Monte Carlo on the manifold and MD refinement for binding pose prediction of protein-ligand complexes: 2017 D3R Grand Challenge.

Mikhail Ignatov[1,2,3,7], Cong Liu[2], Andrey Alekseenko[6,7], Zhuyezi Sun[5], Dzmitry Padhorny[1,2,3,7], Sergei Kotelnikov[1,2,3,4,7], Andrey Kazennov[4,7], Ivan Grebenkin[6,7], Yaroslav Kholodov[6,7], Istvan Kolosvari[5], Alberto Perez[2], Ken Dill[2], Dima Kozakov[1,2,3]

1. Department of Applied Mathematics and Statistics, Stony Brook University, NY, USA
2. Laufer Center for Physical and Quantitative Biology, Stony Brook University, NY, USA
3. Institute for Advanced Computational Sciences, Stony Brook University, NY, USA
4. Moscow Institute of Physics and Technology, Moscow, Russia
5. Department of Biomedical Engineering, Boston University, MA, USA
6. Innopolis University, Innopolis, Russia
7. Institute for Computer-Aided Design, Russian Academy of Sciences, Moscow, Russia

Abstract

Manifold representations of rotational/ translational motion and conformational space of a ligand were previously shown to be effective for local energy optimization. In this paper we report the development of the Monte-Carlo energy minimization approach (MCM), which uses the same manifold representation. The approach was integrated into the docking pipeline developed for the current round of D3R experiment, and according to D3R assessment produced high accuracy poses for Cathepsin S ligands. Additionally, we have shown that (MD) refinement further improves docking quality. The code of the Monte-Carlo minimization is freely available at https://bitbucket.org/abc-group/mcm-demo.

Introduction

Manifold-based representations of rotational and translational degrees of freedom of a ligand with respect to its receptor were shown to be effective for local optimization of energy function[1]. Manifold is a topological space, which locally resembles Euclidean space around each point. Each point can be assigned a space of vectors (tangent space), which intuitively denote the possible directions on the manifold. The basic idea of the approach is to map the tangent space at a current point on a rotational manifold to the neighboring points using exponential map[2]. The map is used to locally transform $SO(3) \times R^3$ group product into R^6 Euclidean space and treat energy as a 6 variable function. Thus, energy minimization can be done using any Euclidean gradient based function optimization algorithm, such as L-BFGS[3]. This results in unconstrained local optimization, which leads to efficient energy minimization algorithms[1,4]. Such local straightening of space also provides insights into the geometry of protein association landscapes[5].

This manifold approach can be extended to fully flexible local optimization by adding internal degrees of freedom to the above framework. Flexible ligands and receptor amino acid sidechains can be represented as a forest graph structure, i.e., set of tree graphs. Each degree of freedom can be seen as an S^1 manifold (circle) and again, using the concept of the exponential map, can be incorporated into the above manifold optimization approach by expanding R^6 to R^{6+d} , where d is a number of rotatable bonds in the molecule[4].

This representation can be used for medium-range Monte Carlo (MC) sampling, such as ligand pose optimization within the binding site. In this work we report implementation of such Monte Carlo Minimization approach and its application to ligand pose prediction as a part of our docking pipeline in the latest round of D3R. In addition, we have studied effects of short MD refinement on the quality of the prediction.

In this round of D3R the docking part of the challenge consisted of pose prediction for 24 Cathepsin S ligands. The experiment contained two stages. In the first stage (1A) the groups were required to predict the poses of the ligands without any additional information, in the next stage (1B) the organizers provided crystal structures of the complexes with ligands removed.

Cathepsin S system was previously well studied[6], hence multiple bound ligand structures were available in the PDB. To account for this, we have developed the following protocol. First, the ligand was aligned to the closest known bound ligand in the proposed binding site, then manifold MCM was performed and accepted poses were clustered. Additionally, for the stage 1B short MD simulations were run starting from the aligned poses and the results were added to the MCM output. Each conformation in the resulting set was minimized and reranked using Vina-based score[7] and 5 best poses were used for submission.

The designed protocol demonstrated good performance on a large number of targets. In this paper we describe our algorithm, evaluate its performance on the provided targets and demonstrate its effectiveness in small molecule docking. The code of the MCM approach is freely available at https://bitbucket.org/abc-group/mcm-demo.

Methods

Implementation of the Manifold Monte Carlo sampling approach

Method Overview

Here we provide a brief description of the methodology used for Monte Carlo based optimization of ligand docking poses. The early prototype of the protocol was tested in the previous D3R round[8]. The source code of the implementation is available at https://bitbucket.org/abc-group/mcm-demo.

The underlying assumption used in our modeling approach is that the changes in covalent bond-lengths and bond-angles can be neglected, and the molecule can be viewed as a set of rigid molecular clusters interconnected with rotatable bonds. Configurational state of any molecule can then be described in terms of 6 rigid body and d internal degrees of freedom, where d is the number of torsions associated with rotatable bonds. Implementation-wise, an aggregate of rigid and rotatable elements forming a molecule is represented as a torsion tree data structure[4] (see figure 1). In general, this type of representation is common for small molecule docking methods[7,9].

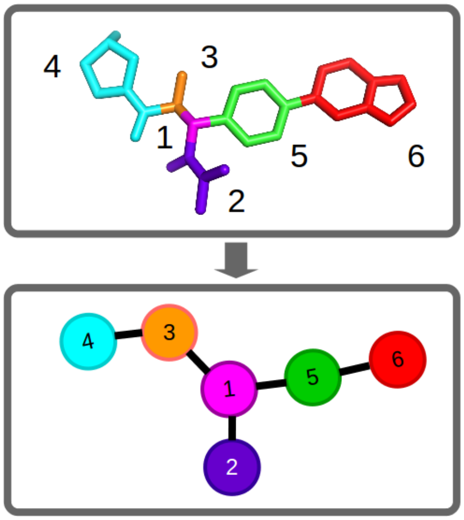


Figure 1. Representation of a small molecule as a torsion tree. The ligand is decomposed into a set of rigid fragments interconnected by rotatable bonds, thus forming a torsion tree.

One distinctive feature of our docking method is related to its ability to perform minimization of molecular poses in internal and rigid body manifold coordinates, as opposed to performing full-atomic minimization. This becomes possible due to the fact that the space of rigid body motions, $SO(3) \times R^{\prime}$, the space of internal torsional motions, $T^{-} = \{S^{+} \times S^{+} \times \ldots \times S^{+}\}_{-}$, and$_{45678}$ their direct product $SO(3) \times R^{\prime} \times T^{-}$ are all Riemannian manifolds, or locally Euclidean topological spaces, and thus energy minimization can be formulated directly on this space as a manifold optimization problem. We have previously shown that although the global geometry of this manifold is not trivial, the local Euclidean property allows to use conventional optimization techniques once suitable parametrization has been chosen, while smaller dimensionality provides a significant speed advantage over all-atom minimization approaches[1,4].

Energy Function

The method includes a variety of potential energy functions, including bonded ($E_{:;<-}$) and Van der Waals ($E_{=->}$) terms, analytical continuum electrostatics model ($E_{?@7}$)[10], knowledge-based hydrogen bonding ($E_{A:;<-}$). In D3R 2017, we also used a quadratic geometric restraint potential ($E_{B7;6}$) to constrain the positions of certain atoms (see MCM configuration below for details). The total interaction energy used for small molecule docking is computed as a linear combination of these individual terms:

$$E_{:;<-} = w_{:;<-} E_{:;<-} + w_{=->} E_{=->} + w_{?@7} E_{?@7} + w_{A:;<-} E_{A:;<-} + w_{B7;6} E_{B7;6} \qquad (1)$$

where $w_{:;<-}$, $w_{=->}$, $w_{?@7}$, $w_{A:;<-}$, $w_{B7;6}$ denote the corresponding weights. All weights were kept equal with hydrogen bonding disabled, which was proved to be effective during our previous studies:

$$w_{:;<-} = w_{=->} = w_{?@7} = w_{B7;6} = 1.0, \; w_{A:;<-} = 0.0 ,$$

$E_{.;<-}$ includes standard bonded terms, such as bond, angle, dihedral and improper terms,$E_{=->}$ is computed using linearized repulsion with 1-4 interaction scaling factor and $E_{?@7}$contains Coulomb interaction and a non-polar solvation term. Force field parameters were obtained from GAFF[11] Amber[12] and charges were assigned according to AM1-BCC protocol[13] using antechamber module.

MCM Steps

When applied to ligand docking, the general Monte Carlo minimization protocol is performed as a series of consecutive steps, each composed of 3 stages (see figure 2). The first stage is perturbation of the ligand conformation. We currently use a simplistic move set consisting of random rigid body moves and random perturbations of all ligand torsions. The second stage is sliding: ligand is being slid into contact with the receptor along the line connecting the geometrical centers of the ligand and the binding site. In this stage, the direction of ligand sliding is towards the protein if there is not sufficient contact between receptor and ligand and away from the protein if ligand and receptor clash. The last stage is local energy minimization performed with a manifold-based optimization algorithm. The pose generated in these 3 stages is accepted or declined using the Metropolis criterion.

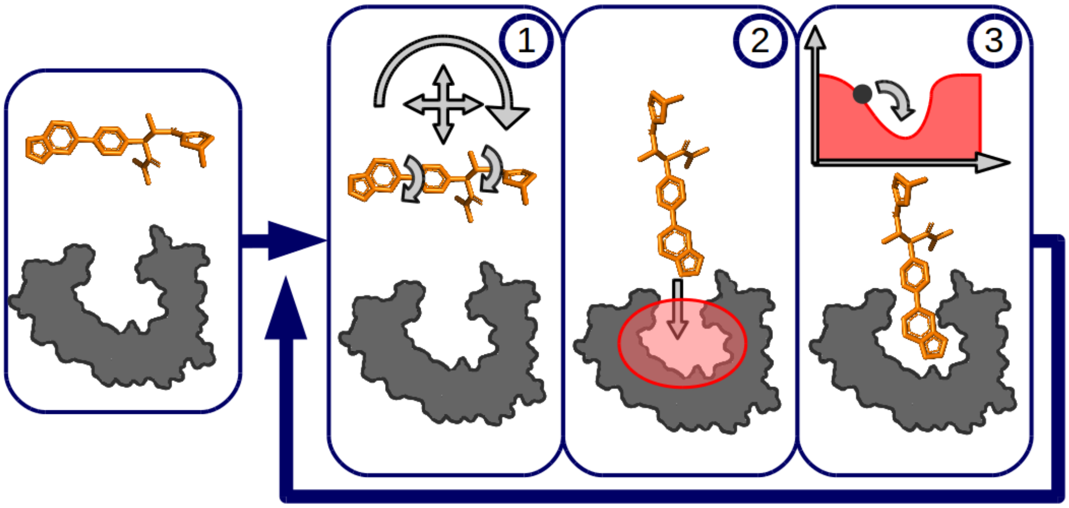


Figure 2. The general scheme of the Monte Carlo minimization protocol. Each MCM simulation consists of a series of steps each divided into 3 stages. In the first stage (1), a trial ligand pose is generated using random rigid body and torsional perturbations. In the second stage (2), receptor and ligand are slid towards/ away from each other to bring the molecules into contact or resolve severe clashes. In the third stage (3), the trial pose is refined using the local manifold-based minimization algorithm. The resulting binding conformation is accepted or declined based on the Metropolis criterion.

Challenge targets

Current D3R challenge included prediction of binding poses of 24 small organic molecules in complex with Cathepsin S, a lysosomal cysteine protease which plays a role in degrading proteins into peptides for presentation[6] on Major Histocompatibility Complex (MHC). In stage 1A the participants were provided with 2 Cathepsin S crystal structures and 24 ligands in SMILES format. In stage 1B the D3R team released 24 complex structures with ligand removed, but everything else present, including crystal water. The objective in the both parts was to predict binding poses for each of the 24 ligands, given provided receptor structures.

Protocol steps

For the 2 stages we used slightly different pose prediction protocols. Each protocol consisted of 3 steps: Receptor and ligand preparation, Pose prediction and Ranking (see figure 3).

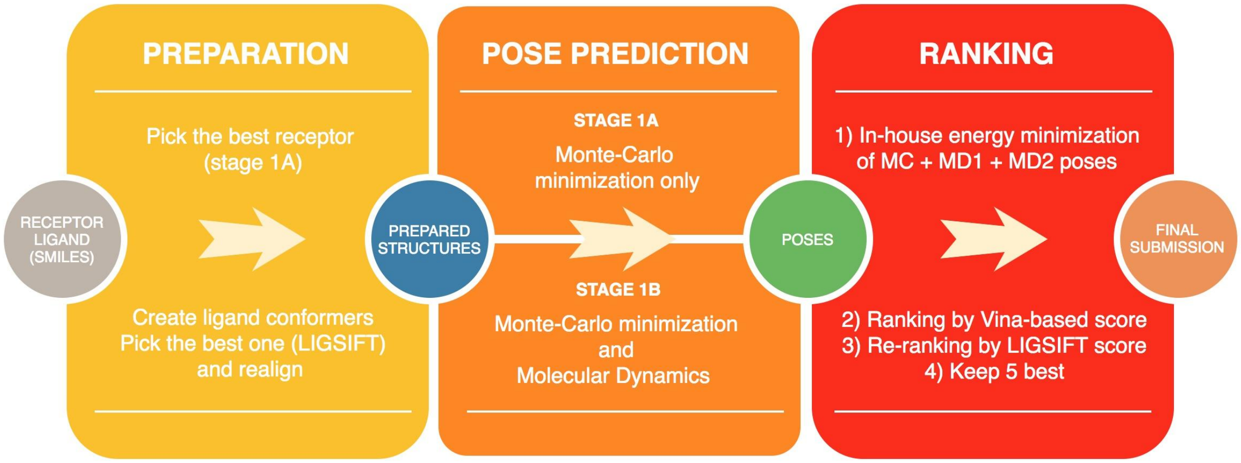

PREPARATION
Pick the best receptor (stage 1A)
Create ligand conformers Pick the best one (LIGSIFT) and realign
RECEPTOR LIGAND (SMILES)
PREPARED STRUCTURES
POSE PREDICTION
STAGE 1A
Monte-Carlo minimization only
STAGE 1B
Monte-Carlo minimization and Molecular Dynamics
POSES
RANKING
1) In-house energy minimization of MC + MD1 + MD2 poses
2) Ranking by Vina-based score
3) Re-ranking by LIGSIFT score
4) Keep 5 best
FINAL SUBMISSION

Figure 3. Block-scheme of the protocol.

Molecule preparation

Receptor preparation

In stage 1A two representative receptors with bound SO4 and DMSO were provided by the organizers. Using BLAST[14], we identified 31 structures of Cathepsin S in the Protein Data Bank and manually examined them to eliminate those which did not contain any ligand or had a modified environment in proximity to the binding site. For each of the 24 targets in D3R 2017 we found the most structurally similar ligand among the remaining complex structures using RDkit[15] package, reducing the final receptor set to 4 co-crystal structures: 3iej, 3kwn, 3mpe, 3mpf. The purpose of this selection was to choose a receptor with the most appropriate interface sidechain conformations for each target, however, as the visual examination suggested, in all of the discovered Cathepsin structures interface sidechains remained relatively fixed. The most suitable co-crystal structure (one of the 4) was assigned to each target, which we will further denote as reference.

Ligand conformers generation

For each of the 24 targets 1000 conformations were generated from SMILES strings using RDkit. Each conformer was aligned to the corresponding bound ligand from the reference co-crystal structure using LIGSIFT[16] and the most similar conformation (according to LIGSIFT score) was used for further processing. Most of the target ligands had scaffold very similar to their assigned reference bound ligands. One of the common features among most of the molecules was a central single or double ring. Since most of the reference bound ligands shared the location of the ring(s) in the binding site, we realigned the ab initio conformations to best fit the rings and the general scaffold of the reference bound ligand.

Pose prediction methods

While the starting ligand conformations and scoring scheme remained the same for both stages, we used slightly different protocols for the docking step. In both stages for docking we used Monte-Carlo minimization algorithm using the starting conformation obtained in the ligand generation step. However, in stage 1B we supplemented it with MD simulations as an alternative docking method.

MCM configuration

For the 2017 D3R challenge, the pre-aligned ligand structures were used as starting points for the MCM algorithm. Rigid body translation steps of up to 0.1 Angstrom, rotations of up to 5.7 degrees and torsional perturbations up to 1.57 radian were used. Additionally, harmonic restraints were imposed on the central ring of a ligand to keep it in the binding pocket. While the sidechain packing is generally important for unbound docking[9], visual inspection of the template X-ray structures suggested, that the sidechain conformations in the binding site remained relatively stable given various ligands. Therefore, we disabled sidechain flexibility throughout the simulation. For each target, simulations of 2000-10000 steps were performed, the accepted poses were clustered using Butina clustering algorithm provided with RDkit with 2.0Å RMSD threshold. Best-scoring poses (scored using MCM energy function from equation 1) in each cluster were retained for a final re-ranking step (see Ranking).

Molecular dynamics simulations

In the stage 1B, where crystal structures of Cathepsin S corresponding to each of the 24 ligands were provided, MD simulations were used as a second approach for pose prediction. The resulting MD poses were mixed with the MC predictions and the combined set was considered for scoring. Two different MD protocols were used.

MD protocol 1

A simple protocol was devised to accommodate the deadlines. A single structure coming from the MC protocol was used to carry out restrained simulations. Restraints were imposed based on the structure of the receptor and solvent molecules as given by the organization. Charge state was determined based on the ligand structure and charges for MD assigned through the antechamber module in Amber[12] using the AM1-BCC protocol[13]. The system is solvated with a TIP3P[17] water box using tleap and a 5Å buffer region between proteins and crystallographic waters and the edge of the box. The system was neutralized using Na+ or Cl- as needed [18,19]. Ligand parameters come from the GAFF force field[11] and protein parameters from FF14SB[20].

The MD protocol included a multistage minimization and equilibration protocol described previously[21] for 2.05 ns. MD production runs were carried through 100 ns with at 2fs timestep, at 298.15 K and 1atm. Hard restraints (50 kcal/mol Å ^{2}) were applied to protein heavy atoms and crystallographic waters. Soft restraints (0.5 kcal/mol Å ^{2}) were used for the ligand. This keeps poses close to the docking conformation but allows a certain degree of relaxation. Finally, DBSCAN[22] is used to cluster each trajectory (distance cut-off of 1.5Å and population cut-off 20). The centroid of the most populated cluster of each target is extracted and used further for energy minimization.

MD protocol 2

Molecular dynamics simulations were performed using the 2017-4 GPU version of Desmond[23]. We used the OPLSAA_2005 force field and SPC water in our simulations. Every simulation started with the standard Desmond relaxation protocol as defined in the Maestro GUI.

The production runs were configured NPT using Nose-Hoover chain with a 1 ps relaxation time for thermostat (single temperature group), and Martyna-Tobias- Klein barostat with 2 ps relaxation time and isotropic coupling. We utilized a RESPA integrator with ∆t = 2.5 fs for bonded and near nonbonded interactions and ∆t = 7.5 fs for far nonbonded interactions. The particle-mesh Ewald algorithm was used with periodic boundary conditions to compute long-range electrostatic interactions with the real space cutoff set to 9 Å for both electrostatic and van der Waals interactions. Water molecules were constrained with SHAKE. An aggregate of 1 μs production sampling was accomplished on each system by running 10 independent 100 ns simulations starting with different random initial velocities. The resulting trajectories were concatenated and subjected to clustering. A greedy clustering algorithm, which finds nearest neighbors within a certain radius, uses pairwise fitted interface root mean square deviation (RMSD) matrices as distance measures. The clustering radii were tailored to individual trajectory with respect to pairwise RMSD distribution, based on our previous experience[24]. The clustering radii we eventually applied ranged from 1.0 A to 2.5 A. The top 10 cluster centers with the largest cluster populations were supplied as suggested models coming from these MD simulations and were further evaluated. Due to the time limits we ran simulations only for the first 8 targets.

Ranking

MCM and MD (in the stage 1B) conformations were combined into a single set and each was relaxed by our energy minimization protocol with fixed C-alpha atoms using L-BFGS algorithm. Both minimized and non-minimized conformations were ranked by affinity values produced by AutoDock Vina[7] (--score_only flag), which are computed using intermolecular energy terms only. Five poses with the best affinity scores were subsequently ranked by similarity to the pose of the native ligand using LIGSIFT scores.

Results and discussion

The results of our docking protocol for stages 1A and 1B are provided in figure 4 and figure 5. Our method demonstrates good performance for the majority of the targets, excluding several ligands, for which our submission did not contain low RMSD poses. Here we describe some of the successfully predicted cases as well as those, for which low RMSD poses were not present in the final submission.

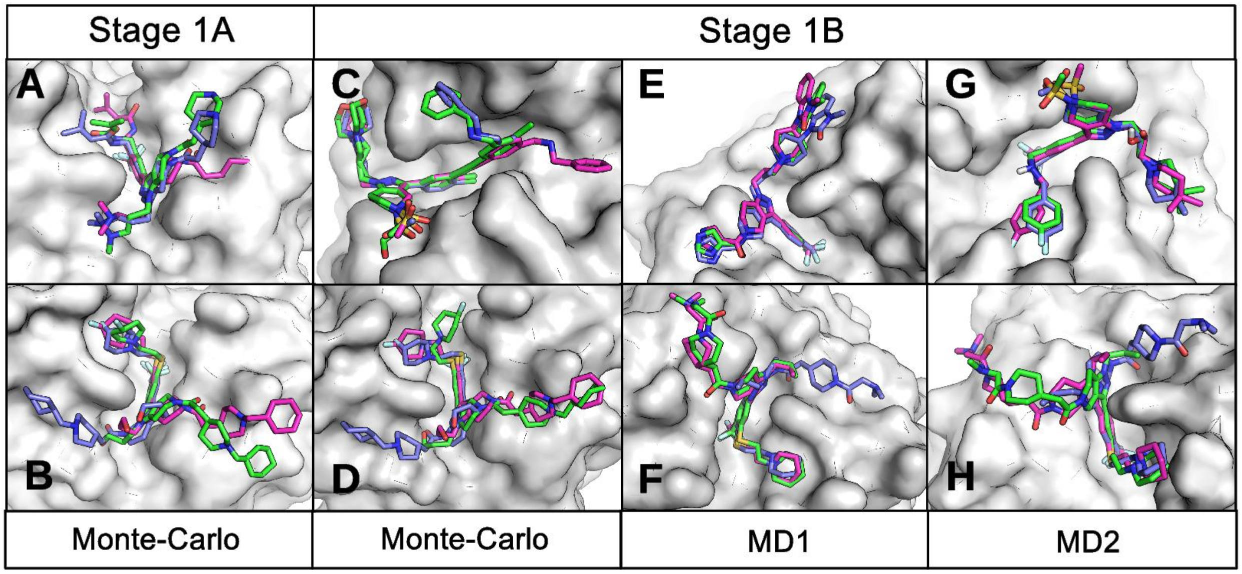


Figure 4. Predictions for different targets. Starting ligand conformation - magenta, native conformation - purple, final prediction - green. Pairs of cases for each protocol contain an example of successful prediction and an example, where the protocol failed to find a near-native conformation.

(A - target 16, B - target 9): result of MCM, stage 1A.
(C - target 15, D - target 9): result of MCM, stage 1B.
(E - target 3, F - target 7): result MD1, stage 1B.
(G - target 8, H - target 7): result MD2, stage 1B.

MCM

The Monte-Carlo protocol has corrected a large number of poses towards the native conformation in both of the stages, being however, more efficient in the stage 1A.

One successful example is target 16 from stage 1A (figure 4A). In the starting structure (magenta) piperidine group and the opposing isopropyl tail had orientation different from the native one (purple). MCM has fixed the starting conformation by placing the tails at the orientation close to the native one. Another case from the stage 1B is target 15 (figure 4C). It had improperly oriented phenyl tail, which was corrected by MC algorithm, which rotated it by about 180 degrees.

However, there are 3 cases (targets 7, 9 and 14) where all submitted structures in both stages have relatively high RMSD due to a common cause. Target 9, for instance, shows in both A and B stages (figures 4B and 4D) 180 degrees flip of nearly the whole (except 4-fluoropiperidine tail) ligand molecule (green) with respect to the native structure (purple). The reason is that the starting pose (magenta) as well as the reference bound ligand had a binding mode reverse to the native one and spring-like restraints used in the protocol prevent MC from placing the target into a near-native state, which would require a 180-degree rotation.

MD protocol 1

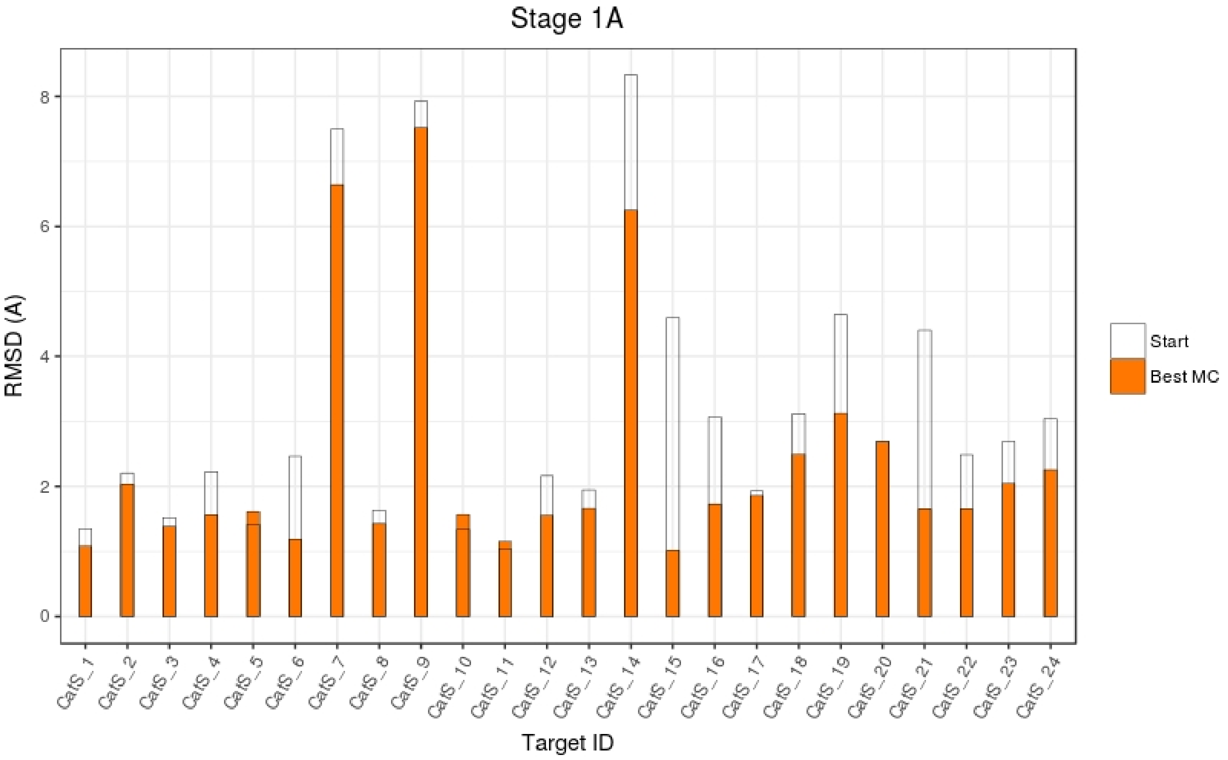
The MD protocol was designed to correct small issues with the structures. Mostly in terms of sidechain flexibility in the receptor and ligand reorientation. Hence, in figure 5B for most targets with an initial RMSD lower than 5Å there is refinement (compare the grey lines to the blue and orange ones). The yellow bars represent the centroid of a cluster, a single structure to represent them all, and hence is often not the best we can pick up in the ensemble. Minimizing this structure results in further improvement -- usually better, than just minimizing the starting pose. In the initial generation of the starting poses, the receptor is kept rigid. The ligand preparation step used to create the starting conformations results in overlap of atoms. These steric clashes are corrected with the minimization and equilibration protocol described in methods. Subsequence MD leads to refinement in some cases, the best case being target 3 (see figure 4E).

Due to the positional restraints, higher reorientation was not possible. Hence, we can see in figure 5 that whenever we start from a pose that is far away from the correct position we are never able to recover from the initial error (targets 7, 9 and 14). Figure 4F exemplifies this for target 7, where the docking site was correct for half of the ligand, but the other half was rotated roughly 180 degrees with respect the right pose.

In future we plan to have a more flexible approach. Instead of placing cartesian restraints on the ligand position we will satisfy a subset of contacts found in the initial ligand/receptor conformation. This way the method has a chance to refine targets where the initial poses are only partially correct (targets 7, 9 and 14). We have had success with this approach outside of the D3R competition, where we were not bound by stringent deadlines using the MELD approach[25,26].

MD protocol 2

As described previously, we ran 100 ns simulations and generated representative structures with clustering. Such short MD simulations would not have allowed drastic transformation of conformations but would serve as refinement of the starting structures. The results (see figure 5, stage 1B) demonstrate the effects and limitations of such MD simulations. For instance, in target 7 (figure 4H) and target 8 (figure 4G), poses generated by MD refinement were ranked highly according to our scoring scheme and were included in the final submission. However, in target 7, none of the top ranked models had the correct binding mode as the native structure. It seems that the starting structure has rotated 180 degrees away from the X-ray structure. Following MD refinement, the predicted models all ended in a similar binding mode as the starting structure (magenta), overlapping poorly with the correct pose (purple). On the other hand, MD simulations performed well with target 8. Figure 4G shows that the starting structure for MD (magenta) was in the same orientation as the crystal structure (purple). MD refinement reproduced the correct binding mode and also further adjusted the positioning of branches and the conformations of rings.



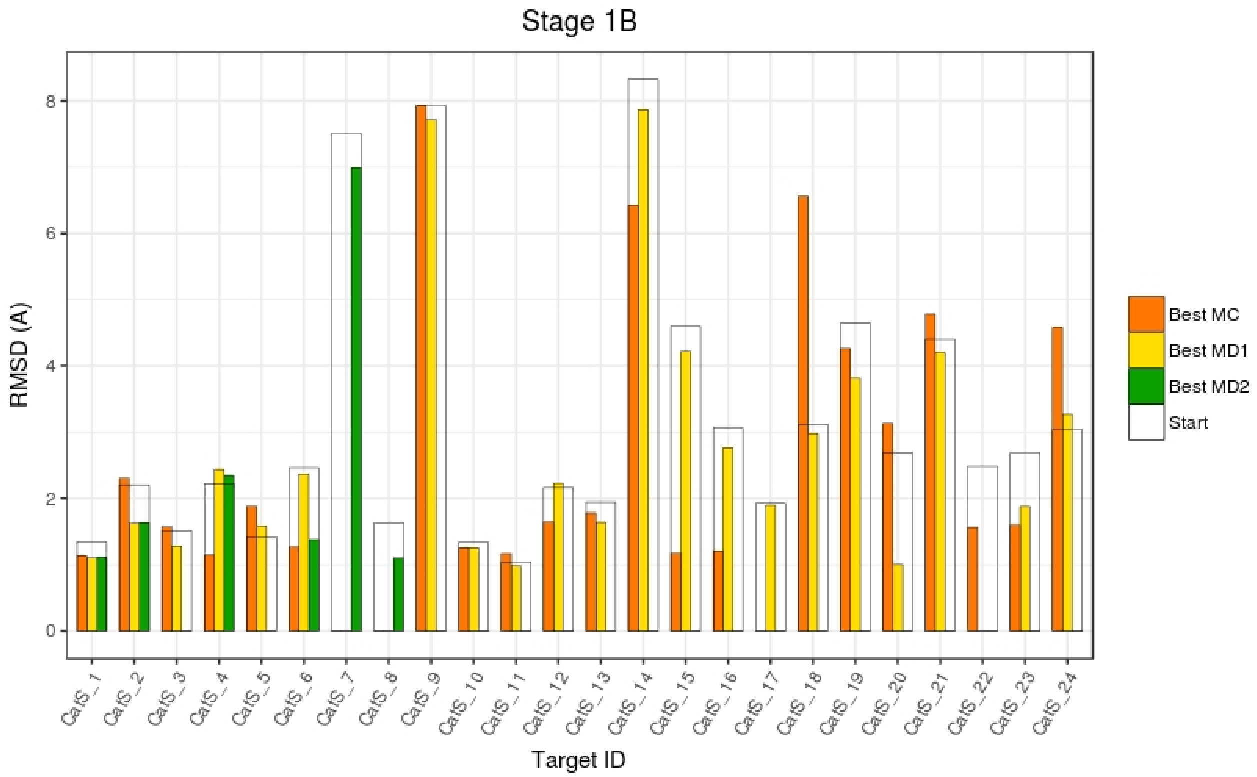


Figure 5. RMSD of predicted poses for stages 1A and 1B. Stage 1B: If orange, yellow or green column is absent, it means that the corresponding method was not represented among the top5 poses in the submission of the particular target.

Summary

Overall, MCM performed well in stage 1A finding a lower RMSD pose in almost every case (figure 5A), with median of closest RMSD values 1.52Å (figure 6). In stage 1B we were provided with receptor X-ray structures, which combined with using both MC and MD protocols further improved the accuracy of the models by 0.3Å to 1.23Å (median of closest RMSD values). MD simulations were designed in order to refine the ligand's starting conformation, other than strongly perturb it, which resulted in overall improvement of the predicted set (figure 5B). For some cases in stage 1B, MC protocol requires some adjustments to account for the presence of water molecules.

Conclusion

In this work we report implementation of manifold MCM based approach and its application to the latest round of D3R Grand Challenge. Additionally, we have studied the effect of MD refinement on the submission accuracy. The designed protocol was placed among the top performers by median overall RMSD (closest among top 5 poses for each target) in the current challenge (see figure 6). For the majority of the targets our submission included many predictions below 2Å RMSD (1A: 17/24, 1B: 17/24). However, some targets (7, 9 and 14) had binding mode different from that of ligands in the reference co-crystal structures, leading to incorrect orientation of aligned starting poses. In future, we plan to improve our protocol by using a number of different starting poses, which could potentially allow to overcome this issue.

MD protocols introduced some improvement, but the moves were localized, hence starting MD with multiple MC poses instead of running it in parallel could potentially improve the results as well. Additionally, the pose scoring can be improved[27,28], since in many cases our lowest RMSD pose was within top 5 models, rather than top 1.

The code of the manifold MCM is available at https://bitbucket.org/abc-group/mcm-demo.

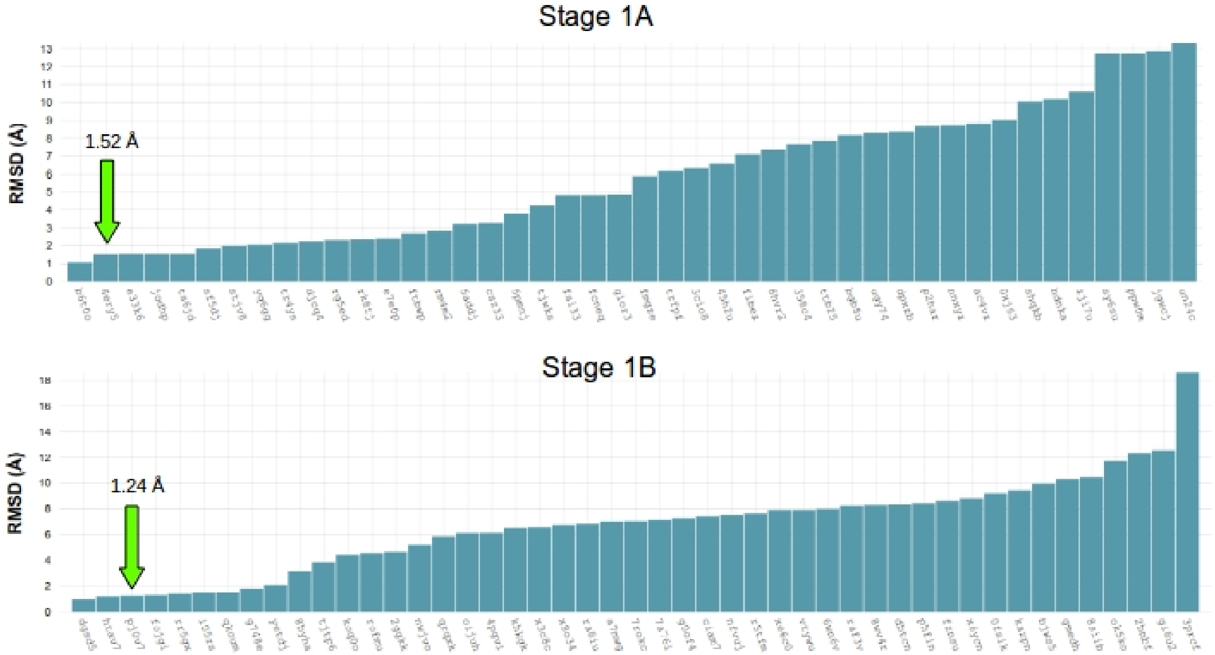


Figure 6. Submissions sorted by median of lowest RMSD values for each target (official results released by D3R organizers). The green arrow indicates our submission.

Acknowledgements

This work was supported by Grants NIH R21 GM127952, NIH R01 GM12581301, NSF AF 1816314, NSF AF 1645512 and RSF No 14-11-00877.

References

1. Mirzaei, H. et al. Rigid Body Energy Minimization on Manifolds for Molecular Docking. J. Chem. Theory Comput. **8,** 4374–4380 (2012).

2. Hermann, R. Differential Geometry, Lie Groups, and Symmetric Spaces (Sigurdur Helgason). SIAM Rev. **22,** 524–526 (1980).

3. Liu, D. C. & Nocedal, J. On the limited memory BFGS method for large scale optimization. Math. Program. **45,** 503–528 (1989).

4. Mirzaei, H. et al. Energy Minimization on Manifolds for Docking Flexible Molecules. J. Chem. Theory

Comput. **11,** 1063–1076 (2015).

5. Kozakov, D. et al. Encounter complexes and dimensionality reduction in protein-protein association. Elife **3,** e01370 (2014).
6. Wilkinson, R. D. A., Williams, R., Scott, C. J. & Burden, R. E. Cathepsin S: therapeutic, diagnostic, and prognostic potential. Biol. Chem. **396,** 867–882 (2015).
7. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem. NA–NA (2009).
8. Padhorny, D. et al. Protein–ligand docking using FFT based sampling: D3R case study. J. Comput. Aided Mol. Des. **32,** 225–230 (2017).
9. Meiler, J. & Baker, D. ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. Proteins **65,** 538–548 (2006).
10. Schaefer, M. & Karplus, M. A Comprehensive Analytical Treatment of Continuum Electrostatics. J. Phys. Chem. **100,** 1578–1599 (1996).
11. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case, 'Development and testing of a general amber force field'Journal of Computational Chemistry(2004) 25(9) 1157-1174. J. Comput. Chem. **26,** 114–114 (2005).
12. Case, D. A. et al. AMBER 2016 (University of California, 2016). Google Scholar
13. Jakalian, A., Bush, B. L., Jack, D. B. & Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. J. Comput. Chem. **21,** 132–146 (2000).
14. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. **215,** 403–410 (1990).
15. Landrum, G. RDKit: Open-source cheminformatics. Available at: http://www.rdkit.org.
16. Roy, A. & Skolnick, J. LIGSIFT: an open-source tool for ligand structural alignment and virtual screening. Bioinformatics **31,** 539–544 (2015).
17. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. **79,** 926–935 (1983).
18. Joung, I. S. & Cheatham, T. E., 3rd. Determination of alkali and halide monovalent ion parameters for

use in explicitly solvated biomolecular simulations. J. Phys. Chem. B **112,** 9020–9041 (2008).

19. Li, P., Song, L. F. & Merz, K. M., Jr. Systematic Parameterization of Monovalent Ions Employing the Nonbonded Model. J. Chem. Theory Comput. **11,** 1645–1657 (2015).
20. Maier, J. A. et al. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J. Chem. Theory Comput. **11,** 3696–3713 (2015).
21. Hornak, V., Okur, A., Rizzo, R. C. & Simmerling, C. HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. Proc. Natl. Acad. Sci. U. S. A. **103,** 915–920 (2006).
22. Ester, M., Kriegel, H.-P., Sander, J., Xu, X. & Others. A density-based algorithm for discovering clusters in large spatial databases with noise. in Kdd **96,** 226–231 (1996).
23. Bowers, K. et al. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. in ACM/IEEE SC 2006 Conference (SC'06) (2006). doi:10.1109/sc.2006.54
24. Kozakov, D., Clodfelter, K. H., Vajda, S. & Camacho, C. J. Optimal clustering for detecting near-native conformations in protein docking. Biophys. J. **89,** 867–875 (2005).
25. Morrone, J. A. et al. Molecular Simulations Identify Binding Poses and Approximate Affinities of Stapled α-Helical Peptides to MDM2 and MDMX. J. Chem. Theory Comput. **13,** 863–869 (2017).
26. Morrone, J. A., Perez, A., MacCallum, J. & Dill, K. A. Computed Binding of Peptides to Proteins with MELD-Accelerated Molecular Dynamics. J. Chem. Theory Comput. **13,** 870–876 (2017).
27. Grudinin, S., Kadukova, M., Eisenbarth, A., Marillet, S. & Cazals, F. Predicting binding poses and affinities for protein - ligand complexes in the 2015 D3R Grand Challenge using a physical model with a statistical parameter estimation. J. Comput. Aided Mol. Des. **30,** 791–804 (2016).
28. Yan, C., Grinter, S. Z., Merideth, B. R., Ma, Z. & Zou, X. Iterative Knowledge-Based Scoring Functions Derived from Rigid and Flexible Decoy Structures: Evaluation with the 2013 and 2014 CSAR Benchmarks. J. Chem. Inf. Model. **56,** 1013–1021 (2016).