to the lowest value of the target energy function? Several independent runs (50 runs by default) of the genetic algorithm are performed to reveal the reliability of the found solution. Then, all 50 solutions (ligand poses) are clustered in respect with their positions: two ligand poses are included in one cluster if RMSD between coordinates of all their atoms is less than a given value (1 Å by default). The clusters are ranked in respect with energies of respective ligand poses and the solution of the global optimization problem corresponds to the ligand pose with the lowest energy from the first cluster. The cluster analysis helps to estimate reliability of the solution: the high population of the first cluster and a low number of separate clusters indicate the high reliability of the found solution of the global optimization problem. It means that in several absolutely independent runs of the genetic algorithm practically one and the same ligand pose corresponding to the lowest energy of the protein-ligand complex is found. In other words, the high population of the first cluster indicates convergence of independent runs of the genetic algorithm to the unique global minimum. In another utmost case when for 50 runs 50 different clusters are found docking should be considered as failed. With the default parameters SOL docks a ligand on one computing core for one to several hours.

Parallel versions of SOLGRID and SOL modules (26) are created on the base of MPI (message passing interface) allowing to perform calculations, generation of the grid of potentials and the global energy optimization, on many hundreds of cores: the time of the grid generation and fifty runs of the genetic algorithm with the default parameters can be reduced to less than 1 minute. The multi-processor performance of the SOLGRID module is useful for the optimization of the position of the docking cube in the active site of the target protein when a fast generation of the grid is needed. The multi-processor performance of SOL is usually used when docking with standard parameters is failed. In virtual screening of large ligand databases (dozens and hundreds of thousands of molecules) it is more effective to run SOL on the Lomonosov supercomputer (27) distributing ligands over hundreds and thousands computing cores and docking one ligand per one core. Certainly some auxiliary scripts and programs are created to queue up respective jobs and to analyze the docking results.

Applications

The SOL program was used for CSAR2011-2012 benchmark (12,28) together with other docking programs Gold, AutoDock, AutoDock Vina, ICM-VLS, Glide and others which were used by different research groups. The area under the ROC curve (AUC) was used as a measure of the reliability of the predicted inhibitor affinity. For the highest reliability of predictions, the AUC value is equal to 1, and for the worst reliability AUC is equal to 0.5. The AUC value shows the docking ability to find inhibitors among a large number of inactive compounds. AUC values were obtained for three target proteins: Chk1, LpxC, Urokinase. SOL was the best in AUC calculations for LpxC and Urokinase both, but it did not demonstrate a good result for Chk1 (28).

SOL has been successfully used at the initial stage of new low molecular weight direct inhibitors of different target proteins including experimentally confirmed inhibitors of thrombin (29,30), urokinase (uPA) (31,32), and the blood coagulation factor Xa (33,34). It should be noted here that a bad solvation of many newly synthesized compounds is one of most important obstacles for experimental testing of docking predicted inhibition activity of ligands.

FLM supercomputer docking program

The name of this program (14–17) is the abbreviation of "finding local minima" and reflects its goal which is to find all low energy minima, including the global minimum, of a protein-ligand complex. This is a gridless docking program which is not used the preliminary calculated grid of protein-ligand interaction potentials. In the course of docking the energy of any protein-ligand configuration is calculated directly in the frame of a given force field. FLM uses the MMFF94 force field (21) without simplifications.

Algorithm and implementation

Protein-ligand complexes have a very complicated multi-dimensional energy surfaces and the search and the search for the low energy minima is performed by the process of random throws of a flexible ligand into the docking area covering the active site of the rigid target protein followed by the optimization of the energy of the protein-ligand system from these random ligand poses using the L-BFGS gradient algorithm (35,36) by varying Cartesian coordinates of all ligand atoms. The initial ligand poses are obtained by random continuous translations and rotations of the ligand as a rigid body and by random continuous variations of ligand torsions. The only restriction is that the ligand geometrical center should be inside a sphere of a given radius (8 Å by default). If the ligand center moves out of the sphere in the optimization process the obtained minimum will not be included in the low energy minima set. Special attention is paid to the uniqueness of the ligand poses corresponding to the minima selected into the low energy minima set. The measure of the minima uniqueness is RMSD between the respective ligand poses calculated in the course of docking over heavy ligand atoms without chemical symmetry. Two minima will be considered different if RMSD and the difference of their energies are less than given values. The size of the low energy minima set is restricted by an input integer parameter and can be a sufficiently large number, e.g. several thousand – by default it is 8192 = 2^{13}. This set consists of the global energy minimum and every successive energy minimum above it. After finishing docking the set of found low energy minima is inspected on uniqueness again but taking into account chemical symmetry and by calculating RMSD over all ligand atoms.

FLM performs a massive parallel search of low energy minima using Message Passing Interface (MPI) and this search continues a given period of time. Basically, there is no the program termination criterion except the performance time and FLM can work as long as possible using as many as computing cores as available. The latter is defined by FLM good scalability with the number of cores growth. There are two versions, FLM-0.05 and FLM-0.10, working in the frame of MMFF94 either in vacuum or using the PCM continuum model (37) to take into account water solvent. Certainly, FLM-0.10 is much slower than FLM-0.05 and the latter needs about 20000 CPU x hours to dock one ligand and to find almost all low energy minima by performing several hundred thousand local optimization.

Applications

FLM can be useful during the hit-to-lead optimization when several ligands should be compared accurately on their ability to bind with a given target protein. However, the most interesting application of this program is to use it for the validation of new docking algorithms. For example, FLM was used for the verification of the TT-docking algorithm in (15). Another application of FLM is a comparison of different energy functions for docking. It was shown in (14) that the use of the MMFF94 force field with the PCM solvent model resulted in much better docking positioning accuracy than in the case when no solvent was taken into account. Later, the docking accuracy was compared for the CHARMM and MMFF94 force fields, for PM6-D3H4X (38) and PM7 (39) semiempirical quantum chemical methods. It was shown (40,41) that CHARMM is much better than MMFF94 but PM7 with the COSMO solvent model (42) is much better than CHARMM and it is slightly better comparing with PM6-D3H4X with COSMO. So, the best energy function for docking is the semiempirical quantum-chemical PM7 method together with the COSMO solvent model. This model as well as the PM7 method are realized in the MOPAC package (43) where the use of MOZYME module allows to calculate the whole protein-ligand complexes. Unfortunately, to use this quantum chemical energy function in the docking procedure is still impossible due too large computer resources needed. But, these findings bring to the idea of quasi-docking which is a two-step procedure. Firstly, a sufficiently broad spectrum of low energy minima is found in the frame of the given force field. Secondly, all these low energy minima are recalculated using a quantum-chemical method with an implicit solvent model. Results of quasi-docking when MMFF94 is used at the first step and PM7 with COSMO are used at the second step are presented in (17,44). The number of low energy minima which should be found at the first step is a specific number for each protein-ligand but for most of test complexes this number is equal to 4096. The energy (MMFF94 in vacuum) band occupied by these minima can reach several dozen kcal/mol for some complexes (17).

SOL-P supercomputer docking program

This supercomputer program (15,18,19) is also the gridless generalized docking program of the new generation. As the FLM program does SOL-P performs the search for low energy minima spectra of molecular systems, in particular, the protein-ligand complexes. However in opposite to the FLM program SOL-P uses for the search a more keen algorithm of the global optimization and also SOL-P is able to perform docking of flexible ligands into the target protein with moveable atoms. The energy of protein-ligand system depends on variables describing translations and rotations of the ligand as a whole rigid body, internal rotations of ligand molecular groups around ordinary chemical bonds (torsions) and Cartesian coordinates of selected target protein atoms and all these variables are treated simultaneously and equally in the global optimization process at that. For example, if the ligand has 10 torsions and 10 protein atoms must be treated as moveable the number of the independent variables describing the protein-ligand energy surface will be equal to d = 6 + 10 + 3 x 10 = 46 where 6 is the number of ligand translations and rotations as a rigid body. The energy of molecular systems is calculated in the frame of the MMFF94 force field as SOL and FLM programs do.

Algorithm and implementation

SOL-P uses the TT-docking algorithm the idea of which is as follows. The continuous energy function depending on d ligand and target protein variables is converted into a *d*-dimensional array (a tensor) using a discretization grid in the space of the variables, and modern methods of tensor analysis are applied to find the largest in magnitude element of the tensor. If the grid is fine enough the solutions of the continuous and discrete global optimization problems will be close to one another. The docking problem which is the global minimization problem can be easily transformed to the global maximization problem and we find that it is convenient to apply the magnitude maximization to the following functional (18):

$$f(x, E_*) = \exp\{100 \operatorname{arccot}[E(x) - E_*]\} \quad (1)$$

where *E(x)* is the dimensionless MMFF94 energy for the given conformation *x* of the protein-ligand complex, E_* is the global minimum found on the previous iteration.

The number of entries of a d-dimensional tensor grows exponentially in *d*, and if *d* is large it will be impossible to use the list of entries for practical needs. For example, the number of tensor entries will be huge, $100^{15} = 10^{30}$, for only 100 points at each dimension and *d* = 15. As a means to fight with this "so-called" *curse of dimensionality* the Tensor Trains (TT) decomposition for *d*-dimensional tensors was introduced ten years ago (45). TT-format is such a decomposition in which the initial real-valued *d*-dimensional tensor $A \in \mathbb{R}^{n_1 \times n_2, \ldots \times n_i}$ of the size n_i along the i-th dimension is reduced to *d* tensors of the dimension 3:

$$A(i_1, \ldots, i_d) \approx \sum_{(\alpha_1=1, \ldots, \alpha_{(d-1)}=1)}^{(r_1, \ldots, r_d)} G_1(i_1, \alpha_1) G_2(\alpha_1, i_2, \alpha_2) \ldots G_{(d-1)}(\alpha_{(d-2)}, i_{(d-1)}, \alpha_{(d-1)}) G_d(\alpha_{(d-1)}, i_d) \quad (2)$$

where numbers r_1, …, r_{d-1} are called TT-ranks of the tensor; for convenience, dummy ranks $r_0 \equiv r_d \equiv 1$ are also introduced. The 3-dimensional tensors $G_i \in R^{r_{i-1} \times n_i \times r_i}$ are called cores or carriages of the tensor train.

If TT-ranks are reasonably small, the TT-format will possess very useful properties (45,46): only $\sum_{i=1}^{d} n_i r_{i-1} r_i \;\; O(dnr^2)$ computer memory cells are required (n = max (n_i), r = max (r_i), i = 1, 2, …., d) to store the tensor, operations on tensors are reduced to standard matrix operations, and most of operations on tensors are performed in *O*(dnr^3) arithmetic operations or even faster. The TT-approximation of a tensor can be constructed in a robust way using TT-SVD (Singular Value Decomposition) method (45). However, the TT-SVD method needs all the elements of the tensor, but for a large tensor to calculate all tensor elements is practically impossible. But, there is a fast method, named TT-Cross, of the large tensor approximation utilizing only a small number of tensor elements (47). It finds the TT-decomposition of a tensor evaluating only *O*(dnr^2) elements and performing just *O*(dnr^2) arithmetic operations. TT-Cross exploits the well-known matrix cross interpolation method (48) applied to selected submatrices of the unfolding matrices of the given tensor. Unfolding matrices $A_k(i_1 \ldots i_k, i_{k+1} \ldots i_d) \in R^{n^k \times n^{d-k}}$ of the given tensor $A(i_1, \ldots, i_d) \in R^{n_1 \times \ldots \times n_d}$ contain the same elements of the initial tensor $A(i_1 \ldots i_d)$ but reordered, and TT-rank r_k is just the rank of the matrix A_k . The matrix cross interpolation method approximates a matrix $B(i, j) \in R^{m \times n}$ using only *O*((m+n)r) of its elements and performing just *O*($(m+n)r^2$) operations, where *r* is the approximation rank. The matrix cross interpolation method performs the search of the largest in magnitude matrix element, uses the found element to perform the Gauss elimination and repeats operations with the obtained matrix until the stopping criteria is met. So, the matrix cross interpolation method could be used as a simple global optimization method as it finds the largest in magnitude element among all evaluated elements of the matrix. Great advantage of the method is that it does not evaluate all matrix elements but only a small portion of them. Similarly, the TT-Cross method iteratively improves the sets of interpolation points searching for submatrices of unfolding matrices of larger volume (determinant in modulus) and consequently containing the elements of larger magnitude. Therefore TT-Cross can be used as a base for the TT global optimization method which uses the same search strategy as TT-Cross, but in more parallel way, and does not explicitly construct the TT-approximation of the whole tensor. To reduce the number of evaluations, the maximal rank is bounded by r_{max}. After the rank limitation iterations could possibly never converge and the maximal iterations number parameter is introduced.

At each iteration the following operations of the TT global optimization are performed: (i) the generation of submatrices of unfolding matrices using a set of tensor elements, (ii) the interpolation of submatrices using the TT-Cross method with rank $\leq r_{max}$ and as a result a set of interpolation points obtained for each submatrix contains elements with large values in modulus, (iii) these sets of interpolation points (protein-ligand conformations) are extended by the local optimization and projections of optimized interpolation points to the tensor are added to interpolation point sets, (iv) updating of each set of interpolation points of the unfolding matrix is made by merging the interpolation points of the previous unfolding matrix and ones of the subsequent unfolding matrix, (v) transition to step (i) using the obtained interpolation point (protein-ligand conformations) as the tensor elements for the next iteration step.

The basic optimal parameters of TT-docking are (18): the discretization grid in the domain of each variable $n = 2^{16}$, the maximal rank is bounded by $r_{max} = 4$, and the number of iterations in the process of optimization equals to 15. More details of the TT global optimization method can be found elsewhere (15,18,19).

The SOL-P program contains a set of modules performing the TT search for low energy minima, selecting only unique minima, performing additional local optimization of the protein-ligand conformations corresponding to selected unique low energy minima by the accurate L-BFGS gradient method and, finally, selecting again only unique minima and ranking them on their energy. In the process of TT-docking the ligand geometrical center (the ligand center of mass for equal ligand atoms' masses) can move inside the docking cube of a given size, by default the cube edge is equal to 10 Å, the center of the cube is situated either at the geometrical center of the native ligand crystallized with the target or can be selected based on other considerations. The selected protein atoms can move inside their small cubes with the edges 1.0 Å by default and centered at the initial protein atom positions, e.g. which are taken from Protein Data Bank. The docking cube size as well as small cubes sizes are input parameters of the program. More detailed description of SOL-P can be found in (18,19).

The TT docking algorithm is 10 times faster than the genetic algorithm (49) with approximately the same reliability of finding the ligand pose corresponding to the lowest energy of the protein-ligand complex. The comparison was made between TTDock and SOL programs for the same test set of protein-ligand complexes, for the same preliminary calculated grid of potentials in the frame of the same MMFF94 force field but it finds the same ligand poses with the lowest energy. For the rigid protein SOL-P is two orders of magnitude faster that the FLM program, these programs find the same global energy minima for a test set of protein-ligand complexes but some low energy minima are missed from the spectrum of low energy minima found by the SOL-P program comparing with the minima spectrum found by FLM (15). SOL-P can cope with docking of flexible ligands having a large number of torsions into a rigid protein as well as with docking of flexible ligands into protein having some moveable atoms. The latter is demonstrated in (18,19) where a successful docking is demonstrated with up to 157 degrees freedom: unsuccessful docking into a rigid target protein becomes successful when several dozen of protein atoms become moveable. Below we present some results of docking very flexible ligands and docking into proteins with moveable atoms.

Applications

Docking ligands with a larger number of torsions.

Docking highly flexible ligands is still a challenging task in the field of modern molecular modeling. In the same time, some flexible molecules, for example, oligopeptides are of high interest as potential therapeutics. Ability to handle them computationally determines a success rate of projects involved in developing peptide-like bioactive compounds. Most popular docking programs are able to dock ligands having no more than 10 torsions because of high dimensionality of the energy surface on which the search for a ligand bound configuration is performed. The TT-docking algorithm intrinsically aims to work with systems of high dimensionality, and to test its ability to dock very flexible ligands SOL-P is applied for docking oligopeptides and other kinds of flexible and branching non-peptide ligands in the same manner. All crystal complexes selected for docking have good resolution (better than 2 Å), do not contain any cofactor molecules (including metal ions) and have no gaps near the binding site. Protein structures and oligopeptide molecules are protonated at pH 7.4 using the Aplite program. Special attention was paid to the 1B9J complex. We found that the native ligand of this complex carried one positively charged group which was not involved in the interaction with any residues in the ligand crystal conformation. The global energy minimum found by SOL-P in this complex corresponds to the ligand pose where this ligand charged group is closely situated near the negatively charged carboxylic moiety of Glu32 and this ligand pose is quite different from the crystallized one: RMSD ≈ 6.7 Å. This is a simple consequence of positive and negative charges attraction in the absence of screening water solvent effect damping Coulomb interactions: SOL-P works without taking into account a water solvent. To avoid this side effect we neutralized the carboxylic moiety of Glu32 and re-docked the 1B9J native ligand. The difference between the ligand pose corresponding to the global energy minimum and the ligand crystallized pose become small: RMSD ≈ 1.2 Å. Non-peptide ligands are protonated at pH 7.4 with the Avogadro program (50). We also perform docking of these flexible ligands by the SOL program to compare efficiency of SOL-P with a program which has common features with popular docking programs (the grid approximation and the genetic algorithm of the global energy optimization) and uses the same force field as SOL-P does. The root-mean-square deviation (RMSD) between all atoms of the crystallized native ligand pose and the docking pose corresponding to the least energy of the protein-ligand complex is selected as an accuracy metric. Results of docking ligands with a large number of torsions are shown in Table 1.

As can be seen from Table 1, if RMSD cutoff is set to 2.3 Å that justified for such large ligands, SOL-P copes quite well with docking highly flexible ligands except one complex (2FLE), whereas SOL is able to dock correctly only one ligand which has the least number of torsions. Analyzing results for 2FLE complex we find that in the prepared protein structure two hydrogen atoms connected to backbone nitrogen atoms are placed incorrectly disturbing the planar configuration of peptide bonds in which they are involved and interfering with ligand hydrogen atoms when the ligand is in the crystallized position. In Figure 1 these two incorrectly added protein hydrogen atoms are designated by red letters and for the sake of clarity only a central part of the ligand in its crystallized position is shown.

Table 1. Results of docking ligands with a large number of torsions by SOL and SOL-P. N_{tor} stands for a number of torsions. The indicated charge corresponds to the total charge of a ligand at pH 7.4. Note: all oligopeptides carried, at least, two charged groups – the positively charged N-terminus and the negatively charged C-terminus.

#	PDB ID	Sequence of an oligopeptide	Charge	N_{tor}	RMSD after SOL, Å	RMSD after SOL-P, Å
1	6DQQ	AAAA	0	10	1.08	1.24
2	1B9J	KLK	+2	15	8.55	1.22
3	1OLA	VKPG	+1	18	9.94	1.28
4	6DTH	RPPGF	+1	18	10.13	2.12
5	3LIN	Non-peptide	0	20	6.36	0.97
6	6DTG	YLGANG	0	22	10.01	1.04
7	2OLB	KKK	+3	23	9.56	1.37
8	6DQU	GIINTL	0	25	8.49	2.28
9	2FLE	Non-peptide	0	25	8.99	13.78 (1.74)
10	1EC3	Non-peptide	0	25	12.29	2.19

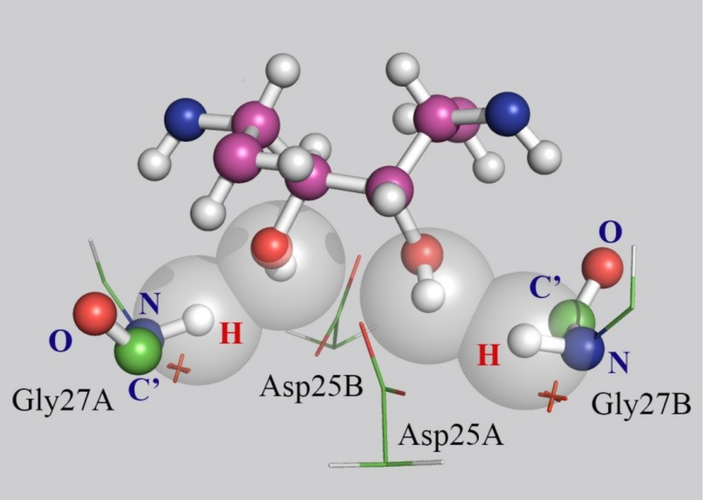

Figure 1. The intersection of Van der Waals radii of two hydrogen atoms of the protein (shown as red letters 'H's) and two hydrogen atoms of the ligand in its crystal state (shown in "balls and sticks" representation: carbon atoms shown in magenta colour, nitrogen atoms – in blue, oxygen atoms – in red, hydrogen atoms – in white).

The mentioned two protein hydrogen atoms are out of the planes (defined by protein atoms designated by blue letters in Figure 1) of the respective peptide bonds and this occurs possibly due to interactions of these hydrogen atoms with neighboring protein atoms of two residues, Asp25A and Asp25B, which are also shown in Figure 1. If a docked ligand pose is close to the ligand crystallized position the energy of the protein-ligand system will be high due to the intersection of Van der Waals radii of the protein and ligand hydrogen atoms. This is the reason of unsuccessful docking for the 2FLE complex in the rigid protein model. We tried two approaches to fix the problem. Firstly, a manual correction of hydrogen positions in the protein is done which does not lead to successful docking. Secondly, we use docking with SOL-P and add mobility to these two hydrogen atoms (other protein atoms are fixed) during docking. This has resulted in success and the final RMSD value for 2FLE complex is equal to 1.74 Å.

In the protein-ligand configuration corresponding to the global energy minimum found by docking with the two moveable protein hydrogen atoms these atoms are observed to be in the planes of respective peptide bonds and the above mentioned Van der Waals radii do not intersect (see Figure 2).

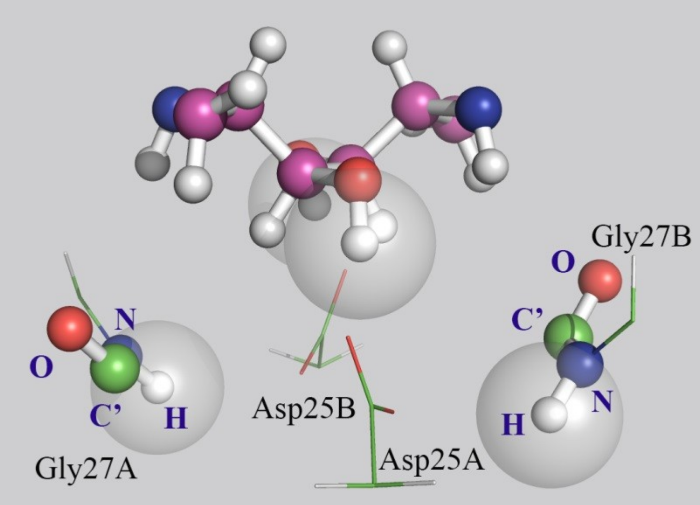

Figure 2. Positions of selected protein and ligand atoms of the 2FLE complex in the conformation corresponding to the global energy minimum found by docking with the two moveable protein hydrogen atoms (see text). The protein hydrogen atoms returned back in the planes of peptide bonds are marked by blue letters. As in Figure 1, the ligand is truncated for clarity.

Summing up, it can be noted that the SOL-P program can perform well in docking flexible ligands having up to 25 torsions. One of the main limitations of applying SOL-P as well as other docking programs can be insufficient quality of the full atomic (including hydrogen atoms) model of the target protein resulting in the presence of untypical distorted conformations to be reproduced. This effect should be taken into account when creating the full-atomic target model. The use of SOL-P with moveable protein atoms which are close to the ligand position, e.g. to the crystallized native ligand position, could be one possible solution of this problem.

Docking flexible ligands into proteins with moveable atoms.

Besides docking oligopeptides and branching ligands, mobility of separate protein atoms during docking is also a daunting task in modern computational chemistry. Each moveable protein atom increases a number of degrees of freedom and dimensionality of the energy space on which the search for a ligand bound configuration is performed. Accounting for mobility of protein atoms is of great importance for some protein targets and virtual screening projects.

The SOL-P program is able to perform docking with moveable protein atoms and this ability was studied recently (18) where SOL-P was tested by using 30 high quality protein-ligand complexes. As in the case of docking ligands with a large number of torsions, all complexes were split into a ligand and a protein and the estimation of docking accuracy relied upon reproducibility of native ligand crystallized poses after their docking into the corresponding protein structures. Both ligands and proteins were protonated at pH 7.4

For almost one third of complexes SOL-P coped with docking without considering mobility of protein atoms. To cope with the rest ones, we apply SOL-P using mobility of some protein atoms. The selection of moveable protein atoms is based on their proximity to the native ligand crystallized pose and it is made by the Mark-PMA program. For 4 complexes docking succeeds when protein flexibility is added. Features of these complexes and results of docking are presented in Table 2.

Table 2. Complexes from the test set for which docking results are dramatically improved after considering mobility of some protein atoms. N_{tor} stands for the number of ligand torsions, N_{lig} is the number of ligand atoms (including hydrogen atoms), $RMSD_{standart}$ is the root-mean-square deviation for all ligand atoms between the crystallized native ligand pose and the docked ligand pose in the global energy minimum after docking without moveable protein atoms, $RMSD_{moveable}$ is the root-mean-square deviation for all atoms between the crystallized ligand pose and the docked ligand pose in the global energy minimum after docking with moveable protein atoms. The number of these atoms is indicated as $N_{mov\ prot}$.

PDB ID	N_{tor}	N_{lig}	$RMSD_{standart}$	$RMSD_{moveable}$	$N_{mov\ prot}$
1J01	6	35	2.35	0.000017	15
1LQD	8	61	5.25	0.00668	17
1O3P	6	46	10.99	1.83	28
3CEN	7	50	7.59	1.59	13

To reveal reasons underlying improvement of docking results, we have visually analyzed docking poses and displaced positions of proteins' atoms. The main factor of the docking improvement is the improved structure of H-bonding: all displaced protein atoms are related to the hydrogen bond donors/acceptors. Ligand poses close to the ligand crystallized pose seem to have more favorable H-bonds in the protein conformation obtained after docking with moveable protein atoms. These H-bonds are less favorable when ligand poses close to the ligand crystallized pose are placed into the rigid protein without any adjustment of protein atoms. For instance, improved geometry of H-bonds between the benzamidine moiety of the ligand and the carboxylic group of Asp189 is observed for 1LQD complex – see Fig.3 in (51). The distance between the hydrogen atom in the amidine and the charged oxygen atom is reduced: from 1.9 Å observed for a complex with accounting the initial positions of the oxygen atoms of the carboxylic group to 1.6 Å observed for a complex after docking with moveable protein atoms. Actually, this change is quite small but, as can be seen from Table 2, it can lead to dramatic improvement in docking results.

We also apply semiempirical calculations to confirm that the configuration of the complex with displaced protein atoms after docking in SOL-P (denote it as M-complex) is better energetically than a configuration with the initial state of the protein (denote it as R-complex). For both M-complex and R-complex, heat of formation is estimated by a single SCF calculation with the PM7 method. Moreover, calculations were performed both in vacuum conditions and in water environment (solvent effects were modelled in the frame of the COSMO model). All calculations were done in MOPAC2016. The results of the calculations are listed in Table 3.

Table 3. Results of semiempirical calculations for the 1LQD complex aimed to confirm that the M-complex configuration is energetically more favourable than the R-complex configuration. Explanations what M/R-complex stands for can be found in the text above. Δ equals to the heat (M-complex) minus heat (R-complex), where heat (M/R-complex) is the heat of formation of the complex in the output file of MOPAC.

Type of calculation	**Heat of formation in MOPAC for R-complex, kcal/mol**	**Heat of formation in MOPAC for M-complex, kcal/mol**	**Δ, kcal/mol**
1SCF in vacuo	-19787.31	-19791.49	-4.18
1SCF + COSMO	-23030.61	-23032.65	-2.04

The consideration of results presented in Table 3 allows confirming that even the very slight change in positions of protein atoms belonging to the hydrogen bond donors/acceptors makes ligand poses near the ligand crystallized conformation to possess lower energy that is crucial for proper ranking of found ligand poses which are found in the docking procedure. Thus, the model of accounting for protein flexibility which is implemented in SOL-P can improve the energy estimation during docking by adjusting positions of protein atoms involved in H-bonding. The TT global optimization method is capable of successful handling of high dimensionality emerged from considering additional degrees of freedom related to moveable protein atoms.

Cross-docking with moveable protein atoms.

Cross-docking is a type of docking when the crystallized ligand from one complex is docked into the same protein which is taken from another complex containing the same protein but another crystallized native ligand. In other words, the ligand is docked into different protein conformations obtained from other crystal structures. The main purpose of the cross-docking approach is the study of the effects of induced fit upon binding and estimation of ability of model relied upon the certain protein structure to reproduce experimentally determined poses of chemically diverse ligands. The latter task is immensely important for quality control of the protein model intended to use in the structure-based virtual screening campaign. Cross-docking implies the superimposition of protein structures to the reference protein structure. The native ligands undergo the same transformation. After that, docking of "rotated" ligands can be assessed standardly in the terms of RMSD value calculated between the docking pose which corresponds to the global minimum and the ligand quasi-crystal ("rotated") pose.

We apply SOL-P to cross-docking using two pairs of complexes containing oligopeptides and also three pairs of protein-ligand complexes with ligands of non-peptide nature. The first pair includes 1B9J and 1OLA complexes which both contain oligopeptide-binding protein (OppA) of *S.typhimurium*. For the study we take the protein structure from the 1B9J complex in which Glu32 is neutralized (see Section "Docking ligands with a larger number of torsions"). Complexes of the second pair, 6DQQ and 6DTG, contain the same protein – OppA but of *H.influenzae*. Additional details about selected complexes can be found above in Table 1. Characteristics of three other pairs of complexes with non-peptide ligands selected for cross-docking are presented in Table 4. It is worthy to note that for all these complexes successful native docking is achieved by the SOL program which utilizes the grid approximation and the genetic algorithm as well as by the gridless supercomputer FLM docking program. In Table 4, column 6 the RMSD values between the native ligand crystallized pose and the docking pose with the best energy are shown. The values in parentheses corresponds to results of the repeated docking by SOL with heightened parameters of the genetic algorithm (population size = 3 x 10^6, number of generations = 1500).

Table 4. Characteristics of complexes with non-peptide ligands for the cross-docking test. N_{lig} – the number of ligand atoms (including hydrogen atoms), N_{tor} stands for the number of ligand torsions, Q_{lig} is the ligand charge.

PDB ID	Protein	N_{lig}	N_{tor}	Q_{lig}	RMSD after SOL, Å	RMSD after FLM, Å
1MRX	HIV-1 Protease group M subtype B (isolate BRU/LAI)	74	11	0	1.51	0.62
1MSM		78	12	0	11.13 (0.71)	0.95
3NU3		70	13	0	1.47	1.34
4LL3		75	13	0	11.16 (0.76)	1.04
2ZDM	Trypsin, Bos Taurus	59	9	1	1.16	0.96
2ZDN		58	9	1	1.98	0.59

The FLM program copes with native docking for all complexes: all calculated global minima are close to the corresponding native poses optimized in the frame of the MMFF94 force field in vacuum. SOL also copes with docking of all native ligands into the proteins they crystallized with either with standard docking parameters or with heightened ones.

The total procedure of preparation includes the following steps. At the beginning, the protein from one complex in each pair is superimposed to another protein and vice versa. Superimposing is conducted with our Super-impose-proteins program. To confirm integrity of aligned proteins, RMSD value for all atoms between the reference protein and the superimposed protein is calculated. For pairs under the consideration this value does not exceed 2 Å that confirms fitness of crystal structures to each other. The corresponding ligands are then superimposed too by the same transformation which is found at the previous step of proteins' superimposing. For example, the ligand from 1B9J is superimposed to the protein from 1OLA and vice versa; the ligand from 6DQQ is superimposed to the protein from 6DTG and vice versa. Transformed ligand conformations are denoted as quasi-native poses for the sake of brevity. Prior to docking in SOL-P, the quasi-native pose of every ligand is locally optimized in the MMFF94 force field with being placed in the active site of the corresponding protein. We check that this optimization do not cause quasi-native ligands to move more than 2 Å from initial coordinates in order to guarantee that optimization in SOL-P will not add bias into results. Moreover, we visually check that when being placed in the corresponding proteins quasi-native ligands have no clashes with protein atoms. After confirming that all quasi-native ligands satisfy these checks, they are used for cross-docking by SOL-P. Results of this docking are listed in Table 5.

Table 5. Results of cross-docking by SOL-P for pairs of complexes. RMSD value is calculated between the locally optimized quasi-native pose in the MMFF94 force field in vacuum and the docking pose corresponding to the global energy minimum. The INON index is found as the lowest number in sorted by the energy list of minima founded by SOL-P so that the RMSD of such a minimum is less than 2 Å relatively to the optimized quasi-native pose.

Ligand from	Protein from	Protein name, organism	RMSD, Å	INON
1B9J	1OLA	OppA, *Salmonella typhimurium*	6.709125	12
1OLA	1B9J		0.974301	1
1MRX	1MSM	HIV-1 protease	10.785163	2
1MSM	1MRX		2.416591	6
2ZDM	2ZDN	Trypsin, Bos taurus	7.128926	inf
2ZDN	2ZDM		2.510446	inf
3NU3	4LL3	HIV-1 protease	8.248767	6
4LL3	3NU3		4.517536	81
6DQQ	6DTG	OppA, *Haemophilus influenzae*	1.560041	1
6DTG	6DQQ		2.884729	6

As can be seen from Table 5, cross-docking for most cases is unsuccessful. The worst-cases correspond to INON=inf for complexes 2ZDN and 2ZDM. In these cases there are no minima with a corresponding ligand pose near (with small RMSD from) the optimized quasi-native ligand pose among all low energy minima found by SOL-P. To improve results, we use docking with moveable protein atoms. Two approaches for the selection of moveable protein atoms are used. The first one is applying the Mark-PMA program (18) which marks protein atoms to be moveable relying upon their proximity to the native crystallized ligand conformation. It was shown previously (18) that in many cases successful docking is reached when from 25 to 35 protein atoms became moveable using marking by this program. However, in this approach mobility of some protein atoms does not result in docking improvement but docking time can increase noticeably. Therefore the second approach is also tested. In the frame of this approach the selection of protein moveable atoms is made by hands taking into account their role in protein-ligand binding. Results of cross-docking with moveable protein atoms are listed in Table 6.

Table 6. Results of cross-docking with moveable protein atoms for different pairs of complexes. RMSD value was calculated between the locally optimized quasi-native pose in the MMFF94 force fiels in vacuum and the docking pose corresponding to the global minimum. N_{mov} stands for the total number of moveable protein atoms.

Ligand from	Protein from	N_{mov} selected by Mark-PMA	RMSD, moveability by Mark-PMA	INON, moveability by Mark-PMA	N_{mov} selected by hand	RMSD, moveablity by hand	INON, moveablity by hands
1B9J	1OLA	32	6.669763	30	17	6.649110	11
1OLA	1B9J	30	2.430996	7	16	2.624906	7
1MRX	1MSM	28	0.245360	1	19	3.589922	2
1MSM	1MRX	29	1.289056	1	23	1.897655	1
2ZDM	2ZDN	28	0.921806	1	18	0.919140	1
2ZDN	2ZDM	26	2.271176	3	18	1.899394	1
3NU3	4LL3	28	1.607883	1	19	8.382071	7
4LL3	3NU3	25	0.921040	1	21	4.278851	2
6DQQ	6DTG	30	1.236382	1	11	1.167734	1
6DTG	6DQQ	34	0.761266	1	25	1.304007	1

It is clearly seen when comparing Table 5 and Table 6 that for most complexes cross-docking is improved and becomes successful after taking into account mobility of neighboring protein atoms and this improvement is higher when automatic marking moveable atoms is applied. These findings are mainly justified by changes in positions of atoms involved in H-bonding during docking with moveable protein atoms because there are no dramatic differences in protein structures within each pair and no steric clashes observed for quasi-native poses of rotated ligands when placing into the corresponding protein structures. Consider results of cross-docking with automatic selection of protein atoms to be moveable. They show that for three pairs unsuccessful docking is obtained even after adding mobility to some protein atoms. In the case of 1OLA/1B9J pair (a ligand from 1OLA is docked into a protein from 1B9J) accuracy of positioning becomes worse after cross-docking with moveable protein atoms. The failure is probably related to facts that the native protein from 1B9J is modified (neutralization of Glu32) and the ligand from 1OLA is not able to bind properly because of the modification. On the contrary, the protein from 1OLA does not contain neutralized Glu32 and thereby allows the ligand from 1B9J to stick to the active site in the wrong way which we have observed earlier performing docking of oligopeptides (see Section "Docking ligands with a larger number of torsions"). Addressing these issues might possibly lead to more adequate results.

In the case of unsuccessful docking of the ligand from 2ZDN into the protein from 2ZDM we find that the docking pose of the ligand both after docking without moveable protein atoms and with moveable protein atoms quite correctly reproduces the ligand crystallized conformation excluding the tail ligand moiety containing a cyclopentyl group (see Figure 3). It is also manifested in the fact that the RMSD values only slightly differ from 2 Å (2.51 Å – in the case of docking without protein flexibility, 2.27 Å – after docking with moveable protein atoms). And besides, the RMSD estimation is actually carried out with considering not only ligand atoms but protein atoms as well. If calculate RMSD using only ligand atoms, one obtain difference being 2.016 Å between the quasi-native pose of 2ZDN ligand and its docking pose found after docking with moveable protein atoms. This RMSD value is very close to "ideal" boundary – 2.0 Å.

Study (52) which describes obtaining crystal structures of 2ZDM and 2ZDN confirms conformational flexibility of the tail moiety of the ligand from 2ZDN in MD simulations. With regard to successful docking for this pair when applying hand marking moveable protein atoms, it was found that "successful" pose of 2ZDN ligand after this docking is immensely similar to the docking pose after positioning with an automatically selected set of moveable protein atoms (see Figure 3). In that context, the weakness of the RMSD concept for estimating docking results is illustrative. Given all these facts, we can conclude that, despite RMSD being slightly larger than 2 Å SOL-P manages to dock the ligand from 2ZDN in the meaningful way during our cross-docking simulations.

Summing up, one can note that adding protein flexibility in the cross-docking procedure helps to improve accuracy of positioning of quasi-native ligands: for six initially failed complexes the RMSD value is reduced to less than 2 Å after including protein flexibility in the docking process. Considering approaches for selecting protein atoms to be moveable and obtained results, it can be concluded that the first (automatic) approach to choose moveable protein atoms results in more accurate positioning in docking but at the expense of some increase (approximately from 0.5 to 2 hours) of docking time comparing with the second approach.

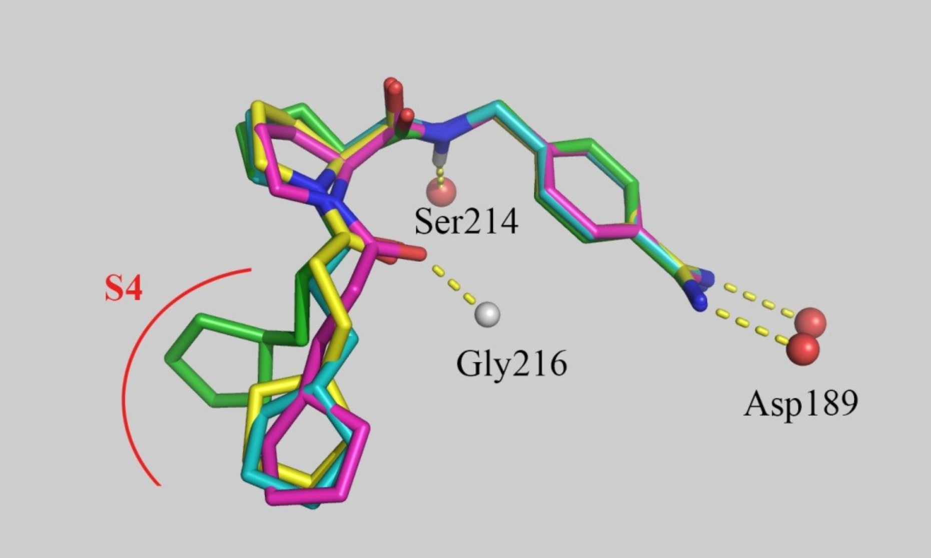

Figure 3. Results of cross-docking for the ligand from 2ZDN into the protein from 2ZDM. 4 different ligand poses are shown: the quasi-native pose (carbon atoms – in green color), the pose after docking with no moveable protein atoms (carbon atoms – in cyan color), the pose after docking with moveable protein atoms selected by Mark-PMA (purple-colored carbon atoms), and the pose after docking with moveable protein atoms chosen by hand (yellow-colored carbon atoms). Protein atoms involved in H-bonding with the ligand are shown as spheres. Nitrogen atoms are colored in blue, oxygen atoms – in red, hydrogen atoms – in white. For sake of clarity, most hydrogen atoms of the ligand are removed.

CONCLUSIONS

Three docking programs developed at Lomonosov Moscow State University are described in the present work: SOL, FLM and SOL-P. Their performance is based on the docking paradigm connecting docking with the global optimization problem: the ligand is bound near the global energy minimum of the protein-ligand complex. The first program, SOL, is used successfully for the computer aided structural based drug design during almost 15 years. The MMFF94 force field is used in SOL for the calculation of the energy of protein-ligand complexes and the genetic algorithm is used for the global optimization. The preliminary calculated grid of potentials of interactions of ligand probe atoms with the target protein is used in SOL and they are Coulomb, van der Waals interatomic interactions and desolvation potentials appeared from the simplified form of the implicit Generalized Born solvent model. The SOLGRID module generating the grid as well as the docking SOL module are parallelized and adapted for the supercomputer calculations. SOL is used for initial docking of native ligands into new targets and then for large ligand databases virtual screening.

FLM and SOL-P belong to the new generation of generalized docking programs designed for running on supercomputers. They use the MMFF94 force field both without any simplifications and fitting parameters. Their task is to find the whole spectrum of low energy minima for a given energy function. FLM can perform long time on available computing resources until the found pool of low energy minima reaches the saturation, i.e. until the pool ceases to be updated. For the first time we investigated carefully spectra of low energy minima of several dozen protein-ligand complexes and checked the feasibility of the docking paradigm. Low energy minima sets found by FLM can be used as reference sets to compare different docking algorithms and energy functions. The quasi-docking procedure approaching quantum chemical docking is realized on the base of several thousand low energy minima found by FLM. It is shown that the PM7 quantum chemical semiempirical method together with the COSMO solvent model is one of the best candidates for energy calculations in the docking procedure.

The new TT-docking algorithm is realized in the SOL-P program. This algorithm is based on the recently developed so-called tensor train (TT) global optimization method which in its turn is based on the TT approximation of large multi-dimensional tensors. The main advantage of this TT-docking method is its ability to perform docking with a large number of degrees of freedom, i.e. to find low energy minima spectra on the energy surface with very large number of dimensions. This property of the TT docking algorithm opens up the possibility for docking molecules with a large number (> 20) of torsions and for docking flexible ligands into proteins with mobile atoms. In the TT-docking procedure all degrees of freedom, i.e. all variables describing ligand and protein conformations, are treated equally and simultaneously during the global energy optimization process. Some examples of successful docking of ligands with a larger number of torsions, docking into the protein with moveable atoms as well as cross-docking are presented.

Finally, the supercomputer docking programs briefly described in this work open the road to higher accuracy of docking: to the higher positioning accuracy as well as to the high accuracy of the protein-ligand binding energy calculation. This will certainly results in higher effectiveness the rational drug design.

ACKNOWLEDGEMENTS

The manuscript is presented at the 12th Asian Federation for Medicinal Chemistry (AFMC) International Symposium (AIMECS 2019), 08-11 September, 2019, Istanbul, Turkey. 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 (27,53).

REFERENCES

1. Sulimov VB, Sulimov A V. Docking: Molecular modeling for drug discovery. Moscow: AINTELL; 2017. 348 (in Russian).

2. Sadovnichii VA, Sulimov VB. Supercomputing technologies in medicine. In: Sadovnichii VA, Savin GI, Voevodin VV, editors. Supercomputing Technologies in Science. Moscow: Moscow University Publishing; 2009. p. 16–23.

3. Sliwoski G, Kothiwale S, Meiler J, Lowe Jr. EW. Computational methods in drug discovery. Pharmacological reviews [Internet]. 2013;66(1):334–95. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24381236.

4. Sulimov VB, Kutov DC, Sulimov A V. Advances in docking. Current Medicinal Chemistry [Internet]. 2019;25(42):1–25. Available from: http://www.eurekaselect.com/node/165105/article.

5. Pagadala NS, Syed K, Tuszynski J. Software for molecular docking: a review. Biophysical Reviews. 2017/05/17. 2017;9(2):91–102.

6. Yuriev E, Holien J, Ramsland PA. Improvements, trends, and new ideas in molecular docking: 2012-2013 in review. Journal of Molecular Recognition. 2015/03/27. 2015;28(10):581–604.

7. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research [Internet]. 1999/12/11. 2000;28(1):235–42. Available from: https://www.ncbi.nlm.nih.gov/pubmed/10592235.

8. Allen WJ, Balius TE, Mukherjee S, Brozell SR, Moustakas DT, Lang PT, et al. DOCK 6: Impact of new features and current docking performance. Journal of Computational Chemistry [Internet]. 2015/04/29. 2015;36(15):1132–56. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25914306.

9. Brozell SR, Mukherjee S, Balius TE, Roe DR, Case DA, Rizzo RC. Evaluation of DOCK 6 as a pose generation and database enrichment tool. Journal of Computer-Aided Molecular Design [Internet]. 2012;26(6):749–73. Available from: https://doi.org/10.1007/s10822-012-9565-y.

10. Trager RE, Giblock P, Soltani S, Upadhyay AA, Rekapalli B, Peterson YK. Docking optimization, variance and promiscuity for large-scale drug-like chemical space using high performance computing architectures. Drug Discovery Today [Internet]. 2016/06/30. 2016;21(10):1672–80. Available from: http://www.sciencedirect.com/science/article/pii/S1359644616302410.

11. Sulimov AV, Kutov DC, Katkova EV, Kondakova OA, Sulimov VB. Search for approaches to improving the calculation accuracy of the protein-ligand binding energy by docking. Russian Chemical Bulletin [Internet]. 2017;66(10):1913–24. Available from: https://doi.org/10.1007/s11172-017-1966-6.

12. Sulimov AV, Kutov DC, Oferkin IV, Katkova EV, Sulimov VB. Application of the docking program SOL for CSAR benchmark. Journal of Chemical Information and Modeling [Internet]. 2013/07/09. 2013;53(8):1946–56. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23829357.

13. Romanov AN, Kondakova OA, Grigoriev FV, Sulimov AV, Luschekina SV, Martynov YB, et al. The SOL docking package for computer-aided drug design. Vol. 9, Numerical Methods and Programming. 2008. p. 213-233 (in Russian).

14. Oferkin IV, Katkova EV, Sulimov AV, Kutov DC, Sobolev SI, Voevodin VV, et al. Evaluation of docking target functions by the comprehensive investigation of protein-ligand energy minima. Advances in Bioinformatics [Internet]. 2015 [cited 2015 Dec 23];2015(Article ID 126858, 32 pages). Available from: https://www.ncbi.nlm.nih.gov/pubmed/26693223.

15. Oferkin IV, Zheltkov DA, Tyrtyshnikov EE, Sulimov AV, Kutov DC, Sulimov VB. Evaluation of the docking algorithm based on tensor train global optimization. Bulletin of the South Ural State University, Ser Mathematical Modelling, Programming & Computer Software [Internet]. 2015;8(4):83–99. Available from: https://mmp.susu.ru/article/en/362.

16. Sulimov AV, Kutov DC, Sulimov VB. Parallel supercomputer docking program of the new generation: finding low energy minima spectrum. In: Voevodin V, Sobolev S, editors. 4th Russian Supercomputing Days, RuSCDays 2018. Cham: Springer International Publishing; 2019. p. 314–30. (Communications in Computer and Information Science; vol. 965).

17. Kutov DC, Sulimov AV, Sulimov VB. Supercomputer docking: Investigation of low energy minima of protein-ligand complexes. Supercomputing Frontiers and Innovations [Internet]. 2018;5(3):134–7. Available from: http://superfri.org/superfri/article/view/248.

18. Sulimov AV, Zheltkov DA, Oferkin IV, Kutov DC, Katkova EV, Tyrtyshnikov EE, et al. Evaluation of the novel algorithm of flexible ligand docking with moveable target-protein atoms. Computational and Structural Biotechnology Journal [Internet]. 2017;15:275–85. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28377797.

19. Sulimov AV, Zheltkov DA, Oferkin IV, Kutov DC, Katkova EV, Tyrtyshnikov EE, et al. Tensor train global optimization: Application to docking in the configuration space with a large number of dimensions. In: Voevodin VV, Sobolev SI, editors. 3rd Russian Supercomputing Days, RuSCDays 2017 [Internet]. Cham: Springer International Publishing; 2017. p. 151–67. (Communications in Computer and Information Science; vol. 793). Available from: https://doi.org/10.1007/978-3-319-71255-0_12.

20. Sulimov V, Romanov A, Grigoriev F, Kondakova O, Sulimov A, Bryzgalov P, et al. Web-oriented system Keenbase for virtual screening and design of new ligands for biological macromolecules. Application for new drugs searches. In: Saint-Petersburg international workshop on nanobiotechnologies. Saint-Petersburg; 2006. p. 33–4.

21. Halgren TA. Merck molecular force field [Internet]. Vol. 17, Journal of Computational Chemistry. 1996. p. 490–641. Available from: http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1096-987X(199604)17:5/6%3C490::AID-JCC1%3E3.0.CO;2-P/abstract.

22. Halgren TA. MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. Journal of Computational Chemistry [Internet]. 1999 May 1;20(7):730–48. Available from: https://doi.org/10.1002/(SICI)1096-987X(199905)20:7%3C730::AID-JCC8%3E3.0.CO.

23. Beachy MD, Chasman D, Murphy RB, Halgren TA, Friesner RA. Accurate ab Initio Quantum Chemical Determination of the Relative Energetics of Peptide Conformations and Assessment of Empirical Force Fields. Journal of the American Chemical Society [Internet]. 1997;119(25):5908–20. Available from: http://dx.doi.org/10.1021/ja962310g.

24. Forli S, Huey R, Pique ME, Sanner M, Goodsell DS, Olson AJ. Computational protein-ligand docking and virtual drug screening with the AutoDock suite. Nature protocols [Internet]. 2016;11(5):905–19. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4868550/.

25. Liebeschuetz JW, Cole JC, Korb O. Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test. J Comput Aided Mol Des [Internet]. 2012/05/10. 2012;26(6):737–48. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22569591.

26. Oferkin IV, Sulimov AV, Kondakova OA, Sulimov VB. Implementation of parallel computing for docking programs SOLGRID and SOL. Noviye viychi. Vol. 12, Numerical Methods and Programming. 2011. p. 9-23 (in Russian).

27. Voevodin VV, Antonov AS, Nikitenko DA, Shvets PA, Sobolev SI, Sidorov IY, et al. Supercomputer Lomonosov-2: Large scale, deep monitoring and fine analytics for the user community. Supercomputing Frontiers and Innovations [Internet]. 2019;6(2):4–11. Available from: https://superfri.org/superfri/article/view/278.

28. Damm-Ganamet KL, Smith RD, Dunbar Jr. JB, Stuckey JA, Carlson HA. CSAR Benchmark Exercise 2011−2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series,. J Chem Inf and Model. 2013;53:1853–70.

29. Sulimov VB, Romanov AN, Kondakova OA, Sinauridze EI, Butylin AA, Gribkova IV, et al. New thrombin inhibitors: Molecular design and experimental discovery. In: 5th Anniversary Congress of International Drug Discovery Science & Technology 2007, IDDST 2007, 7-13 November 2007. Xi’an, China; 2007. p. 145.

30. Sinauridze EI, Romanov AN, Gribkova IV, Kondakova OA, Surov SS, Gorbatenko AS, et al. New synthetic thrombin inhibitors: Molecular design and experimental verification. PLoS One [Internet]. 2011/05/24. 2011;6(5):e19969. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21603576.

31. Sulimov VB, Katkova EV, Oferkin IV, Sulimov AV, Romanov AN, Roschin AI, et al. Application of molecular modeling to urokinase inhibitors development. BioMed Research International [Internet]. 2014/06/27. 2014;2014:625176. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24967388.

32. Beloglazova IB, Plekhanova OS, Katkova EV, Rysenkova KD, Stambol’skii DV, Sulimov VB, et al. Molecular modeling as a new approach to the development of urokinase inhibitors. Bulletin of Experimental Biology and Medicine [Internet]. 2015;158(5):700–4. Available from: https://doi.org/10.1007/s10517-015-2839-3.

33. Sulimov VB, Gribkova IV, Kochugaeva MP, Katkova EV, Sulimov AV, Kutov DC, et al. Application of molecular modeling to development of new factor Xa inhibitors. BioMed Research International [Internet]. 2015 [cited 2015 Oct 21];2015(Article ID 120802, 15 pages). Available from: https://www.ncbi.nlm.nih.gov/pubmed/26484350.

34. Ilin I, Lipets E, Sulimov A, Kutov D, Shikhaliev K, Potapov A, et al. New factor Xa inhibitors based on 1,2,3,4-tetrahydroquinoline developed by molecular modelling. Journal of Molecular Graphics and Modelling [Internet]. 2019;89:215–24. Available from: http://www.sciencedirect.com/science/article/pii/S1093326318305576.

35. Byrd R, Lu P, Nocedal J, Zhu C. A Limited Memory Algorithm for Bound Constrained Optimization. SIAM Journal on Scientific Computing [Internet]. 1995;16(5):1190–208. Available from: https://doi.org/10.1137/0916069.

36. Zhu C, Byrd RH, Lu P, Nocedal J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM Transactions on Mathematical Software. 1997;23(4):550–60.

37. Sulimov VB, Mikhalev AY, Oferkin IV, Oseledets IV, Sulimov AV, Kutov DC, et al. Polarized continuum solvent model: Considerable acceleration with the multicharge matrix approximation. International Journal of Applied Engineering Research. 2015;10(24):44815–30.

38. Rezac J, Hobza P. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. J Chem Theory Comput [Internet]. 2012/01/10. 2012;8(1):141–51. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26592877.

39. Stewart JJ. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. J Mol Model [Internet]. 2012/11/29. 2013;19(1):1–32. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23187683.

40. Sulimov AV, Kutov DC, Katkova EV, Sulimov VB. Combined docking with classical force field and quantum chemical semiempirical method PM7. Advances in Bioinformatics [Internet]. 2017;2017(Article ID 7167691, 6 pages). Available from: https://www.ncbi.nlm.nih.gov/pubmed/28191015.

41. Sulimov AV, Kutov DC, Katkova EV, Ilin IS, Sulimov VB. New generation of docking programs: Supercomputer validation of force fields and quantum-chemical methods for docking. Journal of Molecular Graphics and Modelling [Internet]. 2017;78:139–47. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29055806.

42. Klamt A, Schuurmann G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. Journal of the Chemical Society, Perkin Transactions 2 [Internet]. 1993;(5):799–805. Available from: http://dx.doi.org/10.1039/P29930000799.

43. Stewart JJP. MOPAC2016 [Internet]. Colorado Springs, CO, USA: Stewart Computational Chemistry; 2016. Available from: http://openmopac.net.

44. Sulimov AV, Kutov DK, Il'in IS, Sulimov VB. Doking s kombinirovanniym primeneniev silovovo pola i kvantovo-himicheskovo metoda. Biomeditsinskaya himiya. 2019;65(2):80–5 (in Russian).

45. Oseledets I, Tyrtyshnikov E. Breaking the Curse of Dimensionality, Or How to Use SVD in Many Dimensions. SIAM Journal on Scientific Computing [Internet]. 2009;31(5):3744–59. Available from: https://doi.org/10.1137/090748330.

46. Oseledets I. Tensor-Train Decomposition. SIAM Journal on Scientific Computing [Internet]. 2011;33(5):2295–317. Available from: https://doi.org/10.1137/090752286.

47. Oseledets I, Tyrtyshnikov E. TT-cross approximation for multidimensional arrays. Linear Algebra and its Applications [Internet]. 2010;432(1):70–88. Available from: http://www.sciencedirect.com/science/article/pii/S0024379509003747.

48. Goreinov S, Tyrtyshnikov E. The maximal-volume concept in approximation by low-rank matrices. Contemporary Mathematics. 2001;268:47–51.

49. Zheltkov DA, Oferkin IV, Katkova EV, Sulimov AV, Sulimov VB, Tyrtyshnikov EE. TTDock: a docking method based on tensor train decompositions [Internet]. Vol. 14, Numerical Methods and Programming. 2013. p. 279-291 (in Russian). Available from: http://num-meth.srcc.msu.ru/zhurnal/tom_2013/pdf/v14r131.pdf.

50. Hanwell MD, Curtis DE, Lonie DC, Vandermeersch T, Zurek E, Hutchison GR. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. Journal of Cheminformatics [Internet]. 2012;4(1):17. Available from: https://doi.org/10.1186/1758-2946-4-17.

51. Sulimov A, Kutov D, Ilin I, Zheltkov D, Tyrtyshnikov E, Sulimov V. Supercomputer docking with a large number of degrees of freedom. SAR and QSAR in Environmental Research [Internet]. 2019 Oct 3;30(10):733–49. Available from: https://www.tandfonline.com/doi/full/10.1080/1062936X.2019.1659412.

52. Brandt T, Holzmann N, Muley L, Khayat M, Wegscheid-Gerlach C, Baum B, et al. Congeneric but still distinct: how closely related trypsin ligands exhibit different thermodynamic and structural properties. Journal of molecular biology [Internet]. 2011;405(5):1170—1187. Available from: https://doi.org/10.1016/j.jmb.2010.11.038.

53. Sadovnichy V, Tikhonravov A, Voevodin V, Opanasenko V. "Lomonosov": Supercomputing at Moscow State University. In: Contemporary High Performance Computing: From Petascale toward Exascale. Boca Raton, United States: Boca Raton, United States; 2013. p. 283–307.