Proposal for KAIST Grand Challenge 30 Project

<table>
<tr><td rowspan="2">1. Title</td><td colspan="2">Korean</td><td colspan="3"></td></tr>
<tr><td colspan="2">English</td><td colspan="3">Finding Research Hypothesis in Natural Sciences</td></tr>
<tr><td rowspan="11">2. Principal Investigator (PI)</td><td rowspan="2">Name</td><td>Korean</td><td></td><td rowspan="2">Department</td><td rowspan="2">Knowledge Service Engineering / Computing</td></tr>
<tr><td>English</td><td>Aviv Segev</td></tr>
<tr><td colspan="2">Major Research Field</td><td colspan="3">Knowledge Modeling</td></tr>
<tr><td colspan="5">The list of on-going research projects as of March 2016</td></tr>
<tr><td colspan="2">Funding Institution</td><td>Title</td><td colspan="2">Funds for this year</td></tr>
<tr><td colspan="2">Korean Science Foundation</td><td>BK21</td><td colspan="2">15,000,000 KRW</td></tr>
<tr><td colspan="2"></td><td></td><td colspan="2">KRW</td></tr>
<tr><td colspan="2"></td><td></td><td colspan="2">KRW</td></tr>
<tr><td colspan="5">The list of research projects funding from KAIST (during 2009-present)</td></tr>
<tr><td colspan="2">Category
(e.g. KIHRHRP/EEWS/etc.)</td><td>Title</td><td colspan="2">Period
(yyyy/mm - yyyy/mm)</td></tr>
<tr><td colspan="2">---</td><td>---</td><td colspan="2"></td></tr>
<tr><td colspan="2">---</td><td>---</td><td colspan="2"></td></tr>
<tr><td colspan="2">---</td><td>---</td><td colspan="2"></td></tr>
</table>

3. Project Summary	

a. Provide the global research trends concerning the proposed research theme including an overview and general solutions.
b. Specify how the nature of the proposed research satisfy the requirements of eligibility(b).
c. No need to write research necessity and expected outcome.
d. Write in plain language within 2~3 pages.

Scientific data in natural science is expanding much faster than scientific research areas can analyze. Search technologies can find many scientific documents and knowledge models, but they cannot organize the extracted documents and knowledge models within them to suggest research hypotheses to the researcher. We propose to build a hypothesis

generator based on existing search methodologies, databases representing known scientific structures, and published research documentation. The research will develop a tool that will be used by researchers at KAIST and worldwide to suggest possible hypotheses in expanding research areas. The goal is to develop a search and analysis tool in a predefined research area such as chemical analysis of molecular structures. Once the tool proves useful in suggesting an interesting hypothesis that has not been considered before in the research area and that is approved as interesting by KAIST researchers, the hypothesis generation area will be expanded to a larger research area such as chemistry. In the future, after the successful ability of the tool is proved, other areas such as biology and physics can be considered.

Previous work has implemented hypothesis generation and experimentation by a laboratory robot system [1]. The robot was developed to work in a very limited predefined area of determination of gene function using deletion mutants of yeast (Saccharomyces cerevisiae) and auxotrophic growth experiments.

Other work developed a prototype system that mines the information contained in the scientific literature and attempts to create experimentally testable hypotheses [2]. The application of the approach was based on mining published literature to identify new protein kinases that phosphorylate the protein tumor suppressor p53.

Our experience is within the area of analyzing technology based on data sources such as research articles, published books, news, and search topics in order to predict future trends of technology [3]. We analyzed future research communities based on citation networks [4]. However, the goal of suggesting a specific set of hypotheses in other diverse areas such as chemistry, biology, or physics is a new and interesting challenge.

The goal of our proposal is to integrate the described approaches and enable the generation of possible hypotheses and the suggestion of experiments based on previously developed knowledge modeling databases and research publications. The area of hypothesis generation will be expanded well beyond existing sample examples to create a tool which could be used in research fields such as chemistry, biology, and physics. The tool will serve as a research assistant, allowing the choice of relevant and possible hypotheses while eliminating previously performed experiments or experiments where the result could be

predicted based on existing literature and datasets.

The main challenge of the research is to build a tool based on technology that is general enough to be implemented in different research fields and yet has a set of rules that can be tailored to each specific area. These specified sets of rules could be predefined by the system developers, defined and modified by each researcher, or learned automatically from existing literature.

The research addresses the most fundamental question in basic sciences, which is how we analyze everything which has been done before, to prevent from repeating the same mistakes and to identify new and promising directions. The problem is not unique to any specific research area and therefore poses a global challenge. Many of the existing tools from computer science, which are used for searching, extracting, and analyzing data, can be utilized. These tools are usually used for analyzing data designated for computer use where here we propose using similar approaches for research scientists in basic sciences. It is usually hard to build basic tools that could be used throughout research fields since new search or analysis methods in computer science are based on prototyping in a specific area. The development of a system for hypotheses generation has multiple requirements across different research areas. Although commercialization of such a system is possible in the long run, the advantage of building a leading tool for research analysis, developed and supported by KAIST, has a much greater promotional value for universities worldwide.

Research period : 8 years

Amount of funding requested : 20 million won (experiment)

[1] King, R. D.; Whelan, K. E.; Jones, F. M.; Reiser, P. G. K.; Bryant, C. H.; Muggleton, S. H.; Kell, D. B.;Oliver, S. G. (2004). "Functional genomic hypothesis generation and experimentation by a robot scientist". Nature 427 (6971): 247–252.

[2] Spangler, S.; Wilkins, A. D.; Bachman, B. J.; Nagarajan, M.; Dayaram, T.; Haas, P.; Regenbogen, S.; Pickering, C. R.; Comer, A.; Myers, J. N.; Stanoi, I.; Kato, L.; Lelescu, A.; Labrie, J. J.; Parikh, N.; Lisewski, A. M.; Donehower, L.; Chen, Y. and Lichtarge. O. (2014). "Automated hypothesis generation based on mining scientific literature". In Proceedings of the 20th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD '14). ACM, New York, NY, USA, 1877-1886.

[3] Segev, A.; Jung, S. and Choi, S.; Analysis of Technology Trends Based on Diverse Data Sources, IEEE Transactions on Services Computing, 8(6), pp. 903-915, 2015.

[4] Jung, S. and Segev, A., Analyzing Future Communities in Growing Citation Networks, Knowledge-Based Systems, 69, pp. 34-44, 2014.

Applicant : Aviv Segev (signature)