Studying the Cohesion Evolution of Genes Related to Chronic Lymphocytic Leukemia Using Semantic Similarity in Gene Ontology and Self-Organizing Maps

Efstratios Kontopoulos[1], Theodoros Moysiadis[2], Maria Tsagiopoulou[2], Sándor Darányi[3], Peter Wittek[3,4], Nikos Papakonstantinou[2], Stavroula Ntoufa[2], Georgios Meditskos[1], Kostas Stamatopoulos[2], Ioannis Kompatsiaris[1]

[1] Information Technologies Institute, Thessaloniki, Greece
{skontopo, gmeditsk, ikom}@iti.gr

[2] Institute of Applied Biosciences, Thessaloniki, Greece
{moysiadis.theodoros, m.tsagiopoulou, npapakonstantinou, sntoufa, kostas.stamatopoulos}@certh.gr

[3] University of Borås, Sweden
Sandor.Daranyi@hb.se

[4] ICFO – The Institute of Photonic Sciences, Spain

Abstract. A significant body of work on biomedical text mining is aimed at uncovering meaningful associations between biological entities, including genes. This has the potential to offer new insights for research, uncovering hidden links between genes involved in critical pathways and processes. Recently, high-throughput studies have started to unravel the genetic landscape of chronic lymphocytic leukemia (CLL), the most common adult leukemia. CLL displays remarkable clinical heterogeneity, likely reflecting its underlying biological heterogeneity which, despite all progress, still remains insufficiently characterized and understood. This paper deploys an ontology-based semantic similarity combined with self-organizing maps for studying the temporal evolution of cohesion among CLL-related genes and the extracted information. Three consecutive time periods are considered and groups of genes are derived therein. Our preliminary results indicated that our proposed gene groupings are meaningful and that the temporal dimension indeed impacted the gene cohesion, leaving a lot of room for further promising investigations.

Keywords: Chronic Lymphocytic Leukemia, Gene Ontology, Semantic Similarity, Semantic Drift, Self-Organizing Maps.

1 Introduction

A significant and continuously growing body of work on biomedical text mining is aimed at uncovering meaningful associations between biological entities, including genes (e.g. [1]-[3]). This line of research promises to deliver new scientific insights, uncovering hidden links between genes involved in critical pathways and processes.

C*hronic lymphocytic leukemia* (*CLL*), the most common adult leukemia in the West, is a chronic, incurable malignancy of mature B lymphocytes. CLL displays remarkable clinical heterogeneity linked to the underlying biological complexity that, despite recent progress, remains insufficiently characterized and understood [4]. Moreover, CLL has not been subjected to analysis with text mining tools, though such approach holds great promise for uncovering hidden knowledge within the wealth of published biomolecular information from CLL patient profiling studies.

Towards addressing this vital challenge, in this paper we propose a novel methodology for studying genes and their potential relationship to CLL. Based on a set of CLL-related PubMed abstracts, we are deploying a state-of-the-art software tool for generating groups of genes belonging to three consecutive time periods. After measuring the semantic similarity of genes within the groups, we investigate the types of explicit and implicit information that can be derived from each group. This information could be valuable in characterising groups of interest. Furthermore, and since the groups are expected to evolve over time, we are also investigating the impact inflicted on the groupings by the temporal dimension, mainly focusing on the temporal evolution of the cohesion among CLL-related genes and the relevant extracted information. Temporal refers to the comparison of the results among the three periods. All our implementations are based on established open-source software tools.

Our preliminary results indicate that, compared to random group generation, the groupings we considered within each time period are indeed meaningful. Naturally, in some cases this difference is more emphatically reflected and statistically validated. Furthermore, the temporal dimension indeed impacted the gene groupings. In particular, derived groups of genes with biological importance were not consistent throughout the three time periods considered, and in some cases the changes are due to explicit biological explanation.

2 Related Work

A significant body of work exists on the area of biomedical text mining, namely the application of information extraction and text mining approaches on large amounts of biomedical literature with the goal of uncovering novel associations between biological entities of interest such as genes. For instance, in [1] the authors carried out automated extraction of explicit and implicit biomedical knowledge from publicly available gene and text databases, in order to create a gene-to-gene co-citation network for 13,712 named human genes by automated analysis of titles and abstracts in over 10 million MEDLINE[1] records.

In [2] the authors presented a method for creating a network of gene co-occurrences from the literature (article abstracts) and for partitioning it into communities of related genes, expected to be related by function. They noted that the derived gene communities are not meant to perfectly reproducing biological reality.

A couple of other related approaches are [3] and [5]: The former explored the utility of Latent Semantic Indexing (LSI) to automatically identify conceptual gene rela-

[1] https://www.nlm.nih.gov/pubs/factsheets/medline.html

tionships from titles and abstracts in MEDLINE citations, while the latter presented GenCLiP, a tool that searches gene lists to identify functional clusters of genes based on up-to-date literature profiling.

LSI is based on globally optimal dimensionality reduction, namely, singular value decomposition. If we reduce the dimensions to two, we gain visual insight on the layout of the data, but a globally optimal embedding disregards the local topologies of the data points. Complete isometric feature mapping (Isomap [6]), t-Distributed Stochastic Neighbor Embedding (t-SNE [7]), and self-organizing maps (SOM [8]), among many other techniques, provide a two-dimensional embedding that preserves the local topology, at the expense of distorting the global alignment of data points. SOM is particularly useful in online updates.

The above approaches aim to facilitate researchers to swiftly search for known interactions and to provide insight into unexplored connections. This is also one of our aims in this paper. However, we strongly believe that it is even more important to detect potential semantic drifts within the groups of genes, and, if such drifts occur, to evaluate them. Thus, we are not solely interested in gene groupings, but rather in revealing implicit information and assessing occurring drifts. To the best of our knowledge, this novel problem has not been tackled yet in biomedical sciences.

3 Semantic Similarity and Applications in Biomedicine

Semantic similarity is a measure for calculating the likeness of the meaning or semantic content of a set of terms or documents. One can use dictionaries, semantic networks, thesauri and, more recently, ontologies as the underlying resource for measuring semantic similarity between terms.

The various metrics for calculating semantic similarity are grouped in four categories: (a) *Path-based* metrics, where the similarity between two terms depends on their relative position in the underlying taxonomy, as well as on the length of the path linking the concepts (e.g. [9, 10]); (b) *Content-based* metrics are based on the information content available for each concept; the more common information two concepts share, the more similar they are (e.g. [11, 12]); (c) *Feature-based* metrics are based on the properties of the underlying ontology for obtaining a similarity value. The more common characteristics two concepts have, the more similar they are (e.g. [13]); (d) *Hybrid* metrics combine the ideas presented above (e.g. [14]).

Especially in the biomedical domain, measuring semantic similarity constitutes a core process towards information extraction and information retrieval. Due to its vast popularity, the *Gene Ontology* (*GO*) [15] typically plays the role of the underlying ontology model [16, 17]. Some of the most popular tools for calculating the semantic similarity in GO terms include *GOssTo* [18], *GOSemSim* [19] and *G-SESAME* [20]. The latter is a very popular set of online tools and was deployed in this work as the tool for measuring semantic similarities of genes in GO.

4 Methodology

Based on previous preliminary work of ours in a different domain [21], we are proposing a novel methodology for studying the temporal evolution of cohesion among CLL-related genes, which is composed of the following steps (see Fig. 1):

1. A set of input documents relevant to CLL are fed to *Somoclu* [22], a massively parallel open-source implementation of SOMs.
2. With the help of an accompanying input set of gene terms, Somoclu generates a two-dimensional (2D) map of gene clusters, based on the terms' co-occurrence in the set of input documents.
3. The semantic similarities of the gene terms residing in the groups are measured using the *G-SESAME* set of online tools (see previous subsection). G-SESAME was chosen due to its novel *Aggregate Information Content* (*AIC*) semantic similarity metric that relies on the information content of all ancestor terms of a GO term [23]. AIC implicitly reflects the GO term's location in the GO graph and effectively represents how humans use this term and its ancestor terms to annotate genes.
4. A set of semantically meaningful gene groupings is derived, along with potentially useful implicit information.

The above series of steps is clearly not restricted to CLL, and can be easily applied for input documents relevant to any malignancy of interest.

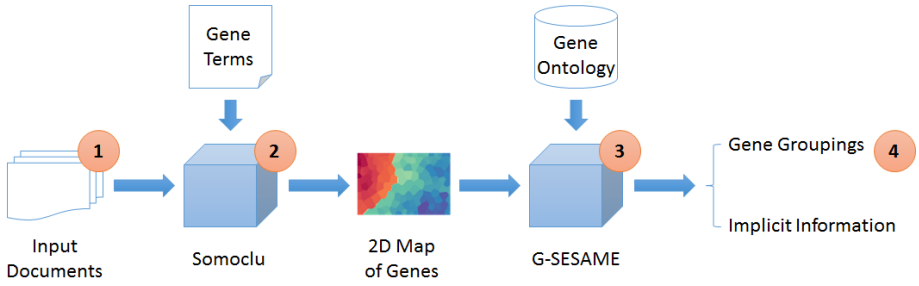

Fig. 1. Proposed methodology steps.

We build a vector space model by random indexing of the evolving text collection and use Somoclu to train a self-organizing map (step 2) that is able to closely track the changes. Random indexing captures term co-occurrence patterns similarly to an ordinary term-document matrix, but the resulting vector space is inherently lower dimensional – we used two hundred dimensions. The SOM training procedure used Euclidean distances to map points to the 2D grid. The input corpus is split into time periods (see next subsection) and a vector space is built for each period. We use *Emergent Self Organizing Maps* (*ESOMs*) [24] to project the vector space to a 2D surface where clusters and shifts are more apparent. Because ESOMs reproduce the local but not the global topology of the high-dimensional space [8], the derived clusters are locally meaningful and consistent on a neighbourhood level only.

5 Experiment Design

In order to study the temporal evolution of the semantic cohesion among CLL-related genes, we used a dataset of PubMed abstracts on CLL (downloaded in April'16), which we divided in three periods, after selecting two milestones with significant impact to the bibliography that followed them. The milestones are: (a) The end of the *Human Genome Project* (*HGP*) [25], which signalled the complete sequencing of the human DNA for the first time and revealed a big part of today's known genes; (b) The first appearance of *Next-Generation Sequencing* (*NGS*) [26] technologies in CLL, which allowed an entire human genome to be sequenced within a single day. Thus, the resulting three periods are: *Period 1* – From 3/1949 (first CLL publication) to 5/2004 (the end of HGP): 4390 abstracts; *Period 2* – From 6/2004 (first publication after HGP) to 7/2010 (appearance of NGS methods): 2567 abstracts; *Period 3* – From 8/2010 (first NGS publication for CLL) to 4/2016: 3456 abstracts.

The list of all known human genes, which served as the underlying vocabulary for Somoclu, was downloaded from the HUGO Gene Nomenclature Committee (HGNC)[2]; besides the "official" gene symbols, the latter also includes a list of synonymous symbols referring to each gene, which were also considered in our analysis.

5.1 CLL Gene Groupings

Based on the input corpus and list of gene terms, Somoclu generates 2D maps of the derived clustering of genes for the three periods (see Fig. 2 – term locations are indicated with white dots – term labels are omitted here). As already mentioned, clusters are locally meaningful and consistent on a neighbourhood level only.

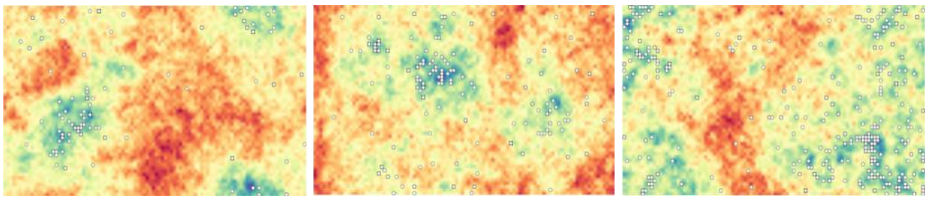

Fig. 2. 2D maps of gene groupings per period, visualized by Somoclu. First period is on the left, second period in the middle and third period is on the right.

Drift indications are also highlighted in the visualization: content splitting tendencies are indicated by the ridge wall width and height around the clusters, so that the method yields an overlay of two aligned contour maps in change, i.e. content structure vs. tension structure. Thus, in the figure, the lighter coloured areas (blue to green) stand for mostly inactive content: points assigned to these areas cluster nicely, their distance in the original high-dimensional space was low. The darker areas (yellow to red) indicate tensions: the two-dimensional embedding is distorted, adjacent points refer to distant locations in the original high-dimensional space. Content can still be

[2] http://www.genenames.org/

mapped in the latter, but it's more likely to drift than not. Overall, Fig. 2 indicates a lot of turbulence in all periods, in the sense that the embedding undergoes rapid and abrupt changes, reflecting similar tendencies in the distances in the original space.

Genes grouped within the same group (or mapped onto the same neuron in the grid used by the algorithm) indicate an inherent interrelatedness, e.g. they are related to the same functionality, but also potentially hold additional important information regarding unexpected relationships between genes, which was not previously determined. For instance, two characteristic neurons in our experiment were: (1) Neuron [*ccr7, ccr2, cxcr2, cxcr1*] in the first period, referring to two families (CCR and CXCR) of chemokine receptors; (2) Neuron [*tlr10, tlr6, tlr2, tlr1*] in the third period that includes genes belonging to the same family (TLR), which showed significant association with B cell physiology and CLL pathogenesis.

Furthermore, and even more importantly, the spatial behaviour of genes from period to period is of great significance. For instance, imagine two genes that are close to each other in the first two periods but drift apart in period 3. This could possibly imply that some underlying event taking place between periods 2 and 3 resulted to the genes moving away from each other and that further investigation would be needed.

5.2 Statistical Evaluation

In order to assess the groupings derived from Somoclu, we conducted a statistical evaluation, based on the groupings derived in each period. The aim was to assess whether proximity of genes in the toroid plane indicates a strong underlying semantic similarity. We focused on neurons containing more than one genes. To evaluate the consistency of our approach in each period, we have checked whether the assignment of genes within a randomly selected neuron was better than random assignment.

The key idea was to simulate random assignment by the empirical probability distribution of the average semantic similarities corresponding to randomly generated neurons based on our data. Thus, for each neuron, we computed the average semantic similarity with G-SESAME, and determined their empirical probability distribution.

For the actual groups that emerged from Somoclu's analysis, we calculated the AIC semantic similarities among all terms within each neuron and computed their average. Then, we compared the empirical probability distributions of the actual and random groups both graphically and based on statistical tests. In particular, the non-parametric Mann-Whitney test was selected, since the sample sizes where not large enough in every case to consider a parametric t-test. In addition, we performed a 1-sided test (since the actual mean was higher than the random in all 6 cases) to evaluate the alternative hypothesis that actual groups will exhibit higher values of average similarities compared to the random groups. In case the null hypothesis of equality was rejected in most cases, our expectation that the actual assignment is better than a random assignment would have been validated.

Towards the same direction, we computed the 90%, 95% and 99% percentiles, based on the probability distribution of the random groups, and computed the respective percentage of actual group averages that took value beyond these thresholds. Our expectation was validated in each case where the percentage was higher than the

complementary probability (10%, 5% and 1%, respectively). To further consolidate the results, these comparisons were performed, based on the exact binomial test. The confidence level in all tests was set to 5%.

5.3 Discussion and Assessment of Results

Regarding the grouping of genes, we observe in Table 1 that in all cases the mean of averages is higher in actual groups compared to random groups. In half the cases this observation is emphatically validated through the Mann-Whitney test (p-value<0.05). In two other cases the non-rejection of the null hypothesis was rather marginal and in only one case (F3) the p-value was large (0.526). This validates our expectations regarding higher values of average AIC semantic similarity within the actual groups and shows that the groupings considered are meaningful.

Table 1. The number of the groups and the corresponding mean of the average semantic similarities are displayed per state, as well as the derived p-value of the distribution comparison according to the Mann-Whitney test. The upper panel refers to C and the bottom panel to F.

GO category	Period	Size		Mean		P-value (1-sided)
		Random	Actual	Random	Actual	
C	1	93	23	0.380	0.395	0.174
	2	151	45	0.401	0.424	0.011
	3	220	93	0.399	0.421	0.023
F	1	93	23	0.402	0.437	0.202
	2	151	45	0.382	0.423	0.018
	3	220	99	0.389	0.392	0.526

Table 2. The 90%, 95% and 99% percentiles, based on the probability distribution of the random groups, together with the corresponding p-values of the exact binomial test that compares the percentage of actual group averages to the complementary probability 10%, 5% and 1%, respectively. The upper panel refers to C and the bottom panel to F.

GO category	Period	90%	P-value	95%	P-value	99%	P-value
C	1	0.477	0.408	0.512	0.105	0.536	1.000
	2	0.527	0.947	0.551	0.665	0.611	0.363
	3	0.506	0.021	0.537	0.002	0.627	0.014
F	1	0.542	0.684	0.597	0.320	0.716	0.206
	2	0.509	0.012	0.541	0.006	0.667	1.000
	3	0.526	0.667	0.576	0.555	0.650	0.630

Table 2 further validates the results of Table 1, since statistically significant differences in the distributions are mostly reflected in significant results in Table 2, however, this is not always the case (C2). This could be due to many reasons, e.g. note that the number of actual groups in C2 is small in comparison to the respective number of random groups. All the above are also reflected in a smooth representation of the empirical probability distributions for the actual and random case in Table 1.

In order to biologically evaluate the drift between the three periods, the set of genes in the map was analysed per period, based on KEGG pathway analysis, a common methodology for discovering the enriched molecular pathways that correspond to a set of genes, the latter given as an input. The results revealed that 53 out of 80 pathways were common in all three periods, most of them with a critical role in CLL, such as the B cell receptor, NF-kappaB, p53, PI3K-Akt and signalling pathways [27]. This indicates the similarities between the three periods, but also highlights their differences, since a number of 27 pathways emerged in only one or two of the periods. For instance, MAPK and Fc epsilon RI signalling pathway are key pathways in CLL but appeared in only two of the three periods

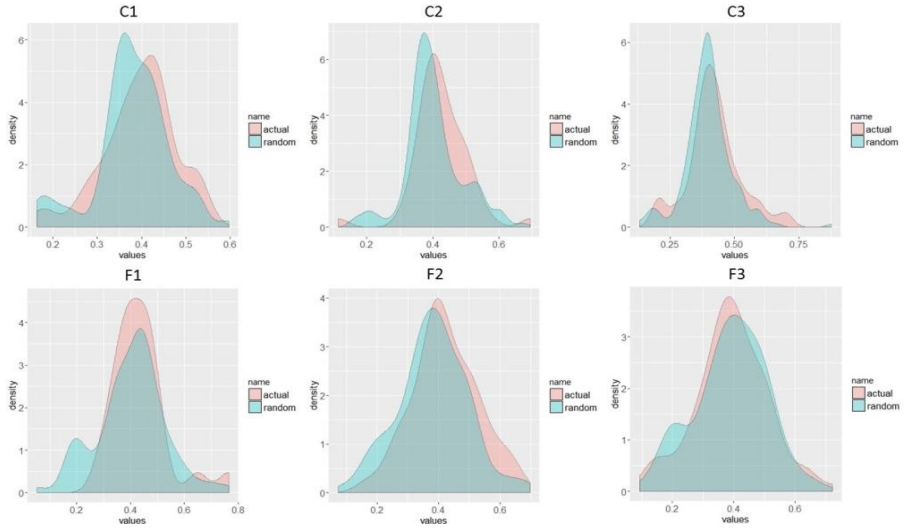

Fig. 3. Smoothed empirical probability distributions for the actual and random groups for C, F for all three periods. Pink represents actual and light blue represents random distribution.

The next step was to focus on a neuron level and explore the differences drifting between periods. Neurons with high values of average semantic similarity were of particular interest. One characteristic example is neuron [*trim8, zbtb4, tp53inp1, e2f5*] that exhibited the highest average semantic similarity in the third period, when the cellular component category was considered. None of these genes appeared in the first two periods, but all of them are transcription factors and show common regulation.

Another example is neuron [*wee1, cllu1, per2, per1, clock*], which also exhibited high similarity in the third period under the cellular component category as well. The *wee1, per2, per1* and *clock* genes are regulatory elements of the circadian clock, while the *cllu1* gene is overexpressed in CLL and serves as a prognostic marker. In the first period only *wee1* appeared, while in the second period *cllu1, per2* and *per1* were found in the same neuron. This observation could be possibly explained by the fact that the circadian clock in cancer was well documented from 2009 and onwards, which almost coincides with the end of the second period. This coexistence could suggest an association of circadian clock and the overexpression of *cllu1* in CLL.

Finally, we note that no neurons were unchangeable (same set of genes) in all three periods. Thus, it would be of interest to expand the groups considered, in order to include the neighbourhoods of neurons (e.g. at a radius of 2 or 3 cells). Another justification is the fact that the three periods did not overlap. Thus, well studied genes in one period may be missing from the literature in the next period, and even more specifically the interest for relation among genes in a specific period could have faded out in the next one, or could have been inexistent in the previous.

6 Conclusions and Future Work

The paper proposed a novel methodology for studying genes and their potential relationship within the field of CLL, which can be easily transferred to other disease settings as well. The methodology capitalizes on GO-based AIC semantic similarity and a state-of-the-art ESOM implementation for generating groups of genes belonging to consecutive time periods. Our experiments[3] verified the validity of the methodology and demonstrated that the temporal dimension indeed impacted the gene groupings.

This is largely a preliminary investigation with room for further explorations. In particular, instead of three consecutive periods (1, 2 and 3), it would be interesting to consider three overlapping periods (1, 1+2, 1+2+3). Such a consideration would maintain all the information derived from past periods, although obsolete information would also be maintained. Another interesting parameter is the number of periods, which could be increased, under the constraint that the number of abstracts per period is not greatly decreased. This would enable to collect more information in terms of drifting and to statistically evaluate it. Of a more immediate interest would be to study characteristic grouping examples already derived, such as neuron [*wee1, cllu1, per2, per1, clock*] (see Section 5.3), in terms of implied information, based on GO. A subsequent expert biological assessment could reveal valuable novel knowledge.

Acknowledgments. This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under Grant Agreement Number FP7-601138 PERICLES.

References

1. Jenssen, T. K. et al. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1), pp. 21-28.
2. Wilkinson, D. M., Huberman, B. A. (2004). A method for finding communities of related genes. *National Academy of Sciences*, 101, pp. 5241-5248.
3. Homayouni, R., Heinrich, K., Wei, L., Berry, M. W. (2005). Gene clustering by latent semantic indexing of MEDLINE abstracts. *Bioinformatics*, 21(1), pp. 104-115.
4. Fabbri, G., Riccardo D.-F. (2016). The molecular pathogenesis of chronic lymphocytic leukaemia. *Nature Reviews Cancer*, 16, pp. 145-162.

[3] All related resources are available at: https://github.com/skontopo/swat4ls2016

5. Huang, Z. X., Tian, H. Y., Hu, Z. F., Zhou, Y. B., Zhao, J., Yao, K. T. (2008). GenCLiP: a software program for clustering gene lists by literature profiling and constructing gene co-occurrence networks related to custom keywords. *BMC Bioinformatics*, 9(1), pp. 308.
6. Tenenbaum, J.; Silva, V. & Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), pp. 2319-2323.
7. Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(85), pp. 2579-2605.
8. Kohonen, T. Self-Organizing Maps. Springer, 2001.
9. Leacock, C., Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2), pp. 265-283.
10. Wu, Z., Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (pp. 133-138). Association for Computational Linguistics.
11. Jiang, J. J., Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.
12. Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.
13. Sánchez, D., Batet, M., Isern, D., Valls, A. (2012). Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications*, *39*(9), pp. 7718-7728.
14. Dong, H., Hussain, F. K., Chang, E. (2009). A hybrid concept similarity measure model for ontology environment. *OTM Confederated Int. Conf.*, pp. 848-857, Springer.
15. Ashburner, M. et al. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1), pp. 25-29.
16. Couto, F. M., Silva, M. J., Coutinho, P. M. (2007). Measuring semantic similarity between Gene Ontology terms. *Data & Knowledge Engineering*, 61(1), pp. 137-152.
17. Lord, P. W., Stevens, R. D., Brass, A., Goble, C. A. (2003). Semantic similarity measures as tools for exploring the gene ontology. *Pacif. Symp. on Biocomputing* 8(4), pp. 601-612.
18. Caniza, H. et al. (2014). GOssTo: a stand-alone application and a web tool for calculating semantic similarities on the Gene Ontology. *Bioinformatics*, 30(15), pp. 2235-2236.
19. Yu, G. et al. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7), pp. 976-978.
20. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. Chen, C. F. (2007). A New Method to Measure the Semantic Similarity of GO Terms. *Bioinformatics*, 23: 1274-1281.
21. Wittek, P. et al. (2015). Monitoring term drift based on semantic consistency in an evolving vector field. *IJCNN-15, International Joint Conference on Neural Networks*, pp. 1-8.
22. Wittek, P., Gao, S. C., Lim, I. S., Zhao, L. (2015). Somoclu: An efficient parallel library for self-organizing maps. *arXiv:1305.1422*.
23. Song, X., Li, L., Srimani, P. K., Yu, P. S., Wang, J. Z. (2014). Measure the semantic similarity of go terms using aggregate information content. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, *11*(3), 468-476.
24. Wittek, P., Darányi, S., Liu, Y. H. (2014). A vector field approach to lexical semantics. In *Proceedings of QI-14, 8th International Symposium on Quantum Interaction*, pp. 78-89.
25. International Human Genome Sequencing C. Finishing the euchromatic sequence of the human genome. *Nature*. 2004;431(7011):931-45.
26. Schuster SC (January 2008). Next-generation sequencing transforms today's biology. *Nat. Methods*. 5.
27. Burger JA, Chiorazzi N. (2013). B cell receptor signaling in chronic lymphocytic leukemia. *Trends in Immunology*, 34(12), pp. 592-601.