D3R Grand Challenge 2015: Dataset Description and Instructions

Revised October 17, 2015

Overview

The Challenge involves protein-ligand datasets for two targets: Heat Shock Protein 90 (HSP90) and mitogen-activated protein kinase kinase kinase kinase (sic; MAP4K4). For each target, Stage 1 of the Challenge is to predict the ligand poses of the available crystal structures and also to predict or rank the potencies of all ligands, including those for which crystal structures are not available. After Stage 1 has closed, all available co-crystal structures will be made public, and the Stage 2 Challenge is to repeat the affinity predictions or rankings with the additional disclosed ligand-pose information.

The HSP90 Challenge dataset, which was generated and generously provided by AbbVie and Prof. Heather Carlson and CSAR colleagues at the University of Michigan, comprises:

- IC50 data for 180 compounds in three chemical series: aminopyrimidines, benzimidazolones, and benzophenone-like
- potency range 5.2 nM to 50 uM, and 33 compounds with no detectible activity,
- 8 co-crystal structures, resolution 1.6 – 1.95 Å, including examples from each chemical series; two are unblinded (4YKR and 4YKY).

Three subsets have been selected, of 4, 5 and 10 ligands, which are chemically similar and lend themselves to the calculation of relative binding affinities by alchemical methods, such as free energy perturbation.

The MAP4K4 Challenge dataset, which was generated and generously provided by Genentech, comprises

- 30 co-crystal structures of chemically varied ligands
- co-crystal structure resolutions range from 1.59 – 3.04 Å, with 26 below 2.6 Å
- IC50 data for 18 of these ligands, potency range from 3 nM to 10 uM
- 25 ligands have MW <350 kDa.

Because of the variety of conformations of the crystal structures, the smaller number of compounds with measured IC50 values, and the fact that more than half of the compounds are fragment-like in size (MW < 350 kDa), this challenge focuses largely on cross-docking pose-prediction.

For both datasets, we are providing prepared protein structures, compounds in the form of canonical SMILES strings, and information on the experimental conditions for the crystallography and binding measurements. The SMILES strings and ligand ID numbers for ligands with co-crystal structures are listed in Appendix below.

Due dates

Your Stage 1 predictions (poses + ligand affinities or rankings) must be uploaded to the D3R website before midnight on November 16th for HSP90 and December 16th for MAP4K4. The experimental ligand-protein poses and the corresponding IC50s will be released immediately after Stage 1 closes for the respective target.

Your Stage 2 predictions (ligand affinities or rankings) are due by midnight on 1st February, 2016. The experimental ligand-protein IC50s will be released immediately after Stage 2 closes.

Computational methods allowed
You may use any method(s) you like to generate your predictions; *e.g.*, docking and scoring, MM-PB(GB)/SA, FEP, quantum-based methods, *etc*.

Anonymous *versus* public participation
When you sign up for the challenge, you are given the option of participating anonymously. Anonymous participation means that we may report on your predictions and methods, but your identity will not be disclosed. Public participation means we may also disclose who you are. Please note that, although we are committed to protecting the identity of anonymous participants, we cannot make any guarantees.

For each stage, you may use the D3R website to change your anonymous/public status until the stage has closed. However, after the stage has closed, you may not change your anonymous/public status.

D3R Workshop in March 2016
Participants are invited to share and discuss their results, as well as the D3R project more broadly, at the first D3R workshop, which is scheduled for March 9-12, 2016, at UC San Diego, La Jolla, CA. Note that the workshop immediately precedes the ACS National Meeting in San Diego, whose theme is *Computers in Chemistry*.

Selection of protein structures provided
The information packet for this Challenge includes protein crystal structures for use as potential starting points in your calculations. There is no guarantee that they are correct or optimal. You are free to modify them and/or use other structures from the PDB instead of these. However, when you upload your pose predictions, your structures must be superimposed with one of the structures we provided, to facilitate evaluation. The following subsections identify the structures provided and explain why they were chosen.

HSP90
Our intent is to provide at least one protein structure that should be suitable for docking compounds in each of the three chemical series in the challenge. The four protein structures provided, 2JJC, 2XDX, 4YKR and 4YKY, were solved at the following pH values, respectively: 6.8, 6.0, 8.0, 8.0. The rationale for the choice of these structures follows.

The first two structures, extracted from the PDB, are high quality structures of HSP90 with bound ligands in the same aminopyrimidine series present in the challenge series. They exemplify “closed” and “open” conformational states that can be adopted by HSP90 when compounds of this class are bound.

Although there are two PDB entries, 3OW6 and 3OWD, of HSP90 with bound ligands in the same benzimidazolone series present in the Challenge series, it was not clear that they would afford a feasible challenge, due to substantial conformational differences in the binding site. Therefore, one of the blinded challenge structures is provided -- and thereby disclosed (4YKR). Similarly, there are no PDB entries for HSP90 with ligands in the benzophenone-like chemical series, so one of the otherwise blinded structures is provided (4YKY). The two disclosed structures were selected for their relatively high quality, based on crystallographic resolution and the conformity of their conformational features with those observed for related features in the Cambridge Structural Database of small molecule crystal structures.

All four crystal structures include either one or two water molecules, chosen based on the pattern of conserved waters across many publicly available structures of HSP90 with bound ligands. However, there is no guarantee that these waters must be included, or that they are sufficient for success in these calculations.
The compound IDs and SMILES strings for the ligands to be docked in the pose-prediction stage of the HSP90 challenge are provided in Appendix A, at the end of these instructions. Potencies are to be computed or ranked for all 180 ligands. Please note a discontinuity in the compound numbering convention, hsp90_62, hsp90_67 and hsp90_109 aren't included in the list as they have been excluded from the Challenge.

MAP4K4
The goal, again, is to provide structures that provide a fair starting point for predicting the poses of the Challenge ligands. The ligands in this dataset are chemically diverse and do not fall into clear-cut chemical series. In addition, the crystal structures of this protein, both those in the PDB and those in the Challenge dataset, show considerable conformational diversity in the binding site. For example, the P-loop (residues 32 to 38) can adopt either a "closed" or an "open" conformation, as exemplified by PDB entries 4OBO and 4U44, respectively. In addition, as one may infer from the fact that residues 174-184 are not resolved in 4OBO, there can be considerable flexibility in parts of the binding pocket other than the P-loop.

In-house docking of the Challenge ligands into 4OBO and 4U44 showed no clear disadvantage to using the public structures instead of exemplary closed and open structures from the challenge set, aside from the trivial self-docking cases. Therefore, we provide PDB entries 4OBO and 4U44 as starting points for this Challenge. These structures are relatively well-resolved and were both solved at pH 8.3; the challenge structures were solved at pH 6.5.

No alteration to the number of water molecules included in the crystal structures has been made, and it is up to you to decide which, if any, water molecules should be included in your calculations. MAP4K4 crystallizes with two kinase monomers per unit cell, and only one of the two monomers contains the bound ligand. For the sake of clarity, we have deleted the apo form of the monomer from each PDB file.

Poses are to be predicted for all 30 compounds, and potencies are to be predicted or ranked for the 18 compounds listed in Appendix B, at the end of these instructions. Please note a discontinuity in the compound numbering convention, MAP10 and MAP24 compounds are not included in the lists and have been omitted from the Challenge.

Binding assay methods
The HSP90 binding assays[1] were carried out in Tris buffer (pH 7.5) with 150mM NaCl at room temperature, using a FRET-based assay. The buffer contained geldanamycin-biotin, streptavidin-APC, Eu-labeled anti-6XHis antibody, and HSP90. The compounds were diluted with the buffer and, after excitation of Eu in the complex, binding was measured as the ratio of the intensity at the emission of APC to that of Eu, where a decrease in fluorescence energy transfer would occur when the added compound displaces geldanamycin-biotin.

The activity of purified human MAP4K4 kinase domain on a peptide substrate was monitored using an ATP consumption assay[2]. For activation, the protein was pre-incubated at room temperature in 50mM HEPES buffer with 10 mM $MgCl_2$ at pH 7.2, in the presence of ATP. This

step was followed by incubation with the compound and peptide substrate in the same buffer at room temperature. Unreacted ATP was then quantified to determine activity.

Preparation and format of protein and ligand structures provided

Schrodinger Inc's Prep Wizard was used to add hydrogens to the protein structures and run energy-minimization with the OPLS2005 force field, in the presence of the bound crystallographic ligand, with only hydrogens free to move. Histidine side-chains were protonated according to the crystallization pH, and asparagine and glutamine side-chain amides were checked for plausible rotamers. One asparagine rotamer was adjusted: Asn51 in 4YKY.pdb was rotated to adopt the hydrogen bonding pattern consistently observed in HSP90-ATP analog and other Abbvie-solved structures, using the electron density map to guide the positioning of the heavy atoms. No subsequent energy-minimization was employed. The protein structures are provided in PDB file format.

No attempt was made to set optimal ligand protonation or tautomer states for the ligands, or to generate alternative tautomer states. It is up to you to choose and set these states for your calculations. The compounds are provided in the form of SMILES strings and SDfiles.

We again emphasize that the protein and ligand structures are not guaranteed to be correct or optimal for the Challenge calculations. Note, too, that ligand and/or protein protonation states might change upon binding. Participants are encouraged to critically examine the structures provided, using the context provided by other crystal structures available in the PDB.

Predicted Aggregators

All compounds have been subjected to Open Eye FILTER Version 2.5.2.4 (http://www.eyesopen.com/filter) to identify known and predicted aggregators. None of the Challenge compounds are known aggregators. However, several compounds are "predicted" as aggregators using the QSAR model within FILTER. Those compounds are HSP90_140, HSP90_79, MAP05, MAP11, and MAP19. We will provide measured affinity data for these compounds but whether they are true aggregators or not has not been tested. Open Eye notes that this QSAR model is very aggressive in its predictions.

Submitting your predictions

Predicted structures must be submitted in the form of PDB files in the same coordinate frame as those provided in the downloads. You may submit up to five predicted poses for each ligand. If you submit more than one pose for a given ligand, then a docking score or energy should be provided for each pose. For compound potencies, we anticipate accepting one or more of the following for each challenge set:

- predicted potencies, in units of nM, for each compound
- relative potencies for each compound, where the ligand with the lowest ID number is arbitrarily set to a potency of 1
- an ordinal ranking of ligands, where 1 indicates maximum potency (e.g., lowest IC50).

A template file will be provided for submitting affinity or ranking predictions, and detailed instructions for uploading your predictions for each stage will be provided in the coming weeks.

Evaluation of predictions
Pose predictions will be evaluated based on, at minimum: symmetry-corrected RMSD to crystallographic conformations, ligand-protein contacts, and overlap of predicted and experimental electron densities. We also anticipate evaluating predicted conformational changes of the protein binding site.

Affinity predictions/rankings will be evaluated based on, at minimum, accuracy of ranking.

Pending items, error reports, questions
We will email you during the Challenge regarding the following pending items:

- Templates and instructions for uploading your predictions
- Workshop details

We will also email you if necessary to share additional information or changes to the Challenge.

Please feel free to contact us if you notice any errors in the information provided or have questions about D3R Grand Challenge 2015: drugdesigndata@gmail.com

References

(1) Huth, J. R.; Park, C.; Petros, A. M.; Kunzer, A. R.; Wendt, M. D.; Wang, X.; Lynch, C. L.; Mack, J. C.; Swift, K. M.; Judge, R. A.; Chen, J.; Richardson, P. L.; Jin, S.; Tahir, S. K.; Matayoshi, E. D.; Dorwin, S. A.; Ladror, U. S.; Severin, J. M.; Walter, K. A.; Bartley, D. M.; Fesik, S. W.; Elmore, S. W.; Hajduk, P. J. "Discovery and Design of Novel HSP90 Inhibitors Using Multiple Fragment-bsed Design Strategies". *Chemical Biology & Drug Design* **2007**, *70*:1.

(2) Brunko, M. Tahir, S.K., Song, X., Chen, J., Ding, H., Huth, J. R., Judge, R. A., Madar, D. J., Park, C. H., Park C-M.; Petros, A.M., Tse, C.; Rosenberg, S. H.; Elmore, S. W. "N-Aryl-benzimidazolones as novel small molecule HSP90 inhibitors". *Bioorganic & Medicinal Chemistry Letters* **2010***, 20:7503.*

(3) Crawford, T. D.; Ndubaku, C. O.; Chen, H.; Boggs, J. W.; Bravo, B. J.; Delatorre, K.; Giannetti, A. M.; Gould, S. E.; Harris, S. F.; Magnuson, S. R.; McNamara, E.; Murray, L. J.; Nonomiya, J.; Sambrone, A.; Schmidt, S.; Smyczek, T.; Stanley, M.; Vitorino, P.; Wang, L.; West, K.; Wu, P.; Ye, W. "Discovery of selective 4-Amino-pyrimidine inhibitors of MAP4K4 using fragment-based lead identification and optimization." *Journal of Medicinal Chemistry* **2014**, *57*:3484.

(4) Gobbi, A.; Giannetti, A.M.; Chen, H.; Lee, M.-L. "Atom-atom-path similarity and sphere exclusion clustering: tools for prioritizing fragment hits." *Journal of Cheminformatics* **2015**, 7:11.

(5) Ndubaku, C.O.; Crawford, T.D.; Chen, H.; Boggs, J.W.; Drobnick, J.; Harris, S.F.; Jesudason, R.; McNamara, E.; Nonomiya, J.; Sambrone, A.; Schmidt, S.; Smyczek, T.; Vitorino, P.; Wang, L.; Wu, P.; Yeung, S.; Chen, J.; Chen, K.; Ding, C.Z.; Wang, T.; Xu, Z.; Gould, S.E.; Murray, L.J.; Ye, W. "Structure-based design of GNE-495, a potent and selective MAP4K4 inhibitor with efficacy in retinal angiogenesis." *ACS Medicinal Chemistry Letters* **2015**, 6: 913.

Appendix A. HSP90 compounds for pose-prediction

hsp90_40	CCc1cc(N2C(=O)Nc3ccccc23)c(O)cc1O
hsp90_175	Oc1cccc(c1)C(=O)c2ccc(O)cc2O
hsp90_164	Oc1ccc(cc1)C(=O)c2cc(Cl)ccc2O
hsp90_73	Cc1cc(nc(N)n1)c2ccc(cc2Cl)[N+](=O)[O-]
hsp90_179	Nc1nc(N)nc(n1)c2ccccc2Cl
hsp90_44	Oc1cc(O)c(cc1Cl)N2C(=O)Nc3cc(CNS(=O)(=O)c4cccnc4)ccc23

Appendix B. MAP4K4 compounds for potency predictions.

MAP01	COc1cc(ccc1O)-c1cnc(c2nccn12)Nc1ccc(cc1)N1CCN(C)CC1
MAP02	CN1CCN(CC1)c1ccc(cc1)-c1cnc2[nH]c3c(nc(nc3c2c1)-c1cnn(c1)C)C
MAP03	Nc1ncnc2sc3c(c12)CCCC3
MAP04	Cc1cc(c2c(c(sc2n1)C#N)N)C
MAP05	NC(=O)c1ccccc1NC(=O)c1ccc(o1)-c1cccc(c1)Cl
MAP06	CC(=O)NCC1Cc2cc(cc(c2O1)-c1ccncc1)Cl
MAP07	Clc1ccccc1-c1cc2cc(ncc2[nH]1)NC(=O)C1CC1
MAP08	Cn1c(cc2cnc(cc12)NC(=O)C1CC1)-c1ccccc1Cl
MAP09	FC(F)(F)Cn1c(cc2cc(ncc12)NC(=O)C1CC1)-c1ccccc1Cl
MAP11	N#Cc1cccc(c1)-c1cc2c(ncnn2c1)Nc1ccncc1
MAP12	CNC(=O)c1ccc-2c(c1)OCCc1cc(sc12)-c1ccnn1-c1ccc(cc1F)F
MAP13	CNC(=O)c1ccc-2c(c1)OCCn1nc(cc12)-c1ncnn1-c1ccc(cc1F)F
MAP14	CN(C)CCC(=O)Nc1cc(cc(c1)-c1ccc2ncnc(c2c1)N)F
MAP15	Cc1c[nH]nc1C1CCCN(C1)c1ccc2ncnc(c2n1)N
MAP16	Nc1ncnc2ccc(nc12)N1CCC(C1)Cc1ccccn1
MAP17	CNC(=O)c1ccc-2c(c1)OCCn1nc(c(c12)C)-c1ncnn1-c1ccc(cc1F)F
MAP18	Nc1ncc(c2ccc(cc12)-c1cccc(c1)F)C(=O)N[C@H]1CCC[C@@H](O)C1
MAP19	Cc1c[nH]nc1-c1cccc(c1)-c1ccc2ncnc(c2n1)N