

# 해부학적 구조를 이용한 유전자 정보 통합 기법

양기철

목포대학교, 멀티미디어공학과

## Integration Scheme of Gene Information based on Anatomical Structure

Gi-Chul Yang

Department of Multimedia Engineering, Mokpo National University

**요약** 생물학자들은 특정 암이나 선천성 질병을 이해하는데 핵심정보를 제공할 수 있는 유전자관련 연구를 진행하고 있다. 하지만 생물학적 실험은 실험당시의 여러 가지 요소나 상황의 차이 또는 해석의 차이에 의해 서로 다른 결과를 생성하기도 한다. 따라서 현존하는 연구 결과들은 서로 상이한 정보를 제공할 수 있다. 유전자 정보의 통합을 통하여 이러한 불일치를 찾을 수 있다. 유전자 정보들이 불일치가 없이 통합 된다면 생물학자들은 어떤 유전자 정보를 알기 위해서 여러 연구 결과를 검토하지 않아도 되어 시간과 노력을 절감할 수 있게 된다. 이를 위하여 본 논문에서는 서로 다른 연구에 의해 구축된 유전자 정보를 하나의 정보로 통합 및 확장하는 기법을 소개한다.

**주제어** : 유전자 정보, 유전정보통합, 해부학적 온톨로지, 생물정보학

**Abstract** Biologists are pursuing genetics related researches that can provide the core information to understand a certain cancer or inherent diseases. However, biological experimentations can produce different results by the difference of various elements or environments at the time of experimentation and/or difference of interpretations. Therefore, currently existing research results can possibly provide different information. These inconsistency can be found through integration of gene information. Biologists can save their time and efforts to find certain gene information if the gene information is integrated without inconsistency. An efficient gene integration and augmentation scheme of gene information generated through different researches is introduced in this paper.

**Key Words** : Gene Information, Genetic Information Integration, Anatomical Ontology, Bioinformatics

### 1. 서론

세계 여러 곳의 생물학자들은 배아의 발달에 있어서 어떤 유전자가 어떻게 영향을 미치는지를 연구하고 있다.

이러한 연구는 특정 암이나 선천성 질병을 이해하는데 핵심정보를 제공할 수 있는 흥미롭고 중요한 연구이다.

실험에 의하면 유전자는 각각의 3차원 패턴으로 발현한다. 이러한 3차원 패턴은 문장으로 기술하기 어렵고,

Received 15 November 2014, Revised 16 December 2014

Accepted 20 February 2015

Corresponding Author: Gi-Chul Yang

(Mokpo National University)

Email: gcyang-umkc@hanmail.net

ISSN: 1738-1916

© The Society of Digital Policy & Management. All rights reserved. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

실험결과 어떤 유전자가 어떻게 발현 했는지 찾기가 어렵다. 이러한 문제를 해결하기 위해 연구자들은 유전자 활동의 패턴을 저장하여 쉽게 찾아볼 수 있도록 데이터 베이스를 만든다.

이를 위하여 배아를 각 발전 단계별로 나누고 단계별 3차원 이미지를 만든 다음 각각의 유전자와 그 발현 패턴을 연결시켜 보여줄 수 있도록 한다. 또한 생물학자들은 배아의 영상뿐만 아니라 해부학적 명칭을 사용하기를 원하기도 한다. 따라서 배아의 각 부분에 이름을 붙이고 이를 배아 영상에 일치시키는 작업도 수행한다. 이를 위해서는 각 단계별 배아를 자르고 사진을 찍어 재조합하는 과정을 거치고 이를 3차원 형태로 컴퓨터에 저장하여야 한다.

또한 텍스트 데이터를 데이터베이스에 저장하기 위하여 해부학적 이름을 정하고 3차원 영상에 부위별로 이름을 연결하여 저장한다. 이러한 해부학적 이름은 TREE형태로 조직되는데, 이 TREE는 PART-OF 계층구조를 형성하여 해부학적 구조를 쉽게 설명할 수 있도록 한다. 이것은 일종의 해부학적 온톨로지이다.

대부분의 유전자 정보는 이러한 해부학적 부위에 발현하는 유전자들을 해부학적 구조로 정리해 놓은 것들이 많다. 하지만 현존하는 연구 결과 들은 다른 예에서도 보듯이 상이한 정보를 제공할 수 있다. 따라서 유전자 정보의 통합을 통하여 이러한 불일치를 배아 전체에서 찾을 수 있다. 배아 전체에 나타나는 유전자 정보들이 통합 된다면 생물학자들은 하나의 유전자 정보를 알기 위해서 여러 연구 결과를 검토하지 않아도 되어 시간과 노력을 절감할 수 있게 된다. 이를 위하여 본 논문에서는 서로 다른 연구에 의해 구축된 유전자 정보를 하나의 정보로 통합 및 확장하는 기법을 소개한다.

2장에서는 유전자 정보 통합의 필요성을 본 연구에서 예로 사용하고자하는 데이터베이스 EMAGE와 GXD와 함께 소개하고, 3장에서는 K. McLeod와 A. Burger의 논증 기법을 소개한다. 4장에서는 이러한 논증 기법을 이용하여 유전자 정보가 서로 상이할 때 이를 통합할 수 있는 기법을 소개한다. 끝으로 5장에서는 결론과 함께 앞으로의 연구 방향을 제시 한다.

## 2. 유전자 정보 통합의 필요성

컴퓨터와 인터넷의 발달로 여러 가지 정보가 넘쳐나

는 가운데 생물학자들도 넘쳐나는 온라인 정보를 접하고 있다[1]. 생물학적 실험은 실험당시의 여러 가지 요소나 상황의 차이 또는 해석의 차이에 의해 서로 다른 결과를 생성하기도 한다. 이러한 결과들이 같은 목적을 위해 사용된다면 많은 혼란을 야기 할 것이다. 예를 들어 생물학자들이 쥐의 배아에서 어떤 유전자가 언제 어떻게 발현 하는지를 연구 했다고 가정하자. 그 결과 하나의 실험에서는 신경계 외배엽에서 Hoxb1 유전자가 발현 했다는 결론을 도출하고, 다른 실험에서는 신경계 외배엽에서 Hoxb1 유전자가 발현하지 않았다는 결론을 도출 했다면 어떤 실험 결과를 신뢰 할 것인가?

이런 문제를 해결하기 위해서 K. McLeod와 A. Burger는 추론 규칙과 논증을 이용한 시스템을 만들었다[2]. K. McLeod와 A. Burger의 시스템은 사용자의 질의 (e.g., expressed ('Hoxb1', 'EMP:151'))에 답을 해줄 뿐 전체 정보를 수정하거나 재구성해 주지는 않는다. 본 논문에서는 K. McLeod와 A. Burger의 연구 결과를 활용하여 서로 다른 연구에 의해 구축된 유전자 정보를 하나의 정보로 통합 및 확장하는 기법을 소개한다.

또한 K. McLeod와 A. Burger의 논증 시스템을 이용하여 어느 연구 결과가 더 신뢰할 수 있는 지를 밝히면 통합된 유전자 정보는 한층 더 신뢰할 수 있고 연구자들의 요구에 부합한 정보를 갖게 된다. 따라서 서로 다른 연구에서 생성된 유전자 정보가 통합되면 연구자들은 배아에서 유전자 발현과정을 찾는데 시간과 노력을 절감할 수 있게 된다.

## 3. K. McLeod와 A. Burger의 논증 시스템

K. McLeod와 A. Burger의 논증 시스템은 웹기반의 시스템으로 ASPIC에서 제공하는 추론엔진을 사용한다 [3]. ASPIC에서는 여러 응용에서 사용할 수 있는 논증 이론을 표준화하려고 노력하고 있다. ASPIC의 추론엔진은 Dung의 이론[4]에 기반을 두고 있으며, JAVA를 이용하여 구현되어 GUI나 Java API를 통하여 사용할 수 있다. GUI를 사용할 때 추론엔진의 입력은 PROLOG와 유사한 First-Order Logic의 형태를 갖는다. 또한 ASPIC의 추론엔진은 다음과 같은 TREE형태의 추론규칙을 갖는다.

Conclusion

Premise 1

Premise 3

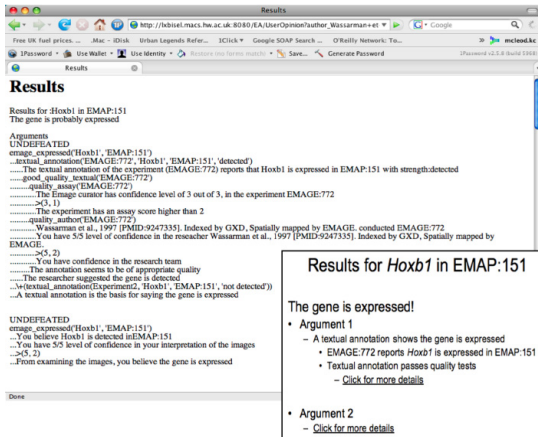
Rule: Premise 1 <- Premise 3

Premise 2

Rule 1 : Conclusion <- Premise 1 & Premise 2

위의 예를 설명하면 다음과 같다. Rule 1이 결론을 생성하는데 기여하는데 Rule 1은 Premise 1과 Premise 2가 동시에 '참'이어야만 결과가 '참'이라는 것을 나타낸다. 그리고 Premise 1이 다른 논증의 결론이 되고 Premise 2가 '참'이 되려면 Premise 3가 '참'이어야만 한다.

다음[Fig. 1]은 K. McLeod와 A. Burger의 논증 시스템이 웹상에서 실행되는 모습을 보인 것이다[5].



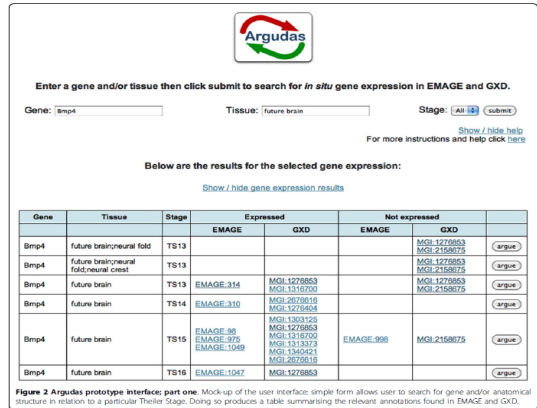
[Fig. 1] K. McLeod & A. Burger's Reasoning System[5]

[Fig. 1]은 논증시스템이 신경계 외배엽에서 Hoxb1 유전자가 발현했다는 정보가 더 신뢰성이 있는 정보라는 결론을 내린 것을 보이고 있다. [Fig. 1]에서 오른쪽 아래에 보이는 것은 웹상에 나타난 정보보다 간소화 된 것으로 생물학자들이 더 선호하는 방식이다.

4. 유전자 정보 통합 및 확장

현재는 서로 다른 유전자 정보 데이터베이스를 통합하는데 [Fig. 2]와 같은 도구를 사용한다. [Fig. 2]는 현재 사용하고 있는 유전자 정보 통합 시스템인 Augudas의

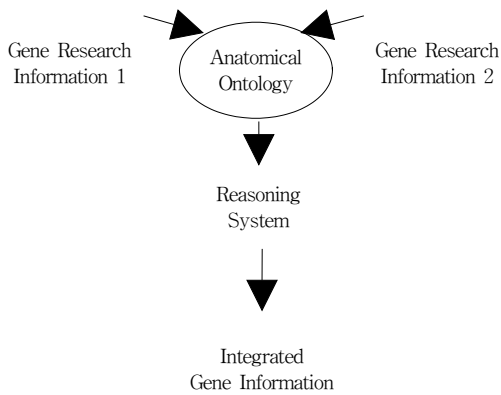
인터페이스 화면이다[5].



[Fig. 2] Argudas Prototype Interface

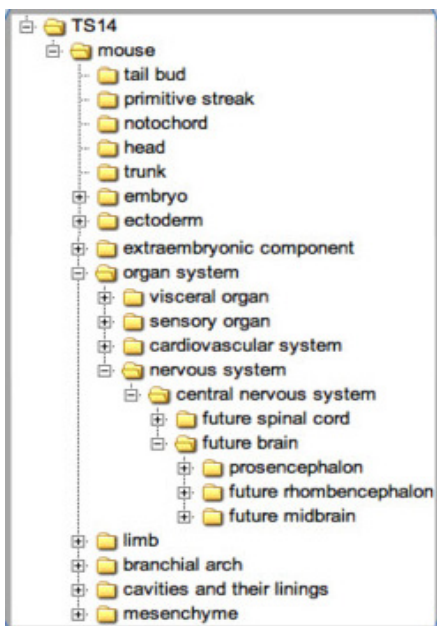
이는 서로 다른 두 데이터베이스 EMAGE와 GXD를 수동으로 통합 할 수 있다. 본 논문에서는 해부학적 구조를 이용한 유전자 정보 통합 및 확장 기법 (IAGIAS: Integration & Augmentation of Gene Information based on Anatomical Structure)을 소개한다. IAGIAS는 배아의 해부학적 구조를 밝히고 해부학적 구조의 이름을 TREE 형태의 온톨로지로 구축한 다음, 두 개의 유전자 연구 정보를 각각 하나의 해부학 구조 온톨로지에 대응시킨다. 이때 하나의 해부학 구조에 서로 상이한 유전자 정보가 나타나면 이는 K. McLeod와 A. Burger의 논증 시스템을 이용하여 어느 정보가 더 신뢰성이 있는 지를 밝혀 이를 최종 정보로 해부학 구조에 대응시킨다. 이러한 과정은 아래 [Fig. 3]에 나타나 있다.

본 논문에서 예로 사용하는 두 개의 유전자 발현 정보 데이터베이스는 EMAGE와 GDX이다. EMAGE는 쥐의 배아에서 나타나는 유전자 발현 정보를 모아놓은 데이터베이스[6]이고, GDX도 이와 유사한 데이터베이스[7,8]이다. 이들에는 서로 상이한 정보가 포함되어 있다. 따라서 이러한 상이한 정보들은 하나의 올바른 정보로 통합되어야 한다. 또한 하나의 데이터베이스에는 있으나 다른 데이터베이스에는 없는 정보도 있을 수 있다. 이러한 정보도 통합 데이터베이스에는 기록되게 되어 통합 데이터베이스를 활용하는 사용자는 유전자 발현 정보를 검색하기 위해 여러 데이터베이스를 검색하지 않아도 정확한 정보를 얻을 수 있게 된다.



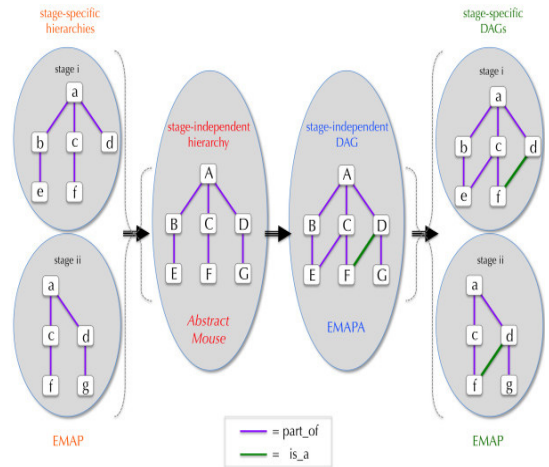
[Fig. 3] Flow Diagram of IAGIAS

다음 [Fig. 4]는 e-Mouse Atlas (EMA)[9]에 기반을 둔 해부학적 온톨로지의 일부이다. EMAP ontology는 쥐 배아의 발현 과정에 따른 해부학적 구조를 기반으로 하는 온톨로지이다. 서로 다른 유전자 정보 통합 시 이러한 해부학적 구조에 기반한 온톨로지를 사용하면 통합이 용이하고 빠지거나 중복되는 유전자 정보를 발견하기 쉬워진다.



[Fig. 4] EMAP ontology: Mouse developmental anatomy and gene expression data.

[Fig 5]는 쥐 배아의 발현 단계별로 EMAP와 EMAPA 온톨로지에서 각 단계에 종속된 표현과 각 단계에 비종속적인 표현의 예를 보여준다[10].



[Fig. 5] EMAP and EMAPA ontologies provide stage-specific and stage-independent representations of mouse embryo anatomy[9].

여기에는 배아 발달의 각 단계에 종속된 부분과 종속되지 않은 부분에 대한 구분이 되어있고 'Part\_of' 관계와 'is\_a' 관계가 구분되어 보여진다.

IAGIAS의 다음 단계는 논증 시스템을 적용하는 단계이다. 논증 시스템은 서로 다른 연구 결과에서 유전자 정보가 서로 다르게 밝혀졌을 때 이를 해결하기 위해서 필요하다. 이를 위한 논증 시스템은 앞에서 언급한 K. McLeod와 A. Burger의 논증 시스템[2]를 이용할 수 있다. 이러한 논증 시스템의 결과로 [Fig. 2]에서 보듯이 최종적인 통합 유전자 정보를 얻을 수 있다. 이렇게 하여 얻은 통합 유전자 정보는 상충되는 정보가 제거된 통합 유전자 정보이다.

다음은 이러한 작업을 위한 IAGIAS 알고리즘이다.

IAGIAS

GED := Gene Expression Database

APN := Anatomical Part Name (Ontology)

MBA := McLeod & Burger's Argumentation System

```

Do i = 1 to N
  Extract APN(i) from GED(i)
  Build Integrated APN

IAPN := Integrated APN

Do j = 1 to M
  Mapping GED(j) information to IAPN
  IF Conflict is found
    THEN call MBA
    
```

본 논문에서 제안하는 유전정보 통합 기법 IAGIAS를 따르면 서로 다른 연구에서 발견한 유전자 발현 정보를 통합하는 것은 물론이고 통합 시 나타날 수 있는 유전자 정보의 불일치를 배아 전체에서 찾을 수 있다. 배아 전체에서 나타나는 유전자 정보들이 불일치 없이 통합된다면 생물학자들은 하나의 유전자 정보를 알기 위해서 여러 연구 결과를 검토하지 않아도 되어 시간과 노력을 절감할 수 있게 된다.

## 5. 결론

본 논문에서는 세계 여러 곳에서 연구된 유전자 연구 정보를 해부학적 온톨로지를 이용하여 통합 및 확장할 수 있는 기법 IAGIAS를 제안 하였다. IAGIAS는 여러 개의 유전자 발현 데이터베이스로 부터 해부학적 구조의 이름을 통합하여 TREE 형태의 온톨로지로 구축한 다음, 복수 개의 유전자 연구 정보를 각각 하나의 통합된 해부학 구조 온톨로지에 대응 시킨다. 이때 하나의 해부학 구조에 서로 다른 데이터베이스로 부터 상이한 유전자 정보가 나타나면 이는 K. McLeod와 A. Burger의 논증 시스템을 이용하여 어느 정보가 더 신뢰성이 있는지를 밝혀 이를 최종 정보로 선택하고, 이를 통합된 해부학 온톨로지상의 구조 이름에 대응 시킨다.

IAGIAS는 같은 목적의 유전자 정보들 가운데 서로 다르게 보고되고 있거나 일부분만 보고되고 있는 정보를 발견하고 통합할 수 있는 간단한 자동화 기법이다. IAGIAS를 이용하여 통합된 유전자 정보 데이터베이스를 구축하면 연구자들은 정확한 유전자 발현 정보를 찾

는데 시간과 노력을 절감할 수 있을 것이다.

본 논문에서는 정보기술을 이용한 유전학적 연구정보의 통합과 관리 방법을 제시하였다. 이처럼 최근에는 정보기술과 의료 및 생물 등 다양한 분야의 융합에 의한 연구[11, 12, 13]와 건강관련 연구[14, 15]가 많이 수행되고 있다. 앞으로도 정보기술과 타 분야의 융합 연구가 활성화되기를 기대한다.

## REFERENCES

- [1] Bateman, A., Editorial. Nucleic Acids Research, Vol. 35,(Database Issue), D1-D2. 2007.
- [2] McLeod, K. & A. Burger, Towards the use of argumentation in bioinformatics: a gene expression case study, Vol.24 ISMB i304-i312., 2008.
- [3] ASPIC, Theoretical framework for argumentation, deliverable d2.1., 2004
- [4] Dung, P.M., On the acceptability of argumentation and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence, Vol. 77, 321-358, 1995.
- [5] McLeod, K., Gus Ferguson & A. Burger, Argudas: lessons for argumentation in biology based on a gene expression use case, BMC Bioinformatics, 13(Suppl 1):S8, 2012
- [6] Davidson, D., The mouse atlas and graphical gene-expression database. Semin. Cell Dev. Biol. Vol. 8, 509-517., 1997.
- [7] Ringwald, M. et at., The mouse gene expression database (GXD), Nucleic Acids Research, Vol. 29, 98-101, 2001.
- [8] Finger JH, Smith CM, Hayamizu TF, McCright IJ, Eppig JT, Kadin JA, Richardson JE, Ringwald M: The mouse Gene Expression Database (GXD): 2011 update. Nucleic Acids Res, 39(Database issue): D835-D841, 2011.
- [9] <http://www.emouseatlas.org/emap/home.html>
- [10] Terry F Hayamizu, Michael N Wicks, Duncan R Davidson, Albert Burger, Martin Ringwald and Richard A Baldock, EMAP/EMAPA ontology of

- mouse developmental anatomy, Journal of Biomedical Semantics. Vol.4., No.15, 2013,
- [11] Lee, Y-J., The Effect of Information Conditions on Mental Health among Elderly, The Journal of Digital Policy & Management, v.11, no.10, 17-29, 2013.
- [12] Han, D., M. Kang, Study on smart contents development directions for children with autistic disorder to enhance adaptive behavior, The Journal of Digital Policy & Management, v.11, no.10, 123-131, 2013.
- [13] Sanghyuk Lee, Grouping DNA sequences with similarity measure and application, Journal of the Korea Convergence Society, v.4, no.3, 35-41, 2013.
- [14] Weon Kim, Min-Ho Kim, Gyu-Beom Shim, Moon-Ju Shin, A study on the satisfaction of Health Examination for National Health Insurance service -Target of medical examinee in Busan-, Journal of the Korea Convergence Society, v.4, no.2, 1-8, 2013.
- [15] Kyung-hee Kang, Moo-Sik Lee, Moon Sook Shim, Kwang-Hwan Kim, A Study on Awareness of Organization of Public Health Center based on Educational Background and Position, Journal of the Korea Convergence Society, v.2, no.1, 37-44, 2011.

양 기 철(Yang, Gi-Chul)



- 1986년 8월: University of Iowa, Department of Computer Science, MS
- 1993년 5월: University of Missouri -Kansas City, Computer Science, Ph.D.
- 2002년 1월 ~ 2002년 8월: Heriot-Watt University, Visiting Scholar
- 1993년 9월 ~ 현재 : 목포대학교 멀티미디어공학과 교수
- 관심분야 : 인공지능, 휴먼-컴퓨터 상호작용, 정보검색
- E-Mail : gcyang@mokpo.ac.kr