台灣農業研究 (J. Taiwan Agric. Res.) 57(3):161–174 (2008)

水稻白葉枯病抗性基因之蛋白質胺基酸組成的多變數分析[1]

魏夢麗[2] 鍾依涵[3] 呂椿棠[2] 呂秀英[2,4]

摘　要

魏夢麗、鍾依涵、呂椿棠、呂秀英。2008。水稻白葉枯病抗性基因之蛋白質胺基酸組成的多變數分析。台灣農業研究 57:161–174。

蛋白質中的胺基酸組成隱藏甚多訊息，但其組成特性會因不同物種或物種內不同基因而異，故蛋白質的胺基酸組成分析成爲生物資訊研究的重要課題之一。胺基酸組成特性是由多個物種或基因 (觀測值) 與 20 種胺基酸之頻率 (變數) 所構成的資料矩陣來決定，這種多維資料形式所含之訊息，最適合利用多變數分析 (multivariate analysis) 之統計技術來解析。爲促進水稻白葉枯病抗病基因 (*Xa*) 蛋白質序列的結構研究，有必要針對所有已完全定序的 *Xa* 基因進行蛋白質之胺基酸組成分析。基此，本研究以 NCBI 公共資料庫內已知序列的 *Xa1*、*xa5*、*xa13*、*Xa13*、*Xa21*、*Xa26* 及 *Xa27* 等基因共 17 條蛋白質序列爲供試材料，綜合運用集群分析 (cluster analysis) 及對應分析 (correspondence analysis)，來檢測不同 *Xa* 基因之蛋白質胺基酸組成的變異形式。結果顯示，根據胺基酸組成比例，可將 *Xa* 基因及其家族分成六群，各群基因之蛋白質序列中各有偏好的胺基酸。*Xa1*、*Xa21*、*Xa26* 及 *Xa27* 基因之蛋白質序列中皆以白胺酸 (leucine) 出現頻率最高；*xa13*、*Xa13* 及 *Xa27* 的丙胺酸 (alanine) 出現較多但絲胺酸 (serine) 較少；*xa13* 及 *Xa13* 也有較多的纈胺酸 (valine)；*xa5* 出現麩胺酸 (glutamic acid) 和酥胺酸 (threonine) 的頻率遠高於其他基因；所有 *Xa* 基因皆含有高比例的疏水性胺基酸。本研究揭示出多變數分析之統計技術，可有效檢測出 *Xa* 基因間蛋白質之胺基酸組成的變異形式。

關鍵詞：水稻、*Xa* 基因、蛋白質結構、胺基酸組成及屬性、集群分析、對應分析。

前　言

胺基酸是組成蛋白質的單位小分子，在蛋白質中出現的只有 20 種，其支鏈在大小、形狀、電荷、形成氫鍵的能力和化學活性方面都存在差異。正由於這 20 種胺基酸的差異，以及它們的各種組合變化的結果，影響最後所形成蛋白質的構造與特性，因此造就了結構多變、功能紛繁的蛋白質。

---

1. 行政院農業委員會農業試驗所研究報告第 2324 號。接受日期：2008 年 7 月 23 日。
2. 本所作物組助理研究員、助理研究員與研究員。台灣 台中縣 霧峰鄉。
3. 私立朝陽科技大學生物技術研究所碩士。台灣 台中縣 霧峰鄉。
4. 通訊作者，電子郵件：iying@wufeng.tari.gov.tw；傳真機：(04)23390528。

蛋白質的長條胺基酸序列，爲其一級構造 (primary structure)，是蛋白質最終構成的根本，各級構造的訊息都決定於胺基酸的序列，即蛋白質摺疊所含之訊息揭示在其胺基酸殘基的線性序列中 (Shi *et al*. 2005)。Wang & Wang (1999) 指出，胺基酸的分類和組成與蛋白質的分類相關，由此推論出胺基酸的分類和序列中應含有蛋白質二級結構的訊息，胺基酸的合理分類是進一步研究蛋白質的分類及預測等工作的前提及基礎。Villar & Koehler (2000) 分析序列總長度在 50 個胺基酸以下的小型蛋白質之胺基酸組成，發現某些種類的胺基酸之出現比例較高，但相對其他疏水性 (hydrophobic) 胺基酸則出現較少，此一組成之趨勢特徵有利於胜肽 (peptide) 與其目標物間的反應。Chien & Chiang (2001) 對含 20 個胺基酸以下且形成分子內雙硫鍵之自然胜肽進行胺基酸組成分析，發現從其所存在的胺基酸偏好性，似乎反應出蛋白質胜肽中形成雙硫鍵時的結構及功能上的一些特性。由此可見，蛋白質中的胺基酸組成隱藏甚多訊息。在生物演化過程中可能由於突變或基於蛋白質結構和功能所需，蛋白質中的胺基酸組成會產生變異，於是造成胺基酸組成特性在不同物種或物種內不同基因間的差異 (Tang *et al*. 2004; Bharanidharan & Gautham 2005; Yampolsky & Stoltzfus 2005)。因此蛋白質的胺基酸組成分析，成爲生物資訊研究的重要課題之一。

胺基酸組成特性是由多個物種或基因 (觀測值) 與 20 種胺基酸之頻率 (變數) 所構成的資料矩陣來決定，這種多維資料形式所含之訊息，最適合利用多變數分析 (multivariate analysis) 之統計技術來解析。多變數分析是一種綜合分析方法，能夠在多個研究對象和多個指標相互關聯的情況下分析出它們的統計規律。集群分析是多變數分析方法的一種，係利用一般邏輯程序，能根據多個變數間或研究對象的相似性與相異性，客觀地將相似者歸在同一群內，其一般原則是使同一集群內的個體差異性最小，而不同集群間的個體則具有最大差異性 (Leps & Smilauer 1999)；集群分析之結果最後以樹狀圖 (dendrogram) 的分支狀況來判斷群數。對應分析爲 Benzecri 於 1973 年所提出，又稱爲相互平均法 (reciprocal averaging, Hill 1973)，用於分析簡單二維及多維表格的行與列之間的對應關係，是一種廣泛運用到生態學 (尤其是植物生態學) 的多變數統計分析技術，現已大量應用到其他各方面的研究領域 (Beh 2004)，包括基因組分析上之各種應用 (Gupta & Ghosh 2001; Tekaia *et al*. 2002; Tan *et al*. 2004)。對應分析對於數據的要求比較簡單，不需要有嚴格的線性關係，也無須限定所處理的資料是連續或離散變數，只需要是一個列聯表 (contingency table) 的資料形式，即可將多維之變數或樣本於一個低維空間上以一個序列分佈圖 (ordination diagram) 呈現；藉由序列分佈圖，不但可用來檢測變數間或樣本間的關係，也能從各變數或樣本落在序列分佈圖上之距離遠近，來輔助集群分析之樹狀圖的分群結果判斷，倘發現點間有群團現象表示可能是同一群。因此，綜合運用集群分析與對應分析，將可瞭解不同基因間之蛋白質序列中胺基酸組成的相似與相異程度，並據此判斷基因間的群聚關係。

水稻白葉枯病 (rice bacterial blight) 是由黃單孢桿菌 (*Xanthomonas oryzae pv. Oryzae*, *Xoo*) 引起的一種細菌性維管束病害，爲世界性主要稻作病害之一。目前雖然已有藥劑能針對此病害進行防治，但育成抗病品種爲解決此病害最有效之方法 (Chang 1995)，因此瞭解水稻白葉枯病抗性基因 (*Xa*) 的結構，有利於水稻的分子育種 (Bai *et al*. 2003; Monosi *et al*. 2004; Zhou *et al*. 2004)。至 2007 年 3 月 31 日截止，日本的 NIG (National Institute of Genetics) 之 Oryzabase 資料庫 (http://www.shigen.nig.ac.jp/rice/oryzabase/top/top.jsp, Yamazaki & Jaiswal 2005) 內已登記命名 32 個 *Xa* 基因，其中已完全定序的有 *Xa1*、*xa5*、*xa13*、*Xa13*、*Xa21*、*Xa26* 及 *Xa27* 7 個基因 (Song *et al*. 1995; Yoshimura

et al. 1995, 1998; Anjali & McCouch 2004; Gu *et al*. 2005; Sun *et al*. 2006)。植物抗病機制中，有兩套基因系統先後作用，一爲抗病基因 (resistant gene, R gene)，二爲防衛基因。R 基因能特異性的識別病原菌對應之無毒 (Avr) 基因的產物，兩者可直接或間接地相互作用而刺激植物的訊號傳遞系統，最後誘導出植物防衛基因的表達，使植物表現出抗病性，同時 R 基因能不斷進化，可對病原菌不斷突變所產生的生理小種產生抗性 (Wang *et al*. 2002)。植物 R 基因的表現蛋白質結構分爲五類 (Baker *et al*. 1997)：(1) 含有核苷酸結合位置 (nucleotide binding sites, NBS) 的細胞質受體類蛋白 (cytoplasmic receptor-like proteins) 和多白胺酸重複 (leucine-rich repeat, LRR) 序列區域，簡稱爲 NBS-LRR；(2) 絲胺酸-酥胺酸激酶 (serine-threonine kinase)；(3) 含有大量細胞質外 LRR 區域的穿膜受體 (transmembrane receptor)；(4) 含細胞外 LRR 區域之穿膜受體和細胞內絲胺酸-酥胺酸激酶 (intracellualar serine-threonine kinase) 區域；及 (5) 其他。R 基因的發現，使水稻的抗病機制在分子層次研究上有了突破性的發展，且由於水稻的 *Xa* 基因在演化上的保守性與差異性，揭示出 *Xa* 基因可能存在較複雜的分子機制 (Guo *et al*. 2005)。已知 *Xa1* 爲上述第一類 R 基因，而 *Xa21* 和 *Xa26* 爲第四類 R 基因，至於 *xa5*、*xa13*、*Xa13* 及 *Xa27* 之抗病防禦系統的生化途徑異於 *Xa1*、*Xa21* 和 *Xa26*，其表現蛋白質結構尚未清楚，目前被歸於其他之第五類 R 基因 (Yoshimura *et al*. 1998; Khush & Angeles 1999; Anjali & McCouch 2004; Gu *et al*. 2005; Sun *et al*. 2006)。爲有利於瞭解不同 *Xa* 基因之蛋白質序列的結構差異，因此本研究利用多變數分析法之集群分析與對應分析，來探討這些已定序的 *Xa* 基因之蛋白質胺基酸組成的變異形式。

材料與方法

Xa 基因序列之蒐集與下載

本研究透過 NCBI (National Center for Biotechnology Information) 之 GenBank 資料庫 (http://www.ncbi.nlm.nih.gov/Genbank/index.html, Guisez *et al*. 1993) 下載已定序之 *Xa* 基因的編碼區序列 (coding domain sequence, CDS) 和蛋白質序列，這些已知 *Xa* 基因的相關資訊整理如表 1，其 CDS 長度介於 318-5,409 bp 之間，來源包含有印度型水稻 (*Oryza sativa indica* caltivar-group)、日本型水稻 (*japonica*)、長花藥野生稻 (*Oryza longistaminata*)。其中 *Xa1* 分別於 1999 年和 2004 年登錄有兩條序列 (本文簡稱 *Xa1_1999* 和 *Xa1_2004*)，各位於第 4 條與第 10 條染色體；*xa5* 爲一隱性基因，位於第 5 條染色體上；*Xa13* 及 *xa13* 分別爲顯、隱性基因，皆位於第 8 條染色體上；*Xa21* 基因家族位於第 11 條染色體，包含七個成員分別爲 *Xa21-A1*、*Xa21-A2*、*Xa21-B*、*Xa21-C*、*Xa21-D*、*Xa21-E* 及 *Xa21-F*；*Xa26* 基因家族位於第 11 條染色體，包含 4 個成員分別爲 *Xa26-MRKa*、*Xa26-MRKb*、*Xa26-MRKc* 及 *Xa26-MRKd* (本文簡稱 *Xa26a*、*Xa26b*、*Xa26c* 及 *Xa26d*)；*Xa27* 位於第 6 條染色體。所蒐集共 17 條基因之序列中 *Xa21-A2*、*Xa21-C*、*Xa21-F* 及 *Xa26d* 等 4 個基因只有 CDS 序列，在 NCBI 中尚未登錄有完整的蛋白質序列，故本研究將此四條 CDS 另利用 BioEdit 生物資訊軟體 (Biological Sequence Alignment Editor for Win95/98/NT/2K/XP; http://www.mbio.ncsu.edu/BioEdit/bioedit.html, Hall 2001) 將其轉譯成蛋白質序列，合計共 17 條蛋白質序列。

多變數分析在 *Xa* 基因之蛋白質胺基酸組成的應用

爲探討各 *Xa* 基因及其家族間胺基酸組成之相似及相異程度，首先計算各 *Xa* 基因及其家族之蛋白質序列中各種胺基酸所佔比例，然後利用 STATISTICA 統計軟體 (Statsoft Inc. 2002) 以非加權

表 1. Oryzabase 及 NCBI 資料庫中已定序之 *Xa* 基因的資訊摘要 (至 2007 年 3 月 31 日止)

Table 1. Information summary of known sequences in rice *Xa* genes from Oryzabase and NCBI databases (update time: March 31, 2007)

Label	Gene symbol	Submitted year	Accession no. in NCBI	Definition	Gene length	Chr. no	CDS		Protein sequence	
							Location	Length	Accession no.	Length
Xa1_1999	*Xa1*	1999	AB002266	*Oryza sativa* (*indica* cultivar-group) mRNA for *Xa1*	5910	4	113..5521	5409	BAA25068	1802
Xa1_2004	*Xa1*	2004	NM_194685	*Oryza sativa* (*japonica* cultivar-group) putative bacterial blight resistance protein *Xa1*-like protein, mRNA	4206	10	1..4206	4206	NP_919667	1401
xa5	*xa5*	2004	AY643716	*Oryza sativa (indica* cultivar-group) transcription factor IIA gamma subunit (TFIIAy) mRNA, partial cds.	318	5	1..318	318	AAV53715	106
xa13	*xa13*	2006	DQ421394	*Oryza sativa (indica* cultivar-group) cultivar IRBB13 disease resistant allele *xa13* (*xa13*) gene, complete cds.	924	8	1..924	924	ABD78942	307
Xa13	*Xa13*	2006	DQ421395	*Oryza sativa (indica* cultivar-group) cultivar IR64 disease resistant allele *XA13* (*Xa13*) gene, complete cds.	924	8	1..924	924	ABD78943	307
Xa21-A1	*Xa21*	1997	U72725	*Oryza longistaminata* receptor kinase-like protein gene	8416	11	4771..7384,7676..8052	2991	AAB82755	996
Xa21-A2	*Xa21*	1998	U72727	*Oryza longistaminata* receptor kinase-like protein, pseudogene sequence	5940	11	2151..4764,4872..5248	2991	-	996 [z]
Xa21-B	*Xa21*	1995	U37133	*Oryza sativa* (*indica* cultivar-group) receptor kinase-like protein gene	3921	11	1..2677,3521..3921	3078	AAC49123	1025
Xa21-C	*Xa21*	1998	U72723	*Oryza longistaminata* receptor kinase-like protein gene, pseudogene	19639	11	15118..17720, 17827..18201	2978	-	992 [z]
Xa21-D	*Xa21*	1997	U72726	*Oryza longistaminata* receptor kinase-like protein	13341	11	2367..4205	1839	AAB82753	612
Xa21-E	*Xa21*	1997	U72724	*Oryza sativa* (*indica* cultivar-group) receptor kinase-like protein gene	9424	11	2819..5260	2442	AAB82756	813
Xa21-F	*Xa21*	2005	U72728	*Oryza longistaminata* receptor-like kinase protein, pseudogene sequence	7204	11	1683..4352,5147..5547	3071	-	1023 [z]
Xa26a	*Xa26*	2006	DQ355952	*Oryza sativa* (*indica* cultivar-group) receptor kinase MRKa	3401	11	14225..17135,17240..17625	3297	ABD36511	1098
Xa26b	*Xa26*	2006	DQ355952	*Oryza sativa* (*indica* cultivar-group) bacterial blight resistance protein *Xa26*	3631	11	21725..24659,24765..25141	3312	ABD36512	1103
Xa26c	*Xa26*	2006	DQ355952	*Oryza sativa* (*indica* cultivar-group) receptor kinase MRKc	3456	11	34496..34911,35026..37951	3342	ABD36513	1113
Xa26d	*Xa26*	2006	DQ355952	*Oryza sativa* (*indica* cultivar-group) cultivar Minghui 63 receptor kinase pseudogene	11396	11	2021..3423,9305..10839, 13028..13416	3327	-	1109 [z]
Xa27	*Xa27*	2005	AY986491	*Oryza sativa* (*indica* cultivar-group) *Xa27* gene, *Xa27*-IR24 allele	2393	6	1590..1931	342	AAY54163	113

[z] Protein sequence translated from CDS using BioEdit software.

配對算術平均法 (unweighted pair group method using arithmetic average, UPGMA) 進行集群分析，以求得歸群樹狀圖。UPGMA 是集群分析的演算方法之一，以群間個體所有成對距離的平均值爲判斷，較不受離群值影響 (Mohna *et al*. 1992; Mumm *et al*. 1994)。接著，同樣再利用 STATISTICA 統計軟體進行對應分析，由其行列之間的對應關係，求得序列分佈圖，以呈現基因間蛋白質胺基酸組成之關係。最後綜合集群分析之樹狀圖和對應分析之序列分佈圖的結果，來共同決定群數。

各群 *Xa* 基因之蛋白質胺基酸的組成及屬性比較

根據上述分群結果，計算各群 *Xa* 基因蛋白質序列中各種胺基酸的組成比例，並利用 Excel 軟體繪製雷達圖，以進行各群 *Xa* 基因間蛋白質胺基酸組成之差異性比較。再根據胺基酸支鏈之性質 (Mathews *et al*. 2000，整理如表 2)，計算各群基因之蛋白質序列中不同屬性胺基酸所佔比例，並藉由雷達圖之呈現，來探討各群 *Xa* 基因間之蛋白質胺基酸屬性的差異。

結　果

根據 17 個 *Xa* 基因及其家族之蛋白質序列中胺基酸組成比例的集群分析，由所得之樹狀圖 (圖 1A) 可發現，在歐氏距離 13.2 時，*Xa* 基因及其家族可被分成四群：*Xa1_1999*、*Xa1_2004*、*Xa21-C*、*Xa21-F*、*Xa26d*、*Xa21-A1*、*Xa21-A2*、*Xa21-B*、*Xa21-D*、*Xa21-E*、*Xa26a*、*Xa26b* 及 *Xa26c* 爲同一群，*xa13* 和 *Xa13* 自成一群，*xa5* 和 *Xa27* 則各自成群。進而由對應分析所得之序列分佈圖上各基因落點之距離遠近，大致也可將 *Xa* 基因及其家族分成四群，此與集群分析之結果一致 (圖 1 B)。然而，由於考慮到 *xa5*、*xa13*、*Xa13* 及 *Xa27* 之蛋白質序列的長度 (106-307 base) 相較於其他 *Xa* 基因甚短，因此移去此四個基因再重新進行集群分析及對應分析。根據集群分析之樹狀圖 (圖 2A)，結果發現於歐氏距離 8.3 時，13 個 *Xa* 基因及其家族可進而被分成三群，其中 *Xa1_1999*、*Xa1_2004* 及 *Xa21-C*、*Xa21-F*、*Xa26d* 從前述群聚結果 (圖 1A) 中被分離出來而成爲兩群 (圖 2 A)。檢視對應分析之序列分佈圖，也得到一致的分群結果 (圖 2 B)。綜合圖 1 及圖 2 的兩次分群結果，*Xa* 基因之蛋白質胺基酸組成可區分成六群 (表 3)：第 I 群爲 *Xa1_1999* 及 *Xa1_2004*，第 II 群爲 *Xa21-A1*、*Xa21-A2*、*Xa21-B*、*Xa21-D*、*Xa21-E*、*Xa26a*、*Xa26b* 及 *Xa26c*，第 III 群爲 *Xa21-C*、*Xa21-F* 及 *Xa26d*，第 IV 群爲 *xa5*，第 V 群爲 *xa13* 及 *Xa13*，第 VI 群爲 *Xa27*。

表 2. 胺基酸屬性分類

Table 2. Classification of amino acid attribution

Non-Polar	Polar (Hydrophilic)		
(Hydrophobic)	Uncharged	Positively charged	Negatively charged
Alanine (Ala)	Cysteine (Cys)	Histidine (His)	Aspartic acid (Asp)
Phenylalanine (Phe)	Glycine (Gly)	Lysine (Lys)	Glutamic acid (Glu)
Isoleucine (Ile)	Asparagine (Asn)	Arginine (Arg)	
Leucine (Leu)	Glutamine (Gln)		
Methionine (Met)	Serine (Ser)		
Proline (Pro)	Threonine (Thr)		
Valine (Val)	Tyrosine (Tyr)		
Tryptophane (Trp)			

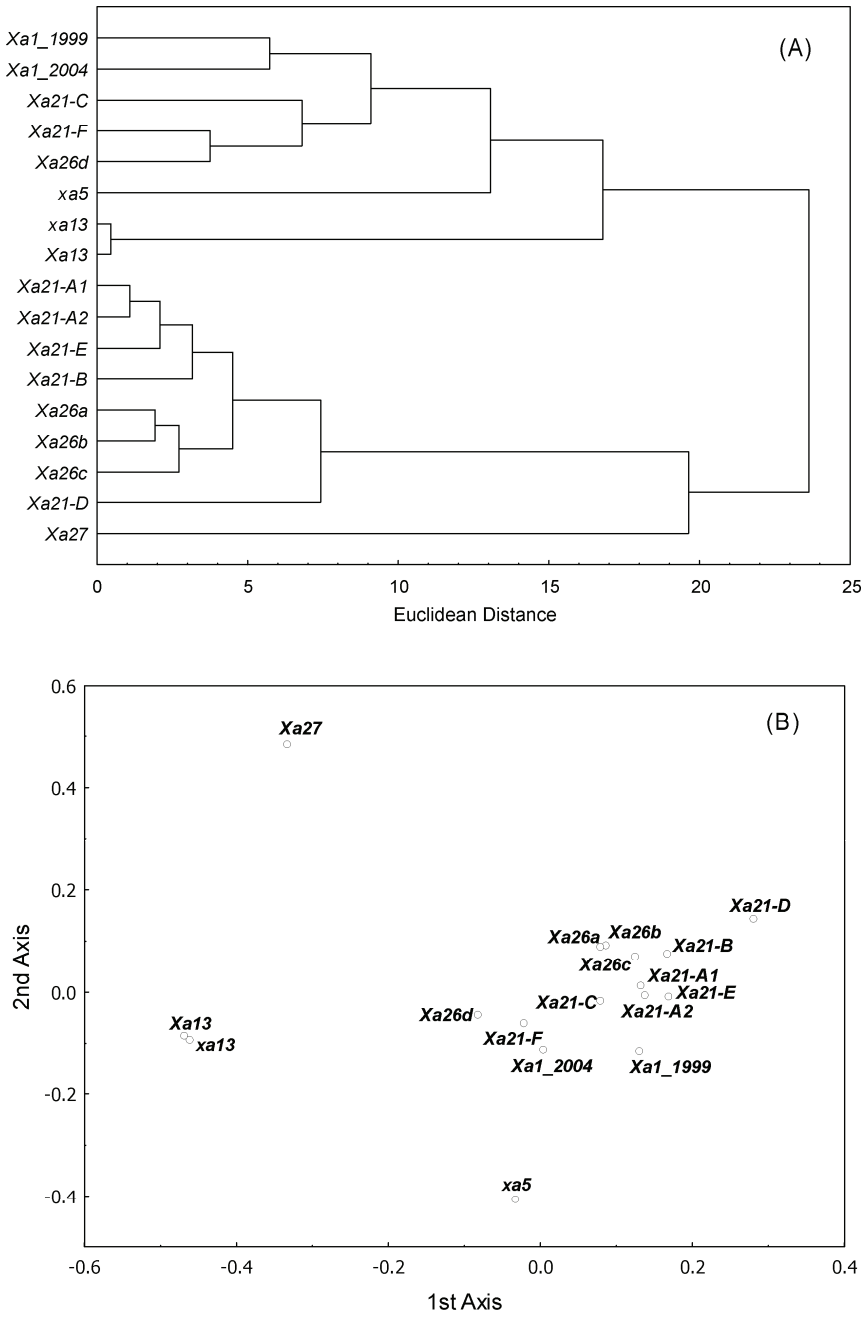

圖 1. 17 個 *Xa* 基因及其家族根據蛋白質胺基酸組成的分群結果。(A) 集群分析採 UPGMA 法之樹狀圖；(B) 對應分析之序列分佈圖。

Fig. 1. Grouping results of 17 *Xa* genes and their families according to the amino acid composition in proteins. (A) Dendrogram of cluster analysis using UPGMA, (B) Ordination diagram of correspondence analysis.

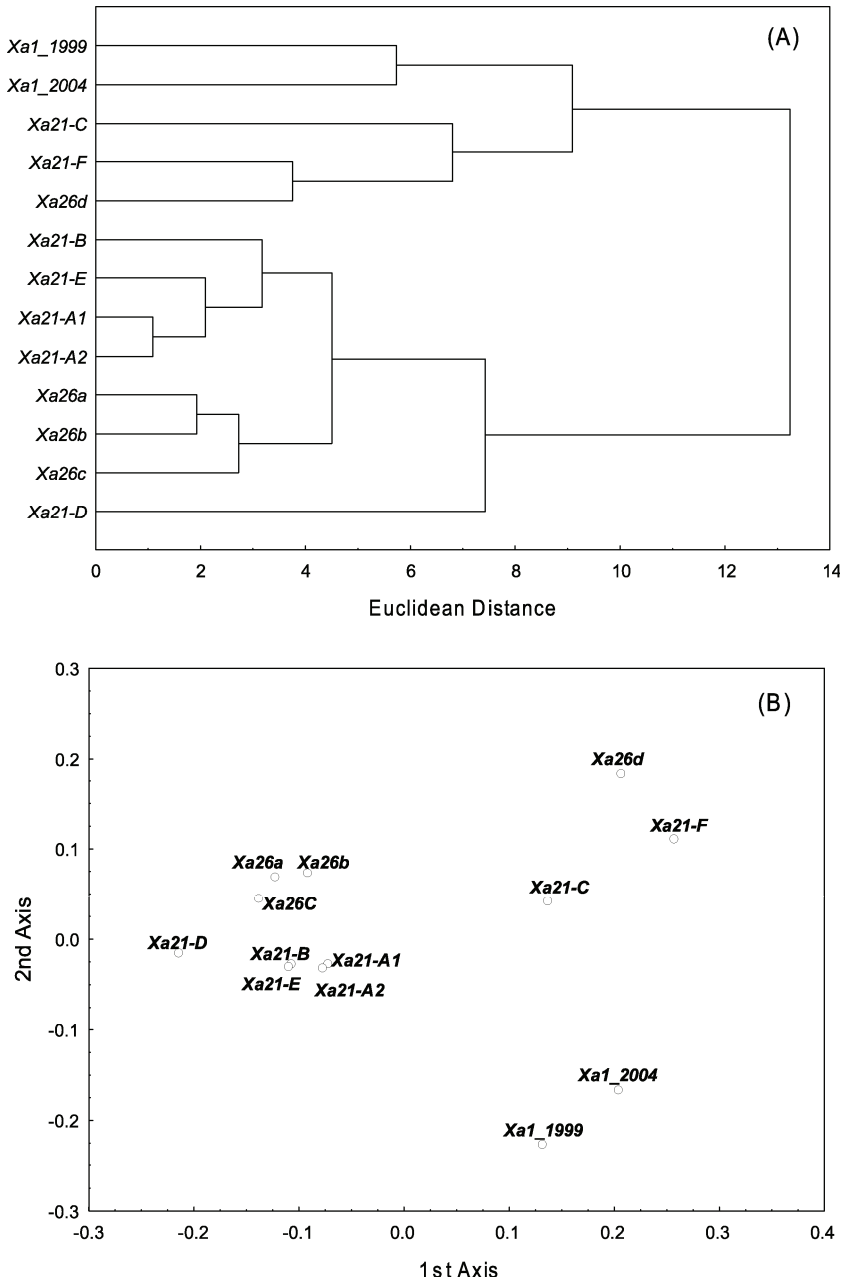

圖 2. 12 個 *Xa* 基因及其家族 (不含 *xa3*、*xa13*、*Xa13* 及 *Xa27*) 根據蛋白質胺基酸組成的分群結果。(A) 集群分析採 UPGMA 法之樹狀圖；(B) 對應分析之序列分佈圖。

Fig. 2. Grouping results of 12 *Xa* genes and their families (without *xa3*, *xa13*, *Xa13* and *Xa27*) according to the amino acid composition in proteins. (A) Dendrogram of cluster analysis using UPGMA, (B) Ordination diagram of correspondence analysis.

表 3. *Xa* 基因之蛋白質胺基酸組成比例的分群結果

Table 3. Grouping results of *Xa* genes determined from the percentage compositions of amino acids in proteins

Group	Genes							
I	*Xa1_1999*	*Xa1_2004*						
II	*Xa21-A1*	*Xa21-A2*	*Xa21-B*	*Xa21-D*	*Xa21-E*	*Xa26a*	*Xa26b*	*Xa26c*
III	*Xa21-C*	*Xa21-F*	*Xa26d*					
IV	*xa5*							
V	*xa13*	*Xa13*						
VI	*Xa27*							

由各群 *Xa* 基因表現蛋白質之胺基酸組成比例的雷達圖 (圖 3)，可清楚看出各群 *Xa* 基因對表現蛋白質的胺基酸使用確實有差異。分屬第 I、II、III 群的 *Xa1*、*Xa21* 及 *Xa26* 的蛋白質序列中皆一致出現最多的白胺酸 (Leu)，其次爲絲胺酸 (Ser)，但三群基因間在蛋白質序列中的其他胺基酸之組成比例仍多少有異，其中 *Xa1* 的麩胺酸 (Glu) 出現較多而 *Xa21* 則以甘胺酸 (Gly) 和天門冬醯胺酸 (Asn) 較多。相對於其他基因，*xa13*、*Xa13* 及 *Xa27* 的丙胺酸 (Ala) 出現較多但 Ser 較少；*xa13* 及 *Xa13* 也有較多的纈胺酸 (Val)；*xa5* 出現 Glu 和酥胺酸 (Thr) 的頻率遠高於其他基因。第 VI 群之 *Xa27* 之蛋白質序列中出現 Leu 之頻率是所有基因中之最高；所有 *Xa* 基因之蛋白質序列出現色胺酸 (Trp) 的頻率都甚少。

再由各群 *Xa* 基因之表現蛋白質的不同屬性胺基酸所佔比例之雷達圖 (圖 4)，可發現各群 *Xa* 基因的蛋白質序列中皆以疏水性 (hydrophobic) 胺基酸出現較多，尤其以第 V 及 VI 群之 *xa13*, *Xa13* 及 *Xa27* 出現疏水性胺基酸之頻率最多，各高達約 58%，而第 IV 群之 *xa5* 的疏水性胺基酸雖亦多，但含量爲 38%，僅略高於無電荷胺基酸所佔比例 (36%)。*Xa1*、*Xa21* 及 *Xa26* 之蛋白質的胺基酸屬性分布較相近。所有 *Xa* 基因蛋白質序列的親水性 (hydrophilic) 胺基酸中以無電荷胺基酸之比例較高 (25–39%)，而正電荷及負電荷胺基酸的含量都甚少，約佔 4–15%。

討　論

根據本研究利用多變數分析進行 *Xa* 基因蛋白質序列中胺基酸組成比例的分群結果 (圖 1、圖 2 及表 3)，顯示第 I 群爲屬第一類 R 基因的 *Xa1*，而同屬第四類 R 基因的 *Xa21* 和 *Xa26* 及其基因家族則被分成第 II、III 群，屬其他之第五類 R 基因的 *xa5*、*xa13*、*Xa13* 及 *Xa27* 分別歸於第 IV、V、VI 群。由分群結果可以發現這些 *Xa* 基因之表現蛋白質各有其胺基酸組成特性 (圖 3)，*Xa1*、*Xa21* 及 *Xa26* 都含有豐富的 Leu，這是因爲它們的表現蛋白質皆含有 LRR 序列區域結構 (Song *et al*. 1995; Yoshimura *et al*. 1998; Sun *et al*. 2006)；此外，這三個 *Xa* 基因也含有豐富的 Ser，尤其是 *Xa21* 及 *Xa26* 的抗病機制含有細胞外 LRR 區域之穿膜受體和細胞內絲胺酸-酥胺酸激酶區域 (Song *et al*. 1995; Sun *et al*. 2006)，因此出現 Ser 頻率也較 *Xa1* 爲高；而 *Xa1*、*Xa21* 及 *Xa26* 之間在其他胺基酸 (如 Glu、Gly 和 Asn) 的所佔比例仍多少有異。*xa13* 是 *Xa13* 顯性基因在啓動子 (promoter) 區域內序列上之突變所產生的完全隱性基因，只有在同型 (homozygous) 狀態下對黃單孢桿菌菲律賓生理小種 *Xoo* race 6 (strain PXO99) 能產生抗性 (Chu *et al*. 2006)，經本研究分析結果 *xa13* 和 *Xa13* 的表現

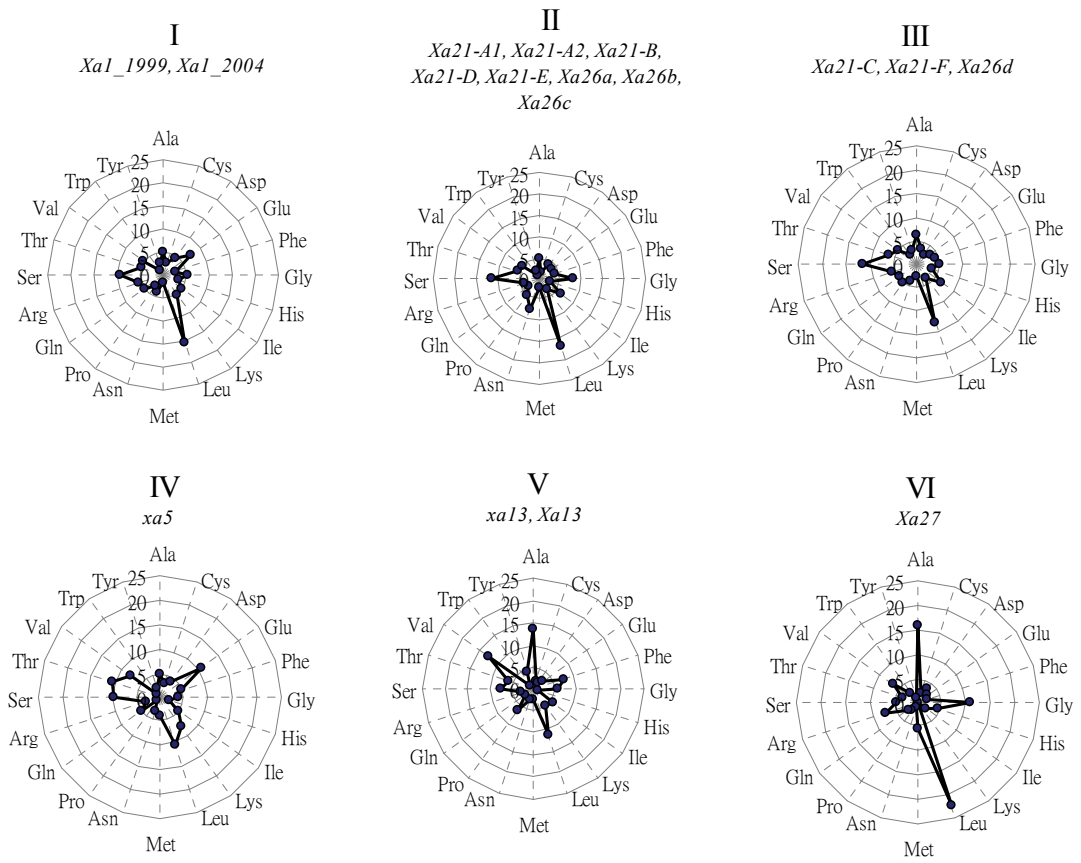

圖 3. 各群 *Xa* 基因之蛋白質胺基酸組成差異的雷達圖。

Fig. 3. Polar plot showing the compositional differences among amino acids in proteins for each *Xa*-gene group.

蛋白質之胺基酸組成被歸於同一群，經檢查兩者的蛋白質序列之編碼，發現幾乎相同，其差異僅在第 298 位置上 *xa13* 編碼爲 Thr，而 *Xa13* 編碼爲 Ala。*xa5*、*xa13*、*Xa13* 及 *Xa27* 4 個基因之蛋白質序列都不具有 LRR 結構，其 R 基因的表現蛋白質結構目前尚未清楚 (Yoshimura *et al*. 1998; Khush & Angeles 1999; Anjali & McCouch 2004; Gu *et al*. 2005; Sun *et al*. 2006)，但從本研究結果發現，它們的胺基酸組成特性明顯不同。*Xa27* 蛋白質序列中有非常多的 Leu 且含量遠高於其他 *Xa* 基因，按此可能與 *Xa27* 之產物內含 3 個核定位信號 (nuclear localized signal, NLS) 模組之 C 端保守區域和一個轉錄活化區的細胞核局部性第 III 型作用器有關 (Gu *et al*. 2005)。由此顯見，*Xa* 基因的蛋白質胺基酸組成類別與蛋白質結構類別相關，但同類 R 基因間的蛋白質胺基酸組成仍有差異。

胺基酸支鏈的極性 (親水性) 或非極性 (疏水性)，對形成蛋白質的整體構造與性質非常重要；疏水性強的胺基酸，通常居於蛋白質的內部，再藉由疏水性所引發之摺疊 (folding) 過程中扮演了

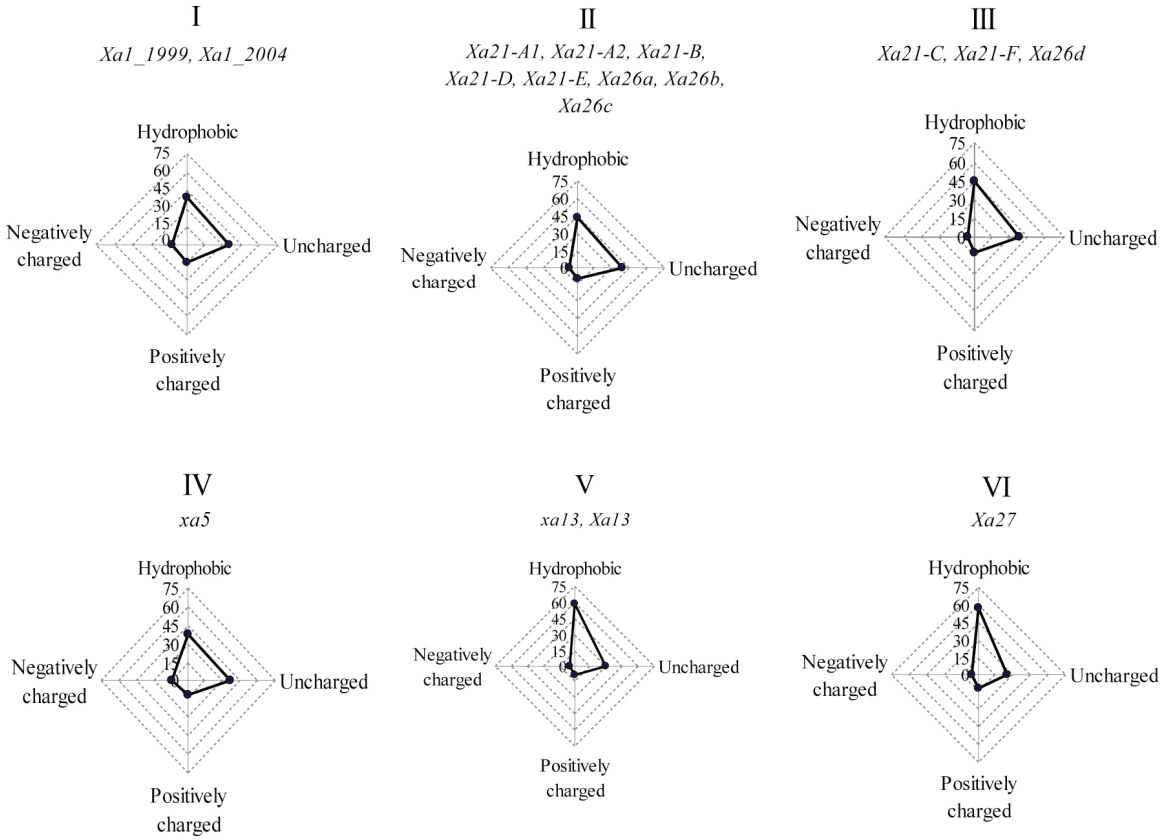

圖 4. 各群 *Xa* 基因之蛋白質胺基酸屬性差異的雷達圖。

Fig. 4. Polar plot showing the attributional differences among amino acids in proteins for each *Xa*-gene group.

重要角色 (Dill 1985; Nauchitel & Somorjai 1994; Villar & Kauvar 1994)。Mathews *et al*. (2000) 指出疏水性胺基酸多，有利於維持蛋白質的完整三級結構。藉由多變數分析進行 *Xa* 基因蛋白質胺基酸組成之分群結果，可以檢視出各群基因在胺基酸屬性分布的相似與相異程度，結果發現 7 個 *Xa* 基因之蛋白質皆含有高比例的疏水性胺基酸，但各群基因間的比例仍有差異，其中 *xa13*、*Xa13* 及 *Xa27* 基因之蛋白質序列出現疏水性胺基酸的頻率最高，而以 *xa5* 基因蛋白質最少，*Xa1*、*Xa21* 及 *Xa26* 基因表現蛋白質的胺基酸屬性分布則較相近 (圖 4)。

誌　　謝

本研究承蒙行政院農業委員會農業生物技術國家型科技計畫 (計畫編號：94 農科-5.2.1-農-C1(8)) 及國家科學委員會 (計畫編號：NSC95-2317-B-055-004) 補助經費，特致謝忱。

引用文獻 (Literature cited)

Anjali, S. I. and S. R. McCouch. 2004. The rice bacterial blight resistance gene *xa5* encodes a novel form of disease resistance. Mol. Plant-Microbe Interact. 17:1348–1354.

Bharanidharan, D. and N. Gautham. 2005. Amino acid variation in cellular processes in 108 bacterial proteomes. Arch. Microbiol. 184:168–174.

Baker, B., P. Zambryski, B. Staskawicz, and S. P. Dinesh-Kumar. 1997. Signaling in plant-microbe interactions. Science 276:726–733.

Bai, J., L. A. Pennill, J. Ning, S. W. Lee, J. Ramalingam, C. R. Webb, B. Zhao, Q. Sun, J. C. Nelson, J. E. Leach, and S. H. Hulbert. 2003. Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. Genome Res. 12:1871–1884.

Beh, E. J. 2004. A Bibliography of the Theory and Application of Correspondence Analysis. Vol. II-By Publication. School of Quantitative Methods and Mathematical Sciences, Univ. Western Sydney. Australia. 101 pp.

Benzecri, J. P. 1973. Lanalyse des donnees. II. Lanalyse des correspondances. Dunod. Paris, France. 619 pp.

Chang, S. J. 1995. Physiological and genetical studies on bacterial blight (*Xanthomonas oryzae* pv. *oryzae*) resistance in rice (*Oryza sativa* L.). PhD. Grad. Inst. Agron. National Chung-Hsing Univ. Taichung. 172 pp.

Chien, W. J. and F. T. Chiang. 2001. The preference of amino acids in cyclic peptides an application of bio-information in elementary biochemistry. J. Chaoyang Univ. Technol. 6:617–628. (in Chinese with English abstract)

Chu, Z., M. Yuan, J. Yao, X. Ge, B. Yuan, C. Xu, X. Li, B. Fu, Z. Li, J. L. Bennetzen, Q. Zhang, and S. Wang. 2006. Promoter mutations of an essential gene for pollen development result in disease resistance in rice. Genes Dev. 20:1250–1255.

Dill, K. A. 1985. Theory for the folding and stability of globular protein. Biochemistry 24:1501-1509.

Gu, K., B. Yang, D. Tian, L. Wu, D. Wang, C. Sreekala, F. Yang, Z. Chu, G. L. Wang, F. F. White, and Z. Yin. 2005. R gene expression induced by a type-III effector triggers disease resistance in rice. Nature 435:1122–1125.

Guisez, Y., J. Robbins, E. Remaut, and W. Fiers. 1993. Folding of the MS2 coat protein in *Escherichia coli* is modulated by translational pauses resulting from mRNA secondary structure and codon usage: A hypothesis. J. Theor. Biol. 162:243–252.

Guo, S. W., Y. Zhang, L. H. Sun, and D. Y. Gao. 2005. Progress on the research of bacterial blight resistance in rice. Chin. Agric. Sci. Bull. 21:339–344. (in Chinese with English abstract)

Gupta, S. K. and T. C. Ghosh. 2001. Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. Gene 273:63–70.

Hall, T. 2001. BioEdit version 5.0.6. North Carolina State Univ. 192 pp.

Hill, M. O. 1973. Reciprocal averaging: An eigenvector method of ordination. J. Ecol. 61:237–249.

Khush, G. S. and E. R. Angeles. 1999. A new gene for resistance to race 6 of bacterial blight in rice, *Oryza sativa* L. Rice Genet. Newsl. 16:92–93.

Leps, J. and P. Smilauer. 1999. Multivariate Analysis of Ecological Data. Faculty of Biological Science. Univ. South Bohemia, Ceske Budejovice. 110 pp.

Mathews, C. K., K. E. van Holde, and K. G. Ahern. 2000. Biochemistry. Addison Wesley Longman, Inc. USA. p.126–160.

Mohna, F. I., P. Shen, S. C. Jong, and K. Orikono. 1992. Molecular evidence supports the separation of *Lentinula edodes* from *Lentinus* and related genera. Can. J. Bot. 70:2446–2452.

Monosi, B., R. J. Wisser, L. Pennill, and S. H. Hulbert. 2004. Full-genome analysis of resistance gene homologues in rice. Theor. Appl. Genet. 109:1434–1477.

Mumm, R. H., J. Hubert, and J. W. Dudley. 1994. A classification of 148 U.S. maize inbreds: II. Validation of cluster analysis based on RFLPs. Crop Sci. 34:852–865.

Nauchitel, V. V. and R. L. Somorjai. 1994. Spatial and free energy distribution patterns of amino acid residues in water soluble proteins. Biophys. Chem. 51:327–336.

Shi, X. H., X. R. Liu, L. Luo, W. B. Liu, and J. Xu. 2005. A research of amino acid order based on amino acid classification. J. Biomath. 20:491–495.

Song, W. Y., G. L.Wang, L. L. Chen, H. S. Kim, L. Y. Pi, T. Holsten, J. Gardner, B. Wang, W. X. Zhai, L. H. Zhu, C. Frauquet, and P. Ronald. 1995. A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*. Science 270:1804–1806.

Statsoft Inc. 2002. STATISTICA: The small book chinese version. USA. 144 pp.

Sun, X. L., Y. L. Cao, and S. P. Wang. 2006. Point mutations with positive selection were a major force during the evolution of a receptor-kinase resistance gene family of rice. Plant Physiol. 140:998–1008.

Tan, Q., K. Brusgaard, T. A. Kruse, E. Oakeley, B. Hemmings, H. Beck-Nielsen, L. Hansen, and M. Gaster. 2004. Correspondence analysis of microarray time-course data in case-control design. J. Biomed. Inform. 37:358–365.

Tang, H., J. G. J. Wyckoff, J. Lu, and C. I. Wu. 2004. A universal evolutionary index for amino acid changes. Mol. Biol. Evol. 21:1548–1556.

Tekaia, F., E. Yeramoan, and B. Dujon. 2002. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: A global picture with correspondence analysis. Gene 297:51–60.

The Rice Chromosome 10 Sequencing Consortium. 2003. In-depth view of structure, activity, and evolution of rice chromosome 10. Science 300:1566–1569.

Villar, H. O. and L. M. Kauvar. 1994. Amino acid preferences at protein binding sites. FEBS Lett. 349:125–130.

Villar, H. O. and R. T. Koehler. 2000. Amino acid preferences of small, naturally occurring polypeptides. Biopolymers 53:226–232.

Wang, A. J., C. X. Zhu, and F. J. Wen. 2002. Molecular biology of the resistance to blight disease in rice. J. Shandong Agric. Univ. (Nat. Sci.) 33:101–106. (in Chinese with English abstract)

Wang, J. and W. Wang. 1999. A computational approach to simplifying the protein folding alphabet. Nat. Struc. Biol. 6:1033–1038.

Yamazaki, Y. and P. Jaiswal. 2005. Biological ontologies in rice databases. An introduction to the activities in Gramene and Oryzabase. Plant Cell Physiol. 46:63–68.

Yampolsky, L. Y. and A. Stoltzfus. 2005. The exchangeability of amino acids in proteins. Genetics 170:1459–1472.

Yoshimura, S., A. Yoshimura, N. Iwata, S. R. McCouch, M. L. Abenes, M. Baraoidan, T. W. Mew, and R. J. Nelson. 1995. Tagging and combining bacterial blight resistance genes in rice using RAPD and RFLP markers. Mol. Breed. 1:375–387.

Yoshimura, S., U. Yamanouchi, Y. Katayose, S. Toki, Z. X. Wang, I. Kono, N. Kurta, M. Yano, N. Iwata, and T. Sasaki. 1998. Expression of *Xa1*, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. Proc. Nat. Acad. Sci. 95:1663–1668.

Zhou, T., Y. Wang, J. Q. Chen, H. Araki, Z. Jing, L. Jiang, J. Shen, and D. Tian. 2004. Genome-wide identification of NBS genes in *japonica* rice reveals significant expansion of divergent non-TIR NBS-LRR genes. Mol. Genet. Genomics 271:402–415.

Multivariate Analysis of Amino Acid Composition in Proteins from Rice Bacterial Blight Resistance Genes [1]

Meng-Li Wei[2], Yi-Han Chung[3], Chun-Tang Lu[2], and Hsiu-Ying Lu[2,4]

Abstract

Wei, M. L., Y. H. Chung, C. T. Lu, and H. Y. Lu. 2008. Multivariate analysis of amino acid composition in proteins from rice bacterial blight resistance genes. J. Taiwan Agric. Res. 57:161–174.

Much information is stored in amino acid composition of proteins, while amino acid compositional features vary among species and among genes within species; thus, the analysis of amino acid composition of proteins becomes one of the important topics in bioinformatics research. The total amino acid usage is determined by obtaining a data matrix of multiple species or genes (observations) and 20 amino acid frequencies (variables). The multivariate analysis is well suited for exploring this multidimensional information. To accelerate the analysis of protein structure in rice bacterial blight resistance gene (*Xa*), it is essential to examine the amino acid composition of all completely sequenced *Xa* genes. Thus, using a total of 17 protein sequences of rice bacterial blight resistance genes, i.e., *Xa1*, *xa5*, *xa13*, *Xa13*, *Xa21*, *Xa26*, and *Xa27* collected from NCBI, as test data, the pattern of variation of amino acid composition in *Xa* genes was detected by the complementary use of correspondence analysis and cluster analysis. The results showed that *Xa* genes were divided into six groups according to their percentage compositions of amino acids. The *Xa1*, *Xa21*, *Xa26*, and *Xa27* genes were found to have leucine mostly. The *xa13*, *Xa13*, and *Xa27*genes had fewer serine but more alanine and glycine. The *xa13* and *Xa13* genes also contained much valine. The *xa5* gene had much glutamic acid and threonine compared to other *Xa* genes. All of *Xa* genes had high-frequency hydrophobic amino acid in proteins. This research showed that multivariate statistical technique is useful for detecting the variation pattern of amino acid composition of proteins among *Xa* genes.

Key words: Rice (*Oryza sativa* L.), *Xa* genes, Protein structure, Composition and attribution of amino acids, Cluster analysis, Correspondence analysis.

---

1. Contribution No.2324 from Agricultural Research Institute, Council of Agriculture. Accepted: July 23, 2008.

2. Respectively, Assistant Researcher, Assistant Researcher and Senior Researcher. Crop Science Division, ARI, Wufeng, Taichung, Taiwan, ROC.

3. Master, Graduate Institute of Biotechnology, Chaoyang University of Technology, Wufeng, Taichung, Taiwan ROC.

4. Corresponding author, e-mail: iying@wufeng.tari.gov.tw; Fax: (04)23390528.