BIOINFORMATICS ERRATUM
Vol. 26 no. 6 2010, pages 853–854
doi:10.1093/bioinformatics/btq091

Erratum

Biological network comparison using graphlet degree distribution

Nataša Pržulj
Department of Computing, Imperial College London, SW7 2AZ, UK, E-mail: natasha@imperial.ac.uk

Bioinformatics, 23 (2): e177. (2007)

In Pržulj (2007), the formula for GDD agreement has a typo. The j-th GDD agreement was defined [formula (5) in Pržulj (2007)] as

$$A^j(G, \pm H) = 1 - D^j(G, H), \quad (1)$$

where $\pm H$ is a typo and it should be just:

$$A^j(G, H) = 1 - D^j(G, H) \quad (2)$$

Next, $D^j(G, H)$ was mistakenly claimed to be between 0 and 1. In fact, this distance is between 0 and $\sqrt{2}$. Since it can be greater than 1 (consider, e.g. $D^1(G, H)$ where G is 5-node cycle and H is 5-node path, see Fig. 1), formula (2) can result in negative values.

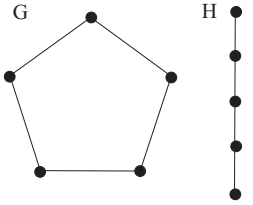

Fig. 1. An example of two graphs G and H for which $D^1(G, H) > 1$. For orbit number 1 (see Fig. 1 in Pržulj (2007)), using formulas (1)–(4) from Pržulj (2007) we can calculate for graph G that $N_G^1(2) = 1$, and for all $k \neq 2$, $N_G^1(k) = 0$, while for graph H, $N_H^1(1) = \frac{4}{4.5}$, $N_H^1(2) = \frac{0.5}{4.5}$, and for all $k > 2$, $N_H^1(k) = 0$. Then $D^1(G, H) = \frac{\sqrt{32}}{4.5} > 1$.

Therefore, we correct the formula for $D^j(G, H)$ by simply dividing it by $\sqrt{2}$:

$$D^j(G, H) = \frac{1}{\sqrt{2}} \left(\sum_{k=1}^{\infty} \left[N_G^j(k) - N_H^j(k) \right]^2 \right)^{\frac{1}{2}} \quad (3)$$

$D^j(G, H)$ defined in this way is guaranteed to be between 0 and 1. All other formulas from Pržulj (2007) remain correct. Below we prove that $D^j(G, H)$ defined in this way is between 0 and 1.

Proof. According to formula (3) from Pržulj (2007), $N_G^j(k) = \frac{S_G^j(k)}{T_G^j}$. Therefore, we can rewrite formula (3) above as

$$\begin{aligned} D^j(G, H) &= \frac{1}{\sqrt{2}} \left(\sum_{k=1}^{\infty} \left[N_G^j(k) - N_H^j(k) \right]^2 \right)^{\frac{1}{2}} \\ &\leq \frac{1}{\sqrt{2}} \left(\sum_{k=1}^{\infty} \left[\left[N_G^j(k) \right]^2 + \left[N_H^j(k) \right]^2 \right] \right)^{\frac{1}{2}} \\ &= \frac{1}{\sqrt{2}} \left(\sum_{k=1}^{\infty} \left[\left[\frac{S_G^j(k)}{T_G^j} \right]^2 + \left[\frac{S_H^j(k)}{T_H^j} \right]^2 \right] \right)^{\frac{1}{2}} \\ &= \frac{1}{\sqrt{2}} \left(\sum_{k=1}^{\infty} \left[\frac{S_G^j(k)}{T_G^j} \right]^2 + \sum_{k=1}^{\infty} \left[\frac{S_H^j(k)}{T_H^j} \right]^2 \right)^{\frac{1}{2}} \\ &\leq \frac{1}{\sqrt{2}} (1+1)^{\frac{1}{2}} = 1 \end{aligned}$$

since, according to formula (2) from Pržulj (2007), $T_G^j = \sum_{k=1}^{\infty} S_G^j(k)$. ■

The upper bound of 1 for $D^j(G, H)$ is reachable, for example, if G is a 5-node cycle and H is a 3-node path.

We reanalyzed all the results from Pržulj (2007) and this correction does not affect them qualitatively; there is only a small quantitative difference in the results. Figure 2 presents the results of the same analysis as Figure 3 from Pržulj (2007), but performed with the formula corrected as described above. As shown in Figure 2, the model ordering remains unchanged, with GEO-3D model being superior to other models.

As in Pržulj (2007), to gauge the range of this agreement measure, we computed the average agreements between various model (i.e. theoretical) networks. For example, when comparing networks of the same type that are of the same size and are generated with the same parameters (ER versus ER, ER-DD versus ER-DD, SF-BA versus SF-BA, or GEO-3D versus GEO-3D), we found that the mean GDD agreement is the smallest for two SF networks (0.86 ± 0.01) and the highest for two GEO-3D networks (0.95 ± 0.002). To verify that our agreement measure can give low values for networks that are very different, we also constructed a straw-man model graph called a circulant and compared it with some actual PPI network data. A circulant graph is constructed by adding chords to a cycle on n nodes so that the i-th node on the cycle is connected to the $[(i+j) \bmod n]$-th and $[(i-j) \bmod n]$-th node on the cycle. Clearly, a large circulant with an equal number of nodes and edge density as the data would not be very representative of a PPI network and indeed we find that the agreement between such a circulant, with chords defined by $j \in \{5, 10, 15, 20, 25, 30\}$, and the data is < 0.26.

The author apologizes for these mistakes.

ACKNOWLEDGEMENTS

The author thanks Zi Wang, a PhD student of Prof Gesine Reinert and Dr Charlotte Deane, Department of Statistics, University of Oxford, for noticing that the interval was wrong in the original paper; he also noticed that the upper bound of 1 can be achieved. Many thanks to Oleksii Kuchaiev, Computer Science, Univeristy of California, Irvine, for helping with the proof and the data analysis using the corrected formula in this erratum.

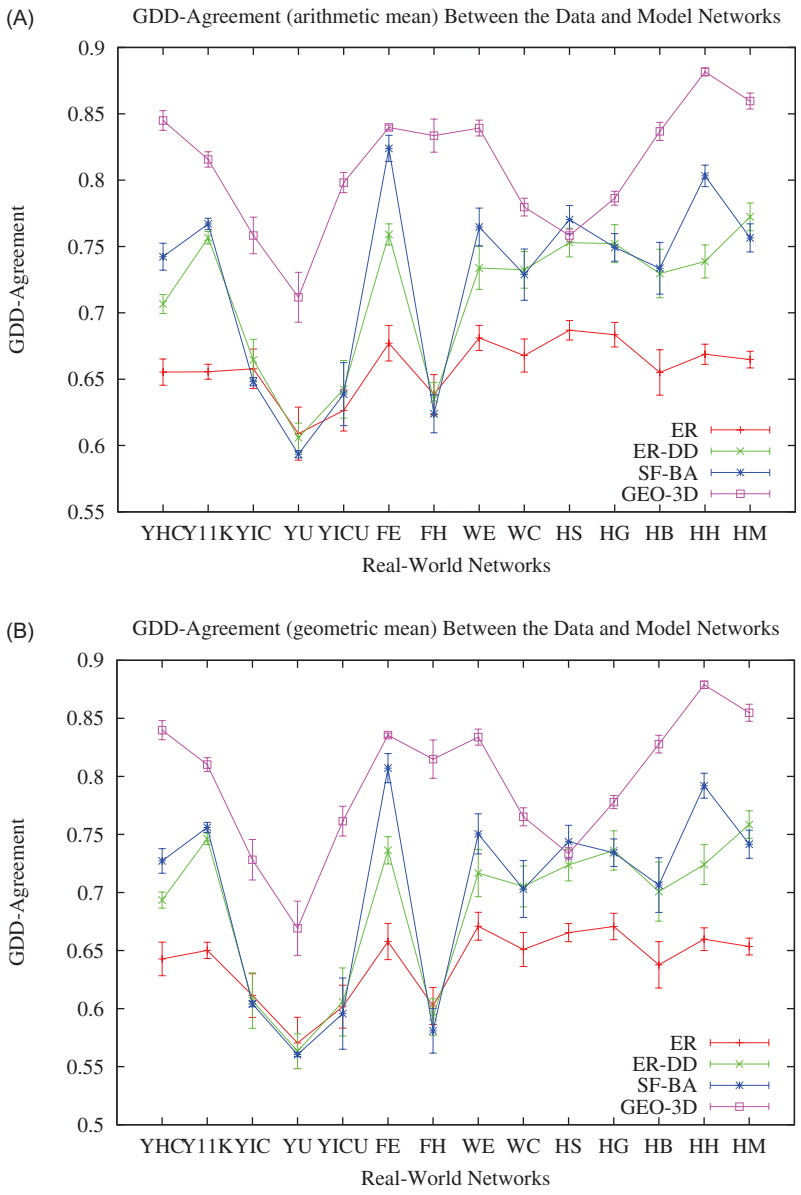

Fig. 2. Agreements between the 14 PPI networks and their corresponding model networks from Pržulj (2007). Labels on the horizontal axes are described in Section 2.1 of Pržulj (2007). Averages of agreements between 25 model networks and the corresponding PPI network are presented for each random graph model and each PPI network, i.e. at each point in the figure. As described in Section 2.3 of Pržulj (2007), the agreement between a PPI and a model network is based on the: (**A**) arithmetic average of *j*-th GDD agreements; and (**B**) geometric average of *j*-th GDD agreements.

REFERENCES

Pržulj,N. (2007) Biological network comparison using graphlet degree distribution. *Bioinformatics*, **23** (2): e177–e183.