

PosMed: a biomedical entity prioritisation tool based on statistical inference over literature and the Semantic Web

Norio Kobayashi¹, Yuko Makita¹, Manabu Ishii¹, Akihiro Matsushima¹,
Yoshiki Mochizuki¹, Koji Doi¹, Koro Nishikata¹, David Gifford¹,
Terue Takatsuki², Hiroshi Masuya², and Tetsuro Toyoda¹

¹ Integrated Database Unit, Advanced Center for Computing and Communication,
RIKEN, Wako, Japan

{`norikobayashi, ymakita, manabui, amatsus, ym, kdoi, koro, gifford, toyoda`}@base.riken.jp

² Technology and Development Unit for Knowledge Base of Mouse Phenotype,
BioResource Center, RIKEN, Tsukuba, Japan
{`takatter, hmasuya`}@brc.riken.jp

Abstract. Positional MEDLINE (PosMed) is a web application that quickly prioritises biomedical entities such as genes and diseases based on statistical significance of associations between these and a user-specified keyword by employing our original search engine named General and Rapid Association Study Engine (GRASE). GRASE search is modelled as an extension of SPARQL search with statistical analysis, which enables searching over semantic data including not only linked datasets but also significant extracted semantic links over multiple biomedical documents. PosMed was originally implemented for *in silico* positional cloning studies by prioritizing genes. Further applications include bioresource search with associated genetic functions or ontologies, and functional interpretation of gene variants found from exome sequencing of personal genomes. PosMed is available at <http://database.riken.jp/PosMed/>.

Keywords: Linked data prioritisation, Statistical search, Text mining, Omics analysis

1 Introduction

In the life sciences field, a Semantic Web approach that employs machine-readable linked data prepared from conventional various omics datasets has been studied to understand biomedical phenomena. However, because the task of generating semantic links for our biomedical knowledge is too expensive, and such knowledge is described by a vast amount of human-readable biomedical literature, this semantic technology is still not widely adopted by biologists.

For practical use of published biomedical data on the Semantic Web, especially use of data difficult to utilise due to lack of semantic links, it is beneficial to reinforce acquisition of such data by supplying a hybrid methodology combining not only inferences over that knowledge described as linked data but also

knowledge supported by statistical significance over a vast number of multiple raw documents.

Our implementation of this methodology is the search engine named GRASE [1]. To confirm the problem solving abilities of GRASE for the life sciences, we developed a simple but effective graphical user interface for GRASE called PosMed [2] and in 2005 published this service to be accessible by a user's web browser. We started with mouse and human gene prioritisation for *in silico* positional cloning, and so far extended datasets and the service for intelligent bioresource search and exome analysis for the next generation sequencing. The rest of this paper presents a computational model of GRASE search and problem solving examples using PosMed with our latest datasets.

2 Statistical search model of GRASE

GRASE search is modelled as an extension of SPARQL search with statistical analysis, which enables searching over semantic data including not only datasets in Resource Description Framework (RDF) but also significant extracted semantic links over multiple biomedical documents including MEDLINE abstracts.

Direct search (keyword \rightarrow entity) The GRASE search engine quickly prioritises biomedical entities such as genes, diseases, drugs and mouse strains based on statistical significance of associations with a user-specified keyword. More concretely, for each entity GRASE generates a 2×2 contingency table $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ consisting of the number of documents (a) where both the the keyword and the entity appear, (b) where the keyword appears but the entity does not appear, (c) where the keyword does not appear and the entity does appear, and (d) where neither the keyword nor the entity appear, then applies the Fisher's exact test to the contingency table to compute a P -value for the significance of the test.

Inference search (entity \rightarrow entity) GRASE further infers other entities from the result of direct search by applying semantic links described by RDF triples and statistically extracted co-citation relationships over two entities e_1 and e_2 appearing in a common document by applying Fisher's exact test against 2×2 contingency table $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ consisting of the number of documents (a) where both e_1 and e_2 appear, (b) where e_1 appears but e_2 does not appear, (c) where e_1 does not appear but e_2 does appear, and (d) where neither e_1 nor e_2 appear. Therefore, an entity can be searched via a search path $keyword \rightarrow entity_1 \rightarrow entity_2$ and its significance computed as $1 - (1 - p_d)(1 - p_i)$, where p_d and p_i are P -values of direct search and inference search respectively.

3 Practical applications of PosMed

Since 2005, we have been extending datasets in PosMed to make it possible to follow an ever-changing trend of biomedical applications. Datasets currently introduced in PosMed are shown at <http://database.riken.jp/PosMed/>.

3.1 *In silico* positional cloning

A typical application of PosMed is searching genes with user-specified keywords and chromosomal intervals suggested by linkage analysis. So that inference searches can be performed such as mouse gene–drug inference and mouse gene–human gene inference, currently PosMed supports the following datasets:

- up to 352,000 entities including not only genes in mouse, human, rat, *Arabidopsis* and rice, but also drugs, metabolites, diseases and mouse strains associated with document sets including up to 9,870,000 documents from MEDLINE abstracts, OMIM, gene annotation, molecular interaction, Open Biomedical Ontologies (OBO) [3] including Gene Ontology, Mammalian Phenotype, Human Disease Ontology and Plant Ontology, and
- up to 828,000 semantic links including homologue genes and mouse strain–gene relationships.

In order to realise quick response, the datasets listed above are distributed over 11 computers and these work in parallel.

PosMed was used to prioritise genes in the RIKEN large-scale mouse ENU mutagenesis project and contributed to successful identification of 65 responsible genes [4]. PosMed is also used worldwide and successfully narrowed candidate genes responsible for a specific function after QTL analysis [5].

3.2 Bioresource search in mice and *Arabidopsis*

One conventional problem for a mouse bioresource database is that knockout strains are not used when the targeted gene has an unknown function and no observed phenotype. We introduced 19,885 mouse strains registered in the International Mouse Strain Resource (IMSR) [6] to discover wider resources than by simple keyword search over mouse strain catalogues and this accelerated bioresource utilisation, especially for those having fewer phenotypic annotations.

PosMed successfully connects these functionally unknown genes to known genes using molecular interactions, pathway information and co-citations and as a result enables suggestion of unobserved phenotypic bioresources. PosMed not only allows users to retrieve mouse bioresources directly with the user’s keywords described in bioresource annotations, but also inferentially through corresponding documents for mouse and human genes, diseases, drugs, ontologies, pathways, metabolites, molecular interactions and MEDLINE abstracts.

As an extension to other species, we newly introduced 7,207 *Arabidopsis* bioresources and 823 *Arabidopsis* phenotype observations extracted by human literature curation, so that PosMed inferentially discovers *Arabidopsis* bioresources as well through corresponding documents for genes, phenotypes, ontologies, co-expressions, molecular interactions and MEDLINE abstracts.

3.3 Functional interpretation of gene variants

PosMed can also be applied to functional interpretation of genetic variants detected by exome sequencing studies using a next generation sequencer. Since

exome sequencing studies usually find several hundreds or thousands of genetic variants by comparing samples and controls, prioritisation of the candidate genes using PosMed is an effective method for further functional analysis.

Users can upload a tab-separated values file with gene IDs and their descriptions. PosMed prioritises genes listed within files by statistical relevance between the user's keywords and each gene, and displays ranked genes together with user uploaded descriptions and associated documents.

4 Discussion and conclusion

We proposed a Semantic Web data search methodology and tool that extends conventional graph search like SPARQL with statistical text mining over a vast number of documents. Not only discovering discoveries documents related to biomedical entities when given a query as does the service GoPubMed, our PosMed also supports biomedical entity prioritization. Among several software tools available to prioritise positional candidate genes, PosMed was evaluated as sepecially highly effective in comparison with two other similar tools GeneSniffer and SUSPECTS [7]. We expect our prioritisation tools to effectively assist further practical life science studies, making the most of the data extensibility of the Semantic Web.

References

1. Kobayashi, N., Toyoda, T.: Statistical search on the Semantic Web. *Bioinformatics*, 24(7), pp. 1002–1010 (2008)
2. Makita, Y., Kobayashi, N., Yoshida, Y., Doi, K., Mochizuki, Y., Nishikata, K., Matsushima, A., Takahashi, S., Ishii, M., Takatsuki, T., Bhatia, R., Khadbaatar, Z., Watabe, H., Masuya, H., Toyoda, T.: PosMed: Ranking genes and bioresources based on Semantic Web Association Study. *Nucleic Acids Res.*, 41(Web Server issue), pp.W109–W114 (2013)
3. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J.; OB Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, 25(11), pp.1251-1255 (2007)
4. Masuya, H., Yoshikawa, S., Heida, N., Toyoda, T., Wakana, S., Shiroishi, T.: Phenosite: a web database integrating the mouse phenotyping platform and the experimental procedures in mice. *J. Bioinform. Comput. Biol.*, 5, pp.1173–1191 (2007)
5. Kato, N., Watanabe, Y., Ohno, Y., Inoue, T., Kanno, Y., Suzuki, H., Okada, H.: Mapping quantitative trait loci for proteinuria-induced renal collagen deposition. *Kidney Int.*, 73, pp.1017–1023 (2008)
6. Eppig, J.T., Strivens, M.: Finding a mouse: the International Mouse Strain Resource (IMSR). *Trends Genet.*, 15(2), pp.81-82 (1999)
7. Thornblad, T., Elliott, K., Jowett, J., Visscher, P.: Prioritization of Positional Candidate Genes Using Multiple Web-Based Software Tools. *Twin Res. Hum. Genet.*, 10(6), pp.861–870 (2007)