

# 行政院國家科學委員會專題研究計畫 成果報告

## 研發人體免疫缺失病毒治療的最佳化藥物配方篩選系統 研究成果報告(精簡版)

計畫類別：個別型  
計畫編號：NSC 100-2221-E-009-143-  
執行期間：100年08月01日至101年07月31日  
執行單位：國立交通大學生物科技學系(所)

計畫主持人：何信瑩

報告附件：出席國際會議研究心得報告及發表論文

公開資訊：本計畫可公開查詢

中華民國 101 年 10 月 30 日

中文摘要：本計畫為「研發人體免疫缺失病毒治療的最佳化藥物配方篩選系統」之一年期研究計畫，主要目的是發展一套可依據病人生理特性，協助醫生決定使用人體免疫缺失病毒(HIV)治療的最佳藥物配方之電腦篩選系統。本篩選系統需建立三個藥物治療模型，分別是雞尾酒療法、進入抑制劑及成熟抑制劑。藥物治療模型之研發程序分為三個重要步驟：(1)蒐集並建立病毒序列和病人生理表現之資料庫，以便於擷取作為HIV個人化醫療的生物醫學特徵；(2)整合HIV病毒序列特性與生物醫學特徵來設計藥物治療模型，並將模型設計轉換成數個參數最佳化問題；(3)發展一套以高效能繼承式雙目標基因演算法(IBC GA)作為核心的建模與預測演算法。

本最佳化藥物配方篩選系統主要分為三個子系統，各子系統間相輔相成，簡述如下：

1. 病毒序列及病患生理現象的資料庫，並將相對應之基因型和生理型資料整合。
2. 雞尾酒療法、進入抑制劑及成熟抑制劑等三個整合的藥物治療模型。
3. 個人化醫療的決策系統，藉由利用HIV病毒與生理現象資料庫及三個藥物治療模型而來。

我們結合IBC GA與支持向量機分類器，已建立進入抑制劑之藥物治療的初步模型與病毒序列資料庫，並驗證此模型能有效地進行藥物篩選。此成果與經驗可幫助另二個藥物篩選模型的建立，進一步加速整合篩選系統的建立與驗證。本年度已建造序列資料庫與建立完整進入抑制劑及成熟抑制劑的藥物治療模型為主要目標，並於後續研究中整合完整的藥物篩選系統。

中文關鍵詞：人體免疫缺失病毒、雞尾酒療法、個人化醫療、基因演算法、數學建模、參數最佳化、藥物配方篩選系統。

英文摘要：This is a one-year project: Developing an Optimized Recipe Selection System of HIV Therapy. The objective is to develop a recipe selection system of HIV therapy according to physiological properties of patients to help doctors decide the medicine usage. The recipe selection system needs to establish three medicine therapy models, HAART, entry inhibitor, and maturation inhibitor. The procedure of developing the therapy models consists of the three important steps: 1) establish databases of HIV sequences and physiological properties of patients to extract potentially medical features for individualized HIV

therapy ; 2) integrate characteristics of HIV sequences and medical features to design the therapy models, and transform the model design into a number of parameter optimization problems, and 3) develop a set of modeling and prediction algorithms based on a high-performance inheritable bi-objective optimization genetic algorithm (IBCGA) to solve these optimization problems.

The optimized recipe selection system of HIV therapy mainly consists of three subsystems which are fully cooperated, described below.

1) The databases of integrating HIV sequences and physiological properties of patients.

2) Three integrated therapy models of HAART, entry inhibitor, and maturation inhibitor.

3) The decision system of individualized HIV therapy by utilizing the three therapy models and databases of HIV and physiological properties.

We have established a prototype of the ‘entry inhibitor’ therapy model and a database of HIV sequences based on IBCGA and a support vector machine based classifier. The prototype model has been verified that can screen out effective recipes. The achievement and experience can help we design and establish the other two therapy models of HAART and maturation inhibitor. Therefore, the optimized recipe selection system can be fast established and validated.

英文關鍵詞： HIV、HAART、Individualized Medicine、Genetic Algorithm、Modeling、Parameters Optimization。

# 前言

人體免疫缺失病毒(HIV)的防治與愛滋病的治療在流行病學中一直都是重要的議題，但由於病毒本身的快速變異，使得病毒在體內的致病性、致毒性、感染力等特性都會隨著改變。此外在病人的治療的過程中，病毒的變異也會產生抗藥性，使得必須改變治療的藥物，所以一個能快速並正確的判斷何種藥物適用於目前病毒的藥物篩選系統是非常重要的[1]。愛滋病的早期治療是以阻止病毒在細胞內的反轉錄與降低相關酵素的活性為主，因而有雞尾酒療法(Highly Active Antiretroviral Therapy)的產生。但近年來對於病毒的防治除了現有的治療法外，還有兩方面備受重視，一個是抑制病毒進入細胞的進入抑制劑，另一個是使病毒蛋白質無法摺疊成功能性蛋白的成熟抑制劑。這兩種抑制劑都已被實驗證實等有效的輔助 HAART 的治療，因此本計畫將以這兩種抑制劑為主，建立快速並具有高準確率的藥物篩選模型，以達到輔助治療與個人化醫療的目的。

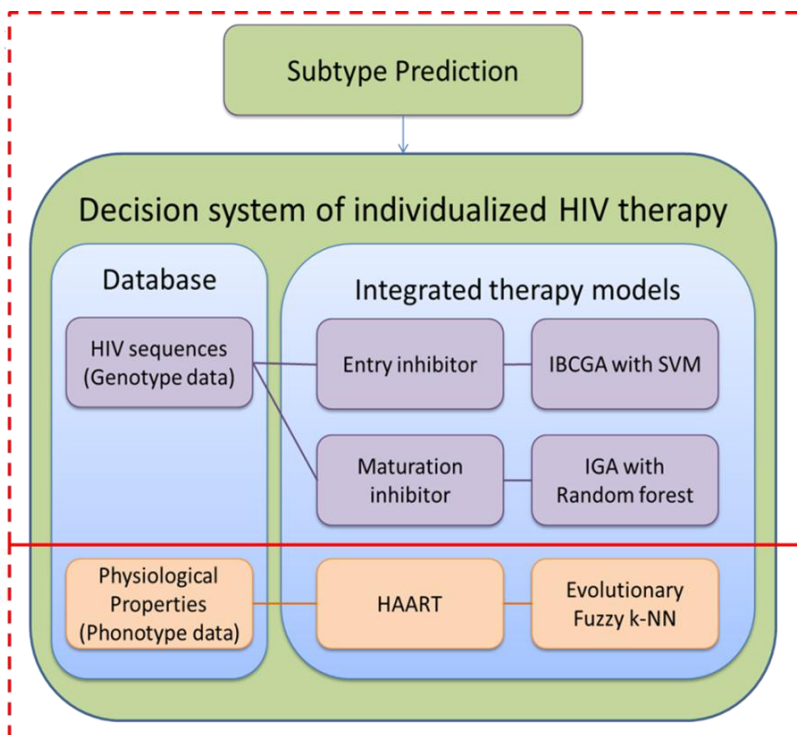
## 研究目的

本計畫利用過去多年所開發最佳化演算法之技術，將研究成果擴展至人體免疫缺失病毒防治之研究，利用繼承式雙目標基因演算法結合分類器，並整合不同的病毒序列特徵，發展一套人體免疫缺失病毒治療的最佳化藥物配方篩選系統。除了利用生物資訊的方法建立最佳化藥物配方篩選系統外，建立人體免疫缺失病毒資料庫也是本計畫目標之一，藉由建立病毒資料庫可以建構各亞型之專屬篩選系統、依照不同地區病例修正系統並完成個人化醫療之目的。

本年度計畫預計分為完成內容如下，A、建立病毒序列資料庫。B、依病毒不同複製時期分別建立藥物篩選模型。而在本年度計畫完成後預計會再將此技術應用到 HAART 的藥物篩選上，並結合已完成的藥物篩選模型，已完成個人化的最佳化藥物配方篩選系統。圖一為本計畫研究架構圖，包含本年度預計完成目標及未來預定工作。

本年度預計完成目標

未來預定工作



圖一、計畫研究架構圖。包含本年度預計完成目標和後續研究預定完成目標。

## 文獻探討

在人體免疫缺失病毒治療的問題中，分別對於病毒感染的三個時期作分析，並判斷所使用之藥物。

病毒感染的三個時期包括融合期、反轉錄期和活化複製期，由於本計畫主要針對融合期與活化複製期作為研究的目標，因此以下便對這兩個時期的研究、以及病毒亞型的相關研究及治療方法相關研究的情況作介紹。

## 不同亞型差異之相關研究

在許多研究指出，病毒亞型對於病毒感染的機制會有差異，並會影響病毒體中酵素的活性，更進一步使藥物判斷的準確度降低。在近期的研究中，主要是以人體免疫缺失病毒基因中 env 和 gag 區域中的 RNA 序列作為不同亞型的依據。像 Myers 的研究[2]便是利用 env 區域的基因組來做分類，並可以將 A、B、C、D、F 等亞型分類。而在 Louwagie 研究[3]中分析 gag 區域的 RNA 基因，更加強了分類能力，使得除了 E 亞型外大部分的亞型都能被分類。不同亞型對於病毒酵素的影響也是生物學家重要的研究課題之一，目前已有不少研究亞型對於反轉錄酶、蛋白酶及整合酶的相關研究[4-7]。

## 病毒融合期防治之相關研究

病毒融合期的防治主要為阻止病毒的膜上糖蛋白 gp120 和目標細胞結合為主。Gp120 會和目標細胞上的 CD4 受體和趨化素協同受體連結，並產生型態變化，使人體免疫缺失病毒能附著於目標細胞膜上。目前研究主要是利用 V3 環狀序列作為判斷連接何種協同受體的主要依據[8]，並且藉由協同受體的預測，可精確使用藥物來阻止病毒進入細胞，進而抑制人類免疫缺失病毒的細胞感染及擴散[9]。除了早期的 11/25 電荷規則[10]、位置加權矩陣(Position Specific Score Matrix, PSSM)[8]和演化式類神經網路(Evolved neural networks)[11]等方法來處理病毒的建模及預測外，近年來主要是以利用支援向量機器(Support vector machines)結合不同的序列特徵值，如胺基酸組成、結構特性等作為預測的方法，並且都有不錯的結果[9, 12]。

## 活化複製期之相關研究

而除了進入抑制劑外，成熟抑制劑也成為了輔助治療愛滋病的研究重點之一[13, 14]。成熟抑制劑在近年的研究中證實可以有效的輔助雞尾酒療法[14]，但是成熟抑制劑的療效會受到病毒本身抗藥性的影響，而組成病毒衣殼(capsid)的 p24 序列為影響抗藥性的主要片段，所以此序列對判斷病毒抗藥性對於在治療中是否加入成熟抑制劑有很大的決定作用[13]。

## 研究方法

### 資料蒐集及分類

本計畫中預計從兩方面著手收集人體免疫缺失病毒的相關資料，包括人體免疫缺失病毒的序列資料以及受感染病人之生理表現資料。在病毒的序列資料方面，目前的人體免疫缺失病毒研究中，最常被使用的資料庫為 Los Alamos National Laboratory HIV Sequence Database (<http://www.hiv.lanl.gov/>)，本計畫所使用的大部分病毒資料即從此資料庫獲得。此外也有從 GenBank[15]、Trofile phenotypic assay[16]以及其他研究論文[17-21]等來源蒐集相關的序列資料。收集到資料後做系統化的整理，依照 HIV 病毒的基因組分類，並把建立藥物篩選系統所需要的序列部分分別建成資料組，以供後續實驗使用。

### 建立藥物篩選模型

本計劃的核心為設計優良的藥物篩選模型來分析病毒序列資料，並結合多個藥物篩選系統與預測器，判斷出最適合的藥物組合，於是本計畫提出了以繼承式雙目標基因演算法為基礎，配合病毒中不

同蛋白質序列來結合分類器，建立準確的藥物篩選模型。本計畫中預計建立病毒亞型預測模型和兩組藥物篩選模型，其中兩組藥物篩選模型分別為融合期藥物篩選模型和成熟抑制劑抗性預測模型。繼承式雙目標基因演算法及分類器說明如下。

## 繼承式雙目標基因演算法

當使預測準確度最大化時而選擇小數量的有用特徵是雙目標 0/1 組合最佳化的問題。一個有效的繼承式雙目標基因演算法(IBC GA)[22]可被用來解決上述的最佳化問題。IBC GA包含了智慧型基因演算法和繼承的機制。智慧型基因演算法使用分割與合併的策略，並且透過直交表設計實驗來解決大量參數最佳化的問題，具有收斂速度快、精確度高的優點[23]。

智慧型基因演算法可有效地探測以及發現 $C(n,r)$ 的搜尋空間，而IBC GA更可透過繼承機制從 $C(n,r)$ 空間上的解，來有效地解決 $C(n,r\pm 1)$ 搜尋空間的問題[24]。因此IBC GA可以在每一次循環中對不同的解空間獲得一個完整而且高品質的解集合， $r$ 的範圍會隨著所使用的分類器或特性質的不同而改變[25]。以結合支持向量機器(SVM)為例，IBC GA的染色體編碼設計包含了兩個部分：二元基因與參數基因。二元基因被用來做為特徵值選擇，數量隨著所使用的特徵不同而改變，而參數基因用來調整支持向量機器的參數。本計畫中便依照藥物篩選系統所使用的不同特徵而有不同的二元基因組合，當使用胺基酸物化特性資料庫(AAindex)[26]中的特徵值時，則包含了 $n=531$ 個二元基因，但使用序列胺基酸組成(amino acid composition)時，則只有 $n=400$ 個二元基因組合。而結合SVM時，會根據所選擇的核心函數(kernel function)不同而有不同的參數基因，一般常用的SVM參數為 $C$ 和 $\gamma$ ， $C$ 和 $\gamma$ 的編碼在本計畫中均為16個值，分別為 $[2^{-7}, 2^{-6}, \dots, 2^8]$ 。

## 支持向量機器(Support Vector Machine, SVM)

支持向量機器(SVM)是用來處理二元分類問題的機器學習模型，SVM是找出能將兩類別資料分開最大距離的超平面(hyperplane)，並利用此超平面建立二元的分類器。當處理多類別的問題時，SVM會將所有類別視為個別的二元分類問題：建立子分類器將資料轉換成只有兩類的情況，個別處理子分類器並利用投票的方式獲得最後的結果。SVM提供許多不同的核心函數(kernel function)用於將資料轉換成高緯度搜尋空間，以能夠找到適當的超平面來區分資料，如本計畫中所使用的非線性轉換核心函數RBF(radial basis function)，其定義如下：

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0$$

核心函數參數 $\gamma$ 決定樣本是如何轉換成高維度搜尋空間，成本參數 $C > 0$ 用來調整分類錯誤的損失程度。調整 $C$ 和 $\gamma$ 兩個參數可用來決定一組擁有最佳的預測效率的分類器。支持向量機器的優點在於不容易對一個問題過度適應化(over fitting)。因為支持向量機器旨在找出一個能使兩類別資料分開最大距離的超平面，並依此可避免過適化之問題，並減少在預測資料時的錯誤。

## 模糊分類器與模糊 K-NN 分類器

目前的模糊系統大都是處理變數較少的問題，對於多變數高緯度分類器的問題，仍難以有效的解決，因為參數的增加會導致模糊規則數量的增加，尤其是大量參數時規則的數量更是成指數成長(curse of dimensionality)。因此在設計模糊系統分類器時，傳統的錯誤嘗試法和專家知識設計法難以達成模糊系統的最佳化設計，而基因演算法在近年已被證實非常適合設計高效能的模糊系統。

吾人過去成功應用了智慧型基因演算法來設計模糊分類器解決不同的生物資訊問題[27, 28]，模糊分類器利用模糊規則來對不同特徵作分類。由於其方法為以規則為基礎之方法(rule-based method)，因此最後分類結果將具有可解讀性[29, 30]。

K-NN為K-近鄰分類器(k-nearest neighbor, K-NN)，主要是根據同類型的資料應該會擁有類似的參

數與特性並會聚集在一起。傳統KNN常用的距離函數是歐基里德距離(Euclidean length)，讓資料提升至高度空間並以其中的點來表示，則這些點可以獲得較近的距離。模糊K-NN分類器(fuzzy k-nearest neighbor, 簡稱FKNN)是在傳統K-NN分類器中加入了模糊理論，以研究具有模糊現象存在的事實，並根據不清晰訊息，透過近似推理過程而得到正確結果。FKNN會取最接近測試資料的k個訓練資料點，並藉由這些訓練資料之歸屬程度的距離來決定測試資料的類別，而即使屬於同一類別的資料也會依據歸屬函數 (membership function)而有不同的歸屬程度，將其歸屬程度從0和1擴展到介於0~1之間之實數值。

## 結果與討論

本計畫研發人體免疫缺失病毒治療的最佳化藥物配方篩選系統先建立病毒序列資料庫，再藉由所建立之資料庫完成一套系統化的病毒藥物篩選系統。以下根據本計畫所完成之工作項目逐步進行成果陳述與討論。

### 已完成建立不同片段之病毒序列資料庫

本計畫從不同資料庫及相關研究中蒐集 HIV 病毒的蛋白質序列資料，並進一步的將 HIV 病毒的序列資料依照所轉譯的功能不同將基因組做分類，以三組主要的病毒序列基因資料組：env、gag 與 pol 將序列作初步的分類。而除了利用轉譯的功能不同將序列依照基因組分類外，也針對本計畫藥物篩選系統中所使用的序列如 V3 環狀序列片段、p24 序列片段等作進一步的篩選，以建立相對應的預測模型。

表一、本計畫中三組主要的病毒序列資料組數量分布情形。

|                   | env    | gag    | Pol    |
|-------------------|--------|--------|--------|
| Sequences numbers | 12,285 | 12,056 | 10,515 |

表一中列出的是 HIV 病毒中三組主要基因組的序列數量，而表二則進一步列出各基因組中建立藥物篩選系統相關部分區域的序列片段數量。由於各研究論文中所使用的序列片段不同，因此各基因組和區域序列所組成的資料庫會有些許數量上的差異。

表二、各基因組中區域片段序列數量分布表。

| Genome       | Env     |       | Gag |       | Pol      |       |
|--------------|---------|-------|-----|-------|----------|-------|
| Area/Numbers | Gp120   | 1,656 | P17 | 1,164 | P51      | 529   |
|              | Gp41    | 1,570 | P24 | 1,279 | P31      | 435   |
|              | V3 loop | 1,225 | P7  | 1,169 | Protease | 4,828 |

### 已完成亞型分類器

由於不同亞型會影響病毒在各時期的蛋白質組成與酵素活性，並間接或直接影響到病毒感染力的強弱和施予藥物，因此在本研究中便先個別利用三個主要的基因組來建立亞型分類模型，並比較在不同亞型中各基因組的分類能力與影響的物化特性。由於 HIV 病毒最常見的兩種亞型為 B 與 C 亞型，因此本計畫便以這兩種亞型為例，比較不同基因組在亞型之間的差異。下表表三列出在不同基因組中所得到的最佳訓練模型與其獨立測試的準確度結果。由表三發現在建立訓練模型時，利用我們的智慧型基因演算法可將模型的準確率幾乎都達到 100%。而做獨立測試時也都有 90% 以上的預測準確率，

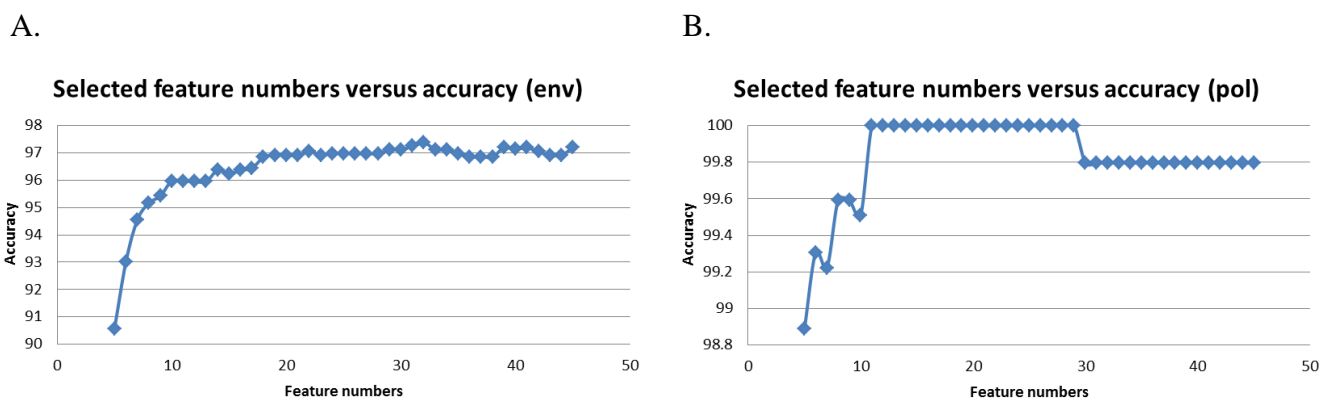


並且在 pol 與 gag 兩種基因組所建成的預測模型，其分類準確率甚至達到了 98% 以上，顯示我們所建立的預測模型確實可將不同的亞型分類。

表三、利用不同基因組建立 HIV 病毒亞型分類模型，各訓練模型與獨立測試結果之準確率。

|                  | Env      |        | Gag      |      | Pol      |        |
|------------------|----------|--------|----------|------|----------|--------|
|                  | Training | Test   | Training | Test | Training | Test   |
| Overall Accuracy | 97.41%   | 92.86% | 100%     | 100% | 100%     | 98.10% |
| Sensitivity      | 97.52%   | 90.61% | 100%     | 100% | 100%     | 98.66% |
| Specificity      | 97.25%   | 96.06% | 100%     | 100% | 100%     | 96.75% |

而三個主要的亞型分類器中，除了 env 所建構的分類器是使用了 32 個物化特性外，另外兩個分類器分別使用了 15 和 11 個物化特性，顯示藉由智慧型計算所挑選出的物化特性確實是能有效的分類不同亞型的病毒。下圖二以 env 與 pol 的訓練模型為例，將不同物化特性數量與準確率的關係呈現出來。



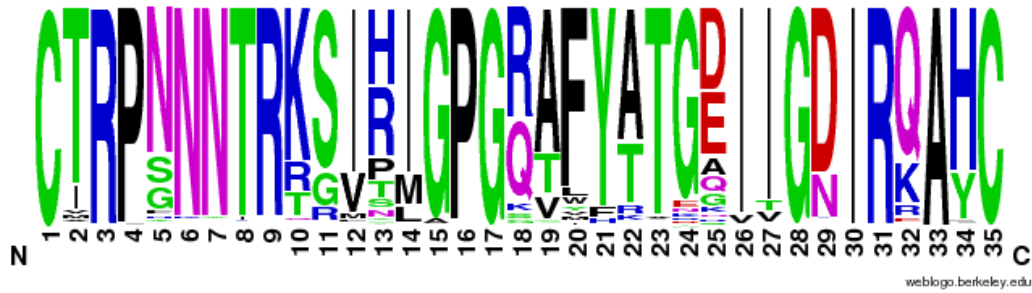
圖二、不同數量的物化特性與相對應的準確率。A 為利用 env 基因組所建成的亞型預測模型，B 為使用 pol 基因組建成的亞型預測模型。

由於繼承式演算法可以在範圍內尋找最佳值，因此範圍的設定與結果是否已達到平衡都會影響模型的準確率與可信度。而由圖 2 可看到建立的亞型預測模型都是已經達到穩定狀態，並且當使用過多的物化特性時，容易導致過度訓練(over training)，以圖 2B 為例，當使用超過 30 個特性時反而使得準確率下降，表示模型已經有過度訓練的情形產生。本計畫所完成之亞型分類器為目前使用胺基酸特性的分類器中最好的，並且所使用的特性不會導致產生過度訓練的模型，符合實際的使用。

### 已完成進入抑制劑篩選模型

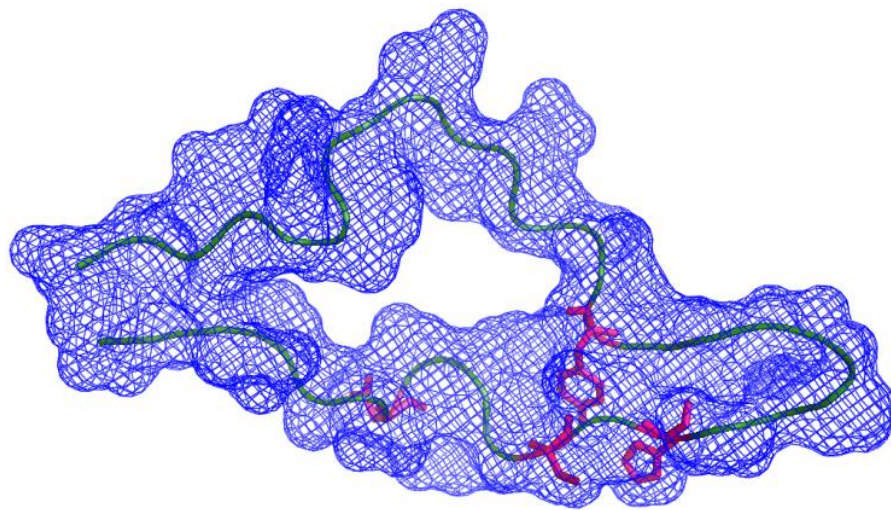
準確並快速的預測 HIV 病毒所使用協同受體，可以有效幫助篩選病人所適合的進入抑制劑，並進一步抑制 HIV 病毒在細胞間傳播。目前用來預測協同受體種類是以病毒膜蛋白上的 V3 環狀序列為主，本研究也是以此段序列建造預測模型，並在進一步地找出能快速並簡單判斷協同受體的規則。在初步成果我們已經建立目前最好的進入抑制劑藥物篩選模型，本研究再對 V3 序列作進一步的探討，包含胺基酸組成、物化特性組成等等以找出一套可快速決定藥物使用的規則。下圖三為 V3 序列經由 alignment 後所得到的氨基酸分布圖，其中可以發現有不少位置的氨基酸組成是非常穩定的，如 N 端的 CTRP 組成和中段的 GPG，都是相對穩定的區域，但這些位置反而會造成分類的困難。





圖三、V3 環狀序列中各位置的氨基酸分布圖。

而我們利用先前所找出的物化特性組，對 V3 序列中的氨基酸做主效果分析，以找出那些位置的氨基酸對篩選進入抑制劑有比較大的作用，而本研究最後發現有四個胺基酸對分類不同種類的病毒具有較大的影響力，這四個位置的氨基酸分別為 site 302、site 306、site 308 和 site 317。圖四為這四個胺基酸在 V3 序列上的位置示意圖。



圖四、四個重要胺基酸位置示意圖。圖中四個粉紅色的樹枝狀結構氨基酸便為實驗中所找出的重要位置，分別為 site302、site 306、site 308 和 site 317。

這四個位置除了 site 306 在之前有被使用於分類病毒外，其餘三個胺基酸位置是由本計畫所發現可用來決定病毒所使用之協同受體的重要位置。為了瞭解這些位置的氨基酸是如何影響協同受體的使用，本計畫結合了三個不同的氨基酸物化特性：帶電量、極性與親疏水性，研究這四個位置，並且找出一組含有四條規則的組合，可以有效的判斷出病毒是否利用 CXCR4 這個協同受體進入細胞。以下為本計畫所找出的四條規則：

- i. 當 site 317 為帶極性胺基酸並且 site 306 同為極性胺基酸及 site308 不為帶負電胺基酸時，則此序列屬於 CXCR4 型病毒。
- ii. 當 site 302 不為極性胺基酸並且 site 308 為親水性胺基酸時，則此序列屬於 CXCR4 型病毒。
- iii. 當 site 306 不為極性胺基酸並且 site 308 為帶負電胺基酸時，則此序列屬於 CXCR4 型或雙性型病毒(可同時使用 CXCR4 與 CCR5 協同受體的病毒)。
- iv. 當 site 302 為 positive 胺基酸時，則此序列屬於 CXCR4 型或雙性型病毒。

表四將上述的四條規則列為表格，並將各規則的訓練模型與獨立測試的準確度列出。表五是將本計畫中所找到的規則和之前常使用的 11/25 電荷規則所比較的結果，我們所得到的規則組不管在適用

資料的覆蓋率和準確率，都比 11/25 電荷規則要高的許多，顯示本計畫所找到的規則組能有效地輔助篩選病毒的進入抑制劑。

表四、藉由分析各位置的胺基酸特性所找出的四條分類規則。由於使用 CXCR4 的序列較難被區分出來，因此本研究便著重於利用規則找出使用 CXCR4 協同受體的病毒。

| Tropisms | Site 302 | Site 306    | Site 308     | Site 317 | Training | Independent test |
|----------|----------|-------------|--------------|----------|----------|------------------|
| X4       |          | Polar       | Not negative | Polar    | 70.4%    | 45.0%            |
|          | Nonpolar | Hydrophilic |              |          | 83.3%    | 64.7%            |
| X4 or RX |          | Nonpolar    | Positive     |          | 80.6%    | 86.4%            |
|          | Positive |             |              |          | 100%     | 88.6%            |
| Total    |          |             |              |          | 76.9%    | 72.0%            |

表五、本計畫中所得到的分類規則與 HIV 研究常用的 11/25 電荷規則比較表。無論是訓練模型或是獨立測試結果，我們的規則所得到的準確率都比 11/25 電荷規則高 15% 以上。並且資料覆蓋率也都高出許多。

|                 | Training accuracy      | Cover rate of all training data | Test accuracy          | Cover rate of all test data |
|-----------------|------------------------|---------------------------------|------------------------|-----------------------------|
| <b>Our rule</b> | <b>76.9% (120/156)</b> | <b>53.1% (156/294)</b>          | <b>72.0% (126/175)</b> | <b>46.3% (175/378)</b>      |
| 11/25 rule      | 61.6% (69/112)         | 38.1% (112/294)                 | 51.6% (64/124)         | 32.8% (124/378)             |

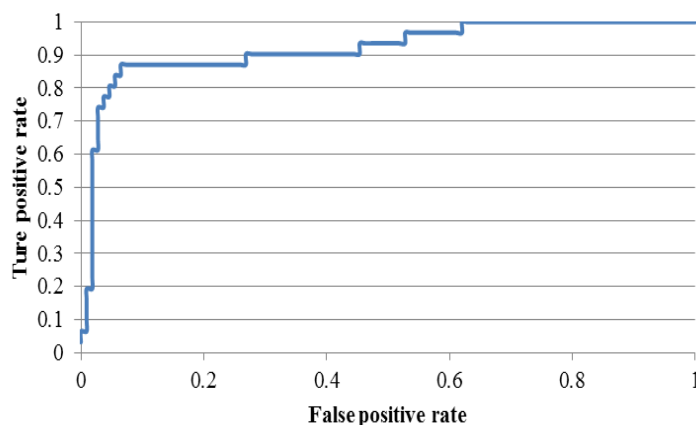
### 已完成成熟抑制劑預測分類器

成熟抑制劑是近年來新使用在輔助 HAART 治療的方法。在本計畫中利用論文及資料庫所取得的 p24 蛋白質序列資料建立成熟抑制劑的預測模型。使用 IBCGA 挑選最佳物化特性組所建立的模型，其預測準確率可達到 92.09%，在與其他研究比較時，本計畫在使用單一分類器時具有最高的 AUC (area under the ROC curve)，並且與多重分類器的隨機森林 (Random Forest, RF) 分類器相比，AUC 只相差了 0.006，顯示本計畫挑選的物化特性組有很好的辨別能力，來判斷病毒是否會被成熟抑制劑抑制。表六列出與其他研究相比的結果，圖五為我們所建立的成熟抑制劑預測模型的 ROC 曲線。

表六、比較不同研究中，成熟抑制劑訓練模型所獲得的 AUC

| Method        | Feature                     | AUC          |
|---------------|-----------------------------|--------------|
| Linear model  | Hydrophobicity              | 0.826        |
|               | Molecular weight            | 0.811        |
| Decision tree | Hydrophobicity              | 0.815        |
|               | Molecular weight            | 0.841        |
| ANN           | Hydrophobicity              | 0.841        |
|               | Molecular weight            | 0.839        |
| RF            | Hydrophobicity              | 0.927        |
|               | Molecular weight            | 0.923        |
| <b>SVM</b>    | <b>Selected feature set</b> | <b>0.921</b> |

AUC = 0.921

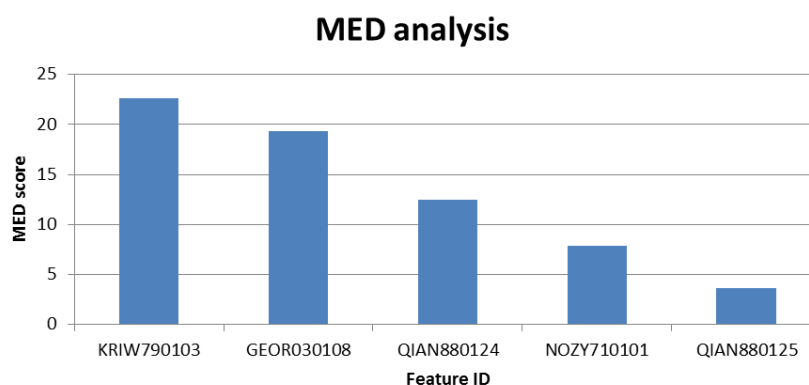


圖五、本計畫建立之成熟抑制劑預設模型的 ROC curve。

我們進一步的利用主效果分析來分析 IBCGA 所挑選出來的 5 個物化特性，並找出在特性組中對於分類貢獻較大的特性，發現有兩個特性的貢獻度較大：Side chain volume 和 Linker propensity，這兩個特性都是關於結構的特性。因為 p24 主要會形成病毒的衣殼(capsid)，為保護病毒內部 RNA 與酵素的結構，而成熟抑制劑便是破壞此結構，讓病毒的 RNA 形成不正確的連結，或是讓酵素失去作用。利用 IBCGA 挑出的特性幾乎都是與結構相關，而對分類最有貢獻的兩個特性也是結構特性，表示本計畫所得到的特性確實能有效的區分病毒是否會受到成熟抑制劑的影響。表七為所挑選出的五個物化特性與相對應的主效果分析分數，圖六為主效果分析中物化特性的排名狀況。

表七、成熟抑制劑預測模型中所挑選出的物化特性組與各物化特性所對應的主效果分析分數。

| Feature ID | Feature description  | MED score |
|------------|--|-----------|
| KRIW790103 | Side chain volume  | 22.59559  |
| GEOR030108 | Linker propensity from helical (annotated by DSSP) dataset | 19.33991  |
| QIAN880124 | Weights for beta-sheet at the window position of 4         | 12.44026  |
| NOZY710101 | Transfer energy, organic solvent/water                     | 7.8405    |
| QIAN880125 | Weights for beta-sheet at the window position of 5         | 3.658905  |



圖六、主效果分析結果，在物化特性組中具有最大分類能力的特性 ID 為 KRIW790103，特性的敘述是 Side chain volume。

本計畫所建立的亞型預測模型和兩個藥物篩選預測模型與目前的方法相比都是準確率最高的預測模型。而使用 IBCGA 挑選最佳的物化特性組，使得模型使用最少的特徵數便能有相當好的結果，並且所挑選出的物化特性經由相關文獻探討後，也都有相關的生物實驗證實為實際會影響藥物連結，或是影響相關病毒蛋白質的特性，代表吾人的 IBCGA 方法對於建立 HIV 的藥物篩選系統是相當適合並有效。

## Reference

1. Seclen, E., et al., *High sensitivity of specific genotypic tools for detection of X4 variants in antiretroviral-experienced patients suitable to be treated with CCR5 antagonists*. J Antimicrob Chemother, 2010. **65**(7): p. 1486-92.
2. Myers, G., K. Maclnnes, and B. Korber, *The emergence of simian/human immunodeficiency viruses*. AIDS Res Hum Retroviruses, 1992. **8**(3): p. 373-86.
3. Louwagie, J., et al., *Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides*

- evidence for multiple genotypes. *AIDS*, 1993. **7**(6): p. 769-80.
4. Loemba, H., et al., *Genetic divergence of human immunodeficiency virus type 1 Ethiopian clade C reverse transcriptase (RT) and rapid development of resistance against nonnucleoside inhibitors of RT*. *Antimicrob Agents Chemother*, 2002. **46**(7): p. 2087-94.
  5. Kantor, R., et al., *Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: results of a global collaboration*. *PLoS Med*, 2005. **2**(4): p. e112.
  6. Bar-Magen, T., et al., *Comparative biochemical analysis of HIV-1 subtype B and C integrase enzymes*. *Retrovirology*, 2009. **6**: p. 103.
  7. Iordanskiy, S., et al., *Subtype-associated differences in HIV-1 reverse transcription affect the viral replication*. *Retrovirology*, 2010. **7**: p. 85.
  8. Jensen, M.A., et al., *Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences*. *J Virol*, 2003. **77**(24): p. 13376-88.
  9. Sing, T., et al., *Predicting HIV coreceptor usage on the basis of genetic and clinical covariates*. *Antivir Ther*, 2007. **12**(7): p. 1097-106.
  10. Fouchier, R.A., et al., *Simple determination of human immunodeficiency virus type 1 syncytium-inducing V3 genotype by PCR*. *J Clin Microbiol*, 1995. **33**(4): p. 906-11.
  11. Lamers, S.L., et al., *Prediction of R5, X4, and R5X4 HIV-1 coreceptor usage with evolved neural networks*. *IEEE/ACM Trans Comput Biol Bioinform*, 2008. **5**(2): p. 291-300.
  12. Sander, O., et al., *Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage*. *PLoS Comput Biol*, 2007. **3**(3): p. e58.
  13. Heider, D., J. Verheyen, and D. Hoffmann, *Predicting Bevirimat resistance of HIV-1 from genotype*. *BMC Bioinformatics*, 2010. **11**: p. 37.
  14. Li, F., et al., *Determinants of activity of the HIV-1 maturation inhibitor PA-457*. *Virology*, 2006. **356**(1-2): p. 217-24.
  15. Benson, D.A., et al., *GenBank*. *Nucleic Acids Res*, 2010. **38**(Database issue): p. D46-51.
  16. Whitcomb, J.M., et al., *Development and characterization of a novel single-cycle recombinant-virus assay to determine human immunodeficiency virus type 1 coreceptor tropism*. *Antimicrob Agents Chemother*, 2007. **51**(2): p. 566-75.
  17. Keele, B.F., et al., *Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection*. *Proc Natl Acad Sci U S A*, 2008. **105**(21): p. 7552-7.
  18. Dybowski, J.N., D. Heider, and D. Hoffmann, *Prediction of co-receptor usage of HIV-1 from genotype*. *PLoS Comput Biol*, 2010. **6**(4): p. e1000743.
  19. Naganawa, S., et al., *Net positive charge of HIV-1 CRF01\_AE V3 sequence regulates viral sensitivity to humoral immunity*. *PLoS One*, 2008. **3**(9): p. e3206.
  20. Garrido, C., et al., *Evaluation of eight different bioinformatics tools to predict viral tropism in different human immunodeficiency virus type 1 subtypes*. *J Clin Microbiol*, 2008. **46**(3): p. 887-91.
  21. Huang, W., et al., *Vertical transmission of X4-tropic and dual-tropic HIV-1 in five Ugandan mother-infant pairs*. *AIDS*, 2009. **23**(14): p. 1903-8.
  22. Ho, S.Y., J.H. Chen, and M.H. Huang, *Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications*. *IEEE Trans Syst Man Cybern B Cybern*, 2004. **34**(1): p. 609-20.

23. S.-Y. Ho, L.-S.S., and J.-H. Chen, *Intelligent Evolutionary Algorithms for Large Parameter Optimization Problems*. IEEE Trans. Evolutionary Computation, 2004. **8**(6): p. 19.
24. Ho, S.Y., et al., *Design of accurate classifiers with a compact fuzzy-rule base using an evolutionary scatter partition of feature space*. IEEE Trans Syst Man Cybern B Cybern, 2004. **34**(2): p. 1031-44.
25. Tung, C.W. and S.Y. Ho, *POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties*. Bioinformatics, 2007. **23**(8): p. 942-9.
26. Kawashima, S., et al., *AAindex: amino acid index database, progress report 2008*. Nucleic Acids Res, 2008. **36**(Database issue): p. D202-5.
27. Ho, S.Y., et al., *Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis*. Biosystems, 2006. **85**(3): p. 165-76.
28. Huang, W.L., et al., *Accurate prediction of enzyme subfamily class using an adaptive fuzzy k-nearest neighbor method*. Biosystems, 2007. **90**(2): p. 405-13.
29. Bhasin, M. and G.P. Raghava, *Analysis and prediction of affinity of TAP binding peptides using cascade SVM*. Protein Sci, 2004. **13**(3): p. 596-607.
30. Cao, Y., et al., *Prediction of protein structural class with Rough Sets*. BMC Bioinformatics, 2006. **7**: p. 20.

## 相關研究成果

### Journal paper

H.-L. Huang, I.-C. Lin, Y.-F. Liou, C.-T. Tsai, K.-T. Hsu, W.-L. Huang, S.-J. Ho, and **S.-Y. Ho\***, "Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties", *BMC Bioinformatics*, 12(Suppl 1):S47, 2011. (SCI)

### International Conference papers

1. H.-L. Huang, Y.-F. Liou, H.-C. Lee, W.-L. Huang, and S.-Y. Ho\*, 2012, "Designing predictors of bioluminescence proteins using an efficient physicochemical property mining method," The 6th International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2012), Shanghai, China, May 17-20, 2012. (EI)
2. H.-L. Huang, H.-C. Lee, Y.-F. Liou, M.-C. Li, and S.-Y. Ho\*, 2012, "Designing predictors of carbohydrate-binding proteins using informative physicochemical properties," Proceedings of 2012 International Conference on Bioinformatics and Biomedical Technology (ICBBT 2012), Singapore, Feb. 26-28, 2012.

# 國科會補助專題研究計畫項下出席國際學術會議心得報告

日期：101 年 09 月 30 日

|        |   |         |                              |
|--------|---|---------|------------------------------|
| 計畫編號   | NSC 100-2221-E-009-143-   |         |                              |
| 計畫名稱   | 研發人體免疫缺失病毒治療的最佳化藥物配方篩選系統  |         |                              |
| 出國人員姓名 | 何信瑩   | 服務機構及職稱 | 交通大學生物資訊研究所<br>教授兼任生物科技學院副院長 |
| 會議時間   | 101 年 5 月 17 日至<br>101 年 5 月 20 日   | 會議地點    | 中國上海                         |
| 會議名稱   | (中文)2012 年第六屆生物資訊與生醫工程國際會議<br>(英文)The 2012 6 <sup>th</sup> International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2012)   |         |                              |
| 發表論文題目 | (中文)1. 使用計分卡設計醣結合蛋白質的預測器<br>2. 使用物化特性方法設計螢光蛋白質的預測器<br>(英文)1. Prediction of Carbohydrate-Binding Proteins Using a Scoring Card Method<br>2. Designing predictors of bioluminescence proteins using an efficient physicochemical property mining method |         |                              |

## 一、參加會議經過

本次參加會議是 2012 年第六屆生物資訊與生醫工程國際會議(iCBBE 2012)，由國際電子電機學會 IEEE Engineering in Medicine and Biology Society 所贊助舉辦，協辦單位包含美國 The Gordon Life Science Institute 與中國大陸九所大學。吾人所發表之論文題目一共有兩篇，一篇是「Prediction of Carbohydrate-Binding Proteins Using a Scoring Card Method」，主要是使用基因演算法找出最佳化的計分卡 (scoring card)，設計出醣結合蛋白質的預測器與分析蛋白質序列在醣結合所扮演的功能與腳色；另外一篇是「Designing predictors of bioluminescence proteins using an efficient physicochemical property mining method」，主要是使用計算智慧的最佳化技術來從 531 個物化特性中挑選出一組最佳的小集合，結合 support vector machine 分類器，設計出螢光蛋白質的預測器，並分析這些蛋白質序列的物化特性在螢光發光所扮演的功能與腳色。

第六屆生物資訊與生醫工程國際會議自五月十七日至五月二十日在中國上海的上海寶隆美爵酒店舉行。吾人本次投稿之兩篇論文屬於生物資訊蛋白質資料分析



領域的口頭論文報告，會議安排論文報告之時間在五月十九日早上。由於本次會議不是於放假日舉辦，考量開會上課之因素，因此選擇星期五（十八日）搭長榮航班於桃園機場前往中國上海，直飛上海浦東參加會議及論文報告，並於二十日搭長榮航班由虹橋飛往松山機場回到國內。目前生物資訊與生醫工程為近年學者爭相研究之領域，相關國際會議有增加的趨勢。此會議口頭報告學者來自眾多國家，包含義大利、美國、臺灣、中國、以色列、泰國、日本、摩洛哥、沙烏地阿拉伯、巴基斯坦、俄國、捷克、伊朗、德國、土耳其、馬來西亞、波蘭、愛爾蘭、英國、法國、卡達、加拿大、印尼、西班牙.....等，其中中國學者人數有大幅增加之趨勢。另外，大會也安排四個 keynote speech 讓來自各個國家的學者互相交流、聯誼，促進了與會學者日後的學術交流機會。

大會安排四個重要的演講，包括（一）法國國家科學研究院 Athel 教授演講 Metabolic modeling: A Necessary Tool for Biotechnology；（二）美國康乃爾大學 Ann 教授演講 Fenton Oxidation of Contaminants using Nanomagnetite；（三）美國哈佛醫學院 Chou 教授演講 An NMR view of membrane transporters: application to mitochondrial carriers；（四）以色列理工學院 Daniel 教授演講 A general overview of medical robotics，令參與者更深入瞭解此項領域的重大研究發展潮流，大大提高了對這方面研究的瞭解。本人認為，目前生物資訊與生醫工程很重視分析工具、奈米元件與生醫應用這幾大類的研究領域，這是未來的趨勢。

接下來是論文口頭報告，吾人兩篇論文發表之時段都是探討以生物資訊角度進行 Gene expression 或 protein 方面之研究成果論文報告，此時段的論文研究報告大多是來自臺灣、中國、日本與美國方面的學者，此一現象表示臺灣在此方面的成就仍然是名列前茅的。會議中會議報告進行之過程非常順利與和諧，也吸收了很多新的知識。

吾人所投稿的口頭報告論文是主題是使用計分卡（scoring card）設計醣結合蛋白質的預測器，介紹此計分卡由智慧型基因演算法算出最佳化的一份計分卡，因此，只要經過此簡單之計分卡即可以達到以往使用繁複計算方法才能達到的準確度，大幅降低一般研究生物資訊計算所需的複雜度。另外一篇是使用基因演算法的最佳化技術從 531 個物化特性挑選出一組最佳解，結合 support vector machine 分類器，設計出螢光蛋白質的預測器並進一步分析螢光蛋白質序列的物化特性在螢光發光所扮演的功能。每一位口頭論文發表為 30 分鐘，其中 25 分鐘是論文發表，5 分鐘是問題討論。過程中有多位學者提出問題，過程十分順利，本人在場與參與學者交換心得，且有一位英國學者對此方法特別感興趣，會後來函希望能取得計分卡之工具。第一篇論文由共同作者黃慧玲教授上台發表，第二篇論文由共同作者李華錦博士後研究員上台發表，而我則是全程在台下參與整個過程。本次會議引發很多合作的靈感，吾人深知學術上的合作將是一個很重要的一步。此外，我也結識不少同行並留下以後學術上合作的機會，本人認為收穫成果豐碩。

除了生物資訊領域之外，我對其他生物醫學或是生物影像也非常有興趣，因此主動參加其他時段的 session，過程中學到很多收穫與新知。在日益蓬勃生物資訊發展之下，讓來自眾多不同地區的學者共聚一堂討論可以有效率地精準吸收新知。



圖一：論文發表會場（左起黃文玲教授、徐禮燊教授、李華錦博士後研究員、我、何信璋教授與黃慧玲教授）



圖二、會後會場照片（左起黃文玲教授、徐禮燊教授、我、何信璋教授與黃慧玲教授）

## 二、與會心得

本人跨領域生物資訊已經超過八年，且研究基因演算法已近十年，每年均參加多場生物計算方面的國際會議，故本次會議重點 - 生物資訊與生醫工程特別感興趣，由其對很多研究主題及發展方向也關心，希望對提高台灣的學術聲望及研究能量提升有所貢獻。

此 iCBBE 2012 包含傳統基因分析、生醫訊號處理分析、演算法、蛋白質架構、計算生物學、生醫元件、生醫影像處理、臨床醫學工程、生醫工程之新技術.....等等的口頭報告；海報發表則多在生資計算、生醫工程方面。本次會議所演講之內容讓與會者更易於瞭解、掌握研究趨勢，本人認為此會議藉由國際電子電機學會 (IEEE) 贊助推廣生物資訊與生醫工程研究的努力是值得讚賞和學習。

## 三、考察參觀活動(無是項活動者略)

#### 四、建議

近年來國家與學校無不積極鼓勵年輕學者參與會議目的是希望藉由參與大型國際會議，能快速瞭解相關研究領域的進展，學習國際研究經驗與交流，提高國人的研究水能量。參加國際生物資訊或是生醫工程相關會議是非常重要的事情，在會議中不但可以吸收相關研究的最新發展現況與資訊，認識結交具有此領域深入研究的學者，彼此交換研究的心得，並找到跨領域的學者國際合作，在跨領域的生物資訊研究更是重要。希望國家與學校單位能多在補助年輕學者出國，提升研究的品質。

#### 五、攜回資料名稱及內容

Proceedings of the 2012 6<sup>th</sup> International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2012), Shanghai, China, May 17-20, 2012. (含紙本與光碟)

#### 六、其他

# 國科會補助計畫衍生研發成果推廣資料表

日期:2012/10/24

|           |                                      |
|-----------|--------------------------------------|
| 國科會補助計畫   | 計畫名稱: 研發人體免疫缺失病毒治療的最佳化藥物配方篩選系統       |
|           | 計畫主持人: 何信瑩                           |
|           | 計畫編號: 100-2221-E-009-143- 學門領域: 生物資訊 |
| 無研發成果推廣資料 |                                      |

100 年度專題研究計畫研究成果彙整表

| 計畫主持人：何信瑩                     |             | 計畫編號：100-2221-E-009-143- |                 |            |      |                                     |  |
|-------------------------------|-------------|--------------------------|-----------------|------------|------|-------------------------------------|--|
| 計畫名稱：研發人體免疫缺失病毒治療的最佳化藥物配方篩選系統 |             |                          |                 |            |      |                                     |  |
| 成果項目                          |             | 量化                       |                 |            | 單位   | 備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等） |  |
|                               |             | 實際已達成數（被接受或已發表）          | 預期總達成數（含實際已達成數） | 本計畫實際貢獻百分比 |      |                                     |  |
| 國內                            | 論文著作        | 期刊論文                     | 0               | 0          | 100% | 篇                                   |  |
|                               |             | 研究報告/技術報告                | 0               | 0          | 100% |                                     |  |
|                               |             | 研討會論文                    | 0               | 0          | 100% |                                     |  |
|                               |             | 專書                       | 0               | 0          | 100% |                                     |  |
|                               | 專利          | 申請中件數                    | 0               | 0          | 100% | 件                                   |  |
|                               |             | 已獲得件數                    | 0               | 0          | 100% |                                     |  |
|                               | 技術移轉        | 件數                       | 0               | 0          | 100% | 件                                   |  |
|                               |             | 權利金                      | 0               | 0          | 100% | 千元                                  |  |
|                               | 參與計畫人力（本國籍） | 碩士生                      | 0               | 0          | 100% | 人次                                  |  |
|                               |             | 博士生                      | 0               | 0          | 100% |                                     |  |
|                               |             | 博士後研究員                   | 0               | 0          | 100% |                                     |  |
|                               |             | 專任助理                     | 0               | 0          | 100% |                                     |  |
| 國外                            | 論文著作        | 期刊論文                     | 1               | 1          | 100% | 篇                                   |  |
|                               |             | 研究報告/技術報告                | 0               | 0          | 100% |                                     |  |
|                               |             | 研討會論文                    | 2               | 2          | 100% |                                     |  |
|                               |             | 專書                       | 0               | 0          | 100% | 章/本                                 |  |
|                               | 專利          | 申請中件數                    | 0               | 0          | 100% | 件                                   |  |
|                               |             | 已獲得件數                    | 0               | 0          | 100% |                                     |  |
|                               | 技術移轉        | 件數                       | 0               | 0          | 100% | 件                                   |  |
|                               |             | 權利金                      | 0               | 0          | 100% | 千元                                  |  |
|                               | 參與計畫人力（外國籍） | 碩士生                      | 0               | 0          | 100% | 人次                                  |  |
|                               |             | 博士生                      | 0               | 0          | 100% |                                     |  |
|                               |             | 博士後研究員                   | 0               | 0          | 100% |                                     |  |
|                               |             | 專任助理                     | 0               | 0          | 100% |                                     |  |

|  |          |
|--|----------|
| <p>其他成果<br/>(無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)</p> | <p>無</p> |
|--|----------|

|   | 成果項目            | 量化 | 名稱或內容性質簡述 |
|---|-----------------|----|-----------|
| 科<br>教<br>處<br>計<br>畫<br>加<br>填<br>項<br>目 | 測驗工具(含質性與量性)    | 0  |           |
|   | 課程/模組           | 0  |           |
|   | 電腦及網路系統或工具      | 0  |           |
|   | 教材              | 0  |           |
|   | 舉辦之活動/競賽        | 0  |           |
|   | 研討會/工作坊         | 0  |           |
|   | 電子報、網站          | 0  |           |
|   | 計畫成果推廣之參與(閱聽)人數 | 0  |           |



# 國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

## 1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

達成目標

未達成目標（請說明，以 100 字為限）

實驗失敗

因故實驗中斷

其他原因

說明：

## 2. 研究成果在學術期刊發表或申請專利等情形：

論文： 已發表  未發表之文稿  撰寫中  無

專利： 已獲得  申請中  無

技轉： 已技轉  洽談中  無

其他：（以 100 字為限）

本子計畫衍生成果包括多目標建模及最佳化篩選特徵值與物化特性探討，研究成果豐碩。合計已發表相關期刊論文 1 篇，即將投稿 2 篇期刊論文及已發表 2 篇國外會議論文。

## 3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

本計畫在執行期間成果豐碩。透過最佳化演算法挑選物化特性組，所建立的 HIV 藥物篩選模型，除了改進現有藥物篩選模型的預測準確度，使得篩選系統更加可信外，也可以避免藥物篩選模型使用過多特徵值而產生過適化的問題，建造出來的模型會符合一般的狀況，適用於大多數的情況，而所挑選出的物化特性也被相關生物研究驗證，實際會影響相關藥物的連結或是病毒蛋白質的特性。準確的進入抑制劑與成熟抑制劑預測模型的建立，可以有有效的輔助雞尾酒療法，並且能依照病人體內的病毒不同，快速選擇出適合個別病患的藥物，以達到個人化醫療的目標。本計畫亦對所挑選出的物化特性做分析，確認了這些特性都有相關研究佐證，會影響病毒的藥物連結位置或抗藥性。而進一步的發現這些特性除了可以在分類上有所貢獻外，對於相關藥物的設計也具有相當大的意義，可將這些發現應用於進入抑制劑與成熟抑制劑的藥物設計研究中。