International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5, June 2019

# Scoring Function Enrichment and Optimization Techniques for Docking in Heterogeneous Parallel Platforms

**Abhishek K, S. Balaji**

***Abstract:Ever since the discovery of micro-organisms and the ways to be safe from their infection, the scientific community has been intrigued on knowing further to the depths of the various domains of what we call today as life sciences. The major game changer event was the discovery of proteins which were thought to be the most fundamental building blocks. Further study of proteins was greatly hampered due to various factors like non-availability of techniques to purify proteins in large quantities and the computational power and the associated costs. However, with the advent of the distributed and the parallel processing techniques and High-Performance Computing (HPC) it has become possible to perform many in-silico experiments. Furthermore, with the exponential reduction in cost of computing, it has become possible to perform docking studies and related experiments. With advances in areas like Graphical Processing Unit (GPU) computing and prediction algorithms the drug design and drug discovery techniques have seen an exponential growth in the traction that they have received from the academia and industry as well. Today, it wouldn't be wrong to say that docking study has become one of the most vital parts of computational proteomics and drug discovery. It is used in predicting the orientation of preferred molecule with another protein or ligand to create a stable complex. This information can be further used to design the scoring functions which are important in determining the likelihood of binding between a molecule with another ligand or a protein molecule. To facilitate this process further and to identify the molecules from their respective databases and study their structure, in-silico techniques like virtual screening are employed. From the afore-mentioned information, it is trivial that Scoring Function (SF) is the heart of the docking process and the performance of the SF is very critical to the docking study. In our previous studies we have presented our work on different techniques [1] [2].***

***Keywords:Docking, FFTs on GPU, Spherical Transforms Protein-Ligand Docking***

## I. INTRODUCTION

Molecular recognition through the protein–ligand interactions is fundamentally the 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 [8]. 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 afore-mentioned 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 [6]. 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. Protein-protein interactions are very central to any biological function, so much so that it has become imperative to gain the knowledge of the structure of the target complex in such studies. Although there exist several techniques like NMR, X-ray Crystallography which help to fetch the structure of the target complex, actuating all the structures of interest remains a challenge due to various factors which involves the tradeoff between cost and efficiency.

**Manuscript published on 30 June 2019.**
* Correspondence Author (s)
**Abhishek K*,** Research Scholar-Jain University, Dept. of Information Science & Engineering., Jyothy Institute of Technology, Tataguni, Bengaluru-560082, India
**S. Balaji,** Centre for Incubation, Innovation, Research and Consultancy, Jyothy Institute of Technology, Tataguni, Bengaluru-560082, India

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

In this respect, it can be said that the in-silico techniques employed to predict the target complex (known as protein-protein docking) has gathered a lot of traction from the scientific community with the hope that it will provide the required structural information that the in-vivo / in-vitro methods fail to provide.

In general, protein-protein docking is characterized by the 3-D (three dimensional) structure of the target complex by leveraging the information from the unbound monomers. In most of the literature, it has become the de-facto standard to assume that all the information that can be leveraged in a docking process is nothing but the co-ordinates of the monomers sans any data which refers to the binding sites of any protein.

The works in [17], [18], [19] give a very comprehensive summary of several docking methods that have been proposed over time. One of the daunting challenges that the scientific community is facing is the computational complexity of the problem and also the high degree of the freedom the target complex system comes with. Hence, it becomes imperative to be more agile and adapt techniques that can scale up so much so that we can reach a solution in real time.

The general approach towards this problem is two-fold in the sense that:

1. We must perform a holistic search and then predict the probable candidates. This usually relies on techniques like scoring function and assumption of a rigid body model to reduce the search space.
2. This stage is where we refine the results or outcome of the preceding stage and refine the probable candidates that have been predicted in phase 1. This phase includes techniques that are more computationally intense since they deal with pose ranking and the structural parameters. The initial phase of the search and predicting the probable candidates is very crucial for the overall success of the docking approach.

Surface feature point matching techniques as explained in [7-11] and the techniques based on the energy minimization as discussed in [27-29] and also the algorithms which perform a global search that leverages Fast Fourier Transforms (FFT) as discussed in [30-32] have been leveraged and been experimented with for performing a holistic search and for predicting the probable candidates which is the first phase as mentioned above. Though there have been lot of studies conducted on the afore-mentioned techniques and algorithms, these algorithms suffer with the traditional tradeoff between the computational complexity (time) and accuracy of the predictions. The works described in [30],[33] leverage the concept of FFTs where the authors claim that FFTs help them achieve an equilibrium between the time complexity and the accuracy which means that the scoring algorithm can be designed with agility and also we can achieve a fair accuracy.

One other approach to overcome the trade-off as suggested in the works [34-35] leverages the spherical aspects of the protein molecules unlike the other works which leverage the FFTs. The FFTs being very efficient though, cannot tap the multi-dimensional aspect of a molecule and hence we base our current work on the spherical aspects.

Ritchie's work [34], [35] uses a radial basis function which is reported to be successful in optimizing the time complexity of the docking algorithm and that makes it a very promising method. This work further states that, as the distance from origin **r** increases owing to its radial basis function, the accuracy of field expression drastically reduces. This intuitively means that it becomes increasingly difficult to apply it on larger protein molecules. To overcome these shortcomings the authors have proposed to leverage the spherical harmonics and modified Legendre polynomials in combination which forms the radial basis function. This means that there is no decay [36] for **r** which is the distance from origin.

Scoring Functions (SF) are mathematical models in computational chemistry which are used to predict the non-covalent interaction. This interaction is also called as binding affinity. Binding affinity is essential for docking. Scoring functions are also used to predict the strength of intermolecular interactions. Commonly used molecules include drug and biological target. Scoring functions are trained against data which consists of determined binding affinities between molecules similar to the unpredicted molecules.

For currently used methods aiming to predict affinities of ligands for proteins, the following must first be known or predicted:

- **Protein tertiary structure** – Structure of the protein atoms in three-dimensional space. Protein structures are determined by experimental techniques such as X-ray crystallography or solution phase NMR methods or predicted by homology modelling.
- **Ligand active conformation** – three-dimensional shape of the ligand when bound to the protein
- **Binding-mode** –Binding mode is the positioning of the two binding associates relative to each other in the complex. The above information gives the three-dimensional structure of the complex. The SF can then estimate the strength of the association between the two molecules in the complex using one of the methods given below. The SF itself may be used to predict binding mode and the active conformation of the small molecule in the complex, or alternatively a simpler and faster function may be utilised within the docking run.

There are four general classes of scoring functions:

- **Force field** – affinities are estimated by adding the strength of intermolecular van der Waals and electrostatic interactions between all atoms of the two molecules in the complex. The intramolecular energies (also referred to as strain energy) of the two binding partners are also frequently included. Since the binding normally takes place in the presence of water, the desolvation energies of the ligand and of the protein are sometimes taken into account using implicit solvation methods such as GBSA or PBSA can be utilized [40].


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

- **Empirical** – based on counting the number of various types of interactions between the two binding partners. Counting may be based on the number of ligand and receptor atoms in contact with each other or by calculating the change in solvent accessible surface area (ΔSASA) in the complex compared to the uncomplexed ligand and protein. The coefficients of the scoring function are usually fit using multiple linear regression methods. These interaction terms of the function may include for example:
  - hydrophobic — hydrophobic contacts (favorable),
  - hydrophobic — hydrophilic contacts (unfavorable) (Accounts for unmet hydrogen bonds, which are an important enthalpic contribution to binding. One lost hydrogen bond can account for 1–2 orders of magnitude in binding affinity),
  - number of hydrogen bonds (favorable contribution to affinity, especially if shielded from solvent, if solvent exposed no contribution),
  - number of rotatable bonds immobilized in complex formation (unfavourable conformational entropy contribution).
- **Knowledge-based** –It is based on observations of intermolecular close contacts in large 3D databases (such as the Cambridge Structural Database or Protein Data Bank) which are used to derive "potentials of mean force". It is founded on the assumption that close intermolecular connections between certain types of atoms or functional groups that occur frequently than one would expect by a random distribution are likely to be energetically positive and therefore contribute favourably to binding affinity.
- **Machine-learning** – Classical scoring functions use structure as the input, machine-learning scoring functions are considered by not assuming a predetermined functional form for the relationship between binding affinity and the structural features describing the protein-ligand complex. The functional form is inferred directly from the data. Machine-learning scoring functions have been found to outperform classical scoring functions at binding affinity prediction of diverse protein-ligand complexes. This has also been the case for target-specific complexes although the advantage is target-dependent and mainly depends on the volume of relevant data available [33].Machine-learning scoring functions perform at least as well as classical scoring functions at the related problem of structure-based virtual screening [8].

### *1.1 Ligand Flexibility*

It is important to select reasonable confirmations for proper docking. Confirmations can be generated in the presence of binding cavity or using dihedral angle which employs fragments. Force field energy is employed to select proper orientations.

### *1.2 Receptor Flexibility*

Computational capacity has increased dramatically over the last decade making possible the use of more sophisticated and computationally intensive methods in computer-assisted drug design. However, dealing with receptor flexibility in docking methodologies is still a thorny issue. The main reason behind this difficulty is the large number of degrees of freedom that have to be considered in this kind of calculations. Neglecting it, however, leads to poor docking results in terms of binding pose prediction.

Multiple static structures experimentally determined for the same protein in different conformations are often used to emulate receptor flexibility. Alternatively, rotamer libraries of amino acid side chains that surround the binding cavity may be searched to generate alternate but energetically reasonable protein conformations. Finally, hybrid scoring functions have also been developed in which the components from two or more of the above scoring functions are combined into one function.The ease of access to high performance computing resources and the decrease of the computational cost have boosted the development of computational techniques in proteomics

Figure 1 shows the overall process. The process of structure-based drug design is an iterative one and often proceeds through multiple cycles before an optimized lead goes into phase I clinical trials. The first cycle includes the cloning, purification and structure determination of the target protein or nucleic acid by X-ray crystallography, NMR, or homology modelling. Using computer algorithms, compounds or fragments are positioned into a selected region of the structure. These compounds are scored and ranked based on their steric and electrostatic interactions with the target site and the best compounds are tested with biochemical assays. In the second cycle structure determination of the target in complex with a promising lead from the first cycle, one with at least micromolar inhibition in vitro, reveals sites on the compound that can be optimized to increase potency. Additional cycles include synthesis of the optimized lead, structure determination of the new target: lead complex, and further optimization of the lead compound. After several cycles of the drug design process, the optimized compounds usually show marked improvement in binding and, often, specificity for the target.

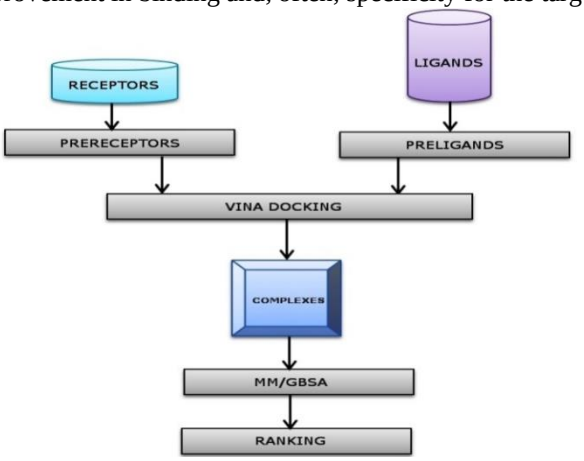


**Figure 1: Overall Docking Process**


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

In this work, we present an empirical study of docking process and the various techniques to improve the scoring function which include evaluation of the virtual screening pipeline based on the HPC techniques thus facilitating and effective rescoring techniques. Furthermore, we have extended our study from FFTs to impact of hyper parameter tuning of the spherical polar transform function to improve the efficiency of overall docking process.

### *1.3 The Affinity of Protein Ligand Docking*

Computer-aided drug design uses computational chemistry to realize, enhance, or study drugs and related biologically active molecules. The fundamental goal is to predict whether a given molecule will bind to a target and if so how strong. Molecular mechanics are most often used to predict the conformation of the small molecule and to model conformational changes in the biological target that may occur when the small molecule binds to it. Also, knowledge-based scoring function may be used to provide binding affinity estimates. These methods use linear regression, neural nets or other statistical techniques to obtain binding affinity equations by fitting new affinities to derived interaction energies between the small molecule and the target. Semi-empirical, ab initio quantum chemistry methods, or density functional theory are often used to provide optimized parameters for the molecular mechanics calculations and also provide an estimate of the electronic properties (electrostatic potential, polarizability, etc.) of the drug candidate that will influence binding affinity [10].

Ideally, the computational method should be able to predict affinity before a compound is synthesized and hence in theory only one compound needs to be synthesized. The reality however is that present computational methods are imperfect and provide at best only qualitatively accurate estimates of affinity. Therefore, in practice, it still takes several iterations of design, synthesis, and testing before an optimal molecule is discovered. On the other hand, computational methods have accelerated discovery by reducing the number of iterations required and in addition have often provided more novel small molecule structures. Drug design with the help of computers may be used at any of the following stages of drug discovery:

1. Hit identification using virtual screening (structure- or ligand-based design)
2. Hit-to-lead optimization of affinity and selectivity (structure-based design, QSAR, etc.)
3. Lead optimization: optimization of other pharmaceutical properties while maintaining affinity

To overcome the insufficient prediction of binding affinity calculated by recent scoring functions, the protein-ligand Interaction and compound 3D structure information are used for analysis.

## II. EFFECT OF FFTS ON SCORING FUNCTION

### *2.1 The Concept*

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 [7]. 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, Potential Static Charge (PSC), free energy etc. Our work leverages the multiple FFT calculations which are used to calculate multiple effects in the paper “A geometric approach to macromolecule-ligand interactions” [17]

In ligand rotation process, the atomic coordinates of a ligand are updated according to a given rotation matrix. The process is independent for each atom and it can be fully parallelized. We mapped the atomic coordinates onto a GPU. In ligand voxelization process, it is imperative to set a suitable PSC score, electrostatic interaction values, and free energy scores for the ligand voxel model during this process. Ligand voxelization calculates the distance between the coordinates of an atom and each grid, before assigning a value to each grid within the van der Waals radius of the atom [41-46]. The assignment process can be parallelized for each atom. The PSC score and the free energy score of a ligand has only binary states (0 or 1), and the electrostatic interaction value of a grid is calculated as the cumulative sum of the values of all adjacent atoms, thus the calculation order for each atom can be exchanged freely. Therefore, we processed the atoms in parallel and mapped them onto a GPU. Thus, multiple atoms were processed simultaneously on different GPU cores in this process. In FFT processes (P3, P5, P7), single precision complex 3-dimentional FFT is performed using the NVIDIA cuFFT library to map the FFT calculations onto a GPU. In convolution process, the output of FFT of receptor voxel is complex conjugated and multiplied by the output of FFT of ligand voxel. The convolution can be independent for each grid, thus we mapped them onto a GPU. In identifying the best solutions process, the best docking pose was selected according to the docking score. This process was also implemented on a GPU using reduction.

### *2.2 Implementation on GPU (HPC)*

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 leveraged all the cores for computation. The general schematic for the GPU processing is as shown in Figure 2.


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

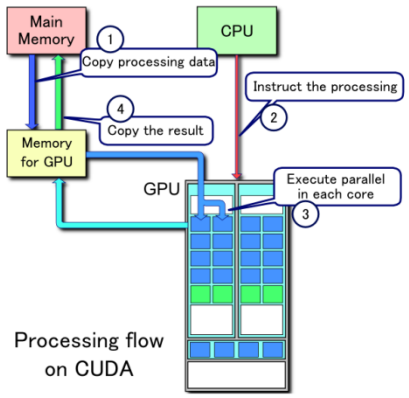


**Figure 2: GPU Processing Schematic**

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 multi-conformer docking algorithms or rigid-body docking methods, to find out the ligand flexibility, a single confirmation or multi-confirmation library is used [7]. ZiyiGuo; Brian Y. Chen et. al. [7] 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 [8-10] suggest that the algorithms work on the pharmacophore which is used as a protein representation that guides the docking. The study also suggests generating an initial ligand conformation and uses the same to derive a ligand pharmacophore [11].

The efficiency of any docking programme is determined by two components that complement each other: (a) methods employed in exploring the conformational space of the target and (b) the scoring function used to evaluate the docking process. A scoring function as studied in [12-14] suggests that it should 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 [3]. Smith R.D et.al [17] suggest that accurate prediction of binding mode is very critical in docking studies and the former of the afore-mentioned parameters is very critical in determining the binding mode [10]. 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 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 afore-mentioned 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 [11]. 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 the case of CPU implementation [4].

## III. MOLECULE AS A RIGID BODY

*3.1 Scoring Functions Using Spherical Transforms*

Scoring Function is characterized by the interaction energy between two molecules. In the present work, we determine the scoring function in terms of dot product of the two scalar fields that are associated with the molecules.

Let f1(x), f2(x), f3(x),……,fNs(x) be the scalar field functions for molecule A and g1(x), g2(x), g3(x),……, gNs(x) be the scalar field functions for molecule B. The scoring function would be formulated as follows:

$$E(T^A, T^B) \equiv \sum_{i=i}^{N_S} w_i \int f_i^{T^4}(x) g_i^{T^B}(x) dx \quad \text{(eq 1)}$$

where$w_i$ represents the weight of the $i^{th}$ term,

Tx – rotational or translational operation on a field for molecule x and

fT(x) is the field generated by applying the operation T to x.

*3.2 Leveraging Radial Basis Functions*

By representing the scalar fields in terms of orthogonal basis functions, the score computation becomes much faster. The basis function B can be expressed as

$$B_{k,n,l,m}(x) = B_{k,n,l,m}(r,\theta,\varphi) \equiv S_{k,n}(r) Y_{l,m}(\theta,\varphi) \quad \text{(eq 2)}$$

where,

Sk,n(r) represents the radial part of the basis function and

$S_{k,n}(r)Y_{l,m}(\theta,\varphi)$– normalized spherical harmonics which is the angular part of basis function.

As discussed earlier the radial part r of the basis function is split into multiple intervals Ik of widths 'a'; that is,

Ik ≡ [ka,(k+1)a) for k=0,1,…, then

Sk,n(r) for each region can be defined as

Sk,n(r) = 0 if r∉[ka,(k+1)a)

and

$$\int_0^\infty S_{k,n}(r) S_{k,n'}(r) r^2 dr = \delta_{nn'} \text{ ---------- (eq 3)}$$

By leveraging the Gram-Schmidt process we can satisfy the afore-mentioned conditions:

$$S_{k,n}(r) \equiv \begin{cases} \sqrt{\frac{8}{N_{k,n}^2 a^3}} h_{k,n} \left(\frac{2}{(a)} r - 2k - 1\right), \text{ r} \in [\text{ka}, (\text{k+1})\text{a}) \\ 0, \text{ otherwise} \end{cases} \quad \text{(eq 4)}$$

hk,n(x) – orthogonal polynomials characterized using Gram-Schmidt Process.

The weight function and the intervals used for the Gram-Schmidt process are (x+2k+1)2 and [-1,1], respectively.

Nk,n – norm of the polynomial function. It now becomes easy to realize that the orthonormality

$$\int_0^\infty S_{k,n}(r) S_{k',n'}(r) r^2 dr = \delta_{kk'} \delta_{nn'} \quad \text{(eq 6)}$$


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

The orthonormality of the combined functions can be similarly expressed as

$$\int S_{k,n}(r)Y_{l,m}(\theta,\emptyset)S_{k',n'}(r)Y_{l',m'}(\theta,\emptyset)dx = \delta_{kk'}\delta_{nn'}\delta_{ll'}\delta_{mm'} \quad \text{(eq 7)}$$

There can exist in some regions where the radial basis function can have non-zero values which are as shown in Figure 3.

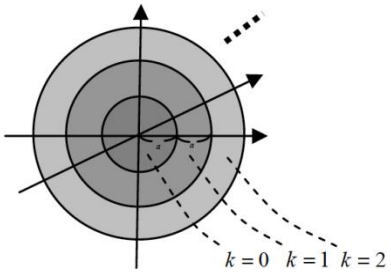


**Figure 3: Regions of Radial Basis Functions**

*3.3 Fast Rotational and Translational Operators*

In order to achieve better performance, we need to obtain the following transformed coefficients $a^{T^A}_{k,n,l,m}$ and $b^{T^B}_{k,n,l,m}$ which facilitate configuration space search. The original coefficients characterize the new transformed coefficients and hence computing the re-expansion of the fields become much faster.

*a) Rotational Operation on Coefficients*

Let $a_{k,n,l,m}$ be the original coefficients and let R be the rotational operator on the field. Let $a^R_{k,n,l,m}$ be the new rotational coefficients. As previously mentioned, we can derive the rotational coefficients using the original coefficients as follows:

$a^R_{k,n,l,m} = \sum^{l}_{m'=-l} a_{k,n,l,m'} R^l_{mm'}(R^{-1})$; Rlmm'(R) represents the rotational matrices for real spherical harmonics.

*b) Translational Operation on Coefficients*

Let $a^{S_{\Delta Z}}_{k,n,l,m}$ represent the coefficients of a translated field. Note that SΔz intuitively indicates that the translation operation has been applied along the Z-axis.

The new translated coefficients $a^{S_{\Delta Z}}_{k,n,l,m}$ with an offset of (0,0,Δz) can be determined as follows:

$$a^{S_{\Delta Z}}_{k,n,l,m} = \sum_{k'n'l'} a_{k',n',l',m'} \int_0^{\infty}\int_0^{\pi} S_{k',n'}(r')S_{k,n}(r)P^{|m|}_{l'}(cos\theta')P^{|m|}_{l}(cos\theta)r^2 sin\theta\, d\theta\, dr \equiv \sum_{k'n'l'} a_{k',n',l',m} O_{k',k,n',n,l',|m|}(\Delta z) \quad \text{(eq 8)}$$

$P^m_l(x)$ is the Legendre Polynomial

$O_{k',k,n',n,l',|m|}(\Delta z)$ are the overlap integrals during the translation.

It is very important to note here that the overlap $O_{k',k,n',n,l',|m|}(\Delta z)$ can be calculated using the numerical integration methods and are calculated in advance at each step and stored in a lookup table. This indeed is a very practical approach since they are independent of scalar fields.

## IV. RESULTS

*4.1 Results of GPU Implementation*

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:

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

$= FFT_{inv}\,[FFT\,[R(v)] * FFT[L(v]]$ --- (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 $O(N^3 logN)$ using (2) which intrinsically uses FFT. It is very important to note here that the FFT can be computed in parallel using the GPUs.

Figure 4 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 2 shows the measurements of the 3600 rotations on each of the selected GPUs. Whereas Figure 5 shows the performance on select GPUs.

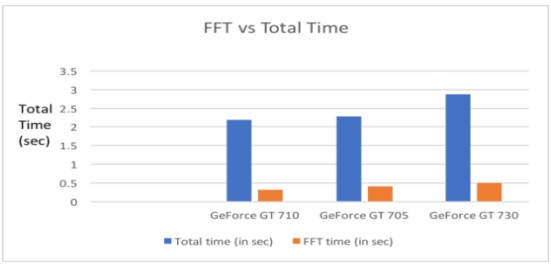


**Figure 4: FFT vs. Total Time**

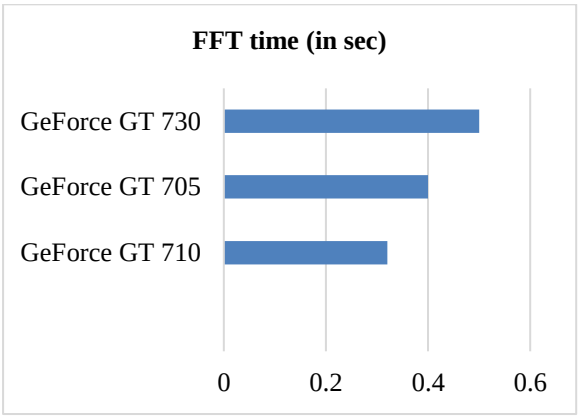


**Figure 5: GPU Performance**


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

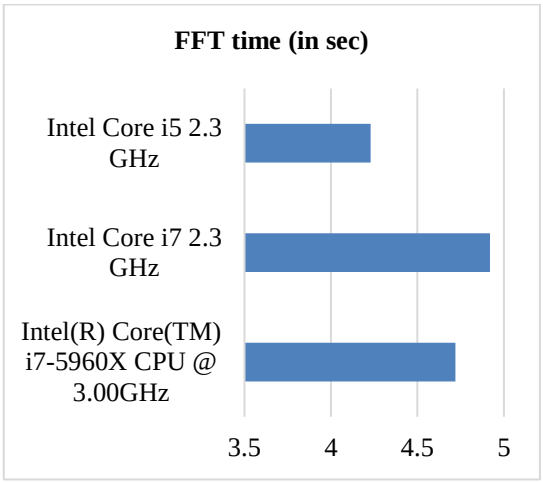


**Figure 6: CPU Performance**

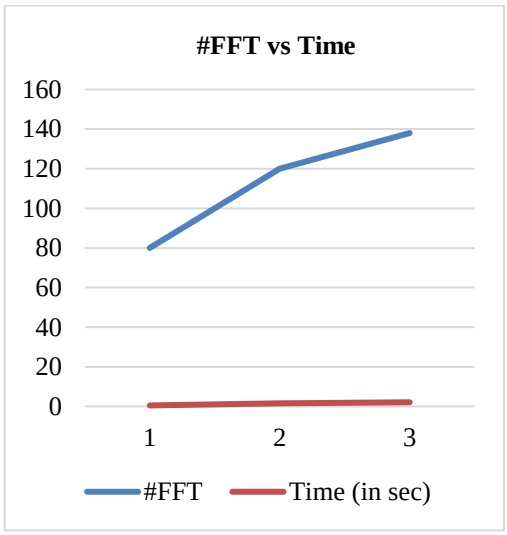


**Figure 7: #FFT vs. Time (in sec)**

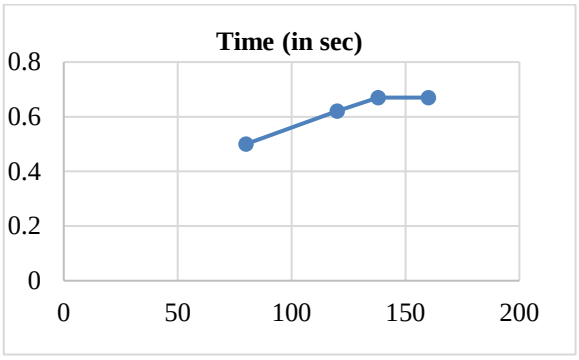


**Figure 8: GPU Cluster Performance**

Figure 5 shows the measurements of the FFT computation time on the afore-mentioned GPUs whereas Figure 6 shows the measurement of FFT computation on CPU. We can observe from Figure 5 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 6). We can perform docking of heavier molecules and test for scalability in future.

*4.2 Results of Spherical Polar Transforms*

We have used NVIDIA GeForce GT 710 and measured the computation time required on the GPU. The computation times are as shown in Table 1.

**Table 1: Computation Time on GeForce GT 710**

| Mol. A | Mol. B | Computation Time (in sec) |
|---|---|---|
| 1AKZ | 1UGI(A) | 42.4 |
| 1BRA | 6PTI | 51.3 |
| 1SUP | 3SSI | 61.1 |
| 3PTN | 6PTI | 67.3 |

We further measured the computation on Intel Core i7 2.3GHz and the computation times are as shown in Table 2.

**Table 2: Computation Time on Intel Core i7 2.3 GHz**

| Mol. A | Mol. B | Computation Time (in sec) |
|---|---|---|
| 1AKZ | 1UGI(A) | 89.2 |
| 1BRA | 6PTI | 103.8 |
| 1SUP | 3SSI | 138.3 |
| 3PTN | 6PTI | 115.7 |

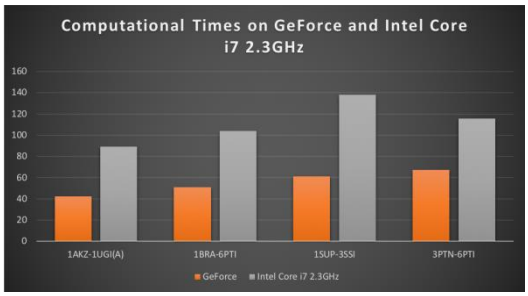


**Figure 9: Computational Times on GPU and CPU**

## V. PERFORMANCE EVALUATION ON BENCHMARK SET

Figure 10 and Table 3 show the total docking calculation time results for the dataset. The algorithm was parallelized previously using OpenMP and it provided good acceleration with multicores, as reported in our previous study.

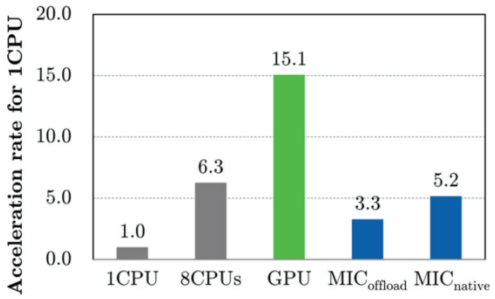


**Figure 10: Acceleration Rate CPU vs. GPU**

**Table 3: Docking Calculation Times for 100 Proteins**

| | 1CPU | 8CPUs | GPU |
|---|---|---|---|
| Total docking time [hour] | 10.2 | 1.5 | 0.72 |

With this dataset, it achieved a 6.3-fold speed up using eight CPU cores. GPU has accelerated the protein docking calculations. Using a GPU, the docking calculations were 15.1 times faster than the calculations with a CPU core alone. With a GPU, the acceleration was more than double that obtained with eight CPU cores, that is, a CPU socket.


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

By contrast, the acceleration rates were increased by 3.3-fold and 5.2-fold with the MIC offload mode and MIC native mode, respectively, which were much lower than the improvements obtained with the GPU.

## VI. CONCLUSION

This study of docking on high performance computing environments shows high scalability. Also, it is found 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. We have further extended out previous work [1] where we did an empirical study on porting the entire FFT pipeline for docking onto the GPU and in the present work we have explored the option of porting the spherical transforms on the HPC platform. It can be noted from (eq. 8) that the computation of the overlap is independent of the scalar fields and can be done using numerical methods and hence it is pre-computed at each step and stored in a look up table for future computational references. This step significantly reduces the computational complexity and hence the computational time has significantly improved.

## REFERENCES


1. K. Abhishek, S.Balaji "Throughput optimization and FFT parallelization for Protein-Ligand docking using GPU Cluster" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-6S, April 2019 pp 324-327 Impact factor: 5.15 Unpaid Scopus Indexed Journal
2. K. Abhishek, S.Balaji "Hybrid Parallelization of Protein-Ligand Docking using Fast Fourier Transformations and Rigid Body conformation",International Journal of Innovative Technology and Exploring Engineering (IJITEE),D1S0032028419/19, Volume-8 Issue-4S2 March, 2019
3. KhushbooBabaria;SanyaAmbegaokar;Shubhankar Das;HemantPalivela, "Algorithms for ligand based virtual screening indrugdiscovery", International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 10.1109/ICATCCT.2015.7457004, 2016
4. P. B. Jayaraj;K. Rahamathulla;G. Gopakumar, "A GPU Based Maximum Common Subgraph Algorithm forDrugDiscovery Applications", IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), 10.1109/IPDPSW.2016.65, 2016
5. AjinkyaNikam;Akshay Nara;Deepak Paliwal;S. M. Walunj, "Acceleration ofdrugdiscoveryprocess on GPU", Green Computing and Internet of Things (ICGCIoT), 10.1109/ICGCIoT.2015.7380432,2015
6. MajidRastegar-Mojarad;RavikumarKomandurElayavilli;DingchengLi;Rashmi Prasad;Hongfang Liu, "A new method for prioritizingdrugrepositioning candidates extracted by literature-baseddiscovery", IEEE transactions on Bioinformatics and Biomedicine (BIBM), 10.1109/BIBM.2015.7359766, 2015
7. ZiyiGuo;Brian Y. Chen, "Predictingprotein-ligandbinding specificity based on ensemble clustering", IEEE transactions on Bioinformatics and Biomedicine (BIBM), 10.1109/BIBM.2015.7359858, 2015
8. Peng Chen;ShanShan Hu;Jun Zhang;XinGao;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
9. 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 toProtein-LigandBinding Sites Prediction", IEEE Transactions on NanoBioscience, 10.1109/TNB.2015.2394328,2015
10. Nishamol P H,Gopakumar G, "Multi-target Drug Discovery Using System Polypharmacology –State of the art",978-1-4799-1823-2, 2015
11. Hossam M. Ashtawy;Nihar R. Mahapatra, "A Comparative Assessment of Predictive Accuracies of Conventional and Machine LearningScoringFunctionsfor Protein-Ligand Binding Affinity Prediction", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2351824,2015
12. 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
13. Daniel Li;Brian Tsui;Charles Xue;Jason H. Haga;Kohei Ichikawa; Susumu Date, "Protein Structure Modeling in a Grid Computing Environment", 1109,2013
14. Ginny Y. Wong;Frank H. F. Leung;S. H. Ling, "PredictingProtein-LigandBinding Site Using Support Vector Machine withProteinProperties", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10.1109/TCBB.2013.126,2013
15. PratyushaRakshit;AmitKonar;ArchanaChowdhury;Eunjin Kim;Atulya K. Nagar, "Muti-objective evolutionary approach ofliganddesign forprotein-liganddocking problem", 2013 IEEE Congress on Evolutionary Computation, CEC.2013.6557576,2013
16. AnkurDhanik,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
17. 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
18. D.W. Richie, "Recent Progress and Future Directions in Protein-Protein Docking",Current Protein and Peptide Science, vol. 9, pp. 1-15, 2008.
19. G. R. Smith and M. J. Sternberg. Prediction of protein-protein interactions by docking methods. CurrOpinStructBiol, 12(1):28–35, 2002.
20. D. W. Ritchie. Recent progress and future directions in protein protein docking. Curr Protein PeptSci, 9(1):1–15, 2008.
21. I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. J MolBiol, 161(2):269–88, 1982.
22. M. L. Connolly. Shape complementarity at the hemoglobin alpha 1 beta 1 subunit interface. Biopolymers, 25(7):1229–47, 1986.
23. R. Norel, S. L. Lin, H. J. Wolfson, and R. Nussinov. Shape complementarity at protein-protein interfaces. Biopolymers,34(7):933–40, 1994.
24. R. Norel, S. L. Lin, H. J. Wolfson, and R. Nussinov. Molecular surface complementarity at protein-protein interfaces: the critical role played by surface normals at well placed, sparse, points in docking. J MolBiol, 252(2):263–73, 1995.
25. R. Norel, D. Petrey, H. J. Wolfson, and R. Nussinov. Examination of shape complementarity in docking of unbound proteins. Proteins, 36(3):307–17, 1999.
26. J. Fernandez-Recio, M. Totrov, and R. Abagyan. Icm-disco dockingby global energy optimization withfully flexible side-chains. Protein,52(1):113–7, 2003.
27. M. Zacharias. Protein-protein docking with a reduced protein modelaccounting for side-chain flexibility. Protein Sci, 12(6):1271–82,2003.
28. J. S. Taylor and R. M. Burnett. Darwin: a program for dockingflexible molecules. Proteins, 41(2):173–91, 2000.
29. E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C.Aflalo, and I. A. Vakser. Molecularsurface recognition:determination of geometric fit between proteins and their ligands bycorrelationtechniques. ProcNatlAcadSci U S A, 89(6):2195–9,1992.
30. H. A. Gabb, R. M. Jackson, and M. J. Sternberg. Modelling proteindocking using shape complementarityelectrostatics and biochemicalinformation. J MolBiol, 272(1):106–20, 1997.
31. R. Chen and Z. Weng. Docking unbound proteins using shapecomplementarity, desolvation, andelectrostatics. Proteins, 47(3):281–94, 2002.



*© Copyright: All rights reserved.*

32. B. S. Duncan and A. J. Olson. Applications of evolutionaryprogramming for the prediction of protein-protein interactions. InLawrence J. Fogel, Peter J. Angeline, and Thomas Baeck, editors,Evolutionaryprogramming V : proceedings of the Fifth AnnualConference on Evolutionary Programming, pages 411–417. MITPress, Cambridge, MA, 1996.
33. D. W. Ritchie and G. J. Kemp. Protein docking using spherical polarfourier correlations. Proteins,39(2):178–94, 2000.
34. D. W. Ritchie, D. Kozakov, and S. Vajda. Accelerating and focusingprotein-protein docking correlations using multi-dimensionalrotational FFT generating functions. Bioinformatics, 24(17):1865–73,2008.
35. K. Sumikoshi, T. Terada, S. Nakamura, and K. Shimizu. A fastproteinprotein docking algorithm using series expansion in terms ofspherical basis functions. Genome Inform, 16(2):161–73, 2005.
36. K. Sumikoshi, T. Terada, S. Nakamura, and K. Shimizu. A fastproteinprotein docking algorithm using series expansion in terms ofspherical basis functions. Genome Inform, 16(2):161–73, 2005.
37. C. H. Choi, J. Ivanic, M. S. Gordon, and K. Ruedenberg. Rapid andstable determination of rotation
38. matrices between spherical harmonicsby direct recursion. Journal of Chemical Physics, 111:8825–8831,1999.
39. J. Ivanic and K. Ruedenberg. Rotation matrices for real sphericalharmonics. direct determination by recursion. J. Phys. Chem.,100(15):6342–6347, 1996.
40. C. Zhang, G. Vasmatzis, J. L. Cornette, and C. DeLisi. Determinationof atomic desolvation energies fromthe structures of crystallizedproteins. J MolBiol, 267(3):707–26, 1997.
41. R. Mendez, R. Leplae, L. De Maria, and S. J. Wodak. Assessment ofblind predictions of protein-protein interactions,2001
42. J. Thompson, D. Higgins, and T. Gibson, "ClustalW: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice", Nucleic Acids Research, Vol. 22, No. 22, 1994, pp. 4673-4680.
43. S. Needleman, and C. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of two Proteins", Journal of Molecular Biology, 1970, vol. 48, pp. 443–453.
44. T. Smith, and M. Waterman, "Identification of Common Molecular Subsequences", Journal Molecularly Biology, 1981, 147, pp.195-197.
45. GenBank, ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/A. Darling, L. Carey, and W. Feng, "The design, implementation,and evaluation of mpiBLAST", In Proceedings of the Cluster World Conference and Expo, in conjunction with the 4th International Conference on Linux Clusters: The HPC Revolution,2003.
46. P. Borovska, V. Gancheva, G. Dimitrov, K. Chintov, S. Gurov, "Parallel Performance Evaluation of Multithreaded Local Sequence Alignment", Proceeding of International Conference on Computer Systems and Technologies, CompSysTech'11, Vienna, Austria, 2011 (under print).
47. P. Borovska, O. Nakov, V. Gancheva, I. Georgiev, "Parallel Multiple Alignment of the Influenza Viru A/H1N1 Genome Sequences on a Heterogeneous Compact Computer Cluster", Proceedings of the 9th WSEAS International Conference on Software Engineering, Parallel and Distributed Systems (SEPADS'10), Cambridge, UK, 2010, pp. 50-55.


*© Copyright: All rights reserved.*