Marcin PACHOLCZYK
Politechnika Śląska

# FINITE ABSORBING MARKOV CHAIN AS A MODEL OF PROTEIN-LIGAND INTERACTION

**Summary.** The Stochastic Roadmap Simulation (SRS) and finite absorbing Markov chain theory is applied to build a model of protein-ligand binding process. The time to escape (TTE) from a funnel of attraction around binding site, a computational quantity, is evaluated as a measure of binding affinity. The results based on PDBBind CoreSet (release 2008) show statistically significant correlation between experimental binding affinity and calculated TTE. Presented approach performs best for ligands with small number of internal degrees of freedom (rotatable bonds).

# SKOŃCZONY, POCHŁANIAJĄCY ŁAŃCUCH MARKOWA JAKO MODEL INTERAKCJI BIAŁKO-LIGAND

**Streszczenie.** W pracy wykorzystano technikę symulacji metodą map stochastycznych i teorię łańcuchów Markowa do stworzenia modelu procesu wiązania ligandu przez białko. Oceniono wielkość obliczeniową - czas ucieczki ze stożka przyciągania wokół miejsca wiążącego, jako miarę powinowactwa białko-ligand. Wyniki uzyskane dla danych ze zbioru PDBBind CoreSet (wyd. 2008) wykazują istotną statystycznie korelację pomiędzy eksperymentalną miarą powinowactwa i obliczonym czasem ucieczki. Prezentowane podejście wykazuje najlepsze własności dla ligandów o małej liczbie wewnętrznych stopni swobody (wiązań obrotowych).

## 1. Introduction

Analysis of protein - small molecule interactions is crucial in the discovery of new drug candidates and lead structure optimization. Small biomolecules (ligands) are highly flexible and may adopt numerous conformations upon binding to the protein. Scoring functions are traditionally used in many docking protocols and have key impact on a quality of structure-based virtual screening. A correct scoring function should be able to guide search algorithm to find and recognize native-like docking poses. In ideal case scoring function should be able to predict binding affinity. Despite extensive research, scoring remains a major challenge in structure-based virtual screening [14]. The Stochastic Roadmap Simulation (SRS) and finite absorbing Markov chain theory is applied to build a model of protein-ligand binding process

[1, 8]. The time to escape (TTE) from a funnel of attraction around binding site, a computational quantity, is evaluated as a measure of binding affinity.

## 2. Model of Protein-Ligand Interaction

The model of electrostatics associated with Poisson-Boltzmann equation (PBE) is far more accurate in this case than simple Coulombic models and incorporates features such as location dependent dielectric constant and mobile ions contribution to the electrostatic potential (natural environment for proteins is usually salty aquatic solution). Protein is considered a rigid body limited by solvent accessible surface [3]. In order to solve linear PBE on 3D grid (Fig. 1) computer program DelPhi was used [10]. The configuration was set to 1Å grid resolution, protein and solution dielectric constants equal to 4 and 80 respectively and physiological salt concentration. The electrostatic grid was supplemented by van der Waals interactions calculated using typical Lennard-Jones 12-6 potential [6]. The total protein-ligand interaction model takes the following form:

$$E = E_{PL}^{elec} + E_{PL}^{vdW} + E_{L}^{elec} + E_{L}^{vdW} \tag{1}$$

where $E_{PL}^{elec}$, $E_{PL}^{vdW}$ are protein-ligand interactions and $E_{L}^{elec}$, $E_{L}^{vdW}$ are internal ligand electrostatic and van der Waals interactions respectively. The internal ligand interactions were calculated using Coulomb's law (assuming dielectric constant of water solution) in electrostatic part and Lennard-Jones potential in van der Waals part

.

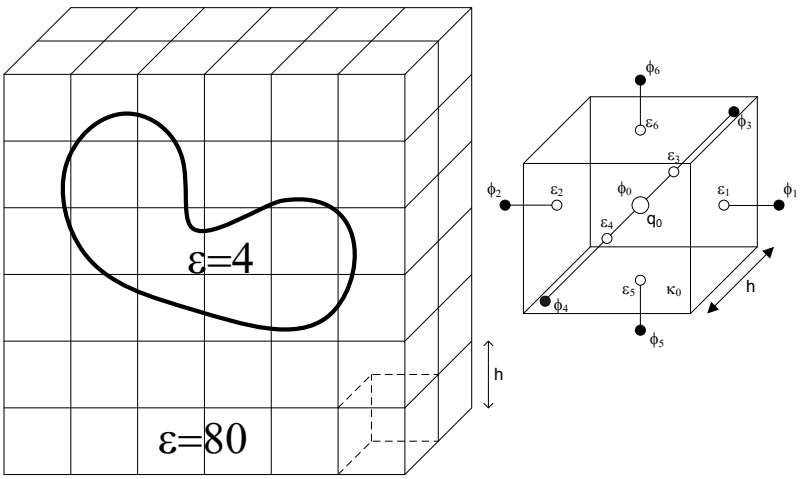


Fig. 1. Grid discretization associated with PBE model of electrostatics

Ligand is a small molecule with a limited number of conformational degrees of freedom (up to 50). First, selected terminal atom was assigned three Cartesian coordinates ($x$, $y$, $z$), which describe the location of the ligand in space, and two angles ($\alpha$, $\beta$) describing the orientation of the base bond. Second, one dihedral angle $\psi$ was assigned for each single order non-terminal bond (conformational degrees of freedom, see Fig. 2). The structures of rings are assumed constant. The complete set of the above coordinates is referred to as ligand pose.

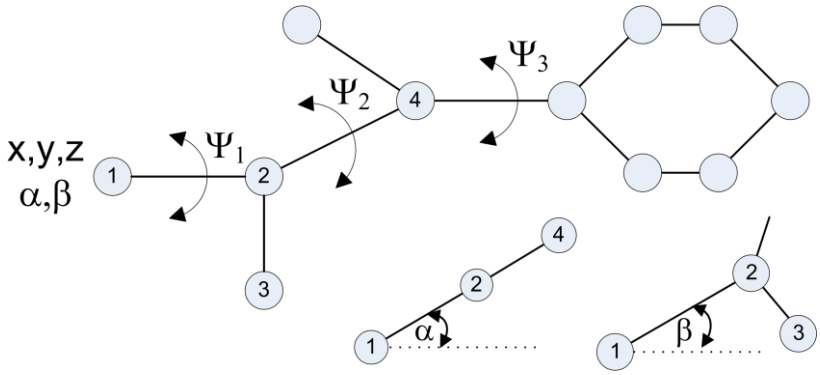


Fig. 2. Assignment of ligand degrees of freedom

## 3. Stochastic Roadmap Simulation

Stochastic Roadmap Simulation was proposed as efficient general framework for analysis of molecular motion and has been used with success for protein ligand interaction analysis [1, 8] and prediction of kinetic parameters in computational protein folding problem [2]. Let's briefly describe the idea of Stochastic Roadmap Simulation (SRS) first introduced by Apaydin et al. [1]. Each node of a roadmap represents one pose of a ligand. Formally, each pose of $n$ parameters is represented by a vector $\boldsymbol{q}$. The set of all possible poses forms the conformational space $C$. SRS assumes that the interactions are described by an energy function $E(\boldsymbol{q})$, which depends only on the pose $\boldsymbol{q}$ of the ligand. A pathway in $C$ represents motion of the ligand around protein. A roadmap may be considered a directed graph $G$ encoding many pathways in $C$. Each node of a roadmap is a randomly selected pose $\boldsymbol{q}$ from $C$ with associated energy $E(\boldsymbol{q})$. Each directed edge between two nodes $v_i$ and $v_j$ has associated weight, which is equal to the probability of transition between the two nodes. In order to construct a roadmap the algorithm samples $r$ poses, randomly and independently from $C$ (Fig. 3). Then for each node $v_i$ one finds $k$ nearest neighbors of that node according to selected metric (i.e. RMSD or Euclidean). After that a transition probability $P_{ij}$ is computed for every pair of neighboring nodes (Fig. 4). $P_{ij}$ calculation is based on difference in energy:

$$\Delta E_{ij} = E\left(v_i\right) - E\left(v_j\right) \tag{2}$$

between nodes $v_i$ and $v_j$ and assigned according to the formula:

$$P_{ij} = \frac{1}{N_i} e^{-(\Delta E_{ij}/k_B T)}, \quad \Delta E_{ij} > 0 \tag{3}$$

or

$$P_{ij} = \frac{1}{N_i}, \quad \Delta E_{ij} \le 0 \tag{4}$$

where $k_B$ - Boltzmann constant, $T$ - system temperature, $N_i$- number of neighbors of node $v_i$. The self-transition probability is defined as:

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij} \tag{5}$$

which ensures that the transition probabilities from any node sum up to 1.

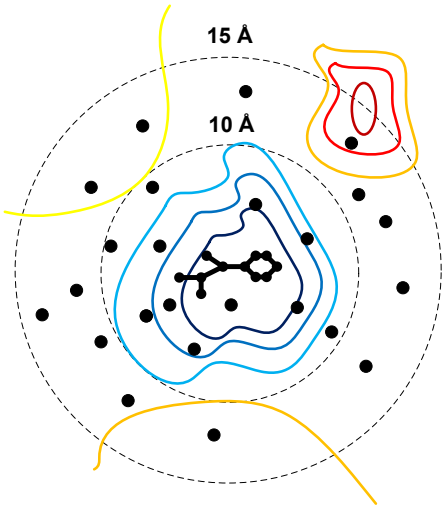


Fig. 3. Funnel of attraction around binding site, dots represent sampled ligand configurations

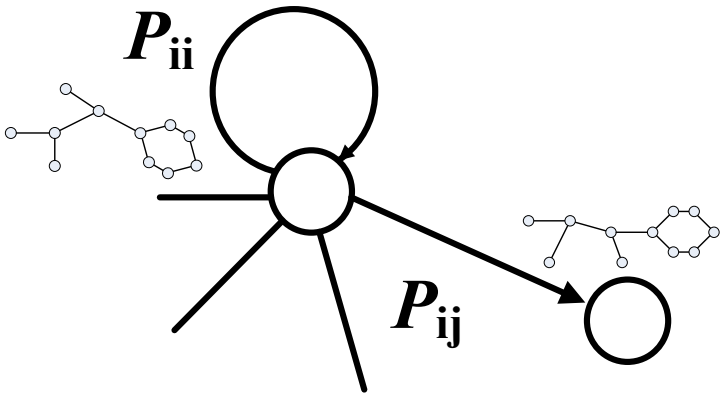


Fig. 4. Map building and assignment of transition probability

## 4. The Time to Escape and The Time to Absorption

Although it is possible to perform a simulation on a roadmap, which corresponds to a discrete version of the standard Monte Carlo method (discretization is defined by a roadmap) Apaydin et al. (Apaydin et al. 2003) suggest that usually it is not needed to generate individual trajectories on a roadmap but rather evaluate a parameter of interest. The time to escape (expressed as a number of simulation steps) from the funnel of attraction around the protein binding site is given as an example. Apaydin et al. propose the escape time as a measure of affinity of a ligand to a putative binding site. The funnel of attraction $F_i$ is defined as the set of poses within 10 Å RMSD of the bound pose (Fig. 3).

The directed graph $G$ with assigned transition probabilities $P_{ij}$ between nodes can be regarded as finite absorbing Markov chain (FAMC). The time to escape was calculated as mean or expected time to absorption in FAMC. FAMC in this case has a number of transient states which once left are never again entered and single absorbing state which once entered is never again left. The states represent various poses of ligand inside protein binding site. The nodes within 10 Å RMSD of the starting pose are considered transient while nodes sampled further away (up to 15 Å) - absorbing

states (Fig. 3). The starting pose should be usually that of the ligand bound to the protein which is assumed known form x-ray crystallography or docking. In general there can be more than one absorbing state but since it is not considered in which particular pose ligand left the binding site, it is possible to group all absorbing states into single state by summing transition probabilities from all transient states connected with any absorbing state. FAMC with single absorbing state has the following transition probability matrix:

$$\mathbf{P} = \begin{Vmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & 1 \end{Vmatrix} \tag{6}$$

where $r$-1 × $r$-1 matrix **Q** groups transition probabilities among $r$-1 transient states and **R** is a $r$-1 × 1 vector of probabilities of absorption starting from given transient state.

Expected value of the time to escape (7) defined as mean time to absorption starting from any transient state $X_0 = i$:

$$\tau_i = \mathrm{E}\left[T \middle| X_0 = i\right] \tag{7}$$

can be easily calculated using the first step analysis technique [1], from Markov chain theory [9] by solving the following linear system of equations (8):

$$\tau_i = 1 + \sum_{j=1}^{r-1} Q_{ij}\tau_j \tag{8}$$

$$Q_{ij} = P_{ij} \quad \text{for} \quad 0 \le i, j < r \tag{9}$$

where $\tau_i$– time to escape starting from $i$-th node. Alternatively the problem can be solved using fundamental matrix ***N*** approach [5]:

$$\mathbf{N} = \left(\mathbf{I} - \mathbf{Q}\right)^{-1} \quad \bar{\tau}_i = \sum_{j=1}^{r-1} N_{ij} \tag{10}$$

Both techniques are numerically equivalent and require computation of inverse of **Q** matrix. The mean time to escape was calculated using internal *Matlab 2010b* routines.

## 5. Experimental Binding Affinity and the Mean Time to Escape

The described approach was applied to enzyme-inhibitor complexes with experimentally determined affinity data deposited in the PDBBind database (release 2008) CoreSet [13]. The CoreSet consists of 210 structurally diverse protein-ligand crystallographic complexes with recorded affinity constant $K_i$ ($K_d$). The set was further divided according to ligand molecular properties i.e. molecular weight, number of rotatable bonds, net charge (Gasteiger), lipophilicity (AlogP) and binding affinity. For every protein-ligand complex 100 roadmaps of 1000 nodes (ligand poses) was generated and the time to escape, averaging the results over the 100 roadmaps, was calculated. The results show significant correlation between the computed mean time to escape and experimentally determined binding constant $K_i$ ($K_d$). The obtained Pearson's correlation coefficient $R$=0.39 for the whole dataset (Fig. 5).

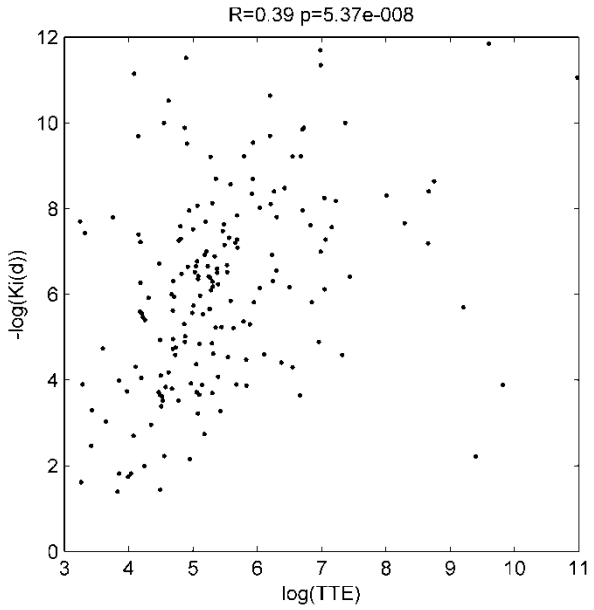


Fig. 5. Overall correlation for the whole PDBBind CoreSet (release 2008)

In author opinion the proposed scoring procedure should not be directly compared with popular scoring functions used in docking as the time to escape is averaged over many ligand poses, while scoring functions evaluations are based on single protein – ligand conformation. However, in a recent study performed on similar test set of 195 protein-ligand complexes (PDBbind CoreSet release 2013) Pearson's correlation coefficient ranges from $R$=0.221 to $R$=0.614 for the 20 scoring functions evaluated in terms of binding affinity prediction [7].

Although a correlation between the time to escape and equilibrium dissociation constant $K_d$ is reported, the physical concept of the time to escape is closer to protein-ligand complex residence time related to dissociation rate constant $k_{off}$. In the closed systems (in vitro assays) under constant ligand concentration $K_d$ and $k_{off}$ often strongly correlate. In the open system of human body, however, the residence time of protein-ligand complex begins to play more important role than ligand binding affinity alone [4, 11, 12].

The highest correlation was observed for ligands with small number (up to 3) of rotatable bonds (Fig. 6) $R$=0.67 ($p$=6.15e-11) and ligands of low (less than 300 D) molecular weight (Fig.7) $R$=0.57 ($p$=1.34e-8).

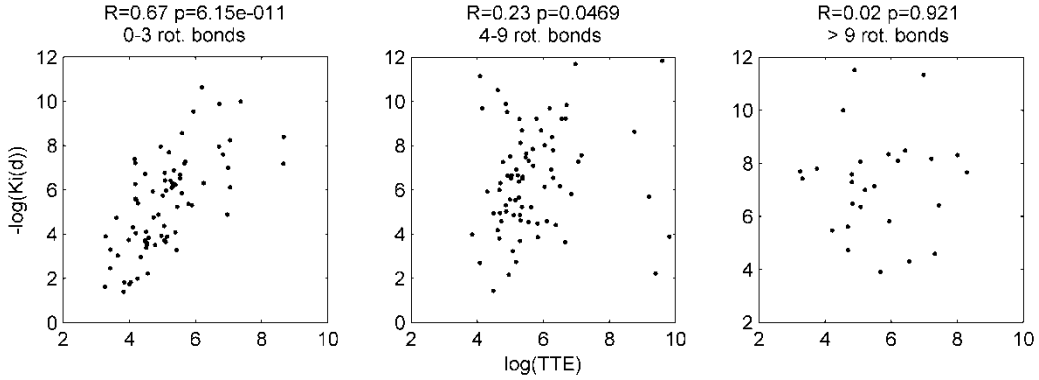


Fig. 6. Correlation for ligands with different number of rotatable bonds

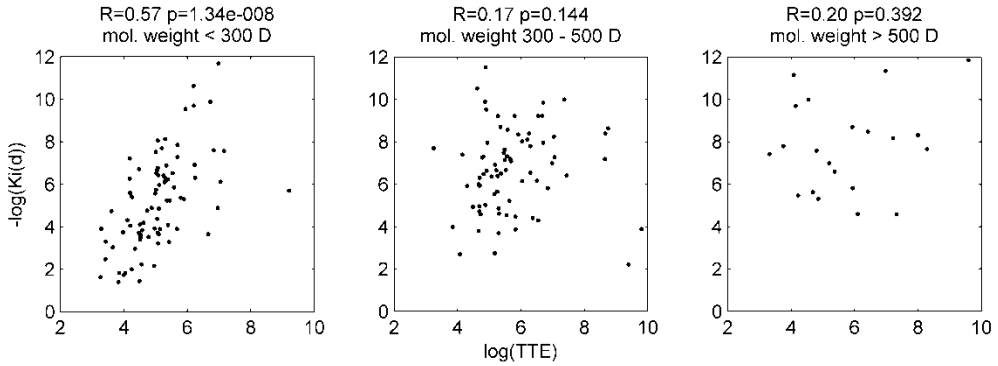


Fig. 7. Correlation for ligands of different molecular weight

Insignificant difference was observed in correlation concerning net charge of the ligands (Fig. 8), however for positively charged ligands $R$=0.53 ($p$=2.72e-4) while $R$=0.33 ($p$=4.05e-3) and $R$=0.43 ($p$=4.72e-4) for negatively charged and neutral ligands respectively.

Correlation coefficient $R$=0.35 ($p$=6.38e-4) for hydrophilic and $R$=0.4 ($p$=1.14e-4) ligands (Fig. 9).

Significant correlation was not observed for ligands with large number of rotatable bonds (Fig. 6) and ligands of higher molecular weights (Fig. 7) with high (pM) affinity (Fig. 10) which is also attributed to large number of internal degrees of freedom.

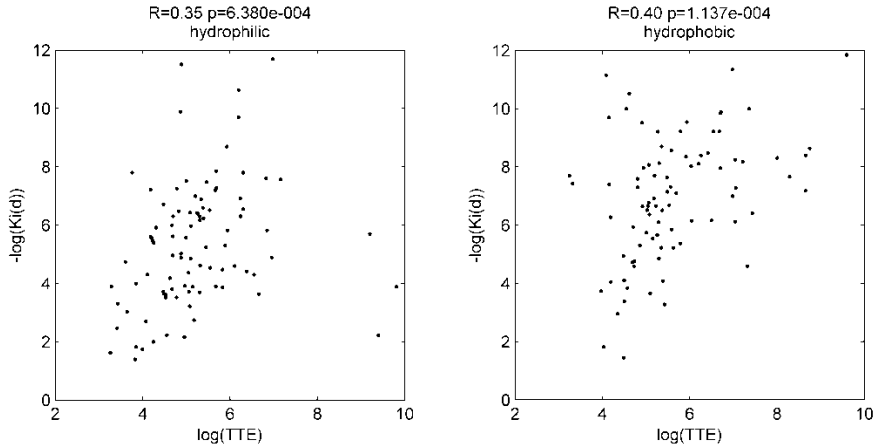


Fig. 8. Correlation for hydrophilic and hydrophobic ligands

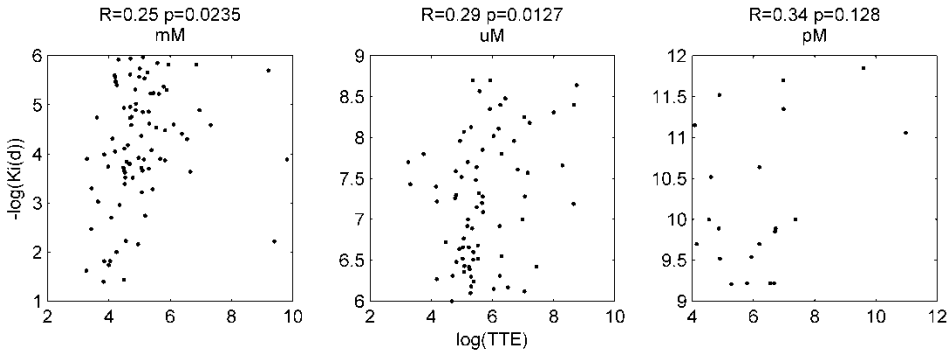


Fig. 9. Correlation for mM, uM and pM affinity ligands

Proposed approach apparently performs best for ligands with small number of rotatable bonds (internal degrees of freedom). The reason for decrease of performance is due to the fact that a map of 1000 nodes (ligand poses) is probably too small to capture the complex nature of molecular motion in high dimensional space. Unfortunately due to connectivity issues at the map building stage it was impossible to use larger maps in current implementation of SRS. Encouraged by interesting properties of the presented approach there is work in progress on new implementation which allows us to create maps with millions of nodes.

REFERENCES

1. Apaydin MS, Brutlag DL, Guestrin C, Hsu D, Latombe JC, Varma C.: Stochastic roadmap simulation: An efficient representation and algorithm for analyzing molecular motion. J Comp Biol 10, 2003, p. 257-281.
2. Chiang TH, Apaydin MS, Brutlag DL, Hsu D, Latombe JC: Using Stochastic Roadmap Simulation to predict experimental quantities in protein folding kinetics: Folding rates and phi-values. J Comp Biol 14(5), 2007.p. 578-593
3. Connolly ML: Solvent-Accessible Surfaces of Proteins and Nucleic-Acids. Science 221, 1983, p. 709-713.
4. Copeland RA, Pompliano DL, Meek TD: Drug–target residence time and its implications for lead optimization. Nat Rev Drug Discov 5, 2006, p. 730–739.
5. Kemeny JG, Snell JL.: Absorbing Markov Chains. In: Finite Markov Chains, Springer-Verlag, New York, 1983, p 43-68.
6. Leach AR.: Empirical Force Field Models: Molecular Mechanics. In: Molecular Modelling. Principles and Applications. Pearson Education Limited, Essex, 2001, p 207.
7. Li Y, Han L, Liu Z, Wang R.: Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results. J Chem Inf Model 54(6), 2014, p. 17.
8. Pacholczyk M., Kimmel M.: Exploring the landscape of protein-ligand interaction energy using probabilistic approach. J Comp Biol 18(6), 2011, p. 843-850.
9. Taylor HM, Karlin S.: Markov Chains: Introduction. In: An Introduction to Stochastic Modelling, Academic Press, San Diego, 1998, p 95-198.
10. Rocchia W., Alexov E., Honig B.: Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions. J Phys Chem B 105, 2001, p. 6507-6514.
11. Swinney DC: Biochemical mechanisms of drug action: what does it take for success? Nat Rev Drug Discov 3, 2004, p. 801-808.
12. Tummino PJ, Copeland RA: Residence time of receptor–ligand complexes and its effect on biological function. Biochemistry 47, 2008, p. 5481-5492.
13. Wang R, Fang X, Lu Y, Wang S.: The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. J Med Chem 47(12), 2004, p. 2977-2980.
14. Yuriev E, Ramsland PA: Latest developments in molecular docking: 2010-2011 in review. J Mol Recognit 26(5), 2013, p. 215-239