國立臺灣大學電機資訊學院資訊工程學系

碩士論文

Department of Electrical Engineering
College of Computer Science and Information Engineering
National Taiwan University

Master Thesis

使用旋轉切面影像進行蛋白質活性基比對與鍵結

Matching and Docking of Protein Active Sites with Spin Images

張席維

Chang, Hsi-Wei

指導教授：歐陽明 博士

Advisor: Ming Ouhyoung, Ph.D.

中華民國 97 年 6 月

June, 2008

摘要

本篇論文展示了一種以 Spin Image 做到快速的蛋白質 Receptor to Receptor，Ligand to Receptor 的比對或配對之技術，其結果爲在短時間內留下較有可能的蛋白質，並藉由去除掉大部分不可能的樣本，減少實驗錯誤樣本所需要的時間，以期能增進新的藥物，抗體研發的效率。

本系統進行的程序上主要分爲兩部分：之前的資料準備以及線上的資料比對。在資料準備上，我們建構一個蛋白質活動基的資料庫系統，皆以 Spin Image 爲基礎儲存。而在線上比對的部分，又分爲表面模型建構、Spin Images 建構、Spin Image 比對等三個步驟。給定一個未知的蛋白質，可由資料庫中的活動基資料，藉由全體比對而得到可能的相似活動基位置、或者可配對的位置。

藉由找出相似的活動基位置、或幾何上可能結合的受體位置，可幫助生物學家們在尋找有效藥物的過程中給予適當的建議與方向。

在結果的部分，建構一個一萬網點的蛋白質全體影像，時間約在一天左右。而線上比對一活動基與上述蛋白質全體影像，所花的時間約爲兩分鐘。因此以現有 PDB 資料庫結構規模約 50,000 筆來說，在資料庫事先建構完成的情況下，對一個蛋白質做完整的搜索約需要 70 天。多核心電腦上使用 multi-thread 技術似乎是一個加快上述計算之有效方法，因爲 50,000 筆資料比對正可以分群來比對，符合平行分散計算之原則。

Abstract

This paper demonstrates a rapid method using spin images to compare or dock the receptors between proteins by searching the possible protein collections in a reasonable period of time. By getting rid of impossible proteins, we can reduce the time consumed in experimenting on obviously wrong ones. We expect it can improve the efficiency of discovering new drugs and antibodies in the future.

The process of our system mainly consists of two stages. The first one is off-line ligand database construction and the second one is on-line data comparison. In the data construction stage, we build a database of protein active sites, storing them in spin images. In the on-line comparison stage, we will go through the following three steps: surfaces construction, spin image construction, and comparison of spin images. Given an unknown protein, we can give a full-set comparison with known active sites in the database to find the possible position that similar ones may reside, or further, to guess the possible position that may be a receptor site.

By finding the position of similar active sites, or the geometrically possible docking position of receptors, we can give some suitable suggestions or directions, and may help biologists in searching effective drugs.

Finally, in results, constructing a full set of spin images of a protein which had 10,000 vertices in its mesh takes about 1 day to generate. Comparing an active site with the spin images mentioned above would take about 2 minutes. As a result, under the well constructed database, a worst case of full-set searches would take about 70 days, based on the size of current PDB, about 50,000 protein structures that had been discovered. Using multi-threads technology for multi-core computers is one way to speed up the full-set search, since the 50,000 structures can be easily decompiled into clusters for parallel search.

目　錄

口試委員會審定書…… 0

中文摘要…… 1

英文摘要…… 2

目錄…… 4

Chapter 1 Introduction…… 5

1.1 Problem…… 5

1.2 Related Works…… 6

Chapter 2 Algorithm and Implementation…… 7

2.1 Algorithm Overview…… 7

2.2 Surfaces…… 8

2.3 Spin Images…… 11

2.4 Spin Image Generation…… 14

2.5 Comparison Functions…… 15

2.5.1 Linear Correlation Coefficient…… 15

Chapter 3 Result and Discussion…… 18

3.A Receptor to Receptor Matching…… 18

3.A.1 以 Elephant, Starfish model 爲例…… 18

3.A.2 以蛋白質 DHFR 爲例…… 21

3.B Receptor to Ligand Docking…… 25

Chapter 4 Conclusion and Future works…… 32

Chapter 5 Reference…… 33

Chapter 1

Introduction

1.1 Problem

在科技醫學日新月異的今日，許多新出現的病菌，病毒，甚或癌症等等無不威脅著現在的人們；而藥物的研發速度，卻已漸漸的跟不上病菌，病毒的變種速度。

翻開古典的藥物史，幾乎所有的藥物都在 Trial and error 之中發現，甚或發現藥物本身就是個意外。而事實上藥物如何對抗病毒，爲何藥性有效一直是個謎。直到雷射解析分子結構的時代來臨，一些藥物運作的原理才漸漸被人們所了解。舉個例子來說，著名的愛滋病治療藥物 AZT，它的一部份結構與 DNA 組成物質: Thymidine 相似，但是尾端的結構，卻與其大不相同。由於愛滋病病毒會使用 Thymidine 來做自我複製的行動，當其「誤食」了 AZT 之後，因爲無法成功複製，也無法分開，便因而「死亡」。有一部分的抗癌藥物，也是類似的原理。而其中 AZT 與 Thymidine「相似」的部份，也就是欺騙了病毒那部分，是爲其 Active ligand（與病毒 binding site/ receptor 結合，Ehrlich，1897）。

爲了簡化問題，我們從組成生物體的基本組織「蛋白質」出發。但儘管是蛋白質這個相對簡單的結構，光在 RCSB PDB Databank 中已紀錄的數量就有將近五萬筆，現有的配對方法，動輒數小時至一天以上的時間，要做完資料庫裡的比對可說是不合理的事情，更不用說一個新的病毒產生，要花多少時間去尋找對治的蛋白質結構了。因此爲了能夠幫助對抗這些

頑強的敵人，我們期望在蛋白質層級上提供有效率的演算法，幫助預測其 Receptor 與 Ligand 的配對可能。

1.2 Related Works

我們使用的方法，在 Bioinformatics 之中被歸類於 Structure Biology。這是一個近年來才被重視的領域，甚至還不到 50 年。而其重要的中心思想爲：若兩個 ligand 能夠產生反應，則其結構必有某種程度上的配合，畢竟分子在空間中總是佔有著自己的一席之地。

相關的研究近年來已漸受重視，這邊列舉一些研究的內容與方向參考。

在蛋白質功能預測方面，完整的 Docking 已經尋之有年，一些例子有 AutoDock [RGAD06]、ZDOCK [CLW03]、RDOCK [LCW03]、GEMDOCK [JC04] 等等，在研究上目前大致分爲兩個方向，其一爲使用互補表面來做幾何上的對應，而另一派的方法則是使用模擬 Docking 過程去計算兩者的作用能量而達到效果。

互補表面方面，主要想法是使用一些特徵與幾何表面去描述蛋白質本身，再對它們做 Docking 的動作。大部分的情況下，Solvent-accessible surfaces 是最常被使用的表面（後述），亦有只使用蛋白質主鍊上的原子做處理，或使用 Fourier shape descriptor 去表示其表面的方法。無論使用何種方法，使用表面來做幾何的 Docking 具有相對上較爲快速且穩定的好處。

而使用模擬的方法，相對來說複雜了一些。簡單的想法是將兩個實驗的蛋白質分開一定的距離，模擬其可能運動的每一小步，直到最後 Docking 成功。在進行的每一小步包括了移動、轉動等等（甚至是分子內化學鍵的轉動），而在每次的運動中，皆影響一個系統能量的變數。就如同其他的模擬演算法一般，最後慢慢使熵減少而達成 Docking task。這部分的好處是相對幾何直接對應的方法來說較爲靈活，而在接近 docking 的過程中與實際的情況較相似；當然，缺點在執行效率過差，在越大的行動空間中其複雜度成長驚人，現在也已有一些方式嘗試解決這個問題。

而由於時間效率一直是 3D-Docking 很大的問題，除了實際 Docking 之外，在判斷蛋白質 Active ligand 功能上，亦有許多人提出了不同的想法。如 Volumetric Models [TRJ06]，使用的是對蛋白質凹洞（有可能爲 Receptor 處）建構出來之 Volumetric Models，以一個 3D grids 組合來代表其蛋白質的凹洞內部結構。給定一個 active site 位置，使用 knowledge-based 的方法建構出在其凹洞中的模型，之後使用一個 grid matching 的演算法進行與其它 active sites 的比對。Fingerprints [DJJJJWA06] 的方法，則是使用 Frequent Subgraph Mining 的技術，從 SCOP 之中取出相對應家族的「指紋」。給定一個目標的蛋白質，即可藉由 subgraph isomorphism 的快速演算法找出其「指紋」並藉此找出其對應的功能。

而最後，有許多使用 Data mining、Machine learning 的方法也都在嘗試之中，這是一個生命力旺盛的領域，而我們的 spin image approach 亦希望能在這個時空下能夠有一些些的貢獻。

Chapter 2

Algorithm and Implementation

2.1 Algorithm Overview

在我們的實驗中，主要針對於 Spin Image 的產生與使用此技術所增進的比對效益。而這是由於 Spin Image 在比對上的速度極快（$O(n^2)$ 的時間複雜度。）的緣故。以下爲整個演算法流程：

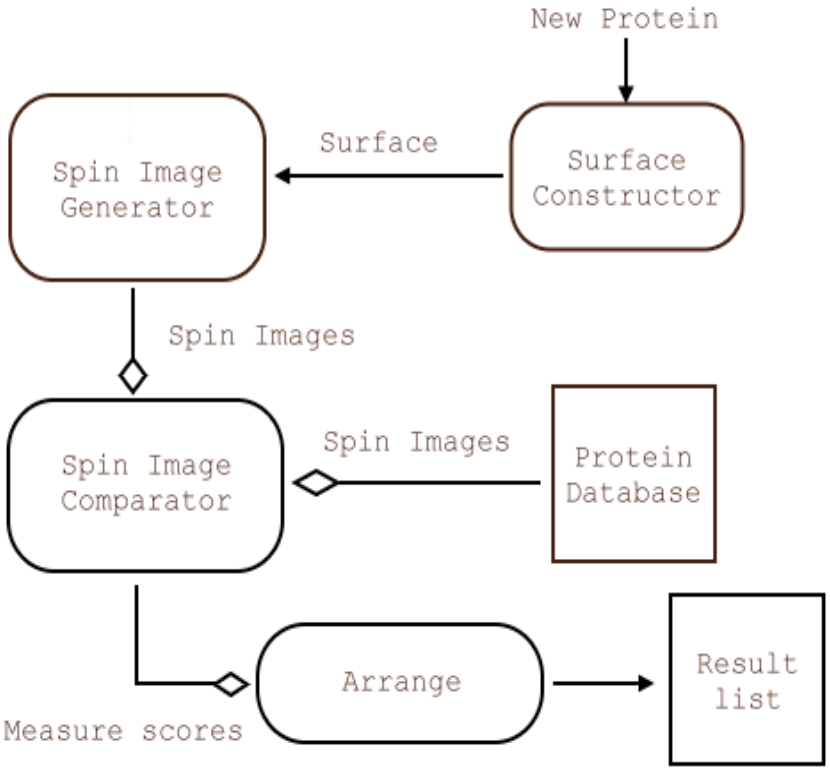
1) 取得蛋白質的表面資料

2) 取得可能之 Receptor (Inhibitor) 之位置

3) 產生對應之 Spin Image

4) 與資料庫中之 Spin Images 比對（Docking or Matching），產生結果。

我們將一一走過每個步驟。



蛋白質比對鍵結運作流程圖

2.2 Surfaces

在 Surface 的處理上，給定一個蛋白質的結構，目前有三種定義方式：Van der Waals surface，Solvent accessible surface，Contact/re-entrant surface。

Van der Waals surface，在建構上非常的容易（甚至不需要預先建構），也是最直接的表面。其定義爲蛋白質組成結構中，每個原子表面的集合（即以每個原子半徑形成的球體集合）。此種表面的優點在於它的簡單，而缺點在於其會產生許多無法 docking 的空隙，不但不具意義，更會發生擾亂評分階段不必要的差距。

Solvent accessible surface (SAS)，是我們目前採用的表示方法。它的想法在於使用一個原子大小的球型探針去碰觸蛋白質，而表面即是這個球型探針圓心繪出的集合。使用 SAS 的好處在於這個表面其實是個 marginal surface，以致於在這個 surface 被否決的例子可信度會比較高，再一次的提到，我們對 recall 的要求。

Contact/re-entrant surface，是較新的表示法，想法與 SAS 類似，差別在於 SAS 取的是探針的圓心，Contact/re-entrant surface 則是取探針與原子接觸的點，當探針一次與多個原子接觸時，取探針表面接觸點間形成的區塊。原則上應該是最接近蛋白質外觀的表面，但由於這種緊貼著蛋白質的表面在比對上的容錯較差（儘管它較爲準確），因此未被我們納入使用的名單之上。

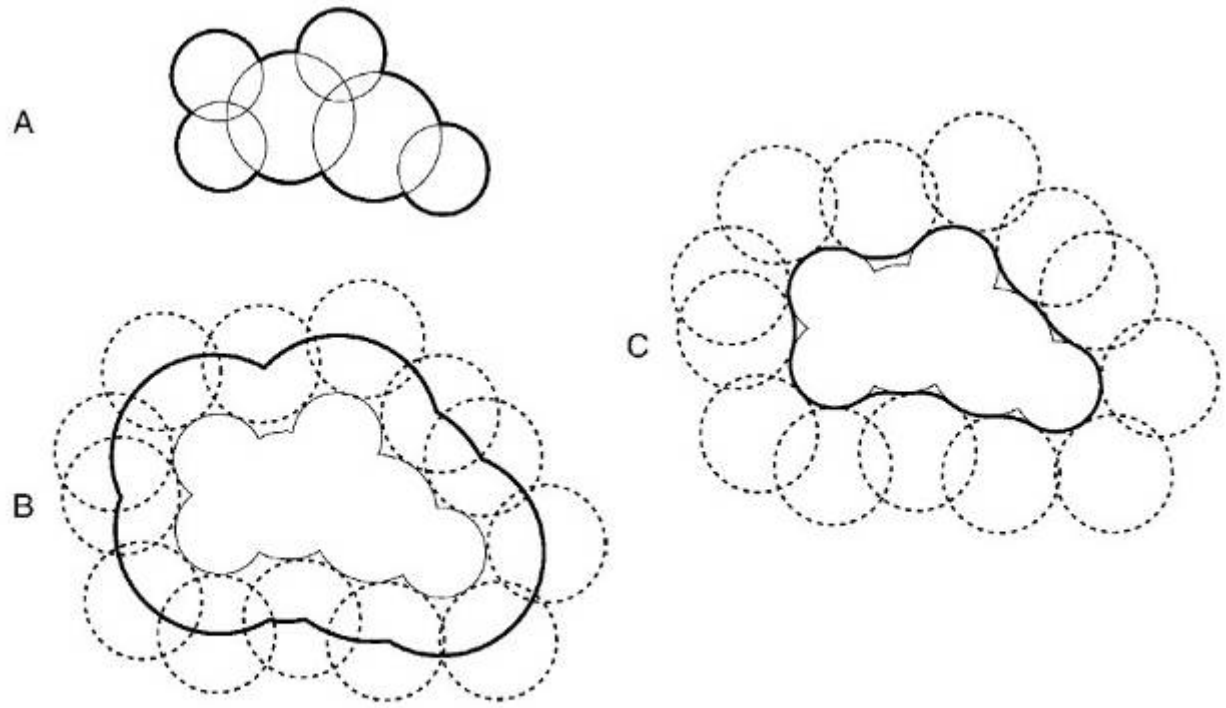
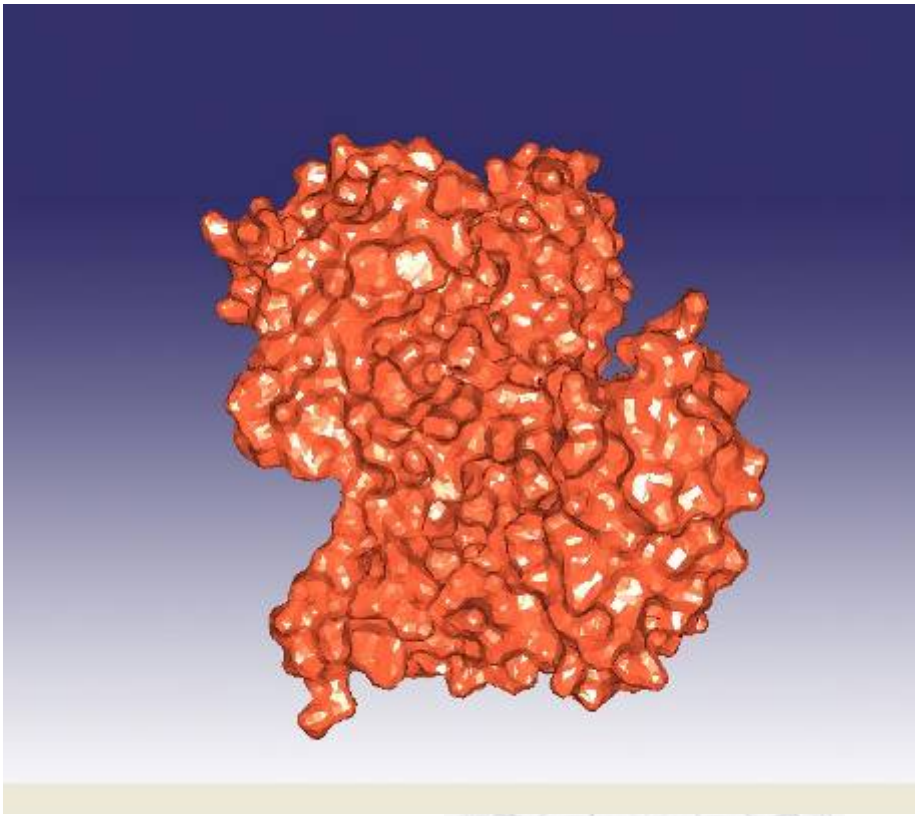


Fig. 1. Cross-sections of the (A) van der Waals surface, (B) SAS and (C) contact/re-entrant surface.

SAS 的實作並非來自我們，而是使用 Connolly [M83] 完成的演算法與函式庫。

在表面取得完成之後，另一個問題是如何找到可能的 Receptor 位置，這邊我們使用了 Brandy 在 2000 年提出的 Putative Active Sites with Spheres (PASS)。PASS 從蛋白質表面的凹陷深度與曲率來判斷其爲 Receptor 的可能性，我們在此完成了製作 Spin Images 的前置作業。



此爲 SAS 的一個模型例子。

2.3 Spin Images

什麼是 Spin Image 呢？其定義爲：

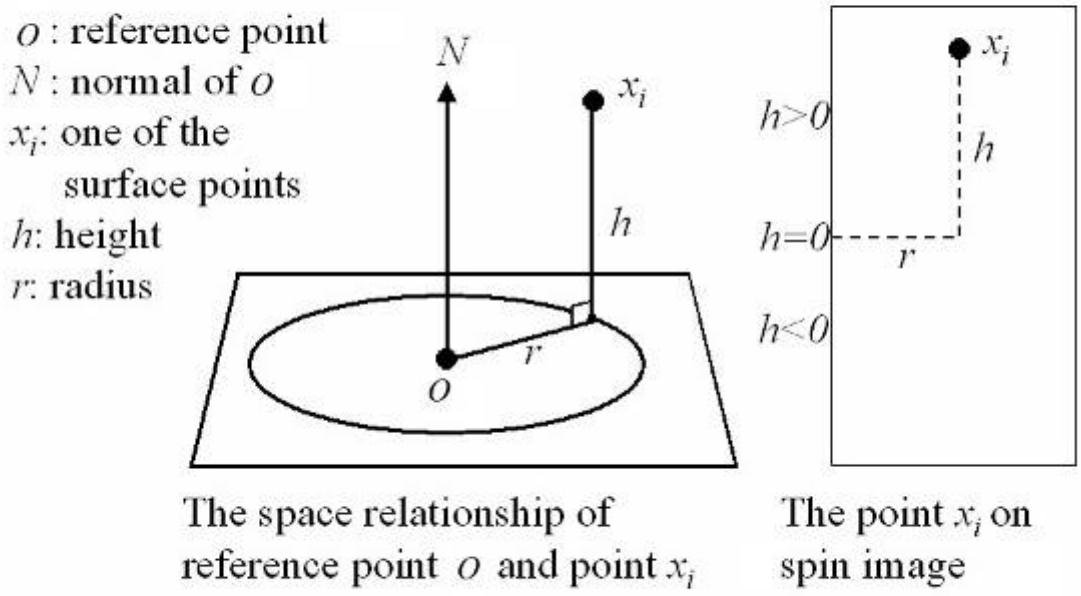
Given 3D model M, vertices V belongs to M, an origin O, and a normal N.

The spin image of M on (O, N) is a 2D Image (H, R, D) (denotes X, Y, depth), where

H is the distance between V and Plane(O, N),

R is the distance between V and Line(O, N), and

D is the density of V which mapped on (H, R).



The space relationship of reference point o and point x_i

The point x_i on spin image

2.4 Spin Image Generation

產生 Spin Image 的流程如下：

1) 取一點 O 以及向量 N 得一平面 E(O,N) 與直線 L(O,N)

2) 對於 3D 圖形上的任意的點 v 求得與平面 E 距離 h 與直線 L 距離 r

3) Spin Image I 之上 (r,h) 深度加一

4) 重覆直到 3D 圖形之所有點都紀錄完成

如何產生 Spin Image 的方式會影響它的效率。最簡單的實作方式就如同上面，產生 3D 圖形上的點之後對所有的點計算 E, L 的距離。但是實際上使用一點方法可以讓產生圖形的速度更快。

考慮直接進行運算的流程，求得 Distance(v, E) 需要計算以下公式：

$$\text{Distance}(v, E) = \left| \frac{(ax_v + by_v + cz_v + d)}{\sqrt{a^2 + b^2 + c^2}} \right| \text{ Where } E : ax + by + cz + d = 0$$

而 Distance(v, L) 則需要：

1) 取一平面 e 過 v 其 normal 爲 L。

2) 取 L 與 e 交點 k。

3) Distance(v, L) = Distance(v, k)

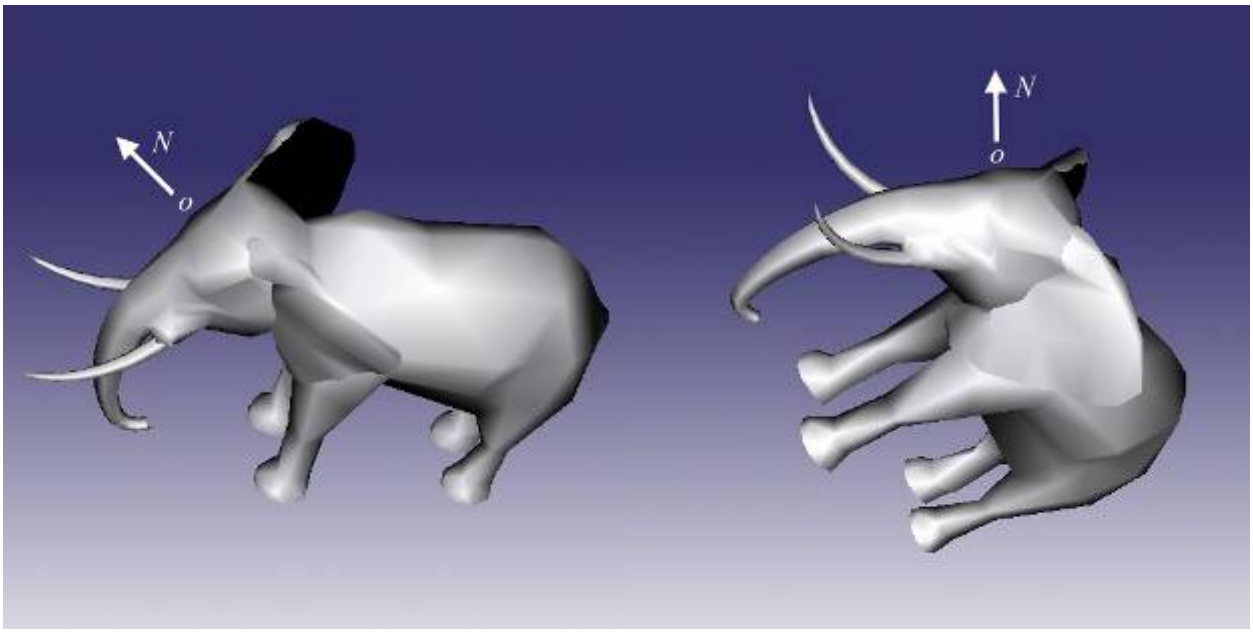
光每個點的計算量就需要超過二十個浮點數乘法、除法和取平方根。這個情況在實行上是極爲不妥的。但是「換個角度來看」，問題可能就簡單多了。

想像一個 3D 圖形在經過旋轉後會成爲什麼樣子。我們將已知的 N 做一個旋轉（使用 Quaternion）使其指向 Z-axis，在產生 M 表面點時先位移至 O，並對 surface mesh 做與 N 相同方向的旋轉，而這個在 mesh 上的旋轉只使用總共 31 個加法與乘法。在整個旋轉完成後，計算距離的方式便成爲：

$$\text{Distance}(v, E) = |z|$$

$$\text{Distance}(v, L) = \sqrt{x_v^2 + y_v^2}$$

每個點的計算量減少到總共四個計算，約增加了五倍以上的速度。



將模型轉到適當的角度，便可以較爲簡單的計算出其圖形。

以上可以看出每個點的 z 值即我們所求的 h。

2.5 Comparison Functions

而在比對上，什麼叫做相似？而什麼不是？以 Spin Image 上的一點來看，每個點都是一個環的集合，由於 r 值代表著其在 3D 中這個環的半徑，當 r 越大，其所含括的區域越廣，資料的損失也越大，其參考價值相對的就較低了。但儘管如此，幸運的是以我們的蛋白質結構來看，無論是表面的凹陷或者凸出處（即可能爲 Receptor/Active Site 的地方），其規模都還在可以容許的範圍之內。

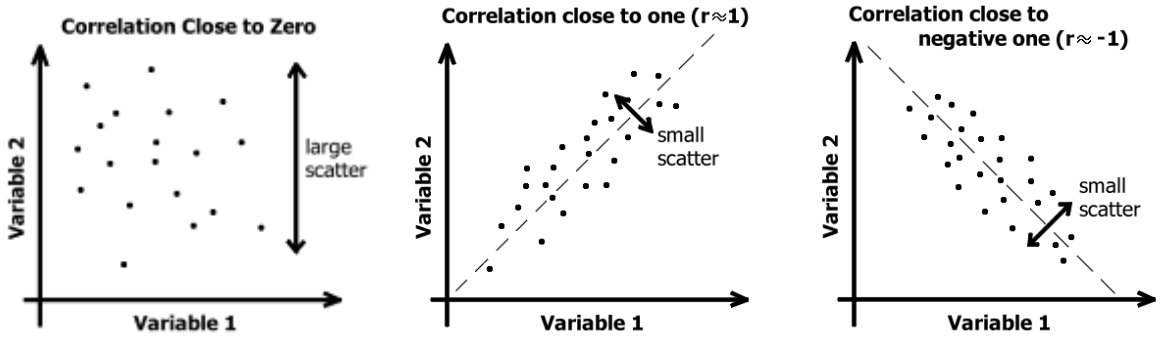
由於 Spin Image 在不同結構下相同的機率極低，並且因爲我們只對刪去大部分不可能的情況有興趣；我們可以知道同樣或相似的結構所求出的 Spin Image 必然會有相似性，符合我們希望高 recall 的需求。

儘管是同樣的蛋白質，因爲取樣點的不同、法向量的不同，甚或在取得 surface 的過程亦有可能產生差異性，因此我們使用線性相關係數（Linear correlation coefficient）來比對兩張 Spin Images 有多相近。對於一個 Receptor 我們將其 mesh 上每個點的 Spin Image 存成一個 Spin Image Stack（offline）。當一新的蛋白質之 Binding site 進來，我們針對其 mesh 上的點做 random sampling，取 N 個點（在時間許可的情況下，亦可對所有點做），得到它們的 Spin Images。使用這些 Spin Image 到我們的 Protein database 之中比對，只要與其中一張圖形的線性相關係數達到一定的信任指數 k，便可得其爲可能的 Binding site（或相似的 Receptor）。以下，我們會詳細的說明線性相關係數的計算方法（對統計熟悉的讀者可以跳過這段）。

2.5.1 Linear Correlation Coefficient

線性相關係數是個介於 1 ~ -1 之間的的數字。它代表著一組相關的點與一條直線的相近

程度，這個數字越接近 0 就表示這些相關的點集合越接近一條直線（也因此叫做「線性」相關係數）。



以上三張圖分別表示相關係數在三種不同的數值上，兩份資料的分佈情形，數值越接近 1，則其分佈越趨近於由左下至右上角的直線；數值越接近 -1，則偏向由左上至右下角的斜線；而當數值越趨近於 0，則兩數列之間越接近於隨機分佈，相關性也越低。

計算方法如下：

設兩數列

$$X:(x_1,x_2,x_3,...,x_n)$$
$$Y:(y_1,y_2,y_3,...,y_n)$$

1) 計算平均值

$$Xa=\frac{1}{n}\sum_{i=1}^{n}x_i$$
$$Ya=\frac{1}{n}\sum_{i=1}^{n}y_i$$

2) 計算兩者的標準差

$$\sigma_x=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-x_a)^2}$$
$$\sigma_y=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-y_a)^2}$$

3) 計算兩者協方差

$$\sigma_{xy}{}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-x_a)(y_i-y_a)$$

4) 最後線性相關係數 r 即爲：

$$r = \frac{\sigma_{xy}^{\ 2}}{\sigma_x \sigma_y}$$

從以上我們可以看到計算線性相關係數的步驟複雜度爲 $O(N^2)$（N 爲一個維度的長度）。爲了增加比對的效率，我們只計算與檢索蛋白質活動基重疊的部分（直覺的，亦即大於目標的部分皆不計算），另外，由於我們要比對的目標是兩種不同的蛋白質，只計算重疊的部分可以降低蛋白質額外部分產生的誤差。當越多的點被計算在內時，我們相信比對的結果將會越有說服力，而值得注意的是線性回歸係數本身之 Variance 亦可用來當做其信心指數的參考。因此我們使用一個相似度計量 C 來結合線性回歸係數 R 與其 Variance [AM99]。當計算兩個 Spin Images P, Q 時，N 爲兩者重疊的資料量，則 C 定義爲：

$$\mathrm{C}(P,Q) = \mathrm{atanh}^2(\mathrm{R}(P,Q)) - \frac{\lambda}{N-3}$$

這種改變變數的做法，使用 Hyperbolic arctangent function 能將我們原來的線性相關係數，改變成更具有統計意義的分佈。而更重要的特性是，經過這樣子的轉換之後，R 的 Variance 轉變成爲 $\frac{1}{n-3}$，由簡單的計算即能完成。最後，λ 是一個加權值，經過經驗的計算我們暫定爲 3。

Chapter 3

Results and Discussion

我們的演算法產生 Spin image 的複雜度爲

$O(n^3)$, Where n is the amount of units sampled in one dimension

可以理解成複雜度隨著解析度增加，而兩張 spin images 比對的複製度爲 $O(n^2)$。

我們主要的目的分爲兩類問題。其一（A）是相似功能之 Receptor（或 inhibitor）之比對；其二（B）是 Receptor/Inhibitor 之間的 Docking。

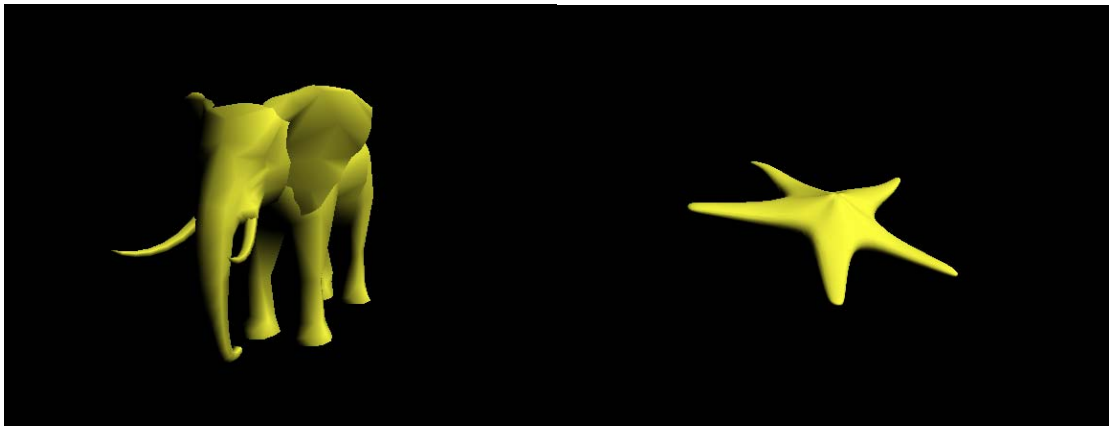
以下所有結果運行環境：Intel Centrino 1.4GHz, DDR-333 1.23GB, WindowsXP SP2, JRE 1.6.0_05-b13。

3.A Receptor to Receptor Matching

在實驗的部分，我們分爲幾個例子來討論，蛋白質的取樣爲 Dihydrofolate Reductase (DHFR) 的檢索，在簡單的例子上，我們先使用一個容易判斷的大象模型來實驗。

3.A.1 以 Elephant, Starfish model 爲例

在最開始的實驗我們以一個大象模型與海星模型，嘗試驗證我們實驗的正確性。由於蛋白質本身的結構複雜，較不易看出實驗的結果，我們便從肉眼可輕易判別的模型開始。



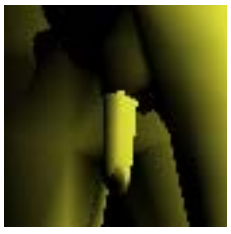
模型 Elephant(左) 與 Starfish(右)

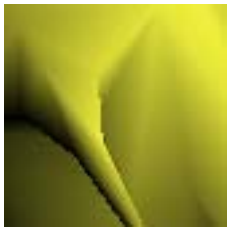
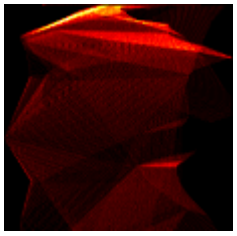
[產生 Spin Image 的時間]

Model Name	Vertices	Faces	Spin Images	Used Time(s)	Avg. Time(s)
Elephant	623	1,148	623	10,384	16.67
Starfish	1,890	3,776	1,000	16,522	16.52

Spin Image 的比對時間，平均每張圖 0.003 秒。

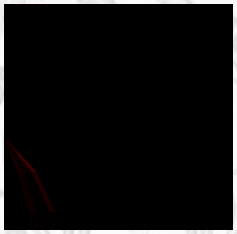
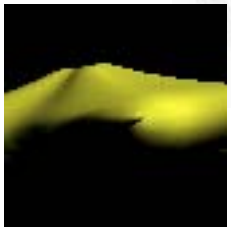
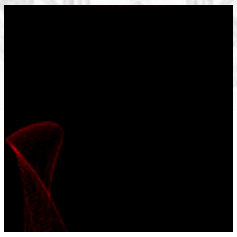
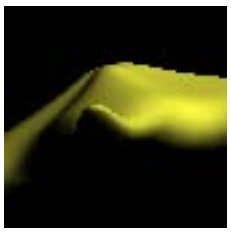
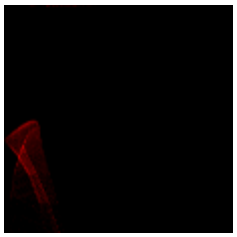
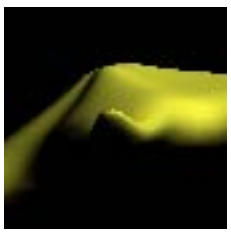
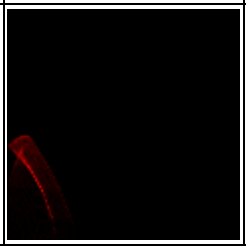
以下爲使用 Elephant 模型與自己的比較。每個 Reference Point 皆擁有自己的號碼與相對應的法向量。而每個 Reference Point 皆有屬於它自己的 Spin Image，藉由判斷其 Spin Image 的相似程度，我們便可以知道最符合檢索目標在模型上的位置。以下是實驗結果。

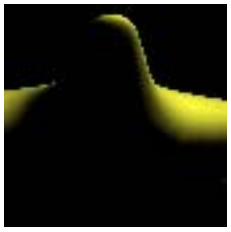
Model, R.P. No.	3D Image	Spin Image	Measure Score	Hit
Elephant, 44 大象左牙頂點 (Query)			N/A (Best Match)	O
Elephant, 88 大象右牙頂點			11.9744 (1st candidate)	O*
Elephant, 41 大象左牙頂邊			0.0523 (2nd candidate)	O

Elephant, 580 大象屁股			-0.0025 (Worst Match)	X

在模型自我的比較之中，大象左牙頂點與大象自己其它位置比較，可得到最佳 Reference Point 爲右牙的頂點。這個結果（*）因爲大象模型左右對稱的緣故，是正確且合理的。而在次佳的結果亦反應爲左牙近頂點的另一 Reference Point。在這個例子中我們亦取出最差的結果來比較（留意屁股並未包含尾巴尖端的部分）。

而接下來，仍然以大象左牙爲目標，嘗試在海星模型上尋找相似目標。在我們預期中，海星最佳解應該落在腳部（與象牙最像處）。

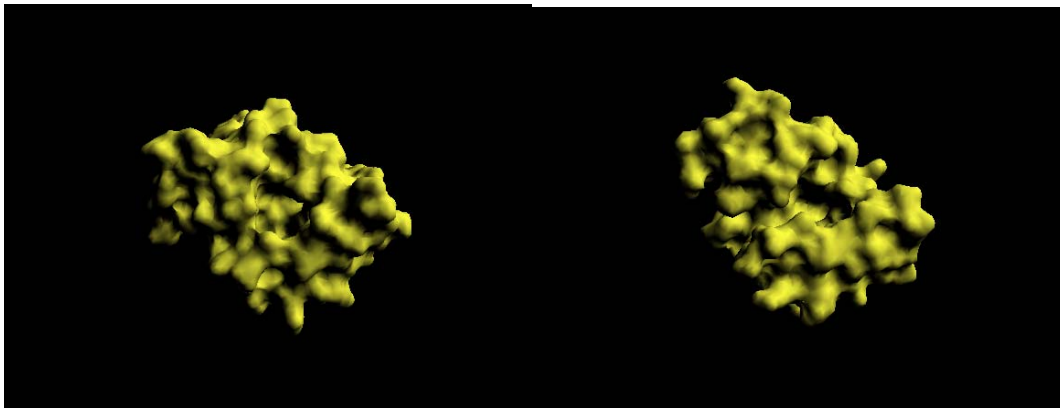
Model, R.P. No.	3D Image	Spin Image	Measure Score	Hit
Elephant, 44 大象左牙頂點 (Query)			N/A (Best Match)	O
Starfish, 399 海星腳 1			0.1686 (1st candidate)	O
Starfish, 761 海星腳 2			0.1629 (2nd candidate)	O
Starfish, 621 海星腳 2			0.1541 (3rd candidate)	O

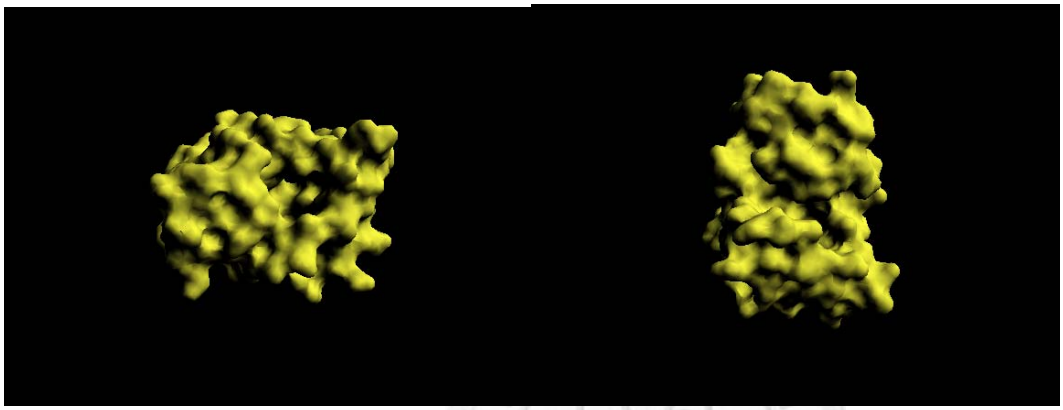
Starfish, 580 海星肚			-0.0025 (Worst Match)	X

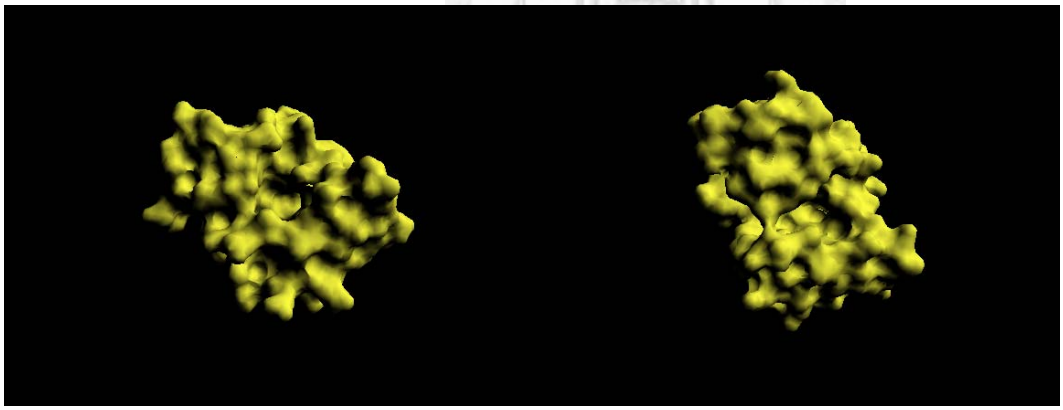
在與海星的比對之中，如我們預期的，腳部尖端的部份被截取了出來。而在海星五隻腳當中，有曲度的兩隻腳（1,2）分數比其他腳爲高，表示象牙的曲度亦有在 Spin Image 之中表現出來。

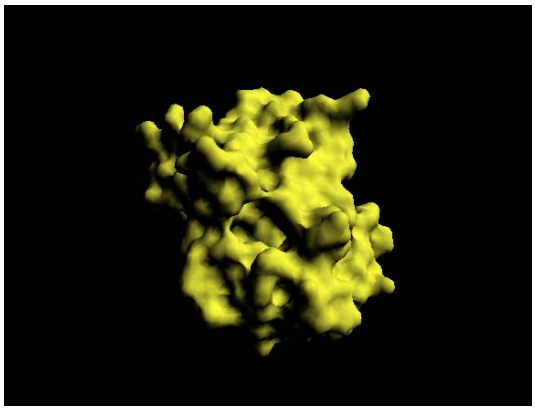
3.A.2 以蛋白質 DHFR 爲例

鑒於實驗時間的不足，在實際蛋白質比對上我們做了一些隨機取樣。DHFR 分子之中網點數平均約在 10,000 個點左右，此次我們在每個分子中皆取 200 個樣本點來進行實驗，結果如下：







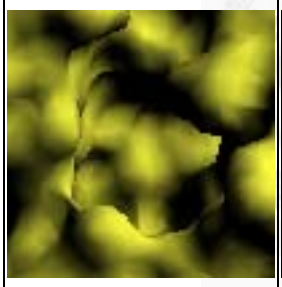
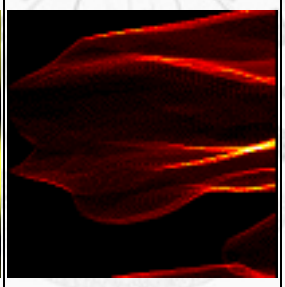
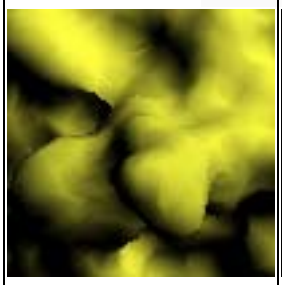
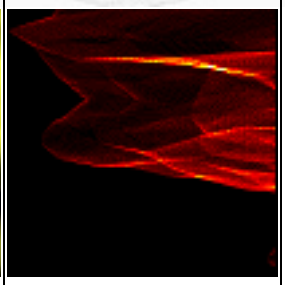
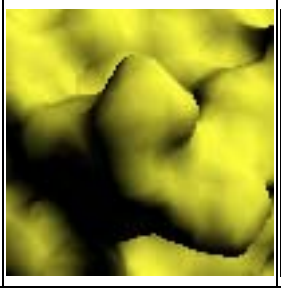
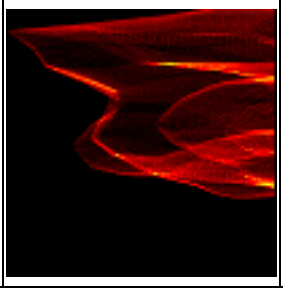


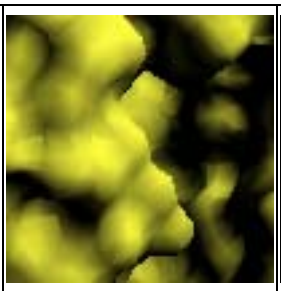
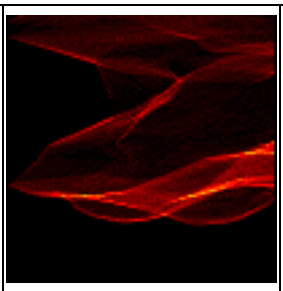
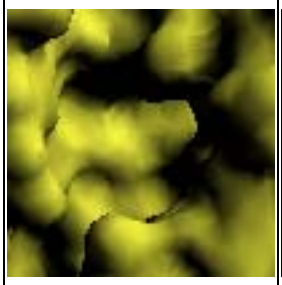
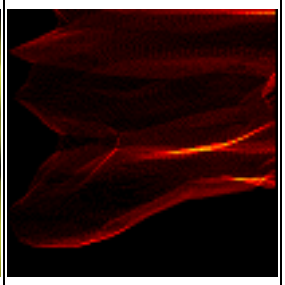
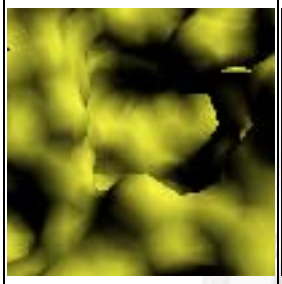
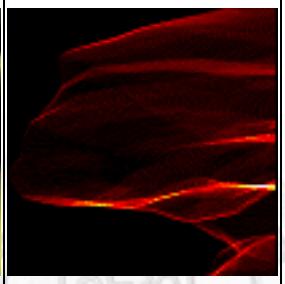
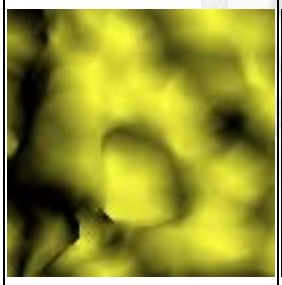
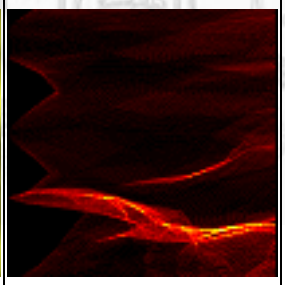
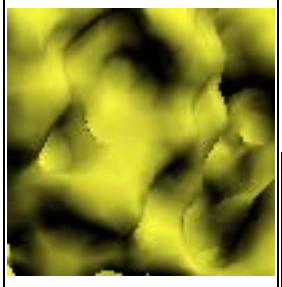
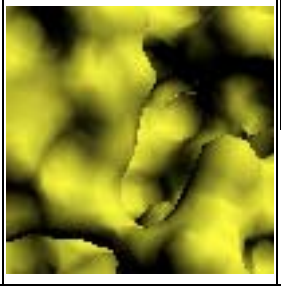
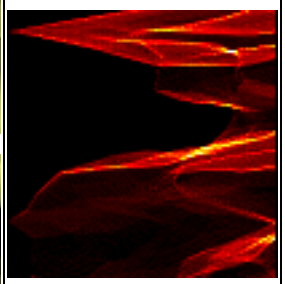
由上至下分別爲：
1DDRA、1DDRB、
1DHJA、1DHJB、
4DFRA、4DFRB、
8DFR

[產生 Spin Image 的時間]

Model Name	Vertices	Faces	Spin Images	Used Time(s)	Avg. Time(s)
1DDRA	9,584	19,168	200	2,954	14.77
1DDRB	9,940	19,872	200	1,915	9.575
1DHJA	9,917	19,838	200	1,770	8.85
1DHJB	9,774	19,544	200	1,777	8.885
4DFRA	9,706	19,412	200	1,749	8.745
4DFRB	9,909	19,818	200	1,750	8.75
8DFR	11,817	23,650	200	1,974	9.87

P.S. 平均時間小於 Elephant 模型是因爲蛋白質產生 Spin Image 時只做活性基規模的局部區域生成，其 Spin Image 的大小爲 Elephant 模型之一半。每次一對之比對時間約 0.001 秒。

Model, R.P. No.	3D Image	Spin Image	Measure Score	Hit
1DDRA, 201 (Query)			N/A (Best Match)	O
1DDRB, 51			0.1820 (1st candidate)	X
1DHJA, 74			0.1770 (2nd candidate)	X

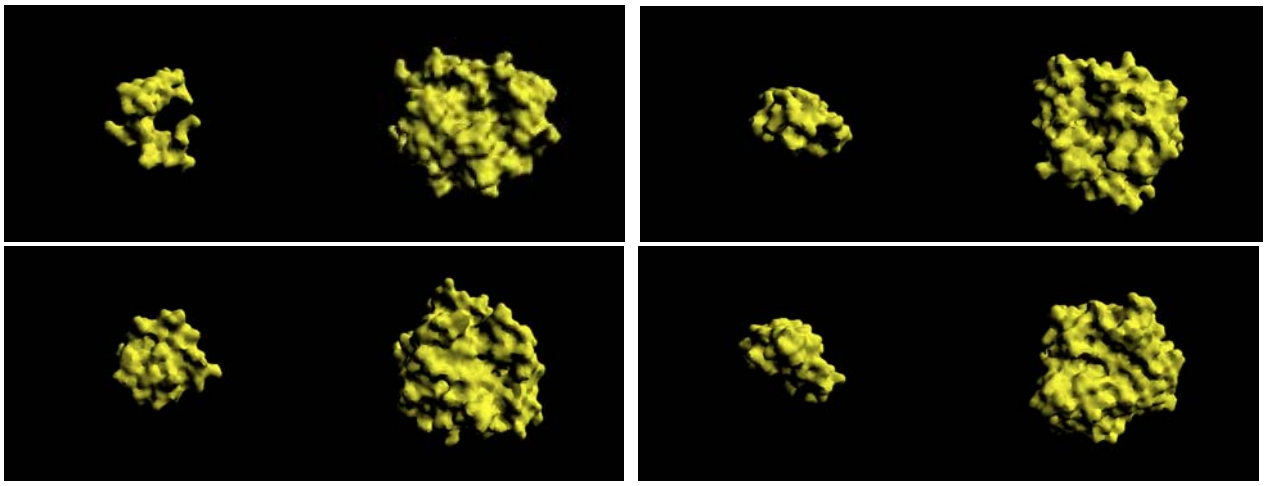
1DDRB, 125			0.1148 (3rd candidate)	X
1DDRB, 201			0.1053 (4th candidate)	O
4DFRA, 201			0.0985 (5th candidate)	O
4DFRA, 194			-0.0099 (Worst Match)	X
4DFRB, 47	 		0.0684 (Special)	O*

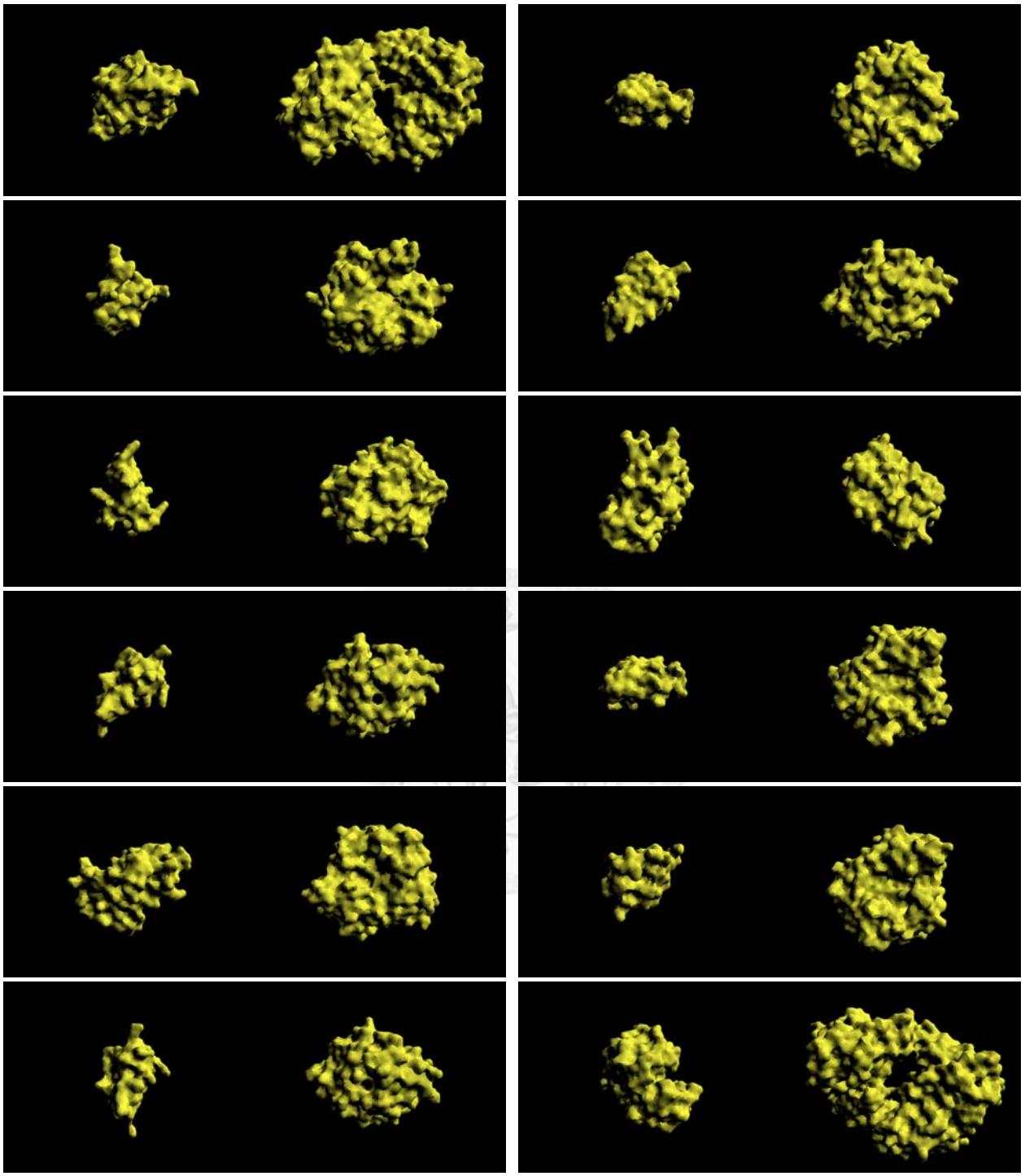
在蛋白質的比對方面，由於其結構差別較大，加上取樣數略爲不足，結果不容易由肉眼判斷其相似程度。但在大略形狀方面，最佳解之第四、五個可以明顯的看到與檢索蛋白質凹洞（即 Receptor 所在之處）相似的凹洞（分數分別爲 0.1053、0.0985）；另外，在特殊的例子（*）之中，雖然未在最好的排名，但其實在稍做旋轉之後（下方的圖），即可看見被遮住之凹洞，揭示我們也許需要將這個現象加入考量（其分數爲 0.0684）；至於排除方面，最差的結果（4DFRA, 194）可輕易的看出其不適性。

最後，我們回到最佳解一、二、三。儘管這個檢索原來是要進行相同的比對，但意外的，我們在比對的過程中也將 Docking 可能人選尋找出來了（第一、二解），在事後的研討，由於我們產生的 Spin Image 過程中使用的 H 爲距離而非「距離向量」（即皆爲正數），因此突出的 Inhibitor 與凹陷的 Receptor 所產生的 Spin Image 會是類似的（如果它們有機會接合的話），但這也顯示了我們的做法仍然略嫌粗糙，還有進一步改進的空間。

3.B Receptor to Ligand Docking

由於取得 H 時以距離取得會產生無法辨識其爲凹陷處或凸出處的問題。因此在接下來的實驗中，我們將 H 以距離向量代入來改進程式，並在比較時將 ligand 圖形的每行做 reversion 形成相反圖形來做比對。此次的 docking 我們使用總共十六組 receptor/ligand 配對進行實驗。





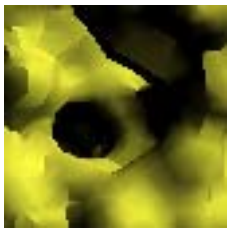
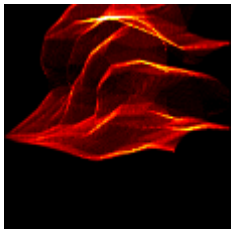
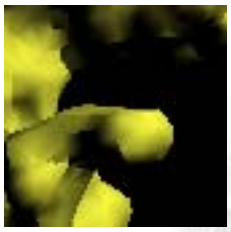
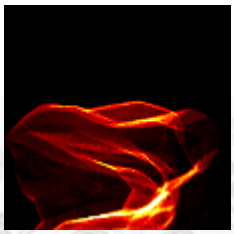
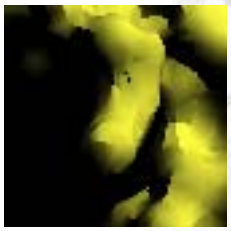
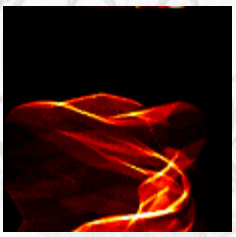
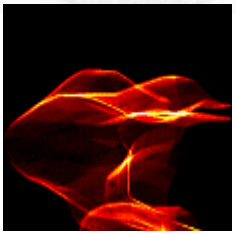
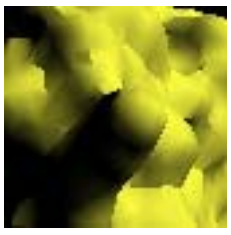
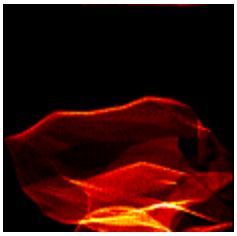
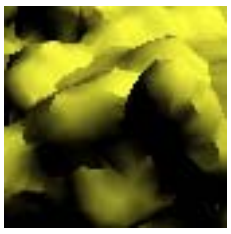
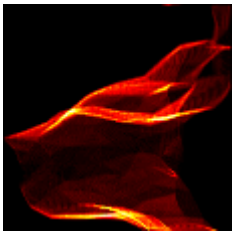
以上依序（由上到下、左至右）爲：1ABIL/1ABIH,1ACBI/1ACBE,1CHOI/1CHOE,1CSEI/1CSEE,
1FDLY/1FDLLH,1TECI/1TECE,1TGSI/1TGSZ,1TPAI/1TPAE,2KAII/2KAIAB,2MHBB/2MHBA,
2PTCI/2PTCE,2SECI/2SECE,2SICI/2SICE,2SNII/2SNIE,2TGPI/2TGPZ,3HFLY/3HFLLH

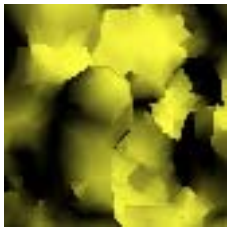
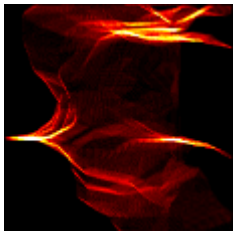
[產生 Spin Image 的時間]

Model Name	Vertices	Faces	Spin Images	Used Time(s)	Avg. Time(s)
1ABIL	3,098	6,192	1,000	15,336	15.34
1ACBI	4,489	8,974	1,000	16,003	16.00
1CHOI	3,715	7,424	1,000	15,792	15.79
1CSEI	4,331	8,658	1,000	18,452	18.45
1FDLY	7,319	14,632	1,000	20,442	20.44
1TECI	4,256	8,504	1,000	15,545	15.55
1TGSI	4,228	8,460	1,000	16,161	16.16
1TPAI	4,331	8,642	1,000	18,002	18.02
2KAII	3,850	7,696	1,000	15,042	15.04
2MHBB	9,968	19,922	1,000	18,644	18.64
2PTCI	4,262	8,502	1,000	15,333	15.33
2SECI	4,418	8,828	1,000	15,010	15.01
2SICI	7,373	14,722	1,000	18,001	18.00
2SNII	4,409	8,814	1,000	20,131	20.13
2TGPI	4,201	8,392	1,000	15,123	15.12
3HFLY	7,440	14,870	1,000	19,337	19.34
Model Name	Vertices	Faces	Spin Images	Used Time(s)	Avg. Time(s)
1ABIH	15,153	30,290	1,000	20,013	20.01
1ACBE	13,803	27,580	1,000	16,571	16.57
1CHOE	12,881	25,730	1,000	15,980	15.98
1CSEE	12,197	24,361	1,000	15,663	15.66
1FDLLH	23,964	47,936	1,000	25,301	25.30
1TECE	12,402	24,782	1,000	20,109	20.11
1TGSZ	12,590	25,146	1,000	19,344	19.34
1TPAE	12,559	25,086	1,000	16,097	16.10
2KAIAB	14,244	28,444	1,000	17,111	17.11
2MHBA	9,541	19,070	1,000	15,042	15.04
2PTCE	12,457	24,886	1,000	15,550	15.55
2SECE	12,062	24,106	1,000	20,196	20.20
2SICE	12,164	24,314	1,000	19,724	19.72
2SNIE	12,673	25,326	1,000	15,221	15.22
2TGPZ	13,297	26,570	1,000	15,404	15.40
3HFLLH	24,306	48,618	1,000	25,472	25.47

前十六組為 ligands 後十六組為 receptors

在此，我們先選擇 1TPAE 做為例子，對所有 ligands 做 Spin Image 的比較，觀察實驗結果，以 receptor 爲主，我們截出其 binding pocket 之 spin image 爲檢索對象（一般而言，binding pocket 爲大的凹陷孔洞或突出處）：

Model, R.P. No.	3D Image	Spin Image	Measure Score	Hit
1TPAE, 870 (Query)			N/A (Best Match)	O
1ABIL, 57			0.2357 (1^{st} candidate)	X
1ABIL, 178			0.2184 (2^{nd} candidate)	X
2PTCI, 190			0.2075 (3^{rd} candidate)	X*
1TPAI, 946			0.2064 (4^{th} candidate)	O
2TGPI, 562			0.2037 (5^{th} candidate)	X

1CSEI, 22			-0.0023 (Worst Match)	X

在 receptor 1TPAE 的實驗中，其相配對的 ligand 爲 1TPAI。上面的結果 1TPAI 出現在第四個位置，出現在前 25%，算是很不錯的結果。值得一提的是 2PTCI（*），雖然不是正確答案，但是從 3D 圖像看來其與 1TPAI 非常相似，我們可認爲其是合理的結果。以下列出十六組 receptor/ligand 之 correlation matrix：

A	B	C	D	E	F	G	H
1ABIL 1ABIH	1ACBI 1ACBE	1CHOI 1CHOE	1CSEI 1CSEE	1FDLY 1FDLLH	1TECI 1TECE	1TGSI 1TGSZ	1TPAI 1TPAE
I	J	K	L	M	N	O	P
2KAII 2KAIAB	2MHBB 2MHBA	2PTCI 2PTCE	2SECI 2SECE	2SICI 2SICE	2SNII 2SNIE	2TGPI 2TGPZ	3HFLY 2HFLLH

理想的狀況中，以下的表格最佳分數應落於對角線上，同爲 A 對 A、B 對 B、C 對 C、……P 對 P。結果如下表，有 56.25% 的最佳解都可以在 Top 3 candidates 中找到。

（R：Receptors、L：Ligands）以 Receptor 爲 Query，Measure Score 爲所有比較最大值。

R\L	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	0.23	0.17	0.19	0.16	0.15	0.20	0.21	0.19	0.17	0.19	0.19	0.19	0.19	0.19	0.19	0.12
B	0.20	0.21	0.19	0.17	0.14	0.19	0.19	0.19	0.19	0.19	0.20	0.19	0.16	0.18	0.18	0.13
C	0.21	0.19	0.19	0.19	0.16	0.16	0.18	0.18	0.20	0.17	0.16	0.17	0.16	0.18	0.20	0.12
D	0.21	0.18	0.18	0.20	0.15	0.16	0.18	0.20	0.20	0.16	0.17	0.20	0.17	0.19	0.19	0.12
E	0.22	0.18	0.20	0.20	0.17	0.17	0.19	0.19	0.19	0.19	0.20	0.19	0.20	0.17	0.19	0.14
F	0.19	0.19	0.19	0.19	0.16	0.20	0.20	0.19	0.15	0.18	0.21	0.18	0.18	0.19	0.18	0.12
G	0.19	0.17	0.19	0.15	0.14	0.18	0.19	0.18	0.16	0.17	0.20	0.17	0.16	0.19	0.19	0.13
H	0.24	0.19	0.18	0.16	0.14	0.16	0.19	0.20	0.15	0.18	0.21	0.16	0.16	0.20	0.18	0.11
I	0.22	0.19	0.18	0.15	0.17	0.16	0.16	0.17	0.20	0.18	0.21	0.18	0.17	0.19	0.19	0.13
J	0.20	0.16	0.17	0.16	0.13	0.17	0.17	0.19	0.15	0.19	0.18	0.19	0.18	0.18	0.17	0.15
K	0.20	0.17	0.19	0.15	0.14	0.18	0.18	0.20	0.16	0.18	0.22	0.18	0.18	0.19	0.19	0.12
L	0.19	0.18	0.17	0.16	0.16	0.18	0.19	0.19	0.18	0.17	0.20	0.18	0.16	0.20	0.18	0.14
M	0.22	0.19	0.19	0.18	0.13	0.16	0.20	0.18	0.17	0.18	0.19	0.18	0.16	0.19	0.19	0.13
N	0.23	0.20	0.18	0.17	0.14	0.16	0.19	0.20	0.18	0.18	0.20	0.18	0.18	0.17	0.21	0.13
O	0.22	0.19	0.19	0.18	0.12	0.18	0.17	0.19	0.15	0.16	0.20	0.19	0.18	0.19	0.20	0.12
P	0.19	0.17	0.19	0.15	0.16	0.17	0.18	0.17	0.18	0.18	0.15	0.17	0.19	0.17	0.17	0.14

（黃色、綠色、藍綠：Top 1、2、3 candidate）

由上表可知，在前三名的比對結果之內，能夠藉由 Spin Image 找出相配對之 ligand 爲 56.25%，而若將範圍提高到前五名，幾乎所有的 ligand 都能夠被找到。值得一提的是，點數太多的 ligand 如：1FDLY、2MHBB、3HFLY，表現得都不好，可能是因爲取樣不足，ligand 的活動基沒有被取到，造成無法與配合的 receptor 配對成功。而最小的 1ABIL 看起來形狀最怪異，但卻意外的所有分數都很高，也許是因爲表面變化較多，因而與不同的 receptor 都能夠有不錯的比較結果。

而在 Spin Image 的產生方面，由於這次實驗的 Spin Image 仍然是 200x200 的規模，大部分產生時間皆與 Elephant 模型差不多，約在 10~20 秒之間，模組的大小與產生的時間雖然有相關但是成長得並不快，也驗證了影響我們演算法的主要元素還是 Spin Image 的規模。

Chapter 4

Conclusion and Future works

本次的實驗中未能夠建構完整的資料庫，算是比較可惜的地方。但 Spin Image 在資料的比對上確實既快速又簡單。尚在研發中的精確比對或接合的方法（如：ZDOCK）在一個結構上約需要九小時的運算時間（300 取樣結構；2.2GHz, Linux），相對於我們的方法約兩分鐘（1.4GHz, WindowsXP, JVM1.6），可說是在前導的處理上，節省不少的時間。

系統本身在建構蛋白質全體影像會花較多的時間，一個一萬網點的蛋白質全體影像約需要 24 小時的計算時間（1.4GHz, WindowsXP, JVM1.6），在上面的實驗中我們做了隨機的取樣，嘗試在短時間內獲得可接受的實驗結果。實際上的測試中，藉由觀察到 C、E、J、L、M、N、P 組的結果，超過 7,000 個面之 ligands（總數共四個）皆未在前三名的結果中比對到正確的 receptors，3,000 個面至 7,000 個面之 ligands（總數共十二個）則只有三組未得到正確結果，因此，我們可以發現取樣在四分之一以上，應是不過於影響結果的極限。而在 Spin Image 生成方面，若能加入 GPU 的幫助，甚而進行平行化的運算，相信可以有更長足的效能增進。

另外，在取圖比例上，由於我們處理的蛋白質組成元素大小固定，因此取圖比例過大並不會增加其精確度，而找到最適比例，以期在不影響其精確度的情況下減少產生與比對 Spin Image 的時間，可以是一個增進效率的部份。

在未來的工作上，建立起整個蛋白質資料庫會是一個龐大但必要的工程，另外在 Binding sites 的預測上，應該可加入電子電位圖的概念，以期能夠在篩選結果之中增加建議的參考價值。

Chapter 5

Reference

[J05] Jeng-Sheng Yeh: 3D Protein Retrieval Based on Pocket Modeling and Matching, Ph.D. dissertation, CSIE, National Taiwan University, Taipei, Taiwan, 2005.

[M83] Michael L. Connolly: Solvent Accessible Surfaces, http://www.netsci.org/Science/Compchem/feature14e.html (2008/03/20), 1983.

[TRJ06] Thomas A. Funkhouser, Roman A. Laskowski, and Janet M. Thornton: Protein Function Prediction by Matching Volumetric Models of Active Sites, Automated Function Prediction Meeting (AFP), San Diego, CA, August 2006.

[RGAD06] Ruth Huey, Garrett M. Morris, Arthur J. Olson, David S. Goodsell: A semiempirical free energy force field with charge-based desolvation, Journal of Computational Chemistry, 28, 1145-1152, 2006.

[CLW03] Chen R, Li L, Weng Z: ZDOCK: An Initial-stage Protein-Docking Algorithm, Proteins 52, 80-87, 2003.

[LCW03] Li L, Chen R (joint first authors), Weng Z: RDOCK: Refinement of Rigid-body Protein Docking Predictions. Proteins 53, 693-707, 2003.

[JC04] J.-M. Yang and C.-C. Chen: GEMDOCK: A generic evolutionary method for molecular docking, Proteins: Structure, Function and Bioinformatics, 55, 288-304, 2004.

[DJJJJWA06] Deepak Bandyopadhyay, Jun Huan, Jinze Liu, Jan Prins, Jack Snoeyink, Wei Wang, Alexander Tropsha: Structure-based function inference using protein family-specific fingerprints, 2006.

[AM99] Andrew E. Johnson and Martial Hebert: Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes, IEEE Transactions on pattern analysis and machine intelligence, 21, No. 5, 1999.

[AM98] Andrew Edie Johnson and Martial Hebert: Surface Matching for Object Recognition in Complex 3-D Scenes, 1998.

[DXYM03] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen and Ming Ouhyoung: On Visual Similarity Based 3D Model Retrieval, *Computer Graphics Forum (Eurographics 2003 Conference Proceedings)*, 22, No. 3, 223-232, 2003.

[PG00] P. Lindstrom and G. Turk: “Image-Driven Simplification”, ACM Transactions on Graphics, 19, No. 3, 204-241, July 2000.