RESEARCH ARTICLE

Cancer Diagnosis Epigenomics Scientific Workflow Scheduling in the Cloud Computing Environment Using an Improved PSO Algorithm

Sadhasivam N*, Balamurugan R, Pandi M

Abstract

Objective: Epigenetic modifications involving DNA methylation and histone statud are responsible for the stable maintenance of cellular phenotypes. Abnormalities may be causally involved in cancer development and therefore could have diagnostic potential. The field of epigenomics refers to all epigenetic modifications implicated in control of gene expression, with a focus on better understanding of human biology in both normal and pathological states. Epigenomics scientific workflow is essentially a data processing pipeline to automate the execution of various genome sequencing operations or tasks. Cloud platform is a popular computing platform for deploying large scale epigenomics scientific workflow. Its dynamic environment provides various resources to scientific users on a pay-per-use billing model. Scheduling epigenomics scientific workflow tasks is a complicated problem in cloud platform. We here focused on application of an improved particle swam optimization (IPSO) algorithm for this purpose. **Methods:** The IPSO algorithm was applied to find suitable resources and allocate epigenomics tasks so that the total cost was minimized for detection of epigenetic abnormalities of potential application for cancer diagnosis. **Result:** The results showed that IPSO based task to resource mapping reduced total cost by 6.83 percent as compared to the traditional PSO algorithm. **Conclusion:** The results for various cancer diagnosis tasks showed that IPSO based task to resource mapping can achieve better costs when compared to PSO based mapping for epigenomics scientific application workflow.

Keywords: Cancer diagnosis- genomics- gene expression-particle swam optimization- scientific workflow- scheduling

Asian Pac J Cancer Prev, **19 (1)**, 243-246 Submmission Date:10/21/2017 Acceptance Date:12/01/2017

Introduction

Genomics is defined as the study of entire genomes of organisms, including extra-chromosomal DNA such as the mitochondrial genetic material. This field includes intensive efforts to determine the entire DNA sequence of organisms, using fine-scale genetic mapping and DNA sequencing with current and emerging technologies. In contrast, investigating the roles of single genes is a primary focus of genetics. Single gene research does not fall into the definition of genomics unless the aim is to verify the effect that a gene may have on the entire genome's networks and pathways. Genomics has been the main focus in molecular biology, especially after the completion of the sequencing of genomes from several organisms. Genomics tools have already helped in the understanding of several aspects of the genome of cancer cells when compared to normal controls. One important example is the identification of the gene HER2/neu (ErbB-2), which is an oncogene mapped to human chromosome that is over-expressed, or amplified, in ~30% of breast cancer tumors (Slamon et al., 1989). Identification of this molecular characteristic culminated in the development of the drug trastuzumab (Herceptin®) (Paik et al., 2008). Breast cancer patients that are HER2/neu (ErbB-2) positive have increased survival rates when treated with this drug.

Recently, cancer genomes were sequenced and compared with normal cells for leukemia, breast, lung, and other tumor types, using second-generation DNA sequencing technologies (Ley et al., 2008; Stephens et al., 2009; Lee et al., 2010; Pleasance et al., 2010). The purpose was to identify mutations that could give rise to new biomarkers and new therapies for these types of cancers. In addition, the 1,000 Genome Project was recently launched (Butler et al., 2010) with the objective of sequencing the genome of thousands of individuals in a small period of time. In parallel, companies are starting to provide whole genome sequencing services, with the aim of understanding the individual's susceptibility for diseases, including different cancers (Kaye, 2008; Pandi and Premalatha, 2015). Medical science workflow applications consists of large number of tasks which are either complex or simple, also there exists a huge amount of data transfer between the tasks. The tasks of the workflow applications are interdependent and are normally represented using directed acyclic graphs

*Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Erode, India. *For Correspondence: sadhasivamn82@gmail.com*

(DAG). Cloud platform is currently considered as the cost-effective distributed computing platform for scientific workflow applications (Hoffa et al., 2008; Sadhasivam and Thangaraj, 2017). Currently, most of the cloud service providers provide the resources to the customers using different billing models and charge the usage either on hourly basis or on minute basis. So scheduling of medical science workflow applications requires the user to properly select the resources (Sahni and Vidyarthi, 2015). However, since there are large number of resources which are dynamic so the workflow application scheduling to cloud resources can be efficiently addressed by using meta-heuristics. The medical workflow application Epigenomics which is used at USC Epigenome center for analyzing the genome sequencing of the human body. It contains sequence of automated operations to collect short DNA segments using high-throughput gene sequencing machines and it uses MAQ software and IlluminiaSolexa genetic analyzer (Gil et al., 2007; Deelman et al., 2015). The relationship between the tasks of the Epigenomic workflow application is modelled using eight levels, with each level containing tasks that perform certain functions. Figure 1 shows the workflow structure of the Epigenomics workflow application. This application is a data intensive workflow application and contains both large and small size tasks.

Materials and Methods

Problem Formulation

The problem can be formulated to identify a task to resource mapping instance P, such that when calculating the total cost incurred using each compute resource PC, the high cost among all the computer resources is minimized. The primary objective of this work is to derive the metaheuristic optimization for mapping the Epigenomics tasks to the compute resources such that the total cost is minimized.

$$C_{total}(P)_j = C_{exe}(P)_j + C_{tx}(P)_j \quad (1)$$

$$Cost(P)_j = Max(C_{total}(P)_j) \quad (2)$$

$$Minimize\ (Cost(P)_j) \quad (3)$$

Where

C_{exe} = The total execution cost for Epigenomics tasks
C_{tx} = The communication cost between the resources

Epigenomics workflow scheduling for Cancer Diagnosis using Particle Swam Optimization

Particle swarm optimization (PSO) is a population based optimization technique inspired by social behavior of bird flocking (Kennedy and Eberhart, 1995; Balamurugan et al., 2017). It combines self-experiences with social experiences and uses a number of particles that represent a swarm moving around in the search space looking for the best solution. In PSO, each single solution is a "particle" in the search space. All of particles have fitness values which are evaluated by the fitness function to be optimized and velocities which direct the moving of the particles. PSO is initialized with a group of random particles. In every iteration, each particle is updated by the two "best" values. After finding the two best values, the particle updates by velocity and position equations.

Particle velocity and position renewal of PSO

$$V'_{id} = V_{id} + C_1 rand1()(P_{idb} - X_{id}) + C_2 rand2()(P_{gdb} - X_{id}) \quad (4)$$

$$X'_{id} = X_{id} + V'_{id} \quad (5)$$

Particle Swarm Optimization Steps

1. Initialize the particles with random solutions

2. Evaluate the fitness value of each particle's according to the objective function

3. If the current fitness value is better than the previous Pbest , Set current fitness value as new Pbest

4. Otherwise keep Previous Pbest

5. Choose the global best particle (best particle of all pbest particles)

6. Calculate particles' velocities according equation (4)

7. Update particle's new positions according equation (5)

Repeat from step 2 until stopping criteria are satisfied.

Improved PSO for Epigenomics scientific workflow application scheduling in Cancer Diagnosis

In the standard PSO algorithm, the convergence speed of particles is fast, but the adjustments of cognition component and social component make particles search around entire solution. So, the whole swarm will be trapped into a local optimum; and the capacity of swarm jump out of a local optimum is rather weak. In order to avoid being trapped into a local optimum, the new PSO adopts a new information sharing mechanism. The proposed method one can not only record the best positions an individual particle and the whole swarm have experienced, one can also record the worst positions.

Particle velocity and position renewal of EPSO

$$V'_{id} = V_{id} + C_1 rand1()(P_{idb} - X_{id}) + C_2 rand2()(P_{gdb} - X_{id}) \quad (6)$$

$$V'_{id} = V_{id} + C_1 rand1()(X_{id} - P_{idw}) + C_2 rand2()(-X_{id} - P_{gdw}) \quad (7)$$

$$X'_{id} = X_{id} + V'_{id} \quad (8)$$

Results

The result is experimented and analyzed with the cloudsim and it consists of 10 resources with different processing speed from Amazon EC2 services. The test has been conducted for the task scheduling problem from 10 processors with 50 tasks. The experimental parameter settings of PSO and IPSO algorithms are shown in Table 1.

Figures 3 plots the computation cost computed by PSO and IPSO over the 50 number of iterations for different sizes of cancer diagnosis Epigenomics workflow applications such as Eigenomics_24, Eigenomics_46,

Table 1. Parameters and Its Value for PSO and IPSO

Parameter description	Parameter value
Size of Swarm	50
Self-recognition coefficient c1	2
2 Social coefficient c2	2
Weight(w)	0.9
Iterations	50

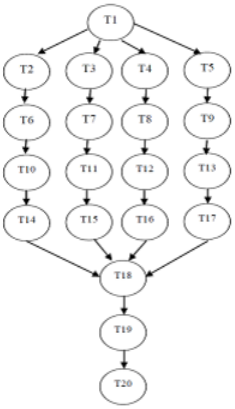

Figure 1. Sample Structure for Epigenomics Scientific Workflow

Eigenomics_100 and Eigenomics_997 applications respectively. Initially, the particles are randomly initialized. Therefore, the initial total cost is always high. This initial cost corresponds to the 0th iteration. As the algorithm progresses, the convergence is drastic and it finds a global minima very quickly. The average number of iterations needed for the convergence is seen to be 30-35, for this application environment. It displays that IPSO usually had better average completion time values than PSO.

Comparative Analysis of PSO and IPSO

Table 2 plots comparison of optimal total cost between PSO based resource selection and IPSO algorithms when varying total data size of a workflow. IPSO achieves 10.39 percentages of improvements for Epigenomics_24 application with 24 tasks processed than the PSO algorithm. For Epigenomics_46 application with 46 tasks and Epigenomics_100 application with 100

```
For each particle
  Initialize particle
End
Do
  For each particle
    Calculate fitness value
    If the fitness value is better than the best fitness value (Pidb) in history
      set current value as the new Pidb
  End
  Choose the particle with the best fitness value of all the particles as the
Pgdb
  For each particle
    Calculate particle velocity according equation (6) (7)
    Update particle position according equation (8)
  End
```

Figure 2. Pseudocode of IPSO Algorithm

Table 2.Comparison of Computation Cost with Various Data Size for PSO and IPSO

Size of Data	PSO	IPSO	Percentage of Improvement
Epigenomics_24	31.19	28.01	10.19%
Epigenomics_46	31.32	28.22	9.89%
Epigenomics_100	33.03	30.96	6.26%
Epigenomics_997	58.7	54.69	6.83%

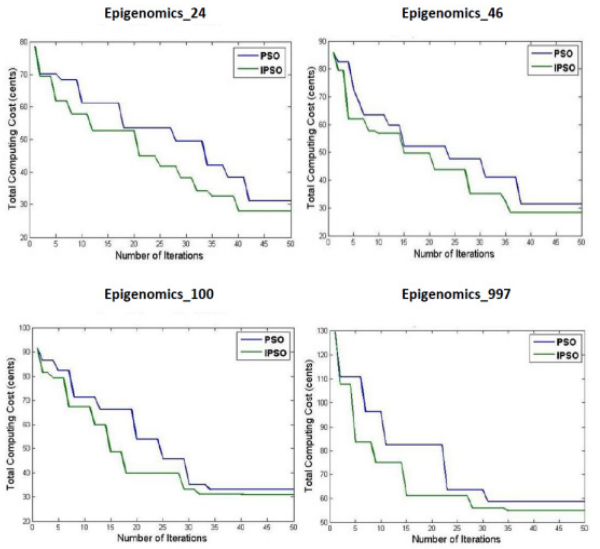

Figure 3. Cancer Diagnosis Epigenomics Workflow Applications Computation Cost of PSO and IPSO

tasks, the proposed IPSO method attains 9.89 and 6.26 percentage of improvements respectively. Whereas for Epigenomics_997 application with 997 tasks the proposed IPSO method returns 6.83 percentage of improvements in optimal total computation cost. Clearly, IPSO based mapping has much lower cost as compared to that of the existing PSO based mapping.

Discussion

The burgeoning fields of genomics and epigenomics comprise essential facets of modern cancer research. A single genes and groups of genes from the same pathway have been identified as differentially methylated in cancers, and some have been used as molecular biomarkers in order to identify patients with a better or a worse prognosis .Based on the information, growing evidence indicates that new epigenomic tools will increasingly affect the way to monitor and manage cancer in the future. The Epigenomics scientific workflow application is essentially a data processing pipeline to automate the execution of the various genome sequencing operations or tasks. Cloud platform is a popular distributed computing platform for deploying large scale of Epigenomics scientific workflow applications. Scheduling Epigenomics scientific workflow tasks are a complicated problem in cloud platform. In the proposed method IPSO is achieved better results compared with existing methods. The performance of

IPSO is analyzed with different cancer diagnosis tasks.

Statement conflict of Interest

The authors whose names are listed immediately below certify that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Author names:

1. Sadhasivam N
2. Balamurugan R
3. Pandi M

References

Balamurugan R, Natarajan AM, Premalatha K (2017). Cuckoo search with mutation for biclustering of microarray gene expression data. *Int Arab J Inf Technol*, **14**, 300-6.

Beroukhim R, Mermel CH, Porter D (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899-905.

Butler D (2010). Human genome at ten: science after the sequence. *Nature*, **465**, 1000-1.

Dalgliesh GL, Furge K, Greenman C (2010). Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature*, **463**, 360-3.

Deelman E, Vahi K, Juve G, et al (2015). A workflow management system for science automation. *Future Gen Comp Sys*, **46**, 17-35.

Gil Y, Deelman E, Ellisman M, et al (2007). Examining the challenges of scientific workflows. *IEEE Comput*, **40**, 24–32.

Hoffa C, Mehta G, Freeman T, et al (2008). On the use of cloud computing for scientific workflows. IEEE Fourth Int Conf in eScience, pp 640-5.

Kaye J (2008). The regulation of direct-to-consumer genetic tests. *Hum Mol Genet*, **17**, 180–3.

Kennedy J, Eberhart R (1995). Particle swarm optimization. Proceedings of IEEE Int Conf on Neural Networks. pp 1942-8.

Lee W, Jiang Z, Liu J (2010). The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature*, **465**, 473–7.

Ley TJ, Mardis ER, Ding L (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, **456**, 66–72.

Paik S, Kim C, Wolmark N (2008). HER2 status and benefit from adjuvant trastuzumab in breast cancer. *N Engl J Med*, **358**, 1409–11.

Pandi M, Premalatha K (2015). Clustering microarray gene expression data using enhanced harmony search. *Int J Bio Inspir com*, **7**, 296-6.

Pleasance ED, Cheetham RK, Stephens PJ (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**, 191–6.

Pleasance ED, Stephens PJ, O'Meara S (2010). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, **463**, 184–90.

Sadhasivam N, Thangaraj P (2017). Design of an improved PSO algorithm for workflow scheduling in cloud computing environment. *Intell Autom Soft Co*, **23**, 493-500.

Sahni J, Vidyarthi D (2015). A cost-effective deadline-constrained dynamic scheduling algorithm for scientific workflows in a cloud environment. *IEEE Trans Cloud Comput*, **4**, 5065-82.

Slamon DJ, Godolphin W, Jones LA (1989). Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science*, **244**, 707–12.

Stephens PJ, McBride DJ, Lin ML (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**, 1005–10.