

Amandeep S. Sidhu
Sarinder K. Dhillon (Eds.)

Advances in Biomedical Infrastructure 2013

Proceedings of International Symposium
on Biomedical Data Infrastructure
(BDI 2013)

Editor-in-Chief

Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Amandeep S. Sidhu and Sarinder K. Dhillon (Eds.)

Advances in Biomedical Infrastructure 2013

Proceedings of International Symposium
on Biomedical Data Infrastructure
(BDI 2013)

 Springer

Editors

Dr. Amandeep S. Sidhu
Curtin Sarawak Research Institute
Curtin University
Miri
Malaysia

Dr. Sarinder K. Dhillon
Institute of Biological Sciences
University of Malaya
Kuala Lumpur
Malaysia

ISSN 1860-949X

ISBN 978-3-642-37136-3

DOI 10.1007/978-3-642-37137-0

Springer Heidelberg New York Dordrecht London

ISSN 1860-9503 (electronic)

ISBN 978-3-642-37137-0 (eBook)

Library of Congress Control Number: 2013933103

© Springer-Verlag Berlin Heidelberg 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Due to the emerging demands of huge amounts of biomedical data, new and improved data management capabilities are required for supporting a wide range of applications. Current Biomedical Databases are independently administered in geographically distinct locations, lending them almost ideally to adoption of intelligent data management approaches. As a result next generation of information infrastructure and data integration capabilities are needed to ensure increasing infrastructure agility required for high-throughput biomedical research.

The goal of this book is to focus on research issues, problems and opportunities in Biomedical Data Infrastructure identifying new issues and directions for future research in Biomedical Data and Information Retrieval, Semantics in Biomedicine, and Biomedical Data Modeling and Analysis. The book will become a useful guide for researchers, practitioners, and graduate-level students interested in learning state-of-the-art development in biomedical data management. The content of this book is at an introductory and medium technical level.

There are 13 chapters presented in this book. Individual chapters cover approaches and methodologies for Biomedical Data Integration, Gene Expression Data Analysis, Content Based Image Retrieval, Semantic Determination of Cancer Stages, Kinetic Modeling for Systems Biology, and Drug Discovery. Book Chapters included here were presented at the First International Symposium on Biomedical Data Infrastructure (BDI 2013). All submissions were evaluated on their originality, technical soundness, significance, presentation, and interest to the symposium attendees.

We hope that bioinformatics students will use the book material as a guide to acquire basic concepts and theories of biomedical data management. Bioinformatics practitioners will find valuable lessons in the book for building similar biomedical systems in future and will find rewarding research data management questions to address in their research.

January 2013
Kuala Lumpur, Malaysia

Amandeep S. Sidhu
Sarinder K. Dhillon

Contents

Integrative Approaches for Drug Discovery – PPAR Gamma as a Case Study	1
<i>Meena Kishore Sakharkar</i>	
Biomedical Informatics and the Future of Medicine	3
<i>Jean-Pierre A. Kocher</i>	
Inferring <i>E. coli</i> SOS Response Pathway from Gene Expression Data Using IST-DBN with Time Lag Estimation	5
<i>Lian En Chai, Mohd Saberi Mohamad, Safaai Deris, Chuii Khim Chong, Yee Wen Choon</i>	
Framework for Biodiversity Information Retrieval in Malaysia	15
<i>Sarinder K. Dhillon, Baldeve Paunoo, Amandeep S. Sidhu</i>	
Using Ant Colony Optimization (ACO) on Kinetic Modeling of the Acetoin Production in <i>Lactococcus Lactis C7</i>	25
<i>Nor Farhah Binti Saidin, Chuii Khim Chong, Yee Wen Choon, Lian En Chai, Safaai Deris, Rosli M. Illias, Mohd Shahir Shamsir, Mohd Saberi Mohamad</i>	
Semantic Rule-Based Determination of Cancer Stages from Free-Text Radiology Reports	37
<i>Sangsoo Nam, Heung-Seon Oh, Jong-Beom Kim, Sung-Hyon Myaeng, Jinwook Choi</i>	
Using Particle Swarm Optimization for Estimating Kinetics Parameters on Essential Amino Acid Production of <i>Arabidopsis Thaliana</i>	51
<i>Siew Teng Ng, Chuii Khim Chong, Yee Wen Choon, Lian En Chai, Safaai Deris, Rosli M. Illias, Mohd Shahir Shamsir, Mohd Saberi Mohamad</i>	
Content-Based and Similarity-Based Querying for Broad-Usage Medical Image Retrieval	63
<i>Christopher Town</i>	

Inferring Gene Networks from Gene Expression Data Using Dynamic Bayesian Network with Different Scoring Metric Approaches	77
<i>Masarrah Abdul Mutalib, Lian En Chai, Chuii Khim Chong, Yee Wen Choon, Safaai Deris, Rosli M. Illias, Mohd Saberi Mohamad</i>	
Malaysian Parasite Database Infrastructure	87
<i>Sarinder K. Dhillon, Nur-Imtiazah Shuhaimi, Susan Lim Lee Hong, Amandeep S. Sidhu</i>	
Prediction of Vanillin Production in Yeast Using a Hybrid of Continuous Bees Algorithm and Flux Balance Analysis (CBAFBA)	101
<i>Leang Huat Yin, Yee Wen Choon, Lian En Chai, Chuii Khim Chong, Safaai Deris, Rosli M. Illias, Mohd Saberi Mohamad</i>	
Identifying Gene Knockout Strategy Using Bees Hill Flux Balance Analysis (BHFBA) for Improving the Production of Ethanol in <i>Bacillus Subtilis</i>	117
<i>Yee Wen Choon, Mohd Saberi Mohamad, Safaai Deris, Rosli M. Illias, Lian En Chai, Chuii Khim Chong</i>	
A Hybrid of Artificial Bee Colony and Flux Balance Analysis for Identifying Optimum Knockout Strategies for Producing High Yields of Lactate in <i>Echerichia Coli</i>	127
<i>Seet Sun Lee, Yee Wen Choon, Lian En Chai, Chuii Khim Chong, Safaai Deris, Rosli M. Illias, Mohd Saberi Mohamad</i>	
Author Index	139

Integrative Approaches for Drug Discovery – PPAR Gamma as a Case Study

Meena Kishore Sakharkar

University of Tsukuba, Japan

meena.sak.gn@u.tsukuba.ac.jp

Abstract. The pharmaceutical industry is spending increasingly large amounts of money on the discovery and development of novel medicines, but this investment is not adequately paying off in an increased rate of newly approved drugs by the FDA. Accumulated knowledge on genomic information, systems biology, and disease mechanisms provide an unprecedented opportunity to elucidate the genetic basis of diseases, and to discover novel therapeutic targets from genomic data. With hundreds to a few thousand potential targets available in the human genome alone, and the rise in the role of multi-drug therapies for complex diseases, there is an urgent need to understand the relationships between diseases and genes, and drugs and targets. These data can further be used for mapping cellular pathways and gene networks underlying the onset of disease and the possible mechanisms of pharmacological treatments that ameliorate the specific disease phenotype and help understand the relationships between diseases, genes, drugs, targets, and phenotypes. One key multi-disease target is PPAR Peroxisome proliferator-activated receptor γ (PPAR γ). PPAR-gamma is a nuclear receptor and plays important roles in breast cancer cell proliferation. The complexity of the underlying biochemical and molecular mechanisms of breast cancer and the involvement of PPAR γ in breast cancer pathophysiology is unclear. We have carried out computational prediction of the Peroxisome Proliferator Response Element (PPRE) motifs in 2332 genes reported to be involved in breast cancer in literature. A total of 178 genes were found to have PPRE (DR1/DR2) and / or PACM (PPAR-associated conserved motif) motifs. We further analysed the protein-protein interaction networks, disease gene networks and gene ontology to identify novel key genes for experimental validation. Four transcriptional targets of PPAR-gamma - MnSOD (ROS balance), NHE1 (pH maintenance), PGK1 (Glycolysis) and PKM2 (Glycolysis/metabolic regulator) were validated in vitro and PPAR-gamma ligands were found to repress these genes in two breast cancer cell lines MDA-MB-231 and MCF-7 and cause apoptosis. These findings have implications in breast cancer therapeutics and will also help in understanding the molecular mechanisms by which PPAR γ regulates the cellular energy pathway.

Biomedical Informatics and the Future of Medicine

Jean-Pierre A. Kocher

Mayo Clinic, Minnesota, USA
kocher.jeanpierre@mayo.edu

Abstract. One year ago Mayo Clinic decided to significantly invest in three newly-created centers: the Center for Individualized Medicine (CIM), the Center for the Science of Healthcare Delivery (CSHD) and the Center for Regenerative Medicine (CRM). The development of these three centers has been prioritized by Mayo as a strategic investment into the future of the Clinic and the future of medicine. Biomedical informatics at Mayo Clinic, including the four fields of Medical Information, Bioinformatics, Biostatistics and Medical Imaging, will play a significant role in the successful implementation of these three centers. As one of the Program Directors of the Center for Individualized Medicine, I will discuss the contribution of biomedical informatics to the future success of this center. I will provide an overview of the systems that we have developed and will describe how these systems leverage our biomedical informatics expertise to accelerate translational research activities, enhance medical practice and support the deployment of individualized medicine at Mayo Clinic.

Inferring *E. coli* SOS Response Pathway from Gene Expression Data Using IST-DBN with Time Lag Estimation

Lian En Chai, Mohd Saberi Mohamad*, Safaai Deris,
Chuii Khim Chong, and Yee Wen Choon

Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai 81310, Johor, Malaysia
{lechai2, ckchong2, ywchoon2}@live.utm.my, {saberi, safaai}@utm.my

Abstract. Driven to discover the vast information and comprehend the fundamental mechanism of gene regulations, gene regulatory networks (GRNs) inference from gene expression data has gathered the interests of many researchers which is otherwise unfeasible in the past due to technology constraint. The dynamic Bayesian network (DBN) has been widely used to infer GRNs as it is capable of handling time-series gene expression data and feedback loops. However, the frequently occurred missing values in gene expression data, the incapability to deal with transcriptional time lag, and the excessive computation time triggered by the large search space, are attributed to restraint the effectiveness of DBN in inferring GRNs from gene expression data. This paper proposes a DBN-based model (IST-DBN) with missing values imputation, potential regulators selection, and time lag estimation to address these problems. To assess the performance of IST-DBN, we applied the model on the *E. coli* SOS response pathway time-series expression data. The experimental results showed IST-DBN has higher accuracy and faster computation time in recognising gene-gene relationships when compared with existing DBN-based model and conventional DBN. We also believe that the ensuing networks from IST-DBN are applicable as a common framework for prospective gene intervention study.

Keywords: Dynamic Bayesian network, missing values imputation, time-series gene expression data, gene regulatory networks, network inference.

1 Introduction

In the post-genomic era, aided by the breakthroughs in technology, researchers have begun to shift the research paradigm from the classical reductionism to the modern holism, wherein biological systems and experimental design are viewed as a whole instead as collections of parts [1]. One of the innovations conceived in such era, the

* Corresponding author.

DNA microarray technology, which is capable of representing the expression of thousands of genes under various circumstances (otherwise known as gene expression profiling), has allowed the development of numerous new experiments for exploring into the complex system of gene expression and regulation [2]. Since its conception, various organisms and mammalian cells have been profiled, such as *S. cerevisiae* [3], human cancerous tissue [4], and *E. coli* [5]. The consequent output, commonly known as gene expression data, comprises immense information such as the robustness and behaviours denoted by the cellular system under diverse situations [6], assists us in understanding the underlying mechanism of gene expression and regulation.

From a computational perspective, a GRN can be represented as a directed graph containing nodes (genes) and edges (interaction/relationship). In recent years, various computational methods have been developed to infer GRNs from gene expression data. Among them, Bayesian network (BN) [7], which uses probabilistic correlation to distinguish relationships between a set of variables, was popular in GRNs inference. This is mainly due to several factors: BN is capable of working on local elements, assimilating other mathematical models to avert data overfitting, and merging prior knowledge to fortify the causal relationships. Nonetheless, BN also has two disadvantages: it is unable to deal with time-series gene expression data and construct feedback loops.

From a biological perception, feedback loops actually embody the homeostasis procedure in living organisms. Hence, to take account of the feedback loops, researchers have developed the dynamic Bayesian network (DBN) [8] as a replacement to tackle BN's weaknesses. However, the scattering missing values commonly found in gene expression data could affect more than 90% of the genes and subsequently negatively influencing downstream analysis and inferring approaches [9]. Furthermore, in identifying gene-gene relationships, conventional DBN generally comprises all genes into the subsets of potential regulators for each target gene, and thus instigated the large search space and the excessive computational time [10]. To address the two problems, Chai *et al.* [11] suggested a three-step DBN-based model (ISDBN) with missing values imputation and potential regulators selection, and the proposed model showed better performance than conventional DBN in GRNs inference.

Yet, ISDBN and conventional DBN is still not adept enough to effectively take account of the transcriptional time lag, in which a time delay exists before the target genes are being expressed into the system. This shortcoming hampers the accuracy of DBN-based approaches in GRNs inference. To solve this problem, we proposed to further improve the aforesaid DBN-based model with time lag estimation (IST-DBN) which would take account of the transcriptional time lag based on the time difference between the initial changes of expression level of potential regulators and their target genes.

2 Methods

Essentially, IST-DBN involves four main steps: missing values imputation, potential regulators selection, time lag estimation and DBN inference. Fig. 1 illustrates the schematic overview of IST-DBN.

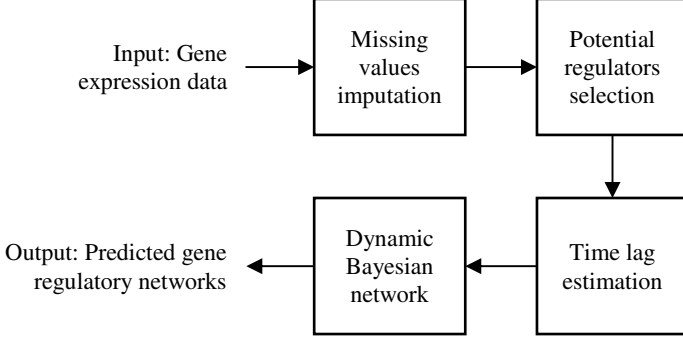


Fig. 1. Schematic overview of IST-DBN

2.1 Missing Values Imputation

Missing values in gene expression data can occur for numerous reasons. For example, small contaminations would corrupt the microarray slides at multiple spots as they are very tiny and packed together. These questionable spots are then labelled as missing after scanning and digitalising the microarray slides. Many imputation methods have been established to impute missing values by exploring and utilising the underlying expression data structure and pattern. In particular, based on the local similarity structure, LLSimpute imputes missing values by constructing a linear combination of similar genes and target genes with missing values through a similarity measure [12]. This method entails two steps. Firstly, k genes are selected by the L_2 -norm, where k is a positive integer that expresses the number of coherent genes to the target gene. As an example, to impute a missing value g found at x_{1j} in a $m \times n$ matrix \mathbf{X} , the k -nearest neighbour gene vectors for x_{1j} ,

$$\mathbf{v}_{s_i}^T \in \mathbf{X}^{1 \times n} \quad 1 \leq i \leq k \quad (1)$$

are computed, whereby the gene expression data is defined as a $m \times n$ matrix \mathbf{X} (m is the number of genes, n is the number of observations), and x_1 signifies the row of the first gene with n observations. s_i is a list of k -nearest neighbour genes vectors, which actually corresponds to the i -th row of the transpose vector \mathbf{v}^T . The following step implicates regression and estimation of the missing values. A matrix, $\mathbf{A} \in \mathbf{X}^{k \times (n-1)}$ wherein the k rows of the matrix contains vector \mathbf{v} , and two vectors, $\mathbf{b} \in \mathbf{X}^{k \times 1}$ and $\mathbf{w} \in \mathbf{X}^{(n-1) \times 1}$, are then formed. The vector \mathbf{b} encloses the first element of k vectors \mathbf{v}^T , whereas vector \mathbf{w} comprises $n - 1$ elements of vector x_1 . A k -dimensional coefficient vector \mathbf{y} is subsequently computed such that the least square problem is minimised as

$$\min_{\mathbf{y}} |\mathbf{A}^T \mathbf{y} - \mathbf{w}|^2 \quad (2)$$

Let \mathbf{y}^* to denote the vector wherein the square is minimised such that

$$\mathbf{w} \simeq \mathbf{A}^T \mathbf{y}^* = \mathbf{y}_1^* \mathbf{a}_1 + \mathbf{y}_2^* \mathbf{a}_2 + \cdots + \mathbf{y}_k^* \mathbf{a}_k \quad (3)$$

where $\mathbf{a}_i \in \mathbf{A}^{k \times 1}$, and thus, the missing value g could be imputed as a linear combination of coherent genes such that

$$g = \mathbf{b}^T \mathbf{y} = \mathbf{b}^T (\mathbf{A}^T)' \mathbf{w} \quad (4)$$

where $(\mathbf{A}^T)'$ exists as the pseudoinverse of \mathbf{A}^T [12].

2.2 Potential Regulators Selection

In most occurrences, the expression level of regulators (also known as TFs, transcriptional factors) would vary before or simultaneously with their target genes [13]. By exploiting this information, we formulated an algorithm which would shrink the search space by confining the number of potential regulators for each target gene. Firstly, a threshold for categorising the status of gene expression values (e.g. up- or down-regulation) is determined through either experiments or the average expression level of the genes. In this paper, the threshold for up-regulation and down-regulation are decided based on the baseline cut-off of the gene expression values. As such, for the *E. coli* dataset used in this paper, the threshold is determined as ≥ 1.4 for up-regulation and ≤ 0.7 for down-regulation. The gene expression values are successively categorised into one of the three states: up-, down- and normal regulation. The three states specify whether the expression value is greater than, lower than or similar to the threshold. Subsequently, the precise time units of initial up-regulation and down-regulation of each gene are chosen, and genes with preceding fluctuations in expression level are encompassed into the subset of potential regulators against genes with later expression fluctuations. As genes with significantly late expression fluctuations could have involved a large number of potential regulators, the maximum time gap for preceding expression fluctuations is constrained to five time units. This is to avert choosing potential regulators for a target gene from the entire gene expression dataset. To further elucidate this algorithm, let's assume two hypothetical genes: gene P and gene R . Gene P encountered an initial expression change at time T_1 prior to the initial expression change of gene R at time T_2 , hence gene P is included into the subset of potential regulators for gene R (Fig. 2). The same procedure applies to other up- or down-regulated genes which satisfy the criteria.

2.3 Time Lag Estimation

Transcriptional time lag is the time interval between the expression of the regulators and the expression of their target genes. Remember the two hypothetical genes, P and R ,

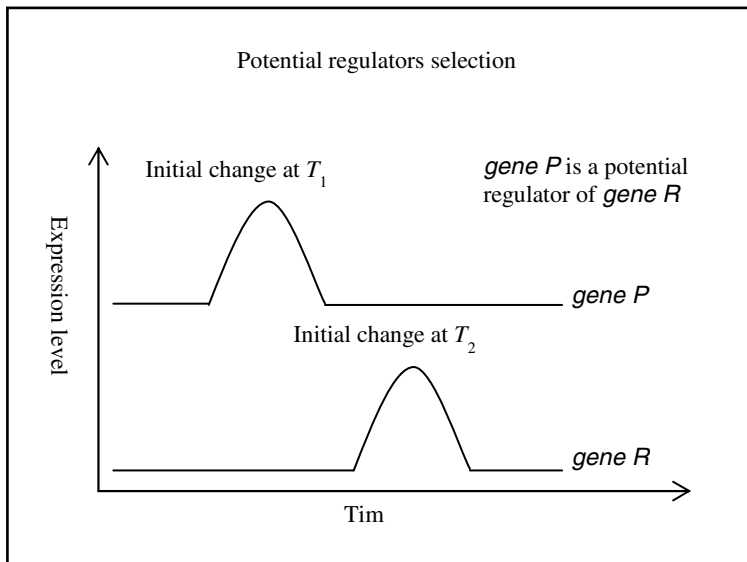


Fig. 2. Schematic overview of potential regulators selection

whereby gene *P* regulates gene *R*. Gene *P* starts expression change at time T_1 and gene *R* has an expression change at T_2 . The time difference between T_1 and T_2 is regarded as the transcriptional time lag. In inferring GRNs from gene expression data, conventional DBN aligns regulator-gene pairs based on the statistical analysis of their probabilistic strength between time units. Nonetheless, DBN usually pairs up regulators with their target genes by only one time unit, although the actual transcriptional time lag could have been multiple time units. With such cases, IST-DBN takes consideration of the real transcriptional time lag by coupling up potential regulators and their target genes based on the time difference between their initial expression fluctuations. For a target gene, potential regulators are categorised into different groups based on the time lag (e.g. groups of one, two or three time units), mostly due to the fact that a target gene could have numerous regulators acting upon it in dissimilar time unit.

2.4 Dynamic Bayesian Network

DBN infers time-series gene expression data by observing the values of a set of variables at diverse time units. DBN inference typically involves two steps: parameter learning and structure learning. In parameter learning, the joint probability distribution (JPD) of the variables is calculated based on the Bayes theorem. Let's assume a microarray dataset with m genes and n observations, such that we have a $m \times n$ matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_m)$ wherein each row, vector $\mathbf{x}_m = (\mathbf{x}_{m1}, \dots, \mathbf{x}_{mn})$ embodies a gene expression vector observed at time t . The temporal vectors chain relationship is

defined as a *first-order Markov chain* in which only forward edges are permitted. The JPD of the model has the overall form of:

$$P(\mathbf{x}_{11}, \dots, \mathbf{x}_{mn}) = P(\mathbf{x}_1)P(\mathbf{x}_2|\mathbf{x}_1) \dots P(\mathbf{x}_i|\mathbf{x}_{i-1}) \quad (5)$$

Based on the earlier threshold, the expression values acquired from preceding steps are discretised into three categories: -1, 0 and 1, which correspond to down-, normal and up-regulation respectively. Each set of potential regulators is subsequently distributed into smaller subsets. For instance, in a set of potential regulators comprising gene X , and gene Y , the subsets would be $\{X\}$, $\{Y\}$ and $\{X, Y\}$. Each of the subset and the target gene are then arranged into a data matrix with their discretised expression values. The conditional probabilities of each subset of potential regulators against their target genes are then calculated. The following step is to look for the optimal network structure through a scoring function based on the Bayesian Dirichlet equivalence (BDe). The final results are then imported into GraphViz (<http://www.graphviz.org>) for network visualisation and analysis.

3 Result and Discussion

3.1 Experimental Data and Setup

The experimental data involved in this paper is the *E. coli* SOS response pathway gene expression data [14]. The *E. coli* SOS response pathway is a DNA restoration system which reacts to damaged DNA by pausing cell cycle and triggering DNA repair [15]. In normal situation, the SOS genes are negatively regulated through the binding of the repressor protein, *lexA* to the promoter region of these genes. When DNA is damaged, DNA polymerase is blocked and single-stranded DNA (ssDNA) start to accumulate. The sensor of DNA damage, the *recA* protein, activates by binding to these ssDNA. After being activated, the *recA* protein initiates the self-cleavage of the *lexA* repressor. This would cause a drop in *lexA* level and in turn the SOS genes are de-repressed. This remains until the damage is restored, wherein the level of activated *recA* falls, *lexA* amasses and represses the SOS genes again. This dataset comprises 8 genes observed at uniformly spaced 50 time units with 6 minutes apart, and also 11.5% missing values (184 out of 1,600 observations).

The DBN inferring part of IST-DBN is applied under the framework of BNFinder [16], while the missing values imputation, potential regulators selection and the time lag estimation are applied in MATLAB environment. To assess the performance of IST-DBN, the accuracy and computation time of the proposed model is compared against ISDBN and DBN (characterised by BNFinder). The accuracy is evaluated by comparing the results of the three models to the reputable *E. coli* SOS response pathway by Ronen *et al.* [14]. All three models are executed using the same hardware configuration (3.2GHz Intel Core i3 computer with 2GB main memory) to ensure a fair assessment of computation time. Table 1 summarises the results, wherein the first

row denotes the network inferred by IST-DBN, the second row denotes the network inferred by ISDBN, and the third row denotes the network inferred by DBN. An edge shows a relationship between the two linked genes. ‘Correctly inferred relationships’ represents the number of relationships which are found in both inferred and established networks, ‘sensitivity’ is the rate of correctly inferred relationships, and ‘specificity’ relates to the rate of correct inference that no relationship exists between two genes.

3.2 Experiment Results

IST-DBN succeeded in identifying all ten relationships (*lexA*–*recA*, *lexA*–*polB*, *lexA*–*umuD*, *lexA*–*uvrY*, *lexA*–*uvrA*, *lexA*–*uvrD*, *lexA*–*ruvA*, *lexA*–*lexA*, *recA*–*recA*, and *recA*–*lexA*) (Fig. 3), whereby ISDBN correctly identified nine relationships and DBN only recognised eight relationships. Both IST-DBN and ISDBN outperformed DBN in this category, and this is because the effectiveness of DBN was hampered by numerous missing values in the original gene expression data. Also, through the alignment of regulator-gene pairs based on actual transcription time lag, the causal correlation between pairs with greater transcriptional time lag are strengthened, due to the fact that IST-DBN reported lesser false positives when compared with ISDBN and DBN (3 against 6 and 5). IST-DBN registered 100% sensitivity and 83.33% specificity compared to ISDBN’s 90% sensitivity and 66.67% specificity. Conversely, DBN reported 80% sensitivity and 72.22% specificity. The perfect sensitivity of IST-DBN and the relatively significant difference in percentage with the other two models is obviously attributed to the relatively small dataset, but we expect that IST-DBN would still outperform ISDBN and DBN on larger dataset. All three models were capable of identifying at least two self-regulatory loops: *recA*, which senses DNA damage and subsequently self-activate by binding to ssDNA; and *lexA*, which undergoes self-cleavage after initiated by the relatively high level of activated *recA*.

Four probable situations arise when an edge exist between two genes: correct direction and regulation type, correct direction but incorrect regulation type, misdirected but correct regulation type, and misdirected and wrong regulation type. IST-DBN was able to revise an incorrect relationship type in ISDBN. However, IST-DBN also contains an incorrect regulation type while ISDBN showed two wrong regulation type and one misdirected edges, and conventional DBN reported three incorrect regulation type and one misdirected edges. In regard to the computation time, IST-DBN demonstrated a computation time of 7 minutes and 56 seconds while ISDBN showed a computation time of 8 minutes and 43 seconds. On the contrast, DBN recorded 15 minutes and 17 seconds. As the dataset used in this study was relatively small, the computation time for IST-DBN and ISDBN do not differ drastically, although DBN suffers from longer computation time which is caused by a larger search space. We expect that the computation time difference between DBN and the other two models would be much more radical with a larger dataset.

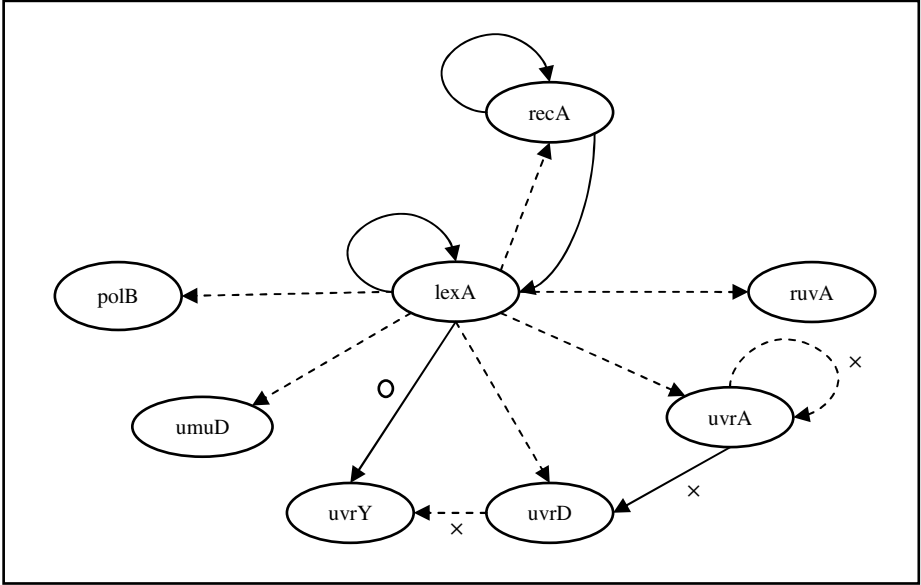


Fig. 3. Inferred *E. coli* SOS response pathway using IST-DBN. Dash edges (---) indicate down-regulations and straight-lined edges (—) indicate up-regulations. A cross denotes an incorrect inference; a circle denotes an incorrect regulation type; an edge without any attachment is a correct inference.

Table 1. The results of experiment study

Model	Correctly predicted relationships	Sensitivity	Specificity	Computation time (HH:MM:SS)
IST-DBN	10	100.00%	83.33%	00:07:56
ISDBN	9	90.00%	66.67%	00:08:43
DBN	8	80.00%	72.22%	00:15:17

4 Conclusion

Traditional DBN has been troubled by three main problems: the missing values commonly found in gene expression data, the comparatively large search space due to encompassing all genes as potential regulators against target genes, and the absence of a method to consider transcriptional time lag. ISDBN was put forth by Chai *et al.* [11] to tackle the first two problems: Missing values are imputed based on linear grouping of analogous genes, and the search space is diminished by restricting to certain potential regulators which fulfill the criteria. Nevertheless, this model is unable to deal with transcription time lag and thus, we proposed an enhanced version of ISDBN with time lag estimation (known as IST-DBN) to solve the third problem. Rather than

pairing up with the default one time unit, IST-DBN utilises the actual time difference between expression changes to align regulator-gene pairs. Therefore, IST-DBN is capable of seizing most of the probabilistic connection between genes that possess transcriptional time lag greater than one time unit. Based on the *E. coli* SOS response pathway dataset, IST-DBN presented encouraging results in regards to accuracy and computation time when matched against ISDBN and traditional DBN. We are interested to apply IST-DBN to other datasets, for instance, *S. cerevisiae* or *A. thaliana*, to examine the performance consistency of IST-DBN.

Acknowledgments. We would like to thank the Malaysian Ministry of Science, Technology and Innovation for supporting this research by an e-science research grant (Grant number: 06-01-06-SF1029). This research is also supported by UTM GUP research grant that was sponsored by Universiti Teknologi Malaysia.

References

1. Lee, W.P., Tzou, W.S.: Computational methods for discovering gene networks from expression data. *Brief Bioinform.* 10(4), 408–423 (2009)
2. Jornsten, R., Wang, H.Y., Welsh, W.J., Ouyang, M.: DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics* 21(22), 4155–4161 (2005)
3. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9, 3273–3297 (1998)
4. Muro, S., Takemasa, I., Oba, S., Matoba, R., Ueno, N., Maruyama, C., Yamashita, R., Sekimoto, M., Yamamoto, H., Nakamori, S., Monden, M., Ishii, S., Kato, K.: Identification of expressed genes linked to malignancy of human colorectal carcinoma by parametric clustering of quantitative expression data. *Genome Biol.* (4), 21 (2003)
5. Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sanchez-Solano, F., Santos-Zavaleta, A., Martinez-Flores, I., Jimenez-Jacinto, V., Bonavides-Martinez, C., Segura-Salazar, J., Martinez-Antonio, A., Collado-Vides, J.: RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.* 34, 394–397 (2005)
6. Karlebach, G., Shamir, R.: Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Bio.* 9(10), 770–780 (2008)
7. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyse expression data. *J. Comp. Biol.* 7, 601–620 (2000)
8. Murphy, K., Mian, S.: Modelling gene expression data using dynamic Bayesian networks. Technical report, Computer Science Division, University of California, Berkeley (1999)
9. Ouyang, M., Welsh, W.J., Geogopoulos, P.: Gaussian mixture clustering and imputation of microarray data. *Bioinformatics* 20(6), 917–923 (2004)
10. Jia, Y., Huan, J.: Constructing non-stationary dynamic Bayesian networks with a flexible lag choosing mechanism. *BMC Bioinformatics* (11), 27 (2010)
11. Chai, L.E., Mohamad, M.S., Deris, S., Chong, C.K., Choon, Y.W., Ibrahim, Z., Omatu, S.: Inferring gene regulatory networks from gene expression data by a dynamic bayesian network-based model. In: Omatu, S., De Paz Santana, J.F., González, S.R., Molina, J.M., Bernardos, A.M., Rodríguez, J.M.C. (eds.) *Distributed Computing and Artificial Intelligence. AISC*, vol. 151, pp. 379–386. Springer, Heidelberg (2012)

12. Kim, H., Golub, G., Park, H.: Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics* 21(2), 187–198 (2005)
13. Yu, H., Luscombe, N.M., Qian, J., Gerstein, M.: Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet.* 19, 422–427 (2003)
14. Ronen, M., Rosenberg, R., Shraiman, B.I., Alon, U.: Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci.* 99, 10555–10560 (2002)
15. Radman, M.: Phenomenology of an inducible mutagenic DNA repair pathway in *Escherichia coli*. *Basic Life Sci.* 5A, 255–367 (1975)
16. Wilczynski, B., Dojer, N.: BNFinder: exact and efficient method for learning Bayesian networks. *Bioinformatics* 25(2), 286–287 (2009)

Framework for Biodiversity Information Retrieval in Malaysia

Sarinder K. Dhillon¹, Baldeve Paunoo¹, and Amandeep S. Sidhu²

¹ Institute of Biological Sciences, Faculty of Science,
University of Malaya, 50603 Kuala Lumpur, Malaysia

² Curtin Sarawak Research Institute,
Curtin University, Sarawak, Malaysia

Abstract. A wealth of information exists on Malaysia's biodiversity resources and associated knowledge. This may be in form of specimens, literature, and electronic databases. The problem currently faced by Malaysia in regards to all this information is basically one of dissemination and hence retrieval of this information in a networked environment. Presently there is no system to link the scattered data so as to facilitate exchange of data amongst the different databases available in the country. In this paper we developed an indigenous framework for integrating distributed heterogeneous and relational biodiversity databases to facilitate biodiversity data sharing among the scientific contemporaries especially in Malaysia. This proposed approach is expected to encourage interested members of the research community to work together while maintaining the privacy of their data. Simplicity is also stressed with the intention of offering a system which is adaptable to wide spectrum users within the scientific community in Malaysia and the rest of the world.

Keywords: Data Integration, Information Retrieval, Biodiversity.

1 Introduction

Bioinformatics deals with computational management and analysis of biological information (Buttler et al., 2002; Nilges et al., 2002). Biological information and its management is a huge area of interest as it consists of data and information from many areas such as medical, genetics, biodiversity, environment, biotechnology and many more. In this paper we focus on biodiversity data and information.

The Convention on Biological Diversity (CBD, 2005) stated that biological diversity - or biodiversity - is the term given to the variety of life on Earth and the natural patterns it forms. The biodiversity seen today is the fruit of billions of years of evolution, shaped by natural processes and, increasingly, by the influence of humans. It forms the Web of life of which humans are an integral part and upon which they so fully depend. It is the combination of Life forms and their interactions with each other and with the rest of the environment that has made Earth a uniquely habitable place for humans.

During the last decade there has been an increasing interest in gathering and analyzing biodiversity information for the scientific administration of natural resources. Many initiatives around the globe arose to sustain the biodiversity resources and communicate biological information using computational tools. Examples include the Canadian Clearing-House Mechanism (Secretariat of the Convention on Biological Diversity, 2005), the Japan's Global Taxonomy Initiative (GTI) (Ando and Watanabe, 2003), Inter-American Biodiversity Information Network (IABIN, 2004a) and Global Biodiversity Information Facility (GBIF, 2004).

However, as far as communication is concerned, biodiversity information clusters are usually disperse, unreported and in some cases inaccessible. There is an emerging need to overcome this setback by looking at issues concerning communication of biodiversity information (biodiversity informatics). There is a global need to link all the disperse information clusters in remote and distributed medium. Technological advances within distributed network communications for biodiversity informatics offer a window of opportunity for gathering and maintaining information repositories about biodiversity, analyzing this information, reporting and visualizing it (Blum, 2000). However, poor results have been obtained so far (Schnase, 2003). This is because each researcher collects data in a way suitable to their interests and collaboration between the biology scientists and the information technology community is still immature. The lack of a common or standard format adopted for data representation, makes it difficult to access consistent data. Also some individuals and institutions working in biodiversity are reluctant to publish the collected information.

Reflecting the concerns discussed above, it can be concluded that there is a need to integrate the different views and versions of taxonomic data, making it available in simple formats, with friendly interfaces to be shared among the scientific community and to bring them together to work as a team to achieve their respective goals, without expecting them to export their collected information to a centralized data warehouse.

2 Malaysian Perspective on Biodiversity Information

Malaysia has been identified as one of the world's twelve mega-diversity areas with extremely rich biological resources. There are over 15,000 known species of higher plants, 300 species of mammals, 254 species of breeding birds, 198 species of amphibians, 379 species of reptiles, over 150,000 species of the invertebrates, and over 4,000 species of marine fishes and 449 species of freshwater fishes in Malaysia (Burhanuddin, 2000). Despite these facts, Malaysia does not have a central physical body for storing natural history collections while the virtual repositories are maintained by individuals or organizations disparately. This is explained further in the following subsections.

Due to the historical background, it also contains specimens collected in Malaysia. University of Malaya (UM) has "Rimba Ilmu" Botanical Garden (Wong, 2005) and Zoology Museum which are two important physical bodies that store the vast biodiversity specimens collected by scientists in the university. Besides Raffles Museum, some of the Malaysian (especially Sarawak) specimens are kept in Natural History

Museum in London and Natural History Museum at Tring. Sarawak is famous with its rich diversity of biological specimens which date back to the time of A.R Wallace. Wallace, who arrived in Sarawak at the invitation of Rajah Sir James Brooke on Nov 1 1854, spent fifteen months exploring and collecting an enormous 25,000 specimens, including 2,000 beetle species, 1,500 moth species and 1,500 other insect orders along the Sarawak River valley from Santubong to Bau as well as the peat swamps of Simunjan. The collections, which he sold to private collectors and institutions in the United Kingdom to finance his travels in the region, are now kept at the Natural History Museum in London and Natural History Museum at Tring.

Besides UM, University Science Malaysia (USM) has also taken steps towards research in biodiversity. It adopted the concept of “the university in a garden” to promote the preservation of green areas as integral to the development of the intellect and thus enhancing the spirit and practice of nature conservation. University Putra Malaysia (UPM) too plays a very active role in biodiversity conservation in the country. The Institute of BioScience in UPM is a center of excellence for biological research, including biodiversity. The Institute of Medical Research (IMR, 2006), a deserving institute which existed for about 106 years now, has been actively involved in biodiversity research especially medical related organisms, bacteria, virus, protozoas, parasites and pathogens. It is one of the oldest institutions in Malaysia which has a physical repository of specimens. Other institutions, for example FRIM (Forest Research Institute of Malaysia) and MARDI (Malaysian Agricultural Research and Development Institute) have huge biodiversity collections in Malaysia. The above organizations and institutions explain about the physical distributed data warehouses that records, stores and analyzes biodiversity information. However, these physical entities work individually with minimum collaboration among each other.

While the physical repositories are essential to store the various biodiversity specimens, the virtual repositories are equally important. They can be used to manipulate, share and disseminate the data. In Malaysia, several initiatives have been undertaken to digitize their biodiversity information. In 2003, the Institute of Biological Sciences at UM started an initiative named Integrated Biological Sciences Initiative (IBDI) (Sarinder et al., 2005), in which relational biodiversity databases were developed and museum collections catalogued. This initiative resulted in the digitization and subsequent electronic availability of vast amount of biodiversity data in UM. Palm Oil Research Institute of Malaysia (PORIM) maintains an oil palm database, accessible to registered internet users only. In addition, the Forest Research Institute of Malaysia (FRIM) provides Web users limited database access to its huge forest resource collections (Merican et al., 2002).

From the discussion above, it can be said that a wealth of information exists on Malaysia's biodiversity resources and associated knowledge. This may be in form of specimens, literature (such as: unpublished reports, and books, monographs and scientific papers), and electronic databases. Undoubtedly, the problem faced by Malaysia in biodiversity information is basically one of dissemination and hence retrieval of information in a networked environment. Presently there is no system to link the scattered data so as to facilitate exchange of data amongst the different databases

available in the country. Besides the issues discussed above, there are other impediments which can be summarized as follows:

1. Some institutions have information documented in the form of databases, while other institutions are looking into using database technologies.
2. Databases are in heterogeneous formats.
3. Few databases can be accessed through the Web, while many are only available offline.
4. Some of these databases are well-structured, whereas others are largely project or species specific and are mainly unstructured or semi-structured.
5. There is no meta-data (data-dictionary) to be followed as a standard for these databases.

As a result, there are challenges to tackle the issues regarding biodiversity information dissemination and retrieval which we will address in this paper.

3 Communication Architecture for Biodiversity Information Retrieval (CABIR)

We propose a novel Communication Architecture for Biodiversity Information Retrieval (CABIR) system as a significant improvement from the previously mentioned DiGIR system. Apart from the issues discussed above, at the moment Malaysia does not have an indigenous system to link and integrate the local biodiversity databases. The proposed system will not only integrate existing local biodiversity databases but will link the integrated information with global biodiversity resources using international standards accepted by the community. The proposed system will address the issue of heterogeneity by supporting all the data formats biodiversity data is available in Malaysia. Also the proposed system will provide data security for database owners by allowing them to keep and maintain their own data and to choose information to be shared and linked. Our roadmap to development of CABIR system is shown in Figure 1.

Data for CABIR is gathered from existing biodiversity databases and from researchers in various biodiversity fields. A data format for CABIR was developed based on the analysis of existing data that needs to be integrated. This format will be used to standardize naming conventions for the new biodiversity databases built after this research. The data format can be altered or added according to the requirements of a database owner. It is important to note that the key fields in this format is inline a global standard which is Darwin Core V2, therefore databases built using this format can be shared with other biodiversity databases in the world.

The hierarchical design and the work flow of the Database Integration System used for CABIR is shown in Figure 2. The provider and XML wrapper (highlighted box) reside in the remote machine where the database is stored. The query based portal, resource and results reside in the application machine. Users will only see the search page as the rest are hidden from users.

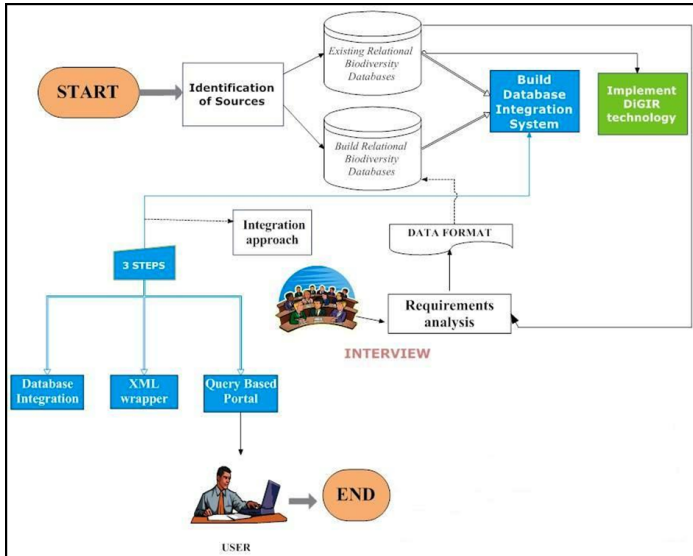


Fig. 1. Roadmap to development of CABIR System

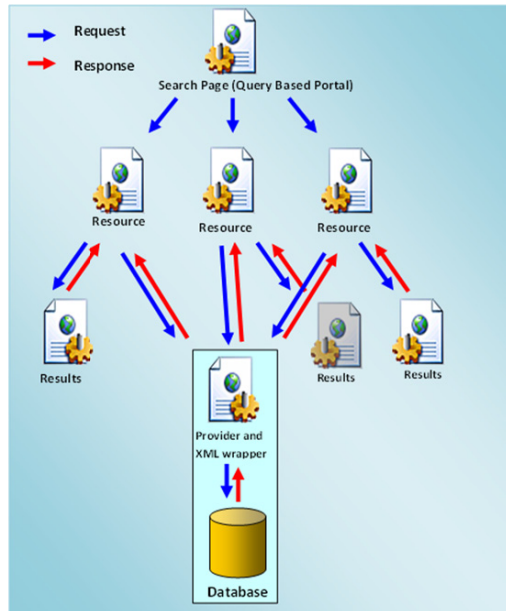
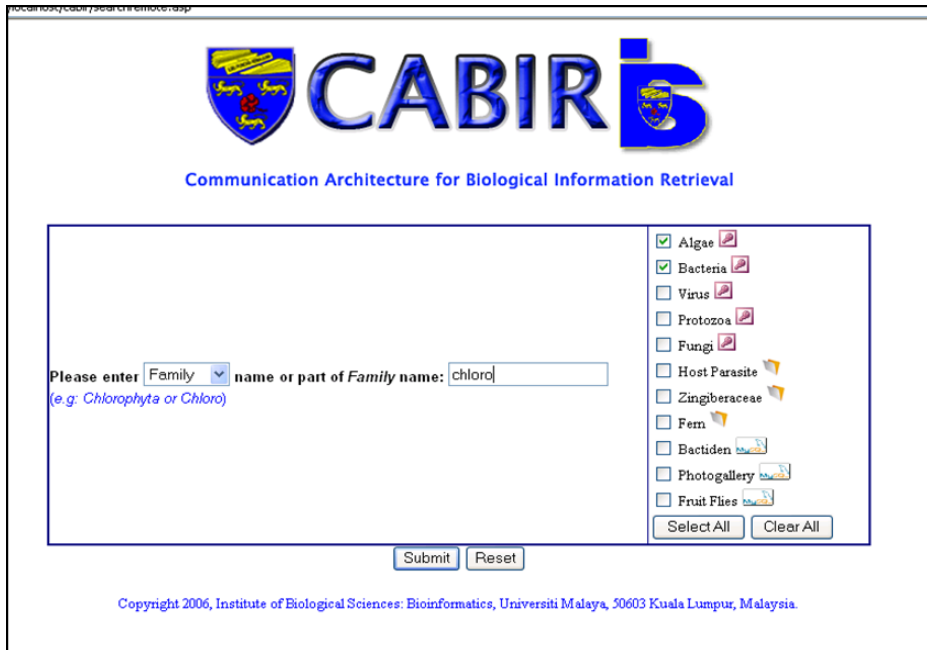


Fig. 2. Hierarchical view of Database Integration System used for CABIR

Figure 3 shows the main page of the CABIR system. There is a simple form for collecting user query. There are thirteen databases at the right column of the form, from which a user can choose to send query. Among them, five are Microsoft Access databases (Algae database, Bacteria database, Virus database, Protozoa database, and

Fungi database), three are FileMaker databases (Host Parasite database, Zingiberaceae database, and Fern database), four more are MySQL databases (Bactiden database, Photogallery database, Fruit Flies database and Biodiversity sample database) and one DB2 database (Virus sample database). User can choose multiple databases concurrently. At the left column, there is one drop down box of fields which are Family, Genus, and Species. Next to the drop down box is the text box for user to type the query of the selected field.



Communication Architecture for Biological Information Retrieval

Please enter name or part of Family name:

(e.g. Chlorophyta or Chloro)

- Algae
- Bacteria
- Virus
- Protozoa
- Fungi
- Host Parasite
- Zingiberaceae
- Fern
- Bactiden
- Photogallery
- Fruit Flies

Copyright 2006, Institute of Biological Sciences: Bioinformatics, Universiti Malaysia, 50603 Kuala Lumpur, Malaysia.

Fig. 3. Main page of CABIR

In order to search for a Chlorophyta family in algae and bacteria databases, a user just needs to type Chloro or chloro (see Figure 3). This is because CABIR was designed to be case insensitive. Users need to be aware of case sensitivity only when searching from databases on FileMaker DBMSs. The result of a search is displayed in Figure 4. There is one header for every database which describes the source of database (location of database). There is a bar below the header which shows the field user is searching from and number of records found. Each record is segregated by the alternate green and white background. The query results for “chloro” were obtained in 0.09 seconds.

CABIR has a simple and user friendly interface. Users can select fields to search from a drop down menu, key in the search in the textbox and select the database they want to query. There is a help link that will take users to the help screen which will assist users on using the system. The results of search are presented in a manner that is easy to understand and has features that can assist a user such as timer and total

count of records found in each database. Besides that, users can jump to records in different databases by clicking on the hyperlinks in the results page. For each record, only Family, Genus and Species are displayed to avoid congestion in the screen. If users want more details of a specific record, they can navigate through the links.



Fig. 4. Results of search

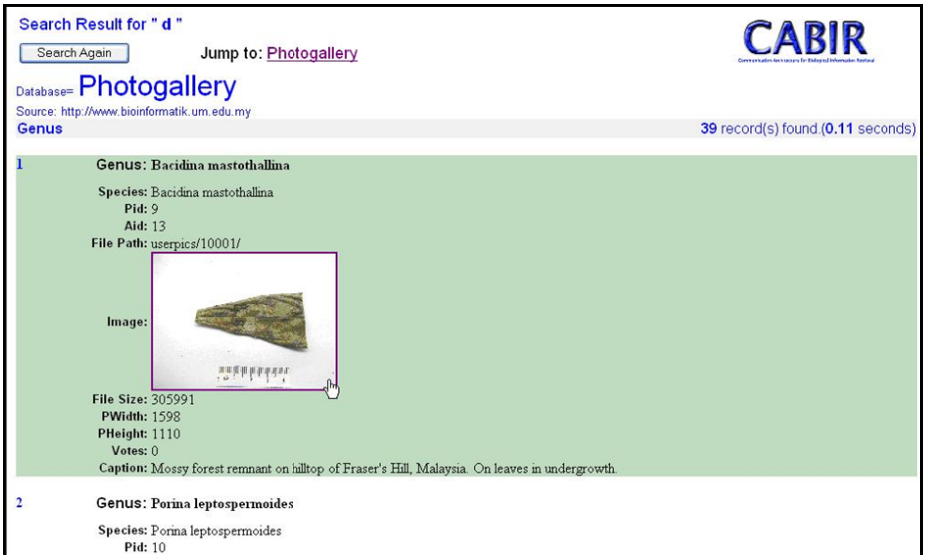


Fig. 5. Image displayed in results from Photogallery database

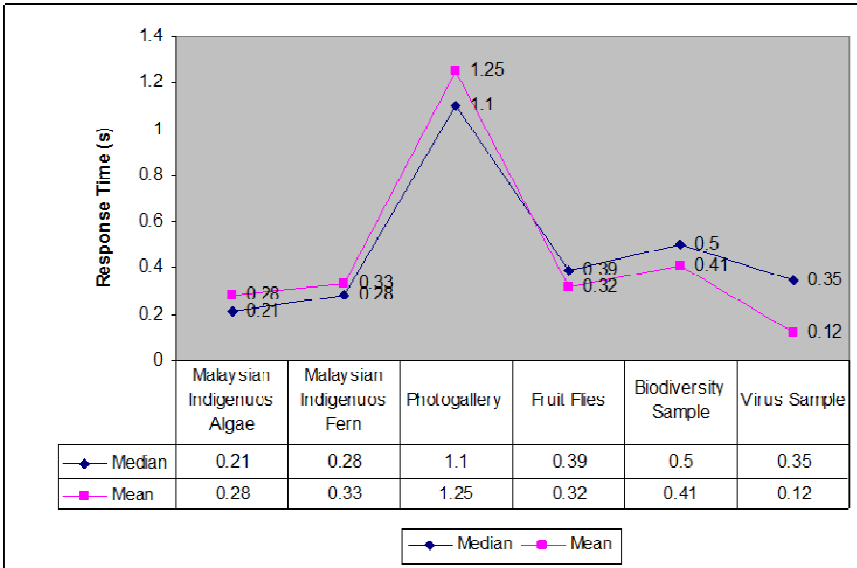


Fig. 6. Response time mean and median plots for six databases using the CABIR system

CABIR is also able to retrieve images from Bactiden and Photogallery databases (see Figure 5). The images are shrunk to a smaller size for faster download and ease of view. However these images can be easily zoomed out for a clearer view by clicking them. The result of the query search on these databases was also very quick (39 records in 0.11 seconds).

We used a performance test method that follows a standard model used by Roderic (2005). For each database, the species name was searched 100 times and response time was recorded from the time the query was made until the time the results were returned. Response time mean and median for six databases using the CABIR system are also plotted in Figure 6.

4 Discussion

In this paper we showed that CABIR is able to integrate the different views and versions of taxonomic data, making it available in simple formats, with friendly interfaces to be shared among the scientific community. Due to the powerful programming tool and techniques, the underlying details of CABIR are concealed from users. Users do not need to know where the databases reside, the structure of the system, Database Management System used to build databases and the programming behind the system. The information retrieval is similar to a system accessing a local database to ensure its simplicity.

CABIR's user interface is made simple and straightforward. A user does not need much effort in understanding the functionality of the system. Thus it will suit a wide spectrum of users. Every page is designed to be as simple and as compact as possible.

CABIR system is designed to support search in multiple heterogeneous databases simultaneously. As providers of CABIR, databases owned by scientists are secured. Furthermore, database owners do not need to export their data in order to share their data. Instead they can store their data in their own hosts, while allowing them to be shared. Thus, they are able to protect their databases.

CABIR works well on biodiversity databases and it may also work for data in other biological domains and for a variety of different data sets outside the biological domain. This is because CABIR's components are independent of each other, especially the data format. Therefore, different data formats can be used with CABIR. CABIR also has generic characteristics of existing database integration systems.

Therefore, CABIR satisfies the need of a generic database integration system. In addition, it is made simple with powerful underlying facilities making it very suitable for the needs of the scientific community. In summary, CABIR will be used as one of the standard approaches for biodiversity data integration.

References

1. Merican, A.F., Othman, R.Y., Ismail, N., Cheah, K.P., Mok, L., Yin, Y.K.C., Kaur, S.: Development of Malaysian Indigenous Microorganisms Online Database System. *Asia Pac. J. Mol. Biol. and Biotech.* 10
2. Ando, K., Watanabe, M.: Global Taxonomy Initiative (GTI) and Taxonomy. National Institute of Evaluation and Technology and National Institute of Environment (2003)
3. Anthony, Leslie: The quest for data integration. Incyte Genomics, Inc. (2002), <http://www.incyte.com/insidegenomics/> (accessed September 5, 2010)
4. Buttler, D., Coleman, M., Critchlow, T., Fileto, R., Han, W., Liu, L., Pu, C., Rocco, D., Xiong, L.: Querying multiple bioinformatics data sources: can semantic Web research help? *ACM SIGMOD Record* 31(4) (2002)
5. Burhanuddin, M.: Biodiversity and Information in Malaysia. In: Workshop on Biodiversity Research and Information in Asia Oceania (2000), http://www.sp2000ao.nies.go.jp/english/whats_new/year_2000/abstract.html (accessed November 4, 2011)
6. Blum, S.: Overview of biodiversity informatics (2000), http://www.calacademy.org/research/informatics/sblum/pub/biodiv_informatics.html (accessed February 12, 2011)
7. CBD: Convention on Biological Diversity (2005), <http://www.biodiv.org> (accessed January 10, 2011)
8. GBIF: Global Biodiversity Information Facility, Copenhagen, Denmark (2004), <http://www.gbif.org> (accessed November 2, 2012)
9. IABIN: Inter-American Biodiversity Information Network (2004a), <http://www.iabin.net/english/index.shtml> (accessed November 8, 2010)
10. IMR: Institute of Medical Research (2006), <http://www.imr.gov.my/> (accessed October 8, 2011)
11. Nilges, M., Linge, J.P.: Bioinformatics, Paris, France (2002), http://www.pasteur.fr/recherche/unites/Binfs/definition/bioinformatics_definition.html (accessed June 10, 2010)

12. Sarinder, K.K.S., Majid, M.A., Lim, L.H.S., Ibrahim, H., Merican, A.F.: Integrated Biological Database Initiative (IBDI). In: Proceedings of International Conference on Biogeography and Biodiversity Wallace in Sarawak – 150 Years Later, Kuching, Malaysia (2005)
13. Schnase, J., Cushing, J., Frame, M., Frondorf, A., Landis, E., Maier, D., Silverschatz, A.: Information technology challenges of biodiversity and ecosystems informatics, pp. 339–345 (2003)
14. Secretariat on the Convention on Biological Diversity. Clearing-House Mechanism (CHM), Canada (2005), <http://www.biodiv.org/chm/default.aspx> (accessed January 10, 2011)
15. Wong, K.M.: Rimba Ilmu Botanical Garden, University Malaya, Kuala Lumpur (2005), <http://rimba.um.edu.my> (accessed July 15, 2010)

Using Ant Colony Optimization (ACO) on Kinetic Modeling of the Acetoin Production in *Lactococcus Lactis C7*

Nor Farhah Binti Saidin¹, Chuii Khim Chong¹, Yee Wen Choon¹, Lian En Chai¹,
Safaai Deris¹, Rosli M. Illias², Mohd Shahir Shamsir³, and Mohd Saberi Mohamad^{1,*}

¹ Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia
{nfarhah3, ckchong2, ywchoon2, lechai2}@live.utm.my,
{safaai, saberi}@utm.my

² Department of Bioprocess Engineering, Faculty of Chemical Engineering,
Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia
r-rosli@utm.my

³ Department of Biological Sciences, Faculty of Biosciences and Bioengineering,
Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia
shahir@fbb.utm.my

Abstract. Ant colony optimization (ACO) is a population-based meta-heuristics that can be used to find approximate solutions based on ant's behavior patterns for solving difficult combinatorial optimization problems. To apply Ant colony optimization (ACO), the artificial ants incrementally build solutions by moving on the graph as their results. The solution construction process is stochastic and is biased by a pheromone model. It is a set of parameters associated with graph components which either nodes or edges whose values are modified at runtime by the ants. Best value of kinetic parameters from the experimental data can be obtained by implementing the ant colony optimization (ACO). Model development that can represent biochemical systems is one of the hallmarks of systems biology. Scientists have been gathering data from actual experiments but there is a lack in computer models that can be used by scientists in analyzing the various biochemical systems more effective. However it is also time consuming and expensive and rarely produce large and accurate data sets when carried out in wet lab. Parameter estimation is used to adjust the model to reproduce the experimental results in the best possible way for a set of experimental data. *Lactococcus Lactis C7* will be used as the dataset and functions as the benchmark dataset for the experiments. The results were gathered by conducting some steps in SBToolBox. The result and discussions sections which include the comparison on the performances in term of computational time, average of error rate, standard deviation and production of graph. ACO shows a better result compared to the other algorithms like Simulated Annealing (SA) and Simplex algorithms. The method used in this research also can be used for other datasets as well.

* Corresponding author.

Keywords: Ant Colony Optimization (ACO), meta-heuristics, Lactic Acid Bacteria, Ant Behavior.

1 Introduction

For Biological investigations, there are several good reasons for inserting or adding kinetic modeling to improve skills. There are a lot of information and data contained in dynamic biological data than can be extracted by using simple procedures. If the collection of data that has just been collected contains an unknown answer, then an investment in computing can give a competitive advantage both in new knowledge and in productive experimental design.

Performing a lab experiment is costly as lab experimentation requires a substantial budget. Besides that, it rarely produces accurate data sets as because of the limit of accuracy of the measuring instruments. It is possible to calculate the limits of accuracy of different measuring instruments. Errors which are not quantifiable also occur in experiments and actual effect cannot be calculated.

Manually following the tradition, sometimes the experimental data is accepted as true, and theories have been modified to make the next prediction match the data. Meanwhile in diagram, the pathway leading from a mismatch to a new hypothesis is label as "imagination." This is purely human activity and computer helps nothing here. However, that modeling still helps because it enables human to know with precision what your imagination must accomplish. Finally, the advantage of the model for the design of informative experiments can be used.

Parameter estimation task is well known as an optimization problem that minimizes an objective function in measuring a generalized gap between the experimental data and model predictions [1]. Usually, researchers will use the Euclidean distance which is commonly referred to as least-squares error criterion. There are two objective functions commonly used which are concentration error based objective function and slope error based objective function. The concentration error based objective function is a straightforward method which calculates the sum of squared distances between the metabolite measurements and the predictions. Meanwhile, the slope error based objective function employs the decoupling techniques and uses the slope information for evaluating fitness of the function. The most effective methods for parameter estimation from time series data can be classified into gradient based methods, stochastic search algorithms and other algorithms.

Two main approaches, deterministic estimation and stochastic estimation are the two common approaches in parameter estimation. Deterministic optimisation can performs effectively in small scale linear optimisation problem. In contrast, the stochastic optimisation performance well with nonlinear problem but with no fixed setting for the control parameters; this effects the accuracy of the estimation [2]. In other words, when most of the interactions in system biology fail to be clearly studied and show its attribute, the implementaion of stochastic optimisation is very useful in such condition [3].

Global optimization method is one of the ways in parameter estimation. It needs high computational time complexity as well but they can find better values and the time

taken is not as much as experiments [4]. Ant Colony Optimization (ACO) algorithm is one of the global optimization methods that are appropriate to be used for the parameter estimation from experimental results. According to Alonso *et al.* [5], the ACO algorithm produced better results on many NP-hard problems compared to other evolutionary algorithms.

There are previous works that involved other algorithms such as Genetic Algorithm (GA) and Simulated Annealing (SA) on parameter estimation. According to Kikuchi *et al.* [6], using conventional simple genetic algorithm (SGA), inferring parameter values of small network but with very limited number of parameters, and the convergence rate is very low. There are two problems of SGA which are early convergence in the fast stage of search and evolutionary stagnation in the last stage.

For the SA, it is physically inspired method. The global and local search of the SA depends on the temperature. Temperature will decrease automatically, switching from global to local search. According to Gonzales *et al.* [7], he adapted the SA from the S-system parameter estimation from time series data.

This paper focuses on the parameter estimation in *Lactococcus lactis* C7 by estimating the parameter values of pyruvate metabolism by using Ant Colony Optimization (ACO) algorithm. This is because there are no researches that have been conducted on parameter estimation using ACO. The production that will be focused on is Acetoin production. There are also other substrates that can influence the production of acetoin such as acetolactate (acLac).

2 Materials and Methods

In this section, we describe the details of the proposed ACO into SBToolbox in MATLAB to estimate the parameter value. The dataset used is *Lactococcus lactis* C7. The parameter value for acetoin production is the main target to be estimated. From the previous work, the algorithms that are commonly used for parameter estimation are Genetic Algorithm (GA), Simulated Annealing (SA) and Simplex. However, for this dataset, there are no experiments ever conducted for the implementation of ACO in parameter estimation. Thus, we propose ACO to estimate the parameter in this research.

2.1 Datasets

This research uses *Lactococcus lactis* C7 as its dataset. The dataset was obtained from the BioModel database which was in SBML format (<http://biomodels.caltech.edu/BIOMD000000017>). In order to make it used in MATLAB, the SBML file (.xml) of the dataset needed to be converted into a format that could be read. This dataset was obtained from a study conducted by Marcel *et al* [8] to understand the regulations of pyruvate metabolism using a model based on measured kinetic parameters. This model was simulated by estimating the kinetic parameter values. The initial values of the kinetic parameters for experimental and simulated algorithms is shown in Table 1 for acetoin. There were 12 kinetics parameter to be estimated for acetoin.

Table 1. The Initial Value for experimental and simulated algorithm based on kinetic parameters of acetoin production

Kinetic parameters	Experi-mental Value	Simulated Value		
		ACO	SA	Simplex
R9_V_9	106	152.1	201.0231	152.1
R9_Kaclac_9	10	11.3	14.8018	11.3
R9_Kacet_9	100	80.4	122.5118	80.4
R10_V_10	200	495	483.9463	495
R10_Kacet_10	5	0.5	5.4955	0.5
R11_V_11	105	87.1	78.9765	87.1
R11_Keq_11	1400	1113.5	854.3025	1113.5
R11_Kacet_11	0.06	0.1	0.00852	0.1
R11_Knad_11	0.02	0	0.305	0
R11_Kbut_11	2.6	2.4	1.8348	2.4
R11_Knadh_11	0.16	0.2	0.1539	0.2
R14_k_14	0.0003	0	0.0004	0

Figure 1 shows the reactions included in the model to describe the distribution of carbon from pyruvate in *L. lactis*. Numbers in circles indicate the following enzymes or steps: 1, 'lumped' glycolysis; 2, LDH; 3, pyruvate dehydrogenase; 4, phosphotransacetylase; 5, acetate kinase; 6, acetaldehyde dehydrogenase; 7, alcohol dehydrogenase; 8, ALS; 9, acetolactate decarboxylase; 10, acetoin efflux; 11, acetoin dehydrogenase; 12, ATPase; 13, NOX; 14, non-enzymic acetolactate decarboxylation; 15, pyruvate formate lyase, which is considered not to be active under aerobic conditions; 16, chemical conversion to diacetyl, not included in the model. Substrates and products that were clamped in the model are indicated in italics.

2.2 Ant Colony Optimization (ACO)

This algorithm is inspired by the behavior of ants during the colony searching for the shortest path. The pheromones deposited by ants attract other ants which then will increase the pheromones. ACO is a probabilistic technique which solves the computational problems that have ability to reduce finding good path through nodes in graph. According to Zuniga *et al.* [9], he adapted ACO for S-system models by other nodes that were connected to it. They called the algorithm as discrete ACO. Meanwhile, the continuous ACO enhanced the aggregation pheromone system (eAPS) for parameter

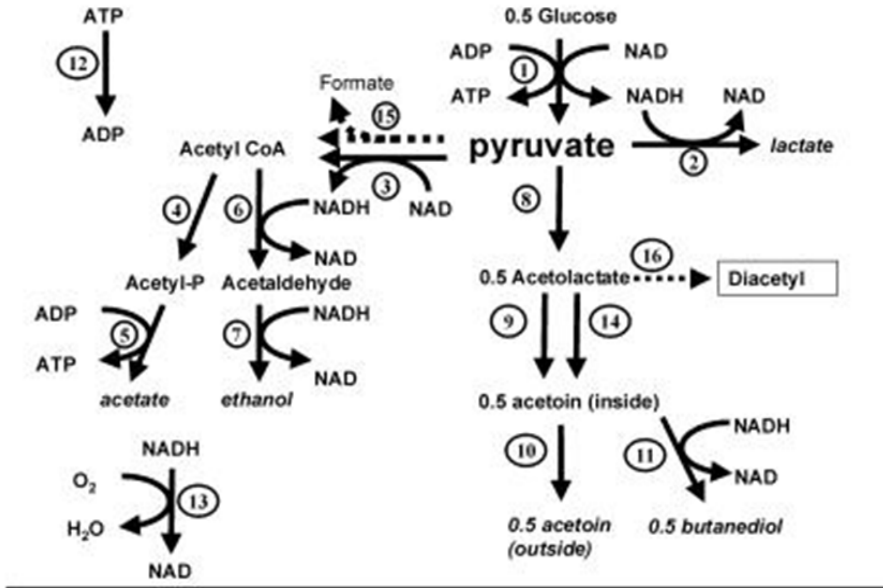


Fig. 1. Pyruvate metabolisms in *Lactococcus lactis* C7 (Marcel et al., 2002)

task involving S-system. Figure 2 shows the foraging of ants, Figure 3 shows the pseudocode of Ant Colony Optimization (ACO), and Figure 3 shows variable used in ACO's pseudocode.

2.3 System Biology Toolbox (SBToolBox)

MATLAB is a programming environment for algorithm development, data analysis, visualization, and numerical computation. MATLAB can solve technical computing problems faster than by using traditional programming languages, such as C, C++, and FORTRAN. The toolbox used is the Systems Biology Toolbox for MATLAB which offers systems biologists open and user extensible environment, in which to explore ideas, prototype, share new algorithms, and build applications for the analysis and simulation of biological systems. It features a wide range of functions. The Ant Colony Optimization (ACO) will be integrated into the SBToolBox(MATLAB) to obtain the best value of kinetic parameters from the experimental data.

2.4 COPASI

COPASI is one of the softwares used for simulation and modeling. Simulation and modeling have become a standard approach to understand complex biochemical processes. The parameter values gathered from the implementation of the algorithm in MATLAB will be used in this software to get the data needed for the evaluation of the metabolites; which is acetoin production that needs to be analyzed. In this research, COPASI was used in the process of evaluation of the algorithms performances.

```

Init pheromone  $\tau_{ij}$  ;
repeat for all ants i: construct solution(i);
    for all ants i: global pheromone update(i);
    for all ants edges: evaporate pheromone;
        ( $\tau_{i-j} := (1-\rho) \tau_{i-j}$ )

construct_solution(i):
init ant;
while not yet a solution:
    expand the solution by one edge probabilistically according
    to the pheromone;
        ( $\tau_{\rho i-j} / \sum_{\rho i-j} \tau_{\rho i-j}$ ;)

global_pheromone_update(i):
for all edges in the solution:
    increase the pheromone according to the quality;
        ( $\Delta\tau_{j-i} := 1/\text{length of the path stored}$ )

```

Fig. 2. Pseudocode of Ant Colony Optimization (ACO)

```

( $V=\{0, \dots, N\}, E=\{i \rightarrow j\}$ ) = directed acyclic graph
N = food source
 $\tau_{ij}$  = initial pheromone
 $\rho$  = pheromone deposited

```

Fig. 3. Variables used in ACO's pseudocode

2.5 Performance Measurement

For the evaluation, four performance measurements were used to evaluate the performance of Ant Colony Optimization (ACO) algorithm. The evaluation was conducted based on the time series data that had been generated. The time series data contain measurement result, y , and simulated results y_i for the algorithm. The first performance measurement used was the average of the error rate of the algorithm. The average of error rate is calculated by using the following equations:

$$e = \sum_{i=1}^N (y - y_i) \quad (1)$$

$$A = \frac{e}{N} \quad (2)$$

The algorithm has a better performance if the average of error rate is lower. The second performance measurement used was by calculating the standard deviation for 50 runs. The performance of the measurement was calculated using the equation below. If the values of the standard deviations are closer to zero, the performance of the algorithm is more accurate.

$$STD = \sqrt{\frac{e}{N}} \quad (3)$$

Production graph were also constructed. The production graph would be the third performance measurement to be used. The performances would be evaluated by comparing the simulated line and the experimental line. The closer the simulated line to the experimental line, the performances of the algorithm is more accurate and better. Last but not least, the performance measurement used was computational time. The shortest time taken for a run would be the best algorithm for the computational time.

3 Results

This section discusses on the performance of the algorithms by measuring and evaluating the standard deviation, average error rate, time and the accuracy of the algorithm. To clearly see the performance of the algorithms, a comparison between algorithms had been made, comparing the performance of Simulated Annealing (SA) and ACO. This research focuses on the acetoin production in *Lactococcus lactis* C7. Acetoin is very important in the fermentation and production of food products. Acetoin

is used as a food flavoring (in baked goods) and as a fragrance. It is also important in giving butter its characteristic taste. There are other substrates/metabolites that influence the production of acetoin such as acetolactate (aclac). For this paper, the metabolite that we focused on to estimate the parameter values is acetolactate (aclac) and acetoin production itself.

3.1 Acetoin Production

Figure 4 shows the graph production of acetoin in the pyruvate metabolism. The output graph shows four lines which represent experimental, simulated ACO, simulated simplex and simulated SA.

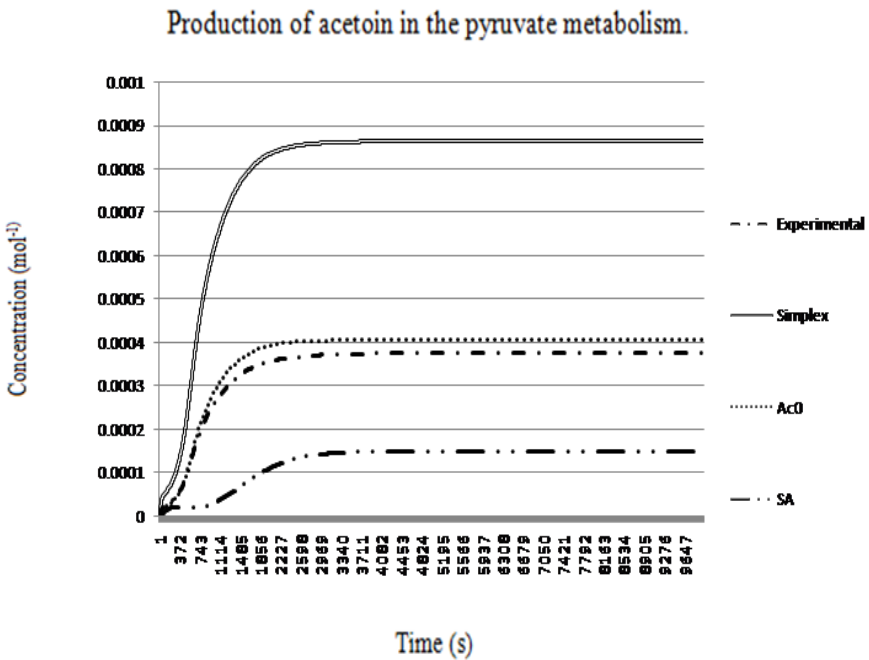


Fig. 4. Production of acetoin in pyruvate metabolism

Table 2 shows the average error rate of ACO and Simplex. The average error rate of ACO is 0.1033E-08, SA is 0.31598E-08 and the average error rate for Simplex is 2.1482E-08. For the standard deviation, the result of standard deviation using ACO is 0.7743E-08, SA is 4.3543E-08 and standard deviation for Simplex is 111.9990E-08. For measurement in term of time, the ACO computational time is 239.8957, SA is 240.7595 and Simplex is 1460.0173.

Table 2. 2 Comparison of average error rate and standard deviation using ACO, SA and Simplex for the production of acetoin in pyruvate metabolism

Algorithms \ Measurements	ACO	SA	Simplex
Average Error Rate	0.1033E-08	0.31598 E-08	2.1482E-08
Standard Deviation	0.7743E-08	4.3543E-08	111.9990E-08
CPU Time (seconds)	239.8957	240.7595	1460.0173

Note: Shaded column represents the best results.

4 Discussion

Figure 4 shows the graph production of acetoin in the pyruvate metabolism. The output graph shows four lines which represent experimental, simulated ACO, simulated simplex and simulated SA. The simulated line of ACO is closer to the experimental line if compared to simulated SA and simulated Simplex. This shows that the performance of ACO in parameter estimation is better than SA and Simplex.

Table 2 shows the average error rate of ACO and Simplex. The average error rate of ACO is 0.1033E-08, SA is 0.31598E-08 and the average error rate for Simplex is 2.1482E-08. From the average error rate, it shows that ACO has the smallest values of average error rate if compared to Simplex and SA. A low average error rate means the performance of the algorithm is good because the errors involved are few.

For the standard deviation, the algorithm is good in performance and consistent if it has a standard deviation that is close to zero. In Table 2, the result of standard deviation using ACO is 0.7743E-08, SA is 4.3543E-08 and standard deviation for Simplex is 111.9990E-08. The value of standard deviation of ACO is the closest to zero if compared to the standard deviation of Simplex and SA. This shows that performance of ACO is good, consistent and better than others. For measurement in term of time, the ACO computational time is 239.8957, SA is 240.7595 and Simplex is 1460.0173. ACO shows a better result than Simplex and SA by having a slight difference with SA and extreme difference with Simplex.

From all the results that have been discussed, ACO has the best performances. This is contributed by the advantages of ACO in finding good solution, enabling it to perform better for parameter estimation. This is because it builds a solution using local solutions, by keeping good solutions in memory. A completed tour is analyzed for the optimality by calculating the strength of the trails of pheromone. Each ant leaves a trail of pheromones when it explores the solution landscape. This trail is meant to guide other ants. More pheromone on trail increases the probability of trail being followed

and it will be repeated until most ants select the same tour on every run/cycle (convergence to solution). ACO updates the pheromone only for the (local or global) best ants. Besides that, the collective interaction through indirect communication, called stigmergy can also be influential to lead to good solutions.

5 Conclusion and Future Work

As conclusion, the performance of ACO in parameter estimation is quite good and better than Simplex and SA after being implemented into SBToolbox in MATLAB. The comparison graph shows that the simulated line of ACO is the closest to experimental line compared to other algorithms. Besides that, the result generated by ACO is more consistent because the standard deviation value is the closest to zero compared to Simplex and SA and the computational time is also the shortest compared to others. The attempt to develop algorithms inspired by observing ant behavior has the ability to find good paths based on probabilistic strengths of pheromones and has become the field of ant colony optimization (ACO). Thus, this makes ACO able to perform better compared to other algorithms in term of computational time and accuracy. Therefore, it proves that ACO can solve problems in a cheaper way without using wet lab experiments and with short computational time. Besides that, ACO, which is a global optimization method, is appropriate to be used for the parameter estimation from experimental results because of the performances shown. For future work, we can increase the number of program run in MATLAB to achieve more accurate results for the performance of the algorithm.

Acknowledgments. We would like to thank Malaysian Ministry of Science, Technology and Innovation for supporting this research by an e-science research grant (Grant number: 06-01-06-SF1029). This research is also supported by UTM GUP research grant that was sponsored by Universiti Teknologi Malaysia.

Competing Interests

The authors declare that they have no competing interests.

References

1. Maksat, A., Yves, F.N., Jaap, A.K., Joke, G.B.: Systems biology: parameter estimation for biochemical models. *FEBS Journal* 276, 886–902 (2009)
2. Lei, M.U.: Parameter Estimation Methods for Biological Systems. Master Thesis, The University of Saskatchewan, Canada (2010)
3. Liberti, L., Kucherenko, S.: Comparison of deterministic and stochastic approaches to global optimisation. *International Transactions in Operational Research* 12(3), 263–285 (2005)

4. Kahng, D.S., Lee, D.: Clustering Parameter Values for Differential Equation Models of Biological Pathway. In: Proceedings of the Second International Symposium on Optimization and Systems Biology (OSB 2008), Lijiang, China, October 31-November 3, pp. 265–270 (2008)
5. Alonso, S., Cerdón, O., Viana, I.F.D., Herrera, F.: Integrating Evolutionary Computation Components in Ant Colony Optimization Evolutionary Algorithms: An Experimental Study, pp. 148–180 (2004)
6. Kikuchi, M., Hajime, K., Kobayashi, S.: Moving target detection from infrared images using genetic algorithms. *Systems and Computers in Japan* 34(7), 76–86 (2003)
7. Gonzalez, O.R., Kuper, C., Jung, K., Naval, P.C.J., Mendoza, E.: Parameter estimation using simulated annealing for S-system models of biochemical networks. *Bioinformatics* 23(4), 480–486 (2007)
8. Marcel, H.N.H., Marjo, J.C.S., Dirk, E.M., Jeroen, H., Michiel, K., Iris, I.V.S., Roger, B., Hans, V.W., Jacky, L.S.: Metabolic engineering of lactic acid bacteria, the combined approach: kinetic modelling, metabolic control and experimental analysis. *Microbiology* 148, 1003–1013 (2002)
9. Zuñiga, P.C., Pasia, J., Adorna, H., del Rosario, R.C.H., Naval, P.: An ant colony optimization algorithm for parameter estimation and network inference problems in S-system models. In: Proceedings of International Conference on Molecular Systems Biology 2008 (ICMSB 2008), Manila, Philippines, pp. 105–106 (2008)

Semantic Rule-Based Determination of Cancer Stages from Free-Text Radiology Reports

Sangsoo Nam¹, Heung-Seon Oh¹, Jong-Beom Kim¹,
Sung-Hyon Myaeng¹, and Jinwook Choi²

¹ Department of Computer Science,
Korea Advanced Institute of Science and Technology
{sangsoo.nam, ohs, jongbeomkim, myaeng}@kaist.ac.kr

² Department of Biomedical Engineering,
Seoul National University
jinchoi@snu.ac.kr

Abstract. Cancer staging provides a basis not only for determining proper treatments for a patient but also for making future national-wide health plans. Despite its benefits, it is very difficult to obtain staging data because it is not commonly performed for all cancer patients or simply not collected. Moreover, it requires medical experts to do the analysis, incurring expensive cost. In this paper, we propose a method for semantic rule based determination of a cancer stage, which is considering a semantic type (e.g. body part, organ, or organ component). Compared to previous work, our work is unique in that we utilize radiology reports instead of pathological report since they are more available. Moreover, we argue that a rule-based approach is more suitable for cancer staging than machine learning because the international staging protocols specify certain conditions for determining stages. Since semantic type of words should be considered to determine the cancer stage and construct rules, we construct rules using MetaMap, which provides a meta-thesaurus of UMLS (Unified Medical Language System) for medical text. Based on our semantic rules, TNM (Tumor Nodes Metastasis) stages are determined for 275 reports of liver cancer. From our experiments, whole performance are highly incremented than machine learning approach.

Keywords: Cancer staging, Radiology reports, Semantic rule-base.

1 Introduction

In the medical domain, cancer staging aims at classifying the status of a patient's cancer into a set of defined stages based on certain criteria. It has significant benefits such as deciding proper treatments for a patient and analyzing national wide health statistics to make future health plans. These benefits motivated the definition of international standard protocols, including the TNM[1] (Tumor Nodes Metastasis), Okuda[2] and CLIP[3] (Cancer of the Liver Italian Program). Among them, the TNM standard of the Union for International Cancer Control (UICC) is widely adopted for

both research and practice because it was superior about prognostic predictive powers for survival[4]. TNM protocol determines the stage of cancer using three stages: T, N, M stage. Each stage is related to the primary tumor, lymph node metastasis, and distant metastasis respectively. In this paper, we utilize the modified version of the UICC (MUICC)[5].

Even though we can anticipate the benefits from cancer staging, it is very difficult to obtain staging data because it is not officially performed for all cancer patients or sometimes such data are simply not collected. Moreover, it requires medical experts to do that, incurring additional cost. However, it is possible to predict a cancer stage without medical experts because it can be determined solely based on medical reports. Clinical reports contain analysis results from radiological images while pathological reports have information from biopsy. As such, cancer staging is divided into clinical staging and pathological staging. Although pathological reports are more specific, reliable, and structured, they are more difficult to obtain than clinical reports.

Several researchers have attempted to deal with automatic cancer staging. Kovalerchuk[6] suggested statistical diagnostic rules to distinguish malignant and benign. McCowan et al.[7][8] proposed a system with binary SVM for lung cancer. Recent work [9] proposed a symbolic rule-based classification for improving generalizability to other types of cancer. These approaches are handling about pathological reports. However, we focus on clinical reports because they are assumed to be more practical and important in some aspects. First, an image is taken for most patients but a biopsy is performed only to the patients who had a surgery. Especially, Stage 4 patients do not usually receive surgery so that pathological reports are not available. Second, determining M stage of a patient depends on analyzing an image in most cases.

Sentences in radiology reports are usually so short and informal that there are not sufficient features to use for machine learning methods. We argue that it is more suitable to construct and use rules than applying machine learning if certain conditions such as MUICC summarized in Table 1 are available in advance. We can determine each stage of TNM by just checking out whether some of the conditions are met by rules. Since semantics of words are important for checking the conditions, we construct rules considering semantics of words through MetaMap, a program providing a meta-thesaurus of UMLS (Unified Medical Language System) for medical text.

Table 1. Summary of the modified UICC protocol

Aspect	Stage	Conditions
T: Primary Tumor	X	Primary tumor cannot be assessed
	1,2,3,4	Increasing size, number, and vascular invasion of tumor
N: Lymph Nodes	0	No regional lymph node metastasis
	1	Involvement of regional lymph nodes
M: Distant Metastasis	0	No distant metastasis
	1	Distant metastasis

In this paper, we propose a method based on semantic rules for cancer staging. For an input document, our method first splits a document into sentences and performs negation filtering. Then, abbreviations are replaced with full forms with a lexicon automatically constructed from MEDLINE abstracts. Finally, a cancer stage is determined by semantic rules, which is considering a semantic type of text.

To analyze clinical reports, abbreviation resolution is a critical task because abbreviations are ambiguous and omnipresent in clinical reports. We can construct a lexicon with abbreviation and full form pairs for a topic of interest, but it requires efforts involving medical experts. Several attempts have been made to alleviate the problem. Yu et al. [10] proposed a set of rules for mapping abbreviations to full forms. Schwartz et al. [11] proposed simpler rules by utilizing heuristics. Chang et al. [3] proposed a method based on logistic regression to score candidates of an abbreviation. Sohn et al. [12] proposed a set of strategies based on several specific rules to identify abbreviation and full form pairs.

Previous works extracted abbreviation and full form pairs well. However, abbreviations can have a different meaning on the different context or domain. Pakhomov et al. [13], Stevenson et al. [14] and other researchers have attempted to solve this word sense ambiguity problem with machine learning. However, we argue that ambiguity can be reduced more easily if we confine the domain. To construct a domain-specific lexicon, we constructed an abbreviation lexicon from a set of documents retrieved by a query that is a representative of a domain.

The contribution of this paper can be summarized as follows: 1) Utilizing radiology reports for automatic determination of cancer staging, instead of pathology reports due to its practicality and importance; 2) Invention of a semantic rule based cancer staging method with MetaMap; 3) Suggestion to solve the abbreviation ambiguity problem through construction a domain specific lexicon for abbreviations from the literature.

The rest of this paper is organized as follows. In Section 2, we briefly explain related work. In Section 3, our semantic rule based method is introduced in detail. Experimental results are described in Section 4. Finally, we will conclude with a discussion and future work in Section 5.

2 Related Work

The method in this article determines a stage of cancer by semantic rules for free-text medical reports. The following subsections briefly describe the context of related work.

2.1 Classification of Cancer Staging

McCowan et al. [7] regarded cancer staging as a common text classification problem and proposed a system with binary SVMs in document level. It first processed documents by mapping terms to UMLS lexicon and detecting negation. Then, binary SVMs were utilized to determine each stage of T and N. For experiments, 718 pathological reports of lung cancer patients were utilized.

In McCowan et al. [8], a system was proposed with two-step binary SVMs in sentence level. It attempted to combine various evidences by analyzing in sentence level. At the first step, a system checked out whether a document has T and N stages by passing through binary SVMs individually. At the second step, a number of detailed binary SVMs for different evidence were applied to each sentence. Then, the final stage of T and N were determined by combining those evidences from sentence-level SVMs. Experiments were performed with 710 pathological reports. The results showed that sentence-level analysis leads to performance improvements than document-level analysis. However, it requires much more annotation for every sentence in documents.

Nguyen et al. [9] proposed a symbolic rule-based classification based on SNOMED CT – Encoded CAP cancer checklist for reducing human effort and improving generalizability to other types of cancer. Their result showed slightly lower performance than previous approaches but they achieved generalizability. However, their approach is not adaptable for radiology reports since CAP cancer checklist was proposed for pathology reports.

2.2 Abbreviation Resolution

Numerous abbreviations are used in the medical reports for efficiency. However, it is often hard to know the full form of abbreviation without domain knowledge. To alleviate this problem, medical experts construct a domain lexicon but it needs expensive cost. Thus, several researchers have attempted to do automatically.

Yu et al. [10] and Schwartz et al. [11] proposed a set of patterns to construct a lexicon from MEDLINE abstracts. A set of rules was made carefully and manually. Since most of literature have specific rules of formatting abbreviations such as *full form (abbreviation)* or *abbreviation (full form)*, their methods performed well. Between them, the method by Schwartz [11] is known as more simple and fast since it was utilizing heuristic ways. It showed 0.82 and 0.96 in recall and precision respectively.

Chang et al. [16] assumed that possible abbreviations are inside parentheses, e.g. *full form(abbreviation)*. Then, they focused on finding relevant full form utilizing logistic regression. Features such as uppercase and letter position were used. Sohn et al. [12] proposed several strategies to identify the full form of abbreviation. For example, first letter strategy can be used for *computer assisted tomography* to capture abbreviation *CAT*. Strategy was determined based on the pseudo precision score for each abbreviation. The pseudo precision score was a statistical rate of the expected chance satisfaction for the strategy over the actual satisfaction.

These works were focused on extraction of abbreviation and full form pairs. However, abbreviation can have various meaning, so word sense ambiguity problem also arises. Pakhomov et al. [13] proposed a Maximum Entropy (ME) classifier to solve this. Surrounding words and sentence-level contexts were used as features.

In addition, Stevenson et al. [14] added more features such as UMLS Concept Unique Identifiers (CUIs), Medical Subject Headings (Mesh). Then, they compared the performance of learning algorithms including Vector Space Model (VSM), Naïve Bayes (NB), and Support Vector Machine (SVM).

2.3 MetaMap

MetaMap is an available program providing meta-thesaurus of UMLS (Unified Medical Language System)[17] for biomedical text [18]. MetaMap is widely used for medical text pre-processing because it provides useful functionalities such as acronyms/abbreviations detection, negation detection, and word sense disambiguation. It outputs concept score, concept name, preferred name, a semantic type of concept such as *a body part, organ, or organ component*, a source of the UMLS, e.g. MeSH (Medical Subject Headings) and SNOMEDCT (Systematized Nomenclature of Medicine -- Clinical Terms). Fig. 1 shows its human-readable output for the input text *lung metastasis*. Through this output, we can infer that lung is *a body part, organ, or organ component*.

```

Phrase: "lung metastasis"
Meta Candidates (14):
  1000 lung metastasis (Secondary malignant neoplasm of lung) [Neoplastic Process]
  861 Metastasis (Neoplasm Metastasis) [Neoplastic Process]
  861 metastasis (Metastatic Neoplasm) [Neoplastic Process]
  827 Metastases (Secondary Neoplasm) [Neoplastic Process]
  789 metastatic (Metastatic to) [Functional Concept]
  789 metastatic (metastatic qualifier) [Quantitative Concept]
  761 Metastat [Organic Chemical, Pharmacologic Substance]
  694 Lung [Body Part, Organ, or Organ Component]
  694 Lung (Entire lung) [Body Part, Organ, or Organ Component]
  638 Pulmonary (Pulmonary--:Point in time:^Patient:-) [Clinical Attribute]
  638 Pulmonary (Pulmonary (qualifier value)) [Qualitative Concept]
Meta Mapping (1000):
  1000 lung metastasis (Secondary malignant neoplasm of lung) [Neoplastic Process]

--- Score      -- Concept Name      --- Preferred Name      --- Semantic Type
    
```

Fig. 1. Human readable output of MetaMap for the input text ‘lung metastasis’

3 Method

The proposed method is shown in Fig. 2 where each sentence is first detected in radiology reports and then TNM stage is determined by semantic rules. A sentence is discarded if it has negative expressions. The rest of section describes the detailed steps of our method: sentence detection, negation filtering, abbreviation lexicon, TNM determination.

3.1 Sentence Detection

In radiology reports, sentences detected by a trained tokenizer or punctuation such as period and question mark are not effective for our method even if the results are correct. The reason is that such sentences are too short and informal. To deal with this, we utilized a heuristic where two sentences are regarded as a single one if

the latter follows the former, starting with a special character such as non-digit and non-alphabet. Fig. 3 shows these examples. In most cases, the latter sentence is directly related to the former one.

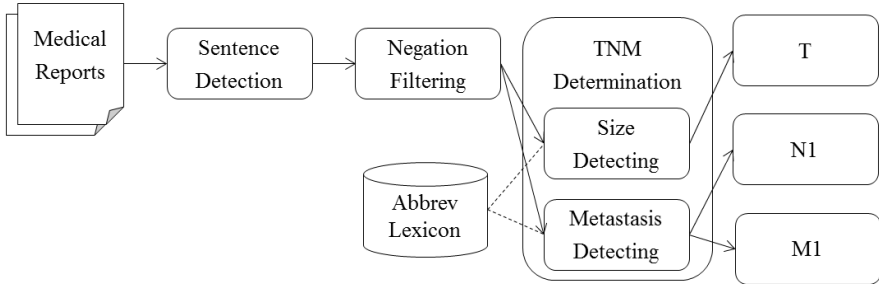


Fig. 2. Overview of proposed method

At least 3 high/low masses with fat component and capsular enhancement
 - 4.8cm in S8, 2.5cm in S8, 3.5cm in S2.
 - 2-3cm low/low nodule in S7 and left lateral segment: DN>

a few low attenuating nodules (>1.5cm) in S4, 7, 8 with focal arterial enhancement (arrow)
 --> early HCC or DN with HCC foci, more likely

A 2.2cm arterial enhancing/portal washout nodule in S3
 --> HCC, most likely.

Fig. 3. Sentence examples which start with a special character

3.2 Negation Filtering

Due to the characteristic of clinical reports, most sentences are about medical findings. In previous work, negative expressions are detected and replaced with a certain symbol by using NegEx [19]. We assume that negative findings are not effective for cancer staging because MUICC protocol only has the conditions with respect to positive findings. Negative sentences are discarded by two strategies. First, a sentence containing negations detected by MetaMap is filtered out. Second, a sentence containing a Korean negative expression such as “ $\frac{\text{아니}}$ ” and “ $\frac{\text{없어}}$ ” are also discarded.

3.3 Abbreviation Lexicon

Regardless of the types of medical documents, abbreviations are widely and frequently used because of efficiency in writing and effectiveness in practice. Unfortunately, an ambiguity problem arises in text processing since many abbreviations often have several different meanings. Moreover, new ones are created and introduced continually. Medical experts commonly construct an abbreviation lexicon for each domain to reduce this ambiguity when people read reports. However, it needs expensive human efforts.

Several researchers ([10], [12], [16], [20]) proposed to automatically construct a lexicon. Since they only focused on extraction of abbreviation and full form pairs,

word sense ambiguity problem still remains depending on different context or domain. Learning algorithms ([13][14]) have been utilized to alleviate this problem but we argue that it can be reduced more easily if we confine the domain. To extract abbreviation and full form pairs, Schwartz’s method is known as simple, fast and effective. The main idea of the algorithm is to extract preceding words associated with words in parenthesis as a pair of abbreviation and full phrase. Thus, we use the Schwartz’s method to construct a domain-specific lexicon. To construct a domain specific lexicon, we first choose a domain representative keyword as a query to the PubMed database. Because we focus on liver cancer, “*hepatocellular carcinoma*” is chosen empirically. By using the query, a set of MEDLINE abstracts is retrieved from PubMed database. Then, abbreviations with full forms defined in documents are extracted by the algorithm induced in [11]. Extraction results also have ambiguity. Rather than resolving one step further among several candidate phrases, we just select the most frequent phrase as a correct meaning based on the our assumption.

3.4 TNM Determination

In TNM, T (Tumor) stage is determined by three factors: size, number, and vascular invasion. Among these factors, we only focus on the size factor to determine T because it is difficult to extract the number of tumors and vascular invasion factors from radiology reports. Regular expressions are utilized to extract numerically described size –revealing tokens only (e.g. 1cm, 0.5mm). However, it is not guaranteed that they are attributes of liver cancer because many sizes are associated with different organs. To ensure appropriate associations, we retain sentences containing the size and the cancer name, *hepatocellular carcinoma* in our work. For example, *2cm* in the first sentence below is associated with a liver cancer but not the expression *14cm* in the second sentence.

1. 2cm size mass in S8 and it is likely Hepatocellular carcinoma.
2. LC with splenomegaly(14cm)

An N (Lymph nodes) stage is determined if metastasis to lymph nodes exists. M (Distant metastasis) is determined if metastasis to other organs exist. Since both are determined by metastasis, we assume that the word metastasis or metastatic is important evidence to find N1 and M1. Thus, we set both words as clue words and check k-size surrounding terms of clue words for N and M. When checking surrounding terms, we should consider the semantic of a word or phrase such as *a lymph node* or *a body part, organ, or organ component*. Since MetaMap provides a meta-thesaurus, we utilize it to find the semantic of word or phrase especially for a preferred name and a semantic type. If a sentence has a clue word and preferred names containing *lymph node* from k-surrounding terms, it is determined as N1. The procedure for M is similar to N except that we check semantic types of k-size surrounding terms from MetaMap are one of *Body Part, Organ, and Organ Component*. It is determined as M1 if any sentence of a document meets both of the conditions. About k-size surrounding terms, k can be any positive number, and we set it to 2 empirically for our experiment.

4 Experiments

4.1 Data

To evaluate the system, a corpus of de-identified medical reports with corresponding stage data was obtained from 275 liver cancer patients, followed by a research ethics approval. The radiology reports were collected over one-year period. The corpus was compiled from radiology reports for patients who were diagnosed with liver cancer for the first time in Seoul National University Hospital. TNM stages were assigned by a medical expert who had several years of experience in related fields. Table 2 shows data statistics of our test collection. It is relatively small but has similar distribution compared to the previous work.

Many reports are written in a mixture of Korean and English (see Fig. 4) although our analysis showed most technical terms were written in English. Thus, we decided not to translate the Korean text to find their semantics using MetaMap.

Table 2. Data statistics

Stage	T1	T2	T3	T4	TX
# of Reports	80	120	55	6	14

Stage	N0	N1	Stage	M0	M1
# of Reports	261	15	# of Reports	262	13

[Original Text]

Liver의 다른 portion에는 HCC의 evidence는 없고 Lt. lobe lateral segment의 upper portion에 enhance가 되는 portion이 있으나 washout되는 증거 없어 아마도 arterioportal shunt로 생각이 됨.

[Translated]

There is no evidence of HCC in other portion of Liver. Enhanced portion is found in upper portion of Lt. lobe lateral segment, but no evidence about washout. It is likely arterioportal shunt.

Fig. 4. An example report written in mixed language

4.2 Performance Measures

The measure for performance of our approach is given by precision, recall, and F1-score. Precision is the fraction of relevant instances that are retrieved, while recall is the fraction of retrieved instances that are relevant. F1-score is a measure that combines precision and recall with their harmonic means. Each measure can be calculated for each category. Then, results can be averaged across all patients to give micro-average results.

A confusion matrix (see Table 3) is also used to visualize the performance of an algorithm. It is a two-dimensional table of frequency counts according to classified or predicted class labels and actual class labels. Each classified result can be categorized: true positive (TP), true negative (TN), false positive (FP), false negative (FN). The results in false positive and false negative represent those incorrectly classified.

Table 3. Confusion matrix for baseline and test collection

	T1	T2	T3	T4	TX	Total
T1	0	0	0	0	0	0
T2	80	119	53	6	14	272
T3	0	1	2	0	0	3
T4	0	0	0	0	0	0
TX	0	0	0	0	0	0
Total	80	120	55	6	14	275

	N0	N1	Total
N0	260	15	275
N1	0	0	0
Total	260	15	275

	M0	M1	Total
M0	262	13	275
M1	0	0	0
Total	263	13	275

* Row and column values mean the numbers of matched in the baseline and in the test collection, respectively.

5 Results

To validate our method, we chose document-level and sentence-level SVMs as a baseline. However, because our data is written in a mixture of languages, text processing applied in previous work is not applicable. For that reason, we set our baseline as multi-class SVMs with a bag-of-words approach and 10-fold cross validation. Abbreviations were resolved based on our lexicon.

Table 4 shows the performance comparison between the baseline and our method. The baseline performance was 0.278, 0.918 and 0.929 corresponding to T, N, and M, respectively. Compared to the baseline, our method achieved performance improvements for all of the stages. The baseline performance of N and M shows relatively higher value than T but it is mainly due to the number of data about no lymph node metastasis(N0) and no distant metastasis(M0). About 95% of data are about N0 and M0. There was no correctly classified result about N1 and M1 in the baseline, shown in Table 3. Our analysis shows that the low performance of baseline can be caused by two reasons. First, there are numerous noisy data to train and classify for machine learning since sentences in radiology reports are short and informal. Second, the number of reports was not enough to find proper parameters.

Table 4. Micro average F1 for each stage

Stage	BaseLine	Our Method(Improvement)
T	0.278	0.636(+35.8%)
N	0.918	0.977(+5.9%)
M	0.929	0.986(+5.7%)

Our method achieved 0.636, 0.977 and 0.986 for T, N, and M, respectively. We obtained improvements of approximately 36% for T stage and 6% for both N and M stage. Even though the amount of increment N and M performance appear small, the

performance of N1 and M1 were highly increased, from 0 to 0.720 and 0.833 for N1 and M, respectively. Table 5 shows the incremented number of matches compared to Table 3. This increment is important because it shows that our method can find N1 (Lymph node metastasis) and M1 (Distant metastasis) stages even though the amount of the data is small. For classification of T, the proposed method increased the performance especially about T1 and T2 stages. Among the size, number, and vascular invasion factor of T, we only focused on the size factor to determine T. Since the stages of T3 and T4 are related to the size of the tumor but also with vascular invasion and the number of tumors, there was less performance improvement than the T1 and T2 stages.

Table 5. Confusion matrix for proposed method and test collection

	T1	T2	T3	T4	TX	Total
T1	52	5	5	1	0	64
T2	5	78	25	2	1	111
T3	0	2	6	1	0	9
T4	0	0	0	0	0	0
TX	23	35	19	2	13	92
Total	88	120	55	6	14	275

	N0	N1	Total
N0	259	6	265
N1	1	9	10
Total	260	15	275

	M0	M1	Total
M0	261	3	264
M1	1	10	11
Total	263	13	275

* Row and column value means the number of matched in the proposed method and in the test collection respectively.

As a pre-processing in our method, all abbreviations were converted to full forms, and sentences which have a negative expression were filtered out. Since we assumed that abbreviation resolution was a critical task, we validated usefulness of a lexicon by comparing the performance with and without a proposed lexicon. Since MetaMap provides an abbreviation mapping algorithm, abbreviations were handled basically by MetaMap when we find the semantics of words and phrases for N and M stages without using a proposed lexicon.

Table 6 shows its results. With a proposed lexicon, our method performs better for all stages than without the lexicon. Especially, there is a big difference about N1 since MetaMap outputs *Lobular Neoplasia* and *MLPH gene* for *LN* which stands for *lymph node* in the liver cancer domain.

Our lexicon also was compared with medical expert's lexicon. Compared results showed 0.70 (52/72) in accuracy. The accuracy was not high but most of incorrectly resolved abbreviations did not matter for determining TNM stages (e.g. *HA(hours ago)*, *CA(calcium)*, and *GS(general surgery)*). Thus, the performance between our lexicon and medical expert's lexicon is not that different. Approximately 0.1 F1 score increment was shown for each stage with medical expert's lexicon throughout the experiment.

Table 6. Use of a lexicon: F1 score for each stage

Stage	w/ lexicon	w/o lexicon
T1	0.727	0.727
T2	0.675	0.681
T3	0.187	0.158
T4	0	0
TX	0.245	0.238
N0	0.986	0.975
N1	0.720	0.235
M0	0.992	0.990
M1	0.833	0.800

Through our analysis results, we identified two major limitations in our method. First, clue words, ‘*metastasis*’ and ‘*metastatic*’, sometimes didn’t appear in the sentences containing information to decide N1 and M1 stage. Since clue words were the first condition to determine N1 and M1 in our method, false negative cases occurred. Second, our lexicon had a coverage problem even though it was extended from a large amount of literature. For example, *BLL(Bilateral Lower Lobe)* is commonly used in radiology reports but could not be found in the abbreviation lexicon even though the exact phrase occurs in the literature. Moreover, some abbreviations were not found in the literature because they are informal. For example, *met*s stands for *Metastasis*.

6 Conclusion and Future Work

In this paper, we proposed a semantic rule based method for cancer staging. Because of practicality and importance, we focused on radiology reports rather than pathology reports. To the best of our knowledge, this is the first attempt to utilize radiology reports for automatic analysis of cancer staging. We argue that because sentences in radiology reports are usually so short and informal, it is more suitable to construct and use rules than machine learning if certain rules and conditions are given in advance. Since semantics of words and phrases should be considered, in addition to the conditions, to construct rules, we apply the MetaMap to extract semantics of words on the medical reports.

Abbreviations are widely used in clinical reports but they create ambiguity problems. We constructed a domain specific lexicon with our belief that abbreviations would be less ambiguous in a specific domain, which was proven in our experiment. Our method was quite simple without human efforts but showed the better result than the baseline. Especially, there was an important improvement in determining lymph node metastasis (N1) and distant metastasis (M1) despite of the small amount of data.

Our future work will concentrate on three areas. First, the number of tumors and vascular invasions are also important features to build rules for T. Since both features are related with higher stages such as T3 and T4, we can achieve better performance

for higher stages by considering both features. Second, we empirically set k as 2 for checking a preferred name or a semantic type of k surrounding terms for N and M stages. It was to avoid an incorrect association between metastasis and an organ due to a long sentence. Since it was an empirical approach, we also plan to find a true association such as metastasis and an organ for better accuracy. Last, since domain representative keywords were selected manually for constructing a lexicon, whose effectiveness was shown through our experiments, we will investigate on a method for extracting them automatically to reduce human efforts.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology

References

1. Varotti, G., Ramacciato, G., Ercolani, G., Grazi, G.L., Vetrone, G., Cescon, M., Del Gaudio, M., Ravaioli, M., Ziparo, V., Lauro, A., Pinna, A.: Comparison between the fifth and sixth editions of the AJCC/UICC TNM staging systems for hepatocellular carcinoma: multicentric study on 393 cirrhotic resected patients. *European Journal of Surgical Oncology* 31(7), 760–767 (2005)
2. Okuda, K.: Natural History of Hepatocellular Carcinoma and Prognosis in Relation to Treatment. *CANCER* 56, 918–928 (1983)
3. Talian, L.I.I., Lip, P.R.C.: A New Prognostic System for Hepatocellular Carcinoma: A Retrospective Study of 435 Patients. *HEPATOLOGY* 28(3), 751–755 (1998)
4. Lu, W., Dong, J., Huang, Z., Guo, D., Liu, Y., Shi, S.: Comparison of four current staging systems for Chinese patients with hepatocellular carcinoma undergoing curative resection: Okuda, CLIP, TNM and CUPI. *Journal of Gastroenterology and Hepatology* 23(12), 1874–1878 (2008)
5. Ueno, G., Tanabe, S.: Prognostic performance of the new classification of primary liver cancer of Japan (4th edition) for patients with hepatocellular carcinoma: a validation analysis. *Hepatol Res.* 24(4), 395–403 (2002)
6. Kovalerchuk, B., Vityaev, E., Ruiz, J.F.: Design of consistent system for radiologists to support breast cancer diagnosis. In: *Proc. Joint Conf Information Sciences*, vol. 2, pp. 118–121 (1997)
7. McCowan, I., Moore, D.: Classification of cancer stage from free-text histology reports. *Engineering in Medicine and 1*, 5153–5156 (2006)
8. McCowan, I., Moore, D., Nguyen, A., Bowman, R.V., Clarke, B.E., Duhig, E.E., Fry, M.J.: Collection of cancer stage data by classifying free-text medical reports. *Journal of the American Medical Informatics Association* 14(6), 736 (2007)
9. Nguyen, A.N., Lawley, M.J., Hansen, D.P., Bowman, R.V., Clarke, B.E., Duhig, E.E., Colquist, S.: Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *Journal of the American Medical Informatics Association, JAMIA* 17(4), 440–445 (2010)
10. Yu, H., Hripesak, G.: Mapping abbreviations to full forms in biomedical articles. *Journal of the American Medical Informatics Association*, 262–272 (2002)
11. Hearst, M.A., Schwartz, A.S.: A simple algorithm for identifying abbreviation definitions in biomedical text. In: *Pacific Symposium on Biocomputing*, vol. 8, pp. 451–462 (2003)

12. Sohn, S., Comeau, D.C., Kim, W., Wilbur, W.J.: Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics* 9, 402 (2008)
13. Pakhomov, S.: Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In: *The Association for Computational Linguistics (ACL)*, pp. 160–167 (July 2002)
14. Stevenson, M., Guo, Y., Amri, A.A.: Disambiguation of biomedical abbreviations. In: *Proceedings of the Workshop on BioNLP*, pp. 71–79 (June 2009)
15. International Health Terminology Standards Development Organisation. SNOMED Clinical Terms User Guide, <http://www.ihtsdo.org/snomed-ct/>
16. Chang, J.: Creating an online dictionary of abbreviations from MEDLINE. *Journal of the American Medical Informatics Association* 9(6), 612–620 (2002)
17. NIH, Unified Medical Language System (UMLS), <http://www.nlm.nih.gov/research/umls/>
18. Aronson, A.R., Lang, F.-M.: An overview of MetaMap: historical perspective and re-cent advances. *Journal of the American Medical Informatics Association, JAMIA* 17(3), 229–236 (2010)
19. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics* 34(5), 301–310 (2001)
20. Schwartz, A.S., Hearst, M.A.: A simple algorithm for identifying abbreviation definitions in biomedical text. In: *Pacific Symposium on Biocomputing*, vol. 8, pp. 451–462 (2003)

Using Particle Swarm Optimization for Estimating Kinetics Parameters on Essential Amino Acid Production of *Arabidopsis Thaliana*

Siew Teng Ng¹, Chuii Khim Chong¹, Yee Wen Choon¹, Lian En Chai¹, Safaai Deris¹, Rosli M. Illias², Mohd Shahir Shamsir³, and Mohd Saberi Mohamad^{1,*}

¹ Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia
evelyn_siewteng@hotmail.com,

{ckchong2, ywchoon2, lechai2}@live.utm.my, {safaai, saberi}@utm.my

² Department of Bioprocess Engineering, Faculty of Chemical Engineering, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia
r-rosli@utm.my

³ Department of Biological Sciences, Faculty of Biosciences and Bioengineering, Universiti Teknologi Malaysia, Skudai 81310, Johor, Malaysia
shahir@fbb.utm.my

Abstract. Parameter estimation is one of nine phases in modelling, which is the most challenging task that is used to estimate the parameter values for biological system that is non-linear. There is no general solution for determining the nonlinearity of the dynamic model. Experimental measurement is expensive, hard and time consuming. Hence, the aim for this research is to implement PSO into SBToolbox to obtain optimum kinetic parameters for simulating essential amino acid metabolism in plant model *Arabidopsis Thaliana*. There are four performance measurements, namely computational time, average of error rate, standard deviation and production of graph. PSO has the smallest standard deviation and average of error rate. The computational time in parameter estimation is smaller in comparison with others, indicating that PSO is a consistent method to estimate parameter values compared to the performance of SA and downhill simplex method after the implementation into SBToolbox.

Keywords: Parameter Estimation; PSO; SBToolbox; *Arabidopsis Thaliana*.

1 Introduction

It is complex to understand the regulation, structure and organization of the underlying biological system because it needs quantitative assessment and reliable understanding of the system functions.

* Corresponding author.

Modeling is a process to transform the symbol model into a numerical model which enables us to understand the model deeply. It converts the biological system into a simple analogue that is easier to analyze, interrogate, predict, extrapolate, manipulate, and optimize than the biological system itself. There are 9 phases in mathematical modelling as shown in Figure 1 according to Chou and Voit[1]. At molecular level, the variables represent the concentration of chemical species such as protein, mRNA and so on. With the known pathway structure, we are able to write down the equation, which depends on several parameters. The parameters might be the reaction rate, production and decay coefficient, approximation or reduction that is satisfied by the structure of the system. Normally, the parameters are unknown. The measurement, if done experimentally, is expensive, hard and time consuming.

Estimation of parameter values is one of the steps in the modelling process. Parameter estimation helps to determine appropriate numerical parameter values that can convert the symbolic model into a numerical model and makes the latter consistent with experimental observations [1]. Among the nine phases, parameter values estimation is the most challenging task. This is due to the previous phases of parameter estimation that will affect the difficulties of the estimation. Examples are like the selection of modelling framework, the size and complexity of the hypothesized model and so on. It will be easier if the model is an explicit linear model that uses linear regression methods. Nevertheless, as soon as the model becomes nonlinear, many of these methods will become inapplicable [1].

In addition, biological model is nonlinear and dynamic. Hence, parameter estimation is complex because there is no general solution exists due to the model's nonlinearity. It is easier to analyze if it is a linear model since linear regression methods are used.

The model above describes the specific phenomena of biological system. It contains parameters that can alter the model behavior and it can be measured directly or inferred from the data. Parameter estimation is the process to determine appropriate numerical parameter values that can convert the symbolic model into a numerical model and makes the latter consistent with the experimental observations [1].

Optimization is a scientific discipline that deals with the detection of optimal solutions for a problem, among other alternatives. Optimization models the actual problem by building a proper mathematical function, or called as objective function. Among all feasible solutions where the solution fulfils all the constraints, global optimization tends to find the optimal one [2]. To estimate the parameter in a system, it is necessary to identify the objective function. Then, the objective function will be minimized by using appropriate optimization methods.

In order to simulate the biological system, parameter estimation is the most important phase because with complete and accurate set of parameter value, the system can be characterized. However, it is not always possible to measure these values in wet lab experiments due to high demands on cost and time, since there is no existing general solution to determine the nonlinearity of the dynamic model. Non-linear system is any problem that cannot be written as a linear combination of independent components and thus the result is not directly proportional to the input.

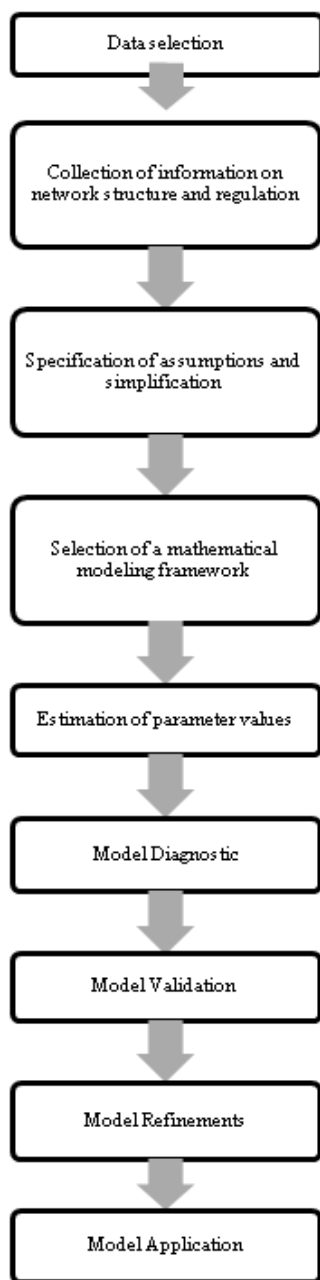


Fig. 1. Mathematical modelling [1]

As a result, it is difficult to obtain and researchers need to spend more time to solve the system since it needs to carry out the experiment within unknown time in order to get the best result. Furthermore, there are certain parameters which have no appropriate measurement method yet [3]. Exploration of several optimization techniques to minimize cost function is necessary to obtain the optimal value. Based on the research by Syed Murtuza Baker et al on the estimation of the kinetic parameters of upper part of glycolysis process [3], comparison of several methods were performed and the result stated that SA took the longest time in order to converge to the best solution. Even though GA was able to complete the estimation in a shorter time, it tended to be stuck in local minima. Moreover, PSO was able to produce better result compared to other methods.

There are several optimization methods in the SBToolbox such as Genetic Algorithm (GA)[4], Simulated Annealing (SA)[5], downhill simplex method[6] and so on. However, there has been no implementation of Particle Swarm Optimization (PSO) [7] to estimate kinetic parameters to simulate the essential amino acid metabolism in plant model *Arabidopsis Thaliana* yet. Furthermore, most of the parameter estimations used other algorithms such as SA, GA, EP (Evolutionary Programming) [3] and so on, and completed the set of kinetics parameters for aspartate metabolism by using appropriate method to estimate the kinetic parameter of aspartate metabolism which was not presented.

PSO is one of the methods based on swarm intelligence to estimate the kinetic parameter values. The concept of PSO is that the particles will fly in limited number of directions and have flying experience by their own or with their companion along the search space in certain velocity; and they are expected to fly to the best position.

In this research, PSO is proposed and implemented into SBToolbox in MATLAB to estimate the parameter values of aspartate metabolism in plant model *Arabidopsis Thaliana*. This method is inspired by bird flocks, fish schools and animal herds when foraging. The significance of the study is that there is no implementation of Particle Swarm Optimization (PSO) into SBToolbox to estimate kinetic parameters to simulate essential amino acid metabolism in plant model, *Arabidopsis Thaliana*, yet. PSO is a consistent method in estimating parameter values. It takes a shorter time to converge to the best value. It has the ability to find the optima in fast pace. Besides that, very few parameters are needed to adjust in order to obtain the optimal value. PSO is computationally inexpensive in terms of memory requirements and speed [8].

2 Method

Previous works have implemented GA, SA, downhill simplex method, and so on in parameter estimation. In this paper, we propose PSO as a new approach for parameter estimation. In this section, the details of the proposed Particle Swarm Optimization for estimating parameter values are discussed. The steps involved to obtain optimal parameter values are summarized in Figure 2.

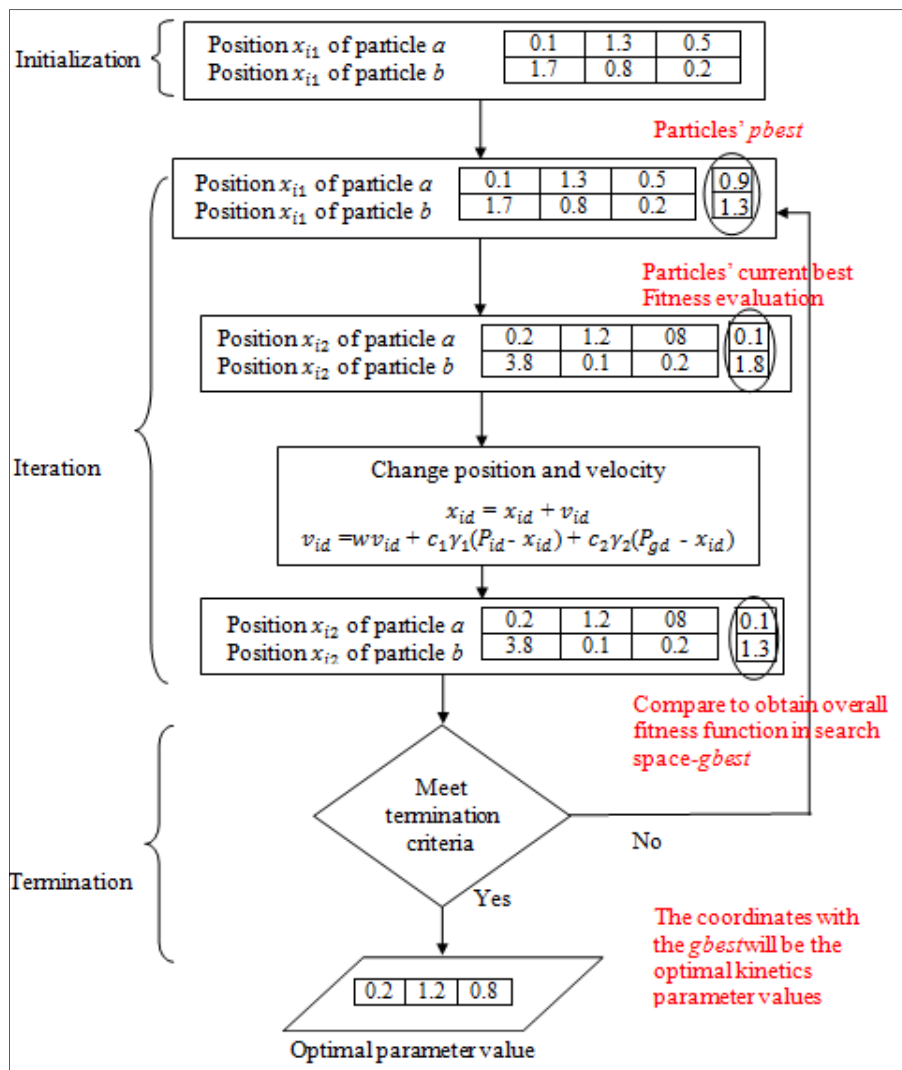


Fig. 2. Three steps involved to estimate parameter values using PSO

2.1 Initialization

Initially, the population array of particles with random positions and velocities on D dimensions in search space was initialized. Then, we defined the number of iterations, inertia weight, positive constant and swarm size. In this study, the inertia weight was 1.0, the positive constant was 2.0, and the number of iteration was 100. Next, the desired optimization fitness function in d variables for each particle was evaluated.

2.2 Iteration

In this part, a loop function was used to search and update the best position. There were two values being updated if best values were found in each iteration which were global best- gbest and best solution (fitness solution)- pbest value.

Initially, the particles' fitness evaluation was compared with particles' pbest. If current value is better than pbest, then set pbest value is equal with the current value and the pbest location equal to the current location in d-dimensional space. Then, we compared fitness evaluation with the population's overall previous best. If current value is better than gbest, then the gbest is reset to the current value. After being updated using Equation 1 and 2, the optimization fitness function in d variables for each particle was evaluated again.

$$x_{id} = x_{id} + v_{id} \quad (1)$$

$$v_{id} = wv_{id} + c_1\gamma_1(P_{id} - x_{id}) + c_2\gamma_2(P_{gd} - x_{id}) \quad (2)$$

Where $X_i = (X_{i1}, X_{i2}, \dots, X_{iD})$: ithparticle's position in search space, $V_i = (V_{i1}, V_{i2}, \dots, V_{iD})$: ith particle's velocity, $P_i = (P_{i1}, P_{i2}, \dots, P_{iD})$: Best position of the ith, $P_g = (P_{g1}, P_{g2}, \dots, P_{gD})$: Best position in the whole swarm, $i = 1, 2, \dots, m$, indicates each particle in one population. $d = 1, 2, \dots, D$, indicates the number of dimension, c_1, c_2 : Acceleration constant representing the pulling of each particle toward pbest and gbest. γ_1, γ_2 : Random number between 0 and 1, $v_{id} \in [-v_{max}, v_{max}]$, v_{max} : maximum velocity decided by user and w = inertia weight that provides the balance between global and local exploration and exploitation to find a sufficient optimal solution.

2.3 Termination

The loop continues until a criterion is met where optimum parameter values are obtained or a maximum number of iteration is reached.

2.4 Dataset

In this research, the dataset used was the aspartate metabolism [9] of Arabidopsis Thaliana. In this research, the kinetic parameters for Lysine, Threonine and Isoleusine were estimated using PSO in SBToolbox [10]. There were 9 kinetic parameters, 16 kinetics parameters, 6 kinetic parameters respectively. Table 1 shows the list of kinetic parameters that needed to be estimated, experimental values, the kinetic parameters values estimated using SA, simplex and PSO.

Table 1. List of kinetic parameters with measured kinetic parameter values for Lysine

Kinetic parameter	Measured kinetic parameter values	SA	Simplex	PSO
Vdhdps1_DHDP S1_k_app_exp	1	0.7019	0.9384	0.4726
Vdhdps1_DHDP S1_Lys_Ki_app_exp	10	10.1627	12.0480	10
Vdhdps1_DHDP S1_nH_exp	2	1.8208	1.9279	1.7768
Vdhdps2_DHDP S2_k_app_exp	1	1.0846	10	1
Vdhdps2_DHDP S2_Lys_Ki_app_exp	33	33.3325	34.5784	32.0637
Vdhdps2_DHDP S2_nH_exp	2	2	20	0.9687
VlysTRNA_Lys_ tRNAs_Lys_Km	25	15.0701	22.8179	35.1274
VlysKR_LKR_k cat_exp	3.1000	0.3430	3.1305	10.0065
VlysKR_LKR_L ys_Km_exp	13000	121600	12350	60575

3 Result and Discussion

In this study, PSO was implemented into SBToolbox in MATLAB to estimate parameter value. Three algorithms; SA, downhill simplex method and PSO were used to estimate the parameters and the result produced by two algorithms were compared. To evaluate the consistency and accuracy of both algorithms, the average of error rate and standard deviation were compared. There were 50 runs for estimating all the kinetic parameters and the formulas used to calculate the standard deviation are as follow:

$$e = \sum_{i=1}^N (y - y_i)^2 \quad (3)$$

$$A = \frac{e}{N} \quad (4)$$

$$STD = \sqrt{\frac{e}{N}} \quad (5)$$

The Equation 3 and 4 were used to calculate the error rate and average of error rate. Then, the standard deviation was obtained using Equation 5, where y_i is simulated results, y is measurement result, e is error rate, A is average of error rate and N is the number of sample. This equation was used to compare the performance of PSO with other methods. The best performance among the methods could be the method with the lower average of error rate and the standard deviation value close to 0 which indicated that PSO was able to produce high accuracy result.

After the discussion on the performance of PSO in estimating kinetics parameter of three amino acids, this section discusses and compares the performance of the three methods including PSO, SA and downhill simplex method. Based on Table 2, the standard deviation values of SA and downhill simplex method did not get close to 0 compared to standard deviation value of PSO. The values were 0.0733, 0.1211 and 0.0113 respectively. Meanwhile, the standard deviation values were 0.0733, 0.1211 and 0.0113 which PSO had the value that was the closest to 0. Based on Figure 3, the simulated line produced by PSO that was the closest to experimental line compared to SA and downhill simplex method. Having the smallest average of error rate, standard deviation value closer to 0 and simulated line closest to experimental line shows that PSO is a more consistent method to estimate parameter values compared to SA and downhill simplex method. In addition, the computational time for PSO to estimate 9 kinetics parameters was 315.9816 seconds which took a shorter time to complete compared to SA which took 4834.0581 seconds and 585.9037 seconds for downhill simplex method. The smaller average of error rate, standard deviation value closer to 0 and simulated line closest to experimental line shows that PSO is a more consistent method to estimate parameter values compared to SA and downhill simplex method. In addition, the computational time for PSO to estimate 9 kinetics parameters was 315.9816 seconds which took a shorter time to complete compared to SA and downhill simplex method. In addition, the computational time for PSO to estimate 9 kinetics parameters was 315.9816 seconds which took a shorter time to complete compared to SA which took 4834.0581 seconds and 585.9037 seconds for downhill simplex method. The smaller average of error rate, standard deviation value closer to 0 and simulated line closest to experimental line shows that PSO is a more consistent method to estimate parameter values compared to SA and downhill simplex method. In addition, the computational time for PSO to estimate 9 kinetics parameters was 315.9816 seconds which took a shorter time to complete compared to SA and downhill simplex method. We have conducted 50 runs with three algorithms and the STD values are shown in Table 2. The results showed that PSO has the lowest STD value; this indicates that the different between each run is small and this proved that it is a reliable estimation algorithm.

PSO had the smallest average of error rate, standard deviation values closer to 0 and the simulated line closer to the experimental line. The results obtained show that PSO outperformed SA and simplex in estimating kinetics parameters of Lysine, threonine and Isoleucine. It also shows that PSO is the most consistent method used in this research. The use of GA to estimate the kinetics parameters easily gets stuck in local minima and as a result, the accuracy of the kinetics parameters values will be low. This can be solved by using PSO due to the inertia weight taken into account in

PSO which was able to avoid being stuck into local minima by increasing the global search ability. The inertia weight produced the balance between the local and global exploration and exploitation. The computational time used to estimate the kinetics parameters is higher by using other algorithms and this can be solved by using PSO, proven by the short time taken in this research. This is the result of PSO which is inspired by bird flocking, fish schooling etc which does not require generation of new population for each iteration, which is time-consuming, but each particle from the same population will fly to better solution in each iteration. Hence, this decreases the time complexity. Furthermore, the steps involved in PSO are less complex compared to other algorithms such as GA which need to undergo selection, mutation and crossover. Besides that, the appropriate acceleration constant in PSO is able to ensure each particle fly towards pbest and gbest, which then lets PSO be able to converge to the best solution faster compared to other algorithms. If the constant value is too low, the particle will tend to fly away from the best solution; at the same time the high value of acceleration constant will make the particle pass the target.

Table 2. Comparison of average of error rate, standard deviation and execution time in seconds between SA, downhill simplex method and PSO for Lysine production from *Arabidopsis Thaliana*

Method \ Feature	SA	Downhill simplex method	PSO
Computational time (second)	4834.0581	585.9037	315.9816
Average of error rate	0.0318	0.1520	0.0057
Standard deviation	0.0733	0.1211	0.0113

Note: Shaded column represents the best results.

PSO had the smallest average of error rate, standard deviation values closer to 0 and the simulated line closer to the experimental line. The results obtained show that PSO outperformed SA and simplex in estimating kinetics parameters of Lysine, threonine and Isoleucine. It also shows that PSO is the most consistent method used in this research. The use of GA to estimate the kinetics parameters easily gets stuck in local minima and as a result, the accuracy of the kinetics parameters values will be low. This can be solved by using PSO due to the inertia weight taken into account in PSO which was able to avoid being stuck into local minima by increasing the global search ability. The inertia weight produced the balance between the local and global exploration and exploitation. The computational time used to estimate the kinetics

parameters is higher by using other algorithms and this can be solved by using PSO, proven by the short time taken in this research. This is the result of PSO which is inspired by bird flocking, fish schooling etc which does not require generation of new population for each iteration, which is time-consuming, but each particle from the same population will fly to better solution in each iteration. Hence, this decreases the time complexity. Furthermore, the steps involved in PSO are less complex compared to other algorithms such as GA which need to undergo selection, mutation and crossover. Besides that, the appropriate acceleration constant in PSO is able to ensure each particle fly towards pbest and gbest, which then lets PSO be able to converge to the best solution faster compared to other algorithms. If the constant value is too low, the particle will tend to fly away from the best solution; at the same time the high value of acceleration constant will make the particle pass the target.

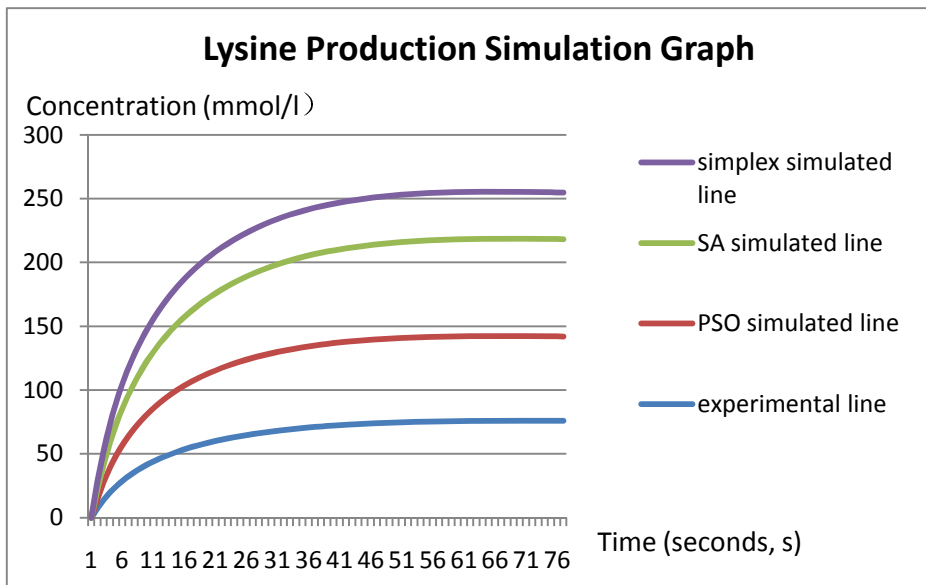


Fig. 3. Comparison of simulated line of SA, downhill simplex method and PSO with experimental line for Lysine production

4 Conclusion

In conclusion, the performance of PSO in estimating parameter values is better than SA and downhill simplex method after the implementation of PSO into SBToolbox in MATLAB. The simulated results generated by PSO are more consistent, as the standard deviation value is closer to 0 compared to SA and downhill simplex methods. The graph also shows that the simulated line of PSO is closer to experimental line. Moreover, the computational time to estimate parameter values for SA and downhill simplex method are longer compared to PSO. This is due to PSO which applies inertia weight to obtain a balance between the local and global exploration and exploitation to

avoid getting stuck into the local minima. In addition, PSO takes a shorter time to converge to best solution. Besides that, the acceleration constant that is taken into account in the equation ensures that each particle is pulled towards the pbest and gbest positions. In this research, value 2 was applied. In conclusion, Parameter Estimation through experiment is time consuming, hard and expensive. However, the implementation of PSO into SBToolbox manages to reduce the computational time for parameter estimation. It also reduces the complexity and the cost needed to use to estimate the kinetics parameters since the estimation only involves the use of computer. For future work, the number of run may be increased to ensure the accuracy of the method and more different weight parameters can be implemented to enhance the performance of PSO.

Acknowledgments. We would like to thank Malaysian Ministry of Science, Technology and Innovation for supporting this research by an e-science research grant (Grant number: 06-01-06-SF1029). This research is also supported by UTM GUP research grant that was sponsored by Universiti Teknologi Malaysia.

References

1. Chou, I.C., Voit, E.O.: Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Mathematical Biosciences* 219, 57–83 (2009)
2. Parsopoulos, K.E., Vrahatis, M.N.: *Particle Swarm Optimization and Intelligence*. Hersey, New York (2010)
3. Baker, S.M., Schallau, K., Junker, B.H.: Comparison of different algorithms for simultaneous estimation of multiple parameters in kinetic metabolic models. *Journal of Integrative Bioinformatics* 7(3), 133 (2010)
4. Houck, C.R., Joines, J.A., Kay, M.G.: A genetic algorithm for function optimization: a Matlab implementation. Technical Report: NCSU-IE-TR-95-09, North Carolina State University, Raleigh, NC (1995)
5. Aarts, E., Korst, J.: *Simulated Annealing and Boltzmann Machines. A Stochastic Approach to Combinatorial Optimization and Neural Computing*. John Wiley & Sons, New York (1989)
6. Nelder, J., Mead, R.: The downhill simplex method. *Computer Journal* 7, 308–313 (1965)
7. Eberhart, R., Shi, Y.: Particle swarm optimization: developments, applications and resources. In: *Proc. Congress on Evolutionary Computation*, Service Center, Piscataway, NJ, Seoul, Korea (2001)
8. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: *Proc. of IEEE International Conference on Neural Networks (ICNN)*, Perth, Australia (1995)
9. Curien, G., Bastien, O., Robert-Genthon, M., Cornish-Bowden, A., Cárdenas, M.L., Dumas, R.: Understanding the regulation of aspartate metabolism using a model based on measured kinetic parameters. *Mol. Syst. Biol.* 5(271), 1–32 (2009)
10. Schmidt, H., Jirstrand, M.: *Systems Biology Toolbox for MATLAB: A computational platform for research in Systems Biology*. *Bioinformatics* 22(4), 514–515 (2005)

Content-Based and Similarity-Based Querying for Broad-Usage Medical Image Retrieval

Christopher Town

University of Cambridge, Computer Laboratory,
15 JJ Thomson Avenue,
Cambridge CB3 0FD, UK
cpt23@cam.ac.uk

Abstract. Health-related information, much of it consisting of images, is being predominantly accessed online by diverse groups of users ranging from medical professionals and researchers to students and the general public. This paper argues that broad-usage medical image retrieval is best approached as a sub-domain of generic image search. We discuss how search over a diverse corpus of biomedical and healthcare related images can benefit from a modern content-based image retrieval (CBIR) system based upon general photographic content classification techniques. The system features a flexible query language based upon a generic image concept ontology which can utilise both metadata (where available) and automatically extracted image content descriptors. Furthermore, the system supports both text-based querying as well as similarity-based searching and is thus well suited to iterative refinement of initial search results without the need for specialist knowledge of relevant keywords.

Keywords: medical images, content-based image retrieval, similar image search, ontological query languages, healthcare information systems.

1 Introduction and Related Work

An ever larger amount of healthcare related information is being searched and accessed via online resources such as health related websites, wikis, internet search engines, blogs, forums, and social media [13]. Images play an important role in a variety of tasks related to healthcare, including (self)diagnosis, medical record keeping, teaching and research, treatment planning, and lifestyle management [10].

1.1 Medical Image Retrieval

Medical image retrieval systems have traditionally been reliant upon manually generated high-quality metadata and optimised for particular interface modalities and retrieval requirements. Many such systems are poorly suited to cope with the huge diversity of online healthcare information and the differing needs of groups such as medical practitioners, students, healthcare administrators, patients, and carers [12].

One of the strongest trends in image retrieval research in recent years has been a move to more advanced methods for image analysis, indexing, and retrieval [5]. Modern computer vision and pattern recognition techniques allow image search to be increasingly based on content rather than purely on metadata or context, and this has given rise to the notion of content-based image retrieval (CBIR).

As noted in [10] and [12], the adoption of CBIR techniques for medical image retrieval tasks has been hampered by a variety of factors which have limited its impact as an aid for diverse needs such as patient information, diagnosis, clinical care, biomedical research, and education. Automated image analysis and CBIR techniques are starting to make an important contribution to biomedical image retrieval [11], but are best established in specialist domains where image analysis is an inherent part of professional practice, such as radiology and pathology [2]. A related application is computer-aided diagnosis (CAD) [1] where image retrieval facilitates medical imaging experts by providing a computational “second opinion”.

It is apparent that most biomedical CBIR systems are targeted at specialist users [12] and are not well suited for the much broader set of stakeholders who have regular or occasional medical image retrieval needs. There has been a strong trend towards user participation in all areas of the healthcare delivery process [13]. At the same time, recent research [7] has shown that the three most common sources of online information used by physicians and the general public are (in decreasing order of usage)

- General search engines (e.g. Google, Bing, Yahoo!)
- Medical research databases (e.g. Pubmed) and websites providing health information
- Wikipedia

1.2 Query Paradigms for Broad-Usage Biomedical Image Search

Despite significant research efforts into provision of customised medical image retrieval interfaces ([6], [2], [1], [11]), both medical professionals and general users are accustomed to keyword-based image search [12]. However, recent research [22] has revealed that initial keyword based query formulation can pose significant problems to medical professionals. In particular, the quality and availability of metadata varies substantially and query formulation and query refinement [6] are problematic when searching for visual biomedical information.

One solution is to combine multiple search modalities [17]. In order to minimise user effort, there has been growing interest in providing image similarity search systems [16] which offer a “query by example” paradigm in which images themselves serve as queries. Despite some early research efforts [14] on creating medical CBIR systems that enable effective query-by-example search, the problem of encoding effective image-to-image similarity measures has primarily been approached in a domain specific manner.

Examples of such methods are probabilistic feature clustering [8] and part-supervised concept classification [15], but most such schemes are very sensitive to the composition of the dataset and work best on relatively homogenous datasets for applications such as histology or radiology. Scalability has also been an issue, although this is being ameliorated by the increasing usage of technologies such as grid computing ([3], [19]) and index partitioning [9] .

1.3 Proposed Approach

In order to make biomedical image retrieval effective and accessible for the widest possible audience, and in order to overcome the “chicken and the egg” conundrum of how a user may obtain an initial image to serve as the basis for a similarity query, this paper argues that general broad-usage biomedical image retrieval is best approached by a system offering two complementary search modalities:

- *Exploratory search*: the initial search is conducted using a query language interface which internally parses the query both syntactically and lexically in order to provide search results that are based both on metadata and on an automated content classification of all the images in the index.
- *Similarity-based refinement*: the user performs relevance assessment of a set of search results and can query the system for additional results which are similar (based on visual attributes, content classification, and annotations) to those images he or she deems most relevant.

We describe a system called *Imense Picturesearch* which is based on automated analysis and recognition of image content. It features a range of image processing and analysis modules, including image segmentation, region classification, scene analysis, object detection, and face recognition methods. In addition to language-based querying, the system provides image similarity search to allow rapid query refinement. We describe the core features of this system and demonstrate its use on a general corpus of several million photographic images. The image data contains a subset of healthcare related image content but is not specifically oriented towards biomedical retrieval, thus demonstrating the usage of such a system for healthcare related image search within a broader corpus, as would be the case for biomedical image retrieval on the internet.

2 Content-Based Image Retrieval System

This paper describes aspects of a content-based image retrieval system called Imense Picturesearch¹. It consists of a novel CBIR system featuring automated analysis and recognition of general photographic image content, and an ontological query language. The underlying technology is being used by several large commercial image collections with up to 25 million images, but this paper will

¹ <http://picturesearch.imense.com>

limit its discussion to the technology and content available via the main Imense Picturesearch website.

At present, this website provides an index of over 3 million stock photography images from around 40 different image providers, many of which are repositories comprising multiple sub-collections. This image corpus is extremely heterogeneous as the images differ greatly in their content, intended usage, and the availability and quality of metadata. An estimated 100,000 of these pictures are related to biomedical and healthcare themes. This is likely to reflect the composition of images available on the internet as a whole: images depicting biomedical and healthcare themes are a large but minority share of all images and are not always clearly identifiable as such based on metadata or website context, thus motivating a need for automated content inference and content-aware image retrieval functionality.

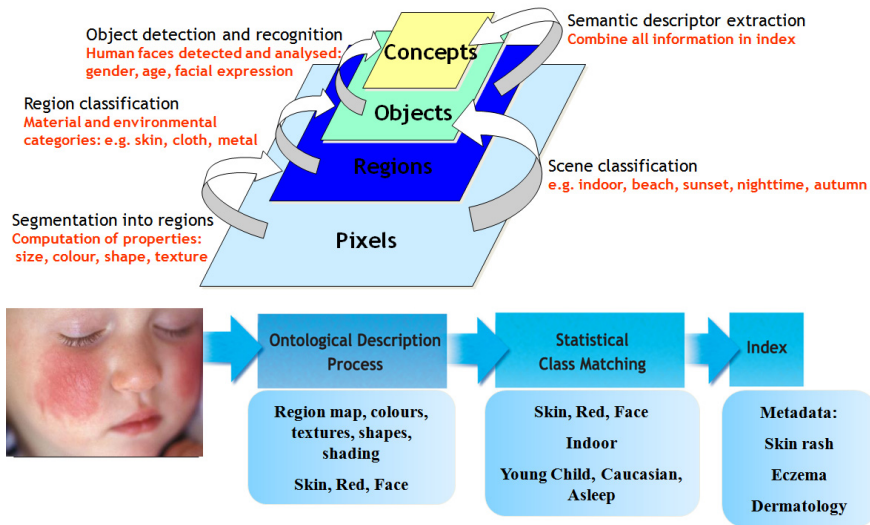


Fig. 1. Diagrammatic overview of the image analysis and recognition processes carried out by the Imense Picturesearch system

2.1 Image Content Inference and Indexing

The Imense Picturesearch system performs a range of image analysis procedures comprising recognition of visual properties, such as colour, texture and shape; recognition of materials, such as grass or sky; detection of objects, such as human faces, and determination of their characteristics; and classification of scenes by content, for example indoor, forest or sunset. The system uses semantic and linguistic relationships between terms to interpret user queries and retrieve relevant images on the basis of the analysis results. Moreover, the system is extensible,

so that additional image classification modules or image context and metadata can be integrated into the index. Some of the underlying concepts are discussed in [21], [18], and [19].

As illustrated in figure 1, image content analysis and index generation consists of the following main steps:

- *Image segmentation*: In order to identify salient parts of an image corresponding to objects or object parts, the image is automatically segmented into a covering set of non-overlapping regions and attributes such as size, colour, shape, and texture are computed for each region. The number of segmented regions depends on image size and visual complexity, but has the desirable property that most of the image area is usually contained within a few dozen regions which closely correspond to the salient features of the picture.
- *Region classification*: Segmented regions are then automatically classified according to a predefined set of material and environmental categories, such as “skin”, “cloth”, “hair”, “metal”, etc. Sophisticated statistical machine learning methods are employed to yield a highly reliably probabilistic classification of the image. This may be regarded as an intermediate level semantic representation which serves as the basis for subsequent stages of visual inference and composite object recognition. The spatial relationships between major regions are also encoded using a graph based representation.
- *Scene classification*: A second stage of classifiers is applied to analyse image content at a higher semantic level. Examples of scene categories include “indoor”, “office”, “text”, “illustration”, “closeup/macro”, etc.
- *Object detection and recognition*: The image analysis also includes detectors for common objects. For example, human faces are automatically detected and classified according to personal attributes such as gender, age, and facial expression.
- *Index generation*: Once all image analysis stages have been applied, then all the information from the various classifiers and recognisers is combined into a special indexing format which supports fast content based image retrieval.

2.2 Grid-Based Image Indexing

One of the drawbacks associated with CBIR is the increased computational cost arising from tasks such as image processing, feature extraction, image classification, and object detection and recognition. Consequently CBIR systems have suffered from a lack of scalability, which has greatly hampered their adoption for real-world public and commercial image search. At the same time, paradigms for large-scale heterogeneous distributed computing such as grid computing, cloud computing, and utility based computing are gaining traction as a way of providing more scalable and efficient solutions to large-scale computing tasks.

In order to provide scalability to the vast image collections that are being accumulated, we have also made use of grid processing technology. Image analysis is well suited to grid computing since the processing stages are intrinsically sequential and take up to 10 seconds of single core CPU time for high-resolution images. In order to benefit from parallelisation, it was decided to parallelise at the granularity of single images or small subsets of images. Each image can therefore be processed in isolation on the grid, and such processing takes no more than a few seconds.

As detailed in [20], our grid processing has thus far been restricted to GridPP [4], which is the UK part of a very large scale (over 120,000 CPU) processing grid set up by the international particle physics community. In order to minimise overheads, several hundred images are automatically agglomerated into a batch which is then submitted for processing via the Ganga job submission and control framework, with the results of image processing and analysis being passed back to the submission server upon successful completion.

Each job was given a list of several hundred images to process, which were downloaded to a worker node. After processing, results were uploaded to a grid storage element. Ganga, running on the submission machine, continually monitored job status, and automatically retrieved the outputs of completed jobs. The status of jobs at each site was checked every 10 minutes and new jobs submitted via the gLite workload-management system

Over 25 million high resolution images have been processed and indexed using this approach thus far. Grid computing techniques are increasingly being utilised for biomedical image analysis ([13], [3], [2]).

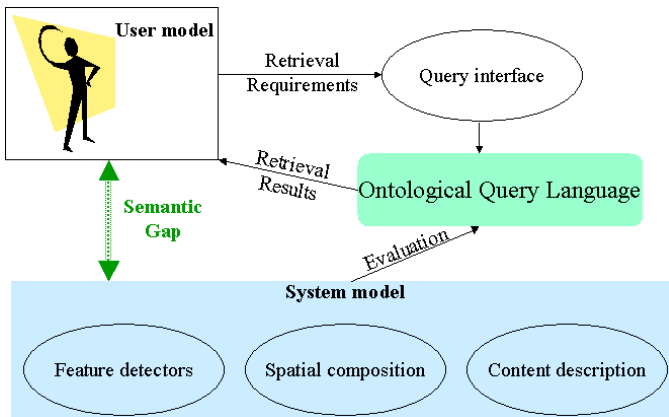


Fig. 2. Model of the retrieval process using an ontological query language to bridge the semantic gap between user and system notions of content and similarity

2.3 Ontological Query Processing

CBIR systems are particularly prone to creating a wide *semantic gap* between human and computer capabilities for interpreting image content. One approach that has been proposed for bridging this gap is the use of an ontological query language, as shown in figure 2 ([21], [19]).

User queries are parsed into a canonical representation which is then linked to automatically recognised image content in accordance with the retrieval need expressed by the query. The underlying ontology encompasses relational information about concepts and attributes pertaining to automatically recognised image content, as well as knowledge about the structure and meaning of natural language queries expressed in English. The relevance of each image in a collection with respect to a given user query is assessed probabilistically while taking into account both the reliability and salience (as it pertains to the query) of all information available for that image.

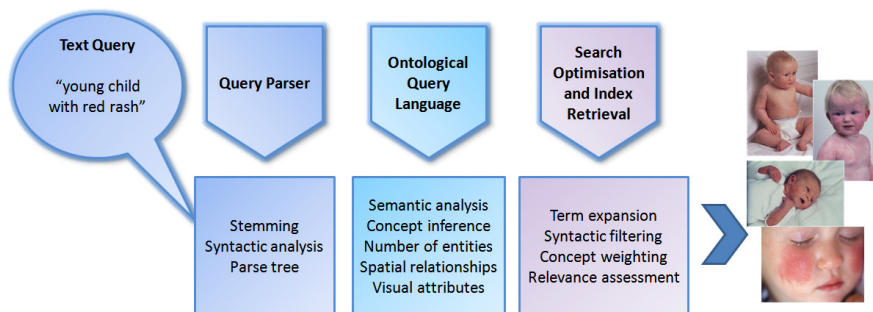


Fig. 3. Example of the ontological query processing and retrieval process

A key strength of this approach to image analysis is that it is based on concepts rather than particular keywords. For this reason the system is in principle not tied to a specific natural language. Furthermore, the content classifiers are probabilistic, which means that they give some indication of the degree to which a given concept is applicable to a particular image or part of an image. This property is very useful for ensuring a high precision for search results, since images can be ranked with respect to how well they match the search terms in the query rather than the mere presence or absence of a given keyword.

In addition, since the classifiers pertain to the content of the image itself, we are also able to provide search over the visual and spatial composition of the actual picture, which is something that is very difficult to realise by means of keywords. For example, we can cater for queries such as “green centre purple background” (which will be interpreted differently from “purple centre green background”) or “3 people in the forest” or “patient in bed”. Figure 3 illustrates this process.

2.4 Similarity Search

As described in section 1.3, our system also offers search based on image similarity. Having selected a relevant image on the basis of search results obtained from an initial textual query, the user is presented with an ordered list of images that are deemed to be similar to it.

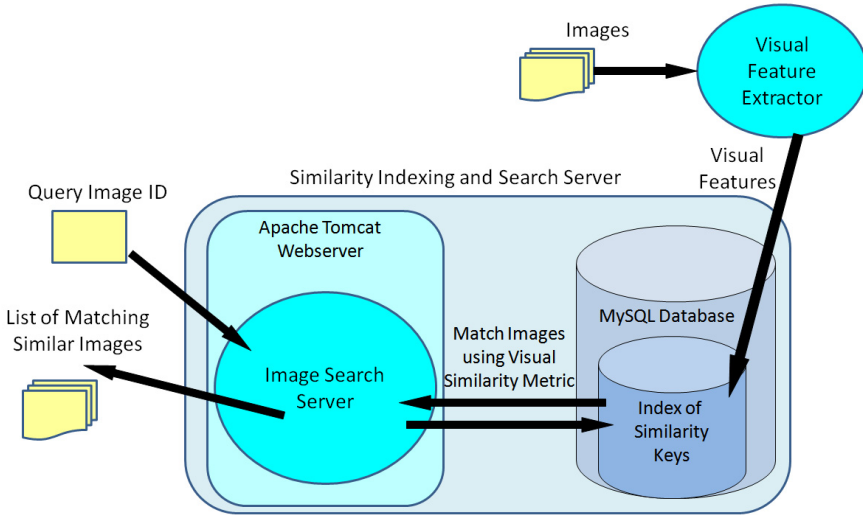


Fig. 4. Similarity search indexing and querying system

Similarity search is performed on the basis of the following (c.f. section 2.1):

- *Visual characteristics*: colours, shapes, texture, and composition
- *Semantic attributes*: image region and scene type classification and object types such as faces
- *Metadata and query context*: associated image annotations and tags, and the original query (if any) that was in effect when the query image was selected

Internally the system performs automated statistical query expansion, and the resulting query is applied as a relevance filter to potentially re-rank the results of the visual image similarity query. We also determine which of the metadata keywords associated with the chosen image are particularly relevant. This is done using a bimodal statistical relevance model to reduce the emphasis on words which are either too common and thus lack specificity, or too rare and thus lack generality. The former includes both common “stop words” such as “and”, “the”, “of”, but also other words that are too frequent in a given collection to be sufficiently discriminative. The latter category (rare words) will include terms that are not relevant to the content of an image, such as the name of

a specific photographer or the unique ID given to the image. Keywords whose relative frequency falls between these two extremes are given higher rankings (for example, “soccer” would be relevant to discriminate amongst images tagged as “football”, and “football” has selective power against other sports images).

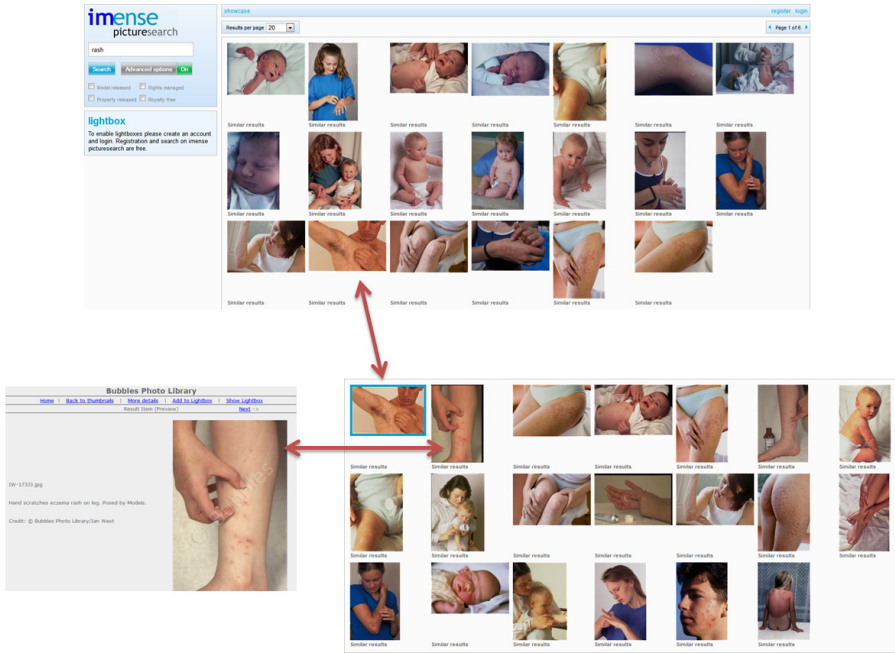


Fig. 5. Search examples. A patient wishing to self-diagnose a skin condition may start with a keyword search for ‘rash’ and utilise image similarity search to further explore the subset of images that are visually similar to those results that appear most relevant. Inspection of individual images may then suggest other terms to narrow down the search further, such as ‘eczema’. Only a small number of top search results is shown. The red arrows illustrate how a particular image can be selected to initiate a similarity based search for that image.

The Imense Similarity Search system consists of the following main components:

- *Visual Feature Extractor*: The visual attributes and content classification described in section 2.1 is compressed into a feature vector of about 1 kilobyte per image. This representation of images in terms of compact “similarity keys” is compact enough to allow efficient in-memory index representation using data structures such as hash tables or MySQL database tables.
- *Similarity Hashing function*: This function optionally generates a hash key from a similarity key as a means of providing a partially inverted index. This allows an index to be partitioned by visual content for added efficiency and

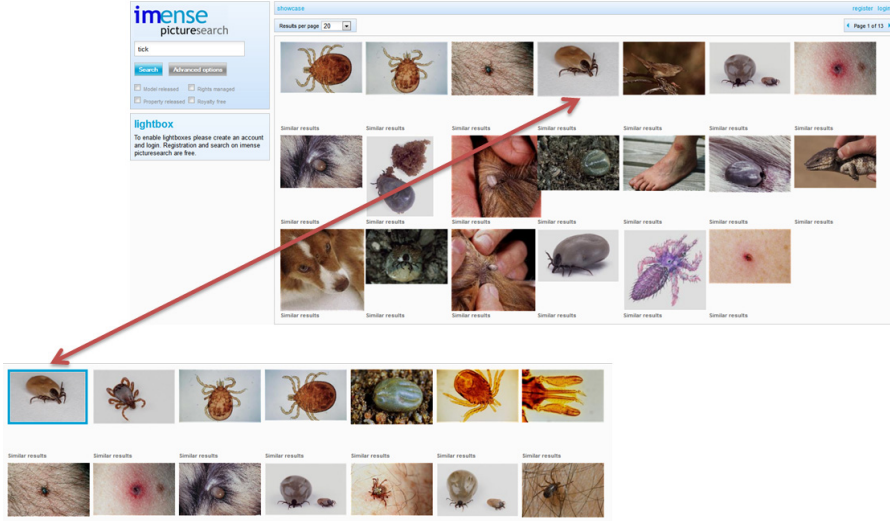


Fig. 6. In this example, similarity search is used to investigate parasitic ticks. The quality of results is enhanced through the use of image content classification into image types such as 'closeup/macro'.

flexibility in very large image collections, thus allowing similarity search to be restricted to a relevant subset of the index. The scheme we have implemented is a carefully tailored form of locality sensitive hashing [9] based on a projection of the feature vectors of each image in order to partition the set of images as evenly as possible.

- *Similarity Metric*: A function that can be used to compute a number representing the visual distance (similarity score) between images. The result of such a search is an ordered list of the best matching (most similar) images for a particular query.
- *Similarity Search Server*: This provides an API that allows other software modules or client servers to send image search queries to a dedicated similarity search server implemented using Java servlets running under Apache Tomcat. The server performs visual similarity searches against the index and returns ordered results.

The diagram in figure 4 illustrates how these components can be used to facilitate visual similarity search.

3 Example Use Cases

As studies of medical image search behaviour have shown ([22], [12], [13], [6]), most users wishing to find healthcare related images are primarily accustomed to formulating very short queries consisting of a few keywords. Single word queries

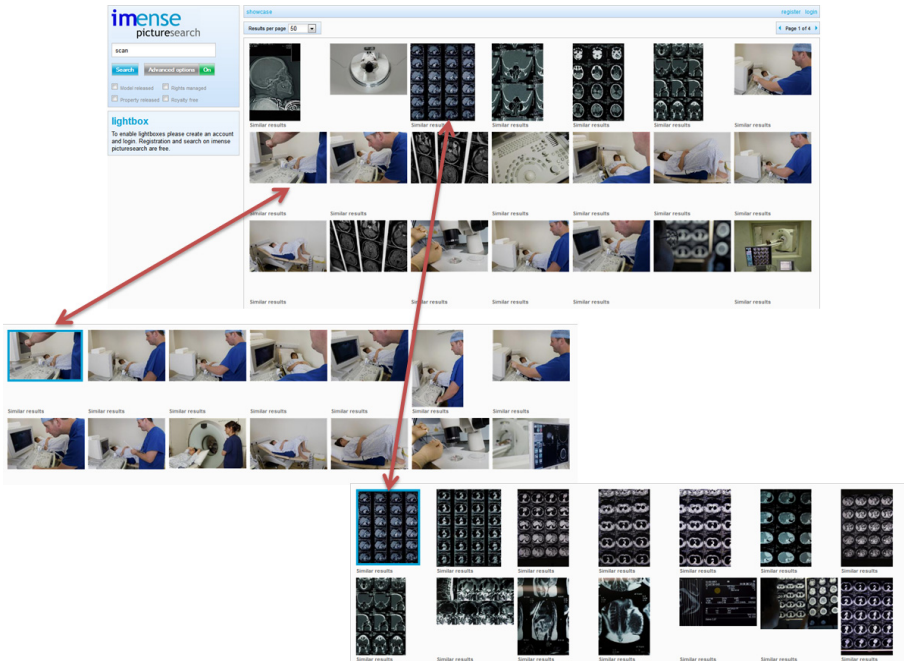


Fig. 7. This image search example concerns a medical procedure. Having performed a search for 'scan' the user can use similarity search to investigate particular medical scanning techniques or the types of imagery they produce.

are by far the most common, with users usually forced to perform manual query refinement or expansion based on their interpretation of the search results. This is especially challenging for users who have little prior knowledge about the biomedical domain they are researching, but even medical doctors may struggle with effective query composition [12].

The CBIR system described in section 2 can make this kind of search both more efficient and more productive. Initial exploratory search, whether on the basis of a single keyword or a more complex query, is performed by harnessing both image metadata as well as automated analysis and classification of image content (see section 2.1). The query parser mediates between user requirements and the capabilities of the content analysis system as discussed in section 2.3.

Having obtained initial search results, the user can then utilise image similarity search (section 2.4) to quickly find pictures related to those of the search results that appear to be most relevant. This can be done by simply clicking a link next to each image. Furthermore, the user can inspect images and their associated metadata and perform manual query refinement. Both of these processes may be combined and iterated to efficiently and selectively refine search within a large and very diverse corpus of images.

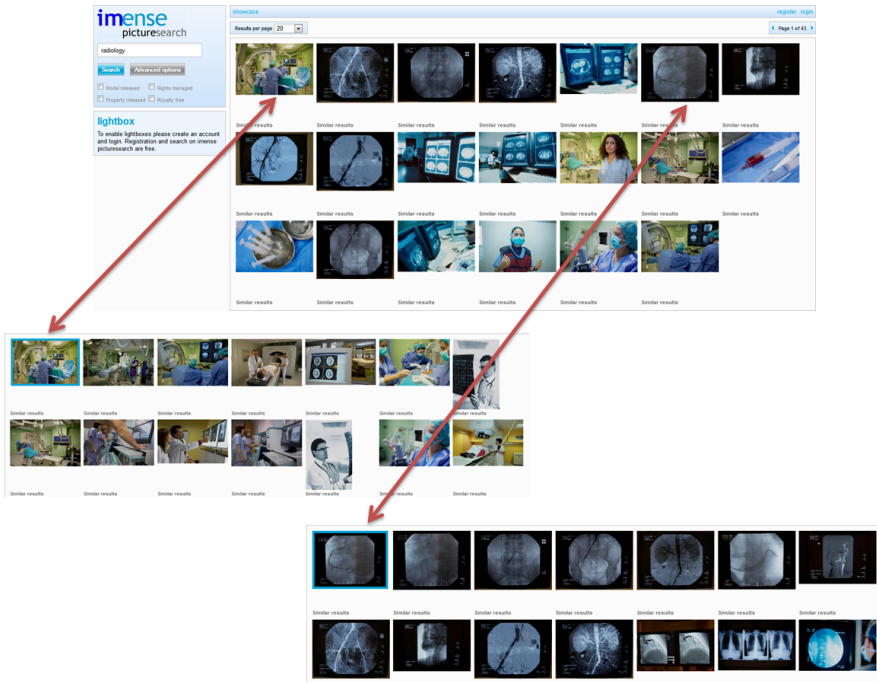


Fig. 8. Following on from the previous example, in this usage case the search concerns the domain of radiology, which can be visually explored using similarity search to obtain images pertaining to particular techniques or the images they produce.

Figures 5, 6, 7, and 8 illustrate this process for different example search scenarios. For reasons of space, only the top two or three rows of search results are shown in each sub-figure.

4 Conclusion

Biomedical image search systems are starting to benefit from content-based image retrieval (CBIR) techniques. However, most such systems are designed for very specialised usage and are not well suited to the very diverse groups of users who have medical image search needs, ranging from medical professionals to students and patients. At the same time, the internet has become the primary resource for healthcare related information. The vast bulk of data on the internet consists of images and video rather than text, very little of which has been adequately described using textual metadata.

Consequently there is great scope for systems that are able to perform broad-usage biomedical image search on the basis of an automated analysis of the actual content of images, thus allowing users to search “inside the picture” just as they have become accustomed to being able to search within other documents.

This paper described a system called Imense Picturesearch which is based on automated analysis and recognition of image content. It features an ontological query processor to optimise query efficiency without users requiring knowledge of the underlying content indexing system. Moreover, it provides an efficient “query-by-example” search functionality based on a generic image similarity search modality. This enables users to quickly refine image search queries through simple selection of relevant images from an initial exploratory search. We argue that such a system is well suited to broad-usage biomedical image retrieval, as exemplified by healthcare related retrieval scenarios over a diverse image corpus.

References

1. Aggarwal, P., Sardana, H.K.: Enhancements in medicine by integrating content based image retrieval in computer-aided diagnosis. In: Second International Conference on Digital Image Processing, pp. 75461X–75461X. International Society for Optics and Photonics (2010)
2. Akgül, C.B., Rubin, D.L., Napel, S., Beaulieu, C.F., Greenspan, H., Acar, B.: Content-based image retrieval in radiology: current status and future directions. *Journal of Digital Imaging* 24(2), 208–222 (2011)
3. Batko, M., Falchi, F., Lucchese, C., Novak, D., Perego, R., Rabitti, F., Sedmidubsky, J., Zezula, P.: Building a web-scale image similarity search system. *Multimedia Tools and Applications* 47(3), 599–629 (2010)
4. Britton, D., Cass, A.J., Clarke, P.E.L., Coles, J., Colling, D.J., Doyle, A.T., Geddes, N.I., Gordon, J.C., Jones, R.W.L., Kelsey, D.P., et al.: Gridpp: the uk grid for particle physics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367(1897), 2447–2457 (2009)
5. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)* 40(2), 5 (2008)
6. Deserno, T.M., Güld, M.O., Plodowski, B., Spitzer, K., Wein, B.B., Schubert, H., Ney, H., Seidl, T.: Extended query refinement for medical image retrieval. *Journal of Digital Imaging* 21(3), 280–289 (2008)
7. Hanbury, A.: Medical information retrieval: an instance of domain-specific search. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1191–1192. ACM (2012)
8. Iakovidis, D.K., Pelekis, N., Kotsifakos, E.E., Kopanakis, I., Karanikas, H., Theodoridis, Y.: A pattern similarity scheme for medical image retrieval. *IEEE Transactions on Information Technology in Biomedicine* 13(4), 442–450 (2009)
9. Kulis, B., Grauman, K.: Kernelized locality-sensitive hashing for scalable image search. In: 12th International Conference on Computer Vision, pp. 2130–2137. IEEE (2009)
10. Long, L.R., Antani, S., Deserno, T.M., Thoma, G.R.: Content-based image retrieval in medicine: retrospective assessment, state of the art, and future directions. *International Journal of Healthcare Information Systems* 4(1), 1 (2009)
11. Müller, H., Deserno, T.M.: Content-based medical image retrieval. In: *Biomedical Image Processing - Methods and Applications*, pp. 471–494. Springer (2011)
12. Müller, H., Despont-Gros, C., Hersh, W., Jensen, J., Lovis, C., Geissbuhler, A.: Health care professionals’ image use and search behaviour. In: *Proceedings of Medical Informatics Europe (MIE 2006)*, pp. 24–32 (2006)

13. Müller, H., Geissbuhler, A.: Medical multimedia retrieval 2.0. *Methods of Information in Medicine* 3(suppl. 1), 55–63 (2008)
14. Petrakis, E.G.M., Faloutsos, A.: Similarity searching in medical image databases. *IEEE Transactions on Knowledge and Data Engineering* 9(3), 435–447 (1997)
15. Rahman, M.M., Antani, S.K., Thoma, G.R.: A classification-driven similarity matching framework for retrieval of biomedical images. In: *Proceedings of the International Conference on Multimedia Information Retrieval*, pp. 147–154. ACM (2010)
16. Shapiro, L., Atmosukarto, I., Cho, H., Lin, H., Ruiz-Correa, S., Yuen, J.: Similarity-based retrieval for biomedical applications. In: *Case-Based Reasoning on Images and Signals*, pp. 355–387 (2008)
17. Torjmen, M., Pinel-Sauvagnat, K., Boughanem, M.: Methods for combining content-based and textual-based approaches in medical image retrieval. In: *Evaluating Systems for Multilingual and Multimodal Information Access*, pp. 691–695 (2009)
18. Town, C.: Ontological inference for image and video analysis. *Machine Vision and Applications* 17(2), 94–115 (2006)
19. Town, C.: Ontology based image and video analysis. In: *Computer Vision, Nova*, pp. 303–328 (2011)
20. Town, C., Harrison, K.: Large-scale grid computing for content-based image retrieval. In: *Aslib Proceedings*, vol. 62(4/5), pp. 438–446. Emerald Group Publishing Limited (2010)
21. Town, C., Sinclair, D.: Language-based querying of image collections on the basis of an extensible ontology. *International Journal of Image and Vision Computing* 22(3), 251–267 (2004)
22. Tsirikika, T., Müller, H., Kahn Jr., C.E.: Log analysis to understand medical professionals' image searching behaviour. In: *Proceedings of the 24th European Medical Informatics Conference (MIE 2012)* (2012)

Inferring Gene Networks from Gene Expression Data Using Dynamic Bayesian Network with Different Scoring Metric Approaches

Masarrah Abdul Motalib¹, Lian En Chai¹, Chuii Khim Chong¹, Yee Wen Choon¹, Safaai Deris¹, Rosli M. Illias², and Mohd Saberi Mohamad^{1,*}

¹ Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai 81310, Johor, Malaysia
{masarrah2, lechai2, ckchong2, ywchoon2}@live.utm.my, safaai@utm.my, saberi@utm.my

² Department of Bioprocess Engineering, Faculty of Chemical Engineering, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia
r-rosli@utm.my

Abstract. Inferring gene networks can be defined as the process of identifying gene interactions from experimental data through computational analysis. The aim is to infer gene network from gene expression data using dynamic Bayesian network (DBN) with different scoring metric approaches. The previous method, Bayesian network has successfully identified those gene networks but there are some limitations. Hence, DBN is able to infer interactions from a data set consisting time series rather than steady-state data. This research is conducted in order to construct and implement gene network and to analyze the effect by applying a different scoring metric approach for modeling gene network. In order to achieve the goals, a discrete model of DBN is used with different scoring metric approaches which are BDe and MDL. The *S. cerevisiae* cell cycle pathway is used for this research. To ensure the gene networks are biologically probable, this research employs previous annotation relative to the dataset. By having all of these implementations, this research is able to identify the effect of different scoring metric approaches, identify biologically meaningful gene network within the gene expression datasets and display the results in convenient representations.

Keywords: Dynamic Bayesian network, missing values imputation, gene expression data, gene regulatory networks, network inference.

1 Introduction

Dynamic Bayesian network (DBN) is well defined as a Bayesian network (BN) that represents sequences of variables. DBN can construct cyclic regulations using time delay information. DBN uses time series data for constructing causal relationships

* Corresponding author.

among random variables. Friedman *et al.* [1] first applied DBN to the analysis of gene networks. They constructed a discrete DBN model and used the Bayesian Dirichlet equivalence (BDe) scoring metric for learning networks. Ong *et al.* [2] also used a discrete DBN model but combined it with prior biological knowledge and current observations to model the tryptophan metabolism in *E. coli*. They utilized a repetitive EM (Expectation-maximization) algorithm to compute scores in learning network structure. On the other hand, to avoid data loss due to discretization, Kim *et al.* [3] developed a continuous DBN model with non-parametric regression model based on *B*-splines to take into account of linear dependencies. To select the optimal network, Kim *et al.* [3] subsequently defined a scoring metric known as $\text{BNRC}_{\text{dynamic}}$ based on the Laplace approximation.

Inferring gene networks can be defined as the process of identifying gene interactions from experimental data through computational analysis. Gene expression data from microarray are typically used for this purpose. The aim is to infer gene network from gene expression data using DBN with different scoring metric approaches. In addition, network visualization tools are available to indicate the network surrounding a gene of interest by extracting information from experimental data sets, such as Cytoscape [4]. We evaluated the efficiency of each scoring approach through the analysis of the *S. cerevisiae* gene expression data.

2 Materials and Methods

In previous works, researchers used BN which could not model a feedback loop because it did not have loops or cycles. In this section, we describe the details of the DBN-based model for inferring GRNs from gene expression data. In essence, the proposed model consists of three main steps: missing values imputation, construct gene network and evaluating network structures using scoring metric with respect to the given data. The following sub-sections discuss in detail for each of the three main steps.

2.1 Experimental Data and Missing Values Imputation

After all of the possibly used method and techniques identified, this is the stage where the researcher develops and implements a computational model based on the techniques in the previous steps. The model is implemented using BNFinder software [5]. This software allows for BN reconstruction from experimental data. Besides that, it supports DBN and if the variables are partially ordered, this also applies for static BN. It is written in python, and distributed under GNU GPL Library version 2.

The experimental study is based on the *S. cerevisiae* cell cycle time-series gene expression data [6]. However, the dataset contains missing values which must be processed. Conventional methods of treating missing values include repeating the microarray experiment which is not economically feasible, or simply replacing the missing values by zero or row average. A better solution is to use imputation algorithms to estimate the missing values by exploiting the observed data structure and expression pattern. In view of this, we applied the k-nearest neighbor method (kNN) imputation algorithm [7] that is the most fundamental and simple classification

methods, and should be one of the first choices for a classification study when there is little or no prior knowledge about distribution of the data.

2.2 Construction of Gene Networks

The DBN is used to construct gene networks, hence producing directed acyclic graphs (DAGs). For this research, we used Cytoscape for visualizing complex network and integrating these with any type of attribute data.

After the gene networks have been constructed, the performance of the gene networks constructed using DBN is evaluated. To evaluate the gene performance, the networks constructed are compared with the sub-networks constructed by Dejori [8]. Dejori [8] has also implemented BN to construct gene networks from *S. cerevisiae* dataset which is the same dataset in this research. Therefore, the sub-networks constructed by Dejori [6] are the benchmarks for this research.

We compared both of the methods by calculating True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). True Positive is the number of edges that exist in both network constructed by Dejori [6] and in the research. True Negative (TN) is the number of edges that do not exist in both networks (Dejori and this research). False Positive (FP) is the number of edges that exist in this research, but do not exist in the network by Dejori [8], while False Negative (FN) is the number of edges that exist in Dejori [8], but do not exist in this research.

2.3 Evaluating Network Structures

This research applies different scoring metric approaches in order to get the best network structures. The scoring metric approaches used to test in this research are the BDe score and the MDL score.

The BDe scoring criterion originates from Bayesian statistics and corresponds to posterior probability of a network given data. BDe uses Bayesian analysis to evaluate a network given a dataset. The Dirichlet distribution is a multinomial distribution that describes the conditional probability of each variable in the network, and has many properties that are useful for learning.

The MDL scoring criterion originates from information theory and corresponds to the length of the data compressed with the compression model derived from the network structure. Besides that, MDL provides the criterion for the selection, prediction and estimation of models. The purpose of MDL is to discover regularities in observed data. Generally, both BDe and MDL scores were originally designed for evaluating discrete variables.

3 Result and Discussion

The sub-networks that are chosen to be compared are YPL256C sub-network and YOR263C sub-network. TP, TN, FP, and FN are calculated to evaluate the performance of the sub-networks constructed from this research.

3.1 YPL256C Sub-network

Fig. 1 shows the YPL256C sub-network that is constructed by Dejori [8]. It can be seen that, the network consists of 12 nodes (genes) and 9 directed edges. However, the node for YGR108W does not form any edges with other nodes in the network.

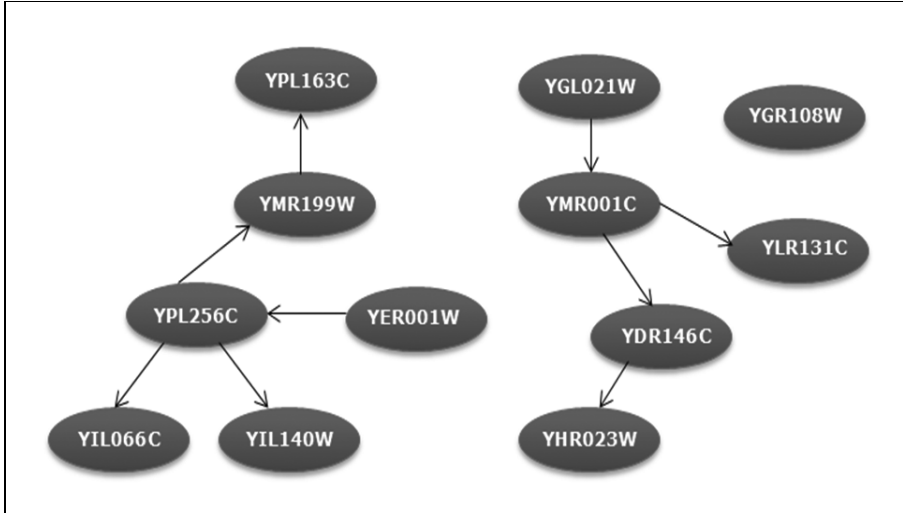


Fig. 1. YPL256C sub-network constructed by Dejori [8]

As shown in Fig. 1, there is two directed edge from gene YPL256C to gene YIL066C and YIL140W. It shows that there is a causal dependency between these three genes. The functions of gene YPL256C are encoding for G1-cyclin which involves in regulation of the cell and activates Cdc28p kinase to promote the G1 to S phase transition. A YIL066C gene is a minor isoform of the large subunit of ribonucleotide-diphosphate reductase which is involved in DNA replication. Whereas, the YIL140W gene is an integral plasma membrane protein that is required for axial budding in haploid cells and has potential to Cdc28p substrate. Therefore, a causal dependence of YIL066C and YIL140W from YPL256C is biologically logical since their functions are correlated.

As we look further, gene YGL021W contains characteristic motifs for degradation via the APC pathway and phosphorylated in response to DNA damage which is quite similar to A1k2p and to mammalian haspins. Gene YGL021W regulates YMR001C with multiple functions in mitosis and cytokinesis through substrate phosphorylation, also functioning in adaptation to DNA damage during meiosis. An unexpected result is the gene YGR108W does not connect any edge with other nodes. However, it does form edges with other nodes in the research done by Spellman *et al.* [6].

Fig. 2 shows the YPL256C sub-network that is constructed in this research using BDe and MDL scoring metric approaches. It is very clear that both networks consist of 12 nodes and 24 edges. It shows a different number of edges obtained in this research as compared to Dejori [8]. Through this research, we can see that several edges in the network are from cyclic regulation and have at least one directed edge

with other nodes. The network done by Dejori [8] does not show any cyclic regulation and the gene YGR108W failed to construct with any edge. About 20 new edges had been identified in this research. It is two times more compared to the result obtained by Dejori [8]. Hence, it is proven that DBN implemented in this research are able to construct cyclic regulation and form more potential edges between genes in a sub-network.

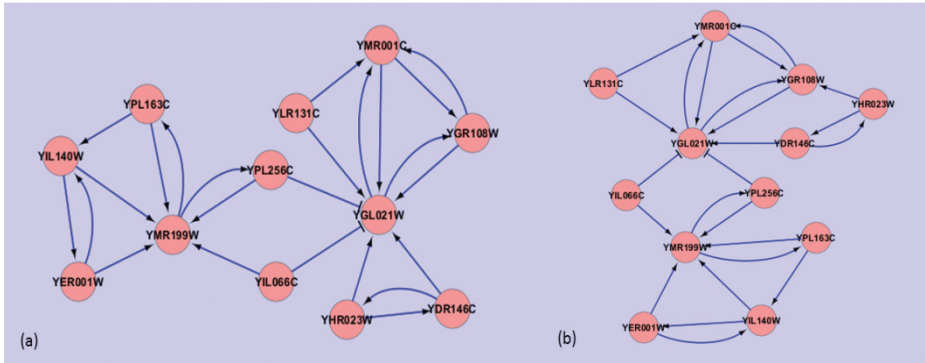


Fig. 2. YPL256C sub-network constructed with (a) BDe, (b) MDL scoring metric approaches

Table 1 shows the comparison of edges formed in YPL256C sub-network between Dejori [8] and this research. True Positive (TP) is the number of edges that exists in both network constructed by Dejori [8] and this research. False Negative (FN) is the number of edges that exist in Dejori sub-network, but does not exist in network of this research. False Positive (FP) is the number of edges that exists in network of this research, but does not exist in Dejori [8]. True Negative (TN) is the number of edges that does not exist in both networks constructed by Dejori [6] and in this research. The sensitivity for this sub-network is 44% whereby 4 directed edges that exist in the network by Dejori [8] have been captured in this research as well. However, there are about 5 directed edges exist in Dejori [8] but it does not exist in the network of this research. The missing edge is between gene YPL256C to YIL140W and YIL066C, gene YER001W to YPL256C, gene YMR001C to YLR131C and YDR146C respectively.

Table 1. Result of YPL256C sub-network

Condition	Number of Edges	Statistical Measures
TP	4	Sensitivity 44.44%
FN	5	
FP	20	Specificity 84.96%
TN	113	

The specificity for this sub-network is approximately 84.96%. Gene YLR131C regulates both genes of YMR001C and YGL021W. Gene YLR131C encodes for transcription factor that activates transcription of genes expressed in the G1 phase of the cell cycle. On the other hand, gene YMR001C is involved in regulation of DNA replication which encodes a protein. Furthermore, gene YIL066C is expressed only after DNA damage occurred in order to cope with the function of YMR199W. Gene YMR199W encodes for G1-cyclin which involved in regulation of the cell cycle. Therefore, it is biologically logical for YIL066C regulating the expression of YMR199W.

3.2 YOL263C Sub-network

In this study, we compared the YOR263C sub-network obtained from this research and YOR263C sub-network by Dejori [8]. Fig. 3 shows the YOR263C sub-network that is constructed by Dejori [8]. It can be seen that, the network consists of 8 nodes (genes) and 6 undirected edges. The undirected edge between YOR263C and YOR264W are the most conspicuous features in the sub-network because both genes are located next to each other on the DNA strand of chromosomes XV. However, the biological and molecular for both genes are still unknown. Gene YNR067C and YGL028C is another feature with high confidence level that is the undirected edge. YNR067C is a daughter cell-specific secreted protein with similarity to glucanases and it degrades cell wall from the daughter side causing daughter to separate from mother. The function of YNR067C is still currently unknown. The function of YGL028C is known to be a soluble cell wall protein and play a role in conjugation during mating based on its regulation by Ste12p. It also has an undirected edge with YER124C which may regulate cross-talk between the mating and filamentation pathways and deletion affects cell separation after division and sensitivity to alpha factor and drugs affecting the cell wall. Gene YGL028C is related to YLR286C which is an endochitinase required for cell separation after mitosis. YER124C has undirected edge with two nodes (YLR286C, YGL028C), and both nodes are functionally related to cell wall biogenesis, therefore it can be assumed that it is involved in cell wall biogenesis. Gene network constructed using BN have provided a testable prediction of an unknown gene function.

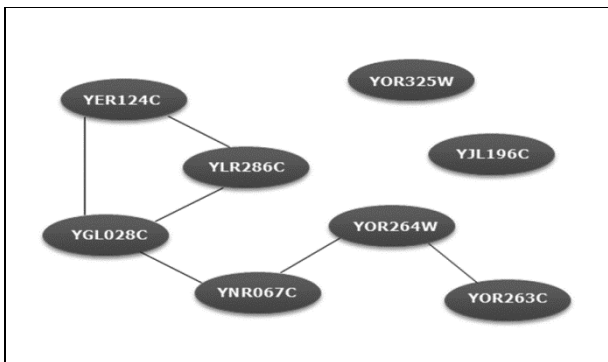


Fig. 3. YOL263C sub-network constructed by Dejori

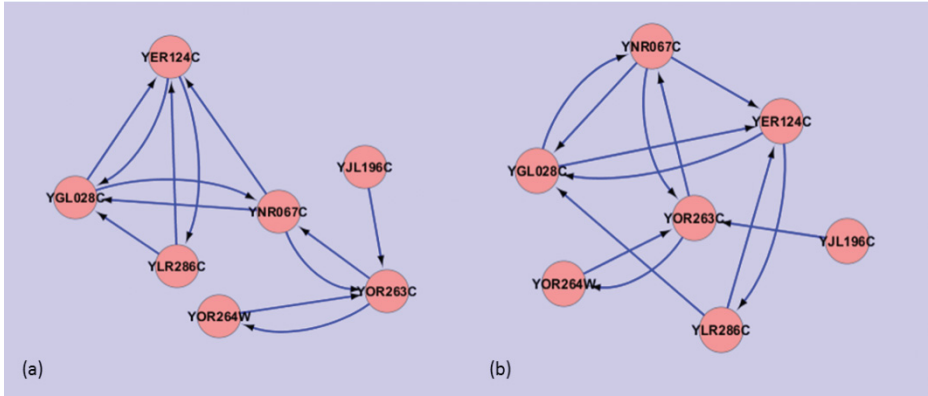


Fig. 4. YOL263C sub-network constructed with (a) BDe, (b) MDL scoring metric approaches

Fig. 4 shows the YOR263C sub-network that is constructed in this research using BDe and MDL scoring metric approaches. It is very clear that both networks consist of 7 nodes and 13 edges. They show a different number of edges obtained in this research as compared to the result obtained by Dejori [8]. Through this research, we can see that several edges in the network are form cyclic regulation and have at least one directed edge with other nodes, while the network done by Dejori [8] does not show any cyclic regulation. The main difference between this research and Dejori [8] are that they can show the interactions between genes clearer. As we can see in the Dejori [8] sub-network, the edge formed between YOR263C and YOR264W cannot show which gene is regulating another. However, this research shown clearly that YOR263C is regulating YOR264W and it is a cyclic regulation. It means that the expression level of YOR264W is depending on YOR263C and YNR067C as well. About three new edges have been identified in this research.

Table 2. Result of YOR263C sub-network

Condition	Number of Edges	Statistical Measures
TP	5	Sensitivity 83.33%
FN	1	
FP	3	Specificity 80.00%
TN	12	

Table 2 shows the comparison of edges in YOR263C sub-network between the network constructed by Dejori [8] and this research. The sensitivity of YOR263C sub-network is approximately 83.33%. There are about 5 cyclic edges formed in this sub-network. The specificity for this sub-network is approximately 80%. This shows that the DBN implemented in this research is capable of uncovering more potential edges, interactions and cyclic regulation between genes compared with the study by Dejori [8].

3.3 Performance of Scoring Metrics

Table 3 summarizes the computation time comparison between scoring metric approaches of YPL256C sub-networks. MDL excels in speed as it had a computation time of 1 minute and 10 seconds while BDe took approximately 2 minutes. This concurs with the finding of Vinh *et al.* [9] which discovered that BDe is more time-consuming than MDL. However, both scoring metric approaches obtained the same network results (24 edges and 12 nodes) and accuracy (as summarized in Table 1). On the other hand, Table 4 shows the computation time comparison between scoring metric approaches of YOR263C sub-networks. Both scoring metric approaches gave roughly the same computation time which is 1 second. This is probably due to the fact that YOR263C has a smaller network structure compared to YPL256C. Both scoring metric approaches also computed the same network results (13 edges and 7 nodes) as well as accuracy (refer to Table 2). The experiment with YPL256C showed that MDL has an advantage in computation time without compromising the accuracy for network inference.

Table 3. YPL256C: Comparison of computational time between scoring metrics

Sub-network	Scoring Metric Approaches	Computation Time (HH:MM:SS)
YPL256C	BDe	00:02:01
	MDL	00:01:10

Table 4. YOR263C: Comparison of computational time between scoring metrics

Sub-network	Scoring Metric Approaches	Computation Time (HH:MM:SS)
YOR263C	BDe	00:00:01
	MDL	00:00:01

Table 5. Network scores between scoring metrics for YPL256C and YOR263C sub-networks

Scoring Metric	YPL256C	YOR263C
BDe	470.257	342.084
MDL	704.546	504.177

Table 5 shows the network scores obtained by both scoring metrics for YPL256C and YOR263C sub-networks respectively. Lower score are said to have optimal network structure. In both sub-networks, BDe performed better than MDL. Nevertheless, this scoring advantage did not influence much on the inference of

optimal network structure as both scoring metric approaches obtained the same network structure for YPL256C and YOR256C.

4 Conclusion

DBN has been widely utilized by researchers in gene networks inference from gene expression data as it is robust, able to handle feedback loops and the temporal aspect of time-series data. To learn the optimal network structure, BDe or MDL scoring metric are often employed in the DBN model. This research is conducted to analyze the influence of both scoring metrics on gene networks inference using DBN. Based on the experiments done on two *S. cerevisiae* cell cycle sub-networks YPL256C and YOR263C, we found that MDL has faster computation speed in larger network structure but BDe has an edge in representing exactness of statistical interpretation. Therefore, we suggest using MDL in exceptionally large networks as exponentially increased computation time would negate the statistical advantage of BDe. BDe is more suitable for smaller networks or in such circumstance whereby accuracy is much sought after. For future work, we would like to apply different scoring function that satisfies the score equivalence property.

Acknowledgments. We would like to thank the Malaysian Ministry of Science, Technology and Innovation for supporting this research by an e-science research grant (Grant number: 06-01-06-SF1029). This research is also supported by UTM GUP research grant that was sponsored by Universiti Teknologi Malaysia.

References

1. Friedman, N., Murphy, K., Russell, S.: Learning the structure of dynamic probabilistic networks. In: Proc. 14th Conference on the Uncertainty in Artificial Intelligence, San Mateo, pp. 139–147 (1998)
2. Ong, I.M., Glasner, J.D., Page, D.: Modelling regulatory pathways in *E. coli* from time series expression profiles. *Bioinformatics* 18, S241–S248 (2002)
3. Kim, S., Imoto, S., Miyano, S.: Dynamic bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. In: Priami, C. (ed.) CMSB 2003. LNCS, vol. 2602, pp. 104–113. Springer, Heidelberg (2003)
4. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504 (2003)
5. Wilczynski, B., Dojer, N.: BNFinder: exact and efficient method for learning Bayesian Networks. *Bioinformatics* 25(2), 286–287 (2009)
6. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9, 3273–3297 (1998)
7. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B.: Missing value estimation methods for DNA microarrays. *Bioinformatics* 17(6), 520–525 (2001)

8. Dejori, M.: Analyzing Gene Expression Data with Bayesian Networks. Graz University of Technology, Austria (2002)
9. Vinh, N.X., Chetty, M., Coppel, R., Wangikar, P.P.: GlobalMIT: Learning Globally Optimal Dynamic Bayesian Network with the Mutual Information Test (MIT) Criterion. *Bioinformatics* 27(19), 2765–2766 (2011)
10. De Campos, L.M.: A Scoring Function for Learning Bayesian Networks based on Mutual Information and Conditional Independence Tests. *J. Mach. Learn. Res.* 7, 2149–2187 (2006)

Malaysian Parasite Database Infrastructure

Sarinder K. Dhillon¹, Nur-Imtiazah Shuhaimi¹, Susan Lim Lee Hong¹,
and Amandeep S. Sidhu²

¹ Institute of Biological Sciences, Faculty of Science,
University of Malaya, 50603 Kuala Lumpur, Malaysia

² Curtin Sarawak Research Institute,
Curtin University, Sarawak, Malaysia

Abstract. The Malaysian Parasite Database is set up to collect, digitize, collate, integrate and analyse available literatures on Malaysian parasites to be shared and disseminated through an integrated database system for the generation of knowledge. We adopted the data warehouse approach which is convenient and meets the purpose for the Malaysian Parasite Database. The data includes information on parasite specimens collected besides their taxonomy, biology, ecology, hosts and DNA which will be regularly updated. These data, which will be regularly updated are initially obtained from literatures and researchers collections which are then digitized and added into the database. We plan to integrate this database with other parasite host databases in order to provide a detailed and comprehensive information on indigenous parasites. This database is envisaged to be dynamic with regular incorporation of new analytical methods and novel use. In this paper, we present the infrastructure of the Malaysian Parasite Database and using a data warehouse approach which uses wrappers for a structured data extraction from related public databases and a structured vocabulary for data integration. The current and future implementations of the proposed infrastructure will be hosted on a cloud environment.

Keywords: Database, Data Integration, Parasite, Data Warehouse, Ontology.

1 Introduction

The problem of management of biological data is as old as the data themselves. It is not only the flood of information and heterogeneity that make the issues of information representation, storage, structure, retrieval and interpretation critical. There also has been a change in the community of users. In the middle 1980s, fetching a biological entry on a mainframe computer was an adventurous step that only few dared. Now, at the end of the 1990s, thousands of researchers make use of biological databanks on a daily basis to answer queries, such as to find sequences similar to a newly sequenced gene, or to retrieve bibliographic references, or to investigate fundamental problems of modern biology [1]. New technologies, of which the World Wide Web (WWW) has been the most revolutionary in terms of impact on science, have made it possible to create a high density of links between databanks. Database systems today

are facing the task of serving ever increasing amounts of data of ever growing complexity to a user community that is growing nearly as fast as the data and is becoming increasingly demanding. The current scope of databases ranges from large-scale archiving projects to individual, private, specialized collections serving the needs of particular user communities. These include the following. Whereas the large primary databases collect and collate information from literature and from the scientific community, specialized data collections integrate, via curatorial expertise, information from a multiplicity of primary sources, including sequence, structure, function, evolutionary relationships and bibliographic references. Rigid database classification has become obsolete, and users choose according to individual needs from the rich WWW-accessible data.

Malaysia, being the 12th largest in biodiversity in the world, information in the digital media is still in its infancy. It is increasingly important to develop novel approaches to understand and manage our living environment, allowing for the development of reliable and science-based management strategies. This can be achieved by managing the biodiversity resources electronically and sharing the data on biodiversity records. Due to an absent integrated infrastructure and inadequate long-term data archiving the state of information management in biodiversity is currently very unsatisfactory. Thus we require some efforts to foremost manage the data in a planned infrastructure. Several initiatives are underway to digitize biodiversity information. [2] started an initiative to develop relational biodiversity databases and catalogue museum collections. This initiative resulted in the digitization and subsequent electronic availability of vast amount of biodiversity data in University Malaya. The Palm Oil Research Institute of Malaysia (PORIM) maintains an oil palm database, accessible to registered internet users only. In addition, the Forest Research Institute of Malaysia (FRIM) provides web users limited database access to its huge forest resource collections [3]. FRIM has almost 2000 records of botanical and entomological collections in their database [4]. University Kebangsaan Malaysia (UKM) also set up an initiative in 2001 to manage biodiversity electronically as well as to promote exchange of data which can be accessed via an online portal (<http://biodiversity.ukm.my>) [5]. Another group from University Technology Malaysia designed a conceptual data model for biodiversity which is known as BiDaM [6]. However, some of these databases have become obsolete due to poor management and no long term plans of data archival. There is no proper ICT infrastructure in place to manage and integrate these databases.

Globally there are a number of dedicated network dealing with specific interests. For example Zebrafish Information Network (<http://www.zfin.org>) is a web based community resource that provides information related to the genetic and developmental data on zebrafish. ZFIN is currently implemented as a relational database management system by IBM [7]. The Reptile Information Network (<http://www.reptileinfo.com>) is a database of reptiles and amphibians that was established as a platform in information sharing. This is a platform for virtual communications between herpetoculturists and herpetologists not only to gain information on the biology of the animal but also conservation matters. The Marine Life Information Network (<http://www.marlin.ac.uk/>) is concerned with information related to marine life in Britain and Ireland. MarLIN provides information

on species, habitats, ecosystems liaison and is related to several databases to achieve its aim as information provider on marine resources to support marine environmental management, conservation and education. The Mouse Resource Web Browser (MRB) is an online registry of mouse resources which is stored in a database using the relational model [8]. In the parasitology domain, the MonoDb (<http://www.monodb.org>) provides information to parasitologists on the known species of monogeneans. Information access in MonoDb is limited to textual-based searching and static image gallery. Host-parasite database (<http://www.nhm.ac.uk/research-curation/research/projects/host-parasites/database/index.jsp>) is another example providing host-parasite information but limited to browsing. SuperIDR [9] is another parasite database which is used as a teaching tool for parasitology [10]. Meanwhile, EKEY (<http://digitalcorpora.org>) is a web-based system that provides taxonomic classification, dichotomous key, text-based search and combination of shape and text-based search which taking into account fish shape outlines and textual terms. To date, there is no database system in the parasitology domain which covers aspects of taxonomy, genomic, ecology, medical, biology, host and publications that can serve as a one stop knowledgebase in this domain.

While, a wealth of biodiversity information exists in Malaysia, there is a paucity of dedicated and specific biodiversity databases. Since parasites are integral part of the ecosystems and can be found on all living organisms, their diversity will outnumber the free-living organisms if we consider that each species can harbour two parasite species [11]. Although substantial work has been done by Malaysian parasitologists [12] in the country, nothing has been done to collate the vast amount of data on the taxonomy, biology, ecology of the parasites and recently DNA data. There is also no information on the locations of type and voucher specimens of species found in this region. Museum collections are in great shambles as there are no central depositories for them nor are there any laws (in Malaysia) instituted to protect the type-specimens. There is a need to have reliable list on museum catalogues online which can be used by taxonomists and other researchers.

Currently, there is no online information on Malaysian parasitofauna. To obtain knowledge about a parasite, biologists often need to go through an information gathering process, navigating between the public databases available freely to them. The pool of data in this domain is mostly scattered and often stored in heterogeneous formats. There is a need to develop specific generic and indigenous dynamic database on parasites especially since such information will be critical in managing health of our natural resources and indirectly human health. The objective of paper is to propose and present an infrastructure to manage, sustain and to disseminate information on Malaysian parasitofauna by collating the information from published records of parasites in Malaysia into a digital format. Besides data management, this system will also enable users to source information, to generate knowledge and ensure continuity of knowledge.

2 Database Content and Design

A characteristic of biology is that, since the number of sources for any particular subject is high, the data integration solution should be scalable with sources. The first

generation of solutions for data integration employs a series of non-interoperable and non-scalable quick fixes to translate data from one format to another. This means writing programs usually in a programming language like Perl, [13] in order to access, parse, extract and transform necessary data for particular applications. If one of the data sources changes the formats, all of the programs involved with this data source must be upgraded. Upgrades are inevitable because changes in web page services and schema are common for biological data sources. The second generation of data integration solutions provides a more structured environment for flexible, scalable and robust integration. Many efforts have been made in this area [14]. They can be divided into three major categories according to architectures: link-driven approach, view integration approach and warehousing approach.

In the link-driven federation approach, the user can switch between data sources using system-provided links. Here, a user starts from some point of interest in a data source and then can jump to related data sources through system created links. The user has to still interact with individual sources; only the interaction is easier through convenient links and does not involve the data sources directly. SRS [15], GeneCards [16] and LinkDB [17] are examples of this approach. The link-driven approach is very convenient for non-expert users because of single point-and-click user interface. The downside of the link-driven approach is that it does not scale well and has no across-source capabilities. When a new data source has to be added to the system, links connecting its entries and those of all other data sources have to be created. If a data source changes, the link building has to be redone. Moreover, a join between two data sources is not possible in link-driven approaches.

In view integration, a virtual global schema in a common data model is created using data source description. Queries on common data model are then automatically reformatted to source level queries. There are two approaches – global-as-view where the global schema is defined as view over the local sources, and local-as-view where a global schema is defined beforehand and local sources are described as views over global schema. DiscoveryLink [18], K2 [19] and its predecessor Kleisli [20] are examples of this approach. The advantages of the view integration approach are that the latest content of the sources is always returned by the system, operations across data sources can be specified in the query language, no storage is needed at the middle-ware, and adding new data sources is easier than with the link-driven approach. The disadvantage is that since sources are accessed instantaneously during query execution, in the event of a source being down, a query may return fewer matches or even fail. Also, to build a common integrated view, great expertise in the system is needed.

The warehousing approach can be described as view integration where the global schema is materialized i.e., an instance is created locally. The data from different sources is downloaded, cleaned of erroneous entries, categorized into meaningful structures (manually or automatically curated) and formatted for suitable analysis. Usually, the instance is stored in a database (e.g., relational database) and can be queried with a database query language (e.g., SQL). GUS [19] is an example of the warehousing approach. The biggest advantage of a warehouse for data integration is that the downloaded source data can be manipulated into suitable formats and annotated to facilitate integration and analyses. Execution of queries is usually very fast because

there is no outside dependency. On the other hand, the maintenance costs of warehousing are high because the downloaded data from the sources have to be kept fresh (i.e., synchronized with data source).

In this paper, we propose a data warehouse approach as an infrastructure for Malaysian Parasite Database.

2.1 Data Gathering

Initially, dedicated gathering of data is required from published literature and records to enrich and populate the database. The publications which are usually hard-to-come-by journals will also be digitized and incorporated into the database. Once the database is populated, it has to be regularly updated and the data verified by experts hosting the database. Next the database will have to be published online and data security has to be in place to secure data and prevent threats. Basically the following procedures will be used to develop the database to make it functional.

- a) Survey of any existing biological database systems
- b) Study on necessary features
- c) Collect and prepare an inventory of datasets and digitization of data
- d) Generate the entities within the system and determine relationships between each entity
- e) Develop an indigenous relational database
- f) Develop standards and protocol for information databases including data processing and information retrieval.

2.2 Development of Modules

The Malaysian Parasite Database will consist of a data warehouse for authenticated and updated information on parasites. Parasitologists are engaged in the verification and preservation of information source and as well as to enrich and update the database from time to time.

The Malaysian Parasite Database will have various relational modules for ease of data management and retrieval. Using such modular system, it will allow new modules be added with ease whenever necessary (Fig. 2).

- a) Parasite Taxonomy and Biology - all information pertaining to the parasites will be listed: parasite identity, description, collection locality, location of specimens, hosts.
- b) Parasite Ecology includes information on distribution patterns and population size of parasites.
- c) Parasite DNA data module will be incorporated for parasite species identifications and phylogeny.
- d) Reference List together with the digitized literature will be incorporated.
- e) Parasite Museum Collection ascension numbers of the parasite specimens deposited in local and renowned safe museums will be noted.

- f) Host Databases (Fauna): Hosts databases such as Fish, Turtle and other animal databases will be developed and integrated to help researchers in obtaining related information.

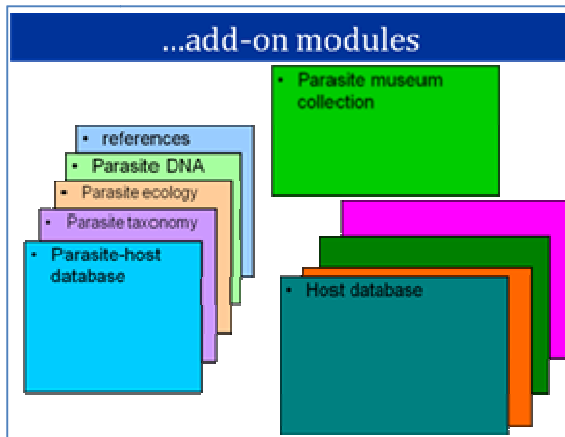


Fig. 1. Add on modules for Malaysian Parasite Database

2.3 Structured Data Extraction

A wrapper will be implemented for automatic extraction of relevant information from other public databases and the web. Given a number of online resources, it automatically finds patterns and grammars and store them in a temporary database. The data obtained from other resources is then verified by domain experts for authenticity.

2.4 Data Integration Using a Structured Vocabulary

Data extracted from multiple sites or even databases typically have different data formats. Thus, we need to match data columns which are represented differently or using different names in other databases. Additionally, values that are semantically identical but represented differently in different web sites need to be integrated using a structured vocabulary. Here we used the Taxonomic Data Working Group (TDWG) (<http://www.tdwg.org>) structured vocabulary as a database schema. Extracted data from other public databases and web sites will be mapped to this schema during the integration process. Data integration in this paper is done using ontologies (Fig. 4).

The TDWG strongly suggests the deployment of Life Science Identifiers (LSID) vocabulary that has been widely used in biodiversity and offers a wide coverage of concepts, which are suitable to annotate the taxonomic information of an organism. The entities and attributes in the proposed database are named according to the Taxonomic Data Working Group (TDWG) Life Sciences Identifiers (LSID) which is a globally accepted standard in biodiversity.

The Malaysian Parasite database physical design consists of 16 entities encompassing information on taxonomic classification of parasite species and its host,

DNA, locality, museum collections, biological information, medical, ecology, authorship and references. These entities are named as Taxonomy, Phylum, Class, Order, Family, Genus, Host, Parasite, Molecular, Ecology, Medical, GIS, Museum Collection, Biology, Author and Reference. Parasite is the main entity represented by the ScientificName as the primary key and nine foreign keys to relate the Author, Molecular, Medical, Museum, Taxonomy, Biology, Reference and Ecology entities. Taxonomy includes ScientificName, Phylum, Class, Order, Family, Genus and each of these has a unique identifier to represent a parasite species and a parasite host species. Each of these attributes is also denoted as entities which contain a unique identifier. Molecular contains the genetic information of the parasite species which is represented by Type of Gene, GenomeLocation, DNASequences, RNASequences attributes. Host entity contains information of the taxonomic classification of host of each of the parasite which is represented by Kingdom, Phylum, Class, Order, Family, Genus, Species, Author, Synonym and ScientificName attributes, Medical entity contains information on diseases on parasites and is represented by Disease, AffectedArea, Symptom, Vector and Treatment attributes. Ecology contains the ecological information about the parasite host and is represented by Locality, PopulationSize, DistributionPattern attributes. Ecology entity is linked with the GIS entity to provide maps for exact locality which is represented by Map, Satellite and Terrain attributes. The Museum Collection entity contains information on museum records which is represented by VoucherNumber, CatalogNumber, IdentifiedBy, StateProvince, DateCollected, Holotype and Paratype attributes. Parasites also require biological information and represented by LifeStage, Habitat and Attributes in the Biology entity. The entities and attributes for the proposed database are denoted using an Entity-Relationship diagram in Fig. 3.

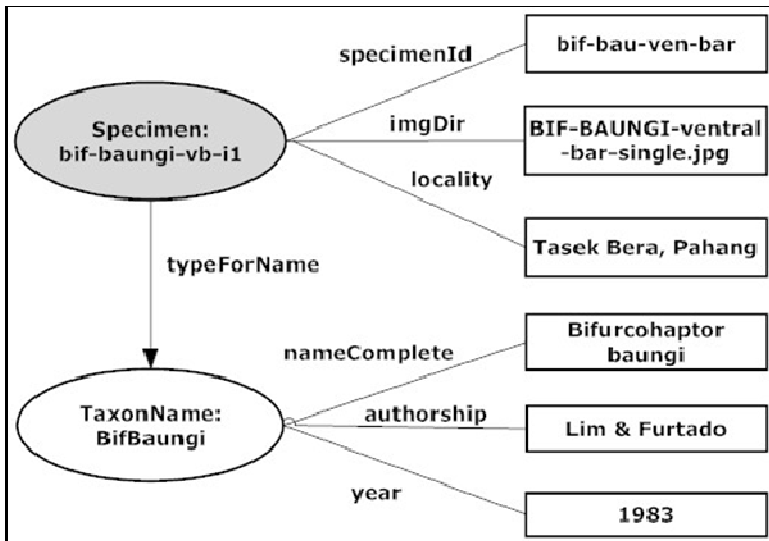


Fig. 2. Example of Ontology

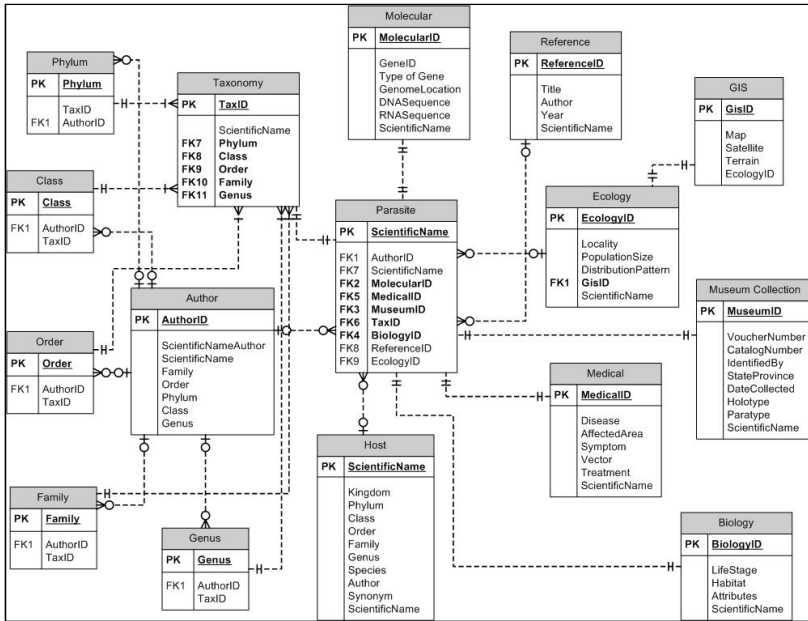


Fig. 3. Entity-Relationship diagram

2.5 Data Analysis

Malaysian Parasite Database System will be adapting the On-Line Analytical Processing (OLAP) technology approach for analysis purposes. The analysis system operates using ROLAP model referring to Relational On-Line Analytical Processing. ROLAP model is designed to allow analysis of data using a multidimensional data model where the system will access the data from a relational database. The main structure of OLAP is the OLAP data cube which can be considered as similar as a table in relational database system. All data cubes will be recorded and saved in XML format which means the cube is saved physically in XML files using analysis manager tool. For querying or analyzing data from the cubes, MDX (Multidimensional Expressions) query is enabled to extract the data. MDX query is a query language used for OLAP analysis which operates in the same way as the SQL language used in relational databases.

3 Database Construction

The proposed architecture in this paper will be implemented using a relational database system PostgreSQL (<http://www.postgresql.org/>). Custom-made parsers will

be developed to integrate data on taxonomy, DNA, biology, ecology, references, museum specimen collections as well as the parasite host in the database. All parsers will be developed in Perl using standard modules, such as BioPerl and DBI. The Web interface will be designed using the standard Perl modules DBI and CGI, with automatic generation of standard SOAP APIs to databases that allow direct database access (Fig. 4). The current and future implementations will be done on Microsoft Azure Cloud Platform (<http://www.windowsazure.com/>). There are numerous opportunities for data services to move to the Cloud. It is an ideal environment for High Performance and Data Intensive applications [21] like Biological Databases with large number of complex query requests in a day.

The vast amount of data in the database can be catalogued, integrated, correlated, shared, queried, analyzed and searched using a query system which is being developed. The infrastructure of a Malaysian Parasite Database System is presented in Fig. 5.

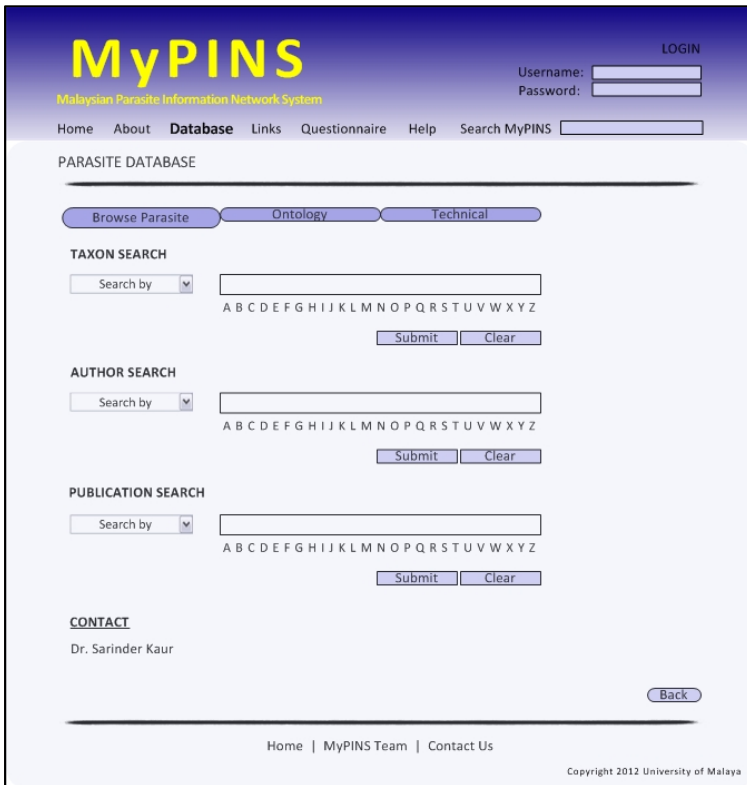


Fig. 4. Example of Screen Design

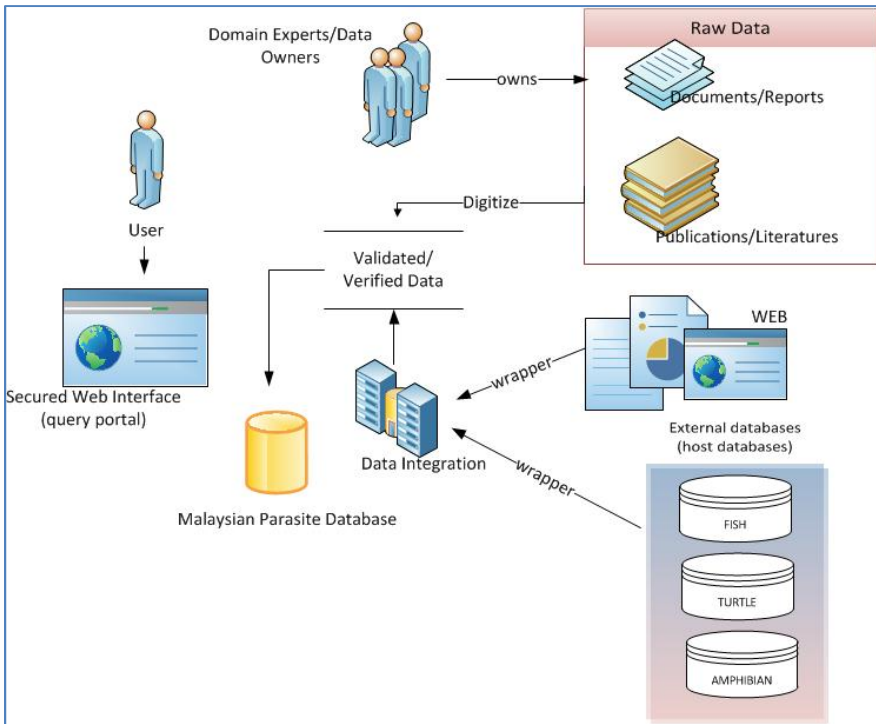


Fig. 5. Architecture of Malaysian Parasite Database

4 Conclusions and Future Work

The Malaysian Parasite Database is a indigenous web based information architecture which encompasses a wide coverage on information on parasites, such as the taxonomy, biology, DNA, ecology, references, museum specimen collections as well as the host. This database containing Malaysian parasitofauna could be linked with related Malaysian or international databases for data sharing and dissemination.

It will provide a platform for sharing parasites species information at both local and global levels through the world wide digital medium. This system is able to display taxonomic details of the identified parasite and show any other information such as diseases caused by the parasites. Such databases will also identify gaps in knowledge and reveal research opportunities. However we also need to protect the intellectual property of the contributors of data therefore data security is a serious issue that needs to be addressed before the database goes on the World Wide Web. We will be proposing the parasite image database as part of the current infrastructure which includes features of pattern recognition to allow researchers in conducting automated species identifications. Other expert systems and decision support systems will be built as front end applications using the current database as a backend.

Biological data must be described in context rather than in isolation [22]. Hence, many databases provide multiple links to other resources, but efficient use of these links requires intelligent retrieval systems. The proposed infrastructure is the concept of a warehouse, or a centralized data resource that manages a variety of data collections translated into a common format [23]. Linking the community of databases through common semantics is impossible because of their extreme heterogeneity of this approach. The 'middleware' approach affords a chance to uncouple data access from data management and to allow for remote retrieval beyond the simple scripts fetching data from external databases. The most prominent industrial standard for a client-server based middleware is Common Object Request Broker Architecture or CORBA [24] as defined by the Object Management Group OMG. CORBA is a distributed object architecture that allows objects to communicate across networks through defined interfaces using the syntax of the Interface Definition Language (IDL). The object management architecture of CORBA specifies an application-level communication infrastructure. Several CORBA-based applications have already appeared. [25] suggest a set of interface definitions for molecular biology to access a simple but realistic data bank of Sequence Tag Sites. The European Commission supports a project to provide CORBA access to a set of public databases (EMBL, SWISS-PROT, PIR, TRANSFAC, and several others).

We will use an alternative middleware approach for database interconnection in future. For example, the Jade software system [26] establishes a connection between the database servers and the application programs, and organizes data exchange through standardized relational tables and parameters. Information retrieved on the data server side is transformed into these tables with the help of a specialized application called Jade adapter. This approach will help in forming a loose federation of autonomous biomedical databases on the web. In order to make all this data really useful, tools that will access and retrieve exactly the information user needs will be developed. This will form basis to develop a representation of the underlying semantics of biodiversity knowledge in a form suitable for integrating all data sources helping to design biodiversity ontology specifically designed to integrate data sources by dynamically retrieving requested information from diverse data sources.

Databanks have always been under competing pressures to provide data quickly and completely, but also to aim for optimal data quality. For suspect data (that do not meet data quality control) one possibility is to withhold until data have been corrected (assuming that this is possible) and the other is to release data with a suitable warning. We tend to prefer the latter regime and in future we will build semantic framework for this database in such a way that the community input will be included in decision making.

References

1. Koonin, E.V., Galperin, M.Y.: Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Current Opinions in Genetic Development* 7, 757–763 (1997)
2. Sarinder, K.K.S., Majid, M.A., Lim, L.H.S., Ibrahim, H., Merican, A.F.: Integrated Biological Database Initiative (IBDI). In: *Proceedings of International Conference on Biogeography and Biodiversity Wallace in Sarawak – 150 Years Later, Kuching, Malaysia* (2005)

3. Merican, A.F., Othman, R., Ismail, N., Cheah, K.P., Mok, L., Yin, Y.K.C., Kaur, S.: Development of Malaysian Indigenous Microorganisms Online Database System. *Asia Pacific Journal of Molecular Biology and Biotechnology* 10(1), 69–72
4. Guan, S.L., Kirton, L.G.: Biological Collections at the Forest Research Institute Malaysia. In: 22nd Pacific Science Congress, Kuala Lumpur (2011)
5. Napis, S., Salleh, K.M., Itam, K., Latiff, A.: Biodiversity Databases for Malaysian Flora and Fauna: An Update. In: Proceedings of Internet Workshop, National Institute of Informatics, Tokyo, Japan and High Quality Internet Study Group of Information Processing Society of Japan, IPSJ (2001)
6. Kamruzzaman, A.Z.M., Selamat, H., Wahid, M.T.: Conceptual Design of Biodiversity Data Model (BiDaM) using Object Relational and Event Based Approach. In: Proceedings of the Postgraduate Annual Research Seminar 2005, pp. 77–81 (2005)
7. Sprague, J., Bayraktaroglu, L., Clements, D., Conlin, T., Fashena, D., Frazer, K., Haendel, M., Howe, D.G., Mani, P., Ramachandran, S., et al.: The Zebrafish Information Network: the zebrafish model organism database about project success: an exploratory study. *Nucleic Acids Research* 34, 581–585 (2006)
8. Zouberakis, M., Chandras, C., Swertz, M., Smedley, D., Gruenberger, M., Bard, J., Schughart, K., Rosenthal, N., Hancock, J.M., Schofield, P.N., Kollias, G., Aidinis, V.: Database, vol. 2010 (2010)
9. Murthy, U., Fox, E.A., Chen, Y., Hallerman, E., da Silva Torres, R., Ramos, E.J., Falcão, T.R.C.: Superimposed Image Description and Retrieval for Fish Species Identification. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonias, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 285–296. Springer, Heidelberg (2009)
10. Kozievitch, N.P., da Silva Torres, R., Andrade, F., Murthy, U., Fox, E., Hallerman, E.: A teaching tool for parasitology: Enhancing learning with annotation and image retrieval. In: Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) ECDL 2010. LNCS, vol. 6273, pp. 466–469. Springer, Heidelberg (2010)
11. Lim, L.H.S.: Diversity of monogeneans in Southeast Asia. *International Journal for Parasitology* 28(10), 1495–1515 (1998)
12. Lim, L.H.S.: Parasites as indicators of present and past ecology of the environment. In: Tuen, A.A., Das, I. (eds.) Proceedings of the Wallace in Sarawak - 150 Years Later. An International Conference on Biogeography and Biodiversity, pp. 223–224 (2005)
13. Tisdall, J.D.: Mastering Perl for bioinformatics. O'Reilly, Sebastopol (2003)
14. Williams, N.: Bioinformatics: How to Get Databases Talking the Same Language. *Science* 275, 301–302 (1997)
15. Etzold, T., Argos, P.: SRS: An Indexing and Retrieval Tool for Flat File Data Libraries. *Computer Application of Biosciences* 9, 49–57 (1993)
16. Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet, D.: GeneCards: encyclopedia for Genes, Proteins, and Diseases. Weizmann Institute of Science, Bioinformatics Unit and Genome Center, Rehovot, Israel (1997)
17. Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y., Kanehisa, M.: DBGET/LinkDB: an Integrated Database Retrieval System. In: Pacific Symposium of Biocomputing. PSB Electronic Proceedings, Hawaii (1998)
18. Haas, L., Schwarz, P., Kodali, P., Kotlar, E., Rice, J., Swope, W.: DiscoveryLink: A system for integrated access to life sciences data sources. *IBM Systems Journal* 40, 489–511 (2001)
19. Davidson, S., Crabtree, J., Brunk, B., Schug, J., Tannen, V., Overton, C., Stoeckert, C.: K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources. *IBM Systems Journal* 40, 512–531 (2001)

20. Buneman, P., Davidson, S., Hart, K., Overton, C., Wong, L.: A Data Transformation System for Biological Data Sources. In: 21st International Conference on Very Large Data Bases (VLDB 1995). Morgan Kaufmann, Zurich (1995)
21. Lu, W., Jackson, J., Ekanayake, J., Barga, R.S., Araujo, N.: Performing Large Science Experiments on Azure: Pitfalls and Solutions. In: Proc. CloudCom, pp. 209–217 (2010)
22. Karp, P.D.: A strategy for database interoperation. *Journal of Computational Biology* 2, 573–583 (1996)
23. Ritter, O.: The integrated genomic database. In: Suhai, S. (ed.) *Computational Methods in Genome Research*. Plenum, New York (1994)
24. Ben-natan, R.: CORBA. McGraw Hill, New York (1995)
25. Achard, F., Barillot, E.: Ubiquitous distributed objects with CORBA. In: *Pacific Symposium on Biocomputing*, pp. 39–50 (1997)
26. Stein, L.D., Cartinhour, S., Thierry-mieg, D., Thierry-mieg, J.: JADE: An approach for interconnecting bioinformatics databases. *Gene* 209, 39–43 (1998)

Prediction of Vanillin Production in Yeast Using a Hybrid of Continuous Bees Algorithm and Flux Balance Analysis (CBAFBA)

Leang Huat Yin¹, Yee Wen Choon¹, Lian En Chai¹, Chuii Khim Chong¹,
Safaai Deris¹, Rosli M. Illias², and Mohd Saberi Mohamad^{1,*}

¹ Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia
{lhyin2, ywchoon2, lechai2, ckchong2}@live.utm.my,
{safaai, saberi}@utm.my

² Department of Bioprocess Engineering, Faculty of Chemical Engineering,
Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia
r-rosli@utm.my

Abstract. Most food and beverage is containing artificial flavor compound. Creation of artificial flavors is not an easy step and it is hardly ever completely effective. In this paper, we introduce an *in silico* method in optimization of microbial strains of flavor compound synthesis. Previously, there are several algorithms such as Genetic Algorithm, Evolutionary Algorithm, OptKnock tool and other related techniques are widely used to predict the yield of target compound by suggesting the gene knockouts. The used of these algorithms or tools is able to predict the yield of production instead of using try and error method for gene deletions. Nowadays, without using *in silico* method, the direct experiment methods are not cost effective and time consumed. As we know, the cost of chemical is expensive and not all flavorist able to afford the cost. However, the main limitations of previous algorithms are it failed to optimize the prediction of the yield and suggesting unrealistic flux distribution. Therefore, this paper proposed a hybrid of continuous Bees algorithm and Flux Balance Analysis. The target compound in this research is vanillin. The aim of study is to identify optimum gene knockouts. The results in this paper are the prediction of the yield and the growth rate values of the model. The predictive results showed that the improvement in term of yield which may help in food flavorings.

Keywords: Bees Algorithm, Flux Balance Analysis, Yeast, Optimization.

1 Introduction

Vanillin is normally used as food ingredient or flavor compound. In order to get vanillin by traditional method of obtaining vanillin is from cured seed pods of the Vanilla

* Corresponding author.

planifolia (natural vanillin) and via chemical synthesis. As we know, yeast is considered to be a workhorse of the biotechnology industry for the production of many value-added chemical, alcoholic beverages and biofuels [1]. Thus, in this research, *S. cerevisiae* model is used in vanillin. Figure 1 showed the *de novo* biosynthetic pathway in *S. cerevisiae* for vanillin production. In order to meet the aim of prediction of the vanillin yield by *in silico* method, the algorithm introduced in this paper is a hybrid of Continuous Bees Algorithm and Flux Balance Analysis which able to predict a set of gene knockouts. The contributions of this paper are three fold. First, up to our best knowledge, this method is first used in prediction of biochemical production where no other researchers used this method before. Second, this prediction algorithm implemented in this research is able to predict the gene knockouts in the large number of reactions in *S. cerevisiae* model. Third, the experimental results shown that the prediction algorithm in this research had given a set of relatedness deletion whereby the experimental technique in wet lab can be avoided before the expected result is confirmed. This will contribute in term of cost efficiency where the materials for experiment are expensive.

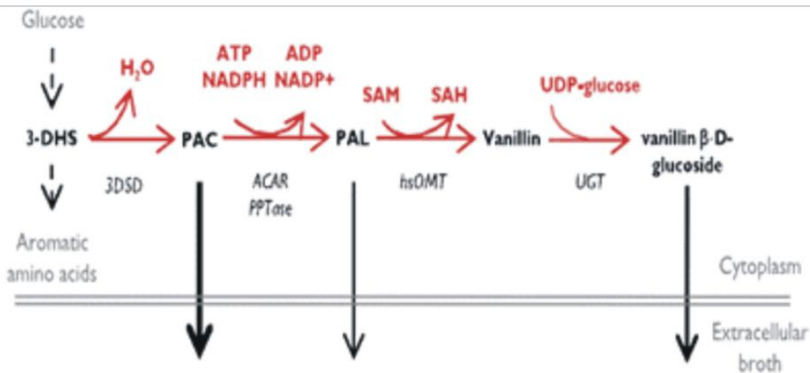


Fig. 1. The *de novo* biosynthetic pathway in *S. cerevisiae* for vanillin production.

Basically, the prediction of biochemical compounds is predicted by several algorithms such as Genetic Algorithm, Evolutionary Algorithm, OptKnock tool and OptGene which are widely used. Unfortunately, there is some limitation of those techniques. In this paper, the limitation of binOptGene is identified. In binOptGene, the representation of population is in binary variable where representation will form a set of “individual” representing a particular mutant. However, in this method the main problem is number of invalid individuals in population is larger and consequently negatively affects the convergence. It happened due to use of penalty functions after evaluation of individuals. Besides that, binOptGene also will suffer of several problems which causes by used of Genetic Algorithm in binOptGene or OptGene itself. One of the limitation of Genetic algorithm is stop criterion of the algorithm is not clear in every problem. In addition, it is tendency to converge towards local optima

rather than global optimum of the problem. In the measurement of the fitness in a single right/wrong problem, Genetic algorithm is failed to solve the problem efficiently.

In order to solve limitations, the usage of Bees algorithm is as an optimization algorithm. The Bees algorithm is known as a new population-based search algorithm [2]. This algorithm is able to search optimum solutions in large search space. However, in Bees algorithm the representation method of population is difference from binOptGene where it represented in integer number. In this way, number of genes to be deleted can be directly imposed by changing the size of the individuals. Besides that, Flux Balance Analysis (FBA) is used as an approach that widely used for studying and analyzing biochemical networks, in particularly the genome-scale metabolic network constructions that have been built in the past decade [3]. Flux Balance Analysis also known as a constraint-based modeling approach in which the stoichiometry of the underlying biochemical network constrains the solution [4]. Constraints applied in Flux Balance Analysis are represented in two (2) ways: Firstly, as equations that balance reactions input and secondly output and as inequalities that impose bounds on the system. Basically, this approach is used a mathematical modeling approach for analyzing the flow of metabolites in metabolic network.

2 A Hybrid of Continuous Bees Algorithm and Flux Balance Analysis (CBAFBA)

In this section, we describe the details of the proposed a hybrid algorithm, hybrid of CBAFBA. In CBAFBA algorithm, there are 3 main parts is explained in next subsections: initialization, neighborhood search, and assignment of the remaining bees for random search and obtained the solution of CBAFBA. The Figure 2 shows the flow chart of CBAFBA.

In next subsection, we describe the dataset used in this proposed. Measurement of the evaluation of the result obtained is the optimization of metabolic production method used to determine the growth rate is included in Flux Balance Analysis. The function is defines as below:

$$\begin{array}{l} \text{Maximize } Z \\ \text{Subject to} \end{array} \quad \sum_{j=1}^N S_{ij} v_j = 0, \quad i = 1, \dots, M \quad (1)$$

Thermodynamic and capacity constraints can be added as below:

$$\alpha_{-}(j) \leq v_j \leq \beta_{-} j, \quad j = 1, \dots, N \quad (2)$$

The Z is the linear objective function which to be minimize or maximize from particular metabolic engineering design objective to maximization of cellular growth of vanillin. The v_j corresponds to the rate of reaction j and S_{ij} is the stoichiometric coefficient. The different optimal solution obtained when different objective function is applied in optimization function.

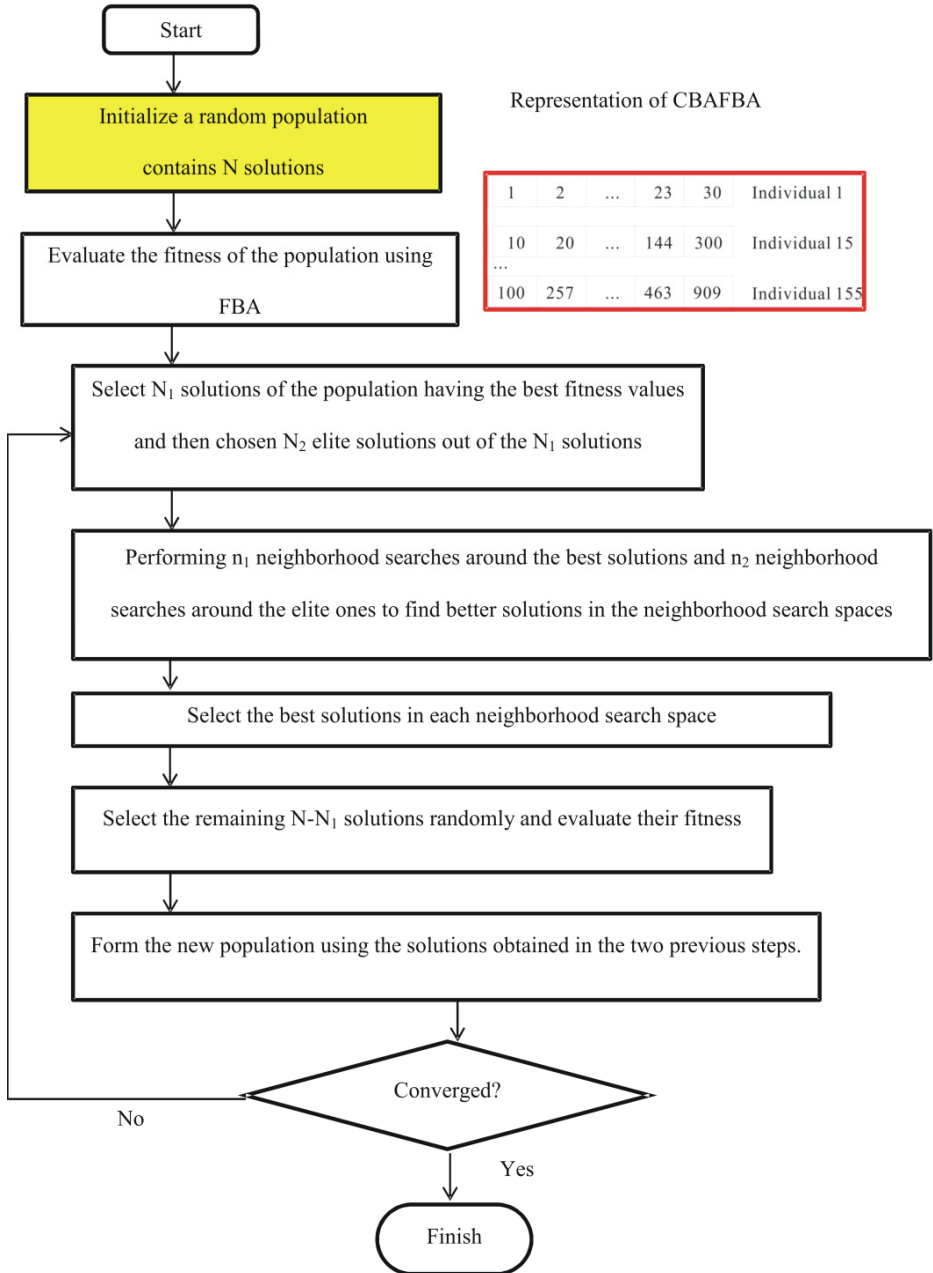


Fig. 2. The flow chart of CBAFBA

The difference of CBAFBA compare to binOptGene [5] and BAFBA [6] is the representation strategies in initialization of population. In binOptGene, the represented in binary which may cause the population become larger and consequently negatively affect the convergence. In CBAFBA, the population is represented in integer number [7]. The representation is showed in next section. Figure 3 shows the overall of OptGene algorithm which the representation of population is in binary variable. Besides that, the CBAFBA is able to search in global population where it formed a new population of each iteration had been completed. However, in binOptGene which applied Genetic algorithm is tendency to convert into local optimal rather than search for global optimal of the problem [8].

2.1 Initialization

Initialization of CBAFBA is the first step which used to create a random set of list which represent as population. The size of population can be set according to the size of dimension or search space needed. The bigger search space will cause the computational time increases. Therefore the search space is reducing by model pre-processing phase. In this initialization of population, the representation of individual is in integer number. The individuals are composed of integer numbers representing only the genes to be deleted. Therefore, it is based on the relative order in of metabolic model.

2.2 Neighborhood Search

In neighborhood search stage, there are 3 steps of Bees algorithm is executed. The selection of sites which has higher fitnesses, recruitment of bees for selected sites and selection of fittest bee from each search are the step in Bees algorithm. In order to select the sites with higher fitnesses, the sorting function is created whereby it sorts the fitnesses and positions of the population. The sorted list is in descending order.

After the population is sorted according to it fitness, the recruitment stage is begin. This recruitment stage is sending the bees around the fittest site and evaluated the fitness. The fitness of this research is based on the Flux Balance Analysis. In order to prevent very small values of the production at set growth rate, there is a comparison between minimum productions of population with a fixed minimum production.

2.3 Assignment of the Remaining Bees for Random Search and Obtained the Solution of CBAFBA

In this stage, CBAFBA is assigned the remaining bees for random search. The remaining bees are used to find the potential new solutions. The searched of potential new solutions is done around the search space. Again, in this stage the fitness by Flux Balance Analysis is used which will used the minimum production compare with fixed value. After the calculation of the fitness, the list of production is sorted in order to identify the best production. At final of each run, the prediction of the gene knock-out list will be generated after the CBAFBA algorithm is completely computed.

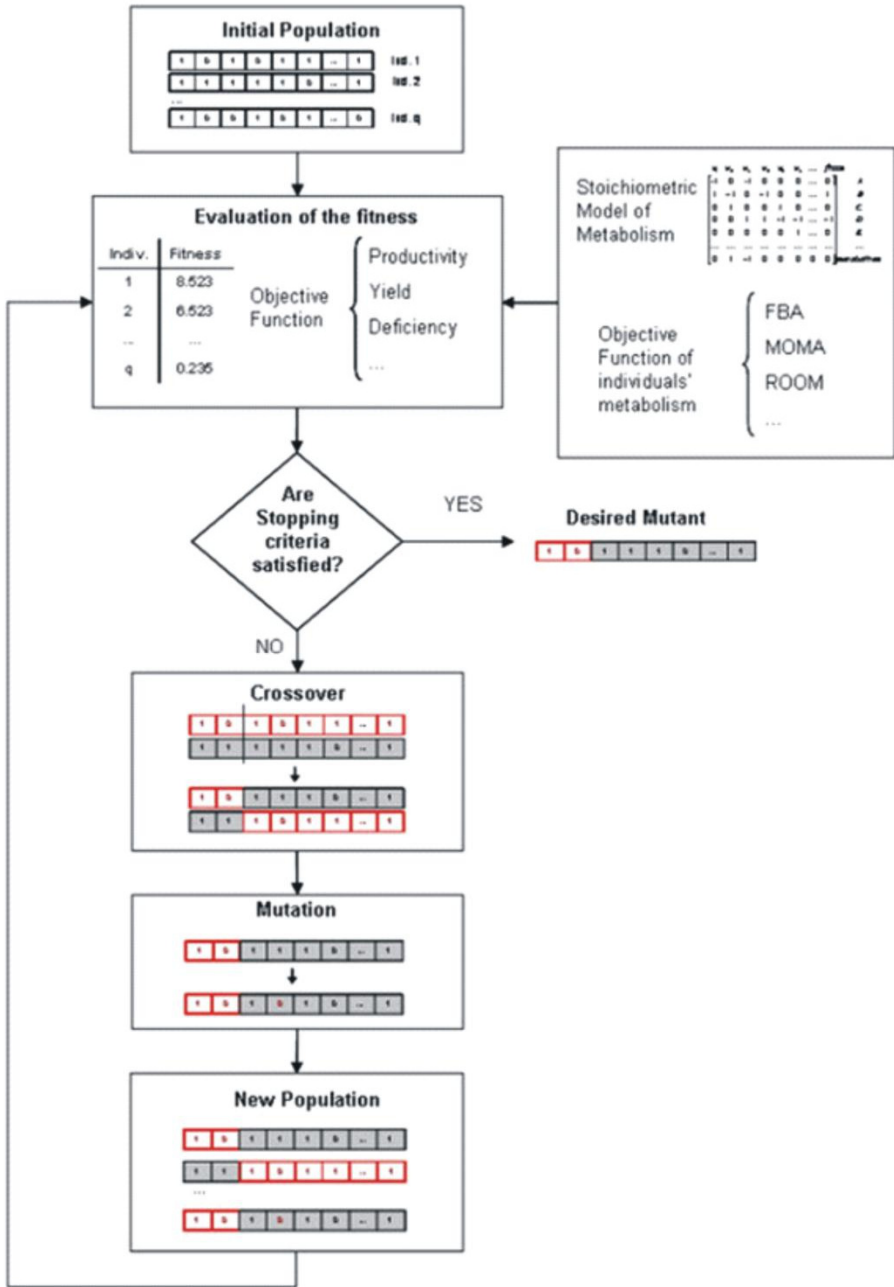


Fig. 3. The overall of OptGene algorithm

3 Results

3.1 Dataset

A model of *S. cerevisiae* dataset is used in the computational algorithms. Basically, *S. cerevisiae* is a type of baker's yeast. This dataset is originally obtained from Kyoto Encyclopedia of Gene and Genomes (KEGG). Yeast dataset from KEGG is then converted into System Biology Markup Language (SBML) format. In the model of *S. cerevisiae* dataset, all the pathways in Baker's yeast are included. From the abundant of pathways included in the dataset, several pathways is excluded or removed to reduce unrelated pathways and minimized the computational time during prediction process in on going. Thus, the model pre-processing is needed. The model pre-processing is include reduce dead-end reactions whereby problem size considerably small compare to initial model.

In this paper, the stoichiometric simulations provide an estimation of possible range of flux values for every reaction in the network. Due to the existence of a large number of alternatives flux routes or path-ways in genome-scale metabolic models require the use of optimization or computational methods to predict the alternative deletion of genes which will help to improve the production. The used of FBA are guaranteed to be optimal, but not necessarily unique due to the existence of a large number of pathways involves [8]. In this paper, the vanillin is the product to be predicted.

3.2 Vanillin Production

Selection of Target Reaction in Vanillin Prediction by Bees Algorithm and Flux Balance Analysis

Here, a core substrate which can contribute to vanillin production is L-phenylalanine which shows in Figure 4. In Figure 4, the formation or production of vanillin is known as biotransformation of aromatic acids. Generally, the production of vanillin increased when the production of L-phenylalanine increased. In the key reaction, phenylalanine is deaminated to transcinnamic acid, which catalyzed by phenylalanine ammonia lysase [9]. The transcinnamic acid then undergoes a chain reaction until reached the vanillin production which show in Figure 4.

In addition, the L-phenylalanine used as precursor to vanillin production in biosynthesis of alkaloid de-rived from shikimate pathway. The L-phenylalanine is synthesis from prehenate where prephenate dehydratase is an enzyme involved in the process. Next, the L-phenylalanine is involved in the synthesis of caffeoyl-CoA. There are series of process and enzymes involved in the synthesis of vanillin from caffeoyl-CoA. One of the enzymes involved is caffeoyl-CoA O-methyltransferase which catalyze the conversion of caffeoyl-CoA into feruloyl-CoA.

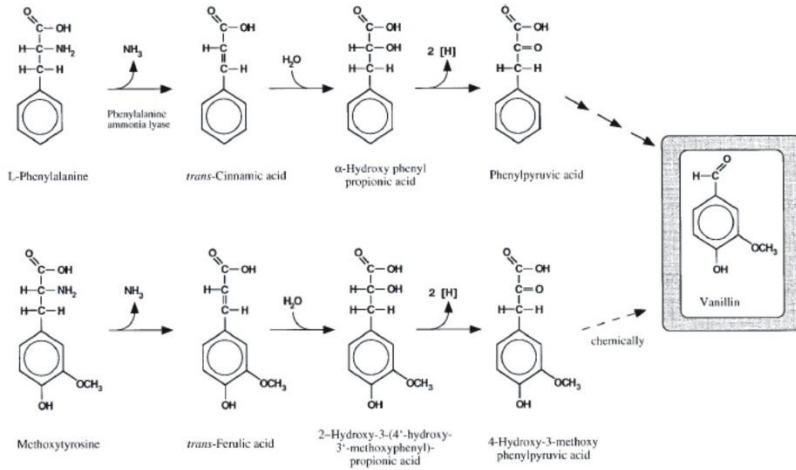


Fig. 4. Main contribution of vanillin production is L-phenylalanine

Therefore, the selection of synthesis of L-phenylalanine is contributed to vanillin production. In order to select the substrate and target reaction, glucose and prephenate dehydratase had been selected, respectively. The purpose prephenate dehydratase been chosen is due to this enzyme will affect the production of L-phenylalanine, a substrate to produce vanillin. Besides that, the purpose of glucose reaction been chose is due to the main substance of mostly biosynthesis in any pathways.

Selection of Gene Knockouts List Based on Growth Rate Prediction Using Flux Balance Analysis

Prediction strategies described in this work are based on the assumption that microbial cells would evolve in higher growth rates and biochemical production. As we knows, knockout mutants that force the coupling between biomass and biochemical production allow researcher to use growth rate as a selective pressure and find adaptively evolved strains with improved growth rates and production capabilities [10].

Figure 5 shows Average of frequency for predicted result cases in each growth rate values from same biochemical product which is vanillin. Based on that result, the higher growth rate is 1.7023 which indicates the cellular cell is alive and able to produce the desired product at optimal rate. Basically, the bio-chemical production would increase along with cellular growth rate [10].

On the other hand in this research, the minimum growth rate is $-5.4019\text{e-}013$ where this set of genes knockout is eliminate. The reason of the elimination is due to the cellular cell is unable to live or cell is death. The elimination is also because of the deletion of lethal gene. This deletion may cause the cellular cell die and fail to produce desire biochemical production.

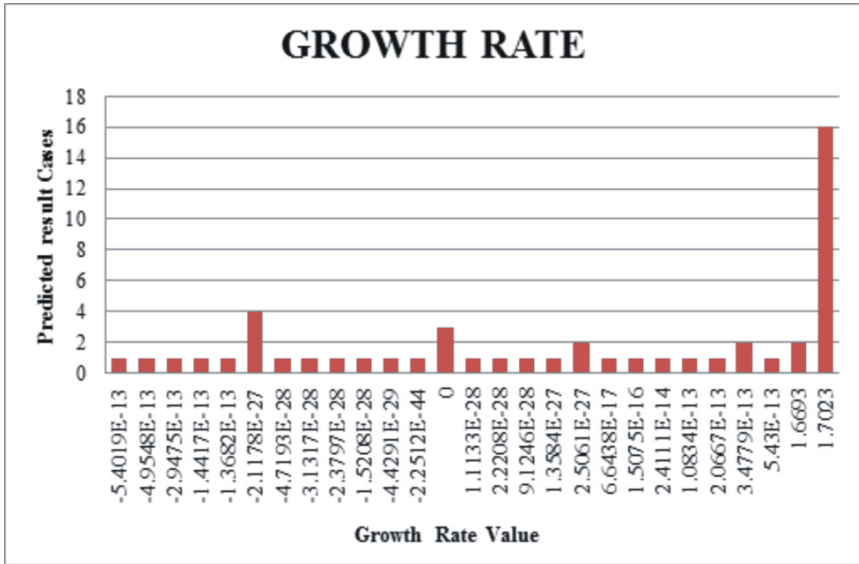


Fig. 5. Average of frequency for predicted result cases in each growth rate value

Selection of Genes Knockout List Based on Production Rate

In general, the knockout strategy consists of two approaches; reaction-based deletion and gene-based deletions. In this section, the genes-based deletion approaches are preferred compare to reaction-based deletion. As we know, the relationship between genes, proteins and reactions is not one-to-one. In other words, a metabolic reaction usually carries out by one or more enzymes where each of it can comprise to produce multiple gene products (proteins) [10]. The removal of multiple genes may affect the additional reactions by removal of additional reactions [10]. The worst scenario for removal of additional reactions is the reaction for desired product been removed incidentally.

After applied the proposed method in yeast model, the result shown that there are 3 mutants are obtained. Table 1 summarizes three of the identified gene knockout strategies for L-phenylalanine (i.e mutant A, B, and C). Based on the result obtained from CBAFBA, maximum yield is 0.19466 for three mutants in the phenylalanine target reaction. Here, the result for mutant A suggests that removal of shikimate pathway reaction from the network. In mutant A, ARO4 gene is being removed which it is encoded into 3-deoxy-D-arabino-heptulosonate 7-phosphate synthetase (DAHPS). Experimentally from web lab, DAHPS is used for condensation of erythrose-4-phosphate and phosphoenolpyruvate to 3-deoxy-d-arabino-heptulosonate-7-phosphate (DAHP) [11].

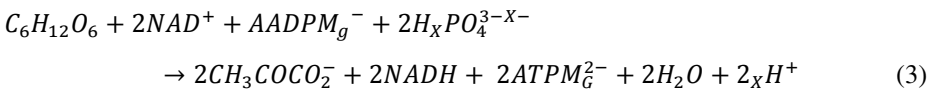
The removal in mutant A is less intuitive strategy which focuses on inactivating phosphoenolpyruvate (PEP) consuming reactions rather than eliminating competing

by product mechanism. With this strategy, some researches assuming that the maximum biomass yield could be attained. Note that the predicted yield in CBAFBA is assumed only by the theoretical maximum where further experiment from wet lab is needed for proofing purpose. Figure 6 reveals the flux distribution of the mutant A.

Table 1. List knockout for vanillin case study which control by L-phenylalanine compound

<i>L-Phenylalanine</i>			
Mutant	Gene	Knockouts	Enzyme
A	ARO4	PEP + D-erythrose 4-phosphate + H ₂ O = DHAP + phosphate	3-deoxy-D-arabino-heptulosonate 7-phosphate synthetase
B	BDH1	(R,R)-Butane-2,3-diol + NAD ⁺ = (R)-Acetoin + NADH + H ⁺	(R,R)-butanediol dehydrogenase
C	ARO10	Ehrlich pathway	2-isopropylmalate synthase

Second gene deletion is removal of butanoate metabolism pathway whereby CBAFBA suggested deletion of (R,R)-butanediol dehydrogenase. The gene involved in encoding (R,R)-butanediol dehydrogenase in *S. cerevisiae* is BDH1. This enzyme is involved in the formation of (R)-acetoin from (R,R)-butane-2,3-diol. In order to obtain (R)-acetoin in this reaction, the (R,R)-butanediol dehydrogenase is dependent on NAD⁺ which is the co-enzyme for butanoate metabolism [12]. Theoretically, this reaction is assumed to be affected by the glycolysis pathway in order to produce PEP. In the glycolysis pathway, NAD⁺ is used as a co-enzyme in the conversion of glyceraldehyde-3-phosphate into 3-phospho-D-glyceroyl-phosphate. The following balanced equation shows that the oxidation of glucose to pyruvate [13].



The strategy assumed by CBAFBA in mutant B shows that the production yield is 0.19466 and the growth rate is 1.7023 respectively after the BDH1 gene is deleted. Figure 7 shows the glycolysis pathway in *S. cerevisiae*.

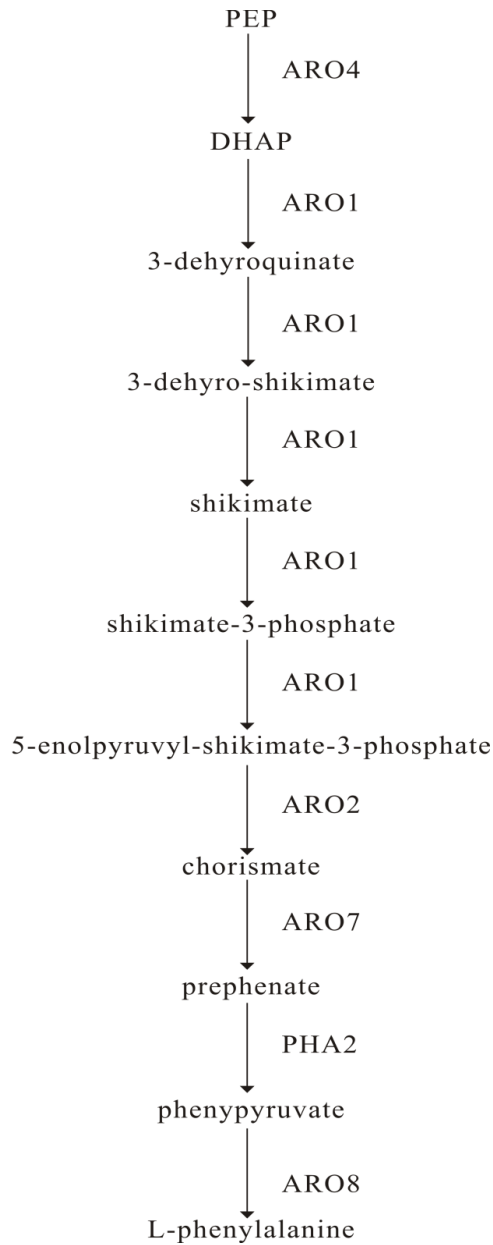


Fig. 6. The flux distribution of the mutant A

Third mutant (mutant C) suggests that deletion of ARO10 gene. The deletion of ARO10 gene is directly remove Ehrlich pathway of *S. cerevisiae*. The Ehrlich pathway is chemical reactions and pathways involved in the catabolism of amino acids to produce alcohols with one carbon less than the starting amino acid. The catabolism

process is involving the breaking down of molecules (amino acids) into smaller units (fusel alcohols). In *S. cerevisiae*, amino acids that assimilated by the Ehrlich pathway is taken up slowly throughout the fermentation time [14]. The amino acids usually taken up from Ehrlich pathway are valine, leucine, isoleucine, methionine, and phenylalanine. This will affect the production of the L- phenylalanine, where the yield will decrease if the Ehrlich pathway fails to be removed.

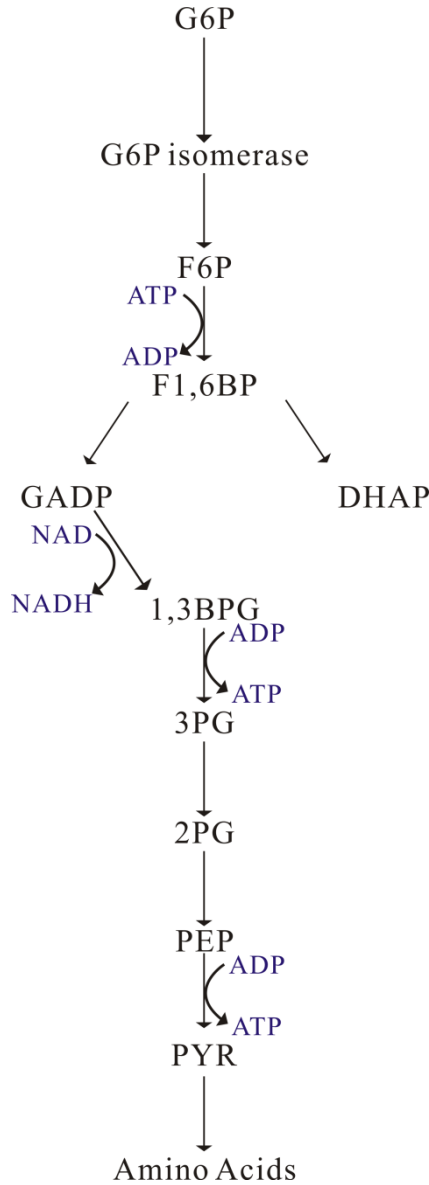


Fig. 7. The glycolysis pathway in *S. cerevisiae*

Besides that, in aerobic glucose-limited chemostat, phenylalanine is used as sole nitrogen sources, where phenylalanine is converted predominantly to fusel acids and only very low concentrations of fusel alcohols are formed [14]. However, without glucose-limited chemostat of *S. cerevisiae*, growth is pre-dominantly fermentative, and when phenylalanine is the sole nitrogen source, it is converted into a mix-ture of phenylethanol and phenylacetate [14]. This proves that the deletion of ARO10 is assumed to be increasing the phenylalanine yield. Figure 8 shows overall of the Ehrlich pathway.

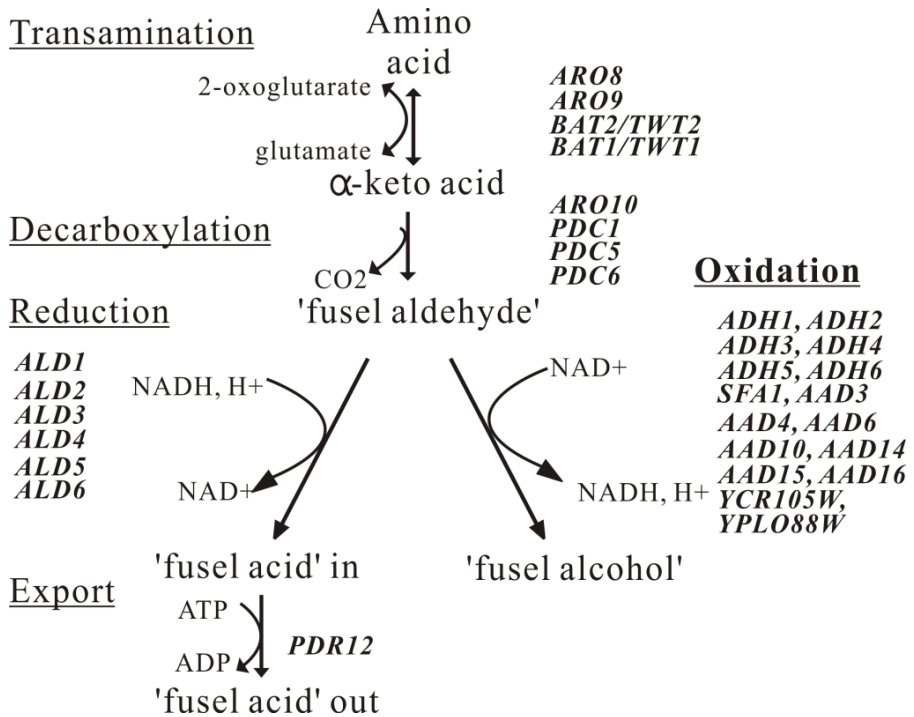


Fig. 8. Overall of the Ehrlich pathway

Table 2 shows the biomass overall yield of vanillin in batch cultivation after simulated by OptGene tool. The production of vanillin β-D-glucoside (VG) in table 2 shows the minimum yield (μ_{max}) is 0.10 (VG1) and the maximum yield is 0.2 (VG2). All stains (VG0, VG1, VG2, VG3, and VG4) are involved in either the removal or overexpression of pyruvate decarboxylase (PDC) and glutamate dehydrogenase (GDH). Based on Brochado and his colleagues, they believe that by in silico analysis, PDC was found as a target to increase formation of VG considering both respiratory and respire-fermentation reference flux distribution. By comparing with the proposed method in this research, the formation of vanillin is acceptable. The μ_{max} of this research is reached 0.19466 where slightly less than μ_{max} for VG2 in Table 2. However, the μ_{max} of this research is higher than μ_{max} of VG0, VG1, VG3,

and VG4 in table 2. Based on Brochado and his colleagues [8], they agreed that VG4 strain showed significantly improved in cellular fitness compare VG2 strain. This is due to the yield of biomass on substrate (glucose) is higher compare other strains.

Table 2. Biomass Overall Yield of Vanillin in Batch Cultivation [7]

Strains	Engi- neered Genotype	μ_{\max}	Y_{Sx}	$Y_{S\text{ Ethoh}}$	$Y_{S\text{ gly}}$
VG0		0.14	0.10	0.23	0.05
VG1	gdh1 Δ	0.10	0.07	0.25	0.03
VG2	pdh1 Δ	0.20	0.14	0.23	0.07
VG3	pdh1 Δ gdh1 Δ	0.11	0.10	0.27	0.05
VG4	pdh1 Δ gdh1 Δ \uparrow GDH2	0.17	0.17	0.25	0.07

Overall, suggested genes deletion by proposed method are included ARO4, BDH1 and ARO10 genes which can contribute to formation of L-phenylalanine, precursor for biotransformation of aromatic amino acids to produce vanillin. In theory, the more L-phenylalanine compound is produced, the more vanillin.

4 Conclusion

As a conclusion, our proposed CBAFBA which predicts the gene knockouts by *in silico* method showed to perform better in terms of time and cost-effective. The strategies applied in CBAFBA could lead to chemical production in *S. cerevisiae*. This is done by ensuring that the drain towards the metabolites/compounds necessary for growth resources such as carbons and energy must be accompanied. However, it should be noted that our proposed is deal with the reactions in the model not the real experiment. Therefore, the experiments are needed to carrier out to validate the deletion technique suggested by *in silico* technique. Specifically, CBAFBA is pinpoints which reactions needed to remove from a metabolic network, which can realized and contribute to yield the product by gene deletions where it associated with the identified the functionality. Reminder, it is important to note that the suggested gene

deletion strategies must be interpreted carefully. For instance, in many cases the deletion of gene in one branch of a branched pathway is equivalent to the significant up regulation in the other [15]. Lastly, the suggested set of gene(s) deletions is not always uniquely specified. Thus, the technique of identification of most economical gene set accounting for enzyme and multifunctional enzyme needs to be made.

Acknowledgments. We would like to thank Malaysian Ministry of Science, Technology and Innovation for supporting this research by an e-science research grant (Grant number: 06-01-06-SF1029). This research is also supported by UTM GUP research grant that was sponsored by Universiti Teknologi Malaysia.

References

1. Vargas, F.A., Pizzarro, F., Perez-Correa, J.R., Agosin, E.: Expanding a dynamic flux balance model of yeast fermentation to genome-scale. *BMC Systems Biology* 5, 75 (2011)
2. Pham, D.T., Ghanbarzadeh, A., Koc, E., Otri, S., Rahim, S., Zaidi, M.: The Bees Algorithm-A novel tool for complex optimisation problems. *Intelligent Production Machine and Systems* (2006)
3. Orth, J.D., Thiele, I., Palsson, B.O.: What is Flux Balance Analysis. *Nat. Biotechnol.* 28(3) (2010)
4. Kauffman, K.J., Prakash, P., Edwards, J.S.: *Advances in Flux Balance Analysis*, vol. 14, pp. 491–496. Elsevier (2003)
5. Patil, K.R., Rocha, I., Forster, J., Nielsen, J.: Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics* 6, 308 (2005)
6. Choon, Y.W., Mohamad, M.S., Deris, S., Chong, C.K., Chai, L.E., Ibrahim, Z., Omatu, S.: Identifying Gene Knockout Strategies Using a Hybrid of Bees Algorithm and Flux Balance Analysis for in silico Optimization of Microbial Strains. In: *The 9th International Symposium on Distributed Computing and Artificial Intelligence (DCAI 2012)*. University of Salamanca, Spain (2012)
7. Chaturvedi, D.K.: *Soft Computing Techniques and its Applications in Electrical Engineering*. SCI, vol. 103, pp. 363–381 (2008)
8. Brochado, A.R., Matos, C., Moller, B.L., Hansen, J., Mortensen, U.H., Patil, K.R.: Improved vanillin production in baker's yeast through in silico design. *Microbial Cell Factories* 9, 84 (2010)
9. Priefert, H., Rabenhorst, J., Steinbüchel, A.: Biotechnological production of vanillin. *Appl. Microbiol. Biotechnol.* 56, 296–314 (2001)
10. Kim, J.H., Reed, J.: OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. *BMC Systems Biology* 4, 53 (2010)
11. Helmstaedt, K., Strittmatter, A., Lipscomb, W.M., Braus, G.H.: Evolution of 3-deoxy-d-arabino-heptulosonate-7-phosphate synthase-encoding genes in the yeast *Saccharomyces cerevisiae* (2005)
12. González, E., Fernandez, M.R., Marco, D., Calam, E., Sumoy, E., Parés, X., Dequin, S., Biosca, J.A.: Role of *Saccharomyces cerevisiae* Oxidoreductases Bdh1p and Ara1p in the Metabolism of Acetoin and 2,3-Butanediol. *Applied and Environmental Microbiology*, 670–679 (2010)

13. Lane, A.N., Fan, T.W.M., Higashi, R.M.: Metabolic acidosis and the importance of balanced equation. *Metabolomics* 5, 163–165 (2009)
14. Hazelwood, L.A., Daran, J.M., van Maris, A.J.A., Pronk, J.T., Dickinson, J.R.: The Ehrlich Pathway for Fusel Alcohol Production: a Century of Research on *Saccharomyces cerevisiae* Metabolism. *Applied and Environmental Microbiology*, 2259–2266 (2008)
15. Burgard, A.P., Pharkya, P., Maranas, C.D.: OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering* 85(7) (2003)

Identifying Gene Knockout Strategy Using Bees Hill Flux Balance Analysis (BHFBA) for Improving the Production of Ethanol in *Bacillus Subtilis*

Yee Wen Choon¹, Mohd Saberi Mohamad¹, Safaai Deris¹, Rosli M. Illias²,
Lian En Chai¹, and Chuii Khim Chong¹

¹ Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia

{ywchoon2, lechai2, ckchong2}@live.utm.my, {saberim, safaaideris}@utm.my

² Department of Bioprocess Engineering, Faculty of Chemical Engineering,

Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia

r-rosli@utm.my

Abstract. *Bacillus subtilis* strains can be manipulated to improve product yield and growth characteristics. Optimization algorithms are developed to identify the effects of gene knockout on the results. However, this process is often faced the problem of being trapped in local minima and slow convergence due to repetitive iterations of algorithm. In this paper, we proposed Bees Hill Flux Balance Analysis (BHFBA) which is a hybrid of Bees Algorithm, Hill Climbing Algorithm and Flux Balance Analysis to solve the problems and improve the performance in predicting optimal sets of gene deletion for maximizing the growth rate and production yield of desired metabolite. *Bacillus subtilis* is the model organism in this paper. The list of knockout genes, growth rate and production yield after the deletion are the results from the experiments. BHFBA performed better in term of computational time, stability and production yield.

Keywords: Bees Algorithm, Hill Climbing, Flux Balance Analysis, *Bacillus subtilis*, Optimization.

1 Introduction

Microbial strains optimization has become popular in genome-scale metabolic networks reconstructions recently as microbial strains can be manipulated to improve product yield on desired metabolites and also improve growth characteristics [1]. Reconstructions of the metabolic networks are found to be very useful in health, environmental and energy issues [2]. The development of computational models for simulating the actual processes inside the cell is growing rapidly due to vast numbers of high-throughput experimental data.

Many algorithms were developed in order to identify the gene knockout strategies for obtaining improved phenotypes. The first rational modeling framework (named OptKnock) for introducing gene knockout leading to the overproduction of a desired

metabolite was developed by Burgard *et al.*, 2003 [3]. OptKnock identifies a set of gene (reaction) deletions to maximize the flux of a desired metabolite with the internal flux distribution is still operating such that growth is optimized.

OptKnock is implemented by using mixed integer linear programming (MILP) to formulate a bi-level linear optimization that is very promising to find the global optimal solution. OptGene is an extended approach of OptKnock which formulates the *in silico* design problem by using Genetic Algorithm (GA) [4]. Meta-heuristic methods are capable in producing near-optimal solutions with reasonable computation time, furthermore the objective function that can be optimized is flexible. SA is then implemented to allow the automatic finding of the best number of gene deletions for achieving a given productivity goal [5]. However, the results are not yet satisfactory.

A hybrid of BA and FBA was proposed by Choon *et al.*, 2012 [6], it showed a better performance in predicting optimal gene knockout strategies in term of growth rate and production yield. Pham *et al.*, 2006 [7] introduced Bees Algorithm (BA), is a typical meta-heuristic optimization approach which has been applied to various problems, such as controller formation [8], image analysis [9], and job multi-objective optimization [10]. BA is based on the intelligent behaviours of honeybees. It locates the most promising solutions, and selectively explores their neighbourhoods looking for the global maximum of the objective function. BA is efficient in solving optimization problems according to the previous studies [7, 10]. However, due to the dependency of BA on random search, it is relatively weak in local search activities [11]. Hence, BHFBA is proposed to improve the performance of BAFBA as Hill climbing algorithm is a promising algorithm in finding local optimum. This paper shows that BHFBA is not only capable in solving larger size problems in shorter computational time but also improves the performance in predicting optimal gene knockout strategy than previous works. In this work, we present the results obtained by BHFBA in two case studies where *B. subtilis* (*Bacillus subtilis*) iBsu1103 model is the target microorganisms [12]. In addition, we also conduct a benchmarking to test performance of the hybrid of Bee algorithm and Hill climbing algorithm.

2 Bees-Hill Flux Balance Analysis (BHFBA)

In this paper, we propose BHFBA in which BAFBA is only applied to identify optimal gene knockout strategies recently. Fig. 1 shows the flow of BAFBA while Fig. 2 shows our proposed BHFBA. The important steps are explained in the following subsections. Both BHFBA and BAFBA are using binary variables rather than continuous variables. The main difference between BHFBA and BAFBA is the neighbourhood search part, BHFBA improves the operation by combining hill climbing algorithm into BAFBA.

2.1 Model Pre-processing

The model is pre-processed through several steps based on biology assumptions as well as computational approaches to reduce the search space as while as increase the accuracy. Lethal reactions such as the genes that are found to be lethal *in vivo*, but not *in silico*, should be removed to improve the quality of the results. The results are

invalid if a lethal reaction is deleted. The following are the details of computational pre-processing steps to the model [5].

a. Fluxes that are not associated with any genes, such as the fluxes related to external metabolites and exchange fluxes that represent transport reaction should not be involved in the process. These fluxes do not have a biological meaning thus they should not be knocked out.

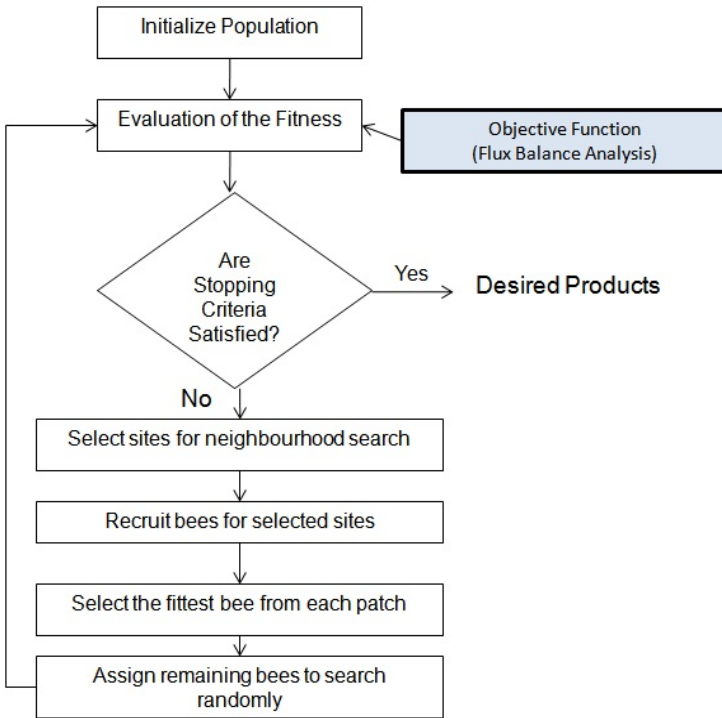
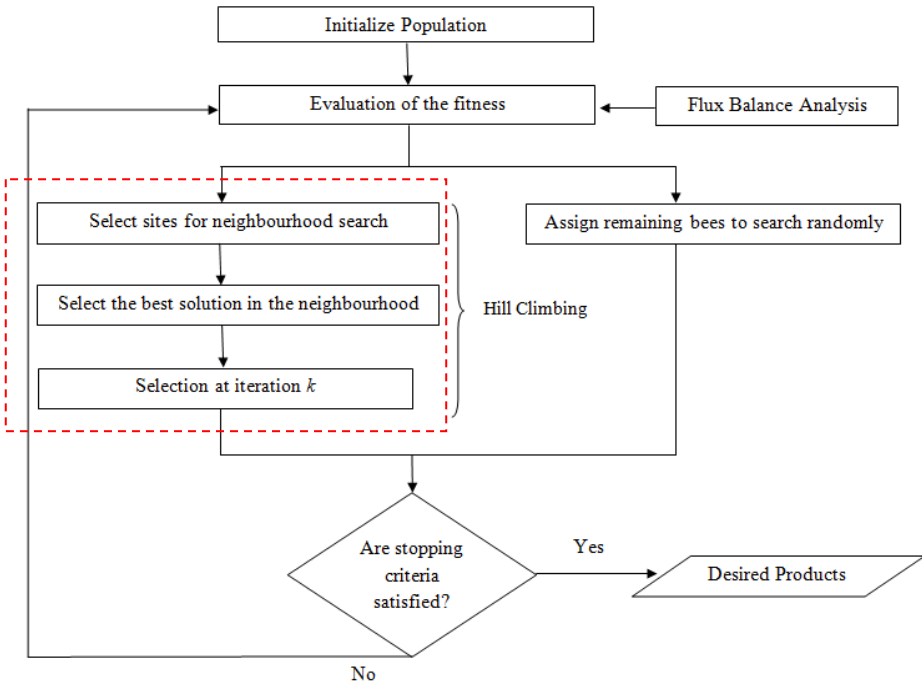


Fig. 1. BAFBA Flowchart

b. Essential genes that cannot be deleted from the microorganism's genome need to be removed. The search space for optimization is reduced due to that these genes should not be considered as targets for deletion. A linear programming problem is defined by setting the corresponding flux to 0, while maximizing the biomass flux for each gene in the microorganism's genome. If the biomass flux result from the Linear Programming algorithm is zero (or near zero) then the gene is marked as essential. This biological meaning of this fact is that the microorganism is unable to survive without this gene. This process does not suggest any changes to the model like the previous one, but provides favorable information for the optimization algorithms. With the help of biologists, the list of essential genes can be manually edited to include genes that are known to be essential *in vivo*, but not *in silico*.

c. Given the constraints of the linear programming problem, the fluxes need to be removed if the fluxes cannot exhibit values different from 0. Two linear programming are solved for every reaction in the model: the first is to define the flux over that

reaction as the maximization target, while the second is to set the same variable as minimization target. If the objective function is 0 for both problems, then the variable is removed from the model.



Note: Red-dotted box is Hill Climbing algorithm which is newly hybridized into BA.

Fig. 2. BHFBA Flowchart

2.2 Bee Representation of Metabolic Genotype

One or more genes can be discovered in each reaction in a metabolic model. In this paper, each of those genes is represented by a binary variable indicating its absence or presence (0 or 1), these variables form a ‘bee’ representing a specific mutant that lacks some metabolic reactions when compared with the wild type (Fig. 3.)

2.3 Initialization of the Population

The algorithm starts with an initial population of n scout bees. Each bee is initialized as follows: assume that a reaction with n genes. Bees in the population are initialized by setting present or absent status to each gene randomly. Initialization of the population is done randomly so that all bees in the population have an equal chance of being selected. The result might not truly reflect the population if it is done with bias setting.

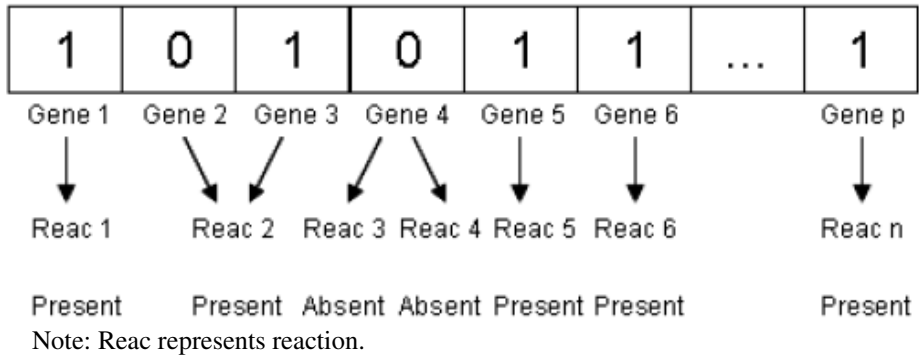


Fig. 3. Bee representation of metabolic genotype

2.4 Scoring Fitness of Individuals

Each site is given a fitness score that determines whether to recruit more bees or should be abandoned. In this work, we used FBA to calculate the fitness score for each site and the equation is as follow:

Maximize Z, where

$$Z = \sum_i c_i v_i = \mathbf{c} \cdot \mathbf{v} \tag{1}$$

where \mathbf{c} = a vector that defines the weights for of each flux.

Cellular growth is defined as the objective function Z, vector \mathbf{c} is used to select a linear combination of metabolic fluxes to include in the objective function, \mathbf{v} is the flux map and i is the index variable (1, 2, 3, ..., n). After optimizing the cellular growth, mutant with growth rate more than 0.1 continues the process by minimizing and maximizing the desired product flux at fixed optimal cellular growth value. Hence, we can enhance yield of our desired products at fixed optimal cellular growth. Production yield is the maximum amount of product that can be generated per unit of substrate. The following shows the calculation for production yield:

$$\text{Production yield} = \frac{(\text{production rate}_{\text{production}})}{(\text{consumption rate}_{\text{substrate}})} \tag{2}$$

(mmol/mmol)(gm/gm)

where mmol = millimole and gm is gram.

We used Biomass-product coupled yield (BPCY) as the fitness score in this work, the calculation for BPCY is as follow:

$$\text{BPCY} = \text{product yield} * \text{growth rate} \text{ (mmol(mmol*hr}^{-1}\text{))(gm (gm * hr}^{-1}\text{))} \tag{3}$$

where mmol is millimole, hr is hour and gm is gram.

2.5 Neighbourhood Search (Hill Climbing Algorithm)

The algorithm carries out neighbourhood searches in the favored sites (m) by using Hill climbing algorithm. Hill climbing is an iterative algorithm that starts with an arbitrary solution to a problem, then attempts to find a better solution by incrementally changing a single element of the solution. In this paper, the initial solution is the m favored sites from the population initialized by BA. The algorithm starts with the solution and makes small improvements to it by adding or reducing a bee to the sites. User defined the value of initial size of patches (ngh) and uses the value to update site (m) which is declared in the previous step to search in neighbourhood area. In this paper, m is equal to 15 and ngh is equal to 30, the values are obtained by conducting a small number of trials with the range of 10 to 25 and 20 to 35 respectively. This step is important as there might be better solutions than the original solution in the neighbourhood area.

2.6 Randomly Assigned and Termination

The remaining bees in the population are sent randomly around the search space to scout for new feasible solutions. This step is done randomly to avoid overlooking the potential results that are not in the range. These steps are repeated until either the maximum loop value is met or the fitness function has converged. At the end, the colony generates two parts to its new population – representatives from each selected patch and other scout bees assigned to perform random searches.

3 Results and Discussion

In this work, we use *E.Coli* and *B.subtilis* models to test on the operation of BAFBA. The *E.Coli* model is a small-scale model of the central metabolism of *E. coli* [12]. It is a modified subset of the iAF1260 model, and contains 134 genes, 95 reactions, and 72 metabolites. We use *E.coli* core model in this work because this model is useful for testing new constraint-based analysis methods, since the results of most constraint-based calculations are easier to interpret on this smaller scale. The second model is *B.subtilis* iBsu1103 model [13] which includes 1437 reactions associated with 1103 genes. We pre-process this model and the size is reduced to 763 reactions. The experiments are carried out by using a 2.3 GHz Intel Core i7 processor and 8 GB DDR3 RAM computer.

Table 1 and Table 2 summarize the results obtained from BHFBA for succinic acid production from *E.coli* and ethanol production from *B.subtilis*. Succinic acid is one of the intermediates of the TCA cycle and is a chemical to be used as a feedstock for the synthesis of a wide range of other chemical with several industrial applications. Besides, as a metabolite from the central carbon metabolism, succinic acid represents a good case study for identifying metabolic engineering strategies. Ethanol is a volatile, flammable, colourless liquid, and it is a promising biofuel. Ethanol is currently used as an alternative fuel for gasoline worldwide. As shown from the results, this method has produced better results to the previous works in term of growth rate and BPCY meanwhile potential genes which can be removed are identified [5][10].

Table 1. Comparison between different methods for production of Succinic acid in *E.coli*

Method	Growth Rate (1/hr)	BPCY	List of knockout genes
BHFBA	0.7988	0.93656	PTAr**, RPE, SUCD1i
BAFBA [10]	0.62404	0.66306	FUM, PTAr**, TPI**
SA + FBA [5]	N/A	0.39850	ACLD19*, DRPA, GLYCDx, F6PA, TPI**, LDH_D2, EDA, TKT2, LDH_D-
OptKnock [3]	0.28	N/A	ACKr, PTAr**, ACALD*

Note: The shaded column represents the best result. N/A – Not Applicable. * Common genes for all methods. ** Common genes in either 2 methods. BPCY is in gram (gram-glucose.hour)⁻¹.

Table 1 shows that BHFBA performed better than the previous works with growth rate 0.7988 and BPCY 0.93656. Knocking out succinate dehydrogenase (SUCD1i) interrupts the formation from succinic acid to fumarate. Without the conversion from succinic acid to fumarate, production yield of succinic is improved. Next, phosphotransacetylase (PTAr) is removed, according to Burgard *et al.*, 2003[3], the mutants can grow anaerobically on glucose by producing lactate. In the next step, ribulose-5-phosphate-3-epimerase (RPE) is suggested to knockout. This knockout involves the inflow reaction of ammonium. As stated in Bohl *et al.*, 2010 [14], the utilization of nitrate as electron acceptor and ammonium source under anaerobic conditions can improve succinate production. Figure 4 shows the comparison among the methods in term of growth rate and BPCY.

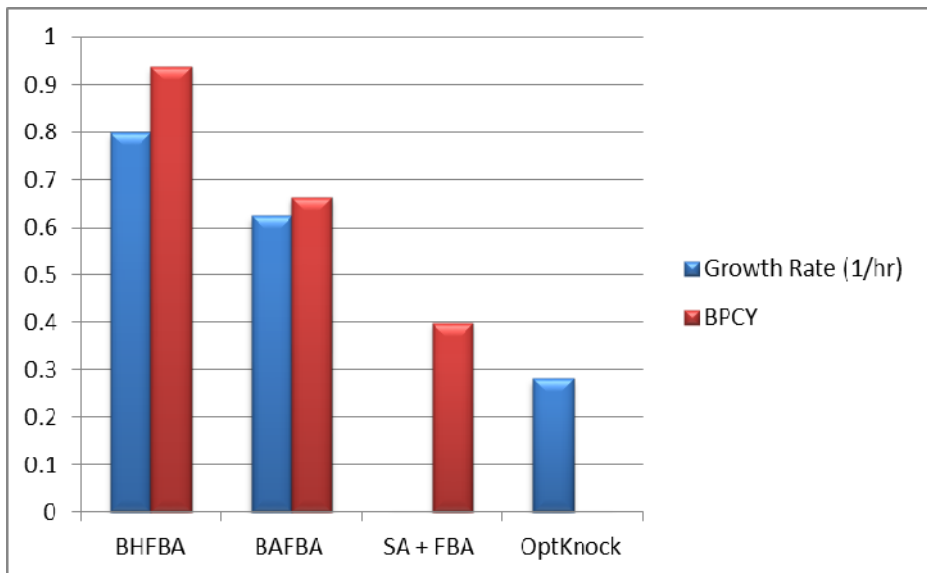

Fig. 4. Growth rate and BPCY comparison among available methods

Table 2. Comparison between different methods for production of ethanol in *B.subtilis*

Method	Growth Rate (1/hr)	BPCY	List of knockout genes
BHFBA	122.9089	1.15680e+05	ALAD_L, GPDH, LDH_L
BAFBA	122.8861	1.1154e+05	ALAD_L, LDH_L, XYLI1, inosose 2,3-dehydratase

Note: The shaded column represents the best result. N/A – Not Applicable. * Common genes for all methods. ** Common genes in either 2 methods. BPCY is in gram (gram-glucose.hour)⁻¹.

Table 3. Comparison between average computational time of BHFBA and BAFBA for 1000 iterations

Method	Computation Time (s)
BHFBA	7028
BAFBA	22515

Table 2 shows the results of BHFBA and previous works. BHFBA obtained a better growth rate and BPCY that are 122.9089 and 1.15680e+05 respectively. In the experiment of Kim *et al.*, 2012, deletion of NADH-dependent glycerol-3-phosphate dehydrogenase 1 (GPDH) showed a slight improvement in ethanol yield. As stated in Kim *et al.* (2012), lactate dehydrogenase (LDH_L) plays a key role in fermentative metabolism in metabolic engineering of *B.subtilis* for ethanol production. The deletion of LDH_L inhibits the conversion from pyruvate to lactate therefore more pyruvate is decarboxylated to acetaldehyde and further converted to ethanol.

Table 3 shows the computational time comparison between BHFBA and BAFBA for 1000 iterations. The average computational time of BHFBA improved 69% of the BAFBA result for 1000 iterations.

In addition, since BA and Hill Climbing algorithm is a new hybrid algorithm. Hence, we conducted a benchmarking to test performance of a hybrid of BA and Hill Climbing algorithm (BH). As BA is looking for the maximum, the functions are inverted before the algorithm is applied. The De Jong, Martin & Gaddy, and Griewangk functions are used in this paper. Table 4 shows the mathematical representation of the functions. Table 5 shows mean and standard deviation (STD) of the three functions, De Jong, Martin & Gaddy, and Griewangk, tested on both original BA and BH.

Table 4. Mathematical representation of De Jong and Beale functions

Name	Mathematical representation
De Jong	$\max F = (3905.93) - 100(x_1^2 - x_2)^2 - (1 - x_1)^2$
Martin & Gaddy	$\min F = (x_1 - x_2)^2 + ((x_1 + x_2 - 10) / 3)^2$
Griewangk	$\min F = 1 / (0.1 + (\sum(x(1,i)^2 / 4000)) - \sum(\cos(x(1,i) / \sqrt{i}) + 1))$

As seen from the results, both BHFBA and BH performed better than other algorithms. It can be concluded that the capability of Hill Climbing algorithm in finding local optimum improved the performance of the original BA. The original BA with the problem of repetitive iterations of the algorithm in local search where each bee keep searching until the best possible answer is reached. Our proposed BHFBA solved the problem by implementing Hill Climbing algorithm into the local search

part. Hill Climbing algorithm is a powerful local search algorithm which attempts to find a better solution by incrementally changing a single element of the solution until no further improvements can be found, the search process is recorded so the process is not repeated. Furthermore, one of the advantages of Hill Climbing algorithm is it can return a valid solution even if it is interrupted at any time before it ends.

Table 5. Obtained fitness value of both De Jong and Beale functions

Function	Mean		STD	
	BA	BH	BA	BH
De Jong	3.91e+03	3.90e+03	0.000504	4.79e-13
Martin & Gaddy	11.1083	11.1111	0.002797	0
Griewangk	-0.5263	-0.5263	5.76765E-09	0

4 Conclusion and Future Works

In this paper, BHFBA is proposed to predict optimal sets of gene deletion to maximize the production of desired metabolite. BHFBA improves the performance of BAFBA as Hill climbing algorithm is a promising algorithm in finding local optimum. Experimental results on *B.subtilis* iBsu1103 model obtained from literature showed that BHFBA is effective in generating optimal solutions to the gene knockout prediction, and is therefore a useful tool in Metabolic Engineering [12]. In the future, to improve the performance of BHFBA we are interested in applying an automated pre-processing function in BHFBA to refine the genome-scale metabolic model. We are also interested in applying other fitness functions in BHFBA such as minimization of metabolic adjustment (MOMA) and regulatory on/off minimization (ROOM) to further improve the performance of BHFBA. Besides that, BA employs many tunable parameters which are difficult for the users to determine so it is important to find ways to help the users choose suitable parameters.

Acknowledgement. Institutional Scholarship MyPhD provided by the Ministry of Higher Education of Malaysia finances this work. We would like to thank Malaysian Ministry of Science, Technology and Innovation for supporting this research by an e-science research grant (Grant number: 06-01-06-SF1029). This research is also supported by UTM GUP research grant that was sponsored by Universiti Teknologi Malaysia.

References

1. Feist, A.M., Herrgård, M.J., Thiele, I., Reed, J.L., Palsson, B.Ø.: Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* 7, 129–143 (2009)
2. Chandran, D., Copeland, W.B., Sleight, S.C., Sauro, H.M.: Mathematical modeling and synthetic biology. *Drug Discovery Today Disease Models* 5(4), 299–309 (2008)

3. Burgard, A.P., Pharkya, P., Maranas, C.D.: OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strains optimization. *Biotechnol. Bioeng.* 84, 647–657 (2003)
4. Patil, K.R., Rocha, I., Förster, J., Nielsen, J.: Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics* 6, 308 (2005)
5. Rocha, M., Maia, P., Mendes, R., Pinto, J.P., Ferreira, E.C., Nielsen, J., Patil, K.R., Rocha, I.: Natural computation meta-heuristics for the in silico optimization of microbial strains. *BMC Bioinformatics* 9, 499 (2008)
6. Pham, D.T., Ghanbarzadeh, A., Koç, E., Otri, S., Zaidi, M.: The bees algorithm – a novel tool for complex optimization problems. In: *Proceedings of the Second International Virtual Conference on Intelligent Production Machines and Systems*, July 3-14 (2006)
7. Choon, Y.W., Mohamad, M.S., Deris, S., Chong, C.K., Chai, L.E., Ibrahim, Z., Omatu, S.: Identifying Gene Knockout Strategies Using a Hybrid of Bees Algorithm and Flux Balance Analysis for in silico Optimization of Microbial Strains. In: *The 9th International Symposium on Distributed Computing and Artificial Intelligence (DCAI 2012)*. University of Salamanca, Spain (2012)
8. Pham, D.T., Darwish, A.H., Eldukhri, E.E.: Optimisation of a fuzzy logic controller using the bees algorithm. *International Journal of Computer Aided Engineering and Technology* 1(2), 250–264 (2006)
9. Olague, G., Puente, C.: The honeybee search algorithm for three-dimensional reconstruction. In: Rothlauf, F., Branke, J., Cagnoni, S., Costa, E., Cotta, C., Drechsler, R., Lutton, E., Machado, P., Moore, J.H., Romero, J., Smith, G.D., Squillero, G., Takagi, H. (eds.) *EvoWorkshops 2006*. LNCS, vol. 3907, pp. 427–437. Springer, Heidelberg (2006)
10. Pham, D.T., Ghanbarzadeh, A.: Multi-objective optimisation using the bees algorithm. Paper. In: *Proceedings of the Third International Virtual Conference on Intelligent Production Machines and Systems*, July 2-13 (2007)
11. Cheng, M.Y., Lien, L.C.: A Hybrid Swarm Intelligence Based Particle Bee Algorithm for Benchmark Functions and Construction Site Layout Optimization. In: *Proceedings of the 28th ISARC*, Seoul, pp. 898–904 (2011)
12. Orth, J.D., Fleming, R.M.T., Palsson, B.Ø.: *Reconstruction and Use of Microbial Metabolic Networks: the Core Escherichia coli Metabolic Model as an Educational Guide*. ASM Press, Washington, DC (2009)
13. Henry, C.S., Zinner, J.F., Cohoon, M.P., Stevens, R.L.: iBsu1103: a new genome-scale metabolic model of *Bacillus subtilis* based on SEED annotations. *Genome Biology* 10, 69 (2009)
14. Bohl, K., de Figueiredo, L.F., Hadicke, O., Klamt, S., Kost, C., Schuster, S., Kaleta, C.: CASOP GS: Computing intervention strategies targeted at production improvement in genome-scale metabolic networks. In: *The 5th German Conference on Bioinformatics 2010*, September 20-22 (2010)
15. Kim, J.W., Chin, Y.W., Park, Y.C., Seo, J.H.: Effects of deletion of glycerol-3-phosphate dehydrogenase and glutamate dehydrogenase genes on glycerol and ethanol metabolism in recombinant *Saccharomyces cerevisiae*. *Bioprocess Biosyst. Eng.* 35, 49–54 (2012)

A Hybrid of Artificial Bee Colony and Flux Balance Analysis for Identifying Optimum Knockout Strategies for Producing High Yields of Lactate in *Echerichia Coli*

Seet Sun Lee¹, Yee Wen Choon¹, Lian En Chai¹, Chuii Khim Chong¹, Safaai Deris¹,
Rosli M. Illias², and Mohd Saberi Mohamad^{1,*}

¹ Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia

{sslee6, ywchoon2, lechai2, ckchong2}@live.utm.my,
{safaai, saberi}@utm.my

² Department of Bioprocess Engineering, Faculty of Chemical Engineering,
Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia

r-rosli@utm.my

Abstract. The advent of genome-scale models of metabolism has laid the foundation for the development of computational procedures for suggesting genetic manipulations that lead to overproduction. Previously, for increasing the production of Lactate in *E. coli*, a traditional method of chemical synthesis was being used, this always lead the products are far below their theoretical maximums. This is not surprise as the cellular metabolism is always competing with the chemical overproduction. Besides, several optimization algorithms often get stuck at a local minimum in a multi-modal error. In this research, a hybrid of Artificial Bee Colony (ABC) and Flux Balance Analysis (FBA) is proposed for suggesting gene deletion strategies leading to the overproduction of Lactate in *E. coli*. In this work, the ABC is introduced as an optimization algorithm based on the intelligent behavior of honey bee swarm. As for the evaluation of fitness part, each mutant strain is evaluated by resorting to the simulation of its phenotype using the FBA, together with the premise that microorganisms have maximized their growth along natural evolution. This is the first research that successfully combined ABC and FBA for identifying optimum knockout strategies. The successfully created hybrid algorithm is applied to the *E. coli* model dataset.

Keywords: Artificial Bee Colony, Flux Balance Analysis, Lactate, Gene KnockOut, *Echerichia Coli*.

1 Introduction

There is a genetic technique called gene knockout where the one of the organism's genes is being made to inoperative, just like to knock out the specific gene from the

* Corresponding author.

organism. This technique is a platform for human to learn about how a gene functions based on the sequenced gene. Researchers draw inferences from the difference between the knockout organism and normal individuals.

Besides, the term also refer as creating a new organism as “knocking out” a gene, this is essentially opposite of a gene knocking. Double knockout has the meaning of two genes being knocked out at the same time. The same meaning goes to triple knockout and quadruple knockout which describes the 3 and 4 genes being knocked out simultaneously.

Succinate and its derivatives have been used as common chemicals to synthesize polymers, as additives and flavoring agents in foods, supplements for pharmaceuticals, or surfactants. Currently, it is mostly produced through petrochemical processes that can be expensive and have significant environmental impacts. In fact, the knockout solutions that lead to an improved phenotype regarding the production of Succinates are not straightforward to identify since they involve a considerable number of interacting reactions.

Lactate and its derivatives have been used in a wide range of food-processing and industrial applications like meat preservation, cosmetics, oral and health care products and baked goods. Additionally, as lactate can be easily converted to readily biodegradable polyesters, it is emerging as a potential material for producing environmentally friendly plastics from sugars [1].

Several microorganisms have been used to commercially produce lactate [2], such as *Lactobacillus* strains. However, those bacteria also have undesirable traits, such as a requirement for amino acids and vitamins which complicates acid recovery. *E. coli* has many advantageous characteristics as a production host, such as rapid growth under aerobic and anaerobic conditions and simple nutritional requirements. Moreover, well-established protocols for genetic manipulation and a large knowledge on this microbe's physiology enable the development of *E. coli* as a host for production of optically pure D- or L-lactate by metabolic engineering [3].

The first approach to suggest gene deletion strategies was the OptKnock algorithm, where mixed integer linear programming (MILP) is used to reach an optimum solution. An alternative approach was proposed by the OptGene algorithm that considers the application of Evolutionary Algorithms (EAs), EAs are a meta-heuristic optimization method, and they are capable of providing solutions in a reasonable amount of time.

Unfortunately, for the above approaches, they may often get stuck at a local minimum in a multi-modal error. Based on this, above algorithms might not perform well in global and local optimization which will lead to local minimum and inefficiently used for multivariable and multimodal functions optimization [4]. Therefore, a combination of Artificial Bee Colony (ABC) and Flux Balance Analysis (FBA) has been looked into for identifying the gene knockout strategies for obtaining high yields of Succinate in *E. coli*. The developed algorithm is evaluated in term of the production of biochemical in *E. coli*.

The successfully created hybrid algorithm has contributed to the gene knockout field where it can design the experiment protocol so that biochemical production will be increased. Before this, there is no research is being carried out for the hybrid of

these two algorithms. Moreover, the newly formed hybrid algorithm is applied on the *E. coli* dataset.

2 Methods

2.1 Hybrid of Artificial Bee Colony and Flux Balance Analysis

In this section, we describe the details of the proposed ABCFBA in which ABC is newly combined with FBA to identify optimal gene knockout strategies. In essence, the proposed algorithm consists of five main steps:

1. Initialize population
2. Employed phase
3. Onlooker phase
4. Memorize the best
5. Scout phase

Figure 1 shows the flow of ABCFBA. Figure 2 shows the comparison of ABCFBA and ABC, rectangles that in red color showed the difference steps between the original ABC and ABCFBA. The flow chart on the left side indicates the original ABC algorithm while the flow chart on the right side is the ABCFBA. As compare with the original ABC, this study's method has integrated the FBA into ABC for the purpose of fitness calculation which main for identifying optimum knockout strategies in *E. coli* model.

Originally, the ABC is main for food foraging of honey bees, therefore, its fitness calculation is the nectar amounts calculation while ABCFBA is focusing on the gene knockout identification, so its fitness calculation step will be replaced by FBA.

Based on Edwards and Palsson [5], FBA was developed to analyze the metabolic capabilities of a cellular system based on the mass balance constraints. The mass balance constraints in a metabolic network can be represented mathematically by a matrix equation as follow:

$$S \cdot v = 0 \quad (1)$$

The matrix S is the $m \times n$ stoichiometric matrix, where m is the number of metabolites and n is the number of reaction in the network. The vector v represents all fluxes in the metabolic network, including the internal fluxes, transport fluxes and the growth flux.

For the *E. coli* metabolic network, the number of fluxes was greater than the number of mass balance constraints; thus, there were multiple feasible flux distributions that satisfied the mass balance constraints, and the solutions were confined to the null space of the matrix S . as follow:

$$\alpha_i \leq v_i \leq \beta_i \quad (2)$$

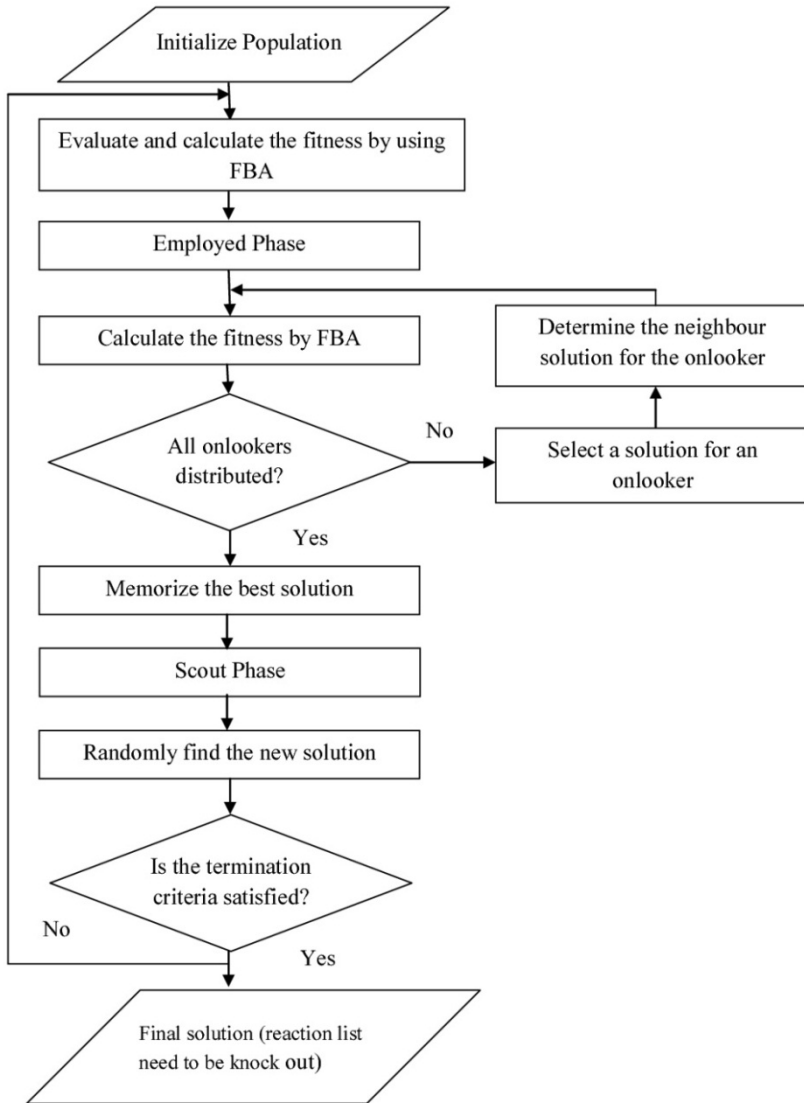


Fig. 1. Flow of ABCFBA

The linear inequality constraints were used to enforce the reversibility of each metabolic reaction and the maximal flux in the transport reactions. The reversibility constraints for each reaction are indicated online. The transport flux for inorganic phosphate, ammonia, carbon dioxide, sulfate, potassium, and sodium was unrestrained ($\alpha_i = -\infty$ and $\beta_i = \infty$).

The transport flux for the other metabolites, when available in the in silicon medium, was constrained between zero and the maximal level ($0 \leq v_i \leq v_{i,max}$). The $v_{i,max}$ values used in the simulations are noted for each simulation. When a

metabolite was not available in the medium, the transport flux was constrained to zero. The transport flux for metabolites capable of leaving the metabolic network was always unconstrained in the net outward direction.

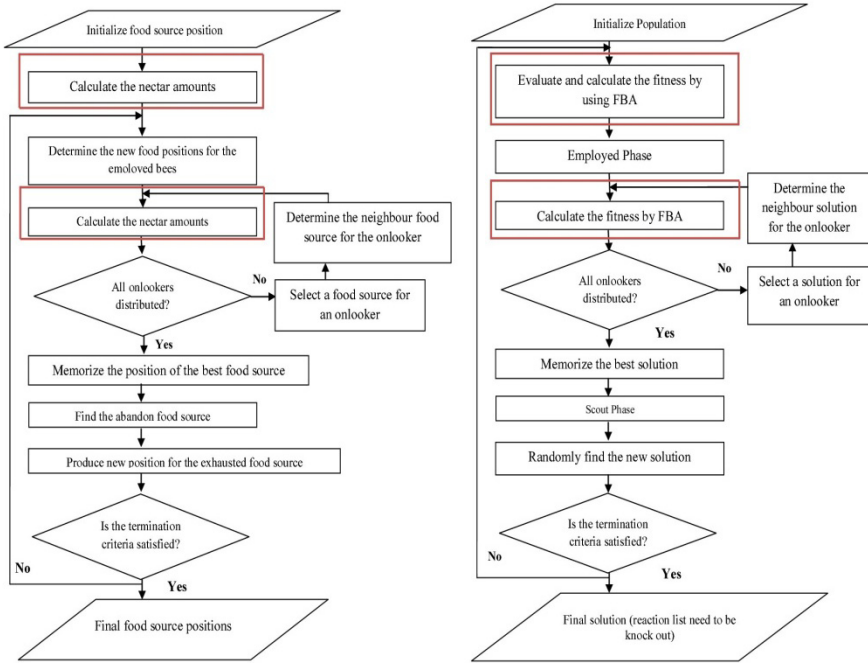


Fig. 2. Comparison of ABCFBA and ABC

The intersection of the nullspace and the region defined by the linear inequalities defined a region in flux space that we will refer to as the feasible set and the feasible set defined the capabilities of the metabolic network subject to the imposed cellular constraints. It should be noted that every vector v within the feasible set is not reachable by the cell under a given condition due to other constraints not considered in the analysis. The feasible set can be further reduced by imposing additional constraint and in the limiting condition where all constraints are known, the feasible set may reduce to a single point.

A particular metabolic flux distribution within the feasible set was found using Linear Programming (LP). LP identified a solution that minimized a metabolic objective function, and was formulated as shown below:

$$\text{Minimize } -Z \tag{3}$$

where $Z = \sum c_i v_i = \langle c \cdot v \rangle$

The vector c was used to select a linear combination of metabolic fluxes to indicate in the objective function. Herein, c was defined as the unit vector in the direction of

the growth flux, and the growth flux was defined in terms of the biosynthetic requirement.

$$\sum_{all\ m} dm \cdot Xm \xrightarrow{V_{grow}} \text{Biomass} \quad (4)$$

where dm is the biomass composition of metabolite Xm , and the growth flux was modeled as a single reaction that converts all the biosynthetic precursors into biomass.

As compare to other well-known metabolic modeling approaches, FBA is different in term of accuracy, this is because instead of predicting the metabolic behavior, it defines the 'best' the cell can do. FBA assumes that the regulation is such that metabolic behavior is optimal but not directly considers regulation or the regulatory constraints. Therefore, the results are generally consistent. However, it is only valid for a system that has evolved toward optimally.

In mutant strains, the regulation of the metabolic network has not evolved to operate in an optimal fashion. Because of this, it will cause a problem when coupling to highly parallel experimental programs, such as large-scale mutation studies.

FBA is an effective tool for the analysis of metabolic networks. FBA can complement the uncertainly and incompleteness of metabolic data, and thereby provide a better characterization of cellular phenotypes. Recent advances in FBA include the prediction of flux distribution of engineered cells, investigation of a cellular objective and the design of a mutant strain with desired properties.

Although the development of analytic techniques has facilitated the generation of dynamic profiles of metabolites, such data sets are not accurate enough for generating large-scale kinetic models. FBA has its pro and con in analyzing the biological network.

Initialize Population

The system start with create a population with the matrix of 95x500, since there are 95 reactions in the E. coli model and the dimension of the matrix where it must more than the number of reactions which is 500. This matrix was essentially create with all value 0's, then the value 1's were randomly distributed among them. The 1's represent those reaction that will be knockout while the 0's represent those reaction that cannot be knockout.

After the population has been created, each line of the columns, the population of the possibility of the reaction knockout, will be the inputs of the FBA for calculating the fitness. The system will return growth rate which determine whether the cell still survive after the deletion occurs where the value must more than 0.1. Another value that will return is the minimum production which represents the minimum production of biochemical after the deletion occurs where it must be more than $-1e-3$ to prevent the very small values from being considered as improvement.

Employed Phase

As for this stage, it is performing a job of randomly creating a new population where it is near the original population. For the 500 populations that created from the first

stage, the system will randomly create another 500 populations. Then, the greedy algorithm will be applied so that those with the smaller value of fitness will be abandoned. The new generated population will be formed with the better fitness values.

The greedy algorithm is based on the evaluation of a pre-defined maximum number of solutions that are obtained in the neighborhood of the best ones found and by using exhaustive search when no local search can be performed.

There might be some original populations that have the higher fitness value than newly formed possibilities. This showed the current solution cannot be improved. Therefore, the control parameters, trial, will increase by 1. Otherwise, it will remain as 0.

Onlooker Phase

The onlooker phase basically is randomly generating other neighborhoods, but it has a little bit different with employed phase. This phase started with calculating the probability value p for the fitness, fit_i values by using the formula:

$$P_i = \frac{fit_i}{\sum_{i=1}^{CS} fit} \quad (5)$$

The highest values of that specify possible reaction knockout will be the input of this phase, the system will randomly generate another new population and compare with the old one. If newly formed has higher fitness than the older formed, it will replace the older, vice-versa situation happens on the other hands.

Then, the system will recalculate the value p to decide the next population that will be replaced or remained. This phase will iterate till 500 times. Those populations that cannot be replaced will increase the trial value by 1 while those that have been replaced will set the trial to 0.

This will result the good potential population will become better while the bad population will be abandoned forever as the p values of good populations will keep increasing while the bad populations' p values will keep decreasing since the fitness value is divided by the sum of all the fitness values for every population.

Memorize the Best

After going through the three phases, the 500 populations will be the input for this stage. The best population will be selected based on the fitness value by using Greedy Selection algorithm. Only one population will remain as a result where it represents the best reactions knockout list in terms of highest growth rate and highest yields of target biochemical productions.

Scout Phase

If the population cannot be improved where its predetermined number of trials has exceeded the limit=100, the population is considered exhausted, it will be abandoned. The

Employed bee will immediately transform become the scout bee, then it will randomly generate another new population, evaluate and calculate the new fitness.

Then the phase will loop back to the calculation of fitness, go through the employed phase, onlooker phase, and memorize the best phase again and again. The system will repeat the cycle until it satisfies the termination criteria which is the maxCycle more than 200.

Finally, the result obtained has the best fitness value where it is the best knock out reaction list in the model in term of local and global search in ABC algorithm.

3 Results and Discussion

3.1 Experimental Data

In this research, the *E. coli* K-12 stoichiometric model [5] dataset is being used. The dataset can be found in BioModels Database, KEGG and system biology research group. The datasets is in SBML format.

This model basically contained 26 fields. Since the research is mainly focus on the biochemical productions which results from deletion of certain pathways or reactions that happened inside the cell. Therefore, the fields that used in this research are rxns, lb, ub and rxnNames.

For rxns field, it contained of 95 reactions (see Figure 3) which are the input data set and based on the knockout number, the system will identify which reactions can be deleted as to result a high yield of biochemical productions.

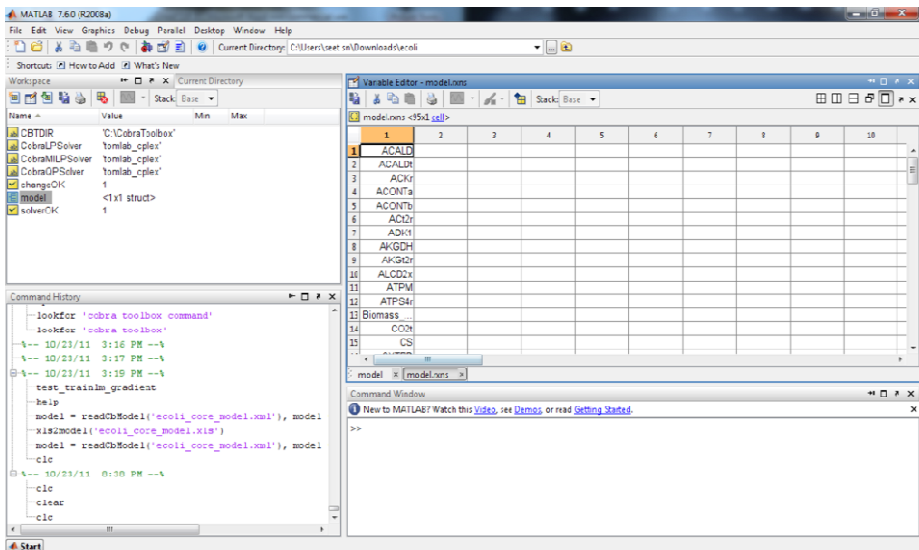


Fig. 3. 95 Reactions in the rxns field

While for lb and up, they stands for the lower boundary and upper boundary of the 95 reactions respectively. Last but not least, the rxnNames is the full names of the 95 reactions that can be understand by human language.

The target reaction in this research is lactate. Table 1 shows 3 sets of knockout list as result after 50 runs. The deletion of gene *adhE* which formed enzyme Alcohol dehydrogenase (ethanol) will increase the production of Lactate to 18.0738 mol per hour. The growth rate of the *E.coli* is 0.1186 indicates the cells still survived after the deletion. According to Q. Hua *et al.* [6], mutation in gene such *adhE* in the anaerobic environment, the lactate secretion will significantly increase. Gene *adhE* is catalyzing the reduction of acetyl-CoA to ethanol. After the deletion of *adhE*, the more highly reduced fermentation byproduct ethanol cannot be produced, NAD⁺ regeneration will mainly depend on the reduction pathway of pyruvate to lactate. Therefore the Lactate production will keep on increasing.

The deletion of genes *ackA* and *adhE* has the same lactate production as previous deletion, 18.0738 mol per hour. Although the lactate production remains the same but the growth rate for this deletion is higher than the previous deletion, which is 0.1253 if compare to 0.1186. L. Zhou. *et al.* [7]'s study stated in strain B0013, acetate is the main byproduct, the encoding gene (*ackA*) was initially deleted to reduce acetate yield and to increase lactate yield. Acetate kinase catalyzes the conversion of pyruvate via acetyl coenzymeA (CoA) and acetyl-phosphate to acetate. By deleting gene *ackA*, the main pathway for acetate production in *E. coli* has been restricted. Since the inhibitor of lactate production was disappear, then lactate will be produced significantly.

Table 1. KnockOuts list for the target reaction of Lactate in *E.coli*

KnockOuts	Enzyme	Lactate (gram-glucose.hour) ⁻¹	Growth Rate (h ⁻¹)
1 NAD + 1 ETOH <==> 1 NADH + 1 H + 1 ACALD	Alcohol dehydrogenase (ethanol)	18.0738	0.1186
ACTP + ADP <== > AC + ATP 1 NAD + 1 ETOH <==> 1 NADH + 1 H + 1 ACALD	Acetate kinase Alcohol dehydrogenase (ethanol)	18.0738	0.1253
FADH2 + Fumarate <==> FAD + SUCC	Fumarate reductase	18.0738	0.1253

The deletion of gene *frdA*, Fumarate reductase, has generated 18.0738 mol of lactate per hour, and the growth rate is 0.1253. Both biomass and growth rate achieved the same amount with the second deletion even through both deletions are not the same. Based on the study of Y. Zhu *et al.* [8], accumulation of succinate was prevented by knockout of gene *frdA*. During the anaerobic respiration, menaquinol-fumarate

oxidoreductase (QFR) is used for succinate production. Since the production of succinate being prevented, the lactate production increase significantly.

For the obtained results, three of them are having the same Lactate production 18.0738 with different growth rate. All the obtained results have proved with the wet laboratory results that the predicted knockout list has increased the biochemical production in the industry. This also proved that the newly formed hybrid algorithm has good performance in identifying the gene knockout list.

Overall, the obtained results are consistent. This is because ABC algorithm has the advantages of simple, high robustness, fast convergence, high flexibility and fewer control parameters. Hence, it solved the multidimensional and multimodal optimization problems.

4 Conclusions

As a conclusion, our proposed hybrid algorithm showed a better performance than the previous gene knockout tools such as OptKnock and OptGene in term of the gene knockout identification for producing high yields of succinate and lactate in *E.coli*. ABC algorithm has the advantages of simple, high robustness, fast convergence, high flexibility and fewer control parameters. In the future work, another new data set was suggested to put in as to test the feasibility of this newly develop algorithm. Besides, other intelligent optimization algorithms like ants colony, particle swarm optimization (PSO) were encouraged to replace the artificial bee colony algorithm, so that by comparing these algorithms, a better algorithm will be found.

Acknowledgements. We would like to thank Malaysian Ministry of Science, Technology and Innovation for supporting this research by an e-science research grant (Grant number: 06-01-06-SF1029). This research is also supported by UTM GUP research grant that was sponsored by Universiti Teknologi Malaysia.

References

1. Hofvendahl, K., Hahn-Hagerdal, B.: Factors Affecting the Fermentative Lactic Acid Production from Renewable Resources. *Enzyme and Microbial Technology* 26(2-4), 87–107 (2000)
2. John, R.P., Nampoothiri, K.M., Pandey, A.: Production of L(+) Lactic Acid from Cassava Starch Hydrolyzate by Immobilized *Lactobacillus delbrueckii*. *J. Basic Microbiol.* 47(1), 25–30 (2007)
3. Chang, D.E., Jung, H.C., Rhee, J.S., Pan, J.G.: Homofermentative Production of D- or L-Lactate in Metabolically Engineered *Escherichia coli* RR1. *Appl. Environ. Microbiol.* 65(4), 1384–1389 (1999)
4. Karaboga, D., Basturk, B.: A Powerful and Efficient Algorithm for Numerical Function Optimization: Artificial Bee Colony (ABC) Algorithm. *Journal of Global Optimization* 39(3), 459–471 (2007)

5. Edwards, J.S., Palsson, B.O.: Metabolic Flux Balance Analysis and The In Silico Analysis of *Escheichia coli* K-12 Gene Deletions. *BMC Bioinformatics* 1, 1 (2000)
6. Hua, Q., Joyce, A.R., Fong, S.S., Palsson, B.O.: Metabolic Analysis of Adaptive Evolution for In Silico-Designed Lactate-Producing Strains. *Biotechnology and Bioengineering* 95(5), 992–1002 (2006)
7. Zhou, L., Zuo, Z.R., Chen, X.Z., Niu, D.D., Tian, K.M., Bernard, A.P., et al.: Evaluation of Genetic Manipulation Strategies on D-Lactate Production By *Escherichia Coli*. *Current Microbiology* 62(3), 981–989 (2010)
8. Zhu, Y., Eiteman, M.A., DeWitt, K., Altman, E.: Homolactate Fermentation by Metabolically Engineered *Escherichia coli* Strain. *Applied and Environmental Microbiology* 73(2), 456–464 (2006)

Author Index

- Chai, Lian En 5, 25, 51, 77, 101, 117, 127
Choi, Jinwook 37
Chong, Chuii Khim 5, 25, 51, 77, 101, 117, 127
Choon, Yee Wen 5, 25, 51, 77, 101, 117, 127

Deris, Safaai 5, 25, 51, 77, 101, 117, 127
Dhillon, Sarinder K. 15, 87

Hong, Susan Lim Lee 87

Illias, Rosli M. 25, 51, 77, 101, 117, 127

Kim, Jong-Beom 37
Kocher, Jean-Pierre A. 3

Lee, Seet Sun 127

Mohamad, Mohd Saberi 5, 25, 51, 77, 101, 117, 127
Musalib, Masarrah Abdul 77
Myaeng, Sung-Hyon 37

Nam, Sangsoo 37
Ng, Siew Teng 51

Oh, Heung-Seon 37

Paunoo, Baldeve 15

Saidin, Nor Farhah Binti 25
Sakharkar, Meena Kishore 1
Shamsir, Mohd Shahir 25, 51
Shuhaimi, Nur-Imtiazah 87
Sidhu, Amandeep S. 15, 87

Town, Christopher 63

Yin, Leang Huat 101