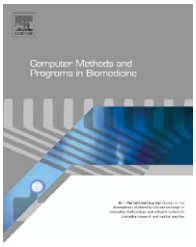
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 108 (2012) 168–175

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Circular Cone: A novel approach for protein ligand shape matching using modified PCA

Shuangjian Zhang[a,*], *Jun Du*[a], *Liang Zhang*[b], *Cheng Zeng*[c], *Qiao Liu*[a], *Tao Zhang*[c], *Gang Hu*[a]

[a] *School of Mathematical Sciences, Nankai University, 300071 Tianjin, PR China*
[b] *College of Software, Nankai University, 300071 Tianjin, PR China*
[c] *College of Life Sciences, Nankai University, 300071 Tianjin, PR China*

ARTICLE INFO

Article history:
Received 22 November 2011
Received in revised form
24 February 2012
Accepted 28 February 2012

Keywords:
Shape matching
Pocket
Ligand
New pharmaceuticals
Virtual screen
Circular Cone

ABSTRACT

Nowadays in modern medicine, computer modeling has already become one of key methods toward the discovery of new pharmaceuticals. And virtual screening is a necessary process for this discovery. In the procedure of virtual screening, shape matching is the first step to select ligands for binding protein. In the era of HTS (high throughput screening), a fast algorithm with good result is in demand. Many methods have been discovered to fulfill the requirement. Our method, called "*Circular Cone*", by finding principal axis, gives another way toward this problem. We use modified PCA (principal component analysis) to get the principal axis, around which the rotation is like whirling a cone. By using this method, the speed of giving score to a pocket and a ligand is very fast, while the accuracy is ordinary. So, the good speed and the general accuracy of our method present a good choice for HTS.

1. Introduction

It is well-known that virus resistance to first-line antibiotics poses a serious threat to public health. For this reason many pharmaceutical and biotechnology institutes are aggressively pursuing novel ways to kill or inhibit viruses. Toward this end, drug design is to develop a small molecular antibiotic against resistant strains. Recent estimates suggest that it takes up to 13.5 years and 1.8 billion U.S. dollars to bring a new drug to the market [1]. The involvement of genomics [2], proteomics [3], bioinformatics [4] and efficient technologies like combinatorial chemistry [5], high throughput screening (HTS) [6], virtual screening, ADMET screening [7] and structure-based [8] drug design serves to expedite as well as economize the modern day drug discovery process. In the past few decades, as drug design is becoming increasingly more significant, its first and fundamental step, shape matching, also called shape complementarity, is also researched hotly as a matter of course. Shape matching is essential in predicting which molecules can bind to a given binding site of a protein with known 3D structure, which is important to decipher the protein function and is useful in drug design.

Two basic steps of shape matching are the representation of the system and the calculation of their similarity. To the first step, the choice of descriptor is often domain-specific and could vary largely from one application to another, facing

* *Corresponding author*. Tel.: +86 155 220 06763.
E-mail addresses: kelvin.sj.z@gmail.com (S. Zhang), dujun166@mail.nankai.edu.cn (J. Du), 542673977@qq.com (L. Zhang), zengcheng@mail.nankai.edu.cn (C. Zeng), liuqiao051@mail.nankai.edu.cn (Q. Liu), zhangtao@nankai.edu.cn (T. Zhang), huggs.hg@gmail.com (G. Hu).
0169-2607/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2012.02.011

the dilemma of either being too coarse (ignoring information) or too complex (redundant and unstable). As the problem of matching rigid closed shapes is generalized into partial or articulated shapes, developing a shape representation for matching and recognition becomes significant. Usually shapes have been presented as curves [9], medial axes [10], shock structures [11], sampled points [12] and so on. (i) For the representation about curves [13], one characteristic of most existing curvature-measurement techniques is the assumption that there is a unique curvature that can be measured at each point. Mokhtarian and Mackworth have introduced the curvature scale-space image as a tool for representing planar curves [14]. This representation is computed by convolving a path-based parametric representation of the curve with a Gaussian function with varying standard deviation. (ii) For the axis-based representations, the idea of decomposing a shape into primitives and building up its description in a frame that expresses the links between these primitives was first made explicit by Marr and Nishihara [15] and has been one of most promising guidelines for recognition. Shaked and Bruckstein had used "pruning medial axes" to define the representation of shape matching [16]. (iii) For the shock-based representation, it is a representation derived from viewing the medial axis as singularities formed during propagation from boundaries, for example, Blum's grassfire is the shock tree or shock graph [17]. (iv) As for points-based representation, matching is typically done using an assignment algorithm [18] when a shape is represented using a point set [19]. These methods have the advantages of not requiring an ordered boundary point. But if the similarity of two points is based on a local measure, the matching process does not necessarily capture the coherence of shapes in that the relationship among portions of shape may not be fully captured in the match.

With regard to the second step, the calculation of their similarity, current shape matching algorithms can be divided into two categories: global and local approaches. Global methods compared the shape of the input objects by defining a global matching cost and optimization algorithm for finding the lowest cost. One of the most popular methods for global shape matching is the shape context proposed by Belongie et al. [12]. Their algorithm uses randomly sampled points as shape representation and is based on a robust shape descriptor – the shape context – which allows formulating the matching step as a correspondence problem. The shape context is the basis for different extensions considering geodesic distances as proposed by Ling and Jacobs [20] or the point ordering as shown by Scott and Nowak [21]. However while such global matching methods work well on most of the standard shape retrieval datasets, it is difficult to handle part deformations, strong articulation or occlusions, for instance, occlusions may lead to matching errors for the context based COPAP framework [21]. These problems are handled well by purely local matching methods as, e.g. proposed by Chen et al. [22], which accurately measure local similarity, but in contrast fail to provide a strong global description for robust shape alignment. In addition, local matching is incoherently more complex than global matching, since global invariants could save a lot of computations for global matching.

In the pharmaceutical development process there are plenty of data to process, though slow methods would make this process excessively long. Thus, speed is a significant factor for shape matching. Many mainstream methods try to improve the speed of matching. In this article, our main contribution is to improve the speed with quality comparable to current methods. We choose the classical method distributed molecular surface (DMS) [23] to signify the surface. Particularly, we only present the pockets of the protein because ligands only combine with the pockets. Therefore we could use fewer points on the surface, which will reduce the complexity of our later calculation. In addition, these critical points will be used in modified principal component analysis (modified PCA) to find principal axes, which enhances the speed of our method. Based on the representation we have chosen, we calculate the complementarity by local shape matching, but it is not a complete local matching method. Firstly we find two principal axes, and then we naturally put two axes together to compare, with some rotations to figure out whether the two shapes match well. Before comparison, some points of these two shapes should be selected, considering the number of all the points is considerable, which gives inconvenience to the calculation and slows down the speed. What's more, not all of these points are useful. So we select some points according to the method given later. And the method of comparison is also given in Section 2. Here we choose some interesting sampled points (called *right points*) of the surface, and compare the contribution of the distances between *right points* and the special point (called *end point*), the intersection of principal axes and the surface. Finally, the most significant is that, our method is proven to have greatly enhanced the speed and insured the accuracy at the same time by the convincing results of enough tests.

2. Method

2.1. Dataset

We use the Kahraman dataset, proposed by [24], which consists of 100 protein crystal structures in complex with one of ten ligands (AMP, ATP, PO4, GLC, FAD, HEM, FMN, EST, AND, NAD). The result will be discussed after Section 2.

We get the representation of the surfaces, and then calculate their principal axes and superimpose the principal axes together. Secondly, we analyze the similarity of the two surfaces to estimate whether the protein and the small molecule can bind to each other and give scores. The whole process is shown in Fig. 1. The details are shown as follows.

2.2. Get surface of pocket and ligand

Here we use DMS to get the surface of pocket and ligand. DMS is an open source program for calculating the molecular surface, which is defined by Richards [25]. The data we get are the coordinates of points representing the surfaces.

These points, which represent the pocket and ligand, are of great importance in that the following processes are all based on these data. Fig. 2(a) shows these points representing the surface of a ligand.

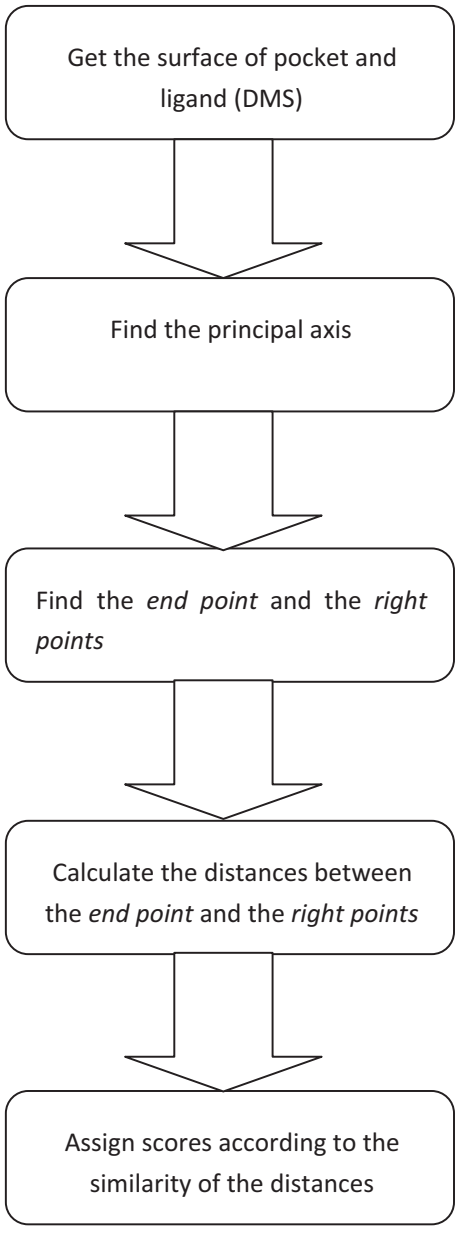


Fig. 1 – The process of the "Circular Cone" method.

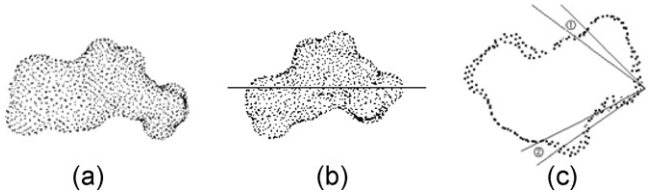


Fig. 2 – Results of each step. (a) The points given by DMS, representing the surface of a ligand. (b) The principal axis of a ligand given by modified PCA. (c) The rays emitted from the *end point* to get the *right points*. ①, ② show the selected points (*right points*) in the interlayer (sectional drawing).

2.3. *Find the principal axis*

With the points representing the surfaces of the pocket and the ligand, we would like to estimate the similarity between them. Their principal axes will coincide with each other if they have a high similarity. At first, we find their principal axes by principal component analysis (PCA) [26]. It is now mostly used as a tool in exploratory data analysis and for making predictive models.

However, the transformations given by PCA include an act of normalizing the data. After these transformations, the scale of the surface has been changed, which would change the shape of the original model. We give back the normalization after the principal axes have been found by taking the inverse matrix of each transformation. Fig. 2(b) shows the principal axis of a ligand given by modified PCA.

2.4. *Find the end point and right points*

After we get the principal axis, we select important points by the following steps. We take one of each principal axis, let the *end point* be the origin, send out two rays, respectively, along 30° and 35° with the axis, then make a rotation by 360°, thus we have two circular conical surfaces and both of them have intersections with the surface of the pocket, we take the interlayer between them, then we get the *right points* from the interlayer and analyze them. Some results are shown in Fig. 2(c). Then, we change 30 to 45° and 60° to ensure the accuracy rating of the result of our arithmetic. Considering that the angle we send out cannot be too small or too large, we choose the angle between 30° and 60°, after repeated experiments, we choose 30°, 45° and 60° to describe our model.

2.5. *Calculate the distances between the end point and the right points*

For the pocket, we divide the conical interlayer equally into 36 parts, 10° each. For each part, we calculate the arithmetic mean value of Euclidean distances from the *right points* of each part to the *end point*, and call the value *pvalue*. And the same goes for the ligand. Then we get the two sets of *pvalues*, each has 108 *pvalues* (each 36 *pvalues* of 30°, 45° and 60° conical interlayers). If the pocket and the ligand have a high similarity, the distances from the *right points* and the *end point* of the pocket and the ligand will also have a high similarity. So, we can use the similarity of distances to describe the similarity of pocket and ligand.

2.6. *Assign scores according to the similarity of the distances*

In this step, we compare the two sets of data representing the distances that we get from the former step. We calculate root-mean-square deviation (RMSD) [27] of the two sets with the rotation to get 36 values. Then we take the minimum values to be the score of the pocket and the ligand, which represents the similarity of the two geometric surfaces. Note that the lower score represents the higher similarity.

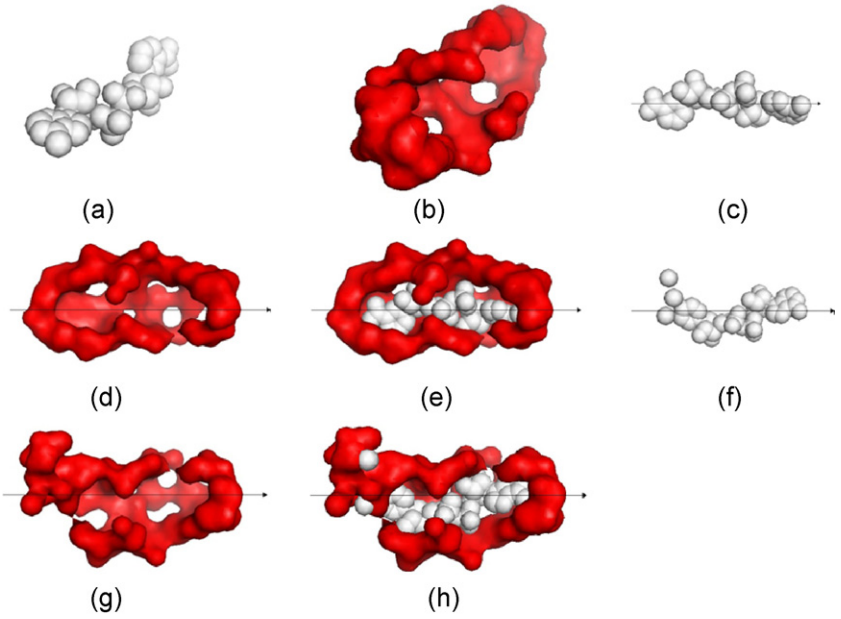


Fig. 3 – The structure of 1ej2_NAD with some results of PCA and modified PCA. (a) The structure of the ligand of 1ej2_NAD. (b) The structure of the pocket of 1ej2_NAD. (c) The principal axis of the ligand given by modified PCA. (d) The principal axis of the pocket given by modified PCA. (e) The combination of the ligand (c) and pocket (d) while the axes coincide. (f) The principal axis of the ligand given by PCA. (g) The principal axis of the pocket given by PCA. (h) The combination of the ligand (f) and pocket (g) while the axes coincide. These figures were generated by PyMOL [29].

3. Results

After doing a variety of experiments on the Kahraman dataset, we acquire a sequence of results to analyze advantages and disadvantages of our method. They indicate the accuracy, adaptability and speed of our approach, as well as shortcomings. And we draw some case analyses and comparison as follows to illustrate the results.

3.1. *Case analysis*

We choose 1ej2_NAD as the first example. 1ej2_NAD is the crystal structure of methanobacterium thermoautotrophicum nicotinamide mononucleotide adenylyltransferase with bound NAD^+ [28]. The ligand and pocket of it is shown in Fig. 3(a) and (b), respectively.

Fig. 3(c) shows the principal axis of the ligand, while Fig. 3(d) shows the principal axis of the pocket, and they both indicate that we have found the principal axes precisely. Fig. 3(e) is the matching figure of them, in which their principal axes coincide with each other. Apparently, they match pretty well.

On the other hand, Fig. 3(f) and (g), respectively, shows each principal axis of the ligand and the pocket, they are both distortedly represented by PCA. Fig. 3(h) shows the matching figure of them like Fig. 3(e) does, but in Fig. 3(h), there are some atoms of the ligand passing through the interior part of the pocket, which contradicts the facts. Thus we discover that it is not proper for PCA to present the axes of these structures because they change the shape of the original structures. This leads to a failing matching when their principal axes are overlapped. However, our modified PCA overcomes this deficiency, which basically ensures the accuracy of our later calculation.

Generally, our modified PCA has quite a wide adaptability, including some semicircular ligands which were considered inadaptable for PCA. For instance, the ligand of 1s7g_NAD is semicircular but it matches well with the pocket by using modified PCA. 1s7g_NAD is the structural basis for the mechanism and regulation of Sir2 enzymes [30]. The results are shown in Fig. 4.

There are also some cases which match well but have lower scores. For example, the ligand PO4, shown in Fig. 5, has only five atoms, whose surface has large differences with the pockets. As a result, it has low scores in most situations. Another type of these cases happens mainly because they match well in local but differ much in a global situation. Therefore, it is not accurate to find the principal axes to compare their complementarity, and that is what we should take into account in future research.

3.2. *Comparison with another method*

We compare our method with ShaEP [31] on the Kahraman dataset. ShaEP, a method for rigid-body superimposition and similarity evaluation of ligand-sized molecules, is capable of identifying a substantial number of active compounds in a database of druglike molecules [31]. Additionally, ShaEP overlays drug-sized molecules on a subsecond timescale, allowing for the screening of large virtual libraries [31]. These two advantages of ShaEP are also our goals. Thus we adopt this

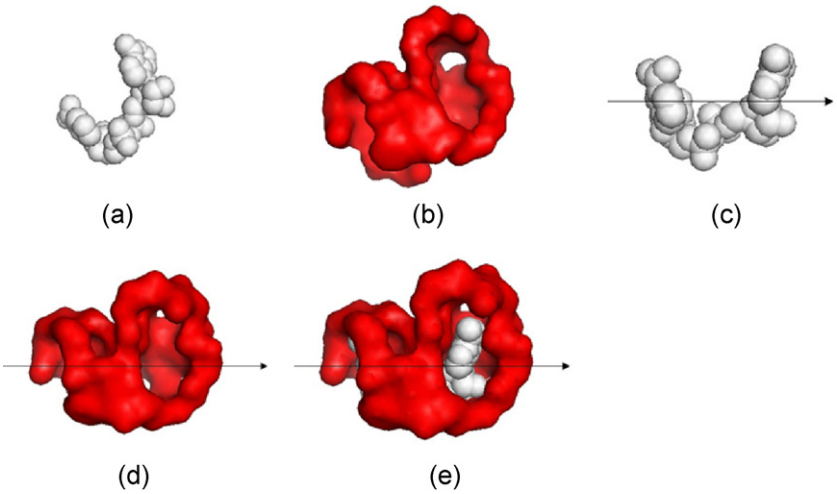


Fig. 4 – The structure of 1s7g_NAD with some results of modified PCA. (a) The structure of the ligand of 1s7g_NAD. (b) The structure of the pocket of 1s7g_NAD. (c) The ligand of 1s7g_NAD with the principal axis in black. (d) The pocket of 1s7g_NAD with the principal axis in black. (e) The combination of the pocket (c) and the ligand (d) of 1s7g_NAD while their principal axes coincide. These figures were generated by PyMOL.

comparison. The performance of both methods is evaluated on the basis of AUC score and speed, which is measured by the average processing time per structure comparison.

AUC score is computed as follows. Consider a set of pockets (P_1, …, P_N) and a similarity measure S. To each pocket P_i, we rank all the pockets according to their similarities to P_i with descending order, using method S. Then we draw the ROC curve for this pocket, with points on XY-plane whose X-axes are the number of pockets binding the different ligand among the top n pockets and Y-axes are the number of pockets binding the same ligand among the top n pockets, when n varies from 0 to N. We align the points next to each other to make up the ROC curve. Divide the area under the ROC curve by the product of the number of pockets binding the different ligand among these N pockets and the number of pockets binding the same one among these N pockets, and we get the AUC score of this pocket. Thus AUC score ranges from 0 to 1.0, and a better method would have a higher AUC score. For example, to each pocket P_i, an "ideal" method will rank all pockets binding the same ligand as P_i on the top of the list, leading to an AUC score equal to 1.0. As a whole, the quality of method S is measured by the AUC scores of these pockets.

In the Kahraman database, there are 100 pockets in total. Each pocket is proposed to compare with all the 100 pockets in the dataset. Then 100 scores are obtained, which present the similarity of pockets. By these 100 scores, we get the AUC score of each complex, which presents the accuracy of our method on this particular pocket. On the Kahraman dataset, we get 100 ROC curves and 100 AUC scores. We also get 100 ROC curves and 100 AUC scores of ShaEP. Fig. 6 shows the ROC curves of FAD. Tables 1–3 show the comparison between ShaEP and our method *Circular Cone* with AUC score.

Meanwhile, we run the program of both methods on the same computer (a 2.00 GHz Pentium(R) Dual-Core CPU, T4200, RAM 2.00 GB, running Windows 7), and calculate the average processing time per structure of both. Table 4 shows the comparison between ShaEP and *Circular Cone* with average processing time per structure.

In Table 1, AUC scores in bold present the portion on which our method is better than ShaEP, in italics present the portion on which ShaEP is better than ours, the others present the portion on which the two methods behave almost the same (the difference in AUC scores is less than 0.05). We could find that *Circular Cone* works better than ShaEP on such ligands as AMP, ATP and AND, with the average AUC scores of *Circular Cone* 0.10 higher than that of ShaEP, while ShaEP does better than *Circular Cone* on FMN.

From Table 2, when t is equal to 0.05, in 45 cases, the AUC scores of *Circular Cone* are better than those of ShaEP; and in 15 cases, the AUC scores of ShaEP are better than *Circular Cone*; in

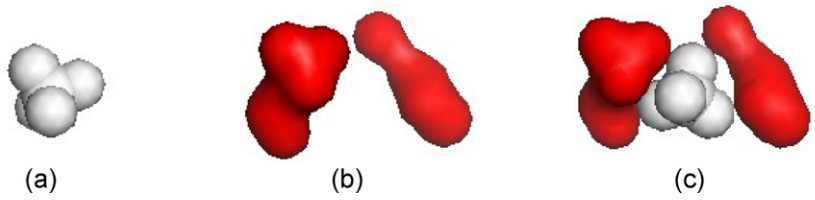


Fig. 5 – The structure of 1a6q_PO4. (a) The structure of the ligand of 1a6q_PO4. (b) The structure of the pocket of 1a6q_PO4. (c) The binding structure of the ligand (a) and the pocket (b). These figures were generated by PyMOL.

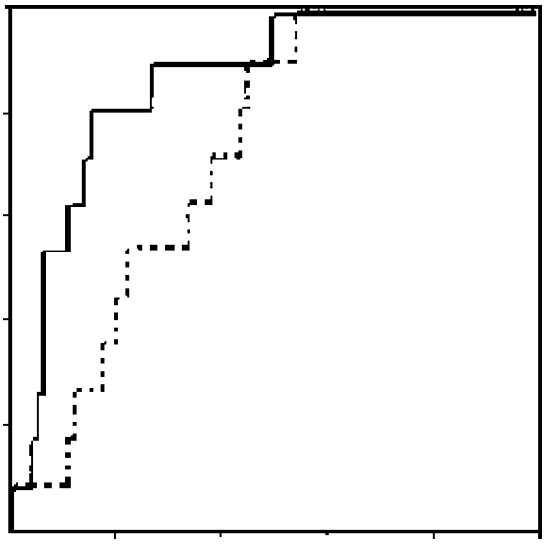


Fig. 6 – The ROC curves of FAD (the solid line above is ROC curve of our method *Circular Cone* and the other one is that of ShaEP).

Table 1 – The comparison of two methods with AUC score (ligands).

I	II	III	IV	V	VI	VII	VIII
AMP	9	**0.78**	**0.87**	**0.54**	**0.75**	**0.68**	**0.81**
ATP	14	**0.71**	**0.81**	**0.55**	**0.69**	**0.62**	**0.77**
FAD	10	**0.74**	**0.89**	0.54	0.49	**0.64**	**0.72**
FMN	6	0.82	0.58	0.58	0.48	0.65	0.53
GLC	5	0.84	0.81	**0.72**	**0.77**	0.78	0.78
HEM	16	**0.77**	**0.83**	0.51	0.51	**0.67**	**0.73**
NAD	15	**0.60**	**0.67**	0.30	0.23	0.49	0.52
PO4	20	1.00	0.99	0.92	0.91	0.97	0.96
AND	2	**0.77**	**0.92**	**0.61**	**0.90**	**0.69**	**0.91**
EST	3	0.91	0.94	**0.59**	**0.78**	**0.77**	**0.85**

(I) Ligands in the Kahraman dataset. (II) The number of complexes for each ligand. (III) The highest AUC score of ShaEP on each ligand. (IV) The highest AUC score of *Circular Cone* on each ligand. (V) The lowest AUC score of ShaEP on each ligand. (VI) The lowest AUC score of *Circular Cone* on each ligand. (VII) The average of AUC scores of each ligand of ShaEP. (VIII) The average of AUC scores of each ligand of *Circular Cone*.

Table 2 – The comparison of two methods with the difference of AUC score (complexes).

t	Diff[a] ≤ −t	−t < Diff < t	Diff ≥ t
0	37	0	63
0.01	31	8	61
0.02	28	15	57
0.03	24	22	54
0.04	20	32	48
0.05	15	40	45
0.06	13	46	41
0.07	10	53	37
0.08	7	57	36
0.09	7	60	33
0.10	6	61	33

[a] Diff donates the result of AUC score of *Circular Cone* subtracts that of ShaEP.

Table 3 – The comparison of two methods with the difference of AUC score (ligands).

I	II	Diff[a] ≤ −0.05	−0.05 < Diff < 0.05	Diff ≥ 0.05
AMP	9	0	3	6
ATP	14	0	1	13
FAD	10	2	1	7
FMN	6	5	1	0
GLC	5	0	5	0
HEM	16	1	5	10
NAD	15	4	5	6
PO4	20	2	18	0
AND	2	0	0	2
EST	3	1	1	1
Total	100	15	40	45

(I) Ligands in the Kahraman dataset. (II) The number of complexes for each ligand.

[a] Diff donates the result of AUC score of *Circular Cone* subtracts that of ShaEP.

the other 40 cases, the AUC scores of both methods are almost the same.

From Table 3, we could find that *Circular Cone* works better than ShaEP on such ligands as AMP, ATP and HEM, while ShaEP does better than *Circular Cone* on FMN.

It could be found from Table 4 that the average processing time per structure of ShaEP is 895.5 ± 439.5 ms with a median of 687.5 ms while that of *Circular Cone* is 132 ± 26 ms with a median of 122.5 ms. Moreover, most of the average processing times per structure of *Circular Cone* are less than fifth of that of ShaEP, which indicates that the program of *Circular Cone* is faster than that of ShaEP.

From the result above, we could find that there are indeed some cases on which ShaEP works better than *Circular Cone*, at the same time, there are more cases on which *Circular Cone* does better. Additionally, *Circular Cone* is faster than ShaEP.

4. Discussion

According to the results, the AUC is ranged from 0.23 to 1.00, when some structures have higher scores and some have

Table 4 – The comparison of two methods with average processing time per structure.

I	II	III	IV
AMP	9	640	123
ATP	14	735	128
FAD	10	1335	158
FMN	6	752	122
GLC	5	552	107
HEM	16	919	148
NAD	15	907	150
PO4	20	456	106
AND	2	600	113
EST	3	590	111

(I) Ligands in the Kahraman dataset. (II) The number of complexes for each ligand. (III) The average processing time per structure of ShaEP (ms). (IV) The average processing time per structure of *Circular Cone* (ms).

lower ones. For example, according to Table 1, the structure of PO4 obtains the highest score of both methods. PO4 only has 5 atoms, making the volume of this molecule much smaller than that of the other structures. So it is easy for both methods to exclude, just from their sizes, pockets combined with other ligands, which leads to high AUC scores of both methods.

Another example comes from NAD, a structure of 44 atoms. It could be readily considered by shape that NAD and its pocket could be combined well to make a new complex. However, from our test, this situation obtains a lower AUC score. Our method works better in situations where more parts of the ligands are enfolded by the pockets, such as most complexes of ATP; while in this database, most of the pockets of NAD have big openings.

In our method, since the results of PCA depend on the scaling of the variables, it could not be confirmed that the axis got from PCA is exactly the principal axis of the original model after the normalization. So, it is no wonder that the accuracy of our method is not perfect. However, as we could conclude from Tables 1–3, in most situations, our method performs not worse than the other one, although in few situations, our method does not perform that well. Thus, results on this dataset show that the accuracy of our method is comparable.

Generally the method with a higher AUC score is supposed to have a lower speed. Compared with the ShaEP, our method has a high speed as well as good scores. Finding principal axis first is one factor that contributes to the high speed, which avoids a lot of rotations. Moreover, *end point* acts as another factor that contributes to the high speed, which avoids some translations. Additionally, using fewer calculations is also another factor.

Our method may be improved in the following ways. First of all, there are some structures on which our method performs worse than ShaEP. As illustrated above, our methods works better in situations in which more parts of the ligands are enfolded by the pockets. As to some other situations, which we could find from Tables 1 and 3, slowing down the speed and comparing more carefully would be a good choice. Secondly, flexibility, which is important to virtual screening, is another factor our method has not taken into consideration. Thirdly, scores given by our method lack an absolute standard, without which we could not know from the score whether a protein could match a ligand or not. At this moment, we could only use the scores to give a rank indicating the higher rank corresponding to a higher possibility of matching.

5. Conclusions

Shape matching is a widely studied topic because of its practical merit in relation to drug design. In this study, we have achieved a new method to measure the similarity between the proteins and the ligands. In this method, pockets and ligands are represented by points on their molecular surfaces. Then we obtain the principal axes to achieve global matching and reduce the degree of freedom. And we implement local matching with rotations to ensure the accuracy of the method. Finally, we calculate the similarity of the proteins and the ligands with a high speed. In the era of HTS, the speed is of great importance. So, the high speed of *Circular Cone* provides a good choice for HTS.

Acknowledgments

The support of NSFC 11101226 and NUSIP 101005541 is gratefully acknowledged. We thank Lei Wu for his assistance in programming.

REFERENCES

[1] S.M. Paul, D.S. Mytelka, C.T. Dunwiddie, C.C. Persinger, B.H. Munos, S.R. Lindborg, A.L. Schacht, How to improve R&D productivity: the pharmaceutical industry's grand challenge, Nature Reviews Drug Discovery 9 (2010) 203–214.
[2] C. Debouck, B. Metcalf, The impact of genomics on drug discovery, Annual Review of Pharmacology and Toxicology 40 (2000) 193–208.
[3] J. Burbaum, G.M. Tobal, Proteomics in drug discovery, Current Opinion in Chemical Biology 6 (4) (2002) 427–433.
[4] J.G. Gatto, The changing face of bioinformatics, Drug Discovery Today 8 (9) (2003) 375–376.
[5] P.T. Corbett, J. Leclaire, L. Vial, K.R. West, J.L. Wietor, J.K.M. Sanders, S. Otto, Dynamic combinatorial chemistry, Chemical Reviews 106 (9) (2006) 3652–3711.
[6] J. Bajorath, Integration of virtual and high-throughput screening, Nature Reviews Drug Discovery 1 (2002) 882–894.
[7] H. van de Waterbeemd, E. Gifford, ADMET in silico modelling: towards prediction paradise? Nature Reviews Drug Discovery 2 (2003) 192–204.
[8] G. Schneider, U. Fechner, Computer-based de novo design of drug-like molecules, Nature Reviews Drug Discovery 4 (2005) 649–663.
[9] T.B. Sebastian, P.N. Klein, B.B. Kimia, On aligning curves, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (1) (2003) 116–125.
[10] A. Kuijper, O.F. Olsen, Describing and matching 2D shapes by their points of mutual symmetry, in: 9th European Conference on Computer Vision, Graz, Austria, May 7–13, LNCS 3953 (Part III) (2006) 213–225.
[11] T.B. Sebastian, P.N. Klein, B.B. Kimia, Recognition of shapes by editing their shock graphs, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (5) (2004) 550–571.
[12] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (4) (2002) 509–522.
[13] F. Mokhtarian, A.K. Mackworth, A theory of multiscale, curvature-based shape representation for planar curves, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (8) (1992) 789–805.
[14] H. Asada, M. Brady, The curvature primal sketch, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI 8 (1) (1986) 2–14.
[15] D. Marr, H.K. Nishihara, Representation and recognition of spatial-organization of 3-dimensional shapes, Proceedings of the Royal Society of London. Series B. Biological sciences 200 (1140) (1978) 269–294.
[16] D. Shaked, A.M. Bruckstein, Pruning medial axes, Computer Vision and Image Understanding 69 (2) (1998) 156–169.
[17] K. Siddiqi, B.B. Kimia, A shock grammar for recognition, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1996, pp. 507–513.

[18] S. Gold, A. Rangarajan, A graduated assignment algorithm for graph matching, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (4) (1996) 377–388.
[19] D.G. Kendall, D. Barden, T.K. Carne, H. Le, Shape and Shape Theory, John Wiley & Sons, Chichester, 1999.
[20] H. Ling, D.W. Jacobs, Using the inner-distance for classification of articulated shapes, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 2 (2005) 719–726.
[21] C. Scott, R. Nowak, Robust contour matching via the order-preserving assignment problem, IEEE Transactions on Image Processing 15 (7) (2006) 1831–1838.
[22] L. Chen, R. Feris, M. Turk, Efficient partial shape matching using Smith–Waterman algorithm, in: Proceedings of NORDIA Workshop at CVPR, 2008, pp. 1–6.
[23] Conrad Huang, http://www.cgl.ucsf.edu/chimera/docs/UsersGuide/midas/dms1.html.
[24] A. Kahraman, R.J. Morris, R.A. Laskowski, J.M. Thornton, Shape variation in protein binding pockets and their ligands, Journal of Molecular Biology 368 (1) (2007) 283–301.
[25] F.M. Richards, The interpretation of protein structures: total volume, group volume distributions and packing density, Journal of Molecular Biology 82 (1) (1974) 1–14.
[26] K. Pearson, On lines and planes of closest fit to systems of points in space, Philosophical Magazine 6th series 2 (1901) 559–572.
[27] E.A. Coutsias, C. Seok, K.A. Dill, Using quaternions to calculate RMSD, Journal of Computational Chemistry 25 (15) (2004) 1849–1857.
[28] V. Saridakis, D. Christendat, M.S. Kimber, A. Dharamsi, A.M. Edwards, E.F. Pai, Insights into ligand binding and catalysis of a central step in NAD^+ synthesis: structures of methanobacterium thermoautotrophicum NMN adenylyltransferase complexes, Journal of Biological Chemistry 276 (2001) 7225–7232.
[29] W. DeLano, PyMOL: an open-source molecular graphics tool, CCP4 Newsletter on Protein Crystallography 40 (2002).
[30] J.L. Avalos, J.D. Boeke, C. Wolberger, Structural basis for the mechanism and regulation of Sir2 enzymes, Molecular Cell 13 (5) (2004) 639–648.
[31] M.J. Vainio, J.S. Puranen, M.S. Johnson, ShaEP: molecular overlay based on shape and electrostatic potential, Journal of Chemical Information and Modeling 49 (2) (2009) 492–502.