Chapter 4

4 Cluster Analysis of Cancer Data Using Similarity Measures

A number of supervised and unsupervised algorithms are available in statistics and machine learning literature for clustering microarray data but the algorithms are restricted in their ability to evaluate the results of a clustering algorithm in the light of biologically meaningful clusters. If two gene sequences are similar, then it is expected that their genetic expressions are similar and that they are similarly annotated in the Gene Ontology (GO) databases. Hence a comparison of the expression level similarity of two gene sequences against their corresponding similarity of annotation in the GO can establish this fact. Semantic similarity has now become a valuable tool for validating the results drawn from biomedical studies such as gene clustering and gene expression data analysis. The work in this chapter borrows from the previous work on meta-ensembles using cancer datasets where the output of several clustering algorithms are subsequently fed to a consensus building process in order to generate a stable set of cluster results. Next, these cluster results are further refined through a sequence of biological validation process for each gene pair of a given cluster using semantic similarity and sequence similarity. The approach has been tested on several benchmark cancer datasets in an attempt to provide a more accurate biological analysis of the clusters and the results have been found to be satisfactory.

4.1 Introduction

Bioinformatics resources hold a lot of information generally in the form of sequences. The annotation of the sequences is done to express its understanding, usually in the form of a natural language that is understandable by humans but may not be accessible computationally. Therefore, ontologies are used to express this natural language annotation in a form that is computationally accessible as well as understandable by humans. Gene Ontology (GO) [259] is one of the most important ontologies within the bioinformatics community that provides annotation for GO terms.

Clustering gene expression data is a powerful tool in bioinformatics to reveal biologically relevant information. It has been used for the purpose of grouping genes or proteins having similar expression patterns, leading to the possibility of sharing common biological pathways [260,261,262,263,264]. Out of the many clustering methods, deciding which clustering method to use and also to determine the number of clusters that are most appropriate for the data can be a daunting task. Taking a single algorithm with one parameter setting and expecting the results to indicate the proper structure of the data will seldom provide the correct analysis. It is also the observation that a supervised or an unsupervised approach alone generally cannot provide accurate analysis of gene expression microarray data whereas the hybrid approach tries to combine the benefits of both supervised and unsupervised learning to improve accuracy. Ensemble methods, which combine the output of several algorithms, are able to improve the robustness and stability of the analysis, but the need of the hour is the correct interpretation of data.

Ideally, the resulting clusters should not only have good statistical properties (compact, well-separated, connected and stable), but also should give results that are biologically relevant. Hence, it is advisable to have runs of multiple approaches of clustering algorithms and only then a comparison of the partitions can result in reliable conclusions drawn from the cluster analysis.

The application of semantic similarity concepts to the Gene Ontology [265,266,267] has fuelled prospects that a well defined annotation of GO terms will bear a close relationship to the functional similarity of the terms that are represented. In other words, it is reasonable to state that gene sequences with similar expression patterns are likely to be similarly annotated within the ontology [268].

4.2 Related Work

For this task, eight clustering algorithms have been considered, which are representative of the partitioning, hierarchical and model-based approaches and their main features are summarized below.

4.2.1 Partitioning Approach

The algorithms considered under partitioning approach are

***K*-means** [269]: It is a simple and fast centroid-based clustering algorithm where the data is initially partitioned into *k* pre-defined number of clusters. *K*-means provides baseline results that are used to compare with when new clustering algorithms are developed. To detect the optimal number of clusters, the algorithm is run repeatedly with different values of *k*. This algorithm is not suitable to detect clusters of arbitrary shapes.

PAM (Partitioning Around Mediods) [23]: PAM is a *k*-medoid method where the medoid is a representative object for each cluster and is selected by using dissimilarity values and an iterative optimization approach. In this algorithm, every non-selected object is grouped with its nearest medoid. The average dissimilarity between an object and the medoid of its cluster determines the quality of the partition.

CLARA (Clustering LARge Applications) [23]: CLARA is also a *k*-medoid method following the same principle as PAM. This algorithm finds the medoid from a sample of the dataset instead of the entire dataset and then applies PAM to this sample. The remaining objects are then classified using partitioning principles.

4.2.2 Hierarchical Approach

This approach generates a hierarchical series of nested clusters which are represented by a tree-structure called a dendrogram. The desired number of clusters can be obtained by cutting the dendrogram at an appropriate level. The algorithms under this approach used are

Hierarchical [23]: The hierarchical algorithm selected for this experiment is AGNES (AGglomerative NESting) which is an agglomerative clustering algorithm. The algorithms proceeds by placing each object initially in its own cluster and then the clusters are successively joined together in order of their proximity. The proximity of any two clusters is determined by a dissimilarity matrix and it can be based on a variety of agglomeration methods such as *single linkage*, *complete linkage* and *average linkage*.

DIANA (DIvisive ANAlysis) [23]: DIANA is a divisive hierarchical algorithm where the observations are initially placed in a single cluster. The clusters are then partitioned successively until each cluster contains a single observation. The highest average dissimilarity among all the observations acts as a point of division.

SOTA (Self-Organizing Tree Algorithm) [270]: SOTA is an unsupervised network with a divisive hierarchical binary tree structure. It uses a fast algorithm and is suitable for clustering large number of objects. The algorithm was originally proposed for phylogenetic reconstruction and later it has been successfully applied to cluster microarray gene expression data [271].

4.2.3 Model-Based Approach

The following two algorithms under the model-based approach are considered.

SOM (Self-Organizing Map) [31]: SOM is a popular among computational biologists since it is non-susceptible to noisy data. It is based on neural networks and is capable of generating intuitive cluster patterns of high-dimensional datasets. The algorithm takes as its inputs the initial number of clusters and the grid structure of the neuron map.

Model-based Clustering [272]: This is an approach used when an object may exhibit membership in more than one cluster. It attempts to find the optimum alignment between the given data and a statistical model consisting of a finite mixture of Gaussian distributions. The group membership of an object is estimated using the maximum likelihood algorithm.

4.3 Validation

Validation is an essential step to be carried out for the purpose of evaluation of the clustering results.

4.3.1 External Validation

External validation incorporates prior knowledge of solutions to the problem being addressed to evaluate the effectiveness of a model. Semantic similarity, which is a comparison of biological entities in terms of similarity in the meaning of their annotations, and sequence similarity are described next.

4.3.1.1 Semantic Similarity

The grouping of genes or proteins depending on their differential expression profiles is the most common method used to interpret microarray data. This method is derived from the understanding that genes expressed in an organized manner are likely to be involved in the same biological processes or may have a similar function [262]. In other words, if two gene sequences are similar, it is expected that their genetic expressions are similar, and that they are similarly annotated in the GO [268]. Therefore, it is reasonable to assume that gene sequences with similar expression patterns might have similarly annotated profiles, i.e., that the expression correlation might relate to the semantic similarity [267,268,273].

Evidences have also shown that the GO can reflect the functional similarity of gene sequences by the closeness of the terms that they represent [267,274,275,273]. In order to establish how two GO terms relate to one another, the concept of semantic similarity as implemented in GO (involving several approaches) can be exploited. Semantic similarity is used as a measure in natural language processing and has been extended to measure the degree of similarity between the terms in the GO structure [274]. This measure can also be directly converted to a measurement of the similarity between two nucleotides/proteins.

Comparison of GO Terms

Two main types of approaches for comparing terms in a graph-structured ontology such as GO have been proposed: edge-based, which use the edges and their types as the data source; and node-based, in which the nodes and their properties are the main data sources. A third approach adopts the hybrid approach involving edge-based and node-based approaches for calculating the semantic similarity. The approaches and the measures proposed are discussed next.

Edge-Based: This approach counts the number of edges in the graph path between two terms [276]. The distance between two terms is taken to be either the shortest path or the average of all paths and then converted into a similarity measure. The similarity measure is based on the common path technique that calculates the length of the path from the lowest common ancestor of the two terms to the root node [277].

Pekar and Staab [278] proposed a measure which is based on the length of the longest path between two terms' lowest common ancestor and the maximum common ancestor depth (root), and on the length of the longest path between each of the terms and that common ancestor. This measure was first applied to GO by Yu *et al.* [279]. Cheng *et al.* [280] proposed a weighted maximum common ancestor depth measure to reflect depth of the terms. A non-weighted maximum common ancestor depth measure was proposed by Wu *et al.* [281] and then an adjustment of this measure was proposed by Wu *et al.* [282] who introduced a term specificity involving the distance to the nearest leaf node and the distance to the lowest common ancestor.

Node-based: The node-based approach uses the *Information Content* (IC) of a GO term to compare the properties of the terms involved, along with their ancestors and descendants. The IC is determined by the terms' frequency of occurrence in annotations, i.e. a rarely used term will contain a greater amount of information [283].

Another approach adopted in node-based is the Most Informative Common Ancestor (MICA) approach that considers only the common ancestor with the highest IC [284]. The Disjoint Common Ancestors (DCA) approach considers all disjoint common ancestors, i.e. the common ancestors that do not subsume any other common ancestors [285].

Lord *et al.* [265] introduced several node-based related semantic similarity metrics for use with GO, originally developed for WordNet, that are based on the MICA of two GO terms, whereas a measure proposed by Resnik [284] measures similarity between two terms as the IC of their MICA. Lin [286] presented an information-theoretic definition of similarity based on a probabilistic model. Jiang and Conrath [287] proposed a measure involving a mixed approach of edge-based method and information content calculation of node-based method along with other factors such as local density, node depth and link type, whereas Schlicker's semantic similarity [275] is a combination of Resnik's and Lin's similarity measure.

Hybrid: Wang *et al.* [288] developed a hybrid measure in which the semantic similarity between two GO terms *A* and *B* is calculated by summing the semantic values of all common ancestors to each of the terms and then dividing by the total semantic contribution of each term's ancestors to that term. Othman *et al.* [289] proposed a hybrid distance measure in which each edge is weighted by node depth, node link density and the difference in IC between the nodes linked by that edge.

Software Used For Deriving Semantic Similarity Measures

There have been quite a few software tools available for calculating the semantic similarity of terms in ontology such as GOSemSim [290], seGOsa [291], DOSim [292] and many others. Even though DOSim provides support for different semantic measures, the measuring of the similarity between human genes is done in terms of diseases. Hence for the purpose it is preferred to use GOSemSim developed by Yu *et al.,* for calculating semantic similarity between GO terms. This tool was developed as a package for the statistical computing environment *R* within the Bioconductor project. GOSemSim depends on a number of packages provided by Bioconductor, such as package GO.db to obtain GO terms and relationships and package org.Hs.eg.db to obtain annotations of gene sequences for human. The main advantage of these Bioconductor implementations is the possibility of integration between the semantic similarity and other packages. Moreover, GOSemSim also provides support for Lin [286], Jiang and Conrath [287] and Wang [288], the semantic similarity measures selected by us for the experiments.

Some of the semantic similarity measures that have been developed for use with GO are shown in *Table* 4.1 along with their formula and the software in which they are available.

Table 4.1: Summary of few Semantic Similarity Measures

Measure	Formula	Software Package		
		GOSemSim	seGOsa	DOSim
Resnik [284]	$sim_{Resnick}(t_1, t_2) = -\ln[p_{ms}(t_1, t_2)]$	✓	✓	✓
Lin [286]	$sim_{Lin}(t_1, t_2) = \frac{2 \text{ x} \ln[p_{ms}(t_1, t_2)]}{\ln(p(t_1) + \ln(p(t_2)}$	✓	✓	✓
Jiang and Conrath [287]	$sim_{Jiang}(t_1, t_2) = 1 - \min(1, \{\ln p(t_1) + \ln p(t_2) - 2 \text{ x} \ln[p_{ms}(t_1, t_2)]\})$	✓	✓	✓
Schlicker et al. [275]	$sim_{Schlicker}(t_1, t_2) = \frac{2 \text{ x} \ln[p_{ms}(t_1, t_2)]}{\ln(p(t_1) + \ln(p(t_2)} \text{ x} \left(1 - p_{ms}(t_1, t_2)\right)$	✓	×	✓
Wang et al. [288]	$sim_{Wang}(A, B) = \frac{\sum_{t \in T_A \cap T_B} S_A(t) + S_B(t)}{\sum_{t \in T_A} S_A(t) + \sum_{t \in T_B} S_B(t)}$	✓	×	✓

Here:

- t, t_1 and t_2 are GO terms
- $p(t)$ represents probability of term t
- p_{ms} is the minimum subsumer
- $S_A(t)$ is the semantic value of GO term *t* related to term *A*
- $S_B(t)$ is the semantic value of GO term *t* related to term *B*

Critique for Selecting Lin, Jiang and Conrath, and Wang Semantic Measures

- Edge-based methods are based on the assumption that all edges represent uniform distances [293] and all nodes in the taxonomy are evenly distributed having similar densities. These assumptions are found not to be true in real taxonomies [265] since some GO branches may be very deep, some terms may have many children terms and some edges may cover a large conceptual distance as compared to others.

- Node-based methods apply concepts borrowed from information science [267]. The information content of a GO term is determined by its frequency of occurrence in the annotations. The lower the probability, the more information a node contains and in a hierarchical taxonomy such as the GO, probabilities increase as one goes higher within the taxonomy. It has been shown by [284] that node-based measures provide better results than their edge-based counterparts as they are not affected by factors such as irregular link density, varying conceptual distance and uneven distribution of nodes. Hence node-based methods are given preference over edge-based methods while calculating semantic similarities.

- A drawback of the Resnik measure is that it does not differentiate between two terms if their subsumer is the same and it loses a part of the information contained in the *structure* of the taxonomy by only concentrating on the information content of a term [268]. By contrast, Lin measure and Jiang and Conrath measure take the information content of the two terms as well as the minimum subsumer into consideration.

- Moreover, Wang et al. evaluated measures proposed by Resnick, Lin and Jiang and Conrath, and tested these measures against gene co-expression data using linear correlation [288]. They pointed out that the distance of a term from the closest common ancestor may not accurately represent the semantic difference between two GO terms, since two terms nearer to the root of the ontology and sharing the same parent should have larger semantic difference than those far away from the root and having the same parent.

- For the purpose of calculating the semantic similarity measures of the gene-pairs, the semantic measures proposed by Lin, Jiang and Conrath, and Wang have been selected for conducting the experiments.

- The semantic measures selected are available as Bioconductor components, distributed as *R* packages, which facilitate the analysis of genomic data by associating it with biological metadata provided by GenBank, Entrez genes and PubMed databases.

4.3.1.2 Sequence Similarity

A sequence similarity search compares a query sequence to a larger database of sequences to find alignments between the query and database that reflect similarities between the two. A sequence alignment provides a measure of relatedness between nucleotide or amino acid sequences. Since all known species, whether yeast, mice, or humans appear to be related to each other, a sequence similarity search may uncover an unknown gene of one species having the same functionality as a gene belonging to a different species. A high sequence similarity score usually implies significant functional or structural similarity between the DNA, RNA or amino acid sequences.

Sequence similarity analysis involves several factors such as how to score individual matches across sequences, whether to perform global or local searches, type of algorithm to use and evaluation of the results to determine the statistical significance of an alignment score, i.e., is this alignment better than could be expected between any two random sequences.

Scoring Model: Sequence alignment is the procedure of comparing two (pair-wise alignment) or more (multiple-alignment) sequences by searching for a series of individual characters or character patterns that are present in the same order in both the sequences. Portions of a sequence may be related even though mutations or changes may have occurred between the sequences in one of two ways. The first type of change is called a *substitution*, where a base of one type is substituted by another type and is evaluated using substitution matrices. The second type of change is based on *gap*, where an insertion or deletion of a base is made to the original sequence to improve the alignment between sequences but the gap should be kept to a minimum number. Gap penalty scores are negative and depend upon on the length of the gap.

Alignment algorithms: Alignment algorithms are used to determine the optimal alignment of a pair of sequences based on a particular scoring mechanism. Sequence alignment determines the correspondences between substrings in the sequences such that the similarity score is maximized. Fundamentally, two different alignment problems exist. Global Alignment finds the best alignment from start to end of both sequences (with provision for gaps). Local Alignment is used to find sub-sequences of a sequence that have the best alignment. The Needleman-Wunsch algorithm [294] is a dynamic algorithm that can be used to analyze global alignments. The Smith-Waterman algorithm [295] modifies the Needleman-Wunsch algorithm to allow it to search for local alignment sub-sequences within a larger sequence and have been often found to work better for sequence similarity searches. Although these dynamic algorithms are guaranteed to find optimal alignment matches, they are computationally expensive, especially for large datasets. Heuristic algorithms such as BLAST [296] have been developed to reduce the computational burden on evaluations. It will find most of the results and in less time but will miss a small fraction of the results generally found by an optimal approach.

4.3.2 Internal Validation

For internal validation, selection of validity indices should be done based on the suitability for the particular data conformations at hand [297]. A brief summary of the validity measures is presented in *Table* 4.2.

Table 4.2: Summary of Measures used for Internal Validation

Measure	Features	Formula
Connectivity [193]	• connectivity indicates the degree of connectedness of the clusters • it has a value between zero and ∞ and should be minimized	$Conn(C) = \sum_{i=1}^{N}\sum_{j=1}^{L} x_{i,nn_{i(j)}}$ where: • partition C = {C_1,...,C_K} of *N* observations into *K* disjoint clusters • $nn_{i(j)}$ is the *j*th nearest neighbor of observation *i*, • L is the number of neighbors that contribute to the connectivity measure
Dunn index [197]	• it is the ratio of the smallest distance between observations, not in the same cluster, to the largest intra-cluster distance • The Dunn index has a value between zero and ∞ and should be maximized	$Dunn\ (C) = \frac{min_{C_k,C_l \in C, C_k \neq C_l}\left(min_{i \in C_{k,} j \in C_{l,}} dist(i,j)\right)}{max_{C_m \in C_,} diam(C_m)}$ where: • $diam(C_m)$ is the maximum distance between observations in cluster C_m.
Silhouette width [195]	• it is the average of each observation's silhouette value that measures the degree of confidence in the clustering of an observation. • range of values vary from 1 to -1	$S_i = \frac{b_i - a_i}{\max\ (b_i, a_i)}$ where: • a_i is the average distance between *i* and all other observations in the same cluster • b_i is the average distance between *i* and the observations in the "nearest neighboring cluster"

One of the measures used is connectivity that indicates the extent to which observations are placed in the same cluster as their nearest neighbors in the data space [193]. The Dunn index [197] and silhouette width [195] measures are used for the purpose of determining the how compact is a cluster and also how much separation exists between the clusters, which is obtained by measuring the distance between the cluster centroids. The cluster homogeneity, indicating compactness, is arrived at by inspecting the intra-cluster variance. The details of each measure are discussed in *Section* 2.7.

Table 4.3: Summary of Stability Measures

Measure	Features	Formula
APN: Average proportion of non-overlap [298]	• measures the average proportion of observations not placed in the same cluster under both the cases • has a value between 0 and 1 with values close to zero indicating highly consistent clustering results	$APN(C) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\left(1 - \frac{n(C^{i,l} \cap C^{i,0})}{n(C^{i,0})}\right)$ • $C^{i,0}$ represent the cluster containing observation *i* using the original clustering • and $C^{i,l}$ represent the cluster containing observation *i* where the clustering is based on the dataset with column *l* removed
AD: Average Distance [298]	• measures the average distance between observations placed in the same cluster under both cases • has a value between zero and ∞, and smaller values are preferred	$AD(C) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\frac{1}{n(C^{i,0})n(C^{i,l})}\left[\sum_{i\in C^{i,0}, j\in C^{i,l}} dist(i,j)\right]$ • $C^{i,0}$ represent the cluster containing observation *i* using the original clustering • and $C^{i,l}$ represent the cluster containing observation *i* where the clustering is based on the dataset with column *l* removed • $dist$ is Euclidean
ADM: Average distance between means [298]	• measures the average distance between cluster centers for observations placed in the same cluster under both cases • has a value between zero and ∞, and smaller values are preferred	$ADM(C) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M} dist\left(\bar{x}_{C^{i,l}}, \bar{x}_{C^{i,0}}\right)$ • $\bar{x}_{C^{i,0}}$ is the mean of the observations in the cluster which contain observation *i* using the original clustering • $\bar{x}_{C^{i,l}}$ is the mean of the observations in the cluster which contain observation *i* where clustering is based on the dataset with column *l* removed • $dist$ is Euclidean
FOM: Figure of merit [185]	• measures the average intra-cluster variance of the observations in the deleted column where the clustering is based on the remaining columns • has a value between zero and ∞, and smaller values are preferred	$FOM(l,C) = \sqrt{\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in C_k(l)} dist(x_{i,l}, \bar{x}_{C_k(l)})}$ • $x_{i,l}$ is the value of the i^{th} observation in the l^{th} column • $\bar{x}_{C_k(l)}$ *is* the average of cluster $C_k(l)$ • $dist$ is Euclidean

4.3.3 Stability Measures

The stability measures are a special version of internal measures which include the average proportion of non-overlap (APN) [298], the average distance (AD) [298], the average distance between means (ADM) [298], and the figure of merit (FOM) [185]. They evaluate the stability of the clustering result of the entire dataset by comparing it with the clusters obtained by removing one column at a time [298,185]. These measures are ideally suited for highly correlated data. The stability measures are summarized in the *Table* 4.3 given below.

4.3.4 Biological Measures

Biological validation evaluates the ability of a clustering algorithm to produce biologically meaningful clusters from microarray data. Two measures available for these purposes are the biological homogeneity index (BHI) [299] and biological stability index (BSI) [299], the summary of which is given in *Table* 4.4. The BHI measures the average proportion of gene pairs that are clustered together having matching biological functional classes. The BSI measure inspects the consistency of clustering for genes with similar biological functionality. By removing one sample at a time, the cluster membership for genes with similar functional annotation is compared with the cluster membership using all the samples.

Table 4.4: Summary of Biological Measures

Measure	Features	Formula
BHI: Biological Homogeneity Index [299]	• measures the average proportion of gene pairs that are clustered together having matching biological functional classes • has a value between 0 and 1 with values close to 1 are preferred	$BHI = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_j(n_j-1)}\sum_{x\neq y\in D_j} I\left(C(x)=C(y)\right)$ • x, y are genes that belong to the same statistical cluster D • $C(x)$ is a functional class containing gene x • $C(y)$ is a functional class containing gene y • k is the number of statistical clusters • $n_j = n(Dj \cap C)$ is the number of annotated genes in cluster Dj
BSI: Biological Stability Index [299]	• inspects the consistency of clustering for genes with similar biological functionality • has a value between zero and 1, with larger values corresponding to more stable clusters	BSI $= \frac{1}{F}\sum_{i=1}^{F}\frac{1}{n(C_i)(n(C_i)-1)\rho}\sum_{j=1}^{\rho}\sum_{x\neq y\in C_i}\frac{n(D^{x,0}\cap D^{y,j})}{n(D^{x,0})}$ • F is the total number of functional classes • $D^{x,0}$ is the statistical cluster containing observation x based on all the data • $D^{y,j}$ is the statistical cluster containing observation y when column j is removed

4.4 Motivation

To improve the accuracy of cancer data classification, the method can contribute significantly. The existing clustering based unsupervised methods usually suffer from significantly high false alarms. Also, the groups of genes identified by most of the methods are often found to be biologically irrelevant, one of the reasons being the choice of a suitable proximity measure.

Our method borrows from the previous work on ensemble of classifiers [300]. The approach makes use of several base clustering algorithms to generate individual cluster results followed by building an appropriate consensus based on their individual responses. If a pair of genes is identified by a majority of the algorithms as belonging to same cluster based on their expression similarity, and if they are also found to be similar semantically, then they are considered as belonging to the same cluster. These cluster results are then further refined through a sequence of biological validation process for each gene pair of a given cluster.

Analysis of the clusters of genes on the basis of the expression profiles is not without its share of problems arising from noise, missing values and also inaccurate estimation of the data and should not be dependent only on the expression data but rather, should incorporate some prior knowledge of the data. Lord [274] found that the semantic similarity calculated from annotations correlates well with the sequence similarity and hence, the role of semantic similarity and sequence similarity becomes very important in cancer data classification.

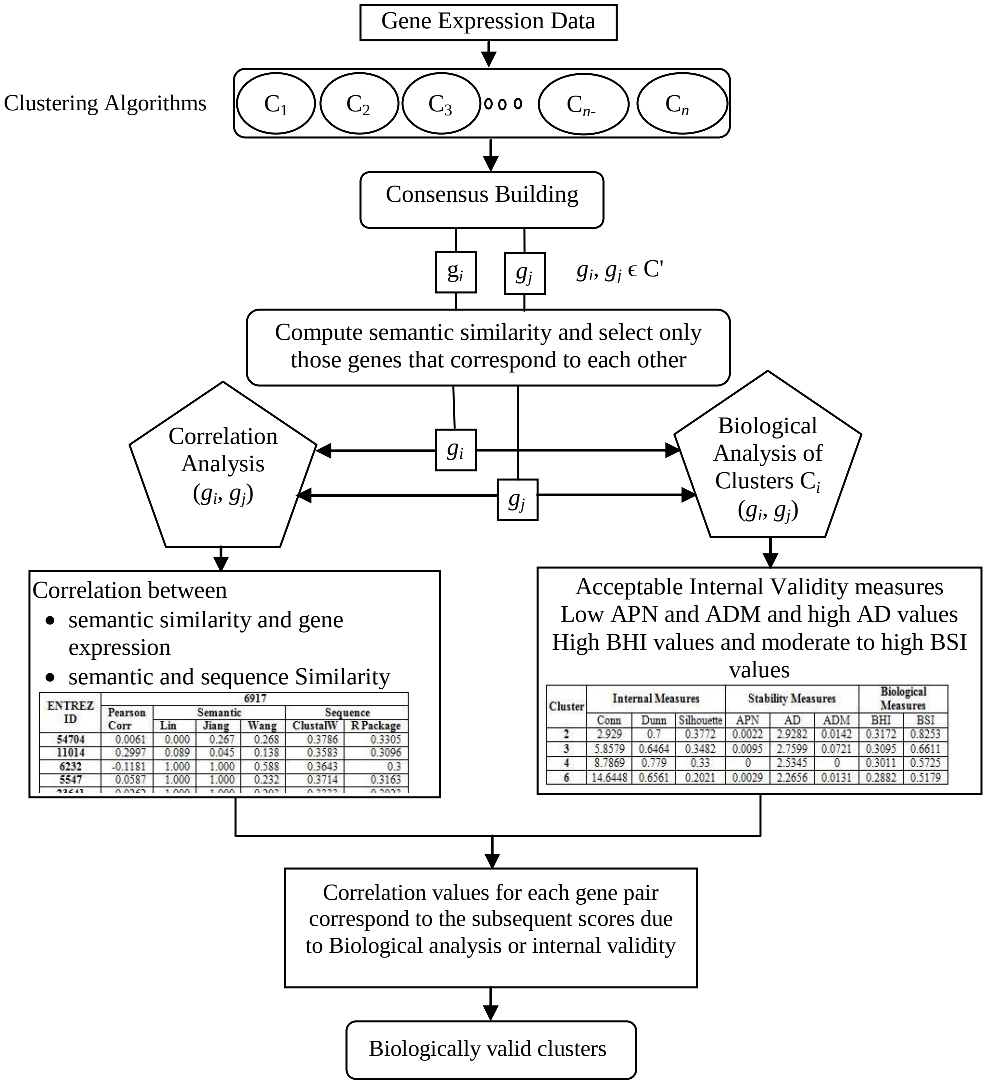
After generating the clusters, external validation in the form of semantic and sequence similarities is performed on several benchmark datasets using GOSemSim [290] package of the Bioconductor project [301], Clustal W [302] and statistical computing environment *R* software package. The cluster results are validated from compactness point of view, followed by two biological validation measures, biological homogeneity index (BHI) [299] and biological stability index (BSI) [299], to evaluate the capability of a clustering algorithm in generating biologically meaningful clusters.

4.5 Methodology

The primary purpose of the work in this chapter is to improve upon the analysis of the results that were obtained from a previous work done on cancer datasets involving supervised ensemble methods [300]. Analysis of cancer datasets [3] demands high accuracy and hence it is essential to validate the results of clustering with the help of external and internal validation.

The similarity of two genes is obtained by computing the correlation between their expression data and the external validation is performed with the help of semantic and sequence similarity measures. Significant quantitative relationship between GO-based gene similarity and expression correlation of pairs of genes were detected. It is further observed that the semantic similarity calculated from GO correlates well with sequence similarity [265]. The internal validity measures assess the quality of a given clustering, based solely on the data themselves and is carried out by visual inspection of stability and consistency of the results. To carry out internal validation, the internal measures such as connectivity [193], Dunn index [197] and silhouette width [195] and stability measure (a special case of internal measures) are used. Finally, biological measures which are incorporated as biological homogeneity index (BHI) [299] and biological stability index (BSI) [299] are employed to evaluate the stability and consistency of a clustering algorithm's ability to produce biologically meaningful clusters.

A conceptual framework used to establish a biological validity of clusters in terms of external and internal validity measures is shown in *Figure* 4.1. Individual cluster results are generated by several clustering algorithms, which are subsequently fed to a consensus building process in order to generate a stable set of cluster results. The consensus building process determines a unanimous decision based on the responses (cluster results) generated by a set of base clustering algorithms, identified after an exhaustive experimentation with a large number of benchmark datasets. The individual responses (i.e., cluster results) are then combined using a weighted majority voting mechanism, where the weight of a classifier is decided based upon its previous performance. After the consensus has been built, a gene pair (g_i, g_j) is picked up from the modified cluster results C'_i. Though a pair of genes in a cluster obtained from the clustering algorithms may be correlated based on expression similarity, they may not be so semantically. The semantic similarity for each gene pair (g_i, g_j) is computed using the GOSemSim [290] package and as a filtering process, only those genes that show a correspondence to other genes semantically are retained.


Gene Expression Data
Clustering Algorithms
C1
C2
C3
Cn-
Cn
Consensus Building
gi
gj
gi, gj ϵ C'
Compute semantic similarity and select only those genes that correspond to each other
Correlation Analysis (gi, gj)
gi
gj
Biological Analysis of Clusters Ci (gi, gj)
Correlation between
semantic similarity and gene expression
semantic and sequence Similarity
Acceptable Internal Validity measures
Low APN and ADM and high AD values
High BHI values and moderate to high BSI values
Correlation values for each gene pair correspond to the subsequent scores due to Biological analysis or internal validity
Biologically valid clusters

ENTREZ ID	6917					
	Pearson Corr	Semantic			Sequence	
		Lin	Jiang	Wang	ClustalW	R Package
54704	0.0061	0.000	0.267	0.268	0.3786	0.3305
11014	0.2997	0.089	0.045	0.138	0.3583	0.3096
6232	-0.1181	1.000	1.000	0.588	0.3643	0.3
5547	0.0587	1.000	1.000	0.232	0.3714	0.3163

Cluster	Internal Measures			Stability Measures			Biological Measures	
	Conn	Dunn	Silhouette	APN	AD	ADM	BHI	BSI
2	2.929	0.7	0.3772	0.0022	2.9282	0.0142	0.3172	0.8253
3	5.8579	0.6464	0.3482	0.0095	2.7599	0.0721	0.3095	0.6611
4	8.7869	0.779	0.33	0	2.5345	0	0.3011	0.5725
6	14.6448	0.6561	0.2021	0.0029	2.2656	0.0131	0.2882	0.5179

Figure 4.1: Biological validity of clusters in terms of External and Internal validity measures

Thereafter, the cluster results are then further refined through a sequence of biological validation process for each gene pair of a given cluster. Moreover, the final result is dependent on both expression similarity (used during clustering) and semantic similarity. A gene pair (g_i, g_j) having a high gene similarity and high semantic similarity is an indication that the gene sequences might have similarly annotated profiles which can be confirmed if the pair wise sequence score is also high.

4.6 Dataset Description

The experiments were performed on the cancer datasets. The first dataset on which the computations were performed is the gene expression data of the breast cancer microarray study from van't Veer *et al.* [304] known as VEER1. The second and third datasets are the Lymphoma and the embryonal tumour of the Central Nervous System (CNS) dataset, both from Kent Ridge Biological Dataset Repository [258]. The datasets are summarised in *Table* 4.5.

Table 4.5: Cancer datasets used for the experiments

Dataset	No. of Genes	No. of Samples	Source
Breast Cancer	4948	78	VEER1 from van't Veer et al. [304]
Lymphoma	4026	96	Kent Ridge Biological Dataset Repository [258]
CNS	7129	60	Kent Ridge Biological Dataset Repository [258]

Quantile normalization is used to remove systematic variation followed by minimum p-value criterion to select genes above a certain threshold value.

4.7 Results

The experiments were carried out on three datasets, namely on (a) Breast Cancer dataset, (b) Lymphoma dataset and (c) Embryonal Tumours of the Central Nervous System (CNS) dataset. Only the results of the experiments involving semantic similarity, sequence similarity, internal measures, stability measures and biological measures for the breast cancer dataset is shown in this chapter. The results of all the three datasets used in the experiment are given in the *Appendix* of this thesis.

4.7.1 Results of External Validation

Next, the results of the external validation are discussed.

4.7.1.1 The Pair-Wise Gene Expression Similarity Matrix

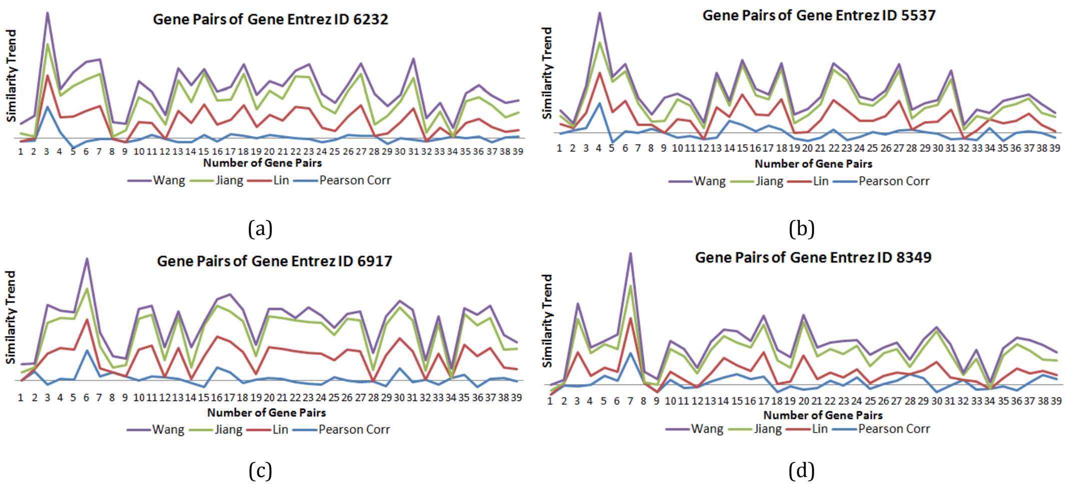
As has been pointed out [268], if two gene sequences are similar, then their genetic expressions are similar and they should be similarly annotated in the GO. To confirm this observation, the pair-wise gene expression similarity needs to be calculated which is done using Pearson Correlation [305] for Breast Cancer dataset, Lymphoma dataset and Embryonal Tumours of the Central Nervous System (CNS) dataset. These expression values will be used subsequently for comparison with the semantic similarity values for the corresponding pair of genes in *Section* 4.7.1.3.

4.7.1.2 The Pair-Wise Semantic Similarity Matrix

It has been shown by Lord [274] that sequence similarity is more tightly correlated with *molecular function* aspect of GO, followed by *cellular component* and then *biological process*. Since the semantic similarity will be compared with sequence similarity, the pair-wise semantic similarity matrix for the Lin, Jiang and Conrath and Wang measures are calculated for the Breast Cancer dataset using the molecular function aspect of GO. The observation has been that the semantic similarity values appear to be correlated for each gene-pair, the reason being that gene sequences with similar genomic expression may also be functionally related [306,307,308]. In all the cases, the average method for calculation of semantic similarity [273] was incorporated.

4.7.1.3 Comparison of Pair-Wise Gene Expression Similarity And Semantic Similarity

A well defined ontology indicates the functional similarity of the terms being represented. In this section the gene expressions of two genes are compared with their corresponding similarity of their annotation in the GO since it is expected that functions that are close in the ontology are expected to be related [268]. Hence it can be assumed the expression correlation may relate to the semantic similarity [267,268,273]. This assumption appears to hold true when the similarity and semantic values are compared for the breast cancer dataset from the graphs for the plot of the values for the genes (Entrez ID 6232, 5547, 6917 and 8349, to name a few). It is noticed that the gene correlation and the semantic similarities Lin, Jiang and Conrath and Wang exhibit a similar graphical trend as shown in *Figure* 4.2 and the underlying relationship between gene correlation and semantic similarity becomes apparent.



(a) (b) (c) (d)

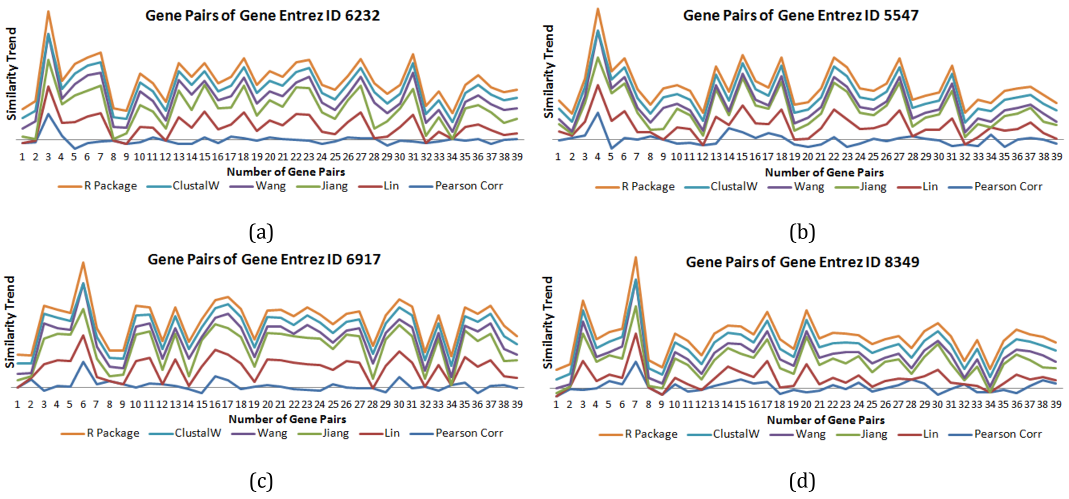
Figure 4.2: Comparison of Gene Expression Similarity and Semantic Similarity for Lin, Jiang and Conrath and Wang measures for the Breast Cancer dataset

4.7.1.4 Pair-Wise Sequence Similarity Matrix

The extent of similarity between the nucleotide/protein sequences of two genes gives the sequence similarity between them. The sequence similarity can be obtained by using an online tool called ClustalW [302,309] that is used to compute the multiple sequence alignment of genes using their nucleotide or protein sequences. The nucleotide sequences of the genes are given as input to ClustalW in FASTA file format. Keeping other parameters like the gap penalty, weight matrix, clustering, etc. to the default values, the progressive multiple sequence alignment of genes is computed pair-wise.

4.7.1.5 Comparison of Pair-Wise Gene Expression Similarity, Semantic Similarity and Sequence Similarity

To validate that the semantic similarity measures used for the calculation were producing appropriate results, they were compared to sequence similarity, since it is expected that highly similar sequences should be highly semantically similar. This expectation stems from the fact that sequence similarity is supposed to be strongly correlated with semantic similarity based on the molecular function aspect of GO [274]. From a biological point of view, the correlation should exist since the sequence of a nucleotide or protein determines its molecular function but does not necessarily relate to the biological process that it is involved in. *Figure* 4.3 presents the pair-wise summary of some of the genes indicating gene expression similarity, semantic similarity for Lin, Jiang and Conrath and Wang measures and sequence similarity scores obtained from ClustalW and *R* Package for the breast cancer dataset. The scores calculated by *R* Package for paired sequence alignment match those obtained from ClustalW. From the graph it can be clearly observed that the pair-wise score for a gene pair shows similar level of scores for semantic and sequence measures indicating correlation between the measures as expected. This common trend is extended to gene expression similarity, semantic similarity and sequence similarity, indicating correlation among them.



(a) (b)

(c) (d)

Figure 4.3: Comparison of Gene Expression Similarity, Semantic Similarity and Sequence Similarity for Lin, Jiang and Conrath and Wang measures for the Breast Cancer dataset

4.7.2 Results of Internal Validation

The internal validity measures obtained for connectivity, Dunn index and silhouette width for the Breast Cancer dataset, using the eight clustering algorithms mentioned earlier, are shown in *Table* 4.6.

It is expected that the connectivity should be minimized, while both the Dunn index and the silhouette width should be maximized and this is depicted in the optimal scores in *Table* 4.6. It is also noticed from the above readings that hierarchical clustering with two clusters performs the best in the case of connectivity and silhouette width and with four clusters in case of Dunn index.

Table 4.6: Scores of Internal Validation Measures for the Breast Cancer dataset

Clustering	**Validation**	**Number of Clusters**				
Algorithm	**Measures**	**2**	**3**	**4**	**5**	**6**
hierarchical	Connectivity	2.929	5.8579	8.7869	11.7159	14.6448
	Dunn	0.7	0.6464	0.779	0.6109	0.6561
	Silhouette	0.3772	0.3482	0.33	0.2283	0.2021
kmeans	Connectivity	5.8579	16.2127	16.546	30.0127	39.7175
	Dunn	0.6464	0.4386	0.4902	0.3318	0.3538
	Silhouette	0.374	0.2763	0.253	0.1119	0.1019
diana	Connectivity	5.8579	5.8579	8.7869	19.7528	19.9194
	Dunn	0.6464	0.6464	0.779	0.5083	0.5083
	Silhouette	0.374	0.3482	0.33	0.1891	0.1882
som	Connectivity	2.929	17.225	42.019	48.2409	59.0298
	Dunn	0.7	0.3534	0.3388	0.2798	0.3083
	Silhouette	0.3772	0.1222	0.0854	0.0121	-0.0197
pam	Connectivity	28.4897	30.6187	33.381	35.3571	37.2361
	Dunn	0.2716	0.2905	0.2995	0.36	0.3837
	Silhouette	0.0902	0.0781	0.0852	0.0706	0.0559
sota	Connectivity	26.4512	35.2464	37.7075	38.2075	41.1448
	Dunn	0.3005	0.2998	0.3341	0.3341	0.3341
	Silhouette	0.0919	0.0873	0.0828	-0.0106	-0.0429
clara	Connectivity	28.4897	30.6187	33.381	35.3571	37.2361
	Dunn	0.2716	0.2905	0.2995	0.36	0.3837
	Silhouette	0.0902	0.0781	0.0852	0.0706	0.0559
model	Connectivity	34.3698	16.2127	16.546	25.0968	25.7079
	Dunn	0.3518	0.4386	0.4902	0.4902	0.5422
	Silhouette	0.1651	0.2763	0.253	0.1901	0.1787

Optimal Scores:

	Score	**Method**	**Clusters**
Connectivity	2.929	hierarchical	2
Dunn	0.779	hierarchical	4
Silhouette	0.3772	hierarchical	2

The plots of the connectivity, Dunn index, and silhouette width are given in *Figure* 4.4 and it appears that hierarchical clustering outperforms the other clustering algorithms under each validation measure, for nearly every number of clusters evaluated, whereas model-based clustering does not perform well on any of the measures. Regardless of the clustering algorithm, the optimal number of clusters seems to be two when utilizing the connectivity and silhouette width. For the Dunn index the best choice for the number of clusters appears to be four.

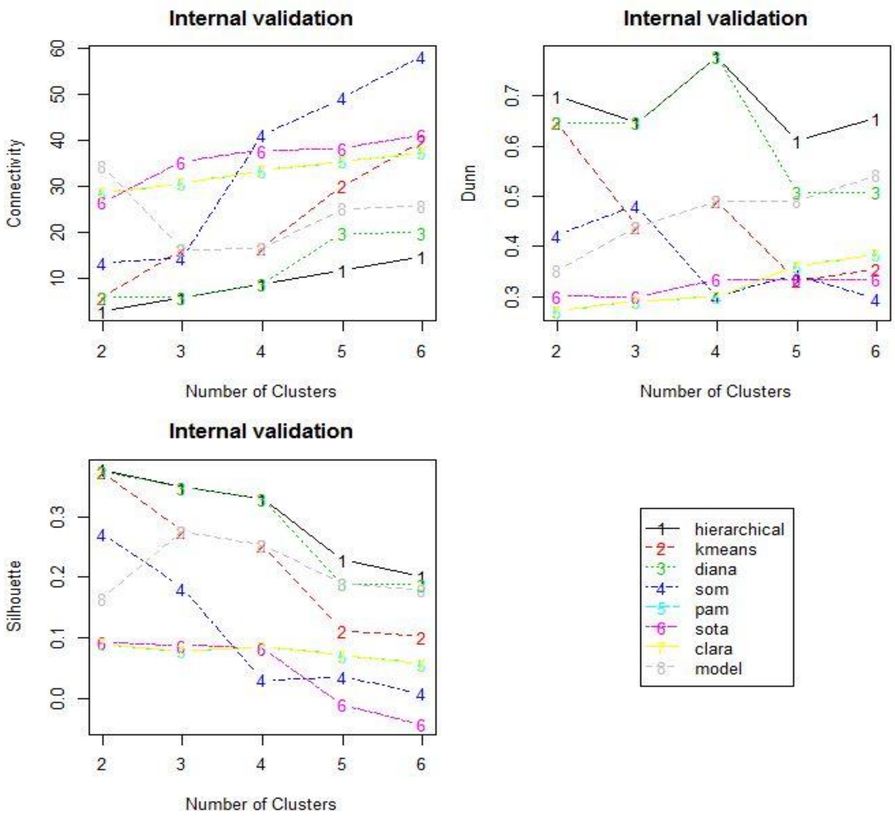


Figure 4.4: The plots of the connectivity, Dunn index, and silhouette width for the Breast Cancer dataset

4.7.3 Results of Stability Measures

The results of APN, AD, ADM and FOM for the Breast Cancer dataset are given in *Table* 4.7.

For the APN and ADM measures, values close to zero are preferred. The optimal scores in *Table* 4.7 shows that hierarchical clustering with four clusters gives the best score, as was also in the case of internal validation. However, for the other two measures model based clustering with six clusters has the best score.

Table 4.7: Scores of Stability Measures for the Breast Cancer dataset

Clustering	**Validation**	**Number of Clusters**				
Algorithm	**Measures**	**2**	**3**	**4**	**5**	**6**
hierarchical	APN	0.0022	0.0095	0	0.0007	0.0029
	AD	2.9282	2.7599	2.5345	2.3969	2.2656
	ADM	0.0142	0.0721	0	0.0039	0.0131
	FOM	0.2535	0.2455	0.2276	0.2245	0.2188
kmeans	APN	0.0062	0.0189	0.0196	0.0188	0.0614
	AD	2.8815	2.7049	2.5194	2.3658	2.2546
	ADM	0.0195	0.1349	0.0772	0.0606	0.2108
	FOM	0.2453	0.2406	0.2297	0.2229	0.2206
diana	APN	0.014	0.0102	0.002	0.0155	0.0074
	AD	2.9087	2.7618	2.5364	2.3801	2.2426
	ADM	0.0821	0.0772	0.0066	0.0882	0.0459
	FOM	0.2468	0.2464	0.2286	0.2235	0.2182
som	APN	0.1159	0.1977	0.2954	0.3578	0.3588
	AD	3.0412	3.0197	2.7384	2.6827	2.5544
	ADM	0.4747	0.7853	0.764	0.9293	0.9193
	FOM	0.248	0.2458	0.2392	0.231	0.2326
pam	APN	0.02	0.0161	0.0204	0.0526	0.1105
	AD	2.9623	2.75	2.5661	2.3923	2.3213
	ADM	0.1316	0.0512	0.0639	0.1034	0.2211
	FOM	0.2572	0.2432	0.2323	0.2228	0.2243
sota	APN	0.1951	0.2074	0.2061	0.1932	0.193
	AD	3.0026	2.9005	2.8139	2.667	2.5647
	ADM	0.4622	0.5552	0.6112	0.6088	0.6153
	FOM	0.257	0.2532	0.2473	0.2388	0.2329
clara	APN	0.02	0.0192	0.0204	0.035	0.072
	AD	2.9623	2.7511	2.5661	2.3873	2.297
	ADM	0.1316	0.057	0.0639	0.0753	0.1456
	FOM	0.2572	0.2435	0.2323	0.2222	0.2227
model	APN	0.0217	0.0085	0.0226	0.0744	0.0776
	AD	2.9038	2.6755	2.5151	2.395	2.2251
	ADM	0.0925	0.0363	0.0606	0.2165	0.1665
	FOM	0.2515	0.2364	0.2301	0.2224	0.2114

Optimal Scores:

	Score	**Method**	**Clusters**
APN	0	hierarchical	4
AD	2.2251	model	6
ADM	0	hierarchical	4
FOM	0.2114	model	6

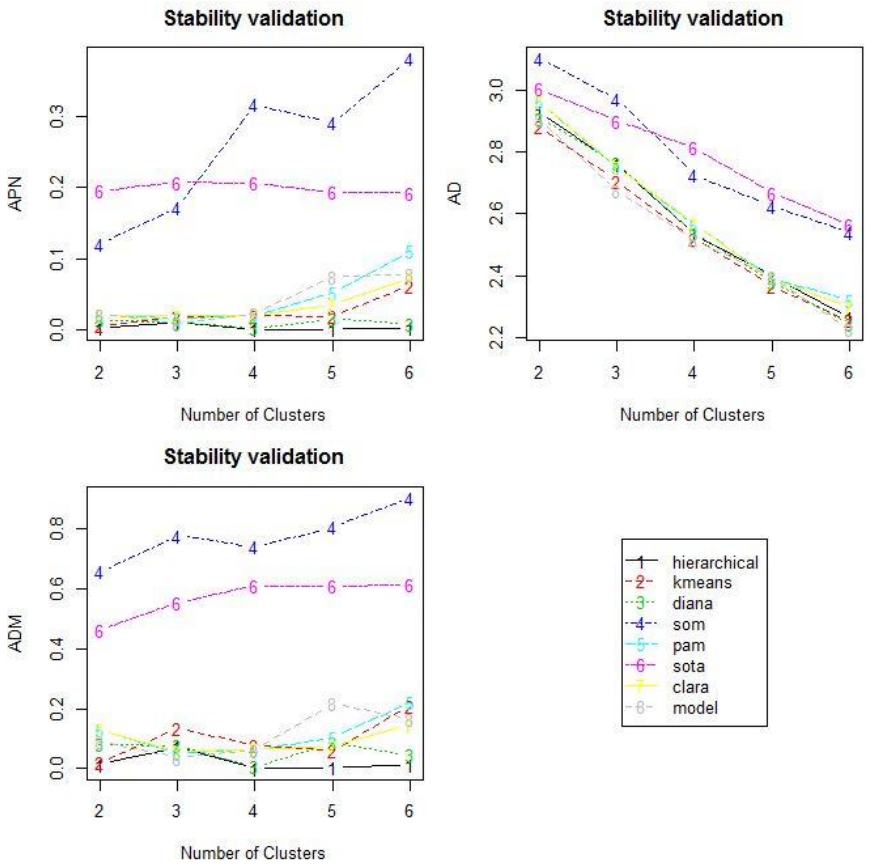


Figure 4.5: The plots of the APN, AD and ADM of stability measures for the Breast Cancer dataset

It is illustrative to graphically visualize each of the validation measures. The plots of the APN, AD, and ADM are given in *Figure* 4.5. The APN measure shows an interesting trend, in that it initially stabilizes from two to four clusters for all the clustering methods except for SOM and SOTA, but marginally increases afterwards. Though hierarchical clustering with four clusters has the best score, Diana with six clusters is a close second. The AD and FOM measures tend to decrease as the number of clusters increases. Here model based clustering with six clusters has the best overall score, though the other algorithms have similar scores. The plot of the FOM measure is very similar to the AD measure, so it has been omitted from the figure. For the ADM measure hierarchical with four clusters again has the best score.

4.7.4 Results of Biological Validation

The BHI and the BSI values were computed for each clustering algorithm in the range of cluster numbers from two to six. *Table* 4.8 shows the scores for the Breast Cancer dataset and it is seen that DIANA has the highest BHI score for six clusters and the highest BSI score is achieved by hierarchical algorithm for two clusters, which indicates that consistency of clustering for genes with similar biological functionality is given by hierarchical algorithm.

Table 4.8: Optimal Scores of BHI and BSI for the Breast Cancer dataset

Algorithm	**Measure**	**Number of Clusters**				
		2	**3**	**4**	**5**	**6**
hierarchical	BHI	0.3172	0.3095	0.3011	0.2966	0.2882
	BSI	0.8253	0.6611	0.5725	0.5329	0.5179
kmeans	BHI	0.4047	0.3639	0.3889	0.3529	0.2879
	BSI	0.6371	0.4925	0.4579	0.3006	0.2511
diana	BHI	0.4047	0.3095	0.3011	0.3861	0.4241
	BSI	0.6574	0.6605	0.5708	0.3361	0.3021
som	BHI	0.3163	0.3156	0.2947	0.2989	0.2361
	BSI	0.7014	0.5614	0.3147	0.2431	0.1773
pam	BHI	0.3052	0.3022	0.2979	0.2853	0.2839
	BSI	0.5904	0.5511	0.4181	0.3831	0.3612
sota	BHI	0.3041	0.2833	0.2125	0.1708	0.1867
	BSI	0.4697	0.4221	0.4131	0.4092	0.3934
clara	BHI	0.3052	0.3022	0.2979	0.2853	0.2839
	BSI	0.5904	0.5501	0.4181	0.3895	0.3725
model	BHI	0.3091	0.3639	0.3889	0.2651	0.2651
	BSI	0.3612	0.4574	0.4493	0.3891	0.3246

Optimal Scores:

	Score	**Method**	**Clusters**
BHI	0.424	diana	6
BSI	0.8253	hierarchical	2

Figure 4.6 shows the plots of BHI for the eight clustering algorithms which reveal that DIANA happens to produce most homogeneous biological clusters based on this dataset and the results are statistically significant when the number of clusters is between four and six.

The plots of BSI are shown in *Figure* 4.7 and hierarchical algorithm seems to be the most stable in its capability of producing clusters using reduced datasets that are biologically alike. Considering both indices, it can be concluded that hierarchical algorithm is the best choice for this dataset to maximize the biological homogeneity and DIANA can be a worthwhile consideration if six clusters are desired.

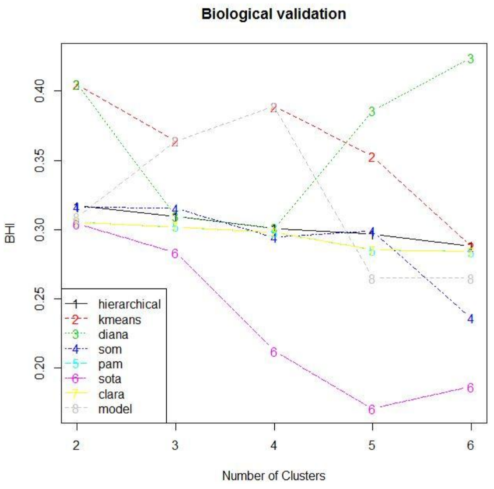


Figure 4.6: BHI plot for Breast Cancer dataset

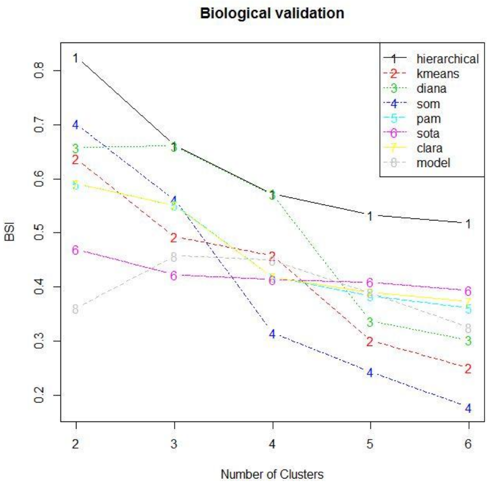


Figure 4.7: BSI plot for Breast Cancer dataset

4.8 Discussion

Existing biological knowledge, such as the GO database, can assist in the cluster validation process of cancer datasets since their analysis can be done with a very high level of accuracy. The clustering results arrived at are validated with the help of GO by using external validity measures such as semantic and sequence similarity measures to ascertain whether the clusters obtained are biologically significant. Internal validity measures such as Dunn index, silhouette width or the homogeneity index are used to evaluate the visual separation of the clusters obtained from a clustering algorithm. Since the dataset is highly correlated, additional stability measures and biological validation measures are used in the form of biological homogeneity index and biological stability index to arrive at a decision whether the clustering results produced by a particular clustering method is biologically significant. In the present scenario, it is noticed that hierarchical algorithm seems to produce the best results.

It is a foregone conclusion that in bioinformatics, clustering gene expression data can reveal biologically relevant information. But it is also all the more important not to depend upon one algorithm with one parameter setting and take the results on face value. Rather, one should compare multiple clusterings with various parameter settings and then take the results as the conveyed structure of the data. Only the comparison of carefully chosen clustering algorithms can result in reliable conclusions drawn from cluster analysis.

The interpretability of the results is highly dependent on the accuracy of the biological annotations being incorporated into the clustering validation. The power of cluster analysis of gene expression data is that it can greatly reduce the search space and thus can lead biologists towards promising presumptions which are worth further biological examination. The verification of these presumptions by biological experiments is not replaceable. The development of biomedical technology will lead to a marked increase in the use of accurate and reliable biological data, which is absolutely essential for statistical analysis such as clustering validation.

In the next chapter, an empirical study of some of the popular protein complex prediction algorithms will be performed with an aim to uncover the limitations and biasness of the algorithms. An ensemble framework for protein complex detection is then proposed where in the initial phase, external information in the form of gene expression data and Gene Ontology will be integrated in the PPI network to purify the network. The base cluster algorithms will generate clusters of protein complexes from the purified PPI data which will become the input to the proposed ensemble for the process of protein complex identification.