A short message from Michela Taufer

Dear Friends,

The last time we shared with you the D@H newsletter was a long time ago, in 2009. In the past two years we have been so thankful for your support to D@H. We went through some challenging times with e.g., hurricane Irene and the recent BOINC client update, but also a lot of good crunching. This newsletter was put together by Boyu Zhang (one of our D@H developers) and will give you some D@H insights, tell you what keep us (and your computers) busy, and reveal to you our plans for the next year. We are putting together exciting work for you, including our new D@H game extension called ExSciTecH that will enable you to submit docking jobs to the D@H server, hopefully starting in spring 2012. We are also interested in hearing from you and getting your feedback with an anonymous survey available at the D@H webpage. The key achievement of the project is the completion of the first phase of our cross-docking simulations for the HIV protein (we will be done for the end of this year). We should be able to move to the next phase of D@H with the cross-docking of a new protein (the Trypsin) very soon. We tell you more about cross-docking and what challenges it raised for D@H in this newsletter.

As usual we want to thank you for your patience and support.

Enjoy this newsletter!

Michela

Outline

I. Docking@Home in numbers

Since Docking@Home (D@H) started in September 2006, our project has been stably growing and crunching jobs. Table 1 below summaries the current status of D@H in numbers. As shown in the table, over the five years of D@H being active, we have formed a big community of volunteers that are devoted and contributing to the important drug design scientific problem.

Table 1: Docking@Home in numbers

Num. of Jobs Distributed/Day	22,960
Num. of Hours Donated by Volunteers/Day	78,508
Num. of Flops Donated by Volunteers/Day (in Billions)	1,552
Num. of Registered Volunteers	45,901
Num. of Active Volunteers	7,105
Num. of Registered Hosts	101,282
Num. of Active Hosts	11,232

II. Self-docking and cross-docking simulations

When we started Docking@Home in 2006, we initially performed protein-ligand docking simulations by docking the ligands into corresponding protein conformations (See Docking Newsletter Issue 3 for details). This allows us to understand the docking process and the protein-ligand interatomic interactions (self-docking). It also allows us to computationally search the large space of potential ligand conformations, reducing the time and cost required to design new drugs by several orders of magnitude.

In the docking process, given a protein, several different types of ligands can dock into one of the protein binding pockets. Each ligand can have different types of atoms and levels of flexibility. The same protein, once docked with the ligands, can assume different conformations. A protein conformation and the docked ligand form a complex.

The docking algorithm normally deals with a flexible ligand; however, because of computing constraints, the protein is simplified into a rigid 3D lattice (or grid map). Each grid map consists of a three dimensional lattice of regularly spaced points surrounding and centered on the active site of a protein (the docking pocket). Each point within the grid map stores the potential energy of a ligand atom due to its interaction with the protein. For example, in a carbon grid map, the value associated with a grid point represents the potential energy of a carbon atom of the ligand at that location due to its interactions with all atoms of the protein receptor. This simplification is mainly due to the computational cost associated to a flexible representation of the protein conformation as well as the computational and storage constraints of the computer systems on which the simulations have to be performed.

Given a ligand, the protein flexibility can still be emulated by considering not only the grid map of the protein conformation originally observed experimentally docking with a given ligand, but also other conformations (3D lattices) of the same protein that have been experimentally observed and documented docking with other ligands. We are using the

LPDB dataset as our source of protein-ligand complexes. This approach of implicitly simulating protein flexibility is typically called cross-docking. The cross-docking approach is used to assess the sensitivity of docking results to minor changes in protein conformation due to a flexible binding pocket. In Docking@Home cross-docking simulations, given a ligand, we dock the ligand into multiple conformations of the same protein, each of which has been simplified with a grid map.

Figure 1 shows a small subset of the HIV cross-docking simulations performed by D@H in which the ligand in complex 1g35 is docked into three different protein conformations from other three complexes (i.e., 1gno, 1hih, and 1htf) to form new D@H complexes.

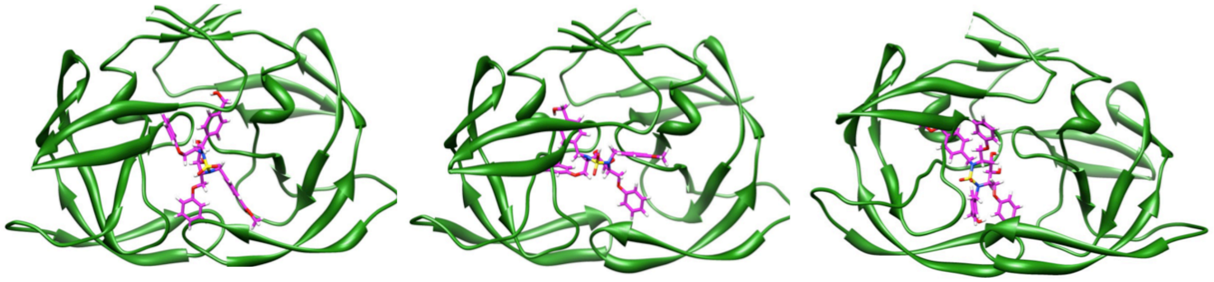


Figure 1: Three complexes with ligand from 1g35 docked into three protein conformations from 1gno, 1hih, and 1htf

Figure 2 shows a small subset of the Trypsin cross-docking simulations that we will perform with D@H in the next phase of our project in which the ligand from the complex 1k1m is docked into three different protein conformations from the complexes 1k1n, 1ppc, and 3ptb to form new complexes.

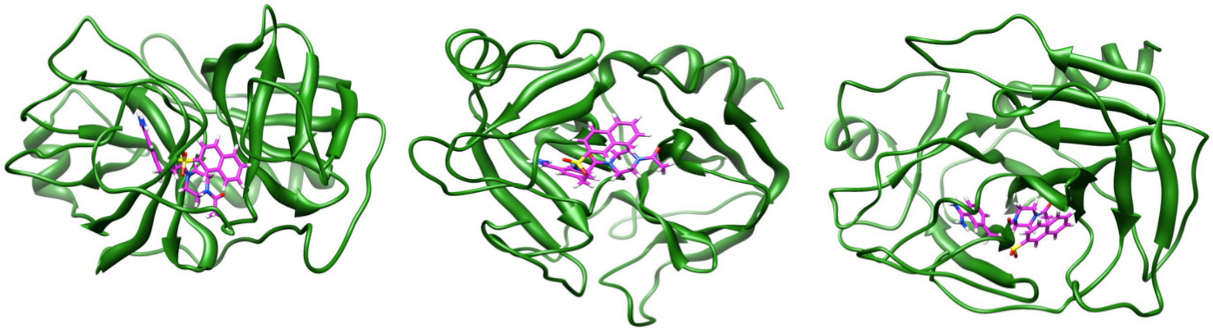


Figure 2: Three complexes with ligand from 1k1m docked into three protein conformations from 1k1n, 1ppc, and 3ptb

Figure 3 shows a small subset of the p38alpha cross-docking simulations that we will perform with D@H in which the ligand from the complex 1bl6 is docked into three different protein conformations from the complexes 1a9u, 1bl7, and 1ouk to form new complexes.

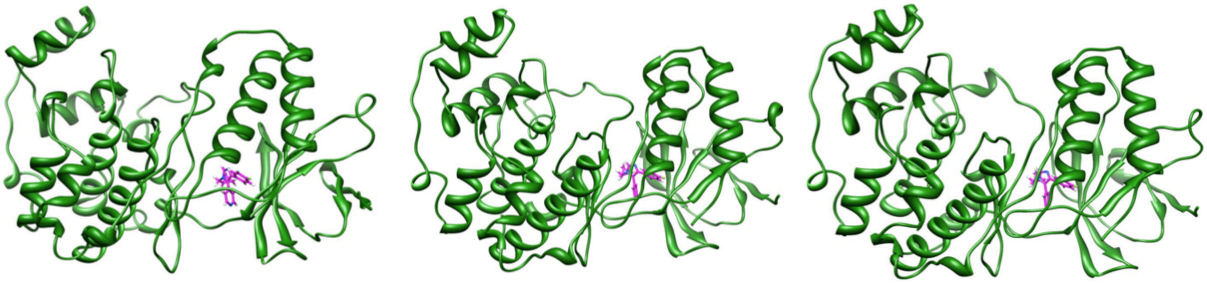


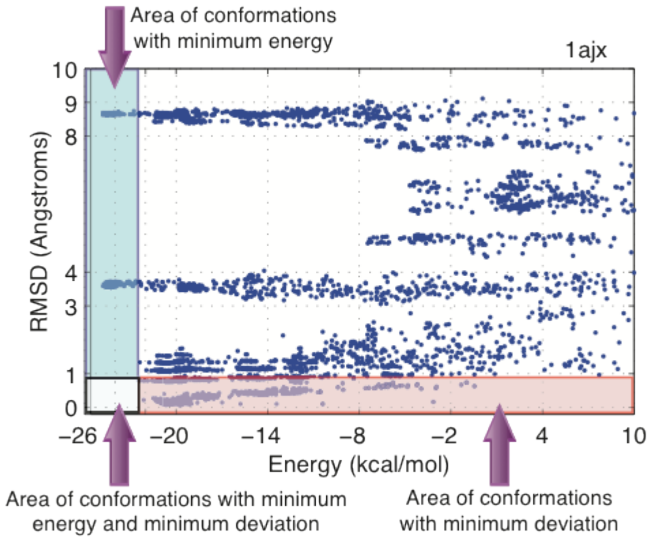
Figure 3: Three complexes with ligand from 1bl6 docked into three protein conformations from 1a9u, 1bl7, and 1ouk

III. D@H results analysis

The docking process is only one of the key steps in drug design. Once the results (ligand conformations) are collected, they need to be analyzed and sort based on their quality. High quality results are those with deviation from the near-native conformation (measured as the root mean square deviation (RMSD) between the given ligand from the crystal structure) smaller than or equal to two Angstroms (Å); however, conformations with RMSD between two and three Å are still considered results of interest. This process of sorting the ligands based on their quality is called scoring. Scoring can be based on energies or geometries of the ligand or of the whole complex. Note that the RMSD is measured in Angstroms (Å) and is calculated by the root square of the average squared difference of all non-hydrogen ligand atoms in the simulated ligand conformation and the ligand atoms in the crystal structure.

Scoring based on minimum energy

While dealing with the scoring, we initially relied on the traditional scoring approach based on energy values: we selected those ligands with lower energy as the more likely near-native conformations. We immediately identified the deficiencies of this approach in terms of accuracy. Figure 4 shows an example for 100,000 ligand conformations (every point in the figure is a ligand conformation) obtained with D@H for one ligand, the 1ajx, of the HIV1 protease. Here, the ligand conformations are scored in terms of their potential energy (x-axis) and their RMSD with respect to the known crystal structure (y-axis).



Area of conformations with minimum energy

Area of conformations with minimum energy and minimum deviation

Area of conformations with minimum deviation

Figure 4: Selecting ligands using energy-based approach

The figure shows three regions of relevance:

(1) The area of conformations with minimum energy, which is the vertical rectangle that goes from -26 to -22 kcal/mol. A ligand conformation with minimum energy does not always have a near-native conformation. Conformations in this area would be selected by a method that only accounts for the energy and chances are that those candidates are not near-native conformations.

(2) The area of conformations with minimum deviation (RMSD). The RMSD is calculated with respect to the crystal structure as explained above. This area is denoted by the horizontal rectangle that goes from 0 to 1Å. Ideally, the global minimum of a scoring function with high accuracy would be in this area. However, the global minimum is not always found. For the discovery of new drugs, the deviation dimension (y-axis) is unknown and cannot be used to select candidate ligand conformations.

(3) The area of conformations with minimum energy and minimum deviation, which is the intersection of the other two areas described before. Ideally this area should be densely populated to increase the opportunity of selecting good candidate ligand conformations. As Figure 4 shows, this may not happen, increasing the level of uncertainty and making harder the selection of near-native ligand candidates.

We observed the same problem across docking results generated with the two different docking algorithms (Algorithm 1 using a implicit representation of water and a distance-dependent dielectric coefficient and Algorithm 2 using a more physically accurate implicit representation of water using Generalized Born model) for the three proteins (HIV, trypsin, p38alpha) and the several ligands considered in D@H. This suggested to us that the scoring problem is independent from the docking method we used. Figure 5 shows an extreme example of this phenomenon for the ligand in complex 1w83 where the scoring function assigns the lowest energy to a set of converging conformations that dock with a significantly different orientation than the crystal structure.

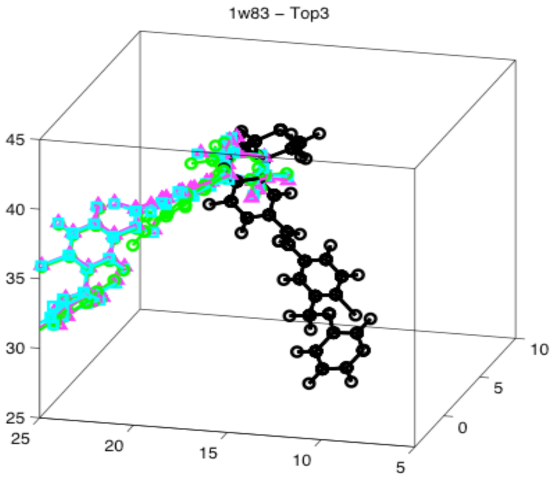


Figure 5: Comparison of ligand structures selected by energy only for 1w83 - crystal structure (black ligand) vs. top three scoring minimum energy (green, blue, and purple ligand)

Figure 5 shows the graphical comparison of the 1w83 (the p38alpha kinase in complex with a small molecule inhibitor) ligand only in the crystal structure (black ligand) versus the top three ligand conformations scoring the minimum energy over the whole set of D@H samples for this complex (green, blue, and purple ligands). This figure shows that the minimum-energy scored structures do not converge to a single solution despite the large number of D@H samples. At the same time, these three results are substantially different among each other and none of them is accurate enough to be called a near-native conformation.

Scoring based on clustering algorithms

Results from energy-based scoring methods raise an important question. Given the inaccuracy of the docking algorithms and millions of collected conformations, how can the scientists select those sampled ligands that are more likely to occur in nature, considering that the energy is not always a reliable scoring metric.

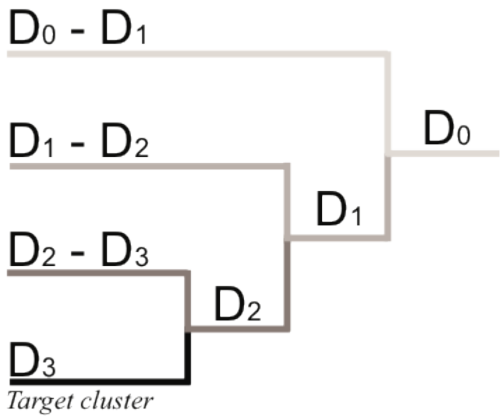


Figure 6: Hierarchical clustering represented as a dendrogram

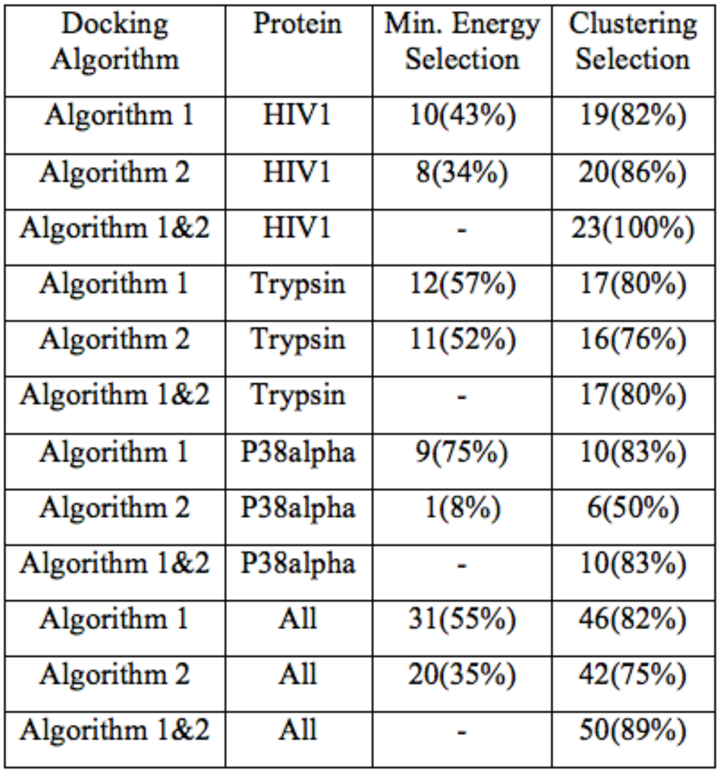
In many data post-processing phase, clustering algorithms are used to narrow down data of interest. Here we propose to use a probabilistic hierarchical framework that combines (1) the capability of dealing with data uncertainty by using a fuzzy c-means partitional clustering with (2) the capability of identifying the number of needed clusters at runtime by using a divisive hierarchical algorithm for which the cluster hierarchy-depth is probabilistically determined based on result variability. Rather than using the energies, we use the geometrical conformations of the ligands as input to our clustering and the RMSD among the D@H resulting ligands as our distance metric. Note that here we refer to the RMSD as a metric to compare resulting ligands among them and we do not refer to the crystal structure that is unknown for us during the scoring process. We also assume that D@H provides us with the sufficient number of samples, and thus the docking simulations converge toward near-native solutions. Our probabilistic hierarchical framework is more sophisticated that simpler clustering method such as the k-mean method. We selected this approach rather than other simpler methods because it is able to perform an effective unsupervised clustering of the large D@H datasets even in the presence of uncertainty and when the number of clusters is unknown a priori.

Figure 6 shows the process of our hierarchical clustering framework. The hierarchical clustering starts with the dataset of all our sampled ligands for a complex D_0 and uses the fuzzy c-means (FCM) to divide the set into two subsets one of which is defined as the compliment of the other ($D_2 = D_1 \cup D_0 - D_1$). Each ligand belongs to each subset with different probability degrees depending on its distance from a randomly chosen center (also called centroid ligand) of that cluster. Since we have two subsets, we select two centroids. Ligands that are not strongly biased to one subset or the other are removed from the two main partitions. Our probabilistic hierarchical framework selects the partition between the two subsets with a probability directly proportional to its size and inversely proportional to the internal variance of the ligands. The division process is iterative and continues until the means of the two partitions (D_{m+1} and $D_m - D_{m+1}$) are equal to each other with a statistical significance of 0.05. At every step, a hierarchy of centroids is kept and it is used to summarize the data space.

In Figure 6, centroids for D_0, $D_0 - D_1$, D_1, $D_1 - D_2$, D_2, $D_2 - D_3$, and D_3 are saved and can be used to analyze and summarize the different dimensions of the dataset. Also, the last cluster is by definition the most compact one (i.e., the larger cluster with smaller internal variance among ligands). Thus, this last cluster (D_3) represents the most reliable consensus obtained from the data. Consequently, the centroid of D_3 can be used as the most likely of the whole data and selected as the near-native conformation.

In order to test if our probabilistic hierarchical clustering is robust and can capture near-native conformations independently from the docking method used, we considered again the two different docking algorithms (Algorithms 1 and 2).

Table 1: Comparison of number of successfully selected ligands per docking algorithms and type of proteins



Docking Algorithm	Protein	Min. Energy Selection	Clustering Selection
Algorithm 1	HIV1	10(43%)	19(82%)
Algorithm 2	HIV1	8(34%)	20(86%)
Algorithm 1&2	HIV1	-	23(100%)
Algorithm 1	Trypsin	12(57%)	17(80%)
Algorithm 2	Trypsin	11(52%)	16(76%)
Algorithm 1&2	Trypsin	-	17(80%)
Algorithm 1	P38alpha	9(75%)	10(83%)
Algorithm 2	P38alpha	1(8%)	6(50%)
Algorithm 1&2	P38alpha	-	10(83%)
Algorithm 1	All	31(55%)	46(82%)
Algorithm 2	All	20(35%)	42(75%)
Algorithm 1&2	All	-	50(89%)

Table 2 summaries the accuracy of the hierarchical clustering for the two docking algorithms. We count the selection of a ligand as success for our clustering selection if the ligand conformation selected has the RMSD less than 2Å with respect to the crystal structure and for the minimum energy selection if the median RMSD of the 100 ligand conformations with lowest energy is less than 2Å. As shown in the table, overall our framework outperforms the naive approach for all the complexes and for each method. With our clustering method we can see that none of the two docking algorithms clearly outperform the other. The results obtained by combining the D@H samples of the two docking algorithms can further strengthen the accuracy of our predictions for the HIV protease for which we observed a hit rate of 100%.

New challenges brought by cross-docking

In cross-docking simulations, when trying to emulate the protein flexibility, the number of cross-docking attempts and related results are much larger than in the simpler self-docking simulations. If we consider M number of ligand conformations, M number of protein conformations, and N number of docking attempts (jobs distributed to the volunteers) for each ligand, we get $M^2 * N$ number of jobs. Note that both N and M are normally very large numbers. Thus it results in very large number of jobs distributed and results datasets collected to score.

This opens two new challenges for the D@H project: first we need to sample more and rely even more on your support; second we need to score much larger datasets.

Sample more results and attract more volunteers

To continue keeping our volunteers engaged and recruit new volunteers we are working on an extension of D@H called ExSciTecH, an interactive, easy-to-use game system to Explore Science, Technology, and Health. Figure 7 shows an overview of ExSciTecH. It extends D@H with a gaming environment in which our volunteers will learn while playing the basics steps of the protein-ligand docking process, which include the ligands docking within proteins to nullify or emphasize the protein's effects in the body. You will also learn how to identify molecules by name or related disease (e.g., HIV and breast cancer), identify molecules (protein or ligand) and atoms by type, and identify molecules by shape, matching docking site, complexity, and degree of flexibility.

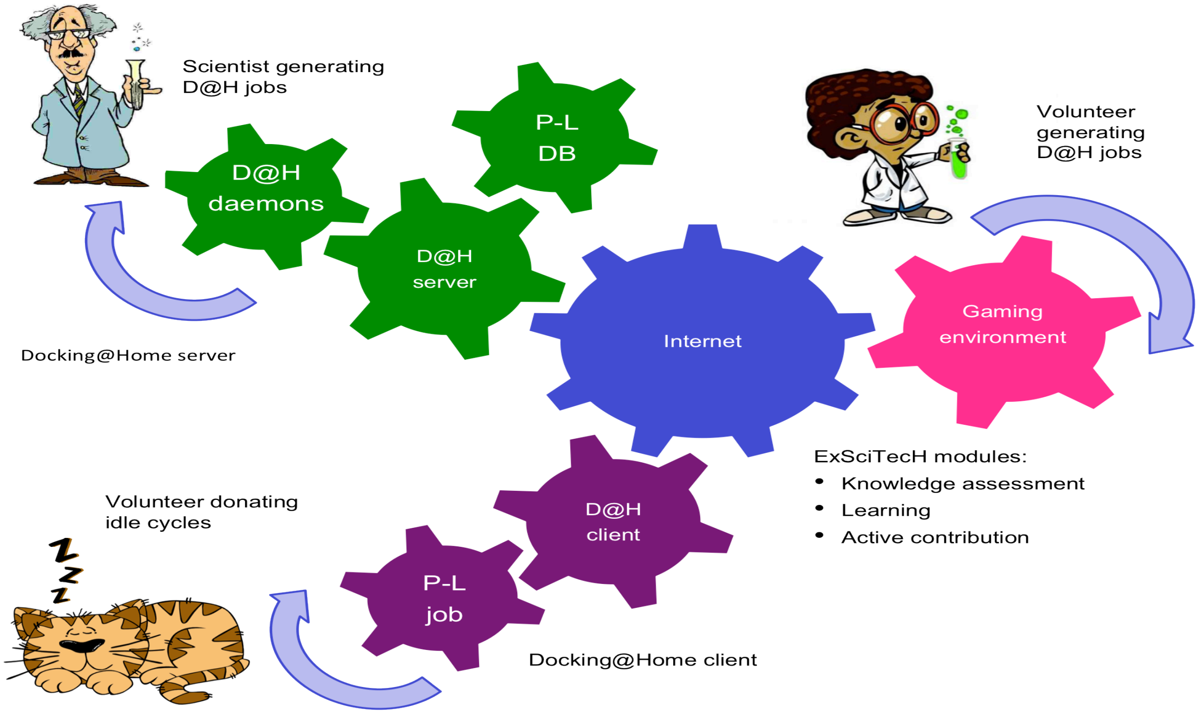


Figure 7: ExSciTecH overview

To host ExSciTecH in D@H and handle the volunteer interactive job generation in a gaming fashion, D@H infrastructure needs some changes. In the process of volunteers "play fun games" and "submit my job", here is what happens under the hood from the D@H server's perspective as shown in Figure 8: volunteers submit the interactive jobs by making queries to the D@H queue on the server; the D@H server validates the simulation before generating jobs and sending them to idle D@H clients; then idle D@H clients execute the jobs and upload the results to the server; D@H server collects the results, validates them and provides the original volunteer with rewards (scores). These interactions between volunteers and D@H server require changes to the server. In addition, volunteers can potentially send erroneous or even damaging jobs, so the D@H server has to be safeguarded from them, and the D@H server should keep track of the participation of the volunteers and the scores of their jobs.

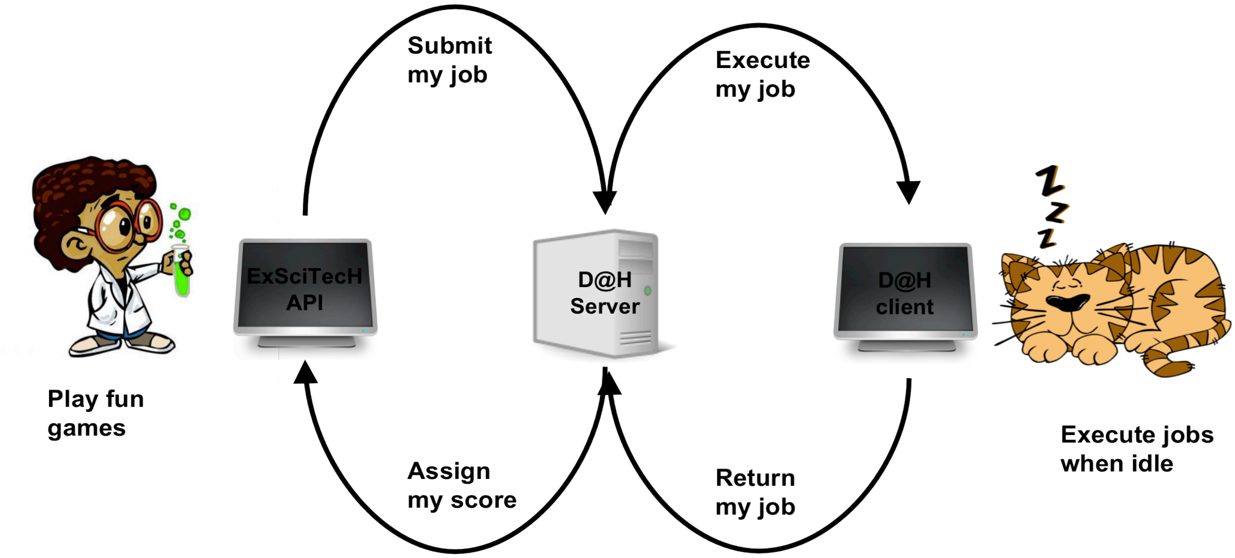


Figure 8: ExSciTecH framework

Figure 9 shows three snapshots of the games, and gives you a flavor of how the ExSciTecH games will look like. In the first snapshot, you identify molecules and the molecule type among atoms, ligands, and proteins; in the second snapshot, you select protein and ligand, which are the complex you want to dock and simulate; in the third snapshot, you submit the job which is a new simulation to the D@H server. In this way, volunteers can participate in the scientific discovery by creating new ligands by editing existing ligands from a database, by finding docking sites in which volunteers dock a ligand in a selected protein, and by finding side effects in which volunteers submit a docking simulation to determine whether a ligand will produce side effects by docking well with the wrong protein.

We expect our beta version of ExSciTecH to be ready for you in spring 2012.

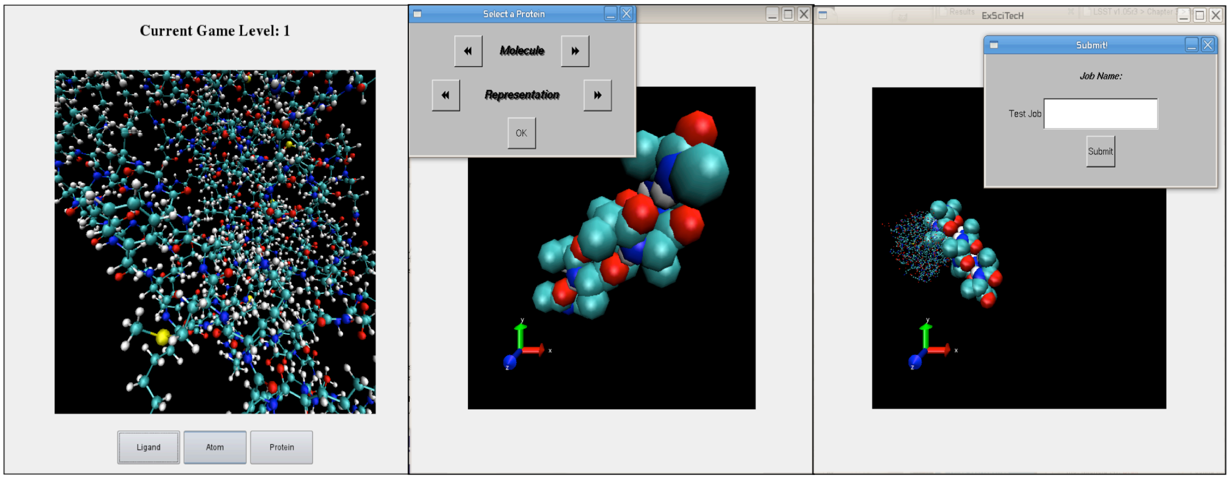


Figure 9: Snapshots of ExSciTecH games

Score over very large datasets

When analyzing cross-docking results, for each ligand, we need to compare among all the complexes formed by docking the given ligand into all the different protein conformations.

Take the ligand 1g35 from HIV protein for example, when analyzing the cross-docking results for it, we need to compare among all the conformations that you generated with you computers for the 25 complexes and this results in a very large dataset. When doing geometry comparison and scoring over such a large dataset, it is difficult to use the same hierarchical clustering method that is used in self-docking, since it poorly scales: the performance of the clustering method degrades significantly while the data frequently swaps between disk and memory. Moreover, this method is difficult to parallel since there are major data dependencies inside: data from previous iterations are used to form the cluster division in the current iteration. To effectively and efficiently analyze the large dataset generated by cross-docking simulations, we have been working on a new, more powerful clustering method that is easy to parallel and scales well. The best-fit parallel model for our method is MapReduce, and we presented the preliminary results in the poster on SuperComputing 2011 (SC11), please refer to the poster in publication section for details. We look forward to share more results with you in the next newsletter, hopefully not too long after this one.

IV. We want to hear from you

Engaging a larger and more diverse group of volunteers is always an important mission for us. To help us learn more about your experiences of D@H, what makes you stay with us and contribute to the scientific problem, what other features you prefer to see in the future, etc., we designed a survey to ask for your feedback. It will be of great help if you take two minutes to share with us what do you think. Together, we can make the volunteering experience even better. The survey is anonymous and can be found from the homepage of D@H: http://docking.cis.udel.edu/. We look forward to hearing from you! And to better understand the purpose of this survey, we interviewed Trilce Estrada, a long time developer of D@H, to ask her why we need you to tell us what is your opinion.

Q&A with Trilce, D@H survey

Q: Who designed the survey?

A: The survey was designed by the social science team of the ExSciTecH project, in particular by Kathleen Pusecker and Manuel Torres from UD and Dr. Joanne Cohoon from UVA.

Q: What is the purpose of the survey?

A: The purpose of the survey is to help us determine what aspects of Docking@Home are useful in keeping our current volunteers engaged in donating cycles and learning about our application, and which aspects prevent new volunteers from joining the project.

Q: How is the survey going to help D@H?

A: The answers will help us to improve the experience of our current volunteers and at the same time, by making our project user-friendlier we hope to attract new volunteers form a diverse background.

Q: How are you going to use the information collected via the survey?

A: We will study the global perception that our volunteers have about the project. Since we care about the privacy of our volunteers, the answers are completely anonymous and we just collect information about the demographics, but not the identity of each volunteer.

Q: What difference can we feel as volunteers?

A: Eventually we will implement improvements to the D@H website, including directions on how to become a D@H volunteer and also better ways of communicating the scientific aspects of our project.

Q: Are there any other comments or information that you would like to share with volunteers?

A: The future direction of Docking@Home will include a game engine where volunteers will be able to play with molecules and at the same time they will learn about chemical complexes and the docking process. But first, with this study we hope to set up the basis for a more user-friendly and inclusive project.

V. The timeline of D@H cross-docking project

The cross-docking simulations for HIV complexes are going to complete in December 2011, we draw a tentative timeline for the simulations and analysis work below.

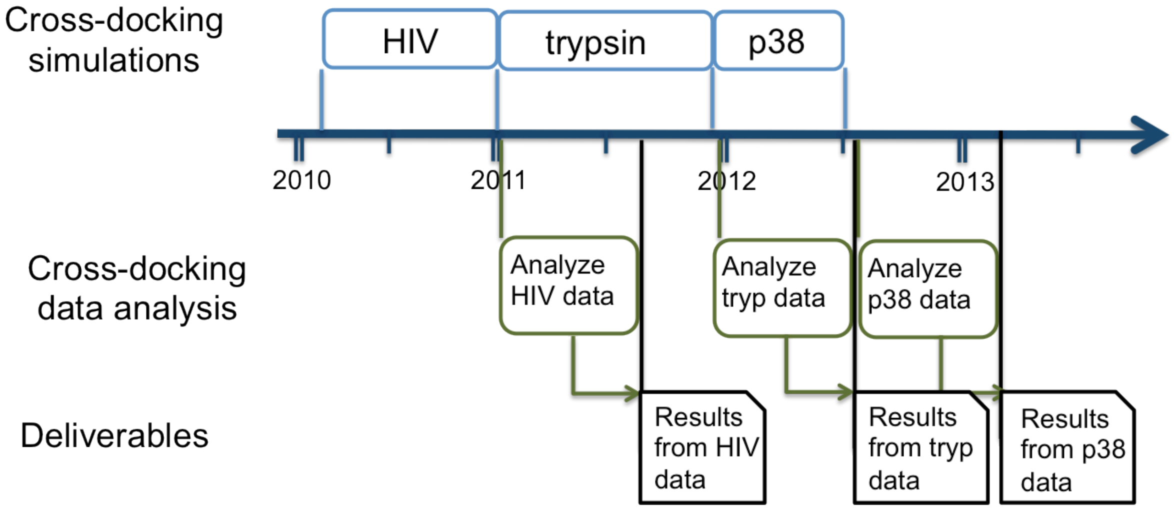


Figure 10: Tentative timeline for cross-docking

As the timeline in Figure 10 shows, we started the cross-docking simulations for HIV protein in early 2010, in which each of the 26 ligand conformations is docked into 25 protein conformations excluding the experimentally observed one and results in 650 complexes. For each complex, we collect at least 20,000 results back. Currently, 550 out of 650 complexes have finished and 100 are left to run. By the end of 2011, the complexes for

HIV should be all finished. And we move on to the next proteins: trypsin and p38alpha, trypsin has approximately the same number of complexes as HIV, and p38alpha has roughly half the size.

At the mean time, after the cross-docking simulations for HIV complete, we are going to perform analysis on the results collected, and share our findings with you.

We are very excited about the study of protein flexibility through cross-docking, stay tuned with Docking@Home!

VI. Publications

With great support from all of you, D@H is able to contribute our knowledge to the community by publishing our findings. Following is a list of recently publications, please feel free to discuss with us if anyone interests you.

T. Estrada, R. Armen, and M. Taufer: Automatic Selection of Near-Native Protein-Ligand Conformations using a Hierarchical Clustering and Volunteer Computing. In Proceedings of the International Conference On Bioinformatics and Computational Biology (ACM-BCB), August 2010, Niagara Falls, NY, USA.

T. Estrada and M. Taufer: Providing Application-Level QoS in Volunteer Computing . In the Proceedings of the 13th IEEE High Performance Computing and Communications (HPCC) Conference. September 2011, Banff, Canada.

T. Estrada, B. Zhang, R.S. Armen, and M. Taufer: Study of Protein-ligand Binding Geometries using a Scalable and Accurate Octree-based Algorithm in MapReduce. Poster in Proceedings of the ACM/IEEE International Conference for High Performance Computing and Communications conference (SC), November 2011, Seattle, Washington, USA.

VII. Thank you

Docking@Home has been actively crunching for several years, and smoothly evolving as the project grows. None of these amazing things would have happened without your help! Many thanks to all of our volunteers from the Docking@Home project team! Thanks for staying with us and we look forward hearing back from you via the survey, let's make the volunteer experience even better in the future!