TECHNOLOGY EVOLUTION ANALYSIS BASED ON SPO USING PATENT DOCUMENTS: A CASE STUDY OF INDUCED PLURIPOTENT STEM CELLS

Hu Z.Y., Wen Y., Liu C.J.
Chengdu Library and Information Center, Chinese Academy of Sciences (CLAS, CAS)
Wei L.
School of Information and Management, Shanxi University of Finance and Economics(SIM, SUFE)
Qin X.C.
Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences (GIBH, CAS)

Leiden

Sep. 11, 2018

Research Background

Knowledge Discovery in Biomedical Literature (KDiBL)

KDiBL

- KDiBL is the process of identifying and extracting the new, useful, potential and understandable patterns from the biomedical literatures in a credible way.

Information extraction

- Information extraction plays an important part in KDiBL , which automatically extracts the specific terms, the corresponding characteristics and the semantic relations among them from the texts as the basic knowledge unit of knowledge discovery. SPO predication is a typical form of extraction.

SPO predication

- SPO predication consists of a Subject argument (noun phrase), an object argument (noun phrase), and the relation that binds them (verb phrase), which can represent science and technology (S&T) information with more details in a simple manner and have been widely applied in KDiBL

Technology Evolution Analysis

Object

- By means of bibliometric analysis, natural language processing, text mining and other technical means to outline the development trend of technological innovation in the field. In the process, the technical connotation of multi-source heterogeneous data in the field will be extracted, identified, analyzed, reconstructed, screened and calculated.

Effect

- It can trace the main technological development paths in the field from the perspective of birth, development and downfall of technology, provide comprehensive decision-making information from the complete vision of the technology chain, help decision-makers to clarify the mainstream of technology development, identify key technologies and emerging technologies, and grasp the future of technology direction of development.

Literature Review

Present State of Research

Scientific Literatures

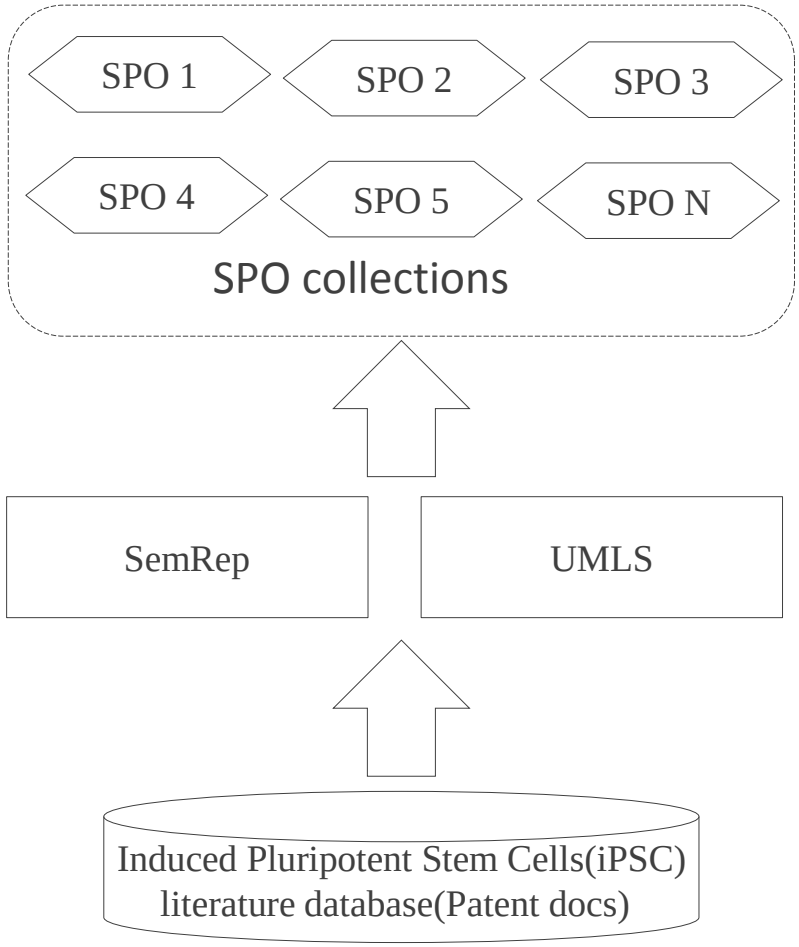
- Mainly focus on the method that can be used to draw technology evolution map of keywords by calculating the distributions of keywords over the documents cluster groups.

Limitations

- **Lack of semantic information**: keywords lack semantic information, knowledge discovery needs automatically extracts information such as entities, entity attributes and semantic relations between entities as the basic knowledge unit.
- **Lack of repeatable**: the performance of various extraction tools is different, and the applicability is different.
- **Single Perspective**: (Technology Evolution Analysis turns into Topic Evolution Analysis) Technology Evolution Analysis is composed of the evolution analysis of technology's problems, solutions, functions and effects, Topic Evolution Analysis is only part of Technology Evolution Analysis.

Methodology

Step 1: Extracting SPO Structures



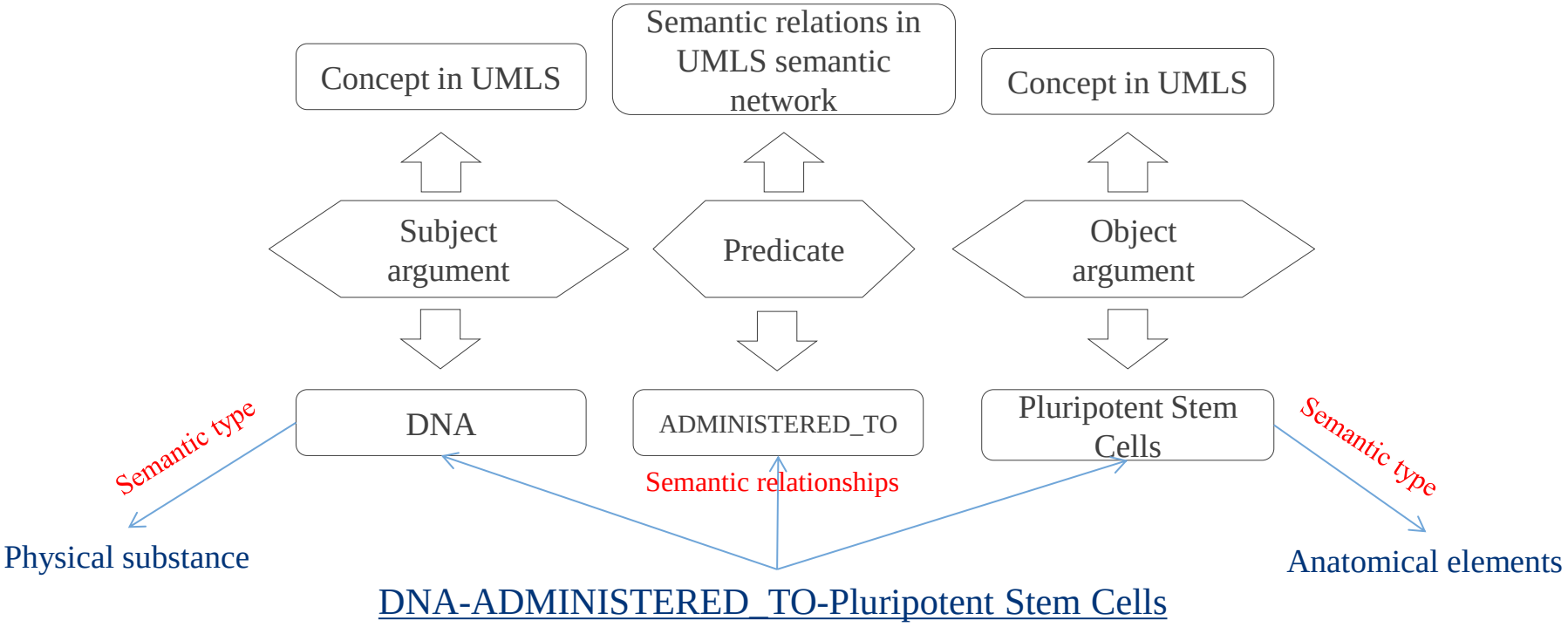
Tools in KDiBL

- Unified Medical Language System， UMLS[1]
  - UMLS is a biomedical integrated super-thesaurus system. It integrates multiple vocabularies in the fields of biomedicine and health, and uses a combination of string-term-concept to standardize the terminology in the biomedical field and provides an interoperable interface for computer processing.
- MetaMap[2]
  - MetaMap is a tool for mapping free words to UMLS concepts. It can mark the UMLS concepts contained in texts. It is widely used in various fields of KDiBL as a basic text processing tool.
- SemRep[3]
  - SemRep is one of the most important achievements of the NLM semantic knowledge representation project. It is a semantic knowledge extraction tool for biomedical literature based on UMLS and MetaMap.

1. NCBI.UMLS® Reference Manual[EB/OL].http://www.ncbi.nlm.nih.gov/books/NBK9676/.
2. ARONSON A R,LANG F.An overview of MetaMap:historical perspective and recent advances[J].Journal of the American Medical Informatics Association,2010,17(3):229-236.
3. ARNOLD P,RAHM E.Semrep:a repository for semantic mapping[EB/OL].https://dbs.uni-leipzig.de/en/publication/title/semrep_a_repository_for_semantic_mapping.

Tools in KDiBL

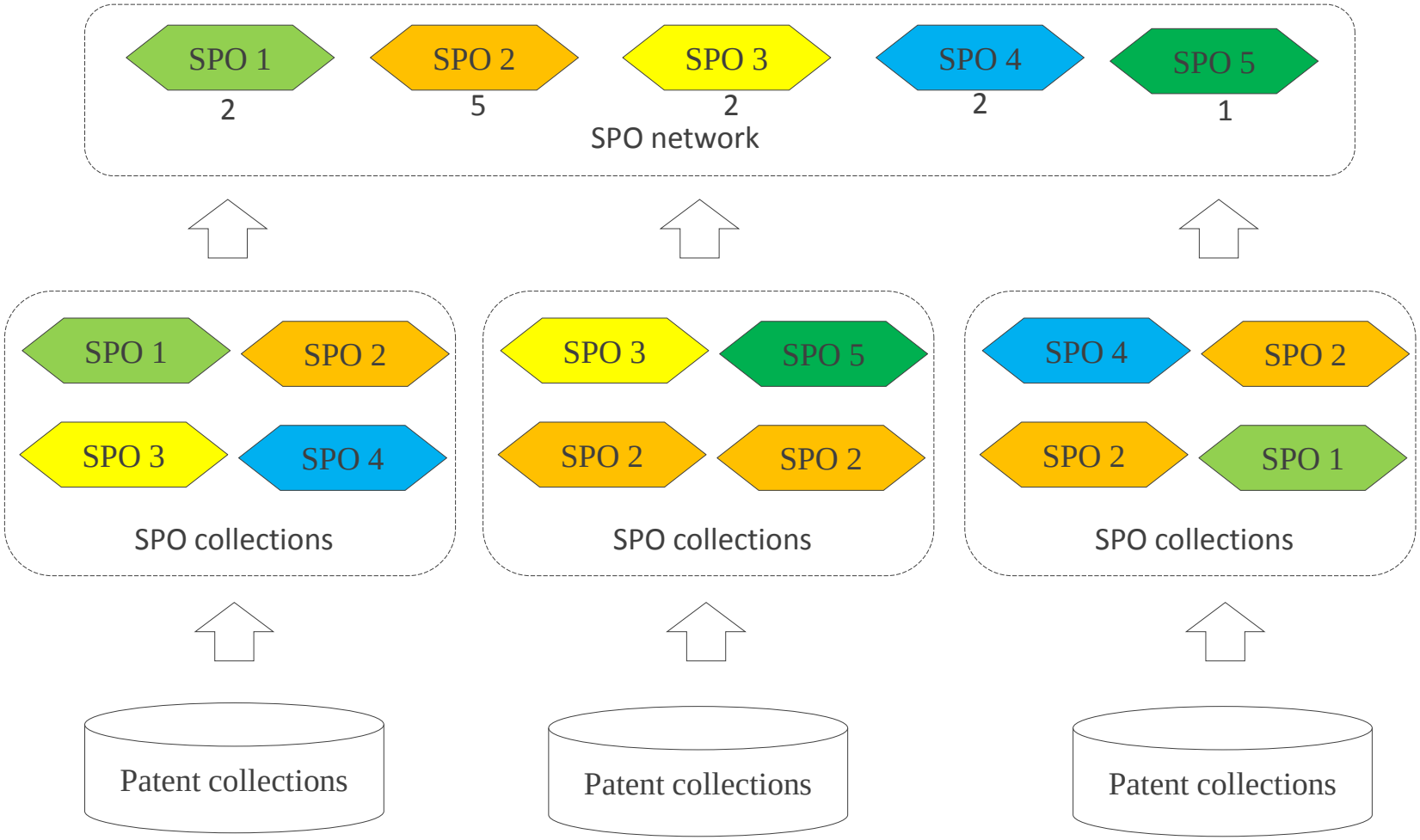
- The UMLS Semantic Network
  - Include **133** semantic types and **54** semantic relationships
- Example



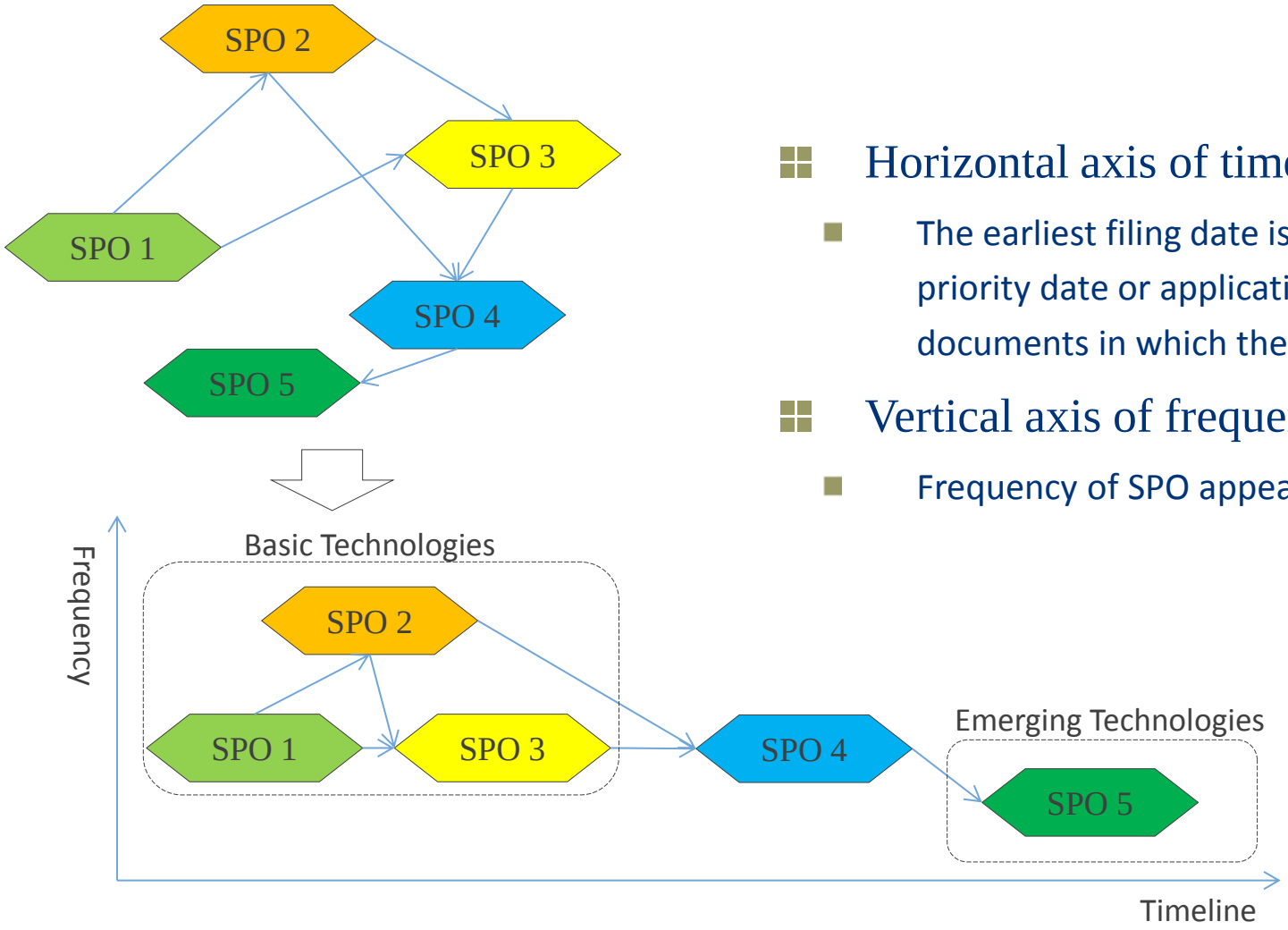
Step 2: Clustering the patent documents

- Fuzzy clustering analysis
  - Clustering analysis using fuzzy mathematics, it shows better than other clustering analysis in terms of runtime and accuracy in some circumstances.
- FCM Algorithm:
  - Standardized data matrix;
  - Establish a fuzzy similarity matrix and initialize the membership matrix;
  - The algorithm starts iterating until the objective function converges to a minimum value;
  - According to the iterative result, the class to which the data belongs is determined by the last membership matrix, and the final clustering result is displayed.

Step 3: Forming semantic network of SPO predications



Step 4: Drawing technology evolution map



- Horizontal axis of timeline:
  - The earliest filing date is the earliest priority date or application date of patent documents in which the SPO appears.
- Vertical axis of frequency:
  - Frequency of SPO appears in each group.

Case Study(Ongoing)

Case Study(Ongoing)

- iPSC patents were selected as a case study, Derwent Innovations Index (DII) as data source and 1,282 patent documents are obtained from 2008 to 2017. Following the above methodology, the technology evolution map of IPSC patents is drawn. A part of the map is shown in Figure 1.

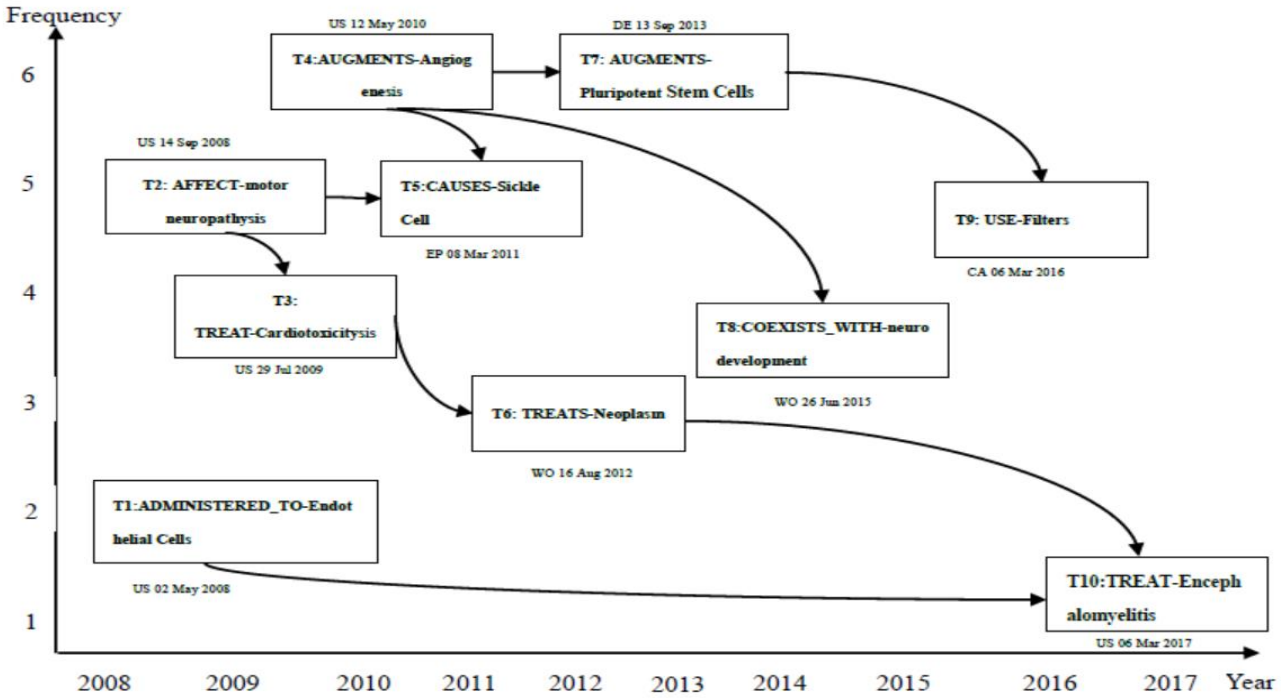


Figure 1

Summary

Advantage

- **Repeatable**: The mentioned tools above make extraction work that have been done repeatable.
- **Cleaner**: That's because traditional extraction tools are casual and These SPO predications extracted by SemRep are cleaner and more formal and can be directly used as the basis of technology evolution analysis.
- **Rich semantic information**: The subject and object of the SPO predications are concepts in UMLS, and the predicate comes from the semantic relationship in the UMLS semantic network.
- **More Perspectives**: our ongoing work didn't go further in the The separate technology evolution maps of problem, solution, function , effect and the combination to a more comprehensive technology evolution map of stem cell.

Summary

- The result indicates that SPO predications which contain more semantic information are more suitable for technology evolution analysis than keywords. SPO predications can be used to generate topics and draw more comprehensive technology evolution map.

中国科学院成都文献情报中心

Chengdu Library and Information Center, Chinese Academy of Sciences

中国科学院广州生物医药与健康研究院

GUANGZHOU INSTITUTES OF BIOMEDICINE AND HEALTH, CHINESE ACADEMY OF SCIENCES

Thank You!

Acknowledgement: The work was supported by the Informationization Special Project of Chinese Academy of Sciences “E-science Application for Knowledge Discovery in Stem Cells” (Grant No: XXH13506-203).