Protein-Protein Docking Using Multi-Dimensional Spherical Basis Functions on High Performance Computing Platform

Abhishek K, S. Balaji

Abstract: ***Docking has become the most important in-silico technique in the process of in-silico drug discovery. The complexes produced because of protein-ligand and the protein-protein interaction are predicted using docking techniques. Hence, it is very important but at the same time it is quite challenging owing to the huge computational costs and the complexity of the computational techniques. In our previous work [1], we had studied the role of the FFTs in increasing the efficiency of Scoring Functions (SF) in heterogeneous parallel processing based virtual screening pipeline for effective rescoring in protein-ligand docking. In this work, we propose multi-layered polar transformation functions to search a multi-dimensional space of a rigid-body model. These functions enhance the efficient use of the spherical co-ordinates to improve the scoring function, thereby improving the overall efficiency of the docking process.***

Keywords: spherical transforms, protein-protein interaction, protein-ligand interaction, multi-dimensional space.

I. INTRODUCTION

Protein-protein interactions are very central to any biological functions, 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. 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 [2], [3], [4] 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 [5], [6], [7], [8], [9] and the techniques based on the energy minimization as discussed in [10], [11], [12] and also the algorithms which performs a global search that leverages Fast Fourier Transforms (FFT) as discussed in [13], [14], [15] 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 aforementioned techniques and algorithms, these algorithms suffer with the traditional tradeoff between the computational complexity (time) and accuracy of the predictions. The works described in [13],[14],[15] 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 of the other approaches to overcome the trade-off as suggested in the works [16], [17], [18] leverage 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.

Revised Manuscript Received on April 15, 2019.

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, Off Kanakapura Road, Bengaluru-560082, India,

Ritchie's work [17], [18] 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 state 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 [19] for *r* which is the distance from origin.

In our present work, we have extended this method by dividing *r* in multiple regions and ported the entire multi-dimensional pipeline onto GPU. It is to be noted that all these regions have a custom radial basis function. To put this into perspective, imagine R^3 space. Now this space will be divided into multiple layers. Each layer has a scalar field which is characterized by the basis functions which in turn is composed of a combination of spherical harmonics and the radial basis function for that particular layer.

The computational pipeline for all these layers will be executed in parallel on an HPC (High Performance Computing) platform. Given that each layer will have its own coefficients which are less in number compared to the entire structure as a whole, the computation will be much efficient and faster.

II. METHODS

A. Scoring Functions

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

Let $f_1(x)$, $f_2(x)$, $f_3(x),\ldots\ldots$, $f_{Ns}(x)$ be the scalar field functions for molecule A and $g_1(x)$, $g_2(x)$, $g_3(x),\ldots\ldots$, $g_{Ns}(x)$ be the scalar field functions for molecule B. The scoring function would be formulated as follows:

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

where w_i represents the weight of the *ith* term,

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

$f^T(x)$ is the field generated by applying the operation T to x.

B. Basis Functions

By representing the scalar fields in terms of orthogonal basis functions, the score computation becomes much faster. The basis function for R^3 can be expressed

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

where,

$Y_{l,m}(\theta,\varphi)$ – 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 I_k of widths 'a' i.e.

$I_k \equiv$ [ka,(k+1)a) for k=0,1,…, then

$S_{k,n}(r)$ for each region can be defined as

$S_{k,n}(r) = 0$ if r∉[ka,(k+1)a)

and

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

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

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

$h_{k,n}(x)$ – 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.

$N_{k,n}$ – norm of the polynomial function. It now becomes easy to realize that the orthonormality

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

The orthonormality of the combined functions can be similarly expressed as

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

There can exist some regions where the radial basis function can have non-zero values which are as shown in Fig. 1.

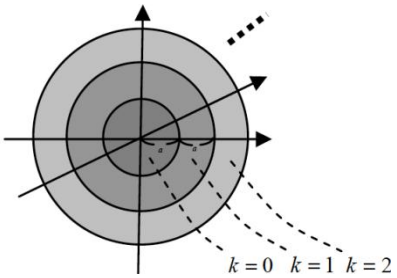

Fig. 1: Regions of Radial Basis Functions

C. Fast Rotational and Translational Operators.

In order to achieve better performance, we need to obtain the following transformed coefficients $a_{k,n,l,m}^{T^A}$ and $b_{k,n,l,m}^{T^B}$ 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.

1. 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 $a_{k,n,l,m}^R$ be the new rotational coefficients. As previously mentioned, we can derive the rotational coefficients using the original coefficients as follows:

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

2. Translational Operation on coefficients

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

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

$$a_{k,n,l,m}^{S_{\Delta Z}} = \sum_{k'n'l'} a_{k',n',l',m'} \int_0^{\infty} \int_0^{\pi} S_{k',n'}(r') S_{k,n}(r) P_{l'}^{|m|}(cos\theta') P_l^{|m|}(cos\theta) r^2 sin\theta \, d\theta \, dr$$

$\equiv \sum_{k'n'l'} a_{k',n',l',m} O_{k',k,n',n,l',l,|m|}(\Delta z)$ ---------- (eq 8)

$P_l^m(x)$ is the Legendre Polynomial

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

It is very important to note here that the overlap $O_{k',k,n',n,l',l,|m|}(\Delta z)$ 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.

III. RESULTS

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

Table I. 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 II. 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 2 below indicates the comparison of the computational times.

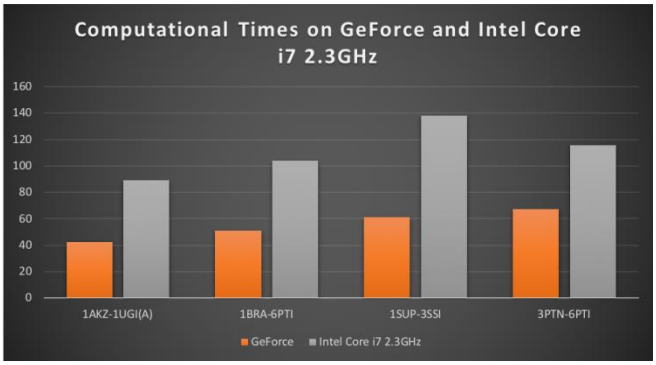

Figure 2 Computational Times on GPU and CPU

IV. CONCLUSION

We have 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. D.W. Richie, "Recent Progress and Future Directions in Protein-Protein Docking", *Current Protein and Peptide Science*, vol. 9, pp. 1-15, 2008.
2. G. R. Smith and M. J. Sternberg. Prediction of protein-protein interactions by docking methods. Curr Opin Struct Biol, 12(1):28–35, 2002.
3. D. W. Ritchie. Recent progress and future directions in protein protein docking. Curr Protein Pept Sci, 9(1):1–15, 2008.
4. I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. J Mol Biol, 161(2):269–88, 1982.
5. M. L. Connolly. Shape complementarity at the hemoglobin alpha 1 beta 1 subunit interface. Biopolymers, 25(7):1229–47, 1986.
6. R. Norel, S. L. Lin, H. J. Wolfson, and R. Nussinov. Shape complementarity at protein-protein interfaces. Biopolymers,34(7):933–40, 1994.
7. 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 Mol Biol, 252(2):263–73, 1995.
8. 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.
9. J. Fernandez-Recio, M. Totrov, and R. Abagyan. Icm-disco docking by global energy optimization with fully flexible side-chains. Protein, 52(1):113–7, 2003.
10. M. Zacharias. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. Protein Sci, 12(6):1271–82, 2003.
11. J. S. Taylor and R. M. Burnett. Darwin: a program for docking flexible molecules. Proteins, 41(2):173–91, 2000.
12. E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and I. A. Vakser. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A, 89(6):2195–9, 1992.
13. H. A. Gabb, R. M. Jackson, and M. J. Sternberg. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol, 272(1):106–20, 1997.
14. R. Chen and Z. Weng. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. Proteins, 47(3):281– 94, 2002.
15. B. S. Duncan and A. J. Olson. Applications of evolutionary programming for the prediction of protein-protein interactions. In Lawrence J. Fogel, Peter J. Angeline, and Thomas Baeck, editors, Evolutionary programming V : proceedings of the Fifth Annual Conference on Evolutionary Programming, pages 411–417. MIT Press, Cambridge, MA, 1996.
16. D. W. Ritchie and G. J. Kemp. Protein docking using spherical polar fourier correlations. Proteins, 39(2):178–94, 2000.

17. D. W. Ritchie, D. Kozakov, and S. Vajda. Accelerating and focusing protein-protein docking correlations using multi-dimensional rotational FFT generating functions. Bioinformatics, 24(17):1865–73, 2008.
18. K. Sumikoshi, T. Terada, S. Nakamura, and K. Shimizu. A fast proteinprotein docking algorithm using series expansion in terms of spherical basis functions. Genome Inform, 16(2):161–73, 2005.
19. K. Sumikoshi, T. Terada, S. Nakamura, and K. Shimizu. A fast protein protein docking algorithm using series expansion in terms of spherical basis functions. Genome Inform, 16(2):161–73, 2005.
20. C. H. Choi, J. Ivanic, M. S. Gordon, and K. Ruedenberg. Rapid and stable determination of rotation matrices between spherical harmonics by direct recursion. Journal of Chemical Physics, 111:8825–8831, 1999.
21. J. Ivanic and K. Ruedenberg. Rotation matrices for real spherical harmonics. direct determination by recursion. J. Phys. Chem., 100(15):6342–6347, 1996.
22. C. Zhang, G. Vasmatzis, J. L. Cornette, and C. DeLisi. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol, 267(3):707–26, 1997.
23. R. Mendez, R. Leplae, M. F. Lensink, and S. J. Wodak. Assessment of capri predictions in rounds 3-5 shows progress in docking procedures. Proteins, 60(2):150–69, 2005.
24. R. Mendez, R. Leplae, L. De Maria, and S. J. Wodak. Assessment of blind predictions of protein-protein interactions