Interpreting Gene Profiles from Biomedical Literature Mining with Self Organizing Maps

Shi Yu, Steven Van Vooren, Bert Coessens, Bart De Moor

Department of Electrical Engineering
University of Leuven, Belgium

Abstract. We present an approach to interpret gene profiles derived from biomedical literature using Self Organizing Maps (SOM). Comparison of different clustering algorithms shows that SOMs perform better in grouping high dimensional gene profiles, even when a lot of noise is present in the data. Qualitative analysis of the clustering results proof that SOMs allow an in-depth interpretation of gene profiles with biological relevance.

1 Introduction

Text mining tools have become indispensable for the biomedical sciences, especially to extract the fundamental textual patterns of biomedical entities inherent in the data of literature. Based on these textual patterns, further analysis is useful to aid the lab research in a more straightforward manner. However, even facing in the results of a well designed text mining system, researchers are still overwhelmed by the paradox of "information overload": the more data we gain from ever growing experiments and researches, the less knowledge we are relatively able to interpret from the data. The large amount of varied tools for data analysis also makes people confused to choose the optimal one for specific problem. In this paper, we present a self organizing maps (SOMs) based application by mining gene profiles from biomedical literature and clustering those profiles with reference to biomedical topics. Though SOMs has been introduced in gene expression clustering analysis, however, the profiles of gene we use in our approach are different from those sequence and clinical based expressions. These profiles are extracted purely from textual resources, for instance, the MEDLINE database of biomedical literature. Hence the application we discuss in this paper represents a text mining approach of bioinformatics study.

2 Gene profiling by mining biomedical literature

Our approach is divided into two stages. The first stage is the extraction of gene profiles by literature mining. A corpus collection of biomedical literature titles and abstracts is investigated We use the TXTGate platform [2] to create textual profiles of genes with different scopes. More details about this mining system are discussed in [2]. The gene profiles we use in this paper are produced by controlled

vocabulary indexing with Gene Ontology. The mining process is conducted by a list of 32,814 human genes and a pruned vocabulary of 9,175 Gene Ontology terms (GO terms). These GO terms are pruned from the standard GO terms by segmentation and indexing: The segmentation process splits the long GO terms, for example, "regulation of DNA recombination" into smaller terms in addition to the original term: "regulation", "of", "DNA" and "regulation DNA recombination". After that, the literature corpus is indexed with all segmented GO terms. Only those terms that occurred in the corpus at least once, are kept. Thus it leads to a $32{,}814 \times 9{,}175$ matrix of gene-by-GO terms. By referencing Gene Ontology, each gene is associated to a certain number (usually $300 \sim 400$) of segmented GO terms. The textual score of a gene is hence represented as a vector, where each elements is represented as the inverse document frequency (IDF) scores of corresponding GO terms in the corpus. The GO term's IDF score is static in a given corpus, however, due to the variety of numbers of associated terms for genes, the resulted gene profiles are different after normalization.

The computational problem of these gene profiles for functional analysis is obvious because of its high dimensionality. This dimensionality curse is severe when comparing two gene profiles by similarity measure. By using latent semantic analysis techniques it is possible to reduce the data into an appropriate size, however, according to our experiments, even in reduce models, distance based measures fail to give prominence to strongly related genes in a view of biomedical topics. And it is also unreliable to retrieve results only from samples because rankings and grouping methods of similarity are very sensitive to noisy data. In this sense, topic based research is a more reasonable solution to interpret the mining results. By grouping gene profiles into biomedical topics, it allows to investigate the boundary of several topics in a feature space concerning of the relationships among genes. Moreover than interpretation, it is also possible to train discriminative classifiers for topic grouping and predicate the functions or features of genes when new data arrives. Thus, our goal in this paper is to find a better method to group gene profiles into biomedical topics and explore some potential factors for the improvement of the research.

3 Gene clustering by self organizing maps

3.1 Candidate gene profiles for clustering in Self Organizing Maps

SOMs has a number of features that make them particularly well in clustering and analysis of gene expression patterns. They are ideally suited to exploratory data analysis, allowing one to impose partial structure on the clusters and facilitating easy visualization and interpretation[5]. In this paper, we introduce this tool in clustering textual mining profiles and compare it with other clustering methods which are widely used in pattern analysis.

We select 36 genes from 5 biomedical topics about genetic and immunological diseases. The selection of topic related genes is guided by Entrez genome database (Table 1). The result datasets of different topics are varied in size thus we have both large clusters and small clusters. We are unaware of whether

the size of topic is an influencing factor in clustering so we continue with the experimental scheme here. We will review this problem in later section.

Table 1. Manual clusters of 36 candidate genes from 5 biomedical topics

Breast Cancer	*Alzheimers*	*Parkinson*	*Leukaemia*	*Wolf Hirschhorn Syndrome*
BRCA1	APOE	CYP2D6	BCL11A	WHSC1
BRCA2	PSEN1	MAPT	BCL11B	WHSC2
BRCA3	PSEN2	PARK2	MLL	WHSC1L1
SNCG	APP	PARK7	TNFRSF8	MSX1
NCOA3	SNCA		BLNK	
TFF1	ACE		NR0B2	
AMPH			NOTCH1	
BCAS2			CD86	
ABCG2			P2RX7	
BRMS1			SEL1L	
SDBCAG84			CDK9	

3.2 Clustering Result and Comparison with Other Methods

The gene profiles are trained in SOMs to generate 5 clusters. The configuration of SOMs is gaussian neighborhood function, square grid neuron map, initial vector randomized and trained for 50000 epochs. Figure 1 shows the clustering result. The qualitative analysis of the clustering result shows that the gene profiles are geometrically grouped into several areas: Alz(*Alzheimers*) genes are gathered at the left side and Wol(*Wolf-Hirschhorn syndrome*) genes are gathered at right side. In the middle part, other topics and their related genes are allocated into continuous distributions. Thus, SOMs present us a intuitive understanding of topic distribution.

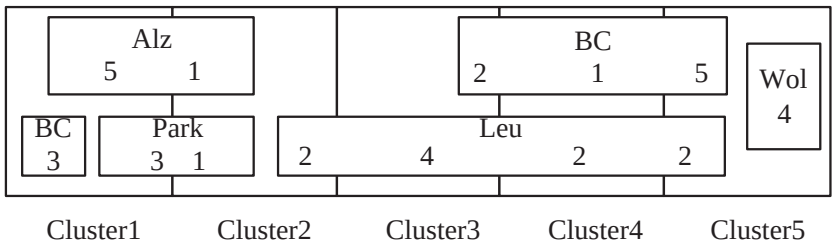

Fig. 1. The SOMs clustering result of genes in Table 1. The number marked in each cluster square represents the number of genes from the topic is grouped in this cluster

After a small qualitative analysis of SOMs in interpreting gene profiles, next step we use cluster evaluation method explaining the effect of SOMs clustering in comparison with other algorithms in grouping gene topics. We try BCUBED[1], which was originally used for evaluating coreference resolution. It uses the complementary of distance measure to calculate the similarity between two clusters, where an "accurate" cluster is manually selected and a "evaluating" cluster is automatically generated by algorithms. The complement of the distance from the automatic cluster away from the manual one is considered as the "fitness" of clustering (Figure 2).

BCUBED

This algorithm counts the number of entities that should be added 1) to the automatic output in order to arrive to the manual one (recall measure) and 2) to the manual output in order to arrive to the automatic one (precision measure). Formally, given N objects:

a) for each object(entity) *i:*

$$precision_i = \frac{number\ of\ correct\ elements\ in\ the\ automatic\ cluster\ containing\ i}{number\ of\ elements\ in\ the\ automatic\ cluster\ containing\ i}$$

$$recall_i = \frac{number\ of\ correct\ elements\ in\ the\ automatic\ cluster\ containing\ i}{number\ of\ elements\ in\ the\ manual\ cluster\ containing\ i}$$

b) $FinalPrecision = \sum_{i=1}^{N} \omega_i * precision_i$

$FinalRecall = \sum_{i=1}^{N} \omega_i * recall_i$

where ω_i are weights to measure the relative importance of each entity. All ω_i should sum to one and they are often chosen as 1/N (1/36 in this case)

Table 2 compares the statistics of BCUBED scores of SOMs, hierarchical clustering and K-Means in clustering 416 genes selected from the dataset to be mentioned in next section. The value in each cell is expressed in the form of "mean(variance)". The BCUBED scores are calculated in 3 different ways. The first row(M1) of the table are the BCUBED scores of pure gene profile clusters between manual cluster, which corresponds to the complement of distance d_1 in Figure 2. The second row(M2) measures the similarity of clustering results of noisy gene profiles between manual cluster. This result is obtained by clustering the pure gene profiles together with the same number of randomly selected genes from the 1,432 most popular genes occurred in literature. Thus, we actually cluster a double number of gene profiles and later remove those random genes when calculating BCUBED scores with manual cluster. Hence, the complement of d_2 can be regarded as the clustering ability in noisy data. Finally, the third row(M3) measures the similarity of pure profile clustering between noisy profiles clustering (the complement of d_3), which evaluates the preservation ability of clustering methods in noisy data. We are especially interested in the comparison

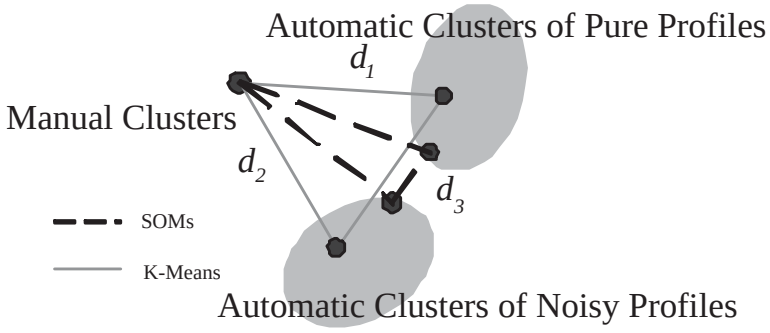

Fig. 2. The geometrical illustration of evaluation in Table 2. It represents the meaning of 3 different measures M1-M3 in Table 2. The two triangles show the difference between KMeans and SOMs clustering according to ANOVA test in Table 2.

between SOMs and KMeans because these two methods show a similar ability in clustering genes in M1 and M2. The HCL methods are obviously faulty because they produce too many singleton clusters, which also explains why they have exceptional high values of recall: BCUBED score doesn't penalize enough to the occurrence of singleton clusters. The ANOVA test shows that SOMs is more robust than KMeans in preserving genes within biomedical topics in noisy gene clustering (F-value 27.67 of Precision and 13.32 of Recall in M3 in Table 2). We can thereby make a conclusion that topic groups clustered in SOMs are insensitive to other irrelevant genes, which is a interesting feature for topic based analysis of gene profiles.

All the repetitive experiments in Table 2 are conducted under a blocking principle: to keep topics in similar size. In this sense the topics are not selected in a fully randomized manner. Combination of large topics with small topics is avoided for clustering. However, according to our results, if the topics are stongly varied in size, for instance, 3 topics having more than 20 genes and 2 others having only 5 genes, SOMs performs much better than other methods in this case. We also compare the clustering results of profiles in full dimensional model and reduced models. We use Lantent Semantic Indexing (LSI) to reduce the model into size of 500, 300 and 200 rather the full size of 9,175 and compare the result. Moreover, we increase both the number of selected topics and the required clustering from 5 gradually to 10 and compare the clustering results again. All these consideration will be planned with carefully designed experiments and analyzed in multiple factors ANOVA analysis, where topic size, model size, topic number are all considered as new level factors, in our future works.

4 Topic map construction of gene profiles

In this section a big map is used for profiles interpretation. In order to avoid the tedious interpretation of fully randomized gene profiles, we again manually select

Table 2. BCUBED evaluation and ANOVA test of clustering results

		SOMs	*HCL Average*	*HCL Complete*	*K-Means*
M1	Precision	0.4165 (0.0044)	0.3932 (0.0093)	0.4275 (0.0091)	0.3762 (0.0040)
	Recall	0.4365 (0.0027)	0.6447 (0.0121)	0.5224 (0.0054)	0.3713 (0.0043)
M2	Precision	0.4072 (0.0069)	0.3607 (0.0081)	0.4278 (0.0064)	0.3927 (0.0063)
	Recall	0.4179 (0.0100)	0.7021 (0.0144)	0.5408 (0.0102)	0.3997 (0.0063)
M3	Precision	0.7284 (0.0059)	0.7027 (0.0181)	0.6588 (0.0058)	0.4135 (0.0084)
	Recall	0.7056 (0.0174)	0.7942 (0.0197)	0.6765 (0.0083)	0.4277 (0.0058)

		F Value(All)	*Pr > F(All)*	*F Value(KM-SOM)*	*Pr > F(KM-SOM)*
M1	Precision	0.34	0.7953	0.78	0.4036
	Recall	9.11	0.0009	2.42	0.1582
M2	Precision	0.49	0.6974	0.06	0.8082
	Recall	9.30	0.0009	0.77	0.4049
M3	Precision	8.75	0.0011	27.67	0.0008
	Recall	7.70	0.0021	13.32	0.0065

profiles with biomedical topic labels from Entrez. A map with 50×80 neurons is used to cluster 416 gene profiles selected from 40 biomedical topics. The map is trained and visualized by ESOM[6]. Due to the capacity and load of training software, the 9,175 feature space of data is reduced by latent semantic methods to 300 terms. The training process iterates 5,000 epochs and costs 60 hours on 64-bit Unix Server with 8G memory. Notice here if the map is for data prediction, the dataset must be kept in full dimension and training will take much longer time. And by now we haven't found handy tools to compare the interpretation of maps trained by full dimensional data between those trained by reduced size data. A small test of predication accuracy, which we use half of the data to train the map and another half labeled data to validate the predication results, does not tell much difference between maps trained in different size of data. We believe that this is not the true nature but also requires us to find better tools to validate the predication error of SOMs. Since the scope of this paper is about interpretation so the works mentioned above will be considered in future works. In Figure 3, it shows that the distribution of neurons and matched data in a well trained map is also well stretched. The dark area represents the activated neuron, deeper color represents the high density of data matches. Inversely, the light area represents the inactivated (gap) neurons between activated neurons and lighter colors stand for bigger gaps. Since the numbers of genes selected from different biomedical topics are varied, the map is biased towards larger topics with more genes selected such as Cro($Crohn'sDisease$), Alb($Albinism$), etc.

We divide our analysis of the map into 2 levels. At the 1st level, we check the matched data at each neuron and mark its labeled biomedical topic on the map.

An intuitive look of this map shows that many genes form local topic groups. Some example groups are marked with white polygons in the map. A re-labeling of these topic groups will produce the 2nd level interpreted map which allows a more intuitive interpretation of gene profiles. We also find that most topic groups are distributed in the edge of map rather than in the center. The center of map seems as a "meeting point" that all kind of genes are mixed together. This coincides to the clustering result in small test in Section 3 and Figure 1. Moreover, we construct a very simple ontology for the 40 selected topics and reference it in our map. If shows that some genes from relevant biomedical topics stay together. For example, Ost(*Osteopetrosis*), Der(*Dermatomyositis*) and Mus(*Muscular Dystrophies*) all belong to "Musculoskeletal Diseases". Correspondingly, they stay close to each other in the centered black line curve in Figure 3. Similarly, genes of "Skin and Tissue disease", for instance, Ehl(*Ehlers-Danlos Syndrome*), Sys(*Systemic Lupus Erythematosus*) and Scl(*Scleroderma*) stay close in the left upper black line curve of the map. Some genes of "immunological disease and blood disease", for example, Aga(*Agammaglobulinemia*), Wis(*Wiskott-Aldrich Syndrome*), Dig(*DiGeorge Syndrome*), Sjo(*Sjogren's Syndrome*) and Sic(*Sickle cell anemia*)stay together in the right lower black line curve.

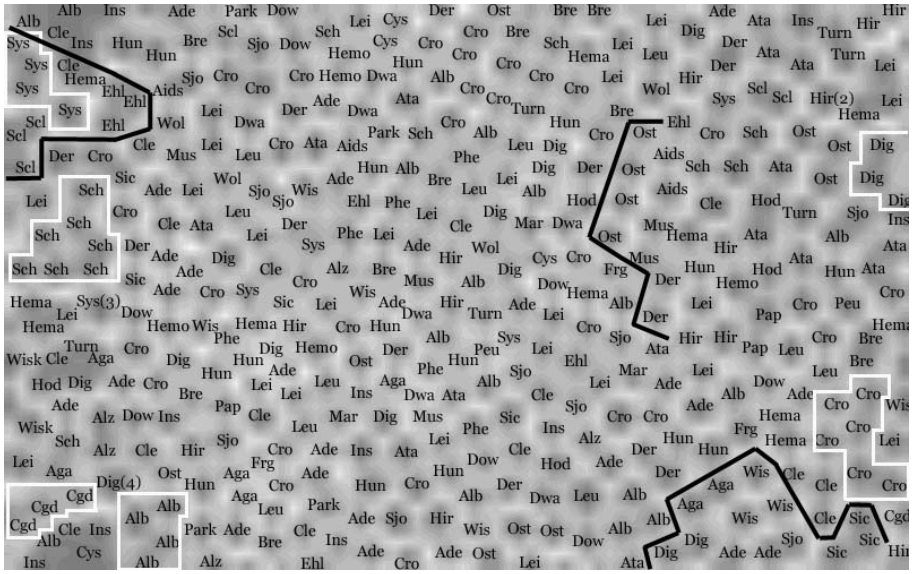

Fig. 3. The 1st level interpretation of a 50X80 neurons map trained with 416 genes selected from 40 biomedical topics.

A possible disadvantage of this study is the inaccuracy of finding designators in biomedical topics. It is possible that some weakly related genes play much more important role in another topic but we manually classify them into an im-

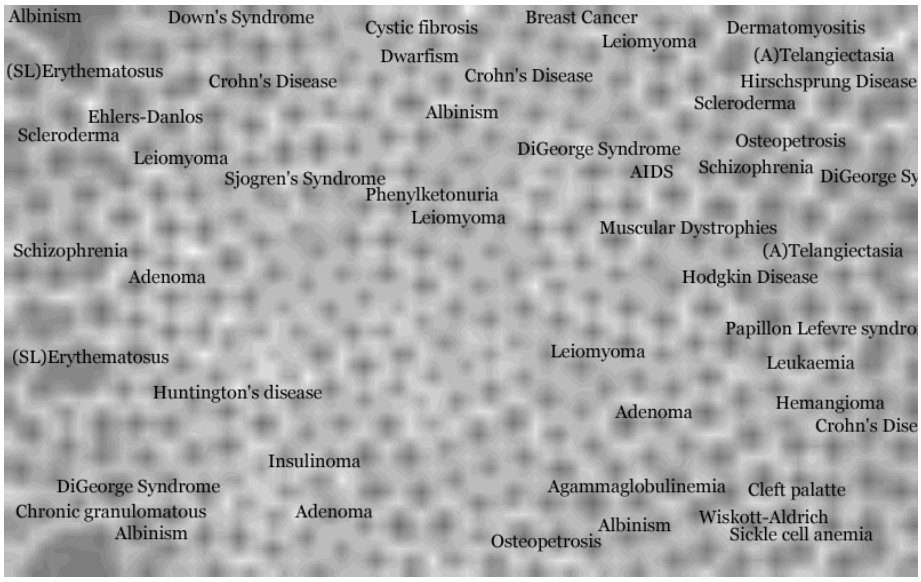

Fig. 4. The 2nd level interpretation of the map in Figure 3 labeled with gene groups

propriate class. Another problem is that there are quite a lot of genes strongly related to multiple topics, for example, gene TP53 is related to Ata(*Ataxia Telangiectasia*) as well as Ade(*Adenoma*). And it is also important to care about the topic selection, for example, Ade(*Adenoma*) actually is a higher level topic than Ins(*Insulinoma*) and Lei(*Leiomyoma*), thus many genes which belong to the sub topics are also related to the higher topics. It is necessary to make sure that the selected topics are representative both in medical categorization and importance. It is also important that the number of their selected genes for SOMs interpretation is almost same to avoid bias in final map interpretation.

At last, it is an interesting hypothesis that the SOMs might produce something like the globe map. As we know, a spatial continuous area if is split in certain area and represented on a 2-dimensional map, this area will occur multiple times at the edges of map. In our map of gene profiles intrepretation, if we fold the left upper corner and left downer corner together, we will see topic areas of Alb(*Albinism*) and Ins(*Insulinoma*)merging together. Similarly, if we do the same thing at right corner, folding up down, topic areas of Hir(*Hirschsprung Disease*) and Ade (*Adenoma*) will come together. It is just a guess that it might be more reasonable to stretch the map on a globe surface (or other kind of spatial shapes) rather than only on 2-dimensional plan surface.

5 Conclusion and Future Works

In this paper we used SOMs to interpret gene profiles mined from biomedical literature. We found that SOMs provide a more reliable and visualized map of

gene profiles according to their features space of GO terms indexing and thus we are able to gain in-depth knowledge about gene relationships and functions. It performs much better than other algorithms we studied to preserve exisiting topics in noisy data. This feature is meaningful to our interpretation task of text mining because it allows us to analyze the results by small samples or incremental study without losing the true knowledge inherent in result. Since our study is still in early stage, the current methodology of our approach is not optimal and much efforts need to be considered in several aspects. We have mentioned some possibilities of improvement in paper and a breif conclusion is made here:

a) To create a refined evaluation method of textual gene clustering. This paper used an coreference evaluation borrowed from text retrieval area to evaluate gene clustering results. Since any evaluation method has the blind point, we need to find a more appropriate method for gene cluster evaluation. A good benchmark dataset is required to conduct this research. Moreover, necessary biological case studies are extremely helpful to prove the validity of the new evaluation method.
b) Some genes analyzed in our experiments are related to multiple biomedical topics. So it is much likely that a different neurons are activated near the data in feature space. However, traditional "winner take all" competitive learning algorithm only allows one neuron to be activated given one instance so it is difficult to present important genes in the trained map. This may be an interesting topic either for the imporvment of learning algorithm of SOMs or finding a post-clustering process, for example, fuzzfying algorithm to allow a single gene occur in multiple clusters.
c) To conduct some well designed experiments to study the influence of multiple level factors in textual gene profile clustering, as mentioned in section 3.
d) Of course, we can also make effort to improve our textual mining results. As mentioned in section 2, the biomedical literature is mined by 9,175 GO terms. Is it possible to construct a more effective and concise vocabulary for textual mining? We can either change the principle of GO segmentation and indexing to only keep the terms meaningful or clustering, which will reduce the computational load much. We can also try other vocabulary instead of GO terms to index the literature. Although these considerations are not directly related to SOMs clustering, it is also possible to try and regard them as factors in statistical investigation of results.

References

1. Baldwin, B., Morton, T., Bagga, A., Baldridge, J., Chandraseker, R., Dimitriadis, A., Snyder, K., Wolska, M.: Description of the upenn cam system as used for coreference. In: Proceedings of the Seventh Message Understanding Conference (MUC-7). Kaufmann, San Mateo CA(1998)
2. Glenisson, P., Coessens, B., Van Vooren, S., Mathys, J., Moreau, Y., De Moor, B.: TXTGate: Profiling gene groups with text-based information. Genome Biology. 5(6) (2004) 1-12

3. Van Hulle, M.: Faithful Representations and Topographic Maps: From Distortion to Information Based Self Organization. John Wiley & Sons Press (2002)
4. Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., Saarela, A.: Self Organization of a Massive Document Collection. IEEE Trans. Neural Networks 11(3) (2000) 574-585
5. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., Golub, T.: Interpreting Patterns of Gene Expression with Self-organizing Maps: Methods and Application to Hematropoietic Differentiation. Genetics 96(6) (1999) 2907-2912
6. Ultsch, A., Moerchen, F.: ESOM-Maps: Tools for Clustering, Visualization, and Classification with Emergent SOM. Technical Report No. 46. Dept. of Mathematics and Computer Science, University of Marburg, Germany(2005)