Electronic Journal of Biotechnology ISSN: 0717-3458
http://www.ejbiotechnology.info
DOI: 10.2225/vol16-issue5-fulltext-3 *REVIEW ARTICLE*

Evaluation of computer methods for biomarker discovery on computational grids

Gustavo Tonini[1] ✉ · **Frank Siqueira**[1]

1 Federal University of Santa Catarina, Department of Informatics and Statistics, Florianópolis, SC, Brazil

✉ Corresponding authors: gustavotonini@gmail.com; frank@inf.ufsc.br
Received January 9, 2013 / Accepted June 27, 2013
Published online: September 15, 2013

Abstract

Background: Discovering biomarkers is a fundamental step to understand and deal with genetic diseases. Methods using classic Computer Science algorithms have been adapted in order to support processing large biological data sets, aiming to find useful information to understand causing conditions of diseases such as cancer.
Results: This paper describes some promising biomarker discovery methods based on several grid architectures. Each technique has some features that make it more suitable for a particular grid architecture. This matching depends on the parallelizing capabilities of the method and the resource availability in each processing/storage node.
Conclusion: The study described in this paper analyzed the performance of biomarker discovery methods in different grid architectures. We have found some methods are more suited for certain grid architectures, resulting in significant performance improvement and producing more accurate results.

Keywords: feature selection, genetics, parallel computing, pattern detection, performance.

INTRODUCTION

It is clear that the cure or, at least, early detection of diseases such as cancer, Alzheimer and Parkinson depends on human understanding of the genetic causing factors and conditions. Therefore, handling genetic code may result in the production of process-blocking substances (Chen et al. 2003).

Genetic research often involves the extraction and processing of huge amounts of genetics and proteomics data (Kim et al. 2008) searching for chemical patterns that represent biological signatures or biological states, also known as biomarkers.

This paper presents an analysis on the most common biomarker discovery methods currently adopted on grid environments. Section Materials and Databases presents the methods and their execution environment. Grid Types describes the main grid environment architectures and types. Performance Analysis analyzes the adequacy of the existing grid environments for executing biomarker discovery methods and section Concluding Remarks concludes this work.

MATERIALS AND METHODS

This section presents three workflow steps of biological analysis information process. First, the samples are collected from organisms or source materials and “digitalized” in a process called sequencing. This process generates data in a format that a computer understands. The second step consists in storing these data in one or more biological databases where researchers have access to do the third and final step: process data in different ways to extract valuable information.

Extraction methods

Recent advances on DNA extraction technologies, such as DNA-chips, allow researchers to map a genome segment in minutes. After mapping, discovery methods (most of them based on Artificial Intelligence) are commonly used to mine these data searching for biomarkers.

A mass spectrometer is an equipment designed to identify the mass-to-charge of particles. This technique enables fast verification of a substance composition. The equipment returns a sequence of mass measures expressed in real numbers. From this numbers, conversion tools can detect the sample composition.

DNA-chips or microarrays are commonly used to compare two or more DNA sequences and identify discrepancies (Lashkari et al. 1997).

Finally, pirosequencing is a promising method that rebuilds DNA synthesis process and can sequence at higher rates with lower costs (Ronaghi et al. 1998).

Biological databases

A vast amount of DNA data is generated every day. These data must be stored and may be shared across different research centers. To accomplish this, databases such as Genbank (substances composition) (http://www.ncbi.nlm.nih.gov/genbank/), BioCyc (biological pathways) (http://biocyc.org/) and Pubmed (medical references) (http://www.ncbi.nlm.nih.gov/pubmed/) allow researchers to submit analyzed datasets.

Most of the biomarker discovery methods query these databases looking for similar patterns and references in medical literature.

These databases are cross-referenced and fed by several sequencers and research staff around the world.

Experimental databases such as Stanford Microarray Database (http://smd.stanford.edu/) and Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) permit researchers without sequencing equipment to obtain samples to study. Stanford Microarray Database contains data from almost 80 thousand experiments. Raw data is published, including previous analysis, if authorized by the patient. It has strong integration with Pubmed and NCI databases.

Discovery methods

Biomarker discovery methods perform pattern search algorithms against biological databases and compare results with data obtained from patients.

A grid architecture can be customized to execute a discovery method, aiming to obtain better performance.

The main method profiles are discussed in the following subsections.

Sequence alignment. The natural way when dealing with DNA data (a sequence of base pair expressed with the C. T. G. A. characters) is trying to match two or more samples and identify differences that indicate a biological condition.

The same concept can be used with proteomic and RNA data. Matching sequence segments can be useful to forensic analysis and studying the evolution of species. Any relevant segment match found can indicate correlations between the state of source and target organisms. This can be used to prove ancestral relationship or identify which mutations happened across generations (useful when trying to identify gene functions).

The main objective of an alignment method is to find the best fit of two or more sequence segments to maximize the matching area. The alignment can be global (trying to cover the whole segment) or local (trying to find specific regions of similarity). To achieve these goals, the algorithm can insert or remove gaps inside the sequences (Hong and Tewfik, 2009). The procedure of inserting/removing gaps is performed in the pre-processing phase.

In this category of methods, BLAST (Basic Local Alignment Search Tool) (Korf et al. 2003) is the best known and effective. A parallel implementation of BLAST (http://www.mpiblast.org/) is available and some experiments demonstrate a sensible performance enhancement in grid environment (Trombetti et al. 2006).

Several types of algorithms can be applied to produce the alignment, such as dynamic programming, string-based alignment, dot-matrix and scoring functions. The efficiency and accuracy of each method depends on the on the characteristics of the source and target sequences (Axelson-Fisk, 2010).

Artificial Intelligence (AI) and non-deterministic methods. Deterministic methods are good when the exact result pattern is known *a priori*. Biomarkers in a genome, for example, can be spread across thousands of random gene positions. This type of method cannot deal efficiently with this scenario because it needs to verify all possible combinations to make any general assumptions (Syarifahadilah et al. 2012).

In this case, an evolutionary algorithm or an auto-learning technique will demand less computer operations to produce results. IA methods try to simulate human intelligence and learning process (Haykin, 2001).

Neural networks (NN) can learn to distinguish patterns depending on input and the frequency of the element occurrences (Azuaje, 2001). The big advantage of using NN is scalability adherence. When more processors are added, the network can be extended adding more decision nodes.

A cluster of GPUs (graphic processing units) was employed by Johnson et al. (2012) reducing in more than 96% the computing time. This case demonstrates a good cost/efficiency ratio.

Support vector machines (SVM) are commonly applied in problems involving mass spectrometry and microarray data (Pang et al. 2006). The objective of this method is to classify one input in a set of categories. Initially the machine is trained with a set of elements associated with categories. After that, the machine works by itself and improves its own knowledge on each hit. SVMs are present in the latest biomarkers discoveries in genomic and proteomic data.

SVM can help in the decision of problems. For example: deciding if a mutation represents a disease biomarker or trying to diagnose which type of cancer one DNA cell sample represents (Figure 1 shows a SVM model that addresses this problem. Each white-leaf node represents cancer type A, while filled-leaf nodes represent cancer type B. Non leaf nodes represent decision points. Numbers inside leaf nodes indicate the number of patients diagnosed with this type of cancer).

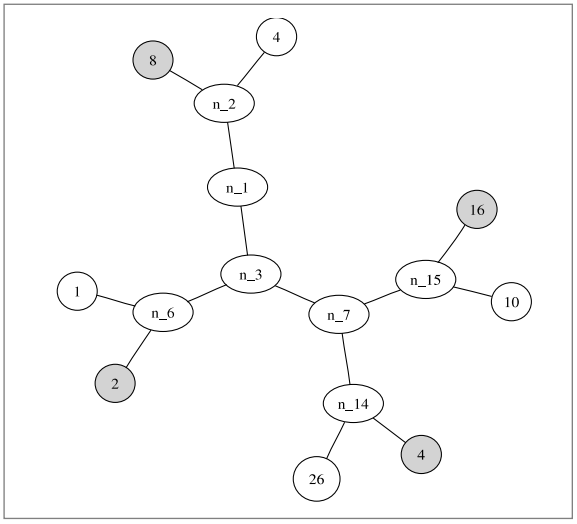

Fig. 1 Example of SVM used on cancer type diagnostics (Pang et al. 2006).

Genetic algorithms (GA) are a good choice to avoid traditional computing in some high complexity problems. The technique tries to simulate evolutionary biology operation, recombining characteristics across generations of species (Goldberg, 1989).

Applications demanding recombination of behaviours or characteristics are suitable for using genetic algorithms, despite the lower parallelization capabilities of this method. Paul and Iba used GA to verify which genes are associated to different types of cancer, based on frequency of appearance (Paul and Iba, 2009). Moreover, the most recent researche with microarray data using GAs showed promising results (Scheubert et al. 2012).

Boolean networks are structures that simulate the possible ways to satisfy a condition. Each network node represents one condition. Depending on wether the condition is satisfied or not, network flow takes one way or another.

Applying this concept on genetics, Boolean networks simulate evolutionary conditions or reactions that cause a situation (causing factors of cancer, for example). It is possible to simulate which specific mutations cause deregulation on the cell division process.

Boolean network traversal algorithms can be subdivided (enabling parallelization). Each process can take a set of ways in the network, since there are not too many cycles.

Statistics. Correlation of a set of variables is a field of statistics that can be applied on pattern discovery problems. Some methods such as Monte Carlo and Single-Value Decomposition are commonly used in biomarker discovery (Lai et al. 2009).

Before applying a statistical method, data must be normalized and converted to the expected base scale (generally a sequence of real numbers) by the method. For this reason, the data obtained from mass spectrometry is most commonly processed using statistical methods.

The basic operation of this type of method is to process samples marking segment group with repetitions trying to associate their occurrence count with some relevant variable, using probability models (Re and Valentini, 2012).

Another strategy typically present in statistical methods is to assign a calculated factor to each sample and verify the pattern match simply by comparing these factors. Samples are considered more similar on how close these factors are to each other (Lazar et al. 2012).

Limpiti et al. (2012) made a frequency analysis on proteomic mass spectrometry data combining several statistical feature selection approaches, obtaining an accurate cancer detection method. A similar approach was used in Cheng et al. (2011) analyzing HBV virus DNA sequences.

Hybrid methods. Different methods can be applied to the same samples to improve pattern matching probability. This is the objective of using a hybrid multiagent approach.

Qabaja et al. (2009) used three different statistical methods to enhance the accuracy of results of cancer-related biomarkers.

Most researchers try to present a verification/validation approach to their research. Using hybrid methods, they can prove the same theory using two or more methods, adding reliability to results. Moreover, if a biomarker is found using two different methods, for example, researches can guide next works focusing on the locus where the first was found.

GRYD TYPES

A computing grid is a collection of computer resources working together to achieve a given goal (Colouris et al. 2005). Depending on the budget, resources and technology, several architectures can be modeled to form infinite possible topologies of grids.

This section discusses the commonly used approaches for building computing grids.

Enterprise grid

In the early 2000's, organizations of different sizes have noticed the potential of grid computing. Then, consortiums of companies, institutions and countries were formed to join efforts and resources to create this type of infrastructure.

CaGrid (http://www.cagrid.org/) is an example of initiative founded in North America for helping in cancer research. The same proposal was implemented in Europe with EGI (European Grid Infrastructure) (http://www.egi.eu/).

Some grids of this type are open to any research groups aiming to develop and execute analysis, depending on the project relevance.

EGI, for example, contains about 330 processing nodes across 57 countries in Europe. 270 thousand processor cores constitute the grid computing power and the total storage capacity is over 139 petabytes. Anyone can submit applications using specific API (Application Program Interface). The EGI Committee chooses which applications will a have access to more resources.

Enterprise grid nodes typically communicate through public national research networks (*e.g.,* Internet 2 in the United States and CLARA in Latin America). In some cases, dedicated links are allocated to connect some nodes, depending on project budget, needs and relevance.

Collaborative computing

Problems demanding massive computer power, such as prime numbers decomposition or proteome folding, take years to execute in a single enterprise or private grid.

Initiatives such as BOINC (Berkeley Open Infrastructure for Network Computing) (http://boinc.berkeley.edu/) intend to join efforts of thousands of small personal computer systems in order to solve problems with great scalability.

Grid nodes have limited resources and network bandwidth available. For this reason, applications must have full scalability capabilities to obtain good performance.

Anyone can download the BOINC software and start collaborating, joining local processing and storage resources to form a very large and open grid. The software downloads tasks from the server, executes them and sends back the obtained results.

The same way, anyone can develop applications using the BOINC API and submit them to the application database. Then, users can choose wich applications they want collaborate with.

Some notable projects running on BOINC are:

-Riesel-Sieve: prime numbers discovery.

-Einsten@Home: find neutron stars and gravitational waves.

-Primegrid: prime numbers decomposition aimed at verifying and improving RSA security algorithm.

-Folding@Home: tries to understand the 3D model of proteins and their relationship with diseases.

-World Community Grid: project related to understanding diseases such as cancer and AIDS.

PERFOMANCE ANALYSIS

Biotechnology problems have quadratic complexity or greater (Karlin and Altschul, 1993). Biological databases grow exponentially.

Collaborative computing grids solve very big and time expensive problems, wich are typically impossible to solve on enterprise grids with low costs. However, there is no performance guarantee, quality of service or end time prevision. These variables depend on node availability.

For example, BOINC has 535 teraflops (Table 1) available, but there is no compromise on when this rate can be reached.

Table 1. BOINC Average processing capacity by processor architecture (Anderson and Fedak, 2006).

CPU Type	Number of hosts	GFLOPS per host	GFLOPS total
Intel	217,278	1.600	347,645
AMD	95,958	1.737	166,679
PowerPC	15,827	1.149	18,185
SPARC	1,035	0.755	781
Others	1,687	1.233	2,080
Total	331,785	1.613	535,169

Enterprise grids can concentrate resources on a specific goal. They have better throughput and a good prevision of task ending time. However, public enterprise grids such as EGI run several applications concurrently. This behaviour can affect specific parts of methods. Depending on algorithm steps precedence, it is better to execute the fundamental steps in periods of low concurrency.

EGI permits to trigger a number of parallel instances of the same application. To obtain performance improvements with this feature, users must verify current grid load to avoid that their processes stay in the queue waiting for resources to become available.

Biomarker detection methods based on AI, such as SVD, have lower parallelizing capabilities because a step depends on results from the previous one. Any recursive algorithm generates vertical parallelization problems; this means that individual tasks cannot be divided in more processes to run on different processors.

Unlike recursive methods, alignment algorithms, such as BLAST can be divided in different ways, as can be seen in Trombetti et al. (2006). One experiment got 19 times of processing time improvement. The same algorithm was modified, using MPI, by Battré and Angulo (2006) resulting in a significant scalability improvement (Figure 2).

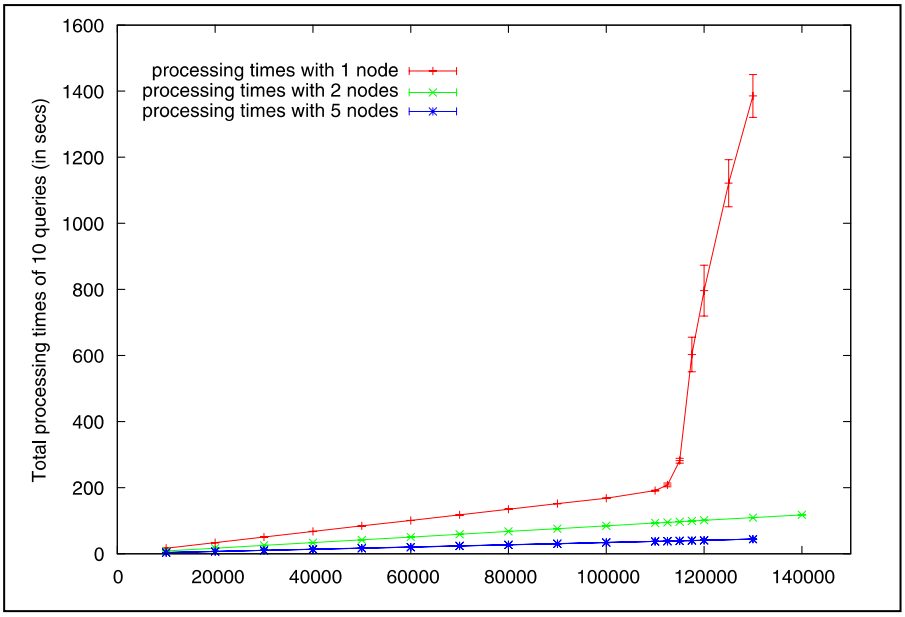

Fig. 2 Processing times using MPI BLAST (Battré and Angulo, 2006).

CONCLUDING REMARKS

In the last decade, a large amount of effort has been put on expanding the capacity of computational grids. Current efforts are concentrated on joining structures from companies and countries, culminating in bigger regional grids.

There is also more and more biological information available to process and make correlations. Network, processing and storage infrastructures are available to perform these jobs. National networks used typically in public/enterprise grids have transfer rates of gigabits per second.

On the other hand, volunteer computing has emerged as a valuable alternative to long tasks, since the problem has a good scalability due to the limitation of node resources.

The perfomance of biomarker discovery methods can be improved if the correct grid type is employed. Project goals are often not achieved if the performance requirements are not satisfied. So, the wrong choice in terms of grid architecture can affect the whole project.

DOI: 10.2225/vol16-issue5-fulltext-3

It is necessary to know the grid features before start writing applications to run on such environment. Factors like concurrency, average load, periods of overload and seasonal demands usually degrade throughput and can make some tasks unable to finish before deadline.

Anyone engaged on studying or developing and proving one promising pattern theory may gain access to the infrastructure and execute his jobs, provided that the purpose is relevant to the scientific community.

The complexity of genetic combinations, mutations and pathways is still far away from human understanding. It will take years to "digitalize" necessary biological information and perform metabolic simulations closer to reality. However, we already use and study information of specific "locus" and obtain important results solving interesting problems, especially about genetic diseases such as cancer.

REFERENCES

ANDERSON, D.P. and FEDAK, G. (2006). The Computational and Storage Potential of Volunteer Computing. In: *Proceedings of the Sixth IEEE International Symposium on Cluster Computing and the Grid*. (16th-19th May, 2006, Singapore). IEEE Computer Society. TURNER, S.T.; LEE, B.S. and CAI, W. eds. p. 73-80. **[CrossRef]**

AXELSON-FISK, M. (2010). Sequence alignment. In: *Comparative Gene Finding*. Springer, London, chapter 3, p. 89-155.

AZUAJE, F. (2001). A computational neural approach to support the discovery of gene function and classes of cancer. *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 3, p. 332-339. **[CrossRef]**

BATTRÉ, D. and ANGULO, D.S. (2006). MPI framework for parallel searching in large biological databases. *Journal of Parallel and Distributed Computing*, vol. 66, no. 12, p. 1503-1511. **[CrossRef]**

CHEN, S.N.; TSAI, J.J.P. and CHEN, W.H. (2003). An intelligent agent-based biomedical literature mining system for cancer-related genes. In: *Proceedings of the Fifth International Symposium on Multimedia Software Engineering*. (10th-12th December, 2003, Taichung, Taiwan). IEEE Computer Society, p. 279-286. **[CrossRef]**

CHENG, C.P.; LEE, P.F.; CHIN, C.Y.; LIU, W.C.; WU, I.C.; CHANG, T.T. and TSENG, V.S. (2011). Discovering clinical biomarkers of chronic hepatitis B by mining mutation hotspots, In*: International Conference on Technologies and Applications of Artificial Intelligence (TAAI*). (11th-13th November, 2011, Chung-Li, Taiwan). p. 51-56. **[CrossRef]**

COLOURIS, G.; DOLLIMORE, J.; KINDBERG, T. and BLAIR, G. (2005). *Distributed Systems: Concepts and Design,* 4th ed. Addison-Wesley, USA. 944 p. ISBN 0321263545.

GOLDBERG, D.E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, USA. 432 p. ISBN 0201157675.

HAYKIN, S. (2001). *Redes neurais: Princípios e Prática.* Porto Alegre, RS, Bookman. 906 p. ISBN 0-13-273350-1.

HONG, C. and TEWFIK, A.H. (2009). Heuristic reusable dynamic programming: Efficient updates of local sequence alignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 4, p. 570-582. **[CrossRef]**

JOHNSON, D.; SHAFER, B.; LEE, J.J. and CHEN, J.Y. (2012). Multi-biomarker panel selection on a GPU. In: *IEEE International Conference on Electro/Information Technology (EIT)*. (6th-8th May, 2012, Indianapolis, Indiana, USA). p. 1-6. **[CrossRef]**

KARLIN, S. and ALTSCHUL, S.F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Sciences of the United States of America,* vol. 90, no. 12, p. 5873-5877. **[CrossRef]**

KIM, K.S.; PIERCE, M.E. and GUHA, R. (2008). SQMD: Architecture for scalable, distributed database system built on virtual private servers. In: *IEEE Fourth International Conference on eScience*. (7th-12th December, 2008, Indianapolis, Indiana, USA). eScience 2008. p. 658-665. **[CrossRef]**

KORF, I.; YANDELL, M. and BEDELL, J. (2003). *Blast*. O'Reilly & Associates, Inc. 362 p. ISBN 0596002998.

LAI, H.; HAN, B.; ZHU, L.; CHEN, Y.; LI, L. and SUTPHEN, R. (2009). Selecting biomarkers for ovarian cancer detection using SVD and Monte Carlo methods. In: *3rd International Conference on Bioinformatics and Biomedical Engineering*. (11th-16th June, 2009, Beijing, China). p. 1-4. **[CrossRef]**

LASHKARI, D.A.; DERISI, J.L.; MCCUSKER, J.H.; NAMATH, A.F.; GENTILE, C.; HWANG, S.Y.; BROWN, P.O. and DAVIS, R.W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America,* vol. 94, no. 24, p. 13057-13062. **[CrossRef]**

LAZAR, C.; TAMINAU, J.; MEGANCK, S.; STEENHOFF, D.; COLETTA, A.; MOLTER, C.; DE SCHAETZEN, V.; DUQUE, R.; BERSINI, H. and NOWÉ, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. In: *Transactions on IEEE/ACM Computational Biology and Bioinformatics*, vol. 9, no. 4, p. 1106-1119. **[CrossRef]**

LIMPITI, T.; ASSAWAMAKIN, A.; INTARAPANICH, A. and TONGSIMA, S. (2012). Time-frequency analysis for cancer detection using proteomic MS-spectra, In: *Fourth Biomedical Engineering International Conference (BMEiCON),* (29th-31st January, 2012, Chiang Mai, Thailand). p. 40-44. **[CrossRef]**

PANG, S.; HAVUKKALA, I. and KASABOV, N. (2006). Two-class SVM trees (2-SVMT) for biomarker data analysis. In: WANG, J.; YI, Z.; ZURADA, J.M.; LU, B.L. and YIN, H. eds. *Advances in Neural Networks - Lecture Notes in Computer Science*. Springer, Chengdu, China, vol. 3973, p. 629-634. **[CrossRef]**

PAUL, T.K. and IBA, H. (2009). Prediction of cancer class with majority voting genetic programming classifier using gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 2, p. 353-367. **[CrossRef]**

QABAJA, A.; ALSHALALFA, M.; ALHAJJ, R. and ROKNE, J.G. (2009). Multiagent approach for identifying cancer biomarkers. In: *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (1st-4th November, 2009, Washington, DC, USA). p. 228-233. **[CrossRef]**

RE, M. and VALENTINI, G. (2012). Cancer module genes ranking using kernelized score functions. *BMC Bioinformatics*, vol. 13, no. 14, p. S3.

RONAGHI, M.; UHLÉN, M. and NYRÉN, P. (1998). A sequencing method based on real-time pyrophosphate. *Science,* vol. 281, no. 5375, p. 363-365. **[CrossRef]**

SCHEUBERT, L.; LUŠTREK, M.; SCHMIDT, R.; REPSILBER, D. and FUELLEN, G. (2012). Tissue-based Alzheimer gene expression markers-comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets. *BMC Bioinformatics*, vol. 13, no. 1, p. 266. **[CrossRef]**

SYARIFAHADILAH, M.Y.; ABDULLAH, R. and VENKAT, I. (2012). ABC algorithm as feature selection for biomarker discovery in mass spectrometry analysis, In: *4th Conference on Data Mining and Optimization (DMO)*, (2nd-4th September, 2012, Langkawi, Malaysia). p. 67-72. **[CrossRef]**

TROMBETTI, G.A.; MERELLI, I. and MILANESI, L. (2006). High performance cDNA sequence analysis using grid technology. *Journal of Parallel and Distributed Computing,* vol. 66, no. 12, p. 1482-1488. **[CrossRef]**

How to reference this article:

TONINI, G. and SIQUEIRA, F. (2013). Evaluation of computer methods for biomarker discovery on computational grids. *Electronic Journal of Biotechnology*, vol. 16, no. 5. http://dx.doi.org/10.2225/vol16-issue5-fulltext-3