Classification of Resources on Knowledge Map in Biology

Jiro Araki[1], Shoko Kawamoto[2],
Asao Fujiyama[3] and Hideaki Takeda[3]

[1] Mitsubishi Research Inst. Inc., Japan
jiro@mri.co.jp
[2] ROIS Transdisciplinary Research Integration Center, Japan
skawamot@nii.ac.jp
[3] National Institute of Informatics, Japan
{afujiyam, takeda}@nii.ac.jp

Abstract. Life is a complex hierarchical system, whereas biology is to investigate life from various perspectives and is subdivided into different disciplines such as molecular biology and biochemistry. Recently vast data and knowledge from several genome projects came to be public on the web. But it is difficult to find a particular data and understand its significance for lack of entire picture of life or biology. In this paper we construct knowledge map to get an entire picture of biology. Efficient construction of knowledge map is realized by using tables of contents in textbooks as structured knowledge. More latest and detailed knowledge is also extracted based on co-occurrence between indices in textbooks and terms in PubMed articles. We show that in addition to enable us to look over biology world, this knowledge map is useful in the classification of resources on the web.

1 Introduction

In life science research, vast amounts of data and knowledges are generated by various huge projects such as genome sequencing and expression analysis, and are public on the web. But we cannot cleverly utilize them because it is difficult to find our targets among a large amount of data and understand each significance for lack of their entire picture.

Thus we try to construct knowledge map to get an entire picture of biology and navigate around resources. We consider that the appropriate knowledge map is a kind of taxonomy tree in which knowledge is classified into a leaf. In addition, each leaf in taxonomy tree has terms associated with its subject in order to classify document resources including their terms.

In this paper, we describe the method to construct knowledge map in Section 2. In Section 3, we evaluate the accuracy and performance of our knowledge map. In Section 4, we demonstrate the classification of resources in biology as an application of our knowledge map.

2 Construction of Knowledge Map in Biology

In this section, we describe the method for construction of the knowledge map.

2.1 Basic Idea

First of all, we consider the requirements of the knowledge map in biology.

- The knowledge map covers almost all subjects in biology.
- Users can access the map from various perspectives and look for cross-cutting relationships on it.
- It includes not only well-established knowledge but also newly-found knowledge.
- From developer's viewpoint, it requires a little time and effort to construct and maintain it.

In order to meet these requirements, we may not construct it from scratch, but utilize as much existing knowledge resources as possible. Of existing knowledge resources, typical well-established knowledge is considered to be accumulated in textbooks. Especially, tables of contents and indices give an overview of knowledge described in textbooks. On the other hand, newly-found knowledge is presumed to be expressed as technical terms and term relationships in published articles.

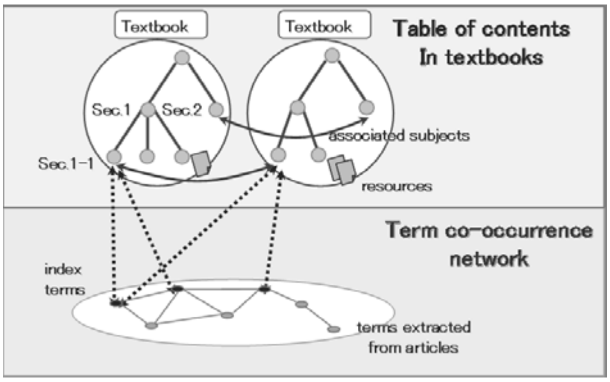

Fig. 1. Overview of knowledge map

2.2 Table of Contents and Indices in Textbooks as Well-established Knowledge

Table of contents in a textbook is a kind of knowledge map. Knowledge is classified and arranged in the form of a hierarchical structure. For example,

in the textbook 'Molecular Biology of the Cell', contents are divided into five parts: I.Introduction to the Cell, II.Basic Genetic Mechanisms, III.Methods, IV.Internal Organization of the Cell and V.Cells in Their Social Context. Part.II is further divided into four sections: 4.DNA and Chromosomes, 5.DNA Replication,Repair, and Recombination, 6.How Cells Read the Genome:From DNA to Protein, 7.Control of Gene Expression. Learners who want to know how cells are synthesized may select Part.II and further Section 6. In this way, table of contents enables readers to navigate around knowledge in a textbook.
Index term is likely a technical term, which is available in last pages of textbooks. Index term belonging to a section has an ability to explain its subjects. For example, the above Section 6 includes various index terms: DNA, protein, transcription, splicing, translation and so on. Learners can find associated sections by selecting known index terms.

We utilize several textbooks as follows:

1. Molecular Biology of the Cell ('Cell' for short)
   This textbook is concerned with cells from molecular perspective.
2. Harper's Illustrated Biochemistry ('Harper' for short)
   This textbook is concerned with biological macromolecules from biochemistry.
3. Wallace's Biosphere: The Realm of Life ('Wallace' for short)
   This textbook is concerned with whole of life and describes it in brief.

We combine these tables of contents and merge index terms to cover many subjects in biology. But index terms from different textbooks have a variety of expressions, for example 'adrenaline' in the Cell is expressed as 'epinephrine' in the Wallace. Therefore, we normalize index terms from different textbooks.

We show the number of sections in table of contents for each textbook in the left part of Figure 2. In the right part of the figure, we show the number of index terms in each textbook and ones included in several textbooks.

	Rank1	Rank2	Rank3	Rank4
Cell	5	25	111	1094
Harper	7	54	462	-
Wallace	7	48	401	399

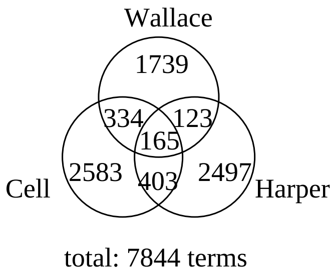

Fig. 2. Number of sections in tables of contents and index terms in three textbooks

2.3 Term Co-occurrence Network as Newly-found Knowledge

Newly-found knowledge is published in articles. For example, in biology, 'gene A interacts with gene B' or 'gene C activates phenomenon D' are expressed in articles. But articles may include only knowledge fragment or wrong knowledge. Additionally, articles express knowledge in the form of free text.
Thus we collect technical terms and term relationships from vast article database and evaluate reliability of them statistically.

Technical Term Extraction We use PubMed[1], which is biomedical article database maintained by NCBI, to extract technical terms and term relationships. PubMed from 1991 to 2003 include about four million abstracts.
Our method to extract technical terms is based on both linguistic and statistical approach as follows.

1. Morphological analysis
   We analyze all abstracts in PubMed by the GENIA Tagger[2] to chunk sentences into one-words and tag them with part of speech tags.
2. Extraction of specific part of speech tag sequences
   We extract one-word or compound terms with the following part of speech tag sequences:
   - 'NN* JJ* NN+ CD*'
   - 'CD* NN+'
   - 'NN+ VBG NN*'
   - 'NN+ VBD NN+'

   NN: Noun; JJ: Adjective; CD: Cardinal number; VBG: Verb,gerund/present participle; VBN: Verb,past participle; *: zero or more; +: one or more.
3. Normalization of extracted terms
   We normalize extracted terms which may be plurals and past participles. The GENIA Tagger outputs also base form candidates of chunked sentences. But because its output includes wrong base forms, we compare base forms candidates with other original forms of extracted terms. If there are same strings with base forms in original forms, we treat them as correct base forms and merge both forms. Otherwise we treat them as specific terms.

4. Statistical filtering of extracted terms
   We statistically filter extracted terms for extracting only technical terms expressing knowledge in biology. Our filter is to delete extracted terms with low frequency (appear in less than ten articles) in PubMed.

We merge extracted terms with index terms in textbook and furthermore MeSH terms[3], which are controlled vocabularies used in NCBI Entrez retrieval system.

We obtained technical terms shown in Table 1.

Table 1. Number of extracted terms from Pubmed articles

	Number of terms
1.chunked words	858,325,873
2.terms with specific POS	44,297,693
3.normalized terms	38,015,595
4.terms filtered statistically	1,522,145

Calculation of Term Co-occurrence frequency We calculate co-occurrence frequency in articles between index terms, extracted terms and MeSH terms with high occurrence frequency. In order to find significant association between terms, we set thresholds for absolute and relative co-occurrence frequency.

$$O_a, O_b \geq 10 \tag{1}$$

$$C_{ab} \geq 5 \tag{2}$$

$$J_{ab} \equiv \frac{C_{ab}}{O_a + O_b - C_{ab}} \geq 0.1 \tag{3}$$

$$\begin{aligned} O_a &: occurrence\ frequency\ of\ term\ a \\ C_{ab} &: co-occurrence\ frequency\ between\ term\ a\ and\ b \\ J_{ab} &: Jaccard\ coefficient\ between\ term\ a\ and\ b \end{aligned}$$

We obtained 372,133 relationships between 314,775 terms. We generate term co-occurrence networks, in which nodes represent terms and edges represent co-occurrence relationships between two terms. Term co-occurrence networks consist of a quite large-scale network, hundreds of medium-scale networks and innumerable small-scale networks as shown in Table 2. The large-scale network includes high-frequent terms such as 'cell' and 'protein', which have many relationships between other terms and result in large-scale networks.

Table 2. Statistics of term co-occurrence networks

	Number of networks	Total number of nodes	Mean number of edges per node	Total number of indices
Large-scale network(50522 nodes)	1	50522	4.0	2062
Medium-scale networks(30 - 400)	464	51655	4.2	460
Small-scale networks(2 - 29)	69654	219659	1.8	1689
All(2-50522)	70119	314775	2.4	4211

In Figure 3, we show an example of medium-scale network which consists of 32 terms, in which elliptic nodes represent index terms, circular nodes are terms extracted from PubMed articles and rectangular node is a MeSH term. Index terms in the Wallace's and Cell's sections about photosynthesis are included in this network. Other terms also are associated with photosynthesis or plant, and MeSH term 'pyruvate, orthophosphate dikinase' plays a critical role on vocabulary control around it.

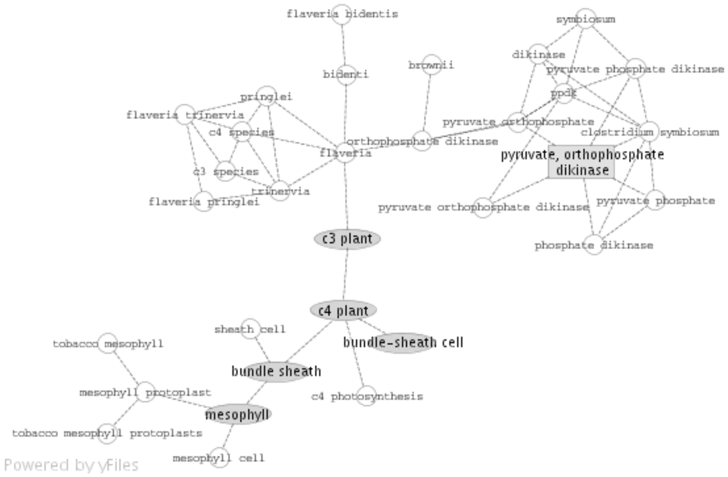

Fig. 3. Example of term co-occurrence network (a middle-scale network)

3 Evaluation of Knowledge Map

We evaluate the accuracy and performance of our knowledge map, especially term co-occurrence networks.

3.1 Accuracy of Geometry

In term co-occurrence network, two terms are laid at a distance of reciprocal of their relationship strength. The accumulation of relationships between two terms makes the whole knowledge space. On the other hand, tables of contents in textbook are accurate knowledge space made by hand. Thus we compare the geometry of term co-occurrence network with tables of contents structures in textbooks.

Here we define the distance between co-occurred terms which are linked directly (Equation 4). We also define the distance between arbitrary terms in the connected network, which may be linked directly or indirectly (Equation 5).

$$D(a,b) \equiv \frac{1}{J_{ab}} - 1 \tag{4}$$

$$d(a,b) \equiv \sum_{c,d\ edge\ in\ shortest\ path\ from\ a\ to\ b} D(c,d) \tag{5}$$

$$sim(a,b) \equiv: \frac{1}{1+d(a,b)} \tag{6}$$

$$\begin{aligned}
D(a,b) &: distance\ between\ co-occurred\ term\ a\ and\ b \\
d(a,b) &: distance\ between\ arbitrary\ term\ a\ and\ b\ in\ a\ connected\ network \\
sim(a,b) &: similarity\ between\ arbitrary\ term\ a\ and\ b \\
J_{ab} &: Jaccard\ coefficient\ between\ term\ a\ and\ b
\end{aligned}$$

Various quantitative measures for hierarchical structure are proposed[4][5]. But it is difficult to measure table of contents structure quantitatively without term similarities, which are to be compared with table of contents structure now. Thus we treat table of contents structure qualitatively, for example two sections are far or close.

In Figure 4, we consider two sections which are thought to be close with each other. One section is 'Energy Conversion: Mitochondria and Chloroplasts' in the Cell and other section is 'Photosynthesis' in the Wallace. Index terms in both sections are laid on a specific region in co-occurrence network.

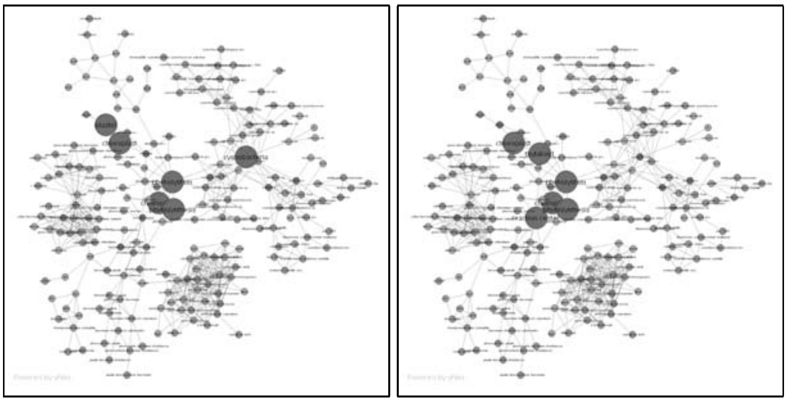

Fig. 4. Index terms in photosynthesis sections of Cell (left) and Wallace (right)

We develop the method to measure similarity between two sections in co-occurrence network (Equation 8).

$$S_a^s \equiv \sum_{b \in s} w(b) \times sim(a,b) \tag{7}$$

$$S(s,s') \equiv \sum_{a \in s'} w(a) \times S_a^s \; / \; normalized \tag{8}$$

$$S_a^s : \; similarity \; between \; term \; a \; and \; section \; s$$
$$S(s,s') : \; similarity \; between \; section \; s \; and \; s'$$
$$w(b) : \; weight \; of \; term \; b \; (\equiv \; log \; \frac{1}{occurrence \; ratio \; of \; b})$$

We compare between arbitrary sections in three textbooks. In Figure 5, two sections which are similar with each other in co-occurrence network are linked by thick lines. This figure shows that sections associated with similar subjects are close in co-occurrence network.

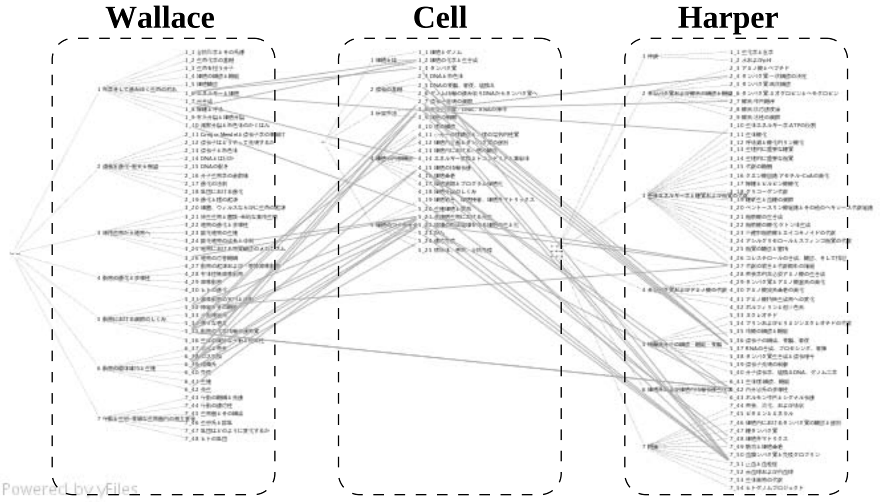

Fig. 5. Comparison between arbitrary sections in three textbooks

3.2 Availability of Term Expansion

Our proposed knowledge map is the combination of textbooks and term co-occurrence network. Textbooks include only fewer terms which are well-established, while term co-occurrence networks include many terms which are not necessarily associated with subjects of textbooks. Thus we expand terms associated with subjects of textbooks from term co-occurrence networks.
Figure 4 shows that terms to be expanded are laid within regions in which index terms

are densely laid. We use Equation7 to measure similarity between term and section. In Figure 6, we show expanded terms associated with 'Energy Conversion: Mitochondria and Chloroplasts' in the Cell as middle-sized nodes.

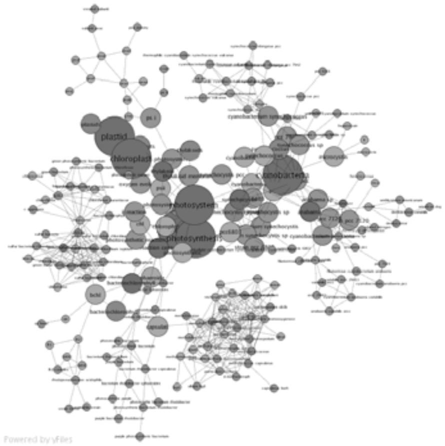

Fig. 6. Term expansion by co-occurrence network

4 Application: Classification of Resources in Biology

We classify resources such as articles and web pages on our knowledge map. In order to determine which sections a resource belongs to, we use the equation analogous to Equation 8.

$$S(s,r) \equiv \sum_{a \in r} w(a) \times S_a^s \; / \; normalized \tag{9}$$

$$S(s,r) : similarity\ between\ section\ s\ and\ resource\ r$$

We classified PubMed articles on our knowledge map. Figure 7, in which the more articles are classified into a section, the larger symbol on it are, shows that recent articles are associated with subjects as follows: immunity, genomics, signal transduction, apoptosis, cancer and so on.

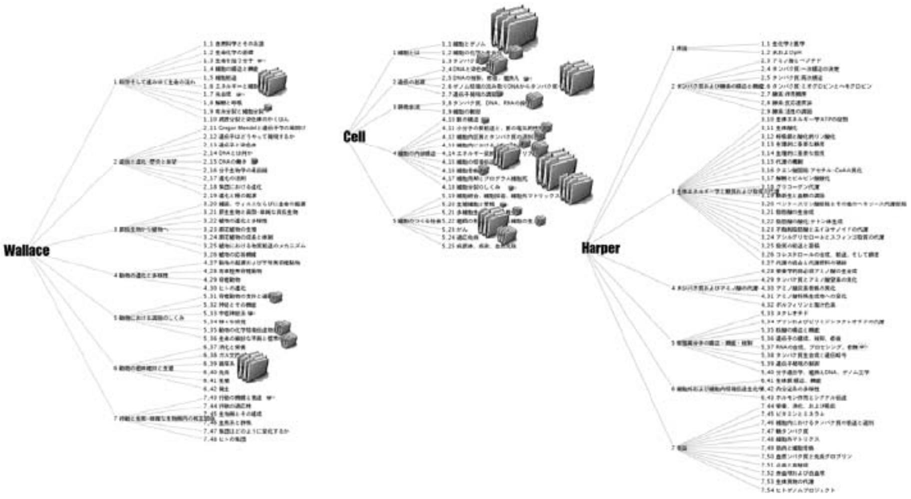

Fig. 7. Classification of PubMed articles on the knowledge map

5 Conclusions

We constructed the knowledge map to get an entire picture of life or biology. We combine textbooks' tables of contents with term co-occurrence networks for the map to include well-established knowledge besides newly-found knowledge. Our term co-occurrence network has good accuracy and performance so that we can classify many resources on out knowledge map.

References

1. PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed).
2. Tsuruoka, Y., Tateishi, Y., Kim JD., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a Robust Part-of-Speech Tagger for Biomedical Text. In the Proceedings of the 10th Panhellenic Conference on Informatics.(2005) (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/).
3. MeSH (http://www.nlm.nih.gov/mesh/)
4. Load, PW., Stevens, RD., Brass, A., Goble, CA.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinfomatics. **19**(10) (2003) 1275-83
5. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res. **11** (1998) 95-130

This article was processed using the LaTeX macro package with LLNCS style