第十七屆旺宏科學獎

成果說明書

參賽編號：SA17-364

作品名稱：大腸桿菌、秀麗隱桿線蟲、烘培酵母菌和阿拉伯芥的粒線體之蛋白質關係演化網路

Protein-related evolution network of mitochondrion existing in Escherichia coli, Caenorhabditis elegans, Baker yeast and Arabidopsis

姓名：方詔陽

關鍵字：粒線體、分群、視覺化

摘要

在細胞中的粒腺體是一個提供細胞能量的胞器，扮演著很特殊的角色。粒線體不僅重要，也在演化上有著重要的地位。透過分析大腸桿菌 (Escherichia coli)、 秀麗隱桿線蟲（Caenorhabditis elegans）、酵母菌 (Yeast)和阿拉伯芥(Arabidopsis thaliana)低等至高等的生物粒線體的組成蛋白質序列，來探討彼此間的演化相似關係。從蛋白質資料網站下載蛋白質序列後，我們選用 MSC(minimum spanning clustering)演算法，利用尋找關係值最小的依據來進行分群。將這些分群過後的蛋白質序列變成節點，利用相似關係值形成距離概念，最後再將這些網路連結資料進行視覺化，形成清楚的蛋白質關係網路圖，再依據粒線體的組成構造和酶的種類(EC number)來分析探討其演化關係。

壹、 研究動機

粒線體是生物產生能量的場所，有自己的遺傳物質，粒線體不僅只提供能量，本身也參與了細胞分化、資訊傳遞核細胞凋亡，扮演著十分重要的角色。在這領域中，有一假說為內共生假說，說明著粒線體並不是一開始就存在於真核細胞內，那這種原本為原核生物的胞器，有著獨特的共生關係，這讓我想去探討粒線體在不同生物間的演化相似關係。而我們挑選了四種由低等至高等的生物，分別為大腸桿菌 (Escherichia coli)、 秀麗隱桿線蟲（Caenorhabditis elegans）、酵母菌 (Yeast)和阿拉伯芥(Arabidopsis thaliana)，利用蛋白質資料庫所表列的氨基酸序列，試圖利用 MSC (minimum spanning clustering)演算法去分析它們在粒線體中蛋白質關聯網在不同物種下演化相似關係[1]，並且深入探討粒線體中不同種酶的演化關聯程度。

貳、 研究目的

在這次的研究，我們想了解粒線體中各構造的蛋白質和粒線體中各種酶在不同物種下的演化相似關系。我們分別採用了大腸桿菌 (Escherichia coli)、 秀麗隱桿線蟲(Caenorhabditis elegans)、酵母菌 (Yeast)和阿拉伯芥(Arabidopsis thaliana)這四種較為指標性的生物，來探討其蛋白質演化相似關系是否符合粒線體的內共生假說。

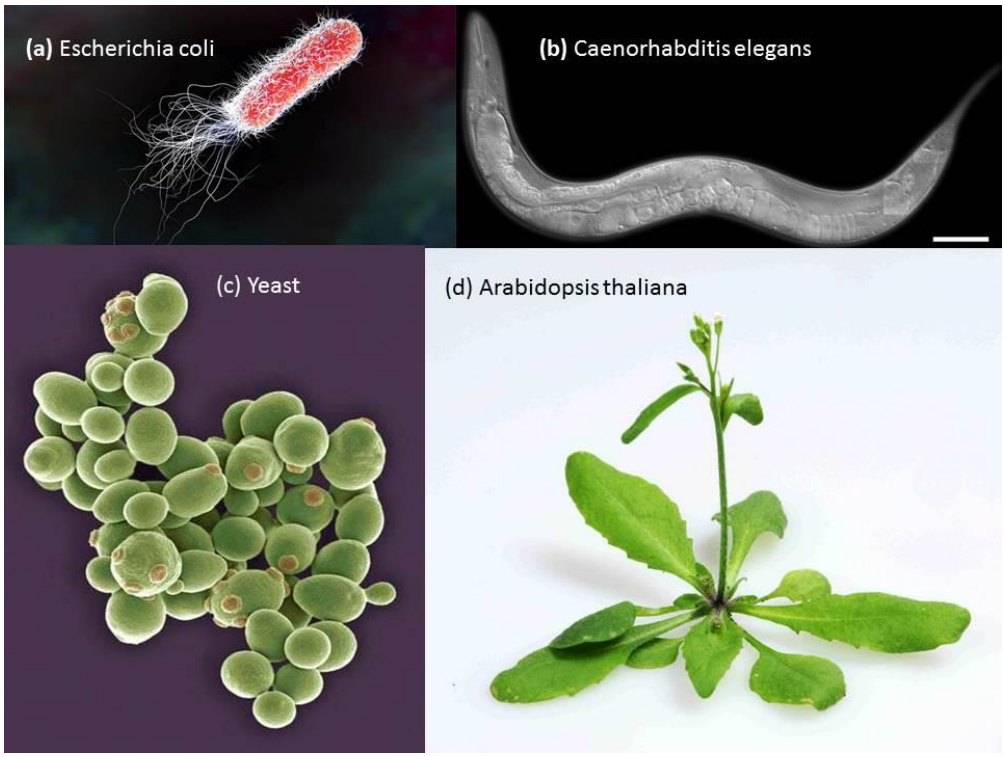


圖 1: 圖為本次研究的四種物種: (a)大腸桿菌 (Escherichia coli)、(b)秀麗隱桿線蟲（Caenorhabditis elegans）、(c) 酵母菌 (Yeast)、(d) 阿拉伯芥(Arabidopsis thaliana)[2-5]

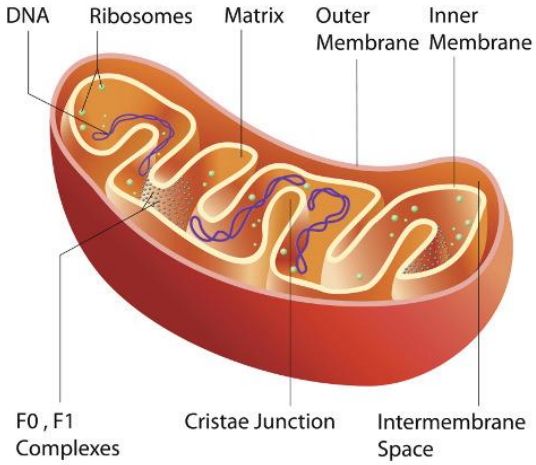


圖 2: 粒線體 （mitochondrion）[6]

參、 研究過程

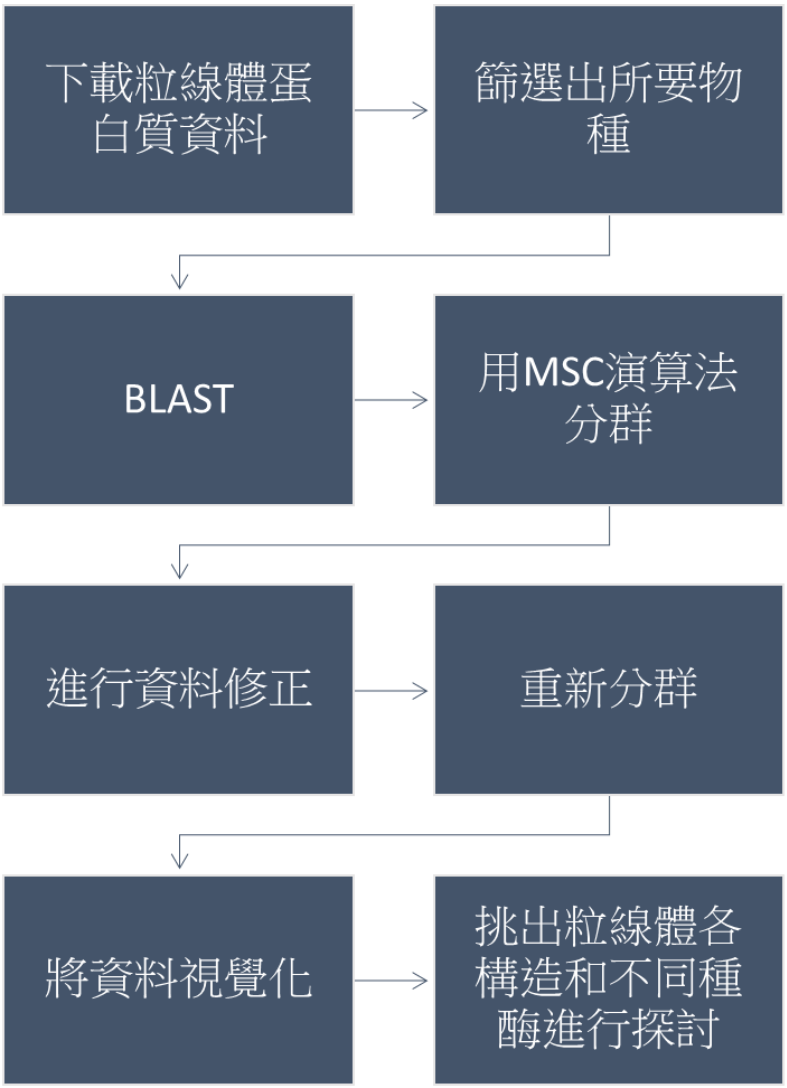


圖 3: 研究方法流程圖

- 利用 Uniprot 蛋白質資料庫網站將粒線體資料下載 [7]。
- 利用 BLAST(Basic Local Alignment Search Tool)比對每一筆粒線體蛋白質序列，並且給出一個關係值 [8]。
- 用 MSC 演算法分群，從中觀察反覆修正資料，剔除偏差分群 [1]。
- 將分群後的資料利用 Cytoscape 軟體進行視覺化 [9]。
- 利用 Matlab 程式探討分析各物種間屬於相同構造或相同種酶之間的關係 [10]。

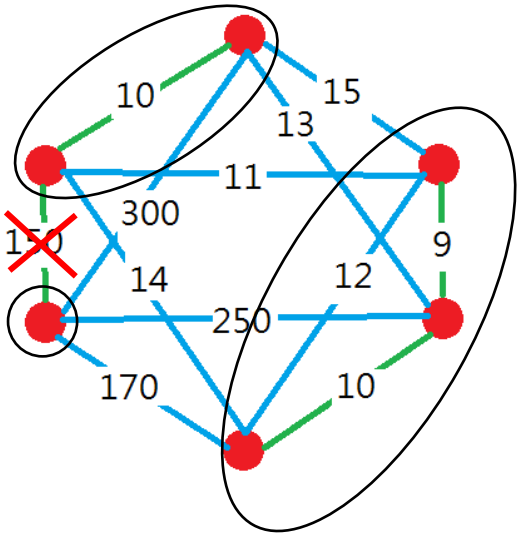


圖 4: MSC (Minimum Spanning Clustering) 演算法的演示圖，將其距離過大的連結刪除，如圖所示，原本兩群被分成了三群。

肆、 研究結論

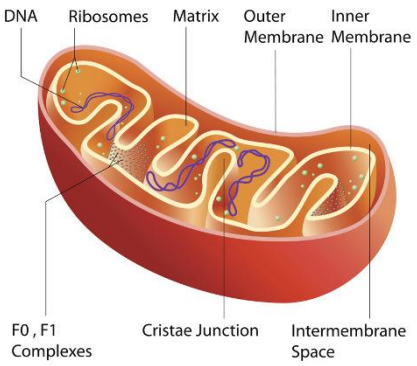
一、透過 Uniprot 下載蛋白質資料後，此研究原始包含 24130 個序列，為了能夠分析我們寫了幾隻程式將要的物種資料挑選出來，共 4004 筆，進行 BLAST 和分群後，由於 MSC 演算法的關係，某些節點間的距離實在太大，因此我們劃出距離分布圖後，訂出截止值，如圖 5，將比截止值大的連結刪除。刪除偏差的距離後，剩餘 2186 點成群，最後找出粒線體各構造中的主要蛋白質分群，就開始繪製網路分布圖。

二、粒線體蛋白質資料視覺化後，發覺到內膜與基質的網路相較於外膜與膜間隙來說，內膜與基質的關聯度相對較高，而我們的結果也恰好支持內共生假說，由於粒線體自帶遺傳物質，所以在演化上，內膜和基質的演化程度相對較小，進而關係較相似；反之外膜和膜間隙則會因生物體的不同而有不同種的外膜和基質，所以外膜和基質的演化程度較大，進而相關性較小。

三、在觀察基質中的蛋白質網路圖時，發現到大部分的蛋白質都有成群，並沒有落單的節點，這與內膜的結果剛好相反，代表大部分的蛋白質都有一定的關聯性，也是如此，我們推論基質中的蛋白質有相當的演化關係，但由於取樣只有 4 種，所以仍需進一步去探討。

四、在這次的分群視覺化後，對於資料的選用和分類要在更佳，這次的資料選用存在著非常不均勻的樣本數目，極有可能會造成分析誤差；在找尋粒線體中各構造的蛋白質在未來應更完整全面，目前的結果資料只能進行部分推論，這些是未來實驗可改進的方向。

表 1:粒線體中各構造的相關介紹

外膜（Outer membrane）	為一 6-7nm 厚的平滑膜，包含許多電荷敏感型（voltage-dependent）離子通道（也叫粒線體孔蛋白），因此小分子可以穿過外膜進入膜間腔，膜上有蛋白質受器，也有許多酵素。
內膜（Inter membrane）	與外膜相比，較薄，向內形成許多 cristae（摺層），顯著增加接觸面積。膜產生摺層的三樣主要功能；產生呼吸傳遞鏈氧化反應、合成 ATP、調節代謝產物的運輸進出基質。
膜間隙（Intermembrane space）	位於內外膜之間，使用來自內膜產生的 ATP，富含特別的酵素包含 creatine kinase，adenylate kinase，cyochrome c（在 apoptosis 中扮演重要角色）
基質（Matrix）	內膜內側所包圍住的區域稱為基質，內含酵素，含重要成分及功能：檸檬酸循環(citric acid cycle, TCA cycle)所需之酵素（除了 succinate dehydrogenase 是嵌在內膜上的）、脂肪酸代謝所需之酵素、粒線體 DNA（mtDNA）、核醣體、轉錄 mtDNA 所需之酵素。
	

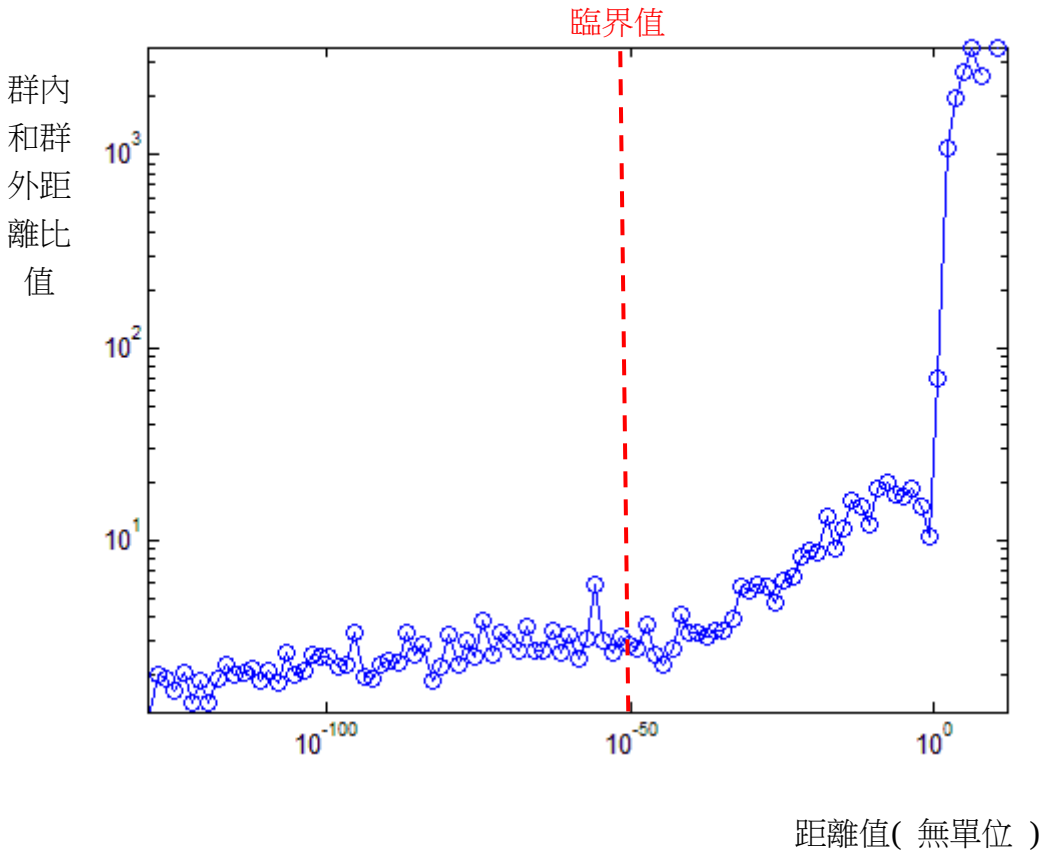


圖 5 此圖為群內和群外蛋白質的距離比值分布圖，縱軸為比值、橫軸為距離數值，能見到在 10^{-50} 後其數值激增，代表在這值之後，大多數比例的距離都屬於群外，那便是偏差值，由於距離太遠，因為相關性極低必須剔除連結，自立成一群。

表 2: 此研究原始包含蛋白質序列，共 4004 條序列

物種	蛋白質序列數量
大腸桿菌 (*Escherichia coli*)	65
秀麗隱桿線蟲 (*Caenorhabditis elegans*)	313
烘焙用酵母菌 (*Baker's Yeast*)	1478
阿拉伯芥 (*Arabidopsis thaliana*)	1323
總共	4004

由圖 5 可得由群內和群外蛋白質的距離比值分布圖，能見到在 10^{-50} 後其數值激增，代表在這值之後，大多數比例的距離都屬於群外，那即為偏差臨界值，由於蛋白質之間距離太遠，相關性極低必須剔除連結。另一種分類研究為 EC number，EC 編號是酶學委員會為酶所製作的一套編號分類法，是以每種酶所催化的化學反應為分類基礎，EC1 (氧化還原酶)催化氧化／還原作用；將氫及氧原子，或電子從一物質轉移至另一物質。 EC 2 (轉移酶): 將官能團從一個物質轉移至另一物質。轉移的基團可以是甲基、醯基、氨基或磷酸。EC 3 (水解酶): 從底物以水解來生成兩個化合物。EC 4 (裂合酶): 非水解的增加或移走底物的某些基團。EC 5 (異構酶): 重整分子內的排列，例如將分子異構化。EC 6 (連接酶): 合成新的 C-O、C-S、C-N 或 C-C 等共價鍵，並同時分解三磷酸腺苷（ATP）來把兩個份子結合 [11]。

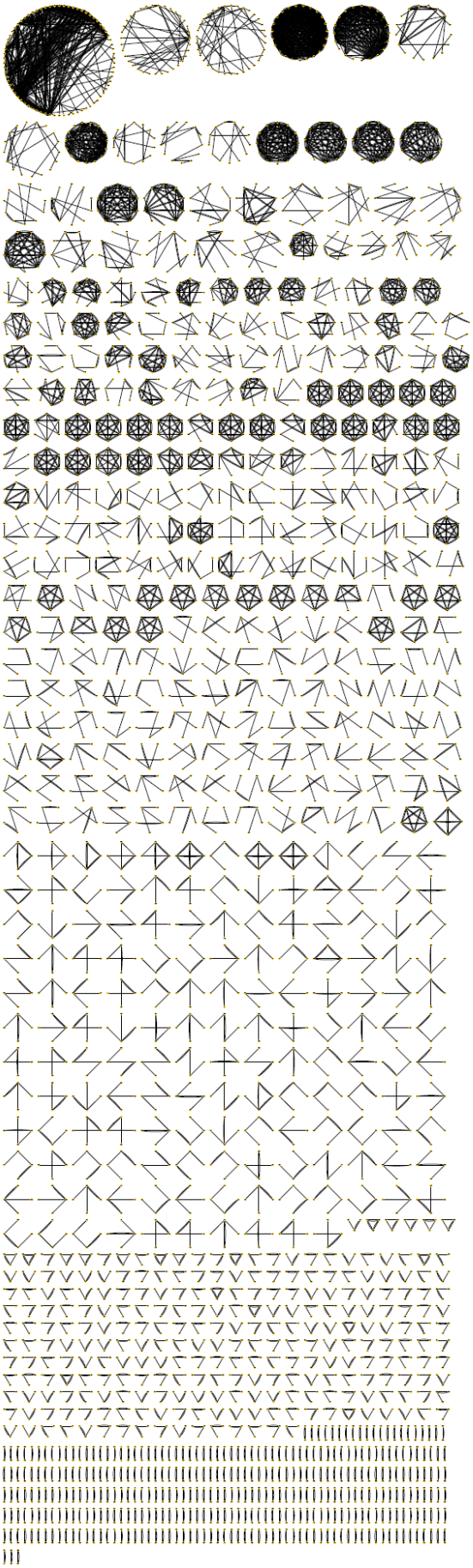


圖 6: 四種目標物種為蛋白質序列(4004 種)分群後的視覺化網路圖，總圖表示

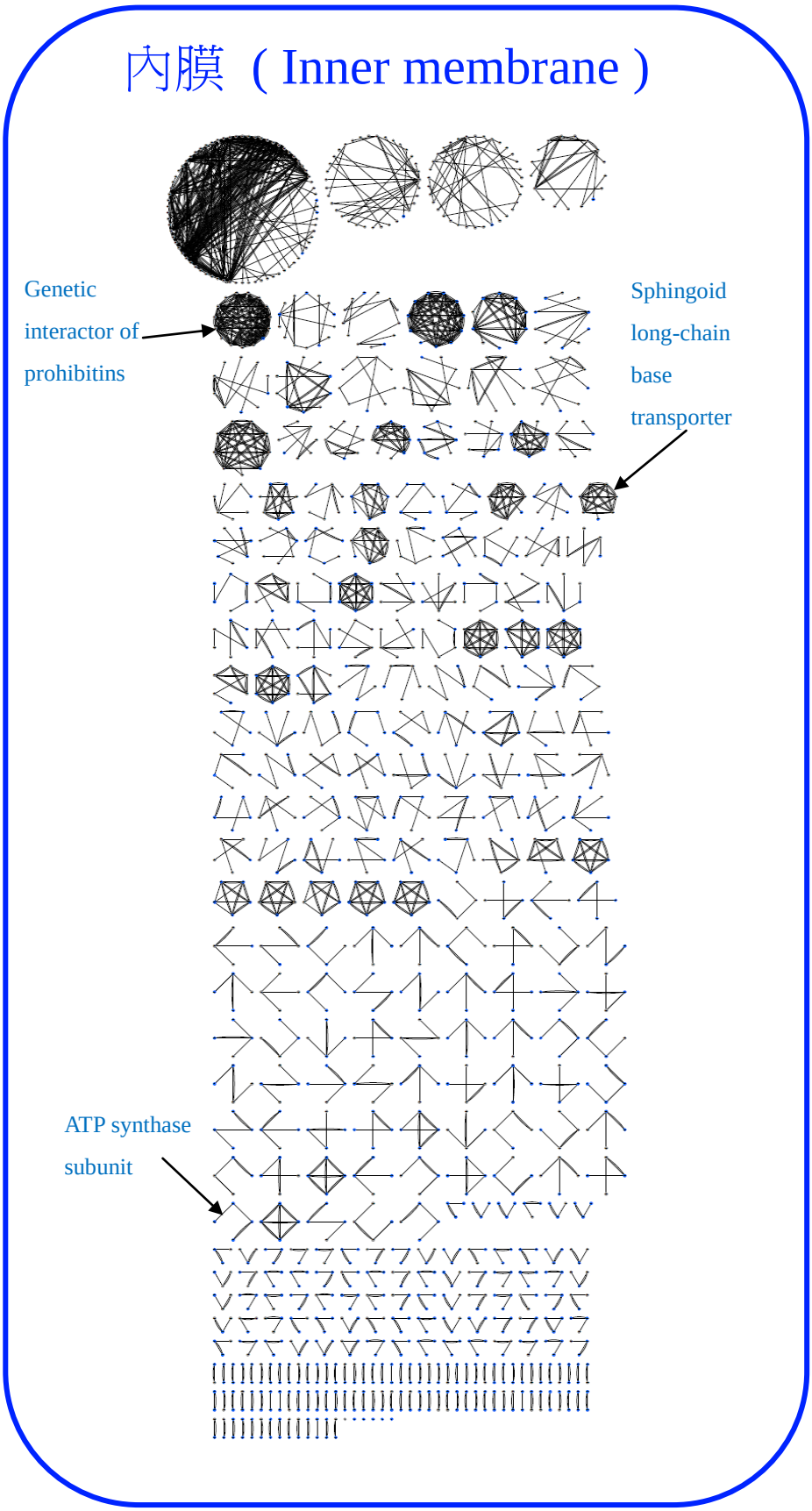


圖 7: 粒線體內膜中的蛋白質分布，藍色點為內膜中蛋白質

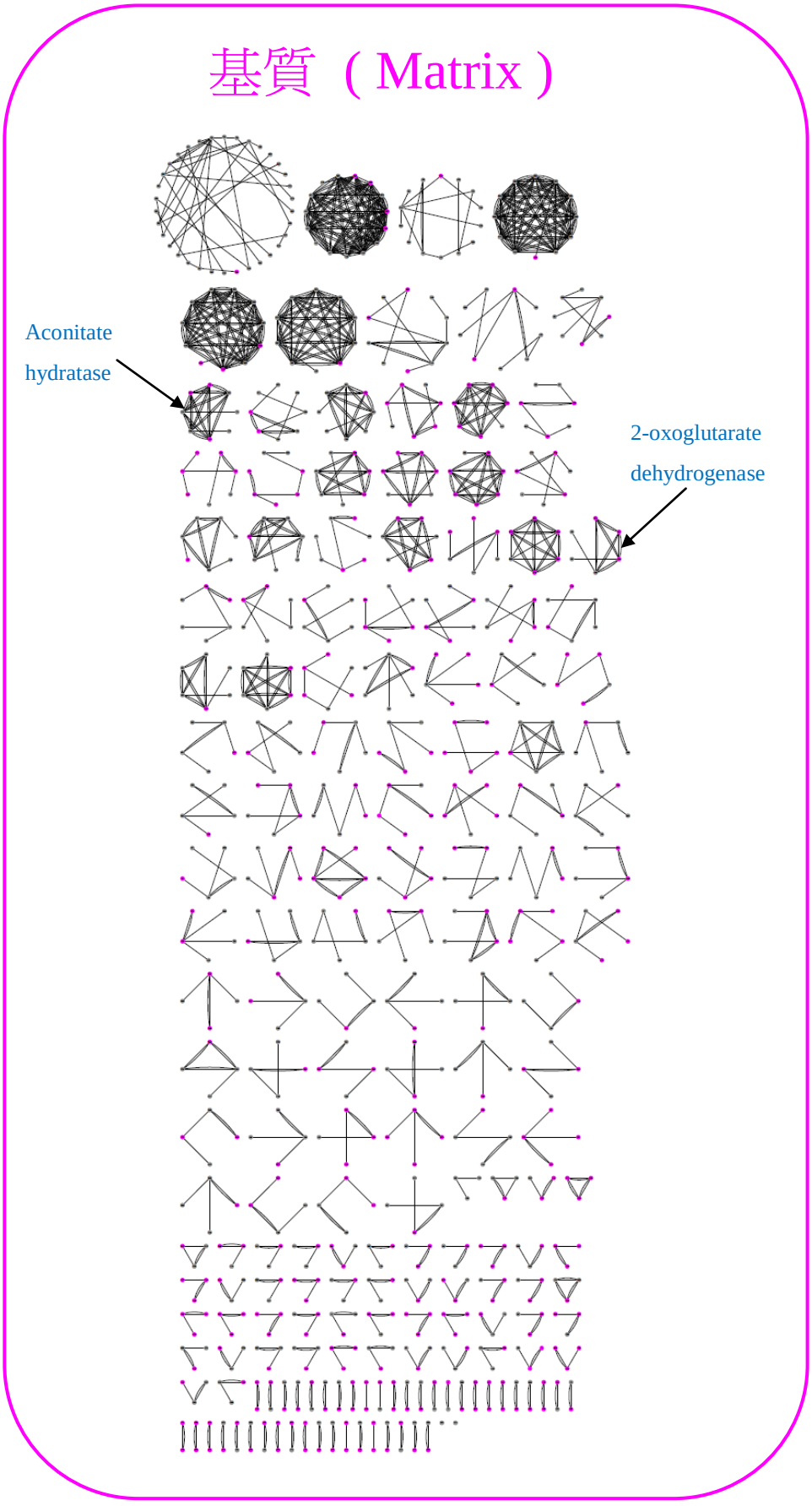


圖 8:　粒線體基質中的蛋白質分布，粉紅色點為基質的蛋白質

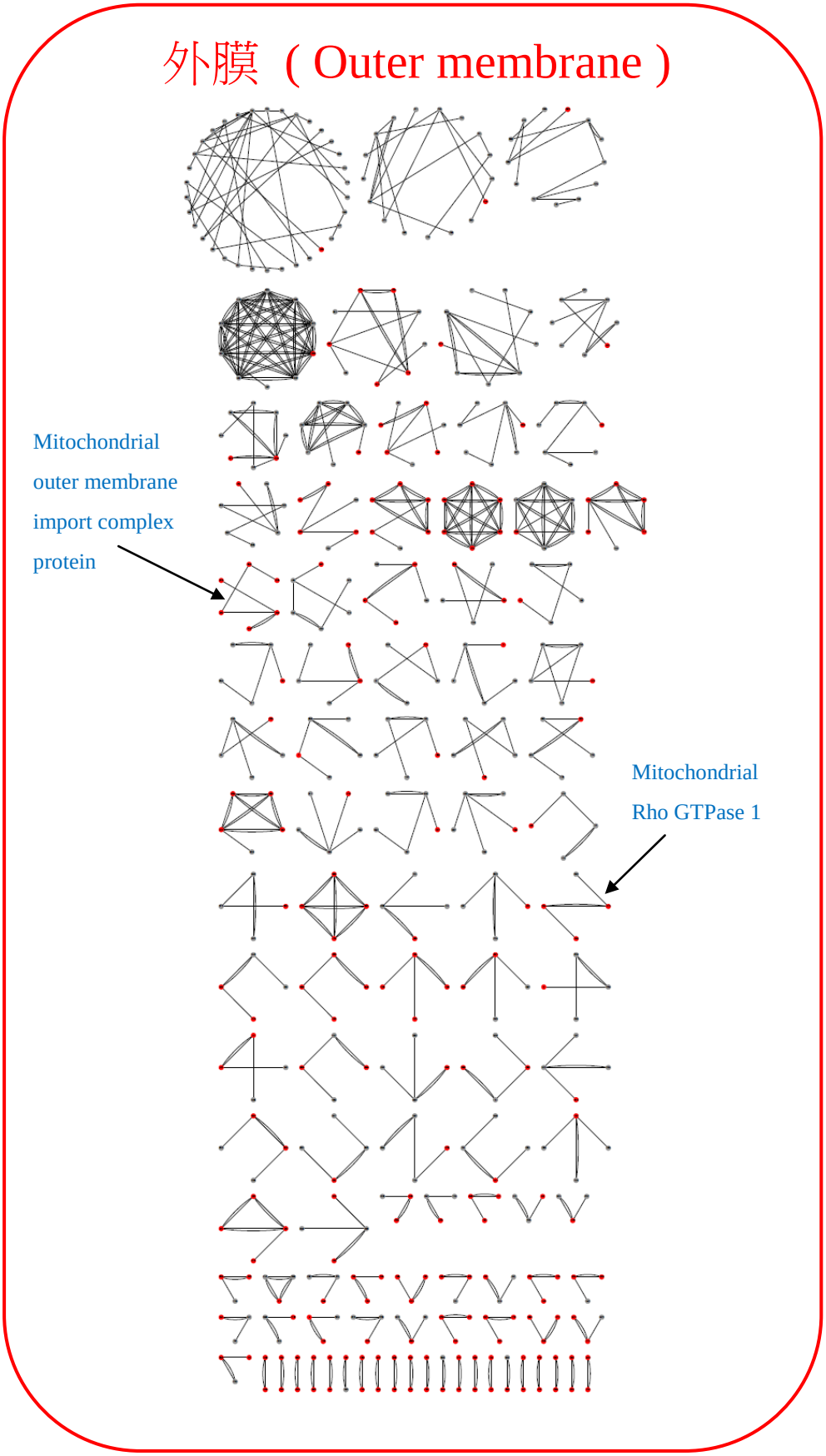


圖 9: 粒線體外膜中的蛋白質分布，紅色點為外膜中的蛋白質

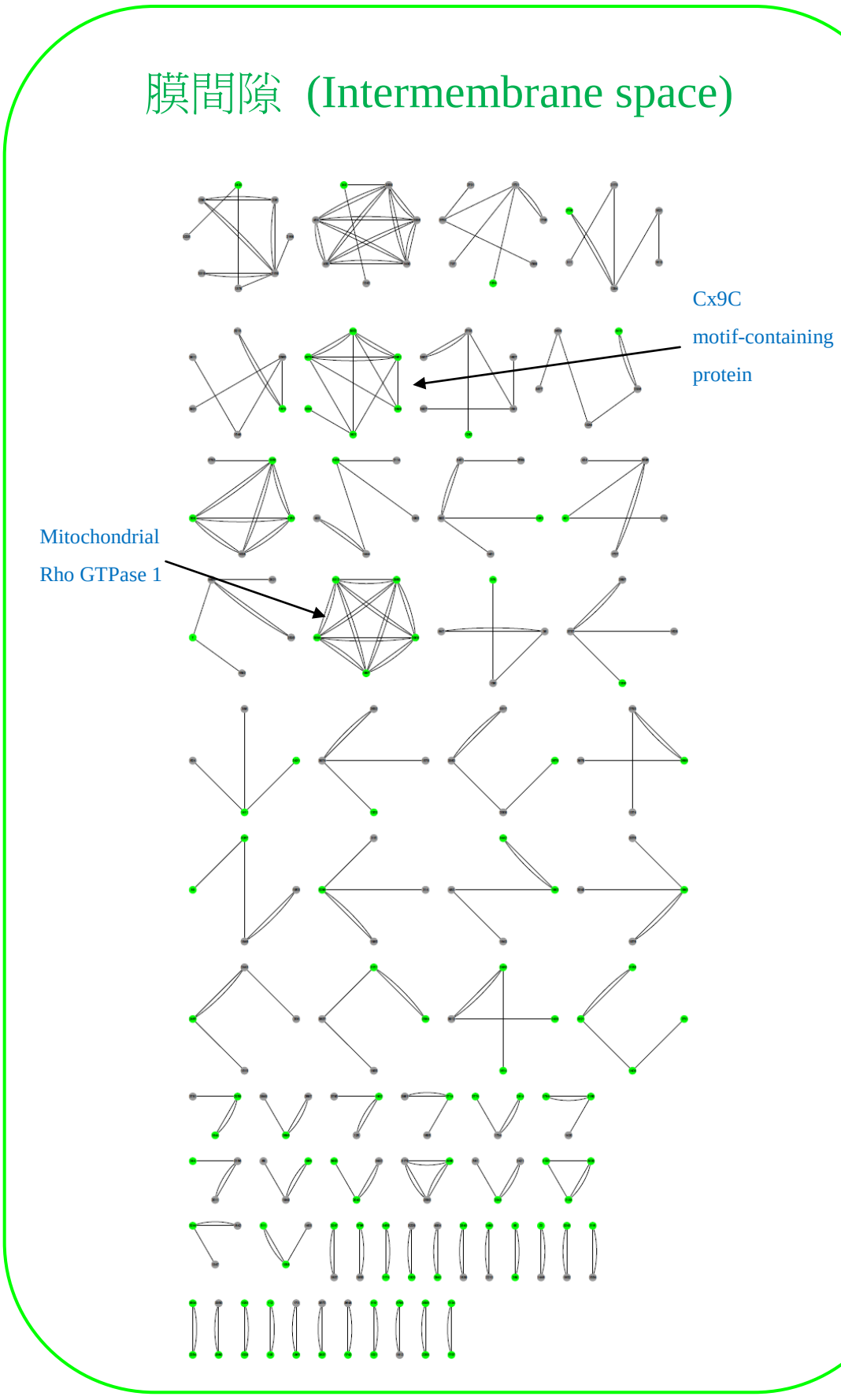


圖 10: 粒線體膜間隙中的蛋白質分布，綠色點為膜間隙中的蛋白質

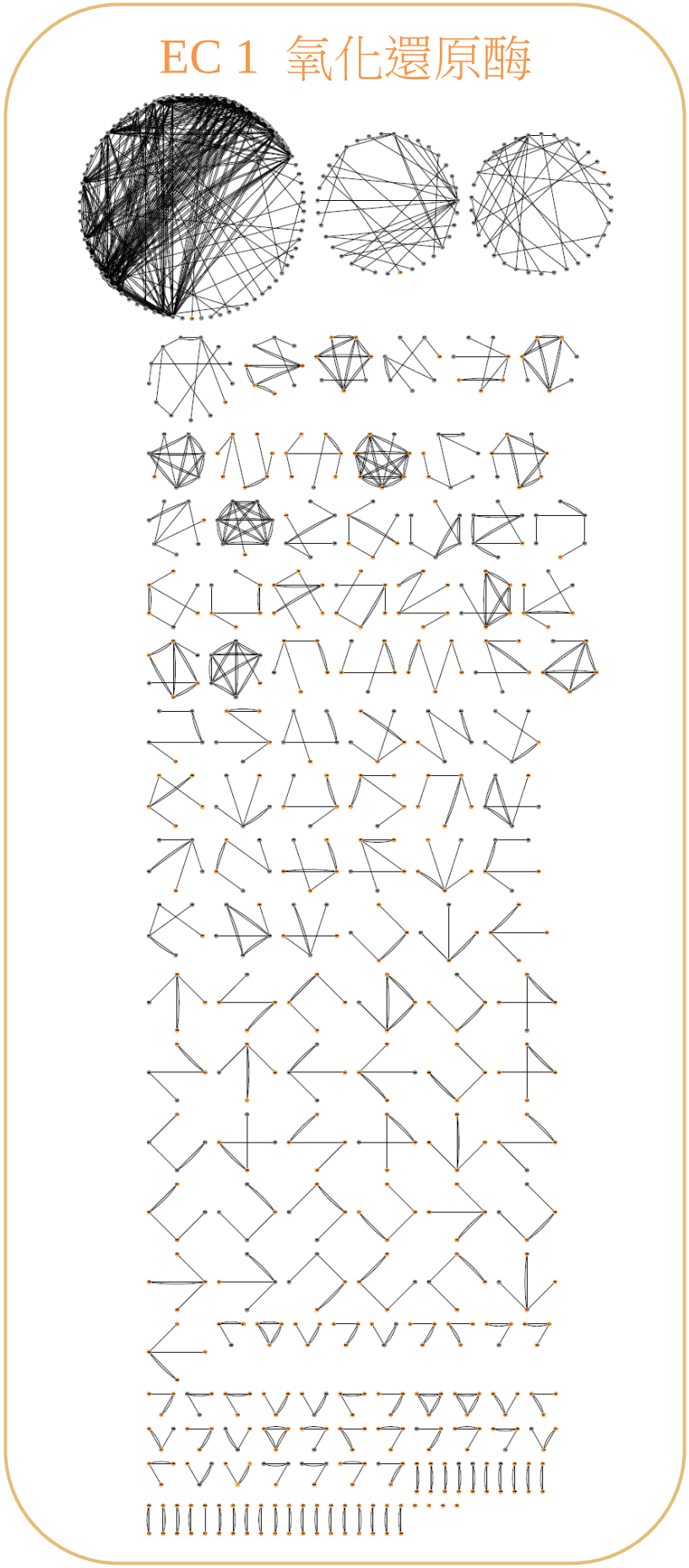


圖 11: 粒線體 EC1 氧化還原酶的蛋白質分布，橘色點為氧化還原酶

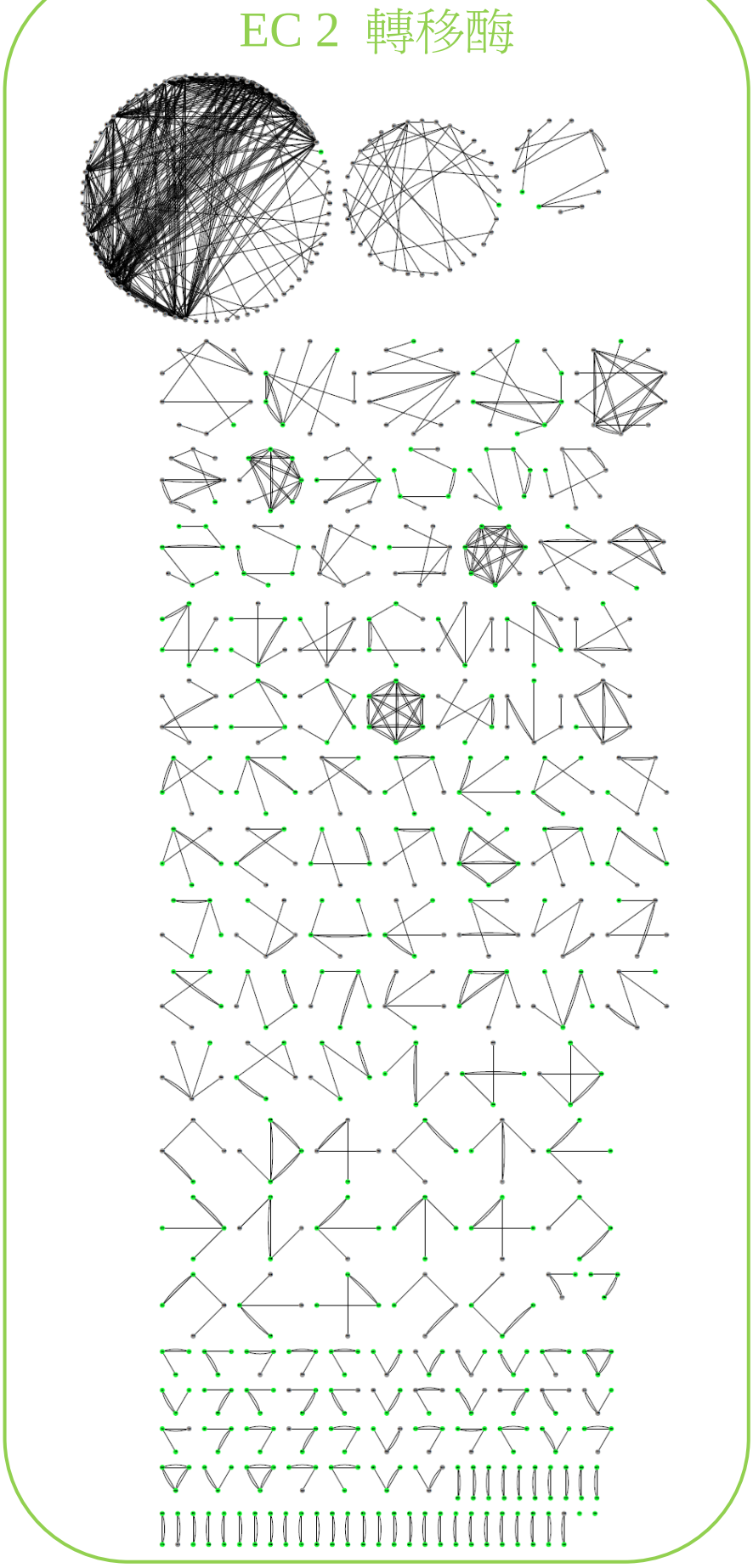


圖 12: 粒線體 EC2 轉移酶的蛋白質分布，綠色點為轉移酶

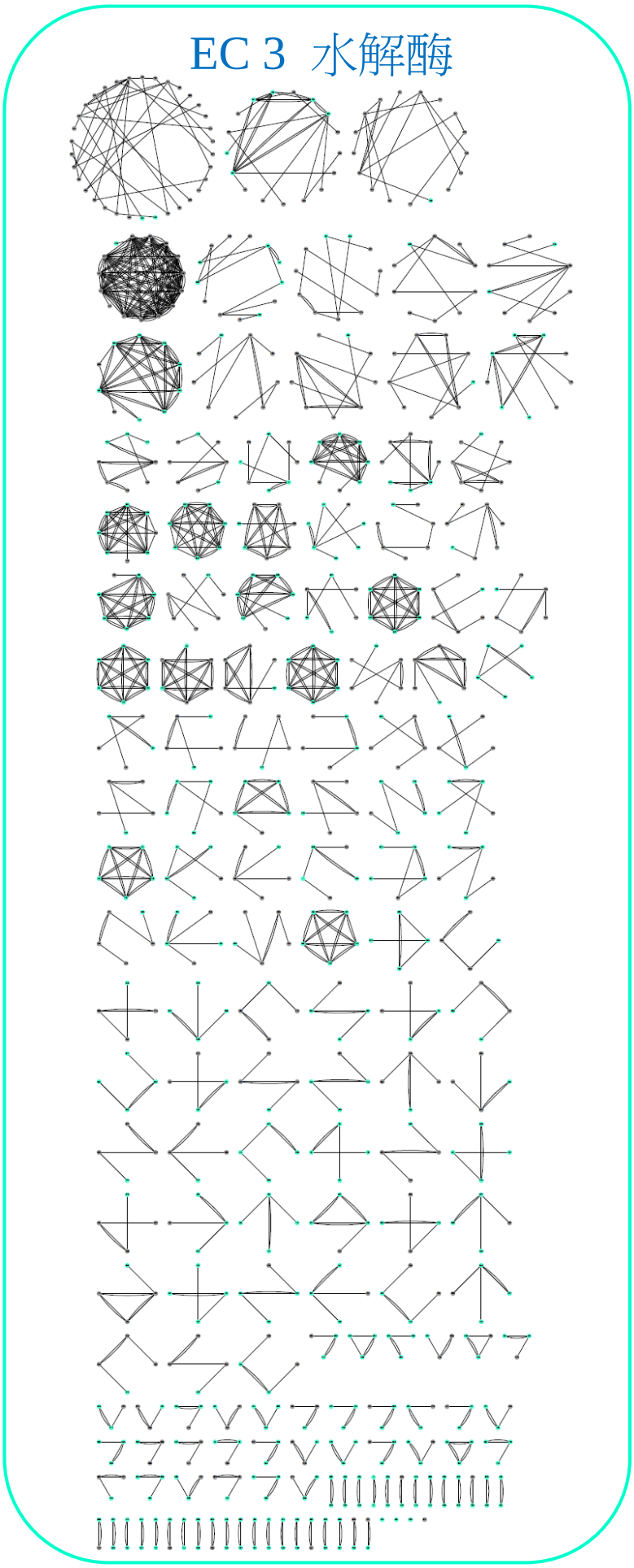


圖 13: 粒線體 EC3 水解酶的蛋白質分布，藍色點為水解酶

EC 4 裂合酶

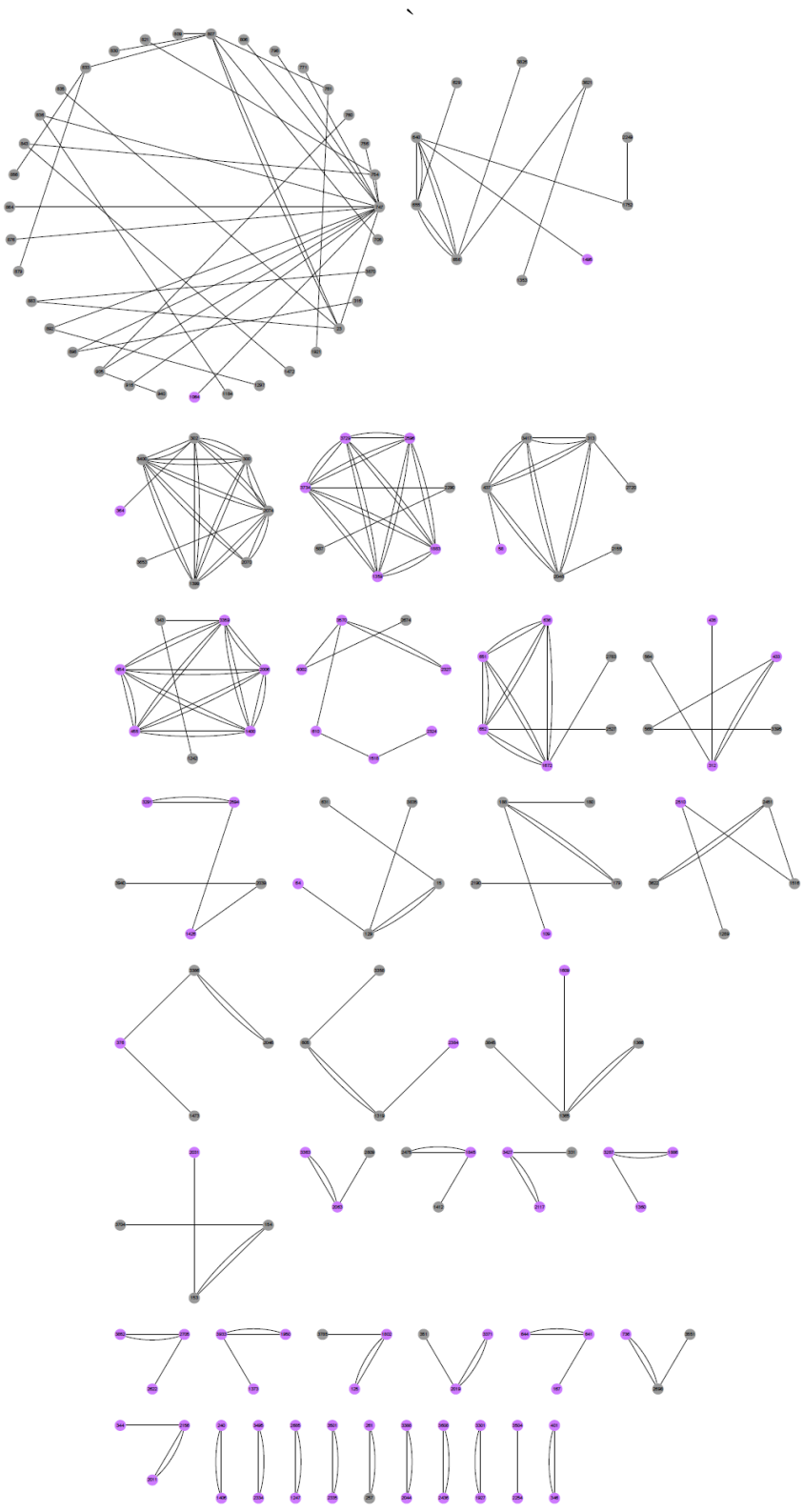


圖 14: 粒線體 EC4 裂合酶的蛋白質分布，粉紅色點為裂合酶

EC 5 異構酶

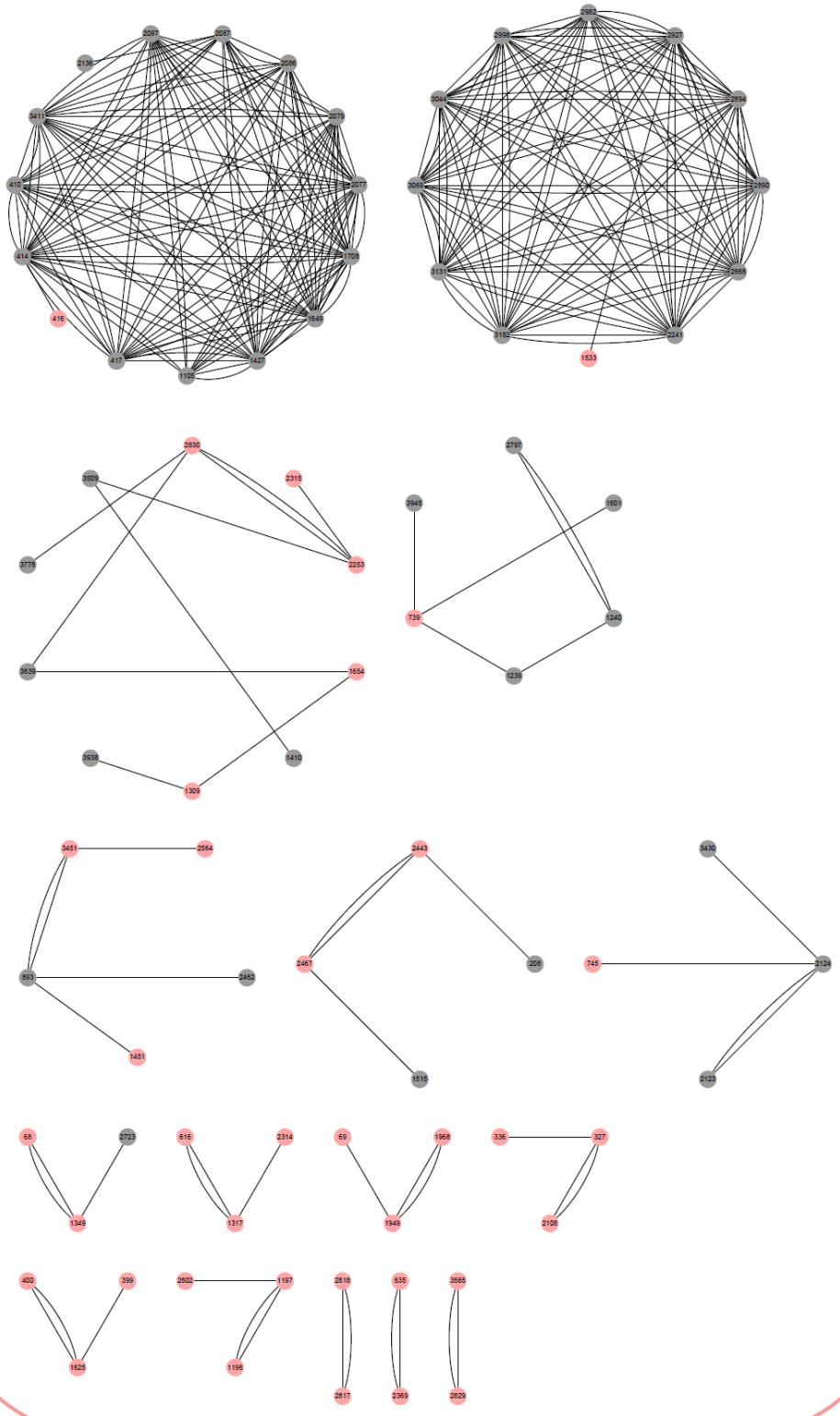


圖 15: 粒線體 EC5 異構酶的蛋白質分布，橘色點為異構酶

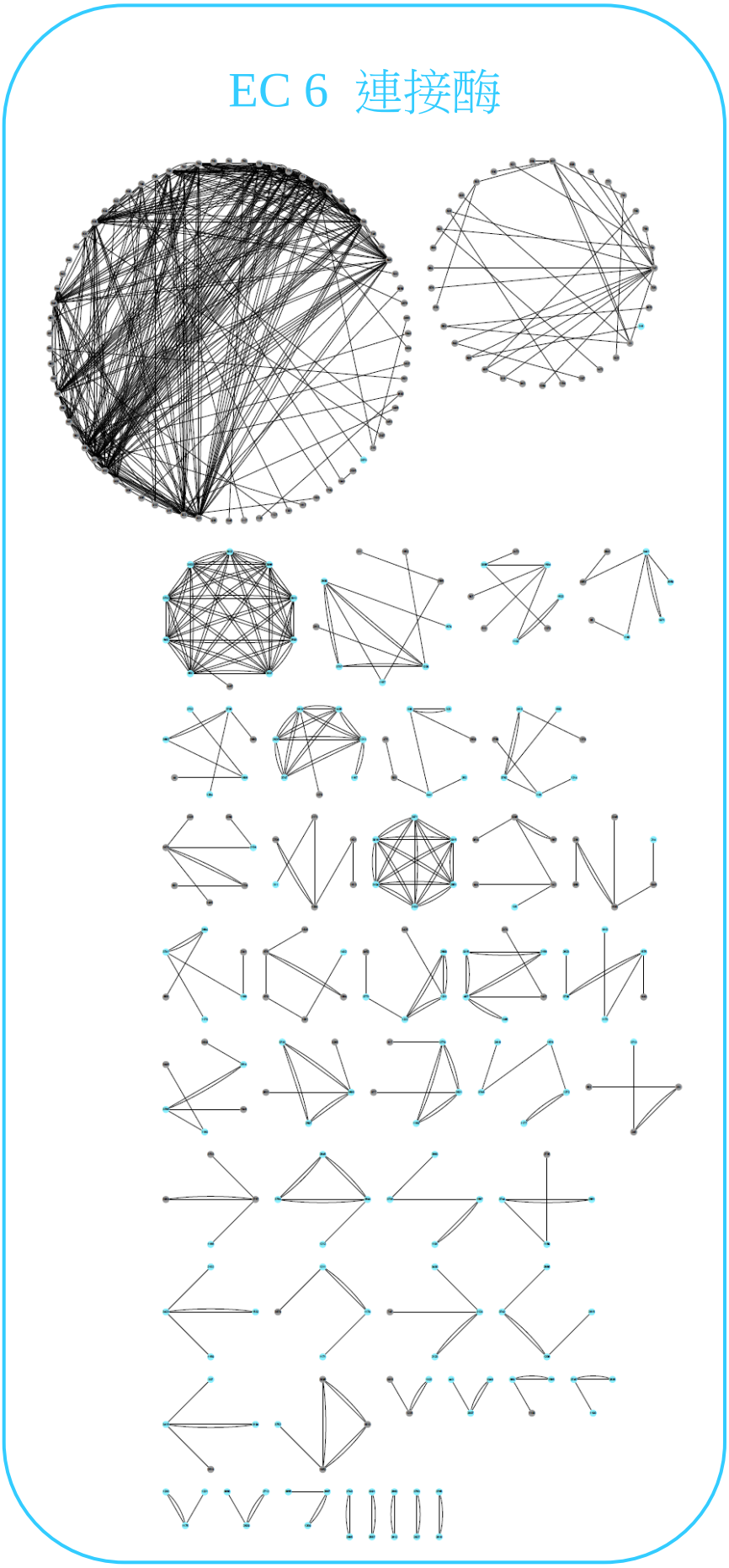


圖 16: 粒線體 EC6 連接酶的蛋白質分布，淺藍色點為連接酶

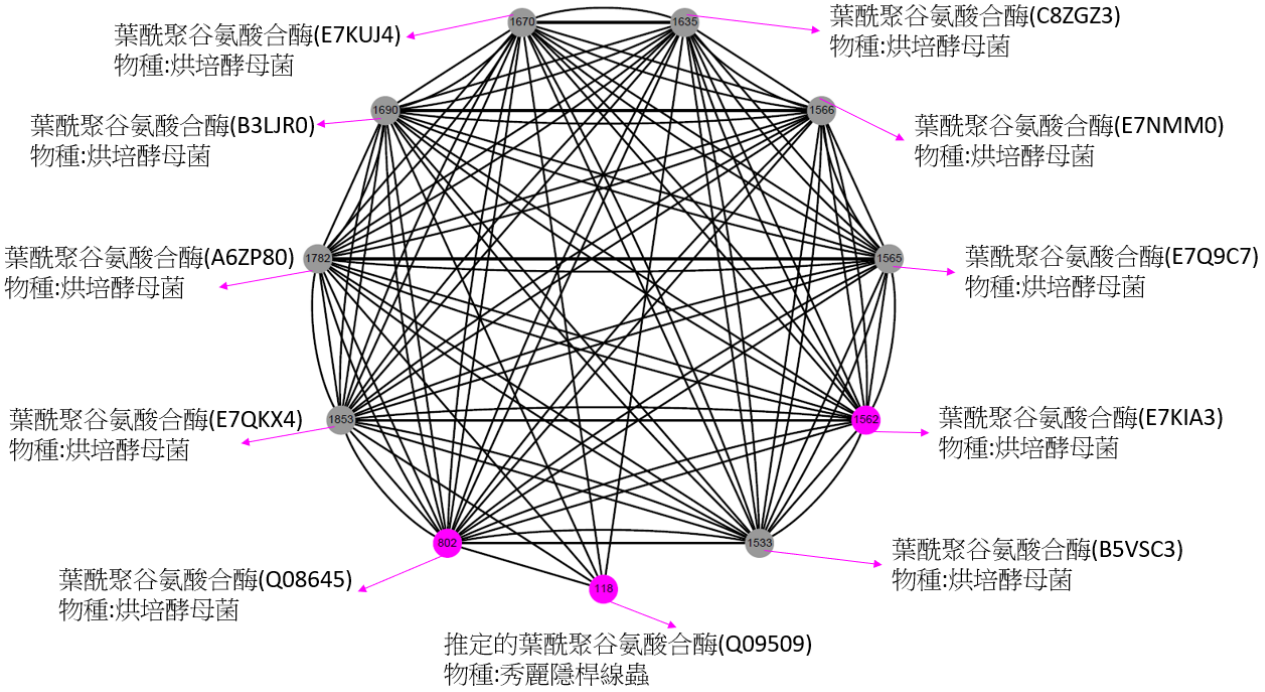


圖 17　基質酶蛋白網路部分關聯關係圖

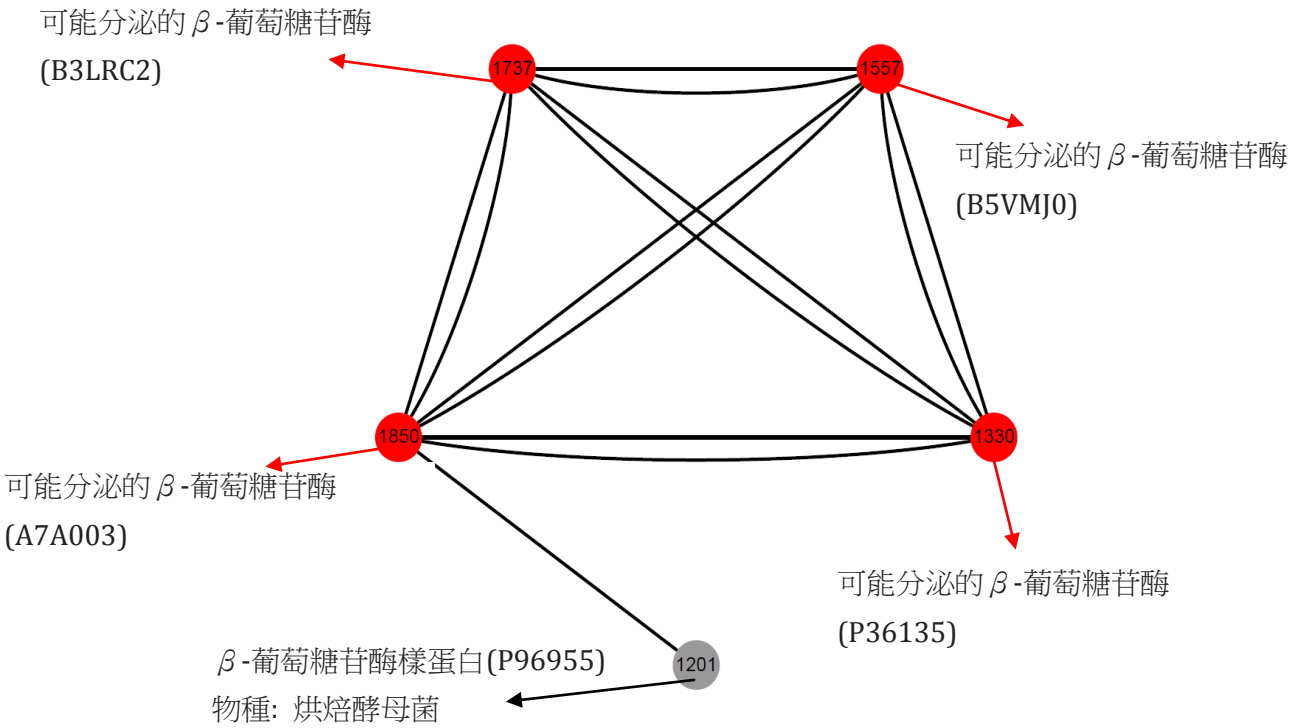


圖 18 外膜蛋白網路部分關聯關係圖

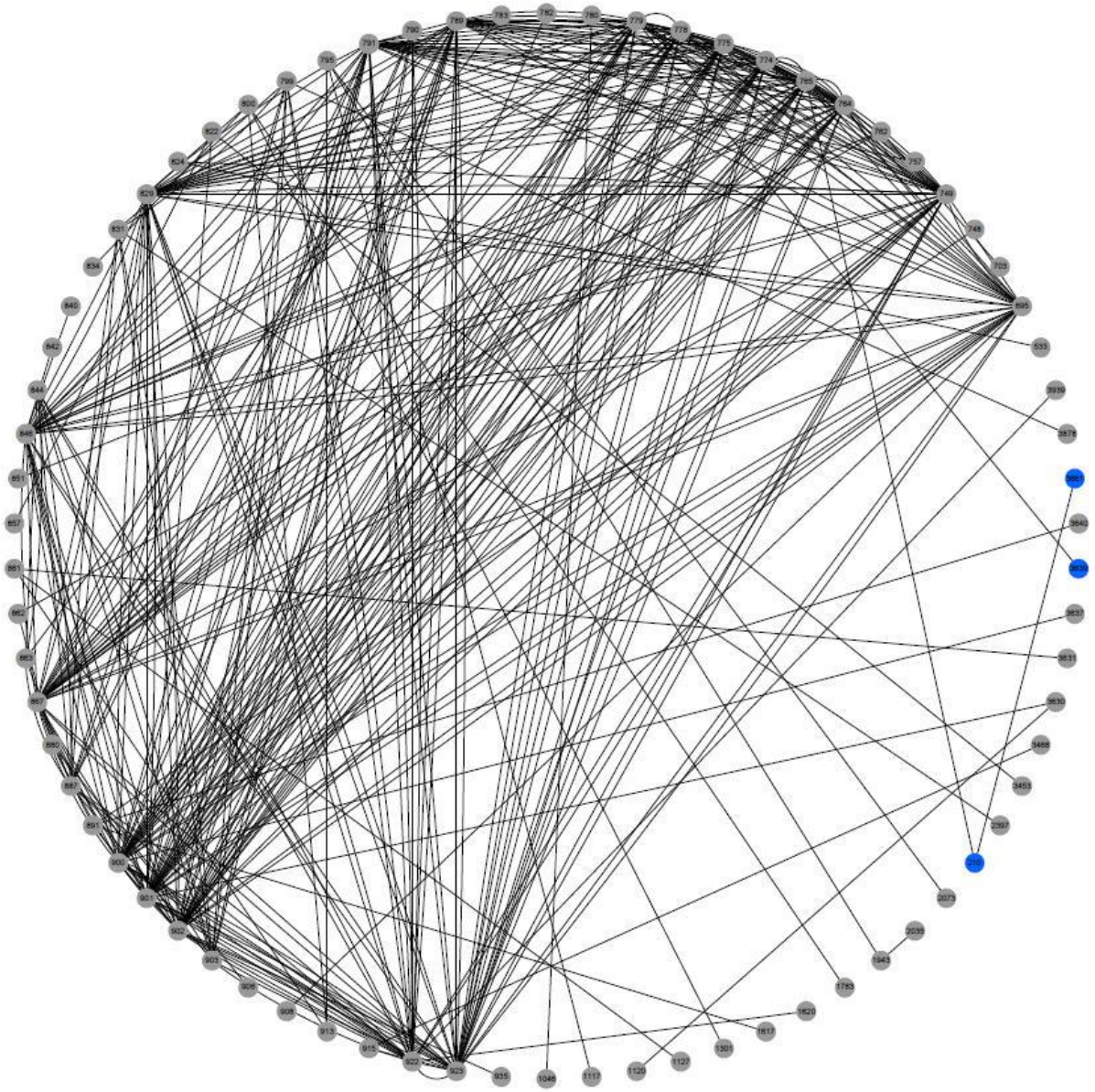


圖 19-1 內膜蛋白網路部分關聯關係圖，圖 7 的第 1 群放大圖，藍色點為內膜蛋白，只有 3 個點，灰色為常見組成蛋白

內膜蛋白網路組成的第 1 群放大圖，共有 74 個節點(3 個藍色點為內膜蛋白，其他為組成蛋白)，包含前 57 個 Arabidopsis thaliana，前 58-59，2 個 Caenorhabditis elegans，前 60-74 等 15 個 Yest(60-64，5 個 Baker's yeast，65-74，10 個 Fission yeast)，這裡沒有 E. coli 節點，詳細編號及基因名稱如下，其中編碼 695-923，連結數相對較高，均屬於含有五肽重複序列的蛋白質 (PPR: Pentatricopeptide repeat-containing protein: Arabidopsis thaliana)，五肽重複序列（PPR）是 35 個氨基酸的序列基序，含有五肽重複序列的蛋白質是植物界常見的蛋白質家族[12]，在阿拉伯芥基因組中已鑑定出約 450 種此類蛋白質，在水稻基因組中已鑑定出另外 477 種蛋白質。阿拉伯芥中的大量相互作用（通常基本上）與線粒體

和其他細胞器相互作用並且它們可能參與 RNA 編輯[12]，而編號 3630:O60142，3631:O42910，3637:O94368，3639:O14275，3640:O94615，為酵母菌的五肽重複序列（PPR），所以第 1 群以五肽重複序列（PPR）占大多數，利用 blast 方法成功的連結不同物種五肽重複序列（PPR）蛋白，3 個藍色點為內膜蛋白(210:Q96326: Arabidopsis，3639:O14275: Fission yeast，3661:Q10262:Fission yeast)。

表 3-1: 內膜蛋白網路第 1 群放大圖每一個點所對應到的蛋白質序列資料(編號:基因名)
210:Q96326，533:Q93Z16，695:Q9SH60，703:Q9LUD6，748:Q84VG6，749:Q9CAM8，757:Q8L6Y7，762:Q9LQQ1，764:Q0WKV3，765:Q9LPX2，774:Q9LQ15，775:Q9CAN6，778:Q9ASZ8，779:P0C7Q7，780:Q9SAD9，782:P0C7Q9，783:Q9FZD1，789:Q9SXD8，790:Q9SI78，791:Q9CAN0，795:Q9LVD3，799:Q9SHK2，800:Q9M9X9，822:Q9SFV9，824:Q9LSQ2，829:Q9SH26，831:P0C7R3，834:Q8GZA6，840:O81028，842:P0C896，844:Q9LUR2，846:Q6NQ83，851:Q8LDU5，857:Q9FNG8，861:Q9LXF4，862:Q9FFE3，863:Q9FLL3，867:Q9C8T7，880:O49436，887:Q9FMD3，891:Q9FMF6，900:Q9SXD1，901:Q9LQ16，902:Q9CAN5，903:Q0WKZ3，906:Q9ZUU7，908:O49558，913:Q9LER0，915:Q94JX6，922:Q3ECK2，923:Q9LQ14，935:P92535，1046:P54150，1117:P92522，1120:P93301，1127:P92525，1301:Q8W4L5，1617:Q1NZ26，1620:Q17439，1783:P33204，1943:P40366，2035:Q02516，2073:P32477，2397:P38266，3453:Q9P6M7，3468:O14178，3630:O60142，3631:O42910，3637:O94368，3639:O14275，3640:O94615，3661:Q10262，3878:O94583，3939:Q9Y807

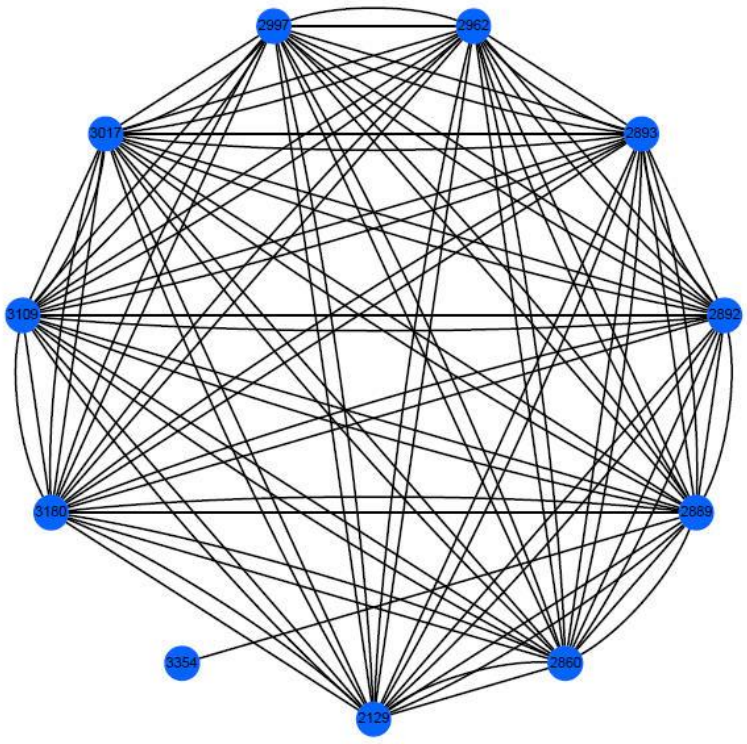


圖 19-2 內膜蛋白網路部分關聯關係圖，圖 7 的第 8 群放大圖，藍色點為內膜蛋白

共有 11 個節點均為 Folylpolyglutamate synthase，包含 11 個 Yest(1-10，1 個 Baker's yeast，12，1 個 Fission yeast)，這裡沒有 E. coli 節點。

表 3-2: 內膜蛋白網路第 8 群放大圖每一個點所對應到的蛋白質序列資料(編號:基因名)
2129:Q08645，2860:B5VSC3，2889:E7KIA3，2892:E7Q9C7，2893:E7NMM0，2962:C8ZGZ3，2997:E7KUJ4，3017:B3LJR0，3109:A6ZP80，3180:E7QKX4，3354:O74742

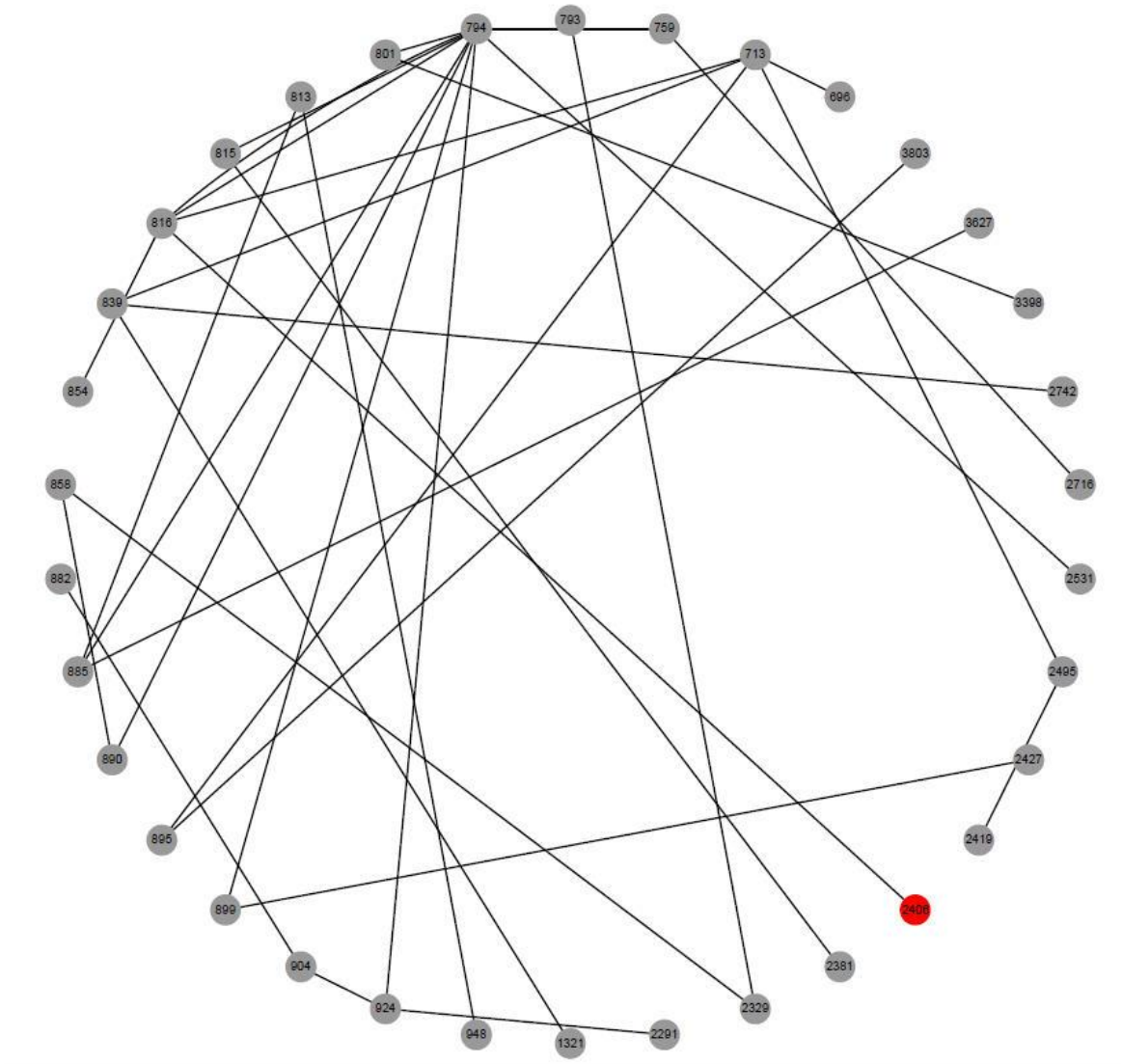


圖 20-1 外膜蛋白網路部分關聯關係圖，圖 9 外膜蛋白網路的第 1 群，紅色點為外膜蛋白

共有 33 個節點，包含前 21 個 Arabidopsis thaliana，前 22-33 等 12 個 Yest(22-31，10 個 Baker's yeast，32-33，2 個 Fission yeast)，這裡沒有 E. coli 節點，灰色為組成蛋白，而 1 個紅色點為外膜蛋白(2406:Q12106) 屬於 MDM10-補充蛋白 1，Baker's yeast。

表 4-1: 外膜蛋白網路第 1 群放大圖每一個點所對應到的蛋白質序列資料(編號:基因名)
696:Q3ECH5，713:Q9LFC5，759:Q9T0D6，793:Q9M3C6，794:Q9FIX3，801:O04491，813:Q9M2A1，815:Q9M316，816:Q0WVK7，839:Q9ZQF1，854:Q9SZ10，858:P0C8Q6，882:Q8VYR5，885:Q9FMQ1，890:Q9FIT7，895:Q9SAA6，899:Q9ZU27，904:Q9SAK0，924:Q9LR67，948:P92531，1321:Q66GI4，2291:Q02773，2329:P07390，2381:P87275，2406:Q12106，2419:Q06820，2427:P40050，2495:Q08818，2531:Q06504，2716:P32342，2742:P34231，3398:O13702，3627:Q10451

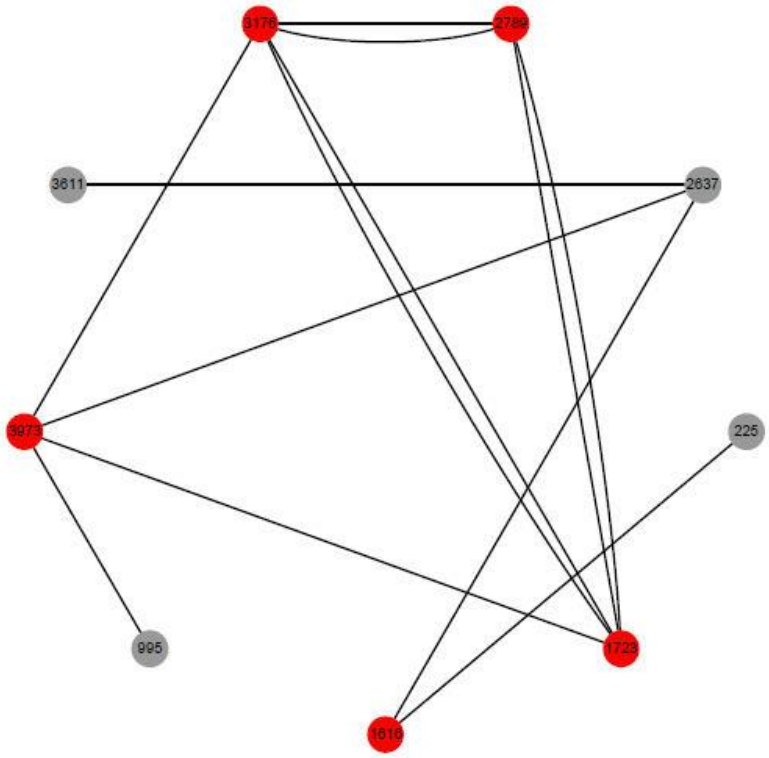


圖 20-2 外膜蛋白網路部分關聯關係圖，圖 9 的第 5 群放大圖，紅色點為外膜蛋白

共有 9 個節點，包含 1 個 Caenorhabditis elegans，6 個 Yeast，2 Arabidopsis thaliana，紅色點為外膜蛋白，1616:Q02335: Caenorhabditis elegans，1723:P07213: Baker's yeast，2789:P38825: Baker's yeast，3176:A6ZRW3: Baker's yeast，3973:O14217: Fission yeast。

表 4-2: 外膜蛋白網路第 5 群放大圖每一個點所對應到的蛋白質序列資料(編號:基因名)
1616:Q02335，1723:P07213，225:Q9M0Y6，2637:P15705，2789:P38825，3176:A6ZRW3，3611:O74991，3973:O14217，995:B7ZWR6

表 5: 總表部分節錄，每一個點所對應到的蛋白質序列資料

index	Entry	Gene names	Protein names
1	Q9ZP06	At1g53240 F12M16.14	Malate dehydrogenase 1, mitochondrial (EC 1.1.1.37) (Mitochondrial MDH1) (mMDH1) (Mitochondr
2	Q9LKA3	At3g15020 K15M2.16	Malate dehydrogenase 2, mitochondrial (EC 1.1.1.37) (Mitochondrial MDH2) (mMDH1) (Mitochondr
3	Q9FMU6	MPT3 AT5 PHT3;1 At5g14040 MUA22_4	Mitochondrial phosphate carrier protein 3, mitochondrial (Mitochondrial phosphate transporter 3) (MP
4	Q9LHE5	TOM40-1 At3g20000 MZE19.5	Mitochondrial import receptor subunit TOM40-1 (Translocase of outer membrane 40 kDa subunit hom
5	O04308	MPPA2 At3g16480 MDC8.11 T02O04.2	Probable mitochondrial-processing peptidase subunit alpha-2, chloroplastic/mitochondrial (EC 3.4.24.6
6	Q9CA93	BAC2 At1g79900 F19K16.14	Mitochondrial arginine transporter BAC2 (Mitochondrial basic amino acid carrier 2) (AtMBAC2)
7	Q42290	At3g02090 F1C9.12	Probable mitochondrial-processing peptidase subunit beta, mitochondrial (EC 3.4.24.64) (Beta-MPP)
8	O48528	OEP163 B14.7 At2g42210 T24P15.12	Outer envelope pore protein 16-3, chloroplastic/mitochondrial (Chloroplastic outer envelope pore prote
9	Q9LVM1	ABCB25 ATM3 STA1 At5g58270 MCK7.14	ABC transporter B family member 25, mitochondrial (ABC transporter ABCB.25) (AtABCB25) (ABC
10	Q9FUT3	ABCB23 ATM1 STA2 At4g28630 T5F17.80	ABC transporter B family member 23, mitochondrial (ABC transporter ABCB.23) (AtABCB23) (ABC
11	Q9FNR1	RBG3 GR-RBP3 MRBP2A ORRM3 At5g61030 MAF19_30	Glycine-rich RNA-binding protein 3, mitochondrial (AtGR-RBP3) (AtRBG3) (Mitochondrial RNA-bin
12	Q9ZU25	At1g51980 F5F19.4	Probable mitochondrial-processing peptidase subunit alpha-1, mitochondrial (EC 3.4.24.64) (Alpha-MP
13	Q9C909	RBG5 GR-RBP5 MRBP2B ORRM4 At1g74230 F1O17.10	Glycine-rich RNA-binding protein 5, mitochondrial (AtGR-RBP5) (AtRBG5) (Mitochondrial RNA-bin
14	Q9SVM8	RBG2 GR-RBP2 GRP2 MRBP1A At4g13850 F18A5.240	Glycine-rich RNA-binding protein 2, mitochondrial (AtGR-RBP2) (AtRBG2) (Glycine-rich protein 2)
15	Q66GP9	NOA1 NOS1 RIF1 At3g47450 T21L8.200	NO-associated protein 1, chloroplastic/mitochondrial (AtNOA1) (Dubious mitochondrial nitric oxide sy
16	P92947	MDAR5 MDAR6 MDHAR6 PMDAR-L PMDAR-S At1g63940 T12P18.4	Monodehydroascorbate reductase, chloroplastic/mitochondrial (EC 1.6.5.4) (Monodehydroascorbate rec
17	Q9FN42	CLPP2 NCLPP7 At5g23140 MYJ24.13	ATP-dependent Clp protease proteolytic subunit 2, mitochondrial (EC 3.4.21.92) (ATP-dependent Clp
18	Q9LIS2	RBG4 GR-RBP4 GRP4 MRBP1B At3g23830 F14O13_2	Glycine-rich RNA-binding protein 4, mitochondrial (AtGR-RBP4) (AtRBG4) (Glycine-rich protein 4)
19	Q9SJY5	PUMP5 DIC1 UCP5 At2g22500 F14M13.10	Mitochondrial uncoupling protein 5 (AtPUMP5) (Mitochondrial dicarboxylate carrier 1)
20	Q9M0G9	ABCB24 ATM2 STA3 At4g28620 T5F17.70	ABC transporter B family member 24, mitochondrial (ABC transporter ABCB.24) (AtABCB24) (ABC
21	Q8W3L1	MFDR ADXR At4g32360 F10M6.10 F8B4.60	NADPH:adrenodoxin oxidoreductase, mitochondrial (Adrenodoxin reductase) (EC 1.18.1.6) (Mitochon
22	Q9SZ45	MICU At4g32060 F10N7.140	Calcium uptake protein, mitochondrial (Mitochondrial calcium uniporter)
23	Q9FWA6	PCMP-E90 MEF13 At3g02330 F11A12.2 F14P3.1	Pentatricopeptide repeat-containing protein At3g02330, mitochondrial (Mitochondrial editing factor 13
24	Q8GYD7	ATG18C At2g40810 T20B5.1	Autophagy-related protein 18c (AtATG18c)
25	Q96252	At5g47030 MQD22.17	ATP synthase subunit delta', mitochondrial (F-ATPase delta' subunit)
26	Q9FT52	At3g52300 T25B15.70	ATP synthase subunit d, mitochondrial (ATPase subunit d)
27	Q0WPK3	ATG18D At3g56440 T5P19_90	Autophagy-related protein 18d (AtATG18d)
28	Q9M1Y0	ATG4B APG4B At3g59950 F24G16.220	Cysteine protease ATG4b (EC 3.4.22.-) (Autophagy-related protein 4 homolog b) (AtAPG4b) (Protein
29	P93298	ATP6-1 AtMg00410; At2g07741	ATP synthase subunit a-1 (F-ATPase protein 6) (P6-1)
30	P92549	ATPA ATP1 AtMg01190	ATP synthase subunit alpha, mitochondrial
31	Q9ZWJ9	ARR2 ARP5 At4g16110 dl4095w FCAALL.297	Two-component response regulator ARR2 (Receiver-like protein 5)
32	Q9FFI2	ATG5 APG5 At5g17290 MKP11_14	Autophagy protein 5 (Protein autophagy 5) (AtAPG5)
33	Q8RUS5	ATG9 APG9 At2g31260	Autophagy-related protein 9 (AtAPG9)
34	Q9SYT0	ANN1 ANNAT1 ANX23-ATH ATOXY5 OXY5 At1g35720 F14D7.2	Annexin D1 (AnnAt1) (Annexin A1)

• • • • •

3980	O60126	top3 SPBC16G5.12c	DNA topoisomerase 3 (EC 5.99.1.2) (DNA topoisomerase III)
3981	Q8TFG7	trm5 SPAPB18E9.01	tRNA (guanine(37)-N1)-methyltransferase (EC 2.1.1.228) (M1G-methyltransferase) (tRNA [GM37] r
3982	P40998	thi2 nmt2 SPBC26H8.01	Thiamine thiazole synthase (Thiazole biosynthetic enzyme)
3983	Q9P544	SPAC1635.01	Probable mitochondrial outer membrane protein porin
3984	O13915	usb1 SPAC23C11.10	U6 snRNA phosphodiesterase (EC 3.1.4.-)
3985	Q09154	rip1 SPBC16H5.06	Cytochrome b-c1 complex subunit Rieske, mitochondrial (EC 1.10.2.2) (Complex III subunit 5) (Ries
3986	O94536	ucp12 SPCC895.09c	Putative ATP-dependent RNA helicase ucp12 (EC 3.6.4.13)
3987	Q9UTF8	ugo1 SPAC1B2.02c	Mitochondrial fusion and transport protein ugo1
3988	Q9UT07	SPAP8A3.10	Protein ups1 homolog
3989	Q10988	uve1 uvde SPBC19C7.09c	UV-damage endonuclease (UVDE) (EC 3.-.-.-)
3990	O74834	ung1 SPCC1183.06	Uracil-DNA glycosylase (UDG) (EC 3.2.2.27)
3991	P32747	ura3 SPAC57A10.12c	Dihydroorotate dehydrogenase (quinone), mitochondrial (DHOD) (DHODase) (DHOdehase) (EC 1.3.
3992	O74560	raf2 clr7 cmc2 dos2 SPCC970.07c	Rik1-associated factor 2 (Cryptic loci regulator 7) (De-localization of swi6 protein 2)
3993	Q10258	met9 SPAC56F8.10	Methylenetetrahydrofolate reductase 1 (EC 1.5.1.20)
3994	Q10308	mug43 SPAC6C3.05	Meiotically up-regulated gene 43 protein
3995	G2TRP8	new18 SPBC30D10.21	Mitochondrial zinc maintenance protein 1, mitochondrial
3996	O94462	puf3 SPAC1687.22c	mRNA-binding protein puf3 (Pumilio homology domain family member 3)
3997	O74806	pth1 SPBC2D10.15c	Probable peptidyl-tRNA hydrolase (PTH) (EC 3.1.1.29)
3998	O74766	SPBC24C6.04	Probable delta-1-pyrroline-5-carboxylate dehydrogenase (P5C dehydrogenase) (EC 1.2.1.88) (L-gluta
3999	P32586	pyp2 SPAC19D5.01	Tyrosine-protein phosphatase 2 (EC 3.1.3.48) (Protein-tyrosine phosphatase 2) (PTPase 2)
4000	O42932	qcr6 SPBC16C6.08c	Cytochrome b-c1 complex subunit 6 (Complex III subunit 6) (Mitochondrial hinge protein) (Ubiquinc
4001	O74433	qcr9 SPCC1682.01	Cytochrome b-c1 complex subunit 9 (Complex III subunit 9) (Cytochrome c1 non-heme 7.3 kDa prote
4002	Q9UTB5	psd2 SPAC25B8.03	Phosphatidylserine decarboxylase proenzyme 2, mitochondrial (EC 4.1.1.65) [Cleaved into: Phosphati
4003	O94526	ptn1 SPBC609.02	Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase ptn1 (EC 3.1.3.67)
4004	P78761	qcr2 SPCC613.10	Cytochrome b-c1 complex subunit 2, mitochondrial (Complex III subunit 2) (Core protein II) (Ubiqui

伍、 討論與運用

一、透過 Uniprot 下載蛋白質資料後，此研究原始包含 4004 序列，進行 BLAST 和分群後，由於 MSC 演算法的關係，訂出截止臨界值，將比截止值大的節點數據刪除，因此所研究蛋白質對象剩下 2186 序列。刪除偏差的距離後，便開始分析蛋白質相關網路分布圖。

二、利用粒線體蛋白質資料視覺化後，發覺到內膜與基質的關聯度相對較高，而我們的結果也恰好支持內共生假說，由於粒線體自帶遺傳物質，所以在演化上，內膜和基質的演化程度相對較小，進而圖形連結關係較高；反之外膜和膜間隙則會因生物體的不同而有不同種的外膜和膜間隙，所以外膜和膜間隙的演化程度較大，進而相關性較小。

三、我們再以 EC 編號來進行分群，將酶大致分為六大類，在這些 EC 編號網路圖中，我們發現屬於 EC1，EC2，EC3，EC6 的酶有較多高連結群的節點，演化程度較少；而 EC4，EC5 的酶具有低連結度的節點群，演化程度較高。

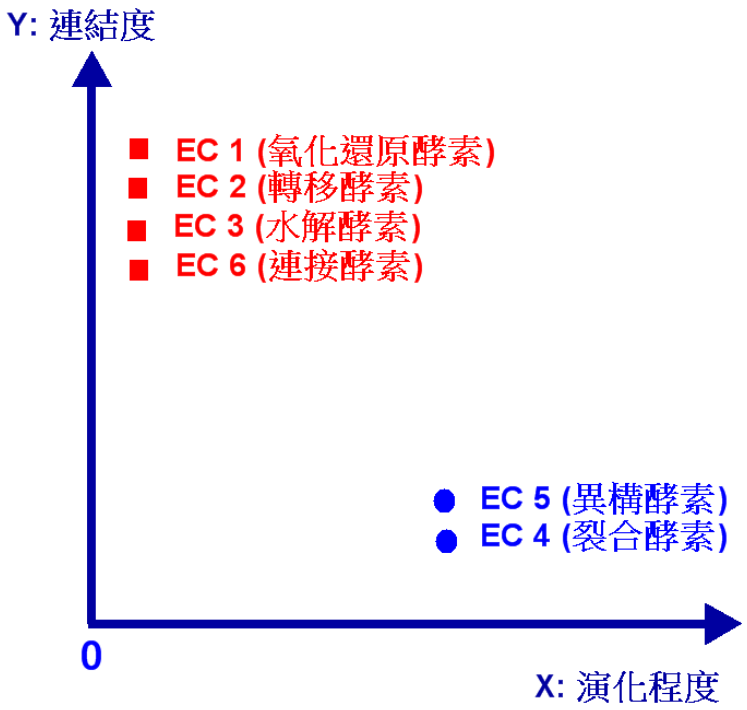


圖 21: 粒線體酵素，根據 EC 編號來進行蛋白網路部分關聯分析

本研究的分群視覺化處理，對於資料選用和分類應效果不錯，但是這次資料選用存在的大腸桿菌、線蟲和烘培酵母菌，阿拉伯芥四種樣本數目，有可能因資料不全造成系統性分析誤差，未來找尋粒線體中的蛋白質演化應需考慮更多的樣本數目以尋求完整性與全面性，生物資訊分析原本是一件耗時又費工的任務，透過本次研究後，也發覺到這個方法是一個有效率而且比實驗快速的方法之一，特別在蛋白質結構方面比對，這個運用在未來生物科技上蛋白質相似度的比對，或者是蛋白質 motif 結構的探討是具有相當好的前瞻性。

陸、 參考資料及其他

[1] Geng-Ming Hu, Te-Lun Mai & Chi-Ming Chen, Visualizing the GPCR Network: Classification and Evolution, Scientific Reports 7, 15495 (2017).

[2] 大腸桿菌(Escherichia coli) https://www.biocote.com/blog/five-facts-e-coli/

[3] 釀酒酵母 (Saccharomyces erevisiae) https://zh.wikipedia.org/wiki/%E9%85%B5%E6%AF%8D#/media/File:S_cerevisiae_under_DIC_microscopy.jpg

[4] 秀麗隱桿線蟲（Caenorhabditis elegans） http://post.queensu.ca/~chinsang/research/c-elegans.html

[5] 阿拉伯芥 (Arabidopsis thaliana) https://www.eurekalert.org/multimedia/pub/159783.php

[6] 粒線體（mitochondrion）https://biologywise.com/mitochondrial-function

[7] UniProt 蛋白質資料庫網站 http://www.uniprot.org/

[8] Blast 程式網站 https://blast.ncbi.nlm.nih.gov/Blast.cgi

[9]Cytoscape 程式網站 http://www.cytoscape.org/index.html

[10] Matlab 程式網站 https://ww2.mathworks.cn/products/matlab.html?requestedDomain=zh

[11] EC number: https://zh.wikipedia.org/wiki/EC編號

[12] Lurin C, Andrés C, Aubourg S, Bellaoui M, Bitton F, Bruyère C, Caboche M, Debast C, Gualberto J, Hoffmann B, Lecharny A, Le Ret M, Martin-Magniette ML, Mireau H, Peeters N, Renou JP, Szurek B, Taconnat L, Small I, "Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis". Plant Cell. 16 (8): 2089–103 (2004).

附錄

1. MSC (Minimum Spanning Clustering) 演算法程式

```
function [ taxo_cell_sort, sorted_min_array_2] = MSC_clustering( Dism, level,threshold_value )


%  ======= Usage ========
%[ taxo_cell_sort, sorted_min_array_3,core_link ] = MSC_clustering( Dism, level,threshold_value )
% @Input parameters
% *Dism: The distance matrix
%       ex:   0  2  3  1
%             2  0  5  3
%             3  5  0  2
%             1  3  2  0
%  * level          : the level-th minimum distance are consideed here.
%  * threshold value : A  threshold value for sparification
% @ output parameter
%  * taxo_cell_sort     : MSC taxonomy table, each cell contains a group of members.
%  * sorted_min_array_3 : the pairwise table for clustering
%  * core_link         : the core link of each clusters
%  * RNM              : Renormalization matrix of core link
%  ======================


     index_th=1;
     if (nargin < 2)
          level=1;
     end
     if (nargin < 3)
          threshold_value=2*max(max(Dism));
          index_th=0;
     end
     clear taxo_cell
     N=max(size(Dism));
     for i=1:N;Dism(i,i)=inf;end;
     if level ==1
          sorted_min_array = find_min_of_matrix(Dism);
     else
          sorted_min_array = find_mins_of_matrix(Dism,level);
     end
     meta_Dis=zeros(length(Dism));
     if index_th ==1
```

```
        meta=size(sorted_min_array);
        index=1;
        for i=1:meta(1)
            if sorted_min_array(i,3) <threshold_value
                sorted_min_array_2(index,1)= sorted_min_array(i,1);
                sorted_min_array_2(index,2)= sorted_min_array(i,2);
                sorted_min_array_2(index,3)= sorted_min_array(i,3);
                sorted_min_array_3{index,1}= num2str(sorted_min_array(i,1));
                sorted_min_array_3{index,2}= num2str(sorted_min_array(i,2));
                sorted_min_array_3{index,3}= num2str(sorted_min_array(i,3));
                meta_Dis(sorted_min_array(i,1),sorted_min_array(i,2))=1;
                if sorted_min_array(i,3) ==0
                    meta_Dis(sorted_min_array(i,1),sorted_min_array(i,2))=1.5;
                end
                index=index+1;
            end
        end
    else
        sorted_min_array_2=sorted_min_array;
    end
    meta_Dis=meta_Dis+meta_Dis';
    for i=1:length(meta_Dis)
        meta_list(i)=0;
        if length(find(meta_Dis(:,i)==2)) ~=0
        meta_list(i)=1;
        end
        if length(find(meta_Dis(:,i)==3)) ~=0
        meta_list(i)=2;
        end
    end
    [taxo_cell,taxo_single_node]= pairclustering_N(sorted_min_array_2,N);
    clear ooxx
    for i=1:length(taxo_cell)
        ooxx(i)=length(taxo_cell{i});
    end
    [A,IXa]=sort(ooxx);
    for i=1:length(IXa)
        taxo_cell_sort{i}= taxo_cell{IXa(length(IXa) -i +1)};
    end
end
```

```
function sorted_min_array = find_min_of_matrix(dist_matrix)
     m=length(dist_matrix);
     min_array=[];
     for i=1:m
          n=min(dist_matrix(i,:));
          minl=find(dist_matrix(i,:)==n);
          l=length(minl);
          for j=1:l
               min_array=[min_array;i,minl(j),dist_matrix(i,minl(j))];
           end
     end
     mm=length(min_array);
     [a,index]=sort(min_array(:,3),'ascend');
     for k=1:mm
          sorted_min_array(k,:)=min_array(index(k),:);
     end
end
function sorted_min_array_out = find_mins_of_matrix(dist_matrix,level)
     sorted_min_array_out =[];
     max_v=max(max(dist_matrix));
     for ii=1:level
          m=length(dist_matrix);
          min_array=[];
          for i=1:m
               n=min(dist_matrix(i,:));
               minl=find(dist_matrix(i,:)==n);
               l=length(minl);
               for j=1:l
                    min_array=[min_array;i,minl(j),dist_matrix(i,minl(j))];
                    dist_matrix(i,minl(j)) = inf;
               end
          end
          mm=length(min_array);
          [a,index]=sort(min_array(:,3),'ascend');
          for k=1:mm
               sorted_min_array(k,:)=min_array(index(k),:);
          end
          sorted_min_array_out=[sorted_min_array_out',sorted_min_array']';
```

```
    end
end


function [ taxo_cell_out,taxo_single_node ] = pairclustering( tablem )
%UNTITLED3 Summary of this function goes here
%   Detailed explanation goes here
    taxo_single_node=[];
    N=max(max(tablem(:,1:2)));
    group_line=zeros(N,1);
    cell_num=0;
    for i=1:length(tablem)
        if group_line(tablem(i,1))~=0 & group_line(tablem(i,2))==0
            group_line(tablem(i,2))= group_line(tablem(i,1));
        elseif group_line(tablem(i,1))==0 & group_line(tablem(i,2))~=0
            group_line(tablem(i,1))= group_line(tablem(i,2));
        elseif group_line(tablem(i,1))==0 & group_line(tablem(i,2))==0
            cell_num = cell_num+1;
            group_line(tablem(i,1))= cell_num;
            group_line(tablem(i,2))= cell_num;
        elseif group_line(tablem(i,1))~=group_line(tablem(i,2))
            meta=find(group_line == max(group_line(tablem(i,1)),group_line(tablem(i,2))) );
            group_line(meta)= min(group_line(tablem(i,1)),group_line(tablem(i,2)));
        end
    end
    GN=unique(group_line);
    if GN(1)==0
        taxo_single_node= find(group_line==0);
        for i=2:length(GN)
            taxo_cell_out{i-1}=find(group_line==GN(i));
         end
    else
        for i=1:length(GN)
            taxo_cell_out{i}=find(group_line==GN(i));
        end
    end
end
```

```
function [ taxo_cell_out,taxo_single_node ] = pairclustering_N( tablem,N )
%UNTITLED3 Summary of this function goes here
%   Detailed explanation goes here
     taxo_single_node=[];
     group_line=zeros(N,1);
     cell_num=0;
     for i=1:length(tablem)
          if group_line(tablem(i,1))~=0 & group_line(tablem(i,2))==0
               group_line(tablem(i,2))= group_line(tablem(i,1));
          elseif group_line(tablem(i,1))==0 & group_line(tablem(i,2))~=0
               group_line(tablem(i,1))= group_line(tablem(i,2));
          elseif group_line(tablem(i,1))==0 & group_line(tablem(i,2))==0
               cell_num = cell_num+1;
               group_line(tablem(i,1))= cell_num;
               group_line(tablem(i,2))= cell_num;
          elseif group_line(tablem(i,1))~=group_line(tablem(i,2))
               meta=find(group_line == max(group_line(tablem(i,1)),group_line(tablem(i,2))) );
               group_line(meta)= min(group_line(tablem(i,1)),group_line(tablem(i,2)));
           end
     end
     GN=unique(group_line);
     if GN(1)==0
          taxo_single_node= find(group_line==0);
          for i=2:length(GN)
               taxo_cell_out{i-1}=find(group_line==GN(i));
          end
     else
          for i=1:length(GN)
               taxo_cell_out{i}=find(group_line==GN(i));
          end
     end
end
```