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

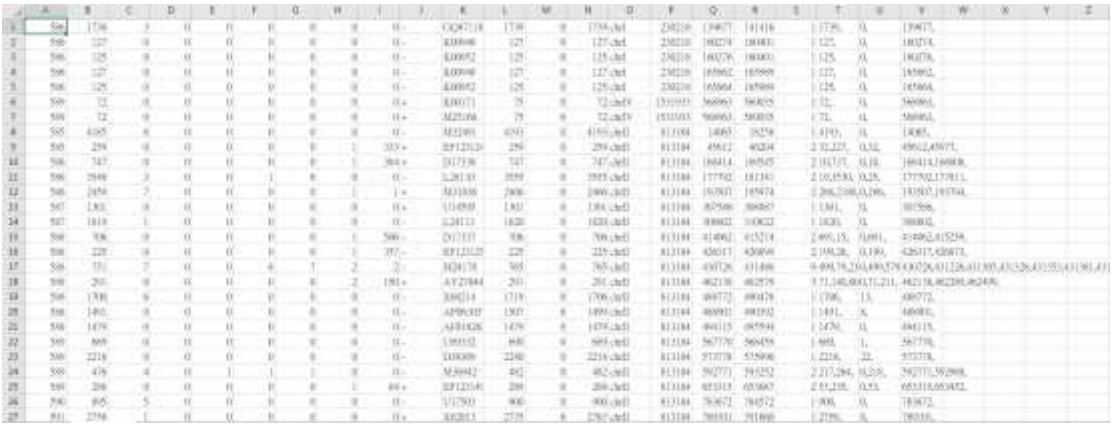
一、安裝環境

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。

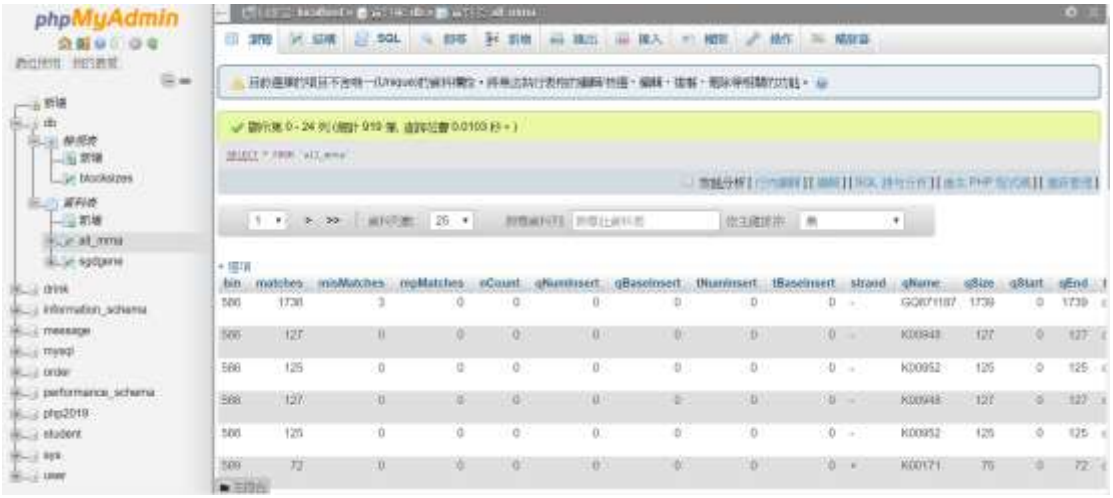


(2)主表格檔案的匯入及欄位的建立

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。

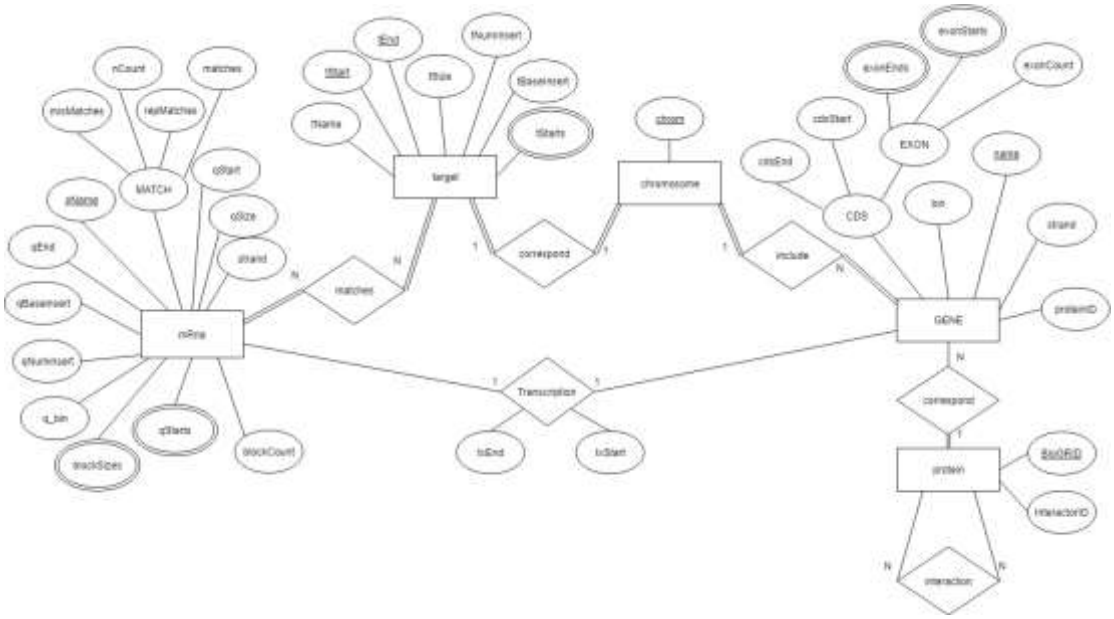
```
CREATE view exonstarts As SELECT name,exonstarts FROM sgdgene;
```

(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。

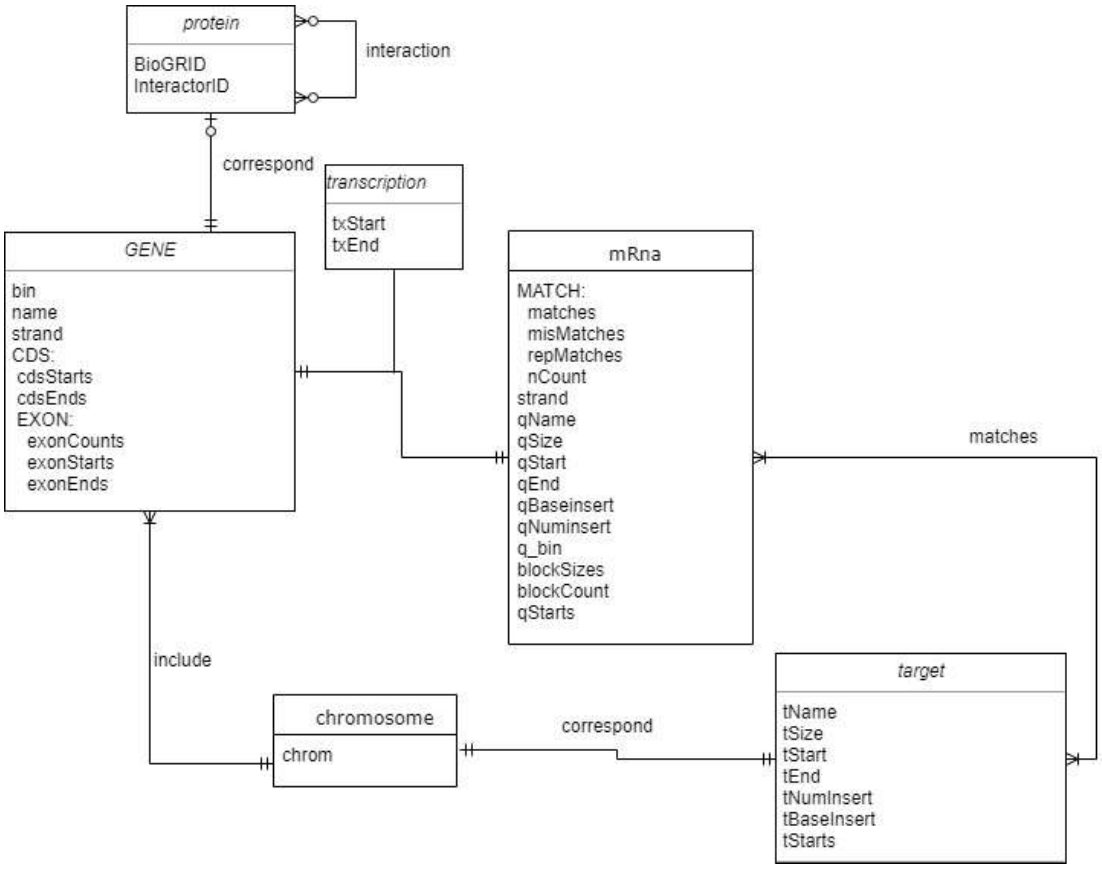
```
CREATE view protein as SELECT biogrid,InteractorA,InteractorB FROM biogrid;
```

四、修改後的圖

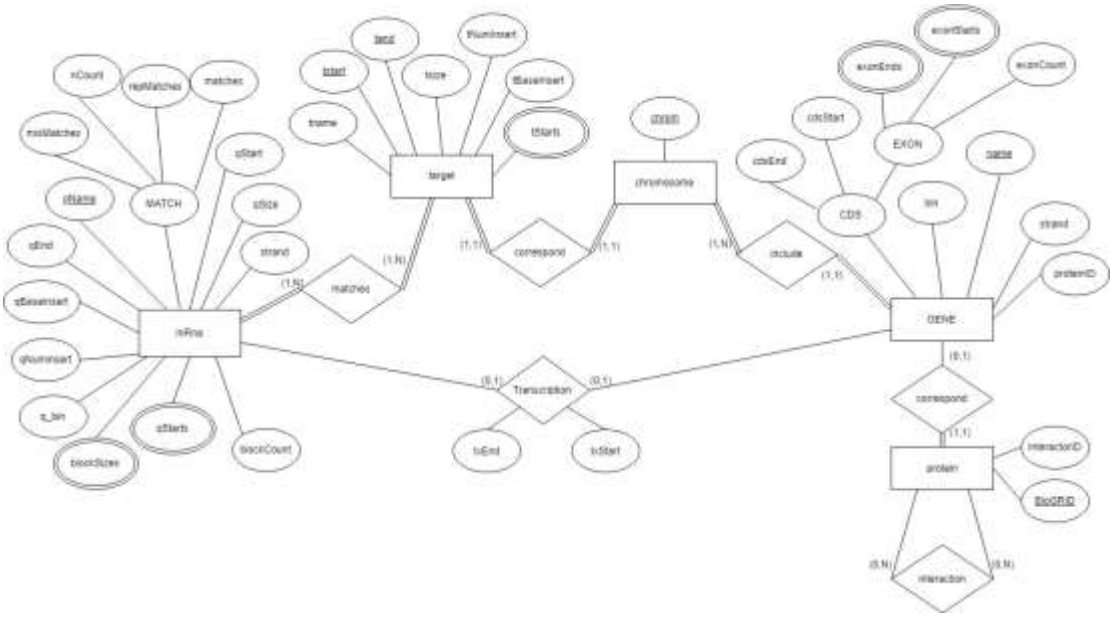
1. Chen’s notation w/ cardinality notation



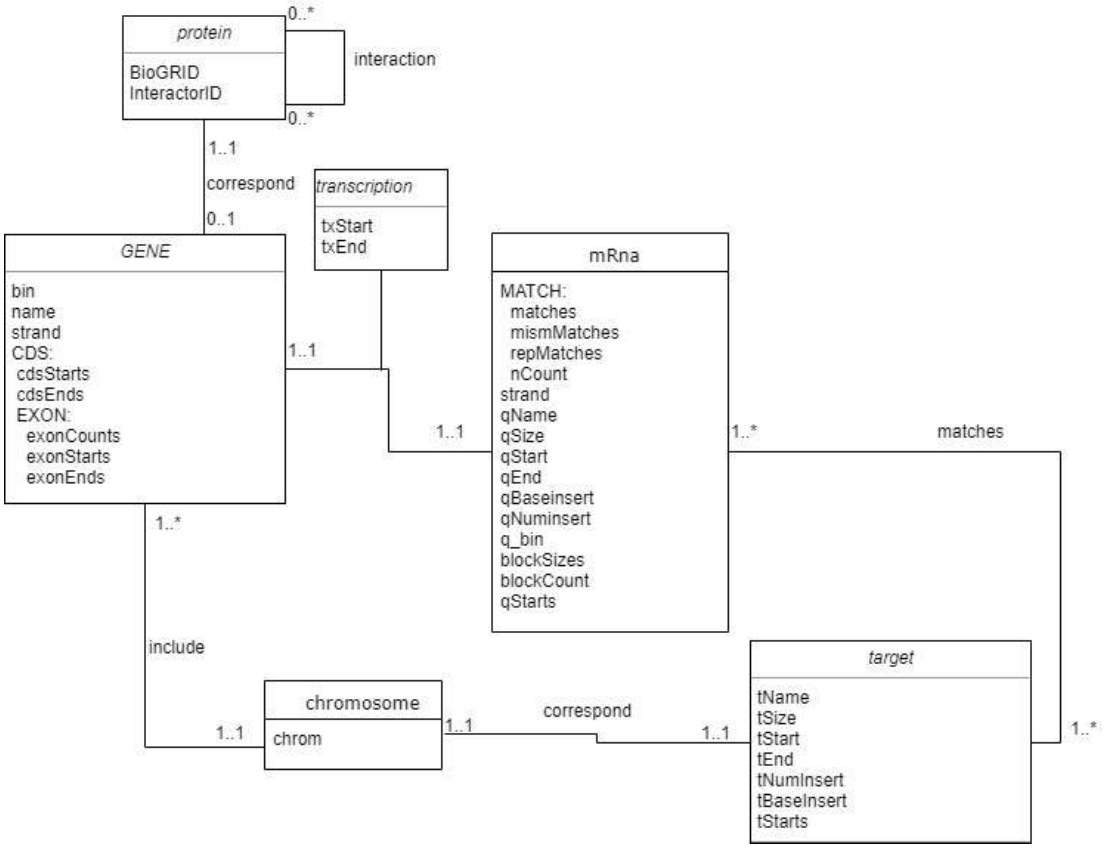
2. UML w/ Crow’s notation



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



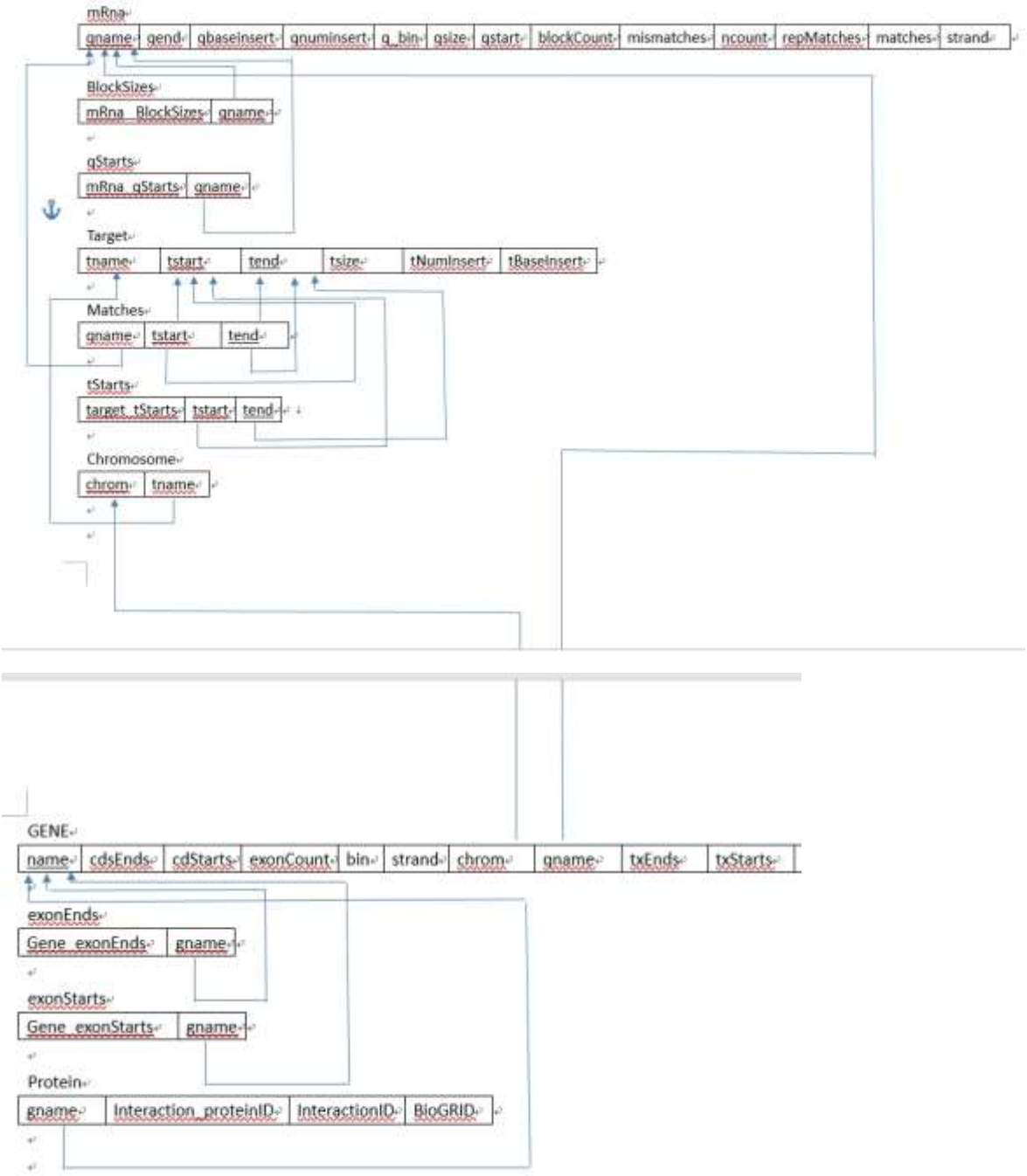
4. UML w/ (min, max) notation



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。

四、互評表

✓ 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