J. Biomedical Science and Engineering, 2016, 9, 7-16
Published Online January 2016 in SciRes. http://www.scirp.org/journal/jbise
http://dx.doi.org/10.4236/jbise.2016.91002

Application of Word Embedding to Drug Repositioning

Duc Luu Ngo[1], Naoki Yamamoto[1], Vu Anh Tran[1], Ngoc Giang Nguyen[1], Dau Phan[1], Favorisen Rosyking Lumbanraja[1], Mamoru Kubo[2], Kenji Satou[2]

[1]Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa, Japan
[2]Institute of Science and Engineering, Kanazawa University, Kanazawa, Japan
Email: ndluu@blu.edu.vn, n-0325@stu.kanazawa-u.ac.jp, tvatva2002@gmail.com, giangnn.bkace@gmail.com, pdaukg@gmai.com, favorisen@gmail.com, mkubo@t.kanazawa-u.ac.jp, ken@t.kanazawa-u.ac.jp

Received 2 December 2015; accepted 18 January 2016; published 21 January 2016

Abstract

As a key technology of rapid and low-cost drug development, drug repositioning is getting popular. In this study, a text mining approach to the discovery of unknown drug-disease relation was tested. Using a word embedding algorithm, senses of over 1.7 million words were well represented in sufficiently short feature vectors. Through various analysis including clustering and classification, feasibility of our approach was tested. Finally, our trained classification model achieved 87.6% accuracy in the prediction of drug-disease relation in cancer treatment and succeeded in discovering novel drug-disease relations that were actually reported in recent studies.

Keywords

Distributed Representation of Word Sense, Discovery of Drug-Disease Relation, Word Analogy

1. Introduction

To develop an effective and highly-demanded drug, hundreds million dollars and 10 or more years for R & D and clinical trial are typically required. Structure-based drug design (SBDD) is actively studied to reduce the cost and time by *in-silico* screening of candidate chemicals [1] [2]; however, still it requires long time for tests on animals and human. Against such a background, a concept of drug repositioning (or drug repurposing, reprofiling, etc.) is attracting much interest and expectation from academic researchers and pharmaceutical companies [3]. One of the famous examples of drug repositioning is a treatment of multiple myeloma by thalidomide that was initially developed for relieving nausea and vomiting in pregnancy. Since drug repositioning means reuse of approved drugs for another purpose, their safety and method of production have already been confirmed.

How to cite this paper: Ngo, D.L., Yamamoto, N., Tran, V.A., Nguyen, N.G., Phan, D., Lumbanraja, F.R., Kubo, M. and Satou, K. (2016) Application of Word Embedding to Drug Repositioning. *J. Biomedical Science and Engineering*, **9**, 7-16.
http://dx.doi.org/10.4236/jbise.2016.91002

Besides biomedical experiments, computational methods are developed for drug repositioning. Most of them adopt network-based algorithms and combination of various databases including gene expression and pathway data [4]. On the other hand, it is also suggested that text mining has much potential for drug repositioning. In biomedical text mining, named entities (genes, proteins, etc.) are recognized and the relations among them are extracted (e.g. "Gefitinib"<inhibit>"EGFR"). Additionally, biomedical ontologies or WordNet [5] are utilized for the sources of semantic information. In this study, we applied word embedding, implemented as word2vec [6]-[8], for efficient representation of semantic information of words in a sufficiently large subset of PubMed abstracts. Through the clustering and classification experiments especially on anti-cancer drugs and cancer-related diseases, it is suggested that the word vectors, generated by word embedding for drugs and diseases, are representing rich semantic information and promising for drug repositioning.

2. Materials and Methods

In this section, data and algorithms are described. Overview of processing pipeline is shown in **Figure 1**.

2.1. Raw Corpus as a Set of Sentences

As a raw corpus, we used a subset of PubMed abstracts downloaded in October 2013, filtered by the keyword "cancer". From 3,099,076 abstracts, 14,847,050 sentences were extracted.

2.2. Parsing

Enju [9] was used for POS recognition of words and conversion into base forms. Since the sentences were extracted from biomedical abstracts, "-genia" option was specified. As a result, part-of-speech (POS) and base form are recognized for each words.

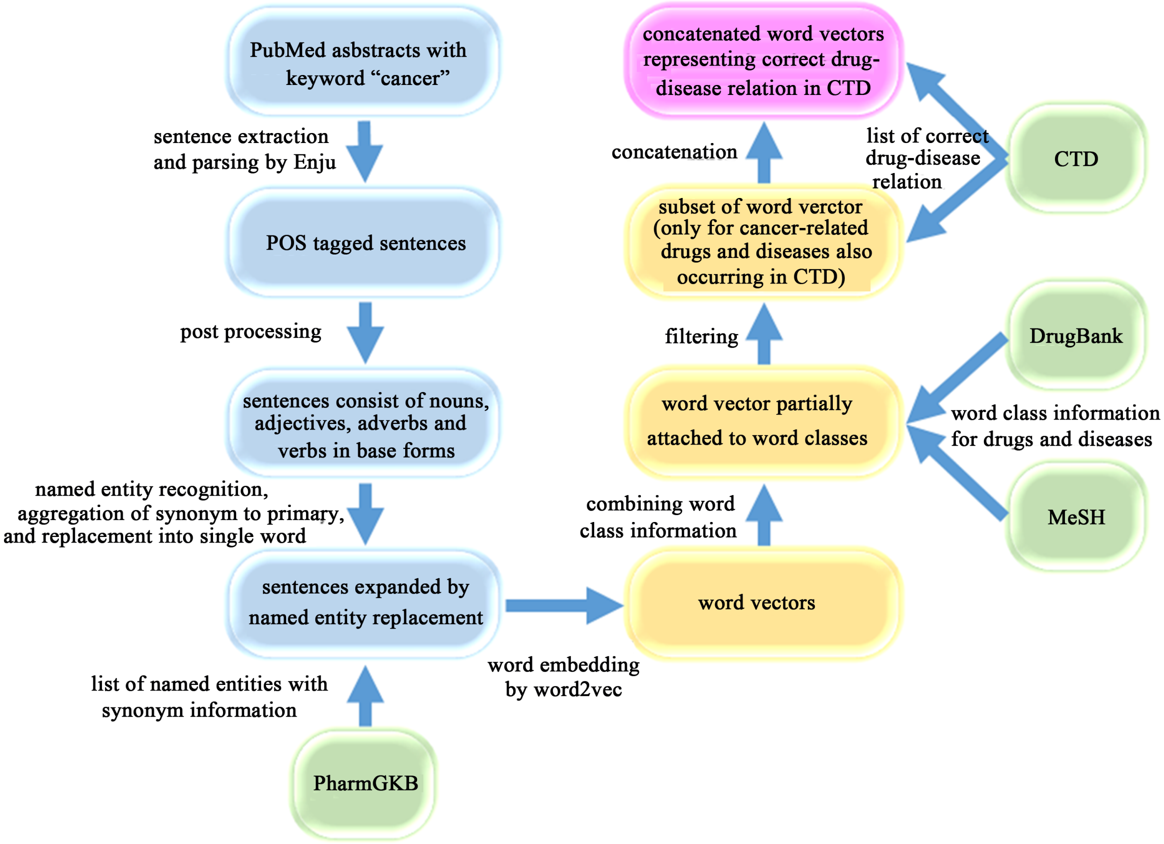
PubMed asbstracts with keyword "cancer"
sentence extraction and parsing by Enju
POS tagged sentences
post processing
sentences consist of nouns, adjectives, adverbs and verbs in base forms
named entity recognition, aggregation of synonym to primary, and replacement into single word
sentences expanded by named entity replacement
list of named entities with synonym information
PharmGKB
word embedding by word2vec
word vectors
combining word class information
word vector partially attached to word classes
word class information for drugs and diseases
DrugBank
MeSH
filtering
subset of word verctor (only for cancer-related drugs and diseases also occurring in CTD)
list of correct drug-disease relation
CTD
concatenation
concatenated word vectors representing correct drug-disease relation in CTD

Figure 1. Overview of processing pipeline. Box colors indicate: light blue for corpus, light green for databases, yellow for word vectors, and pink for concatenated word vectors.

2.3. Post Processing of Parsing Result

So that word2vec can differently treat the same word with different POS categories, they were attached right after the base form of words (e.g. "care" ->"care(V)"). For readability, nouns are kept as is. To simplify the input for word2vec, we removed all words except nouns, adjectives, adverbs, and verbs.

2.4. Named Entity Recognition and Conversion into Single Words

Biological terms typically consist of two or more words. In addition, they have many synonyms. Since word2vec basically treats a sentence as a sequence of words, it is needed to recognize biological synonyms, aggregate them into primary terms, and convert them into single words (e.g. "yolk sac tumor" ->"endodermal sinus tumor" ->"endodermal_sinus_tumor"). In this study, primary names and synonyms of drugs, diseases, and genes were extracted from PharmGKB [10] and used for recognition and aggregation (genes are used only for showing distribution of word vectors). For each converted single words, prefixes indicating their semantic categories were attached for later processing (e.g. "endodermal_sinus_tumor" ->"disease::endodermal_sinus_tumor").

Related to the conversion above, we need to consider about the existence of original single words. Firstly, if a synonym word is aggregated into primary word, the original word disappears and is not used for word embedding. Secondly, if a multi-word term is converted into a single word, all the original single words in the multi-word term disappear. Thirdly, if two multi-word term occur in a sentence with overlapping, it is impossible to replace both of them at once. To avoid there problems, a sentence is converted into the sentences which containing at most one converted word per sentence. For example, if a sentence contains two terms to be converted, three sentences including original one are generated. After all conversion, 14,847,050 sentences are expanded to 45,264,480.

2.5. Word Embedding

In the field of natural language processing and text mining, computational representation of a linguistic unit (e.g. documents, paragraphs, sentences, terms, and words) is essential. The simplest one for document is bag-of-word model, which represents each document as a vector of word frequencies in it. In case of word representation, only the neighboring words in the same sentence are counted. For better analysis, stop-words are removed and raw frequencies are modified by term weighting such as tf-idf. After that, these vectors are used to evaluate the characteristics of the units and similarities between them (vector space model).

One of the serious problems in such a representation and analysis is high dimensionality and sparseness of vectors. For instance, 10 millions of sentences may contain one million of different words, then the dimension of a vector is also one million. In addition, since frequency of word follows Zipf's law, most of the one million of words only occur a few times, which makes the vectors quite sparse. Though there exist traditional algorithms for dimension reduction or compression like Principal Component Analysis (PCA) and Latent Semantic Analysis (LSA), this problem is not fully solved.

Word embedding for distributed representation of word sense is a new approach to this problem. Based on neural network algorithm, reasonably short numerical vectors (e.g. 100 dimensions) are calculated for all words in a set of sentences. Through the application studies, it is proved that the vector space constructed by word embedding represents word senses and distances (similarities) between them quite well. Additionally, in this space of word sense, word analogy works well in some domains. For example, given three words "man", "woman", and "king", word analogy could predict "queen" by calculating $vector(\text{“}man\text{”}) - vector(\text{“}woman\text{”}) + vector(\text{“}king\text{”})$ and searching for the nearest word vector *vector*("queen"). Though word analogy might allow wide variety of applications, the most desired one is discovery of unknown relations.

In this study, we used word2vec software, a de facto standard implementation of word embedding algorithm, with the following parameters by default.

- vector size = 200
- window size = 8
- minimum count of words to be embedded = 1 (*i.e.* all words)
- model = continuous bag of words

As a result, 1,772,186 words were embedded into word vectors (2303 for drugs, 3069 for diseases, 8703 for genes, and 1,758,111 for others).

2.6. Word Classes

For the evaluation of clustering results, ATC codes [11] and MeSH tree numbers [12] were attached to drug and disease names, respectively. ATC codes were extracted from DrugBank [13]. Due to the incompleteness of data annotation, only 1253 drugs out of 2303 and 2745 diseases out of 3069 have such classification.

2.7. Drug-Disease Relations

For the evaluation of difference vectors between drugs and diseases, relations between drugs and diseases occurring in the corpus were extracted from CTD [14]. Only the 12,462 relations with therapeutic evidences were adopted for obtaining trustable results. In the set of relations, the mapping from drugs to diseases is many-to-many. For example, "drug::gefitinib" is related to 17 different diseases, and "disease::lung_neoplasm" is mapped from 60 different drugs.

In order to conduct detailed analysis on cancer-related drugs and diseases, 12,462 extracted drug-disease relations were further filtered so that both of drug and disease names in each relation are attached to an ATC code and a MeSH tree number beginning with "L" (Antineoplastic and immunomodulating agents) and "C04" (Neoplasms), respectively. As a result, 1097 relations consist of 104 anti-cancer drugs and 107 cancer-related diseases were extracted for detailed analysis.

2.8. Clustering

For visual evaluation of word vector quality, we performed hierarchical clustering with cosine distance and Ward's method [15]. Before the clustering, 2303 drugs and 3069 diseases occurring in the corpus were reduced to 1282 and 1051, respectively, since other drugs and diseases did not occur in CTD.

2.9. Classification

Support Vector Machine (SVM) was adopted for learning and predicting possible relations between drugs and diseases. As an implementation, ksvm function included in kernlab package for R software was used with default parameters.

3. Experimental Results

3.1. Distribution of Word Vectors

Figure 2 illustrates the 3D plot of vectors corresponding to 2303 drugs, 3069 diseases, and 8703 for genes. For visualization, the dimension of vector was reduced from 200 to 3 by PCA. In the left panel of the figure, it is

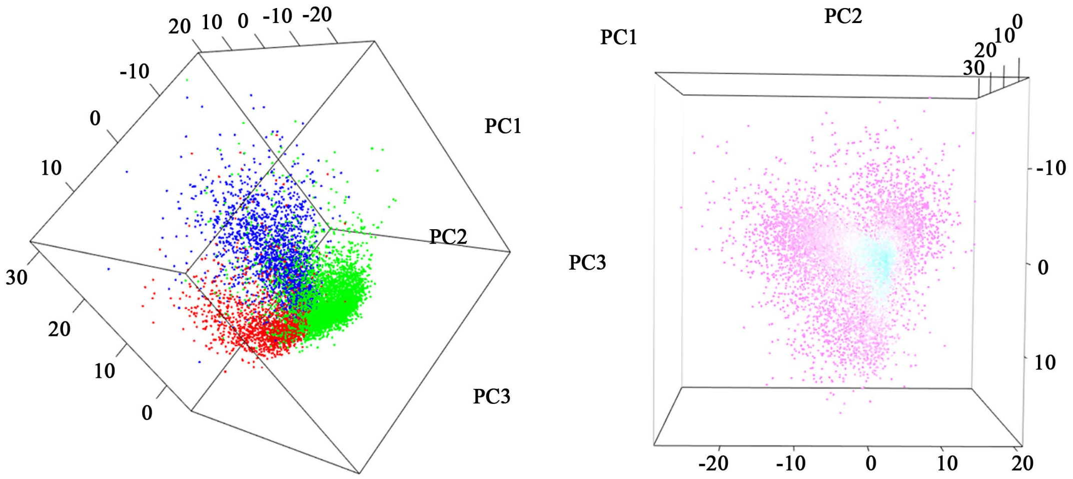

Figure 2. Distribution of word vectors visualized through PCA and 3D plot. Left panel: blue, red, green colors indicate word vectors for drugs, diseases, and genes. Right panel: color gradation from light blue to light pink indicates the frequency of words (from rare to frequent).

shown that the distributions of word vectors in three categories are clearly separated. In the right panel, it is also shown that the frequent words have clear separation, whereas it is relatively difficult to discriminate the categories of rare words.

3.2. Cluster Analysis

Figure 3 and **Figure 4** show the result of hierarchical clustering for drugs and diseases, respectively. In the right panels of them, entire pictures of clustering results for 1282 drugs and 1051 diseases are shown. In the right panel of **Figure 3**, it can be seen that most of the anti-cancer drugs are condensed in the ninth cluster from the top (left panel for more detail). It indicates that the word vectors for drugs well represent the characteristics of corresponding drugs. Also in **Figure 4**, we can see that the seventh cluster from the top contains a number of cancer-related diseases, however, also in sixth and ninth clusters. The difference between these results might come from the fact that diseases can be classified from different perspectives (tissues, mechanism, etc.).

3.3. Applicability of Simple Word Analogy

Though word analogy is quite attractive, it does not always works well. To evaluate the applicability of word analogy to the discovery of new relation between drug and disease, we checked whether most of the displacement vectors between confirmed drug-disease pairs (*i.e.* correct relations) are similar in length and parallel to each other or not. Unfortunately, as shown in **Figure 5**, the displacement vectors have wide range of lengths and directions. It indicates that the simple application of word analogy to drug repositioning cannot achieve high performance.

3.4. Classification of Correct and Incorrect Drug-Disease Relations

Instead of simple application of word analogy, we constructed a classification model using SVM. For all combinations of 104 anti-cancer drugs and 107 cancer-related diseases (*i.e.* 11,128 drug-disease pairs), drug vectors and disease vectors were concatenated and binary class labels (*i.e.* positive or negative) were added according to 1097 correct drug-disease relations extracted from CTD. Due to the imbalance of two classes, 1097 out of 10,031 negative examples were randomly selected so that the numbers of positive and negative examples are balanced.

The result of performance evaluation is shown in **Table 1**. Each accuracy is an average of 100 times 10-fold cross-validation with different subsets of negative examples. In this table, it was revealed that the performance was not so affected by vector size and window size, and the best accuracy was 87.6%. For exploratory use of the classification model to discover candidate drug-disease pairs, it means sufficiently high performance.

Finally, we tested all combinations of 2199 drugs not used in training and 107 cancer-related diseases (in total, 235,293 drug-disease pairs). In case of the classification model trained by 11,128 examples, only 64 test examples were predicted as positive, and all the drugs in the examples were anti-cancer drugs (but not included in 104 anti-cancer drugs used for training). By controlling the degree of class imbalance in training data, it is possible to predict a pair of non-anti-cancer drug and cancer-related disease as positive. For example, using the classification model trained by 1097 positive and 8776 negative examples (degree of imbalance is 1:8), 10 times training and test by 235,293 drug-disease pairs discovered the following candidate drugs for repositioning to cancer treatment, where the numbers indicate how many times they were discovered in 10 times training and test.

Table 1. Accuracy of classifying correct and incorrect drug-disease relations by SVM.

Vector size	Accuracy			
	Window size = 2	Window size = 3	Window size = 4	Window size = 8
50	0.872	0.873	0.875	0.872
75	0.873	0.874	0.874	0.874
100	0.874	0.874	0.874	**0.876**
200	0.874	0.874	0.874	0.874

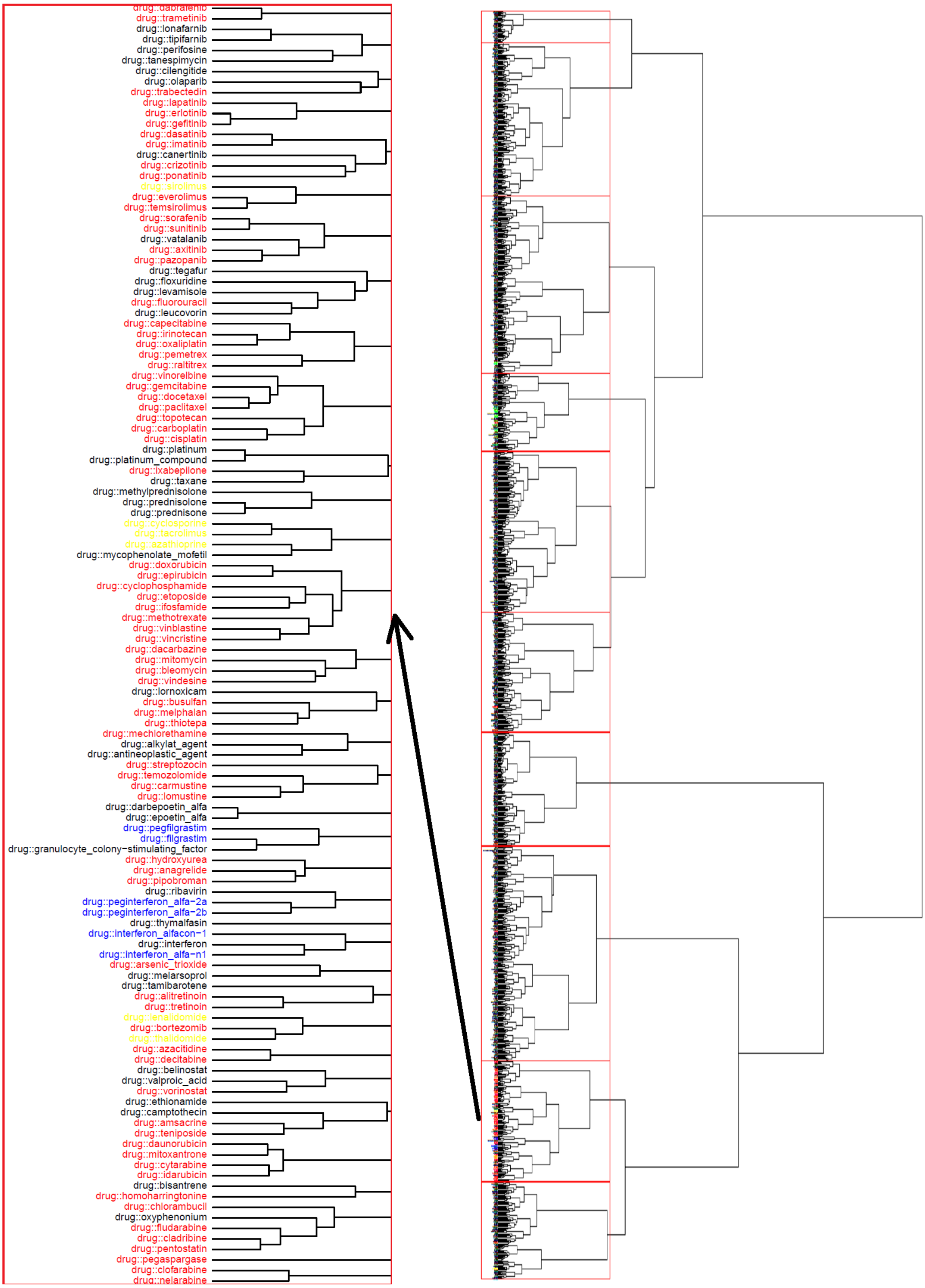

Figure 3. Result of hierarchical clustering on drugs. Red, green, blue, yellow colors for characters indicate that the drugs are classified in ATC codes as “L01: Antineoplastic Agents”, “L02: Endocrine Therapy”, “L03: Immunostimulants”, and “L04: Immunosuppressants”, respectively.

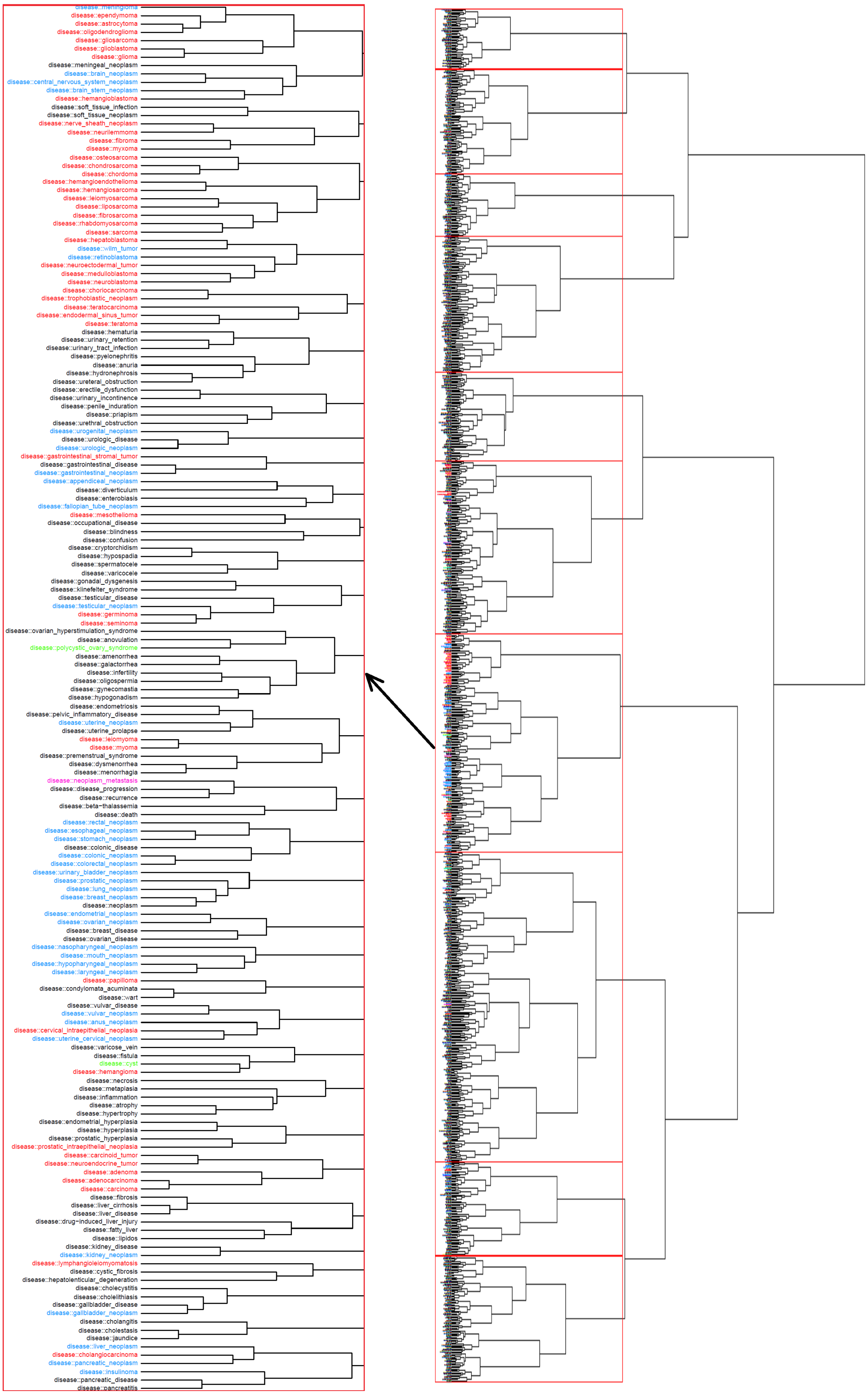
disease::meningioma
disease::ependymoma
disease::astrocytoma
disease::oligodendroglioma
disease::gliosarcoma
disease::glioblastoma
disease::glioma
disease::meningeal_neoplasm
disease::brain_neoplasm
disease::central_nervous_system_neoplasm
disease::brain_stem_neoplasm
disease::hemangioblastoma
disease::soft_tissue_infection
disease::soft_tissue_neoplasm
disease::nerve_sheath_neoplasm
disease::neurilemmoma
disease::fibroma
disease::myxoma
disease::osteosarcoma
disease::chondrosarcoma
disease::chordoma
disease::hemangioendothelioma
disease::hemangiosarcoma
disease::leiomyosarcoma
disease::liposarcoma
disease::fibrosarcoma
disease::rhabdomyosarcoma
disease::sarcoma
disease::hepatoblastoma
disease::wilm_tumor
disease::retinoblastoma
disease::neuroectodermal_tumor
disease::medulloblastoma
disease::neuroblastoma
disease::choriocarcinoma
disease::trophoblastic_neoplasm
disease::teratocarcinoma
disease::endodermal_sinus_tumor
disease::teratoma
disease::hematuria
disease::urinary_retention
disease::urinary_tract_infection
disease::pyelonephritis
disease::anuria
disease::hydronephrosis
disease::ureteral_obstruction
disease::erectile_dysfunction
disease::urinary_incontinence
disease::penile_induration
disease::priapism
disease::urethral_obstruction
disease::urogenital_neoplasm
disease::urologic_disease
disease::urologic_neoplasm
disease::gastrointestinal_stromal_tumor
disease::gastrointestinal_disease
disease::gastrointestinal_neoplasm
disease::appendiceal_neoplasm
disease::diverticulum
disease::enterobiasis
disease::fallopian_tube_neoplasm
disease::mesothelioma
disease::occupational_disease
disease::blindness
disease::confusion
disease::cryptorchidism
disease::hypospadia
disease::spermatocele
disease::varicocele
disease::gonadal_dysgenesis
disease::klinefelter_syndrome
disease::testicular_disease
disease::testicular_neoplasm
disease::germinoma
disease::seminoma
disease::ovarian_hyperstimulation_syndrome
disease::anovulation
disease::polycystic_ovary_syndrome
disease::amenorrhea
disease::galactorrhea
disease::infertility
disease::oligospermia
disease::gynecomastia
disease::hypogonadism
disease::endometriosis
disease::pelvic_inflammatory_disease
disease::uterine_neoplasm
disease::uterine_prolapse
disease::leiomyoma
disease::myoma
disease::premenstrual_syndrome
disease::dysmenorrhea
disease::menorrhagia
disease::neoplasm_metastasis
disease::disease_progression
disease::recurrence
disease::beta−thalassemia
disease::death
disease::rectal_neoplasm
disease::esophageal_neoplasm
disease::stomach_neoplasm
disease::colonic_disease
disease::colonic_neoplasm
disease::colorectal_neoplasm
disease::urinary_bladder_neoplasm
disease::prostatic_neoplasm
disease::lung_neoplasm
disease::breast_neoplasm
disease::neoplasm
disease::endometrial_neoplasm
disease::ovarian_neoplasm
disease::breast_disease
disease::ovarian_disease
disease::nasopharyngeal_neoplasm
disease::mouth_neoplasm
disease::hypopharyngeal_neoplasm
disease::laryngeal_neoplasm
disease::papilloma
disease::condylomata_acuminata
disease::wart
disease::vulvar_disease
disease::vulvar_neoplasm
disease::anus_neoplasm
disease::cervical_intraepithelial_neoplasia
disease::uterine_cervical_neoplasm
disease::varicose_vein
disease::fistula
disease::cyst
disease::hemangioma
disease::necrosis
disease::metaplasia
disease::inflammation
disease::atrophy
disease::hypertrophy
disease::endometrial_hyperplasia
disease::hyperplasia
disease::prostatic_hyperplasia
disease::prostatic_intraepithelial_neoplasia
disease::carcinoid_tumor
disease::neuroendocrine_tumor
disease::adenoma
disease::adenocarcinoma
disease::carcinoma
disease::fibrosis
disease::liver_cirrhosis
disease::liver_disease
disease::drug−induced_liver_injury
disease::fatty_liver
disease::lipidos
disease::kidney_disease
disease::kidney_neoplasm
disease::lymphangioleiomyomatosis
disease::cystic_fibrosis
disease::hepatolenticular_degeneration
disease::cholecystitis
disease::cholelithiasis
disease::gallbladder_disease
disease::gallbladder_neoplasm
disease::cholangitis
disease::cholestasis
disease::jaundice
disease::liver_neoplasm
disease::cholangiocarcinoma
disease::pancreatic_neoplasm
disease::insulinoma
disease::pancreatic_disease
disease::pancreatitis

Figure 4. Result of hierarchical clustering on diseases. Green, yellow, red, dodger blue, pink, blue, and spring green colors for characters indicate that the diseases are classified in MeSH tree numbers as “C04.182: Cysts”, “C04.445: Hamartoma”, “C04.557: Neoplasms by Histologic Type”, “C04.588: Neoplasms by Site”, “C04.697: Neoplastic Processes”, “C04.730: Paraneoplastic Syndromes”, and “C04.834: Precancerous Conditions”.

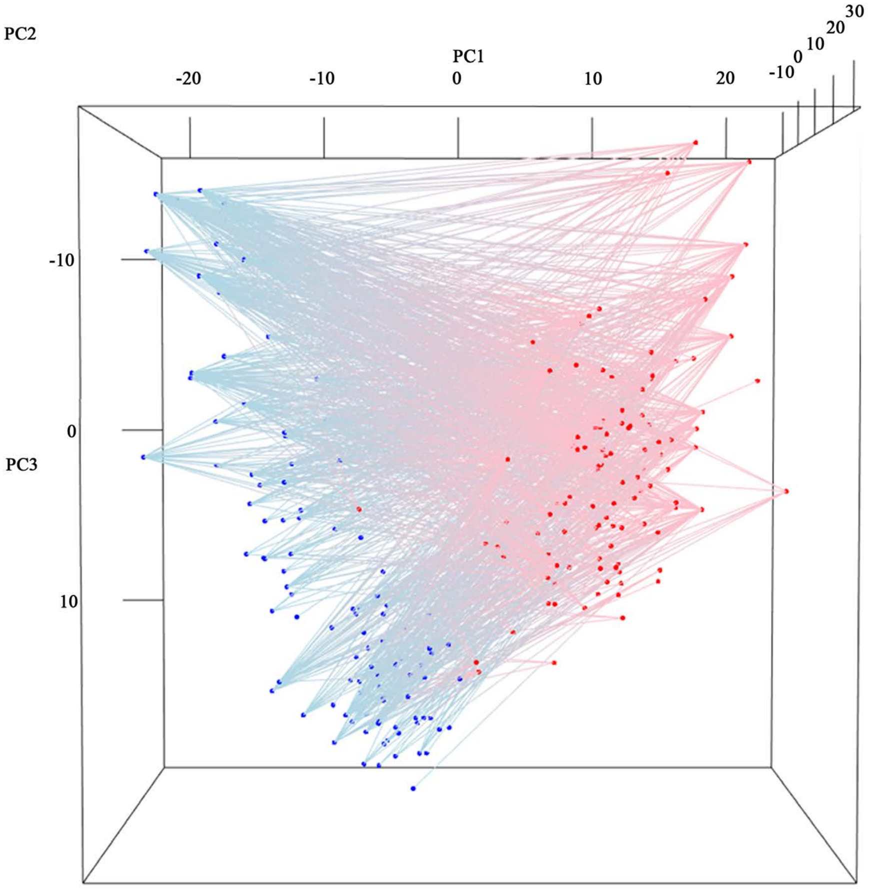

Figure 5. Distribution of displacement vectors for cancer-related drug-disease relations in CTD. Blue and red points represent anti-cancer drugs and cancer-related diseases.

drug::urokinase (10), drug::photodynamic_therapy (10), drug::oxygen (10), drug::nonoxynol-9 (10), drug::nitroglycerin (10), drug::nitrogen (10), drug::l-phenylalanine (10), drug::l-methionine (10), drug::l-glutamine (10), drug::l-cysteine (10), drug::glutathione (10), drug::glucose (10), drug::epoxide (10), drug::enzyme (10), drug::collagenase (10), drug::bisphosphonate (10), drug::amino_acid (10), drug::amide (10), drug::clarithromycin (9), drug::vitamin (8), drug::l-proline (7), drug::vitamin_e (6), drug::xanthophylls (4), drug::phospholipid (4), drug::palifermin (4), drug::ether (4), drug::ethacrynic_acid (4), drug::denosumab (4), drug::egfr_inhibitor (2), drug::pyruvic_acid (1).

Besides too general names like "drug::enzyme" and "drug::amide", it is notable that the above list includes approved anti-cancer drugs (e.g. "drug::denosumab"), anti-cancer drugs under investigation (e.g. "drug:: clarithromycin", "drug::bisphosphonate", and "drug::xanthophyll"), and drugs potentially promote cancer (e.g. "drug::urokinase" and "drug::collagenase"). Especially, it should be emphasized that repositioning of clarithromycin to anti-cancer agent has been reported in 2015 [16], despite the fact that the corpus was downloaded in 2013. Though further screening based on expert's knowledge is necessary, this result demonstrate that the classification of concatenated word vector is a promising approach to in-silico screening of drug-disease relations for drug repositioning.

4. Discussion and Conclusion

One of the reasons why word embedding by word2vec becomes popular is its functionality of word analogy [6]. For example, if a sufficient amount of corpus is converted into word vectors and used in the analogy, it could predict the fourth word "California" from three given words "Chicago", "Illinois", and "Stockton". Since a state for a city is unique, it works well: it readily means that the analogy easily fails for one-to-many relationship (e.g.

predicting "Stockton" from "Illinois", "Chicago", and "California"). About drug-disease relationship, at first we expected that one drug is used for basically one disease. However, as shown in **Figure 5**, it was one-to-many from both sides of drug-disease relation. For another problem like gene-protein relationship, accuracy of word analogy might be high since only one protein is produced from one gene, ignoring alternative splicing.

Although word analogy was not available, word2vec provided significant advantage in the text mining from a large number of biomedical texts in this study. It efficiently encoded more than 1.7 million words into quite short vectors (e.g. 200 dimensions). If we use traditional word frequency and vector space model, one vector for a word is a vector of 1.7 million features with extremely high sparsity. Due to the efficiency of encoding, we could process the whole corpus in reasonable memory space and computation time. Furthermore, the word vectors generated by word2vec seem to well reflect the semantic space of biomedical words. **Figure 2** illustrates that the words in different semantic categories are well separated in case of sufficiently high frequency of occurrences. Also, the results of clustering shown in **Figure 3** and **Figure 4** indicate that similarities among words in the same category are also fine. They might be promising results for further application of word embedding in biomedical text mining.

In this study, it was revealed that word embedding is effective for representing sense of all words in a large number of cancer-related PubMed abstracts. Furthermore, concatenation of word vectors of drugs and diseases well represents their relations and could be used for finding candidate drugs for repositioning by classification. For better performance of classification, various feature selection and over-sampling algorithms [17] will be tested in the future work.

Acknowledgements

In this research, the super-computing resource was provided by Human Genome Center, the Institute of Medical Science, the University of Tokyo. Additional computation time was provided by the super computer system in Research Organization of Information and Systems (ROIS), National Institute of Genetics (NIG). This work was supported by JSPS KAKENHI Grant Number 26330328.

References

[1] Ferreira, L.G., dos Santos, R.N., Oliva, G. and Andricopulo, A.D. (2015) Molecular Docking and Structure-Based Drug Design Strategies. *Molecules*, **20**, 13384-13421. http://dx.doi.org/10.3390/molecules200713384

[2] Bajorath, J. (2015) Computer-Aided Drug Discovery. *F1000Research*, **4**, 630. http://dx.doi.org/10.12688/f1000research.6653.1

[3] Ashburn, T.T. and Thor, K.B. (2004) Drug Repositioning: Identifying and Developing New Uses for Existing Drugs. *Nature Reviews Drug Discovery*, **3**, 673-683. http://dx.doi.org/10.12688/f1000research.6653.1

[4] Emig, D., Ivliev, A., Pustovalova, O., Lancashire, L., Bureeva, S., Nikolsky, Y. and Bessarabova, M. (2013) Drug Target Prediction and Repositioning Using an Integrated Network-Based Approach. *PLoS ONE*, **8**, e60618. http://dx.doi.org/10.1371/journal.pone.0060618

[5] Fellbaum, C. (1998) WordNet: An Electronic Lexical Database. MIT, Cambridge, MA.

[6] Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*. arXiv:1301.3781v1

[7] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) Distributed Representations of Words and Phrases and Their Compositionality. *Proceedings of NIPS*. arXiv:1301.3781v3

[8] Mikolov, T., Yih, W.T. and Zweig, G. (2013) Linguistic Regularities in Continuous Space Word Representations. *Proceedings of NAACL HLT*, 746-751.

[9] Miyao, Y. and Tsujii, J. (2008) Feature Forest Models for Probabilistic HPSG Parsing. *Computational Linguistics*, **34**, 35-80. http://dx.doi.org/10.1162/coli.2008.34.1.35

[10] Whirl-Carrillo, M., McDonagh, E.M., Hebert, J.M., Gong, L., Sangkuhl, K., Thorn, C.F., Altman, R.B. and Klein, T.E. (2012) Pharmacogenomics Knowledge for Personalized Medicine. *Clinical Pharmacology & Therapeutics*, **92**, 414-417. http://dx.doi.org/10.1038/clpt.2012.96

[11] WHO Collaborating Centre for Drug Statistics Methodology (2015) ATC Classification Index with DDDs. WHO Collaborating Centre, Oslo.

[12] Lipscomb, C.E. (2000) Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, **88**, 265.

[13] Wishart, D.S., Knox, C, Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z. and Woolsey, J. (2006)

DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Research*, **34**, D668-D672. http://dx.doi.org/10.1093/nar/gkj067

[14] Davis, A.P., Grondin, C.J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B.L., Wiegers, T.C. and Mattingly, C.J. (2015) The Comparative Toxicogenomics Database's 10th Year Anniversary: Update 2015. *Nucleic Acids Research*, **43**, D914-D920. http://dx.doi.org/10.1093/nar/gku935

[15] Xu, R. and Wunsch, D.I.I. (2005) Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, **16**, 645-678. http://dx.doi.org/10.1109/TNN.2005.845141

[16] Pantziarka, P., Bouche, G., Meheus, L., Sukhatme, V. and Sukhatme, V.P. (2015) Repurposing Drugs in Oncology (ReDO)-Clarithromycin as an Anti-Cancer Agent. *eCancer Medical Science*, **9**, 513. http://dx.doi.org/10.3332/ecancer.2015.521

[17] Dang, X.T., Hirose, O., Bui, D.H., Saethang, T., Tran, V.A., Nguyen, T.L.A., Le, T.T.K., Kubo, M., Yamada, Y. and Satou, K. (2013) A Novel Over-Sampling Method and Its Application to Cancer Classification from Gene Expression Data. *Chem-Bio Informatics Journal*, **13**, 19-29. http://dx.doi.org/10.1273/cbij.13.19