

## The Development of Meta-Information System for Microbial Genome Resources

### 미생물 게놈자원을 위한 메타정보 시스템의 개발

Wonhyong Chung<sup>1</sup>, Jaewoo Yu<sup>2</sup>, Tae-Kwon Sohn<sup>1</sup>, Yong-Ha Park<sup>1,2</sup> and Hongik Kim<sup>1,\*</sup>

<sup>1</sup> proBionic Corp., Daejeon 305-333, Korea

<sup>2</sup> Biological Resource Center, Korea Research Institute of Bioscience and Biotechnology(KRIBB), Daejeon 305-333, Korea

\*To whom correspondence should be addressed. E-mail: hikim@probionic.com

#### Abstract

There are currently about 6000 bacterial species with validly published names, but scientists assume that these may be less than 1% of bacterial species present on the earth. Microbial resource is one of the most important bioresources in bioindustry and provides us with high economic values. To find missing ones, the studies of metagenome, metabolome, and proteome about microbes have started recently in developed countries. We construct the information system that integrates information on microbial genome resources and manages the information to support efficient research of microbial genome application, and name this system 'Bio-Meta Information System (Bio-MIS)'. Bio-MIS consists of integrated microbial genome resources database, microbial genome resources input system, integrated microbial genome resources search engine, microbial resources on-line distribution system, portal service and management via internet. In the future, we will include public database connection and implement useful bioinformatics software for analyzing microbial genome resources. The web-site is accessible at <http://biomis.probionic.com>

#### Introduction

미생물은 현재 약 6천종 정도로만 알려져 있지만 발견된 미생물이 전체 미생물 중의 1% 미만으로 추정되고 있으므로 나머지 약 99%는 아직까지 미발견 상태로 남아있다. 현재의 기술로는 대부분이 배양 불가능한 상태이지만 해양으로부터 지하까지 지구 곳곳에 막대한 양의 세균이 분포하며 종 분화와 기능적 다양화 및 유전적 다양성을 유지하고 있다. 그 동안 생명공학자원으로 활용되어온 미생물들은 전체 미생물의 1~10%에 지나지 않는 배양이 된 것들을 이었으나 앞에서 밝혔듯이 절대 다수를 차지하는 배양되지 못한 미생물들의 다양성을 이용할 수 있다면 학문적, 산업적으로 엄청난 효과를 가져올 수 있다. 따라서 막대한 종류의 알

W. Chung and J. Yu contributed equally to this work.

러지지 않은 미생물들에 대한 체계적인 연구와 이용방법을 개발하여 고부가가치의 자원으로 미생물자원을 확보하려는 노력이 최근 선진국을 중심으로 활발히 이루어지고 있고[1,2], 국내에서도 한국생명공학연구원의 미생물유전체연구단에 의하여 미생물, 메타게놈, 메타볼롬, 및 프로테오믹스에 대한 종합적인 연구 및 지원을 하고 있다. 본 논문에서는 미생물, 메타게놈, 메타볼롬, 프로테오믹스를 합쳐서 미생물유전체자원(Microbial genome resources)이라 부르기로 한다.

### Metagenome

자연계로부터 배양을 거치지 않고 직접 미생물의 유전체를 추출하는 방법이 최근 개발되고 있고 이렇게 추출된 유전체들이 다양한 미생물의 유전체가 섞여있다는 의미에서 메타게놈(Metagenome)이라고 한다. 1998년 Handelsman과 Goodman은 BAC(Bacterial Artificial Chromosome)기법을 Soil metagenome에 사용하여 신약탐색에 응용함으로써 이론에만 머물던 메타게놈을 이용한 연구를 실제로 적용되게 되었고[3] 이후 많은 연구자들에게 주목받게 되었다.

### Metabolome

세포 안에는 단백질(효소를 포함) 이외에 효소에 의해 분해된 대사물이 함유되어 있는데 이들 대사물의 총합을 메타볼롬(Metabolome)이라 하고 생명활동에 있어서 이들의 변동이 중요한 역할을 하고 있다. 메타볼롬에 대한 연구는 아직 초보적인 수준이지만 선진국을 중심으로 데이터베이스의 구축이 활성화 되고 있다. Kyoto 대학의 유전체 연구센터에서 구축하고 있는 KEGG(Kyoto Encyclopedia of Genes and

Genome: [www.genome.ad.jp/kegg](http://www.genome.ad.jp/kegg))와 미국 시카고의 Genomic Inc.에서 구축하고 있는 WIT (<http://wit.mcs.anl.gov/WIT2>)등에서 이러한 연구결과를 데이터베이스화 하고 있다[4].

### Proteome

세포, 또는 개체에서 특정 시간에 발현되는 단백질의 총합을 프로테오믹스(Proteome)이라 하며 유전자의 기능을 밝히는 가장 직접적인 수단으로 활용된다.[5] 현재 유전자의 세포 내 기능을 완전하게 이해하기 위해서는 지노믹스(Genomics)에서 접근할 수 없는 유전자 발현 이후의 최종적으로 완성된 단백질의 생리적 변화를 프로티오믹스(Proteomics)의 방법으로 해결해야만 한다. 세포 내에서 합성되는 프로테오믹스를 추출하는 방법으로 현재 전기영동법(Electrophoresis)와 질량스펙트럼(Mass-spectrum)의 두 가지가 주로 이용되고 있으며[6] 단백질칩(Protein Chip)의 제작을 통한 대량분석방법도 시도되고 있다.

미생물유전체를 연구하는 동안 쏟아지는 방대한 양의 메타게놈, 메타볼롬, 그리고 프로테오믹스 자원은 그들간의 복잡한 연관관계를 내포하고 있지만 현재까지 이를 밝히고 체계적으로 서비스하려는 시도는 없었다. 이에 본 연구팀에서는 미생물유전체자원 통합관리 시스템인 Bio-MIS(Bio-Meta Information System)를 구축하여 미생물유전체 연구성과의 활용도를 획기적으로 높이려고 한다. Bio-MIS는 미생물, 메타게놈, 메타볼롬, 그리고 프로테오믹스의 네가지 종류의 자원에 대한 정보를 데이터베이스로 구축하고 이를 관리 및 서비스하는 포털 시스템을 제공한다. 그리고 연구자들을 위하여 인터넷을 통한 입력시스템, 자원통합 검색시스템, 그리

고 온라인 자원분양 시스템을 개발하여 포함한다. 본 시스템은 미생물유전체자원을 수집하는 연구자와 이를 분석하는 연구자들의 통합서비스 창구의 역할을 함으로써 자원의 통합관리와 원활한 자원 및 실험정보의 공유를 제공할 수 있다. 현재 본 시스템은 통합 데이터베이스와 입력시스템을 구현하여 <http://biomis.probionic.com>으로 서비스하고 있다.

### System Organization

Bio-MIS는 미생물유전체자원 통합 데이터베이스 시스템을 기반으로 미생물, 메타게놈, 메타볼롬, 그리고 프로테오믹 데이터베이스를 구축하고 각각에 대한 입력시스템을 구축하였다. 수집된 데이터의 특성을 파악하여 데이터 필드간의 연관관계를 밝혀내 통합검색 엔진을 제작하고 검색된 자원데이터는 온라인 분양시스템을 통하여 분양과정이 실시간으로 모니터링 된다. 그리고 모든 기능은 Bio-MIS 포털 사이트를 통하여 서비스되고 관리된다. 그림 1은 Bio-MIS의 시스템 구성

도를 보여준다.

### Integrated database system

미생물유전체자원 통합 데이터베이스는 연구자들이 필요로 하는 데이터구조를 제공할 뿐만 아니라 변경되는 구조를 수용할 수 있도록 데이터구조에 관한 메타정보를 통하여 자동으로 데이터베이스가 관리되도록 설계되었다. 본 데이터베이스는 미생물, 메타게놈, 메타볼롬, 그리고 프로테오믹 데이터베이스를 각각 구축하여 데이터의 입출력을 처리하고 각 데이터베이스의 데이터필드 사이의 연관관계를 결정하여 효율적인 통합검색이 이루어지도록 한다.

### Microbial genome resources input system

연구자들이 자신의 연구결과를 편리하게 입력할 수 있는 사용자환경을 제공한다. 이를 위하여 미생물유전체 분야에서 종사하는 연구자들에게 자문 및 검증을 받아 필요한 입력필드를 정의하고 데이터베이스에 적용하였다. 입력시스템에서 입력클라이언트 부분

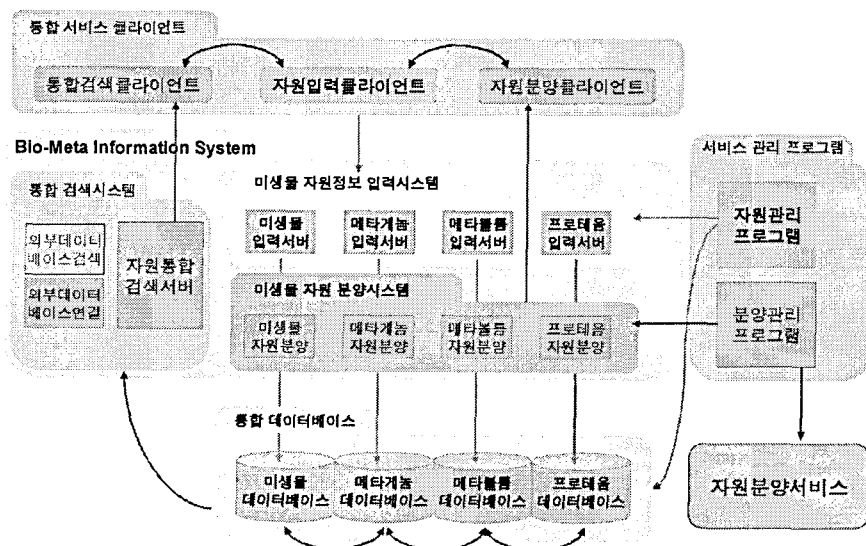


그림 1. 미생물유전체자원 통합 서비스 시스템

은 연구자에게 정보를 입력 받아 서버에 전달하고 입력서버는 정보를 해석하여 데이터베이스에 저장하는 역할을 한다.

#### **Integrated search engine**

미생물유전체자원끼리의 연관관계를 조사하여 효율적인 탐색구조를 구축하고 이를 바탕으로 통합검색서비스를 제공한다. 데이터의 연관관계에 관한 전문적인 지식이 없어도 효율적인 검색을 할 수 있도록 통합검색 클라이언트에서는 논리적 연관관계에 대한 질의를 수행하고 통합검색서버에서는 이 질의를 실제 데이터필드간의 연관관계로 해석하여 검색한다. 그리고 통합검색의 결과는 중요도를 선정하여 출력함으로써 연구자가 방대한 결과 중 유용한 정보를 효율적으로 검색할 수 있도록 한다.

#### **Microbial resources on-line distribution system**

등록된 미생물유전체자원의 보존상태 및 분양정보에 대한 검색과 웹을 통한 온라인 분양신청 절차를 확립하고 분양신청이 담당자에게 이메일 등의 방법으로 실시간으로 통보가 되도록 한다. 신청자는 분양처리과정을 온라인으로 검색할 수 있고 각 과정의 완료를 이메일 등의 방법으로 통보받는다. 분양처리가 완료된 후 분양기록을 보존하여 분양신청자 등이 온라인으로 확인할 수 있도록 한다.

#### **Portal service and Management**

미생물유전체 자원관리 시스템 Bio-MIS는 자원입력, 통합 검색, 분양 등의 기능을 서비스하는 포털로서 연구자에게 웹 기반 사용자환경을 제공한다. 이 포털사이트는 통합데이터베이스와 입력시스템 등을 엮어주

는 서버의 역할, 관리자인증을 통한 자원관리 프로그램과의 연결, 및 회원관리기능의 구현으로 정상적인 사용권한을 가진 사용자만 이용가능 하게하는 기능을 제공한다.

#### **Implementations**

Bio-MIS는 현재 포털사이트, 자원관리프로그램, 자원 입력시스템, 통합 데이터베이스가 구현되었고 통합검색엔진, 온라인 분양시스템에 관한 연구를 진행 중이다. 아래에서는 구현된 내용에 대하여 언급한다.

#### **Integrated microbial genome resources database**

본 데이터베이스는 데이터타입에 대한 메타정보와 데이터 테이블의 구조에 대한 메타정보를 통하여 실제 데이터들을 관리하고 이들 메타정보의 수정으로 데이터베이스의 구조가 동적으로 조절된다. 그리고 3계층 데이터베이스 모델을 적용하여 입력시스템 등에 해당하는 상위계층과 데이터를 저장하는 하위계층 사이에 데이터를 추상화하는 중간계층을 두어 동적인 데이터 구조가 상위계층에서는 고정된 데이터 구조인 것처럼 보이도록 하였다.

주요 메타 데이터 테이블은 선택필드의 정보를 가지는 '데이터 선택 필드 리스트 아이템 테이블(XXX\_CASE)', 실제 데이터레코드가 저장되는 '데이터 저장 테이블(XXX\_DATA)', 데이터베이스의 테이블형태와 필드종류를 결정하는 '데이터 필드정의 및 운영정보 테이블(XXX\_FORM)', 그리고 데이터의 접근권한 등을 관리하는 '데이터 메인 허브 테이블(XXX\_MAIN)'의 네가지로 구성되어 있다. 미생물, 메타게놈, 메타볼롬, 프로테오믹스에 대하여 각각 주요 메타데이터

테이블이 구현되어 있고 그림 2(메타게놈의 데이터 스키마)와 같은 스키마(Schema)를 가진다.

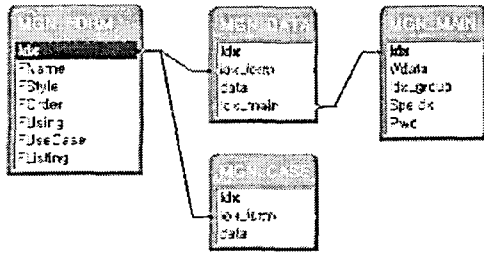


그림 2. 메타게놈 데이터 스키마

### Microbial genome resources input system

Bio-MIS는 미생물, 메타게놈, 메타볼롬, 프로테옴 각각에 대하여 입력 시스템이 구축되어 있으며 각 입력필드는 필수입력필드와 선택입력필드, 어떤 입력이 선택되었을 때 입력되어야 하는 조건선택입력필드 중 하나로 제작되어 사용자의 올바른 입력을 유도한다. 본 시스템은 미생물유전체정보 연구자들의 도움으로 미생물정보 데이터구조를

이름	포지션	순서	*필수구분	시스템필수	필수	의무
공통명	ENR	1	True	True	필수	의무
학명	ENR	2	True	False	필수	의무
분류명	ENR	3	True	True	필수	의무
분류군소	ENR	4	True	False	필수	의무
분류군과	ENR	5	True	False	필수	의무
분류군종	ENR	6	True	False	필수	의무
위양성	ENR	7	True	False	필수	의무
음성	ENR	8	True	False	필수	의무
Genome의 위치	Text	9	True	False	필수	의무
Genome의 크기	Text	10	True	False	필수	의무
Genome의 구조	Combo	11	True	False	필수	의무
Genome의 설명	Text	12	True	False	필수	의무
Genome의 출처	Text	13	True	False	필수	의무
Genome의 크기	Text	14	True	True	필수	의무

(a) 미생물정보 데이터구조

이름	포지션	순서	*필수구분	시스템필수	필수	의무
Genome ID	ENR	1	True	True	필수	의무
Genome Name	ENR	2	True	True	필수	의무
Genome Size	ENR	3	True	True	필수	의무
Genome GC	ENR	4	True	True	필수	의무
Genome GC	ENR	5	True	True	필수	의무
Genome GC	ENR	6	True	True	필수	의무
Genome GC	ENR	7	True	True	필수	의무
Genome GC	ENR	8	True	True	필수	의무
Genome GC	ENR	9	True	True	필수	의무

(c) 메타볼롬정보 데이터구조

14개의 필드(그림 3(a)참조), 메타게놈정보 데이터구조를 10개의 필드(그림 3(b)참조), 메타볼롬정보 데이터구조를 9개의 필드(그림 3(c)참조), 그리고 프로테옴정보 데이터구조를 10개의 필드(그림 3(d)참조)로 정의하고 데이터베이스의 데이터 필드정의 및 운영정보 테이블에 적용하였다.

### Bio-MIS portal site

본 시스템은 미생물유전체자원에 대한 자원 정보 입력, 통합검색, 온라인분양 등의 사용자환경을 제공하는 포털사이트를 Bio-MIS란 이름으로 제공한다. 이것은 Bio-MIS의 목적과 활동방향을 소개하는 '소개', 미생물유전체자원을 검색하는 '유전체자원검색', 미생물유전체자원을 입력하는 '자원등록', 자원분양 등의 서비스를 제공하는 '지원서비스', 그리고 관련자료를 제공하는 '자료실'의 다섯 가지의 메인 메뉴를 가진다. 회원관리기능의 구현으로 자료에 대한 검색 및 입력에 대한 권한을 개인마다 지정할 수 있고, 이

이름	포지션	순서	*필수구분	시스템필수	필수	의무
공통명	ENR	1	True	True	필수	의무
분류명	ENR	2	True	False	필수	의무
분류군소	ENR	3	True	False	필수	의무
분류군과	ENR	4	True	False	필수	의무
분류군종	ENR	5	True	False	필수	의무
분류군과	ENR	6	True	False	필수	의무
분류군종	ENR	7	True	False	필수	의무
분류군과	ENR	8	True	False	필수	의무
분류군종	ENR	9	True	False	필수	의무
분류군과	ENR	10	True	True	필수	의무

(b) 메타게놈정보 데이터구조

이름	포지션	순서	*필수구분	시스템필수	필수	의무
공통명	ENR	1	True	True	필수	의무
Genome ID	ENR	2	True	True	필수	의무
Genome Name	ENR	3	True	True	필수	의무
Genome Size	ENR	4	True	True	필수	의무
Genome GC	ENR	5	True	True	필수	의무
Genome GC	ENR	6	True	True	필수	의무
Genome GC	ENR	7	True	True	필수	의무
Genome GC	ENR	8	True	True	필수	의무
Genome GC	ENR	9	True	True	필수	의무
Genome GC	ENR	10	True	True	필수	의무

(d) 프로테옴정보 데이터구조

그림 3. 입력시스템에 사용된 데이터구조

외에도 EVENT, NEWS, 공지사항, 자료실, FAQ, QNA, Link를 제공한다.

Bio-MIS에서 관리자로 로그인 할 경우 자원 관리 프로그램에 접근할 수 있다. 이것은 통합 데이터베이스에 구현된 동적 데이터구조를 수정하는 기능과 Bio-MIS의 공지사항, 뉴스, 이벤트 등을 추가, 수정, 삭제하는 포탈관리 기능까지 제공한다.

### Conclusion and Future works

산업적으로 막대한 잠재가치를 가지는 미지의 미생물 및 생물자원의 획득에 사용되는 메타게놈, 메타볼롬, 및 프로테옴은 대량의 데이터를 포함하므로 이를 적절히 관리할 수 있는 시스템이 필요해졌다. 본 연구진은 이들 미생물유전체자원의 통합관리 및 입력, 분양시스템으로 Bio-MIS (Bio-Meta Information System)을 개발하였다. 본 시스템은 미생물유전체자원의 효율적인 관리와 함께 자원 사이의 연관관계를 파악하여 연구자들에게 제공함으로써 연구활동 및 상호간의 연구협력이 극대화되도록 설계되었다. 새롭게 등록되는 메타게놈, 메타볼롬, 및 프로테옴 자원은 해당분야의 연구자들이 바로 검색하여 유용성을 판단하고 온라인 자원분양시스템을 통하여 제공받아 신속한 연구에 도움을 줄 수 있다.

지금까지의 진행상황은 미생물유전체자원 통합 데이터베이스와 자원입력시스템, 이들 서비스를 연구자들에게 제공하기 위한 포탈 서비스 및 관리프로그램을 구현한 단계이다. 현재 본 연구진은 통합 데이터베이스의 업그레이드, 통합검색 시스템, 및 온라인 자원분양시스템을 구현하기 위하여 연구 중에 있다. 앞으로는 GenBank와 같은 공용데이터베이스를 이용하여 더 풍부한 정보를 제공

하고 유전체정보 분석알고리즘을 구현하여 통합함으로써 데이터수집, 관리에서부터 분석에 이르기까지의 모든 서비스를 제공하도록 확장시켜 나갈 계획이다.

### Acknowledgements

This work is supported by a grant (MG02-0501-001-0-0-0) from the 21C Frontier R&D Programs of the Ministry of Science and Technology of Korea.

### References

- [1] Yagisawa, M., Trends in exploratory research of microbial metabolite part 1; 50 years from the discovery of penicillin, *바이오사이언스 & 인더스트리*, 2000, 58(1).
- [2] Yagisawa, M., Trends in exploratory research of microbial metabolite part 2; 20 years post-antibiotics, *바이오사이언스 & 인더스트리*, 2000, 58(2).
- [3] Handelsman J, Rondon MR, Brady SF, Clardy J, and Goodman RM., Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products, *Chem Biol.* 1998 Oct;5(10):R245-9.
- [4] Alan H. Fairlamb, Metabolic pathway analysis in trypanosomes and malaria parasites, *Phil. Trans. R. Soc. Lond. B*(2002) 357, 101-107.
- [5] Kahn P., From genome to proteome: looking at a cell's proteins, *Science*, 1995 Oct 20;270(5235):369-70.
- [6] Tsugita A. and Kawakami T., Proteome analysis: the state of the art of the methodology, *Nippon Hoigaku Zasshi*, 1999 Jun;53(2):191-8.