International Journal of Modern Physics C
Vol. 25, No. 2 (2014) 1350092 (9 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0129183113500927

PRESERVED NETWORK METRICS ACROSS TRANSLATED TEXTS

JOSEPHINE JILL T. CABATBAT*, JICA P. MONSANTO
and GIOVANNI A. TAPANG

National Institute of Physics
University of the Philippines-Diliman
Diliman, Quezon 1101, Philippines
**jcabatbat@nip.upd.edu.ph*

Received 21 February 2013
Accepted 16 August 2013
Published 20 September 2013

Co-occurrence language networks based on Bible translations and the Universal Declaration of Human Rights (UDHR) translations in different languages were constructed and compared with random text networks. Among the considered network metrics, the network size, N, the normalized betweenness centrality (BC), and the average k-nearest neighbors, k_{nn}, were found to be the most preserved across translations. Moreover, similar frequency distributions of co-occurring network motifs were observed for translated texts networks.

Keywords: Language networks; language translation; motifs; complex systems.

PACS Nos.: 89.75.−k, 89.75.Fb.

1. Introduction

As a complex system, language, or texts, can be naturally handled as networks with words, being the basic units, taken as nodes and the edges corresponding to the relationship among words. Language networks have been used in attempts at automatic text classification and in the study of different languages.[1,2] Automatic classification of text according to categories such as poetic style,[3] literary movement[4] and forms such as literary and scientific writing[5] were accomplished by employing network metrics and network motifs. Several types of networks can be constructed from texts depending on the particular aspect under study. For example, in a semantic network, the links between nodes represent semantic relations such as synonymy and antonymy.[6,7] Another is syntactic network which links the words based on language syntax.[6,8] Syntactic and semantic networks are language-specific network constructions. Co-occurrence networks, on the other hand, only relate words based on their location in the text (i.e. which word precedes another in a sentence)[6] and would be nonlanguage-specific (an example would be its use in Ref. 9).

A translation of a document is the rendering of an original form from one language to another.[10] Since translations are expected to retain the meaning of the source document,[11] metrics that do not change across different translations may therefore be used as automatic classifiers or as means to quantify similarities in content of different written texts. In this work, network metrics of co-occurrence language networks are calculated and compared for different translations of some texts, specifically the Bible and the UDHR, and for sets of randomly selected unrelated articles.

2. Methodology

Bible and UDHR translations were used because they are among the most translated texts in the world[12,13] and abundant digital copies of their translations abound online. The translations used are summarized in Table 1. Since the Bible is composed of many books with their many chapters, only selected sets of chapters from different books were used as corpus sets. Since the UDHR is composed of 30 short sections, each translation was divided into 10 parts (three sections each). For comparison, 20 randomly selected and unrelated articles in English from Project Gutenberg,[14] a digital archive of full texts of public domain, were also acquired to make up the Random set. Each article was divided into parts. The last sentence to be included in each partition is specified by a set of word counts based on the usual word counts of the Bible chapters. Table 2 summarizes the corpus sets used. The Bible translations were acquired from UnboundBible.org and Biblos.com,[15,16] while the UDHR translations were sourced from the UDHR homepage.[13]

2.1. *Pre-processing and network construction*

All texts were pre-processed so that nonprinting characters are removed and punctuation marks are converted to their word equivalents (e.g. "," to "COMMA" and

Table 1. Bible and UDHR translations used.

Bible	UDHR
Danish Bible (DNS)	Danish (DNS)
Dutch Bible (DUT)	Dutch (DUT)
Finnish Bible (FIN)	Finnish (FIN)
French Bible (FRN)	French (FRN)
German Bible (GER)	German (GER)
Maori Bible (MBF)	Maori (MBF)
Tagalog Bible (TGL)	Tagalog (TGL)
American Standard Version (ASV)	English (ENG)
Basic English Bible (BBE)	
Darby Version (DBY)	
Douay Rheims Version (DRB)	
King James Version (KJV)	
World English Bible (WEB)	
Webster's Bible (WBS)	
Young's Literal Translation (YLT)	

Table 2. Bible, UDHR, and Random sets used.

Set	Co-occurrence networks for
Bible 1	Book 1 (Genesis) chapters 1 to 20
Bible 2	Book 1 (Genesis) chapters 31 to 50
Bible 3	Book 24 (Jeremiah) chapters 1 to 20
Bible 4	Book 4 (Judges 1) chapters 1 to 20
UDHR	UDHR divided into 10 3-section parts
Random	Consists of 20 randomly acquired unrelated articles from Project Gutenberg divided into parts.

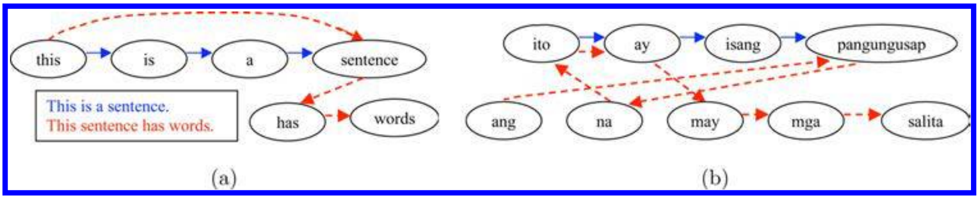


Fig. 1. (Color online) Examples of a co-occurrence network. In (a), all the unique words in the boxed text are taken as nodes in the network. *This* has a blue link to *is* because it precedes *is* in the blue sentence, and *is* has a blue link to *a* because it precedes *a* and so on. (b) Shows the corresponding Tagalog co-occurrence network.

Table 3. Network metrics calculated for the co-occurrence language networks.

Network parameter	
Network Size (N)	Total number of nodes in the network
Diameter (D)	The largest number of links connecting any pair of nodes in the network
Mean Path Length (L)	Average of the shortest paths for all possible node pairs
Average Clustering Coefficient (C)	$C = \frac{1}{N}\sum_{i\in V}\frac{l_i}{k_i(k_i-1)}$, where V is the set of all nodes in the network, l_i is the number of triangles through node i, and k_i is the number of neighbors of node i.
Average Normalized Betweenness Centrality (BC)	$\mathrm{BC} = \frac{1}{N}\sum_{v\in V}\mathrm{BC}(v)$, where $\mathrm{BC}(v) = \frac{1}{(N-1)(N-2)}\sum_{i\neq v\neq j\in V}\frac{\sigma_{ij}(v)}{\sigma_{ij}}$, where σ_{ij} is the number of shortest paths from node i to node j and $\sigma_{ij}(v)$ is the number of shortest paths from node i to node j that include node v.
Average k-nearest-neighbor (k_{nn})	$k_{nn} = \frac{1}{N}\sum_{i\in V}k_{nn_i}$, where $k_{nn_i} = \frac{1}{p}\sum_{n\in P}s_n$ P is the set of nodes that link to node i, s_n is the number of nodes that node n links to, and p is the number of nodes that link to node i.

";" to "SEMICOLON)." Co-occurrence networks were then constructed for each chapter (for the Bible sets) or part (for UDHR and Random set). In constructing this network, all distinct words in the document are represented by nodes and a link from word 1 to word 2 is made when word 1 precedes word 2 in a sentence (illustrated in Fig. 1). The Perl Graph module was used for network construction in this study.[17]

2.1.1. *Metric calculation*

Table 3 summarizes the parameters calculated for each network. The calculations were implemented using the Graph and SocialNetwork::Algorithm modules of Perl.[17]

2.2. *Results*

Similarity in trends for D, L and C are not observed (Fig. 2) across translations in all the corpus sets via visual inspection. This can be attributed to the sensitivity of the parameters on the syntactic differences among languages. For this study, a good metric is defined to be a metric that is conserved across translated texts but are different in random or unrelated texts. Since D, L and C vary for documents with the same content, these metrics cannot be used in comparing similarity in text content.

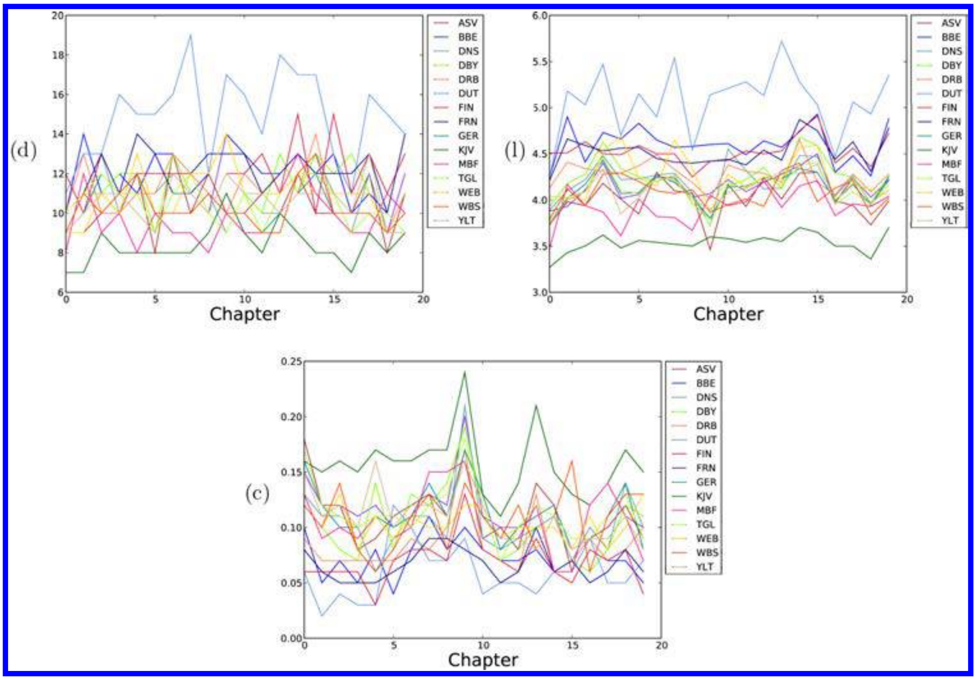
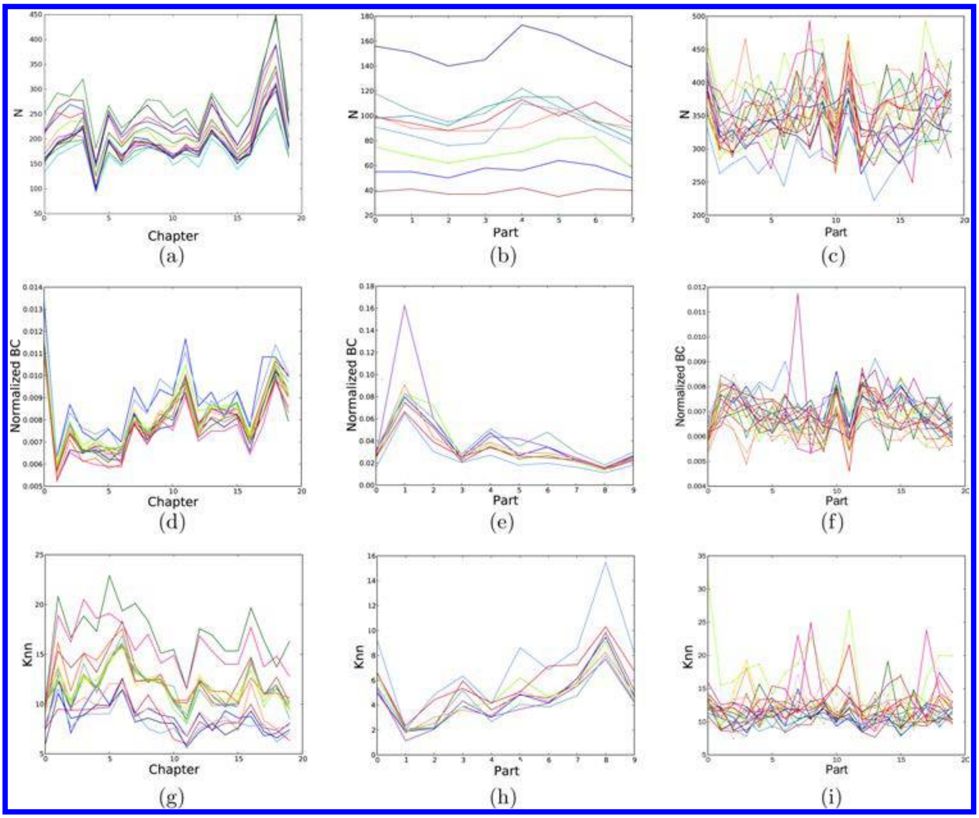


Fig. 2. (Color online) D, L and C trends for Bible1.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 3. (Color online) (a–c) N trends, (d–f) BC and (g–i) k_{nn} trends for Bible 3, UDHR and the Random set (left to right).

The N, betweenness centrality (BC) and k_{nn} on the other hand have similar trends for all corpus sets (Fig. 3). The trends for Random set do not exhibit strong correlation as compared to the translated texts. Since similar trends of these parameter values were observed for different languages, but were absent for the Random set, which are composed of unrelated articles in English (same language structure), the authors infer that the difference is due to the content of the text and not on the structure of the language.

The similar trends in the network size suggest that the rates of introduction of new words or concepts in the text are closer among translated texts than among unrelated texts. The similarity in k_{nn} suggests the preservation of degree distribution and how nodes of varying degrees are connected to each other. As for BC, its preservation across translations implies the importance of the distribution of the central nodes in the representation of text content.

The Pearson correlation coefficient and mutual information among the N, normalized BC and average k_{nn} series were computed using R stats and bioDist

Table 4. Pearson correlation and mutual information among chapters of the combined Bible sets and among the articles in the Random set for the N, BC and k_{nn} values.

	Pearson correlation			Mutual information (nats)		
Set	N	BC	k_{nn}	N	BC	k_{nn}
Bible	0.8967 ± 0.0098	0.9456 ± 0.0041	0.8847 ± 0.0061	0.7800 ± 0.0224	0.6743 ± 0.0119	0.5377 ± 0.0154
Random	0.4620 ± 0.0137	0.5794 ± 0.0074	0.0072 ± 0.0090	0.2258 ± 0.0078	0.3018 ± 0.0067	0.1286 ± 0.0034

packages.[18,19] The Pearson correlation coefficient between variables X and Y is given by

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{1}$$

while the mutual information is given by

$$I(X;Y) = \sum_{y \in Y}\sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}, \tag{2}$$

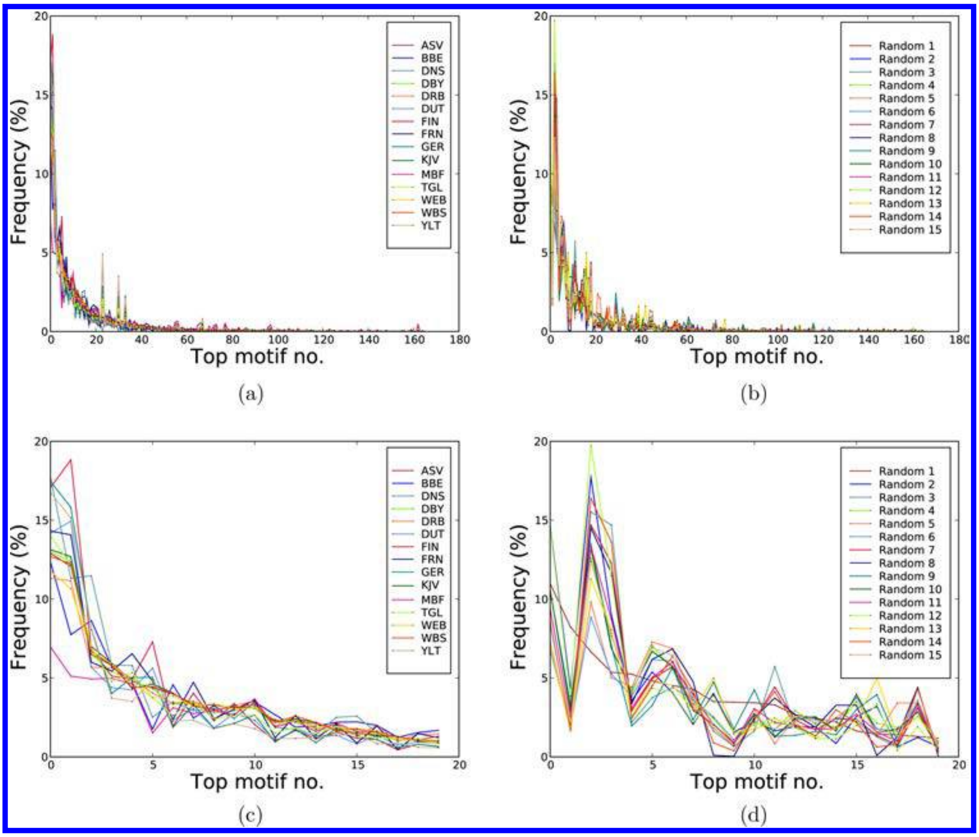
where $p(x,y)$ is the joint probability distribution function of X and Y and $p(x)$ and $p(y)$ are the marginal probability distribution functions.

From Table 4, it can be observed that among the parameters, the N and the k_{nn} gave the highest values for the difference in the correlation and mutual information between the Bible sets and the Random set, making them the better metrics in comparing text content similarity. However, among the parameters, the N, and BC gave the highest values for correlation and mutual information for both the Bible and the Random set. Since translations of a long text tend to be long as well and since the partitions of the Random set was done in such a way that the parts have at most one sentence length of difference, the high correlation and mutual information values for N, and BC suggest the higher sensitivity of these parameters to the length of the text.

2.3. *Network motifs*

Milo *et al.*[20] proposed a more visual structural property of real-world networks: network motifs or sub-networks occur in the network at higher frequencies than in a corresponding random network. Motifs have been used in describing and analyzing the structure of many different real world networks[20–22] and it was observed that networks constructed from texts in different languages have similar motif sets.[21]

FANMOD,[23] a network motif detection tool by Wernicke and Rasche which implements the algorithm RAND-ESU in sampling subgraphs was used to locate the motifs. Since the motifs for 2- to 4- nodes did not show significant difference between the Bible set and the Random set, the results for 5-node motifs are presented. Figure 4 shows the frequencies (percent occurrence counts) of the motifs found in ASV for the different translations of the chapters in Bible 3 and the



(a)

(b)

(c)

(d)

Fig. 4. (Color online) Frequency distribution of all motifs in (a) ASV for all translations of Bible 3 and (b) Random 1 for the other random articles. Frequency distribution of top 20 motifs in (c) ASV for all translations in Bible 3 and (d) Random 1 for the other random articles.

frequencies for 5-node motifs found in one of the random texts (Random 1) for the Random set. It can be seen that the trends in the Bible set are similar while those in the Random set have highly varying trends. To compare the trends, the average value of the quantity $|Q-1|$ where Q is the correlation quality[24] was computed for each corpus set. The correlation quality Q, in the case studied, is obtained using the equation:

$$Q = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} \tag{3}$$

for the data series X and Y with length N.

Table 5 gives the average $|Q-1|$ values for the Bible set and Random set. Although visual inspection would reveal more similar trends in the Bible set, the average $|Q-1|$ values were found to be closer to zero for the Random set than the Bible set, implying that the motif frequency distributions of the articles in the Random set are more similar than those in the Bible set. Upon investigation of the Bible data and

Table 5. Average $|Q-1|$, Pearson correlation coefficient, and mutual information of the motif frequency distributions in the Bible set and the Random set.

Set	$\lvert Q-1\rvert$ (Top 20 motifs)	$\lvert Q-1\rvert$ (All motifs)	Pearson correlation	Mutual information
Bible 3	0.1496	0.1445	0.9585 ± 0.0035	0.2209 ± 0.0058
Random	0.1457	0.1386	0.9246 ± 0.0051	0.2203 ± 0.0049

of Fig. 4, it was found that the motif frequency distribution for the Maori translation is very dissimilar with the other translations and this single translation must have greatly influenced the correlation quality values. In fact, if Maori translation is to be excluded from the calculation of correlation quality, we get the following values for the average $|Q-1|$: 0.1073 (Bible), and 0.1500 (Random) for the top 20 motifs and 0.1076 (Bible) and 0.1427 (Random) for all the motifs. The Pearson correlation coefficient and mutual information were calculated to further compare the distributions. Considering all motifs, both measures gave higher values for Bible set than the Random set (see Table 5).

Common top motifs are found among the Bible sets and among the Random set. However, as in the case for ASV and Random 1, more overlaps in the top motifs are observed in the Bible sets than in the Random set. Furthermore, upon checking the 20 top-occurring motifs in non-ASV translations, we get the following observations: MBF and TGL share the motif 33848 while DUT, FIN and GER share motifs 1084, 33848 and 1082416. Hence, it may also be possible to distinguish English and non-English translations of a document by looking at the different motifs present.

3. Conclusion

Among the network metrics considered for the co-occurrence networks of translated texts, the network size, N, the normalized BC and average k-nearest neighbor, k_{nn}, were found to be the most preserved across translations. Preserved values of these parameters may then allow for comparison of single-document translations and measurement of text content.

The effectiveness of average k_{nn} in reflecting similarity in content may be attributed to the fact that although it is a local property like the clustering coefficient, it considered higher level of neighbors. The BC on the other hand, although a global measure like the average path length and diameter, captures the dependence of nodes with the other nodes in the network and the node's importance and hence also performed well at comparing text content similarity.

Motif frequency distribution is likewise proven as a viable tool for detecting similarity in text content, as the study found dissimilarity in motifs in unrelated texts. Therefore, N, BC, k_{nn}, and motif may be used in quantifying similarities or differences in the content of documents. Since translation is supposed to retain the meaning of texts across languages, these preserved network metrics establish a relationship between network structure and meaning.

Acknowledgments

The authors gratefully acknowledge the Instrumentation Physics Laboratory for usage of its computing facilities.

References

1. J. Hoorn, S. Frank, W. Kowalczyk and F. van der Ham, *Lit. Linguist. Comput.* **14**, 3 (1999).
2. O. Rosso, H. Craig and P. Moscato, *Physica A* **388**, 6 (2009).
3. R. Roxas-Villanueva, M. K. Nambatac and G. Tapang, *Int. J. Mod. Phys. C* **23**, 2 (2012).
4. D. R. Amancio, E. Altmann, O. Oliveira Jr and L. da Fontoura, *New J. Phys.* **14**, 4 (2012).
5. I. Grabska-Gradzinska, A. Kulig, J. Kwapien and S. Drozdz, *Int. J. Mod. Phys. C* **23**, 7 (2012).
6. R. Sole, B. Corominas-Murtra, S. Valverde and L. Steels, *Complexity* **15**, 6 (2010).
7. M. Steyvers and J. Tenenbaum, *Cogn. Sci.* **29**, 1 (2005).
8. H. Liu, *Physica A* **387**, 12 (2008).
9. W. Liang, Y. Shi, C. Tse, J. Liu, Y. Wang and X. Cui, *Physica A* **388**, 23 (2009).
10. *The Columbia Encyclopedia*, 6th edn., 2013, http://www.encyclopedia.com/doc/1E1-translat.html. (Accessed on 12 September 2013).
11. *The Oxford Pocket Dictionary of Current English*, 2009, http://www.encyclopedia.com/doc/1O999-translation.html (Accessed on 12 September 2013).
12. I. Chatzitheodoro, *Translation J.* **5**, 5 (2001).
13. United Nations General Assembly, The universal declaration of human rights, http://www.un.org/en/documents/udhr.
14. M. S. Hart, Project Gutenberg, http://www.gutenberg.org.
15. Biola University, The unbound bible, http://unboundbible.org.
16. Online Parallel Bible Project, Biblos.com, http://biblos.com.
17. J. Hietaniemi, Comprehensive perl archive network, http://www.cpan.org/.
18. R Development Core Team, *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2009).
19. B. Ding, R. Gentleman and V. Carey, bioDist: Different distance measures. R package version 1.10.0, http://www.bioconductor.org/packages/2.11/bioc/html/bioDist.html.
20. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, *Science* **298**, 5594 (2001).
21. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer and U. Alon, *Science* **303**, 5663 (2004).
22. D. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Pinter, U. Alon and H. Margalit, *Proc. Natl. Acad. Sci. USA* **101**, 16 (2004).
23. S. Wernicke, F. Rasche, *Bioinformatics* **22**, 9 (2006).
24. C. Monterola, R. M. Roxas and S. L. Carreon-Monterola, *Complexity* **14**, 4 (2009).