제20회 한국정보처리학회 추계학술발표대회 논문집 제10권 제2호 (2003. 11)

XML 기반의 Rice 60K DNA Chip 데이터베이스 시스템의 구현

박영배*, 안기영*, 남백희**, 이태호**, 최형인***
*명지대학교 컴퓨터공학과, **명지대학교 생명과학과, ***서울대학교 수리과학부
e-mail : parkyb@mju.ac.kr neocom@nate.com, bhnahm@bio.myongji.ac.kr, thlee@bio.myongji.ac.kr, hichoi@math.snu.ac.kr

Implementation of Rice 60K DNA Chip Database system based on XML

Young-Bae Park*, Gi-Young Ahn*, Baek-Hie Nahm**, Tae-Ho Lee**, Choi, Hyeong In***
*Dept. of Computer Engineering, Myong-Ji University
**Division of Bioscience and Bioinformatics, Myong-Ji University
*** School of Mathematical Science, Seoul National University

요 약

본 논문은 Rice 60K DNA Chip 의 실험데이터를 기반으로 한 데이터베이스의 구축과 XML 기반 검색시스템을 설계 및 구현에 대해 설명한다. 본 시스템은 실험 데이터를 저장하기 위하여 RDBMS 를 사용하고 Chip 데이터를 검색하기 위해 XML 기반의 검색시스템을 사용한다. 이를 위해 일반 속성으로 저장될 수 있는 데이터들은 데이터베이스의 테이블의 속성 값으로 저장하고, XML 기반 검색시스템을 통해 검색할 수 있도록 한다. 그리고 BLAST 내용을 기반으로 하는 데이터는 테이블을 별도로 만들어서 검색이 가능하도록 한다.

1. 서론

DNA Chip 은 기존의 분자 생물학적 지식과 눈부시게 발전한 기계 및 전자공학의 기술을 접목해서 만들어졌다. 기계 자동화와 전자 제어 기술 등을 이용하여 적게는 수백 개부터 많게는 수십만 개의 DNA 를 아주 작은 공간에 집적하여 만든 것이다.

즉 DNA chip 이란 유전자 검색용으로서 각각이 특정 유전정보를 나타내는 다양한 서열을 가진 DNA 를 고밀도로 붙여 놓은 것을 말한다. 이러한 DNA chip 은 생물의 특정상태에서 나타나는 유전정보의 발현 양상인 transcriptional profile 을 광범위하게 조사할 수 있다. 이전에 행해졌던 유사한 실험의 경우 유전 물질을 붙이는 매체로 nitrocellulose 막을 사용 하는데 반하여 DNA chip 에서는 유리와 같은 고형체를 사용함으로 DNA chip 은 아주 적은 양의 유전물질을 고밀도로 붙일 수 있게 되었고 동시에 많은 수를 검색할 수 있게 된 것이다. DNA chip 은 붙이는 유전물질의 성격과 길이에 따라 cDNA chip 과 oligonucleotide chip 으로 나누어 질 수 있다. 이들의 차이는 이름에서도 알 수 있듯이 cDNA chip 에는 최소한 200bp 이상의 유전자 (full-length open leading frame 또는 EST)가 붙여져 있고, oligonucleotide Chip 에는 약 25 ~ 70 개의 염기들로 이루어진 oligonucleotide 가 붙여져 있다. 이와 같은 DNA Chip 의 특징은 동시에 최소한 수천개 이상의 유전자 발현 양상을 빠른 시간 안에 조사할 수 있다는 것이다.[1]

본 논문에서는 (주)그린진 바이오텍[1]에서 세계 최초로 개발하여 실험중인 Rice 60K DNA Chip (Oligonucleotide Chip)의 실험데이터를 기반으로 한 데이터베이스의 구축과 XML 기반 검색시스템의 설계

및 구현에 대해 설명한다. 본 시스템은 실험데이터를 저장하기 위하여 XML 데이터 포맷을 지원하는 RDBMS 인 SQL2000 을 사용하고 DNA Chip 데이터를 검색하기 위해 XML 기반의 검색시스템을 사용한다. 이를 위해 일반 속성으로 저장 될 수 있는 데이터들은 데이터베이스의 테이블의 속성 값으로 저장하고 XML 기반 검색시스템을 통해 검색할 수 있도록 하였다. 그리고 Genbank 의 유전자 데이터와의 비교 검색하는 검색방법을 BLAST 검색이라 한다. 이를 기반으로 하는 Rice 60K DNA Chip BLAST 데이터는 테이블을 별도로 만들어서 역으로 실험데이터의 검색이 가능하도록 한다.

본 논문의 구성은 다음과 같다. 제 2 장에서는 본 논문에서 제안하는 XML 기반의 검색시스템의 구조와 구성 모듈에 대해 설명한다. 제 3 장에서는 XML 기반의 검색시스템의 데이터구조에 대해 설명한다. 제 4 장에서는 본 시스템의 구현 및 실험을 한다. 제 5 장에서 결론을 맺도록 한다.

2. 시스템 구조와 모듈

본 장에서는 제안하는 시스템 구조에 대해서 간단히 설명한다. 그림 1 은 본 시스템의 구조를 표현한 것이다.

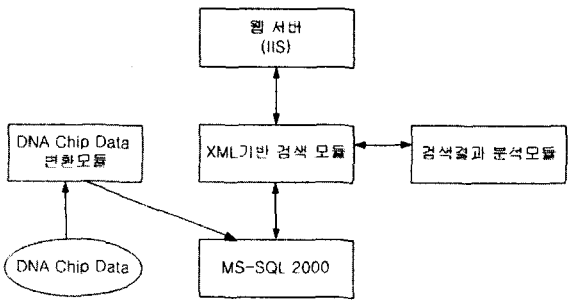


그림 1 XML 기반 검색시스템의 구조

- DNA Chip Data 변환 모듈

본 모듈에서는 텍스트기반의 DNA Chip 의 실험데이터 파일을 파싱(Parsing)하여 SQL2000 의 벌크 로딩(Bulk Loading)포맷으로 변환한다.
그리고 변환된 벌크 로딩 데이터들은 SQL2000 의 벌크 로딩 명령어를 사용하여 빠르게 데이터베이스에 로딩되도록 한다.

- 검색 웹 페이지

검색 웹 페이지는 XML 을 이용하여 보다 다양한 DNA Chip 데이터베이스를 검색할 수 있도록 Chip DB 검색 인터페이스, Slide 검색 인터페이스. Spot 검색 인터페이스를 제공한다. 이를 통해 사용자는 빠르고 폭넓은 검색을 제공 받을 수 있다.

- XML 기반 검색 모듈

본 모듈은 웹 페이지로부터 사용자 입력을 받아 이를 분석하여 DBMS 를 검색하는 기능을 제공한다. 구현언어로는 XML 을 사용한다.

- *DBMS*

본 시스템에 사용된 SQL2000 은 Microsoft 사에서 개발한 RDBMS 이며 URL 쿼리, 템플릿쿼리, xPath 쿼리, Open-XML 등 여러가지 XML 지원기능을 내장하고 있다. 그리고 동사의 IIS 웹서버를 축으로 언어에 구애받지 않는 XML 처리 인터페이스를 제공하고 있다.

- 검색결과분석모듈

본 모듈은 검색한 결과를 XML 형식으로 출력하여 미리 준비한 XSLT 스타일시트와 함께 브라우저에 출력한다. 그리고 각각의 항목을 클릭하면 관련 테이블 전체의 내용이 출력되도록 한다.
사용자는 출력된 결과를 조회하고 동시에 페이지를 자신의 컴퓨터에 저장 할 수 있게 됨으로써 데이터의 재활용도 가능해진다.

3. 테이블의 구성

Rice 60K DNA Chip Database 는 다음의 테이블로 구성되어 있으며 각각의 테이블의 이름과 내용은 다음

과 같다.

그림 2 는 테이블의 구성을 나타낸다

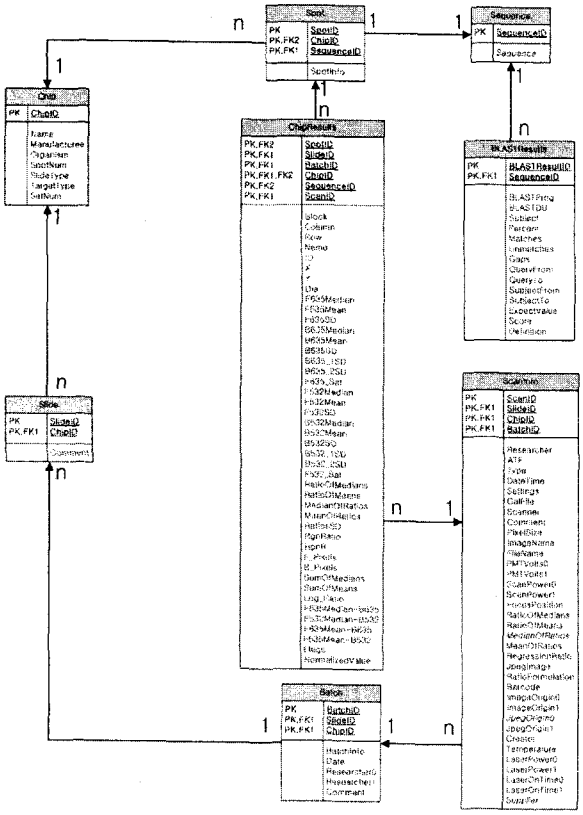


그림 2 테이블 구성

- Chip

DNA Chip 에 대한 name, manufacture, organism 등의 정보가 저장된다.

- Slide

A, B 슬라이드로 구성된 DNA Chip 의 번호를 관리한다.

- Spot

하나의 DNA Chip 은 64896 개의 spot 으로 구성되다. 각각의 spot 은 실제 유전자의 정보를 가지고 있으며 이에 대한 내용이 테이블에 저장된다..

- ChipResults

DNA Chip 을 가지고 실험한 실험데이터들을 저장한다.

- ScanInfo

실험 후 DNA Chip 을 스캐닝하여 그 데이터를 저장한다.

- Batch

실험조건이나 실험시간, 날짜에 대한 내용이 저장된다.

- Sequence

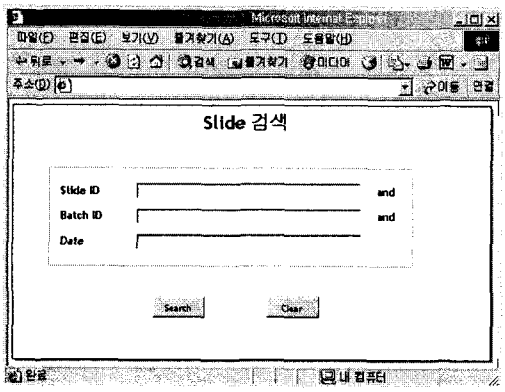
각 Spot 에 대한 유전자정보 서열 데이터를 나열하여 저장된다.

- BLASTResults

Genbank 에 등록된 유전자정보 서열 데이터로서 이를 바탕으로 하여 Chip Result 의 내용과 비교검색을 할 수 있다.

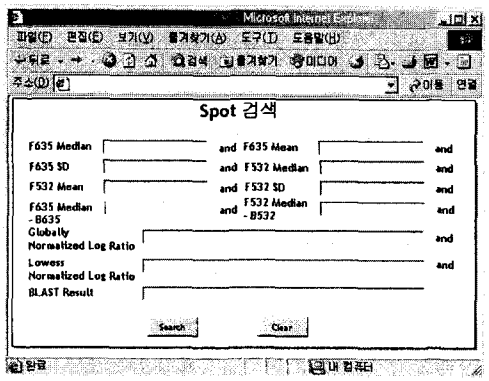
4. 구현 및 실험

- 슬라이드 검색 인터페이스



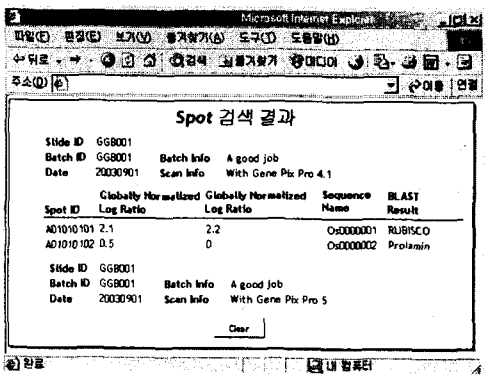
슬라이드를 검색하기 위한 조건을 입력한다.

- Spot 검색 인터페이스



실제 Spot 데이터를 입력하여 검색한다.

- Spot 검색 결과



검색결과에서 SpotID 를 클릭하면 해당 Spot 에 대한 유전자 서열 데이터가 출력되게 된다.

5. 결론 및 향후 연구

본 논문에서는 XML 기반의 Rice 60K DNA Chip 데이터베이스 시스템의 구현에 대한 소개를 했다. 본 시스템은 1) DNA Chip 데이터변환모듈, 2)XML 기반 웹 검색모듈, 3)검색결과 분석모듈로 구성되며 XML 언어를 사용하여 데이터의 일관성을 명확하게 정의하여 보다 빠른 검색결과를 지원한다. 향후 연구로는 범용적인 데이터베이스의 설계와 인터페이스의 사용으로 Arabidopsis 27K Chip 이나 Rice 3K Chip 등의 DNA Chip 데이터베이스에도 적용이 가능하도록 확장한다.

참고문헌

[1] http://www.ggbio.com

[2] http://sqler.pe.kr/sql2k/1303.asp

[3] http://www.oasis-open.org/cover/mageML.html

[4] Feng Tian, David J. Dewitt, Jianjun Chen, Chun Zhang, "The Design and Performance Evaluation of Alternative XML Storage Strategies:

[5] Gavin Sherlock, "Microarray Data Storage and Exchange" Stanford university

[6] 채진석, "XML 기반 과학기술 정보 처리", 지식정보인프라, 한국과학기술정보연구원, 2001

[7] 신시아 기버스, 퍼 잼벡저, 이정근, 오석준, 김종민 역, "Bioinformatics Computer Skills" 한빛미디어, 1999

[8] (주)다산기술 부설 기술연구소 공저, "XML 기반 웹 사이트 개발" 2002

[9] 양진옥,김상수, "웹기반의 Genbank 특허 데이터 검색 시스템의 설계 및 구현" 2001 가을학술발표논문집, 제 28 권 2 호, pp. 43-45, 2001

[10] 황두성, "다양한 웹 데이터를 이용한 특정 유기체의 단백질 상호작용 데이터베이스 개발", 정보처리학회논문지 제 9-D 권, pp.1091-1096, 2002

[11] 정호열, 황미녕, 유영중, 조환규, "cDNA 마이크로어레이 이미지를 위한 그래프 모델과 분석 알고리즘", 정보과학회논문지:시스템 및 이론 제 29 권 제 7 호, 2002