특 집

포스트 게놈시대의 생명정보 데이터베이스 연구 : Issues & Challenges

박 성 희, 류 근 호
충북대학교 데이터베이스 연구실

I. 서 론

생명정보학은 분자생물학 분야의 연구 문제를 해결하기 위해 대용량의 생명정보 데이터를 분석 관리하기 위해 컴퓨터과학, 수학, 통계, 화학 및 물리학 같은 학문을 적용하는 새로운 학문분야이다. 이러한 측면에서 보면, 생명정보 데이터의 저장과 관리를 책임지는 데이터베이스의 역할은 생명정보학 분야에서 더욱 중요해지고 있다.

현재까지 500~1,000 이상의 많은 생명정보데이터베이스가 존재하고 있다. 이중 가장 대표적인 생명 정보 데이터베이스 관리 기관으로 미국의 NCBI(National Center for Biotechnology Information)[Davi00], 일본의 DDBJ(DNA DataBank of Japan)[Tate97], 유럽의 EMBL(European Molecular Biology Laboratory)[Stoe01] 이 있다. 세 기관에서는 생물데이터 저장관리 뿐만 아니라 데이터의 수집과 검색 서비스를 제공한다. 이외에도 단백질 서열 SWISS-PROT[Bair00]과 PIR[Bark01] 및 3차 구조 데이터베이스 PDB[Bern00]가 존재한다. 이들 데이터베이스를 바탕으로 단백질 서열 및 구조 모티프를 유도하고 군집화하여 구축한 단백질 패밀리 데이터베이스와 구조 분류 데이터베이스 등이 있다. 신진대사 경로 관련 데이터베이스 KEGG[Kane97]과 WIT[Selk9], 인간 유전병과 유전체의 돌연변이에 대한 데이터베이스로 OMIM[Scot99]과 HGMD[Coop98]가 있다. 또한 이질적인 생명정보 데이터를 하나의 생명체에 대해 통합된 데이터베이스를 구축한 GDB 등이 있다.

그러나 최근 새롭게 기하급수적으로 증가한 생명정보 데이터의 저장 관리는 생명정보 데이터의 고유의 특성과 이러한 데이터의 특성을 이해하기 위한 생물학적 지식의 요구로 인해 저장관리에 많은 문제점과 어려움이 발생되고 있다. 따라서 이 원고에서는 생명정보 데이터를 관리하기 위해 국내외 기존 구축된 생명정보 데이터베이스를 살펴보고 생명정보 데이터관리 문제점과 연구 방향을 제시한다.

II절에서는 생명정보의 종류 및 특성을 기존 데이터와 비교하여 설명한다. III절에서는 기존의 생명정보 데이터베이스의 특징 및 저장 구조에 대해 기술한다. IV절에서는 III절에서 기술한 생명정보 데이터베이스에서 발생하는 문제점 및 컴퓨터과학자의 입장에서 생명정보 관리에 대한 연구 이슈를 소개한다.

II. 생명정보 데이터의 분류 및 특성

생명정보 데이터의 분석과 저장 관리를 위해서는 데이터의 특성을 이해하고 특성에 맞는 분석 모델의 선택과 이에 대한 제약사항을 추가해야 한다. 특히 HGP 이후 출현한 생명정보 데이터는 그 종류가 다양하고 항상 변화되고 유동적이므로 더욱 그 특성에 대한 분석이 중요하다. <그림 1>에서처럼 생명정보 데이터는 정적 데이터, 동적 데이터, 분석 관련 데이터와 주석 데이터로로 나누어진다.

• **정적 데이터** : 유전 형질에 대한 데이터로서 생

물학적 개체와 개체들 사이의 관계 데이터. 예, 염기, RNA, 단백질, 단백질 상호작용, 신진대사 경로

- **동적 데이터** : 표현 형질에 대한 데이터. 예, 동적인 생물학적 프로세스
- **분석 데이터** : 생물학적 개체와 개체간의 관계를 분석하기 위해 이용될 수 있는 생물학 및 전산학적 방법에 대한 데이터
- **참고문헌 및 주석** : 위에서 기술한 데이터 타입에 대한 과학적 참조 논문 및 문헌 정보와 주석에 대한 정보

정적 생명정보 데이터는 크게 핵산, RNA과 단백질 데이터로 나누어진다. 이들 데이터 중 단백질 데이터는 다시 서열과 2차원이나 3차원의 구조 데이터로 분류될 수 있다. 이외에도 좀더 복잡한 구조의 단백질의 상호작용이나 신진대사 경로, 유전자 조절 경로를 나타내는 네트워크, 유전자 발현정도를 나타내는 DNA chip이나 SAGE 데이터가 있다. 참고문헌 및 주석 정보는 생명정보 데이터에 추가적으로 실험 설명, 데이터의 해석을 위한 주석, 질병에 대한 클리닉 데이터와 문헌정보 등이 해당된다.

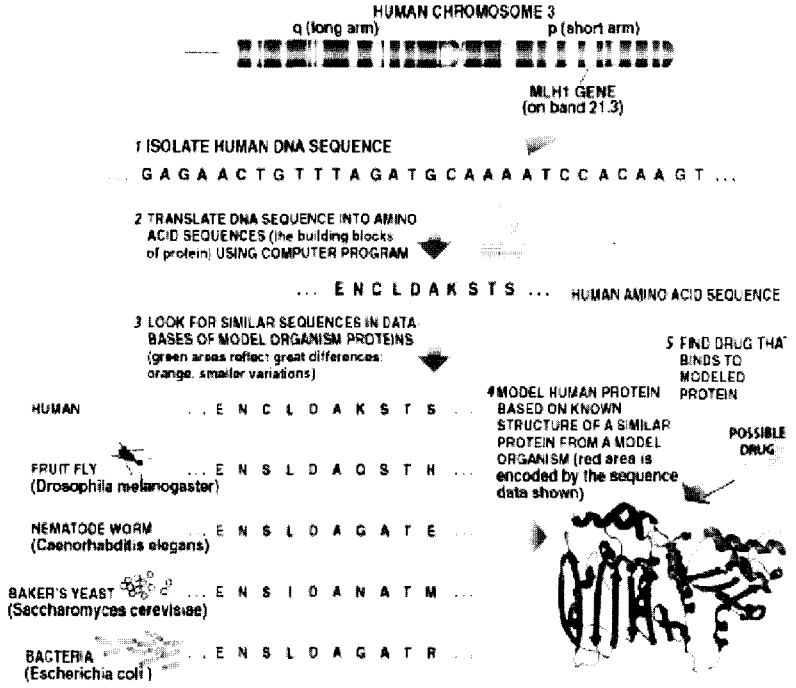
위에서 설명한 다양한 생명정보 데이터의 주요한 특성은 아래와 같이 요약된다.

- 데이터 사이의 숨겨진 연관 관계가 많음
- 데이터 구조 및 상호 관계 복잡
- 계층적이거나 순환적 구조로 표현됨
- 상태가 항상 변화고 유동적임

데이터베이스 측면에서 생명정보 데이터의 특성은 아래와 같다.

- 다차원의 비정형 데이터 타입
- 비정형 질의 요구
- 강력한 패턴 매칭 질의 요구
- 이질적인 데이터 소스
- 연합 데이터베스에 저장되어 데이터가 분산
- 동일 개념에 다양한 용어의 사용

비정형 데이터로서 생명정보 데이터 타입을 살



〈그림 1〉 생명정보 데이터 종류[Gers00]

펴보면 서열데이터는 한정된 알파벳문자의 반복이지만 순서와 위치정보가 고려되어야 하며 이들의 화학적 성질 또한 분석을 위해 중요하다. 현재는 서열데이터를 1차원의 선형의 스트링, time series, probability vectors로 처리하고 있다. 대표적인 비정형 데이터 타입 중 하나는 이미지이다. 이미지 타입을 갖는 데이터는 DNA chip 같은 유전자 발현 데이터와 단백질 동정 데이터인 2D gel image가 있다. 단백질을 구성하는 원자는 3차원의 좌표를 갖으며 유전자 조절경로, 단백질 상호작용, 신진대사 경로와 신호전달 경로와 같은 네트워크는 그래프로 표현된다.

II. 생명정보 데이터베이스 및 데이터 관리 시스템

현재까지 구축된 생명 정보 데이터베이스는 〈표 1〉에서처럼 그 종류가 다양하고 데이터량 또한 기하급수적으로 증가되고 있다. 유전체 서열의 통합관리를 하는 NCBI의 GenBank[Oste01]와 EMBL[Stoe01] 및 DDBJ[Tate97]의 유전체 데이터베이스가 있다. 단백질 관련 데이터베이스에는 서열에 대한 수준 높은 주석 정보를 제공하는 Swiss-Prot[Bair00]과 구조 데이터베이스인

〈표 1〉 생명정보소스 및 데이터 크기

데이터 소스	데이터 크기	데이터베이스	생명정보학 연구 분야
DNA 서열	8.2 백만 서열 (95억 베이스)	GenBank, EMBL, DDBJ	• coding 지역과 non-coding 지역 분리 • Intron과 Exons 식별 • 유전자 구조 예측
단백질 서열	300,000 서열 (대략 각 서열 당 300개 아미노산)	• **1차 데이터베이스** SWISS-PROT, PIR • **복합데이터베이스** OWL, NRDB • **패밀리 데이터베이스** PROSITE, PRINT, Pfam	• 서열 비교 알고리즘 • 다중 서열 정열 알고리즘 • 서열 유사성 검색 • 서열 모티프 식별 기법
분자 구조	13,000 구조 (대략 각 분자구조 당 10000 원자좌표)	• **1차 데이터베이스** Protein Data Bank(PDB) Molecular Modeling Database(MMDB) Nucleic Acids Database(NDB) HIV Protease Database ReLiBase • **구조 분류 데이터베이스** SCOP, CATH, FSSP	• 2차, 1차 구조 예측 • 3차 구조 정열 알고리즘 • 단백질 기하 측정 • 분자 표면과 부피 모양 계산방법 • 분자 상호작용 • 분자 시뮬레이션 (분자간 인력 계산, 분자 이동, 도킹 예측)
유전체	완벽한 유전체 (각 유전체당 1.6백만~30억 베이스)	GDB, ACEDB, Entrez gnome, GeneCensus, COGs	• 반복지역의 특성 발견 • 유전자 구조 배정, 계통 분류 분석 • 유전체 수준의 통계정보 측정 (단백질 함유, 신진대사 경로, 특정 유전자 관련 질병에 대한 계통 분석)

데이터 소스	데이터 크기	데이터베이스	생명정보학 연구 분야
유전자 발현	이스트에 대한 가장 큰 데이터 셋 : 약 6,000 유전자에 대한~20배 점으로 측정	Stanford Microarray Database (SMD)	• 상호관련 발현 패턴 분석 • 발현 데이터와 서열, 구조 및 생화 학 데이터와 매핑
신진대사경로		KEGG, WIT, Ecocys	• 경로 시뮬레이션
문헌 정보	대략 1000만 인용 정보	PUDMed	• 자동화된 그래픽 검색이 가능한 디지털 라이브러리 • 문헌정보 기반 지식 베이스 구축

PDB[Barn00] 데이터베이스가 있다. 이러한 데이터베이스를 기반으로 모티프 발굴과 서열 유사성 검색을 통해 유도되어진 단백질 서열 모티프 및 패밀리 데이터베이스인 PROSITE[Hofm99]와 PRINT, PFam 등이 존재한다. 단백질 3차 구조 도메인의 기능 및 구조가 유사한 단백질을 군집화한 구조분류 데이터베이스로 SCOP[Murz95], CATH[Oren97], FSSP[Holm96]가 대표적이다.

이 절에서는 현재까지 구축되어진 생명정보데이터베이스의 데이터 모델과 관리시스템을 설명하고 이에 해당하는 시스템 및 데이터베이스에 대한 사례를 설명한다.

1. 데이터 모델 및 데이터 관리

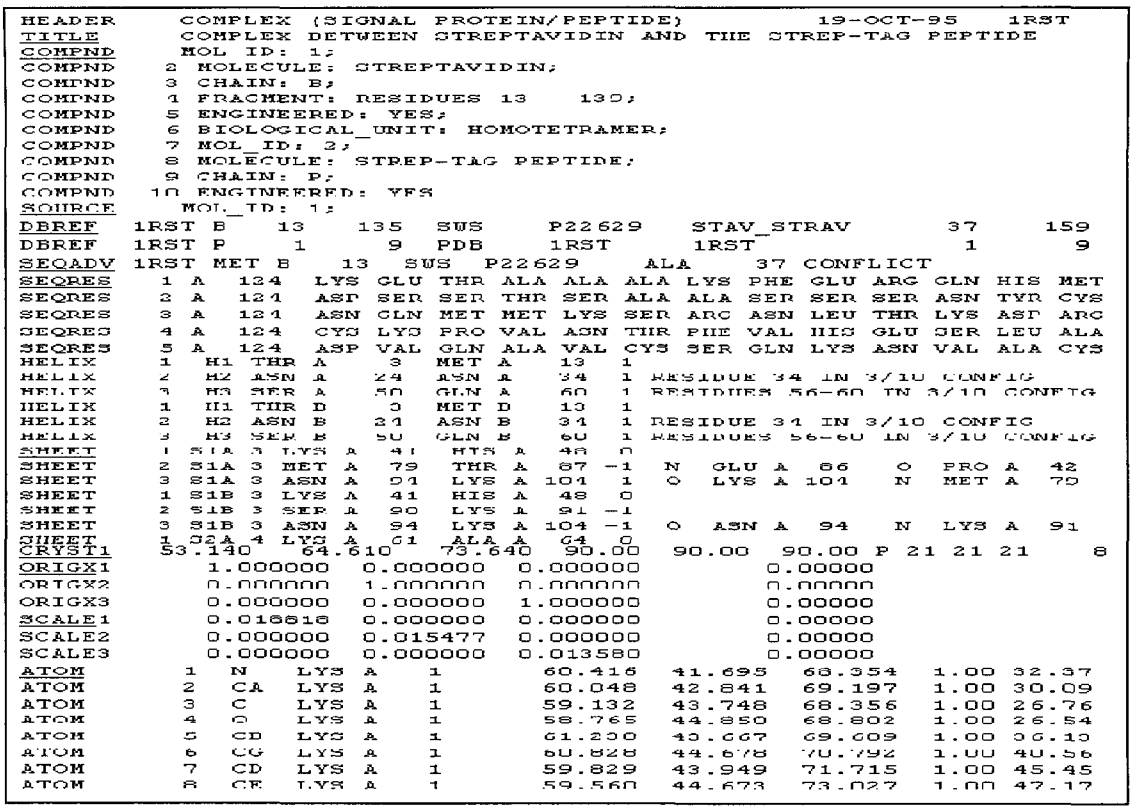
초기의 생명정보 데이터베이스 구축은 단순히 생물 데이터에 대한 접근과 단순한 정보에 대한 검색을 지원하는 기능이 위주였다. 생명정보 데이터베이스에서 사용하는 데이터베이스 관리 시스템은 아래와 같이 분류할 수 있다.

• 플랫 파일의 집합

대부분의 초기 생명정보 데이터베이스는 인덱스된 ascii 텍스트 파일의 집합으로 구축되어졌다. 대부분의 생명정보데이터가 복잡한 구조이어서 플랫파일은 중첩된 레코드를 포함하거나, 집합, 리스트와 같은 복합 데이터 구조의 데이터 타입을 포함한다. 90년대에 중반이후부터 관계형 데이터베이스를 이용하여 데이터베이스를 구축하고 있으나 플랫파일을 관계형 데이터베이스로 포팅하기 위한 비용이 매우 높다. 또한 복합 데이터 타입을 관계형에서 지원하지 않아 더욱 어려운 점이 많다. 그러나 플랫파일은 필드와 필드 값의 구분이 불분명하고 데이터 타입의 불일치하다. 또한 특정 시점에 따라 포맷이 다르며 데이터 값의 범위가 모호하다는 단점을 갖는다. 초기에는 데이터가 변화지 않고 다수 사용자의 접근을 허용하지 않는 가정하에 플랫파일 관리가 가능하였다. 하지만 현재 생명정보 데이터는 매우 빠르게 변화를 반영하는 갱신 관리와 여러 분야의 사용자가 데이터에 접근하여 사용하므로 데이터베이스의 트랜잭션과 병행수행 제어 관리를 이용한 데이터의 일치성 유지가 중요하게 다루어진다.

• 표준 데이터베이스 관리 시스템

90년대 중반부터 관계형, 객체지향, 객체-관계형 데이터베이스 관리시스템을 이용하여 생명정보 데이터베이스를 구축하였다. 데이터간의 상호관계 및 데이터의 복잡성으로 인하여 객체지향 모델이 관계형 모델보다 데이터베이스 모델링을 위해서는 더 적합하다. 그러나 데이터의 규모와 성능 측면에서 관계형 데이터베이스 및 객체-관계형 데이터베이스의 이용이 효율적이다. GenBank



```
HEADER     COMPLEX (SIGNAL PROTEIN/PEPTIDE)          19-OCT-95   1RST
TITLE      COMPLEX BETWEEN STREPTAVIDIN AND THE STREP-TAG PEPTIDE
COMPND     MOL_ID: 1;
COMPND   2 MOLECULE: STREPTAVIDIN;
COMPND   3 CHAIN: B;
COMPND   4 FRAGMENT: RESIDUES 13    139;
COMPND   5 ENGINEERED: YES;
COMPND   6 BIOLOGICAL_UNIT: HOMOTETRAMER;
COMPND   7 MOL_ID: 2;
COMPND   8 MOLECULE: STREP-TAG PEPTIDE;
COMPND   9 CHAIN: P;
COMPND  10 ENGINEERED: YES
SOURCE     MOL_ID: 1;
DBREF   1RST B    13   135  SWS     P22629    STAV_STRAV       37    159
DBREF   1RST P     1     9  PDB     1RST      1RST              1      9
SEQADV  1RST MET B    13   SWS   P22629     ALA     37 CONFLICT
SEQRES    1 A   124  LYS GLU THR ALA ALA ALA LYS PHE GLU ARG GLN HIS MET
SEQRES    2 A   124  ASP SER SER THR SER ALA ALA SER SER SER ASN TYR CYS
SEQRES    3 A   124  ASN GLN MET MET LYS SER ARG ASN LEU THR LYS ASP ARG
SEQRES    4 A   124  CYS LYS PRO VAL ASN THR PHE VAL HIS GLU SER LEU ALA
SEQRES    5 A   124  ASP VAL GLN ALA VAL CYS SER GLN LYS ASN VAL ALA CYS
HELIX     1   H1 THR A     3  MET A    13   1
HELIX     2   H2 ASN A    24  ASN A    34   1 RESIDUE 34 IN 3/10 CONFIG
HELIX     3   H3 SER A    50  GLN A    60   1 RESIDUES 56-60 IN 3/10 CONFIG
HELIX     1   H1 THR B     3  MET B    13   1
HELIX     2   H2 ASN B    24  ASN B    34   1 RESIDUE 34 IN 3/10 CONFIG
HELIX     3   H3 SER B    50  GLN B    60   1 RESIDUES 56-60 IN 3/10 CONFIG
SHEET     1 S1A 3 LYS A  41  HIS A  48  0
SHEET     2 S1A 3 MET A  79  THR A  87 -1  N  GLU A  86   O  PRO A  42
SHEET     3 S1A 3 ASN A  91  LYS A 101  1  O  LYS A 101   N  MET A  79
SHEET     1 S1B 3 LYS A  41  HIS A  48  0
SHEET     2 S1B 3 SER A  90  LYS A  91 -1
SHEET     3 S1B 3 ASN A  94  LYS A 104 -1  O  ASN A  94   N  LYS A  91
SHEET     1 S2A 4 LYS A  61  ALA A  64  0
CRYST1   53.140   64.610   73.640  90.00  90.00  90.00 P 21 21 21     8
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.018816  0.000000  0.000000        0.00000
SCALE2      0.000000  0.015477  0.000000        0.00000
SCALE3      0.000000  0.000000  0.013580        0.00000
ATOM      1   N   LYS A   1      60.416  41.695  68.354  1.00 32.37
ATOM      2   CA  LYS A   1      60.048  42.841  69.197  1.00 30.09
ATOM      3   C   LYS A   1      59.132  43.748  68.356  1.00 26.76
ATOM      4   O   LYS A   1      58.765  44.850  68.802  1.00 26.54
ATOM      5   CB  LYS A   1      61.290  43.667  69.609  1.00 36.10
ATOM      6   CG  LYS A   1      60.828  44.678  70.792  1.00 40.56
ATOM      7   CD  LYS A   1      59.829  43.949  71.715  1.00 45.45
ATOM      8   CE  LYS A   1      59.560  44.673  73.027  1.00 47.17
```

〈그림 2〉 PDB 플랫 파일 예

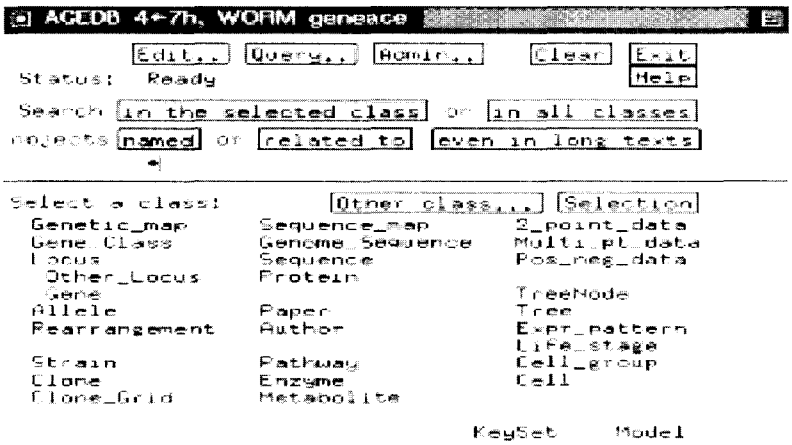
와 PDB에서는 Sybase 관계형 데이터베이스를 이용하고 단백질 서열 데이터베이스인 PIR에서는 객체-관계형 데이터베이스인 Oracle8i를 이용하고 있다. 대장균 박테리아(E. coli)의 유전자와 이러한 유전자에 의해서 발현된 효소와 효소간의 반응과 반응에 의해 발생되는 대사경로를 저장관리하는 Ecocyc 데이터베이스는 객체지향 모델을 이용하고 대상경로 및 유전체의 통합 정보를 검색하기 위한 그래픽 질의 인터페이스를 제공한다.

• 확장 데이터 관리 시스템

이 시스템은 기존의 표준 관계형이나 객체지향 데이터베이스 관리시스템에 특정 생명정보를 저장관리하기 위한 데이터타입, 연산, 인덱스 및 질의 최적화를 위한 비용 함수 등을 확장한 데이터베이스 관리시스템이다. ACEDB는 객체지향 데이터베이스를 기반으로하여 C.elegans 개체에 대한 생명정보 클래스를 모델링하였다. ACEDB 객체는 트리구조로 표현되고 노드는 객체 또는 데이터 값을 나타내고 arc는 속성과의 관계를 나타낸다. 또한 반구조 데이터 모델을 반영하여 비정형적인 데이터를 수용할 수 있도록 하였고 객체에 속성을 추가하여 스키마의 확장을 쉽게 할 수 있다. 트랜잭션과 회복 및 인덱싱을 지원하고 OQL을 확장한 반 구조언어인 LOREL과 유사한 AQL이라는 질의 언어를 제공한다.

• 통합 관리 시스템

생명정보 데이터는 복잡하기 때문에 관계형 데이터 모델로 모델링하기에는 적합하지 않고 분석을 위한 데이터는 이질적 데이터를 통합하여 분석해야된다는 필요성에 의해 통합 관리 시스템에 대한 연구가 시작되었다. 통합 관리 시스템으로는



〈그림 3〉 ACEDB의 주요 클래스

SRS[Etzo96], BioKleisli[Davi99]과 OPM[Mark01]이 있다. 대표적인 통합관리 시스템인 OPM은 버클리대학의 데이터 관리와 연구 그룹에 의해서 제시되었다. 관계형 및 객체지향 데이터베이스를 통합하고 사용자에게 그래픽 질의 인터페이스를 제공하기 위해 객체 모델에 촛점을 둔 생명 정보 통합 툴킷이다. 실제로 객체에 대한 질의와 모델링은 객체지향 모델에 기반하지만 실제로 데이터 관리는 관계형 데이터베이스에 기반한다. OPM은 타입 및 질의 처리에 대해 제한적이지만 질의 그래픽 인터페이스가 뛰어나다.

• XML & 반 구조관리시스템

비정형적이며 빠르게 변화되고 구조가 복잡하다는 생명 정보 데이터는 반구조적 특성을 갖는다. 따라서 반 구조적 데이터베이스 관리시스템 기술을 적용하려는 연구가 시작되고 있다. 특히 GenBank, PIR과 PDB 등에서는 플랫파일에 대한 XML 포맷을 제공하고 있으며 NCBI에서는 BLAST의 결과를 XML 포맷으로 반환하는 BlastXML을 개발하였으며 SWISS-PROT은 단백질 서열의 플랫 파일을 XML로 변환하고 XML 파서를 이용하여 데이터베이스를 구축하는 프로젝트를 진행하였다. XML은 엘리먼트의 중첩된 구조 및 계층적 구조 표현이 쉽고 이러한 계층 구조에 대한 질의 언어가 지원되기 때문에 생명정보와 같은 복잡하고 계층적인 데이터 표현 및 질의에 적합하다. 현재까지 XML 기반 생명정보 마컵언어로는 BSML, BioML, CML과 ProML 등이 존재한다.

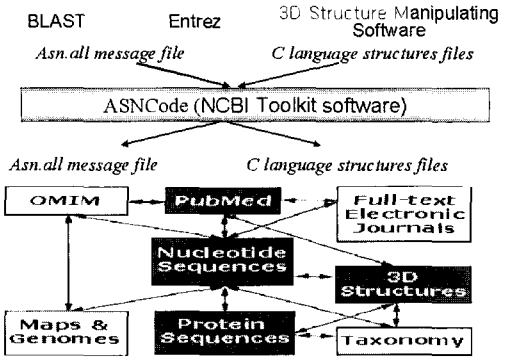
생명정보 관리를 위한 데이터베이스 시스템은 유연적인 데이터 모델, 복잡한 구조에 대한 표현력 있는 질의와 대용량의 데이터 처리 능력을 갖춘 데이터 관리 시스템이 요구된다.

2. 국외 생명정보 데이터베이스 및 관리 시스템

선진 외국에서의 생명정보 데이터 관리에 대한 연구 현황과 생명정보 데이터베이스 구축동향의 분석 내용과 국내의 생명정보 연구에 대한 동향을 기술한다.

• 미국 NCBI의 서열 데이터 관리

NCBI(National Center for Biotechnology Institute)는 유전체 서열 데이터베이스인 GenBank, EST 서열 데이터베이스 dbEST, 단백질 분자 구조 모델링 데이터베이스 MMDB, 인간 유전자 카탈로그와 유전적 변이에 대한 데이터베이스 OMIN, 문헌정보 데이터베이스 PUBMED를 유지하고 있다. 이러한 이질적인 생명 데이터를 통합하기 위해 데이터 사이에 링크가 연결되어 있다. 이러한 데이터베이스로부터 서열 및 유전자 데이터를 검색하기 위한 검색시



〈그림 4〉 NCBI 데이터베이스 및 분석 툴

스템으로 Entrez를 이용하고 서열 유사성 검색 시스템으로 BLAST를 제공한다.

NCBI는 1990년부터 NCBI의 공통 포맷[Oste01]으로 ASN.1 포맷을 이용하며 ASN.1은 단순한 파일포맷이기 보다는 생물학적 데이터를 표현하는 데이터 모델로서 이용된다. ASN.1을 이용하여 서열, 물리적인 지도, 계통 분류 정보, 분자정보와 문헌정보 등를 표현한다. ASN.1 메시지는 BIOSEQ, BIOSEQ-SET, Seq-aligns, PUBs 등이 있고 〈그림 5〉에서는 BIOSEQ에 대한 예를 표현하였다. NCBI에서는 서열 및 생물학적 데이터를 ASN.1 메시지 형태로 인코딩하고 데이터베이스에 저장하기 위해 〈그림 4〉같이 ASNCode 소프트웨어 툴킷을 개발하여 이용한다. 〈그림 4〉처럼 ASNCode는 BLAST 및 Entrz와 같은 검색 시스템과 3D 구조를 분석하는 프로그램에서 생물학 데이터를 이용할 수 있도록 ASN.1 메시지 파일을 C코드 파일로 또는 C 코드 파일을 ASN.1로 자유롭게 변환해준다. 그러나 ASN.1은 이기종간의 컴퓨터 시스템간의 데이터 교환을 위한 이진 언어이다. ASN.1의 사용을 위해서는 데이터를 파싱하고 인코딩하기 때문에 ASN.1 메세지를 위한 전용 파서가 요구되고 데이터에 대한 질의를 지원하지 않으며 확장성이 없다는 단점을 갖는다. 현재는 유전체 서열 데이터를 XML 포맷으로 표현한 BSML(Biopolymar Markup Language)을 이용한다.

유전체 서열 데이터베이스인 Genbank는 2002

```
Bioseq ::= SEQUENCE{
        id      SET OF Seq-id,
        descr   Seq-descr           OPTIONAL,
        inst    Seq-inst,
        annot   SET OF Seq-annot    OPTIONAL }
```

〈그림 5〉 NCBI에 서열정보의 ASN.1 표현

년 8월 18,197,000 서열 레코드를 포함하고 있으며 기하급수적으로 증가하고 있는 추세이다. 이렇게 기하급수적으로 증가하는 서열 및 관련 정보를 포함하는 데이터베이스의 서열 레코드를 식별할 수 있는 식별자가 요구된다. GenBank에서도 초기부터 지금까지 서열 레코드의 식별자를 변경하며 사용하고 있다. 초기 단계에는 서열 레코드와 서열의 기능 및 소스 생명체를 식별하기 위해 생명체의 이름과 유전자 이름을 혼합하여 사용하였다. 그러나 시간이 지남에 따라 유전적 locus 이름의 변경으로 인해서 식별자가 불안정해졌다. 따라서 실험자들이 NCBI에 서열을 제출할 때 Accession Number를 할당받도록 하고 서열 레코드의 식별자로 Accession Number를 사용하였다. Accession Number의 형식은 하나의 영문 대문자와 다섯 자리의 숫자로 구성된다. Accession Number는 Locus 이름보다 안정적이나 Accession으로 검색시 변경된 서열에 대한 검색을 할 수 없다. 따라서 NCBI에서는 서열 레코드의 갱신 시 변경된 서열 데이터도 식별하기 위해 GI(GenInfo Identifier)를 사용하게 되었다. GI는 서열이 GenBank의 ID(integrate DB) 통합 데이터베이스에 로딩될 때 서열에 할당되며 특정 서열이 갱신되면 동일한 Accession number를 가지며 새로운 GI를 할당 받는다. 이렇게 함으로써 서열 데이터의 변경에 대한 이력을 관리할 수 있다.

• 단백질 서열 데이터베이스

Protein Information Resource(PIR: http: // pir.georgetown.edu/)[Bar01]는 1965~1978년 동안 Margaret O. Dayhoff와 연구 구룹에 의해 연구된 결과로써 "Atlas of Protein Se-

```
ENTRY    T48678 #type complete iProClass View of T48678
TITLE    proteasome alpha-1 chain [validated] - Haloferax volcanii
COMPLEX heterodimer; alpha-1 and beta-1 (PIR:T48677) chain
         [validated, MUID:99412283]
FUNCTION #description the predominant peptide-hydrolyzing activity
         of the alpha (1)beta(1)-proteasome is cleavage carboxyl to
         hydrophobic residues [validated, MUID:99412283
```

〈그림 6〉 PIR 데이터베이스의 서열 상태 식별자

quence and Structure" 발표하고 1975년 단백질 서열 superfamily 개념 소개 및 데이터베이스로 조직화하여 현재까지 이어진다. PIR은 altas의 전신인 PIR-PSD(Protein Sequence Database), 중복을 제거한 PIR-NREF(Non-Redundant Reference Protein Database), 통합데이터베이스인 iProClass(a central point for exploration of protein information)를 운영하고 있다.

PIR은 같은 종의 생명체에 대해 서열의 중복을 제거한 데이터베이스로 대표적이다. 단백질 서열에 대한 전문적으로 주석화된 데이터베이스로 실험적으로 결정되거나 문헌적으로 검증 또는 전문적인 주석화된 데이터베이스로부터 얻어진 서열 정보를 이용하여 주석을 삽입한다. 〈그림 6〉처럼 중복을 제거하고 데이터베이스에 저장된 각 서열에 대해 검증된 상태를 나타내기 위해 title, function, complex 필드에 상태 식별자인 validated, similarity, import를 사용한다. validated는 해당 서열의 기능에 대해 실험적인 증명된 것이다. similarity는 유사성 관계에 있는 단백질의 기능을 해당 서열의 기능으로 할당된 경우에 해당된다. import는 서열의 이름이 GenBank, EMBL DDBJ에 의해 입수되었으나 PIR에 의해 증명되지 않은 것을 나타낸다.

• 단백질 구조 데이터베이스

Protein Data Bank[Bern00](PDB : http : // www.rcsb.org/pdb/)는 생물학적인 3차원 고분자 결정 구조를 위한 데이터베이스로서 1971년 부룩헤이번 국립 연구소(BNL)에 의해서 공개되었고 1980년대에 들어서 엑스레이 구조 결정 측정(X-ray crystal structure determination), 세포 자기 잔향(Nuclear Magnetic Resonance-NMR), 저온 전자 현미경(cryoelectron microscopy) 및 이론적 모델링(Theoretical Modelling) 등의 방법을 이용하여 단백질 고분자 결정구조를 빠르게 결정함으로써 고분자 결정 구조의 수가 증가하기 시작했다. 1990년대에 PDB의 3차원 단백질 구조 데이터를 웹을 통해서 접근이 가능하고 공개적으로 PDB 플랫파일과 mmCIF 형식으로 데이터를 배포한다. 또한 고분자 결정 구조 데이터로부터 유도된 서열정보를 FASTA형식으로 제공한다. PDB는 핵심 데이터베이스, 최종 데이터 파일 저장소, POM 데이터베이스, 생물학적 거대분자 결정데이터베이스(BMCD)와 Netscape LDAP 서버와 같은 5개의 컴포넌트로 구성된다. 각 컴포넌트에 대한 상세한 설명은 아래와 같다.

- 핵심데이터베이스(Core Relational Database) : sybase(Sybase SQLserver release 11.0)는 주요한 실험 및 좌표데이터를 제공하는 중앙물리장소
- 최종 데이터 파일 저장소 : 압축된 PDB와 mmCIF 형식파일을 FTP 서버에서 제공
- POM(Property Object Model) : 원자좌표와 유도속성(계산된 2차 구조 및 속성 프로파일)을 포함한 객체에 대한 인덱스를 포함한 데이터베이스
- Netscape LDAP 서버 : PDB의 문서 내용에 인덱스를 만들고 키워드 검색 지원

• 생물학적 거대분자 결정 데이터베이스 : 거대분자, 결정과 요약 데이터와 같은 3가지 분류에 대한 문헌정보를 포함

구조에 대한 검색은 VRML, RasMol과 Chime 프로그램을 이용하여 구조 이미지를 전시하고 상동성 관계의 구조의 부분을 검색할 수 있다. 현재 CORAB 인터페이스와 XML로 거대 분자 구조 표현을 위한 개발이 진행 중이다.

• 단백질 패밀리 데이터베이스

패밀리 데이터베이스는 서열의 특정 패턴을 검색하기 위한 데이터베이스이다. 이러한 데이터베이스로는 Prosite[Hofm99]와 Prints가 있다. Prosite(http : //www.expasy.ch/prosite/)는 스위스 생명정보학 연구소(SIB)에서 운영한다. PSSMs을 이용하고 1020 entries와 Swiss-Prot 서열에 기반한 1358 패턴을 보유하고 있다. 단백질 패밀리와 도메인 데이터베이스로써 지금까지 밝혀진 모티프를 이용하여 새로운 단백질이 속하는 패밀리나 도메인을 찾아내는 것이 가능하다.

Prints(http: //www.bioinf.man.ac.uk/dbbrowser/PRINTS/)는 PSSMs을 이용하여 5701블럭 기반 서열에 대한 990 핑거프린트 엔트리를 보유한다. 핑거프린트는 단백질 패밀리를 특징짓는 모티프의 집합으로 모티프분석에 활용을 위한 패턴 혹은 행렬을 유도한다.

그 외에도 유도 모티프 데이터베이스와 클러스터링 모티프 데이터베이스가 존재한다. 유도 모티프 데이터베이스는 구축되어진 기존의 패밀리 또는 단백질 서열 데이터베이스의 데이터를 결합하여 새로운 단백질 패밀리 데이터베이스를 구축한다. BLOCKS과 Proclass가 이에 해당된다. 클러스터링 모티프 데이터베이스는 SwissProt/TrEMBL과 PIR에서 얻은 단백질 서열을 검사하여 유사한 서열을 단백질 패밀리로 자동적으로 클러스터링한 데이터베이스이다. ProDom, DOMO와 Protomap이 해당된다.

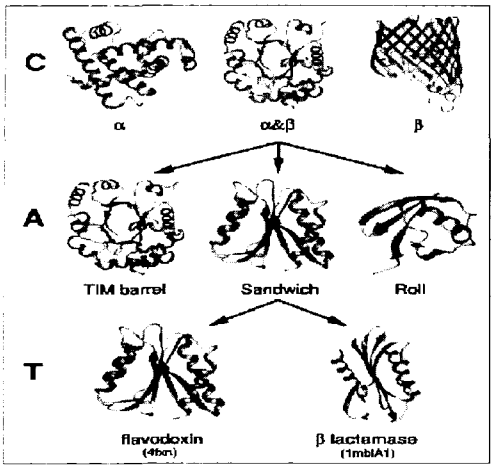
• 단백질 구조 분류 데이터베이스

단백질 3차원 구조는 진화 과정에서 서열보다 진화적 관계가 더 잘 보존되므로 단백질 구조 데이터의 상동관계를 식별해서 진화적 관계를 유추할 수 있다. 따라서 공통된 조상으로부터 진화되어 구조 및 기능의 상동성을 가지는 단백질들을 그룹화하여 분류 데이터베이스를 구축한다. 이러한 분류데이터베이스는 SCOP(Structural Classification Of ProteinsSCOP)[Murz95], CATH (Class Architecture Topology Homology : CATH)[Oren97]와 FSSP(Families of Structurally similar proteins)[Holm96]가 있다.

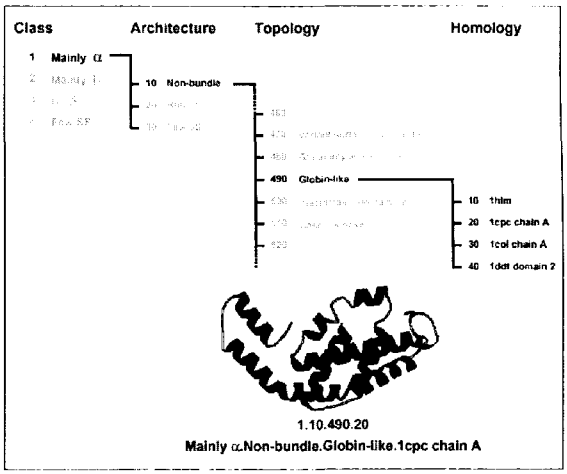
SCOP은 서열 유사성 검색에 기초를 두고 구조 비교와 수작업 조사에 의해 단백질 구조를 군집화하였다. CATH는 반자동 구조 분류시스템으로 단백질 구조 비교 프로그램인 SSAP를 구조 분류에 이용하였다. FSSP(Families of Structurally similar proteins)는 DALI 구조 비교 시스템을 이용하여 완전 자동화된 분류 시스템이다.

〈그림 7〉처럼 CATH의 계층 분류는 단백질 클래스(Class : C-Level), 구조(Architecture : A-Level), 위상(Topology : T-level), 상동성 슈퍼패밀리(Homologous superfamily : H-Level)와 서열 패밀리(Sequence Family : S-Level)로 구성된다. 단백질 클래스(C) 계층은 2차 구조를 이루는 α-helixes와 β-sheet 내용의 유사성 있는 그룹을 형성한다. 〈그림 8〉은 α-helix 2차 구조를 가지며 helix 구조가 non-bundle을 형성하고 globin-like 패밀리에 속하는 단백질 1cpc의 chainA에 해당하는 도메인 구조를 보여준다.

분류데이터베이스는 PDB의 단백질 3차 구조를 도메인으로 분리하고 도메인 안에서 2차 구조 요소의 개략적인 구성, 배치, 위상적 연결에 의해서 클래스와 폴드로 분류하게 된다. 유사한 폴드와 기능을 가진 단백질들을 서열 유사성의 정도에 따라 superfamily와 family로 분류한다. 이렇게 단백질의 구조에 의해서 단백질을 계층적으로 분류한다. 분류 데이터베이스는 기능이 알려지지



〈그림 7〉 CATH 단백질 분류 계층 구조



〈그림 8〉 단백질 1cpc의 구조 계층

않은 단백질의 구조 예측에 이용되고 단백질 구조 분석 및 상동성 구조 추출에 활용된다. 또한 서열 검색에 대한 효율을 평가하는데 활용된다.

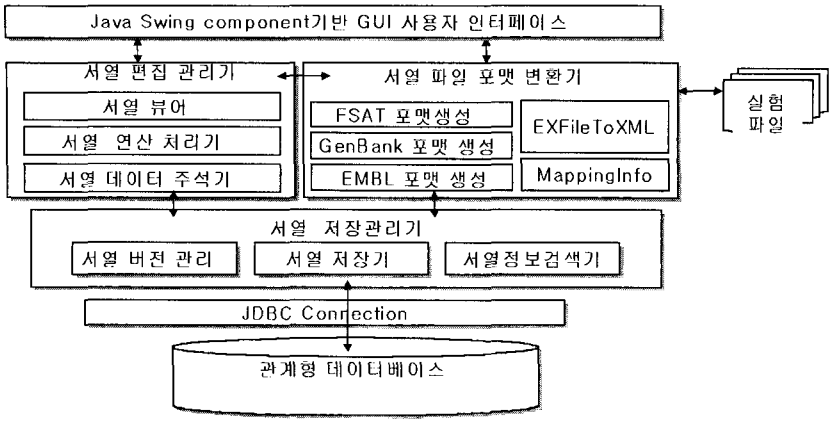
3. 국내 생명정보 데이터베이스

한국 정보 과학 기술 연구원과 대학을 중심으로 1999년부터 현재까지 유전체 염기서열, 단백질 서열 및 구조, 제한효소 및 신약 개발을 위한 화합물 데이터베이스를 구축하였다. 현재까지는 공개용으로 제공되는 플랫파일을 파싱하여 1차 데이터베이스를 구축에 주력하였지만 구축된 1차 데이터베이스를 분석 및 가공하여 유용한 2차 데이터베이스의 구축이 필요하다.

국내의 생명정보학은 기초 단계로 생물학자들에 의해 생산되는 유전체 및 단백질 서열은 국내에서 수집 저장되지 않고 외국의 데이터뱅크로 제출된다. 공개로 제공되는 유전체 관련 플랫파일에 대한 데이터베이스 구축과 분석용 소프트웨어 개발의 시작단계이다. 국내에서 생산되는 가치있는 유전체 및 단백질 데이터의 국내에서 자체적으로 수집, 저장, 분석, 관리를 위한 제반기술에 대한 연구가 요구된다.

지금부터는 아래 〈그림 9〉처럼 국내에서 개발된 서열 데이터 관리 시스템의 구조에 대해서 설명한다. 〈그림 9〉처럼 서열 정보 관리 시스템은 shotgun 시퀀싱 실험을 통해 얻어진 조각 서열을 어셈블리하여 얻은 일치 서열을 편집, 포맷변환 및 저장관리를 위한 시스템이다. 이 시스템은 XML을 기반하여 서열 데이터를 표현 및 연산하는 서열 편집 관리기와 서열 저장 관리기, 서열 포맷 변환기로 구성된다. 제안 시스템은 XML로 복잡하고 계층적인 서열관련 생물학적 지식 및 서열 데이터를 표현하고 서열 편집 및 검증을 위한 서열 연산을 이용하여 서열 데이터의 일치성을 유지한다. 또한 서열의 저장 관리를 위해서는 다차원 모델링 기법인 스타 스키마를 적용하여 서열 정보를 모델링한다. 동일한 DNA 조각에 대한 서로 다른 베이스 구성을 갖는 버전 서열을 정의하고 이를 관리할 수 있는 메커니즘의 제시와 이러한 메커니즘을 능동 데이터베이스의 능동 트리거 규칙을 이용하여 자동적으로 실행할 수 있다.

서열 포맷 변환기는 시퀀싱 실험을 통해 얻어진 실험 파일에 포함된 서열 정보에 대한 연산처리를 돕고 연산 결과를 생물학자가 원하는 포맷으로 저장한다. 서열 편집관리기는 생물학자가 유전체 서열에 대한 편집 및 조작시 서열 데이터의 무결성과 일치성을 지원하기 위한 편집을 위한 연산 처리, 서열 전시와 주석정보를 추가할 수 있는 모듈이다. 서열 저장관리기는 서열 데이터를 데이터베이스 저장하고 검색을 위한 모듈이다. 서열 저장관리기는 새로이 추가되는 서열정



〈그림 9〉 유전체 서열정보 관리 시스템 구조

보 뿐만 아니라 서열 버전에 대한 저장 및 검색도 함께 관리한다.

V. 생명정보 데이터베이스에 대한 연구 방향

후기 생명정보학 분야에서는 이질적인 생물 데이터를 통합 관리하기 위한 통합 데이터베이스 관리 시스템, 생물 데이터 분석을 위한 생명정보 데이터 마이닝 기법, 단백질 구조 예측을 위한 새로운 패턴의 추출 및 분류 기법, 생물학의 대전제인 염기서열로부터 단백질의 3차원 구조 예측을 위한 다양한 분석 방법에 대한 연구가 주류가 될 것이다.

이질적인 생명정보 시스템 통합을 위해서 XML 기술적용은 필수적이다. XML은 응용프로그램과 분산 환경에서 플랫폼 독립적으로 데이터의 상호 운용하고 다른 데이터 포맷과 교환을 위한 상호 교환성을 제공한다. 또한 생물학적 데이터와 같이 복잡하고 계층적 구조 데이터에 대한 표현과 이에 대한 질의를 지원하며 서로 데이터간의 링크를 위한 메커니즘을 지원할 수 있는 장점을 갖는다. Spitzner에 의해 제시된 BSML은 97년부터 NHGRI(National Human Genome Research Institute)에서 유전체 연구 정보를 교환하기 위해 개발된 생물 서열 마컵 언어이다. 현재는 생물 정보 표현을 위한 표준언어로써 이용되며 고유한 생물학 데이터 표현 포맷과 BSML간의 변환기를 개발하여 사용되고 있으며 대부분 온라인 생물학 데이터 저장소에서도 변환계층을 두어 BSML을 지원한다. 현재는 실제로 GenBank, EMBL, DDBJ, Swiss-Prot 같은 유전체 데이터베이스에서도 NCBIBLAST, Clustal multiple alignment와 같은 분석 도구들 사이에서 데이터 변환을 위해 BSML 변화기를 개발하여 이용한다.

또한 XML기술은 인터넷을 통한 이질적인 시스템간의 데이터 교환 및 공유와 XML 기반 응용 프로그램 개발을 위한 공통 인터페이스 및 질의 언어와 같은 개방형 프레임워크을 제공한다는 커다란 장점은 갖는다. 따라서 이러한 XML 기술의 장점은 계속해서 생명정보학 분야에서 XML의 활용을 가속화할 것이다.

비정형의 생명정보 분석 연구를 위해서는 비정형의 다차원 데이터 분석을 위한 대표적인 마이닝 기술인 시공간 데이터마이닝 기술, 제약 사항 기반 및 Outliner 마이닝에 대한 연구를 적용할 수 있다.

데이터베이스에서는 생명정보 분석을 위해서 비정형 데이터 타입에 대한 질의를 지원해야 한다. 이러한 질의에 대한 유형은 다음과 같다.

• 유사성 검색 질의 : 진화적 관계를 고려한 서

열, 단백질 구조, 신진대사의 유사성 질의
- 패턴 질의 : exon, intron, 유전자, 유전자 조절 패턴에 대한 질의
- 대략적 패턴 질의 : Hidden Markov 모델과 같은 확률 패턴 모델에 대한 질의
- 그래프 질의 : 화학적 그래프 구조상의 멱등한 하위 그래프 검색

제시된 유형의 질의를 지원하기 위한 생명정보 데이터베이스에 대한 연구는 크게 두 가지의 접근 방법으로 나누어 볼 수 있다. 첫째는 생명정보 도메인에 적합한 새로운 확장된 DBMS의 개발에 대한 연구이다. 다른 방법은 기존의 DBMS에 관련된 기술을 적용하여 생명정보 데이터를 관리하는 것이다.

새로운 확장된 DBMS의 개발은 II절에서 살펴본 바와 같이 생명정보 데이터의 특성에 적합한 새로운 데이터 타입과 연산자 정의, 인덱스 기법의 개발과 데이터 타입에 대한 질의 처리 최적화 기법의 개발을 포함한다. 데이터 타입의 정의는 서열, 유전자, 주석, 염색체에 대한 속성과 연산을 정의해야 한다. 즉, 서열의 크기나 길이에 대한 정보와 서열의 집합인 염색체를 생성할 수 있어야한다. 연산자의 정의는 특정한 패턴을 포함한 서열을 검색할 수 있는 연산자나 서열간의 진화적 관계의 유사성을 나타낼 수 있는 연산자 등을 SQL에서 사용할 수 있도록 정의해야 한다. 생명정보 데이터에 대한 인덱스는 3차원 구조 비교나 서열 유사성 검색, 하위 분자 구조의 검색 및 구조 사이의 거리에 대한 평가를 실행 시 효율적으로 실행할 수 있어야 한다. 질의 최적화 기법의 개발을 위해서는 질의에 사용된 데이터 타입과 연산자에 대해 질의 수행 계획을 세울 수 있도록 비용함수를 정의하고 이러한 비용함수에서 사용되는 데이터에 대한 통계적 정보를 새롭게 정의해야 한다.

생명정보 데이터 관리를 위한 다른 접근 방법으로써 기존의 비정형 데이터를 관리하는 데이터베이스 기술을 생명정보 데이터에 적용 및 확장하는 것이다. 이러한 적용이 가능한 기술로서는 XML과 같은 반 구조 데이터 관리를 위한 반 구조 데이터 관리 기법 및 XML 기술과 시간, 공간 및 시공간 데이터베이스 이론 등이 있다. 특히 단백질 3차 구조 같이 3차원 좌표를 가지며 공간적 속성을 갖는 생명정보 데이터에 공간 데이터베이스 기술을 적용한 연구가 학계에 발표되고 있다.

V. 결 론

이 원고에서는 갑자기 증가된 생명정보 데이터의 종류와 비정형적인 특성을 살펴보았다. 또한 생명정보 데이터 분석 및 관리를 위해 국외 생명정보 데이터베이스 구축 동향과 국내에서 개발된 서열 유전체 정보 관리시스템의 사례를 기술하였다. 또한 국내외적인 연구 동향을 바탕으로 생명정보 데이터 분석 및 관리를 위한 현재 당면하고 있는 문제점과 이것을 해결하기 위한 연구 방향을 제시하였다.

후기 유전체 시대의 핵심적인 학문이 될 생명정보학 연구는 유전체 서열 분석뿐만 아니라 구조 생물학, 유전체학, 유전자 발현 연구 등과 같은 폭넓은 연구 분야를 포함하고 있다. 이러한 연구 분야의 발전과 생명 과학적 지식 베이스 구축을 위해서는 이 원고에서 기술된 내용과 같이 생명정보 데이터의 관리 및 분석 기술의 확보가 국가적 차원에서 무엇보다도 중요하다.

이를 위해서는 생명과학 관련 학문 분야와 컴퓨터과학의 학문 융합을 통한 기술 공조가 이루어져야 한다. 또한 생명정보 데이터 분석 연구를 위해서는 컴퓨터 기술뿐만 아니라 데이터에 대한 생물학적 이해가 요구되는 어려운 분야이므로 고급 연구 인력의 확보 또한 시급하다.

참 고 문 헌

〔Bark01〕
W. C. Barker, J. S. Garavelli, Z. Hou, H.

Huang, R. S. Ledley, P. B. McGarvey, H. W. Mewes, B. C. Orcutt, F. Pfeiffer, A. Tsugita, C. R. Vinayaka, C. Xiao, L. L. Yeh, C. Wu, Protein Information Resource : a community resource for expert annotation of protein data, Nucleic Acids Research Vol.29, No. 1, 2001.

〔Davi01〕
David W. Mount, "Bioinformatics : Sequence and Genome Analysis" Cold Spring Harbor Laboratory Press, 2001.

〔Mark01〕
Victor M. Markowitz, I-Min A. Chen, Anthony S. Kosky, Ernest Szeto. OPM : Object-Protocol Model Data Management Tools'97. Data Management Research and Development Group Lawrence Berkeley National Laboratory, Berkely, 2001.

〔Oste01〕
J. Ostell, Wheelan, S. J., Kans, J. A. The NCBI data model. Chapter 2 in Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins, 2nd ed., edited by Baxevanis, A. D. and Ouellette, B. F. F. New York : John Wiley & Sons, pp.19-43. 2001

〔Stoe01〕
G. Stoesser, W. Baker, A. V. D Broek, E. Camon, M. Garcia-Pastor, C. Kanz, T. Kulikova, V. Lombard, R. Lopez, H. Parkinson, N. Redaschi, P. Sterk, P. Stoehr, M. Ann T., "The EMBL nucleotide sequence database", Nucl. Acids. Res. 29 : 17-21, 2001.

〔Gers00〕
M. Gerstein. Integrative database analysis in structural genomics, nature structural biology, structural genomics supplement,

〔Bair00〕
A. Bairoch, R. Apweiler, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000", Nucl. Acids. Res. vol. 28 : 45-48, 2000.

〔Bern00〕
H. M. Berman, J. W. Brook, Z. F., Gray G, T. N. Bhat, H. Weissing, I. Shindyalov, P. E. Bourne, "The Protein Data Bank", Nucleic Acids Research, Vol28. No1, 2000.

〔Marc99〕
Marchler-Bauer, A., Addess, K. J., Chappey, C., Geer, L., Madej, T., Matsuo, Y., et al. (1999). Nucleic Acids Res., 27, 240.

〔Hofm99〕
K Hofmann, P Bucher, L Falquet, A Bairoch, "The PROSITE database, its status in 1999", Nucl. Acids. Res. vol. 27 : 215-219, 1999.

〔Scot99〕
Scott AF, Amberger JA, Brylawski B, McKusick VA/SI Letovsky, ed : Online Mendelian Inheritance in Man. Kluwer Press, 1999.

〔Coop98〕
Cooper DN, Ball EV, Krawczak M., The human gene mutation database. Nucleic Acids Res. 1 ; 26(1) : 285-7, 1998.

〔Kane97〕
M. Kanehisa, KEGG : From Genes to Biochemical pathways, Bioinformatics : Database and Systems, stanley letovsky, Kluwer acade, oc publishers, 1999.

〔Oren97〕
Orengo C. A., Michie A. D., Jones D. T.,

Swindells M. B., Thornton J. M, CATH- A hierachic classification of protein domain structures, Structures(5) 1093-1108, 1997.

〔Tate97〕

Y Tateno and T Gojobori, "DNA Data Bank of Japan in the age of information biology", Nucl. Acids. Res. vol. 25 : 14-17, 1997.

〔Etzo96〕

Etzold, T., Ulyanov, A., and Argos, P. SRS : Information Retrieval System for Molecular Biology Data Banks, Methods in Enzymology v. 266, 1996, p.144.

〔Holm96〕

Holm L. Sander C., The FSSP database : fold classification based on structure-structure alignment of proteins, Nucleic Acids Res. Vol. 24, pp206-209, 1996.

〔Murz95〕

Murzin A. G., Brenner S. E., Hubbard T., Chothia C., SCOP : A structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. (247) : 536-540, 1995.

저 자 소 개

박 성 희

1996년 8월 충북대학교 도시공학과 졸업(공학사), 2001년 2월 충북대학교 대학원 전자계산학과 석사(이학석사), 2003년 2월 충북대학교 대학원 전자계산학과 박사과정 수료, 1998년 1월~1998년 12월 : 한국전자통신연구원 컴퓨터소프트웨어 연구소 위촉 연구원, 〈주관심 분야 : Bioinformatics, 반구조 및 XML 데이터베이스, 시공간데이터베이스 등〉

류 근 호

1976년 2월 숭실대학교 전산학과 졸업(이학사), 1980년 2월 연세대학교 공학 대학원 전산전공(공학석사), 1988년 2월 연세대학교 대학원 전산전공(공학박사), 1976년~1980년 : 육군 군수 지원사 전산실(ROTC장교), 1980년~1983년 : 한국전자통신연구소 연구원, 1983년~1986년 한국 방송대학교 전산학과 조교수, 1989년~1991년 : University of Arizona, Research Staff (TempIS 연구원, Temporal DB), 1986년~현재 : 충북대학교 전기전자컴퓨터공학부 교수, 〈주관심 분야 : 시간데이터베이스, 시공간데이터베이스, Temporal GIS, 데이터마이닝, Bioinformatics〉