Shi Rui’s answer to Question 3 **Data Doppelganger effect, part and parcel of data**

Introduction

The doppelganger effect is never unique to biomedical data. As it describes the phenomenon of duplicated or highly similar data leading to biases or errors in analysis, which is a common cause that led to incorrect predications.

Given an interesting example in the local syntax: it is expected by Singapore parents that, a child that graduated from Ai Tong primary school is guaranteed to enter Nanyang Technology University. However, this expectation or rather, prediction, is not always fulfilled. When it fails, the cause may be attributed to the so-called doppelganger effect.

A parent with a data science degree could probably explain the story in this way:

The model, which is the parents’ mind, that used to predict the child’s future university has been ‘trained’ and ‘validated’ based on other children’s profile parents have seen. One possible scenario could be: in the data set, a lot of children that entered NTU are having a same attribute of gradated from Ai Tong primary school. Secondly, the scope or size of the data set could be narrow and limited. Other information that could influence the outcome such as child’s personality, was not included in the data set or not receiving attention from the model. Together, it resulted in many individual data in the set is very similar to each other and led to the data doppelganger effect. In psychological context, there is a term of ‘the filter effect’, which referring to people focus too much on a few attributes but failed to identify the true extinction and resulted in incorrect judgement. It is not the same as data doppelganger but revealed some inspiring interlinks.

In fact, doppelganger effect is never new to social science researchers. Research on doppelganger effect "Social Science & Medicine" published in 2020 indicated that when survey questions are repeated or are highly similar, respondents may provide inconsistent or unreliable answers, leading to biased or inaccurate results [1].

Similarly, another study published in the journal "Personality and Individual Differences" in 2016 examined the doppelganger effect in personality assessments. It derived the same conclusion that when personality trait questionnaires contain highly similar or duplicated items, the accuracy of the results dropped significantly, which eventually, leading to a wrong assessment of personality [2].

Data doppelganger challenges in biological data

However, this could be a bigger challenge in the field of biomedical data analytics due to the highly complexed nature of the biological system of study. One of the topics on the trend of pharmaceutical is personalized medicine, to tailor a drug for the unique individual needs. As a precursor, the individual’s chromosomal DNA needs to be understood and that’s where gene sequencing comes into the picture. Gene sequencing is a process of identifying the actual order of nucleotides (A, C, G, and T) that make up a particular DNA molecule [3]. It is a genius way to ease the sequencing of DNA, the relatively large molecule that usually consists of approximately 3 billion base pairs of DNA, without even mentioned that a large portion of this DNA is made up of repetitive and duplicated sequences [4].

The process can be illustrated briefly as shown in figure 1[5][6]. The first step is to break the DNA into short pieces that contains much lesser base pairs, the so called “Reads”. The Reads are then sent for sequencing by instruments. As Reads are shorter, it is much easier to sequence them. Post the Reads sequencing, it is expected to identify lots of duplication between or even within the Reads. The duplication between reads is considered as an overlap and used to assemble reads to a longer “chain” named contigs. Last but not least, is to reconstruct the DNA using the obtained contigs and the successful reconstruction will obtain the fully understanding of the DNA sequence.

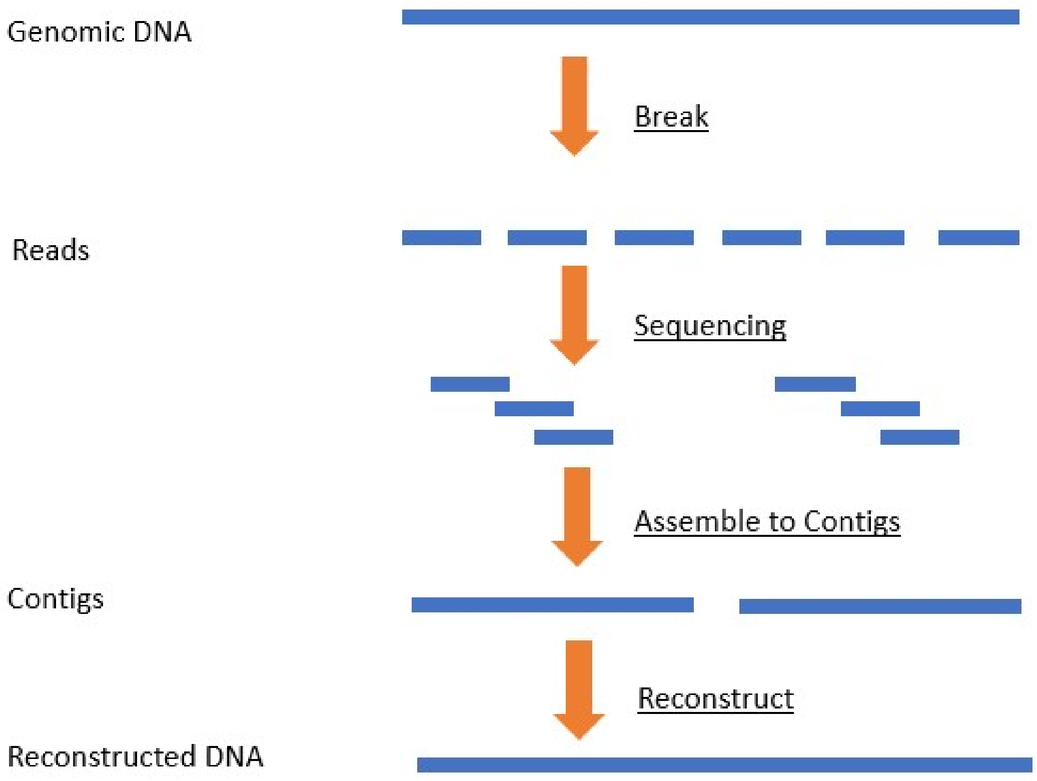

Figure 1 Illustration of Gene sequencing process

Next, zooming into the step of assemble Reads to Contigs, as shown in figure 2 [7]. As the number of Reads are huge, it is relied on a data analytics model to identify which are the right ones to be jointed together based on their overlaps. Given that Reads A and Reads B form the correct pair. But due to the large similarly in the Reads' sequencing, data doppelganger effect occurs. Therefore, Reads C which having a similar overlap is selected to be assembled with Reads A instead of Reads B. This will result in inaccurate Contigs and finally, failed to reconstruct the DNA.

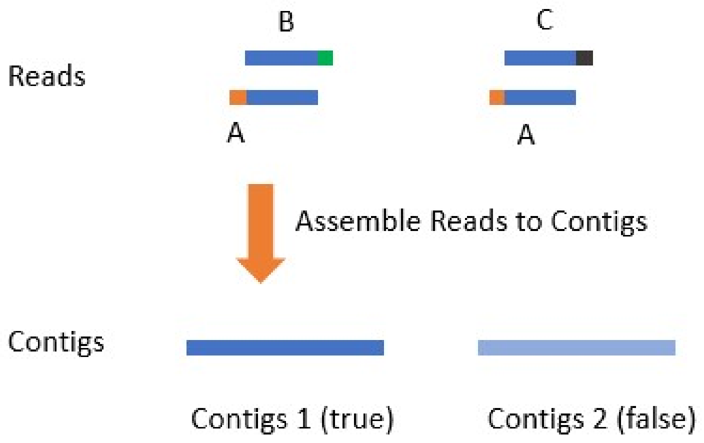

Figure 2 Illustration of assembling Reads to Contigs

Recommendations

There is no easy nor single solution to completely avoid the data doppelganger effect, but a combination of multiple methods may help:

- Identifying the key indicators to better "characterisation" the Reads will help to reduce the likelihood for the model to mix the similar but different ones.
- Remove the same Reads from the data set as much as possible.
- Train the model with a "in-completed" data set for several times, is data set provided each time is slightly different from each other as they should be obtained from the main data set by removing a portion of the data at random. Review the training results for those common answers to look out for and remove duplicates. This will help to minimise the data doppelganger and the model should be trained with the corrected data set afterwards.
- Do reverse error analysis, assemble the Contigs, review the error Contigs to analysis the failures and perform corrections to the training data set for the next training.

Conclusion

The doppelganger effective is inherent in the data science field and nearly impossible to be fully eliminated. However, with the right strategy to detect and reduce it from the data set prior to training and validation is necessary, and suggest significant improvement in the model accuracy.

References:

1. Dodds, P. S., & Rothman, A. J. (2020). The doppelganger effect in survey research: Repeated measures of similar items produce divergent results. Social Science & Medicine 258, 113-115
2. Gao, X., Li, J., Li, H., & Zhang, Y. (2016). The doppelganger effect in personality assessment: An investigation using multidimensional Rasch models. Personality and Individual Differences 94, 35-40.
3. Schuster, S. C. (2008). Next-generation sequencing transforms today’s biology. Nature, 2008(452), 18–24.
4. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., ... & Initial sequencing and analysis of the human genome. (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860-921.
5. Metzker, M. L. (2010). Sequencing technologies - the next generation. Nature Reviews Genetics, 11(1), 31-46.
6. Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. Nature Biotechnology, 26(10), 1135-1145.
7. Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. Genomics, 95(6), 315-327.