iIS 資訊科學簡訊 Update

Newsletter of the Institute of Information Science, Academia Sinica, Taiwan

2016 年

所長的話

資訊科學研究所於 1977 年開始設立籌備處，歷經五年籌備，於 1982 年 9 月正式成立研究所，是中央研究院數理組十個單位之一。本所的研究重點分為八大方向，包含生物資訊、電腦系統、資料處理與探勘、多媒體技術、語言與知識處理、網路系統與服務、程式語言與形式方法、計算理論與演算法。

本所並於 2003 年配合中研院國際研究生院 (TIGP)，開始招收生物資訊學程博士班，致力延攬及培訓此一新領域之國際人材多年有成，已招募 50 餘名學生，對生物資訊的研究有很大的助益。2014 年起，社群網路與人智計算新學程博士班開始招生，目前已招募 20 餘名學生。

所內資深研究人員皆指導或共同指導來自全國各地頂尖大學之碩、博士生，並與這些大學保持緊密的研究合作關係，與國內外資訊業界亦有極為良好之互動關係。

所長：許聞廉

副所長：王新民
劉庭祿

研究群主持人：
施純傑教授
生物資訊實驗室
吳真貞教授
電腦系統實驗室
陳孟彰教授
資料處理與探勘實驗室
廖弘源教授
多媒體技術實驗室
許聞廉教授
語言與知識處理實驗室
陳伶志教授
網路系統與服務實驗室
穆信成教授
程式語言與形式方法實驗室
呂及人教授
計算理論與演算法實驗室

時光飛逝，自上一期通訊發行以來一年很快又過去了。這段期間，本所同仁持續致力於資訊相關領域的研究工作，獲得豐碩的研究成果，除蔡懷寬博士及楊得年博士通過升等為研究員，另有多位同仁榮獲國內外眾多獎項與殊榮，包括前所長及合聘研究員李琳山教授獲選本院第 31 屆院士及第八屆總統科學獎、楊得年副研究員獲得 2015 年中央研究院年輕學者研究著作獎、張原豪副研究員獲得科技部 104 年度吳大猷先生紀念獎、鐘楷閔副研究員獲選本院 105 年度前瞻計畫、葉彌妍副研究員與台大林守德教授、政大蔡銘峰教授共同指導台大學生與中研院助理參加 2016 WSDM Cup，八強中榮獲冠軍、第四、及第八名。

本期通訊的深度報導一介紹宋定懿研究員所執行的生物資訊之蛋白質探索計畫，這是國際合作計畫的一環，宋研究員所屬的台灣團隊認領人體第四號染色體的分析工作，讓台灣再次躍上國際學術的舞台；深度報導二介紹張原豪副研究員的嵌入式檔案系統之空間利用率最佳化研究，利用動態檔尾合併技術，可創造額外 64% 使用空間；實驗室特寫介紹計算理論與演算法實驗室進行中，由鐘楷閔副研究員帶領之古典與 (後) 量子理論密碼學研究；創意發想由蔡懷寬研究員為大家介紹如何運用生物資訊方法，研究生物體內轉錄調控機制；人物焦點則介紹致力於開發新世代語言分析技術的馬偉雲助研究員。

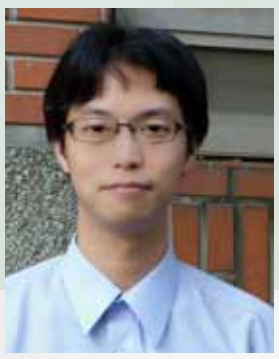
本所希望藉由本期通訊的內容與您交流，並請您一如往常提供寶貴的建議及回響，更竭誠歡迎您抽空蒞臨指教。

名人榜

副研究員葉彌妍博士與台大林守德教授、政大蔡銘峰教授共同指導台大學生與中研院助理參加 2016 WSDM Cup，八強中榮獲冠軍、第四、及第八名。

蔡懷寬博士升等為本所研究員，自 104 年 11 月 16 日起生效。

副研究員鐘楷閔博士獲選本院 105 年度前瞻計畫。



副研究員楊得年博士榮獲 2015 年中央研究院年輕學者研究著作獎。

合聘研究員李琳山教授獲選中央研究院第 31 屆院士，並榮獲第八屆總統科學獎。

副研究員張原豪博士榮獲「科技部 104 年度吳大猷先生紀念獎」。

卓越講座系列

October 2016

Revisiting Control/Data Plane Separation in Software Defined Networking

Giuseppe Bianchi

Professor, University of Roma Tor Vergata

Networking, wireless networks, network security

生物資訊之蛋白體探索：
認領第四號染色體　台灣躍上國際學術舞台

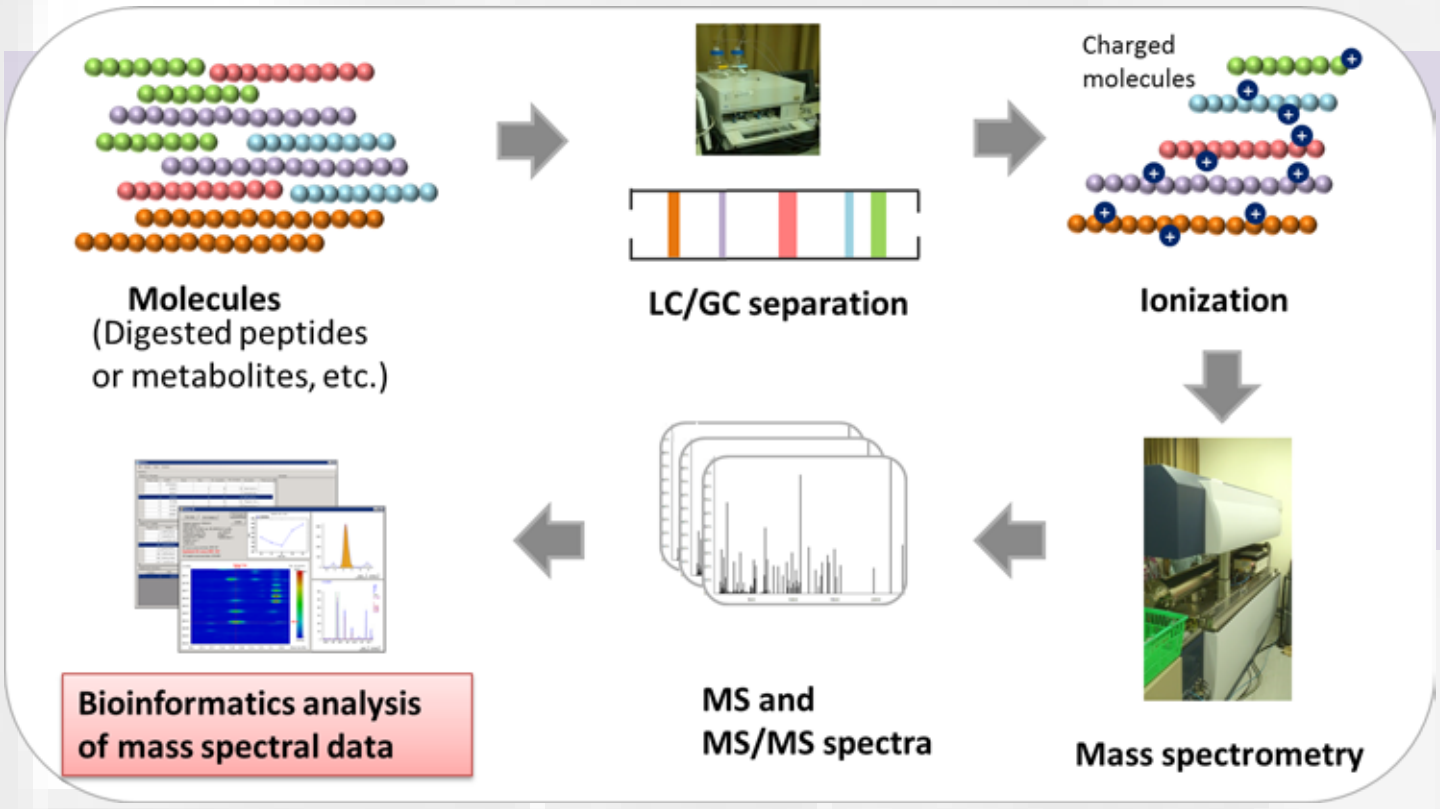
計畫主持人
宋定懿

蛋白質是基因轉譯的最終產物，在生物體內執行各種不同的功能。在醫藥研究上，人類的蛋白質是最主要的藥物標的。因此在後基因體時代，針對生物體進行大規模蛋白質分析的蛋白體學逐漸受到重視。隨著質譜儀器及實驗技術的精進，Aebersold 和 Mann 兩位學者 2003 年在 *Nature* 期刊上發表一篇「奠基於質譜儀資料分析的蛋白體學」重要論文，象徵蛋白體學跨入重要里程碑；接著 2005 年 Ong 與 Mann 兩位學者在 *Nature Chemical Biology* 期刊上發表了另篇論文，宣告蛋白體學研究邁向定量的研究。近十多年來，液態層析搭配質譜儀的實驗技術成為蛋白體學最重要的實驗方法；此方法首先將樣本中的蛋白質分解成胜肽，利用液態層析法分離，再送入質譜儀分析。研究者藉由質譜儀所產生的圖譜大數據，來鑑定生物體、組織或細胞內在不同狀態的蛋白質，及分析不同狀態下蛋白質表現量的差異。但由於質譜儀斷裂機制、樣品的品質及高複雜度、實驗中難免的背景雜訊、數個大分子同時流出或斷裂等因素，造成實驗資料難分析。本實驗室於 2003 年底和本院化學所陳玉如教授長期合作，投入質譜大數據分析的生物資訊研究，提出演算法，開發完成三個不同的蛋白定量軟體 Multi-Q、MaXIC-Q 與 IDEAL-Q，提供學界下載使用。

台灣認領人體第四號染色體

隨著蛋白體學研究的蓬勃發展，人類蛋白體學組織在 2011 年提出類似人體基因體計畫的人體蛋白體計畫 (Chromosome-centric Human Proteome Project, C-HPP)，希望描繪人體從基因轉譯過來的所有蛋白質。此計畫由各國負責不同染色體中蛋白體的分析，台灣團隊由陳玉如教授負責召集，認領第四號染色體以在國際學術舞台發聲；我們實驗室負責台灣團隊的生物資訊研究。2013-2015 年 C-HPP 重要的目標是找尋「失蹤蛋白質」，儘管這些蛋白質的序列已由 DNA、RNA 經過嚴謹的推論得知，但在抗體或質譜實驗上未被看到。失蹤蛋白質未在實驗上被確認，可能的原因很多，如：量少、只在某些過渡的狀態下出現、實驗困難。

陳玉如教授實驗室以多個肺癌的細胞株進行質譜實驗，質譜大數據再用定性軟體（如：Mascot）鑑定出其中的蛋白質。儘管軟體可能鑑定出某些失蹤蛋白質，但這不表示就找到了，必須滿足嚴謹的條件以及使用標靶實驗才算確認。根據軟體產生的蛋白質鑑定結果，我們實驗室以資訊專長擔負嚴謹的資料分析。由於大多數蛋白質定性軟體未提供蛋白質的錯誤發現率，因此我們首先以現有的軟體 (PeptideShaker) 進行蛋白質錯誤發現率的分析，刪除錯誤發現率超過 1％的蛋白質。其次，蛋白質的鑑定是從大量圖譜鑑定出的胜肽推論而來；為求推論的準確性，這些蛋白質必須包



使用質譜儀的蛋白體分析流程。

含至少一條在整個人體蛋白體中的專屬胜肽（即不會出現在其他蛋白質中）被鑑定出來，我們確認是否有專屬胜肽以確保蛋白質的可靠性。我們從實驗資料中可靠地鑑定出 7702 個蛋白質，其中 66% 為膜蛋白。第三，這些蛋白質再和大會提供的失蹤蛋白質名單比對來找出失蹤蛋白質，再針對這些蛋白質的胜肽，進行兩項嚴謹的驗證：(1) 比對 PeptideAtlas 資料庫確認未含這些胜肽，即這些胜肽是新觀察到的；(2) 確定它們並非由其他蛋白質的胜肽產生單點胺基酸變異而來。完成確認後，我們從肺癌細胞中找到 178 失蹤蛋白質，包含 74 個膜蛋白。為了突顯上述嚴謹過程所找到失蹤蛋白質的可靠性，陳玉如教授實驗室針對其中 8 個失蹤蛋白質實驗資料看到的專屬胜肽，以合成產生這些胜肽，再以質譜儀多重反應監測模式驗證它們，因此這些失蹤蛋白質得以確認。此研究成果發表於 *Journal of Proteome Research* (*JPR*) 2015 年主題為 C-HPP 的專刊。此外，我們由資訊專業的角度探討即使在最理想狀況下，即實驗可找到所有蛋白質的胜肽，找尋失蹤蛋白質仍會面臨哪些挑戰。首先，有 77 個失蹤蛋白質沒有專屬胜肽，要確認這些蛋白質非易事。第二，許多高相似度（甚至達 100%）的蛋白質之實驗證據是難以確認。第三，有兩類的蛋白質在轉譯時常發生變異，較難被實驗看到。要尋找失蹤蛋白質，必須另採新的實驗設計或方法。此成果發表於 *JPR* 2015。

蛋白體學研究現在亦發重要，質譜儀大數據的分析處理及蛋白體探索的生物資訊研究更需要資訊專才的投入。歡迎加入！

卓越講座系列

Technology Considerations in Computer Architecture

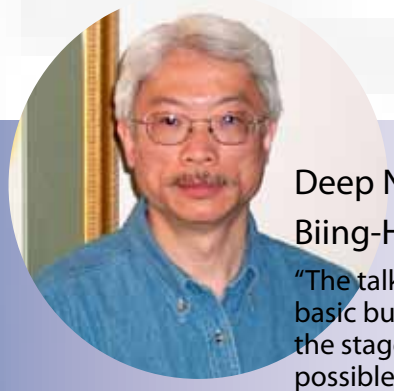
Jean-Luc Gaudiot —September 5, 2016

"….. FPGAs are extremely useful in mobile embedded systems where computing power and energy considerations are major concerns….."

Building Algorithms for the Next Generation Route Planner

Dorothea Wagner — May 17, 2016

"Nowadays, route-planning systems are among the most frequently used information systems. The algorithmic core problem of such systems is the classical shortest-paths problem, which can be solved by Disjkstra's algorithm in almost linear time. However, Dijkstra's algorithm still takes a few seconds in continental-sized graphs, which is too slow for practical scenarios....."



Deep Neural Networks – A Developmental Perspective

Biing-Hwang (Fred) Juang — December 22, 2015

"The talk is given from a developmental perspective with a comprehensive view, from the very basic but oft-forgotten principle of statistical pattern recognition and decision theory, through the stages of problems that are encountered during system design, to key ideas that may lead to possible new advances toward deep learning....."

Building Computing Machines That Sense, Adapt and Approximate

Rajesh K. Gupta — October 27, 2015

"Computing machines today are largely ignorant of the variability in the behavior of underlying components from device to device, chip to chip, and their wear over time — save for thermal sensing in limited energy/power constrained applications....."

Differential Privacy: Theoretical and Practical Challenges

Salil Vadhan — November 10, 2015

"Addressing pressing privacy problems in a variety of domains has attracted the interest of scholars from many fields, including statistics, database management, medical informatics, law, social science, computer security, and programming languages....."

Joint Software-Defined Application-Network Control Plane for Next Generation Real-Time Applications

Klara Nahrstedt — May 25, 2015

"I will argue for a joint software-defined application-network control plane to assist next-generation, real-time applications such as telepresence and teleimmersion. Furthermore, I will discuss OpenSession, the new "Northbound" application-network control plane for multi-stream and multi-site real-time applications, which represents the interaction between the application-level session controller and a Software-Defined Network (SDN) controller. ...."

New Advances in Forensic Identification

Henry Lee — March 23, 2015

"Collection of fingernails is painless, harmless, and convenient. Fingernails can also be found on badly decomposed bodies and body parts, especially in catastrophic incidents. Not only mitochondrial DNA, but also nuclear DNA has been successfully analyzed from fingernail fragments. Fingernail patterns and physical features were extracted by image processing. Features of size, length, and width of fingernails were examined....."

OpenCloud: A Value-added Cloud for Network Operators

Larry Peterson — April 8, 2015

"Network operators are migrating away from purpose-built hardware appliances and moving toward infrastructure that exploits virtualized commodity servers and SDN at the very edge of the Internet, a practice called Network Functions Virtualization (NFV). This talk puts forward a vision for a value-added cloud that demonstrates how network operators can take advantage of cloud technology. It also describes a prototype, OpenCloud, that we are building with Internet2....."

Keeping a Crowd Safe: On the Complexity of Parameterized Verification

Javier Esparza — March 2, 2015

"Many computer systems consist of an arbitrary large number of identical components — a "crowd" — communicating by some means. Examples include distributed algorithms, communication and network protocols, and computer models of biochemical systems. The safety problem for these systems consists of checking that, whatever the size of the crowd, no individual can reach a dangerous state....."

Deep Learning Workshop

December 21–22, 2015

Topics

Deep Neural Networks – A Developmental Perspective

Deep Learning in Computer Vision

Deep Learning in Speech

Frontiers of Communications and Networking Workshop 2015

November 6, 2015

To provide a platform for scholars in the field of communications and networking in Taiwan to interact, collaborate, and share new and trending ideas with each other annually

Topics

Multiuser MIMO Systems: From Rate Adaptation to User Selection ● Some Communication Issues in IoT ● Green Resource Management for Distributed Antenna Systems ● Asynchronous Quorum-based Channel Hopping Schemes for Cognitive Radio Networks ● Enabling Internet of Vehicles with Vehicular Visible Light Communications ● Sharing Experiences on International Academic Services and Research ● Maintaining Competitiveness in the International Academic and Industrial Community

2015 IR Workshop

New Trends and Technologies of Cross-discipline NLP and IR

December 18, 2015

Topics

Jointly Modeling Topics, Events and User Interests on Twitter

A Community-based Method for Valence-Arousal Prediction of Affective Words

Active Learning by Learning

Capture the Great Moment of Social Network Efficiently and Effectively

A Novel Social Influence Model based on Multiple States and Negative Social Influences

Using Non-Verbal Information to Augment Designs of Language-based Interactions

資訊杯桌球賽於 6 月 23-24 日舉行

資訊杯投籃大賽於 6 月 20-21 日舉行

10 月 31 日院區開放活動，人潮踴躍

資訊杯羽球賽於 6 月 22-23 日舉行

計算理論與演算法實驗室

古典與(後)量子理論密碼學

鐘楷閔
副研究員

計算理論與演算法實驗室的理論密碼學團隊主要致力於理論密碼學的研究，研究課題包含古典(非量子)密碼學到(後)量子密碼學。近幾十年來，密碼學的發展已經超越其原本安全傳輸的目的，可以安全的實現許多更複雜的任務。今日，密碼學無所不在的穿梭在我們現今的生活中，並且不知不覺地在使用它，例如：電子郵件、行動電話、SSL、電子商務與電子投票等。

在本文中，我們將試著以淺顯易懂的方式介紹密碼學最近的重大進展。我們將以加密法為例，先簡介其概念，再介紹近十年才被提出與達成的兩種進階加密法：全同態加密 (Fully Homomorphic Encryptions) 以及功能加密系統 (Functional Encryptions)。這兩種加密法都可以允許我們在不知道資料內容的情況下對加密的資料上做運算！我們將藉由幾個實際應用和類比的方式來介紹這些概念。最後，我們將簡述我們團隊近期的研究發展以及我們的研究方式。

在歷史上，密碼學早期主要提供安全的傳輸通道來傳輸私密資料(如軍事或是政府單位)，如同電影 -- 模仿遊戲所講述的故事 [1]。考慮 Alice 想要送一個私密訊息 m 去給 Bob，而訊息的傳輸可能會被一個竊聽者 Eve 所竊聽。因此，Alice 不能直接寄送訊息，否則會被 Eve 所聽到。安全傳輸的目標就是確保 Bob 可以正確地收到 Alice 傳來的訊息，同時確保 Eve 無法學習到訊息的相關內容。舉例來說，在線上消費時，信用卡以及安全碼的資訊必須傳送到銀行，而安全傳輸可以提供訊息的保護以防範中間的竊聽者。

公開金鑰加密法 (public key encryption, PKE) 是一個能達到此目的密碼機制。一個 PKE 主要由三個演算法組成 (KeyGen, Enc, Dec)。一個接收者可執行金鑰生成演算法 KeyGen 來生成一組金鑰對，一把為任何人都可以看到的公鑰 pk、另一把為只有接收者可以擁有的私鑰 sk。任何人可以透過公鑰 pk 來執行加密演算法 Enc 將訊息 m 製成密文 $\mathrm{ct} = \mathrm{Enc}_{\mathrm{pk}}(m)$；擁有 sk 的人可以執行解密演算法 Dec 將 ct 回復成訊息 $m = \mathrm{Dec}_{\mathrm{sk}}(\mathrm{ct})$。

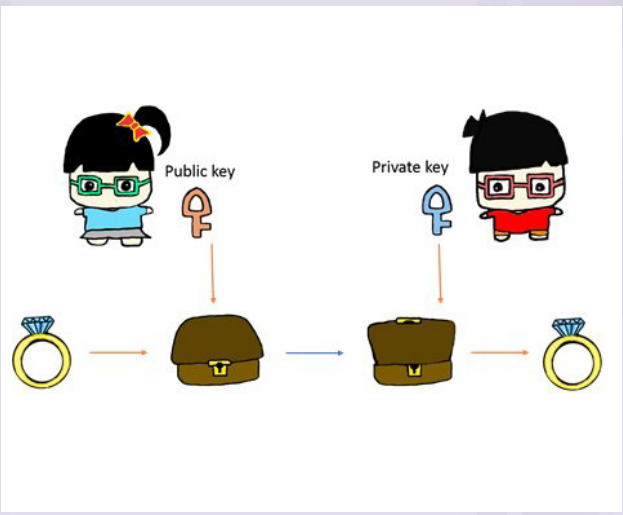
利用 PKE，Bob 可先使用 KeyGen 產生 (pk, sk)，將公鑰 pk 送給 Alice。Alice 可以利用 Bob 的公鑰 pk 來加密想傳送的訊息 m 來產生密文 $\mathrm{Enc}_{\mathrm{pk}}(m)$。只有握有私鑰 sk 的 Bob 可以解密而得到 m。沒有 sk 的人無法學習到訊息 m 的相關資訊。因此，雖然竊聽者 Eve 能學到 ct，因為 Eve 不知道 sk，能保證 Eve 不會學到任何有關 m 的資訊。

如何嚴謹的定義所謂的 “Eve 從 ct 學不到任何有關 m 的資訊” 的安全性其實是非常不容易的事情。 事實上，2012 年獲得圖靈獎 (A. M. Turing Award，資訊科學界的諾貝爾獎) 得主 Shafi Goldwasser 和 Silvio Micali 的獲獎主要貢獻之一就是提出正確的加密法的安全性定義。(有興趣的讀者可參考網路上的資料 [2, 3, 4] 來深入了解密碼學。) 在本文我們將不深入介紹如何嚴謹的定義安全性，但試著以類比的方式(如圖一)來幫助讀者理解：我們可以想像 sk 是一個實體的鑰匙，而 pk 是一個可用來製造箱子的模子，而製造出來的箱子只能以 sk 來開鎖。想像 Alice 要傳送的 “訊息” 是珍貴的珠寶。Alice 想安全的寄送珍貴的珠寶給 Bob，可以利用 Bob 所提供的模子製作出箱子，把珠寶鎖在這箱子裡，將上鎖的箱子寄給 Bob。這樣可以保證只有 Bob 可以開鎖取得珠寶，而無法被其他人竊取。

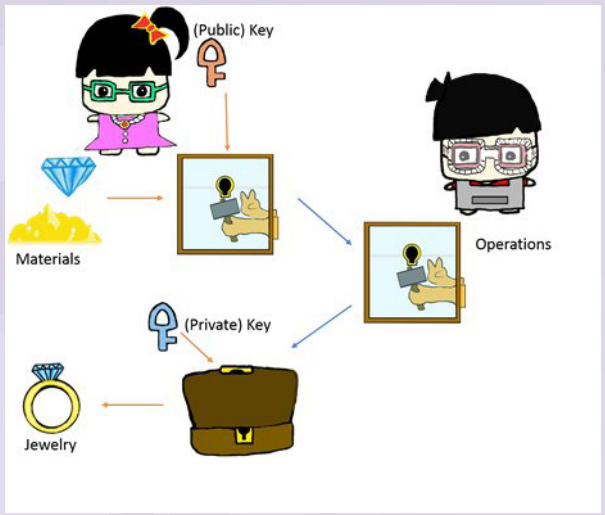
全同態加密法

在現今雲端運算發達的時代，一個常見的應用是外包我們的資料與計算給雲端伺服器 (如：Amazon EC2)。然而，當資料是較敏感的內容時(如：個人隱私資料、病歷紀錄)，外包給無法完全信任的伺服器可能會有損隱私。抽象來說，考慮用戶端想要外包其資料 m 與想計算的函數 f 給伺服器(如:對病歷進行統計分析)，我們是否可以在不洩漏資料 m 給伺服器的情況下讓伺服器幫我們算出答案 $f(m)$?

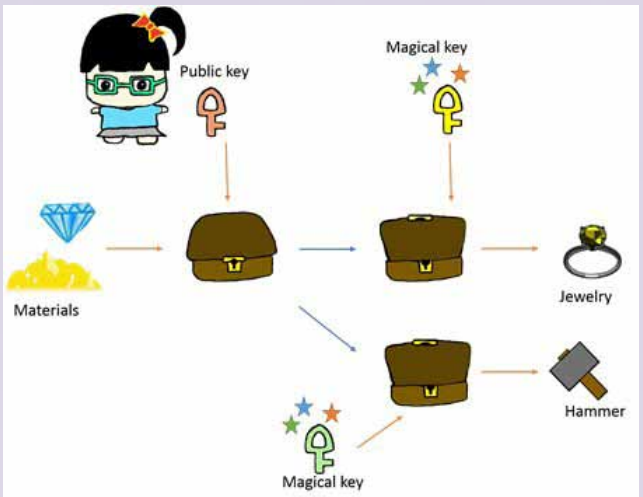
這個看似矛盾的任務，能由所謂的全同態加密法 (fully homomorphic encryption, FHE) 所達成的。FHE 是一個能允許在加密後的密文上做計算的進階加密法。具體來說，FHE 比一般的 PKE 多了一個同態運算演算法 Eval。Eval 演算法可對給定



公開金鑰加密法：安全傳輸（圖一）。



全同態加密法：Alice 的珠寶店（圖二）。



功能加密法：Alice 的魔法鑰匙（圖三）。

的密文 $ct = Enc_{pk}(m)$ 進行給定的函式 f 運算來得到一個新的密文 $ct' = Enc_{pk}(f(m))$，其加密的訊息內容為原始訊息 m 在 f 運算下的結果 $f(m)$。注意到 Eval 演算法不需要使用私鑰 sk，而在沒有 sk 的情況下，執行 Eval 演算法的人雖然能算出 ct'，但仍然無法學到 ct 和 ct' 裡面的訊息。這種神奇的加密法的可能性在九零年代就被提出，但直到 2009 年才由 Craig Gentry 構造出第一個 FHE 加密法 [5]。這是當年密碼學的重大突破。回到外包計算的問題，運用 FHE，用戶端可傳送加密過的密文 $ct = Enc_{pk}(m)$ 給伺服器，而伺服器端可以利用 Eval 演算法來得到 $ct' = Enc_{pk}(f(m))$，再將 ct' 送回給用戶端。用戶端可用 sk 解密而得到 $f(m)$。而伺服器因為不知道 sk，無法學習到關於 m 的任何資訊。

延續之前盒子的類比（如圖二），我們可以想像 FHE 的 ct 是一個裝有原料（原始訊息）的手套箱，可以允許對裡面的元件做組裝（同態運算）。Alice 放入製作珠寶的原料至手套箱並上鎖。並將箱子送給一個工人 Bob 來幫她組裝成珠寶。Bob 可以在看不到原料的情況下幫 Alice 組裝珠寶，將組裝後的手套箱送還給 Alice。Alice 可以解鎖打開箱子得到組裝後的珠寶。

功能加密法

全同態加密法給了我們一個在不解密的情況下可以針對密文（即加密資料）做計算的能力。而“功能加密”(functional encryption, FE) 則允許我們對密文進行精密的存取控制：同一份加密資料給擁有不同“功能金鑰”(functional key) 的使用者，可以解密得到不同的關於原始訊息的資訊。舉例來說，想像一家醫院維護了一個加密的病例資料庫 $ct = Enc_{pk}(m)$，授權不同部門 D_i 去對資料庫進行不同的統計分析 f_i，但基於隱私的考量，不允許部門學習到其特定統計分析結果 $f_i(m)$ 之外的，關於這個資料庫的資訊。FE 可以讓每個部門 D_i 擁有各自對應的功能金鑰 $sk[f_i]$，D_i 可以使用 $sk[f_i]$ 解密 ct 來學到其統計分析的結果 $f_i(m)$，但無法學到 $f_i(m)$ 以外的任何資訊。沿用之前箱子的類比（如圖三），現在箱子可以搭配不同的“魔法鑰匙”，使用不同的魔法鑰匙開箱會得到不同的產品。

研究方向：密碼學的計算模型

注意到前面介紹的兩種進階加密法都和計算有關係，許多其他重要的密法學法展，如安全計算 (secure computation)、程式模糊化 (obfuscation) 也都與計算息息相關。但我們要如何描述一個計算（例如先前舉例的統計分析）？對有資訊科學背景的人來說，很自然的會想到用寫程式的方式，例如用 C++ 或是 Python 等的程式語言去實作計算的演算法。然而，近幾十年密碼學上的發展大多採用電路模型 (circuit model) 來描述一個計算（如圖四）。數位電路主要是由 NOT、AND 或 OR 等邏輯閘組成，通過這樣的組合

（文轉至第 16 頁）

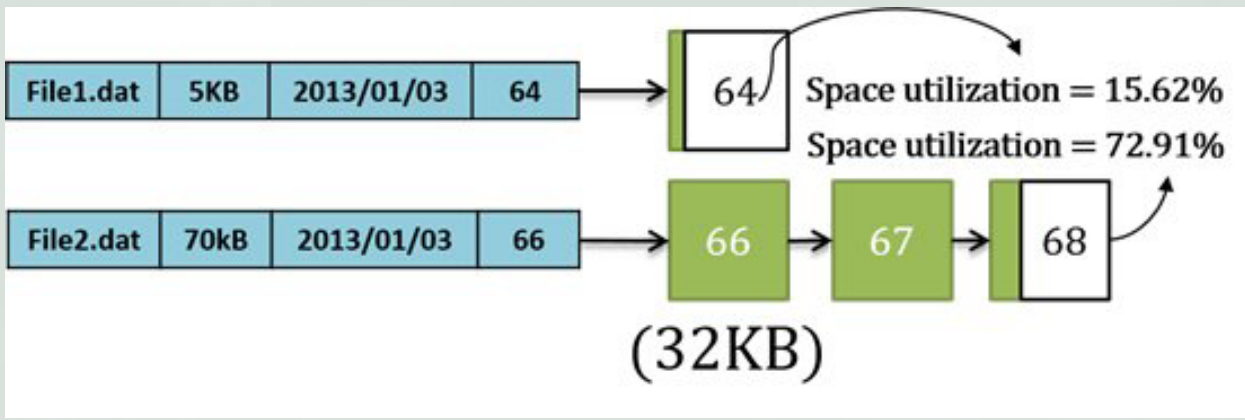
創造額外 64% 使用空間　動態檔尾合併技術

最佳化嵌入式檔案系統之空間利用率

張原豪　副研究員

嵌入式行動運算系統，此系統平台通常是由電池供電，並被廣泛使用在許多日常應用中。傳統上，由於製造成本以及耗能的限制，嵌入式系統平台通常都會配置有限計算能力的處理器、有限空間主記憶體 (RAM) 以及小量的第二儲存裝置。然而，隨著科技的進步，進年來嵌入式行動運算系統具有較好的運算能力，也因此部份嵌入式系統採用嵌入式檔案系統 (Embedded File System) 來簡化複雜的檔案管理程序。舉例來說，安卓 (Android) 檔案系統不再使用日誌型 (Log-based) 檔案管理方式（例如：yaffs2）。取而代之，嵌入式行動運算系統，如：行動電話、嵌入式消費型電子產品，藉由一個簡化的檔案系統來管理在快閃記憶體 (flash memory) 儲存裝置中的檔案。然而，現行的檔案系統為了方便管理儲存空間，通常會將儲存空間配置成多個基本儲存單位，例如：叢集 (Cluster)，而一個基本儲存單位往往是幾千位元組 (Kilobytes)。這樣的空間配置方式導致在儲存小檔案或檔尾時，會有較低的空間利用率問題，而這個問題也成為在設計有限儲存空間的嵌入式儲存系統中的重要議題。

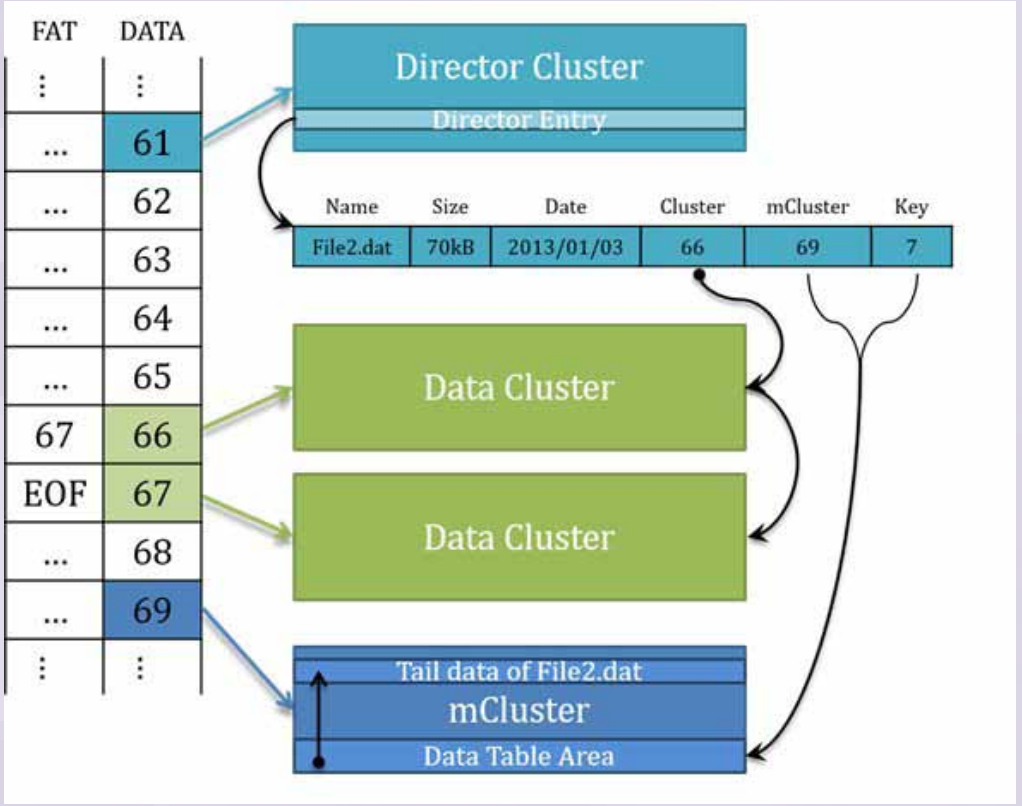
由於大部份的嵌入式檔案系統，無論檔案多小，配置空間給單一儲存檔案都是以一個叢集為最基本的配置單位。其中這類檔案系統著名的有 FAT 和 ext3/ext4 檔案系統，其中 FAT 由於夠簡化，並可以使用在資源限制的嵌入式系統，而 ext4 則是 Linux 作業系統預設的檔案系統。然而，這類以叢集為基本存儲單位來分配空間的檔案系統，應用在嵌入式系統上會使其儲存之空間利用率嚴重下降，特別是應用在感測節點和控制系統，這類應用系統裝置上所有儲存的資料大多是小資料。此外，這樣的狀況會在更大的存儲裝置上更加嚴重。因為當存儲空間上升，叢集的大小也會上升。舉例來說，為了要管理 32GB 以上的儲存空間，FAT 檔案系統將其叢集大小成長至 32KB。這樣一來，空間利用率就會變的更低。舉例來說，當叢集大小為 32KB 時，若有一個小檔案只有 5KB 的內容，在這樣的情況下，此檔案就會浪費 27KB 的空間，而在相同的情況下，一個 70KB 的檔案（由於系統還是會分配一個完整的叢集空間給其檔尾，這樣



低空間利用率問題（圖一）。

就浪費了 26KB 的空間在檔尾的叢集空間中（請參照圖一）。

有鑑於此，我們提出動態檔尾合併技術來解決嵌入式檔案系統空間利用率之問題。其設計的主要目標有兩個方向：(1) 藉由有限的效能負擔來合併檔尾，並優化儲存系統空間利用率，及 (2) 最小化檔尾合併技術的內 / 外部空間碎裂問題。為了達成以上目標，動態檔尾合併技術定義了一個新型態的叢集空間，名為微叢集 (mCluster)，並動態地合併檔尾資料在微叢集的空間中。一個微叢集分為兩個部份，資料區域和資料表，資料區域是用來儲存檔尾資料，而資料表則是用來維護檔尾資料在微叢集裡的位置資訊。此外，微叢集的相關狀態與資訊則是維護在微叢集分配表中，用來進行微叢集的管理與空間分配。藉由微叢集與其分配表，我們所提出的動態檔尾合併技術應用了一個全新的管理機制，叫「微叢集基礎管理機制」，來管理分配微叢集與從微叢集分配位元組層級的空閒空間給檔尾資料。此一動態檔尾合併技術可以利用微叢集這樣的資料結構最小化內部碎裂的問題。與此同時，考量到檔尾大小會在系統執行時產生改變，我們所提出的微叢集基礎管理機制採用了一個動態鏈結的策略，此策略可以不用任何的檔案複製成本，來動態地聚集分散在不同微叢集但屬於相同檔案的檔尾資料。因此，可以有效的解決外部碎裂問題。



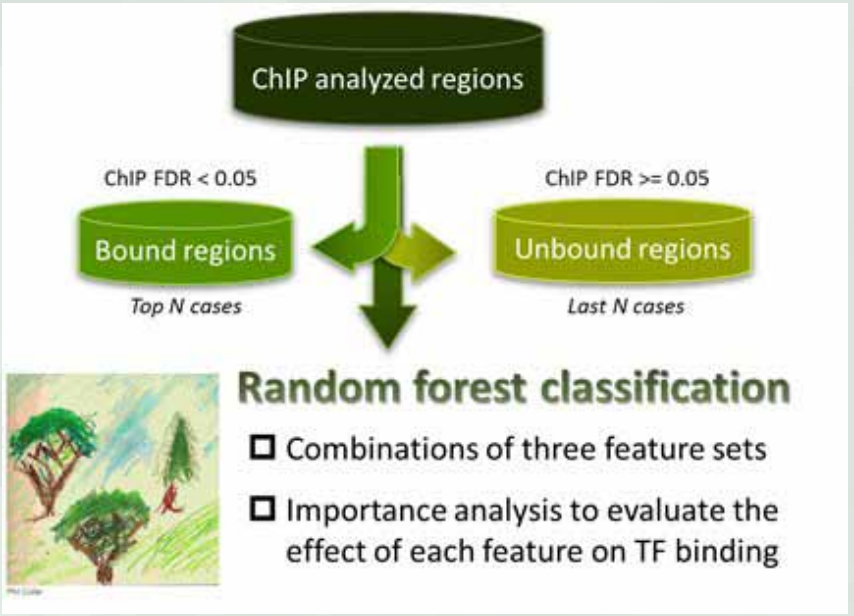
動態檔尾合併技術範例（圖二）。

圖二呈現了一個整合於 FAT 檔案系統的動態檔尾合併技術範例。在此範例，假設一個基本叢集空間是 32KB，而 File1.dat 是一個 70KB 的檔案。File1.dat 中非檔尾的資料會被完整儲存於叢集 64 和 66 當中，而資料儲存於兩個 32KB 的叢集中後，還剩餘 6KB 的檔尾資料，此檔尾資料存在於叢集 69 中，而叢集 69 也就是一個微叢集。為了支援動態檔尾合併技術，原本用來儲存檔案屬性的目錄項 (directory entry) 也利用兩個原本空閒的空間來新增兩個欄位，為「微叢集編號」與「索引值 (Key)」，以標示出檔尾資料儲存的位置。其中微叢集編號是指出那一個微叢集儲存其檔尾資料，而索引值則是用來在微叢集結構內的資料表 (data table) 搜尋其檔尾在此微叢集中所儲存的位置。舉例來說，File1.dat 的目錄項中所記錄的微叢集編號與索引值 (Key)，表示 File1.dat 的檔尾儲存於微叢集 69 (mCluster 69)，並且利用 7 這個值可以在叢集 69 的資料表內尋找到其檔尾儲存的位置。值得注意的是，叢集 69 可以用來儲存不同檔案的檔尾，此外，在空間分配上叢集是利用位元組 (Byte) 為單位，而非一個子叢集。為了驗證動態檔尾合併技術在空間利用率上的效能，我們整合了此技術於 32KB 儲存叢集的 FAT 檔案系統，並且實作在 Linux 系統平台上。從實驗中，我們發現相較於沒有動態檔尾合併技術的檔案系統，動態檔尾合併技術可以省下高達 64% 的空間使用量。

總結來說，在此項研究中，動態檔尾合併技術被引用來優化嵌入式檔案系統之空間利用率。在此技術中，較特別的是一個以叢集基礎管理方式被提出，以最小化效能成本和空間碎裂的方式來動態地合併在微叢集結構中的檔案檔尾資料。與此同時，考量到系統執行時檔案大小的改變，一個動態檔尾鏈結進一步的被提出，此方法可以不用任何資料搬移的負擔，就能動態地聚集分散在不同微叢集但屬於相同檔案的檔尾資料，經由實作 Linux 作業系統的檔案系統中所作的一列系評估及驗證，可以發現此一技術可以在有限的效能負擔下，有效的減少檔案系統的空間浪費。

運用生物資訊方法
研究生物體內轉錄調控機制

蔡懷寬　研究員



應用隨機森林分類器探討轉錄因子結合機制。將分子生物問題與實驗數據轉化為資訊科學擅長解決的分類問題，並從解答過程中更深入的探討分子機制（圖一）。

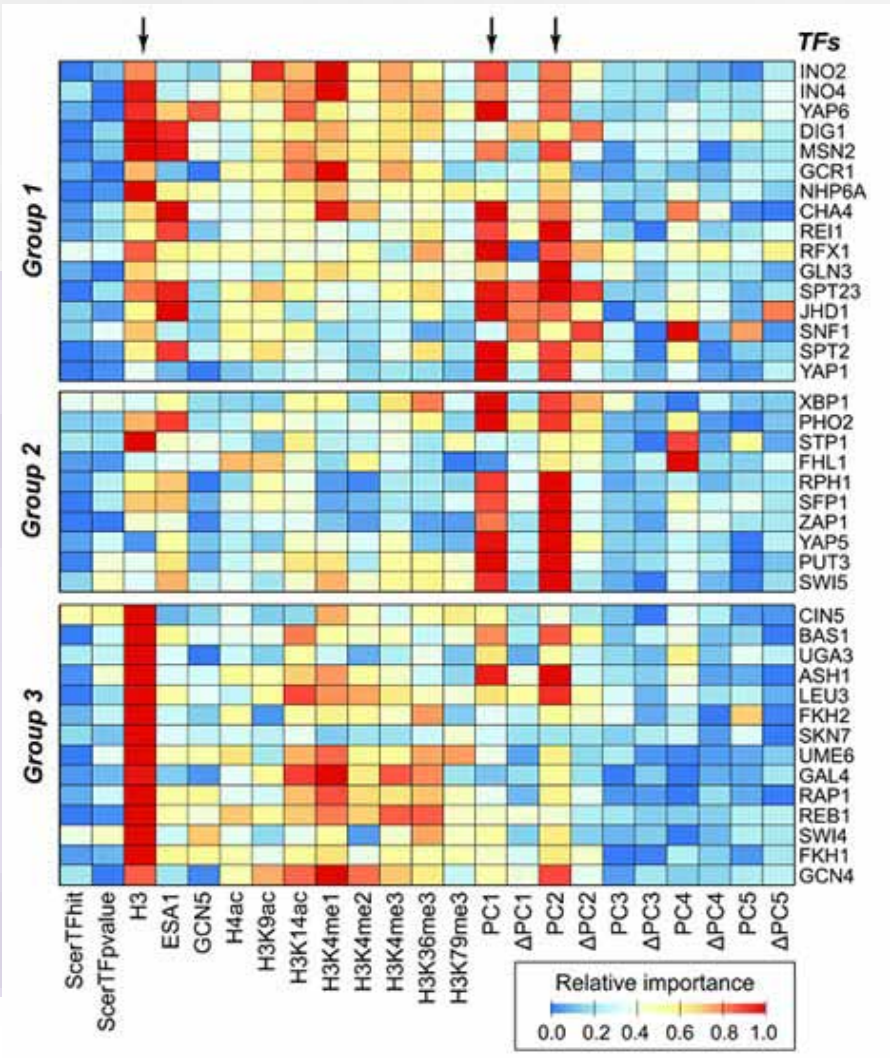
生物體內的轉錄調控機制對生命體的運作極為重要，藉由轉錄因子與其結合位置的動態結合，基因得以精確地對生物過程與周遭環境完成適當反應，因此辨識轉錄因子之結合位置，從而量化細胞內的基因調控情形，一直是生命科學中重要的課題。然而，精確辨識轉錄因子之結合位置是項大難題，因為大多數的結合位置不僅很短（大約 5-20 個鹼基對，但人類的基因體長度超過三十億個鹼基對），而且具有高度退化的特性（簡單來說，轉錄因子可以與類似序列的 DNA 結合，不須完全相同）。隨著生物科技的進步，包括 EMSA、SELEX、PBM 以及 ChIP 等許多生物實驗技術已被發展來偵測辨識轉錄因子之結合位置。這些方法提供生物體內轉錄因子與 DNA 交互作用的證據，以及在生物體外量測轉錄因子與 DNA 結合的親和力。然而到目前為止，由於這些技術仍有高成本與高耗時的問題，該如何在全基因體中大量快速且精準的預測（尋找）轉錄因子之結合位置，仍舊是一項棘手的課題。而生物資訊，正提供了一個解決的可能方向。

目前已有許多計算生物學研究預測轉錄因子的結合位置，這課題可以簡化為特殊序列的尋找問題，也就是從給定的一組生物資料中（例如某轉錄因子目標基因的啟動子序列）尋找特殊的 DNA 序列片段。這些片段通常利用位置權重矩陣 (PWM) 來表示，位置權重矩陣是利用已知的結合位置產生一矩陣，可利用該矩陣來代表結合序列的強度。目前用來尋找轉錄因子結合位置之特殊序列的演算法主要可分為兩種，分別為列舉法以及概率法。列舉法主要精神為分析所有字串出現的頻率，然後找出經常出現的字串作為基礎來生成位置權重矩陣。而概率法之原理則是先將給定的生物字串建立多程序列比對 (MSA)，然後利用機器學習方法（例如 EM 演算法或是 Gibbs

抽樣分析法）最佳化多程序列比對與位置權重矩陣的參數。

近日有賴於高通量實驗技術如次世代定序 (NGS) 的蓬勃發展，數種過去技術無法全面探測的相關資訊得以納入轉錄因子結合之研究，而隨著大量全基因體的資料累積，越來越多的證據顯示 DNA 序列並非決定轉錄因子結合的唯一因素，其中最引人注目的是研究顯示許多偽陽性的轉錄因子結合位置（也就是雖滿足特殊序列的位置加權矩陣但實際上不與轉錄因子交互作用的位置）發生原因是其所在的染色質不具通透性 (inaccessibility)，此外亦有研究指出染色質通透性與轉錄因子結合之親和力有高度相關性，這些研究均顯示染色質通透性可能是控制轉錄因子結合與否的重要指標。而染色質狀態與 DNA 結構特性是目前已知與染色質通透性高度相關的基因體特徵，近來雖然已有研究利用這兩種特徵來預測轉錄因子結合位置，但目前探討各個特徵對於預測的效果付之闕如，且並沒有任何預測方法能成功整合以上所有特徵來預測轉錄因子結合。因此我們以酵母菌為模式生物，以機器學習技術（這裡我們用了隨機森林分類器，如圖一）探討各特徵對轉錄因子結合之影響力。

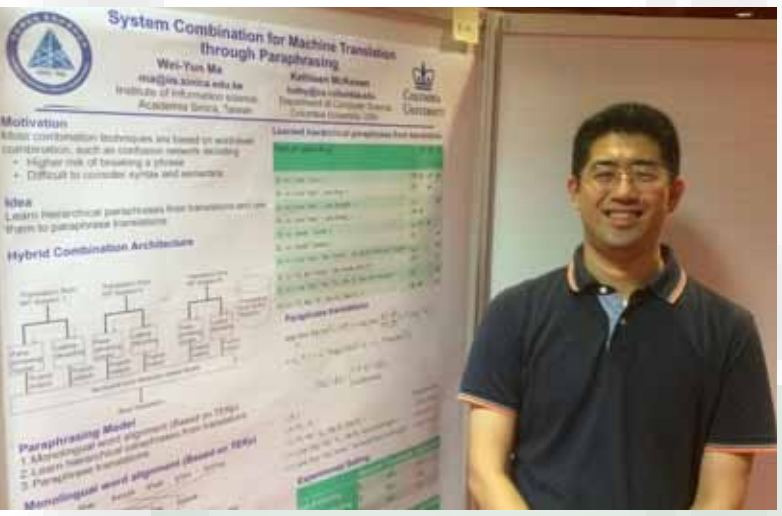
我們發現在預測轉錄因子結合時，單獨使用染色體狀態或 DNA 結構特性建立的分類器會比單用 DNA 序列要好，而同時考慮染色體狀態與 DNA 結構特性會使分類效果顯著提升，充分顯示這兩種特徵互補的效果。進一步觀察則發現，不同特徵對於預測不同轉錄因子的結合位置時大相逕庭，顯示不同的轉錄因子有著相異的結合機制。此外，我們發現三項可直接自 DNA 序列轉換而得的特徵 (nucleosome occupancy, major groove geometry, and dinucleotide free energy，如圖二）對轉錄因子結合有最顯著的影響，而整合該三項特性的資訊方法亦能得到良好的轉錄因子結合預測結果。換言之，對於任何已定序之物種，我們皆有機會能單就其 DNA 序列，預測轉錄因子結合的可能區域。該研究之相關成果發表於 2015 年 PLoS Computational Biology, 11(8), e1004418，共同作者包括了我實驗室的蔡宗曄博士以及美國密西根州立大學的徐新漢博士。



DNA 序列、染色體狀態、與 DNA 結構特性三大類特徵各自對不同轉錄因子結合位置預測的重要性。圖中箭頭所指為對於所有轉錄因子都相對重要的三項特徵（圖二）。

語言分析技術 — 開發者新世代

助研究員　馬偉雲



於 EMNLP 會議發表研究成果。

何因緣進入中研院，並選擇本研究單位

其實早在 2001 年我已經以研究助理的身份加入資訊所，從事自然語言處理的相關研究，而在赴美留學多年後，才又以助研究員的角色回到資訊所的大家庭。記得在研究助理期間，常常藉由所上的演講與交流，感受到大家對於研究的興奮與熱情，那時就埋下日後以做研究為職志的心願。我赴美留學的研究主題仍然是自然語言處理，畢業時正好資訊所徵求自然語言處理的新血，對我來說這是個難得的機會，能主持一個自己的實驗室，並將自己的研究想法付諸實現，是研究人員最興奮的事。而中研院資訊所一直是國內外聲名卓著的研究機構，其研究環境更是提供了許多利於研究的條件，包括：(1) 充分的自由選擇研究題目 (2) 資訊所的研究人員有 40 位之多，分成八大研究群，因此在鑽研自己研究的同時，也有充分的機會跟同研究群或不同研究群的成員合作與切磋。(3) 中研院資訊所對於軟硬體，研究經費，行政等等的支援 (4) 提供許多國內外的重量級教授或研究人員的交流機會。(5) 資訊所並沒有教學上的要求，所以可以專注於研究，但資訊所也同時提供了國際學程的教學機會，讓研究人員也可以有培育英才的機會。

人生歷程

大一開始接觸程式設計，大三在工研院實習一年，為 8051 晶片開發一套語音指令辨識系統，算是研究的啟蒙階段，之後繼續從事連續語音辨識的研究，由於需要使用到語言模型 (Language Model)，那時才很驚訝的發現，原來資訊科學當中還有以“語言”為研究對象的一門學問，當時可以用大開眼界來形容！碩士畢業後我進入中研院資訊所詞庫小組服國防役，做了一系列中文的自然語言處理研究，也觸發了我在這個領域繼續深造的想法。於是在國防役結束後，便赴美國哥倫比亞大學跟隨 Kathleen Mackeown 做自然語言處理的研究。親炙了許多大師的風采，除了在專業知識上的獲得，也學習他們做學問的態度與方法。求學期間也有幸加入幾個美國當時的校際間的大型合作計畫，如 DARPA 所主導的 multilingual QA 計畫以及 NSF 的機器翻譯計畫，觀察到大型研究計畫的規劃與執行層面，同時也有許多機會跟各校的師長同學交流學習。在畢業的前一年有機會赴美國微軟研究院做暑期實習，看到大型企業對研發的投入與重視，更重要地，是深刻感受到自然語言處理早已全面走出象牙塔，各類相關技術為各大企業所利用，落實到其產品與服務。

研究的緣起及目標

人類用語言紀錄知識，彼此溝通。而自然語言處理 (NLP) 就是賦予電腦處理或者理解人類語言的能力的一門學問。人類語言對電腦來說其實相當複雜，不論是從語法或是語義的層面都常常具有高度歧異的現象，舉例來說，“我考試得了鴨蛋”，這裡的“鴨蛋”是“零分”的意思，而不是“鴨”的“蛋”。當電腦能夠正確的辨認出語法或是語義的歧異，其實也就表示某種程度上電腦已經理解了語言。

我們在過去三十多年的發展中，針對不同層次的解歧任務，開發了詞彙分析系統、 斷詞系統與中文語法剖析器等等，同時為了建構知識系統，我們開發了標記語料庫、句結構樹資料庫、詞彙知識庫等基礎建設。這些訓練語料由人工標記產生，數量僅能達到一定的規模。相比於網路上大量未標記的各類文字訊息，數量相當有限。因此如何利用網路上的大量未標記語料來加以訓練是近來 NLP 界的重要目標，而手段就是

赴美國哥倫比亞大學跟隨 Kathleen Mackeown
致力於自然語言處理的研究。

詞庫小組研究團隊。

透過深度學習，藉由大量語料將每一個詞彙訓練出一個向量，再用這些向量進一步來進行不同的任務。這種以向量來表達詞彙知識的作法，跟過去詞庫小組人工所建立的詞彙知識庫（稱為「廣義知網」）有著很大的不同。我們目前的研究方向是如何將「廣義知網」所規範的詞彙知識也能夠轉化為向量的表達方式，並同時參考大量語料，來產生詞彙的向量。以一個孩子的學習過程來做比喻： 一個孩子可以從所每日所接觸的人事物來得到知識，也同時從父母的教導中得到知識，綜合的學習可以發揮更大的成效。

另外，在大數據中的環境中，如何自動化擷取語言知識及一般常識也是我們另一個發展的重點，我們期望由發展出來的語言處理技術配合擷取的知識能自動的分析網際網路中的大量文本，從中抽取知識，並形成一個自動化的學習系統，語文處理系統可經由自動分析學習新知逐日更新知識庫，同時也藉由知識庫的更新增進了語文處理的能力。

對未來資訊所與個人的期許

大數據與人工智慧的應用現在正全面席捲學業與業界，在美國，Google、Facebook、Microsoft、IBM 等世界級的公司以及新創軟體也都投入大量資源在數據採礦與新一代的人工智慧系統的開發。在此潮流中，資訊所已經佔據了相當有利的位置，在各個層面包含多媒體資訊處理技術、資訊理論、社群網路及生物資訊方面等等都有相當傲人的成績，擁有極佳的發展機會。希望未來自己能夠透過與所內同仁的交流與合作，能有更多跨領域的學習與研究，也期待自己能開發出新一代的語言分析技術，壯大資訊所的成果並增進資訊所與個人的國際能見度，並期望所開發的技術能對產業與社會帶來實質的貢獻。

對未來想從事該領域資訊研究的學莘有何建議

對於研究工作來說，我覺得最重要的就是熱情，唯有對特定領域的熱情，才能夠驅使自己把事情做好，甚至廢寢忘食的投入仍然甘之如飴，享受在其中。資訊研究的範疇其實相當廣泛，找到一個自己有熱情並且符合自己個性與興趣的研究領域非常重要。其次，我覺得批判性思考是非常值得學習的技巧，也是台灣學子普遍來說比較不足之處，所謂批判性思考並不是批評，而是以清晰的思路，運用理性與邏輯的方式來嚴謹地分析事物，洞察事情的成因或本質 ，並看出優缺點。其中雖然帶有懷疑與批判精神，但目的是幫助自己更有系統的思考與主動式的學習。批判性思考也可以幫助我們更有好奇心與創造力，在大家都認為理所當然的時候，看出其不尋常的關鍵 ，有時就是一個研究的契機。

最後我想提的是快速學習的能力，目前線上課程 Coursera 與 YouTube 等提供許多高品質的內容，將知識的進入門檻降到非常容易的地步，善用這些資源並且保持積極的學習精神非常重要，能夠快速地學習到需要補強的知識是現代人必備的一種技能。相較於以前，對於一個想學習的人來說，這是一大福音，應該好好把握。

古典與（後）量子理論密碼學

（文轉自第 9 頁）

足以描述所有的計算（我們的電腦就是由電路組成的）。密碼學之所以採用電路模型描述計算是因為在電路模型上較容易去設計密碼學的構造。大致上來說，這是因為我們只需要設計如何處理不同的邏輯閘，以及如何組合它們，相對於複雜的程式語言來說單純許多。

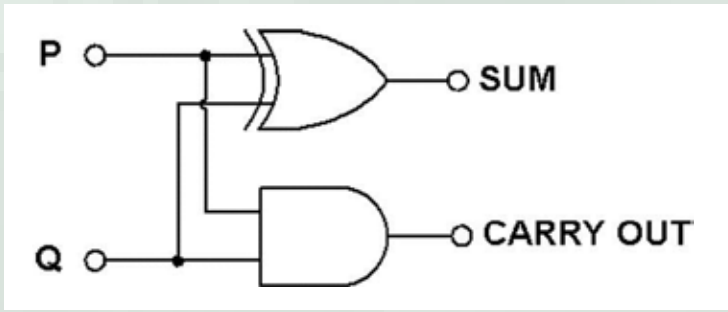
然而，電路模型無法直接表示迴圈或條件分支，當我們將程式轉換成電路時，它們會被展開形成龐大的電路，因此轉換成電路模型往往造成效率上的損失。為了改進此缺點，密碼學近五年來，許多研究學者（包含我們團隊）改以隨機存取機模型 (random access machine model, RAM model) 來描述計算，針對 RAM 計算模型來設計密碼學構造。RAM 模型能直接表示迴圈或條件分支，可視為 C++ 或 Python 等程式語言的抽象模型，能避免上述效率上的損失。

更進一步來說，注意到雲端的大數據運算往往使用平行技術，如 MapReduce 和 GraphLab，來處理大數據資料，然而平行運算這個特色是無法由 RAM 模型所描述的。因此，我們提出進一步使用 PRAM (parallel RAM) 計算模型來設計密碼學構造。簡言之，PRAM 模型是可表示平行運算的 RAM 模型，能描述 MapReduce 和 GraphLab 上的平行運算能力。我們最近的研究主軸之一是針對 PRAM 模型的密碼學發展，如以 PRAM 為計算模型的功能加密法、安全計算、以及程式模糊化構造等。

研究方向：（後）量子密碼學

（後）量子密碼學主要研究量子計算如何影響密碼學。從一個角度來說，量子計算可以有效率的解決因式分解問題。因此，攻擊者可利用量子計算來破解一些由基於因式分解難題所建構的密碼系統。換言之，量子計算給予攻擊者更強的能力。後量子密碼學旨在處理這個問題，探討如何設計量子計算無法有效率破解的計算難題（如基於晶格難題 (lattices problems) 的假設），並利用這樣的假設來設計具量子安全性的密碼架構。另一方面，量子亦能給予誠實方更多的能力，允許去達到一些本來在古典（沒有量子能力的情況下）無法達到的密碼任務。一個最經典的例子是量子金鑰分 配 (Quantum key distribution, QKD)，這是短期內最可能實現的量子密碼應用。世界上很多國家，如中國、日本、瑞士、美國已經耗費上億經費在建置 QKD 網路的硬體原型 [6]。QKD 允許遠距離的兩方通過傳送量子訊息來產生安全的共同秘密金鑰。這樣的任務在古典上是無法達到的。

從這兩個例子可以看出來，量子的能力對於密碼學來說是一把雙面刃，可以給予攻擊者更強的能力來破壞安全性，也可以給予誠實方更多的能力來實現更多的密碼任務。事實上量子密碼學仍處於相對年輕的階段，還有許多還沒有被探索過的可能性。我們對探索各種量子密碼學上新的可能性非常感興趣。目前我們致力於了解量子信息 (quantum information) 在後量子密碼學的影響。也就是說，了解當攻擊者取得額外量子資訊時，可能造成的危害，並發展在這種情況下保證安全性的技術。



半加器的電路示意圖 (P, Q 為兩個輸入位元、SUM 為加總、CARRY OUT 為進位)(圖四)。

實驗室之運作

在實驗室運作上，除了研究本身，我們也希望能提供對理論計算機科學 (Theoretical CS;TCS) 感興趣的學生一個認識不同理論研究領域與學習 / 嘗試進行理論研究的環境。我們鼓勵學生尋找自己有興趣的研究題目。除了各自的研究題目，我們每週有固定的 seminar，報告特定主題的最新研究進展（目前的主題為 Differential Privacy）。我們也鼓勵實驗室成員組織不同主題的 reading group。我們也積極的邀請國際理論研究學者來台訪問進行學術合作，通常會安排學者演講其正在執行的研究，也鼓勵學生與學者進行交流。這些活動都歡迎對 TCS 感興趣的學生來參加 / 旁聽。

最近，我們與台大陳和麟教授、清大廖崇碩教授開始試辦台灣理論日的活動，在為期一天的理論日中，安排三至四個亞洲鄰近區域的講者來介紹理論電腦科學相關的研究，所有活動都是免費開放給所有對理論有興趣的聽眾。活動網址為：http://theoryday.github.io/。我們會透過下面兩個 mailing list 來提供 TCS 相關的活動與演講訊息，歡迎有興趣的讀者訂閱。

Theory-event-announcement:
http://ppt.cc/bpgJF
Theory-talk-announcement:
http://ppt.cc/phI5S

參考資料

[1] The Imitation Game, Wikipedia, https://en.wikipedia.org/wiki/The_Imitation_Game
[2] Dan Boneh, Online cryptography class, Introduction to Cryptography, https://class.coursera.org/crypto-preview
[3] Jonathan Katz and Yehuda Lindell, Introduction to Modern Cryptography, Second Edition (Chapman & Hall/CRC Cryptography and Network Security Series), ISBN: 978-1466570269
[4] Salil P. Vadhan, Lecture note, Introduction to Cryptography (CS 127/CSCI E-127), http://people.seas.harvard.edu/~salil/cs127
[5] Craig Gentry, Fully homomorphic encryption using ideal lattices, STOC 2009.
[6] Quantum key distribution, Wikipedia, https://en.wikipedia.org/wiki/Quantum_key_distribution

中央研究院資訊科學研究所
115 台北市南港區研究院路二段 128 號
發行人：許聞廉
編輯：田力如，曾慧琦
tel: +886-2-2788-3799
www.iis.sinica.edu.tw