Measuring Data Completeness for Microbial Genomics Database

Nurul A. Emran[1], Suzanne Embury[2], Paolo Missier[3], Mohd Noor Mat Isa[4], and Azah Kamilah Muda[1]

[1]Centre of Advanced Computing Technology (C-ACT), University Teknikal Malaysia Melaka, [2]The University of Manchester, [3]The University of Newcastle, [4]Genome Malaysia Institute.
nurulakmar@utem.edu.my,embury@cs.man.ac.uk,paolo.missier@ncl.ac.uk
emno72@gmail.com,azah@utem.edu.my

Abstract. Poor quality data such as data with missing values (or records) cause negative consequences in many application domains. An important aspect of data quality is completeness. One problem in data completeness is the problem of missing individuals in data sets. Within a data set, the individuals refer to the real world entities whose information is recorded. So far, in completeness studies however, there has been little discussion about how missing individuals are assessed. In this paper, we propose the notion of population-based completeness (PBC) that deals with the missing individuals problem, with the aim of investigating what is required to measure PBC and to identify what is needed to support PBC measurements in practice. This paper explores the need of PBC in the microbial genomics where real sample data sets retrieved from a microbial database called Comprehensive Microbial Resources are used (CMR) [1].

Keywords: data completeness, population-based completeness (PBC), completeness measurement

1 Introduction

One type of completeness that has been mentioned in the literature is population-based completeness (PBC) [1]. A population consists of a set of individuals that represent real world entities, whose information are recorded in a data set. For PBC, the concern is to determine whether the data set consists of a complete set of individuals or not, which requires measuring the individuals that are missing relative to a population. The importance of PBC can be seen in the descriptions of many problems in the literature. For example, in bioinformatics, to study the genes that are responsible for certain diseases, a candidate gene set is prepared and validated before more detailed tests are performed to find the disease-causing genes [2]. According to Tiffin *et al.*, because many complex diseases could be linked to multiple gene combinations, determining whether the data set of gene

[1] CMR-http://www.tigr.org/CMR

candidates is complete or not is becoming more important in the analysis for the bioinformaticians in order to produce a more reliable set of (potential) disease-causing genes [3]. Consequently, if some genes are missing from the gene data set used in the analysis, links between those genes and the disease cannot be established.

In this example, completeness of the candidate gene data set used in the analysis is determined by consulting various gene data sources like public genome databases, gene expression databases, data on gene regulatory networks and pathways, as well as biomedical literature to check whether any gene has been missed from the data set [3]; the genes reference population is gathered from multiple sources. The need for PBC is not limited to the example just mentioned where the usage of reference populations is crucial to determine completeness of data sets under measure. However, little is known about how the reference populations are defined. In addition, how PBC is measured is unclear as the measurement method(s) used has not been described formally. Ideally, the reference populations used in PBC measurements are the representation of the real world which we would call the *true populations*. However, using true populations for PBC measurements can be hindered by the lack of knowledge of the individuals of the true populations or by the inaccessibility of the source(s) that provides information about the individuals. The alternatives to true populations are 'approximate' populations, the populations that could be accepted by the application domain's community as 'complete' reference populations in PBC measurements. However, even when approximate populations are adopted, we cannot avoid answering fundamental questions of PBC that unfortunately have not been addressed by any studies in completeness.

The rest of this paper is organised as follows. Section 2 covers the various types of data completeness proposed to date; Section 3 consists of the elements essential for PBC, Section 4 presents the example of PBC. Finally Section 5 concludes the paper.

2 Related Work

Studies in data completeness are not new; they have been conducted since at least the 1970s. During this period, the data completeness problem was well known as the problem of missing information among scholars in the database community [4] as well as among statisticians [5]. For the database community, the early work on completeness largely dealt with the problem of representing missing values (as opposed to 'empty' or undefined values) within the relational tables, where *nulls* were usually assigned for the missing values in the tables [6]. Various representations of null have been used, for example, the @ symbol [7], ω [4] and the use of variables such as x, y and z [8]. The first proposal for a measure of null-based completeness (NBC) was made by Fox, Levitin and Redman [9]; they described a datum as a triple $< e, a, v >$, where v is the value of the attribute a that belongs to an entity e [9]. Nulls were viewed from two levels of granularity: single datum level and at data collection level. At the single datum

level, a *binary measure* was proposed which checks whether a datum has a value or not; at the collection level, the study described the completeness measure as an *'aggregate' measure* that computes the fraction of the data that are null.

The tuple-based completeness (TBC) measure proposed by Motro and Rakov [10] is not only useful for detecting missing tuples, but it also helps to determine whether the tuples are accurate. TBC, in their proposal was viewed from a database level and is defined as an 'aggregate' measure as follows:

$$Completeness(of\,the\,database\,relative\,to\,the\,real\,world) = \frac{|D \cap W|}{|W|},$$

where D is the actual stored database instance while W is the ideal, real world database instance. From this definition, we gain an important insight into completeness which is completeness can be affected by the presence of errors in the data set. W in the definition represents not only a reference data set that is complete, but also a reference data set that is accurate. Nevertheless, because W is very unlikely to be acquired, the measure used the sample of W which came from alternative databases or judicious sampling (where the verification of the samples is made by humans) [11].

Schema-based completeness (SBC) however focuses on "model completeness" where Sampaio and Sampaio defined it as "the measure of how appropriate the schema of the database is for a particular application". From an XML point of view, Sampaio and Sampaio defined SBC as the number of missing attributes relative to the total number of attributes [12].

To the best of our knowledge, the first recorded use of the term 'population' in connection with completeness is in a proposal by Pipino, Lee and Wang [1]. The authors did not provide a formal definition of the PBC measure, but hinted at the presence of this useful concept through an example. In the example, the authors stated that, "If a column should contain at least one occurrence of all 50 states, but only contains 43 states, then we have population incompleteness" [1]. From the example, we observe that there is a data set under measure (from state column) in which its completeness is determined by the number of missing 'individuals' (the states) from a 'reference population' (a set of 50 states). There is a notion of reference population that is used to represent a population that consists of complete individuals. However, details of how PBC measurement is made in practice are missing from the proposal, especially in terms of how the reference populations are acquired and used. The elaboration of the concept of PBC therefore remains an open question for research in terms of the current literature. To continue exploring the notion of PBC, we present the elements essential to measure PBC in the next section.

3 The Elements of PBC

The examples that hinted at PBC given by Pipino, Lee and Wang [1] and by Scannapieco and Batini [13] help us to understand that the authors have a similar concern to each other, which is on the 'individuals' that are missing from a population. They help us to see that completeness is not only about

counting nulls or missing tuples in data sets which receives the most literature coverage. We observe from the examples that, to measure PBC, we need data sets to be measured and 'reference' populations. An explanation of how data sets under measure and their reference populations are used in terms of a formal measurement definition is, however, missing from the literature.

As presented Section 2, Motro and Rakov proposed a TBC measure [10], where the formal definition of the measure is as a *simple ratio method*. We apply a similar form of simple ratio method in our PBC measure and define a basic PBC measurement as:

$$Completeness(D, RP) = \frac{|(D \cap RP)|}{|RP|} \in [0, 1], \tag{1}$$

where D is the data set under measure, and RP is the reference population.

We can see from Equation (1) that measuring PBC is conceptually simple as we only need a data set to measure and a reference population. Nevertheless, to make the measurement workable in practice, we need to know more about the populations. The question of how the reference populations can be acquired is also essential, especially in the context of PBC measurement providers.

3.1 Populations

The term *population* is used widely, especially in statistical studies. Statisticians define a population as the entire collection of items that form the subject of a study and that share common features [14]. Within the statistical studies themselves, the definition of a population however is often specific to the application domain. For example, statistical studies in the biological and ecological domains define a population as a group of organisms of the same species in a given area ([15]). In census studies, a population is defined as the people who inhabit a territory or a state [16]. These items, species or people are the 'individuals' that belong to their defined population. In philosophy, the term *natural kind* is used for "grouping or ordering that does not depend on humans", which is the opposite for the term *artificial kind* used for grouping of arbitrary things made by human [17]. Inspired from the observation of how populations are defined in the literature and from the philosophical domain, we define population as *a set of individuals of a natural kind* and these individuals are the real world individuals (not the artificial individuals created by humans). A question that arises is: what characterises the individuals that are suitable to act as the members of populations for PBC?

Pipino, Lee and Wang pointed out in their example that, the data set that they examined are retrieved from a specific column (states) [1]. This provides us with a hint that only certain attribute of a data set might be of interest and will 'make sense' as the basis of a completeness measure. The instances in state column are therefore the data set under measure that is of interest in terms of its completeness. Thus in the example, the individuals that are suitable to act as the members of a population are a set of states.

3.2 Reference Populations

The notion of reference populations is proposed as an essential element of PBC to represent populations that are 'complete', i.e., that have no missing individuals. The question is, how can we obtain the reference population? In bioinformatics completeness of the human gene population that is used for an analysis for genes that cause diseases is of concerned [3]. Several gene sources such as public genome databases, gene expression databases, data on gene regulatory networks and pathways, as well as biomedical literature were used to retrieve the list of genes that became the reference population [3]. The reference population used in this study is the integration of several sets of genes from a variety of gene sources (identified by the bioinformaticians in the domain).

However, the reference population used in the analysis is not the *true* human gene population unless it consists of all real world human genes that exist. Because there is still a debate on the actual number of human genes among the scientists, and more time is required to discover true human genes, due to the complexity of the gene discovery process [18], the usage of the true gene population is not possible in this example. Therefore, as the alternative, we must find an approximate population to represent the true gene population.

Two forms of reference populations are possible: 1) the *true* populations that consist of all real world individuals that exist, and, 2) the *approximate* populations that are used to represent the true populations and which are more easily available. In addition, we also observe that, the individuals of a reference population may come from multiple sources. Within an application domain, we say that the decision regarding which form of reference population is to be used must be made by the domain experts (e.g., by bioinformaticians) due to their knowledge of the *sources* of the populations. The decision to use approximate populations in the examples above is driven by the costs/difficulties of acquiring the true populations, and the questionable benefits of the small differences in measurements that would result. If this is the case, the approximate populations used must be adequate for determining completeness of data sets within the domain. However, we suspect that the main reason approximate populations are used is that true populations are not feasible to obtain, even though there may be a need to use them.

The situation where there exists a single source that contains a good approximation of the true population is limited however (an exception being the Genbank database[2] that contains the genes with good evidence of their existence). We propose that good approximate populations should be established by integrating individuals from a range of sources. To describe approximate populations established from integrated sources, we adopt the term *universe of discourse (UoD)* or in short *universe*[3]. Conceptually, a universe consists of a collection of approximate populations within an application domain used for PBC

[2] http://www.ncbi.nlm.nih.gov/genbank/

[3] The term *universe* was introduced by Augustus De Morgan in 1846 in formal logics to represent the collection of objects under discussion of a specific discourse [19].

measurements, that is built by integrating individuals from several incomplete sources for the populations.

We use the term *contributing sources* (CSs) for sources that contribute to the reference populations in the universe. The CSs could be in multiple forms, such as databases (private and public) e.g., observation databases from gene regulatory networks and pathways [3] or published literature. As it is crucial to understand (and to manage) the relationship between the CSs and the reference populations for successful integration, we propose a structure called a *population map*. Conceptually, a population map consists of a mapping between a reference population and its CSs. If a reference population is stored as a table, and the CSs are databases, we say that a population map is a mapping between a reference population table and queries over tables on CSs. Note that, as the schema of the universe may not be the same as the schema of the CSs, the designers of the population maps must consider the differences.

4 Example of PBC Measurements in Microbial Genomics

The study of the genomes of microbes, called microbial genomics, helps pharmaceutical researchers gain a better understanding of how pathogens cause disease [20]. By understanding the association between pathogens and diseases, further analyses, such as regarding pathogens' resistance to drugs or antibiotics, can be conducted in search of a cure for specific diseases.

To explain the PBC problems in the microbial domain, we observed the relationships among the subjects of microbial studies that have been documented in the literature and we produced Fig. 1 as the result of these observations. The left side of the figure depicts an ER diagram with five subjects of microbial studies, namely *Microbe, Genome sequence, Gene, Infectious disease* and *Antimicrobial agent/vaccine*, and their relationships. Each relationship between the subjects is related to an analysis within the pathways of microbial genomics (the right side of the figure), shown by a dotted line. The analyses are conducted by the scientists in the wet lab through experiments, or by the bioinformaticians in the dry lab with the support of computational tools [2, 21].

We observe that for every analysis in the microbial study pathway shown in Fig. 1, the scientists/bioinformaticians need to prepare an input data set describing the subjects of interest (e.g., *microbe and gene*) for the analysis. In general, in these analyses, the completeness of the input data set determines the completeness of the analysis result. Therefore, an important question that arises in this domain is regarding the completeness of these input data sets. However, not all information in these input data sets are of interest (in terms of completeness) as scientists often look at specific information that is important to them (i.e., completeness in regards to certain genes or species) - a scenario that hinted PBC problems that are inherent in the multiple stages of analysis in microbial domain as described. The key lesson that we learnt based on the observation in microbial domain is on the applicability of the PBC concept to support answering PBC measurement requests from this domain.

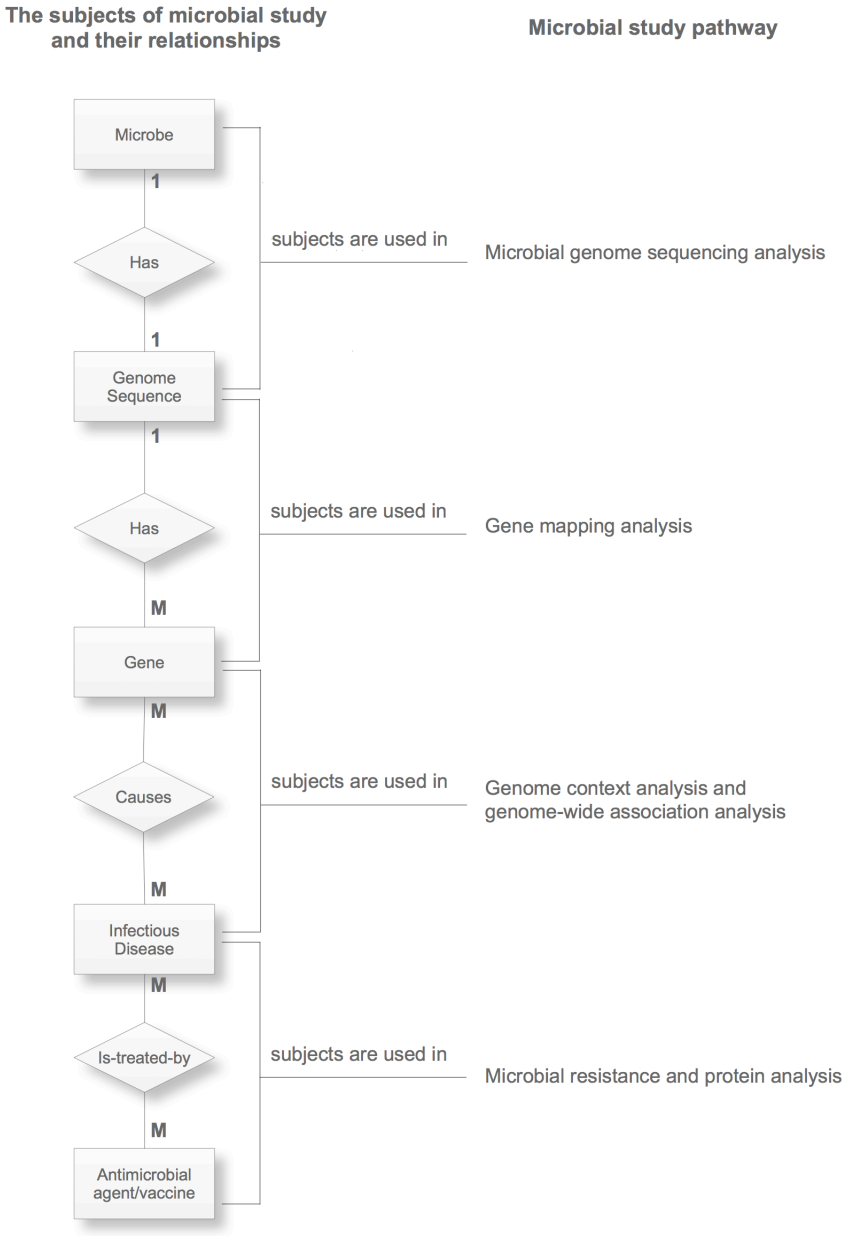

Fig. 1. The Relationship Between the Various Subjects of the Microbial Study and the Analyses in the Microbial Study Pathway

4.1 Answering PBC Measurement Request in Microbial Genomics

In handling PBC measurement requests for the microbial genomics domain, PBC measurement providers need to configure the elements of the PBC model that are specific for this domain. To describe the configuration needed for reference populations, assume that the microbial universe consists of reference populations whose individuals are from databases (CSs) in the microbial domain identified by domain experts. PBC measurement providers need to configure the form of reference population table schema, together with the information that must be stored within the tables. The basic configuration of the PBC model defines the type of reference population table schema (called the POPSCHEMA) to be in the general form of: $\langle I, source, A\rangle$ (where I is identifier attribute(s), *source* is the name of the source attribute and A is the set of attributes other than I and *source*) as as a form of schema to support all types of PBC measurement requests.

Suppose that the general form of the reference population table schema is adopted and the reference candidate gene population is configured as a table called `gene` with schema: $\langle$`geneId`, `source`, `species`$\rangle$. In addition to the reference population, we also need to configure the microbial universe, its CSs and the population maps that are specific for this domain. Based on the basic configuration of the PBC model, two variables can be defined namely UP (the set of reference population tables in the universe and their schema) and CS (the set of CSs in the universe). The following is an example of the instances of the variables configured and stored for PBC measurement in the microbial domain:

- UP={(`gene`,$\langle$`geneId`, `source`, `species`$\rangle$)},
- CS={(`CMR, http://www.tigr.org/CMR,` PM_{CMR})}, where PM_{CMR} is a set of population maps for Comprehensive Microbial Resource (CMR) in the form of:
  {`gene`, `SELECT geneCode, speciesCode FROM microbeGene`}. Every instance of PM_{CMR} consists of the name of the population table (that is equivalent to the name of the reference population it contributes), and the query against the table in the CS.
  For brevity, we only show an instance for each variable.

Assuming that all elements of PBC have been configured for the microbial domain, we will next present one type of PBC measurement requests that can be supported by the PBC configuration. For a request to measure completeness of a candidate gene population relative to the reference candidate gene population that consists of genes coming from certain CSs of the microbial universe only (e.g., CMR and SwissProt), PBC is measured as:

$$Completeness(\langle \texttt{ExtGENE}\rangle, \langle \texttt{gene}, COND\rangle) = \frac{|\texttt{ExtGENE} \cap (\Pi_{\texttt{key(gene)}}(\sigma_{COND}\texttt{gene}))|}{|\Pi_{\texttt{key(gene)}}(\sigma_{COND}\texttt{gene})|},$$

where `ExtGENE` is the external gene data set under measure, `key(gene)` is a function that retrieves `geneId` (the identifier of genes) from `gene`, *COND* is a

conjunction of conditions on `gene` using *source* attribute as the predicate. For example, one condition in *COND* is specified as `source IN ('CMR','SwissProt')` in the query.

This type of request could be driven by the need to use a reference gene population that comes from a preferred source e.g. based on trust/reputation. Because not all CSs chosen by the PBC measurement provider are preferred by the person requesting the measurement, we need to filter the genes by specifying the condition on the `source` predicate in the query. Other specific queries can be specified by adding the required predicate(s) in *COND* (e.g., the analysis may interested in genes for certain microbe species called *S.bongori*).

5 Conclusion

In conclusion, we discovered that defining what is the 'complete' reference data set (the population) to use can be difficult such as in the microbial genomics case. In this paper, the elements of PBC have been defined, and we found that the choice of using true populations (which are the true, complete reference data sets) is often hindered by the lack of knowledge of the true population individuals and technical issues (i.e. accessibility of data sources). Using approximate populations however is complicated by the task of gathering population individuals from multiple data sources that would contribute to the closest approximation of the true populations. How practical is the PBC model is one of the remaining questions that call for further investigation in our future work.

Acknowledgments. The authors would like to thank the financial assistance provided by the Universiti Teknikal Malaysia, Melaka (UTeM) and the Ministry of Higher Education, Malaysia during the course of this research and the members of Information Management Group (IMG), The University of Manchester for their constructive comments.

References

1. Pipino, L.L., Lee, Y.W., Wang, R.Y.: Data quality assessment. Communications of the ACM **45** (2002) 211–218
2. Iles, M.M.: What can genome-wide association studies tell us about the genetics of common disease. PLOS Genetics **4** (2008) 1–8
3. Tiffin, N., Andrade-Navarro, M.A., Perez-Iratxeta, C.: Linking genes to diseases: it's all in the data. Genome Medicine **1** (2009) 1–7
4. Codd, E.F.: Extending the database relational model to capture more meaning. ACM Transactions on Database Systems (TODS) **4** (1979)
5. Reich, D.E., Gabriel, S., Atshuler, D.: Quality and completeness of SNP databases. Nature Genetics **33** (2003) 457–458
6. Zaniolo, C.: Database relations with null values. Journal of Computer and System Sciences **28** (1984) 142 – 166
7. Codd, E.F.: Understanding relations (installment #7). Bulletin of ACM SIGMOD **7** (1975) 23–28

8. Imieliński, T., Lipski, J.: Incomplete information in relational databases. Journal of the ACM **31** (1984) 761–791
9. Fox, C., Levitin, A., Redman, T.: The notion of data and its quality dimensions. Information Processing and Management **30** (1994) 9–19
10. Motro, A.: Integrity = validity + completeness. ACM Transactions on Database Systems **14** (1989) 480–502
11. Motro, A., Rakov, I.: Estimating the quality of databases. In: Proceedings of the Third International Conference on Flexible Query Answering Systems (FQAS), Springer-Verlag (1998) 298–307
12. Sampaio, S.F.M., Sampaio, P.R.F.: Incorporating completeness quality support in internet query systems. In: CAiSE Forum, CEUR-WS.org (2007) 17–20
13. Scannapieco, M., Batini, C.: Completeness in the relational model: a comprehensive framework. In: Ninth International Conference on Information Quality (IQ), MIT (2004) 333–345
14. Knudson, A.: Mutation and cancer: statistical study of retinoblastoma. Proceedings of the National Academy of Sciences of the United States of America **68** (1971) 820–823
15. Hashimoto, C.: Population census of the chimpanzees in the Kalinzu forest, Uganda: Comparison between methods with nest counts. Primates **36** (2006) 477–488
16. Liang, Z., Ma, Z.: China's floating population: new evidence from the 2000 census. Population and Development Review **30** (2004) 467–488
17. Bird, A., Tobin, E.: Natural kinds. In: The Stanford Encyclopedia of Philosophy. Summer 2010 edn. (2010)
18. Science Daily: Human gene count tumbles again. `http://www.sciencedaily.com/releases/2008/01/080113161406.htm` (2008) [Online; accessed 27-June-2011].
19. Maddux, R.: The origin of relation algebras in the development and axiomatization of the calculus of relations. Studia Logica **50** (1991) 421–455
20. Falkow, S.: Who speaks for the microbes? Emerging Infectious Disease **4** (1998) 495–497
21. Fraser, C.M., Eisen, J.A., Salzberg, S.L.: Consanguinity and susceptibility to infectious diseases in humans. Nature **406** (2000) 799–803