진 짜 편 리 합 니 다 !!
발현해석 데이터 해석에 이런 소프트웨어는 어떻습니까?
Microsoft® Access편

데이터 정리나 데이터 해석, 메일, 인터넷 검색 등 컴퓨터는 연구생활에 없어서는 안되는 중요한 도구입니다. 그러나 데이터가 점점 더 다양해지고, 대량의 데이터를 처리해야하는 경우가 생기곤 합니다.
본 기획에서는 실제 실험에서 도움이 될만한 소프트웨어 및 사용상의 팁을 소개해나갈 예정입니다. 첫 회에서는 일반 개인 PC나 연구실에서 데이터 정리 및 해석에 쉽게 사용할 수 있는 데이터베이스 관리 소프트웨어인 Microsoft® Access에 대해 소개하고자 합니다.

어느 대학의 연구실.
이 연구실에서는 최근 Microarray 실험을 시작하였다. 최근 간신히 Excel을 사용하게 된 신입생인 소영에게 담당 교수가 찾아왔다.

교수: 소영! 이 Microarray의 발현비율 데이터(Ratio.xls)와 유전자 정보 데이터(Gene.xls)를 합쳐서 표 하나로 만들어두게. 서둘러 부탁하네....

Excel 파일을 두 개 두고 사라졌다(그림 1)[*1]
서둘러 열어보자 Gene_ID 항목이 공통된 듯 했다.

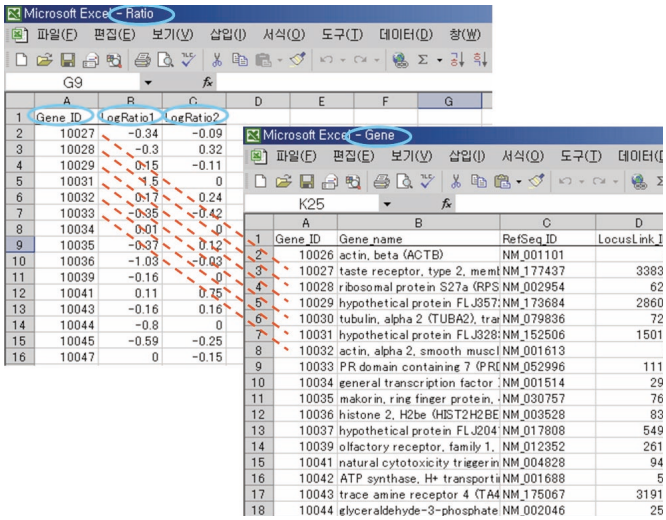


그림 1 데이터 파일(Microarray 발현비율, 유전자 정보)

겨우 Excel을 사용하게 된 소영. 다행이라고 생각했지만...

소영: Gene_ID 수가 일치하지 않네. 숫자도 틀리고. 합쳐서 복사하거나 붙일[*2] 수가 없네...

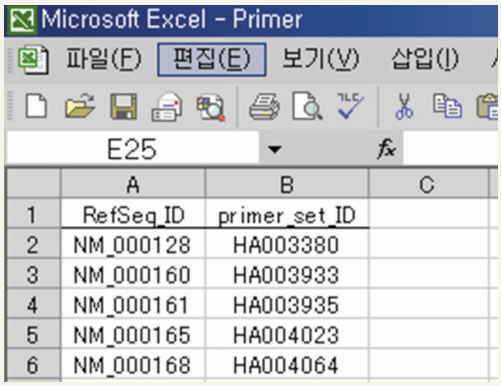
인터넷에서 알아보니 VLOOKUP 함수[*3]를 사용하면 될 것 같아[*4] 서둘러 Excel 데이터를 만들기 시작했지만...

소영: 왠지 계산이 느리군...

그러는 중에 다시 교수님이 찾아왔다.

교수: 소영! 검증용으로 만든 Real time RT-PCR용 Primer 리스트(Primer.xls)[*5]가 있는데(그림 2), 이것도 같이 정리해 주게. 그리고

Microarray 데이터 LogRatio 1의 항목 값이 2 이상이거나 LogRatio 2 항목값이 2,5 이상 중 하나이어야 하며, Primer 리스트에 없는 유전자 데이터만 뽑아 다른 표로 만들어두게. 미안하지만 좀 급해서 30분 후에 찾으러 오겠네. 나중에 Accession(RefSeq ID)으로 리스트를 만들어 다카라에 Real time RT-PCR용 Primer 설계와 합성도 의뢰해두게. 앞으로 여러 가지 데이터를 정리해두지 않으면 안되기 때문에 그렇게 하는걸세...



	A	B	C
1	RefSeq_ID	primer_set_ID	
2	NM_000128	HA003380	
3	NM_000160	HA003933	
4	NM_000161	HA003935	
5	NM_000165	HA004023	
6	NM_000168	HA004064	

그림 2 Primer 리스트

사라지는 교수님. 머리를 감싸며 고민하는 소영.

소영: 더 이상은 이 Excel 파일로는 너무 양이 많아 작동하지 않는데... 게다가 Excel의 오토필터 기능은 같은 열에서는 AND나 OR 관계로 필터 가능하지만, 열 사이는 OR 관계로는 필터가 불가능하고[6]..

그렇게 고심하던 중 컴퓨터나 데이터 해석에 도움을 받고 있는 준호 선배가 들렀다.

소영: 선배님, 도와주세요. 이제 20분밖에 안남았어요...

준호 선배: Excel로는 어렵겠는데... 앞으로도 데이터를 추가해야겠지? 그럴 때는 relational 데이터베이스 소프트웨어를 사용하는거야.

소영: 예?? 그게 뭐죠?

준호 선배는 Windows PC를 작동하면서 이렇게 말했다.

준호 선배: Microsoft® Office가 있는데, 그 중에 Access라는 소프트웨어가 있지[7]. 이걸 사용하면 간단하지. 소영 PC에도 Office가 들어있을걸[8]...
그럼 relational 데이터베이스가 어떤 것인지는 나중에 설명하기로 하고 빨리 Access를 사용해보자. 먼저 3개의 Excel 파일을 불러오고 각각의 표를 작성하는거야(그림 3).

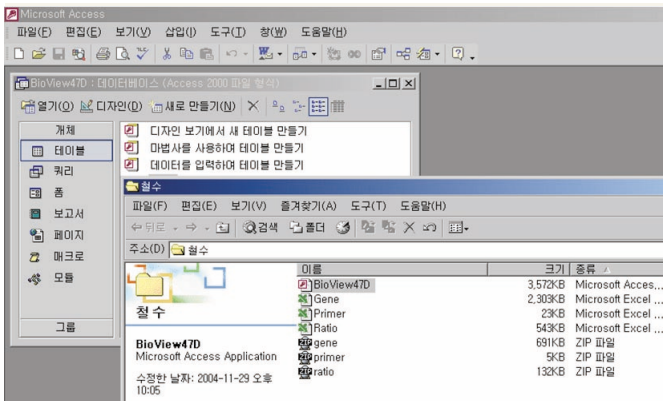


그림 3 데이터베이스 윈도우(불러오기)

다음으로 쿼리를 열고, 3개의 표 [Ratio]와 [Gene]과 [Primer]를 표시하는거야(그림 4).

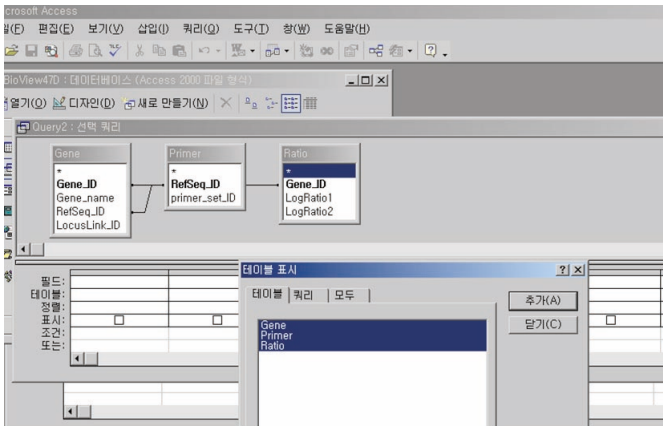


그림 4 쿼리(결합선 설정)

그런 다음 [Ratio]표의 [Gene_ID]에서 [Gene]표의 [Gene_ID]로 drag해서 결합선을 가져오는거야(그림 5).

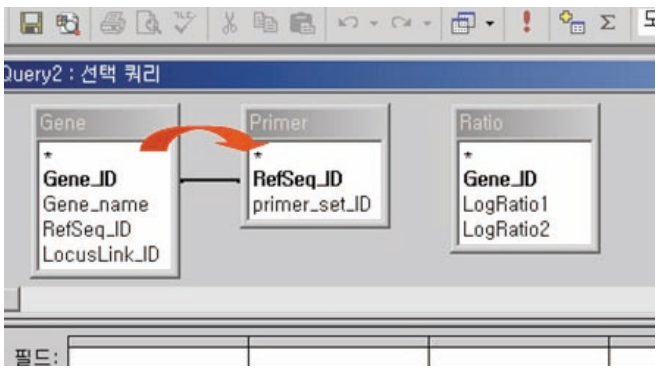


그림 5 쿼리(결합선 설정)

그리고 결합선을 클릭해 조인속성를 2번으로 변경(그림 6).

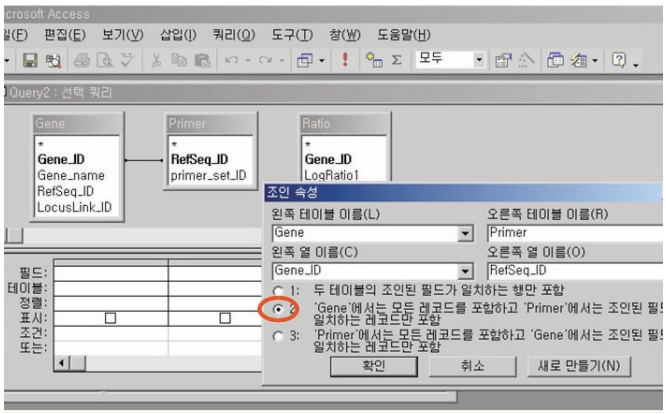


그림 6 쿼리(결합 parameter)

마찬가지로 [Gene]표의 [RefSeq_ID]에서 [Primer]표의 [RefSeq_ID]로 선을 끌어 결합선의 parameter를 변경하면 결합은 완료되지.

남은 것은 표시하고자 하는 항목을 field란에 순서대로 drag 하면서 추가해나가면 돼(그림 7)[*9].

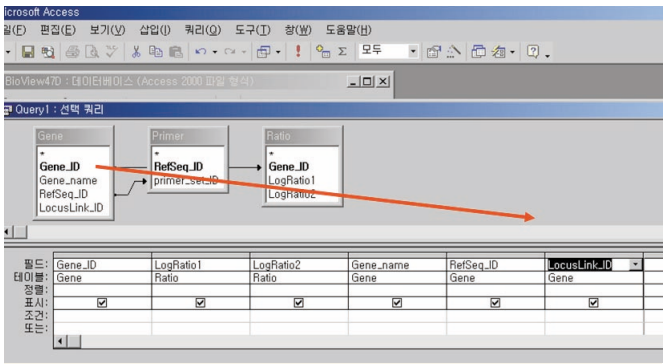


그림 7 쿼리(field 추가)

이것으로 [Ratio] 표에 필요한 데이터를 추가한 표가 완성. 보기 버튼을 누르면 표가 표시되지(그림 8).

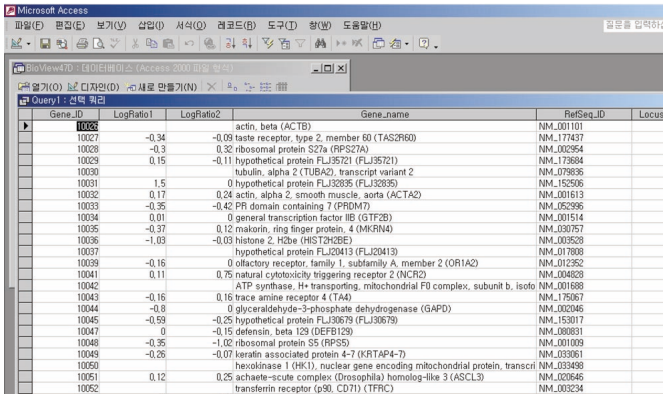


그림 8 Data Sheet 보기(결과 표시)

소영: 와~빠르다! 아직 5분밖에 안지났는데... 게다가 새 데이터가 와도 연결만 하면 추가하는 것도 간단하네.

준호 선배: 저장 버튼을 눌러 "All"로 이름을 붙여 저장하자. 그럼, 다음은 유전자 조건을 해석하면 되지? 전에 만든 쿼리 [All]을 사용하자. 먼저 쿼리 보기를 열어 쿼리 [All]을 불러오기해서 아까처럼 표시하고자 하는 항목을 선택하는거야[*10]. 그런 다음 해석 조건란에 조건을 기입하기만 하면 돼. [LogRatio 1] 열에 "〉=2", [LogRatio 2] 열에 "〉=2.5"로 입력하고 [primer_set_ID] 열에는 "null"로 입력. 이 때 "〉=2"와 "〉=2.5"는 다른 행에, "null"은 각각의 행에 입력하는 것이 포인트! Access에서는 해석조건이 같은 행은 AND의 관계로, 다른 행은 OR의 관계로 처리되지(그림 9)[*11].

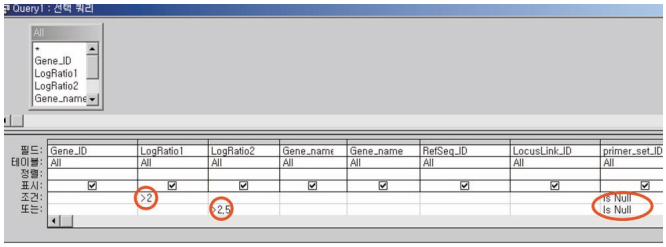


그림 9 쿼리 (해석조건 설정)

소영: 선배님, "null"은 무엇인가요? 그리고 왜 2행이나 적는거죠?

준호 선배: "null"은 "공란"이라는 의미야. 해석조건 관계를 그려봐. AND, OR의 관계나 2행을 적는 이유를 알 수 있으니까[*12].

소영: 이걸로 완성된 건가요?

준호 선배: 맞아. 저장 버튼을 눌러 "Select"라고 이름을 붙여 저장하자. 보기 버튼을 누르면.... 대상 유전자는 41개구나(그림 10). 필요하다면 Excel 파일이나 텍스트파일로 보낼 수도 있지.

그림 10 Data Sheet 보기(결과 표시)

소영: 여유롭게 끝냈네요. Access가 제 PC에 인스톨되어 있는 것은 알고 있었지만 주소록을 만들거나 고객관리를 하거나 사무관련 소프트웨어라고만 생각했습니다. 선배님 감사합니다.

준호 선배: 그럼 이제부턴 직접 잘 해보게...

이 때 교수님이 다시 들렀다. 데이터를 설명하는 소영. 마음 속으로 Access 가이드북[13]을 사야겠다고 생각했습니다.

소영은 새로운 tool을 알아가는 기쁨에 갑자기 Access 실전 편에 돌입했지만, 독자 여러분들에게도 Access의 편리함이 약간은 전해졌는지요? 연구실에서는 Excel 사용이 많을 것으로 생각되는데, Access도 생명공학에서 사용하기에 적합한 성능을 다양하게 갖춘 편리한 소프트웨어입니다[14]. 시료 데이터 외에 지면으로 전하지 못한 상세한 조작에 대해서도 홈페이지(http://www.takara.co.kr)에서 다운로드할 수 있으므로 꼭 한번 보시기 바랍니다.
앞으로도 소영의 연구생활을 바탕으로, 편리한 소프트웨어와 테크닉을 소개할 예정이오니 많은 관심 바랍니다.

*1 : 본 고에서 사용한 데이터는 실제 데이터와는 다르며, 다운로드가 용이하도록 간략화된 것이다.
*2 : 대량의 데이터를 취급할 경우, 복사나 붙이기는 실수의 원인이 될 수 있다.
*3: VLOOKUP(검색값, 범위, 열 번호, 검색의 형태)는 지정된 범위의 좌측 열에서 특정 값을 검색해, 범위 내의 대응 셀 값을 되돌린다.
*4: 이 밖에도 match 함수를 사용하는 방법 등을 생각할 수 있다.
*5: Perfect Real Time 지원시스템 Primer이다.
*6: "필터 옵션 설정"(어드밴스드 필터) 기능으로 가능하지만, 그다지 좋지는 않다.
*7 : 2004년 12월 현재 Office 2003, Access 2003이 최신이다.
*8 : Microsoft Office Professional Edition 2003에는 Access 2003이 포함되어 있는데 Standard, Personal edition에는 포함되어 있지 않다.
*9 : Access의 장점은 그래픽이 강조된 화면에서 데이터베이스를 조작할 수 있는 점에 있다.
*10 : 쿼리는 데이터를 갖고 있지 않으며, 식으로 표와 똑같이 취급된다. 단, "쿼리의 쿼리의 쿼리"라고 하면 동작이 매우 느려질 경우가 있으므로 주의한다.
*11 : 이 밖에 다양한 조건을 설정해 해석할 수 있다.
*12 : Relational 데이터베이스에서는 이처럼 집합연산을 이용해 데이터가 선택된다.
*13 : Access 해설본은 매우 많다. 그러나 사무처리용 책이 많으며 "폼"이나 "레포트"와 같은 입출력 체재를 갖춘 기능 설명이 중심이다. 연구목적으로 특화한 책은 없으므로, 기본적인 해설서 다음에는 "쿼리", "함수" 항목 설명이 충실한 책이 도움이 될 것이다. 인터넷상에도 많은 해설과 팁을 게재하는 사이트가 있다.
*14 : 데이터베이스 관리 소프트웨어와 표 계산 소프트웨어가 다른 것은 당연하지만, 예를 들면 염기서열을 정리하고자 할 때의 문자수 제한(Excel 32,767, Access 65,535)이나 다량으로 데이터가 있을 때의 행수 제한(Excel 65,536, Access 무제한) 등 실용적인 사양에도 차이가 있다.