MAGNIFICENCE OF DATA PROVENANCE FOR BIOINFORMATICS

Muhammad Farhan[1,2], M. Munwar Iqbal [1,2], Syed Muhammad Owais[3], Farhan Ullah[4]

[1]Department of Computer Science and Engineering, University of Engineering and Technology, Lahore
[2]Department of Computer Science, COMSATS Institute of Information Technology, Sahiwal., Pakistan
[3]Department of Computer Science, University of Engineering and Technology, Taxila
farhansajid@gmail.com, munwariq@gmail.com, syedmowais@ciitsahiwal.edu.pk, farhanmscs@ciitsahiwal.edu.pk

ABSTRACT — *Data provenance is important because of the diverse, heterogeneous and vastly distributed data. Proper handling of data for the better results of biosciences is necessary so that the advances in bioinformatics can be utilized for the betterment of humanity. A quality data can guarantee better results of the experiments conducted. Biological data is more sensitive as compared to other data captured or gathered for machines to enhance the technology because biological data is directly concerned with an organism or living being. The main perseverance of producing this article is that sketch the development of actions plans for data cleansing, which derived from the legacy sources or heterogeneous sources for the bioinformatics experiments. Data must be cleaned and validated before loading in confirming experiments. Moreover, general guidelines are discussed as data is cleaned and customized according to the nature of the bioinformatics requirements.*

Keywords*:* Data Provenance, Bioinformatics, Databases, Data quality, Data cleansing

I. INTRODUCTION

Data provenance or source of data is important to know for the bioinformatics experts because it resolves the issues of authenticity and trust. Tracking of data and workflow becomes easy. Provenance is a vast term used for the origin or source of data, results, and workflows, etc. Data provenance is more related the events tracking with respect to their occurrences, initiators, time factor and the purposes of the events. Wikipedia is not considered as an authentic source of information or data, so its provenance is weak [2]. Repositories of research manuscripts are an authentic source of information and knowledge.

Data provenance of biological is a big issue because the diverse and rapid growth of data and sharing of data among different research communities makes it more complex [1]. Different data provenance models proposed to make the quality of data better collection, storage, and annotation. Less research is conducted to focus it on making challenging and next generation of User Interfaces (UIs), and programming tools and languages [3]. It is difficult to make user interfaces for the biological scientists to capture the biological data so more efforts and research should be conducted to overcome this.

Data provenance becomes more important when the scientist wants to conduct the same experiment done before by some other one on the basis of the available data.

Bioinformatics is the merger of biology and computer sciences but from the computer science perspective, it is more related to computer science related issues not the biological complexities and same is the case with a biologist. Problems related to data gathering, analysis and capturingthe data should be an easy task for a biologist and all the computer related issues e.g. good interface design for bioinformatics applications, database query the computer should address . planning and optimization, the syntax of the languages and other issues of the database managements systems scientists and not by the biologists [1].

II. RELATED STUDY

To analyze and grape the trustfulness of data being captured from different sources there comes the need of mapping and measuring tools and units. So, for mapping, there are multiple ways like graphs, tables, and pictorial representations. Data Cleansing process identifies and corrects the errors arise whereas entering the data or processing it; the process includes fixing typos, inconsistent data, eliminate duplicates, spelling mistakes, incorrect or inaccurate data and validating the data is valuable.

Validation of Data comprises inspection of data for accuracy; corroborate data using the set of guidelines appropriate. Certain validation approaches e.g. statistical data, check data type, check zip code or digits for postcode check according to country format like US format, check for constancy and reliability, formatting checks, UK formats, check for the sum of accounts or bills, spelling check, etc. Data Verification is to confirm and cross-check the data for the accuracy; it is good to address verification by various means like web research or from additional sources or by a telephone call

of the data oriented products which are consumed & produced through these applications is necessary for the disambiguating data & allow reusability. Data provenance is a sort of meta-data that relates to the extraction from the history of data product initiating from its origin. Data provenance of products produced through composite conversions processes for instance workflows is of significant and noteworthy to scientists and researchers. One can determine the quality of data on the basis of its inherited data, and origins permit automated re-enactment of origins to update

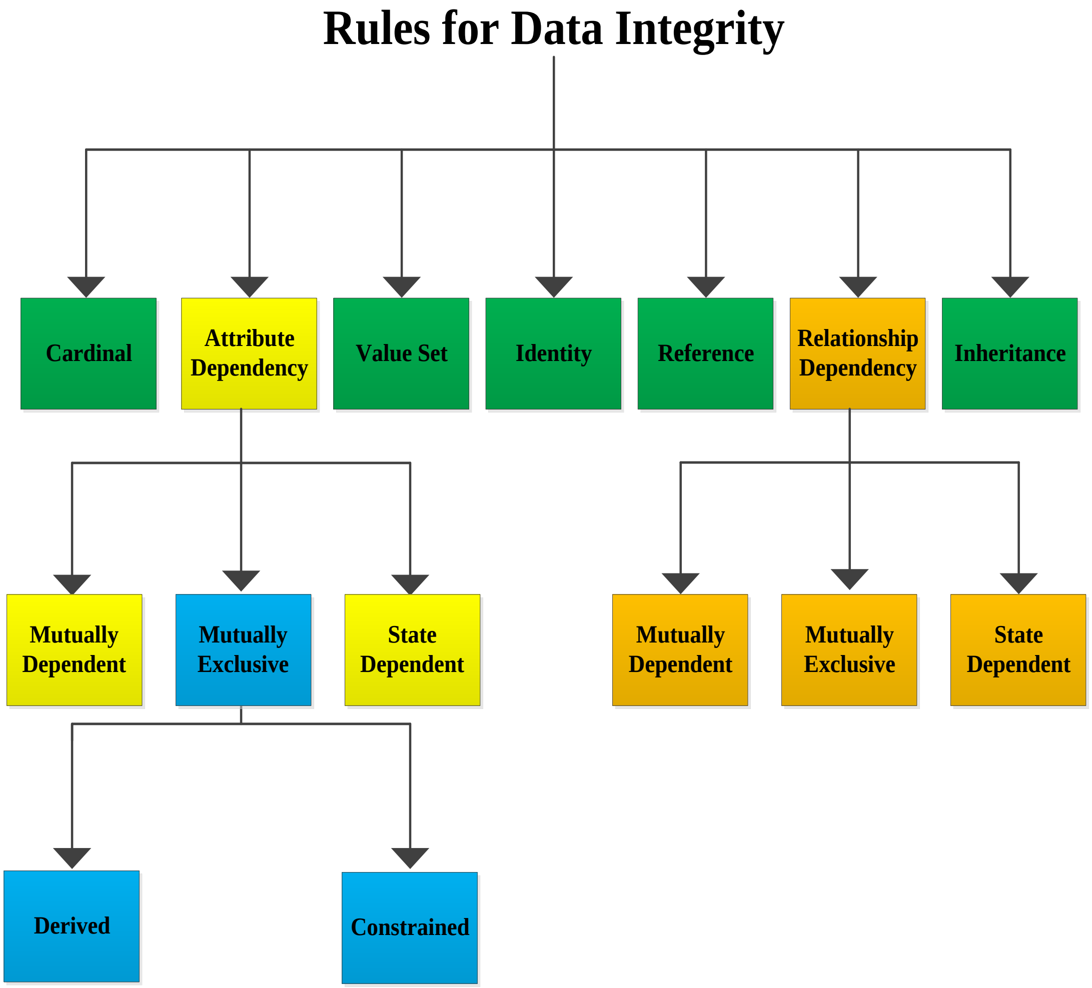

Figure 1: Classifying Integrity Rules

Table 1: Data Quality Constraints

Properties	**Details**
1. Cardinality check	This check is used to validate the record related records. e.g. like each customer has must recorded minimum one linked Order (Cardinality > 0). In case, no order exist against a "customer" record at that time it essential to either altered to "seed" or order must be created. It is a cardinality check and can be complicated by supplementary constraints. Another example, if a "former employee" has left the organization then there must be no record in Payroll database is noticeable, then have not any related salary payments later on the period on which worker gone form organization (Cardinality = 0).
2. Batch totals	It validates the 'Total Cost' attribute by taking the sum of transactions together.
3. Allowed character checks	In it necessary to Numeric field only allow the digits from 0-9, an email id must have one @ symbol & various other essential particulars.
4. Control totals	Control total can be only done on the numeric attributes acts in each record. Total must be meaningful, e.g., calculate the sum of the total fee for a number of students.
5. Cross-system consistency checks	Data item exists in the different system must be ensured its consistency: like that address for a student having same id & same name in both systems may be denoted in a different way in diverse systems. It is essential to be transformed into a mutual format to be matched, for example, at one database system data item may be saved student name filed in distinct Name attribute or field as 'Ali, Khan, Nadir', however another in three, unlike fields, stored: First Name (Ali) separately, Middle Name (Khan) and Last Name (Nadir). For evaluation of these two records, validation engine must require transforming data item of the second system to evaluate the data with the first system, while using SQL query: Last Name || ', ' || Middle Name ||

	subsite(First Name, 1, 1) it may convert data items on the 2nd system must be appeared same as the data item in the first system 'Ali, Khan, Nadir'
6. Consistency checks	If one system data item may be saved as Gender represented as Male or Female. However in the other system it may be stored as Title = "Mr.", or "Miss" then it must be checked that title "Mr." can be only kept against Gender = "M" or "Male."
7. Check digits	An additional number may be added to an integer that calculated the starting digits. Digits system check validates this computation while numbers are entered. E.g. very last number of an ISBN for a journal or is checked by the digits calculated through modulus 100.
8. Data type checks	Data type must be checked while entering the data into fields. If the data is entered wrongly, then the message will be displayed that data type mismatch or an error message. It is ensured that input data must match with the selected data type, For example, In the data item input box that only accepting the numerical data items, the letter 'l' was entered in its place of the number 1, an error message must be displayed.
9. File existence check	Files having specified names would exist. This validation is necessary for the programs which use file handling.
10. Format or picture check	If you have to enter the date can input in a particular format (pattern), for example, date values have to be inputs in a format DD/MM/YYYY and this type of validation should be done by the help of Regular expressions.
11. Hash totals	The batch total can be done on one or additional numeric attributes that seem in each tuple. It has no meanings but only concatenated total, for example, add Phone Numbers together.
12. Limit check	The limit check may be used one limit only, either upper OR lower, for example, password length must at least six characters (=>6).
13. Logic check	It must validate that divisor cannot be input as zero when dividing a number anywhere in a program.
14. Presence check	This Check is significant for facts to how in reality is denoted and also not lost out, for example, Student might be compulsory to ensure their registration number recorded.
15. Range check	Range Checks those in which records lie in a indicated range of values, for example, students birth date must lie in the range between 1 and 31.
16. Referential integrity	Referential integrity is played important Relational database standards tables that may be interconnected through the foreign key of one table & other table's primary key. The records related to primary key attribute are not controlled by inner database method, it must be authenticated. Confirmation of the one table foreign key attribute checks which referencing relation necessarily permanently denote to a valid record in the referenced relation.
17. Spelling and grammar check	Spelling & grammar check must validate the mistakes take placed the word spelling errors and linguistic mistakes.
18. Uniqueness check	Uniqueness Check validates that every fact and figure is distinct. This check may be implemented to numerous fields like Id, Name, Telephone number, Address.
19. Table Look Up Check	This check during validation receipts the fact input entry & matches it to confirmed list of inputs which are saved in a database relation.

III. Application of Provenance

Bioinformatics data in the digital library is classically a huge and diverse in nature of online databases and the form of documents along well-developed software for the traversal of the collection of data. Scholarly resources organize some digital libraries. The Actual question is that how we cite an element of a digital library? Astonishingly this has gained too little consideration. Good standards for citation are missing. Well, prepared databases are built along keys which permit the user uniquely to identify a record in a database table. We can identify a component of a database record by providing attribute name. Consequently, there is typically an undisputed path to any constituent of the database [11].

Forthcoming additions to computer-generated data provenance model comprise preserving a transactional provenance track of modifications to metadata annotations, the application of the model to a distributed web of provenance catalogs employing a similar schema and management of provenance data retention [12].

Management of data is emergent in density by way of large scale applications that yield benefit of loosely bonded resources joined through grid middleware and the plentiful capacity of storage. Meta-data is the description data, backtracking of root causes of errors & make available attribution of data origins. Provenance is also critical for the domain of commerce wherever it may use to mining the data source in data ware-house which tracks the formation of intellectual stuff & delivers a review track for governing purposes. A major feature of the taxonomy classifies origin systems depending upon that why they record provenance. How do they denote, what they designate and store provenance and ways to distribute it [13].

IV. QUALITY OF DATA IN BIOINFORMATICS

Bioinformatics and Genome-based databases store data regarding molecular biological entities like that protein, genes, diseases, etc. The major objective of developing and maintaining similar databases in commercial corporations is their significance in the procedure of drug discovery.

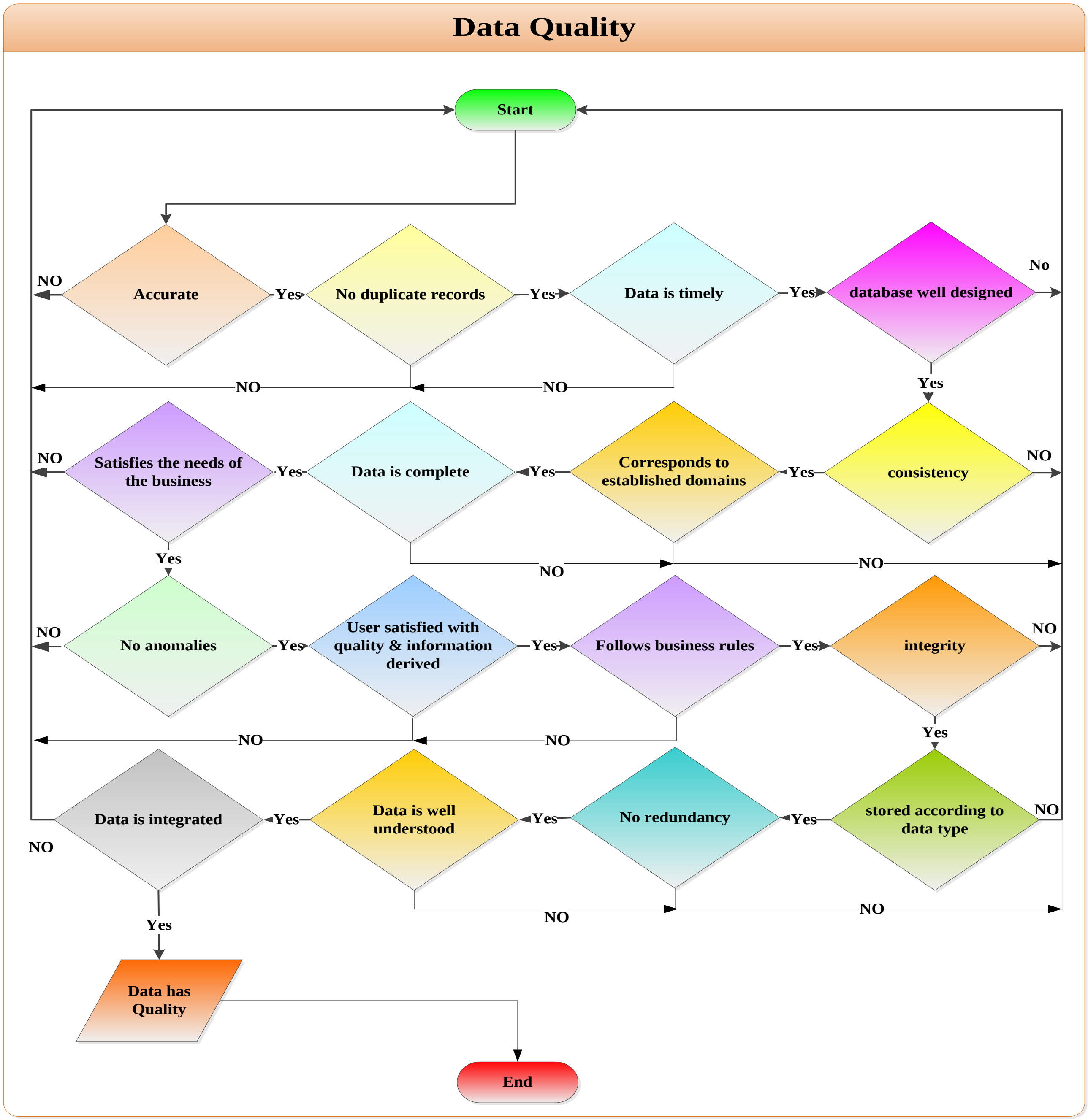

Figure 2: **Data Quality Verification Processes**

Bioinformatics data is analyzed as well as assessed to achieve so-called leads, for example promising structures intended for novel drugs. Following a lead in the course of the procedure of drug testing, development, as well as lastly numerous stages of clinical trials, is very expensive. Therefore, a high-quality fundamental database is of greatest significance. Because of exploratory nature of Bioinformatics databases, commercial as well as public, they are incomplete, inaccurate, outdated as well as in a general poor state [9].

Data of less effective quality in Bioinformatics databases have huge economic as well as medical influence on their users/customers. For example, errors in Bioinformatics data are able to result in inappropriate target selection intended for biological tests or pharmaceutical research. To perform a handful of novel drugs to the marketplace, pharmaceutical corporations spend up billions of dollars in such research. Of thousands of assuring leads derived as of experimental genomic data simply a handful reach clinical trials and only a single drug turns out to be marketable. Clearly, it is of enormous significance to base these far-reaching decisions on high quality data [10].

We can represent the quality of data

$$\sum_{i=1}^{n} \frac{Q_i}{k}$$

Where ***n*** is the total number of quality parameters, ***k*** is the number of activities or checks performed in parallel fashion to ensure the quality of data and ***Q*** is the Quality of Data. The complexity for ensuring the quality of data is linear as we need ***n*** / ***k*** steps or $O(n)$ asymptotic time to perform the steps.

V. THE ROLE OF DATA CLEANSING TOOL

Issues related to the cleaning of data should be handled to improve the quality of data. As the data is too extensive and huge in volume so some automated tools can resolve the issues of cleaning the data. Data cleaning includes the format of data, changing legacy data according to the new requirements, standardizing, finding and removing duplicates in the data, making clusters on the basis of similarity, merging identical records and correcting the structure of the data [7]. BIO-AJAX is a tool for the resolution; detection and duplication taxonomy of organisms that employ prefix harmonizing policy to incorporate diverse terms that depict the identical species [6].

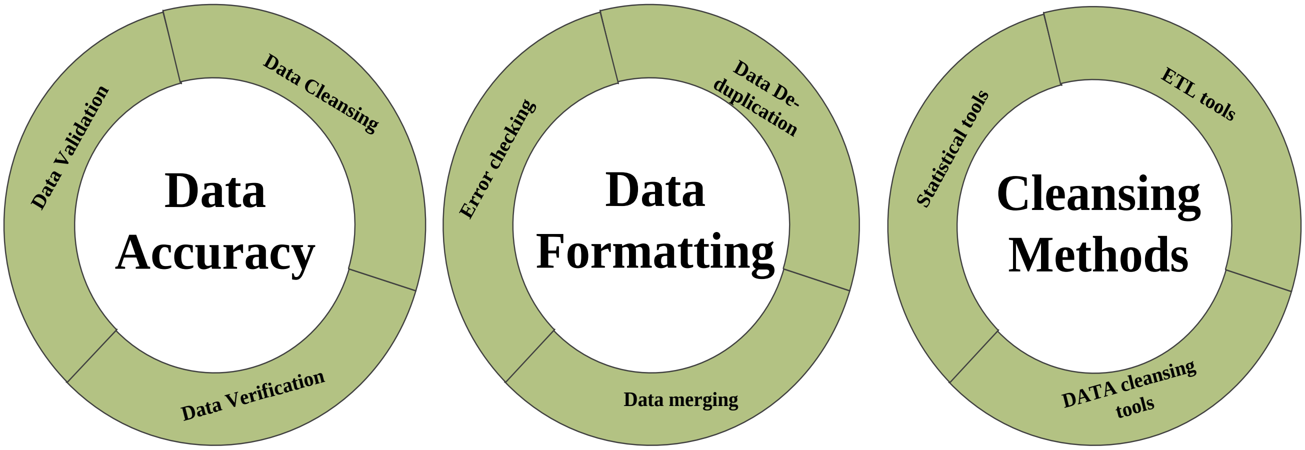

Figure 3: (a) Data Accuracy, (b) Data Formatting, (c) Data Cleansing Methods

Data accuracy involves data cleansing, data validation and data verification [Figure 3]. The process of Data Formatting [Figure 4] is required to process the data according to the specifications, e.g. extracted data from sources can't be imported to database straightforward because it there is a lot of chances to have diverse structure. Checking and formatting are necessary as per required specifications and after that, it can be imported not having any errors e.g. conversion of different image formats among each other as tiff to PDF and BMP to GIF.

Data De-duplication means checking for copies and remove matching tuples. This process is needed to enhance and increases the overall quality of the data, the speed of the service and disk usage. Data updating means the verification of existing data to and updating if any changes or correction are required in it. Updating the data process can save and keep the database up to date. Data methods involve statistical tools, ETL tools, and data cleansing tools [Figure 5].

.

VI. DATA CLEANING CASE STUDY

In the first step, duplicates are identified and then eliminated or removed. Following is a good example to demonstrate this process [Table 2, 3 & 4]. As shown in the following example [Table 2] has unlearned and unsorted data that means duplicates are present. Next process is to sort the records with appropriate field, or it can be according to the primary key. In [Table 3] sorting and detection of the duplicates is done. In [Table 4] final table after removing the duplicates is shown

Table2: With Unclean Bioinformatics' Data

gene	nProbe	cnt	Panova	Early ring S	Early ring S_LogP	Late ring S	Late ring S_LogP	Early trophozoite S	Early trophozoite S_LogP	Late trophozoite S	Late trophozoite S_LogP	Early schizont S
PFC0210c	20	7	1.04E-62	61.6	-3.5	282.3	-6.3	835.3	-8.6	428.6	-8.6	620.2
PFL1030w	20	7	3.90E-37	58.6	-2.1	127.5	-3.8	48.2	-3.6	105.6	-2.3	60.7
PF10_0350	20	7	7.48E-35	7218.5	-10.9	3920.4	-10	2161.7	-9.9	391.1	-9.2	348.8
PFE0065w	20	7	9.72E-34	7962.1	-16.1	2687.9	-15.9	955.1	-14.8	265.8	-12.1	314.3
PF08_0126	20	7	3.51E-32	50.9	-1.8	33.1	-2	101	-8.2	123.7		200.8
PF10_0218	20	7	2.23E-31	43.1	-5.6	39.8				242.3		384.1
PFE1540w	20	4	3.78E-31	8.4	-0.9	21.9	-0.3	21.1	-1.7	45.9		36.4
PFE0070w	20	7	3.16E-29	11527.6	-13.9	2050.2	-12.7	782.5	-11.7	544.6	-8.9	1169.6
PF10_0295	20	7	1.58E-28	78.5	-6.3	62.9	-4.1	49.9	-3.5	95	-7.1	800.2
MAL6P1.27	20	0	8.58E-28		-0.1	0	0	7.1	-0.2	7.2	0	8.8
PF13_0269	20	5	4.49E-27	22.4	-1.2	18.3	-0.3	32.8	-1.8		-2.7	77.5
MAL8P1.16	15	6	9.53E-27	35.1	-0.7	12.2	-0.1	36.6	-1.8	47.6	-2.3	49
PFD0340c	20	0	1.08E-26	0	-0.1	0	-0.5	2.7	0	7.1	-0.5	8.4
PFI0185w	20			20.2	-0.5		0	5.1	0	0.2	0	1.5
PF11_0307	20	1	1.88E-26	2.8	-0.1	0	0	6.2	0	6.6	0	11.2
PF14_0074	20	1	8.92E-26	10.5	-0.2	5.6	0		-0.3	11	-0.3	0
PF13_0201	20	5	1.05E-25	16.2	-0.9	29.2	-0.5		-1.5	24	-0.4	56.6
PF14_0683	20	0	1.13E-25	0	0				-0.3	0.1	0	7.8
MAL7P1.10	20	0	1.80E-25	4			-0.3	4.4	-1.3	3.8	-0.1	4.5
PFE1540w	20	4	3.78E-31	8.4	-0.9		-0.3	21.1	-1.7	45.9	-4.4	36.4
PFE1540w	20	4	3.78E-31	8.4	-0.9		-0.3	21.1	-1.7	45.9	-4.4	36.4
PF10_0295	20	7	1.58E-28	78.5	-6.3	62.9	-4.1	49.9	-3.5	95	-7.1	800.2
MAL8P1.16	15	6	9.53E-27	35.1	-0.7	12.2	-0.1	36.6			-2.3	49
PFI0185w	20	0	1.71E-26	20.2	-0.5	6.6	0	5.1	0	0.2	0	1.5
PFI0185w	20						0	5.1	0	0.2	0	1.5
PFE0065w	20	7	9.72E-34	7962.1	-16.1	2687.9	-15.9	955.1	-14.8	265.8	-12.1	
PF14_0074	20	1	8.92E-26	10.5	-0.2	5.6	0	2.2	-0.3	11	-0.3	
PF14_0683	20	0	1.13E-25	0	0	0	0	0.6	-0.3	0.1	0	
MAL7P1.10	20	0	1.80E-25	4	-0.2	3.4	-0.3	4.4	-1.3	3.8	-0.1	
MAL7P1.10	20	0	1.80E-25	4	-0.2	3.4	-0.3	4.4	-1.3	3.8	-0.1	

VII. USAGE OF D ATA PROVENANCE

The most important usage of data provenance in the decision making especially it is dependent on the reliability, consistency, and credibility of provenance. Affirmations about lineage o f data always have to be assisted through the identity of a reliable person or associations in the direction of make this assertion significant [14]. Several versions of fact delivered by different individuals can also be convoluted in data derivation & will be reconciled and mediated [l5, 16]. Provenance statistics must have an authenticity that was not modified and tampered through authorization provenance by means of digital signatures exists as a solution. Such definite assertions on data provenance will improve its worth and lead to extensive usage of provenance for judgment and decision-making

Table3: After Sorting and Detecting Duplicates in Bioinformatics' Data

gene	nProbe	cnt	Panova	Early ring S	Early ring S_LogP	Late ring S	Late ring S_LogP	Early trophozoite S	Early trophozoite S_LogP	Late trophozoite S	Late trophozoite S_LogP	Early schizont S
MAL6P1.27	20	0	8.58E-28	7.8	-0.1	0	0	7.1	-0.2	7.2	0	8.8
MAL7P1.10	20	0	1.80E-25	4	-0.2	3.4	-0.3	4.4	-1.3	3.8	-0.1	4.5
MAL7P1.10	20	0	1.80E-25	4	-0.2	3.4	-0.3	4.4	-1.3	3.8	-0.1	4.5
MAL7P1.10	20	0	1.80E-25	4	-0.2	3.4	-0.3	4.4	-1.3	3.8	-0.1	4.5
MAL8P1.16	15	6	9.53E-27	35.1	-0.7	12.2	-0.1	36.6	-1.8	47.6	-2.3	49
MAL8P1.16	15	6	9.53E-27	35.1	-0.7	12.2	-0.1	36.6	-1.8	47.6	-2.3	49
PF08_0126	20	7	3.51E-32	50.9	-1.8	33.1	-2	101	-8.2	123.7	-7.3	200.8
PF10_0218	20	7	2.23E-31	43.1	-5.6	39.8	-2.3	129.5	-7.2	242.3	-8.2	384.1
PF10_0295	20	7	1.58E-28	78.5	-6.3	62.9	-4.1	49.9	-3.5	95	-7.1	800.2
PF10_0295	20	7	1.58E-28	78.5	-6.3	62.9	-4.1	49.9	-3.5	95	-7.1	800.2
PF10_0350	20	7	7.48E-35	7218.5	-10.9	3920.4	-10	2161.7	-9.9	391.1	-9.2	348.8
PF11_0307	20	1	1.88E-26	2.8	-0.1	0	0	6.2	0	6.6	0	11.2
PF13_0201	20	5	1.05E-25	16.2	-0.9	29.2	-0.5	31.4	-1.5	24	-0.4	56.6
PF13_0269	20	5	4.49E-27	22.4	-1.2	18.3	-0.3	32.8	-1.8	51.5	-2.7	77.5
PF14_0074	20	1	8.92E-26	10.5	-0.2	5.6	0	2.2	-0.3	11	-0.3	0
PF14_0074	20	1	8.92E-26	10.5	-0.2	5.6	0	2.2	-0.3	11	-0.3	0
PF14_0683	20	0	1.13E-25	0	0	0	0	0.6	-0.3	0.1	0	7.8
PF14_0683	20	0	1.13E-25	0	0	0	0	0.6	-0.3	0.1	0	7.8
PFC0210c	20	7	1.04E-62	61.6	-3.5	282.3	-6.3	835.3	-8.6	428.6	-8.6	620.2
PFD0340c	20	0	1.08E-26	0	-0.1	0	-0.5	2.7	0	7.1	-0.5	8.4
PFE0065w	20	7	9.72E-34	7962.1	-16.1	2687.9	-15.9	955.1	-14.8	265.8	-12.1	314.3
PFE0065w	20	7	9.72E-34	7962.1	-16.1	2687.9	-15.9	955.1	-14.8	265.8	-12.1	314.3
PFE0070w	20	7	3.16E-29	11527.6	-13.9	2050.2	-12.7	782.5	-11.7	544.6	-8.9	1169.6
PFE1540w	20	4	3.78E-31	8.4	-0.9	21.9	-0.3	21.1	-1.7	45.9	-4.4	36.4
PFE1540w	20	4	3.78E-31	8.4	-0.9	21.9	-0.3	21.1	-1.7	45.9	-4.4	36.4
PFE1540w	20	4	3.78E-31	8.4	-0.9	21.9	-0.3	21.1	-1.7	45.9	-4.4	36.4
PFI0185w	20	0	1.71E-26	20.2	-0.5	6.6	0	5.1	0	0.2	0	1.5
PFI0185w	20	0	1.71E-26	20.2	-0.5	6.6	0	5.1	0	0.2	0	1.5
PFI0185w	20	0	1.71E-26	20.2	-0.5	6.6	0	5.1	0	0.2	0	1.5
PFL1030w	20	7	3.90E-37	58.6	-2.1	127.5	-3.8	48.2	-3.6	105.6	-2.3	60.7

Table4: Cleaned and Rectified Bioinformatics' Data

gene	nProbe	cnt	Panova	Early ring S	Early ring S_LogP	Late ring S	Late ring S_LogP	Early trophozoite S	Early trophozoite S_LogP	Late trophozoite S	Late trophozoite S_LogP	Early schizont S
MAL6P1.27	20	0	8.58E-28	7.8	-0.1	0	0	7.1	-0.2	7.2	0	8.8
MAL7P1.10	20	0	1.80E-25	4	-0.2	3.4	-0.3	4.4	-1.3	3.8	-0.1	4.5
MAL8P1.16	15	6	9.53E-27	35.1	-0.7	12.2	-0.1	36.6	-1.8	47.6	-2.3	49
PF08_0126	20	7	3.51E-32	50.9	-1.8	33.1	-2	101	-8.2	123.7	-7.3	200.8
PF10_0218	20	7	2.23E-31	43.1	-5.6	39.8	-2.3	129.5	-7.2	242.3	-8.2	384.1
PF10_0295	20	7	1.58E-28	78.5	-6.3	62.9	-4.1	49.9	-3.5	95	-7.1	800.2
PF10_0350	20	7	7.48E-35	7218.5	-10.9	3920.4	-10	2161.7	-9.9	391.1	-9.2	348.8
PF11_0307	20	1	1.88E-26	2.8	-0.1	0	0	6.2	0	6.6	0	11.2
PF13_0201	20	5	1.05E-25	16.2	-0.9	29.2	-0.5	31.4	-1.5	24	-0.4	56.6
PF13_0269	20	5	4.49E-27	22.4	-1.2	18.3	-0.3	32.8	-1.8	51.5	-2.7	77.5
PF14_0074	20	1	8.92E-26	10.5	-0.2	5.6	0	2.2	-0.3	11	-0.3	0
PF14_0683	20	0	1.13E-25	0	0	0	0	0.6	-0.3	0.1	0	7.8
PFC0210c	20	7	1.04E-62	61.6	-3.5	282.3	-6.3	835.3	-8.6	428.6	-8.6	620.2
PFD0340c	20	0	1.08E-26	0	-0.1	0	-0.5	2.7	0	7.1	-0.5	8.4
PFE0065w	20	7	9.72E-34	7962.1	-16.1	2687.9	-15.9	955.1	-14.8	265.8	-12.1	314.3
PFE0070w	20	7	3.16E-29	11527.6	-13.9	2050.2	-12.7	782.5	-11.7	544.6	-8.9	1169.6
PFE1540w	20	4	3.78E-31	8.4	-0.9	21.9	-0.3	21.1	-1.7	45.9	-4.4	36.4
PFI0185w	20	0	1.71E-26	20.2	-0.5	6.6	0	5.1	0	0.2	0	1.5
PFL1030w	20	7	3.90E-37	58.6	-2.1	127.5	-3.8	48.2	-3.6	105.6	-2.3	60.7

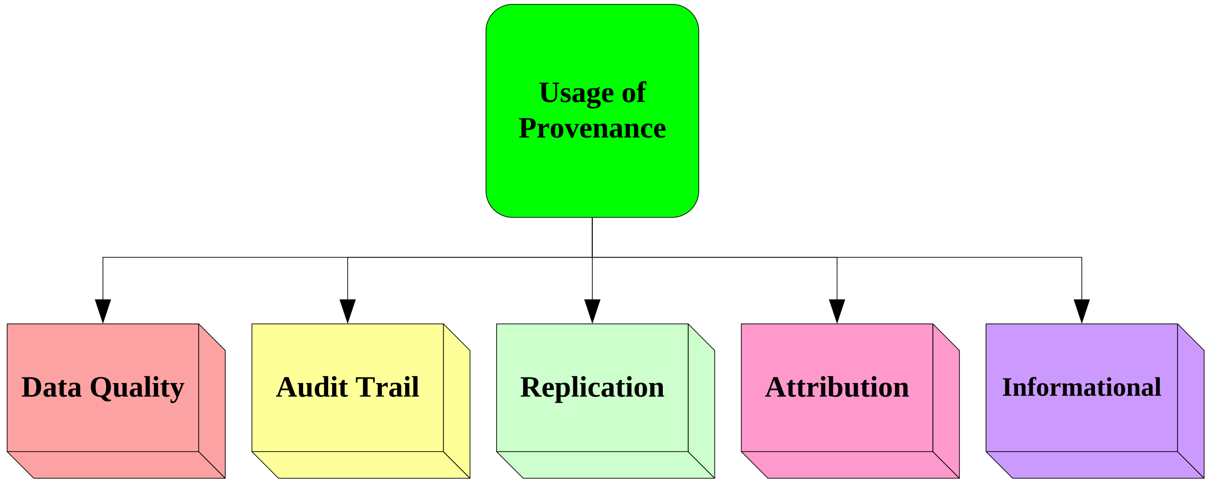

Figure 6: Usage of Data Provenance

Data provenance granularity depends on discipline for which data is collected and application in which used it. The issue of naming datasets uniquely is related to the developing standards to represent provenance, and they can be referred by the provenance [17]. Scaling is also important for efficiently federating the storage, collection & retrieval of provenance is always essential for scaling it to across different communities. Moreover provenance is always used in the data quality, audit and trail replication and attribution for the informational purposes.

VIII. CONCLUSION AND SUGGESTIONS

An enormous amount of data is produced as a result of biological experiments, so it is necessary to make the data clean and properly structured. Data cleaning processing should be carried out to make data clean and issues free so that proper analysis and reports can be generated on which useful results can be extracted in the form of the betterment of humankind. Checking and formatting of data are necessary as per required specifications and after that, it can be imported not having any errors.

The origin of provenance & state the straightforwardness in order for provenance is to be valuable away from a separate group or company. Means to combine provenance facts and figures are necessary for expandable storage, pool and recovery of origin. Development of joint meta-data semantic terms, principles & service interfaces to organized provenance in varied areas will also take part to a broader acceptance of origin and endorse its distribution. The capability to flawlessly embody provenance of data resultant from equally databases and workflows may assist in its portability and manageability. Means to record provenance about deleted or lost data need additional attention. Lastly, a deeper sympathetic of provenance is required to find out innovative and original methods to leverage it.

REFERENCES

[1]. Syed Ahsan, and Abad Shah., "Biological Databanks: Distribution, Heterogeneity, Diversity and Provenance." 7th Workshop on Distributed Data and Structures. Santa Clara, California, January 4-5, 2006

[2]. Ram, S. and J. Liu (2009), "A new perspective on Semantics of Data Provenance", Citeseer

[3]. Buneman, P. and S. B. Davidson (2010). "Data provenance–the foundation of data quality."

[4]. W. Tan P. Buneman, S. Khanna., "Data provenance: Some basic issues", In Proc. of FSTTCS, 2000.

[5]. H. V. Jagadish and F. Olken, "Database Management for Life Sciences Research," in SIGMOD Record, vol. 33, 2004, pp. 15-20.

[6]. K. G. Herbert, N. H. Gehani, W. H. Piel, J. T. L. Wang, and C. H. Wu, "BIOAJAX: An extensible framework for biological data cleaning", Sigmod Record, vol. **33**, no. 2, pp. 51-57,2004.

[7]. Herbert, K.G. and Wang, J.T.L. (2007) 'Biological data cleaning: a case study', Int. J. Information Quality, Vol. **1**, No. 1, pp.60–82.

[8]. Daniele Apiletti, Giulia Bruno, Elisa Ficarra, Elena Baralis (2006), "Data Cleaning and Semantic Improvement in Biological Databases", Journal of Integrative Bioinformatics, **3**(2), pp: 1-11.

[9]. Müller, H., F. Naumann, et al. (2003). Data quality in genome databases, Citeseer.

[10]. Trust, W. (2003). "Sharing data from large-scale biological research projects: a system of tripartite responsibility".

[11]. W. Tan P. Buneman, S. Khanna. Data provenance: Some basic issues. In Proc. of FSTTCS, 2000.

[12]. Zhao, Y., Wilde, M., Foster, I., Applying the Virtual Data Provenance Model, Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW2006), Lecture Notes in Computer Science, Springer, 2006.

[13]. YOGESH SIMMHAN, BETH PLALE, D. G. A Survey of Data Provenance Techniques. Technical Report IUB-CS-TR618, Indiana University, Bloomington, 2005.

[14]. C. A. Lynch, "When documents deceive: Trust and provenance as new factors for information retrieval in a tangled web," in JASIST, vol. **52**, 2001, pp. 12-17.

[15]. D. Pearson, "Presentation on Grid Data Requirements Scoping Metadata & Provenance," in Workshop on Data Derivation and Provenance, Chicago, 2002.

[16]. P. Groth, S. Miles, W. Fang, S. C. Wong, K.-P. Zauner, and L. Moreau, "Recording and Using Provenance in a Protein Compressibility Experiment," HPDC, 2005.

[17]. G. Miklau and D. Suciu, "Enabling Secure Data Exchange," in Data Engineering Bulletin, Special Issue on Data Security and Privacy, 2004.