針對高可靠度蛋白質配體嵌合之最佳化演算法特性分析

研究計畫之背景及目的

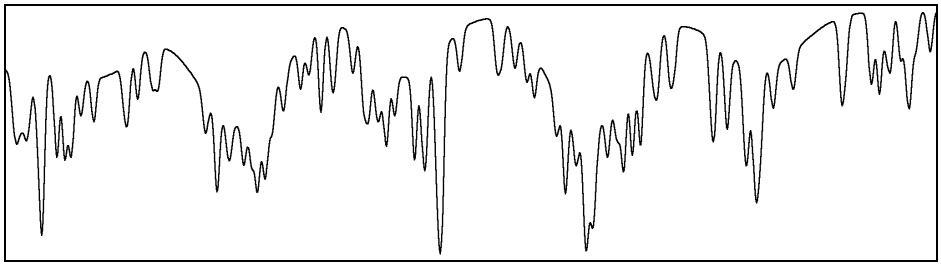
近年來，由於蛋白質資料庫爆炸性且持續性的成長，與蛋白質相關的資訊不斷地被解析出來，其中包含了大量的蛋白質結構，帶來了生物資訊領域於蛋白質結構分析的需求。多年來，利用分析蛋白質分子之間的交互作用來了解進而預測蛋白質功能一直被視為是生物學家與資訊學家共同努力的方向。其中，蛋白質與配體嵌合(protein-ligand docking)工具在電腦藥物輔助設計扮演了很吃重的角色；著名的蛋白質嵌合軟體有 AutoDock、FlexX、以及 DOCK。蛋白質與配體嵌合就是分析分子間的物化特性，並模擬分子間的交互作用得到最符合自然界定律的嵌合形態。蛋白質與配體嵌合方法經過長久的研究及實驗，其產生的預測嵌合形態有一定的參考價值，至少在虛擬藥物篩選的應用上，已經具備相當的成熟度。

此外，雖然蛋白質與配體嵌合已經被廣泛使用於虛擬藥物篩選上，但是還是有其侷限之處。主要原因可歸納為兩個主要的因素：其一為自然界物化特性的分子作用，在蛋白質與配體嵌合的演算法中通常會形成起伏非常強烈的能量函數，使得這類型的演算法有很大的機會產生看似正確的陷阱形態(decoy state)，但是卻不是真正的天然形態(native state)；其二則是當考慮配體的變動性(flexibility)時，將大幅提高演算法本身的維度，使得整體蛋白質與配體嵌合問題的困難度也大幅提升。蛋白質與配體嵌合碰到的陷阱形態問題，主要的關鍵在於從物化特性的角度來看，這些陷阱形態其實有時候也是非常穩定的，只是往往在自然界中只有最穩定的形態，也就是天然形態，才能保存下來。而演算法維度過高的問題，由於配體在自然界中就是具有變動性，傳統的演算法將其視為鋼體(rigid body)來避免高維度的作法，可以說是在軟硬體不足的環境下逼不得已的解決之道，未來擔任蛋白質與配體嵌合核心的最佳化演算法(optimization algorithm)，勢必得應付高維度的搜尋空間，才能真正解決蛋白質與配體嵌合的問題。以下我們先分別介紹這兩個問題與蛋白質與配體嵌合的關聯性，並分別討論各自對於蛋白質與配體嵌合預測所造成的問題；之後再提出我們如何利用現有研究方法同時解決這兩個問題。

一、蛋白質與配體互動能量函式的高起伏特性

一般認為，分子之間的作用力受到許多物化特性的影響，雖然例如靜電力、氫鍵、以及極性等等個別的作用力計算已經被公式化，但是同時考慮多個作用力時之間彼此又會互相影響，使得這些物化特性對於最終作用力的影響程度一直到今天都尚未存在一個統一的標準。雖然在能量函式的發展上仍然有很大的進步空間，但是從許多實驗數據可以看出，這些作用力之間複雜的交互影響，是使得蛋白質與配體互動能量函式具有高起伏特性的重要原因。

圖一是一具有高起伏特性的函式，從能量函式的觀點來看，函式的最低點就是最穩定的狀態，對就是蛋白質與配體嵌合的天然形態。從圖一不難發現，高起伏性的函式具有許多區域最低值(local minimum)，這些區域最低值就是之前提到的陷阱形態。實際上在解決蛋白質與配體嵌合上的問題時，就是將最佳化演算法套用在蛋白質與配體的互動能量函式上，對一般最佳化演算法來說，在許多的區域最低值當中，如何找到真正的全域最低值(global minimum)而不要陷入其他區域最低值將會是一大挑戰，也就是說，傳統的最佳化演算法在處理蛋白質與配體嵌合的問題時，必須小心處理這種高起伏的特性。



圖一、典型具備高起伏特性的函式

多數最佳化演算法在尋找全域最低值的過程中，都會遇到區域最低值帶造成的問題。以傳統基因演算法(genetic algorithm)為例，此演算法雖然利用突變(mutation)效應可以跳脫區域最低值，但是也可以因為過高的突變機率，破壞整個演算法的平衡性，導致演算法的效能跟純亂數的結果差不多。突變效應之所以會造成問題，是由於在一般演化式演算法(evolutionary algorithm)中，突變本身所扮演的角色就是盡量不去利用之前已經取得過的資訊，這樣的設計雖然使得演算法得以跳脫之前陷入的區域最低值的窘境，但是因為每次的突變過程等於是由亂數決定，而非從之前的經驗學習而已，造成資訊累積的速度變慢，連帶會影響最終演算法的效率。

二、配體變動性

隨著硬體以及計算能力的進步，現在配體的變動性已經逐漸變成蛋白質與配體嵌合時必備的條件，傳統將配體視為鋼體的策略已經不能滿足科學家對於模擬真實情況的要求。考慮配體變動性使得嵌合過程更貼近真實自然界的情況。考慮配體的變動性，實際上在配體分子中化學鍵可以任意旋轉的情況下，進行嵌合的模擬，當配體分子多一個可以旋轉的鍵結時，就表示在蛋白質與配體嵌合問題的搜尋空間中多了一個維度，這些維度讓嵌合過程具有更大的彈性，最終的嵌合形態更貼近自然界的情況，但是也同時增加了最佳化演算法的困難度。

目前蛋白質於配體嵌合最主要的應用是虛擬藥物設計，在虛擬藥物設計上，最主要的配體是藥物小分子，基於合成與人體吸收上的考量，現階段的配體分子大小還侷限在一定的範圍以內，從演算法的角度來看，現階段蛋白質與配體嵌合所要面對的搜尋空間，維度大約在六到二十維左右，這也是目前傳統基因演算法所以處理的最大極限。如果能夠提高這個限制，除了可以針對大分子配體進行嵌合模擬，甚至可以考慮蛋白質的變動性，這也是科學家一直以來希望破除的限制。

集合上述兩個問題，我們認知了最佳化演算法在蛋白質與配體嵌合預測上所扮演的重要性。但在本主持人之前的研究經驗中，我們認為現有的最佳化演算法在試圖解決上述兩個問題時，由於無尚未能了解兩個問題本身的特質跟最佳化演算法之間的關聯，而無法找到適當的解決方案。我們認為，這兩個問題在舊有基因演算法的架構上有相當的關聯性，且彼此相互影響，如何找到其中的平衡點是非常困難的事，但是如果能夠發展全新的最佳化機制，或許可以從不一樣的角度去解決問題，這也將是本計畫的重點之一。

本主持人於博士論文中，延續之前於適應性突變演化式演算法的研究，針對蛋白質與配體嵌合提出利用核心密度估計技術改進演化式演算法模型的方法。此核心密度估計技術可應用於分析演化式演算法中母體的分佈情形；這種分析方法不僅提供演化式演算法更完整的分析模型，全面地分析架構亦使得母體落入陷阱形態的機率大幅降低。我們之前利用這個適應性突變演化式演算法有

效地提升許多真實蛋白質與配體嵌合的模擬過程，並大幅提升最終產出天然形態的可靠度，在本計畫中，我們將善用此適應性突變演化式演算法所提出的架構，來做為新的改善方法的基礎。

為解決目前蛋白質與配體嵌合所面臨的挑戰，本計畫希望以核心密度估計為基礎的演化式演算法為基本架構，進一步研究幾個有興趣的問題：
第一個問題是如何利用母體密度分佈的資訊中，正確選擇出最適當的下一代；第二個問題則是進一步母體密度分佈方法本身是否仍有提升的空間；第三個本計畫有興趣的問題是希望把這個最佳化演算法應用到不同領域，現實生活中有許多問題，其資料特性都可以輕易地轉化為最佳化的問題，本計畫之研究將對這一類的問題提供適當的解決方案。

本計畫預計一年完成，預計完的目標如下：
首先根據我們現有的最佳化演算法，配合一些其他我們認為相關的技術，例如單變數分析(univariate analysis)以及直交實驗設計(orthogonal experimental design)，進一步提升適應性突變演化式演算法的效能，進而提升蛋白質與配體嵌合的模擬速度與可靠度。在計畫進行中，我們將同時提供蛋白質與配體嵌合的服務，其預測結果可提供生物學家分析有興趣的蛋白質與配體之間的交互作用。本研究計畫希望利用過去在核心密度估計所累積的基礎與經驗，試圖解決蛋白質與配體嵌合所面臨的困難，同時藉此探討生物分子之間的關係，希望能幫助生物學家更進一步發現與了解重要的生物機能。

研究方法、進行步驟及執行進度

蛋白質與配體嵌合多年來一直是生物學家與資訊學家共同努力的方向。而最佳化演算法在其中也扮演了關鍵性的角色，好的最佳化演算法可以降低陷阱形態的影響，快速地在搜尋空間中找到天然形態的位置，達到良好的嵌合模擬結果。

本研究計畫希望藉由兩個步驟來改進最佳化演算法，其中步驟一是利用單變數分析，將原本分析母體分佈情形的核心密度估計技術，由原本的以個體為基礎，進一步細分為以維度為基礎進行分析，步驟二則利用直交實驗設計並針對適應性突變演化式演算法中比較不成熟的區域最佳化(local optimization)部分進行補強的動作。以下分別詳述演算法步驟。

步驟一、利用單變數分析進行以維度為基礎的母體解析

於步驟一中，我們將利用單變數分析，並搭配一個統計指標來調整適應性突變演化式演算法在各維度母體分佈的情況。變異數是統計學上常用的評量方法，常用來評量一群資料的變動程度。對於母體 P，其變異數 $Var(P)$的定義如下：

$$Var(P) = \frac{1}{|P|}\sum_{x_i \in P}(x_i - \bar{x})^2$$

其中

$$\bar{x} = \frac{1}{|P|}\sum_{x_i \in P} x_i$$

而$|P|$則表示母體 P 中個體的數目。我們將利用這個統計指標來調整下一代母體的分佈。由於一次僅對一維採用單變數分析，所以每個維度將的得到不同的變異數，變動程度較大的維度，可以視

為在該維度上母體分佈非常雜亂，而且尚未找到在該維度上比較好的值；反之在變動程度比較小的情況下，可以想成該維度在母體不斷的演進之下，已經可以辨識出比較好的值域。在這樣的假設之下，應該讓下一代盡量產生在靠近該優良值域附近，在適應性突變演化式演算法的架構中，就是減小該維度的最小擾動係數(α)值。相反的，在變動程度比較大的維度則維持比較大的最小擾動係數。

步驟二、利用直交實驗設計改進區域最佳化演算法

現今絕大部分的演化式演算法都有一個共識，那就是必須針對區域最佳化設計專屬的演算法，而不是直接使用全域最佳化演算法來處理區域最佳化的問題。因應這樣的架構產生了許多區域最佳化演算法例如坡度遞降法(gradient decent)、最陡峭爬山法(steepest-ascent hill climbing)、以及 Solis 跟 Wets 提出的局部搜尋算子等等區域最佳化演算法，結合全域最佳化以及區域最佳化的概念有人稱之為混合策略瀰集進化演算法(memetic algorithm)。這個部分的研究重點將放在專為適應性突變演化式演算法設計一套合適的區域最佳化機制，以彌補之前適應性突變演化式演算法在這方面的不足。在此研究計畫中，我們希望藉由探討適應性突變演化式演算法的特性，以及分析其他區域最佳化演算法的優劣，找出最適合甚至發展全新的區域最佳化演算法，最終能增加適應性突變演化式演算法的整體效能。

在步驟二中，我們將利用直交實驗設計來分析區域的群體，希望能快速找到區域中最優秀的個體並產生較以往更為優秀的下一代。直交表示平衡且不混合，就是統計上的獨立，也就是說直交實驗設計產生的每一個維度中，各個水準(值)出現的次數是相同的。針對一組設計好的因素，所有可能的狀態有 L^n 種，其中 n 為維度，L 為每個因素(維度)的水準數目。使用傳統完全因素實驗需要進行 L^n 次實驗，然而透過直交實驗設計，僅需進行 $L^{\lceil \log_L(n+1) \rceil}$ 次實驗即可獲得近似最佳解(near optimum)，大量節省計算的時間。

因素分析可以評估在評估函數中的因素效果，排名最有效果的因素以及決定各因素最好的水準組合，進而促使評估函數有最佳的結果。直交實驗設計可以縮減因素分析的實驗次數。直交表的實驗次數在單一因素分析時只需 $M = L^{\lceil \log_L(n+1) \rceil}$ 次。假設 y_t 是第 t 次實驗的函數評估值，我們定義 S_{jk} 是 j 因素在水準 k 的主效果：

$$S_{jk} = \sum_{t=1}^{M} y_t \cdot W_t ,$$

其中 W_t 為一個旗標值，若第 t 次實驗中第 j 個因素選用水準為 k，則 W_t 為 1；若否，則 W_t 為 0。若適應函數較大，則較大的主效果值表示對適應函數具有較佳的貢獻度；反之若適應函數較小，則主效果值小者貢獻度較佳。主效果可以顯示因素中水準的各別影響。例如主效果 $S_{j1} > S_{j2}$ 則表示在參數最佳化的問題中，第 j 個因素水準值 1 對於整體最佳化函數的貢獻大於水準值 2。如果相反的情形 $S_{j1} < S_{j2}$，則表示水準值 2 較佳。在各因素間無交互作用的前提下，主效果的分析可以用來推測出全實驗的最佳解。直交因素實驗為一種部分因素實驗方式，可以有效地減少參數設計時的實驗次數，並同時考量實驗因素之間的交互作用。將直交因素實驗後的數據經過主效果分析，便可以將每個維度對於母體的貢獻優劣計算出來，推論在區域中最佳的個體。

以上是本計畫的研究計畫，計畫前半年將針對適應性突變演化式演算法的架構提出調整及改進的方法，後半年則是探討此調整所帶來的影響，並尋找以及開發最合適的區域最佳化演算法來配合

突變演化式演算法並提升整體的效能。本計畫未來的延續研究目標是希望將我們於蛋白質與配體嵌合的研究經驗應用到其他生物資料上。一般來說，生物資訊應用上都會面臨到高維度以及高起伏的特性，所以最佳化演算法能否克服這兩大困難是一個最基本要解決的問題。本研究所針對蛋白質與配體嵌合的特性所設計的演算法，更可應用到許多其他生物資料上，比如說，以微陣列資料進行相關基因偵測的問題，現在也往往同時考量多個基因的趨勢，大大提升了分析時的維度，本計畫所開發的演算法於其他資料的適用性，也是開發演算法的同時，必須審慎思考的問題。

提供網際網路服務：蛋白質與配體嵌合

除了研發演算法，我們計畫提供更為完善的線上服務，幫使用者方便地模擬蛋白質與配體嵌合的過程。此網頁服務將搭配我們所建構的最佳化演算法，將有助於提升服務的回應時間。

預期完成之工作項目及成果

本計畫預計一年完成，前半年將計畫重點放在適應性突變演化式演算法的調整及改進上，我們將檢視現在架構的不足，試圖提出更完善的模型來解決之前的效能瓶頸。後半年的計畫重點則放在研發全新的區域最佳化演算法，來面對生醫問題特有的高維度與高起伏的挑戰，並與現有的適應性突變演化式演算法整合，使整個系統滿足蛋白質與配體嵌合之需求。本計畫預計完成的成果如下：

1. 利用單變數分析發展以維度為基礎的分析架構；
2. 探討生醫問題專有的特性並開發最適當的演算法來配合這樣的特性；
3. 提供網頁分析工作，幫助生物學家模擬生化反應。

本計畫的研究成果預期將提供蛋白質與配體嵌合相關的幾個問題一個整合的空間，將有助於蛋白質與配體嵌合的知識管理。本計畫的成果將可應用於解決其他有類似問題的應用中，進而提升各個領域的研究效率及成果。