Efficient Local ligand-binding site search using Landmark MDS

Sungchul Kim
POSTECH
subright@postech.ac.kr

Lee Sael
SUNY Korea & Stony Brook University
sael@sunykorea.ac.kr

Hwanjo Yu
POSTECH
hwanjoyu@postech.ac.kr

ABSTRACT

In this work, we propose a new local binding site search system, called Fast Patch-Surfer, for extending previous work, Patch-Surfer. Patch-Surfer efficiently retrieves top-k similar proteins based on new representation of proteins capturing features of their local ligand-binding site and newly defined distance function. However, further speed up is needed since in practical setting of computing dissimilarity between proteins, there are possibilities for simultaneous multiple user access on the database. We address this need for further speed up in local ligand-binding site search by exploiting landmark MultiDimensional Scaling (MDS), which is an efficient version of MDS being popularly used for representing high-dimensional dataset. According to the result, using our method, the searching time is reduced up to 99%, and it retrieves almost 80% of exact top-k results.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval—*Content Analysis and Indexing, Indexing methods*; H.3.1 [**Information Systems**]: Information Search and Retrieval—*Information Search and Retrieval*

General Terms

Algorithms, Theory

Keywords

structure-based function prediction; protein surface shape; ligand binding pocket; 3D Zernike descriptor; pocket comparison

1. INTRODUCTION

Due to the increasing number of protein structures of unknown function, computational methods for characterizing protein tertiary structures have been studied. Using typical sequence database searches, it is hard to predict the functions of unknown protein structure [4, 3]. As an alternative approach, structural information of proteins is used as a legitimate analysis strategy [2, 6]. One method of characterizing proteins with structural information is through prediction of ligand binding to a protein, which is a major task of molecular function of proteins. However, the complex nature of protein ligand interactions makes it difficult to predict whether a ligand molecule binds a protein or not.

In previous work, Sael et al. have observed geometric and physicochemical complementarity between the ligand and its binding site in multiple cases. Accordingly, they proposed a method finding ligand molecules which bind to a local surface site of a protein by finding similar local pockets of known binding ligands in the protein structure database.

Sael et al. also suggested a local surface comparison method for more accurate prediction whether a ligand molecule binds to a query protein, called Patch-Surfer [5]. It represents a binding pocket as a combination of segmented surface patches, each of which is characterized by four representative features: 1) geometrical shape, 2) the electrostatic potential, 3) the hydrophobicity, and 4) the concaveness. The shape and the physicochemical properties of surface patches are represented using the 3D Zernike Descriptors (3DZDs). To compare two pockets, patches of given pockets are matched by a modified weighted bipartite matching algorithm and it proposes a similarity function which computes similarity based on the four characteristics. However, computing the similarity of pockets is complex, thus finding the local ligand-binding sites based on the similarity function takes nontrivial amount of time. In addition, there are possibilities for simultaneous multiple user access on the database.

To resolve this problem, we propose an efficient local ligand-binding site search algorithm, called Fast Patch-Surfer by using Landmark MultiDimensional Scaling (LMDS), which is an efficient version of MDS being popularly used for representing high-dimensional dataset. According to the experiments, Fast Patch-Surfer retrieves top-k similar pockets up to 99% faster than the previous method.

2. METHODS

We first provide a brief introduction of Patch-Surfer, which searches a database of known pockets and finds similar ones to the query. Then, the descriptions of Landmark MDS and our algorithm, Fast Patch-Surfer, for more efficient top-k retrieval is provided.

Patch-Surfer method: Patch-Surfer is an alignment free local surface comparison method for predicting a ligand molecule which binds to a query protein. Given a query

DTMBIO'13, November 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2512089.2512092 ...$15.00.

protein structure, a pocket region is extracted based on the computed query protein surface. If the binding pocket of the protein is unknown, binding pocket prediction can be used. The pocket is divided into surface patches represented as four features: geometrical shape, the surface electrostatic potential, the hydrophobicity, and the concaveness. For representation, we exploit 3D-Zernike Descriptor (3DZD). Thus, each pocket is represented as a set of surface patches.

The Patch-Surfer retrieves top-k similar proteins in terms of the characteristics of local binding-sites which are very informative and meaningful. The problem is that a protein is represented a set of 3DZDs (or vectors), not 'a' vector. In addition, computing similarity of even one pair of proteins is quite complex, and the number of proteins in database continues to grow.

Landmark MDS: Metric-preserving dimensionality reduction has been an important task in data analysis and machine learning. Given proximity data which consists of dissimilarity information for all pairs of objects. Multidimensional scaling (MDS) embeds the objects as points in a low-dimensional Euclidean space, while preserving the geometry as precise as possible. Landmark MDS (LMDS) preserves all properties of classical MDS and allows efficient computation as well. Based on a dissimilarity matrix D of l data points, the goal is to embed them in m-dimensional Euclidean space. Specifically, Landmark MDS works in four steps:

1 Select l landmark points.

2 Apply classical MDS to find an embedding of the l landmark points in $\mathbb{R}^m$.

3 Compute the embedding coordinates of the remaining points based on distances to the embedded landmark points.

4 Apply PCA normalization.

Note that the computational complexity of LMDS is $O(nlk+l^3)$ which is much more efficient than classical MDS.

Algorithm 1: Fast Patch-Surfer

Data: A query protein q, Dissimilarity matrix, D
Result: Top-k result, X_k

1 $X_k = \phi$
2 Select l landmark points, L
3 Apply classical MDS and obtain the embedding coordinates of landmark points, Y_l
4 Compute the embedding coordinates of the remaining points, Y
5 Compute the embedding coordinates of the query point, y_q
6 $X_k = GetTopK(y_q, Y_l)$
7 **return** X_k

Fast Patch-Surfer: Based on Patch-Surfer algorithm, we suggest an efficient algorithm, called Fast Patch-Surfer, by exploiting LMDS. Given a query protein, q, our algorithm firstly proceeds LMDS to obtain embedding coordinations of landmark points and other remaining points, $L_x = \{y_1, y_2, \cdots, y_n\}$ where some of them are landmark points. To select landmark points (Line 1), we take two different approaches. After that, embedding coordinates of q is computed. Using the embedding coordinates of and the query, we can find top-k similar proteins by computing Euclidean distance from all candidate points in database (GetTopK() at Line 5). The detailed algorithm is presented in Algorithm 1.

Although this approach still visits every data point (or it cannot reduce the evaluation ratio), the computation of the Euclidean distance between the embedding coordinates is much more efficient than computing similarity between pockets using the way in Patch-Surfer.

According to the result (Table 1), Fast Patch-Surfer reduces the processing time of the Patch-Surfer up to 99%. In addition, the standard deviation of Fast Patch-Surfer is much smaller than that of the Patch-Surfer as well. In accuracy, our method retrieves almost 80% of exact top-k results.

Table 1: The result of Landmark MDS (sec.)

	mean	std.
Patch-Surfer	43.8196	25.4198
Fast Patch-Surfer ($l = 10$)	0.613	0.3571

3. CONCLUSION

In this work, we propose a new local binding site search system for previous work, called Fast Patch-Surfer. We exploit Landmark MultiDimensional Scaling (LMDS), which is an efficient version of MDS being popularly used for representing high-dimensional dataset. We take two different approaches for the selection of landmark points: 1) random selection and 2) greedy selection. According to the result, using our method, the searching time is reduced upto 99%, and it retrieves almost 80% of exact top-k results.

4. ACKNOWLEDGMENTS

This work was partially supported by Mid-career Researcher Program through NRF grant funded by the MEST (No. NRF-2011-0016029). This research was partially supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (No. 2012M3C4A7033344).

5. REFERENCES

[1] F. Abascal and A. Valencia. Automatic annotation of protein function based on family identification. *Proteins*, 53:683–692, 2003.

[2] J.-F. Gibrat, T. Madej, and S. H. Bryant. Surprising similarities in structure comparison. *Curr. Opi. Struct. Biol.*, pages 377–385, 1996.

[3] T. Hawkins, M. Chitale, and D. Kihara. New paradigm in protein function prediction for large scale omics analysis. *Mol Biosyst*, 4(3):223–31, 2008.

[4] D. Lee, O. Redfern, and C. Orengo. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol*, 8(12):995–1005, 2007.

[5] L. Sael and D. Kihara. Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison. *Proteins*, 2012.

[6] I. N. Shindyalov and P. E. Bourne. Protein structure alighment by incremental combinatorial extension (ce) of the optimal path. *Protein Eng.*, pages 739–747, 1997.