Bioinformatics tools analysis framework for quality assessment: an effort analysis incorporation

Jayanthi Manicassamy* and Dhavachelvan P.
*Department of Computer Science, Pondicherry University, Pondicherry, India, jmanic2@yahoo.com, dhavachelvan@gmail.com

Abstract- Bioinformatics merges with few other sciences involving computational biology which deals at molecular level containing enormous tools, techniques and approached used for solving real world biological problems by means of data analysis and interpreting accurate results. As there is enormous growth in biological data's there is a need to manage these data's in databases for global utilization which could be utilized for various functionalities like sequence analysis, functional analysis, structural analysis etc. in computational biology. For various functional interpretations and accruing required result based on the analysis carried on these biological data's for which mining plays a vital role where pattern analysis and text extraction considered to play major part in this area of bioinformatics. Apart from this web also plays a vital role, in the conceptual view it is a means of resource sharing and updating which is an inevitable one today in this area, utilized by most of the tools and techniques. Tools build in this area could not be build by one databases so, database like NCBI, PDB, EMBDL, Medline etc… have been developed to share its resources. Some of the areas in which pattern and text mining is applicable are in Disease identification, Drug discovery, Biomedical Literature etc. At present, development of bioinformatics tools are tremendously increasingly for real-time decision making for which it is vital to evaluate tools. Various quality based assessments have been carried out on bioinformatics tools for which a framework designed that incorporated effort analysis which have been explained theoretically in this paper. The major aim behind this framework design is to fit the tool into apt categorization for carrying out qualitative analysis on the tools. Apart from this analysis based on effort through cost, time and resources have been carried out with could also be considered as a qualitative means. From the analysis made the framework designed proven to be more effective and efficient.

Keywords – Bioinformatics, Pattern Matching, Evaluation, Software Metrics, Standard Evaluation

Introduction

The area of bioinformatics have arisen the needs of biologists to utilize and help interpret the vast amounts of data that are constantly being gathered in various areas and in various modes for biological experiments. It plays a vital role in the various areas of biology which involves, pattern analysis and text mining in this area of bioinformatics with multipurpose text mining in life science to aid the biomedical researcher with a broad range of information needs [5]. Discrimination of cancer [6] evaluation from gene expression, protein sub-cellular localization from experimental data [7] by extracting features from raw images with possibility of applying genetic interactions to predict pathways have been evolved in this field. Various other computational involvements are in sequencers are capable of reading off a sequence of nucleotides in a strand of DNA in biological samples etc. Bioinformatics tools aim to exploit the potential of computers to model and understand complex biological systems and phenomena. Initially, the approach used by most tools was more analytical and engineering-like, separating DNA from all other factors and isolating the sequence for a better understanding of the DNA structure. Today, the genomics approach combines various elements, while taking into consideration their relationships as well as evolutionary aspects in space and time.

Many tools have been developed which has been categorized based on the functionalities are Homology and Similarity tools for gene expression for finding genes with a similar spatial expression pattern which could potentially reveal novel or unknown genes involved in similar processes or pathways [2]. Protein function analysis tools for functional analysis allowing the user to interactively select GO terms according to their significance and specific biological complexity within the hierarchical structure [3]. Structural Analysis Tools for DNA structure prediction and sequence analysis that has been adapted to deal with the experimental riches of complex and multivariate data in biological problem solutions [4]. Tools developed for image annotation allows easy annotation and instant sharing of images which uses a collection of large dataset that spans many object categories, often containing multiple instances over a wide variety of images. Tools including "Database search" like text based search, sequence based search, motif based search, structure based search etc. Since biological resources are real entity that should be kept updated based on the researches that requires vast space, it could not be build by one so database like NCBI, PDB, EMBDL, Medline etc. has made to share its resources. Most of the bioinformatics related applications and tools utilities these Databases. Usage of ontology in bioinformatics as a means of accessing information automatically from large databases for complex alignment of genomic sequences, in clustering of microarray DNA, pattern discovery [8, 9] etc. Biomedical uses various search techniques for mining literatures such as those offered by NCBI, PubMed system, require significant effort on the part of the searcher, and inexperienced searchers for using the systems effectively as experienced for easy and effective extraction [10, 11]. Today tremendous development of web tools and techniques are required to meet the demands in this area of bioinformatics.

Today high throughput and high content screening techniques allow biologists to gather data at an unprecedented rate. However, text mining and pattern analysis of information extraction is the most important techniques used for data analysis at sufficient rate involving various distinct methods [12]. Apart from this it is necessary for development of new of text mining and pattern tools or best utilization of better developed tool to meet its demand in this area of bioinformatics. Thus for best utilization of better tool from set of required functionality tools developed, quality assessment of the tool through quantitative based evaluation and analysis is indispensable. Metrics is a measure that qualifies the characteristics of a product by qualitative means that are being observed directly or indirectly. Pattern based and text mining tools are less commonly evaluated by measuring their contribution to the performance of some task which critical to accurately assessing their basic functionality, but they do not necessarily tell us how well a system will perform in practical applications [1]. Evaluating tools mainly allows to asses the quality. Since quality analyses have been proven very powerful in computer science and other sciences, statistical analysis is also being made to identifying the qualifying levels. Here in this paper for assessing tools quality accurately through various means involving effort a framework have been designed which have been explained theoretically along with the processing steps for carrying out the analysis on various modes.

Method for bioinformatics tools analysis

Now a day's tools in the area of bioinformatics are being developed enormously due to vast increase in biological data's and there is a need for making it computationally. Since these tools are involved in real world problem solving there is need for efficient and effective tool for better decision making. Standards in evaluations have been considered since they are the means through which a product could be analyzed whether it is effective and efficient for usage that would satisfy the need of the user requirements. So the developed tools or for best tools utilizations for better decision making have to be evaluated for their effectiveness and efficiency whether they are upto the user required standard.

Apart from the required features and functionalities of the tools selected quantitative based analyses have to be carried out on each tool. Several researches are undergoing in the area of bioinformatics involving text mining, sequence analysis and pattern matching which had made a way for the development of many algorithms, approaches and tools that have been proposed and developed in bioinformatics. Before moving towards quantitative based analysis of tools it is required to two processes 1) Functionality based tool Selection and 2) Detailed Tool Analysis. These two processes are considered to the method for bioinformatics tools analysis that has been narrated in this section. Fig (1) in this section gives a Framework design of the activities that should be involved for tool analysis.

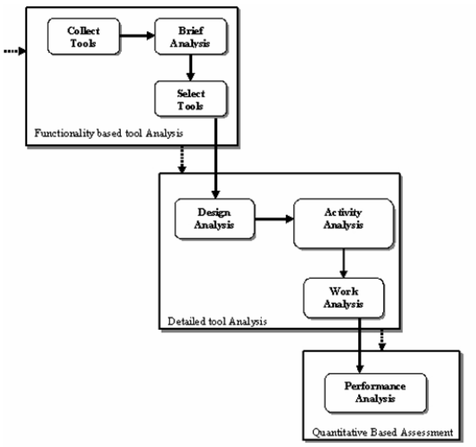

Fig 1- Framework for Bioinformatics Tool Analysis

A. Functionality based tool Analysis

In this functionally based tool analysis is carried out on various tools with functional specific tools as an overview rather than detailed view. Based on the brief analysis carried out users specific functionality with other user required specific tools could be selected on which the next detailed tool analysis process has to be carried out.

B. Detailed tool analysis

In this detailed tools analysis section of the selected tools each tool is selected for carrying out a detailed analysis in which various things should be considered which has been narrated in the following steps that are carried out. This is used for further carrying out quantitative based assessment for identifying better tools from the set of tools selected based on the comparison made performance based analysis.

Design analysis

In tool analysis various thing that should be considered which should satisfy some basic requirements of the tools since mainly the user of the tools are biologist so the tools that are to be developed with at most features and with required functionalities. Where design analysis i.e., user interface plays a vital role along with some of the features and functionalities should be considered that have been represented below. The primary basics that should be made an analysis first are

Whether the tool is of GUI based? (GUI Based)

Does the tool required internet facility for data analysis? (Internet Based)

Does any type of browser supports? (Browser Compactability)
Whether the tool works on any platform? (Platform Independent)
Does any third party support required for effective working of the developed tool? (Support Required)
Does help has been provided? For the tools for novice to know the functionality (Help)
Along with this analysis GUI should have a detailed analysis of navigation of the page including input interface and the result exhibition.

Activities analysis

Bioinformatics tools workflow that involves pattern and mining is becoming more and more complex now a days, involving numerous interacting business objects within considerable processes. Analyzing the tools interaction structure and process of those complexes will enable them to be well understood, redesigned and assess. Analyzing the statistical techniques or other techniques, methods, approached, causal dependencies that have been used in the development of the tool which involves for a particular or set of activities involves in this step.

Work analysis

Individual activities for an event or set of activities are grouped together to perform an overall analysis through practical means of giving the required inputs and flowing through the interior process and workflow activities and extract the appropriate result to certain extent. This work analysis could be made roughly which could be helpful for identifying metrics that are applicable for the specific tool and also significant in the tool assessment.

Tool assessment on quantitative basis

Metrics quantifies some of the characteristics or attributes of a product or process by means of measurement using numerical ratings. Metrics based assessment is made on developed products because it qualifies the characteristics of a product like specificity and precision. Since quantitative means have been proven very powerful in computer science and other sciences, Statistical analysis is also being made to identifying the qualifying levels. Evaluating tools by means of metrics is the one of the major standards used for tools performance, retrieval ability, accuracy, etc. analysis that has been explained in this section for identifying the performance of the tools. There are various other metrics also that have been analyzed from the study carried out on various bioinformatics tools [13-17].

Various metrics that often derived in order to measure performance against a critical success factor; each metric has multiple definitions and ambiguous counting rule. The significance for evaluating metrics for validating tools is, based on this evaluation only tools could be validated through which pitfalls could be identified or assessment for the existing tools could be made which leads to a further improvement of the existing tools or in the development of the new tools by overcoming the pitfalls of the existing ones. Based on the identified metrics two modes of analysis could be followed for carrying out tool evaluation 1) Input, Output based evaluation and 2) Interior processing based evaluation. Here in this paper we mainly concentrated on the first input and output based evaluation that are mostly applicable for performance analysis of pattern based and text mining tools. Apart from this, in this section various modes of performance analysis by quantitative means i.e., by means of metrics have been narrated.

A. Input, Output based evaluation

This evaluation is of direct based where based on the input and the displayed result metrics evaluation is done like motif match, identity metrics etc… Here there is no need of depth analysis and metrics evaluation for the interior process carried in tool for providing the apt result for the given input. Evaluation on each tool should be carried out by giving different input and evaluating the metrics based on the displayed results. Likewise all the selected tools should be evaluated, Finally analysis should be made for utilizing the best tool for effective and efficient usage based on the user specified criteria's and quantitative analysis made on the tools by means of comparison chart. This standard evaluation is found to be most effective and effective standards for evaluating tools quality that have been analyzed.

B. Various quantitative based performance analyses

Quantitative based quality assessment that is to be carried out on the tools based on metrics for evaluating the tools performance could be carried out on various modes of evaluation on pattern based and text mining tools like retrieval ability, accuracy, etc… This evaluation is carried out on each user selected tool finally which undergoes a comparison of all selected tools for better performance and best tool utilization which has been narrated in this section. Tools performance are normally considered to be good, better and low or less which have also been narrated for overall tools performance analysis have also been represented in this section.

Performance control metrics

Precision and recall [15] are the two major metrics that are used in performance control over both, pattern based and text mining tools rather than other tools. Precision is the performance metrics for relevant retrieval of documents for text mining tools where as the recognition of the precised string in a gene or protein or measure the performance of the extraction modules through syntactic analyzes and pattern based. Recall is another performance metrics measures related documents of the retrieves relevant document for text mining tools where as recall for pattern based used to measure performance through syntactic analyzes and also used in classifying catalytic protein structures.

Precision and Recall should be equal or Precision should less than recall with little variation then tool found to have good performance. If recall found to be less than precision than the performance of the tool found to be less. The change of variation in the level of both precision and recall changes the performance level of the tool.

Retrieval ability metrics

Information Retrieval is another performance metrics that represents the system performance in the sense of Information Retrieval for which the correct documents that are been accurately retrieved this is used in text mining or web based tools for performance evaluation [1]. IR metrics should be always good since the high the retrieval ability the tools performance is found to be good if there is recall level is greater than or equal to information retrieval metrics. Since information retrieval metrics can be also considered to be as precision that's the reason behind for performance evaluation along with information retrieval metrics recall is also been considered. The lower the recall the retrieval ability of relevant documents is considered to be less then even if retrieval ability is high the performance of the tool is considered to be less.

Accuracy metrics

Accuracy represents another performance metrics that measures the performance of the system and the accurate analysis and result displayed based on the functionalities carried out on sequences for pattern based and sequence analysis tools [14]. This metrics is also used for classifying catalytic protein structure. Accuracy is same as that of precision the performance of the tool is found to be good if there is good or high accuracy. The performance of the tool is found to be low if it is less than 50%. Accuracy depends on the relevant or related retrieval of specific data's or result based on the given input.

Specificity and sensitivity metrics

Sensitivity and Specificity are another performance metrics mainly used in performance analysis of pattern based tools [16]. Specificity represents tools performance to that particular domain specific and Sensitivity represents tools performance by accurate data extraction. It is considered on the bases that specificity should always be greater than sensitivity and the difference in variation should be little or equal for good performance. If specificity is lesser than sensitivity, that have little difference in variation then the performance of the tool also considered to be better. The tools performance is also considered to be better if specificity is greater than sensitivity with more variations. The tools performance is considered to be low if there is more variation found between specificity and sensitivity.

Performance stability

F-Score or F-measure is another performance metrics used to measure the performance stability of the tools ie., the variation in the tools performance [13]. Evaluation on each tool should be made more the thrice for which variations could be accurately predicted in f-score. For the evaluation carried out on the tools variation in f-score should be identified for performance stability analysis. The more the variation the tools performance stability varies based on the input. If there is less variation than, mostly the performance is stable where the tools performance falls within the evaluated range.

4. Tools based effort analysis

In this section the tool profound to be of effective and efficient in terms of analysis and usage doesn't only fall is the major utilization of the tools. Apart from this tools quality analysis involves in efforts based on time, cost and resources which have been theoretically narrated in this section. Effort level varies for different tools where as, effort based analysis can be carried out on tools for analyzing the tools quality based on design and activities analysis in means of cost, resources and time which profound to also play a vital role for quality analysis.

A. Effort based on time

Effort based on time could be one of the tools quality analyses in terms of data access from the databases and result display. Apart from this time required for evaluating the tools performance.

B. Effort based on cost

Effort based on cost could be a base for tools quality analyses in usage of other resources like usage of third party tools etc. Apart from this cost could be involved in terms of human resources utilization cost for tools evaluation.

C. Effort based on resources

Effort based on resources for tools quality analyses on usage of other resources like usage of third party tools, working environment etc. Apart from this resources could be involved in terms of human resources utilization for tools evaluation.

5. Conclusion

In this paper our aim is to narrate a designed framework along with effort analysis for carrying out quality analysis on tools through means. This approach reduces the time for carrying out accurate processing steps for evaluating tools quality. Quality analysis is significant since, bioinformatics tools are increasingly important for solving real-time biological problems. Other main reason behind for narrating the framework and effort analysis for evaluation on the bioinformatics tools is for the quality implications that could be enhanced on the developed tools. The framework and effort analysis method mentioned in the article found to be effective and efficient in selecting the appropriate quality analysis process for better tools evaluation.

6. References

[1] Dean Cheng, Craig Knox, Nelson Young and Paul Stothard (2008), Oxford Journals, 399-405.

[2] Lydia Ng, Chris Lau, Rob Young, Sayan Pathak and Leonard Kuan (2007), BioMed, 8(2), 7-12.

[3] Hongmei Sun, Hong Fang, Tao Chen, Roger Perkins, and Weida Tong (2006), BioMed, 7(2), 1068-1076.

[4] Jochen Farwer, Martin J. Packer, Christopher A. Hunter (2007), BioMed, 12(4), 595-600.

[5] ob Jelier, Martijn Schuemie, Antoine Veldhoven, Lambert CJ Dorssers and Jan A Kors (2008), Oxford Journal, 6-9.

[6] Tsun-Chen Lina, Ru-Sheng Liua, Chien-Yu Chenc and Ya-Ting Chaoa (2006), Elsevier, 39(12), 2426-2438.

[7] Chung-Chih Lin, Yuh-Show Tsa, Yu-Shi Lin, Tai-Yu Chiu, Chia-Cheng Hsiung, May-I. Lee, Jeremy C. Simpson and Chun-Nan Hsu (2007), ACM Portal, 3374–3381.

[8] Bongshin Lee, Kristy Brown, Yetrib Hathout and Jinwook Seo (2008), ACM Portal, 1026–1028.

[9] Marta Sabou, Chris Wroe, Carole Goble and Gilad Mishne (2005), Citeseer, 190-198.

[10] Jung-jae Kim, Piotr Pezik and Dietrich Rebholz-Schuhmann (2008), ACM portal, 1410–1412.

[11] Yang Jin, Ryan Mark A Mandel,Steven Carroll, Mark Y Liberman, Fernando C Pereira, Raymond S Winters and Peter S White (2006), BMC Bioinformatics, 492-500.

[12] Mathiak B. and Eckstein S. (2004), Proceedings 15th European Conference.

[13] Xiang Xu, Jinyu Wu, Jian Xiao, Yi Tan, Qiyu Bao, Fangqing Zhao and Xiaokun Li (2008), ACM Portal, 1217–1220.

[14] Jayanthi Manicassamy and P. Dhavachelvan (2009), International Journal of Recent Trends in Engineering (IJRTE), 1 (1), 550-555.

[15] Jayanthi Manicassamy and P. Dhavachelvan (2009), ACM International Conference in Advances in Computing Communication & Control (ICAC3’09), 171-176.

[16] Jayanthi Manicassamy and P. Dhavachelvan (2009), International Journal of Computer and Electrical Engineering (IJCEE), 1 (3), 397-402.

[17] Scott Coull, Joel Branch, Boleslaw Szymanski and Eric Breimer (2003), IEEE Explore, 24- 33.

[18] Marti A. Hearst, Anna Divoli, Harendra Guturu, Alex Ksikes, and Jerry Ye (2007), ACM Portal, 2196–2197.