Hybrid Parallelization of Protein-Ligand Docking using Fast Fourier Transformations and Rigid Body conformation

Abhishek.K, S. Balaji

Abstract: ***Docking has become a very important technique catering a wide spectrum of applications and Drug Design / Discovery is one such domain. Docking has become an integral part of proteomics ever since its advent. With the advent of ubiquitous computing and computational proteomics, docking technique has become very invaluable technique for in-silico analysis. Docking is a technique in which the preferred orientation of one molecule to another to form a stable complex is determined or predicted. This knowledge of orientation can be applied to know about binding affinity between two molecules which is known as Scoring Functions (SF). Scoring functions essentially give an insight about how likely one molecule gets bound with another molecule Virtual screening is an in-silico technique which is mainly used in the drug discovery to search libraries of small molecules to identify those structures which are most likely to bind to a target. In virtual screening a large set of libraries of compounds are evaluated to re-score the enrichment or to find out the binding affinity. In this work, we study and evaluate Heterogeneous Parallel Processing Based Virtual Screening Pipeline for Effective Rescoring in Protein-Ligand Docking. We present the study of FFT based docking algorithms that leverages High-Performance Computing platform which is horizontally scalable and enterprise a better performance.***

Keywords: High-performance computing, ligand protein docking, GPU, Multiconformer library, FFT

1. INTRODUCTION

Computational techniques have been a major factor in the proteomics studies in the recent decades. However, with the data that has been collected, it has become even challenging to address the task. Conventional techniques uses only one single computing machinery to do all the heavy-lifting tasks of the experiments and hence would take longer time thus, proving inefficient. However, the newer High-Performance Computing (HPC) models helps us in addressing these challenges very efficiently as they are highly scalable. HPC allows us to exploit the inherent capabilities of a wide range of computational machines to solve difficult problems.

With HPC problems can be broken up into several independent modules which can be categorized and distributed to the computational components that are specifically built to handle such problems.

Molecular recognition through the protein–ligand interactions is a fundamentally important of all the other processes occurring inside the organisms. Transmission of signals that happens due to such molecular complementarity has found to be the driving force in such processes.

The evolution of protein function comprises of the development of highly specific sites for the binding of ligands with the affinity parameters set to mimic the biological function. The “best bind” is said to be in place when ligand binds in the most suitable form so as to find its role in the regulation of biological function.

Docking using Computational approach is used widely for the study of protein-ligand interactions to understand steps towards drug discovery and its development. The process generally begins with a target molecule whose structure is generally known, such as a crystallographic structure [6]. Docking finds its application in the prediction of the bound conformation and to understand the binding free energy of some smaller molecules against the target. Typically, Single docking experiments are beneficial in understanding the function of the target. In virtual screening, a large library of compounds is docked and ranked. The primary idea behind virtual screening is to screen the library of ligands that are present, to identify compounds for experimental testing.

At the heart of any biological process like cell regulation, recognition of antibodies and their corresponding antigens, the transduction of signal, gene expression lays the molecular interactions. These interactions comprise interactions between different proteins, interactions between a drug and a protein (useful in drug design and discovery) etc. To perform their respective biological functions, it is very essential for the formation of stable protein-protein or protein-ligand complexes which, are formed because of the aforementioned molecular interactions.

In order to understand the mode of binding and the affinity of the molecules involved in the molecular interactions, the study of the tertiary structure of the proteins is very important. With the advent of technology, we can obtain the complex structure by X-ray Crystallography and NMR methods. However, obtaining the structures by such methods in most of the cases is challenging and not economically feasible [4].

Revised Manuscript Received on March 02, 2019.

Abhishek.K, Research Scholar-Jain University, Dept. of Information Science & Engineering., Jyothy Institute of Technology, Bengaluru-560082, India

S. Balaji, Centre for Incubation, Innovation, Research and Consultancy, Jyothy Institute of Technology, Tataguni, Bengaluru-560082, India.

With the advent of better computational infrastructure, we can use computational methods like docking to understand these interactions and thus making it a very important approach.

In this work, we present the study of FFT based protein-ligand docking algorithms and leverage Heterogeneous Parallel Processing platform based Virtual Screening Pipeline to study the specific parameters of Effective Rescoring in Protein-Ligand Docking.

2. METHOD

2.1 Idea

Until recently, most of the computational approaches for proteomics were largely constrained due to the non-availability of the scalable computing infrastructure and hybrid computing. With the advent of this kind of infrastructure, it has now become feasible to experiment using these infrastructures. The traditional approach to developing algorithms for protein studies was more monolithic in nature which means that the algorithm would take closer to exponential runtimes as they were sequential. With the parallel processing infrastructure like GPUs becoming feasible it has opened new avenues for the algorithm developers to embrace a more parallel paradigm towards algorithm development [3].

Most conventional docking algorithms use a traditional FFT-based rigid-docking scheme. The performance of this scheme is dependent on factors like electrostatics, the complementarity of the shape, free energy etc. (Ohue et al., 2012, 2014) Have calculated this using one correlation function.

Our work leverages the multiple FFT calculations which are used to calculate multiple effects (Kozakov et al., 2006; Pierce et al., 2011).

2.2 GPU Implementation

We have studied the conventional docking algorithms and have leveraged the GPU computing platform and have performed the extensive study of the algorithms on multiple GPUs using CUDA library. We have mapped the entire pipeline of the docking process on the GPU and leverage all the cores for computation. The general schematic for the GPU processing is as shown in Fig1.

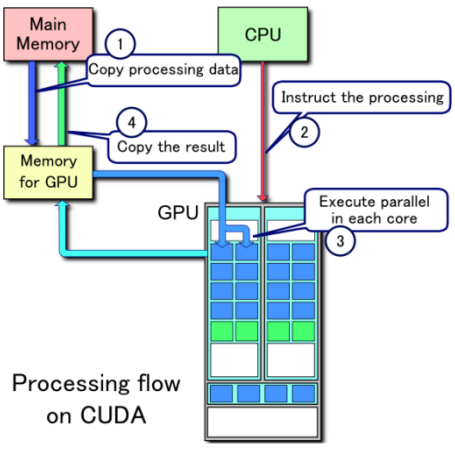

Fig 1: GPU processing Schematic

The main memory is in sync with the GPU Memory and this ensures faster copying of the data. The CPU is involved in the transfer of the instruction pipeline which is executed in parallel on each code of the GPU which in principle contributes for the faster and scalable computation.

2.3 Hybrid CUDA Parallelization

In Multiconformer docking algorithms or Rigid-body docking methods in order to find out the ligand flexibility a single confirmation or multiconfirmation library is used [5].

Ziyi Guo; Brian Y. Chen et. al. [5] have employed an approach which generally docks the small molecules. This approach uses the shape complementarity algorithm or the interaction site matching algorithm.

The studies in [6-8] suggest that the algorithms work on the pharmacophore which is used as a protein representation which guides the docking. The study also suggests generating an initial ligand conformation and uses the same to derive a ligand pharmacophore [9].

The efficiency of any docking programme is determined by 2 components that complement each other which are:
a) Methods employed in exploring the conformational space of the target. b) The Scoring Function(SF) used to evaluate the docking poses.

A scoring function (SF) as studies [10-12] suggests that, should as de-facto assign the best score to the ‘correct pose’. This is the native posed which is observed during the study of crystalline structure of the target. This best score then acts as pivot for the algorithm used of conformational sampling [1].

Smith R.D et.al [15] suggest that accurate prediction of binding mode is very critical in docking studies and the former of the aforementioned parameters is very critical in determining the binding mode [8]. It is trivial to mention that the SF functions should attribute the best scores to the docked poses of the compounds that are highly active that compared to that of the non-binders or non-active / poor binders.

Also, it is to be further noted that in Virtual Screening and lead optimization, it is very critical to extract the potential hits from the huge libraries and the latter of the aforementioned parameters is very critical to it.

Conventional Algorithms have used OpenMP and MPI using a master-slave model (Matsuzaki et al., 2013). In the cluster model the list of protein pairs is fetched by the master node, which is then distributed across the available nodes to the worker processes [7]. The advantage of such a model is that it is fault tolerant and the consistency of the system is maintained unlike a monolithic system.

Our work is implemented on CUDA parallelization. It becomes imperative to optimize the memory utilization. We performed a 1-1 mapping between docking job and the node so that the ligand rotation can be parallelized on the GPU. The docking jobs are distributed by the master node and the worker nodes execute these jobs on the available GPUs by CUDA across the cluster. This scheme of implementation guarantees the fault-tolerance as in case of CPU implementation [2].

3. RESULTS

The docking score, which is also the pseudo-interaction energy score can be determined by the convolution of the FFT and the inverse FFT functions as follows:

$$S(t) = \sum_{v \in V} R(v)L(v+t) \quad (1)$$

$$= FFT_{inv}\,[FFT\,[R(v)] * FFT[L(v]] \quad (2)$$

Where:

R & L – Scoring functions of receptor and ligand respectively in a 3D space V

t – the parallel translation vector in the 3D space

'*' – is defined as the complex conjugation operator

N – is the size of the FFT (2 times grid size)

The algorithm to solve (1) takes about$O(N^6)$, which essentially means takes a longer runtime and larger the size of FFT more the time taken.

This however can be reduced to $O(N^3 \log N)$ using (2) which intrinsically uses FFT. It is very important to note here that the FFT can be computed parallel using the GPUs.

Figure 2 shows the FFT time vs the Total Time taken for docking on different GPUs. We can note that the FFT takes just about an average of 15% of total time for docking.

We have used NVIDIA GeForce GT 710, GeForce GT 705, GeForce GT 730 to check the performance. Figure 1 shows the measurements of the 3600 rotations on each of the selected GPUs. Whereas Figure 3 shows the performance on select CPUs.

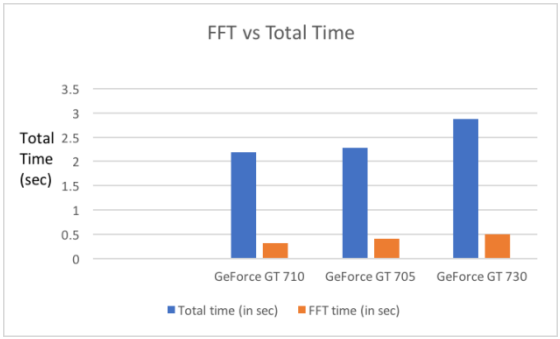

Fig 2: FFT vs. Total Time

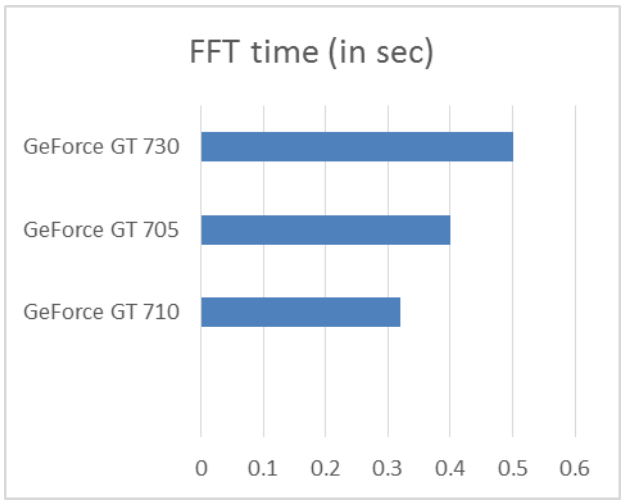

Fig 1: GPU performance

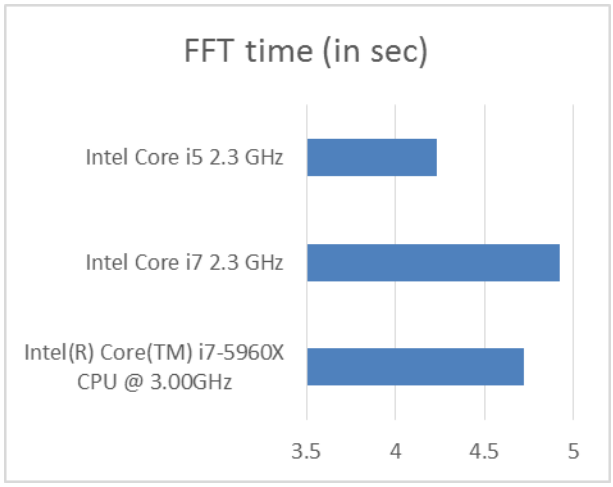

Fig 4: CPU Performance

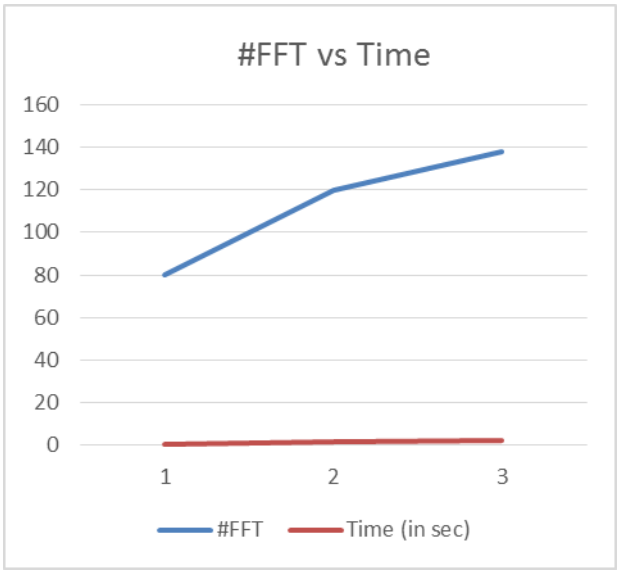

Fig 5: #FFT vs. Time (in sec)

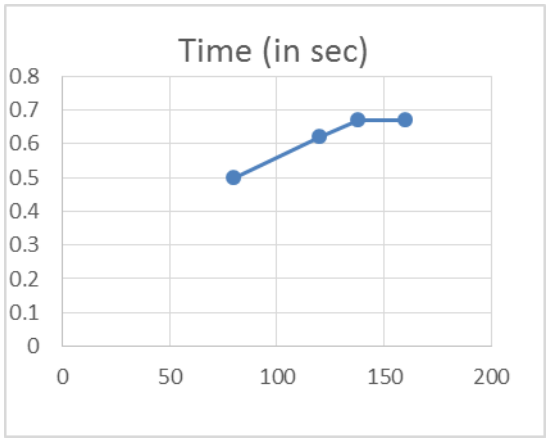

Fig 6: GPU Cluster performance

Figure 3 shows the measurements of the FFT computation time on the aforementioned GPUs whereas Figure4 shows the measurement of FFT computation on CPU. We can observe from Figure4 that on a single node GPU the time taken for the FFT calculation is growing exponentially

whereas it becomes constant on a cluster of 4 (Figure 5). We can perform docking of heavier molecules and test for scalability in future.

4. CONCLUSIONS

This study of docking on high performance computing environments shows high scalability. Also, it is found out that the scoring function can be improved by using the heterogeneous parallel computing environment. Complete leverage of such computing environments helps us build effective virtual pipeline for effective scoring functions.

REFERENCES

[1] Khushboo Babaria; Sanya Ambegaokar; Shubhankar Das; Hemant Palivela, "Algorithms for ligand based virtual screening in drug discovery", International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 10.1109/ICATCCT.2015.7457004, 2016

[2] P. B. Jayaraj; K. Rahamathulla; G. Gopakumar, "A GPU Based Maximum Common Subgraph Algorithm for Drug Discovery Applications", IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), 10.1109/IPDPSW.2016.65, 2016

[3] Ajinkya Nikam; Akshay Nara; Deepak Paliwal; S. M. Walunj, "Acceleration of drug discovery process on GPU" , Green Computing and Internet of Things (ICGCIoT), 10.1109/ICGCIoT.2015.7380432,2015

[4] Majid Rastegar-Mojarad; Ravikumar Komandur Elayavilli; Dingcheng Li;Rashmi Prasad; Hongfang Liu, "A new method for prioritizing drug repositioning candidates extracted by literature-based discovery", IEEE transactions on Bioinformatics and Biomedicine (BIBM), 10.1109/BIBM.2015.7359766, 2015

[5] Ziyi Guo; Brian Y. Chen, "Predicting protein-ligand binding specificity based on ensemble clustering", IEEE transactions on Bioinformatics and Biomedicine (BIBM), 10.1109/BIBM.2015.7359858, 2015

[6] Peng Chen; ShanShan Hu; Jun Zhang; Xin Gao; Jinyan Li; Junfeng Xia; Bing Wang, "A sequence-based dynamic ensemble learning system for proteinligand-binding site prediction", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10.1109/TCBB.2015.2505286,2015

[7] Dong-Jun Yu; Jun Hu; Qian-Mu Li; Zhen-Min Tang; Jing-Yu Yang; Hong-Bin Shen, "Constructing Query-Driven Dynamic Machine Learning Model With Application to Protein-Ligand Binding Sites Prediction", IEEE Transactions on NanoBioscience, 10.1109/TNB.2015.2394328,2015

[8] Nishamol P H,Gopakumar G, "Multi-target Drug Discovery Using System Polypharmacology –State of the art",978-1-4799-1823-2, 2015

[9] Hossam M. Ashtawy; Nihar R. Mahapatra, "A Comparative Assessment of Predictive Accuracies of Conventional and Machine Learning Scoring Functions for Protein-Ligand Binding Affinity Prediction", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2351824,2015

[10] Xiaohua Zhang, Sergio E. Wong, and Felice C. Lightstone, " Toward Fully Automated High Performance Computing Drug Discovery: A Massively Parellel Virtual Screening Pipeline for Docking and Molecular Mechanics /Generalised Born Surface Area Rescoring to Improve Enrichment", Jounal of Chemical Information and Modeling,10.1021,2014

[11] Daniel Li; Brian Tsui; Charles Xue; Jason H. Haga; Kohei Ichikawa; Susumu Date, "Protein Structure Modeling in a Grid Computing Environment", 1109,2013

[12] Ginny Y. Wong; Frank H. F. Leung; S. H. Ling, "Predicting Protein-Ligand Binding Site Using Support Vector Machine with Protein Properties", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10.1109/TCBB.2013.126,2013

[13] Pratyusha Rakshit; Amit Konar; Archana Chowdhury; Eunjin Kim; Atulya K. Nagar, "Muti-objective evolutionary approach of ligand design for protein-liganddocking problem", 2013 IEEE Congress on Evolutionary Computation, CEC.2013.6557576,2013

[14] Ankur Dhanik,John S Mclvlurrayl and Lydia Kavraki," AutoDock-based incremental docking protocol to improve docking of large ligands", 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops, 978-1-4673-2747-3, 2012

[15] Smith, R. D.; Dunbar, J. B.; Ung, P. M. U.; Esposito, E. X.; Yang,C. Y.; Wang, S. M.; Carlson, H. A. CSAR Benchmark Exercise of 2010:Combined Evaluation Across All Submitted Scoring Functions. J.Chem. Inf. Model, 51, 2115−2131, 2012