Challenges in Biopathway Extraction from Literature and Ontology Construction for Biology

박종철 (KAIST 교수)
(042-869-3541(o), -3581(L), park@cs.kaist.ac.kr)

Abstract:

Recent developments ın literature data mining for biology call for the design of a common framework that can be used to assess the performance of the reported systems in a fair and objective way. In this talk, we present an on-going effort to make it possible, in the form of challenges in the extraction of biological pathways and in the ontology construction. We are currently making this effort jointly with Lynette Hirschman (MITRE), Junichi Tsujii (University of Tokyo), Limsoon Wong (KRDL), and Cathy Wu (Georgetown University).

약력

1984.2. 서울대학교 공과대학 컴퓨터공학과 공학사
1986.2. 서울대학교 대학원 컴퓨터공학과 석사
1996.5. 미국 펜실바니아대학교 전산정보학과 박사
1996.6 - 1998.2 미국 펜실바니아대학교 IRCS Research Associate
1998.3 - 현재 한국과학기술원 전산학과 조교수

관심분야

자연언어처리 및 계산언어학
인공지능, 인지과학 및 생물정보학

Challenges in Biopathway Extraction from Literature and Ontology Construction for Biology

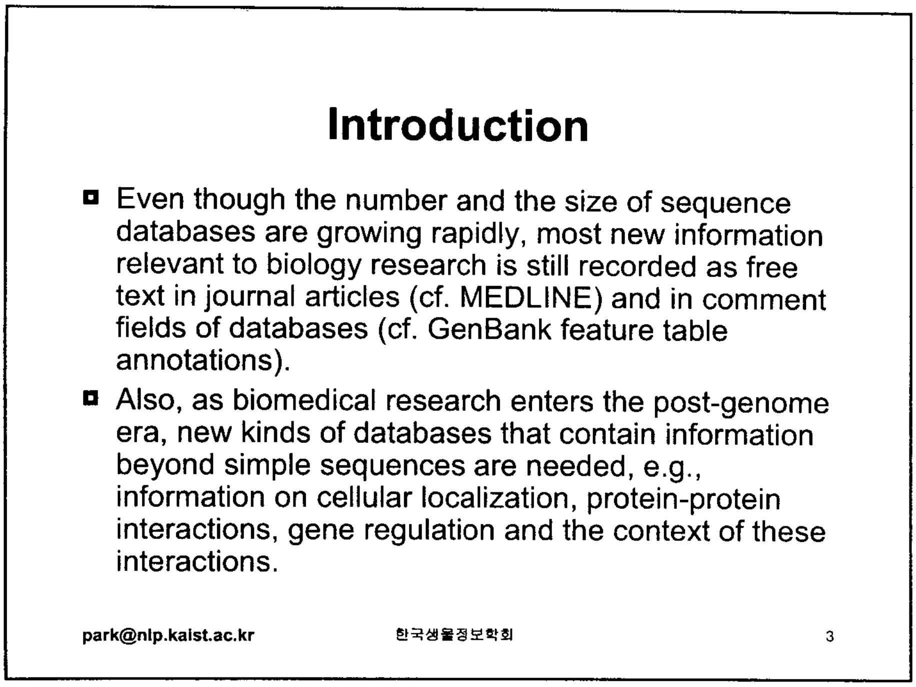
Jong C Park

KAIST

Work in collaboration with L Hirschman (MITRE, USA), J-I Tsujii (U Tokyo, Japan), L Wong (KRDL, Singapore) and C Wu (Georgetown U, USA)

Outline

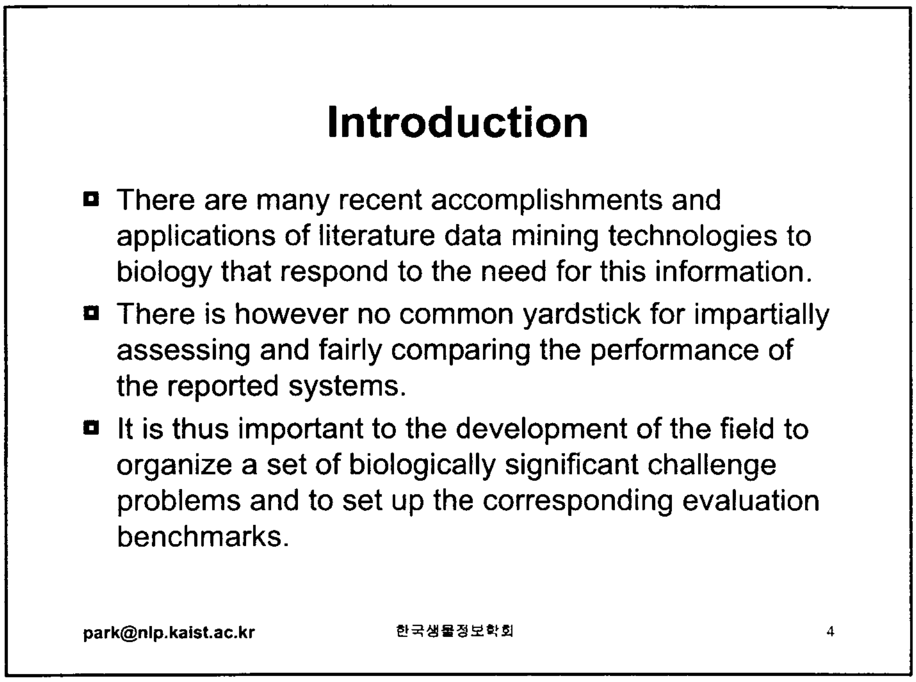
- Introduction
- Example Biopathways
- Example Ontology
- Challenges
- Biopathway Extraction from Literature
- Ontology Construction for Biology
- Our Work
- Conclusion



Introduction

- Even though the number and the size of sequence databases are growing rapidly, most new information relevant to biology research is still recorded as free text in journal articles (cf. MEDLINE) and in comment fields of databases (cf. GenBank feature table annotations).
- Also, as biomedical research enters the post-genome era, new kinds of databases that contain information beyond simple sequences are needed, e.g., information on cellular localization, protein-protein interactions, gene regulation and the context of these interactions.

park@nlp.kaist.ac.kr 한국생물정보학회 3



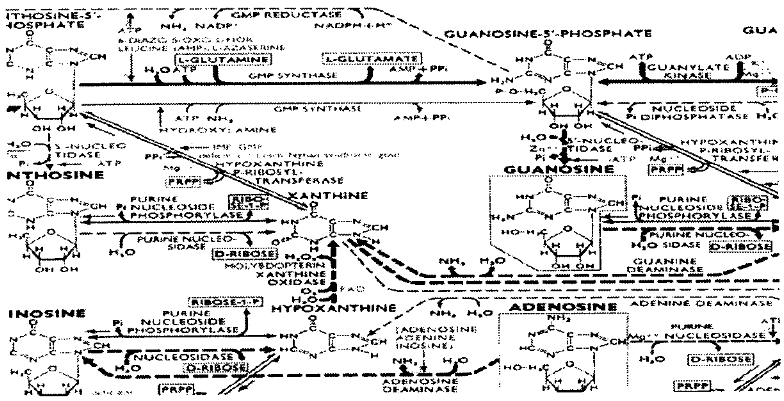
Introduction

- There are many recent accomplishments and applications of literature data mining technologies to biology that respond to the need for this information.
- There is however no common yardstick for impartially assessing and fairly comparing the performance of the reported systems.
- It is thus important to the development of the field to organize a set of biologically significant challenge problems and to set up the corresponding evaluation benchmarks.

park@nlp.kaist.ac.kr 한국생물정보학회 4

Example Biopathways

From Boehringer Mannheim "Biochemical Pathways" wall chart.



Example Ontology

From UMLS Semantic Types

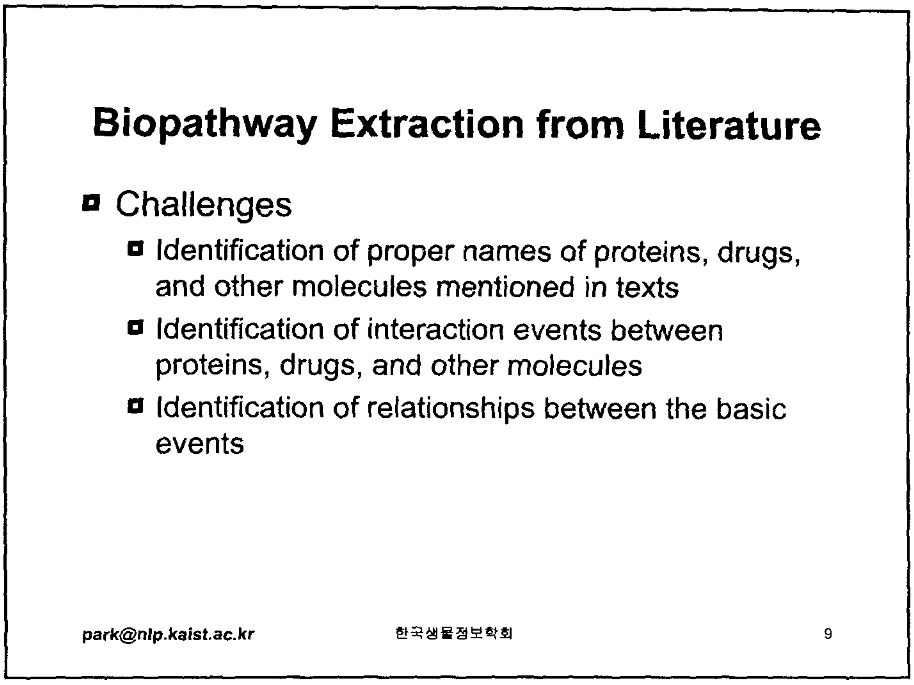
- Entity
  - Physical Object
    - Organism
      - Plant
        - Alga
      - Fungus
      - Virus
      - ...
      - Animal
        - Invertebrate
        - Vertebrate
          - Amphibian
          - ...
          - Mammal
            - Human

Challenges

- Biopathway Extraction from Literature
- Ontology Construction for Biology

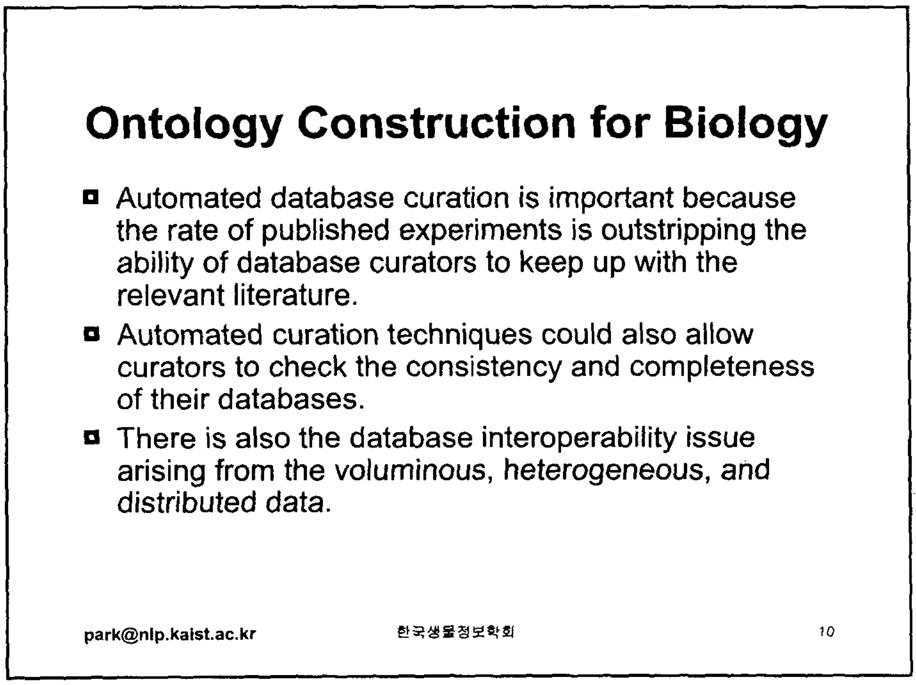
Biopathway Extraction from Literature

- Evaluation Measure
  - Recall and Precision
    - Recall(E) = TP(E)/(TP(E)+FN(E))
    - Precision(E) = TP(E)/(TP(E)+FP(E))
  - Simple Matching Coefficient (SMC)
    - SMC(E) = TP(E)/(TP(E)+FP(E)+FN(E))



Biopathway Extraction from Literature

- Challenges
  - Identification of proper names of proteins, drugs, and other molecules mentioned in texts
  - Identification of interaction events between proteins, drugs, and other molecules
  - Identification of relationships between the basic events



Ontology Construction for Biology

- Automated database curation is important because the rate of published experiments is outstripping the ability of database curators to keep up with the relevant literature.
- Automated curation techniques could also allow curators to check the consistency and completeness of their databases.
- There is also the database interoperability issue arising from the voluminous, heterogeneous, and distributed data.

Ontology Construction for Biology

- There is a growing recognition that the adoption of standard nomenclature, controlled vocabulary, and common ontologies is critical to interoperation and integration of biological data.
- Data integration into a knowledge base system is necessary for answering complex biological questions that may typically involve querying multiple sources.

Ontology Construction for Biology

- An ontology is a semantic model that contains a shared vocabulary and classification of concepts in a domain.
- An ontology is valuable for query expansion in mining scientific literature and integrating data from heterogeneous sources.
- It is also important for consistent database curation where functional conservation is annotated with a common language.

Our Work

- Jong C. Park, Hyun-Sook Kim and Jung-jae Kim, Bidirectional Incremental Parsing for Automatic Pathway Identification with Combinatory Categorial Grammar, Pacific Symposium on Biocomputing (PSB), Hawaii (Big Island), USA, January, 2001.
- Jong C. Park, Using Combinatory Categorial Grammar to Extract Biomedical Information, IEEE Intelligent Systems in Biology, Volume 16, Number 6, pages 62-67, November/December, 2001.
- Jin-bok Lee, Jung-jae Kim and Jong C. Park, Semi-Automatic Extension of Gene Ontology, Human-Computer Interaction, Phoenix Park, Korea, February, 2002.
- Lynette Hirschman, Jong C. Park, Jun-ichi Tsujii, Limsoon Wong, Cathy Wu, Literature Data Mining for Biology, PSB Session, Hawaii (Kauai), USA, January, 2002.
- Lynette Hirschman, Jong C. Park, Jun-ichi Tsujii, Limsoon Wong, Cathy Wu, Achievements and Challenges in Literature Data Mining for Biology, Bioinformatics Journal (submitted), 2002.
- Jung-jae Kim and Jong C. Park, Grammaticality Validation and Unknown Word Handling for Automatic Pathway Identification, Bioinformatics Journal (to be submitted), 2002.

Conclusion

- The time is ripe for common evaluation metrics and challenges for natural language processing applications to biology.
- A challenge evaluation will also give rise to a shared infrastructure, such as annotated training and test data and shared evaluation methods.
- We have discussed two relevant tasks.
  - Extraction of Biological Pathways from Literature
  - Automated Database Curation