行政院國家科學委員會專題研究計畫　成果報告

應用智慧型運算預測主要組織相容複合體與胜鈦鏈之結合特性

計畫類別：個別型計畫
計畫編號：NSC93-2213-E-034-006-
執行期間：93 年 08 月 01 日至 94 年 07 月 31 日
執行單位：中國文化大學資訊科學系

計畫主持人：蔡敦仁

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 94 年 10 月 30 日

行政院國家科學委員會專題研究計畫成果報告

應用智慧型運算預測主要組織相容複合體與胜肽鏈之結合特性

計畫編號：NSC 93-2213-E-034-006
執行期限：93 年 8 月 1 日至 94 年 7 月 31 日
主持人：蔡敦仁 中國文化大學資訊科學系

一、摘要

目前運用智慧型運算(Intelligent Computation)於解決生物資訊(Bioinformatics)領域之研究，大多是進行辨識分類、統計分析或是資料探勘等探討，然而大部分的研究都忽略生物資訊領域具有之問題：在做資料分析時，分析結果嚴重受到資料維度(Data Dimensionality)影響的問題。也就是，當資料維度太低時，屬於不同特性資料將無法區分；當資料維度太高時，區分方式又會有無限多種可能，如此，不同的分類方式都符合要求，但是卻會導致不同的研究結果，將使得結果的可參考性降低。本計畫針對此一問題，嘗試設計發展一個具有分類最佳化(Optimized Classification)能力的智慧型運算模型，以應用於解決生物資訊領域之問題。所進行的生物資訊題目，是針對人體免疫系統中「主要組織相容性複合體」(Major Histocompatibility Complex, MHC)與外來胜肽鏈(Peptides)之間結合的特性之預測，利用分類最佳化的智慧型運算模型來辨識主要組織相容性複合體與胜肽鏈結合的親疏性。在此研究中，我們不僅提出適合於生物資訊領域研究之智慧型運算方法，也對提出 MHC 與 Peptides 間結合特性預測之解法，並進行系統模擬實作，以驗證提出模型之效能。

關鍵詞：智慧型運算、生物資訊、主要組織相容性複合體、胜肽鏈。

Abstract

In this project, we discuss the problems of intelligent computation methods applied on the bioinformatics, and attempt to develop a model to resolve the problem. This method utilizes the optimized classification to solve the problem of data dimensionality which is the key point to analyze the biological data effectively. Based on this model, we also propose the solution for the prediction of the major histocompatibility complex (MHC)-peptide binding affinity. The simulations will be implemented to show that the proposed model improves the effectiveness for the prediction results.

Keywords: Intelligent Computation, Bioinformatics, Major Histocompatibility Complex (MHC), Peptides.

二、緣由與目的

目前運用智慧型運算(Intelligent Computation)於解決生物資訊(Bioinformatics)領域之研究，大多是進行辨識分類、統計分析或是資料探勘等探討，然而大部分的研究都忽略生物資訊領域具有之問題：在做資料分析時，分析結果嚴重受到資料維度(Data Dimensionality)影響的問題。也就是，當資料維度太低時，屬於不同特性資料將無法區分；當資料維度太高時，區分方式又會有無限多種可能，如此，不同的分類方式都符合要求，但是卻會導致不同的研究結果，將使得結果的可參考性降低。本計畫針對此一問題，嘗試設計發展一個具有分類最佳化(Optimized Classification)能力的智慧型運算模型，以應用於解決生物資訊領域之問題。所進行的生物資訊題目，是針對人體免疫系統中「主要組織相容性複合體」

(Major Histocompatibility Complex, MHC)與外來胜肽鏈(Peptides)之間結合的特性之預測，利用分類最佳化的智慧型運算模型來辨識主要組織相容性複合體與胜肽鏈結合的親疏性。在此研究中，我們不僅提出適合於生物資訊領域研究之智慧型運算方法，也對提出 MHC 與 Peptides 間結合特性預測之解法，並進行系統模擬實作，以驗證提出模型之效能。

人體內的免疫系統的反應主要是：免疫細胞會將外來物切成片段(抗原)，然後再利用主要組織相容性複合體 MHC 將抗原呈現在細胞表面，將 MHC 與外來 Peptides 結合分子會呈現給從未接觸過這些抗原的輔助者 T 細胞，並教導 T 細胞認識抗原是外來或危險的細胞，一旦輔助者 T 細胞認識了這些抗原，就會讓 B 細胞去製造抗體，抗體與抗原結合後，便會使得抗原失效[1]。這一連串免疫過程中，最重要的關鍵便是負責呈現抗原的 MHC 分子，本計畫的目的便是找出 MHC 與外來胜肽鏈結合的特性，如此一來便可以預測免疫系統的辨識工作，以提升製藥的效率。

我們利用智慧型運算模型[2]作為資料探勘的工具，智慧型運算模型目前在各領域已被廣泛的運用，而此研究亦可算是資料探勘的一種。然而，用於解決生物資訊領域之研究時，首要解決的就是在做資料分析時，分析結果嚴重受到資料維度影響的問題。我們將提出具有分類最佳化能力的智慧型演算法收集來的資料，利用此模型進行資料探勘，找出 MHC-Peptides 其間的結合力。

為了進一步探討 MHC 與外來胜肽鏈結合的特性，我們研讀過一些相關的文獻[3]-[14]。其中[3]值得在此作一討論。該文作者的動機是為了發展一個預測胜肽鏈與 MHC 第二型分子結合關係的方法，根據實驗結果的結合資料、已知的下錨點及發展出來的演算法去組合會結合的胜肽鏈及智慧型運算模型，使用 HLA-DR4 (B1*0401)作為結合的基座。以「將結合的資料取出，對其執行胜肽鏈演算法，再將這些資料進行智慧型運算模型的訓練以及分類」此方法作者稱為 PERUN。

模型 PERUN 預測 Peptide 各位置的規則是將其與 MHC 結合力的強弱分成：高、中等、低以及零，並以數值化表示之。利用以 Genetic Algorithms 與 Neural Networks 混和方式模型，去預測胜肽鏈與 MHC 分子結合的方式。實驗用的胜肽鏈部分包括 338 組會結合的胜肽鏈以及 312 組不會結合的胜肽鏈，使用進化演算法，並對氨基酸[7]在胜肽鏈中出現的位置給予不同的分數。以介於 0 到 10 的輸出值來訓練智慧型運算模型，其中 0 代表不會結合、6 表示低結合力、8 為中等結合力、10 表示高結合力。最後將最初的 650 組胜肽鏈任意的分配為訓練與測試組，進行測試之後得到的實驗結果。

此篇論文的方法主要是根據氨基酸出現在胜肽鏈中不同位置的機率，給予不同的分數，然而這種方式有幾個可能的問題：

(1) 每次訓練結果的預測能力都不相同。

(2) 只考慮氨基酸在單一位置上出現的機率，是否會遺漏其他重要資訊？

(4) 由於其使用的資料只有 650 組，數量足夠嗎？

(3) 如何恰當地給與每個位置分數？

事實上，上述的幾個問題，其實是生物資訊領域常見的一個大問題：在做資料分析時，分析結果嚴重受到資料維度影響的問題。當資料維度太低時，屬於不同特性資料將無法區分；當資料維度太高時，區分方式又會有無限多種可能，如此，不同的分類方式都符合要求，但是卻會導致不同的研究結果，將使得結果的可參考性降低。本計畫針對此一問題，嘗試設計發展一個具有分類最佳化能力的智慧型運算模型[15]，以應用於解決生物資訊領域之問題。

三、結果與討論

3.1 模型與設計

MHC 分子與外來胜肽鏈結合，便是呈現抗原的關鍵點，因此研究的重點在於找出其結合的特性，外來的胜肽鏈是千變萬化的，那麼有限的人類白血球抗原(Human Leukocyte Antigen, HLA)，HLA 如何辨識出這些數量龐大的外來胜肽鏈？從文獻中可以瞭解到，HLA 與外來的胜肽鏈結合並沒有固定方式或是結合點，所以要得知其下錨點便需要分析出其三維立體圖形，必須花費許多時間才能得到這些資訊，我們認為外來的胜肽鏈內，氨基酸之間似乎存在某種力量，或許這些力量，便是影響其與 HLA 結合的原因，於是我們假設此結合力存在，如果可以找出這種結合力，便可辨識胜肽鏈與 HLA 是否會結合。

首先，胜肽鏈資料的取得，我們以 MHC 第一型中的 HLA-A*0201 為基座，收集與其結合與不會結合的胜肽鏈，其中每組胜肽鏈皆由九個氨基酸所組成(資料來源[16])。接著，利用我們利用逆傳遞學習模型的學習準確率高、修正速度快的優點來學習胜肽鏈的結合力。這是一種避開先前缺陷來訓練智慧型運算模型的演算法，訓練智慧型運算模型就是為了設定最好權重的過程，訓練的目標在於使產生的權重可以讓輸出值盡可能的接近在訓練資料中的各個例子。研究進行步驟如下：

1. 切割胜肽鏈：首先我們切割收集來的胜肽鏈序列。每個胜肽鏈分別由 9 個氨基酸所組成，為了得到每個位置上的氨基酸與前後兩個氨基酸之間結合力的特性，將其以 3 個氨基酸為一個單位切開，稱之為氨基酸序列，以會結合的胜肽鏈部分第一組序列為例，進行切割之後，便可得到七組3個為一單位的氨基酸序列(Figure 1)。

2.去掉重複的序列：原本共有 1266 組胜肽鏈，經過第一步驟之後，共可得到 8862 組氨基酸序列，我們分別去掉重複出現的部分以及資料中未知的氨基酸 ”X” 之後，可得到會結合的部份 2951 組，不會結合的部分 209 組。

3. 資料分析：由於我們以 3 個氨基酸序列為一個單位來進行胜肽鏈內結合力的研究，因此所有的可能性共有 8000 種(氨基酸共有 20 種，每組有三個位置，所以所有可能的組合共有 203 種)。我們將獲得的會結合與不會結合的氨基酸序列進行比對，其中交集的部分共有 171 組，因此會結合的胜肽鏈集合中不出現在不會結合的胜肽鏈集合內的氨基酸序列有 2780 組，不會結合的胜肽鏈集合中不出現在會結合的胜肽鏈集合內的氨基酸序列共有 38 組。

4. 編碼：接下來我們必須對這些三個一組的氨基酸序列進行編碼。首先將氨基酸依照字母的順序排列，給予每個氨基酸一組 20 維的編碼(Figure 2)。接著將會結合與不會結合的氨基酸序列進行編碼，並將會結合的氨基酸序列輸出值設定為 1，不會結合的氨基酸序列輸出值設為 0，因此每組氨基酸序列皆以 60 個輸入以及 1 個的輸出值來表示。

5. 資料取樣：從資料可以發現會結合的氨基酸序列數量遠大於不會結合的部分的，而懸殊的資料量無法訓練出可以正確辨識的智慧型運算模型，而其中未知的胜肽鏈並無顯示結合的可能性，所以我們將其歸類為不會結合，並從中取出資料，來彌補已知資料中不會結合部分的不足。

6. 第一階智慧型運算模型：Figure 3 是第一階智慧型運算模型，包含一層輸入層、兩層的隱藏層、一層輸出層的智慧型運算模型。將資料輸入輸入層之後，便會傳入隱藏層，再與輸出層進行比較，若誤差太大，則回去修改權重，當誤差在範圍值內則停止。並將動量參數設定為 0.2，學習速率參數設定為 0.3，給予第一層隱藏層 40 個神經元，第二層隱藏層 20 個神經元。

7. 訓練：將 4011 組資料輸入，進行智慧型運算模型的訓練。起始參數的設定，共有 4011 組資料，每組資料有 60 個輸入值以及 1 個輸出值，中間可以看到動量和學習速率兩個參數的設定，以及隱藏層中的神經元數目，接著便開始訓練。

8. 檢視訓練過程：訓練過程中，總錯誤值漸漸下降，正確率上升。在訓練至 76%

時，由於正確率的提升開始越來越慢，便終止訓練。

9. 取得胜肽鏈特徵值：將氨基酸序列的資料輸入以訓練好的第一層智慧型運算模型，即可取得胜肽鏈特徵值。

10. 第二階智慧型運算模型：建立第二階的智慧型運算模型。接著再將胜肽鏈特徵值輸入至第二階的智慧型運算模型，進行訓練(Figure 4)，以取得胜肽鏈與 MHC 結合的親疏性。

11. 驗證階段：當第一階智慧型運算模型、第二階智慧型運算模型都訓練完成後，便能用來驗證測試資料。

3.2 結果與討論

訓練第一階智慧型運算模型時，用來訓練的氨基酸序列，輸出值在會結合的部分設定為 1，不會結合的部分設定為 0。故辨識時，若輸入為會結合的資料，智慧型運算模型計算之後應該輸出 1 或極接近 1 的值，若輸入為不會結合的資料，則應該輸出 0 或接近 0 的值。

取得這些特徵值之後便可訓練出第二階的智慧型運算模型，之後只要將未知的胜肽鏈放入這個模組，輸入這兩階智慧型運算模型，便可得知其與 MHC 結合的親疏關係。

我們初步的研究方法以及結果，解決了 MHC-Peptide Binding 的親疏預測問題。首先，這個方法考慮的不再是單一的位置分析，而是一個氨基酸序列與前、後兩者之間關係，並利用第兩階的智慧型運算模型，學習每個氨基酸序列在不同位置上的關係，如此一來可以成功降低訓練資料的維度，並減少資訊的遺失，提高辨識度。其次，我們不使用配分的方式，而是藉由學習其結合力的方式，來進行辨識的工作，對於每個位置來說，此方式較為客觀，無給分的問題，而且過程較為簡易。接著，我們使用的資料量較多，如此，智慧型運算模型可以得到更多正確的資訊，也能更準確地進行辨識的工作。

當在做序列資料分析以及訓練學習時，分析結果嚴重受到資料維度嚴重影響。也就是說，當資料維度太低時，屬於不同特性資料將無法區分；當資料維度太高時，區分方式又會有無限多種可能，如此，不同的分類方式都符合要求，但是卻會導致不同的研究結果，將使得結果的可參考性降低。

我們使用的第一層類神經網路正確率訓練至 76 % 並未訓練達 90 %以上。當我們將類神經網路訓練至 76 %時，其總錯誤值下降的速度變得相當緩慢，可能需要藉由更改參數，或是以更少維度的編碼方式，來加快學習的速度與正確性，以提高其正確率。

不會結合的的資料與會結合的資料的相較之下數量太少。然而不會結合的胜肽鏈資料在數量上的問題，則須待更多的發現，才能補充資料上的不足。

最初我們假設胜肽鏈內存在某種結合力，這種結合力可能是 MHC 與胜肽鏈結合的關鍵，藉此找出兩者間結合的特性，來辨識結合的可能性，從結果中可以發現胜肽鏈內的確存在這種結合力。之後當我們想知道一個未知的外來胜肽鏈是否會與這個基座結合時，只需經過簡單的處理以及編碼之後，放入我們的訓練過的兩層類神經網路中，即可得知其結合的親疏性，而不再需要做複雜的實驗，或是三維結構圖分析，只要幾個簡單的步驟就可以達到不錯的效果。

我們藉由研究免疫系統中 MHC 與這些外來的胜肽鏈結合的資訊，找尋它們之間的結合力。選擇一個特定的 HLA 為基座，以已知的氨基酸序列來訓練類神經網路，之後，便可利用此類神經網路來辨識其他未知的胜肽鏈。研究的結果中顯示已知的資料中會結合部分辨識度相當不錯，高達 99 %，雖然不會結合部分不盡理想，但仍可證明胜肽鏈內的氨基酸序列間的確存在這種結合力，只要能增加更多用來訓

練的資訊或是再提高類神經網路的正確率，便能夠再提升其辨識力，提升辨識度之後，就可以更加精確的辨識這種結合力，如此一來，對於未知的胜肽鏈便不需再做複雜的分析，只要對其進行簡單的整理便可進行辨識的工作，馬上可以知道與人體內 HLA 結合的可能性，如此一來對於製藥方面，便可以對特定的病毒產生特定的抗體，使製藥的過程更有效率。

四、計畫成果自評

本計劃完成之工作項目包括：

- 研究主要組織相容複合體與胜鈦鏈之結合特性。
- 設計一個智慧型運算預測主要組織相容複合體與胜鈦鏈之結合特性。

本計劃之成果對學術研究之貢獻如下：提出一個一個智慧型運算預測主要組織相容複合體與胜鈦鏈之結合特性，並建置相關實作及實驗，驗證其中的效能。計劃的結果具有實驗性價值，可提供後續相關學術研究參考。

五、參考文獻

[1] Honeyman,M.C., Brusic,V. and Harrison L.C. (1997) Strategies for identifying and predicting islet autoantigen T-cell epitopes in insulin-dependent diabetes (IDDM). Ann. Med., 29, 401–404.

[2] Negnevitsky, M. (2002) Artificial Intelligence: A Guide to Intelligent Systems. Addison Wesley.

[3] Brusic, V., Rudy, G.., Honeyman, M., Hammer, J., and Harrison, L. (1998) Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. Bioinformatics, 14(2), 121-130.

[4] Dönnes, P., and Elofsson, A. (2002) Prediction of MHC class I binding peptides, using SVMHC. 2002, BMC Bioinformatics.

[5] Sousa, J, et al (2003). Representation of DNA sequences with virtual potentials and their processing by (SEQREP) Kohonen self-organizing maps, Bioinformatics 19(1) p30-36

[6] Cano, P., and Fan, B. (2001) A geometric and algebraic view of MHC-peptide complexes and their binding properties. 2001, BMC Structural Biology.

[7] Gibbs,A.J. and McIntyre,G.A. (1970) The diagram, a method for comparing sequences. Its use with amino acids and nucleotide sequences. Eur. J. Biochem., 16, 1–11.

[8] Narayanan, A. , Wu, X. ,and Yang, Z. R. (2002) Mining viral protease data to extract cleavage knowledge. 18(1). pp. 5-13

[9] Brusic, V. Rudy, G. and Harrison, L. C. (1998) MHCPEP, a database of MHC-binding peptides：update 1997 . Nucleic Acids Research, 26(1), 368-371.

[10] Adams,H.P. and Koziol,J.A. (1995) Prediction of binding to MHC class I molecules. J. Immunol. Methods, 185, 181–190.

[12] Brusic,V., Rudy,G., Kyne,A.P. and Harrison,L.C. (1996) MHCPEP—a database of MHC-binding peptides: update 1995. Nucleic Acids Res., 24, 242–244.

[13] Brusic,V., Rudy,G. and Harrison,L.C. (1994) Prediction of MHC binding peptides using artificial neural networks. In Stonier,R. and Yu,X.H. (eds), Complex Systems: Mechanism of Adaptation. IOS Press, Amsterdam, pp. 253–260.

[14] Ma,Q., Wang,J.T.L. and Wu,C.H. (1998) Detection of Alu sequences in DNA: a neural network approach. In Proceedings of the Fourth Joint Conference on Information Sciences, Vol. I, pp. 392–395.

[15] 戴文彬, 黃威豪, 賴岱瑛, 沈倍伊. (2003) "Artificial Neural Networks for Affinity Prediction of MHC-Peptide Binding", Technical Reports, Dept. Computer Science, Chinese Culture Univ.

[16] ProPred-I, http://www.imtech.res.in/raghava/propred1/index.html

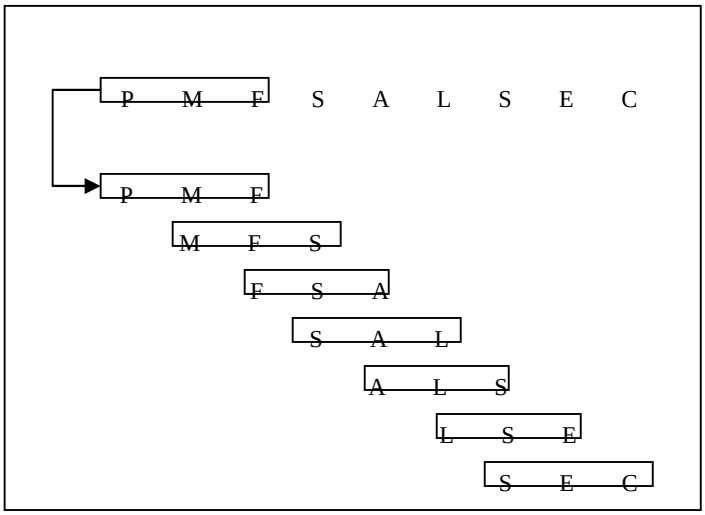

Figure 1 以會結合的胜肽鏈序列中第一序列 PMFSALSEC 為例，可拆成七組氨基酸序列。

amino acid	編碼方式
A	10000000000000000000
C	01000000000000000000
D	00100000000000000000
E	00010000000000000000
F	00001000000000000000
G	00000100000000000000
H	00000010000000000000
I	00000001000000000000
K	00000000100000000000
L	00000000010000000000

amino acid	編碼方式
M	00000000001000000000
N	00000000000100000000
P	00000000000010000000
Q	00000000000001000000
R	00000000000000100000
S	00000000000000010000
T	00000000000000001000
V	00000000000000000100
W	00000000000000000010
Y	00000000000000000001

Figure 2 氨基酸的編碼表。

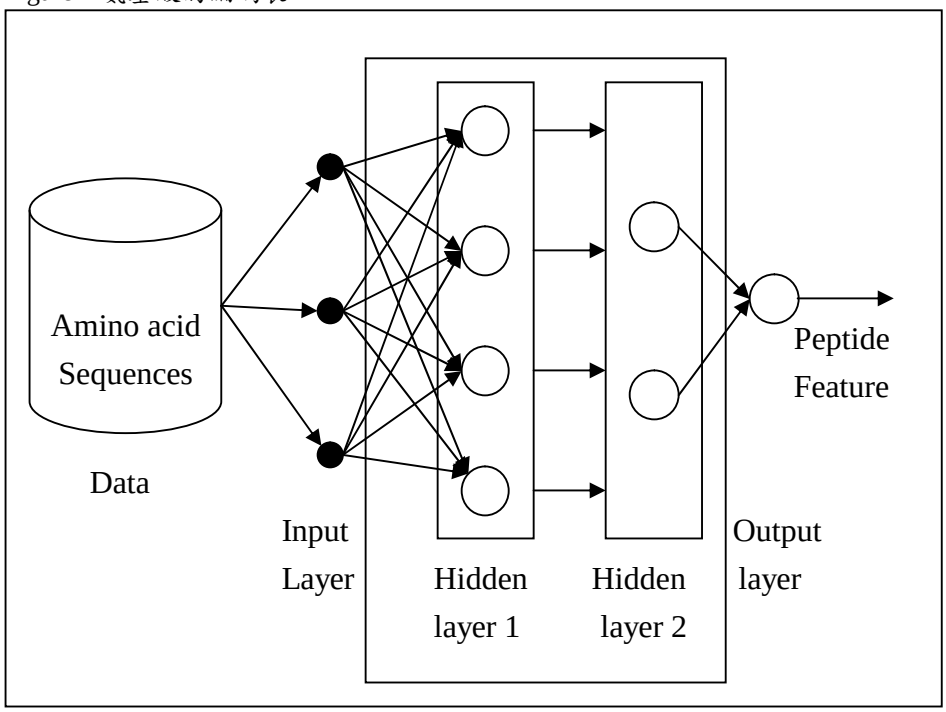

Figure 3 第一階智慧型運算模型。

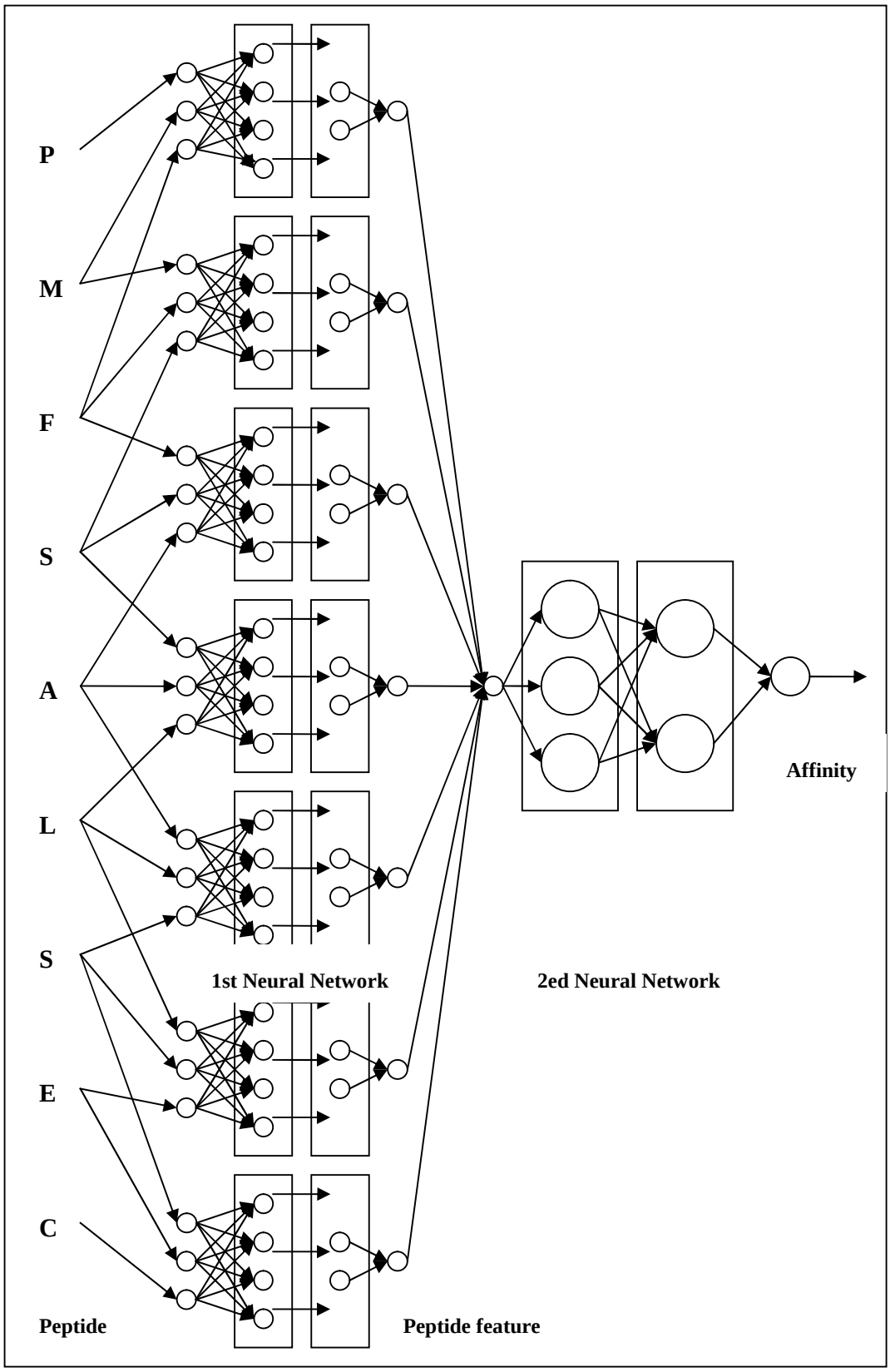

Figure 4 第二階智慧型運算模型。