

Genomic Information Quality

Extended Abstract

Qing Liu

CSIRO Tasmanian ICT Centre, Australia
E-mail: Q.Liu@csiro.au

Xuemin Lin

The University of New South Wales, Australia
E-mail: lxue@cse.unsw.edu.au

1. Introduction

In the last 10 years we have witnessed the sequencing of the human genome and then the explosion in genomic data available in reference public databases and special purpose information products. These advances have profound influence in biology and drug discovery. Biological research has transformed from a purely experimental to an information-driven discovery science.

The quality of genomic information in the public and private database is crucial important. For example, to make a single drug useful, the pharmaceutical companies investigate how to generate proper structure of drug which is derived from experimental genomic data. The whole process of development, testing and clinical trials is extremely expensive. Using incorrect genomic information leads to the whole process failed. Further more, the cost is far more than merely financial. It also have great impact on trust from customers and motivation from workers etc.

In this paper, first, we will study the genomic information process and the errors produced during the process. Second, the key components of information quality theory will be reviewed. The genomic information quality problem is then presented.

2. Genomic Information Processing

The central dogma of molecular biology represents the genomic information process within an organism's cells (see Figure 1). Transcription is the process by which genetic information from DNA is transferred into RNA. During translation, the RNA sequence is translated into a sequence of amino acids as the protein, the building block of the body, is formed. Genomic data is about every piece of information involved in the above process. The data types involved include DNA sequence, gene expression, protein sequence, 3-dimensional protein, structural/functional annotation, metabolic/signalling pathway etc.

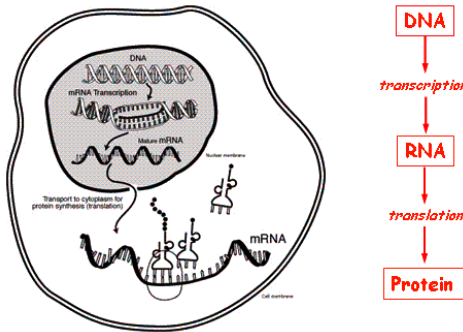


Fig. 1. Central Dogma of Molecular Biology

To understand the information process for drug development and disease prevention, typical discovery process is that biological experiment is designed based on literature and experiments are performed on the living organism by wet-lab biologists. The result is transformed into digital data format using some platforms such as Affymetrix. Then data is analyzed by bioinformaticians using computer software. Biologists re-use the analyzed information to design more relevant and accurate experiment. From the above process, we can see the genomic information production is an inter-dependent process.

Various experiment and analysis steps may be involved based on the biological question studied. The overall process of genomic information production includes four dependent steps:

- Step 1: DNA sequence determination
- Step 2: genome feature annotation
- Step 3: protein sequence determination
- Step 4: protein function annotation

In each step, various errors caused by experiment, analysis etc. may occur. Muller and Naumann¹ provide a good classification of possible errors produced during the genomic information production.

- Experimental errors: unnoticed experimental setup failure or systematical errors
- Analysis errors: misinterpretation of information
- Transformation errors: information is not transferred properly from one representation into another
- Propagated errors: erroneous data is used for the generation of new data
- Stale data: unnoticed changes to base data on which a data item depends and that falsify it

Note analysis error is due to the misinterpretation of information. Incomplete or uncertain domain knowledge may lead to misinterpretation. The complexity of accessing heterogenous genomic data sources, in which the contents are overlapping or even conflicting, is also another major reason of misinterpretation.

Table 1 shows the summary of possible errors in each step. Interested readers may refer¹ for details.

Table 1. Possible Error Produced During Genomic Information Production

Step	Experimental error	Analysis error	Transformation error	Propagated error	Stale data
1	√		√		
2		√		√	√
3	√		√	√	√
4	√	√		√	√

Information Quality Theory

Klein and Rossin mentioned there is no single definition of data quality accepted by researchers and those working in the discipline.² And "fitness for use"³ is widely adopted in the quality literatures.

In the context of information quality assessment, three key components are presented by Ge and Helfert:⁴ identification of quality problem, identification of quality dimension and assessment methodology.

Quality problem: There are three types of quality problems summarized by Garvin:⁵ biased information, outdate information and massaged information.

- Biased information: the content of the information is inaccurate or distorted in the transformation process
- Outdate information: the information that is not up to date for the task
- Massaged information: the different representations of the same information

Quality dimension: It is also confirmed that quality is a multi-dimensional concept. Intuitive, theoretical and empirical methods are three approaches to study the quality.⁶ By intuitive approach, quality dimension are based on the specific application contexts. Theoretical approach define quality dimension by data deficiencies. The data fitness for use to data consumers are the main quality dimension of empirical approach.

Assessment methodology: To assess the quality, Pipino et al. classify the assessment into objective and subjective assessment.⁷ Objective assessment is to measure the extent to which information conforms to quality specifications and references. Subjective assessment is to measure the extent to which information is fitness for use by data consumers.

3. Genomic Information Quality

From the above two sections, a natural mapping could be found from the genomic information error produced during the genomic information generation to the information quality problems (see Table 2).

Table 2. Mapping Between Information Quality Problem and Genomic Information Error

Quality Problem	Genomic Information Error
biased information	transformation error, propagated error, experimental error
outdated information	stale data
massaged information	analysis error

Some work has been done to improve the genomic information quality. Martinez and Hammer⁸ extended an existing data model to include quality metadata. In,⁹ a framework for the declarative specification of a user's personal quality processing requirement is proposed. Data integration approaches are also adopted in this domain.

Many research efforts have been put to apply quality theory to various application context, such as data warehousing, decision making, finance etc. However, very few work has been aimed at applying quality theory to genomic information quality even the problems studied are matched.

One of the fundamental problems with quality measurement in this domain is the lack of agreement on common quality dimension, and of practical instruments for performing quality assessments.

Some progresses have been made to set the foundation towards genomic data quality control standards for gene expression data.¹⁰ However, much efforts are required to address whole genomic information quality.

4. Conclusion

Genomic information is dirty and errors could not be avoided due to the complex information production process. Building genomic data quality standards will substantially reduce analysis cost by reducing the need for replicate experiments and in turn, speed up the drug discovery and disease prevention. Another benefit is that genomic data quality standards will facilitate future technology development. When established standards exist, it is much easier to conduct proof-of-principle studies using new systems.¹⁰

References

1. H. Müller and F. Naumann, Data quality in genome databases, in *Information Quality*, eds. M. J. Eppler and M. Helfert (MIT, 2003).
2. B. D. Klein and D. F. Rossin, *Data Quality* **5** (1999).
3. F. G. Joseph M. Juran and R. Bingham, *Quality Control Handbook* (McGraw-Hill, New York, 1974).
4. M. Ge and M. Helfert, A review of information quality research - develop a research agenda, in *The International Conference of Information Quality*, (Cambridge, Massachusetts, USA, 2007).
5. D. A. Garvin, *Managing Quality* (Free Press, 1988).
6. R. Y. Wang and D. M. Strong, *Journal of Management Information System*. **12**, 5 (1996).
7. L. L. Pipino, Y. W. Lee and R. Y. Wang, *Commun. ACM* **45**, 211 (2002).
8. A. Martinez and J. Hammer, Making quality count in biological data sources, in *IQIS '05: Proceedings of the 2nd international workshop on Information quality in information systems*, (ACM, New York, NY, USA, 2005).
9. P. Missier, S. Embury, M. Greenwood, A. Preece and B. Jin, Quality views: capturing and exploiting the user perspective on data quality, in *VLDB '06: Proceedings of the 32nd international conference on Very large data bases*, (VLDB Endowment, 2006).
10. H. Ji and R. W. Davis, *Nature Biotechnology* **24**, 112 (2006).