第九小組 DBcheckpoint2 報告

組員: A1063329 王新賦、A1063315 許雅涵、A1063306 呂承恩

指導教授:楊子賢教授

一、安裝環境

安裝 appserv(下載網址 https://www.appserv.org/en/),根據網站上的步驟進行安裝,讓我們可以用 phpmyadmin。

二、如何切割表格

(1) 主表格的檔案轉檔

我們採用的是用 phpMyAdmin 去建立資料庫,首先先將 all-mrna.sql 及 sdgGene.sql 轉成 excel 的 xlsx 檔,接著將 all-mrna.xlsx 及 sdgGene.xlsx 分別轉檔為 all-mrna.csv 及 sdgGene.csv,但必須將 xlsx 檔裡的第一行欄位名稱刪掉後,再儲存為 csv 檔。其中 Saccharomyces_cerevisiae 的資料因為沒有 sql 檔,所以須先手動新增所需要的欄位再轉成 csv 檔。

d	28	2		D	10	(F)	0	(#72)	00 1			W	H D		0.	(4))	1	1	1.0	Y	W	- 0	400	
18.3	586	Itte	3		R	P.		- 1	16 -	: CQC11E	1796		1704/64	Ditte	17401	3000		HTB.	10.	DWT.				
2	595	127	- 3	- 10	000	- 10	- 1	- 1	10-	\$300mm	121	- 4	Utfold	2.90218	16074	19000		125	10,	HOPOL.				
4	546	125	- 0	.00	10	- 10	- 0	- 0	00-	8,03047	125	- 0	135 club	236219	160776	180000		125	0,	HODB.				
	596	127	- 0	111	16	10			16-	\$.000H	123		137 det	29019	TRENGT	155900		127,	16	145962.				
3.	516	125	- 0	- 10	0.	. 19		- 0	10-	£00042	125		125-del	Z1027W	100001	115999		125	0,	165864				
6	560	11		.01	00	.10		- 1	10.4	KR0071	26		Think!	1510003	568863	190055		12.	10.	590000				
. 2	589	12	- 0	- 0	0.0	- 10		- 4	H+	M23398	- 73		TZ chrfV	1511003	50063	46000		71.	II.	Money).				
	565	4185	- 6	- 13	0.0	10		- 1	0.5	MEDIN	4993	- 1	4155 jed1	817184	14065	18259		APU.	0.	1065				
. 2	50	219	- 0.	0.0	00	- 30	- 1	- 1	3037+	37121121	259		29 chf1	913194	45612	45334		230227	0.10	10012,49	71.			
11	566	347	- 3	- 11	- 0	11	+	- 1	384+	-007596	.141		747-Jell	913104	18641.4	9545		THEFT.	31,19.	38841438	49636			
11	596	2919		- 0	11	- 1		- 1	0	3/20130	3111		2505 chd1	512184	177192	.01740		Z DIOMERO	0.35	177792.17	78114			
12	546	2456		11.	00.	- 16	- 1	- 1	10	M01006	7400		(146,341)	4131144	160900	165974		134,216	1,0,795.	110802,10	ST41.			
33	587	180	- 0	100	100	1.0	- 10	- 0	III to	(1)-69(0)	1,837	0.	D80(468)	012104	APRIL	(888)		DML	33.	1000				
10	567	1819	- 1	-00	0.0	11		- 1	00-	EDITE	112	- 1	100 cutt	813184	BRHILL	11000		1.000	0.	300				
19.	598	Th.	- 9	- 00	- 6	10		- 1	500 -	D17111	106	- 1	700 (hd)	103014	1149(1)	615214		2400,15	20,691.	1201062.0	5254			
3.6	200	127	- 3		00	- 10	- 1	- 1	307;	1071,0125	128	- 8	215 (44)	813184	1000017	Address		£100,28,	0.190,	428117,40	1790			
17	386	111	. 7	100	0.	+		- 2	2	(8424)78	165	- 4	265 dell	¥13184	400126	001466		9.406,79.2	2,004,000,5	990000000	1218-0113	05,40132	864315930	01361/01
.18	59	- 30	0	100	16	- 6	- 4	- 2	191+	4727944	-201		201 cht1	313384	462136	(6)2575		0.7121403	00,11,21	. 4019.8	2395/824	N)		
.10	595	1700	- 5	. 11	0.	P		- 0	00	309014	1719		1706-Juli	\$17104	10/8772	#90179		1700	11%	-09777				
28	598	1410.	- 0	- 0	0.	.00		- 0.	-00-	APPKKE	1907		18World	813104	488911	4903502		1481	150	agents,				
25	5/6	10%	- 12	- 0	11	11		- 4	16-	AFRICK	1479		(divided)	813194	-pat 15	(85594)		147%	30	d94715.				
22	349	565	- 18	- 0	0.0	- 11		- 1	16-	(1000)2	9.60	- 6	585 dell)	812184	367770	569439		665	1.	3677%				
.23	500	3218	- 9		0.0	. 0		- 2	0	\$10000	. 其後		321V chil1	313104	217/78	535000		1.2226	22.	30370%				
24	500	475	- 4	. 11	1	1.1			· · · III »	M30942	461	- 1	402 dell	83,3184	59(217)	393252		2377,294	11,238.	392171.39	7918.			
23	295	216	. 1	- 00	11	10		1.0	261	31F123540	296		259,041	813184	921313	60,960		2.01200	0.43	16033110.00	SALL			
26	790	96	- 5	. 10	0.0	- 10		- 3	III-	1/37900	900	. 2	900,000	6133184	763672	79672		706	.0,	789672.				
29	931	2316	- 1	0.	U.	- 16		- 6	0.5	300203	-2574		135.441	613194	769103	701696		2196	A	THOUSE.				

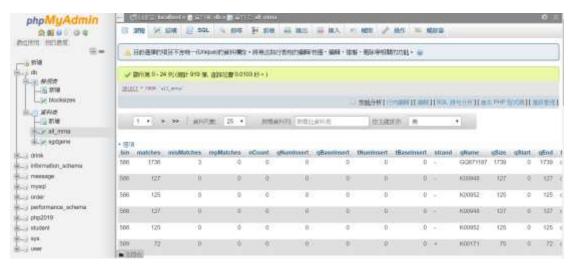
(2)主表格檔案的匯入及欄位的建立

接著在 phpMyAdmin 用匯入功能,先匯入 all-mrna.sql、sdgGene.sql、Saccharomyces cerevisiae.sql,以便先將欄位建立起來。



(3)資料的匯入

將 all-mrna.csv、sdgGene.csv 檔及 Saccharomyces_cerevisiae.csv 檔分別匯入已經建立好的 all-mrna 及 sdgGene 及 Saccharomyces_cerevisiae 的資料表,即會將檔案內的資料值匯入。



(4)表格的切割

依據我們所畫的 ER 圖,並且將他轉為 Relational Model 後,可以得知要建立mRna、BlockSizes、qStarts、Target、Matches、tStarts、Chromosome、GENE、exonEnds、exonStarts、Protein 這些 table,接著依據我們所觀察到的關係去下SQL 指令將 table 連結。

SQL 的程式碼如下:

(a) 建立 view mRna 時也只需要在 all mrna 中查找出來我們需要的即可。

CREATE VIEW mRNA AS SELECT qname,qend,qbaseinsert,qnuminsert,bin,qsize,qstart,blockCount,mismatches,ncount,repMatches,matches FROM all_mrna;

(b) 建立 view blockSizes 時只需要在 all_mrna 中查找出來我們需要的即可。

CREATE VIEW blockSizes AS SELECT blocksizes, qname FROM all_mrna;

- (c) 建立 view qStarts 時也只需要在 all_mrna 中查找出來我們需要的即可。

 CREATE view qstarts AS SELECT qname,qstarts from all_mrna;
- (d) 建立 view target 時也只需要在 all_mrna 中查找出來我們需要的即可。

 CREATE VIEW target AS SELECT tName,tSize,tStart,tEnd,tstarts,tBaseInsert,tNumInsert FROM all_mrna;
- (e) 建立 view matches 時,由於我們需要由 matches 查找特定的 mRna 和 target,所以將 all_mrna 中的 qName、tStart 和 tEnd 查找出來。

 CREATE VIEW matches AS SELECT qname, tstart, tend FROM all mrna;
- (f) 建立 view tStarts 時也只需要在 all_mrna 中查找出來我們需要的即可。

 CREATE view tstarts AS SELECT tstart, tend, tstarts from all_mrna;
- (g) 建立 view chromosome 時需要在 sgdgene 中查找出 chrom,由於要加入 tName 所以要加入 chrom=tname,而 tname 和 chrom 多是重複的值,所以 在前面加入 distinct。

 CREATE VIEW chromosome AS SELECT DISTINCT chrom,tName FROM target,sgdgene WHERE chrom=tName;
- (h) GENE

CREATE view gene AS SELECT s.bin,name,chrom,s.strand,cdsStart,cdsEnd,exonCount,proteinid,qname FROM sgdgene s ,all_mrna a WHERE s.bin=a.bin and s.strand = a.strand;

- (i) 建立 view exonEnds 時也只需要在 sgdgene 中查找出來我們需要的即可。

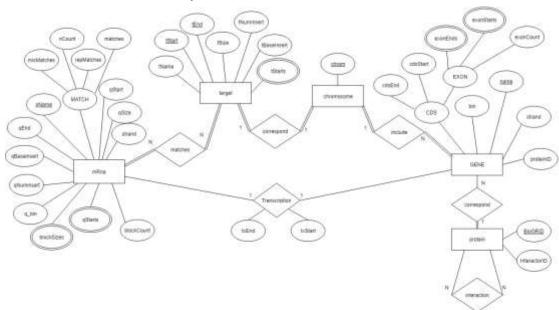
 CREATE view exonends as SELECT name, exonends FROM sgdgene;
- (j) 建立 view exonStarts 時也只需要在 sgdgene 中查找出來我們需要的即可。

 CREATE view exonstarts As SELECT name, exonstarts FROM sgdgene;
- (k) 建立 view protein 時只要把新增的資料表 BioGRID 中的我們所需要的查找 出來即可。

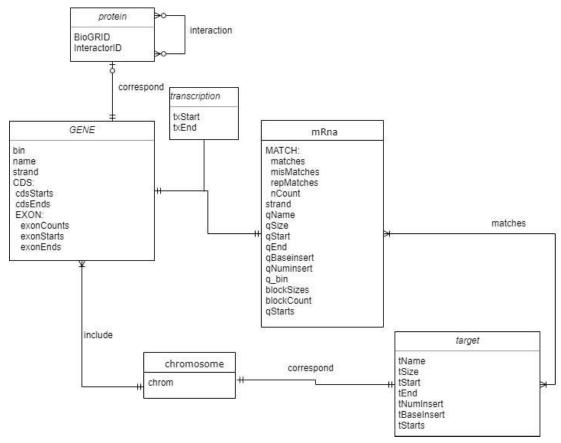
CREATE view protein as SELECT biogrid, InteractorA, InteractorB FROM biogrid;

四、修改後的圖

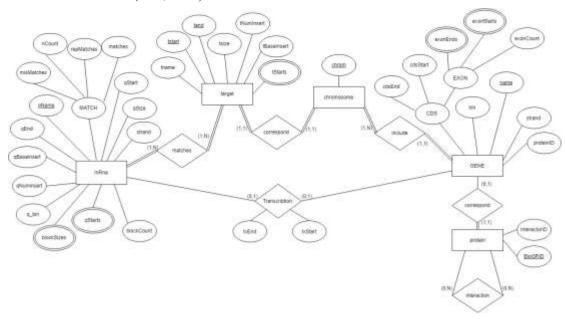
1. Chen's notation w/ cardinality notation



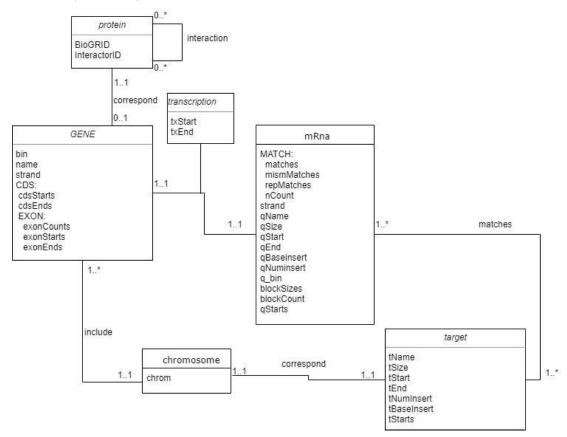
2. UML w/ Crow's notation



3. Chen's notation w/ (min, max) notation



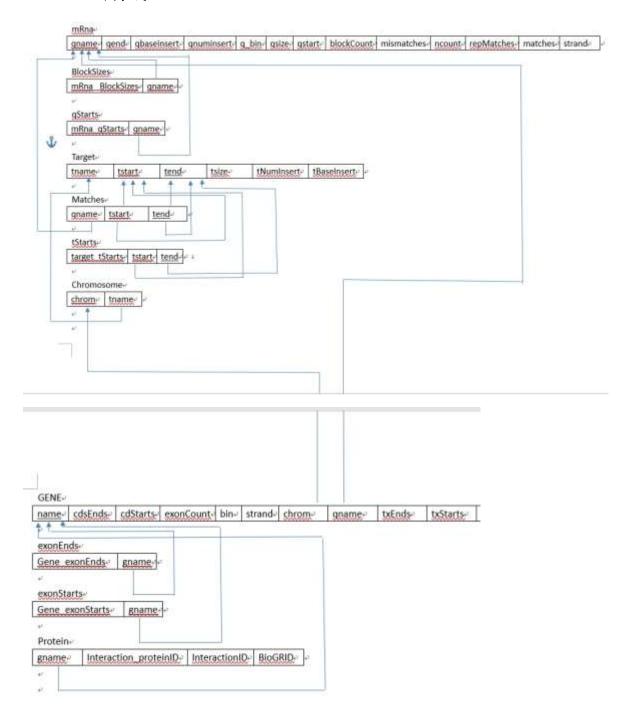
4. UML w/ (min, max) notation



說明:

- 1. 基因裡面包含 name(DNA 的名字)、strand(股)、bin、CDS(編碼區)、proteinID。而 CDS 包含 cdsStarts、cdsEnds、exon,分別為編碼區的開始、結束位置、外顯子。而外顯子又包含 exonCount、exonStarts、exonEnds,分別為外顯子的數量、外顯子開始和結束位置,而編碼區是由外顯子所組成,用來編碼蛋白質的部分。外顯子是基因的一部份,在經過轉錄後仍會被保留下來。
- 2. chromosome 的欄位包含了染色體的名稱 chrom,而基因是被包含在 chrom之中,可以透過查找特定染色體,找到其所包含的基因。
- 3. 基因轉錄成 mRNA,並有記錄轉錄的 txStart(開始)和 txEnd(結束)位置。 mRNA 欄位包含了 bin、qName、qSize、qStart、qEnd、match、qStarts、strand、qBaseInsert、qNumInser、blockCount、blockSizes,分別記錄其查詢名字、大小、查詢的開始位置及結束位置。
- 4. match 包含了 matches、misMatches、repMatches、nCount,可以用來尋找 特定的 Target 值(亦可使用 mRNA 中的其他屬性尋找)。Target 會顯示找到 的對應染色體,而 Target 欄位包含了 tname、tstart、tend、tsize、tstarts、tNumInsert、tBaseInsert,分別記錄其查詢結果所對應到的名字、大小、結果的開始位置及結束位置。
- 5. protein 的欄位包含了 BioGRID(實驗的編號)、InteractorID,而 InteractorID 可以對應到基因中的 name,而在蛋白質交互作用之中 InteractorID 會和另一個 InteractorID 進行實驗。

三、ER 圖轉成 relational model



說明:

每個 entity 都獨立出來當作一個 table, 而其裡面的屬性就是對應到的 table 欄位,若是屬性為複數,就也將其獨立出來當做一個 table。

- 1. mRna 裡有 qname、qend、qbaseinsert、qnuminsert、q_bin、qsize、qstart、blockCount、mismatches、ncount、repMatches、matches、strand 這些欄 位。
- 2. BlockSizes 因為是 mRna 裡的複數屬性,因此需要獨立出來,而裡面有本身的 mRna BlockSizes 欄位,也因為他屬於 mRna,因此要多一欄 mRna 的主鍵(qname),指回去 mRna 的 qname。
- 3. qStarts 同 BlockSizes 因為是 mRna 裡的複數屬性,因此需要獨立出來,而裡面有本身的 mRna_qStarts 欄位,也因為他屬於 mRna,因此要多一欄mRna 的主鍵(qname),指回去 mRna 的 qname。
- 4. target 之中有原本的 tName、tSize、tStart、tEnd、tBaseInsert、tNuminsert 還有用以對應 chromosome 而加入的 chrom。由於 tStarts 是多值,所以將其額外建立一個表格,包含了 tStarts 和 target 中的主鍵 tStart、tEnd。
- 5. 由於 match 是一對多的 relationship,可用以找到特定的 mRna 和 target,所以將 matches 的 relationship 加入 mRna 的主鍵 qName 和 target 的主鍵 tStart 和 tEnd。
- 6. chromosome 由於包含了基因且也可用 target 對應,所以只將 chromosome 的主鍵加入到 target 和基因之中,而 chromosome 只留下 chrom。
- 7. GENE 裡面有 name、cdsEnds、cdStarts、exonCount、bin、strand、chrom、qname 、txEnds 、txStarts、proteinID 這些欄位。其中因為 Gene 與chromosome 為多對一的關係,若是為多對一的關係,就必須在多的那方新增對方的主鍵,因此在 Gene 裡再新增 chromosome 的主鍵(chrom),指回去chromosome 的 chrom。而因為 Gene 與 mRna 為一對一關係,若是為一對一的關係,就在兩個 table 中擇一加入對方的主鍵,而我們是選擇在 Gene裡新增 mRna 的主鍵(qname),指回去 mRna 的 qname,也因為兩者的ralaitonship 有額外的屬性,這些屬性也是可以擇一 table 加入,而我們也是選擇在 Gene 裡新增這些屬性(txEnds 、txStarts)。
- 8. exonStarts 因為是 Gene 裡的複數屬性,因此需要獨立出來,而裡面有本身的 Gene_exonStarts 欄位,也因為他屬於 Gene,因此要多一欄 Gene 的主鍵 (gname),指回去 Gene 的 gname。
- 9. exonEnds 同 exonStarts 因為是 Gene 裡的複數屬性,因此需要獨立出來,而裡面有本身的 Gene exonEnds 欄位,也因為他屬於 Gene,因此要多一欄 Gene 的主鍵(gname),指回去 Gene 的 gname。
- 10. protein 中有 BioGRID 和 InteractorID,因為 InteractorID 會進行交互作用,所以加入了進行交互作用的兩個 InteractorID(IntercatorA、InteractorB),而 protein 要可以對應到基因,所以加入了基因的主鍵 gname 使其可以找到基因。

✓ 王新賦

	評分	理由
王新賦	5	認真討論和和重畫圖
呂承恩	5	分割表格,和畫圖
許雅涵	5	畫圖、切表格、統整 word

✓ 呂承恩

	評分	理由
王新賦	5	畫圖、切表格
呂承恩	5	切表格、用 phpmyadmin 建 表格
許雅涵	5	解讀資料表間的關係、 畫圖

✓ 許雅涵

	評分	理由
王新賦	5	匯入檔案成 excel、畫圖和討論
呂承恩	5	建立沒有 sql 檔的檔案、討 論和切表格
許雅涵	5	Word 的統整、討論如何重 畫圖