Quantitative Based Quality Evaluation for Pattern Based Bioinformatics Tools

Jayanthi Manicassamy and P. Dhavachelvan

Abstract – Today computational biology has merged with few other sciences for solving biological problems where there is a tremendous increment in innovations of new techniques and approaches that have been done day to day. In this field, area of research involves disease identification, sequence analysis, structural analysis, protein-protein interaction etc... in which pattern matching plays a major role. There are various tools that have been developed in pattern matching in the area of bioinformatics for various analysis processes. At present bioinformatics tools are increasingly important to allow real-time decision making for solving biological problems for which performance plays the key role for evaluating the developed tools. Performance of a tool can be evaluated by means both qualitative and quantitative which has been involved in this paper. Quantitative based evaluation represents the quality of the tools for which metric based evaluation like specificity, sensitivity, pattern match etc... are been carried out. Qualitative based evaluation represents features, functionality evaluation together with the performance quantitative assessment. The merit that we are discovering by evaluating the existing products, leads us in the development of new products by avoiding the pitfalls of the existing products. In this work, some of the pattern tools are been considered for evaluating the performance which involves standards evaluation procedure. The results observed during this evaluations made on the specified tools have made an indication for having a standard development cum evaluation of pattern based bioinformatics tools. These evaluations could provide a way for developing and maintaining better systems with enhanced qualities.

Key words — Pattern Matching, Bioinformatics, Performance Evaluation, Software Metrics, Quality Assessment.

I. INTRODUCTION

-Bioinformatics is a field of biology merging with few other sciences like statistics and information technology which involves in the discovery of new tools for data analysis and interpretation of accurate result. Computational biology plays a vital role in the various areas of biology which involves pattern matching as a vital role in bioinformatics. Discrimination of cancer [2] evaluation from gene expression, protein sub-cellular localization from experimental data [3] by extracting features from raw images with possibility of applying genetic interactions to predict pathways have been evolved in this field.

Various other researches are been undergoing on silico detection with conserved patterns of signification mutations

in evolution. Numerous number of visualization tools have

developed for 3-D structures [5] for DNA sequence analysis, protein-protein interaction [6] in cellular activities by integrating different types of life science data's that are been realized. Today high throughput techniques and approaches allow biologists to gather data at an unprecedented rate. However, involvement of pattern functionality is one of the techniques makes inferences from data's at sufficient rate to meet the demands in the area of bioinformatics.

-Metrics is a quantitative measure that qualifies the characteristics of a product that are being observed directly or indirectly expressed in numerical data's. Features and functionality identification and evaluating tools is a qualitative measure together with quantitative measures is the standard procedure involved here in evaluating the tools. A selective of some different pattern tools in bioinformatics is taken for which a detailed analysis have been carried out. Identifying metrics, features and functionalities for each tool that are applicable from a list has been analyzed for pattern tools and evaluated which have been represented in tables. A pictorial representation has also been made for essential features and functionalities that are to be taken into consideration for evaluation.

II. PATTERN BASED BIOINFORMATICS TOOLS, ITS FUNCTIONALITIES AND FEATURES

Several researches are undergoing in the area of bioinformatics involving pattern matching which made way for the development of many algorithms, approaches and tools that have been proposed and developed. A detailed analysis has been carried out on some pattern tools of which selective few tools has been taken for representing the ways to carry out evaluations on pattern based tools in this section. Here we have also represented the generalized qualitative features and functionalities that subsist in bioinformatics tools. Based on this qualitative list evaluation of the taken tools has been carried out as the first standard tool evaluation process on pattern based tools.

The main reason behind for this evaluation process as the first evaluation standard process is that the tools that are been developed or to be developed should satisfy some basic requirements like GUI, user friendly, help facilities etc.... Since, the tools users are mainly biologist so the basic tools functionalities along with other basic functionalities should also be satisfied. Table 1 represented some of the basic qualitative and functionality list of the tools based on the detailed analysis carried out on the tools. We have classified the functionality of tools on three bases like fine grain, coarse grain and scope based. Approaches involved in tools

are classified as fine grain, tools category has been classified as coarse grain and other tools major functionality and features as scope based which has been represented in Figure 1.

A. SubSequer Tool

David He and John Parkinson [7] developed a new web-tool which uses graphical visualization methods borrowed from protein interaction studies to identify and characterize repetitive elements in low-complexity sequences where repetitive protein sequences with a limited amino acid palette are abundant in nature and many of them play an important role in the structure and function of certain types of proteins. However, such repetitive sequences often do not have rigidly defined motifs.

B. CompariMotif Tool

Richard J. Edwards et. al. [4] developed tool for making motif-motif comparisons, identifying and describing

similarities between regular expression motifs. It can identify a number of different relationships between motifs, including exact matches, variants of degenerate motifs and complex overlapping motifs. Motif relationships are scored using shared information content, allowing the best matches to be easily identified in large comparisons. Many input and search options are available, enabling a list of motifs to be compared to within itself to identify recurring motifs or to datasets of known motifs.

C. Whatizit Tool

Dietrich et. al. [10] has developed a tool for researchers in the biomedical research community for literature search that analyze text for contained information and then link them to the corresponding entries in bioinformatics databases such as Swiss-Prot, PubMed data entries and gene ontology concepts.

Table1: Pattern Based Bioinformatics Tools Qualitative and Functionalities Lists

Features & Functionalities	Bioinformatics Tool			
	SubSequer	CompariMotif	Whatizit	Swelfe
	T ₁	T ₂	T ₃	T ₄
License Free	√	√	√	√
Platform Independent	√		√	√
Open source		√		
GUI	√	√	√	√
User Friendly	√	√	√	√
Database usage	√	√	√	√
Query usage			√	√
Automatic Detection		√	√	√
Support Required				
Help	√	√	√	√
Navigate able	√	√	√	√
Graph Theory	√			
Any Browser Type	√	√	√	√

Note: where, shaded cell represents non existence of that particular features or functionalities

D. Swelfe Tool

Anne Laure et. al. [9] has developed tool which associates with web server to find internal repeats at three levels which quickly identifies statistically significant internal repeats in DNA and amino acid sequences and in 3D structures using dynamic programming. The associated web server also shows facilitates visualization of the results.

The work proposed in this paper for evaluating the performance of pattern based tool is a standard procedure for evaluating both qualitative and quantitative analysis for which assessment schemes has been narrated in details in section III.

Figure 1: Pattern Based Tools Major Functionality Representation

III. PATTERN BASED BIOINFORMATICS TOOLS METRICS EVALUATION

Evaluating tools by means of metrics is the second standard process adopted here for evaluating performance of pattern based bioinformatics tool in this paper. Here we have also represented about the pattern based tools metrics which are been considered for quantitative analysis. Table 2 represented the list of metrics that are used for tool evaluation here since metrics is the quantitative measure that qualifies a system together with qualitative measures.

The work proposed in this paper for evaluating the performance of pattern based tool is a standard procedure for evaluating both qualitative and quantitative analysis for which state definition and evaluation schemes have been represented in this section as followed.

A. State Definition

Let ' T_i ' be the set of tools taken for which features, characteristics and metrics are identified. Set $T = \{T_1, T_2, T_3 \text{ and } T_4\}$ and with respect to metrics and features tools can be expressed as

Table 2: Metrics for Pattern Based Bioinformatics Tools

Metrics	Formula	Description
Specificity	$Sp = \frac{TN}{(TN+FP)}$ (1)	where, TN – number of True Negative, FP – number of False Positive. The calculation of TN and FP depends on DNA sequence, protein sequence or usage of mutation. Specificity [11] in general is a standard evaluation metrics that measures the proportion of negative patterns correctly recognized as being negative.
Sensitivity	$Sensitivity (Sn) = \frac{TP}{(TP + FN)}$ (2)	Sensitivity (Sn) metrics [11] measures the proportion of positive patterns correctly recognized genes in a sequence of DNA in general or in a string of DNA. where, TP – True Positive and FN – False Negative.
E-Value	E-value = Expected number of alignments between the text and the pattern that have the same or even better similarity score. (OR) E-value = the number of hits one can expect to get from a search in a database. (3)	Expected value [12] is a statistical measure.
Match position	Match Position = Total number of match positions found in the whole sequence to which it is compared with. (4)	This Match Position [7] metrics used to find the match position in the sequence.
Motif Score	Motif Score = IC * Number of matched positions (5)	where, IC is Information Content Motif score [4] of a motif in a sequence is the maximum match score in over all positions in the sequence.
Pattern Match	Pattern Match = Total number of matches / Number of matches correctly predicted (6)	Pattern match metrics [11] is the universal standard metrics that are applicable for all applications, tools or techniques that are involved in pattern matching
IC _m	IC _m = sum of the component IC _i values in motif comparison. (7)	where, IC _i is the Information Content at position i. IC _m [4] is simply the sum of the component IC _i values which is used in motif comparison of protein sequences.
Length of repeats	Length of the repeats = Total size of the repeat (8)	Length of the repeats [7] is repeat that are been considered for pattern matching in motif comparison.
Density of alignment	Density of Alignment = Alignment Score / Alignment Length (9)	Density of Alignment [11] is used most probably done while going for pattern match
IC match Rank	IC _m match rank = total number of matching positions / the number of exactly matching fixed positions in motifs. (10)	IC _m matches are ranked [4] by the total number of matching positions and then by the number of exactly matching fixed positions which are used in motifs.

$$T_i = T_{M_j} = T_{f_j} = \begin{cases} T_1 = \{(f_{1,1}, f_{1,2}, \dots, f_{1,n}), (M_{1,1}, M_{1,2}, \dots, M_{1,n})\} \\ T_2 = \{(f_{1,1}, f_{1,2}, \dots, f_{1,n}), (M_{1,1}, M_{1,2}, \dots, M_{1,n})\} \\ T_3 = \{(f_{1,1}, f_{1,2}, \dots, f_{1,n}), (M_{1,1}, M_{1,2}, \dots, M_{1,n})\} \\ T_4 = \{(f_{1,1}, f_{1,2}, \dots, f_{1,n}), (M_{1,1}, M_{1,2}, \dots, M_{1,n})\} \end{cases} \quad (11)$$

where, M_{ij} refers to metrics in which i denotes column (tools) and j denotes row (metrics), which in general refers to the existence of that particular metrics, here n represent total metrics that are applicable for that particular tool and f_{ij} represents features, functionalities of the tools.

B. Evaluation Schemes

In this paper we have worked on few tools by carrying out four evaluations on each tool for which we have given different inputs for each evaluation made on each tool. For the evaluation carried out we have represented in Table 3 evaluation scheme 1 as ev1, evaluation scheme 2 as ev2,

evaluation scheme 3 as ev3, evaluation scheme 4 as ev4 and evaluation scheme 5 as ev5. Here the evaluation scheme is mainly for evaluating the performance of the tool by means of quantitative analysis which has been narrated in section IV along with result analysis.

In the evaluation scheme for performance analysis we mainly concentrate on the metrics sensitivity and specificity for analyzing the tools performance rather than other performance metrics. Specificity represents tools performance to that particular domain specific and Sensitivity represents tools performance by accurate data extraction. It is considered on this base 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 is 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 if sensitivity is greater than specificity.

IV. TOOLS EVALUATION AND RESULT ANALYSIS

In this section we have carried out evaluations on each taken tools for evaluating the performance of the tools by Quantitative means. Four evaluations have been carried out on each tools by giving different inputs based on the analyzed result tools performance is evaluated. Tools evaluation values for each evaluation carried out on each tool have been represented in Table 3. Overall summarized of tools quality has been represented here including both qualitative analysis and quantitative analysis. Qualitative analysis involves satisfying basic tools features, functionalities and Quantitative analysis represents performance evaluation results by means of metrics.

A. SubSequer Tool

SubSequer tool [7] uses graphical visualization to identify and characterize repetitive elements in low-complexity sequences where repetitive protein sequences with a limited amino acid palette.

Evaluation Scheme 1

The input given in fasta format for first evaluation input for cytochrome b of *Elephas maximus* is given for which default option of automatic selection of parameters for values for number of wildcard, subsequence size and Percentile Odds Score Cutoff is selected. The inputted sequences have been processed into overlapping subsequences using a fixed-size sliding window. These subsequences were collated and ranked according to how frequently they were observed. The frequency chart is displayed for subsequence size 3 with wildcard 1 of odd score cutoff 7.38. Number of subsequences found 15 with subsequence size 4 with wildcard 1 of odd score cutoff 8.49. Number of subsequences found 6 with subsequence size 4 with 2 wildcard of odd score cutoff 6.29. Number of

subsequences found 15 with subsequence size 5 with 1 wildcard of odd score cutoff 0.0 and number of subsequences found 3, subsequence size 5 with 2 wildcard with odd score cutoff 7.38 and number of subsequences found 6. From this a network link is made between nodes that has overlapping motif context and the number of overlapping nodes are found based on the selection of the nodes. For example subsequence size 3 with 1 wildcard found to have 2 nodes linked to each other found to have total motifs positions as 11 in one node and in other node 8 positions. This has specificity of 0.94. Here match is found at 11 positions which are matching positions, Information content match rank is calculated as the number of match positions found that is 11, repeat size is 3 which are considered for pattern matching and total match is found to be 4 as pattern match. Specificity and sensitivity found to be 0.94 and 0.93. This evaluation result is stored in table 3 of T1 in Ev1. Here the tools performance is found to be good only taking into this result evaluation into consideration.

Evaluation Scheme 2

For the second evaluation input of envelope protein has been given in fasta format with subsequence size 5 with 2 wildcard and 2 percentile odds score cutoff found to have 3 nodes linked in network and the motif total positions of each nodes are found to be 25 from the result displayed. Length of the repeats is found to be 5, pattern match is found to be 5 and specificity, sensitivity found to be 0.88 and 0.88. This evaluation result is stored in table 3 of T1 in Ev2. Here the tools performance is found to be good only taking into this result evaluation into consideration.

Evaluation Scheme 3

For the third evaluation carried out specificity and sensitivity found to be 0.92 and 0.91. Match positions and pattern match found to be 14 and 3. Found 2 be the where, Shaded cell represents that metrics is not applicable for the particular tool.

Table 3: Quantitative Evaluation of Pattern Based Tool

Metrics		Bioinformatics Tools															
		SubSequer				CompariMotif				Whatizit				Swelke			
		T ₁				T ₂				T ₃				T ₄			
		Ev ₁	Ev ₂	Ev ₃	Ev ₄	Ev ₁	Ev ₂	Ev ₃	Ev ₄	Ev ₁	Ev ₂	Ev ₃	Ev ₄	Ev ₁	Ev ₂	Ev ₃	Ev ₄
1.	Specificity	0.94	0.88	0.92	0.92	0.90	0.87	0.92	0.90	0.84	0.90	0.92	0.90	0.85	0.90	0.90	0.91
2.	Sensitivity	0.93	0.88	0.91	0.90	0.87	0.85	0.87	0.85	0.69	0.76	0.74	0.72	0.86	0.94	0.92	0.94
3.	E-Value					4	2	2.9	2								
4.	Match position	11	25	14	9	4	2	5	3					1	2	2	1
5.	Motif Score					4.0	1.52	3.5	2.28								
6.	Pattern Match	4	5	3	2	1	2	1	3	346	384	372	356	1	1	2	1
7.	ICm					4.0	2.0	3.0	2.5					1	2	2	1
8.	Length of repeats													65	93	87	62

9.	Density of alignment	3	5	2	3								12.7	0.64	1.79	11.12
10.	IC match Rank					4.0	2.0	4.0	1.0				1	2	1	2

where,

Shaded cell represents that metrics is not applicable for the particular tool.

density of alignment. This evaluation result is stored in table 3 of T_1 in Ev_3 . Here the tools performance is found to be good only taking into this result evaluation into consideration.

Evaluation Scheme 4

For the third evaluation carried out specificity and sensitivity found to be 0.92 and 0.90. Match positions and pattern match found to be 9 and 2. Found 3 be the density of alignment. This evaluation result is stored in table 3 of T_1 in Ev_4 . Here the tools performance is found to be good only taking into this result evaluation into consideration.

Summary of Result Assessment

Based on the evaluation carried out the SubSequer tool, it satisfies the basic qualitative requirements (that involves satisfaction of basic features and functionalities). Based on the quantitative requirement the tools performance found to be good (metrics based evaluation). Taking into consideration of both qualitative and quantitative means the quality of the tool found to be good.

B. CompariMotif Tool

CompariMotif tool [4] is used for making motif-motif comparisons, identifying and describing similarities between regular expression motifs.

Evaluation Scheme 1

For the inputted motif list, we separated each motif by comma R.S.P.L, GR.[ST].P, ^.[AS][AGS], KE..K,P.P.P, AGS].P.P.P.^.[AGS][GS],FR..[ST].S, [ST].P.[ST]P,Y.C.PG.L and R..S.P.L is given to be searchDB by comparing with the given list, normalized IC cut-off is given as 0.5 and with a searchdb as motif option R.S.P.L will be given to compare with the SLiMDisc dataset name. With the evaluation made 1 pattern match of R.S.P.L is found which consist of matching position of 4. Here R, S, P, L position value of each one is 1.0 so total value of R.S.P.L is 4.0 which are to be compared with the list of motif found to have an exact pattern match of R.S.P.L having positional score 4.0 from which we calculated normalized IC from information content match value 4.0 (R.S.P.L value) divided by lower IC value of the match found from the motif list contains only one match of value R.S.P.L. Normalized IC found to be 1.000, motif score is 4.0 which is calculated by multiplying match position score that is 4.0 with normalized IC. Expected value is 4 calculated by number of alignment made in the text with the pattern to find a match. Specificity is found by calculating true negative repeats of protein motif divided by true negative repeats together with false positive repeats which is found to have specificity of 0.90 and Sensitivity is found to be 0.87. This evaluation result is stored in table 3 of T_2 in Ev_1 . Here

the tools performance is found to be good only taking into this result evaluation into consideration.

Evaluation Scheme 2

For the second evaluation made given the same input of the previous evaluation for motif list which is searched with EML motif Database with the option of both input and searchDB of normalized IC cut-off 1.0. When search is made with the database two matches are found for [ST].P with that of [ST].P.[ST]P and P.[ST]P of Match IC is 2.000, 2.000 of match position 2 for both match, normalized IC 0.760 and 0.722, Score of 1.519 and 1.445. Evaluation procedure is same as that evaluation 1 and here specificity and sensitivity found to be 0.87 and 0.85. This evaluation result is stored in table 3 of T_2 in Ev_2 . Here the tools performance is found to be good only taking into this result evaluation into consideration.

Evaluation Scheme 3

For the third evaluation carried out on this tool specificity and sensitivity found to be 0.92 and 0.87. E-value is 2.9, match positions found to be 5, motif score is 3.5 and pattern match found to be 1. IC match found to be 3.0 and IC match rank found to be 4.0. This evaluation result is stored in table 3 of T_2 in Ev_3 . Here the tools performance is found to be good only taking into this result evaluation into consideration.

Evaluation Scheme 4

For the fourth evaluation carried out on this tool specificity and sensitivity found to be 0.90 and 0.85. E-value is 2, match positions found to be 3, motif score is 2.28 and pattern match found to be 3. IC match found to be 2.5 and IC match rank found to be 1.0. This evaluation result is stored in table 3 of T_2 in Ev_4 . Here the tools performance is found to be good only taking into this result evaluation into consideration.

Summary of Result Assessment

Based on the evaluation carried out the CompariMotif tool, it satisfies the basic qualitative requirements (that involves satisfaction of basic features and functionalities). Based on the quantitative requirement the tools performance found to be good (metrics based evaluation). Taking into consideration of both qualitative and quantitative means the quality of the tool found to be good.

C. Whatizit Tool

Whatizit tool [10] used for researchers in the biomedical research community for literature search that retrieves corresponding entries in bioinformatics databases. Since the tool is literature search tool we consider here specificity as recall and sensitivity as precision. This consideration is done

due to that specificity could be also be said as recall and sensitivity could be considered as precision.

Evaluation Scheme 1

For the first evaluation input text “Cancer” to be searched with PMIDs and pipeline of WhatizitDisease which has retrieved 500 abstract correctly extract. 346 while with relativeness it is found to be 367. Specificity found to be 0.84 which is calculated as correct extracted 346 divided by all correct with relativeness 367. Sensitivity as 0.69 which is calculated as correct extract 346 divided by all extracted 500. This evaluation result is stored in table 3 of T₃ in Ev₁. Here the tools performance is found to be better only taking into this result evaluation into consideration.

Evaluation Scheme 2

For the second evaluation input text “lung cancer” search is made for which 500 documents are filtered and retrieved with match of lung cancer from all the bioinformatics databases. For this 384 found to be exact match while 423 found to be exact with relativeness, here we found to have specificity as 0.90 and sensitivity as 0.76. This evaluation result is stored in table 3 of T₃ in Ev₂. Here the tools performance is found to be better only taking into this result evaluation into consideration.

Evaluation Scheme 3

For the third evaluation carried out specificity and sensitivity found to be 0.92 and 0.74. Found match of 372 of the 500 extracted documents for which 342 found exact extraction made. This evaluation result is stored in table 3 of T₃ in Ev₃. Here the tools performance is found to be better only taking into this result evaluation into consideration.

Evaluation Scheme 4

For this evaluation specificity found to be 0.90, sensitivity found to be 0.72 and pattern match found to be 356. This evaluation result is stored in table 3 of T₃ in Ev₄. Here the tools performance is found to be better only taking into this result evaluation into consideration.

Summary of Result Assessment

Based on the evaluation carried out the Whatizit tool, it satisfies the basic qualitative requirements (that involves satisfaction of basic features and functionalities). Based on the quantitative requirement the tools performance found to be good (metrics based evaluation). Taking into consideration of both qualitative and quantitative means the quality of the tool found to be better.

D. Swelfe Tool

Swelfe tool [9] used to find internal repeats at three levels in DNA and amino acid sequences and in 3D structures using dynamic programming.

Evaluation Scheme 1

For the first evaluation PDB id 1b7fA given as input for 3 level type structural input and from the displayed results we found structure match 1 found. Alignment length is 65 where in a sequence 65 of length of subsequence will be

taken and alignment will be made with that of the pattern should be searched, score of 828.0. Density of alignment is 12.7 which is calculated as alignment score that 828 divided by alignment length 65. Match positions of the repeat 1 match position starts at 124 ends with 189 and that of repeat 2 match position starts from 210 ends with 275. For amino acids sequence match found to be 1 of alignment length 66, repeat 1 position are from 2 to 68, repeat 2 positions are from 88 to 275, score of 112.00. For DNA sequence of match 1 with repeat 1 match positions from 5 to 72, repeat 2 match positions from 263 to 330, length of alignment is 67 with score 19.50 specificity found to be 0.85 and sensitivity found to be 0.86. Pattern match of repeat length 65 is found to be in 1 position. The comparison is made with Drosophila melanogaster from EMBL database for our result has been displayed. This evaluation result is stored in table 3 of T₄ in Ev₁. Here the tools performance is found to be better only taking into this result evaluation into consideration.

Evaluation Scheme 2

For the second evaluation input of amino acid given in fasta format of Serum albumin precursor for organism Bos taurus (Bovine) which is a protein entry made to which 2 matches are found. For this match 1 position is from 16-197 for repeat 1 and 399-590 for repeat 2 with length of alignment 191 of score 91.0. Match 2 repeat 1 positions from 87-175, repeat 2 positions form 356-444, length of alignment is 93.0, score 60.0. This found to have specificity and sensitivity as 0.90 and 0.94. This evaluation result is stored in table 3 of T₄ in Ev₂. Here the tools performance is found to be better only taking into this result evaluation into consideration.

Evaluation Scheme 3

For the third evaluation carried out specificity and sensitivity found to be 0.90 and 0.92. Here match positions and pattern match found to be 2. Length of repeats and density of alignment found to be 87 and 1.79. This evaluation result is stored in table 3 of T₄ in Ev₃. Here the tools performance is found to be better only taking into this result evaluation into consideration.

Evaluation Scheme 4

For the third evaluation carried out specificity and sensitivity found to be 0.91 and 0.94. Here match positions and pattern match found to be 1. Length of repeats and density of alignment found to be 62 and 11.12. This evaluation result is stored in table 3 of T₄ in Ev₄. Here the tools performance is found to be better only taking into this result evaluation into consideration.

Summary of Result Assessment

Based on the evaluation carried out the Swelfe tool, it satisfies the basic qualitative requirements (that involves satisfaction of basic features and functionalities). Based on the quantitative requirement the tools performance found to be good (metrics based evaluation). Taking into consideration of both qualitative and quantitative means the quality of the tool found to be better.

V. CONCLUSION

The aim behind this paper is to provide a standard method

for quality assessment of pattern based tools in bioinformatics. The evaluation of the tools assessment has been carried out on this bases that involves both qualitative analysis of features, functionalities and quantitative analysis of performance metrics. The features that have been represented here could be kept as a base for evaluating tools basic features for all bioinformatics tools. In addition to this basic satisfactory of the tools functionality should be manually considered along with tools features in standard evaluation. In the standard evaluation next follows metrics evaluation which makes an appropriate judgment of tools performance apart from qualitative analysis. The tool evaluation and observation of the result in depth provide an idea for quality assessment pattern based tools. This also provides an idea for the requirement of new tool development with high throughput and better decision making or in the enhancement of the existing tool.

REFERENCES

- [1] ob Jelier, Martijn J Schuemie, Antoine Veldhoven, Lambert CJ Dorssers, Guido Jenster and Jan A Kors, "Anni 2.0: a multipurpose text-mining tool for the life sciences", Oxford Journal, pp 9-6, June 2008.
- [2] Tsun-Chen Lina, Ru-Sheng Liua, Chien-Yu Chenc, Ya-Ting Chaoa, Shu-Yuan Chena, "Pattern classification in DNA microarray data of multiple tumor types", Elsevier, January 2006.
- [3] 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, "Boosting multiclass learning with repeating codes and weak detectors for protein subcellular localization", ACM Portal, September 2007, pp 3374–3381.
- [4] Richard J. Edwards, Norman E. Davey and Denis C. Shield, "CompariMotif: quick and easy comparisons of sequence motifs", ACM Portal, March 2008, pp 1307–1309.
- [5] Joan Herisson, Guillaume Payen and Rachid Gherbi, "A 3D pattern matching algorithm for DNA sequences", Oxford Journal, January 2007, pp 680-686
- [6] Harold J Drabkin, Christopher Hollenbeck, David P Hill and Judith A Blake, "Ontological visualization of protein-protein interactions", BCM, February 2005, pp 6-29.
- [7] David He and John Parkinson, "SubSequer: a graph-based approach for the detection and identification of repetitive elements in low-complexity sequences", ACM Portal, pp 1016–1017, February 2008.
- [8] Timothy L. Baile and William Stafford Noble, "Searching for statistically significant regulatory modules", ACM Portal, pp 16–25, June 2003.
- [9] Anne-Laure Abraham, Eduardo P. C. Rocha and Joel Pothier, "Swelfe: a detector of internal repeats in sequences and structures", ACM Portal, pp 1536–1537, May 2008.
- [10] Dietrich Rebholz-Schuhmann, Miguel Arregui, Sylvain Gaudan, Harald Kirsch and Antonio Jimeno, "Text processing through Web services: calling Whatizit", ACM Portal, pp 296–298, November 2007.
- [11] Scott Coull, Joel Branch, Boleslaw Szymanski and Eric Breimer, "Intrusion Detection: A Bioinformatics Approach", IEEE, pp 24- 33, December 2003.
- [12] Yong Huang, Lingdi Ping, Xuezheng Pan, Guoyong Cai, "A Fast Exact Pattern Matching Algorithm for Biological Sequences", IEEE digital library, pp 8-12., 2008.
- [13] Yang Zhang and Jeffrey Skolnick, "Scoring Function for Automated Assessment of Protein Structure Template Quality ", ACM Portal, pp 702–710, 2004.
- [14] Mathiak B., Eckstein S., "Five steps to text mining in biomedical literature", Proceedings 15th European Conference, 2004.
- [15] Carolina Garrido, Vanessa Roulet, Natalia Chueca, Eva Poveda, Antonio Aguilera, Katharina Skrabal, Natalia Zahonero and Carmen de Mendoza, "Evaluation of Eight Different Bioinformatics Tools To Predict Viral Tropism in Different Human Immunodeficiency Virus Type 1 Subtypes", Journal Of Clinical Microbiology, March 2008, pp 887–891.
- [16] Ian M. Overton, Gianandrea Padovani, Mark A. Girolami and Geoffrey J. Barton, "ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction", ACM Portal, February 2008, pp 901–907.
- [17] Hongmei Sun, Hong Fang, Tao Chen, Roger Perkins, and Weida Tong, "GOFFA: Gene Ontology For Functional Analysis – A FDA Gene Ontology Tool for Analysis of Genomic and Proteomic Data", BioMed, March 2006.
- [18] Lydia Ng, Chris Lau, Rob Young, Sayan Pathak, Leonard Kuan, Andrew Sodt, Madhavi Sutram, Chang-Kyu Lee, Chinh Dang and Michael Hawrylycz, "NeuroBlast: a 3D spatial homology search tool for gene expression", BioMed, July 2007.