

Regular Paper

A GPU accelerated Fragment-Based De Novo Ligand Design by a Bayesian Optimization Algorithm

MOHAMED WAHIB,^{†1} ASIM MUNAWAR,^{†1} MASAHARU MUNETOMO^{†2}
and KIYOSHI AKAMA^{†2}

De Novo ligand design is an automatic fragment-based design of molecules within a protein binding site of a known structure. A Bayesian Optimization Algorithm (BOA), a meta-heuristic algorithm, is introduced to join predocked fragments with a user-supplied list of fragments. A novel feature proposed is the simultaneous optimization of force field energy and a term enforcing 3D-overlap to known binding mode(s). The performance of algorithm is tested on Liver X receptors (LXRs) using a library of about 14,000 fragments and the binding mode of a known heterocyclic phenyl acetic acid to bias the design. We further introduce the use of GPU (Graphical Processing Unit) to overcome the excessive time required in evaluating each possible fragment combination. We show how the GPU utilization enables experimenting larger fragment sets and target receptors for more complex instances. The Results show how the nVidia's Tesla C2050 GPU was utilized to enable the generation of complex agonists effectively. In fact, eight of the 1809 molecules designed for LXRs are found in the ZINC database of commercially available compounds.

1. Introduction

The search for drug molecules with computational methods is often performed by high-throughput docking or to a lesser extent by De Novo drug design approaches. While virtual screening relies on pre-existing compounds, De Novo design approaches generate novel molecules out of building blocks consisting of single atoms or fragments. Due to the huge and non-linear search spaces (Typically, tens of thousands of orientations are generated for each ligand candidate.), global optimization algorithms are usually employed to search the chemical space by generating new molecular structures through probing many different fragments in a combinatorial fashion. Traditionally, related projects have embraced Evolutionary Algorithms (a class of global optimization algorithms inspired from the biological phenomenon of evolution) for this problem as will be shown below. In this research, the choice was to use Bayesian Optimization Algorithm (BOA)¹⁰, an EA that proved to give very good results in complex global optimization problems.

Here, as a main contribution, a novel approach for De Novo Design of agonists is presented. The algorithm utilizes a fragment-based method that generates molecules by joining predocked fragments with linkers. A parallel version of BOA is used to search for feasible solutions. Only the pre-

docked fragments are encoded by the BOA, while suitable linker fragments are efficiently evaluated with a tabu search³) using look-up tables. The fitness function used is a novel combination of force field energy and a measure of the 3D-overlap to known binding mode(s). The energy term consists of intra- and intermolecular contributions. The measure of 3D-overlap enforces a spatial distribution of the atoms of the designed molecule similar to the one in the known binding mode of the agonist(s) without explicitly considering the covalent structure. The algorithm is evaluated on liver X receptors (LXRs)²), presenting a complex optimization problem due to the large number of fragments used. Different fitness function setups are analyzed for their search efficiency. Notably, the algorithm is able to suggest molecules with new scaffolds or substituents that, at the same time, preserve the main binding interaction motifs of known agonists of LXRs.

For complex structures with high order of used fragments (such as LXRs in this case), the massive computation cost expected makes parallel computing a De Facto issue. Therefore, the system proposed utilizes nVidia GPU in order to harness the high computing power of state-of-art GPUs. A major hurdle with utilizing GPU is the complexity of the GPU architecture and the need to carefully optimize and tune any application running over GPU to achieve a highly efficient performance. A second main contribution in this paper is the design of the De Novo drug design algorithm to run efficiently over the SMIT architecture of GPU. The design includes performance optimization strategies we in-

^{†1} Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan

^{†2} Information Initiative Center, Hokkaido University, Sapporo, Japan

roduced earlier in⁹⁾.

The rest of paper is as follows. The following section overviews the algorithm and implementation over GPU while section 3 shows the results and discussion. Finally section 4 concludes.

2. Algorithm and implementation over GPU

2.1 BOA

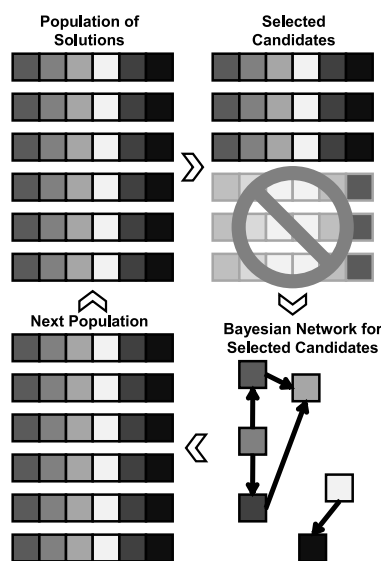


Fig. 1 Bayesian Optimization Algorithm (BOA).

BOA belongs to a class of algorithms known as Estimation of Distribution Algorithms (EDAs)⁸⁾. EDAs are an outgrowth of Genetic Algorithms (GAs). Conventional GAs maintain a population of probable solutions and then they apply genetic operators like selection, mutation, and crossover to find the next population. This process continues until the algorithm finds an acceptable solution. EDAs replace the variation done by the crossover step in conventional GAs with some kind of statistical inference from the existing population to be used to construct the next population. In the case of BOA, variation starts by constructing a Bayesian Network (BN)⁴⁾ as a model of promising solutions after selection. New candidate solutions are then generated by sampling the constructed Bayesian network. Finally, new solutions are incorporated into the population, eliminating some old candidate solutions, and the next iteration is executed unless a termina-

tion criterion is met. Figure 1 shows the important steps in a single iteration of BOA. BOA uses BN to encode the structure of a problem.

2.2 Fragment-based De Novo Ligand Design

De Novo ligand design uses a fragment-based method to generate molecules by joining pre-docked fragments with linkers. To connect pre-docked fragments with linker fragments we use a combination of two stochastic search procedures, BOA and a tabu search. Heavy atom-hydrogen atom vectors are the connection points, which can be selected by the user. Covalent bonds generated by the algorithm for linking fragments are single bonds. The scoring function is a linear combination of two terms with multiplicative parameters as input.

A version of BOA parallelized for GPU runs the main global optimization problem (i.e. searching the structure search space). Every individual contains a single chromosome consisting of multiple genes. As for encoding, and contrary to classic genetic algorithms, the implementation in BOA uses integers as gene values encoding indexes of docked fragments. Hence, the value of each gene ranges from one to the number of docked poses.

Another search algorithm, Tabu search, is used for efficient linking. Linking the encoded fragments for each individual is done by a tabu search. For efficiency reasons (i.e. to avoid conditional branching in the GPU kernel), we built a look-up table containing all distances and angles of all pairs of linker fragment connection vectors. Using cutoff values and the look-up table, all possible connections of fragment pairs of an individual are generated. A connection solution is randomly picked, and the two fragments are joined with the linker defined therein.

Scoring The scoring function implemented is a linear combination, i.e., a weighted-sum, of two terms: a force field-based binding energy E_{ff} and a measure of similarity (Sim_{3D}) to a user-supplied target structure (e.g., a known agonist).

$$S_{total} = w_{ff}E_{ff} - w_{3D}Sim_{3D}$$

where the multiplicative parameters w_{ff} and w_{3D} are input values. The minus signs for the similarity term is used because optimization is performed by minimization of S_{total} while Sim_{3D} grow with increasing similarity. The two scoring terms are evaluated as follows:

- Force field energy function: We utilize nVidia's Bio Workbench¹²⁾, accelerating energy calculation,

to calculate binding energy between the ligand and the receptor protein. The force field-based energy function consists of van der Waals and electrostatic terms. Both intraligand (intra) and ligand/receptor (inter) interactions are taken into account.

$$E_{ff} = E_{inter}^{vdW} + E_{inter}^{elec} + E_{intra}^{vdW} + E_{intra}^{elec}$$

Intrafragment and intralinker interactions as well as fragment-linker interactions between atoms separated by one or two covalent bonds are not evaluated. The potential of the receptor is calculated and stored to be used only for the linkers. The energies of the fragment poses are read in from the MOL2-files to save computational time.

- The 3D structure Similarity Sim_{3D} : between the newly assembled molecule (A) and a user-supplied template molecule (B) is evaluated by

$$Sim_{3D}(A, B) = \frac{S_{AB}}{\max(S_{AA}, S_{BB})}$$

$$S_{XY} = \sum_{i \in X} \sum_{j \in Y} w_{titj} e^{-\gamma r_{ij}^2}$$

where r_{ij} is the distance between two atoms ($i \in$ molecule X, $j \in$ molecule Y), w_{titj} is a matrix whose coefficients reflect the similarity between element types, and γ is a coefficient which acts on the broadness of the distribution of the positions. The 3D similarity Sim_{3D} does not explicitly consider the covalent structure of molecules but relies on the arrangement of atoms in space.

Protein preparation Liver X Receptors (LXRs)

- members of a super family of nuclear hormone receptors and represented by two subtypes, $LXR\alpha$ and $LXR\beta$ - have been shown to be involved in cholesterol homeostasis. Because of the high correlation in binding affinity for the two isoforms, and the high sequence identity in the ligand binding domains (77%), only one isoform was employed for our study. The crystal structure of $LXR\beta$ (PDB code: 1PQ6, [15] 2.4 Å resolution, $R_{free}=0.262$) was selected as a representative receptor structure for the docking of the compound library because of the higher resolution of the crystal structure for the human receptor (2.40 Å for the β isoform compared to 2.90 Å for the highest resolution for a human structure of $LXR\alpha$)¹¹. Subsequently, the term LXR shall refer to the $LXR\beta$ isoform.

Preparation of fragment library The library of fragments, from which the molecules were constructed, was obtained from Molinspiration Chem-

informatics (www.molinspiration.com, March 2011 accession date). The library consisted of 30,000 fragments with one and 30,000 fragments with two connection points occurring in bioactive molecules. CHARMM atom types were assigned, and all fragments were subject to minimization. The connection points defined in the source MOL2-files were used as connection vectors of the fragments, using all possible heavy atom-hydrogen atom vectors. The original and superimposed fragments were deemed identical if the similarity was larger than 0.95. Of the 60,000 fragments in the library only the 13,788 containing less than four rotatable bonds were used. Of these, 6,906 and 6,882 have one and two connection vectors, respectively. They were docked into the receptor binding site with SEED⁷), a program for docking mainly rigid fragments with evaluation of protein-fragment energy and electrostatic desolvation.

2.3 Implementation over GPU

GPU is emerging as one of the most powerful parallel processing devices. GPU is especially well-suited to address problems that can be expressed as data-parallel computations with high arithmetic intensity (i.e. ratio of arithmetic operations to memory operations). Applications that process large data sets can use a data-parallel programming model to speed up the computations. However, although GPUs can offer unprecedented performance gain, implementation of an algorithm over a GPU to take full advantage of this new technology involves a significant complexity of parallelizing across the multiple cores. Memory management over a GPU makes things even more challenging. CUDA¹⁾ is a parallel computing architecture developed by nVidia. CUDA is the compute engine in nVidia's CUDA compatible GPUs, and is accessible to software developers through industry standard programming languages like C. CUDA is widely used for programming nVidia's GPUs for general purpose processing.

The implementation of the algorithm over GPU is shown in figure 2. The figure shows the CPU and the GPU side portions of the algorithm. All the configurations, memory allocations, initializations are performed over the host processor. After the initialization stage (which includes loading the pre-docked fragments and encoding them to a code table), data is transferred to the device. Then the code running at the host side enters a loop. At that loop, breadth-first search detects all the fragments eligible for binding. Next, the codes of the fragments are transferred to the device. At this point the kernel starts running. The kernel uses the frag-

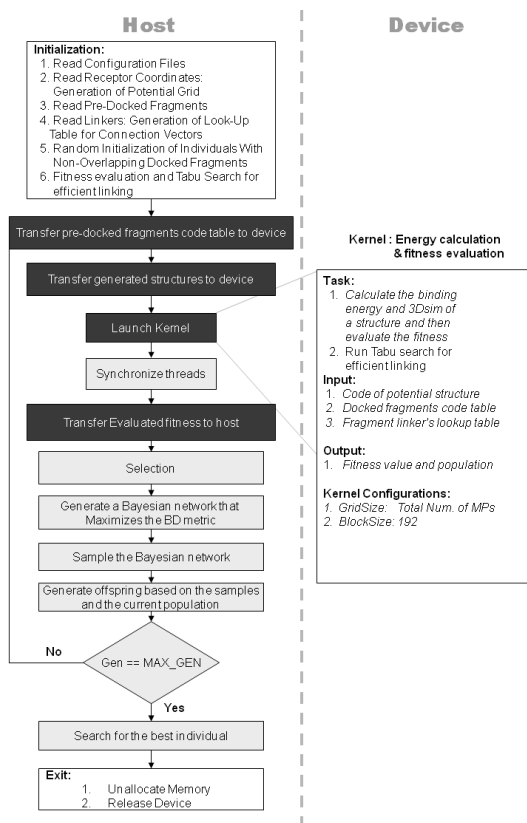


Fig. 2 Flow chart of CPU (host) side and GPU (device) side logic

ments code table and the codes of the fragments in the structure to be evaluated, then decodes the codes to create the structure in thread's registers. Then the kernel runs the force-field energy of the structure. Finally the kernel computes the fitness (S_{total}) after computing Sim_{3D} between the structure and the template then searching for an efficient linking. Next, the computation transfers back to the host after receiving the fitness values. The host then proceeds with the algorithm by selecting candidates from the population to build the Bayesian network. Truncation selection has been used as advised by⁶⁾ to better maintain the building block. Next, the network is constructed using the BD metric as mentioned earlier and the network is sampled to generate new candidate solutions. Lastly, the offspring individuals are inserted into the parent population if no structurally similar parent has a more favorable score to intensify the selection pressure. The cycle is then repeated until a termination criteria is met.

3. Results & discussion

Results given in this section were collected over a system with nVidia Tesla C2050 GPU mounted on a motherboard with Intel®Core™i7 920@ 2.67GHz as the host CPU. C2050 has 3GB of device memory and the total number of processing cores is 448. The maximum amount of shared memory per block is 64KB (16kB was used as cache as advised by CUDA manual) and clock rate is 1.5GHz. We are using Fedora Core 13 as the operating system and CUDA SDK/Toolkit ver. 4.0 with nVidia driver ver. 270.41.19. C2050 is dedicated to computations only. The system has a separate GeForce 8400 GS GPU acting as a display card. The high computational demands of the problem in this paper assures a highly inflated execution time if run in a serial fashion. Therefore the comparison with a serial implementation is skipped due to giving an expected highly inflated values for the serial implementation in favor of the GPU implementation. However, preliminary results over of a serial version of the program running over an Intel i7 920@ 2.67GHz CPU with 4GB memory having Fedora Core 12 as OS showed speedups up to 60x.

Setting of Algorithm Runs Calculations were repeated 10 times for each of three settings (i.e. weighted coefficient phasing as seen below) with distinct random seed numbers for 1000 iterations of the algorithm and 20 iterations of the tabu search per individual. The minimized phenyl acetic acid-based agonist cocrystallized with the protein (PDB code 1PQ6) was used as a target structure. The coefficients of the scoring function terms were set to $\{w_{ff} = 0.02, w_{3D} = 0.98\}$, $\{w_{ff} = 0.06, w_{3D} = 0.94\}$ and $\{w_{ff} = 0.10, w_{3D} = 0.90\}$.

Molecules designed by the algorithm Several of the 100 generated molecules with the most favorable S_{total} when compared with the crystal structure of LXR in the complex with the heterocyclic phenylacetic agonist shows that the generated molecules include key motifs of the target structure, e.g., the two ring systems joined by a linker (table 1, compound 2). Compounds generated by the algorithm in table 1 are a consequence of the enforced 3D-structural diversity within populations during optimization. Analysis of the existence of the generated molecules in the ZINC library⁵⁾ reveals that eight out of 1809 generated molecules are commercially available.

4. Conclusion

This paper presented a system for fragment-based De novo ligand design. A combination of

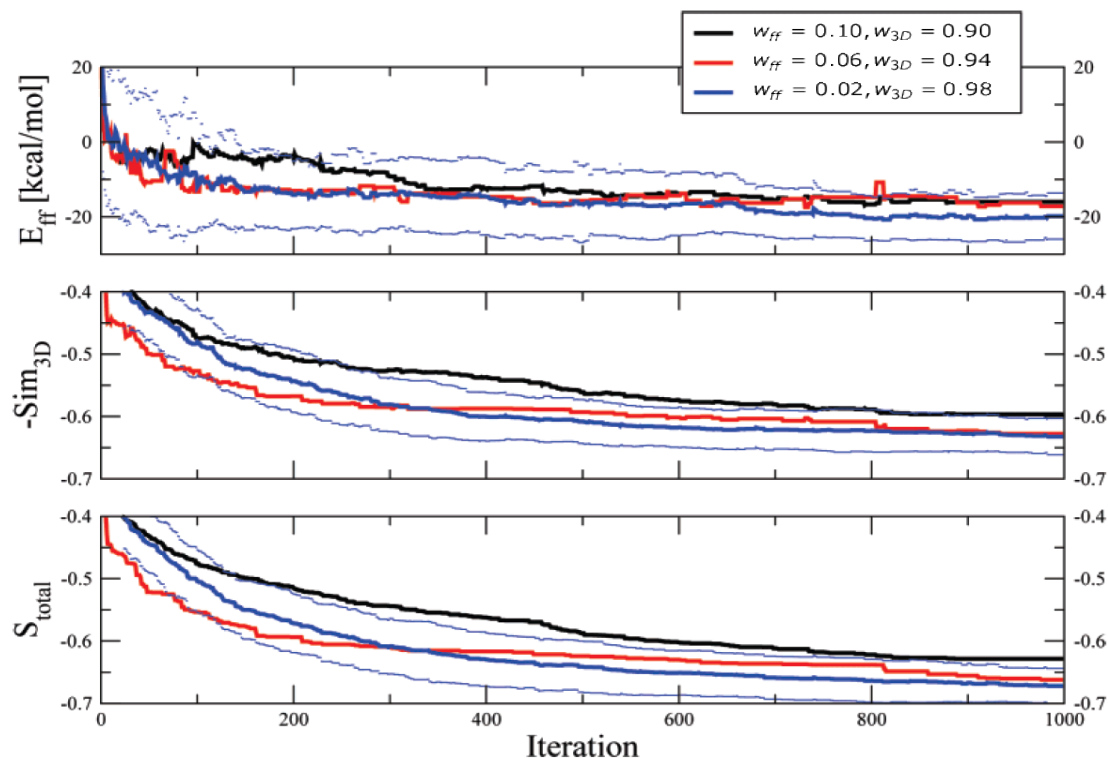


Fig. 3 Evolution of the scoring function terms and the total score of the best individual in each weight setting. The values at each iteration step were averaged over 10 runs. The bold lines are averages, while the thin blue lines are the standard deviation of the run with the least w_{ff}

an evolutionary algorithm (BOA) and tabu search is used for the simultaneous optimization of force field energy and 3D similarity to known agonist(s). Therefore, the design is both binding site-based and ligand-based. Importantly, the relative importance of these two driving forces can be modulated by the user. Due to the proven high computational cost needed to run simulations for complex structures, the entire system was implemented to run over GPU (nVidia's Tesla architecture) in an attempt to accelerate the performance on the GPU instead of high cost of conventional clusters.

In an application to the liver x receptors (LXR_s), 1,809 molecules were generated by the algorithm within the ATP-binding site in less than 16 h on a Tesla C2050 using a library of 14,000 fragments with up to three rotatable bonds. Notably, molecules similar to those generated by the algorithm are commercially available providing further evidence of the usefulness of the proposed system for De novo drug design. The algorithm can generate molecules similar to known LXR agonists. Importantly, by enforcing diversity throughout the op-

timization and by using a 3D-similarity-based scoring function term Sim3D, which does not rely on a covalent structure of the compared molecules, scaffold or linker hopping was observed, retaining the common binding motifs of known LXR agonists.

References

- 1) : nVidia CUDA Programming Guide - CUDA 4.0 SDK Documentation (2011).
- 2) Bonn, M., Sun, T., Ljunggren, S., Ahola, J., Wilhelmsson, H., Gustafsson, A., Jan-Ake and Mats, C.: The three-dimensional structure of the liver X receptor beta reveals a flexible ligand-binding pocket that can accommodate fundamentally different ligands., *Journal of Biological Chemistry*, Vol. 278, No. 40, pp.38821-8 (2003).
- 3) Glover, F. and Laguna, M.: *Tabu Search*, Kluwer Academic Publishers, Norwell, MA, USA (1997).
- 4) Heckerman, D.: A Tutorial on Learning With Bayesian Networks, Technical report, Learning in Graphical Models (1996).
- 5) Irwin, J. J. and Shoichet, B. K.: ZINC - A Free Database of Commercially Available Compounds for Virtual Screening, *Journal of Chemical Infor-*

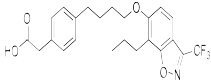
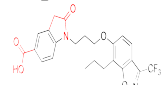
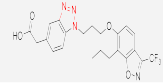
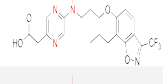
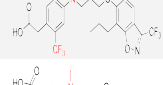
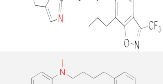
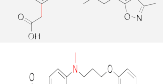
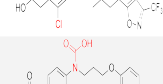
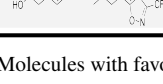
| Structure | Similarity to agonist (PDB code 1PQ6) Sim_{2D} | Scoring S_{Total} | E_{ff} [kcal/mol] | MW [g/mol] |
|---|--|---------------------|---------------------|------------|
| Target (1PQ6)  | (1) | - | -10.7 | 582.5 |
| Generated molecules with optimal score S_{Total} | | | | |
|  | 0.705 | -0.721 | -10.0 | 535.2 |
|  | 0.666 | -0.694 | -10.1 | 547.8 |
|  | 0.659 | -0.687 | -9.7 | 539.0 |
|  | 0.642 | -0.673 | -9.5 | 567.3 |
|  | 0.573 | -0.541 | -10.2 | 559.4 |
|  | 0.512 | -0.525 | -10.5 | 571.5 |
|  | 0.510 | -0.516 | -10.8 | 541.7 |
|  | 0.507 | -0.499 | -10.6 | 589.3 |

Table 1 Molecules with favorable score S_{total} generated in the 10 runs with $w_{ff} = 0.02$

mation and Modeling, Vol. 45, No. 1, pp. 177–182 (2005).

- 6) Lima, C. F., Pelikan, M., Goldberg, D. E., Lobo, F.G., Sastry, K. and Hauschild, M.: Influence of selection and replacement strategies on linkage learning in BOA, *IEEE Congress on Evolutionary Computation*, pp.1083–1090 (2007).
- 7) Majeux, N., Scarsi, M., Apostolakis, J., Ehrhardt, C. and Caffisch, A.: Exhaustive docking of molecular fragments with electrostatic solvation, *Proteins: Struct. Funct. Genet.*, Vol. 37, No. 1, pp. 88–105 (1999).
- 8) Muhlenbein, H., Mahnig, T. and Rodriguez, A.O.: Schemata, Distributions and Graphical Models in Evolutionary Optimization, *Journal of Heuristics*, Vol.5, pp.215–247 (1999).
- 9) Munawar, A., Wahib, M., Munetomo, M. and Akama, K.: Theoretical and Empirical Analysis of a GPU based Parallel Bayesian Optimization Algorithm, *PDAA'09: Proceedings International Work-*

shop on Parallel and Distributed Algorithms and Applications, IEEE Press (2009).

- 10) Pelikan, M., Goldberg, D. E. and Cantu-Paz, E.: BOA: The Bayesian Optimization Algorithm, Morgan Kaufmann, pp.525–532 (1999).
- 11) S, S., T, O., and Norstrom, C. J.M., K, S., D, H., IC, J., K, Z., D, O. and L., J.: Crystal structure of the heterodimeric complex of LXR alpha and RXR beta ligand-binding domains in a fully agonistic conformation, *EMBO Journal*, Vol. 22, pp. 4625–4633 (2003).
- 12) Stone, J.E., Phillips, J.C., Freddolino, P.L., Hardy, D.J., Trabuco, L. G. and Schulten, K.: Accelerating molecular modeling applications with graphics processors, *Journal of Computational Chemistry*, Vol.28, pp.2618–2640 (2007).