International Journal of Computer Theory and Engineering, Vol. 1, No. 3, August, 2009
1793-8201

Multi-Dimensional Visualization of Bioinformatics Tools Classification and Its Transformation – For Quality Assessment

Jayanthi Manicassamy and P. Dhavachelvan

***Abstract*—Bioinformatics aims at computing biology at molecular level where day to day discovery of enormous tools and techniques have been developed. These tools and techniques are involved in solving biological problem which involves data analysis and interpretation of accurate results. Some of the diverse analysis processes of theses developed tools are sequence analysis, functional analysis, structural analysis and similarity and homology analysis. In tools development various factors that are to be taken into consideration where quality plays a vital role from the view point of end users. Quality assessment of the tools is necessary before involving the tool utilization for major purpose. Assessment of tools quality has been considered based on the tools usage considering its features and functionalities. Tools evaluation by means of its features and functionalities is one of the major standards adopted. Visualized representation is more effective than any other modes of representation is profound one which have been adopted in this paper. A visualized model for tools classification and categorization based on its features and functionalities in multi-dimensional form have been symbolized. Here modes of tools classifications and categorization using its features and functionalities multi-dimensionally have been narrated. The adopted standard model will play vital role for pre-primary quality assessment for a large set of tools. Quality assessment is made on ten tools and results have been analyzed. The model adopted here is a standard which could be adopted for any tools in any area.**

***Keywords*-Bioinformatics, Modeling, Multi-Dimension, Quality Assessment, Standards, Tools Evaluation.**

I. INTRODUCTION

Solving biological problems by analysis and computation methods involving statistics, chemistry etc… Many tools have been developed which has been categorized [6] based on the functionalities are Homology and Similarity Tools like 3D spatial homology search [3] for finding genes with a similar spatial expression pattern from gene expression which could potentially reveal novel or unknown genes involved in similar processes or pathways. Protein Function Analysis tool uses Gene Ontology [5] for functional analysis that allows users to interactively select GO terms, according the hierarchical structure with specific biological complexity will be represented. Structural Analysis tool used [4] for DNA structure prediction which allows the calculation of atomic structures of double-helical. Sequence Analysis tools includes database searches like text based search, sequence based search, motif based search, structure based search and various tools have also been developed. Since these tools and techniques in the area of bioinformatics are involved in solving real time biological problem with better decision making. Most of the bioinformatics related applications and tools utilize databases NCBI, PDB, EMBDL, Medline etc… which has been developed to share its resources. There is enormous development of tool for specific performance so it is necessary to carry out for a large set of tools to reduce to make selection of set of better tools based quality assessed. Since these tools are involved in solving real world biological problems, it is necessary carry out quality assessment before utilizing or involving the tool for major purpose.

In this paper a visualized standard model for classifying and categorizing the tools features and functionalities have been represented based on various analysis and study made on bioinformatics tools. Apart from this the model mainly involves the representing the set of tools classifications multi-dimensionally which could be transformed. Primarily the multi-dimensional representation is done in 3-D as a starting point and slowly incremental transformation is made towards n-dimension. This model has been carried out as a pre-primary evaluation on large set of similar functionality tools in order to reduce the number of tools for best usage. In section II we represent the ways for classifying and categorizing based on the detailed study made on tools. In section III sample set of ten bioinformatics tools have been taken for representing classification and categorization of the tools. Representation of visualized multi-dimension of tool classification has been represented and when transformation takes place is explained in section IV. Quality assessment schemes and the result have been narrated in section V.

II BIOINFORMATICS TOOLS CLASSIFICATION

As technology develop day to day innovation of new tool and techniques are also been developing in every area which is not an exception in the area of bioinformatics alone. Various tools that have been developed which could be categorized or classified on various modes, basically based on its acquired features and functionalities. This could be made to a user specific categorized classification which could be a standard and that could be applicable for any tools.

Jayanthi Manicassamy and P. Dhavachelvan are from Department of Computer Science, Pondicherry University, Kalapet, Pondicherry, India.

Here detailed explanation of how a tool could be classified in
a generalized way based on its features and functionalities that could be made applicable for the entire set of existing tools have been narrated.

Bioinformatics tools primarily classifications are made on various modes generally based on its features and functionalities. Based on the classification made it is categorization is made where relatively same features or functionality falls. Likewise various categories with various classifications could be represented. Let Ti be the set of tools, Cj is the tools categorization and CLk is the tools classification. Here i the total number of tools, k is the total number of classification and j is the total number of classified tools category. Set Cj = {C1, C2, C3…..} in consideration with tools and categorization this could be expressed as

$$T_i = T_{C_j} = \begin{Bmatrix} T_1\{C_1, C_2.....\} \\ T_2\{C_1, C_2.....\} \\ . \\ . \\ T_{10}\{C_1, C_2.....\} \end{Bmatrix} \quad (1)$$

where, T_i = {T_1, T_2……T_{10}}and C_j = { C_1(CL_1, CL_2……..CL_k), C_2(CL_1, CL_2……..CL_k), C_3(CL_1, CL_2……..CL_k)…..…..} in which C_1(CL_1, CL_2……..CL_k) ≠ C_2 {CL_1, CL_2……..CL_k}≠ C_3{CL_1, CL_2……..CL_k}……..≠ C_j{CL_1, CL_2……..CL_k}.

Basically tools are classified based on its functionalities followed by features, based on which tools are being categorized. The following below are representation of some of the classification made from some of the tools likewise any number of classifications could be made using any number of tools.

CL_1: Homology and Similarity Tool
This set of tools can be worn out to recognize similarities between sequences of unknown structure, function and database sequences whose structure and function has to be explicated.

CL_2: Protein Functional Analysis Tool
These types of tools are used to compare protein sequence to the secondary or derived protein databases that contain information on motifs, signatures and protein domains.

CL_3: Structural Analysis Tool
These tools allow structural comparison with known structure databases. The function of a protein is more consequence in structure rather than its sequence with structural homology which tends to share functions.

CL_4: Sequence Analysis Tool
This set of tools allows carrying out more detailed analysis of a sequence including evolutionary analysis, identification of mutations, etc... The identification of other biological properties is a trace, which aid the search to elucidate the specific function of the sequence.

CL_5: Algorithm Based Tools
An algorithm is a series of procedure that provides solution to a particular problem. Using algorithms in tool development for performing some specific set of tasks are represented as algorithm based tools.

CL_6: Computational Based Tools
These set are tools are developed to provide solution to problem which uses computational base alone or it uses any one of the other two approaches along with computational approach.

CL_7: Knowledge Based Tools
The computational biological tools are mostly us to provide solution to real world biological problems there is a need of acquire knowledge, based on the datasets extracted from the environment and provide solutions accordingly. Knowledge based tools uses knowledge based approach for providing solution to the problems.

CL_8: Diseases Based Tools
These are the type of tools used for extracting information's about disease like evolution analysis, analysis a sequence of existence of a disease, searching disease related information's etc... The type of tool developed could be for any species.

CL_9: Drug Based Tools
The set of tools that has the functionality of providing information using drug as the base to solve a particular problem like exacting information about drugs, activities of the drugs in the biological system etc…

CL_{10}: Virus Based Tools
Working of the tools on virus like life cycle of the virus with time series analysis, evolutional analysis, viruses based information extraction from the database etc…

CL_{11}: Datasource Based Tools
As biology is tremendously increasingly which has been turned into a data-rich science, there is a need for storing and communicating with this large datasets which has been grown-up tremendously. These are the set of tools that uses databases access for providing solution to biological problems and also used for storing the solution made to the database.

CL_{12}: Alignment Based Tools
These are the set of tools that uses sequence alignment as one of this functionality for performing the task of identifying regions of similarity for functional consequence and structural or evolutionary relationships between the sequences.

Some of the other classifications are windows Based, Linux Based, Image Based, Web-based, Ontology based etc… For easy making for users to select the most appropriate set of tools, classification of the tools based on its features and

functionalities has been carried out based on which some tools categorization have been represented below by considering the above some of the classification.

C_1: Categorized Based on Primary Functionality (Category 1)
The following classifications that fall into this category are Homology and Similarity, Protein Functional Analysis, Structural Analysis and Structural Analysis. Based on the classification made, analysis has been made on each developed tool for its main function which is considered to be the base function. Taking this into account CL_1, CL_2, CL_3 and CL_4 are grouped into one category.

C_2: Categorized Based on Approaches (Category 2)
Algorithm based, computational based and knowledge based tools are categorized by considering used approaches of the tools. Here CL_5, CL_6, CL_7 falls into one category.

C_3: Categorized Based on Features (Category 3)
Classification CL_8 to CL_{12} falls into one category considering features or functionalities of the tools it is classified as Diseases Based, Drug Based, Virus Based, Datasource Based and Alignment Based.

Some the other classification that could be categorized together is Utility, standalone, web-based etc… which is considered as Tools Type (Category 4). Categorized based on technology (Category 5) in which C, R, Perl, PHP, Python, Java etc…. in which the tool has been developed. Windows based, Linux Based, etc… categorized as Platform (Category 6), Category based on sub functionalities (Category 7) are Pattern matching, Text mining etc... In Section III explanation of taken sample ten tools have been narrated along with its classifications and categories

III Considered Tools With Features And Functionalities

Tools in bioinformatics are software programs used for retrieving and analysis of biological data to extract information from them. The tools developed are been made available to the user with at most feature and functionalities integrated together. In this section, for the taken ten tools each has been explained and summarized the tools classification and the category based on the analysis done in Table1. Classifications and categorization of the tools done based on the concepts represented in section II.

A. CGHWeb

Weil Lai et. al [11] developed this web-based tool to apply a number of popular algorithms to a single array CGH profile entered by the user. This generates a heatmap panel of the segmented profiles for each method as well as a consensus profile. This tools interface collects results from multiple algorithms and allows developers to submit their new algorithms and makes possible for the users who are not familiar with programming to ascertain a segmentation profile via multiple methods. The tool developed by using both algorithm and computational approach.

B. GOTreePlus

Bongshin Lee et. al [12] has developed this gene ontology (GO) tool that place over annotation information over GO structures. It can facilitate the identification of important GO terms through interactive visualization of them in the GO structure. Pie chart summarizing an annotation distribution for a selected GO term provides users with a succinct context-sensitive overview of their experimental results.

C. *GENESIS*

Simon Gog et. al [15] developed this web-based tool for three different genome rearrangement problems: Sorting a unichromosomal genome by weighted reversals and transpositions (SwRT), sorting a multichromosomal genome by reversals, translocations, fusions and fissions (SRTl), and sorting a multichromosomal genome by weighted reversals, translocations, fusions, fissions and transpositions (SwRTTl).

D. *HCGene*

Mario Falchi et. al [10] developed this tool for mining complex inbred genealogies that identify clusters of individuals sharing the same expected amount of relatedness is described. Additionally this tool allows reconstruction of sub-pedigrees suitable for genetic mapping in a systematic way.

E. *Phyutility*

Stephen A. Smith et. al [13] developed this network based tool to provide a set of phyloinformatics tools for summarizing and manipulating phylogenetic trees, manipulating molecular data and retrieving data from NCBI. This tool makes easy integration into scripted analyses, and is able to handle large datasets with an integrated database. It performs tree editing on multiple trees and multiple file types which acts as an interface to NCBI's search to fetch its functions.

F. *HGTS*

Hadas Birin et. al [8] developed this network based tool for the evolution of viruses which is very rapidly growing in addition to local point mutations (insertion, deletion,

Table 1: Bioinformatics Tools Classifications and Categories

Categorized Classifications	Bioinformatics Tools									
	CGHWEB	GOTreeplus	GENESIS	HCGene	phyutitlity	HGTS	TOM	BicOverlapper	CSR	Bioshell
	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}
C_1: Category 1										
Homology & Similarity	√			√	√		√			
Protein Functional Analysis		√								√
Structural Analysis								√		
Sequence Analysis			√			√			√	
C_2: Category 2										
Algorithm Based	√	√	√		√	√			√	√
Computational Based	√		√	√		√	√	√	√	
Knowledge Based				√			√	√		
C_3: Category 3										
Disease Based	√						√			
Drug Based										
Virus Based						√				
Datasource Based		√	√	√			√	√		√
Alignment Based					√				√	
C_4: Category 4										
Utility										√
Standalone									√	
Web-based	√		√							
Network Based					√	√				
Ontology Based		√					√			
Others				√				√		

where,

√ Represents the tool belongs to that particular classification, T_1 to T_{10} represents bioinformatics tools and C1 to C4 represents category which contains various classifications.

substitution) it also includes frequent recombination, genome rearrangements and horizontal transfer of genetic materials (HGTS).

G. TOM

Daniele Masotti et. al [18] has developed this ontology based tool to makes use of gene expression information data available on the public repository of Gene Expression Ominbus to extract disease candidate genes given one or two linkage regions.

H. BicOverlapper

Rodrigo Santamaria et. al [16] developed this tool for visualizing biclusters from gene expression matrices in a way that helps to compare biclustering methods, to unravel trends and to highlight relevant genes and conditions.

I. Colorstock, SScolor and Raton (CSR)

Yuri R. Bendan and Ian H. Holmes [7] developed this standalone tool for interactive examination of RNA multiple alignments for covariant mutations which is a useful step in non-coding RNA sequence analysis.

J. Bioshell

Dominik Gront and Andrzej Kolinski [9] developed a new software library for structural bioinformatics. The library contains programs, computing sequence- and profile-based alignments with a variety of structural calculations. This tool is user-friendly which handles various data formats.

In the next section explanation of visualized multi-dimension model and its representation for tools have been explained along with the mode for carrying out transformation in multi-dimension forms.

IV Multi-Dimensional Visualized Classification And Its Transformation

Due to tremendous development of tools for making biology computational it is very hard to judge best tools from a large set of tools that provides high throughput, with at most functionalities and features. To extract few best tools as per requirement from a large set of tools for better and fast evaluation a visualized standard model has been represented. This standard evaluation is a pre-primary evaluation process that should be carried for extracting few best tools from large set of taken tools. In this section, multi-dimension representations for a set of tools have been narrated by using Table 1 as a baseline.

Evaluating by means of its features and functionalities is one of the standards for tools evaluation [1]. In order to provide a flexible and better decision making for making, it is necessary to analysis required features and functionalities. A visualized multi-dimensional functionality, features based classification have been modeled here. The classification of the tools could be made more, on various modes of existence of features and functionality. The ability to make transformation in multi-dimension bases is the identification for quality analysis of the tools. Figure 1 represents generalized visualized model for representing tools classification and categorization in

multi-dimension mode. The proposed model could be applied for any taken set of tools. In this section Figure 2 and Figure 3 have been represented based on the information taken from table 1.

Mostly tools could be divided to three categories taking that into consideration, visualized representation of tools classification in multi-dimension representation we have started representation in 3-D and a slow movement towards n-D is done. It is not necessary always a transformation should done starting from 3-D till the acquired categories to n-D representation. A direct representation of multi-dimension could also be done based on the identified categories. Transformation is necessary or could be done if any functionality, feature classification and categorization have been missed or if the existing tools have been enhanced with additional functionality and features. Any number of categorization and classification could be made.

Here n is the number of categories. As represented in the figure 2 and figure 3 transformation to the next stage in multi-dimension could be made without disturbing the previous stage where, 4-D is the next stage and 3-D is the previous stage in multi-dimensional. In section V a detailed explanation of assessing the tools quality based on its features and functionalities has been represented along with the result analyzed for the taken bioinformatics tools.

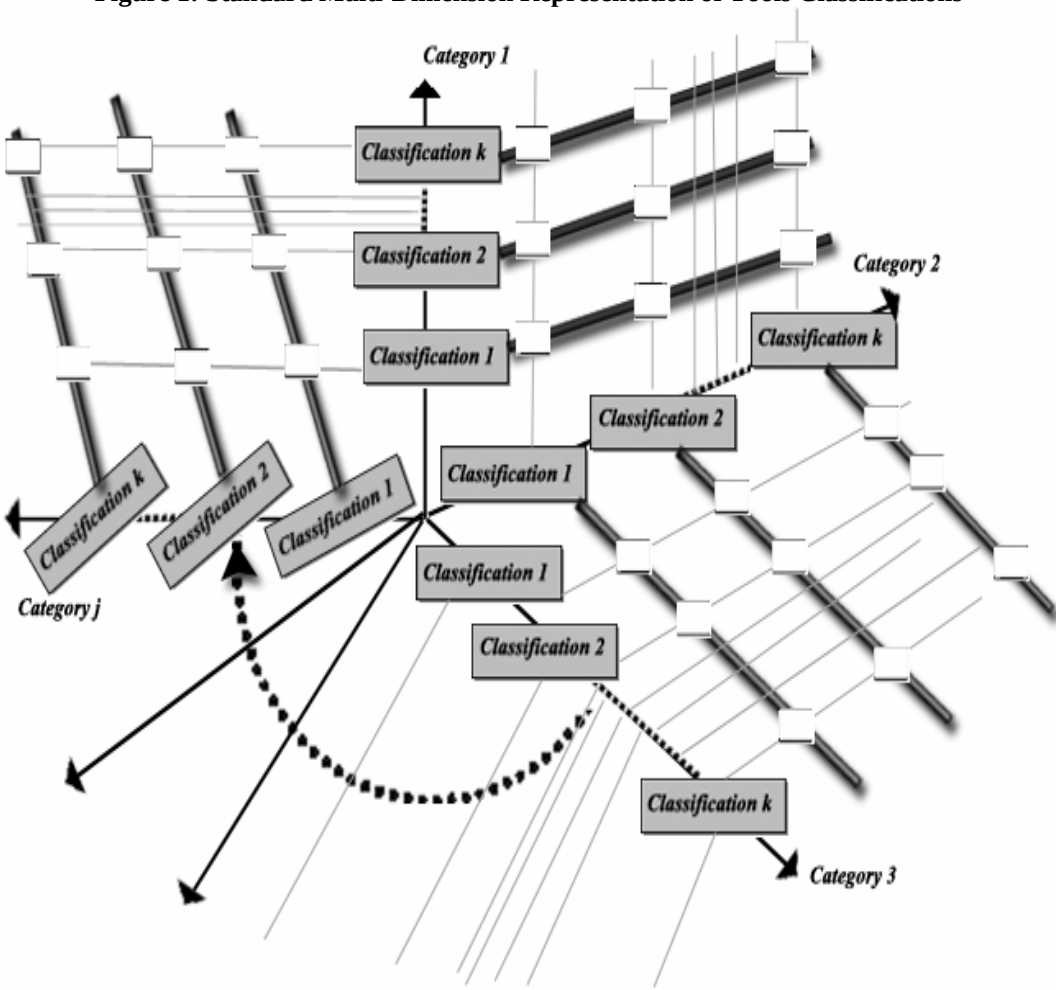
V Quality Assessment And Result Analysis

Assessing tools quality is significant for effective and efficient usage for better decision making due to, vast development of tools in this area. Various standards have been represented for quality evaluation of which quality evaluation by means of tools functionalities and features is the first and primary standard evaluation have been mostly used. Taking this into consideration Multi-dimension visualized model have been proposed for best quality evaluation using tools features and functionalities as a pre-primary step. This model is best utilized for selecting best appropriate few tools from a large set of tools. Following this other standards evaluations could be carried out this reduces the tool evaluation time for the first standard and also provides a best way for selecting best appropriated tools with at most features and functionalities. Since there is tremendous development of tools for performing a particular task it is necessary to select a best tool from the set of tool and evaluates it within stipulated duration.

A. Quality *Evaluation* Schemes

Quality is necessary for the entire developed product

Figure 1: Standard Multi-Dimension Representation of Tools Classifications



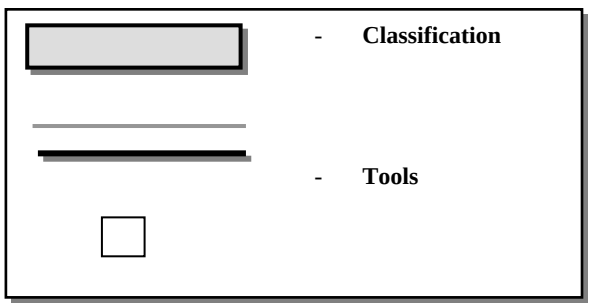


Figure 2: 3-D Classification & Category of Bioinformatics Tools

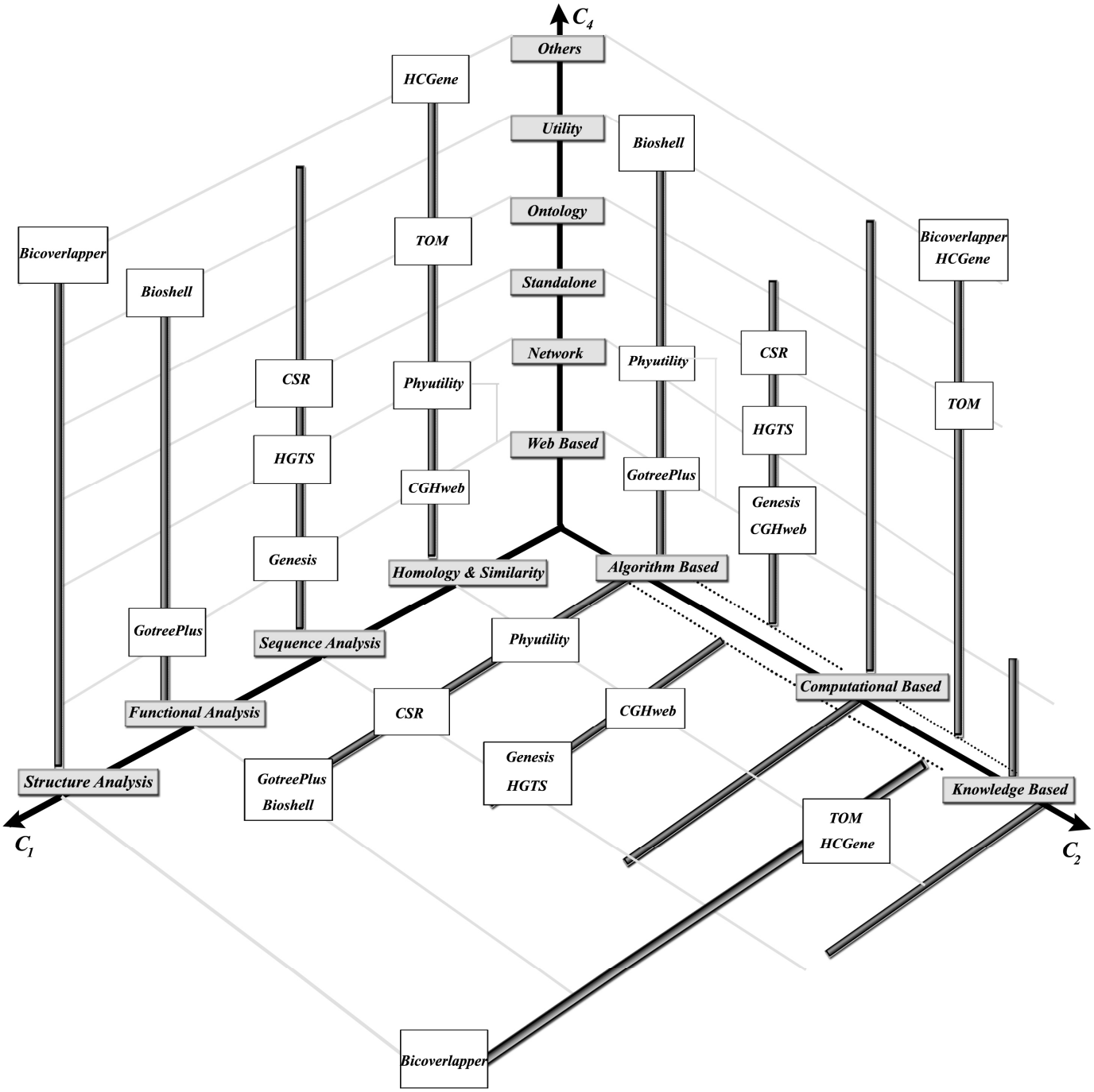


Figure 3: 4-D Classification & Category of Bioinformatics Tools

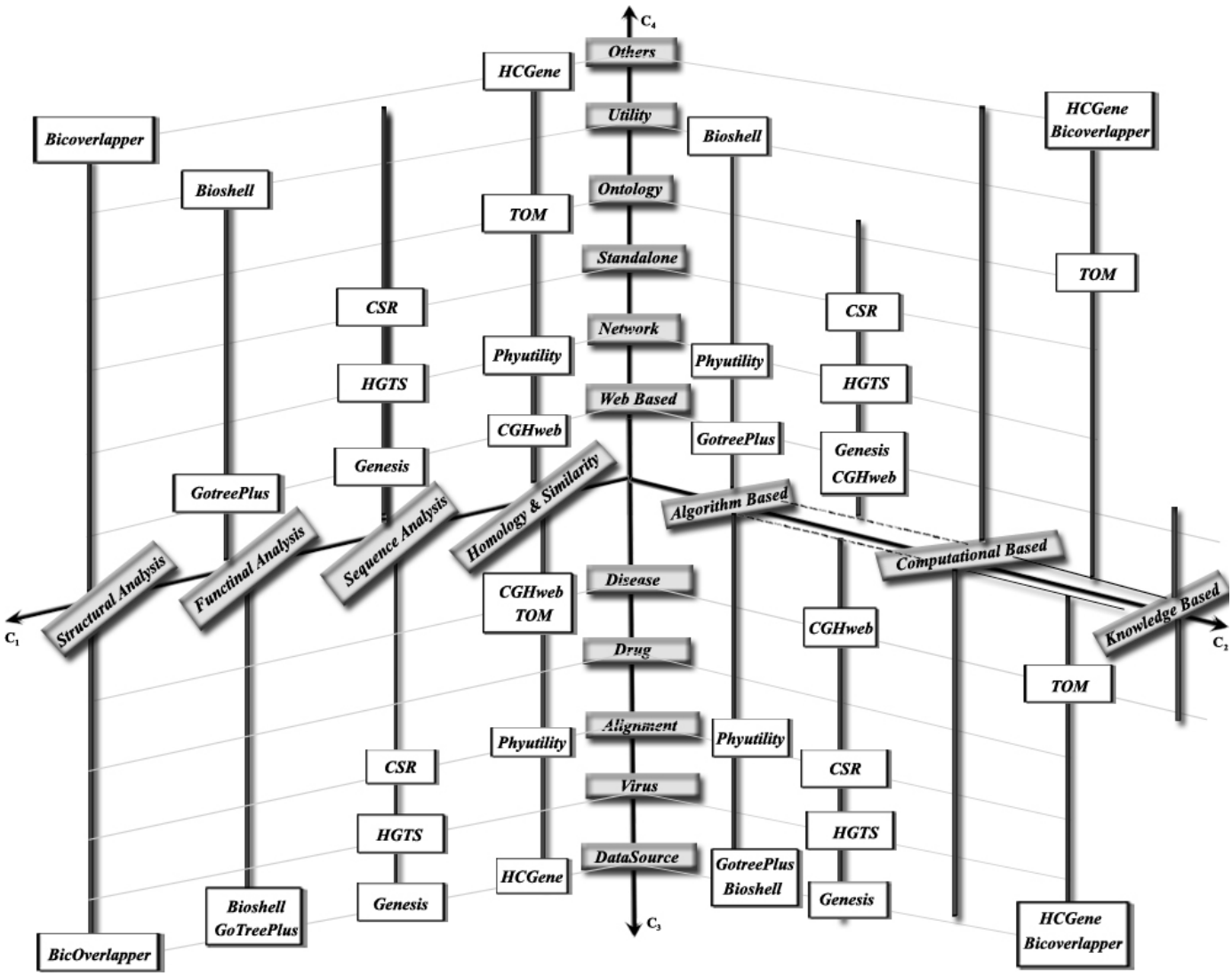


Table 2: Quality Assessment Table

Multi-Dimension	Category	Quality	
3-D	3	Better	
4-D	4		
. . . . . .	. . . . . .	From 5-D to 7-D	Good
n-D	n	Above 7-D	Best

which could be identified by some means of evaluation. Here we have carried out evaluation on a set of ten bioinformatics tools based on its classification and categorization.

This evaluation carried out on the set of tools identifies the better set of tools with user required at most features and functionalities. Further detailed evaluation of tools, based on its performance could be carried out. Here in the evaluation process both classifications and non classified features and functionalities of the tools are considered for evaluating the tools. The more the classifications it represents better the tools quality. Table 2 is considered in evaluating the tools quality for which classification have been made. In addition to this non classified features and functionalities that are essential also considered for quality evaluation.

B. Result Analysis

In this section result analyses for the set of tools which have been taken for evaluation have been explained. From figure 2, figure 3 and based on the quality assessment table (table 2) along with non classified features and functionalities which is manually considered which results for set of selected tools are of Good quality. The tools taken for evaluation could be further classified and up to seven categories which could be represented in 7-D multi-dimension. In general from the analysis made not more than ten valid categories could be made. Here for evaluation other non classified required features and functionalities should be considered manually

VI CONCLUSION

The aim of this paper is to provide a standard model for evaluating large set of tools for selecting only a few appropriate best tools. Classifications and categorization based on tools functionalities and features using standard multi-dimension visualization model has been represented for extracting the best appropriated set of tools for better decision making. Experiments have been carried out on ten bioinformatics tools using this mode for multi-dimension quality assessment for which quality has been judged. This pre-primary quality assessment model makes tools evaluation faster and better. This reduces the large set of tool to certain extent for carrying out further evaluations. From the adopted model it has been found that the more the classification, the better tools quality.

REFREENCES

[1] Jayanthi Manicassamy and P. Dhavachelvan, "Metrics based performance control over text mining tools in bioinformatics", ACM Portal, Pages 171-176, January 2009.
[2] T. W. Lam, W. K. Sung, S. L. Tam, C. K. Wong1 and S. M. Yiu, "Compressed indexing and local alignment of DNA", ACM Portal, pages 791–797, January 2008.
[3] 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.
[4] Jochen Farwer, Martin J. Packer, Christopher A. Hunter, "PREDICTOR: A Web-Based Tool for the Prediction of Atomic Structure from Sequence for Double Helical DNA with up to 150 Base Pairs", ISO Press, 2007.
[5] 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.
[6] Mathiak B., Eckstein S., "Five steps to text mining in biomedical literature", Proceedings 15th European Conference, 2004.
[7] Yuri R. Bendan and Ian H. Holmes, "Colorstock, SScolor, Raton: RNA alignment visualization tools", ACM Portal, pages 579–580, January 2008.
[8] Hadas Birin, Zohar Gal-Or, Isaac Elias and Tamir Tuller, "Inferring horizontal transfers in the presence of rearrangements by the minimum evolution criterion", ACM Portal, pages 826–832, January 2008.
[9] Dominik Gront and Andrzej Kolinski, "Utility library for structural bioinformatics", ACM Portal, pp 584–585, January 2008.
[10] Giorgio Valentini and Nicolo Cesa Bianchi, "HCGene: a software tool to support the hierarchical classification of genes", ACM Portal, pp 729–731, January 2008.
[11] Weil Lai, Vidhu Choudhary and Peter J. Park, "CGHweb: a tool for comparing DNA copy number segmentations from multiple algorithms ", ACM Portal, pp 1014–1015, February 2008.
[12] Bongshin Lee, Kristy Brown, Yetrib Hathout and Jinwook Seo, "GOTreePlus: an interactive gene ontology browser", ACM Portal, February 2008, pp 1026–1028.
[13] Stephen A. Smith and Casey W. Dunn, "Phyutility: a phyloinformatics tool for trees, alignments and molecular data", ACM Portal, pp 715–716, January 2008.
[14] Juby Jacob, Marcel Jentsch, Dennis Kostka, Stefan Bentink and Rainer Spang, "Detecting hierarchical structure in molecular characteristics of disease using transitive approximations of directed graphs", ACM Portal, pp 995–1001, February 2008.
[15] Simon Gog, Martin Bader and Enno Ohlebusch, "GENESIS: genome evolution scenarios", ACM Portal, pp 711–712, January 2008.
[16] Rodrigo Santamaria, Roberto Thero'n and Luis Quintales, "BicOverlapper: A tool for bicluster visualization", ACM Portal, pp 1212–1213, March 2008.
[17] Rodrigo Secolin, Cristiane S. Rocha, Fábio R. Torres, Marilza L. Santos, Cláudia V. Maurer-Morelli, Neide F. Santos, Iscia Lopes-Cendes, "LINKGEN: A new algorithm to process data in genetic linkage studies", Science direct, February 2008.
[18] Daniele Masotti, Christine Nardini, Simona Rossi, Elena Bonora, Giovanni Romeo, Stefano Volinia and Luca Benini, "TOM: enhancement and extension of a tool suite for in silico approaches to multigenic hereditary disorders", ACM Portal, pp 428–429, November 2007.