

なるほど便利!! 発現解析データ抽出に こんなソフトはいかがですか? ~ Microsoft® Access 編 ~

データ整理やデータ解析、メール、インターネット検索など、コンピュータは研究生活に無くてはならないものとなりました。ですが、データはますます多岐にわたり、大量のデータが簡単に発生するようになり、途方にくれることもしばしばです。本連載では実際の研究でお役に立ちそうなソフトウェアや使用上のティップスをご紹介します。サンプルデータや資料をホームページ(<http://www.takara-bio.co.jp/DL/>)からダウンロードして、是非実際に操作しながらお読みください。

初回は、パーソナルでも研究室レベルでも、データの整理や解析に手軽に使えるデータベース管理ソフト Microsoft® Access についてご紹介します。

とある大学の研究室。この研究室では最近、マイクロアレイの実験を始めました。

最近ようやく Excel が使えるようになってきた新人のトヤマ君のところに、ミネヤマ教授がやってきました。

教授：トヤマ君。このマイクロアレイの発現比のデータ (Ratio.xls) と遺伝子情報のデータ (Gene.xls) を合わせて一つのテーブルにしておいて。至急お願いするよ。

Excel ファイルを 2 つおいて立ち去りました(図1)¹。早速開いてみると、どうやら、Gene_ID の項目が共通しているようです。

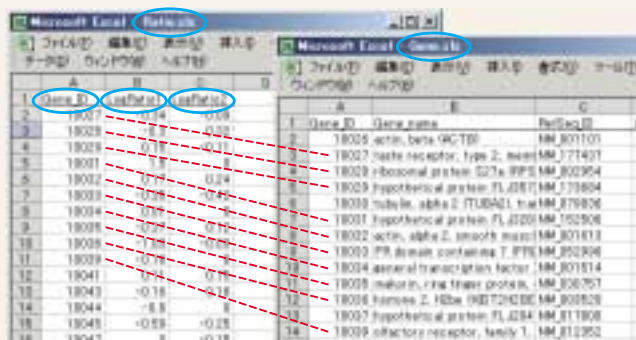


図1 データファイル(マイクロアレイの発現比、遺伝子情報)

ちょっと Excel が使えるようになったトヤマ君。楽勝、楽勝と思いましたが...

トヤマ君：Gene_ID の数が一致しない。数字もとびとびだし。まとめてコピー & ペースト²できないな。

インターネットで調べて何とか VLOOKUP 関数³を使えばできそうだということにたどりつき⁴、早速 Excel データを加工し始めましたが...

トヤマ君：なんだか計算が遅いな...

そこへ、再び教授が現れました。

教授：トヤマ君。検証用に作ったリアルタイム RT-PCR 用のプライマーのリスト (Primer.xls)⁵ があるんだけど(図2)、これも一緒にまとめておいてくれ。それと、マイクロアレイのデータの LogRatio1 の項目の値が 2 以上か LogRatio2 の項目の値が 2.5 以上のどちらかを満たして、なおかつプライマーリストにない遺伝子のデータだけ抜き出して別のテーブルを作っておいてくれ。悪いけれど急いでいるんで 30 分後に取りに来るよ。後でアクセスン(RefSeq ID)でリストを作って、タカラバイオにリアルタイム RT-PCR 用のプライマー設計と合成も依頼しておいて。これからもいろいろなデータをまとめないといけないからそのつもりでね。

立ち去る教授。頭を抱えるトヤマ君。



図2 プライマーリスト

トヤマ君：これ以上はこの Excel ファイルじゃ重すぎて動かないよ。それに、Excel のオートフィルター機能は同じ列では AND(かつ)や OR(または)の関係でフィルタできるけど、列の間は OR の関係ではフィルタできないし⁶。

大ピンチのトヤマ君です。そこへいつもコンピュータのことやデータ解析のことで助けてもらっているサエ先輩が通りかかりました。

トヤマ君：先輩、助けてください。かくかくしかじか…。あと20分しかないんです。

サエ先輩：確かに Excel ではしんどいわね。これからもデータ追加しないといけないでしょ。そういうときにはリレーショナルデータベースソフトを使うのよ。

トヤマ君：??

サエ先輩は近くの Windows PC を起動させて言いました。

サエ先輩：Microsoft® Office が入っているわよね。その中に Access っていうソフトがあるでしょ*7。これを使えば簡単よ。トヤマ君の PC の Office にも入っているんじゃないかな*8。

トヤマ君：あれ？先輩は Mac を使ってませんでしたか？

サエ先輩：そうだけど、Access には Mac 版が無いのよ。

トヤマ君：あと15分しかないですよ。

サエ先輩：それじゃ、リレーショナルデータベースが何かの説明は別の時にするとして、早速 Access を使ってみましょう。

まず、3つの Excel ファイルをインポートしてそれぞれのテーブルを作成するの(図3)。

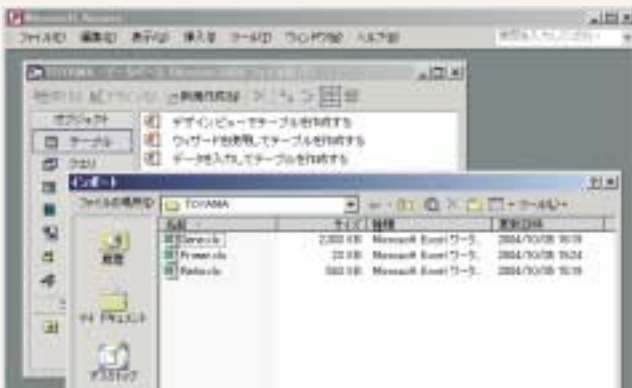


図3 データベースウィンドウ(インポート)

次にクエリデザインビューを開いて、3つのテーブル[Ratio]と[Gene]と[Primer]を表示させるの(図4)。

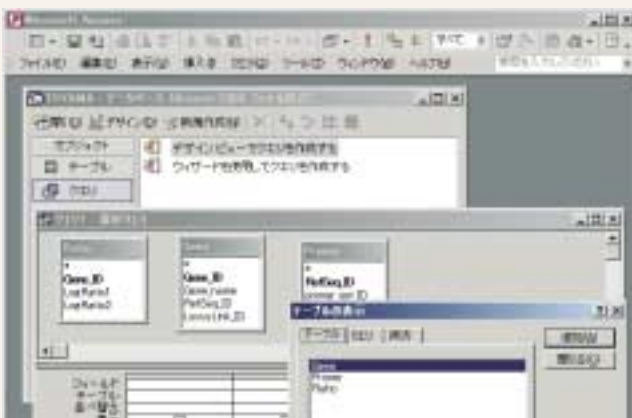


図4 クエリデザインビュー(テーブルの追加)

次につなぎしろになる[Ratio]テーブルの[Gene_ID]から[Gene]テーブルの[Gene_ID]へドラッグして結合線を引くのよ(図5)。

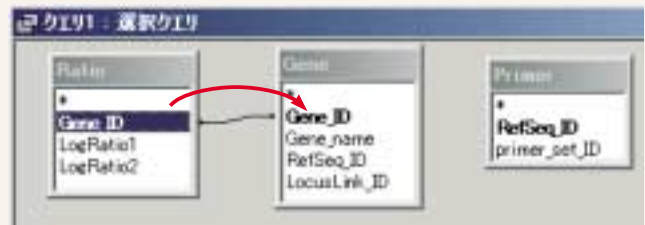


図5 クエリデザインビュー(結合線の設定)

そして、結合線を右クリックしてプロパティを2番に変更(図6)。

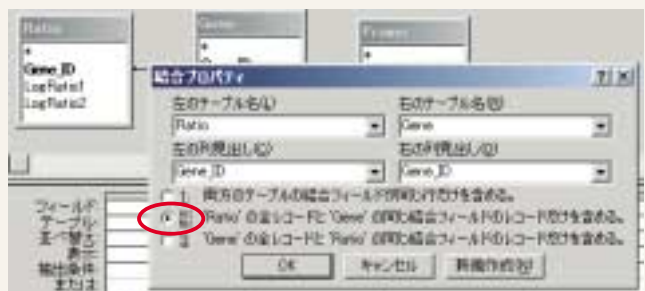


図6 クエリデザインビュー(結合プロパティ)

同じように、[Gene]テーブルの[RefSeq_ID] から [Primer]テーブルの[RefSeq_ID]へ線を引いて、結合線のプロパティを変更すれば結合は完了よ。あとは、表示したい項目をフィールドの欄に順番にドラッグして追加していきだけ(図7)⁹。



図7 クエリデザインビュー(フィールドの追加)

これで、[Ratio]テーブルに必要なデータを追加したテーブルが完成。ビューボタンを押せば、ほらテーブルがサクッと表示されるでしょ(図8)。



図8 データシートビュー(結果表示)

トヤマ君：早い！まだ5分しかたつていませんよ。それに、これなら新しいデータが来てもつなぎしろさえあれば追加するのも簡単ですね。

サエ先輩：保存ボタンを押して、“ All ”とでも名前をつけて保存をしましょう。

さて、次は遺伝子の条件抽出ね。さっき作ったクエリ[All]を使いましょう。まずは、クエリビューを開いて、クエリ[All]を呼び出して、さっきと同じように表示したい項目を選ぶのよ^{*10}。

あとは抽出条件の欄に条件を書き込むだけよ。
[LogRatio1]の列に“ >= 2 ”、[LogRatio2]の列に“ >= 2.5 ”と入力して、[primer_set_ID]の列には“ null ”と入力。この時“ >= 2 ”と“ >= 2.5 ”は別の行に、“ null ”はそれぞれの行に入力するのがポイント！ Access では抽出条件の同じ行は AND(かつ)の関係で、違う行は OR(または)の関係で処理されるのよ(図9)^{*11}。

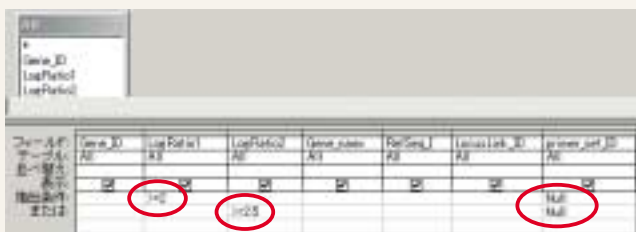


図9 クエリデザインビュー(抽出条件設定)

トヤマ君：先輩、“ null ”って何ですか？それに、どうして2行も書くんですか？

サエ先輩：“ null ”は“ 空欄 ”という意味よ。抽出条件の関係をベン図に書いてごらんねさい。AND、ORの関係や2行書く理由がわかるから^{*12}。

トヤマ君：これで完成ですか？

サエ先輩：そうよ。保存ボタンを押して、“ Select ”とでも名前をつけて保存をしましょう。ビューボタンを押せば・・・対象の遺伝子は41個ね(図10)。必要なら Excel ファイルやテキストファイルとしてエクスポートもできるわよ。

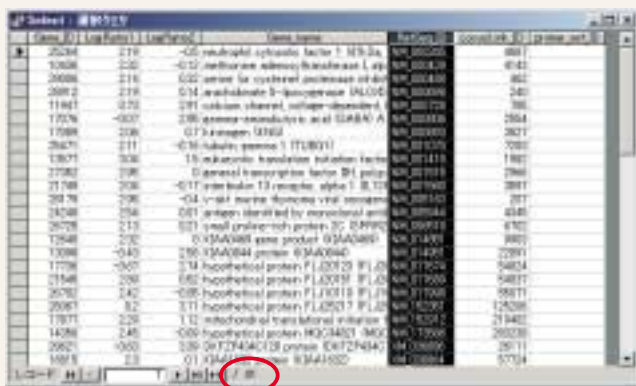


図10 データシートビュー(結果表示)

トヤマ君：余裕で間に合いましたね。Access が自分の PC にインストールされているのは知っていたけど、住所録を作ったり、顧客管理したりする事務系のソフトだとばかり思っていました。先輩ありがとうございました。

サエ先輩：後は自分でがんばってね。

そこへ、教授が再びやってきました。データの説明をするトヤマ君。こころの中で、Access のガイドブック^{*13}を買おうと思うのでした。

いきなり Access 実践編に突入したトヤマ君ですが、読者の皆さんにも Access の便利さの一端が伝わりましたでしょうか。研究室では Excel を使用することが多いと思いますが、Access もバイオサイエンスで使用するのに適した性能をいろいろと持った便利なソフトです^{*14}。

サンプルデータの他、誌面では伝えきれなかった操作の詳細についても、ホームページ(<http://www.takara-bio.co.jp/DL/>)からダウンロードすることができますので、ぜひご覧ください。今後もトヤマ君の研究生生活を追いながら、便利なソフトウェアやテクニックをご紹介してまいりますので、ご注目ください。

- *1 : 本稿で使用したデータは実データとは異なり、ダウンロードしやすいように簡略化したものです。
- *2 : 大量のデータを扱う場合、コピー & ペーストは間違いのもとです。
- *3 : VLOOKUP(検索値、範囲、列番号、検索の型)は、指定された範囲の左端の列で特定の値を検索し、範囲内の対応するセルの値を返します。
- *4 : ほかに、match 関数を使う方法などが考えられます。
- *5 : Perfect Real Time サポートシステムのプライマーです。この研究室ではタカラバイオに遺伝子アクセッションのリストを渡して、バルク対応で設計・合成を委託していました。
- *6 : “ フィルタオプションの設定”(アドバンスドフィルター)機能でできますが、あまりスマートではありません。
- *7 : 2004年12月現在、Office 2003、Access 2003 が最新です。
- *8 : Microsoft® Office Professional Edition 2003 には Access 2003 が含まれていますが、Standard、Personal edition には含まれていません。
- *9 : Access の良さは、グラフィカル画面で直感的にデータベースが操作できる点にあります。
- *10 : クエリはデータをもたず、実態は式なのですが、テーブルと同じように扱えます。ただし、“クエリのクエリのクエリ”なんてやると非常に動作が遅くなる場合があるので要注意。
- *11 : このほか、さまざまな条件を設定して抽出することができます。
- *12 : リレーションデータベースでは、このように集合演算を用いてデータが取り出されます。
- *13 : Access の解説本は星の数ほどあります。しかし、事務処理向けの本が多く、“フォーム”や“レポート”といった入出力の体裁を整える機能の説明が中心になっています。研究目的に特化した本はありませんので、基本的な解説書の次は、“クエリ”、“関数”の項の説明が充実した本が役に立つと思います。インターネット上にも多くの解説や Tips を掲載するサイトがあります。
- *14 : データベース管理ソフトと表計算ソフトで違いがあるのは当然なのですが、例えば塩基配列を整理しようと思ったときの文字数制限(Excel 32,767、Access 65,535)や、多量にデータがあるときの行数制限(Excel 65,536、Access 無制限)など、実用的な仕様にも違いがあります。

Microsoft® Access、Excel、Office は Microsoft® 社の製品です。