Jonathan H. Chan
Yew-Soon Ong
Sung-Bae Cho (Eds.)

Communications in Computer and Information Science 115

Computational Systems-Biology and Bioinformatics

First International Conference, CSBio 2010
Bangkok, Thailand, November 2010
Proceedings

Springer

Communications
in Computer and Information Science 115

Jonathan H. Chan Yew-Soon Ong
Sung-Bae Cho (Eds.)

Computational Systems-Biology and Bioinformatics

First International Conference, CSBio 2010
Bangkok, Thailand, November 3-5, 2010
Proceedings

Springer

Volume Editors

Jonathan H. Chan
King Mongkut's University of Technology Thonburi
School of Information Technology
126 Pracha U-Thit Rd, Bangmod, Thungkru, Bangkok 10140, Thailand
E-mail: jonathan@sit.kmutt.ac.th

Yew-Soon Ong
Nanyang Technological University, School of Computer Engineering
Block N4, 2b-39,Nanyang Avenue, Singapore 639798
E-mail: asysong@ntu.edu.sg

Sung-Bae Cho
Yonsei University, Dept. of Computer Science
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, South Korea
E-mail: sbcho@yonsei.ac.kr

Library of Congress Control Number: 2010937566

CR Subject Classification (1998): J.3, H.2.8, F.1, F.2.2, G.3

ISSN 1865-0929
ISBN-10 3-642-16749-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-16749-2 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 06/3180 5 4 3 2 1 0

Preface

This CCIS volume constitutes the proceedings of the First International Conference on Computational Systems Biology and Bioinformatics (CSBio 2010), held in Bangkok, Thailand, November 3–5, 2010. CSBio 2010 was a follow up to the successful Special Session on Computational Advances in Bioinformatics (CAB 2009) in the 16th International Conference on Neural Information Processing (ICONIP 2009). CSBio will provide an annual forum for international researchers to exchange the latest ideas on advances in the interdisciplinary fields of computational systems biology and bioinformatics. CSBio is proposed to be hosted alternately by King Mongkut's University of Technology Thonburi (KMUTT), Thailand, and Nanyang Technological University (NTU), Singapore. The School of Information Technology (SIT) at KMUTT was the proud host of CSBio 2010. This inaugural conference was launched to coincide with the 15th anniversary of SIT and the 50th anniversary of KMUTT.

CSBio 2010 accepted 19 regular session papers from a total of 48 submissions received on the EasyChair conference system. The authors of the submitted papers covered 16 countries worldwide and there were over 60 authors in the conference proceedings. The technical sessions were divided into five topical categories. Technical highlights included a keynote speech by Michael Brudno (Canada Research Chair in Computational Biology) and plenary talks by Nikhil R. Pal, Sung-Bae Cho, Yaochu Jin, and David W. Ussery. In addition, three tutorials by Kwoh Chee Keong, Stijn Meganck and Philip Shaw were included with CSBio 2010 registration. Furthermore, the 4th International Conference on Advances in Information Technology (IAIT 2010) was collocated with CSBio 2010.

We are indebted to the members of the CSBio 2010 International Advisory Board for their advice and assistance in the organization and promotion of the conference. We are thankful to the Program Committee and additional reviewers for their dedication and support in providing rigorous and timely reviews. Each paper was reviewed by at least three referees and even more reviews were provided in most of the cases.

A special thanks to the Publication Chair, Olarn Rojanapornpun, who worked tirelessly to produce the final proceedings. The organizing committee members would like to express our sincere appreciation to the devoted behind-the-scenes work by Paweena Mongkolpongsiri, Thanyapat Natwaratit, and Kanittha Charoensuk**.** Last but not least, the organizers gratefully acknowledge the contributions and support from all speakers and authors, as well as all other participants and contributors, in enabling this inaugural CSBio conference to have been a success.

November 2010

Jonathan H. Chan
Yew-Soon Ong
Sung-Bae Cho

Organization

Organizer

School of Information Technology (SIT), King Mongkut's University of Technology Thonburi (KMUTT), Bangkok, Thailand.

Technical Co-sponsors

International Neural Network Society (INNS)
IEEE Computational Intelligence Society (IEEE CIS)

Conference Committee

Honorary Chair — Sakarindr Bhumiratana, Thailand

International Advisory Board

Supapon Cheevadhanarak, Thailand
Sung-Bae Cho, South Korea
Igor Goryanin, UK
Yaochu Jin, UK
Nikola Kasabov, New Zealand
Tom Lenaerts, Belgium
Kwong-Sak Leung, Hong Kong
Jens B. Nielsen, Sweden
Kevin Painter, UK
Nikhil R. Pal, India
Kay Chen Tan, Singapore
Sissades Tongsima, Thailand
David W. Ussery, Denmark
Rene Vestergaard, Japan
Lipo Wang, Singapore
Richard Wintle, Canada
Stephen Wong, USA

Local Steering Committee — Borworn Papasratorn, Thailand
Prasert Kanthamanon, Thailand
Narumon Jeyashoke, Thailand
Marasri Ruengjitchatchawalya, Thailand

General Co-chairs — Jonathan H. Chan, Thailand
Yew-Soon Ong, Singapore

Program Chair — Sung-Bae Cho, South Korea

Local Organizing Chairs — Asawin Meechai, Thailand
Kriengkrai Porkaew, Thailand

Special Sessions Chairs	Kyung-Joong Kim, South Korea Stijn Meganck, Belgium
Tutorials Chair	Chee Keong Kwoh, Singapore
Competitions Chair	Ivor Tsang, Singapore
Publicity Chairs	Umaporn Supasitthimethee, Thailand Anju Verma, New Zealand
Publication Chair	Olarn Rojanapornpun, Thailand
Local Arrangement Chair	Wannida Soontreerutana, Thailand
Conference Secretariats	Saowalak Kalapanulak, Thailand Treenut Saithong, Thailand
Webmaster	Chompoonut Watcharinkorn, Thailand

Program Committee

Rafal Adamczak, Poland
Michael Brudno, Canada
Vladmir Brusic, USA
Nachol Chaiyarantana, Thailand
Sally Clift, UK
Jean-Paul Comet, France
Frank Eisenhaber, Singapore
Maoguo Gong, China
David Hardoon, Singapore
Yaochu Jin, UK
Saowalak Kalapanulak, Thailand
Kyung-Joong Kim, South Korea
Chee Keong Kwoh, Singapore
Gary Lee, Singapore
Tom Lenaerts, Belgium
Chee Peng Lim, Malaysia
Dudy Lim, Singapore
Meng-Hiot Lim, Singapore
Bjorn Lindman, Sweden
Jianjun Liu, Singapore
Ferrante Neri, Finland
Chakarida Nukoolkit, Thailand
Nikhil R. Pal, India
Shaoning Pang, New Zealand
Somnuk Phon-Amnuaisuk, Malaysia
Sirirat Pinsuwan, Thailand
Santitham Prom-on, Thailand
Partha Roy, Singapore
Treenut Saithong, Thailand
Jittisak Senachak, Thailand
Kok Yong Seng, Singapore
Kay Chen Tan, Singapore
Ee Chon Teo, Singapore
Swee-Hin Teoh, Singapore
Chuan-Kang Ting, Taiwan
Ivor Tsang, Singapore
David Ussery, Denmark
Anju Verma, New Zealand
Dianhui Wang, Australia
Lipo Wang, Singapore
Bunthit Watanapa, Thailand
David Weiss, Belgium
Stephen Wong, USA
Zexuan Zhu, China

Additional Reviewers

Atthawut Chanthaphan, Edward Chuah, Yi Li, Pornchai Mongkolnam, Pauline Ng, Saowalak Watanapa

Local Sponsors

IEEE Thailand Section
Thailand Chapter of ACM
Software Park Thailand
Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology Association of Thailand (ECTI)
Ministry of Information and Communication Technology (MICT)
National Center for Genetic Engineering and Biotechnology (BIOTEC)
Yip In Tsoi

Table of Contents

Session 1. Modeling and Simulation of Biological Processes

Session 2. Gene Expression Analysis

Session 3. Biological Sequence Analysis and Network Reconstruction

Session 4. Bio-data Visualization and Biological Databases

Session 5. Medical and Biomedical Informatics

A Formal Model for Gene Regulatory Networks with Time Delays

Jean-Paul Comet[1], Jonathan Fromentin[2], Gilles Bernot[1], and Olivier Roux[3]

[1] Laboratoire I3S, UMR 6070 UNS-CNRS, Université de Nice
2000, route des Lucioles, B.P. 121, 06903 Sophia Antipolis CEDEX, France
{bernot,comet}@unice.fr
[2] Labri, Université de Bordeaux, 33 Talence, France
jonathan.fromentin@labri.fr
[3] IRCCyN UMR 6597, CNRS & École Centrale de Nantes
1, rue de la Noë - BP 92 101 - 44321 Nantes CEDEX 03, France
olivier.roux@irccyn.ec-nantes.fr

Abstract. We introduce a hybrid modelling framework for gene regulatory networks as an extension of the René Thomas' discrete modelling framework. We handle temporal aspects through *delays* expressing the time mandatory to pass from a qualitative state to another one. It permits one to build, from a specification expressed in terms of paths, the constraints on the temporal parameters in order to assure the consistency between the hybrid model and the specification.

We illustrate this modelling framework on the simple system of *mucus* production in the bacterium *Pseudomonas aeruginosa*. We show through this example how to build the constraints on the delays parameters for the specification of a cycle in the dynamics.

1 Introduction

Modelling gene regulatory networks aims at deep understanding of their behaviours and thus at some non-obvious predictions [1,2,3,4]. Unfortunately, while available data on the interaction graph between genes are more and more numerous, the kinetic data allowing us to identify the sensible parameters are difficult to obtain experimentally. This parameter identification problem constitutes the cornerstone of the modelling activities. Whereas the quantitive models (differential equations, stochastic models) need a good precision on the available information about the dynamics of the system, qualitative models which focus only on the qualitative features of the dynamics, make easier the parameter identification problem. This comment motivates the development of different methods for which this identification problem is tractable [5,6,7]. For example René Thomas' discrete modelling [8] of gene regulatory networks (GRN) is a well-known approach to study the dynamics resulting from a set of interacting genes. It deals with some *discrete* parameters that reflect the possible targets of trajectories. Those parameters are *a priori* unknown, but they can generally

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 1–13, 2010.

be deduced from a well-chosen set of biologically observed trajectories. Moreover there exists a strong correspondence between modelling by piecewise linear differential equations and such boolean or discrete modellings [9].

Unfortunately this framework neglects the time delay necessary for a gene to pass from one level of expression to another one, whereas information on the time necessary for the system to go from one state to another one is often experimentally available. For example, time spent by the system to cover a whole turn of a periodic trajectory (*e.g.* circadian cycle) is often well known. Such kind of information is not used to face up the parameter identification problem in the "standard" Thomas' framework without delays. This remark motivated several researchers to develop mathematical frameworks [10] or formal frameworks [11,12,13,14,15] where time is explicit. The effect of time delays on the robustness of differential systems becomes also an interesting research perspective [16].

In this article, we propose a new modelling framework which extends the discrete modelling framework of René Thomas by introducing temporal aspects. This modelling framework inherits from the pure qualitative modelling framework the computer aided methods for determination of suitable parameter values, but it introduces a continuous notion of time through the handling of delays. Thus, we propose a hybrid modelling framework where discrete and (temporal) continuous dynamics are mixed. Naturally, these delays are coded by new parameters, *i.e.* delays mandatory for a gene to go from a discrete abstract level to another one, which are not deductible from previous qualitative models. When this framework is viewed as an abstraction of piecewise linear differential equations, as discrete modelling is, some constraints on the delays of the hybrid model can be built to ensure the consistency between the hybrid model and the underlying system of piecewise linear differential equations. In particular, delays of the hybrid models have to satisfy some constraints which can be deduced from the piecewise linear differential equation systems. Nevertheless, kinetic parameters of the PLDE system are generally unknown and the key point of the modelling process lies in identification of these kinetic parameters. Adding delays, the identification problem is more difficult because of the increased number of parameters. Nonetheless much temporal data is available from experiments and because hybrid modelling frameworks preserve powerful computer-aided reasoning capabilities, computer is able to reject a large class of parameter values. To illustrate our hybrid modelling framework, we use as *running example*, an extremely simplified model, representing the production of *mucus* of the bacterium *Pseudomonas aeruginosa* [17,18]. *P. aeruginosa* is an opportunistic pathogen, often encountered in chronic lung diseases such as cystic fibrosis. The main regulator for the mucus production, AlgU, supervises an operon which is made of 4 genes among which one codes for a protein that is an inhibitor of AlgU. Moreover AlgU favours its own synthesis. The mucus production regulatory network can then be simplified into a regulatory graph with two nodes: x represents AlgU, and y its inhibitor [17]. x regulates positively y and also regulates itself, whereas y regulates negatively x. From a biological point of view, it is crucial to determine if the change of behaviours (passing from a state where mucus is not

produced to another one where it is) is mostly due to change of the regulations (mutation) or mostly due to a change of state. We show in this article that it is possible to construct a hybrid model in which the behaviour which does not produce mucus is represented by a limit cycle.

The paper is organized as follows. We first recall in section 2 the principles of the modelling by a system of piecewise linear differential equations and of the discrete modelling of René Thomas. Section 3 is devoted to the definition of the considered hybrid models. In section 4, we sketch how to build a set of constraints on the delays parameters in order to get a hybrid model whose dynamics present a particular path. Finally section 5 is devoted to concluding remarks.

2 Continuous and Discrete Models

PLDE modelling. Modelling a gene regulatory network with a system of piecewise linear differential equations [19], PLDE for short, makes mandatory the knowledge of regulations. In particular, for each regulation, which can involve several regulators, one has to define under which real concentration conditions this regulation is effective. As usual, because regulations are often considered as sigmoidal, we consider only the piecewise differential system, which is built as an approximation of the differential system by replacing sigmoids by steps functions: $s^+_\theta(x) = \begin{cases} 1, & x > \theta \\ 0, & x < \theta \end{cases}$ and $s^-_\theta(x) = 1 - s^+_\theta(x)$ where $\theta \in \mathbb{R}^+$ is the threshold of the sigmoid.

Definition 1 (PLDE). *Let us consider a finite set of positive real variables* $X = \{x_1, x_2, \ldots, x_n\}$ *and let us denote* x *the vector* $(x_1, x_2, \ldots, x_n)$*. A system of piecewise linear differential equations (PLDE) on* X *is defined by:*

$$\dot{x}_i = g_i(x) - \gamma_i x_i \quad with \quad 0 \leq x_i \quad and \quad 1 \leq i \leq n$$

where γ_i *is the degradation rate of variable* x_i *and each* g_i *is a function representing the synthesis rate of variable* x_i *which is supposed to be additive (the synthesis rate is the sum of all effective regulations):*

$$g_i(x) = k_i + \sum_{j \in \mathcal{R}(i)} k_{ij} r_{ij}(x) \tag{1}$$

where

- $k_i \in \mathbb{R}^+$ *and* $k_{ij} \in \mathbb{R}^{+*}$ *are kinetic parameters,*
- *The regulation functions* r_{ij} *are some combinations of step functions:*

$$< r > ::= s^+_\theta | s^-_\theta | 1 - < r > | < r > \times < r >$$

- $\mathcal{R}(i)$ *is the set of possible indices such that* r_{ij} *is a regulation function on* i*.*

The dynamics of a PLDE system is intrinsically related to kinetic parameters. In the rest of the paper, kinetic parameters are indexed by a *set of resources*. Intuitively, the set of resources at a given continuous state is the set of the regulations which are effective at this continuous state.

Definition 2 (Resources). *The set of resources of variable x_i at continuous state x, denoted $\Omega_i(x)$, is the finite set $\Omega_i(x) = \{j \mid r_{ij}(x) = 1\}$.*

Because of the finite number of possible sets of resources, the concentration space of each variable x_i can be partitioned in equivalence classes defined by the same set of resources of variable x_i: x_i^1 and x_i^2 are in the same equivalence class iff $\Omega_i(x_i^1) = \Omega_i(x_i^2)$. These equivalence classes split the concentration space of x_i into open intervals which can be classicaly numbered by 0, 1... : 0 is the *name* of the first interval, 1 denotes the second interval and so on.

We extend this equivalence relation to the concentration space of n dimensions. The principle of the partition is simple: we gather in the same *domain* all the continuous states for which each concentration coordinate is in the same interval. Because of the form of the regulation functions, all domains (*i.e.* equivalence classes) are hyper-rectangular zones. Moreover, since all the continuous states of the same domain are identically situated with regard to the thresholds, they all have the same set of resources: $\forall x \in d$, $\Omega_i(x)$ =constant. So we can define the set of resources of a domain:

Definition 3 (Resources of a domain). *The set of resources of variable x_i in domain d, denoted $\omega_i(d)$ is the set of resources (see Def. 2) of variable x_i at any point x of d: $\omega_i(d) = \{j \mid \forall x \in d, r_{ij}(x) = 1\}$.*

Finally, to simulate a PLDE system, values of kinetic parameters (k_i and k_{ij} in eq. (1)) have to be given. Unfortunately, these parameters are not easy to evaluate *in vivo*, and values obtained *in vitro* are not necessarily transposable for the system *in vivo*. Valuating parameters thus becomes the cornerstone of the modelling process.

Discrete modelling. To overcome these difficulties of parameters valuation, René Thomas first introduced a boolean framework [7] then a discrete formalism [8] which have been proven to be consistent with the PLDE modelling framework [9]. In this section, we sketch this qualitative framework which mimics qualitativelly the continuous framework.

From a qualitative point of view, at a particular point of the concentration space, the dynamics is controlled only by the set of the regulations which are resources. Actually René Thomas did not propose such a rich way to describe the regulations but this discrete modelling framework can be easely extended. Let us first notice, that the regulations do not change inside a same domain class, that is, the differential equation system is linear in each hyper-rectangular zone which define the domains. Then the solutions in each zone are analytically deducible and converge towards a unique *focal point*. Then

- with each domain is associated a qualitative state,
- the coordinate i of the focal point associated to the domain d is given by $((k_i + \sum_{j \in \mathcal{R}(i)} k_{ij} r_{ij}(x))/\lambda_i)_{i \in V}$ for any $x \in d$,
- because of the monotonicity of the solutions of the differential equations, trajectories starting in the domain d go towards the associated focal point until they reach the boundary of d.

– From a qualitative point of view, only the position of the focal point is important. Then, for the domain d we call $K_{i,\omega_i(d)}$ the number of the interval in which stays the coordinate i of the focal point which depends only on the set of resources $\omega_i(d)$ of variable x_i in domain d.

This idea leads to the definition of the discrete transition system.

Definition 4 (Transition system). *The discrete dynamics of a gene regulatory network with n variables is given by the* transition system *defined by:*

– *the set of vertices is the set of equivalence classes of the concentration space, called a discrete states; each equivalence class is represented by a vector of integer $d = (d_i)_{i\in[1,n]}$ where d_i is the number of the interval in which stays the coordinate i of a particular point of the equivalence class,*
– *There exists a transition from the discrete state d to the discrete state d' if*
  • $\exists i \in [1,n]$ *such that* $\begin{cases} d'_i = d_i + 1 & \text{and} \quad K_{i,\omega_i(d)} > d_i \\ d'_i = d_i - 1 & \text{and} \quad K_{i,\omega_i(d)} < d_i \end{cases}$
  • $\forall j \neq i, d'_j = d_j$.

A strategy for determining discrete parameters. Let us observe that the number of different parameters in the discrete modelling framework is finite and that each parameter can take a finite number of values. Thus, by enumaration, all the possible models can be simulated in order to keep only the valuations of parameters leading to a transition system which is consistent with all the available specifications on the behaviour of the biological system. Generally, known behavioural properties can be expressed by a particular qualitative observation of the following class: the saturation of the cell in a particular gene product (resp. the knock-out of a gene) leads to a state where an other specific gene product is present or absent.

This computer aided modelling approach has already been implemented using classical model-checking techniques [18] or symbolic model-checking techniques [20], and then using constraint programming techniques [21]. The observation data are transcripted into temporal logic formulas, a formal representation of a knowledge about the traces of a system which can be handled by computers. In [18], for each possible valuation, the transition system is computed and a procedure of model-checking is performed. This allows one to retain only the valuations that lead to a transition system satisfying the formula. This approach, requiring enumeration of all parameter valuations, has been rephrased for a temporal logic so that a single pass of model-checking gives a symbolic representation of all the models validating the temporal property [20]. The approaches adopted in [21,22,23] use constraints programming. The temporal logic formula is translated into constraints on the discrete parameters of the model. These constraints also symbolically represent all the parameter valuations that lead to transition systems satisfying the formula.

3 Hybrid Modelling

Because real time is ignored in the complete discrete modelling framework, some qualitative behaviours are not distinguishable. For example, an inward spiral is abstracted by the same discrete model than a outward spiral. This remark motivated us to introduce a modelling framework that combines the discrete modelling framework with temporal delays while preserving consistency with PLDE systems.

Syntactical features of hybrid models. We associate with each domain a *temporal zone* which measures the time elapsed in the domain. This zone is represented as a n-dimensional hypercube (where n is the number of variables in the system) whose edges have various lengths. Intuitively, the length of the hypercube in the i-axis represents the mandatory delay for the system to entirely cross the associated domain (in concentration) along the i-axis.

Definition 5 (State graph with delays (SGD)). *Let $G = (g_i(x_i))_{x_i \in X}$ be the regulation schema, which defines the synthesis rate of each variable according to the effectiveness of each regulation (see equation 1). A* State Graph with Delays *(SGD for short) associated with the regulation schema G is a 4-tuple $\mathcal{N} = (X, L, K, D)$ where:*

- *$X = \{x_1, \ldots, x_n\}$ is the set of variables,*
- *$L = \{(l_i(x_i))_{x_i \in X}\}$ is the finite set of domains deduced from G by the equivalence relation on the concentration space; for each $x \in X$, we define the integer b_x as the number of different thresholds describing the different actions of x on its targets,*
- *$K = \{K_{x,\omega}\}_{x \in X, \omega \subset \mathcal{R}(x)}$ is a family of integers such that $K_{x,\omega} \in [0, b_x]$ for any variable x and for any set ω of regulations on x.*
- *$D = D^+ \cup D^-$ is a family of positive real numbers such that:*
  - *$D^+ = \{\delta^+_{x,i,\omega}\}_{x \in X, i \in [0,b_x], \omega \subset \mathcal{R}(x), i \leq K_{x,\omega}}$ with $\delta^+_{x,i,\omega} \in \mathbb{R}^+$ ($[0, b_x]$ being an interval of integers).*
  - *$D^- = \{\delta^-_{x,i,\omega}\}_{x \in X, i \in [0,b_x], \omega \subset \mathcal{R}(x), i \geq K_{x,\omega}}$ with $\delta^-_{x,i,\omega} \in \mathbb{R}^+$ ($[0, b_x]$ being an interval of integers).*

  The subfamily D^+ is called the set of production delays of $\mathcal{N}$ and the subfamily D^- is called the set of degradation delays of $\mathcal{N}$.

Intuitively, for a given domain d, the temporal zone is defined by the product of intervals: $\prod_{x \in X} [0, \delta^+_{x,l(x),\omega_x(d)} + \delta^-_{x,l(x),\omega_x(d)}]$

Running example. Let us now consider the SGD $\mathcal{P} = (X, L, K, D)$ modelling the system of mucus production of *Pseudomonas Aeruginosa*, defined by:

- $X = \{x, y\}$,
- $L = \{(0,0), (1,0), (0,1), (1,1), (2,0), (2,1)\}$,
- $K = \{K_{x,\emptyset} = 0, K_{x,\{x\}} = 2, K_{x,\{y\}} = 2, K_{x,\{x,y\}} = 2, K_{y,\emptyset} = 0, K_{y,\{x\}} = 1\}$
- $D = D^+ \cup D^-$ where
  - $D^+ = \{\delta^+_{x,2,\{x\}}, \delta^+_{x,2,\{x,y\}}, \delta^+_{x,1,\{y\}}, \delta^+_{x,0,\{y\}}, \delta^+_{y,1,\{x\}}, \delta^+_{y,0,\{x\}}\}$ and
  - $D^- = \{\delta^-_{x,2,\{x\}}, \delta^-_{x,2,\{x,y\}}, \delta^-_{x,1,\emptyset}, \delta^-_{x,0,\emptyset}, \delta^-_{y,1,\emptyset}, \delta^-_{y,0,\emptyset}, \delta^-_{y,1,\{x\}}\}$.

This *state graph with delays* is represented in Figure 1.

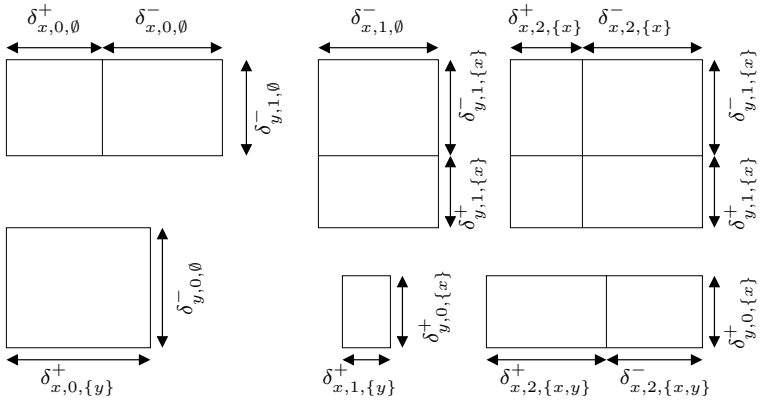


Fig. 1. State graph with delays modelling the system of the mucus production by *Pseudomonas aeruginosa*. Only non-zero delays are drawn.

Semantics of hybrid models: the dynamics. Let us observe that to specify a particular state of a SGD, one needs a couple of values: the first value is a domain, and the second is a point in the associated temporal zone. More formally, for a given SGD $\mathcal{N} = (X, L, K, D)$, a *state* of $\mathcal{N}$ is a couple $\eta = (l, \tau)$ where:

- $l : X \to \mathbb{N}$ is a domain of $\mathcal{N}$ (i.e. $l \in L$).
- $\tau : X \to \mathbb{R}^+$ is a total function s.t. $\forall v \in X, \tau(v) \leq \delta^+_{v,l(v),\omega_v(l)} + \delta^-_{v,l(v),\omega_v(l)}$
  The real number $\tau(v)$ is called the *delay residue* of v at the level $l(v)$.

As we already mentioned, temporal zones allow one to measure the time elapsed in a domain. Intuitivelly, the evolution in the model is twofold:

- inside a domain, the point in the temporal zone evolves in a *linear* way, it measures the time spent in a domain along a given evolution direction.
- to pass from a domain l to another one, it is mandatory that the point in the temporal zone reaches a border. If the point reaches the face for which the delay residue $\tau(v)$ is null (resp. equal to $\delta^+_{v,l(v),\omega_v(l)} + \delta^-_{v,l(v),\omega_v(l)}$), the system leaves the previous domain and enters into the new domain where the concentration level $l(v)$ is decremented (resp. incremented). The face of the temporal zone that is reached defines the new (accessible) domain.

To go further in the formalization of these ideas, we introduce two kinds of delays. The first one is the mandatory time for a variable to allow the system to move from a domain to another one: it is the *moving delay*, see figure 2. Unfortunatelly, this definition is not sufficient to determine if the reached face allows the exit from the domain. Thus, the *cross delay* is introduced.

Definition 6 (moving & cross delays). *Let $\eta = (l, \tau)$ be a state of a* SGD *$\mathcal{N}$,*

- *the* moving delay *of a variable v is given by the function $\mu_\eta : X \to \mathbb{R}^+ \cup \{\infty\}$ defined by*
$$\mu_\eta(v) = \begin{cases} \infty & \text{if } K_{v,\omega_v(l)} = l(v) \\ |\delta^+_{v,l(v),\omega_v(l)} - \tau(v)| & \text{if } K_{v,\omega_v(l)} \neq l(v) \end{cases}.$$

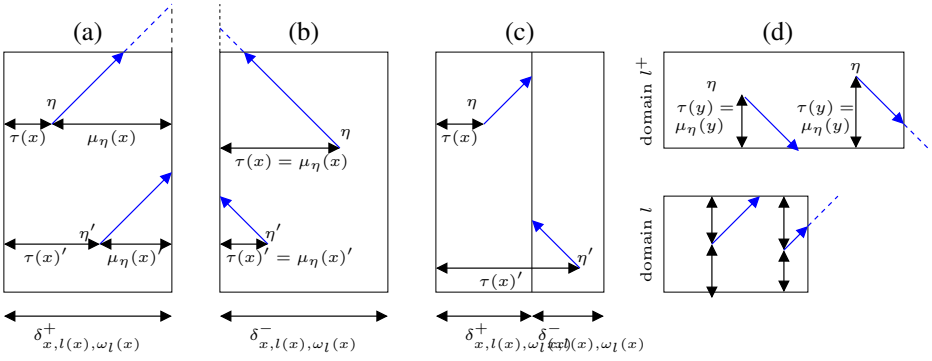


Fig. 2. Moving and cross delays. **(a)** When $l(x) < K_{x,\omega_l(x)}$ the moving delay is $\delta^+_{x,l(x),\omega_l(x)} - \tau(x)$ in the x-direction (horizontal). **(b)** When $l(x) > K_{x,\omega_l(x)}$ the moving delay is $\tau(x)$ in the x-direction (horizontal). **(c)** When $l(x) = K_{x,\omega_l(x)}$, x cannot be responsible for the exit from the domain. Thus the moving delay is $\mu_\eta(x) = \infty$. **(d)** Illustration of cross delays when $\mu_\eta(y) \neq \infty$ and $\bar{\mu}_\eta(y) = \infty$ (y-axis is the vertical one).

– *the* cross delay *of a variable v is given by the function $\bar{\mu}_\eta : X \to \mathbb{R}^+ \cup \{\infty\}$ defined by:*
  - *If ($K_{v,\omega_v(l)} < l(v)$ and $K_{v,\omega_v(l^-)} > l(v) - 1$) or ($K_{v,\omega_v(l)} > l(v)$ and $K_{v,\omega_v(l^+)} < l(v) + 1$) then $\bar{\mu}_\eta(v) = \infty$,*
  - *else $\bar{\mu}_\eta(v) = \mu_\eta(v)$.*

  where domains l^+ and l^- are such that $\forall u \neq v$, $l^+(u) = l^-(u) = l(u)$ and $l^+(v) - 1 = l^-(v) + 1 = l(v)$.

The moving delay of variable v is simply the time necessary for this variable to allow the system to exit from the current domain. If variable v is not able to reach the boundary of the temporal zone, the moving delay of variable v is ∞, see Fig. 2-(c). When $\mu(v) \neq \infty$, the cross delay of variable v can nevertheless be equal to ∞ when v is attracted outside the current domain, but cannot exit in that direction since, beyond the limit of the domain, this variable is immediately attracted again inside the domain. This stands for the notion of sliding modes [19]. Illustration of such a situation is given in Fig. 2-(d).

The temporal evolutions from a state within the temporal zone are linear: directions of these evolutions are given by the following definition.

Definition 7 (Discrete partial derivative). *Given a domain l of a* SGD *$\mathcal{N}$, for any state $\eta = (l, \tau)$ and for any variable v, the* discrete partial derivative *of $\mathcal{N}$ at l with respect to v, $\kappa_l(v)$, is defined by:*

– *if $l(v) < K_{v,\omega_l(v)}$ and $\bar{\mu}_\eta(v) \neq \infty$ then $\kappa_l(v) = 1$*
– *if $\bar{\mu}_\eta(v) = \infty$ then $\kappa_l(v) = 0$*
– *if $l(v) > K_{v,\omega_l(v)}$ and $\bar{\mu}_\eta(v) \neq \infty$ then $\kappa_l(v) = -1$*

We can now define the successor states of a state (see Fig 3) using the function *sign*: $sign(x) = 1$ if $x > 0$, $sign(x) = -1$ if $x < 0$ and $sign(x) = 0$ if $x = 0$.

Definition 8 (Successor). *A state $\eta' = (l', \tau')$ of a* SGD $\mathcal{N}$ *is a* successor *state of the state $\eta = (l, \tau)$ if there exists a variable $x \in X$ such that:*

1. $\forall y \in X,\ \bar{\mu}_\eta(x) \leq \bar{\mu}_\eta(y)$,
2. $l'(x) = l(x) + \kappa_l(x)$,
3. $\forall y \in X,\ y \neq x \Rightarrow l'(y) = l(y)$,
4. $\kappa_l(x) = 1 \Rightarrow \tau'(x) = 0$,
5. $\kappa_l(x) = -1 \Rightarrow \tau'(x) = \delta^+_{x,l'(x),\omega_x(l')} + \delta^-_{x,l'(x),\omega_x(l')}$,
6. $\forall y \in X$ *such that* $y \neq x$ *and* $\kappa_l(y) \neq 0$,
$$\kappa_l(x) \neq 0 \Rightarrow \tau'(y) = \frac{\left(\tau(y) + \mathrm{sign}(\delta^+_{y,l(y),\omega_y(l)} - \tau(y)) \times \mu_\eta(x)\right) \times \left(\delta^+_{y,l'(y),\omega_y(l')} + \delta^-_{y,l'(y),\omega_y(l')}\right)}{\delta^+_{y,l(y),\omega_y(l)} + \delta^-_{y,l(y),\omega_y(l)}},$$
7. $\forall y \in X$ *such that* $y \neq x$ *and* $\kappa_l(y) = 0$,
$$\kappa_l(x) \neq 0 \Rightarrow \tau'(y) = \frac{\left(\tau(y) + \mathrm{sign}(\delta^+_{y,l(y),\omega_y(l)} - \tau(y)) \times \min(\mu_\eta(x), \mu_\eta(y))\right) \times \left(\delta^+_{y,l'(y),\omega_y(l')} + \delta^-_{y,l'(y),\omega_y(l')}\right)}{\delta^+_{y,l(y),\omega_y(l)} + \delta^-_{y,l(y),\omega_y(l)}},$$
8. $\kappa_l(x) = 0 \Rightarrow \left(\forall y \in X,\ \tau'(y) = \delta^+_{y,l(y),\omega_y(l)}\right)$.

If $\kappa_l(x) \neq 0$, then the transition time *from η to η' is $\zeta(\eta, \eta') = \mu_\eta(x)$. If $\kappa_l(x) = 0$, then $\zeta(\eta, \eta')$ is equal to* $\min_{v \in X}(\mu_\eta(v))$.

Note that in item 6 of the previous definition, the computation of new delay residue for variable y depends on the sign of $(\delta^+_{y,l(y),\omega_l(y)} - \tau(y))$. Indeed, if $\delta^+_{y,l(y),\omega_l(y)} < \tau(y)$ (resp. $\delta^+_{y,l(y),\omega_l(y)} > \tau(y)$), the coordinate $\tau(y)$ decreases (resp. increases) towards $\delta^+_{y,l(y),\omega_l(y)}$.
The previous definition covers both of the following cases.

1. Let us first focus on the case where a domain contains its focal point (see Fig. 3-b). Temporal trajectories do not go out of this domain: all the cross delays are equal to ∞, and each discrete partial derivative is null. Thus, we can take for x any element of X (see item 1). Items 2 and 3 imply that $l' = l$. Finally item 8 gives the temporal coordinates of the focal point. Transition time is then the time necessary for each variable y to reach the coordinate f_y of the focal point.
2. We now focus on a domain which does not contain its focal point (see Fig. 3-a). Each temporal trajectory goes out of this domain passing a threshold on one v-axis. This variable v is the one which has the smallest *not-infinite* cross delay (see item 1). Items 2 and 3 imply that l' differs from l on only one coordinate. Items 4 and 5 reset *residue delay* associated with v whereas items 6 and 7 compute the new *residue delays* associated with the other variables (these expressions come from the homothetic transformation). The transition time is then the time to reach the face of the temporal zone, that is the moving delay.

Definition 9 (State space). *The* state space *of a* RND $\mathcal{N}$ *is the (infinite) directed "graph" $S_\mathcal{N}$ the vertices of which are the states of $\mathcal{N}$ and the edges*

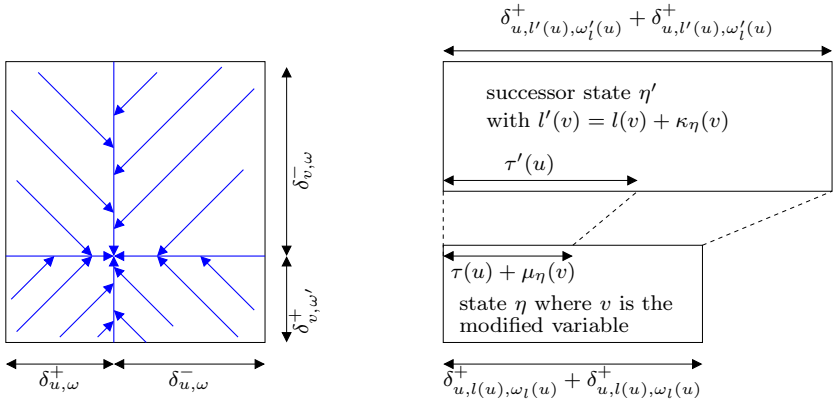


Fig. 3. Illustration of Definition 8. **(a)** the domain contains its focal point and all the cross delays are infinite. **(b)** the domain does not contain its focal point.

of which are the couples (η, η') *such that* η' *is a successor of* η. *Given a path* $p = \eta_0\eta_2\cdots\eta_n$, *the* crossing time *of* p *is defined as* $\tau(p) = \sum_{i=1}^{n} \tau(\eta_{i-1}, \eta_i)$.

Transitions between domains are the transitions of the discrete dynamics in the formalism of René Thomas.

4 Construction of the Delays Constraints

The parameter values of the hybrid model can be straightforwardly deduced from a PLDE system with known parameters. For the discrete part, the parameters correspond to the position of the steady states of the linear differential system of the considered domain, whereas the parameters of family D correspond to the time mandatory for the differential system to cross the domain. Let us just mention that when the v-coordinate of the focal point f of the domain l is inside $l(v)$ then, neither the production delay nor the degradation delay is null: $\delta^+_{v,k,\omega(l)}$ (resp. $\delta^-_{v,k,\omega(l)}$) measures the duration between the time when a trajectory gets into the domain by the face having the smaller (resp. the bigger) v-concentration value and the time when the coordinate v of the focal point f is reached. Whereas from the point of view of PLDE, this time is infinite, for the hybrid model this time is not infinite.

But in general, when modelling a biological regulatory network, we have only a partial knowledge about the form of the regulatory functions (r_{ij}). Specifically, kinetic parameters of the PLDE system are unknown and the key point of the modelling process thus lies in identification of these kinetic parameters. Paragraph about *strategies for determining discrete parameters* of section 2 sketches, in the context of purely discrete modelling, a computer aided method for helping in this task. In our context of hybrid modelling, even if the qualitative parameters $(K_{x,\omega})$ are assumed to be known (or deduced from a computer aided approach), it remains to determine which values of the time delays are actually consistent with known properties of the studied system.

Once again, we start from some knowledge about the dynamics of the studied biological system. This knowledge often comes from experimental observations which are expressed as paths in the discrete transition system. These paths constitute the *specifications* since it determines the set of models which have to be considered. This section sketches how these specifications build up the models, and more accurately a system of parameter constraints.

The principle of the construction of these constraints relies on the enumeration of constraints due to paths of length 2: $\mu_0 \to \mu_1 \to \mu_2$. For a longer path, the constraint is the conjunction of constraints due to each sub-path of length 2.

For sake of readability, we describe here only one situation among twelve[1]. Let us consider the path $\mu_0 \to \mu_1 \to \mu_2$ where the first (resp. second) transition is due to a qualitative increasing of variable i_0 (resp. i_1). Let us suppose moreover that the vector $(c_i)_{i \in V}$ represents the delays residue when entering into μ_1 and that there exists in μ_1 a variable i'_1 which can also increase. In order to allow the global path $\mu_0 \to \mu_1 \to \mu_2$, the following relation has to be satisfied:

$$(d^+_{i_1}(\mu_1) - c_{i_1}) < (d^+_{i'_1}(\mu_1) - c_{i'_1})$$

Processing discrete cycle. The discrete cycles can abstract several different behaviours: fully cyclic temporal trajectories, convergent spirals, divergent spirals, limit cycle, etc. Thus, it is interesting to know more precisely their behaviours in the hybrid modelling. For example, it can be proved that the discrete cycle of *Pseudomonas aeruginosa* – $(0,0) \to (1,0) \to (1,1) \to (0,1) \to (0,0)$ – can abstract different kinds of qualitative behaviours of hybrid models. In other words, from the same purely discrete model with a discrete cycle, it is possible to construct a hybrid model which presents either: (1) a set of convergent spirals or (2) a set of cyclic temporal trajectories which constitute a torus and that we call fully cyclic temporal trajectories or (3) a set of divergent spirals or (4) a limit cycle, that is, a torus of volume null (see Fig. 4).

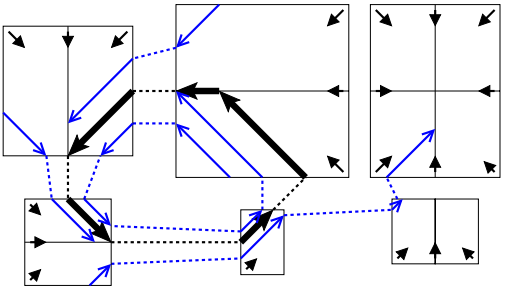


Fig. 4. A particular hybrid dynamics with a limit cycle (in thick black line) modelling the system of the mucus production by Pseudomonas aeruginosa

[1] The other cases are addressed in a similar enough way and the proof can be sent upon request.

5 Conclusion

We developed a new hybrid modelling framework for gene regulatory networks which extends the discrete modelling framework of René Thomas by introducing temporal features through delays handling. These delays express the time mandatory to pass from a qualitative state to another.

On the one hand, this modelling framework inherits from the differential modelling framework, since it is possible to build an hybrid model consistent with the underlying system of piecewise linear differential equations (PLDE). On the other hand, this modelling framework inherits also from the pure qualitative modelling framework, the computer aided methods for determination of suitable parameter values, but it introduces a continuous notion of time through delays handling. When kinetic parameters are not available, it is possible to build some constraints on the new delays parameters in order to get a model satisfying a specification expressed in terms of paths. Finally, adding information about delays in the qualitative framework allows one to distinguish qualitatively different behaviours which are abstracted into a common purely discrete model.

With such hybrid frameworks, systems biology should take advantage of the whole corpus of formal methods from computer science which opens a large horizon of research perspectives. It will be necessary to develop for example algorithms that compute the set of parameter valuations that are compatible with reachability properties. Indeed, hybrid modellings are not the ultimate aim, they are only a guideline for predictions that suggest biological experiments, whose success will be *in fine* the discriminent criterion. In such a perspective, hybrid approaches could constitute a trade-off between expressiveness and computational tractability.

References

1. Ideker, T., Galitski, T., Hood, L.: A new approach to decoding life: systems biology. Annual Rev. Genomics Hum. Genet. 2, 343–372 (2001)
2. Oltvai, Z., Barabási, A.: Systems biology. Life's complexity pyramid. Science 298(5594), 763–764 (2002)
3. Kitano, H.: Computational systems biology. Nature 420(6912), 206–210 (2002)
4. Conti, F., Valerio, M., Zbilut, J., Giuliani, A.: Will systems biology offer new holistic paradigms to life sciences? Syst. Synth. Biol. 1(4), 161–165 (2007)
5. Rashevsky, N.: Mathematical Biophysics: Physico-Mathematical Foundations of Biology. University of Chicago Press, Chicago (1948)
6. Sugita, M.: Functional analysis of chemical systems in vivo using a logical circuit equivalent. Journal of Theoretical Biology 1, 415–430 (1961)
7. Thomas, R.: Boolean formalization of genetic control circuits. Journal of Theoretical Biology 42, 563–585 (1973)
8. Thomas, R.: Regulatory networks seen as asynchronous automata: A logical description. Journal of Theoretical Biology 153, 1–23 (1991)
9. Snoussi, E.: Qualitative dynamics of a piecewise-linear differential equations: a discrete mapping approach. Dynamics and stability of Systems 4, 189–207 (1989)

10. Farcot, E., Gouzé, J.L.: Limit cycles in piecewise-affine gene network models with multiple interaction loops. International Journal of Systems Science 41(1), 119–130 (2010)
11. Siebert, H., Bockmayr, A.: Incorporating time delays into the logical analysis of gene regulatory networks. In: Priami, C. (ed.) CMSB 2006. LNCS (LNBI), vol. 4210, pp. 169–183. Springer, Heidelberg (2006)
12. Ahmad, J., Bernot, G., Comet, J.P., Lime, D., Roux, O.: Hybrid modelling and dynamical analysis of gene regulatory networks with delays. ComPlexUs 3(4), 231–251 (2007)
13. Batt, G., Ben Salah, R., Maler, O.: On timed models of gene networks. In: Raskin, J.-F., Thiagarajan, P.S. (eds.) FORMATS 2007. LNCS, vol. 4763, pp. 38–52. Springer, Heidelberg (2007)
14. Maler, O., Pnueli, A.: Timing analysis of asynchronous circuits using timed automata. In: Camurati, P.E., Eveking, H. (eds.) CHARME 1995. LNCS, vol. 987, pp. 189–205. Springer, Heidelberg (1995)
15. Comet, J.P., Bernot, G.: Introducing continuous time in discrete models of gene regulatory networks. In: Proc. of the Nice Spring school on Modelling and simulation of biological processes in the context of genomics. EDP Sciences, pp. 61–94 (2010) ISBN: 978-2-7598-0545-7
16. Radde, N.: The impact of time-delays on the robustness of biological oscillators and the effect of bifurcations on the inverse problem. Eurasip J. Bioinf. Syst. Biol. (2009)
17. Guespin-Michel, J., Kaufman, M.: Positive feedback circuits and adaptive regulations in bacteria. Acta. Biotheor. 49, 207–218 (2001)
18. Bernot, G., Comet, J.P., Richard, A., Guespin, J.: Application of formal methods to biological regulatory networks: Extending Thomas' asynchronous logical approach with temporal logic. Journal of Theoretical Biology 229(3), 339–347 (2004)
19. de Jong, H., Gouzé, J.L., Hernandez, C., Page, M., Sari, T., Geiselmann, J.: Qualitative simulation of genetic regulatory networks using piecewise-linear models. Bull. Math. Biol. 66(2), 301–340 (2004)
20. Mateus, D., Gallois, J.P., Comet, J.P., Le Gall, P.: Symbolic modeling of genetic regulatory networks. J. of Bioinformatics and Comput. Biol. 5(2B), 627–640 (2007)
21. Fromentin, J., Comet, J.P., Le Gall, P., Roux, O.: Analysing gene regulatory networks by both constraint programming and model-checking. In: EMBC 2007, 29th IEEE EMBS Annual Intern. Conf., pp. 4595–4598. IEEE Press, Los Alamitos (2007)
22. Fanchon, E., Corblin, F., Trilling, L., Hermant, B., Gulino, D.: Modeling the molecular network controlling adhesion between human endothelial cells: Inference and simulation using constraint logic programming. In: Danos, V., Schachter, V. (eds.) CMSB 2004. LNCS (LNBI), vol. 3082, pp. 104–118. Springer, Heidelberg (2004)
23. Corblin, F., Fanchon, E., Trilling, L.: Modélisation de réseaux biologiques discrets en programmation logique par contraintes. Technique et Science Informatiques 26(1-2), 73–98 (2007)

Modelling *fim* Expression in *Escherichia Coli* K12

Patrick de Vries[1], Colin G. Johnson[1], and Ian C. Blomfield[2]

[1] School of Computing and [2] School of Biosciences,
University of Kent, Canterbury, Kent, CT2 7NF
{pd79,c.g.johnson,i.c.blomfield}@kent.ac.uk
http://www.cs.kent.ac.uk

Abstract. Fimbriae are structures in *Escherichia coli*, the expression of which is controlled by the *fim* operon. Understanding this expression is important because the fimbriae are important virulence factors.

This expression can be studied using targeted mutations to the DNA, which can be used to disable binding or transcription of a protein. However, this can be problematic as only the net effect is observed. Turning off expression of a protein may enhance *fim* expression, but deactivating this protein may also repress another protein that functions as an activator of *fim* expression. The net result may be that *fim* expression goes down, so it would seem at first glance that the disabled protein was an activator of *fim* expression and not a repressor.

In order to understand this complex network of interactions, an agent based model of *fim* expression has been created. The subject of this paper is to introduce this model and to use it to disambiguate between a number of hypotheses about this system. Parameters such as binding probability will be optimised using a genetic algorithm. The final model and parameters show a good match to experimental data.

1 Introduction

Fimbriae are hair-like attachments that *Escherichia coli* (*E.coli*) bacteria use to attach themselves to host cells and subsequently enter them. Because of the bacteria's ability to penetrate cells, *E.coli* bacteria infections are very hard to treat and so it is imperative we learn more about the way the fimbriae are regulated.

The main method of investigating the processes within a bacterial cell is by making focused mutations of the DNA. By directed disabling of the production of protein, or by changing binding sites within the DNA new information on protein expression can be gained. However, a mutation can have further effects within the cell then just the process focussed upon. therefore an effect attributed to a DNA fragment or protein can in fact be a different mechanism.

One can try using a computer model to simulate the process, but for this one needs parameters to feed the model, such as binding affinities etc. The aim of this paper is to model this process using the experimental data currently available, i.e. data on replacement mutations.

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 14–25, 2010.

A key regulator of *fim* expression is the protein FimB. The regulation of FimB expression is not well understood at present. One theory is that H-NS (Histone-like nucleoid-structuring) protein [1] represses *fimB* expression and that SlyA—a protein first discovered in *Salmonella*—antagonises H-NS and reduces *fimB* repression [2]. Experimental data [3] concerning these interactions has been produced via replacement mutations, where the binding sites that control this expression are deleted and the consequent behaviour of the system observed. This information will form the core data input for tuning parameters in our model.

2 Previous Work

Previous attempts have been made at modelling aspects of *E.coli*, for example by means of differential equations, either focused on the individual cell and the processes within [4,5], or on the entire population [6]. Using differential equations on the whole population can be a good method for predicting global properties such as cell growth, but since biological systems are inherently not continuous, these models will ignore the stochastic nature of the system. For this reason we can use stochastic agent-based models such as those of Karmakar and Bose[7] and Ramsey et al. [8]. Karmakar and Bose describe a stochastic model for transcription factor-regulated gene expression, however this is limited to a broad conceptual model because the detailed parameters are not matched to any experimental data. Ramsey et al. discuss a modelling environment for stochastic and deterministic models and compare results for complex—but well known—regulatory networks using both a deterministic and a stochastic approach. For this example data is available, but it is not clear what could be done if the data were limited or not available. Parameters necessary for this are binding affinity/probability and extent of interaction between different bound proteins. Usually, binding probability of a protein to the DNA is found by gel-shift experiments [9].

A significant difficulty in understanding *fim* expression is that there is no direct way of measuring binding affinity of the protein SlyA. Normally, when doing these gel shifts at different concentrations of the protein clear bands appear for the parts where the protein is bound to the DNA [10]. For unknown reasons gel shifts with SlyA produce irregular banding. The only band with a consistent location is that of the unassociated DNA. If gel shifts would produce consistent results we could have used a similar method as Valeyev et al. [11] used in their model for calcium-calmodulin interaction.

There are many hypotheses for how *fim* expression is regulated in *E.coli* [12,9]. A main regulatory process in the expression is controlled by a fragment of DNA that can be expelled and reinserted in the opposite direction [13,14]. It can be seen as a switch turning from OFF to ON and back [15,16]. It is also known that the switch is regulated by the proteins FimB and FimE, where FimB is expected to turn the switch from OFF to ON and FimE favours the OFF position [12]. The regulation of FimB is the main focus of this paper.

3 Materials and Methods

Our work consists of two main parts, an agent-based model for the *fim* expression and a parameter optimisation model using genetic algorithms.

3.1 Agent-Based Model

In this case the hypothesis tested will be the assumption that the protein H-NS will act as a repressor for *fimB* expression and the protein SlyA will be acting as an antagoniser of H-NS preventing it from binding to the DNA.

The main components of this system are the regulatory region for the *fimB* gene and the proteins that bind to that region, which are SlyA and H-NS. These proteins and the binding sites are represented by entities in the model, which interact according to the description given in the remainder of this section.

Two SlyA binding sites have been identified, called O_{SA1} and O_{SA2}—there is also a possible third site called O_{SA3}. The sites O_{SA1} and O_{SA3} overlap not only with each other (by one base pair) but also with H-NS binding sites. This is shown in Figure 1.

H-NS represses *fimB* expression and is antagonised by SlyA. FimB in turn will switch the *fim*-switch (*fimS*) ON which will start the translation of the *fim* operon to form the actual fimbriae. While the switch is turned ON FimE will be formed which will stimulate the switch to turn OFF.

Two sites for H-NS binding have been identified to repress *fimB* expression, however in previous hypotheses it was believed that the experimental results could be explained by only having one H-NS site (H-NS1). Therefore in Model 1 only H-NS1 is taken into consideration.

It is unknown how high the binding affinity is of SlyA to any of the three possible binding sites or how strong the effect of SlyA is on the binding of H-NS. Using an agent-based model with a genetic algorithm to supply the binding affinities for the different binding sites and the effects of binding on the repression will circumvent this problem.

A number of variants on the model have been hypothesised, with different assumptions about the interaction between the binding sites. These are given in Table 1. Figure 1a shows model 1 from Table 1, whereas the various interactions in models 2-5 can be understood with reference to Figure 1b. Figure 1 also shows the regions RM40, RM39 and RM42, which are the regions that are replaced by a non-functional DNA fragment in the various replacement mutation experiments.

In total there are several parameters to optimise, subdivided into 5 sets. For example, Model 1 has 15 parameters. The first 5 parameters describe the effects of binding to the different sites O_{SA1}, O_{SA2} and O_{SA3} has on H-NS. The second group contains 4 parameters describing the binding probability of the two protein to their respective binding sites. Two parameters describe at which concentration of FimB or FimE the switch has a 50% probability to turn OFF-to-ON or ON-to-OFF. There are three parameters describing the effects of the replacement mutations on the binding of H-NS and the final parameter gives the effect of repression of *fimB* expression by H-NS (See also Table 2 for a description).

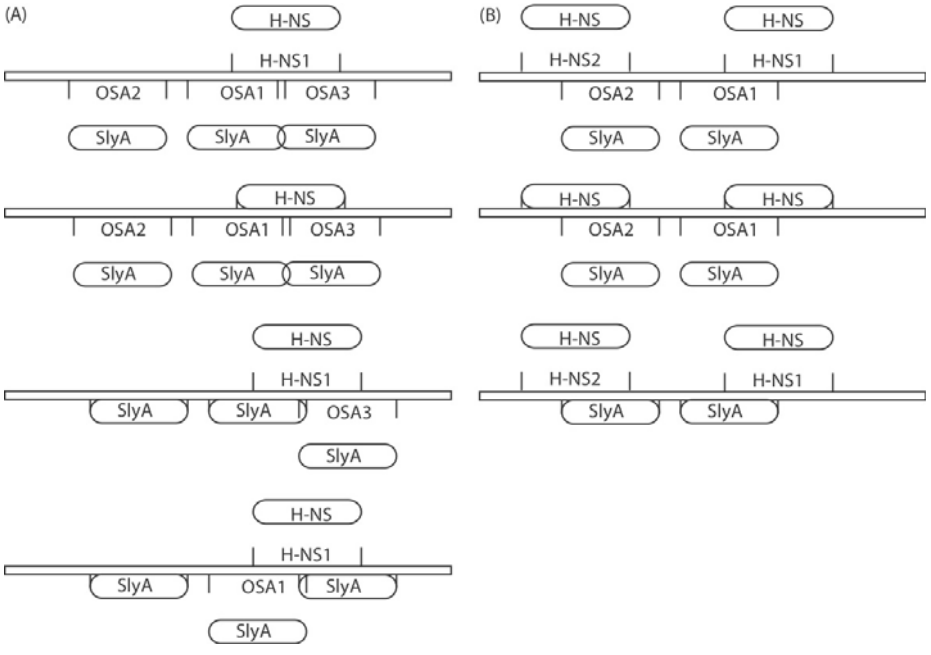


Fig. 1. (a) assumption 1 – Three SlyA binding sites, O_{SA1}, O_{SA2} and O_{SA3} and the overlap with the binding site for H-NS. (b) assumption 2 – Two SlyA binding sites, O_{SA1} and O_{SA2} overlap each with an H-NS binding site. RM39, RM40 and RM42 are replacement mutations targetted to replace respectively O_{SA1}, O_{SA2} and O_{SA3}. H-NS can also be partially replaced by SlyA and the different SlyA sites (O_{SA1}, O_{SA2} and O_{SA3}) can also act independent from each other where the effect on H-NS is reduced.

Table 1. Summary of differences in the 5 models

Model	Summary
1	Assumption 1, where effect of SlyA on H-NS is expected to be the same as the Replacement mutation.
2	Assumption 2. H-NS effect on *fimB* expression is when bound always 100%
3	H-NS has a cumulative effect on *fimB*expression, but H-NS effect can vary.
4	H-NS has an independent variable effect on *fimB* expression, but repression is 100% when both sites are occupied.
5	H-NS only has a variable effect on *fimB* expression, but only when both sites are occupied.

Table 2. Explanation of the different parameters

Sets	Explanation
Set 1	Effect of binding of SlyA to O_{SA1}, O_{SA2} and O_{SA3}
Set 2	Binding probability of SlyA and H-NS to the different sites
Set 3	Concentration of either FimB or FimE at which switching probability is 50%
Set 4	Effect of different replacement mutations
Set 5	Repressing effect of H-NS binding on *fimB* expression

In the simulation the population of *E. coli* bacteria start out as 50 afimbriate cells, growing, dividing and dying for 1000 iterations, where the colony grows to approximately 30,000 cells, consisting of a mixture of the fimbriate and afimbriate types. At each iteration, each cell individually checks the amount of FimB and FimE protein and based on that decides whether to switch the production of fimbriae ON or OFF. FimE promotes the ON-to-OFF switch and is only produced when the switch is turned ON. FimB production depends on the repression by H-NS and production is independent of the switch, although FimB promotes the switch OFF-to-ON. This reflects the best current knowledge about the functioning of the biological system.

The calculation of the binding probabilities of SlyA and H-NS depends on a number of parameters, that are described in Table 2. When H-NS binds it has a maximum effect on the repression of FimB. The effect is reduced when SlyA binds and the repression is reduced by as much as stated in parameter set 1. The binding probability of H-NS can be reduced or enhanced in the case of replacement mutations by values stated in parameter set 5. When SlyA binds to one of the sites it reduces the effect of H-NS inhibiting *fimB* expression by as much as the the respective gene from parameter set 1.

3.2 Parameter Optimisation

The parameter optimisation model starts out with generating a population of solutions for the parameters. In each generation the solutions are tested by running the agent-based model described above on each set of parameters in the population and comparing with experimental data, from which a fitness measure is calculated, which is stored in an output-file.

The experimental data used is concerned with replacement mutations obtained from *switching experiments* as done by Gally et al. [17] and calculated from β-galactosidase experiments [18] (the relation between β-galactosidase and *switching frequency* as shown by El-Labany et al. [19]; data as used in the model can be found in Figure 3F). These mutations include RM39, RM40 and RM42,

where RM39 replaces O_{SA1}, RM40 replaces O_{SA2} and RM42 replaces O_{SA3} (as illustrated in Figure 1). Notice that RM39 and RM42 have a direct effect on the binding of H-NS. Two other mutations include ΔSlyA and ΔH-NS[1].

The initial population of parameters is randomly created by sampling from a uniform distribution within sensible ranges. The algorithm then iterates for 50 generations. In each generation, after they have been tested for their fitness, the best solution is kept for the next generation and the rest of the solutions are generated by means of crossover. The candidates for crossover are selected by tournament selection, where from a random selection of four solutions the two strongest are mixed [20,21]. The new parameter sets are then subjected to random mutation, where one of the parameters is altered, to prevent ending in a local optimum.

Fitness is measured by taking the least square error (LSE) for each model compared with experimental results.

$$\text{LSE} = \sum (F_{i,exp} - F_{i,Model})^2 \tag{1}$$

Where i is the *fimB* expression for each of the different mutants.

For every test of a parameter list a new run of the model is created as described in the previous section. Reaching an optimum in the parameter optimisation is a good sign the hypothesis may be correct, however, the parameters produced should be scrutinised by comparison with what is known from biology in order to check that the hypothesis produced by the optimisation process is biologically realistic.

4 Results

The LSE for the different models is shown in Table 5 and the change in LSE over the generations is shown in Figure 2. The parameter optimisation shows that most of the different assumptions lead to an optimum where the least square error (LSE) is reduced 10-400 fold.

The plots of *fimB* expression for each mutant is plotted in Figure 3, where Model 1 has a problem with modelling the SlyA mutant, but the other models give a close resemblance with experimental data, also shown by having relatively low error values.

A further test is to look at the different parameters collected from these different models. These are shown in Tables 3-4. No experimental data is available to compare these. However we do know that H-NS binds much stronger to the DNA then SlyA and also is known that replacement mutation RM40 makes site H-NS2 closer to the consensus of an H-NS binding site, so may actually enhance binding of H-NS. Replacement mutation RM39 and RM42 tend to make binding of H-NS to site H-NS2 less likely.

[1] In wild type background the absence of H-NS is tested and as with the other experiments either with or without SlyA present.

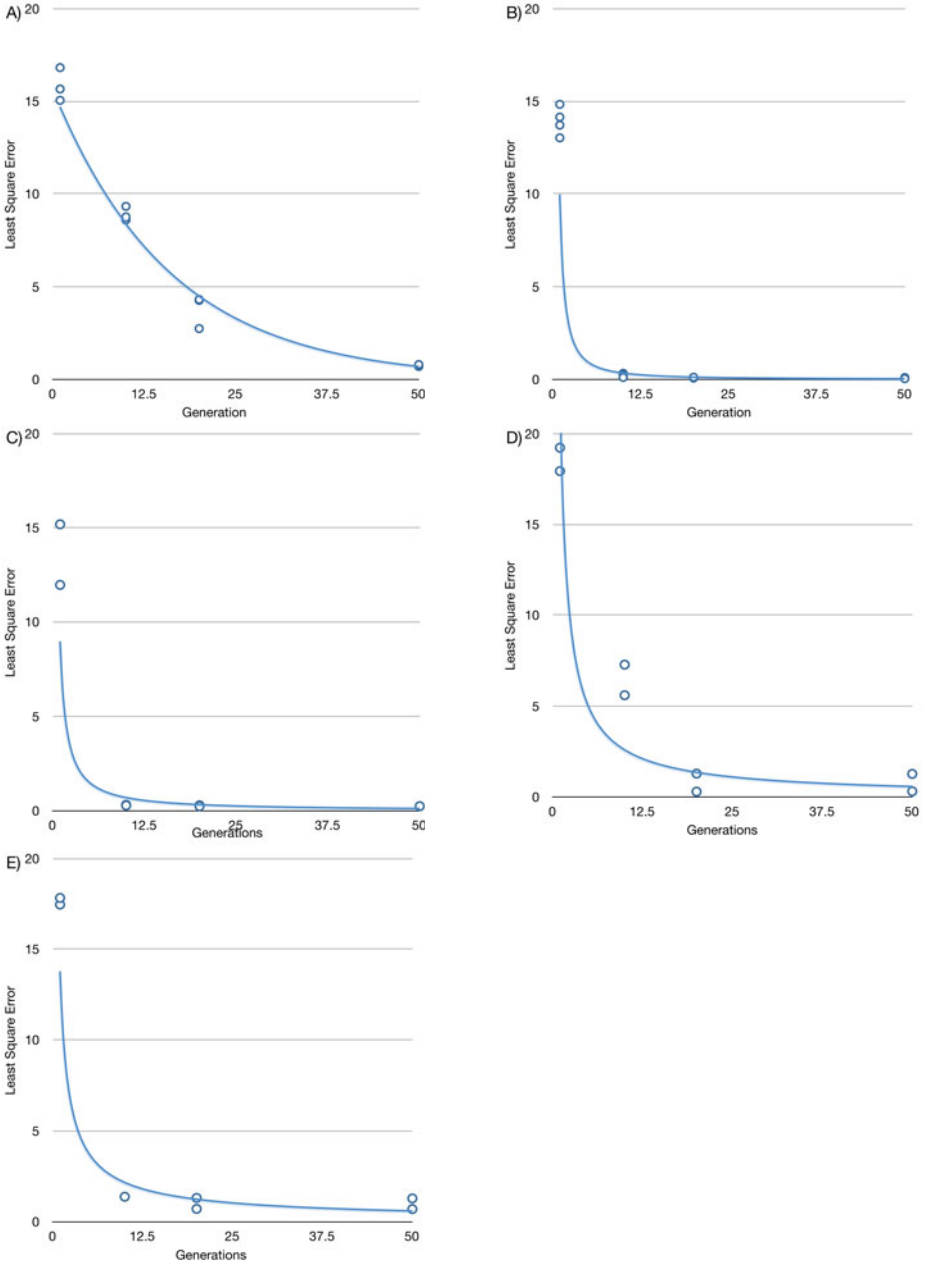


Fig. 2. Error values as a function of generation, samples at initial population, after 10 generations, 20 generations and finally after 50 generations. See A, B, C, D or E for respectively Model 1, 2, 3, 4 or 5.

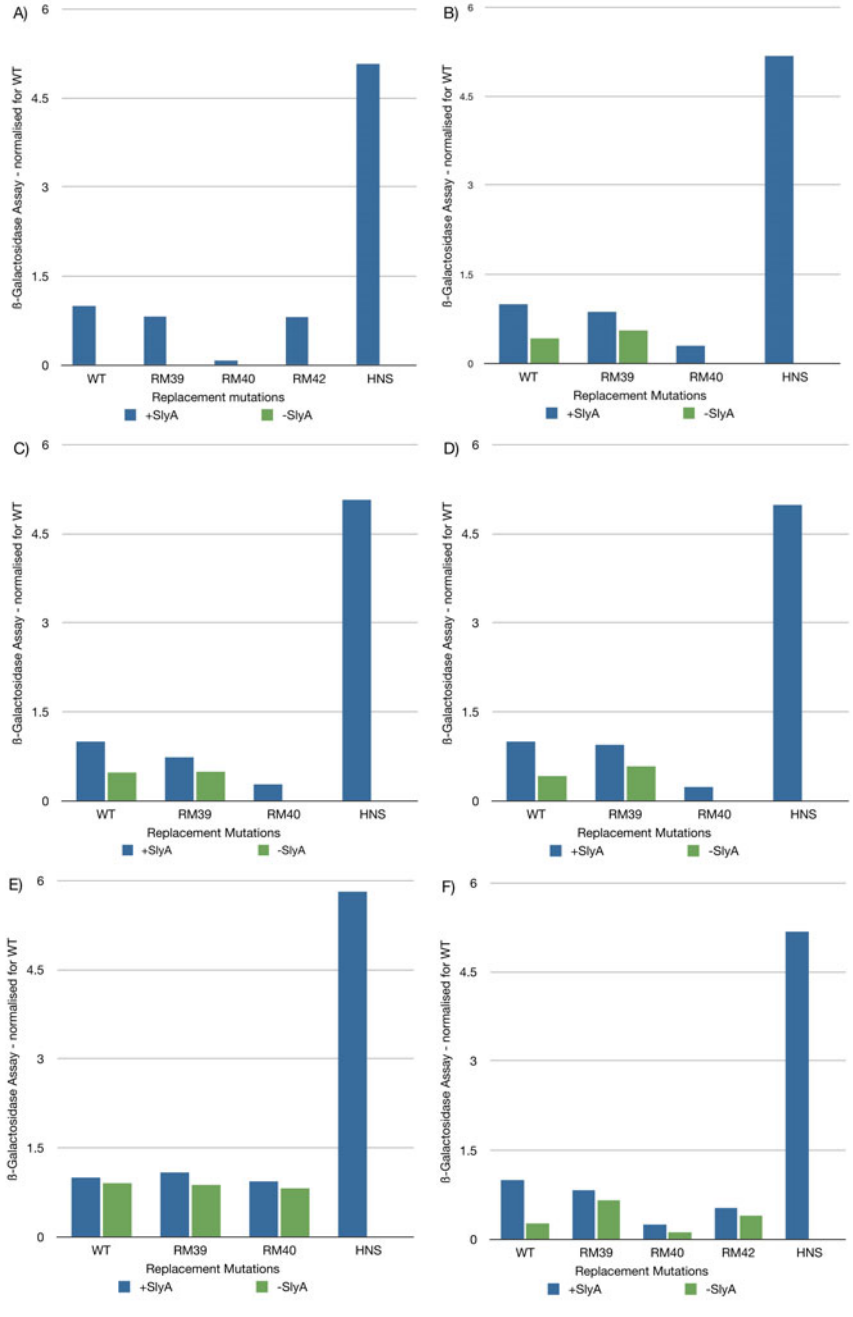


Fig. 3. *fimB* expression for each model. See A, B, C, D or E for respectively Model 1, 2, 3, 4 or 5 and F for experimental values

Table 3. Parameter sets 1 and 2. Parameter set 1 shows the effect of SlyA binding on H-NS repression, and parameter set 2 shows binding probabilities.

	Parameter Set 1					Parameter Set 2				
Model	O_{SA1}	O_{SA2}	O_{SA3}	$O_{SA1\&2}$	$O_{SA2\&3}$	O_{SA1}	O_{SA2}	O_{SA3}	H-NS1	H-NS2
1	29	40	23	53	9	9	58	9	99	-
2	5	15	-	-	-	31	22	-	100	94
3	27	1	-	-	-	9	43	-	100	94
4	13	3	-	-	-	15	33	-	100	94
5	2	12	-	-	-	4	21	-	92	98

Table 4. Parameter sets 3,4 and 5. Parameter set 3 shows the FimB and FimE concentrations in the cell at which the probability of switching ON or OFF the production of fimbriae is 50%, parameter set 4 shows the effect of replacement mutation on the ability of H-NS to repress *fimB* expression (where negative numbers mean that H-NS binding is enhanced), and parameter set 5 shows the extent of H-NS repression of *fimB* expression (note that Model 1 only has one H-NS binding site).

	Parameter Set 3		Parameter Set 4			Parameter Set 5	
Model	[FimB]	[FimE]	RM39	RM40	RM42	H-NS1	H-NS2
1	100	3	0	8	1	100	-
2	25	92	3	-64	-	100	100
3	100	39	3	-71	-	100	100
4	75	39	-88	-49	-	83	82
5	55	19	-8	-47	-	93	100

5 Discussion

In general the model corresponds well with the experimental data as the graphs for *fimB* expression are nearly identical to that of the biological experiments. This is also shown in the small error value in Table 5. Even though Model 1 has a smaller error value, it misses out greatly in predicting *fimB* expression for the SlyA mutants. The real measure however is in the parameters.

Knowing the biological mechanism on which this model is based, we can see that only the parameters of Model 3 seem to make any sense, as the parameters are close to what we were expecting from what we know from the biological system. We know that the H-NS sites will be occupied most of the time. Binding of SlyA is two to ten times weaker than H-NS. It has been shown that RM40 makes the H-NS2 site closer to the H-NS consensus, thus increasing the binding of H-NS to this site (negative value for this parameter). RM39 disturbs binding of H-NS to H-NS1, even though not very strongly it is still a significant effect. O_{SA1} or O_{SA2} have a small, but still significant effect on H-NS repression. This is all in line with what was expected from biological experiments. The other 3 models (2, 4 and 5) differ too much from these expectations to be considered as valid hypotheses, in addition to having a higher error value.

Table 5. Results for the 5 models, Error values

Model	Error
1	0.7920
2	0.0566
3	0.0611
4	0.1013
5	1.1901

At the time of writing it is unclear how the two H-NS sites interact when repressing *fimB* expression. The results of this study would suggest the two H-NS sites act independently, although their effect on the repression is cumulative.

The third SlyA site has only been included in model 1, and its omission from the other models is supported by an examination of the DNA sequence, which shows that the O_{SA3} has less resemblance to the SlyA binding site consensus compared with the other two sites (See Table 6). Further experiments have shown that SlyA seems unable to bind to O_{SA3}.

Table 6. Genetic code of the different binding sites

Binding site	Genetic code
SlyA	TTAGCAAGCTAA
O_{SA1}	TTAGCATGATAA
O_{SA2}	CTAGGGACCTAA
O_{SA3}	ATAGCCACTAA

Further work is needed to investigate the remaining H-NS sites. There are two more sites in the same region on the DNA, although it is accepted that these sites are not under control by any of the SlyA sites mentioned here. A further SlyA site (O_{SA4}) has been identified, but under normal circumstances is not found to be occupied by SlyA, but the site does overlap with the H-NS2 site and under some circumstances its effect on antagonising H-NS is still significant. With existing knowledge of the system and careful analysis of the parameters found, it has been possible to rule out that the resemblance of the model to experimental data is effectuated merely by coincidence.

6 Conclusions

An agent-based model has been presented for the regulation of expression of *fimB* in *E. coli* with regard to the regulatory proteins SlyA and H-NS. A number of hypotheses have been presented for the effect of these proteins on the expression, and optimisation of parameters against experimental data from replacement mutation experiments has been used to disambiguate between these

hypotheses. One hypothesis has clearly been identified as the most likely candidate for explaining the experimental data.

Future work will focus on the influence of other H-NS and SlyA binding sites on the system.

References

1. Dame, R.T., Luijsterburg, M.S., Krin, E., Bertin, P.N., Wagner, R., Wuite, G.J.L.: DNA bridging: a property shared among H-NS-like proteins. Journal of Bacteriology 187(5), 1845–1848 (2005)
2. Corbett, D., Bennet, H.J., Askar, H., Green, J., Roberts, I.S.: SlyA and H-NS regulate trascription of the *Escherichia coli* K5 capsule gene cluster, and expression of *slyA* in *Escherichia coli* is temperature dependent, positively autoregulated, and independent of H-NS. Journal of Biological chemistry 282(46), 33326–33335 (2007)
3. Sohanpal, B.K., Friar, S., Roobol, J., Plumbridge, J.A., Blomfield, I.C.: Multiple co-regulatory elements and IHF are necessary for the control of *fimB* expressionin response to sialic acid and n-acetylglucosamine in *Escherichia coli* K-12. Molecular Microbiology 63, 1223–1236 (2007)
4. Chu, D., Blomfield, I.C.: Orientational control is an efficient control mechanism for phase switching in the *E. coli fim* system. Journal of Theoretical Biology 244(3), 541–551 (2007)
5. Chu, D., Roobol, J., Blomfield, I.: A theoretical interpretation of the transient sialic acid toxicity of a *nanR* mutant of *Escherichia coli*. Journal of Molecular Biology 375, 875–889 (2008)
6. Chen, T., He, H.L., Church, G.M.: Modeling gene expression with differential equations. Pac. Symp. Biocomput., 29–40 (1999)
7. Karmakar, R., Bose, I.: Stochastic model of transcription factor-regulated gene expression. Physical Biology 3, 200–208 (2006)
8. Ramsey, S., Orrell, D., Bolouri, H.: Dizzy: Stochastic simulation of large-scale genetic regulatory networks. Journal of Bioinformatics and Computational Biology 3, 415–436 (2005)
9. Roesch, P.L., Blomfield, I.C.: Leucine alters the interaction of the leucine-responsive regulatory protein (Lrp) with the *fim* switch to stimulate site-specific recombination in *Escherichia coli*. Molecular Microbiology 27(4), 751–761 (1998)
10. Lithgow, J.K., Haider, F., Roberts, I.S., Green, J.: Alternate SlyA and H-NS nucleoprotein complexes control *hlyE* expression in *Escherichia coli* K-12. Molecular Microbiology 66, 685–698 (2007)
11. Valeyev, N.V., Bates, D.G., Heslop-Harrison, P., Postlethwaite, I., Kotov, N.V.: Elucidating the mechanism of cooperative calcium-calmodulin interactions: a structural systems biology approach. BMC Systems Biology 48(2) (2008)
12. Klemm, P.: Two regulatory *fim* genes, *fimB* and *fimE*, control the phase variation of type 1 fimbriae in *Escherichia coli*. The EMBO Journal 5(6), 1389–1393 (1986)
13. Blomfield, I.C., Kulasekara, D.H., Eisenstein, B.I.: Integration host factor stimulates both FimB- and FimE-mediated site-specific DNA inversion that controls phase variation of type 1 fimbriae expression in *Escherichia coli*. Molecular Microbiology 23(4), 707–717 (1997)
14. Adiciptaningrum, A.M., Blomfield, I.C., Tans, S.J.: Direct observation of Type 1 fimbrial switching. EMBO reports 10(5), 527–532 (2009)

15. Gardner, T.S., Cantor, C.R., Collins, J.J.: Construction of a genetic toggle switch in *Escherichia coli*. Nature 403, 339–342 (2000)
16. Wolf, D.M., Arkin, A.P.: 15 minutes of *fim*: Control of phase variation in *E.coli*. OMICS: A Journal of Integrative Biology 6(1), 91–114 (2002)
17. Gally, D.L., Bogan, J.A., Eisenstein, B.I., Blomfield, I.C.: Environmental regulation of the *fim* switch controlling Type 1 fimbrial phase variation in *Escherichia coli* K-12: Effects of temperature and media. Journal of Bacteriology 175(19), 6186–6193 (1993)
18. Miller, J.H.: Experiments in molecular genetics. Cold Spring Harbor Laboratory (1972)
19. El-Labany, S., Sohanpal, B.K., Lahooti, M., Akerman, R., Blomfield, I.C.: Distant *cis*-active sequences and sialic acid control the expression of *fimB* in *Escherichia coli* K-12. Molecular Microbiology 4, 1109–1118 (2003)
20. Mitchell, M.: An Introduction to Genetic Algorithms. The MIT Press, Cambridge (1996)
21. Eiben, A.E., Smith, J.E.: Introduction to Evolutionary Computing. Springer, Heidelberg (1998)

Study of the Structural Pathology Caused by CYP2C9 Polymorphisms towards Flurbiprofen Metabolism Using Molecular Dynamics Simulation

Yuranat Saikatikorn[1], Panida Lertkiatmongkol[2], Anunchai Assawamakin[3], Marasri Ruengjitchatchawalya[4], and Sissades Tongsima[3,*]

[1] Bioinformatics and Systems Biology Program, King Mongkut University of Technology Thonburi, Bankok, Thailand
[2] Department of Biochemistry, Faculty of Science, Mahidol University
[3] Biostatistics and Informatics Laboratory, Genome Institute, National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand Science Park
[4] School of Bioresources and Technology, King Mongkut University of Technology Thonburi, Bankok, Thailand
sissades@biotec.or.th

Abstract. CYP2C9 is one of the major cytochrome P450 enzymes that play a crucial role in metabolic clearance of several drugs in the current clinical used. CYP2C9 has several allelic variant forms each of which arises from single amino acid substitution and could reduce/increase enzyme activities and affect drug metabolism. Mutant alleles may cause serious toxicity in some narrow therapeutic index drugs. CYP2C9*13, one of the CYP2C9 variant forms that is commonly found in Asian population, has a Leu90Pro amino acid substitution that leads to defective drug metabolism in individuals who carry this allele. It has been reported that metabolic activity of CYP2C9*13 was reduced towards some CYP2C9 substrates compared to wildtype. In this study, X-ray crystal structure of human cytochrome P450 2C9 complexed with flurbiprofen (PDB code: 1R9O) was represented to wildtype and the structure of CYP2C9*13 was constructed based on the X-ray crystal structure of CYP2C9-flurbiprofen complex. Herein, molecular docking of CYP2C9*1 and CYP2C9*13 with flurbiprofen was performed in search for flurbiprofen orientation that corresponds to its binding state before undergoing monooxygenation. Subsequently, molecular dynamics simulation was operated to compare binding of flurbiprofen in catalytic cavity of these 2 variants. Substrate access channel of CYP2C9*13 has a dramatic effect on an interaction between the drug and the enzyme. Consequently, this study can lead to an understanding of structural pathology caused by single amino acid change in CYP2C9*13 variant.

Keywords: Cytochrome P450 2C9, CYP2C9*13, Genetic polymorphisms, Flurbiprofen, Molecular dynamics simulation.

1 Introduction

Cytochrome P450s (P450s) are a diverse superfamily group of enzymes that have been found in all kingdoms of life [1]. P450s are heme-containing enzymes, which

* Corrresponding author.

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 26–35, 2010.

have a main function to catalyze the oxidation of organic substances, so called monooxygenation reaction. They are involved in biotransformation for a large number of endogenous compounds, drugs and xenobiotics [2-3]. Moreover, they are able to activate or metabolize chemical carcinogens, degrade several substances and synthesize important compounds such as steroidal hormones and soluble vitamin. More than 40% and 55% of primary amino acid sequence identities are shared among P450 families and subfamilies, respectively [4]. Despite their difference in amino acid similarity, three-dimensional structures of P450 enzymes are generally conserved [5].

Human P450s are membrane bound proteins that are found on the endoplasmic reticulum, few are identified on mitochondria. P450s anchor membrane by their N-terminal α-helix [6]. More than 57 cytochrome P450 enzymes [7] are encoded in the human genome. CYP2C9 is one of the four functional CYP2C genes (including CYP2C8, CYP2C18 and CYP2C19) that locate on the chromosome 10 [8]. One of the most important and abundant 2C subfamily P450 enzymes in the human liver is CYP2C9. It is responsible for numerous metabolic clearances of therapeutic agents approximately 15% in current clinical used [9] which include a wide range of narrow therapeutic drugs and many non-steroidal anti-inflammatory agents (NSAIDs). However, therapeutic treatments are varied among individuals because of lethal effects caused by polymorphic variants of P450 enzymes. CYP2C9 has several variant forms, which arise from single amino acid substitution, resulting in reduce/increase enzyme activities as well as drug metabolism [8-9]. The narrow therapeutic index drugs could cause serious toxicity to the people who carry the mutant CYP2C9 allele(s). A lot of studies on human CYP2C9 polymorphisms in vivo and in vitro have been conducted in order to elucidate the enzyme-drug interactions. In addition, these studies have made a great effort to clarify the influences of CYP2C9 polymorphisms that alter enzyme activities and drug metabolism [9-11].

CYP2C9*13 is a novel CYP2C9 variant form that is commonly found in Asian population [12, 13]. It emerges from a T269C transversion of the CYP2C9 gene, causing a substitution of residue 90 leucine to proline (L90P) [13]. This mutation is located in N-terminal loop that is closed to an entrance for substrate access path. Regarding to allele frequency analysis, the incident of this allele is approximately 1.02% in the Chinese population [12] and 0.6% in the Korea population [14]. Furthermore, numerous studies have measured the catalytic activities of CYP2C9 mutant in comparison to wildtype against various substrates [10, 15-16]. Owing to the important role of CYP2C9 polymorphisms in defective drug metabolism including the serious toxicity in the poor metabolizers carrying CYP2C9*13 allele, it is necessary to describe structural pathology of CYP2C9*13 variant for better understanding on how single amino acid substitution in this allele affects enzymatic activities and influences drugs metabolism.

In 2003, X-ray crystal structure of CYP2C9 both unliganded and complexed with the anti-coagulant drug warfarin (PDB code: 1OG2 and 1OG5) were determined [17]. Later, a crystal structure of CYP2C9 with flurbiprofen bound (PDB code: 1R9O) was investigated in 2004 [18]. With an advantage of these available X-ray crystal structures of CYP2C9, structural analyses of CYP2C9-drug interaction can be attained, especially the effect of single amino acid substitution in mutant alleles. Previously, Zhou et al. determined the structure of CYP2C9*13 and found out that the size of

substrate access channel was altered compared to CYP2C9*1, resulting in difficulty of substrates to enter into active site cavity [19].

In this study, the crystal structure of CYP2C9-flurbiprofen complex (PDB code: 1R9O) was designated as wild type structure while the structure of CYP2C9*13 was constructed based on the X-ray crystal structure of CYP2C9-flurbiprofen complex. The constructed CYP2C9*13 with bound flurbiprofen was implemented to investigate the structural pathology that is induced by single nucleotide polymorphism resulting in reduced metabolic activities. Molecular dynamics simulation can provide the information on a characteristic of specific single amino acid substitution in CYP2C9 polymorphism causing defective enzymatic activities and influences drug metabolism. It gives a more comprehensive on how the single amino acid substitution in CYP2C9*13 has an effect on drug-enzyme complex. The aim of this study is to introduce the molecular dynamics simulation as a tool to explore the underlying structural pathology of CYP2C9 polymorphisms towards ineffective flurbiprofen metabolism. In addition, interaction of enzyme-drug complex can be evaluated by means of pharmacophore model.

2 Methods

2.1 Structure Preparation

Crystal structure of human cytochrome P450 2C9 with flurbiprofen bound (PDB code 1R9O) [18] was obtained from the Brookhaven Protein Databank (http://www.pdb.org/). This structure was solved by X-ray crystallography at 2.0 Å resolution and has a number of missing residues. These missing residues (residues 38-42 and residues 214-220) were resolved by MODELLER9v6 [20]. The CYP2C9-flurbiprofen complex was employed to construct CYP2C9*13 which has mutation of residue 90 leucine to proline (L90P) by using SCWRL3.0 [21].

Additionally, initial atomic coordinates of flurbiprofen were obtained from ChemIDplus database (http://chem.sis.nlm.nih.gov/chemidplus/) in order to perform molecular docking and further molecular dynamics simulation. All computation was carried out on Linux high performance cluster AMD quad cores 2.3 GHz 64 GB memory available at BIOTEC, NSTDA, Thailand.

2.2 Molecular Docking

The molecular interaction between the CYP2C9-flurbiprofen complex was computed by using AutoDock4.0 program [22]. Partial charges of flurbiprofen and CYP2C9 were assigned using Gasteiger method with the aid of AutoDockTools [23].

Affinity maps were generated using AutoGrid program [23] and centered on heme. The maps were manually adjusted following substrate recognition sites of CYP2 family that were previously proposed by Gotoh [24]. Dimension of cubic box was set to be 60 x 60 x 60 grid points and 0.375Å spacing. AutoDock4.0 program was employed to dock ligands into catalytic cavity of CYP2C9 by using Lamarckian genetic algorithm (LGA) consisting 200 runs with 270,000 generations.

Estimated free binding energy of each substrate-protein complex was considered. Dock conformations were clustered and analyzed using AutoDockTools.

2.3 Molecular Dynamics Simulation

Simulations of CYP2C9-flurbiprofen complex both CYP2C9*1 and CYP2C9*13 were performed by using AMBER10.0 package [25]. The SANDER program of AMBER10.0 was used for minimization and MD simulations. The initial structures were first energy minimized for 2,000 steps (1,000 steepest descent [26] and 1,000 conjugate gradient) and then simulated at a temperature of 300 K. The Shake algorithm was applied to all bonds containing hydrogen atoms, and a time step of 2 fs was used. The method of Berendsen was used to couple the system to constant temperature and pressure. The carbon atoms were restrained for 20 ps and followed by an unrestrained simulation of 2.5 ns.

2.4 Pharmacophore Model

LigandScout2.03 [27], a software tool that automatically derives pharmacophores from protein-ligand complexes, was used to determine interaction patterns between CYP2C9 and flurbiprofen obtained from molecular dynamics simulation.

3 Results

3.1 Molecular Docking of CYP2C9*1

The crystal structure of CYP2C9-flurbiprofen complex (PDB code: 1R9O) was docked with flurbiprofen in order to find a suitable conformation of flurbiprofen accessing into interior cavity of CYP2C9*1. The initial structure of CYP2C9-flurbiprofen complex is illustrated in figure 1. Estimated free binding energy of this substrate-protein complex was -6.66 kcal/mol, which is favorable for the drug to bind in this position of the enzyme.

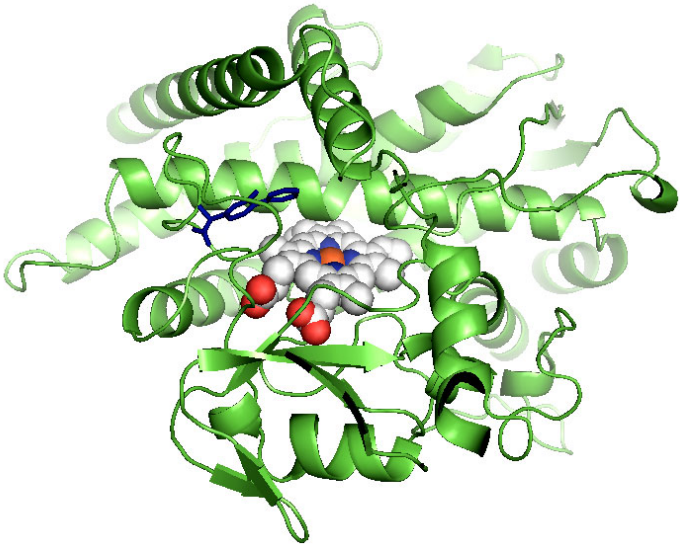


Fig. 1. Initial binding orientation of flurbiprofen obtained from molecular docking of CYP2C9*1. This conformation of flurbiprofen was also applied to CYP2C9*13. Flurbiprofen is represented by blue stick.

3.2 Molecular Dynamics Simulation of CYP2C9*1 and CYP2C9*13

Molecular dynamics simulation was performed to determine enzyme-drug interaction. In addition, it allows an observation of drug motions in substrate access channel of the protein molecule. The conformation of CYP2C9-flurbiprofen complex selected from the molecular docking study was used as a starting structure for simulations. Simulations of both CYP2C9*1 and CYP2C9*13 were performed by using AMBER10.0 package. BC loop of CPY2C9*13 was dramatically altered, and it led to different drug-enzyme interactions between CYP2C9*1 and CYP2C9*13. For CYP2C9*13, Arg108 bends upwards aromatic ring of flurbiprofen, while that of CYP2C9*1 points away from flurbiprofen as demonstrated in figure 2. This slight difference caused a dramatic effect on orientation of flubiprofen in that flurbiprofen is likely to rise up in CYP2C9*13 despite of lying horizontally as observed in CYP2C9*1.

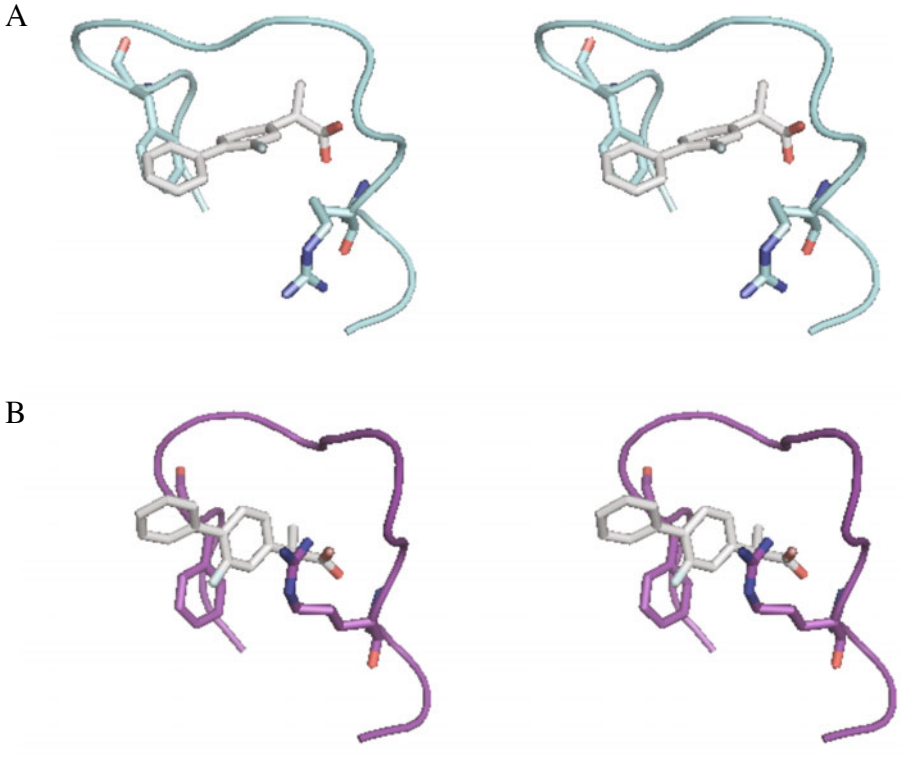


Fig. 2. Stereoview of flurbiprofen binding orientation under BC loop in CYP2C9*1 (A) and CYP2C9*13 (B). Arginine108 of CYP2C9*1 bends away from flurbiprofen, which is in contrary to CYP2C9*13, and leads to additional aromatic interactions of CYP2C9*13-flurbiprofen complex. Phe106 is also implicated in this interaction, reinforcing the aromatic interactions to be stronger.

3.3 Pharmacophore Model of CYP2C9*1 and CYP2C9*13

The conformation of CYP2C9-flurbiprofen complex from the molecular dynamics simulation was analyzed by using LigandScout2.03, which can describe the interaction patterns of drug-enzyme complex.

Pharmacophore modeling of CYP2C9*1-flurbiprofen complex and CYP2C9*13-flurbiprofen complex are shown in figure 3 and figure 4, respectively. In the complex of CYP2C9*1-flurbiprofen, hydrophobic interactions are formed between two aromatic rings of flurbiprofen and non-polar amino acid resides that line the catalytic cavity of the enzyme. These interactions were also observed in CYP2C9*13-flubiprofen complex. Nevertheless, since Arg108 of CYP2C9*13 roars upwards, aromatic interactions between one aromatic ring of flurbiprofen and amino group of Arg108 distinguishingly differentiates drug interaction pattern of CYP2C9*13 from that of CYP2C9*1. Moreover, the aromatic interactions were strengthened by stacking aromatic interactions between aromatic rings of flurbiprofen and Phe106, although Phe106 conformation is similar to that of wild type enzyme.

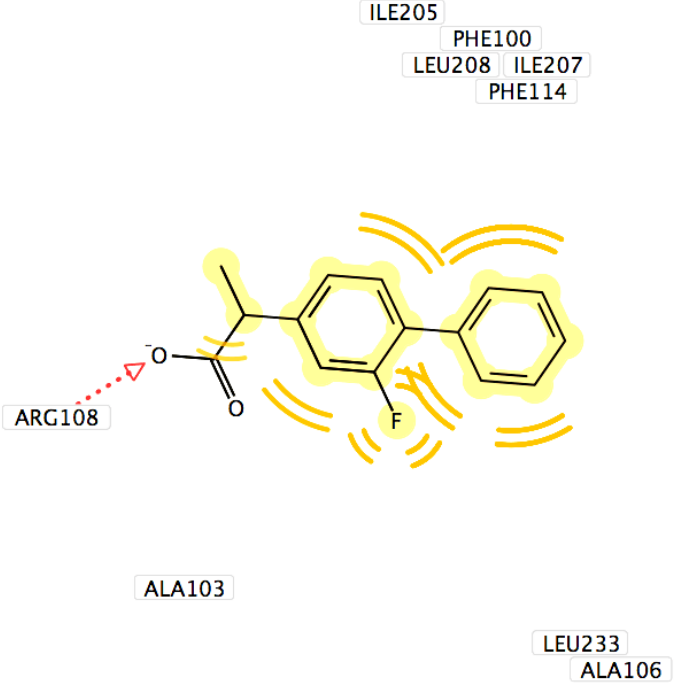


Fig. 3. Pharmacophore modeling of flurbiprofen bound in catalytic cavity of CYP2C9*1, hydrophobic interactions are represented by yellow color and hydrogen bond acceptor represented by a red arrow.

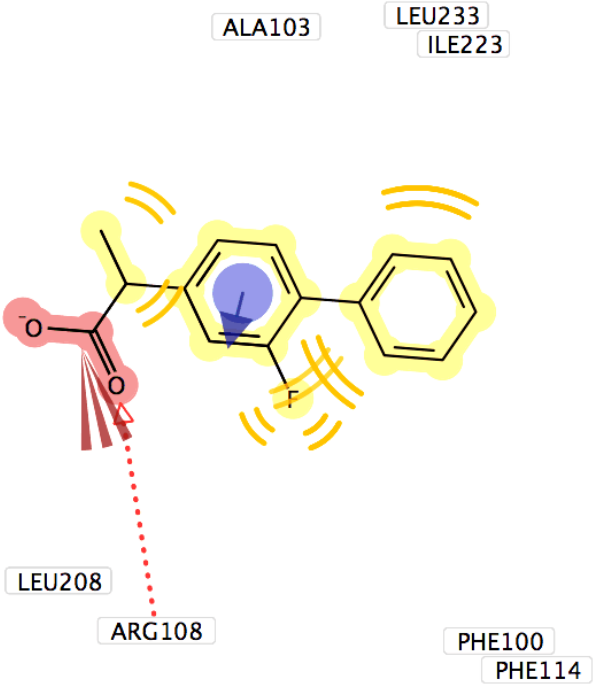


Fig. 4. Pharmacophore modeling of flurbiprofen bound in catalytic cavity of CYP2C9*13. Hydrophobic interactions are represented by yellow color, hydrogen bond acceptor is represented by red arrow, aromatic ring interactions are represented by blue arrow and negative ionizable area is represented by red area.

4 Discussion

CYP2C9 has an important role on both metabolic clearance and the response of a wide range of therapeutic agents. CYP2C9 polymorphisms are associated with reduced enzymatic activity and cause a risk of serious toxicity in poor metabolizers who carry the mutant alleles. Several studies have been indicated that CYP2C9*13, one of the CYP2C9 polymorphism variants caused by a single amino acid substitution of Leu90Pro, exhibits a reduced tolbutamide metabolic activity in some studied CYP2C9 substrates as Michaelis-Menten constant (K_m) of CYP2C9*13 was found to be increased while maximal reaction velocity (V_{max}) was not altered [15]. Surprisingly, V_{max} was reduced in diclofenac metabolism although K_m was also increased [15]. Consequently, drug K_m of CYP2C9*13 tend to be increased, indicating its decline in rate of the reaction. This change in kinetics is probably originated from amino acid substitution that alters the 3D structure of the protein. In an attempt to investigate the structure and metabolism relationship of mutant enzyme, molecular dynamics simulation of lornoxicam as well as diclofenac binding in CYP2C9*13 were performed in comparison to CYP2C9*1 as substrate entrance of CYP2C9*1 is considerably larger than that of CYP2C9*13 [28]. This tremendous change is caused by turnover of

residue 106-108 backbones in CYP2C9*13 [28]. Corresponding to the study by Zhou et al., we also observed the turnover of these residues that results in distinct conformation of Arg108 on BC loop of CYP2C9*13. In addition, they remarked that less hydrogen bonds were formed to stabilize diclofenac and lornoxicam in CYP2C9*13 cavity, affecting distances between the drugs and the heme iron of the mutant enzyme [28]. Herein, we simulated binding of flurbiprofen, which is one of anti-inflammatory drug (NSAIDs) metabolized by CYP2C9, to determine the consequence of this conformational change, which is caused by amino acid substitution. We found that orientation of fluriprofen located below the BC loop of CYP2C9*13 differs from that of CYP2C9*1. To illustrate the interactions more evidently, pharmacophores of flurbiprofen bound in different CYP2C9 variants were constructed and compared. CYP2C9*13-flurbiprofen complex had additional aromatic interactions between aromatic ring of flurbiprofen and amino group of Arg108. These interactions were not observed in CYP2C9*1. Therefore, the aromatic interactions might hinder metabolism rate of flurbiprofen in the mutant enzyme by strengthening the binding of flurbiprofen beneath the BC loop. Consequently, the drug might participate in monooxygenation reaction with difficulty, resulting in reduced metabolic rate. Accordingly, conformation of Arg108 is crucial in binding of flurbiprofen in CYP2C9.

In order to comprehend defective drug metabolism caused by single nucleotide polymorphism, structural insight is demanding. Herein, molecular dynamics simulation may be an alternative approach. It can be applied to investigate structural pathology caused by amino acid substitution of mutant enzymes regardless of the simulation system (Discovery-3 module by Zhou et al. and SANDER program in this study). Furthermore, this study strategy can be applied to other polymorphic variants of CYP2C9 in order to elucidate effects of structural changes that underlie poor metabolic activities in drug clearance among individuals carrying mutant allele(s).

Acknowledgments. This work was supported by the National Center for Genetic Engineering and Biotechnology of Thailand (BIOTEC), School of Information Technology and School of Bioresources and Technology King Mongkut's University of Technology Thonburi. Anunchai Assawamakin was supported by BIOTEC postdoc grant.

References

[1] Sigel, A., Sigel, H., Sigel, R.K.O.: The Ubiquitous Roles of Cytochrome P450 Proteins. In: Metal Ions in Life Science, vol. 3. John Wiley & Sons Ltd, West Sussex (2007)

[2] Guengerich, F.P.: Cytochrome P450 and chemical toxicology. Chem. Res. Toxicol. 21, 70–83 (2008)

[3] Anzenbacher, P., Anzenbacherová, E.: Cytochromes P450 and metabolism of xenobiotics. Cell. Mol. Life Sci. 58, 737–747 (2001)

[4] Nelson, D.R., Koymans, L., Kamataki, T., Stegeman, J.J., Feyereisen, R., Waxman, D.J., Waterman, M.R., Gotoh, O., Coon, M.J., Estabrook, R.W., Gunsalus, I.C., Nebert, D.W.: P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. Pharmacogenetics 6, 1–42 (1996)

[5] Otyepka, M., Skopalik, J., Anzenbacherova, E., Anzenbacher, P.: What common structural features and variations of mammalian P450s are known to date? Biochim. Biophys. Acta. 1770, 376–389 (2007)

[6] Williams, P.A., Cosme, J., Vinkovic, D.M., Ward, A., Angove, H.C., Day, P.J., Vonrhein, C., Tickle, I.J., Jhoti, H.: Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. Science 305, 683–686 (2004)
[7] Zhao, Y., Halpert, J.R.: Structure-function analysis of cytochromes P450 2B. Biochem. Biophys. Acta. 1770, 402–412 (2007)
[8] Zhou, S.F., Zhou, Z.W., Huang, M.: Polymorphisms of human cytochrome P450 2C9 and the functional relevance. Toxicology (2009), doi: 10.1016/j.tox.2009.08.013
[9] Schwarz, U.I.: Clinical relevance of genetic polymorphisms in the human CYP2C9 gene. Eur. J. Clin. Invest. 33, 23–30 (2003)
[10] Yamazaki, H., Inoue, K., Chiba, K., Ozawa, N., Kawai, T., Suzuki, Y., Goldstein, J.A., Guengerich, F.P., Shimada, T.: Comparative studies on the catalytic roles of cytochrome P450 2C9 and its Cys- and Leu-variants in the oxidation of Warfarin, flurbiprofen, and diclofenac by human liver microsomes. Biochem. Pharmacol. 56, 243–251 (1998)
[11] Yasar, U., Eliasson, E., Forslund-Bergengren, C., Tybring, G., Gadd, M., Sjoqvist, F., Dahl, M.L.: The role of CYP2C9 genotype in the metabolism of diclofenac in vivo and in vitro. Eur. J. Clin. Pharmacol. 57, 729–735 (2001)
[12] Si, D.Y., Guo, Y.J., Zhang, Y.F., Yang, L., Zhou, H., Zhong, D.F.: Identification of a novel variant CYP2C9 allele in Chinese. Pharmacogenetics 14, 465–469 (2004)
[13] Rosemary, J., Adithan, C.: The pharmacogenetics of CYP2C9 and CYP2C19: ethnic variation and clinical significance. Curr. Clin. Pharmacol. 2, 93–109 (2007)
[14] Bae, J.W., Kim, H.K., Kim, J.H., Yang, S.I., Kim, M.J., Jang, C.G., Park, Y.S., Lee, S.Y.: Allele and genotype frequencies of CYP2C9 in a Korean population. Br. J. Clin. Pharmacol. 60, 418–422 (2005)
[15] Guo, Y.J., Wang, Y., Si, D.Y., Fawcett, J.P., Zhong, D.F., Zhou, H.: Catalytic activities of human cytochrome P450 2C9*1, 2C9*3 and 2C9*13. Xenobiotica 35, 953–961 (2005)
[16] Lee, C.R., Pieper, J.A., Frye, R.F., Hinderliter, A.L., Blaisdell, J.A., Goldstein, J.A.: Differences in flurbiprofen pharmacokinetics between CYP2C9*1/*1, *1/*2, and *1/*3 genotypes. Eur. J. Clin. Pharmacol. 58(12), 791–794 (2003)
[17] Williams, P.A., Cosme, J., Ward, A., Angova, H.C., Vinkovic, D.M., Jhoti, H.: Crystal structure of human cytochrome P4502C9 with bound warfarin. Nature 424, 464–468 (2003)
[18] Wester, M.R., Yano, J.K., Schoch, G.A., Yang, C., Griffin, K.J., Stout, C.D., Johnson, E.F.: The structure of human cytochrome P4502C9 complexed with flurbiprofen at 2.0-angstrom resolution. J. Biol. Chem. 279, 35630–35637 (2004)
[19] Zhou, Y.H., Zheng, Q.C., Li, Z.S., Zhang, Y., Sun, M., Sun, C.C., Si, D., Cai, L., Guo, Y., Zhou, H.: On the human CYP2C9*13 variant activity reduction: a molecular dynamics simulation and docking study. Biochimie 88, 1457–1465 (2006)
[20] Sali, A., Blundell, T.L.: Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779–815 (1993)
[21] Canutescu, A., Shelenkov, A., Dunbrack, R.: A graph-theory algorithm for rapid protein side-chain prediction. Protein Science 12, 2001–2014 (2003)
[22] Huey, R., Morris, G.M., Olson, A.J., Goodsell, D.S.: A semiempirical free energy force field with charge-based desolvation. J. Comput. Chem. 28, 1145–1152 (2007)
[23] Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K., Olson, A.J.: Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J. Comput. Chem. 19, 1639–1662 (1998)
[24] Gotoh, O.: Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. Journal of Biological Chemistry 267(1), 83–90 (1992)

[25] Case, D.A., Darden, T.A., Cheatham, T.E., Simmerling, C.L., Wang, J., Duke, R.E., Luo, R., Crowley, M., Walker, R.C., Zhang, W., Merz, K.M., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossváry, I., Wong, K.F., Paesani, F., Vanicek, J., Wu, X., Brozell, S.R., Steinbrecher, T., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Mathews, D.H., Seetin, M.G., Sagui, C., Babin, V., Kollman, P.A.: AMBER 10. University of California, San Francisco (2008)

[26] Miyamoto, S., Kollman, P.A.: SETTLE: An Analytical Version of the SHAKE and RATTLE Algorithms for Rigid Water Models. J. Comp. Chem. 13, 952–962 (1992)

[27] Wolber, G., Langer, T.: LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J. Chem. Inf. Model. 45, 160–169 (2005)

3D Structure Modeling of a Transmembrane Protein, Fatty Acid Elongase

Sansai Chumningan[1,2], Natapol Pornputtapong[2], Kobkul Laoteng[3], Supapon Cheevadhanarak[2,4], and Chinae Thammarongtham[3]

[1] Bioinformatics and Systems Biology Program, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
[2] Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi, Bangkhuntien, Bangkok, 10150, Thailand
[3] Biochemical Engineering and Pilot Plant Research and Development Unit, National Center for Genetic Engineering and Biotechnology at King Mongkut's University of Technology, Thonburi, Bangkhuntien, Bangkok, 10150, Thailand
[4] School of Bioresources and Technology, King Mongkut's University of Technology Thonburi, Bangkhuntien, Bangkok, 10150, Thailand
sansai.chumningan@gmail.com

Abstract. Fatty acid elongase is an enzyme responsible for fatty acid chain elongation, a key step in synthesis of long chain fatty acids, including polyunsaturated fatty acids (PUFAs). Currently, the increasing demand has raised the interest in obtaining these PUFAs from alternative sources, e.g. filamentous fungi that are more economical and sustainable. To date, many research on primary structures of fatty acid elongases ELO family, including fugal elongases, revealed several conserved motifs. However, molecular mechanism for their functions is still unclear. In addition to experimental study, computational analysis of elongase structures may provide more insight into their substrate specificities and mechanisms of fatty acid chain elongation. Thus, this work proposes a 3D structural model of elongase of *Mortierella alpina* (BAF97073). This fungal elongase has been reported to be a PUFA-specific elongation enzyme. The model was built by an ab initio membrane-modeling application using ROSETTA 3.1, and was then refined by molecular dynamic simulation. The 7-transmembrane helices of the constructed model folds into an anti-parallel configuration and embeds in the lipid bilayer. The model reveals that all four conserved signature motifs of fatty acid elongase enzymes are located within the juxta-cytosolic transmembrane helix regions. This work also suggests a modeling strategy of this elongase structural model that can be applied to model other transmembrane proteins.

Keywords: PUFAs, Fatty Acid Elongase, Transmembrane Protein, *ab initio* Modeling, ROSETTA.

1 Introduction

Fatty acids, especially polyunsaturated fatty acids (PUFAs), which are primary compounds of complex lipids, play important roles for human health as they are structural

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 36–45, 2010.

components of cell membrane and precursors of biologically active molecules such as prostaglandins, thromboxanes and leukotrienes [1]. Some PUFAs are essential to human since they can not be synthesized by mammalian cells. Then they are needed to be supplementary diets. Plant oils are major sources of PUFAs. Alternatively, some fungi including *Mortierella alpina*, *Mucor rouxii* and *M. circinelloides* are able to produce several essential PUFAs. In filamentous fungi, PUFAs are synthesized by aerobic pathway, which involves an alternating series of desaturation and elongation.

Besides desaturation, fatty acid elongation is another key step for PUFA synthesis. This is responsible for the addition of two carbon units to the carboxyl end of a fatty acid chain. In eukaryotes, fatty acid elongation comprises of four distinct chemical reactions catalyzed by β-ketoacyl-CoA synthase, β-ketoacyl-CoA reductase, β-hydroyl-CoA dehydratase and enoylCoA reductase. The initial condensation reaction catalyzed by β-ketoacyl-CoA synthase (KCS) is rate-limiting step [2]. This enzyme is usually called an "elongase". It is responsible for the fatty acid substrate specificity regarding chain length and pattern of double bonds, whereas the other three enzymes of the elongase system display little or no particular substrate specificity [3]. In many organisms including filamentous fungi, although several copies of elongase encoding genes appear in individual genomes for example GLELO and MAELO from *M. alpina* [4], the enzymes of a certain organisms are different in substrate specificities. Molecular analysis of elongase proteins may gain insight in to mechanisms of fatty acid elongation. The elongation system is mainly performed in endoplasmic reticulum (ER) by membrane-bound enzymes [5]. Fungal elongases are also membrane-bound enzymes. Thus, purification of the enzymes, biochemical characterizations and also 3-dimensional structure determination of the enzymes in the elongation system by conventional techniques are difficult due to their membrane-bound nature.

There are several protein structure prediction methods that can be broadly divided into three categories: 1) homology modeling, 2) threading or fold recognition, and 3) *ab initio*. Fundamentally, the classification reflects the degree to which different methods utilize the information content available from the known structure database. Homology modeling has been immensely successful with soluble proteins [6]. These methods require a homologous protein template based on evolutionary of target and template sequences with percent identity of two sequences basically more than 30%. Nowadays, only few representative atomic-resolution structures of transmembrane proteins are available. Homology modeling does not seem to be a general-purpose approach for transmembrane protein structure modeling. On the other hand, membrane proteins present much higher uniformity of secondary structure (mostly alpha-helical bundles) than soluble protein, and are highly constrained in their conformation because of the presence of membrane lipid bilayer. Thus, fold-recognition method that bases on a principle that there are limited number of fold of protein in nature and many different remotely homologous protein sequence tent to have similar structure may be appropriated for membrane protein prediction [7]. Moreover, it could therefore be expected that de novo or *ab initio* structure prediction, whereby the membrane protein structure is predicted without requirement of homology with other proteins. This method may be a feasible goal for protein with the slightest homology protein in know structure database.

Recently, the computational method has become an alternative method to generate and analyze 3-dimensional models of a number of proteins including those of

VLCFAs elongase family proteins in *Arabidopsis thaliana* [8]. In order to accomplish structural analysis of fatty acid elongases in oleaginous fungi, structure modeling would be required for the first step. In the work a reliable structural model GLELOp of *M. alpine* was constructed.

2 Material and Methods

2.1 Sequence

The amino acid sequence of GLELOp elongase from *M. alpina* (BAF97073) was retrieved from GenBank. The sequence comprises 318 residues.

2.2 Transmembrane Topology Prediction

In order to model transmembrane protein structures, determining their topologies is a key preliminary step to model their structures. Transmembrane regions of elongase were predicted by following tools; TMHMM [9], Phobius [10], TOPpred [11], TMpred [12], SOSUI [13], Octopus [14], and PHDhtm [15]. Based on MetaTM [16], consensus transmembrane regions among predicted results obtained these selected tools was then generated, as TMcons, according to 2 criteria 1) the amino acid residues to be included in a particular transmembrane region had to be predicted, being in such transmembrane region, by at least 4 of 7 tools 2) the TMcons based transmembrane regions have to be 18-24 amino acid residues in length.

2.3 Template-Based Modeling

2.3.1 Template Selection

There are 2 approaches in template selection. 1) Homology searching, this approach searches for suitable homologues in protein structure database. BLAST or Basic Local Alignment Search Tool was utilized for this task. The protein BLAST tool was performed to search for homologues in Protein Data Bank (PDB). To obtain a reliable model, sequence identity of 30% or above between target and template should be considered. 2) Fold recognition (threading), is an alternative approach for finding a template based on minimum folding energy concept. The principle of this method is that there are limited numbers of protein fold in nature, thus many different remotely homologous protein sequences adopt remarkably similar structures. Phyre or Protein Homology/analogY Recognition Engine was used for this step. The algorithm of this selected tool is profile-profile matching which can search template less than 20 percents sequence identity [8].

2.3.2 Homology Modeling

After an appropriated template was obtained, the target-template alignment, the key part of modeling, was required. The alignment was manually adjusted based on secondary structure, transmembrane prediction results for this case. Then, the alignment was used as an input of model building. The model building part was performed by

Modeller program [17]. The best model was selected from first ranked list by DOPE score of all built models. Loop regions in the built model were further refined by loop refinement modeling module.

2.4 *ab initio* Modeling

ab initio modeling is an alternative method to generate 3D models of protein structures. Transmembrane topology obtained from TMcons was used to set initial membrane normal and membrane center vectors in membrane *ab initio* modeling application that implemented in ROSETTA 3.1 [18]. In order to obtain the most reliable structural models, the cycle for *ab initio* modeling or repeating the random formation of fragments was set for 1, 3, 5, 10, and 100 cycles.

2.5 Molecular Dynamics Simulation

In order to refine the built model, Molecular dynamics simulation was performed based on energy minimization by using NAMD program [24]. The protein model was placed in a native-like membrane environment, POPC lipid bilayer. The simulation time step was set to 2 fs/step. RMSDs of the protein model were calculated by using the structure of the first-frame as a reference.

2.6 Model Quality Assessment

Evaluation of model quality was conducted on SWISS-MODEL server [19]. The following methods, Anolea [20], DFire [21], and MolProbity [22] were performed. The helical wheel by HELIQUEST [23] also was used to check rearrangement of residues in transmembrane helices.

3 Result and Discussion

3.1 Secondary Structure of GLELOp Sequence

Although the results of transmembrane topology predictions for GLELOp by the seven selected tools were not identical, however they agreed with each others. The topology of consensus transmembrane regions of GLELOp generated by TMcons is shown in Fig. 1. The result shows that GLELOp is composed of seven transmembrane helices embedded in lipid bilayer.

Since transmembrane topology is required for transmembrane protein 3D structure modeling, accuracy of topology prediction should be assessed. In order to check the accuracy of TMcons approach, transmembrane regions predicted by TMcons of Sur4p (encoded by *ELO2*), an elongase of *S. cerevisiae*, was compared with transmembrane determined regions by dual topology report experiment (DTR) [25]. The overall topology predicted by TMcons agree with the one determined by DTR as shown in Fig. 2. This result demonstrates that TMcons method can predict the reliable transmembrane regions.

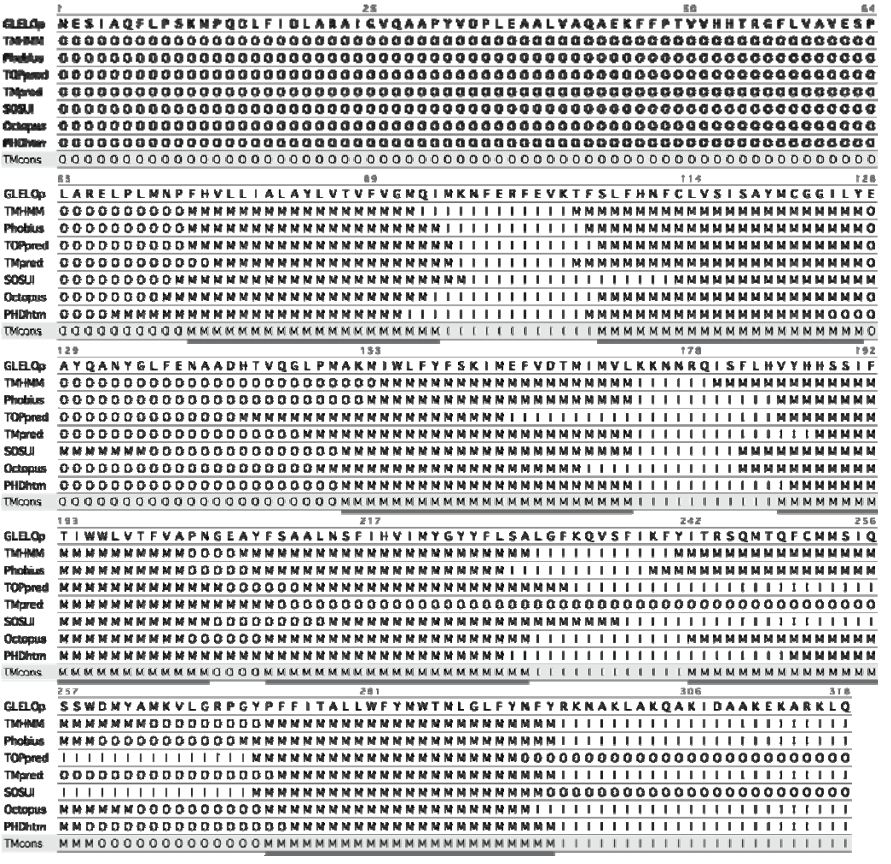


Fig. 1. Transmembrane consensus result of GLELOp from TMcons method, the first line is amino acid sequence, next 7 lines are predictions, and last line is transmembrane consensus result (TMcons). The letters O, I, and M stand for residues located in luminal side, cytosolic side, and transmembrane region, respectively.

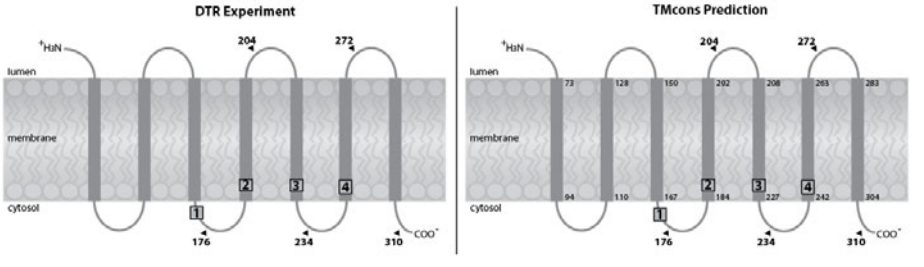


Fig. 2. Comparing Sur4p topology from Dual Topology Report (DTR) experiment with TMcons prediction. The numbered boxes correspond to the approximate positions of the ELO signature motifs. The labeled 1, 2, 3, and 4 are KXXEXXDT, HXXHH, HXXMYXYY, and TXXQXXQ, respectively. The numbers indicate the amino acid positions at which the dual topology reporter (DTR) was inserted.

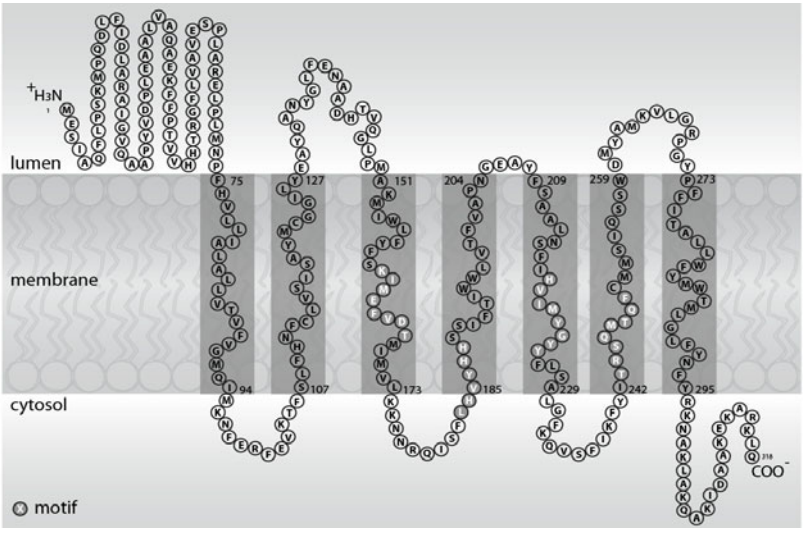


Fig. 3. GLELOp topology depicted based on TMcons

3.2 Modeling

3.2.1 Template-Based Modeling

To select the homology-modeling template, GLELOp sequence was searched by BLAST against PDB database. The GLELOp sequence did not significantly match with any sequences in the database at the time it was analyzed (sequence identity less than 20 percent with short coverage region). Thus, homology-modeling approach was not suitable for modeling of GLELOp. Alternative technique, Fold-recognition method was performed by Phyre server. According to the Table 1, the GLELOp sequence matched to certain known structure protein is 1U19 with the highest estimated precision at 95% and percent identity at 10%.

GLELOp model was built by using atomic coordinate of 1U19 chain-A crystal structure as template. The constructed model was 7-transmembrane helix conformation. For N terminal region, approximately 70 amino acid residues, the model reveals random coil conformation, protruding out of lipid bilayer membrane. Modeling of this region was likely low accuracy as the input alignment illustrates secondary structure element of random coils with unalignable sequences. For template-based method which is based on sequence similarity between target and template, low quality alignment would give unreliable model. For the constructed model, both primary sequence and secondary structure between target and template, particularly in N terminal region, were not well aligned. This is because the available known transmembrane protein structures are limited. An alternative modeling method would be considered.

3.2.2 *ab initio* Modeling

ROSETTA 3.1 was exploited to build a GLELOp structural model by *ab initio* modeling based on consensus transmembrane topology. The transmembrane helices of constructed model fold into anti-parallel configuration. The N terminal was located on

Table 1. Fold-recognition hits by using GLELOp as query

PDB ID	Description	Estimated Precision	% ID
1U19 Chain A	Rhodopsin, Signaling protein, SCOP class 6	95%	10%
2R9R Chain B	Voltage-dependent K+ channel, Membrane protein, SCOP class 3	90%	7%
2R4R Chain A	Adrenocepter, Signaling protein,	90%	7%
1M56 Chain C	Cytochrom c oxidases, Oxidoreductase, SCOP class 6	90%	9%
1FFT Chain A	Ubiquinol oxidase, Oxidoreductase, SCOP class 6	70%	9%

Table remarks:
1. Description containing name of protein, PDB classification, and SCOP class.
2. SCOP class 3 is Alpha and beta protein (a/b)
2. SCOP class 6 is Membrane and cell surface proteins and peptides

opposite site of the C terminal. This agrees with the predicted 2D topology. The model reveals that all four conserved signature motifs of elongase enzymes are located within juxta-cytosolic transmembrane helix regions. Optimization of *ab initio* cycle numbers (1, 3, 5, 10, and 100) suggests that modeling with the larger number of cycle would give more chance to obtain a model with a reliable conformation. Notably, this option will take multiple computation times of 1 cycle. In case of no suitable template could be obtained from either homology searching or fold recognition, *ab initio* modeling using ROSETTA 3.1 is an alternative choice for modeling transmembrane protein. Nevertheless, correct transmembrane topology is required. Experimental determination of transmembrane secondary structure will be greatly beneficial. Otherwise at least the most reliable predicted transmembrane topology is needed.

The constructed model was refined in lipid environment by using molecular dynamics simulation. After 4 ns, RMSD was rather steady with a little fluctuation within a range of 0.75 Å (Fig. 4), suggesting that the model reaches to equilibrium. At this step, the obtained model was ready for further study.

The Ramachandran plot analysis reveals that the main-chain conformations for 97.39% of amino acid residues are within the most favored or allowed regions, indicating that constructed GLELOp model is of good quality. MolProbity score given

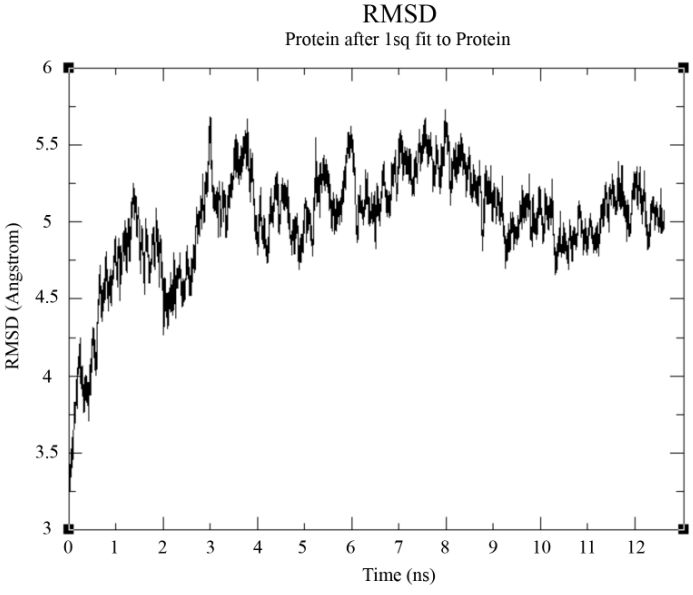


Fig 4. RMSD (y axial) VS Time (x axial) of GLELOp model during MDs

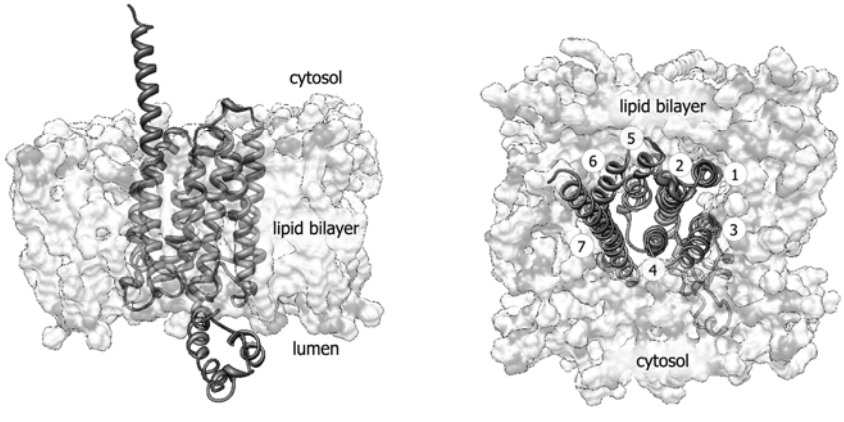


Fig. 5. 3D structural model of M. alpina elongase GLELOp in lipid bilayer environment, transmembrane helices are numbered from N terminal.

96th from 100th percentile that referred to the best among structures of comparable resolution. To assess packing quality of each residue (local assessment), ANOLEA result reveals negative energy values for most amino acid residues in the model indicating favorable energy environment. The global model quality estimation by DFire or all-atom distance-dependent statistical potential method shows energy of -436.54, suggesting that the model is close to the native conformation. The helical wheel of the model exhibits most of the polar amino acid residues distributed in the inner part of

the protein (data not shown). These suggest that the model folds into a reasonable conformation and most amino acid residues are in their nature configuration. The final GLELOp model (Fig. 5) composes of 7-transmembrane helices folded into anti-parallel configuration. The N terminal was located on lumen while C terminal located on cytosol. There are 2 non-membrane-alpha-helix fold out sites of the lipid-bilayer on the luminal site.

4 Conclusion

This work presented a 3D structural model of *M. alpina* elongase GLELOp constructed by an ab initio technique. The model was refined by molecular dynamic simulation in a lipid bilayer environment. The quality of the model is satisfactory as indicated by several model quality assessments, including Ramachandran plot, packing quality, and helical wheel analysis. The modeling strategy of fatty acid elongase protein model can be applied to other transmembrane proteins modeling in case there is a lack of a suitable template structure. Besides an experimental structure, the constructed model provides an alternative choice to explore structural characteristic of a fatty acid elongase at the molecular level. In particular, for studying enzyme and substrate interaction, the proposed model provides a platform for exploring the residues that play important roles in the catalysis of elongase with their substrates.

Acknowledgement

Chumningan, S. would like to thank National Center for Genetic Engineering and Biotechnology, Thailand (BIOTEC), and King Mongkut's University of Technology Thonburi for the scholarship of Bioinformatics program.

References

1. Wettstein-Knowles, P.M.: Waxes: Chemistry, Molecular Biology and Function. In: Hamilton, R.J. (ed.), vol. 6, pp. 91–130. Oily Press, Dundee (1995)
2. Bernert, J.T., Sprecher, H.: An analysis of partial reactions in the overall chain elongation of saturated and unsaturated fatty acids by rat liver microsomes. J. Biol. Chem. 252(19), 6736–6744 (1977)
3. Cinti, D.L., Cook, L., Nagi, M.N., Suneja, S.K.: The fatty acid chain elongation system of mammalian endoplasmic reticulum. Prog. Lipid Res. 31(1), 1–51 (1992)
4. Parker-Barnes, J.M., Das, T., Bobik, E., Leonard, A.E., Thurmond, J.M., Chaung, L.T., Huang, Y.S., Mukerji, P.: Identification and characterization of an enzyme involved in the elongation of n-6 and n-3 polyunsaturated fatty acids. Proc. Natl. Acad. Sci. USA 97, 8284–8289 (2000)
5. Nugteren, D.H.: The enzymic chain elongation of fatty acids by rat-liver microsomes. Biochim. Biophys. Acta. 106, 280–90 (1965)
6. Petrey, D., Honing, B.: Protein structure prediction: inroads to biology. Mol. Cell. 20, 811–819 (2005)

7. Kelley, L.A., Sternberg, M.J.E.: Protein structure prediction on the Web a case study using the Phyre server. Nature protocol 4(3), 363–371 (2009)
8. Joubés, J., Raffaele, S., Bourdenx, B., Garcia, C., Laroche-Traineau, J., Moreau, P., Domergue, F., Lessire, R.: The VLCFA elongase gene family in Arabidopsis thaliana: phylogenetic analysis, 3D modelling and expression profiling. Plant Mol. Biol. 67, 547–566 (2008)
9. Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L.L.: Predicting transmembrane protein topology with a hidden Markov model Application to complete genomes. J. Mol. Biol. 305(3), 567–580 (2001)
10. Käll, L., Krogh, A., Sonnhammer, E.L.L.: A Combined Transmembrane Topology and Signal Peptide Prediction Method. J. Mol. Biol. 338(5), 1027–1036 (2004)
11. von Heijne, G.: Membrane Protein Structure Prediction Hydrophobicity Analysis and the Positive Inside Rule. J. Mol. Biol. 225, 487–49 (1992)
12. Hofmann, K., Stoffel, W.: TMbase - A database of membrane spanning proteins segments. Biological Chemistry Hoppe-Seyler 374, 166 (1993)
13. Mitaku, S., Hirokawa, T., Tsuji, T.: Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. Bioinformatics 18, 608–616 (2002)
14. Viklund, H., Elofsson, A.: Improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. Bioinformatics 24(15), 1662–1668 (2008)
15. Rost, B., Casadio, R., Fariselli, P.: Refining neural network predictions for helical transmembrane proteins by dynamic programming. In: Proc. Int. Conf. Intell. Syst. Mol. Biol., vol. 4, pp. 192–200 (1996)
16. Klammer, M., Messina, D.N., Schmitt, T., Sonnhammer, E.L.: MetaTM - a consensus method for transmembrane protein topology prediction. BMC Bioinformatics 10(314) (2009)
17. Eswar, N., Marti-Renom, M.A., Webb, B., Madhusudhan, M.S., Eramian, D., Shen, M., Pieper, U., Sali, A.: Comparative Protein Structure Modeling With Modeller in Current Protocols in Bioinformatics. John Wiley & Sons, Chichester (2006)
18. Chivian, D., Kim, D.E., Malmstrom, L., Schonbrun, J., Rohl, C.A., Baker, D.: Prediction of CASP6 structures using automated Robetta protocols. Proteins 61(7), 157–166 (2005)
19. Rost, B., Casadio, R., Fariselli, P.: Refining neural network predictions for helical transmembrane proteins by dynamic programming. In: Proc. Int. Conf. Intell. Syst. Mol. Biol., vol. 4, pp. 192–200 (1996)
20. Melo, F., Feytmans, E.: Assessing protein structures with a non-local atomic interaction energy. J. Mol. Biol. 277(5), 1141–1152 (1998)
21. Zhou, H., Zhou, Y.: Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. 11, 2714–2726 (2002)
22. Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B., Snoeyink, J., Richardson, J.S., Richardson, D.C.: MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Research 35 (2007)
23. Gautier, R., Douguet, D., Antonny, B., Drin, G.: HELIQUEST: a web server to screen sequences with specific α-helical properties. Bioinformatics 24(18), 2101–2102 (2008)
24. Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L., Schulten, K.: Scalable molecular dynamics with NAMD. Journal of Computational Chemistry 26, 1781–1802 (2005)
25. Denic, V., Weissman, J.S.: A Molecular Caliper Mechanism for Determining Very Long-Chain Fatty Acid Length. Cell. 130(4), 663–677 (2007)

Sequential Application of Feature Selection and Extraction for Predicting Breast Cancer Aggressiveness

Jonatan Taminau[1], Stijn Meganck[1], Cosmin Lazar[1], David Y. Weiss-Solis[2], Alain Coletta[2], Nic Walker[2], Hugues Bersini[2], and Ann Nowé[1]

[1] Computational Modeling Lab, Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussels, Belgium
http://como.vub.ac.be
{jonatan.taminau,stijn.meganck,cosmin.lazar,ann.nowe}@vub.ac.be
[2] IRIDIA, Université Libre de Bruxelles
Avenue Franklin D. Roosevelt 50, 1050 Brussels, Belgium
http://iridia.ulb.ac.be
{david.weiss,alain.coletta,nic.walker,hugues.bersini}@ulb.ac.be

Abstract. Breast cancer is a heterogenous disease with a large variance in prognosis of patients. It is hard to identify patients who would need adjuvant chemotherapy to survive. Using microarray based technology and various feature selection techniques, a number of prognostic gene expression signatures have been proposed recently. It has been shown that these signatures outperform traditional clinical guidelines for estimating prognosis. This paper studies the applicability of state-of-the-art feature extraction methods together with feature selection methods to develop more powerful prognosis estimators. Feature selection is used to remove features not related with the clinical issue investigated. If the resulted dataset is still described by a high number of probes, feature extraction methods can be applied to further reduce the dimension of the data set. In addition we derived six new signatures using three independent data sets, containing in total 610 samples.

Additional information:
http://como.vub.ac.be/~jtaminau/CSBio2010/

Keywords: Breast Cancer Signatures, Feature Selection, Feature Extraction.

1 Introduction

Breast cancer is a very heterogenous disease and it still remains a challenge to distinguish patients who would need adjuvant chemotherapy from those who don't need it. Using microarray based technology, a number of prognostic gene expression signatures have been proposed recently [1,2,3,4] to guide the clinicians with the selection of patients who should receive such treatments. Those signatures,

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 46–57, 2010.

all developed with a different approach and on different data sets, seem to have similar prognostic performance [5,4] despite their limited overlap in genes.

In this paper six new gene signatures are proposed, all based on correlation based analysis with respect to the survival outcome of breast cancer patients. These signatures were derived using several completely independent studies, aiming to increase the generality of the results, and were validated on another independent test set.

A single microarray can contain tens of thousands of probes; the good part in collecting such rich data sets is the fact that one experiment can accomplish many genetic tests in parallel. However, only few probes are relevant for particular clinical issues and their identification is quite difficult even for clinicians or biologists. The analysis of such big data sets is in general subjected to the well known *curse of dimensionality* or *the empty space phenomenon* [6,7]. In order to deal with this challenge, two main strategies can be used: feature selection and feature extraction, each one of them with their specific algorithms. Previous works on the analysis of microarray data have focused mainly on feature selection methods. Basically a number of genes are selected which are meaningful with respect to some clinical features such as disease outcome [1,2] or histological grade [3], but the dimension of the resulting set of genes is still high, typically around 100 probes or genes. On the other hand, feature extraction methods are mainly used to improve the results of the analysis in situations where one deals with high dimensional data sets and no *a priori* information about the data is known [8]. The result of feature extraction methods is a new representation of the original data in a compressed form, by minimizing the loss of information and by preserving the distribution of the original data.

Sequential application of feature selection and feature extraction methods can also be used for dimension reduction especially when the selection of features still results in a high dimensional data set. In a first instance, feature selection is used to remove those features which are not related with the clinical issue investigated but also features carrying low information (features with low variance). If the resulted data set is still described by a high number of features, extraction methods can be applied to further reduce the dimension of the data set. The increase in prognostic accuracy after applying feature extraction methods is shown both in combination with existing signatures and with our six newly proposed ones.

2 Methods

In this section we describe the data, the different gene signatures and the feature extraction methods used in this study.

2.1 Data

We used a collection of three different and independent breast cancer data sets which were retrieved from the InSilico DB[1] in order to consistently pre-process

[1] `http://insilico.ulb.ac.be/` (manuscript in progress)

Table 1. Different microarray data sets used in this study. The first three data sets are the *training* data from which the CoMo signatures are derived. The fourth data set can be seen as an independent *test* set for validation.

GEO Acc. *InSilicoDB Acc.*	*# Samples*	*Author*	*Ref.*
GSE1456	159	Pawitan	[10]
GSE3494	251	Miller	[11]
GSE11121	200	Schmidt	[12]
GSE7390	198	Desmedt	[9]

and annotate them. As our independent validation set we selected the TRANSBIG data set because it was already used before to test and compare different breast cancer gene expression signatures [5]. For consistency we only selected the Affymetrix study (TBVDX serie [9]).

All four studies, listed in Table 2.1, were computed on the Affymetrix HG-U133 research GeneChip™ and we checked for duplicated samples between all studies by looking at the correlation across samples. No sample pairs with a correlation higher than 0.95 were observed, ensuring the independence of all the different data sets.

2.2 Existing Gene Signatures

Several different prognostic gene signatures for breast cancer aggressiveness have been proposed in recent years [1,2,3]. They have been shown to be advantageous compared to standard clinical guidelines and could therefore reduce the number of patients subject to adjuvant chemotherapy.

Gene70. In [1], genes were identified that were differentially expressed between two groups of patients with differing survival. They used a cut-off of 5 years after diagnosis to check whether someone had developed distant metastasis or not, and divided the patients in two groups accordingly. They used microarrays of Agilent technology and identified a set of 70 relevant probes.

Gene76. With a similar method the Erasmus Medical Center, Netherlands and Veridex LLC, USA identified a set of 76 probes [2] using Affymetrix microarrays. These genes were used to build a risk prediction model taking into account the difference between estrogen receptor positive (ER+) and negative (ER-) patients.

GGI. In [3], they proposed a gene signature that is predictive for the histological grade of the tumor. Since the histological grade is highly correlated with predictive outcome, this can be used as a prognostic signature. Probes were selected based on their ranking with respect to their differential expression between histological grade 1 and 3. GGI was also derived using Affymetrix microarrays.

2.3 CoMo-Signatures

The existing gene signatures mentioned above have few overlap between them and the number of selected genes is quite high (tens of probes). Moreover, these signatures have been selected from a single microarray data set and tested in some cases only on few samples which limits their power to generalize. With the public availability of well annotated large breast cancer data sets in the InSilico DB, new strategies for discovering gene signatures can be executed. In this paper we derive new gene signatures using information from three public microarray data sets and validate them on a fourth independent data set. Correlation based techniques were used to develop these signatures as follows.

For each of the three training data sets individually, the correlation of each gene with the survival time was computed. All genes were ranked according to their correlation and six gene signatures were created as follows:

Intersect Low: The intersection of the top 500 negatively correlated probes per individual training data set (8 probes).
Intersect High: The intersection of the top 500 positively correlated probes per individual training data set (13 probes).
Intersect Both: The union of *Intersect low* and *Intersect high*.
Multiple Low: Any probe that appeared in the top 500 negatively correlated probes in at least two training data sets (256 probes).
Multiple High: Any probe that appeared in the top 500 positively correlated probes in at least two training data sets (152 probes).
Multiple Both: The union of *Multiple low* and *Multiple high*.

All details of the six gene signatures can be found in additional information. The use of three completely independent data sets should offer more robust signatures because it decreases the risk of overfitting by combining different sources of the same signals. However, missing or incorrect probes from a defective study will not show up in a strict intersection, regardless their overall importance in the other studies. Therefore, we developed signatures with two different strategies, the probes in the *Intersect* signatures capture relevant information in all data sets and are therefore assumed to be very important for estimating survival. The *Multiple* signatures are less prone to the absence of relevant probes in a specific study but the size of these signatures grows rapidly.

2.4 Feature Extraction

Following, we describe two different simple feature extraction methods we used in this paper. We focused on PCA and ICA since they are both well understood techniques which have proven to be useful in many applications. PCA is perhaps the most popular feature extraction technique used in a wide range of applications such as face recognition [13], multivariate image segmentation or multidimensional data clustering. On the other side, ICA has only been recently used as a feature extraction tool: it has been successfully applied in applications such as target detection from multi/hyperspectral remote sensing images

[14] and [15] and more recently in bioinformatics [16]. In [17] PCA and ICA are investigated as feature extraction tools for brain tumor classification.

Principal Component Analysis (PCA). PCA consists in finding the eigenvectors as well as the eigenvalues of the covariance matrix of the original gene-expression matrix X [18]. Then the principal components are obtained by projecting every sample from X on the eigenvectors with the highest eigenvalues.

$$S = BX \tag{1}$$

where B is the eigenvectors matrix of the covariance matrix of X and S are the principal components. The dimension reduction is performed by choosing those eigenvectors whose corresponding eigenvalues are greater than a fixed threshold. A general and comprehensive description of the method can be found in [18].

Independent Component Analysis (ICA). ICA is similar to PCA. The orthogonality assumption, inherent to the eigenvectors, is replaced with that of independence between the newly derived features and basis vectors [19]. It can also be employed as a dimension reduction method and it sometimes embeds PCA as a preprocessing step. The independence of two random variables is expressed in various ways explaining the big number of algorithms developed for ICA. In our simulations we used the FastICA algorithm [19]. In [16], the use of ICA is motivated by the fact that the expression of genes is the result of a specific combination of cellular variables. ICA has been used to derive a linear model based on hidden variables called expression modes and the expression of each gene is a linear function of those modes. A recent survey of the use of ICA for feature extraction from microarray data can be found in [20].

2.5 Combining Feature Selection and Feature Extraction

Our goal is to look at the combination of feature selection and feature extraction. Although there is a lot of work on using feature extraction methods on entire microarray data sets [16,21], we believe that combining both methods will have a positive impact on the overall results. We motivate our believe by the fact that a microarray data set encodes rich and various information about many aspects of the cell such as functions, states or processes taking place inside it, some of them being more dominant than another (for instance ER status). This is the reason why a first selection of probes, the most relevant with respect to a particular issue in question (in our case the breast cancer aggressiveness) is mandatory.

Extracted features are harder to interpret in the sense that a particular value of a probe is a linear (or non linear) combination of some basis vectors resulted in the feature extraction process. This is the main drawback of these methods, in the sense that a particular value of a probe is a linear (or non linear) combination of some basis vectors resulted in the feature extraction process. By combining both approaches we render interpretation easier since the extracted features depend on a limited number of probes.

In our approach, we start by only keeping probes inside the signature (both the existing as the CoMo ones) and calculate a set of features from this data set by using both PCA and ICA, resulting in two extra transformed data sets per training set and also for the TRANSBIG test data set. On every data set we then perform a k-means clustering. The assumption is that the set of probes will naturally cluster samples based on their expected survival time. Furthermore, we expect that reducing the noise and dimension by feature extraction might help improve these results. We then perform Kaplan-Meier and Cox proportional hazard ratio analysis for the groups that were identified by the clustering algorithm.

3 Results

In this section we describe our experimental setup and the survival analysis results on all training data sets as well as on the independent TRANSBIG test data set.

3.1 Experimental Setup

All code was written in $\boldsymbol{R}$[2]. All data sets were downloaded from the InSilico DB, which automatically retrieved the original CEL files, performed RMA for each individual data set and normalized them between data sets using Batch Mean Centering [22].

Feature selection, which in this case amounts to probe selection, for the CoMo-signatures was done as explained above. For both the *GGI* and *Gene*76 signatures the *genefu* package[3] was used to identify the probes available in $data(sig.ggi[\text{“}probe\text{”}])$ and $data(sig.gene76[\text{“}probe\text{”}])$ respectively. We did not include the *Gene*70 signature in our analysis since this study was not performed on Affymetrix arrays.

The basic *prcomp* function and $fastICA$ package[4] were used to perform PCA and ICA feature extraction respectively. The number of components for PCA was chosen by removing those eigenvectors whose standard deviation was less than a fifth of that of the first eigenvector. The number of components of ICA was chosen to be the same as that of PCA.

For the CoMo-signatures samples were divided into two groups based on k-means clustering with correlation as a distance metric by using the $Kmeans$ function of the *amap* package[5] with five random restarts. For GGI and Gene76 a similar division was done and also one using the risk score as provided in the *genefu* package.

3.2 Analysis

We performed a cox proportional hazard ratio (HR) analysis on all data sets for each of the six new gene signatures. We performed this analysis on all training

[2] www.r-project.org/
[3] http://cran.r-project.org/web/packages/genefu/
[4] http://cran.r-project.org/web/packages/fastICA/index.html
[5] http://cran.r-project.org/web/packages/amap/index.html

data sets as well to see whether the signatures still captured essential information of the data sets that they were derived from. These results however can give an overly positive view as all information of these data sets was used in the derivation of the signatures.

As can be seen in Table 3.2, for the training data sets, in all but one case the best results (higher HR) are obtained after sequentially performing feature selection and feature extraction. Full details, including 95% confidence intervals and p-values can be found in additional material.

Clearly the highest HR is obtained when using the signatures composed of probes which are both negative and positively correlated with survival which shows that these two sets are complementary for the estimation of aggressiveness.

On the TRANSBIG data set, we see similarly that feature extraction aids the division in good and bad prognosis classes, thereby affirming the results on the training data sets. The improvements are however not always as significant. Applying our strategy on both GGI and Gene76, see Table 3.2, shows similar

Table 2. Hazard ratio's using the different gene signatures on both training and test data sets. The bold numbers indicate the best (highest HR) per data set/gene signature combination.

Signature	*Data set*	*Original*	*PCA*	*ICA*
Intersect low	GSE1456	3.498072	**5.025954**	2.953581
	GSE3494	1.483355	**2.548084**	1.303337
	GSE11121	1.717594	**3.265009**	1.281384
Intersect high	GSE1456	2.803051	**5.197641**	3.061206
	GSE3494	2.086769	**3.082308**	2.225119
	GSE11121	1.977458	**2.633595**	1.987610
Intersect both	GSE1456	4.555611	**6.287580**	4.811514
	GSE3494	2.582587	2.981011	**3.006106**
	GSE11121	3.192547	2.072678	**3.742142**
Multiple low	GSE1456	2.884515	**6.421448**	4.398484
	GSE3494	2.450127	**2.565744**	2.373649
	GSE11121	2.113939	**2.863432**	1.943166
Multiple high	GSE1456	3.836090	**7.170804**	3.913145
	GSE3494	**3.597214**	3.435415	3.263882
	GSE11121	2.397085	1.935239	**2.705435**
Multiple both	GSE1456	8.054912	7.564513	**8.314927**
	GSE3494	2.950295	2.628732	**3.089008**
	GSE11121	3.274170	2.870842	**3.327138**
Intersect low	TRANSBIG	2.184360	**2.767724**	1.219542
Intersect high	TRANSBIG	1.774095	**2.065701**	1.610812
Intersect both	TRANSBIG	2.737137	**2.870929**	2.793816
Multiple low	TRANSBIG	4.968810	4.904974	**5.047237**
Multiple high	TRANSBIG	1.037704	**1.910830**	1.432156
Multiple both	TRANSBIG	3.693934	3.041489	**4.947116**

Table 3. Hazard ratio's using GGI and Gene76 on the test data set. The bold numbers indicate the best (highest HR) per data set/gene signature combination using the k-means approach. The first column shows the HRs based on the signature score and risk classification from GGI and Gene76 respectively.

Signature	*Risk Factor*	*Original*	*PCA*	*ICA*
GGI	*5.091227*	**3.816467**	3.414592	1.438335
Gene76	*5.057281*	2.245084	**3.512873**	3.210784

results although the division based on k-means for GGI outperforms those with feature extraction. With several of the CoMo-signatures we are able to obtain higher HRs than with GGI and Gene76 with our approach. However, none of the results can improve the HR of GGI and Gene76 based on their own risk score.

We also performed a Kaplan-Meier (KM) analysis for each data set/gene signature combination. We show the results for the training sets for the *Intersect low* signature in Figure 1, the other figures can be found online as additional material. Results of both *Intersect low* and *Multiple low* on the TRANSBIG data set can be found in Figure 2. These figures map the corresponding results that were discussed previously for the hazard ratios in Table 3.2.

We created a single sample predictor (SSP) for each of the CoMo-signatures. The SSP was created based on the merged data set combining the three test data sets using Batch Mean Centering [22] by taking the mean for each probe/component for samples with metastasis and samples without. We then assigned each sample in the TRANSBIG data set to a group based on the highest correlation with each of the SSPs. The resulting HRs are given in Table 3.2 and corresponding KM plots can be found in additional information.

The SSP based on the *Multiple Low* signature outperforms both the existing GGI and Gene76 signatures and all the divisions based on k-means of all CoMo-signatures on the TRANSBIG data set.

3.3 Discussion on Feature Extraction Methods

PCA tends to work better in most cases than any of the other. ICA seems to perform poorly in general but works good on the largest collection of genes.

Table 4. Hazard ratios for the TRANSBIG data set based on the SSP created on the merged data set of all three training data sets for all CoMo-signatures

Feat. Extr. Method	*Intersect Low*	*Intersect High*	*Intersect Both*	*Multiple Low*	*Multiple High*	*Multiple Both*
None	2.726886	1.719567	2.489592	**8.988187**	2.728809	3.379779
PCA	2.685891	1.898967	2.328386	**6.276682**	2.345202	3.757589
ICA	1.642725	1.841632	2.727805	**7.193519**	2.855434	3.757589

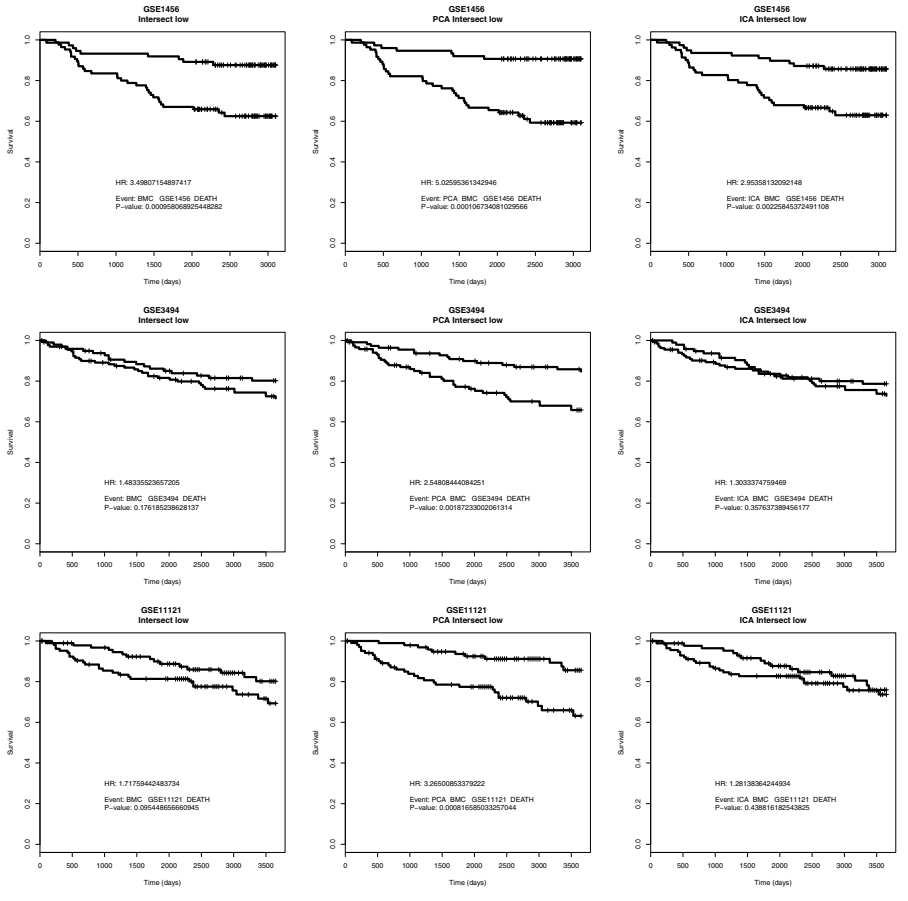


Fig. 1. Comparison of Kaplan-Meier analysis for *Intersect Low* signature. Columns indicate the different feature extraction algorithm used: *none, PCA*, and *ICA* respectively. Rows correspond to the different training data sets: *GSE1456*, *GSE3494* and *GSE11121* respectively.

While a PCA decomposition preserves in the best way possible the original data by implicitly imposing the minimum loss of information, the independence constraints imposed by ICA might give completely misleading results depending on the application. This is why it is not surprising that PCA performs in general better than ICA in this particular application. It is not yet fully understood why feature selection + ICA performs better for gene signatures containing a higher number of probes. One explanation could be the fact that gene signatures containing few probes have relatively independent features and thus the ICA can not improve the results significantly; more than that, removing some features might result in a significant loss of information. On the other hand, the large gene signatures are likely to have several significant correlated features and thus ICA can make a significant improvement.

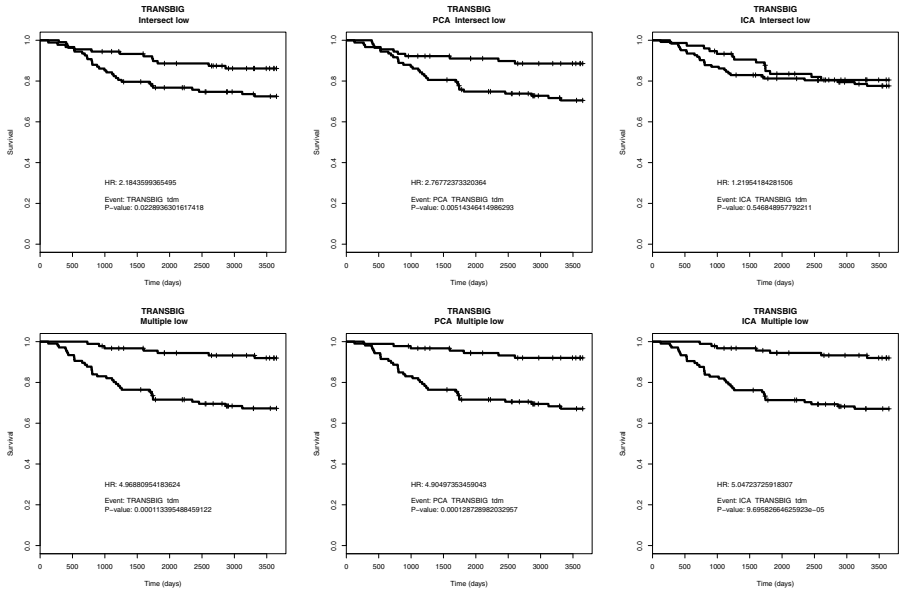


Fig. 2. Comparison of Kaplan-Meier analysis for several CoMo-signatures on the TRANSBIG data set. Columns indicate the different feature extraction algorithm used: *none, PCA*, and *ICA* respectively. Rows correspond to the different signatures: *Intersect Low* and *Multiple Low* respectively.

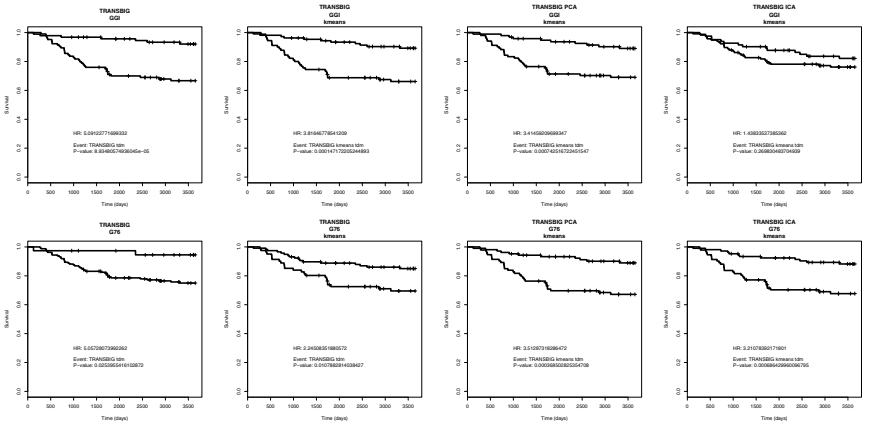


Fig. 3. Comparison of Kaplan-Meier analysis for the TRANSBIG data set for both the GGI and Gene76 signatures. The first column uses the risk factor of the original signals to form risk groups. The following three columns indicate the different feature extraction algorithm used: *none, PCA*, and *ICA* respectively. Rows correspond to the different signatures: *GGI* and *Gene76* respectively.

4 Conclusions

Here we investigate the beneficial aspects of jointly use feature selection and feature extraction methods for breast cancer aggressiveness prediction. For feature selection we have used already existing gene signatures (GGI and Gene76) derived to predict breast cancer aggressiveness but we also derived our own gene signatures from three independent data sets. Further, these signatures have been used as inputs for feature extraction (PCA and ICA) aiming the dimension reduction and prediction improvement. Results show that for this application PCA applied on gene signatures improves the breast cancer aggressiveness prediction in most of the cases.

The promising results of our newly created signatures encourage us to further investigate the use of information of multiple studies in order to derive consistent, robust and predictive signatures.

Acknowledgements

This research is partially funded by the Institute for the encouragement of Scientific Research and Innovation of Brussels (IRSIB).

References

1. van 't Veer, L.J., Dai, H., van de Vijver, M.J., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871), 530–536 (2002)
2. Wang, Y., Klijn, J.G.M., Zhang, Y., et al.: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365(9460), 671–679 (2005)
3. Sotiriou, C., Wirapati, P., Loi, S., et al.: Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J. Natl. Cancer Inst. 98(4), 262–272 (2006)
4. Korkola, J.E., Blaveri, E., DeVries, S., et al.: Identification of a robust gene signature that predicts breast cancer outcome in independent data sets. BMC Cancer 7, 61 (2007)
5. Haibe-Kains, B., Desmedt, C., Piette, F., et al.: Comparison of prognostic gene expression signatures for breast cancer. BMC Genomics 9, 394 (2008)
6. Scott, D., Thompson, J.: Probability density estimation in higher dimensions. In: Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface (1983)
7. Somorjai, R.L., Dolenko, B., Baumgartner, R.: Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics 19(12), 1484–1491 (2003)
8. Bild, A.H., Yao, G., Chang, J.T., et al.: Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439(7074), 353–357 (2006)
9. Desmedt, C., Piette, F., Loi, S., et al.: Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. Clin. Cancer Res. 13(11), 3207–3214 (2007)

10. Pawitan, Y., Bjöhle, J., Amler, L., et al.: Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res. 7(6), R953–R964 (2005)
11. Miller, L.D., Smeds, J., George, J., et al.: An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc. Natl. Acad. Sci. USA 102(38), 13550–13555 (2005)
12. Schmidt, M., Böhm, D., von Törne, C., et al.: The humoral immune system has a key prognostic impact in node-negative breast cancer. Cancer Res. 68(13), 5405–5413 (2008)
13. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cognitive Neurosci. 1(3), 71–86 (1991)
14. Chiang, S.S., Chang, C.I.: Unsupervised hyperspectral image analysis using independent component analysis. In: IEEE International Geoscience and Remote Sensing Symposium, vol. 1(7), pp. 3136–3138 (July 2000)
15. Robila, S.A., Varshney, P.K.: Target detection in hyperspectral images based on independent component analysis. In: SPIE AeroSense, Orlando, Florida, USA, vol. 1(7), pp. 3136–3138 (April 2002)
16. Liebermeister, W.: Linear modes of gene expression determined by independent component analysis. Bioinformatics (January 2002)
17. Luts, J., Poullet, J.B., Garcia-Gomez, J.M., et al.: Effect of feature extraction for brain tumor classification based on short echo time 1h mr spectra. Magn. Reson. Med. 60(2), 288–298 (2008)
18. Jolliffe, I.: Principal component analysis. Springer Series in Statistics (2002)
19. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Netw. 13(4-5), 411–430 (2000)
20. Kong, W., Vanderburg, C.R., Gunshin, H., et al.: A review of independent component analysis application to microarray gene expression data. BioTechniques 45(5), 501–520 (2008)
21. Alter, O., Brown, P.O., Botstein, D.: Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA 97(18), 10101–10106 (2000)
22. Sims, A.H., Smethurst, G.J., Hey, Y., et al.: The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis. BMC Medical Genomics 1, 42 (2008)

On Assigning Individuals from Cryptic Population Structures to Optimal Predicted Subpopulations: An Empirical Evaluation of Non-parametric Population Structure Analysis Techniques

Pornchalearm Deejai[1], Anunchai Assawamakin[2], Pongsakorn Wangkumhang[2], Kanokwan Poomputsa[3], and Sissades Tongsima[2,*]

[1] Bioinformatics and Systems Biology Program, King Mongkut University of Technology Thonburi, Bangkok 10140, Thailand
[2] Biostatistics and Informatics Laboratory, Genome Institute, National Center for Genetic Engineering and Biotechnology, Thailand Science Park, Paholyothin Road, Pathumthani 12120, Thailand
[3] School of Bioresources and Technology, King Mongkut University of Technology Thonburi, Bangkok 10140, Thailand
sissades@biotec.or.th

Abstract. Many algorithms have been proposed to analyze population structures from the single nucleotide polymorphism (SNP) genotyping data of some number of individuals and try to assign individuals to genetically similar groups. These algorithms can be categorized into two computational paradigms: parametric and non-parametric approaches. Although the parametric-based approach is a gold standard for population structure analysis, the computational burden incurred by running these algorithms is unacceptable for large complex dataset. As genotyping platforms incorporating more SNPs, analyzing ever larger and more complex datasets are becoming a standard practice. Hence, the computationally efficient non-parametric methods for analysis of genotypic datasets are needed to reveal the population structure. In this study, we evaluated two leading non-parametric population structure analysis techniques, namely ipPCA and AWclust, on their abilities to characterize the genetic diversity and population structure of two complex SNP genotype datasets (as many as 243855 SNPs). The head-to-head comparisons were conducted on two major aspects: ability to infer the number of genetically related subpopulations (K) and ability to correctly assign individuals to these subpopulations. The experimental results suggested that AWclust could be more suitable when applying to a small and less complex dataset. However, with a large and more complex dataset, ipPCA is a much better choice yielding higher accuracy on assigning genetically similar individuals to the inferred groups.

Keywords: Population genetic, Population genetic structure, parametric-based method, non-parametric-based method.

* Corresponding author.

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 58–70, 2010.

1 Introduction

Population genetics is concerned with the structure of different populations, which can be observed by frequency differences among the populations. Population structure analysis is important to genetic association studies [1-4] and evolutionary investigations [5-7]. Since most studies of human variation focus on sampling from predefined "populations" using culture and/or their geographical origins, these populations may not reflect the underlying genetic relationships [8-9]. Many algorithms have been proposed to analyze population structures from the single nucleotide polymorphism (SNP) genotyping data of some number of individuals and try to assign individuals to genetically similar groups. These algorithms can be categorized into two major computational paradigms: parametric and non-parametric approaches.

Parametric approaches require assumption of genetic model to assign individuals with similar genetic background to a predefined number of subpopulations (K). Such an assignment is carried out based on statistical likelihood using assumptions such as Hardy-Weinberg equilibrium (HWE) and linkage equilibrium (LE) among loci for each population [10,11]. The parametric approach, e.g., STRUCTURE, has been used as standard practice on population structure analyses. Nonetheless, the computational burden incurred by running these algorithms is unacceptable for solving large complex dataset. Furthermore, the statistical estimators of HWE and LE may not hold by any statistical estimators due to randomness in sampling. For this scenario, the non-parametric methods are more appropriate for analyzing population structure than parametric methods. These non-parametric approaches use standard statistical techniques to search for relatedness of genetic signal among data instead of finding the -best-fit likelihood of presumed genetic model. Two most recent reports of the algorithms in this class include ipPCA [12] and AWclust [13].

As SNP genotyping data becomes ever larger, it is increasingly difficult to efficiently analyze population structure by means of parametric statistical techniques due to their computational intensive requirements. The non-parametric approaches are becoming viable tools for researchers to understand population diversity and structure. Both ipPCA and AWclust tools have both advantages and disadvantages, but it is still not clear how these methods differ in their power to analyze population structure and suitability for analyzing large and complex SNP genotype data in terms of individual assignment and estimation of the optimal number of subpopulations (K). Hence, this paper aims to empirically evaluate these two aspects (inferring K and individual assignment) of these non-parametric algorithms. This evaluation was conducted on large complex datasets, 1) worldwide human dataset from Xing et al [14] containing 586 individuals from 28 populations 243855 SNPs and 2) BovineHapMap dataset containing 497 samples from 19 predefined breeds 27203 SNPs. 3) Simulated dataset from program GENOME [15] containing 20 subpopulations 10000 SNPs. These comparison results from of all datasets can suggest researchers to select non-parametric tools to analyze their datasets.

2 Material and Methods

2.1 Dataset

There are one simulated and two real datasets used in this study. For the simulated dataset, we create a population model shown in **Fig. 1** and use the program GENOME [15] to generate the genotypic data under the Wright-Fisher neutral coalescent model (backward in time) [16]. The simulated model contains 20 subpopulations derived from three ancestral populations. The simulated data contains 400 individuals with 10000 SNPs. This simulated model was used to test both AWclust and ipPCA by simulating 30 datasets from this model as the inputs to these algorithms. These simulated datasets with only 10000 SNPs are much less complex when comparing with the real datasets. The first real dataset represents a complex dataset with a large number of subpopulations but with a smaller number of SNP markers. The second group of real dataset represents a very complex dataset both in number of subpopulations and the number of SNP markers. The first real dataset is the SNP genotype of 497 cattle from BovineHapMap Project obtained from 19 different biologically diverse breeds. Due to computational limitation of Gap Statistics, AWclust demarcates the number of maximum inferred subpopulations to 16. In order to perform the experiment, the dataset was reduced to 15 breeds, containing 368 individuals with 27203 SNPs, by dropping individuals labeled as HOL, HFD, JER, and GNS from the original dataset. We also use a complex dataset from [14] to test ipPCA algorithm. This dataset represent a large population from 27 worldwide populations from Africa, Asia, and Europe in which we added our 32 samples from Thai population making up 586 individuals with 243855 SNPs. Tables 1 and 2 present the detail information of these two complex datasets.

Table 1. The number of individuals of each cattle breed in BovineHapMap. The total number of subpopulations is 19 using the three letter labeling as follows: ANG, Angus; BMA, Beefmaster; BRM, Brahman; BSW, Brown Swiss; CHL, Charolais; GIR, Gir; LMS, Limousin; NDA, N'Dama; NEL, Nelore; NRC, Norwegian Red; PMT, Piedmontese; Gertrudis; SHK, Sheko.RGU, Red Angus; RMG, Romagnola SGT, Santa; HFD, Hereford; GNS, Guernsey; HOL, Holstein; JER, Jersey. The asterisk (*) symbol indicates the breeds that were removed from the experiment so as to meet the K=16 limitation imposed by AWclust.

Breed	Count	Breed	Count
CHL	24	RMG	24
GIR	24	SHK	20
HFD *	27	BSW	24
ANG	27	NDA	25
BRM	25	NEL	24
HOL *	53	BMA	24
JER *	28	GNS *	21
LMS	42	NRC	25
PMT	24	SGT	24
RGU	12		
		Total	**497**

Table 2. The number of individuals from 27 subpopulations reported in [14]. The 32 Thai samples (unpublished data) were added to this dataset. The total number of individuals is 586 samples with 243855 SNPs.

Ethnic Group	No. of individuals	Ethnic Group	No. of individuals
Alur	10	CHB	45
Hema	15	JPT	45
Pygmy	25	Luhya	24
Brahmin	25	Tuscan	25
Utah	25	Kung	13
Khmer	5	Pedi	10
Chinese	7	Sotho_Tswana	8
Dalit	13	Stalskoe	5
Irula	24	Iban	25
Japanese	13	Chinese_TW	3
Madiga	10	Tamil	14
Mala	11	Urkarah	18
CEU	60	Veitnam	7
YRI	60	Nguni	9
		Thai	32
		Total	**586**

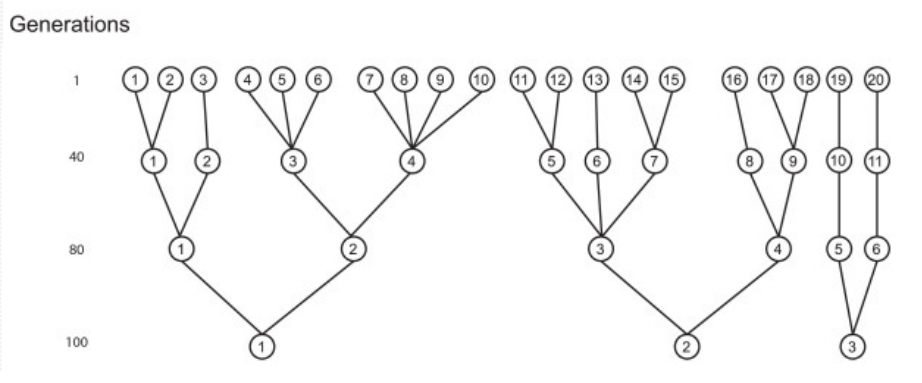


Fig. 1. Population history trees for generating simulated datasets. The GENOME tool [15] was used to generate the simulated datasets.

2.2 Comparison of the Two Non-parametric Algorithms

In this paper, the two non-parametric population structure analysis algorithms were studied in terms of their performance. Both algorithms claimed that they could efficiently operate on dataset on which the parametric STRUCTURE algorithm would be infeasible to operate. The following paragraphs describe fundamental non-parametric techniques deployed in each algorithm.

The ipPCA makes use of an exploratory data analysis technique, called principal component analysis (PCA), to observe common pattern from given genetic data by means of covariance analysis. The algorithm markedly improves resolution of population substructure by an iterative pruning process. It first performs PCA on the dataset and uses fuzzy c-mean algorithm [17] to cluster the PCA result to split the transformed data into two prominent. The process is repeated on each of the split group. The terminating condition is verified, for every run of PCA, using the TW test statistic described in EIGENSTRAT/SmartPCA [18,19]. The default TW p-value threshold used for detecting structure is conservative ($p < 10^{-12}$). This software is publicly available from http://www4a.biotec.or.th/GI/tools/ippca.

The AWclust software calculates the allele sharing distance (ASD) matrix, which represents the underlying genetic distance between every pair of individuals. It performs non-parametric exploration with the SNP data set by generating multidimensional scaling (MDS) 2D/3D plots to get a general idea of how the data clusters and to detect any outliers in the dataset. The MDS plot helps reveal outliers in the dataset and identify clusters and general relationships among individuals. AWclust calculates the Gap statistic for estimating the optimal number of groups based on the sample genetic relatedness. The Gap statistics compares the pooled within-cluster sum of squares with its expectation from a null reference distribution. Hence, the precision of this method requires multiple simulations from the null reference distribution. This process, however, is computationally intensive. The data points are then plotted ranging by cluster sizes and the optimal size maximizes the distance between the observed and expected pooled within-cluster sum of squares. The resulting hierarchical plots may also help interpret Gap statistic plots [13]. This software is publicly available from http://awclust.sourceforge.net/

3 Results

In this section, we present the results obtained by running ipPCA and AWclust algorithms to analyze BovineHapMap SNP dataset. Individual assignment tables were created to report results obtained from the two algorithms. Each column represents the genetically related group inferred by each algorithm. To make the tables more readable, we labeled each group to match the breed name originally given when the samples were first collect.

3.1 ipPCA Analysis

The ipPCA program was used to analyze the dataset with 27203 SNPs of BovineHapMap hosting 15 breeds. By observing the terminal nodes produced by ipPCA, this dataset can be re-organized into K=15 genetically related clusters. This is in concordance with the breed labels previously assigned at the sample collection time. **Fig. 2** presents the individual assignment to the terminal nodes of ipPCA. Most of the assignments agree with the breed labels previously specified, except the 3 samples from the CHL breed. We also experimented on the whole data set of BovineHapMap (with 19 breeds). The ipPCA algorithm was able to predict 19 genetically similar groups (K=19), which is the same as the breeds. The assignment was preformed yielding the assignment accuracy as high as 99.2 percent as shown in **Fig. 3**.

	BMA (24)	BSW (24)	CHL (24)	LMS (40)	NDA (25)	NRC (25)	PMT (24)	RMG (24)	SHK (20)
ipPCA	BMA (24)	BSW (24)	CHL (21)	LMS (40)	NDA (25)	NRC (25)	PMT (24)	RMG (24)	SHK (20)
AWclust K=15	BMA (24)	BSW (24)	CHL (21)	LMS (40)	NDA (25)	NRC (25)	PMT (24)	RMG (24)	SHK (20)
AWclust K=16	BMA (24)	BSW (24)	CHL (21)	LMS (40)	NDA (25)	NRC (25)	PMT (24)	RMG (24)	SHK (20)

Continue

	NEL (23)	SGT (24)	BRM (25)	ANG (27)	GIR (24)	RGU (12)	Others	Others
ipPCA	NEL (23)	SGT (24) CHL (3)	BRM (25)	ANG (27)	GIR (24)	RGU (12)	-	-
AWclust K=15	NEL (23)	SGT (20)	-	ANG (27) RGU (12)	GIR (24) BRM (25)	-	SGT (3) LMS (2) NRC (3)	SGT (1) NEL (1) CHL (3)
AWclust K=16	NEL (23)	SGT (20)	BRM (25)	ANG (27) RGU (12)	GIR (24)	-	SGT (3) LMS (2) NRC (3)	SGT (1) NEL (1) CHL (3)

Fig. 2. Analysis results of reduced BovineHapMap dataset (15 breeds). This figure presents the individual assignment ipPCA (observed at the terminal nodes generated by ipPCA tree) and the individual assignment results obtained from the cut tree of AWclust. Each column represents a genetically similar group, which both algorithms assigned the samples to. The columns labeled "others" represent the extra groups suggested by Gap statistics. The number of samples is shown in parentheses. For demonstration purpose, we tried to put the same assigned breed name in the same column. The "-" symbol indicates no such group name, implying that the samples might be in the same group with other samples. The first row indicates the assignment results done by ipPCA. These results demonstrate that most of the assignments agree with the breed labels previously specified, except the 3 samples from the CHL breed mixing with the SGT one. The second and third rows of the table reports the assignment results of AWclust when setting K=15 and 16 respectively.

We also applied ipPCA to analyze the large and complex dataset [14] combining with our unpublished SNP genotype data of 32 Thai individuals. This combined dataset forms 28 different ethnics groups, geographically distributed around the world. ipPCA was able to infer 15 genetically similar groups (K=15). The individual assignment results observed at each terminal nodes of the ipPCA bifurcation tree are shown in **Fig 5**.

Moreover, we also applied ipPCA framework to analyze the simulated dataset, which was generated from program GENOME [15] and derived from three ancestral populations. To be able to compare with AWclust, the simulated data was reduced to

	BMA (24)	BSW (24)	LMS (40)	NDA (25)	RMG (24)	SHK (20)	HFD (27)	HOL (53)	GNS (21)	JER (28)
ipPCA	BMA (24)	BSW (24)	LMS (40)	NDA (25)	RMG (24)	SHK (20)	HFD (27)	HOL (53)	GNS (21)	JER (28)
AWclust K=16	BMA (24)	BSW (24)	LMS (40)	NDA (25)	RMG (24)	SHK (20)	HFD (27)	HOL (53)	GNS (21)	JER (27)

	ANG (27)	RGU (12)	PMT (24)	CHL (24)	GIR (24)	BRM (25)	NEL (24)	NRC (25)	SGT (24)	Others
	ANG (27)	RGU (12)	PMT (24)	CHL (21)	GIR (24)	BRM (25)	NEL (24)	NRC (25)	SGT (24) CHL (3)	
Continue	ANG (27) RGU (12)	-	PMT (24) CHL (21)	-	GIR (24) BRM (25) NEL (23)	-	-	NRC (25) SGT (3) LMS (2) JER (1)	SGT (20)	SGT (1) NEL (1) CHL (3)

Fig. 3. ipPCA and AWclust results of BovineHapMap dataset with 19 subpopulations. The ipPCA algorithm was able to predict 19 genetically similar groups (K=19), which is the same as the breeds. Using K=16, AWclust was not able to correctly assign the individuals to the pre-allocated groups.

15 subpopulations with 300 individuals, we found that ipPCA was able to infer the correct K with results swinging between K=14 and K=15 (**Fig 6**.). When we test ip-PCA against the full dataset (20 pops, 400 individuals), the classification accuracy was much improved. The individual assignment results observed at each terminal nodes of the ipPCA bifurcation tree are shown in **Fig 7**.

3.2 AWclust Analysis

AWclust calculated the allele sharing distance (ASD) matrix from the raw data and calculate the Gap statistics for estimating the number of K. To predict the optimal K, the module Gap statistics must be performed incrementally; the highest Gap statistics value (y-axis) indicates the most probable K (x-axis). We present the Gap statistics values ranging from K= 1 to 16 in **Fig 4**.

AWclust was applied to analyze 27,203 SNPs of the reduced BovineHapMap dataset. The Gap statistics suggested the optimal K to be either 15 or more than 16 (the maximum inferred K is 16 for AWclust). **Fig 4** presents the Gap statistic results suggesting the value of K to be used in the individual assignment step. Next, AWclust used this K number to create a cut on dendogram plot in order to inform us which individuals belong to what groups (see AWclust user manual for more information on the hierarchical clustering and its dendogram plot). Since it is not certain if inferred K should be 15 or more groups, we tried to create different cuts based on K=15 and 16. The assignment results are tabulated in **Fig 2**. In this figure, the assignment results of AWclust when using K=16 are worse than those results when using K=15. Extra groups were created with mixed individuals from different breeds assigned to the group (see the columns “others” in **Fig 2**).

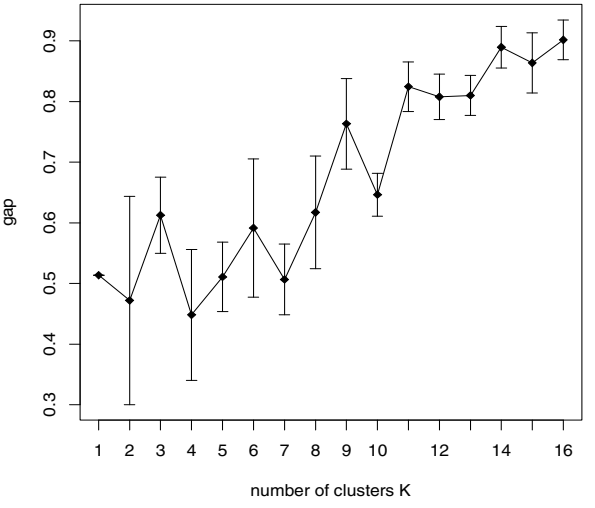


Fig. 4. Gap statistic result from reduced BovineHapMap dataset. The numbers of inferred Ks ranging from 1 to 16 are shown in the graph. The x-axis represents different possible Ks and the y-axis represents the gap value (higher is better).

Since we cannot verify if the Gap statistics can accurately predicting the correct K, for the full set of BovineHapmap data containing 19 breeds, we omitted the Gap statistics step and applied different larger Ks (K=15 to K=18) to test the individual assignment function of AWclust. Similar to the case K=16, larger K values resulting in incorrect cuts on the hierarchical clustering dendogram. **Fig. 3** presents the assignment results on the full BovineHapmap dataset. However, too many mis-assignments, i.e., having a group of combined breeds or the same breed get split to two or more different groups, are observed. Due to the page limitation of this conference, the assignment data when K=17 and 18 are not shown in this paper.

AWclust was deployed to analyze a very large and complex dataset, the 28 worldwide population dataset. Since the number of geographic subpopulations is far greater than the limit which was set for Gap statistic, the experiments on this dataset will only test the individual assignment accuracy. Similar to the BovineHapmap situation, we assume that Gap statistic could infer K to be any value larger than 16. We then used these numbers to create cut trees, which in turn gave us the individual assignment results. **Fig 5** shows the assignment given by AWclust assuming K=15. Unlike, the results shown in ipPCA row, the AWclust assignment tends to group unrelated individuals together. These groups are in the form of mixed populations in which some populations were split and assigned to several other groups. The AWclust assignment results got worst when increasing the number of K (data not shown).

AWclust was tested against the reduced simulated dataset (15 pops with 300 individuals having 10000 SNPs each). The experiment was repeatedly performed for 30 times on both reduced and full simulated datasets. We found that AWclust was able to infer the correct K with 100% accuracy for individual assignment of 15 subpopulations (see **Fig 6**.). On the other hand, when the dataset become more complex, we found AWclust failed to correctly infer K (see **Fig 7.).**

	YRI (60)	Kung (13)	Pygmy (25)	Sotho (8)	Nguni (9)	Pedi (10)	Luhya (24)	Hema (15)	Alur (10)	CEU (60)
ipPCA	YRI (60)	Kung (13)	Pygmy (25)	Sotho (8) Nguni (9) Pedi (10)	-	-	Luhya (24) Hema (15) Alur (10)	-	-	CEU (60) Utah (25) Tuscan (25)
AWclust K=15	YRI (60)	Kung (11)	Pygmy (25)	Sotho (1)	Nguni (5) Sotho (4) Pedi (1)	Pedi (9) Nguni (4) Sotho (4) Kung (1)	Luhya (24) Hema (15) Alur (10)	-	-	CEU (60) Utah (23) Urkarah (17) Tuscan (25) Stalskoe (2)

Continue

	Utah (25)	Tuscan (25)	Urkarah (18)	Stalskoe (5)	Brahmin (25)	Tamil (14)	Madiga (10)	Mala (11)	Dalit (13)	Irula (24)
	-	-	Urkarah (18) Stalskoe (5)	-	Brahmin (25) Tamil (14)	-	Madiga (10) Mala (11) Dalit (13)	-	-	Irula (24)
	-	-	-	Stalskoe (3) Utah (2) Urkarah (1)	Brahmin (24) Tamil (14) Mala (10) Madiga (10) Irula (1) Dalit (13)	-	-	-	Mala (10) Brahmin (1)	Irula (23)

Continue

	Iban (25)	Khmer (5)	Vietnam (7)	Chinese (10)	CHB (45)	Japanese (13)	JPT (45)	Thai (32)
	Iban (25)	Khmer (5) Vietnam (7)	-	Chinese (10) CHB (45)	-	Japanese (13) JPT (45)	-	Thai (32)
	-	-	-	Chinese (10) CHB (45) Vietnamese (7) Thai (1) Khmer (4) JPT (45) Japanese (2)	-	Japanese (11)	-	Thai (31) Sotho (5) Khmer (1) Iban (20)

Fig. 5. Result of analysis with dataset from [14] combining with 32 Thai samples. ipPCA was able to infer 15 genetically similar groups (K=15) and assigned most related individuals to these predicted groups (see [14] for detail discussion on these populations]. AWclust, however, was not able to predict the K due to the Gap statistic limitation. To make it comparable with ipPCA, we set K=15 for AWclust. The AWclust assignment is shown in the row under that of ipPCA. Since there are 28 populations being observed while only 15 groups to assign individuals to, mixed populations of different combinations can be expected. For demonstration purpose, the table was split into three parts to accommodate different mixed-pop combinations.

	POP2 (20)	POP3 (20)	POP4 (20)	POP5 (20)	POP6 (20)	POP7 (20)	POP10 (20)
ipPCA	POP2 (20)	POP3 (20)	POP4 (20)	POP5 (20)	POP6 (20)	POP7 (20)	POP10 (20)
AWclust K=15	POP2 (20)	POP3 (20)	POP4 (20)	POP5 (20)	POP6 (20)	POP7 (20)	POP10 (20)

Continue

POP11 (20)	POP12 (20)	POP13 (20)	POP14 (20)	POP15 (20)	POP1 (20)	POP9 (20)
POP11 (20)	POP12 (20)	POP13 (20)	POP14 (20)	POP15 (20)	POP1 (20) POP9 (6)	POP9 (14)
POP11 (20)	POP12 (20)	POP13 (20)	POP14 (20)	POP15 (20)	POP1 (20)	POP9 (20)

Fig. 6. Analysis results of reduced simulated dataset (15 subpopulations). This figure presents the individual assignment ipPCA (observed at the terminal nodes generated by ipPCA tree) and the individual assignment results obtained from the cut tree of AWclust. Each column represents a genetically similar group, which both algorithms assigned the samples to. The number of samples is shown in parentheses. For demonstration purpose, we tried to put the same assigned subpopulation name in the same column. The first row indicates the assignment results done by ipPCA. These results demonstrate that most of the assignments agree with the subpopulation labels previously specified, except the 6 samples from the population 9 mixing with the population 1. The second row of the table reports the assignment results of AWclust when setting K=15, we found the accuracy of individual assignment has 100%.

	POP1 (20)	POP2 (20)	POP3 (20)	POP11 (20)	POP13 (20)	POP14 (20)	POP15 (20)	POP16 (20)	POP18 (20)	POP19 (20)
ipPCA	POP1 (20)	POP2 (20)	POP3 (20)	POP11 (20)	POP13 (20)	POP14 (20)	POP15 (20)	POP16 (20)	POP18 (20)	POP19 (20)
AWclust K=16	POP1 (20)	POP2 (20)	POP3 (20)	POP11 (20)	POP13 (20)	POP14 (20)	POP15 (20)	POP16 (20)	POP18 (20)	POP19 (20)

Continue

POP20 (20)	POP5 (20)	POP6 (20)	POP7 (20)	POP9 (20)	POP10 (20)	POP12 (20)	POP17 (20)	POP8 (20)	POP4 (20)
POP20 (20)	POP5 (20)	POP6 (20)	POP7 (20)	POP9 (20)	POP10 (20)	POP12 (20) POP17 (1)	POP17 (19)	POP8 (20) POP4 (3)	POP4 (17)
POP20 (20)	POP5 (20) POP6 (20)	-	POP7 (20) POP9 (20) POP10 (20)	-	-	POP12 (20)	POP17 (20)	POP8 (20) POP4 (20)	-

Fig. 7. ipPCA and AWclust results of simulated dataset with 20 subpopulations. The ipPCA algorithm was able to predict 20 genetically similar groups (K=20), which is the same as the population label. Using K=16, AWclust was not able to correctly assign the individuals to the pre-allocated groups.

4 Discussion

Labeling the breed or subpopulation according to their pedigree or ethnics could be inaccurate. Genetic profile of each individual is more appropriate to distinguish subpopulations. Both non-parametric algorithms strive to infer the optimal number of genetically related groups, which may differ from the number of original labels. The number of inferred groups (K) heavily influences the accuracy in assigning individuals to the groups. The following discussion points out advantages and disadvantages on using ipPCA versus AWclust.

Practicality of ipPCA and Comparison with AWclust

In view of practical use, both ipPCA and AWclust tools are convenient to use because both of them provide graphical user interface, which is required by many life science scientists. Since both programs make use of different algorithms, the running time of these tools are also different. AWclust utilizes Gap statistics, which is computational intensive due to iterative statistical inference process. Hence, AWclust demands more computational resource than ipPCA, which makes use of PCA technique. Moreover, the larger number of SNPs dramatically slow down the execution of AWclust while this does not happen to ipPCA since it makes use of singular value decomposition (SVD), which reduces the size of correlation matrix down to the matrix rank (set by the number of individuals). Due to the slowness of Gap statistics, AWclust set a hard limit of the maximum number of inferred K to be 16; this value already doubled the upper limit set in the previous version of AWclust. For this aspect, AWclust is clearly not suitable to perform large-scale population genetic analysis of current genome wide SNP array platform.

Assignment of Individual Samples to Inferred Group K in Real Datasets

AWclust tends to perform better than ipPCA when the number of SNP markers is small and the data contain less variety of individuals (smaller number of inferred K). The simulated results of 15 subpopulations with 300 individuals and 10000 SNPs demonstrate that AWclust consistently yielded correct inferred K and able to re-assign these individuals to their original subpopulations. This experiment was repeatedly performed for 30 times on the same simulated data in order to test the robustness of these two algorithms. However, ipPCA was able to infer the correct K for merely half of all the experiments (swinging between K=14 and K=15). For the real datasets, which contain more SNPs and samples, ipPCA outperformed AWclust on both inferring K and assigning individuals to correct subpopulations. This performance discrepancy stems from the different core algorithms used in these two programs. In other words, the exploratory data analysis in PCA offers better results only when the number of informative attributes, SNPs, is large enough so that the eigenanalysis of PCA can thoroughly explore the variance profile among the input samples. AWclust core algorithms, however, rely heavily on Gap statistics to correctly predict K, which is later used to create a cut point on the hierarchical clustering dendogram. For a not so complex dataset, Gap statistics can correctly predict the optimal K. Furthermore, the hierarchical clustering in AWclust can also produce a decent dendogram based on

allele sharing distance (ASD) matrix. This simple two-step process is nearly deterministic rendering AWclust to outperform ipPCA for a small simulated dataset. On the other hand, for small dataset with not many SNPs, the clustering step, during each iteration of ipPCA, tends to perform inconsistently. It is also worth noting that both non-parametric approaches can discover the genetical differences within populations that the parametric STRUCTURE approach could not see. In particular, the paper [18] used STRUCTURE to analyze BovineHapmap data and it failed to differentiate the three admixed breed, namely NEL, BRM and GIR [20]. These breeds appeared as one group in STRUCTURE view. However, the prior information suggests us that these three breeds are distinct by their nature. Both AWclust and ipPCA were able to put them in three genetically different groups. Although the discussion on STRUCTURE is beyond the scope of this work, we suspected that the genetic model used by STRUCTURE may not be able to work well on this BovineHapMap dataset.

5 Conclusion

This study empirically demonstrated the performance of non-parametric-based population structure analysis methods in the aspect of individual assignment and prediction of the number of genetically similar subpopulations when applying the algorithms to the large complex datasets. The results showed that ipPCA is more suitable when applying to large dataset. This conclusion was derived from the ability to assign individuals to K subpopulations. Furthermore, we observed that the predicted K played a significant role in the overall prediction accuracy performance. AWclust used Gap statistics, which was claimed to be optimal by the authors, can accurately infer the optimal K with small datasets. For the complex dataset with large number of SNP markers, AWclust was not able to predict the correct number of subpopulations. Furthermore, if the number of correct subpopulations was given, AWclust cannot assign individuals to the correct group. Thus, from these real experimental results ipPCA is a better choice for handling complex dataset with large number of SNPs with large variety of subpopulations. However, from the result tested on the less complex simulated dataset, PCA-based technique was not able to accurately observe the overall trend from less number of SNPs for which AWclust outperformed ipPCA. Consequently, researchers can choose the tools to analyze the data based on the number of SNP markers and the number of subpopulations.

Acknowledgments. This work was supported by the National Center for Genetic Engineering and Biotechnology of Thailand, School of Information Technology and School of Bioresources and Technology King Mongkut's University of Technology Thonburi. Anunchai Assawamakin was supported by BIOTEC postdoctoral grant.

References

1. Lander, E.S., Schork, N.J.: Genetic Dissection of Complex Traits. Science 265(5181), 2037–2048 (1994)
2. Risch, N.J.: Searching for Genetic Determinants in the New Millennium. Nature 405, 847–856 (2000)

3. Marchini, J., Cardon, L.R., Phillips, M.S., Donnelly, P.: The Effects of Human Population Structure on Large Genetic Association Studies. Nat. Genet. 36(5), 512–517 (2004)
4. Freedman, M.L., Reich, D., Penney, K.L., McDonald, G.J., Mignault, A.A., Patterson, N., Gabriel, S.B., Topol, E.J., Smoller, J.W., Pato, C.N., Pato, M.T., Petryshen, T.L., Kolonel, L.N., Lander, E.S., Sklar, P., Henderson, B., Hirschhorn, J.N., Altshuler, D.: Assessing the Impact of Population Stratification on Genetic Association Studies. Nat. Genet. 36, 388–393 (2004)
5. Cavalli-Sforza, L.L., Menozzi, P., Piazza, A.: The History and Geography of Human Genes. Princeton University Press, Princeton (1994)
6. Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J., Cavalli-Sforza, L.L.: High Resolution of Human Evolutionary Trees with Polymorphic Microsatellites. Nature 368, 455–457 (1994)
7. Mountain, J.L., Cavalli-Sforza, L.L.: Multilocus Genotypes, a Tree of Individuals, and Human Evolutionary History. Am. J. Hum. Genet. 61, 705–718 (1997)
8. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., Feldman, M.W.: Genetic Structure of Human Populations. Science 298, 2381–2384 (2002)
9. Shriver, M.D., Kennedy, G.C., Parra, E.J., Lawson, H.A., Sonpar, V., Huang, J., Akey, J.M., Jones, K.W.: The Genomic Distribution of Population Substructure in Four Populations Using 8,525 Autosomal SNPs. Hum. Genomics 1, 274–276 (2004)
10. Pritchard, J.K., Stephens, M., Donelly, P.: Inference of Population Structure Using Multilocus Genotype Data. Am. J. Hum. Genet. 67, 945–959 (2000)
11. Purcell, S., Sham, P.: Properties of Structured Association Approaches to Detecting Population Stratification. Hum. Hered. 58, 93–107 (2004)
12. Intarapanich, A., Shaw, P.J., Assawamakin, A., Wangkumhang, P., Ngamphiw, C., Chaichoompu, K., Piriyapongsa, J., Tongsima, S.: Iterative Pruning PCA Improves Resolution of Highly Structured Populations. BMC Bioinf. 10(382) (2009)
13. Gao, X., Starmer, J.D.: AWclust: Point-and-Click Software for Non-parametric Population Structure Analysis. BMC Bioinf. 9(77) (2008)
14. Xing, J., Watkins, W.S., Witherspoon, D.J., Zhang, Y., Guthery, S.L., Thara, R., Mowry, B.J., Bulayeva, K., Weiss, R.B., Jorde, L.B.: Fine-Scaled Human Genetic Structure Revealed by SNP Microarrays. Genome Res. 19, 815–825 (2009)
15. Liang, L., Zollner, S., Abecasis, G.R.: GENOME: a rapid coalescent-based whole genome simulator. Bioinformatics (Oxford, England) 23(12), 1565–1567 (2007)
16. Ewens, W.J.: Mathematical Population Genetics. Springer, Berlin (1979)
17. Bezdec, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
18. Parsons, L., Haque, E., Liu, H.: Subspace Clustering for High Dimensional Data: a Review. ACM SIGKDD Explor. Newslett. 6(1), 15 (2004)
19. Patterson, N., Price, A.L., Reich, D.: Population Structure and Eigenanalysis. PLoS genet. 2(12), e190 (2006)
20. Gibbs, R.A., Tassell, C.V., Weinstock, G., Green, R., Hamernik, D., Kappes, S., Liu, G., Matukumalli, L., Matukumali, A., Sonstegard, T., Silva, M.: Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science 24, 528–532 (2009)

Extended Constraint-Based Boolean Analysis: A Computational Method in Genetic Network Inference

Somkid Bumee[1], Chalothorn Liamwirat[2], Treenut Saithong[1,3], and Asawin Meechai[1,4]

[1] Systems Biology and Bioinformatics Research Laboratory, Pilot Plant Development and Training Institute
[2] Division of Biotechnology, School of Bioresources and Technology
[3] Bioinformatics and Systems Biology Program
[4] Department of Chemical Engineering, Faculty of Engineering,
King Mongkut's University of Technology Thonburi, Bangkok, Thailand
somkid@pdti.kmutt.ac.th, chalothorn09@yahoo.com,
treenut.sai@kmutt.ac.th, asawin.mee@kmutt.ac.th

Abstract. Reconstruction of a genetic network, which describes gene regulation of cellular response processes, has been widely studied by using various approaches. Some of which are computational expensive and require enormous efforts. Herein, we proposed an *extended constraint-based Boolean* to infer genetic network. Our method incorporated the specific constraints for a particular system in addition to the general conceptual constraints of a typical genetic circuit, to improve the performance of the existing constraint-based Boolean algorithm. This method was demonstrated in inference of the genetic network underlying circadian rhythms from microarray time series data. The results showed that the proposed method provides good accuracy, specificity, and precision under the trade-off of computational efforts. Moreover, the resulting network showed that prior knowledge is a useful bias for modeling genetic network. The proposed method is therefore a promising alternative approach for inferring genetic network from high-throughput data, such as microarray.

Keywords: Genetic network, extended constraint-based Boolean, conceptual constraints, specific constraints.

1 Introduction

Relationship between gene in a genetic network is important information in understanding the cellular response processes, which involve the regulation of gene expression [1]. The regulation of gene expression lies on a huge number of components that comprise a genetic network, multiple levels of regulation as well as the elaborated interaction between levels [2]. Though the number of network constituents is a barrier of network reconstruction, the (differential) expression of such components is often employed in network inference. This strategy becomes more and more popular once the measurement of thousands of gene components (or whole genome) can simultaneously be performed with the aid of microarray techniques.

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 71–82, 2010.

In the last decade, availability of high-throughput technologies, allowing the levels of transcripts to be measured for the whole genome at the same time, enables scientists to understand cellular system by reconstructing genetic network [3, 4]. Various computational approaches have been developed on the purpose of genetic network reconstruction, such as Boolean network [5-7], graphical Gaussian model [8], and Bayesian network [7, 9]. Among these approaches Boolean network and Bayesian network methods are mostly used in the context of reconstructing genetic network from microarray data [10]. These two approaches have distinct advantages and disadvantages. Bayesian network provides a more accurate result yet with a huge requirement of prior data and computational efforts in an iterative learning algorithm, while Boolean network is the simpler method to reconstruct genetic network. Under the trade-off of computational effort, Boolean network is considerably a competitive method to Bayesian network.

Boolean network was originally introduced by Kauffman [5,11]. Later, Shmulevich and Zhang used Boolean network to infer genetic network of cell cycle regulation based on gene expression data [12]. In Boolean network, gene expression is simply considered as binary values, ON or OFF, and the regulation between genes is set by Boolean function. Boolean network was then extended to be Probabilistic Boolean network [13]. This model consists of a family of Boolean networks that combine more than one transition Boolean functions. Inferred network which composes of a set of Boolean functions is selected by using the highest score based on probability. In 2007, Martin and colleagues [6] used Boolean dynamics to infer genetic regulatory network. Possible Boolean networks were generated from microarray time series data. The genetic network was then inferred from selection of possible Boolean networks by using steady-state dynamics. However, the result from Boolean network often includes a number of false positive, resulting in a complex inferred network. To resolve such a problem of Boolean network, recently, our group proposed a constraint-based Boolean network to formulate genetic network by taking prior knowledge into account [14]. The prior knowledge in this work, called the *conceptual constraints*, *i.e.*, enzymatic coding genes do not control and regulate regulatory genes, were included in the filtering process before generating Boolean functions. The result showed the achievement of this model to reduce the complexity of the inferred genetic network by eliminating a certain false prediction.

One of the most studied genetic networks is circadian clock system (*i.e.* a genetic circuit generates about 24h rhythm or circadian rhythm) because it is an important system controlling many biological processes in a wide range of organisms, including plants. Circadian clock in plants has mostly been studied in *Arabidopsis thaliana* [15, 16] in which a certain network components and regulations are revealed. The core circadian clock composes of multiple interlocked feedback loops such as interlock with the timing of cab expression 1 (*TOC1*)/CIRCADIAN AND CLOCK ASSOCIATED1 (*CCA1*)/LATE ELONGATED HYPOCOTYL (*LHY*) loop and (Pseudo-response regulator; *PRR5*/*PRR7*/*PRR9*)/*CCA1*/*LHY* loop. Experimental results show that *CCA1* and LHY are partially redundant genes which are negative regulators of *TOC1* [15]. *CCA1* and *LHY* are also positive regulators of two *TOC1* relatives, *PRR7*, and *PRR9* [16], while *TOC1* acts as a positive regulator of *CCA1* and *LHY*.

The circadian clock network has been used for computational method demonstration in various works [7, 14, 17], mainly due to the appropriate size of the network and (microarray) data availability (http://www.ncbi.nlm.nih.gov/geo/). Needham and colleagues [7] inferred a relationship between circadian-clock genes from an initial set of key genes and iteratively learned to increase network members around genes. A circadian clock was also used as a seed for inferring genetic network by finding co-regulation patterns between gene pairs in circadian clock [8]. The genetic network of *Arabidopsis* genome was performed by using an iterative random sampling strategy.

In this work, we extended the previous constraint-based Boolean analysis [14] to acquire a more accurate network inference by focusing on the filtering process. The specific biological constraints derived from prior knowledge of a particular system under study were introduced to the Boolean network in addition to the conceptual constraints. The algorithm takes a set of genes in a standard input format that is easy in the data preparation process. The extended method was demonstrated in inference of a genetic network underlying circadian rhythms in *A. thaliana* using microarray time series data. Finally, the inferred circadian network was validated with literature [15, 16, 18] and the performance of the extended algorithm was evaluated. The genetic network inferred from our algorithm was compared with that of the constraint-based Boolean approach. The results showed that our algorithm, which considers both conceptual and specific constraints, can increase the accuracy, specificity, and precision of the inferred network. Also the degree of complexity of the inferred network is substantially reduced, resulting more understandable results. The proposed method is therefore a promising alternative approach for inferring larger-scale genetic network from high-throughput microarray time-series data.

2 Methods for Reconstructing Constraint-Based Boolean Network of Circadian Rhythms

The overview of methodology is shown in Fig. 1A. Briefly, expression data was pre-processed before discretization. The binary values from discretization step are the inputs for the extended constraint-based Boolean program to generate Boolean functions and types of regulating genes, i.e., activation and inhibition. The output from the constraint-based Boolean program is Boolean relationship which can be visualized by Cytoscape [19].

2.1 Microarray Data and Data Pre-Processing

Gene expression time series datasets from the Affymetrix microarray under diurnal changes of *Arabidopsis* leaves were downloaded from NCBI database (http://www.ncbi.nlm.nih.gov, experiment reference number is GSE8365) [20]. *Arabidopsis* were grown in light/dark cycles for 7 days and then transferred to constant light. After 24 hours in constant light, 12 samples were harvested at four hours intervals over the next 44 hours for RNA extraction and hybridization on Affymetrix microarrays. The expression data were preprocessed using a package of Bioconductor [21, 22].

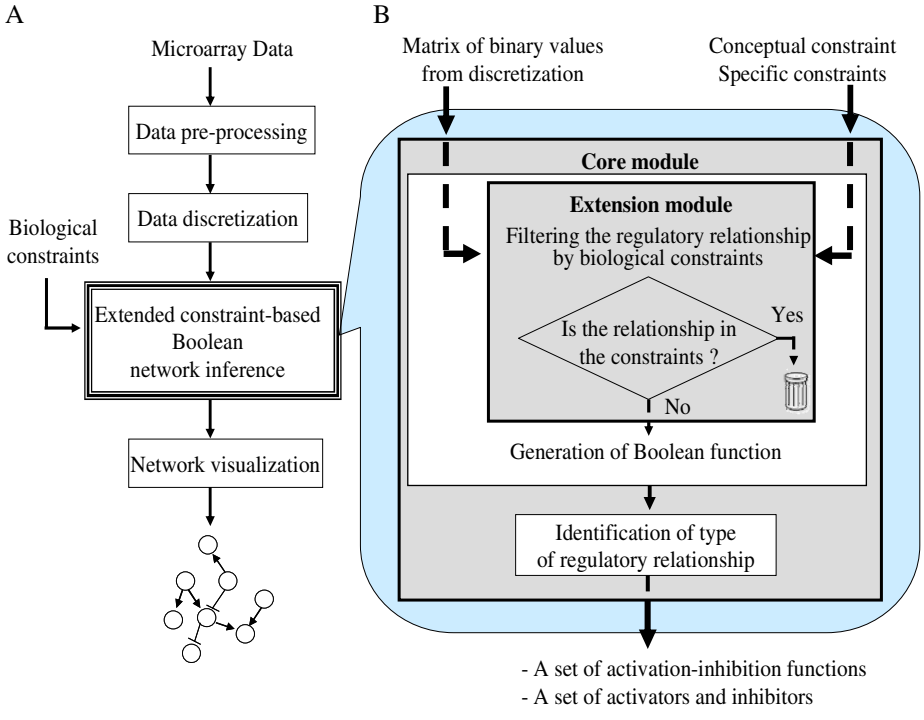


Fig. 1. Overall methodology for genetic network reconstruction by using (A) our constraint-based Boolean algorithm where the extension to the previous method is described in (B)

2.2 Data Discretization

The continuous expression level of gene was discretized into two levels, either '0' or '1', based on the concept of Boolean analysis that, herein, represents the strength of expression, or state, of gene at particular time. In other words, the expression level of gene was converted into either '0' for weak expression or '1' for strong expression. In this work, we used the maximum value of expression level as the simple criteria for discritization that the expression level of gene greater than the determined percentage of the maximum value was discretized into '1', '0' otherwise. The discretized value, $s_{i,t}$, for gene i at time t is defined as Equation (1).

$$s_{i,t} = \begin{cases} 1 & \text{if } x_{i,t} > \text{Max}(\boldsymbol{G}_i) - r \cdot \text{Max}(\boldsymbol{G}_i) \\ 0 & \text{others,} \end{cases} \tag{1}$$

where $x_{i,t}$ is the expression level of gene i at time t, $\boldsymbol{G}_i$ is the set of all expression levels of gene i over the time series, and r is the percentage of the maximum value of expression level of gene i. Here, the expression level greater than 30% of the highest value was converted to 1, *i.e.* r = 0.3. The data matrix of discretized values of all gene expression, *i.e.* $\boldsymbol{S} = [s_{i,t}]$, is called matrix of binary values. It was then passed to extended constraint-based Boolean algorithm to generate Boolean functions

representing the regulatory relationship that is necessary for the construction of the Boolean network.

2.3 Constraint-Based Boolean Network Inference

Our method, called the *extended constraint-based Boolean algorithm* was adapted and implemented based on the previous published works [6, 14]. The algorithm consists of two modules, core and extension modules (Fig. 1B). The core module slightly adapted from the algorithm in Martin et al. [6] includes generation of Boolean function and identification of type of regulatory relationship, i.e. either activation or inhibition. The latter module is filtering the relationship of genes by biological constraints provided by a user. It is the extension we added in order to improve the performance of the classical Boolean algorithm by reduction of the number of relationships considered in the core module. Nevertheless, the network inference can be performed without this module if a set of biological constraints is not submitted.

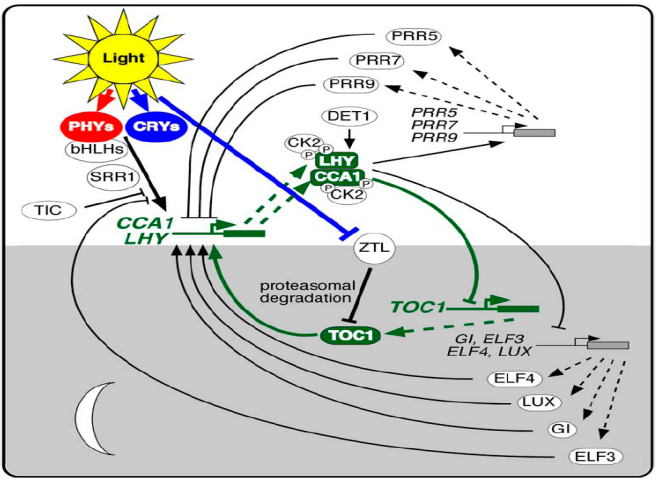
Core module: Generation of Boolean functions and identification of types of regulatory relationship. In brief, at first, the matrix of binary values of all interesting genes is passed into the first module to extract the regulatory relationship between the set of regulating genes and single target gene. The gene expression data, *i.e.* '0' or '1', of the target gene at time t is influenced by the expression strength of the regulating genes at previous time, t-1. This relationship is in the form of Boolean functions having a logic combination of the expression strengths of the regulating genes as an input and the expression strength of target gene as an output. The maximum number of the regulating genes, k, for single target gene is depended on the number of time points of expression data, T, and is defined by Max(k) that $2^k < T$ and $k > 0$.

Each Boolean function is then checked if it is an activation-inhibition function which its logic relationship is in the form of $g_i = (act_1 \text{ OR } act_2 \text{ OR } \ldots \text{ OR } act_{A_i})$ AND NOT $(inh_1 \text{ OR } inh_2 \text{ OR } \ldots \text{ OR } inh_{I_i})$, where g_i indicates the expression strength of the i^{th} target gene, *act* and *inh* indicates the expression strength of activators and inhibitors for the i^{th} target gene respectively, and A_i and I_i are the number of activators and inhibitors for the i^{th} target gene respectively. The activators and inhibitors of each target gene are simultaneously identified in this checking step. The type of regulatory relationship, either activation or inhibition, is finally assigned based on the activation-inhibition function. Other Boolean functions that are not the activation-inhibition function are ignored. Consequently, the output from the algorithm is a set of activation-inhibition functions and a set of activators and inhibitors for each target gene.

Extension module: Filtering the regulatory relationship by biological constraints. In the extension module, the relationship of genes is filtered by a set of biological constraints that are considered as the prior knowledge for a specific system. In this work, two types of biological constraints, *i.e.*, conceptual and specific constraints are added in the program. The conceptual constraint is first added to the previous algorithm [14]. It includes the general concept of the regulation in the transcriptional level, for example, (i) transcription factors directly regulate the gene expression by binding at promoter of genes; and (ii) enzymes and transporters do not regulate the gene expression although some of their downstream products do. Here, this type of

constraints is set based on a presumption that there is not any regulation by products of enzyme-encoding genes within duration of study. The relationship with this type of genes is hence discarded. The specific constraint introduced in this work includes prior knowledge, hypothesis, or existing experimental data indicating the regulatory relationship between the specific set of genes. Here, this set of constraints is specified by pairs of genes having no regulatory relationship which is supported by biological evidences. For example, the relationship between CONSTANTS (*CO*) and phosphoglycerate kinase (PGK) was set as null because the genes have distinct functions in different pathway and there is no experimental data inferring their relationship. The Boolean relationship between these two genes generated was hence

A



B

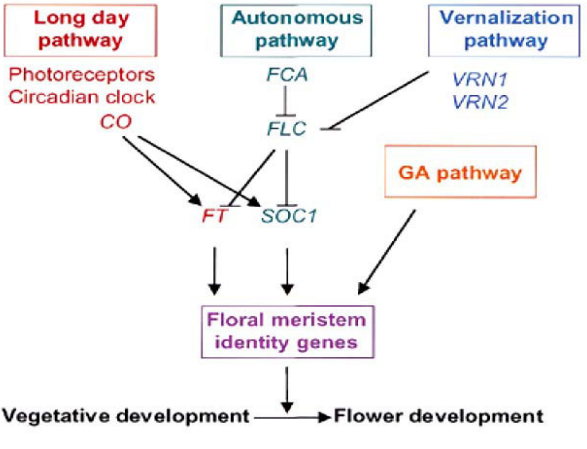


Fig. 2. Gene relationship in (A) circadian clock [18] and (B) flowering pathway [23]

discarded by consideration of that specific constraint. However, the network inference can be performed without this module if a set of biological constraints is not submitted.

2.4 Evaluation of Genetic Network

The obtained genetic network were evaluated by using standard measures that were accuracy (ACC) calculated by (TP+TN)/(TP+TN+FP+FN), specificity (SPC) calculated by TN/(FP+TN), precision or positive prediction value (PPV) calculated by TP/(TP+FP), and false discovery rate (FDR) calculated by FP/(FP+TP), where *TP* refers to correctly inferred edges, either activation or inhibition. *FP* refers to false predicted relationship, including wrong type of regulation. *TN* refers to missing edges in both the inferred and the reference networks. *FN* refers to missing edges, which exist in the reference network, in the inferred network. The network that was used as a reference was based on selected genes in this study [18, 23] (Fig. 2).

Accuracy indicates the percentage of correct predictions. Specificity indicates the percentage of negative predictions which are correctly inferred. Precision indicates the percentage of positive predictions which are correctly inferred. False discovery rate indicates the percentage of false predictions among all predictions.

3 Results and Discussion

3.1 Data Discretization

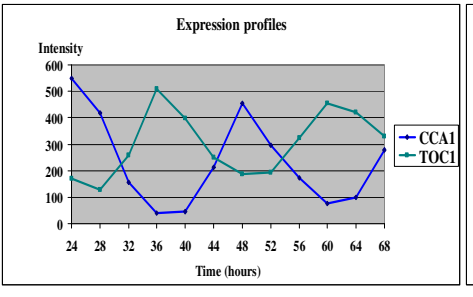
To demonstrate the algorithm, the expression data of fourteen genes were selected from microarray data [20], including seven known core-circadian-clock genes and seven non-circadian genes (*i.e.* genes in glycolysis and flowering pathways). The selected genes and their molecular functions are shown in Table 1.

Table 1. List of genes and functions used in this study [18, 24]

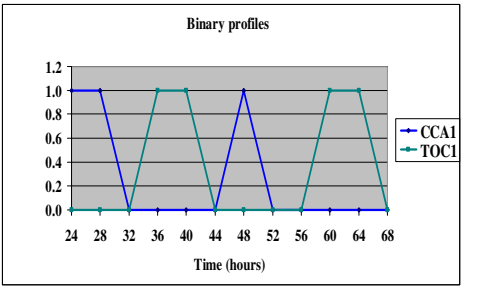
Gene	Function
CCA1	Single Myb domain Transcription factor
ELF4	Transcription factor
GI	Unknown
LHY	Single Myb domain transcription factor
PRR5	Pseudo-response regulator
PRR7	Pseudo-response regulator
TOC1	Pseudo-response regulator
CO	CONSTANTS (CO) promotes flowering under long days
FT	Flowering locus promotes flowering
SOC1	Suppressor of overexpression of CO1
PGI1	Phospho-glucose (Glc) isomerase
TPI	Triosephosphate isomerase
PGK	Phosphoglycerate kinase
PGM	Phosphoglycerate mutase

Max-30%max was used as a threshold for the data discretization in this study. An example of the characteristics of discretized data is shown in Fig. 3. Fig. 3A shows the expression data of two selected circadian clock-genes, *CCA1* and *TOC1*. Based on the discretization method, the expression data of *CCA1* and *TOC1* across time points were discretized into 0 or 1. So, each gene consists of a series of binary values, called binary profile. Fig. 3B shows the binary profiles of such genes. The selection of the discretization method may affect the final result; however the employed discretization method was proven to be the most appropriate one for the system under studied (unpublished data). The matrix of binary values representing binary profiles of a set of genes was an input for the algorithm, see Fig. 3C. The matrix of binary values is delimited text format. The first column is the gene name. The following columns are binary values of each gene across time points. This is a standard format that is convenient for users in the data preparation process.

A



B



C

CCA1	1	0	0	0	0	0	0	0	0	0	0	0
ELF4	0	0	0	0	1	0	0	0	0	0	0	0
GI	0	0	1	1	0	0	0	0	0	0	0	0
LHY	1	0	0	0	0	0	0	0	0	0	0	0
PRR5	0	0	1	1	0	0	0	0	1	1	0	0
PRR7	0	0	1	0	0	0	0	1	1	0	0	0
TOC1	0	0	0	1	0	0	0	0	0	0	0	0
CO	0	0	0	0	0	1	0	0	0	0	0	0
FT	1	0	0	0	0	0	0	0	0	0	0	0
SOC1	0	0	0	0	0	0	0	0	1	1	0	0
PGI1	1	1	1	0	1	0	0	1	1	1	1	1
TPI	0	0	0	1	1	1	0	1	0	0	0	1
PGK	1	1	0	0	1	1	0	0	0	0	1	0
PGM	1	1	1	0	0	0	1	1	0	0	0	0

Fig. 3. Characteristics of data expression profiles (A), example of discretized data (B), and the matrix of binary values (C) of genes in circadian clock, glycolysis, and flowering mechanism

3.2 Genetic Network of Circadian Rhythms

The genetic network of circadian clock was inferred by using our method, the extended constraint-based algorithm with taking consideration of both conceptual and specific constraints into account. The network result by our method was compared with those by the classical Boolean without any biological constraint and the constraint-based Boolean with only conceptual constraints. All these three algorithms can

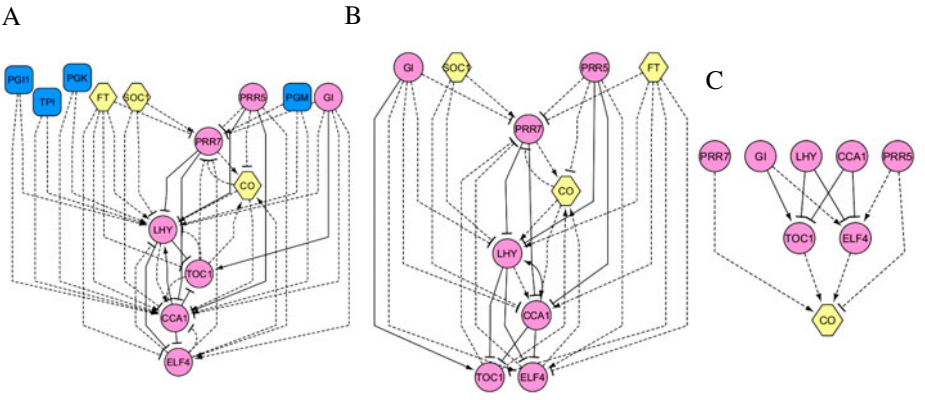


Fig. 4. Genetic networks of 14 genes in circadian rhythm and flowering mechanism using classical Boolean, without consideration of biological constraint *in priori* (A); constraint-based Boolean with conceptual constraints (B); constraint-based Boolean with both conceptual and specific constraints (C). A blue rectangular represents an enzymatic gene; a yellow hexagon represents a flowering gene; a pink circle represents a circadian gene. Black solid lines represent predicted relationship corresponding to known biological knowledge, while broken lines represent predicted relationship that exceeds the current knowledge.

infer directed networks describing the types of gene regulatory relationship, either activation or inhibition (Fig. 4). The node is a gene and the edge is the relationship between genes. The types of the relationships are represented by an arrow and a T-shape arrow for activation and inhibition, respectively.

Figs. 4A-C show the networks inferred by using classical Boolean, constraint-based Boolean with conceptual constraints, and constraint-based Boolean with both conceptual and specific constraints, respectively. All these inferred genetic networks can describe regulations between *TOC1*/*CCA1*/*LHY* and (*PRR5*/*PRR7*/*PRR9*) /*CCA1*/*LHY* loops. All three inferred network show that *CCA1* and *LHY* are inferred as negative regulators of *TOC1* and *ELF4* which corresponds to known biological knowledge [15, 18], while *PRR5* and *PRR7* are inferred as positive regulators of *CCA1* and *LHY* [16] in the inferred network by using classical Boolean and constraint-based Boolean with conceptual constraints. However, among three inferred networks, the two networks from the previously developed methods show significantly higher in complexity and false predictions than the one from our method, indicated by the number of solid and broken line edges. Adding conceptual constraints before generating Boolean function can greatly reduce the complexity of the network (Fig. 4B). That means it can reduce false predicted relationships that are caused by regulations by products of enzyme-encoding genes as shown in Fig. 4A. Fig. 4C shows the inferred genetic network by using our method with adding conceptual and specific constraints before generating Boolean function. The algorithm can identify regulations in the core oscillator of circadian mechanism and also substantially reduce the false predicted relationships. This resulted in less complex inferred network that is reasonable for further analysis and making a biological sense.

3.3 Network Evaluation

The inferred genetic networks were evaluated through a set of coefficients: ACC, SPC, PPV, and FDR. These coefficients allow us to assess the performance of our algorithm in comparison with those of the previously developed methods which are the classical Boolean and the constraint-based Boolean algorithms with conceptual constraints.

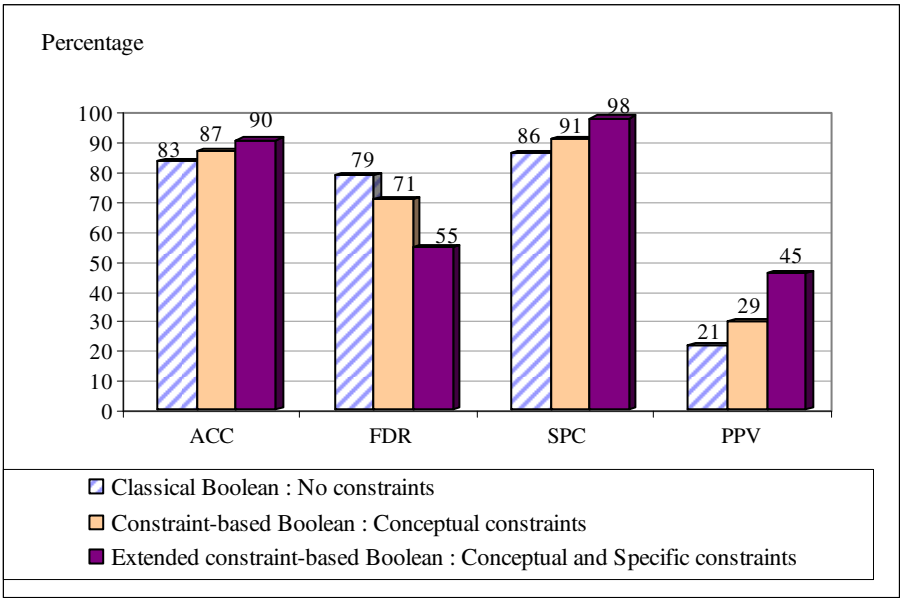


Fig. 5. Comparing the performances of the extended constraint-based Boolean, constraint-based Boolean, and classical Boolean algorithms

Fig. 5 shows that the Boolean network taking into consideration of biological constraints gives better accuracy, specificity, and precision. In comparison with the classical Boolean network, the extended constraint-based Boolean algorithm provides 90% accuracy, 98% specificity, and 45% precision, which are 8%, 13% and 114% improvement, respectively. Moreover, the false discovery rate (FDR) is decreased from 79% to 55% (31% improvement). When considering the Boolean network taking only the conceptual constraints into account, the percent improvement over the classical Boolean network are 4%, 5%, and 38% for accuracy, specificity, and precision, respectively. These results clearly show that taking more consideration of biological constraints in priori can provide better accuracy, specificity, and precision. Besides the improve accuracy, the extended constraint-based Boolean algorithm provides a result with a low level of false prediction. The extension of the constraint-based method by incorporating the specific constraint is thus not only advantage in term of reduction in, but also great decrease in computational burden due to Boolean functions calculation. Not only genetic network inference of circadian clock, the

algorithm was also applied to infer genetic network of galactose pathway using microarray data [25]. The results show that our algorithm provides both high (>70%) accuracy and (>80%) specificity (unpublished data). Therefore, the extended constraint-based Boolean algorithm might be an alternative strategy for genetic network inference. Also, this method might be employed to infer a large-scale genetic network, whose result might be used as seed information for further network analysis or hypothesis development. Although the incorporation of prior knowledge into Boolean network is not yet systematic, this can help scientists to understand simpler genetic network inferred by using this method. However, it will be great to develop it as more systematic approach.

In this work, we have shown the advantages and successes of incorporation prior knowledge (in terms of specific constraint) into the Boolean network though implantation of the constraint in not yet systematic. For the next step, computational technique including systematic incorporation of the constraint will be improved to have the capability of the algorithm to support the large-scale data analysis.

4 Conclusion

The regulation of gene expression lies on a huge number of components that comprise a genetic network. Understanding of this regulation system is often studied by inference of genetic networks from microarray data. We have proposed an algorithm so-called extended constraint-based Boolean algorithm to infer genetic network. The algorithm considers both conceptual constraints of a typical genetic circuit and specific constraints of a particular system before generating Boolean functions. The method was demonstrated in inference of a genetic network underlying circadian rhythms in Arabidopsis thaliana from microarray time series data. The inferred circadian network was validated with literature and the performance of the novel algorithm was evaluated. The resulted network showed that prior knowledge is an useful bias for modeling genetic network. Moreover, the results showed that the proposed method provides good accuracy, specificity, and precision under the trade-off of computational efforts. The proposed method is therefore a promising alternative approach for inferring genetic network from high-throughput microarray time series data. In the future, this method will be applied to infer genetic network from different conditions of microarray data.

References

1. Karlebach, G., Shamir, R.: Modelling and analysis of gene regulatory networks. Nat. Rev. Mol. Cell. Biol. 9, 770–780 (2008)
2. Zhang, S.-Q., Ching, W.-K., Ng, M.K., Akutsu, T.: Simulation study in Probabilistic Boolean Network models for genetic regulatory networks. Int. J. Data Min. Bioinform. 1, 217–240 (2007)
3. Kwon, A.T., Hoos, H.H., Ng, R.: Inference of transcriptional regulation relationships from gene expression data. Bioinformatics 19, 905–912 (2003)
4. Kervestin, S., Amrani, N.: Translational regulation of gene expression. Genome Biol. 5, 359 (2004)

5. Kauffman, S.A.: Metabolic stability and epigenesis in randomly constructed genetic nets. J. Theor. Biol. 22, 437–467 (1969)
6. Martin, S., Zhang, Z., Martino, A., Faulon, J.L.: Boolean dynamics of genetic regulatory networks inferred from microarray time series data. Bioinformatics 23, 866–874 (2007)
7. Needham, C.J., Manfield, I.W., Bulpitt, A.J., Gilmartin, P.M., Westhead, D.R.: From gene expression to gene regulatory networks in Arabidopsis thaliana. BMC Syst. Biol. 3, 1–18 (2009)
8. Ma, S., Gong, Q., Bohnert, H.J.: An Arabidopsis gene network based on the graphical Gaussian model. Genome Res. 17, 1614–1625 (2007)
9. 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)
10. Li, P., Zhang, C., Perkins, E.J., Gong, P., Deng, Y.: Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks. BMC Bioinformatics 8 (suppl. 7), 13 (2007)
11. Kauffman, S.A.: The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press, New York (1993)
12. Shmulevich, I., Zhang, W.: Binary analysis and optimization-based normalization of gene expression data. Bioinformatics 18, 555–565 (2002)
13. Shmulevich, I., Dougherty, E.R., Kim, S., Zhang, W.: Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics 18, 261–274 (2002)
14. Munkung, W., Liamwirat, C., Bumee, S., Meechai, A.: A constraint-based Boolean approach to inferring genetic circuits. In: The 13th International Annual Symposium on Computational Science and Engineering, pp. 427–431 (2009)
15. Alabadí, D., Oyama, T., Yanovsky, M.J., Harmon, F.G., Más, P., Kay, S.A.: Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock. Science 293, 880–883 (2001)
16. Farre, E.M., Harmer, S.L., Harmon, F.G., Yanovsky, M.J., Kay, S.A.: Overlapping and distinct roles of PRR7 and PRR9 in the Arabidopsis circadian clock. Curr. Biol. 15, 47–54 (2005)
17. Du, P., Gong, J., Syrkin Wurtele, E., Dickerson, J.A.: Modeling gene expression networks using fuzzy logic. IEEE Trans. Syst. Man Cybern. B Cybern. 35, 1351–1359 (2005)
18. McClung, C.R.: Plant circadian rhythms. Plant Cell 18, 792–803 (2006)
19. 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)
20. Covington, M.F., Harmer, S.L.: The circadian clock regulates auxin signaling and responses in Arabidopsis. PLoS Biol. 5, e222 (2007)
21. Workman, C., Jensen, L.J., Jarmer, H., Berka, R., Gautier, L., Nielser, H.B., Saxild, H.H., Nielsen, C., Brunak, S., Knudsen, S.: A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol. 3, research0048 (2002)
22. Li, C., Wong, W.H.: Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection. In: Conference Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection, pp. 31–36 (2001)
23. Blazquez, M., Koornneef, M., Putterill, J.: Flowering on time: genes that regulate the floral transition. EMBO reports 2, 1078–1082 (2001)
24. The Arabidopsis Information Resource, `http://www.arabidopsis.org/`
25. DeRisi, J.L., Iyer, V.R., Brown, P.O.: Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278, 680–686 (1997)

Mining LINE-1 Characteristics That Mediate Gene Expression

Naruemon Pratanwanich[1], Apiwat Mutirangura, and Chatchawit Aporntewan[3]

[1] Graduate School
[2] Department of Anatomy, Faculty of Medicine
[3] Department of Mathematics, Faculty of Science,
Chulalongkorn University, Bangkok 10330, Thailand
Naruemon.npr@gmail.com, {Apiwat.M,Chatchawit.A}@chula.ac.th

Abstract. We proposed to use data mining to identify LINE-1 (L1) characteristics that were associated with gene expression in bladder cancer. The data were collected from L1Base and GSE3167. The memory-efficient data structure called FP-Tree was employed to enumerate all frequent item sets. The frequent item sets were then used to produce rules for predicting "down regulation" and "not down." Each rule was assigned a p-value by means of Chi-square test. No statistically significant rules for "down" had been found, in contrast 692 rules for "not down" were significant with odd ratios ranging from 1.68 to 1.98. All the significant rules were concentrated only in 20 characteristics. We were able to infer the L1 characteristics that down-regulated genes. Those characteristics were number of L1 elements in host genes, full-length intactness, number of CpG islands, conserved 5'UTR and mutated ORF2.

Keywords: LINE-1, hypomethylation, gene expression, bladder cancer, and data mining.

1 Introduction

DNA methylation is a cellular process characterized by the attaching of methyl groups (CH_3) to genomic DNA. The global hypomethylation is found in common in aging, cancer, and autoimmune diseases [1]. The majority of DNA methylation occurs at DNA fragments called transposable elements (TEs). There are two main families of TEs: short interspersed element (SINE) and long interspersed element (LINE). In our previous works [2], [3], [4], a pattern of hypomethylation in LINE-1 (L1), which is a subfamily of LINE, was discovered in various types of cancer. The loss of genome-wide L1 methylation may activate L1 elements and result in disrupting gene expression [5]. Our preliminary results, not yet published, showed a strong association between intragenic L1 and gene expression in induced-hypomethylation environments and different cancers. It is known that methylation at a gene promoter can silence the gene by blocking the transcription [6]. However, the alteration of gene expression caused by gene-body methylation remains largely unknown.

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 83–93, 2010.

To unravel the roles of L1 in the transcription process, we conducted data mining on gene expression data and L1 characteristics. Data mining algorithms, such as Apriori [7], search for frequent patterns of L1 characteristics that coexist with up/down regulation of the host genes. We found no statistically significant patterns associated with down regulation, in contrast not-down-regulated genes exhibited a number of common characteristics. Our findings indicated that full-length intactness, number of CpG islands, mutated ORF2, and conserved 5'UTR were L1 characteristics that may be prone to the down regulation of host genes.

2 Materials and Methods

Gene expression data set (GSE3167 bladder carcinoma vs. normal bladder epithelium) was downloaded from NCBI website. Our preliminary result is shown in Fig. 1. We found that genes possessing L1 elements were more frequently up/down regulated in several cancers. But not all the genes with L1 were totally up or down regulated. Consequently, we hypothesized that the transcription process of host genes may be subject to L1 characteristics. Since the most significant p-value was obtained in bladder cancer (Down vs. Not down), we initiated data mining from here. All L1 characteristics, deployed from L1Base [8], are shown in Table 2. Most characteristics were L1 subsequences. Besides the characteristics compiled by L1Base, we added the number of L1 elements and the orientation (sense/antisense of the host genes).

To conduct data mining, we built a two-dimensional table whose rows were genes that possessed L1 elements, and columns were L1characteristics plus the classification ("Down" or "Not down"). If a gene possessed multiple L1 elements, multiple rows were produced as a Cartesian product of a gene and a set of

	Up	Not up
L1	188	731
No L1	3,687	8,452

P-value = 2.84E-10
Odd ratio = 0.59
Lower 95% CI = 0.50
Upper 95% CI = 0.70

	Down	Not down
L1	382	537
No L1	3,377	8,762

P-value = 9.83E-19
Odd ratio = 1.85
Lower 95% CI = 1.61
Upper 95% CI = 2.12

Fig. 1. This figure shows the experiment GSE3167 bladder carcinoma situ vs. normal bladder epithelium. A gene either possesses LINE-1 (denoted by L1) or does not possess LINE-1 (denoted by "No L1"). The up/down regulation of a gene (denoted by "Up" and "Down") is determined by unpaired t-test (p-value threshold is set at 0.01). The entries in the 2x2 tables show the resulting number of genes. The p-values of 2x2 tables are obtained from Chi-square distribution. The 41 tests and 9 controls in the t-test are (GSE71028 to GSE71068) and (GSM71019 to GSM71027).

L1 elements. Rows with missing values were discarded. The characteristics that were of many values (but sparse) were manually clustered, for instance, "mut," "mut 84G/A," "mut 100T/C" were grouped together as "mut."

Weka [9], a software suite for data mining, was used to perform Apriori algorithm. Apriori enumerated all frequent item sets with support greater than or equal to a threshold called minimum support. But only the frequent item sets with the classification ("Down" or "Not down") were of interest, thus we filtered out the frequent item sets that did not contain the classification. In practice, Apriori failed to cope with large data due to memory bloat. Then we turned to a more memory-efficient data structure, FP-Tree [10] (available in RapidMiner [11]), but FP-Tree restricted that all variables (L1 characteristics) must be binominal. Non-binominal variables were automatically converted. The final table consisted of 1,593 rows and 146 columns. All L1 characteristics were summarized in Table 2.

Next, we produced an if-then rule from each frequent item set by placing the classification on the right-hand side of the rule. Each rule was assigned a confidence value. An example is shown in Eq. 1, 2, 3, and 4. C_i denotes a characteristic of L1, and "Down" denotes down regulation. Note that C_i and "Down" are binominal variables, their values are either "yes" or "no."

Basically higher support and higher confidence are always better, but low support may be accepted if the confidence is very high. However, determining a strong rule only from support and confidence could be misleading. A p-value was assigned to each rule by means of Chi-square test. An example of a 2x2 contingency table was shown in Table 1. Other statistics, odd ratio and confident intervals, were conducted as well.

$$\text{A frequent item set: } \{C_1 = \text{yes}, C_2 = \text{yes}, C_3 = \text{no}, \mathit{Down} = \text{yes}\} \tag{1}$$

$$\text{Support} = \frac{\text{number of rows that } \{C_1 = \text{yes}, C_2 = \text{yes}, C_3 = \text{no}, \mathit{Down} = \text{yes}\}}{\text{total number of rows}} \tag{2}$$

$$\text{An if-then rule: } C_1 = \text{yes}, C_2 = \text{yes}, C_3 = \text{no} \rightarrow \mathit{Down} = \text{yes} \tag{3}$$

$$\text{Confidence} = \frac{\text{number of rows that } \{C_1 = \text{yes}, C_2 = \text{yes}, C_3 = \text{no}, \mathit{Down} = \text{yes}\}}{\text{number of rows that } \{C_1 = \text{yes}, C_2 = \text{yes}, C_3 = \text{no}\}} \tag{4}$$

Table 1. This table shows a 2×2 contingency table where a, b, c, and d are numbers of rows

	Consistency to the left-hand side of the rule	Contradictory to the left-hand side of the rule
Down (down = yes)	a	b
Not down (down = no)	c	d

Table 2. A summary of all L1 characteristics. For more details, consult L1Base [8]. The data types (the second column) N, I, B, and R denote nominal, integer, binominal, and real respectively.

Characteristics of LINE-1	Type	Description
L1 Type	N	Type of L1 {FLI_L1, ORF2_L1, FLnI_L1}.
Chromosome	N	The chromosome where L1 is {1, . . . , 22}.
ORF1/ORF2 Conserved	N	The sequence for intactness of ORF1/ORF2 {1}.
ORF1/ORF2 Gaps	I	The number of gaps in ORF1/ORF2 (Min=0/0, Max=57/443, Average=2.36/11.99).
ORF1/ORF2 Frameshifts	I	The number of frameshifts in ORF1/ORF2 (Min=0/0, Max=8/31, Average=1.67/8.11).
ORF1/ORF2 Stops	I	The number of stops in ORF1/ORF2 (Min=0/0, Max=25/72, Average=3.15/8.07).
Ta SSVs	N	The family determining Ta locus of L1 element identified {AAGA, ACAG, ACGA, ACGG, GAGA, GAGG, GCGA, unclassified}.
Ta0/Ta1 SSVs	N	The family determining Ta0/Ta1 locus in ORF2 of L1 element identified. {Ta-0/L1PA2, Ta-1, L1PA5, non canonical}
Ta1-nd/d	N	The family determining Tad/nd deletion in the 5'UTR of L1 element identified {Ta1-d, Ta1-nd, non canonical}.
L1M/L1PA Discrimination	N	A HMM a diagnostic part of ORF2, which discriminates L1M and L1PA families {Mammalian L1M, Primate L1PA}.
Poly-A	I	The length of the pure Poly-A tail, as well as the length of an estimated Poly-A tail (containing mutations). Calculates as well a Kimura distance to a pure poly-A tail of the length of the estimated sequences. Distance is not available if mutation rate is too high.
PolyA Signal, Runx3 Site, Runx3 ASP, SRY Site 1, SRY Site 2, YY1 BoxA+ BoxA, TF nkx-2.5, TF nkx-2.5B, REKG235, ARR260, YPAKLS282, N14, E43, Y115, D145, N147, T192, D205, SDH228,R363, ADD700, HMKK1091, FADD700 SSS1096, I1220, S1259	B	The sequence for intactness of each name specified in ORF1 or ORF2 or 5'UTR {mut, cons}.
find TSDs	I	The number of target-site-duplications flanking L1 element (Min=0, Max=325, Average=15).

continued on next page

continued from previous page		
ORF StartStop	N	The ORFs of L1 for presence of valid methionine start codons and stop codons {mut, cons, ORF1 cons, ORF2 cons}.
G-C Content	R	The %G-C of L1 element in a 50nt window (Min=30.42, Max=44.59, Average=40.10).
ORF1/ORF2/ORF1&2 %A	R	The %A for ORF1/ORF2/ORF1&2 (Min=0.36/0.39/0.39, Max=0.48/0.47/0.47, Average=0.41/0.42/0.42).
ORF1/ORF2/ORF1&2 %T	R	The %T for ORF1/ORF2/ORF1&2 (Min=0.16/0.19/0.19, Max=0.25/0.28/0.28, Average=0.19/0.21/0.20).
ORF1/ORF2 CAI	R	The codon adaptation index of ORF1/ORF2 (Min=0.56/0.56, Max=0.71/0.67, Average=0.66/0.63).
Intactness score	I	The intactness measure for L1 (Min=2, Max=24, Average=16.44).
CPG Islands	I	The number of CPG islands found in L1 (Min =0, Max=2, Average=0.16).
Strand	B	The strand of L1 $\{+, -\}$.
Orientation	B	The orientation of gene $\{+, -\}$.
Number of L1s	I	The number of L1 elements found in the host gene (Min=1, Max=15, Average=3.039).

Although FP-Tree was memory efficient, the memory usage could be up to 16 GB. A 64-bit computer was needed to suffice the high memory demand. The HP Z800 workstation was equipped with dual Intel Xeon E5520 2.26 GHz, 16 GB of memory, Windows 7 Professional 64-bit, and RapidMiner 64-bit (version 4.6). With this setting and minimum support = 0.25, finding all frequent item sets of "Down vs. Not down" could be accomplished in 30 minutes. In the case of "Up vs. Not up," the minimum support was set to be lower than that of "Down vs. Not down" because there were fewer up-regulated genes. Unfortunately, the available 16GB memory was exceeded.

3 Results

We mined "Down vs. Not down." The minimum support was set at 0.25. It is important to note that the maximum support for a frequent item set containing "Down = yes" was 0.44. Setting the minimum support at 0.25 made a rule covering at least 0.25 / 0.44 = 56.82% of all genes that were down regulated. The rules whose supports were less than 0.25 could not explain the data in general, hence rejected.

Table 3 and Table 4 show the top-five rules (sorted by p-value) of "Down = yes" and "Down = no" respectively. No rules for "Down = yes" were significant at p-value threshold 0.05 (adjusted by Bonferroni correction). In contrast, 692

Table 3. The top-five rules of "Down = yes" are listed. There are totally 8,027 rules (Bonferroni adjusted p-value threshold = 0.05 / 8,027 = 6.23E-6). No rules are statistically significant.

Rule 1: L1M/L1PA Discrimination = PrimateL 1PA, TF nkx-2.5B = mut, CPG Islands $\leq$ 0.5, Runx3 Site = mut $\rightarrow$ *Down* = yes
Rule 2: Type = FLnI_L1, L1M/L1PA Discrimination = Primate L1PA, TF nkx-2.5B = mut, CPG Islands $\leq$ 0.5, Runx3 Site = mut $\rightarrow$ *Down* = yes
Rule 3: N147 = cons, TF nkx-2.5B = mut, CPG Islands $\leq$ 0.5, Runx3 Site = mut $\rightarrow$ *Down* = yes
Rule 4: Type = FLnI_L1, N147 = cons, TF nkx-2.5B = mut, CPG Islands $\leq$ 0.5, Runx3 Site = mut $\rightarrow$ *Down* = yes
Rule 5: Type = FLnI_L1, L1M/L1PA Discrimination = Primate L1PA, N147 = cons, TF nkx-2.5B = mut, Runx3 Site = mut $\rightarrow$ *Down* = yes

Rule	Support (max = 0.43)	Confidence	Odd Ratio	Confident Interval	Unadjusted p-value
1	0.27	0.42	0.75	0.61 - 0.92	0.0053
2	0.27	0.42	0.75	0.61 - 0.92	0.0053
3	0.26	0.42	0.76	0.62 - 0.93	0.0088
4	0.26	0.42	0.76	0.62 - 0.93	0.0088
5	0.25	0.42	0.77	0.63 - 0.94	0.0109

Table 4. The top-five rules of "*Down* = no" are listed. There are totally 87,042 rules (Bonferroni adjusted p-value threshold = 0.05 / 87,042 = 5.74E-7). 692 rules are statistically significant.

Rule 1: N147 = cons, number of L1 $\leq$ 2.5 $\rightarrow$ *Down* = no
Rule 2: Type = FLnI_L1, N147 = cons, number of L1 $\leq$ 2.5 $\rightarrow$ *Down* = no
Rule 3: N147 = cons, Y115 = cons, number of L1 = $\leq$ 2.5 $\rightarrow$ *Down* = no
Rule 4: number of L1 $\leq$ 2.5 $\rightarrow$ *Down* = no
Rule 5: Type = FLnI_L1, number of L1 $\leq$ 2.5 $\rightarrow$ *Down* = no

Rule	Support (max = 0.54)	Confidence	Odd Ratio	Confident Interval	Unadjusted p-value
1	0.35	0.63	1.98	1.62 - 2.42	2.46E-11
2	0.34	0.63	1.97	1.61 - 2.41	2.71E-11
3	0.32	0.64	1.96	1.60 - 2.39	4.26E-11
4	0.37	0.62	1.97	1.61 - 2.42	4.74E-11
5	0.37	0.62	1.96	1.60 - 2.40	5.69E-11

rules for "Down = no" were significant at the same p-value threshold. Moreover, the odd ratios of all 692 rules were greater than one (ranging from 1.68 to 1.98).

Without an expert, it is difficult to judge that one rule is better than the others just because its p-value is smaller. However, we can interpret all rules at once by taking an overview. Fig. 2 summarizes the frequency of all L1 characteristics that were found in the significant 692 rules. Although there were hundreds of rules, but all of them were concentrated in only 20 characteristics.

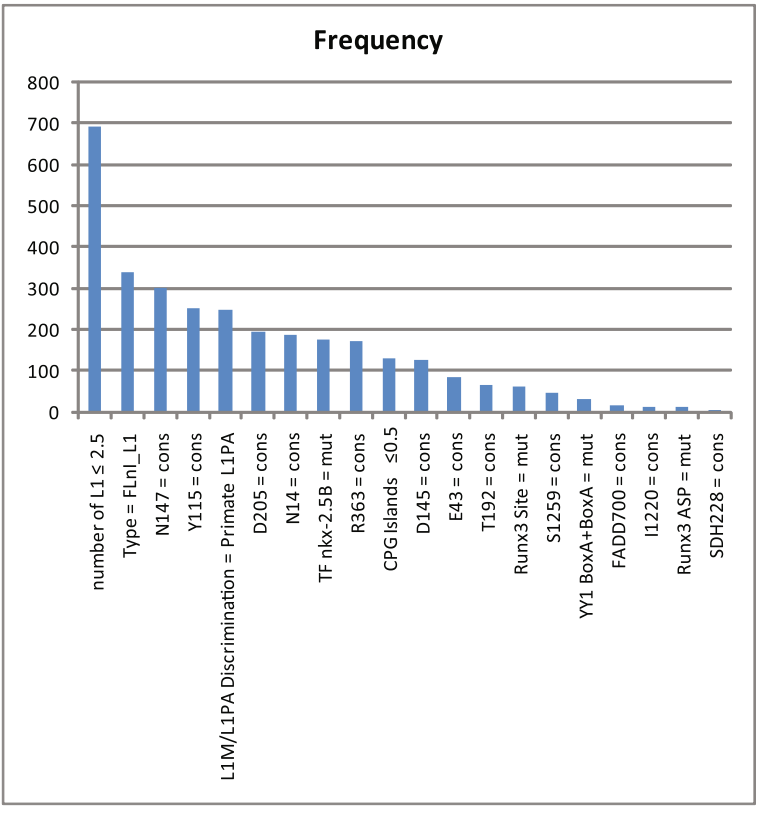


Fig. 2. The frequency of L1 characteristics that were found in the significant rules ("Down = no," Bonferroni adjusted p-value threshold = 5.74E-7)

4 Discussion

At the first glance, the absence of statistically significant rules in "Down = yes" suggests that no L1 characteristics are associated with down regulation. In contrast to the ubiquity of significant rules with high odd ratios in "Down = no," several L1 characteristics exhibit a protective effect on down regulation. In other words, the complementary characteristics (that are not protective) would promote down regulation. A rule, for instance, (C_1 and C_2 and C_3) → "Not down" would imply its complement ($\neg C_1$ or $\neg C_2$ or $\neg C_3$) → "Down." The later rule has never been observed because data mining restricts the operators to "and" ("or" operator is not allowed). As a result, no significant rules for "Down" are observed. Another reason might be that a lower minimum support is required, but lowering the support is not practical due to the memory limit. In the following discussion, we draw the characteristics that promote down regulation by taking the complements of the characteristics found in "Not down."

Let's consider the frequent characteristics in Fig. 2. The number of L1 elements is the hypothetical characteristic that we deliberately added into data mining. RapidMiner converted the number of L1 elements to a binominal variable by cutting at 2.5. "Not down" requires the number of L1 ≤ 2.5 (or equal to 0, 1, 2). This implies that the number of L1 ≥ 3 increases the chance of being down regulated.

The second characteristic is "Type = FLnI_L1," which is the abbreviation for full-length non-intact L1. The L1 elements in this type are not active due to multiple mutations. Some functions that are important for behaving like normal full-length intact L1s (FLI-L1s) may be lost. The non-intactness is required for "Not down." This implies that the down-regulation mechanism may require intact L1 elements.

The number of CpG islands ≤ 0.5 (no CpG islands) can protect down regulation. The CpG islands are likely to be required for transcriptional initiation of L1 elements. No CpG islands may prevent L1 transcription and make L1 inactive. Therefore down-regulation mechanism may require CpG islands for transcription process.

N147, Y115, L1M/L1PA, D205, N14, R363, D145, E43, T192, S1259, FADD-700, I1220, and SDH228, these characteristics refer to aminoacid sequences in ORF2. All these characteristics are of the same ORF and are of the same value (conserved). For L1M/L1PA, we considered the value "Primate L1PA" as conserved (another value is "Mammal L1M"). Note that ORF2 harbors three domains involved in L1 retrotransposition activity. It sounds like if the retrotransposition activity is still active (conserved ORF2), L1 elements cannot be used in down-regulation mechanism because these L1 elements are still capable of damaging genomic DNA. This implies that mutated ORF2 may be required for L1 elements that play a role in down regulation.

TF nkx-2.5B, Runx3Site, YY1 BoxA+BoxA, and Runx3ASP refer to aminoacid sequences in 5'UTR. Surprisingly, all these sequences in 5'UTR are of the same value (mutated). The mutations in L1 5'UTR may disrupt L1 RNA synthesis or called transcription. This is consistent with the number of CpG islands in the sense that the blockade of L1 RNA prevents down regulation. We concluded that L1 RNA may be a regulatory factor and required for down regulation.

Together all frequent characteristics infer that down-regulation mechanism requires the intactness of L1, the transcription of L1, and the disability of retrotransposition activity. The rules for predicting down regulation were invented and shown in Table 5. It can be seen that the number of genes that are consistent with the rules is getting smaller towards more specific rules. Hence those rules cannot be discovered by data mining due to very low support. An important observation is that many numbers of genes that are not down-regulated and consistent with the rules are zero (undefined odd ratio). This suggests that the rules can predict down regulation with no false positives. The p-values do not show significance because of insufficient sample size (there are very few L1 elements with specific characteristics).

Table 5. The rules for predicting down regulation were invented by combining four L1 characteristics. ORF2 is mutated if at least one of the subsequences is mutated.

Rules	Genes with L1				Odd Ratio			Unadjusted p-value
	Consistent with rule		Not consistent		Lower CI	Odd Ratio	Upper CI	
	Down	Not	Down	Not				
A (full-length, intact)	12	13	346	552	0.66	1.47	3.26	3.38E-01
B (CPG islands > 0)	88	114	270	451	0.94	1.29	1.77	1.15E-01
C (mutated ORF2)	268	395	90	170	0.95	1.28	1.73	1.03E-01
D (conserved 5'UTR)	25	32	333	533	0.73	1.25	2.15	4.17E-01
A and B	11	12	347	553	0.64	1.46	3.35	3.68E-01
A and C	3	0	355	565	NA	NA	NA	2.93E-02
A and D	7	10	351	555	0.42	1.11	2.93	8.38E-01
B and C	29	25	329	540	1.10	1.90	3.31	2.04E-02
B and D	24	31	334	534	0.71	1.24	2.15	4.47E-01
C and D	6	1	352	564	1.15	9.61	80.19	1.05E-02
A and B and C	3	0	355	565	NA	NA	NA	2.93E-02
A and B and D	7	10	351	555	0.42	1.11	2.93	8.38E-01
A and C and D	1	0	357	565	NA	NA	NA	2.09E-01
B and C and D	5	0	353	565	NA	NA	NA	4.85E-03
A and B and C and D	1	0	357	565	NA	NA	NA	2.09E-01

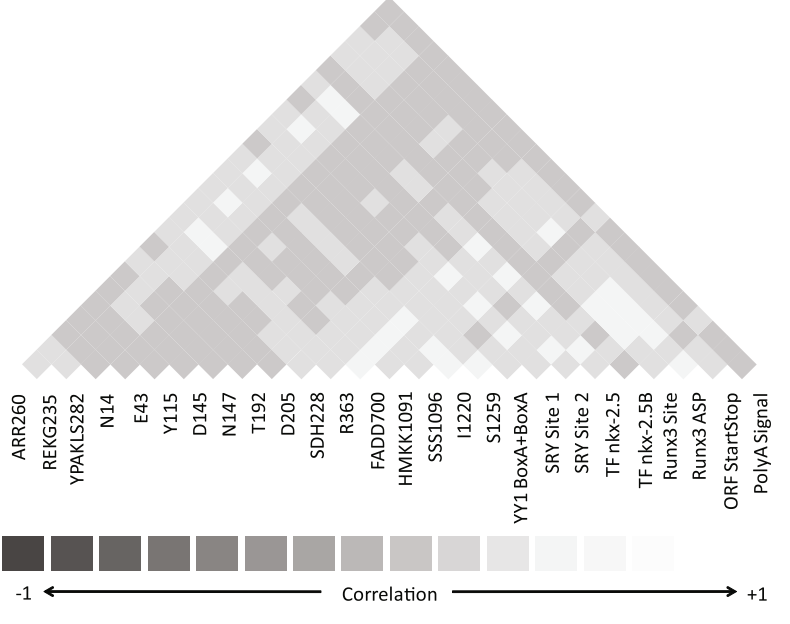


Fig. 3. This figure shows a correlation matrix of L1 characteristics in 5'UTR and ORF2

As many L1 characteristics are subsequences in 5'UTR and ORF2, they may be in tight linkage. To investigate this, we made a correlation matrix as shown in Fig 3. The L1 characteristics seem to be independent to each other. As a result, the discovery of rules consisting of L1 subsequences being "conserved" or "mutated" many times in a row did not happen due to genetic linkage, but the conserved/mutated subsequences and gene expression should be associated indeed.

Note that we cannot discover the association with gene expression by trying L1 characteristics one by one. There must be interactions among L1 characteristics at a certain degree as the rules were composed of multiple characteristics. We found that data mining is a promising tool for bioinformatics, and there are still a plenty of rooms for its applications. We are going to repeat the experiment with other types of cancer. The number of L1 elements in host genes may play a role as well as other L1 characteristics. However, a thorough investigation requires a proper experimental design to take other L1 characteristics into account (not considering the number of L1 elements alone). And it is not convenient to report in this paper.

5 Conclusion

We have mined L1 characteristics that are associated with gene expression in bladder cancer. A total of 692 significant rules has led to the discovery of L1 characteristics that may play an important role in down regulation of the host genes. Those characteristics are number of L1 elements, full-length intactness, number of CpG islands, mutated ORF2, and conserved 5'UTR. Our findings indicate that down regulation mechanism requires both quality and quantity of L1 (intactness and number of L1 ≥ 3), the transcription of L1 (CpG islands > 0 and conserved 5'UTR), and the disability of retrotransposition activity (mutated ORF2).

References

1. Robertson, K.D.: DNA methylation and human disease. Nature Reviews Genetics 6, 597–610 (2005)
2. Chalitchagorn, K., Shuangshoti, S., Hourpai, N., Kongruttanachok, N., Tangkijvanich, P., et al.: Distinctive pattern of LINE-1 methylation level in normal tissues and the association with carcinogenesis. Oncogene 23, 8841–8846 (2004)
3. Phokaew, C., Kowudtitham, S., Subbalekha, K., Shuangshoti, S., Mutirangura, A.: LINE-1 methylation patterns of different loci in normal and cancerous cells. Nucl. Acids Res. 36, 5704–5712 (2008)
4. Subbalekha, K., Pimkhaokham, A., Pavasant, P., Chindavijak, S., Phokaew, C., et al.: Detection of LINE-1s hypomethylation in oral rinses of oral squamous cell carcinoma patients. Oral Oncology 45, 184–191 (2009)
5. Han, J.S., Szak, S.T., Boeke, J.D.: Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429, 268–274 (2004)

6. Clark, D.: Molecular Biology. Elsevier Academic Press, Amsterdam (2005)
7. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2005)
8. Penzkofer, T., Dandekar, T., Zemojtel, T.: L1Base: from Functional Annotation to Prediction of Active LINE-1 Elements. Nucl. Acids Res. 33(Database issue), D498–D500 (2005)
9. WEKA Project, The University of Waikato, `http://www.cs.waikato.ac.nz/~ml`
10. Han, J., Pei, J., Yin, Y., Mao, R.: Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery 8(1), 53–87 (2001)
11. Rapid-I, `http://rapid-i.com`

Mining Regulatory Elements in Non-coding Regions of *Arabidopsis thaliana*

Xi Li[1,2] and Dianhui Wang[1,⋆]

[1] Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, VIC, 3086, Australia
[2] Department of Primary Industries, Bioscience Research Division, Victorian AgriBiosciences Centre, Bundoora, VIC, 3083, Australia
dh.wang@latrobe.edu.au

Abstract. Analysis of regulatory elements (DNA motifs) in non-coding regions is considered as one crucial step to understand the regulation mechanisms of genes with similar expression patterns. With the help of accumulated gene expression data and complete genome sequences, computational approaches have been developed in the past decade to accelerate the mining task. In previous studies, we proposed a DNA motif discovery framework, named as MODEC, which incorporated the evolutionary computation (EC) searching algorithm with data filtering techniques to favor the algorithm performance. With the attempt on exploring real-world motif mining problems, we apply both MODEC and a famous discovery algorithm MEME to predict regulatory elements in different non-coding regions of co-expressed genes from the model plant *Arabidopsis thaliana*. Results from both MODEC and MEME show that the targeted motif patterns can be found in the expected non-coding regions of the co-expressed gene groups. As the preliminary step of this work, we investigate whether different motif patterns can be detected in the specified non-coding regions of co-expressed genes with different functional categories. The similar prediction results from MODEC and MEME demonstrate the potential of MODEC in the field of practical motif discovery.

Keywords: Evolutionary computation, regulatory element, *Arabidopsis thaliana*.

1 Introduction

Regulatory elements (Transcription factor binding sites or DNA motifs) are short and subtle genomic segments that can be recognized by a specified group of proteins (e.g. transcription factors). Most of the time, regulatory elements are found in non-coding regions (referred to promoters, UTRs and introns) of genes. The interaction between transcription factor binding sites (TFBSs) and transcription factors (TFs) determines the transcriptional activity and dominates

⋆ Corresponding author.

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 94–105, 2010.

the regulation level of gene expression. TFBSs from a set of co-expressed genes usually share a common conserved pattern that is recognized by a particular transcription factor which can lead to the same biological function. As a major complement to the traditional wet-lab identification methods (e.g. DNas footprinting [1]), computational algorithms have shown the good potential on the problem solving in terms of time and cost. Due to the high-throughout genome sequencing and microarray technologies, the dramatically increased number of complete genome sequences and large-scale gene expression data have made the computational approaches become available to cope with the growing needs on motif discovery.

Since the motifs usually have nucleotides with low divergence on the binding positions, current computational explorations aim to capture this biological feature by developing advanced motif models and implementing with efficient procedures. The searching algorithms can be classified into exhaustive approaches and heuristic methods. The exhaustive approaches, such as CONSENSUS [2], enumerate all possible combinations from the entire search space to maximize the model quality with the most statistical over-representations. Motifs predicted by the heuristic methods are usually inferred from the probabilistic model with the optimized parameters after a number of searching iterations, such as MEME [3]. According to the performance assessment from [4] and [5], developing advanced algorithms to produce the satisfactory results on both eukaryotic and prokaryotic genomes is still a major assignment for computational biologists.

Evolutionary Computation (EC) techniques have been recently introduced to the domain of motif discovery ([6], [7], [8]), which have shown some promising improvements on prediction accuracy. In our previous work, we proposed an EC-based motif discovery tool named MODEC (MOtif Discovery using Evolutionary Computation) [9], which applied the seed concept to group similar patterns for noise data elimination and employed a combined metric of mismatch model and 3rd-order Markov chain as fitness function to evolve the possible candidates. Comparative studies on eight experimental datasets have demonstrated that MODEC achieves a better prediction performance over than three commonly used algorithms as well as two state-of-the-art GA applications.

In many cases, people target at promoters of the co-expressed genes since regulatory elements usually hide within the promoters. Recent studies have found the appearances of regulatory elements in introns and UTRs ([10],[11]). The theory behind that is the regulatory elements in non-coding regions are usually under a higher selective pressure than non-functional segments during the evolution. Thus, a rich discovery approach should investigate the distributions of any possible functional elements not only in promoters but also in intron and UTRs.

The small flowering species *Arabidopsis thaliana* is well-known as a model plant in molecular biology and genetics [12]. As the first fully sequenced plant genome, *Arabidopsis thaliana* has only five chromosomes with about 157 million base pairs and 27,000 genes. Both the small genome size and rich annotated genome resources [13] have made *Arabidopisis thaliana* ideal for the study of functional regulatory elements in non-coding regions.

In this study, to perform the motif mining task in non-coding regions of co-expressed genes from *Arabidopsis thaliana*, we collect the information of gene expression clusters published by a recent work [14] along with their non-coding sequences from TAIR [13]. MODEC and the well-known EM-based motif discovery tool MEME are chosen to carry out the motif prediction. With the respect to algorithm development, a new motif model quality metric MAR-G is proposed in MODEC, which is used to characterize the motif features from a combination of different perspectives. The prediction results show MODEC is capable of finding expected patterns in target co-expressed gene groups. Also, the patterns predicted by MODEC show similar enriched distributions across different non-coding regions against MEME. Those enriched words along with their distribution studies will help with understanding the gene functions of the co-expressed networks in *Arabidopsis thaliana*.

2 MODEC

2.1 Motif Model Representation

A motif model is usually referred as an abstract representation that summarizes a set of collected subsequences with the identical length k. Such a subsequence can be defined as a k-mer. In here, each k-mer is described as a binary matrix which adopts the k-mer binary encoding approach from [15]. That is, for a single k-mer, i.e., $B_1B_2\cdots B_k$, we denote $e(k\text{-mer}) = [a_{ij}]_{4\times k}$, $a_{ij} = 1$ if $B_j = V_i$, otherwise $a_{ij} = 0$, where $(V_1, V_2, V_3, V_4) = (A, C, G, T)$. For example, the 7-mer AGCGTGT can be encoded as:

$$e(K) = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 1\,0\,0\,0\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0\,0 \\ 0\,1\,0\,1\,0\,1\,0 \\ 0\,0\,0\,0\,1\,0\,1 \end{bmatrix}$$

Position Frequency Matrix (PFM) as one commonly employed representation [16] is a $4\times k$ matrix, where each entry of a given PFM is assigned by the relative frequency of a particular nucleotide $\{A, C, G, T\}$ at the particular position of the aligned sequences. Due to the significance of PFM in motif discovery, we choose it as the motif model representation. Let S be a collection of k-mers $\{K_p : p = 1, 2, \cdots\}$, the PFM model M of S can be given by:

$$M = \frac{1}{|S|} \sum_{K_p \in S} e(K_p), \tag{1}$$

where $|S|$ represents the cardinality of the set S.

2.2 Model Quality Metrics

Relative Model Mismatch Score (RMMS). Model Mismatch Score (MMS) proposed in [15] is a mismatch-based metric that is used to quantify the

conservation property of the motif model. The model mismatch score (MMS) of S can be expressed as follows:

$$\text{MMS} = \frac{1}{|S|} \sum_{K_p \in S} d(K_p, M_s). \tag{2}$$

where $d(\cdot, \cdot)$ is a generalized Hamming distance function that measures the mismatch between a k-mer $\in S$ and the PFM model M_s of S.

As an extension of MMS, relative model mismatch score (RMMS) has been developed recently in [17] and successfully applied in [9]. The strength of RMMS is able to reflect both the conservation property and the rareness of the binding sites with respect to the non-functional sequences. The expression of the RMMS is given by,

$$\text{RMMS} = \frac{1}{|S|} \sum_{K_p \in S} R(K_p, M_s), \tag{3}$$

where

$$R(K_p, M_s) = \frac{d(K_p, M_s)}{d(K_p, M_b)}. \tag{4}$$

Here, $R(\cdot, \cdot)$ represents the relative distance measure of a k-mer to both the PFM M_s of S and the PFM M_b of the background model B. The M_b can be pre-calculated by using the genome sequence of a specific species or constructed by the whole input sequences.

Markov Background Model. In this study, we adopt the 3rd-order Markov chain to calculate the background probability of a k-mer. Examples have been given by [18] which demonstrate the high order Markov chain model has the advantage to take the context dependency of DNA sequences into the measurement of background rareness. The formula of a 3rd-order Markov chain can be defined as:

$$\begin{aligned} p_0(d_1, d_2, \cdots, d_{k-1}) = {} & p(d_1)p(d_2 \mid d_1)p(d_3 \mid d_1, d_2) \\ & \times \prod_{h=1}^{k-3} p(d_{h+3} \mid d_h, d_{h+1}, d_{h+2}) \end{aligned} \tag{5}$$

Model α-ratio Score with GC-content (MAR-G). In [9], a mixture quality metric Model α-ratio (MAR) is developed which incorporates the 3rd-order Markov model into the RMMS. MAR has shown the ability to capture the rareness property of motif comparing with the randomness. In recent studies, people address that GC-content can be an important indicator to a couple of genomic features and some of them closely correlate with the binding activity, such as sequence repetitive region, and CpG islands [19].

The GC-content of a genomic sequence can be simply calculated as:

$$GC = \frac{G + C}{A + T + G + C} \tag{6}$$

The information of GC-content is necessary to be monitored during the motif discovery process. We include the GC-content into MAR and term the extended metric as MAR-G.

Having a dataset D and motif model S, the α-ratio score of a given k-mer K in D is first computed. The formula is given as:

$$\alpha(K) = \frac{log(p_0(K))}{R(K, M_s)}, \tag{7}$$

where $K \in D$, $R(K, M_s)$ is the RMMS of K and $p_0(K)$ is the background probability by the 3rd-order Markov chain of K.

Suppose K is from an input sequence Q of D, we can easily have the GC-content value of Q. Then, the (7) can be further extended as:

$$\alpha(K) = \frac{log(p_0(K))}{R(K, M_s) \times GC_Q}, \tag{8}$$

where $K \in Q$ and $Q \in D$.

Since the coding region of a gene usually shows a higher GC-content than the non-coding regions [19], a given k-mer with a smaller α-ratio indicates it has a greater possibility to be a binding site from the non-coding regions.

With one assumption that the existence of binding sites is independent with each other, the background probability P_s of model S can be expressed as:

$$log(P_s) = \sum_{p=1}^{n} log(p_0(K_p)). \tag{9}$$

where n is the total number of k-mers in S.

Based on (8), the α-ratio of S can be given as below:

$$\alpha(S) = \frac{log(P_s)}{\sum_{K_p \in S} R(K_p, M_s) \times \sum_{Q_p \in S} GC_{Q_p}}. \tag{10}$$

Then by scaling the above ratio using logarithm, we can define the Model α-ratio score with GC-content (MAR-G) of S as:

$$\text{MAR-G}(S) = log(-log(P_s)) - log(\sum_{K_p \in S} R(K_p, M_s)) - log(\sum_{Q_p \in S} GC_{Q_p}). \tag{11}$$

MAR-G can be regarded as a variation of RMMS with the heuristically joint background probability and GC-content. A motif model with a high MAR-G score is supposed to have a high degree of conservation associated with the low randomness. MAR-G is served as the model quality metric in MODEC.

2.3 Algorithm Description

MODEC has three main components: data pre-processing, evolutionary computation process and model post-processing. Detailed algorithms are described in [9]. The overall framework of MODEC is introduced in following.

Data Pre-processing. Two stages are applied here before the core evolutionary searching algorithm, which are Potential Core Collection (PCC) and Search Space Reduction (SSR).

The purpose of PCC is to collect some conserved k-mers together as the initial motif models. The seed concept is applied here to group similar k-mers from the input sequences. We first randomly select a number of sequences from the input datasets. Then, for each k-mer from the chosen sequences, a core is formed which contains a number of k-mers that take the minimal hamming distance to that k-mer. For each core, an alternative pattern AP is produced by the nucleotide with the highest frequency from each column of its PFM. The cores with the same AP are merged. The merged cores are then ranked by MAR-G and the top H cores are used as the seed models.

For each seed model C, a filtering process is applied to remove k-mers that have large α-ratio scores against C. First, we set the filtering threshold δ which is the maximum α-ratio from the k-mers in C. Since C have the potential to contain true binding sites, whose scores are considerably smaller than the random k-mers. Thus, a k-mer that has a greater score than δ will be considered as a background k-mer and be eliminated. The filtering rule is given as:

$$\text{If } \alpha(K) > \delta, \text{Then, } K \text{ is discarded, where } K \in D. \tag{12}$$

The reason to produce multiple seed models is to reduce the opportunity of false dismissal. The SSR is applied to each seed model against the dataset individually. After the process, a number of pools that contain reduced k-mers are kept. The evolutionary process runs through each pool separately to enlarge the chance of optimal solutions.

Fitness Function. In the domain of motif discovery, the fitness function of EC is used to evaluate the predicted motif quality which is the fitness of each chromosome. Though a couple of variations have been proposed recently ([6] and [8]), a fitness function that can perfectly catch the motif features is still a major challenge. In this study, we use MAR-G as the fitness function to measure the chromosomes and attempt to produce sound solutions over generations. The optimal chromosome from the final generation is supposed to maximize the proposed fitness function, that is:

$$\max f(U) = MAR\text{-}G(U), \tag{13}$$

where U is a chromosome.

Sometimes a model with high conservation property is produced by a group of highly repetitive segments. In this study, we apply a complexity score function that is designed to exam the complexity degree of the compositional structure of a given motif model [20]. During the evolutionary process, a chromosome with a lower complexity score than the pre-defined threshold value will be excluded from the population even though it may have a high MAR-G score. The complexity measure is given as:

$$c(M) = \left(\frac{1}{4}\right)^k \prod_{b=A}^{T} \left(\frac{k}{\sum_{j=1}^{k} f(b,j)}\right)^{\sum_{j=1}^{k} f(b,j)}, \tag{14}$$

where $f(b, j)$ is the relative frequency of nucleotide b in position j of a given PFM.

Evolutionary Process. The whole EC process starts with the chromosome construction and population initialization. Each chromosome is represented as a vector $\{vc_1, vc_2, \cdots, vc_n\}$, where vc_i is a k-mer from the i-th input sequence Sq_i and here n denotes the total number of input sequences. Each chromosome is considered as a potential motif model. During the population initialization, each time a chromosome is formulated by a collection of k-mers which take the minimum α-ratio against the PFM generated by a number of randomly chosen k-mers from the seed model.

The conventional roulette-wheel selection is applied to choose parents for reproduction. Two genetic operators *Crossover* and *Replacement* are used to assist the reproduction process and to maintain the evolution in a stable scale. The reproduction will not be terminated until the size of the population is doubled. Both winners-take-all selection and tournament selection are selected in order to maintain the good solutions as well as keep the generations from premature convergence. The detailed description of the evolutionary process is introduced in [9].

Model Post-processing. In MODEC, a model post-processing is developed to finalize the chromosomes from the last generation once the evolution process meets the termination criteria. The detailed process is described in [9], which comprised three components: Merging, Adding and Removing. A short introduction is given here for completeness.

The process starts with Merging that is to group similar chromosomes for reducing the size of solution space. Each chromosome now is termed as a candidate. Candidates with the same alternative pattern (AP) will be first merged together. We then rank the merged models based on RMMS. Starting from the top model in the ranked list, models that have a small information content difference (0.01) against it are merged together. We repeat this process till the end of the list. The problems caused by Merging could be k-mer duplications. To deal with the k-mer duplications, we simply remove all the duplicated k-mers by keeping only one occurrence.

"One k-mer Adding and Removing" is applied after Merging which aims to improve the problem of k-mer mis-assignment and to find weak true binding sites. For a given model M, each time one k-mer that has the smallest α-ratio against M is selected from the pool. This k-mer will be added to M if there is no MAR-G reduction after the adding, which is ΔMAR-G =: MAR-G$_{new}$ − MAR-G$_{old}$ >= 0. The Removing is triggered right after Adding. A k-mer of M that has the largest α-ratio will be removed only if there is non-decrease of MAR-G of M. The "One k-mer Adding and Removing" continues iteratively till no further quality improvement of M.

The model refinement process extends the search ability of finding weak binding sites as well as purifying motifs by discarding false instances.

3 Results

3.1 Data Preparation

To fulfill the motif discovery task in the co-expressed gene groups of *Arabidopsis thaliana*, we collect the gene expression information from ATCOECIS (http://bioinformatics.psb.ugent.be/ATCOECIS/), which holds the analysis results of predicted regulatory elements and their functional categories within the gene co-expression networks of *Arabidopsis thaliana* [14]. In total, there are 19,064 co-expressed clusters. Each of them is formed by a clustering method which groups highly correlated genes based on their expression patterns.

Table 1. Top ten motifs associated with Gene Ontology (GO) description

Motif pattern	Motif Name	GO Label	GO Description
AAACCCTA	TELO	GO:0042254	Ribosome biogenesis
CTTATCCN	Ibox	GO:0015979	Photosynthesis
GGCCCANN	UP1	GO:0042254	Ribosome biogenesis
GCCACGTN	Gbox	GO:0015979	Photosynthesis
GCGGGAAN	E2F	GO:0006260	DNA replication
GACCGTTN	MSA	GO:0007018	Microtubule-based movement
AANGTCAA	Wbox	GO:0050832	Defense response to fungi
CNGATCNA	AGMOTIFNTMYB2	GO:0048193	Golgi vesicle transport
NCGTGTCN	ABA-responsive element	GO:0009737	Response to ABA stimulus
CATGCANN	RYREPEATBNNAPA	GO:0048316	Lipid transport

According to the work from [14], the information of ten most enriched motif patterns is listed in Table 1. Each of them drives the genes with specific functional roles during the expression. For example, the Ibox, Gbox and Wbox can be found in genes involved in photosynthesis, stress response and defense response. The gene clusters which are found to highly correlate with the ten motif patterns are further processed by removing the clusters with the small number of genes less than $10^{1.5}$ or with low co-expressed profiles (the correlation coefficient less than 0.70). After the filtering, we collect top ten gene clusters that have the significant enrichment of the corresponding motif patterns. Overall, 100 clusters are used to serve the motif discovery in this study.

The Arabidopsis Information Resource TAIR (http://www.arabidopsis.org) holds the most up-to-date information of *Arabidopsis thaliana* such as the complete genome sequence, gene expression profiles, and detailed gene annotation. The promoter regions, 3' UTRs and 5' UTRs of genes are collected from TAIR version 9. The promoter regions are further divided into proximal promoter [0, 500] and distal promoter [0, 1000]. The overall properties of four non-coding districts are shown in Table 2.

Table 2. Statistics of the four non-coding districts

District	No. of Sequences	Min Seq. length	Max Seq. length	Nucleotides
3' UTRs	17,162	10	3164	2,096,659
5' UTRs	16,341	10	2435	1,164,447
P. promoters	7839	500	500	3,919,500
D. promoters	7839	1000	1000	7,839,000

3.2 Experimental Discussions

By running the popular EM-tool MEME and the proposed framework MODEC, we collected the prediction results in non-coding regions of the 100 selected gene clusters with the expectation that the top ten motifs can be found as well as some interested patterns.

For both MODEC and MEME, we set the number of output motifs to 10 and length of motif to 10. For MODEC, to enlarge the opportunity of global optimization, we adjusted three key parameters in evolutionary computation which were population size (1000), number of generation (1000) and replacement rate (0.8). For each dataset, we ran MEME and MODEC two times. In each run, we kept top ten predicted models from each algorithm. The predicted words were then compared with the known plant binding sites in ATCOECIS.

The results show that the expected patterns can be predicted by the two algorithms and mainly detected in promoters of the corresponding gene clusters, though some of the predicted patterns show subtle diversities against the targets. As examples to display, the predicted patterns of Gbox (GCCACGTN), Wbox (AANGTCAA), TELO (AAACCCTA) and AGMOTIF (CNGATCNA) are shown in Table 3. We noticed that, in the UTR's, the appearances of the expected patterns are less than those in promoters. It shows the evidence that the functional elements usually locate in the promoters rather than UTRs. The comparisons between promoters and UTRs also demonstrate that AT enriched words such as AATTTTT, AAAAAAT, ATTTTTA are often found in UTRs rather than promoters, while TATATAA can be detected in promoters which is quite likely to be the binding sites of RNA polymerase II (TATA-box). Also, the predicted patterns from MEME tend to be repetitive with high conservation, while those from MODEC prefer to be less conserved but high diversity.

In addition to discover the expected motifs from target gene clusters, we are also interested to provide some preliminary analysis and comparisons on the words predicted in the non-coding regions. All predicted models were ranked by their model quality scores. Some highly repetitive patterns were then filtered out. The top five most enriched words (8-mers) found in UTRs and promoters by MEME and MODEC are given in Table 4. We find that the distributions of those words are with a high degree of similarities. Since the genome of *Arabidopsis thaliana* is believed to be AT-rich, it is not surprised to discover patterns with high A/T occurrences. Many common patterns such as AAAACAAA and TCTTTTTC are found by both algorithms in the same district, while some of

Table 3. Sequence logos for four predicted motifs

Motif	MEME	MODEC
Gbox		
Wbox		
TELO		
AGMOTIF		

Generated by WebLogo (http://weblogo.berkeley.edu/logo.cgi)

Table 4. Top five most enriched words in non-coding regions by MEME and MODEC

MEME			MODEC		
3' UTRs	5' UTRs	Promoters	3' UTRs	5' UTRs	Promoters
TTTTGTTT	AAAACAAA	TCTTTTTC	TTTAATTT	AAAACAAA	TCTTTTTC
GTTTCTTT	TTCTCTCC	GAAAAAGA	TTTTAATT	TCAAATCA	ATTTTTTA
AAGAAGAA	AAAGAAAA	AGGCCCAA	TTTTGTTT	AAAAGAAA	TATAAAAT
TTTCTTCT	AAAACCAA	TTGGGCTT	TTCTCTTC	AAAACTAA	AGGCCCAA
AAAACAAA	AAAGCAAA	AAACCCTA	TATATTTA	TTTTTACT	TTGGGCTT

them are partially similar, such as AAAGCAAA (by MEME) and AAAACAAA (by MODEC) in 5' UTRs, which can be a complementary resource for further process by using a combinational ensemble analysis.

4 Conclusion

This work applies our recently proposed framework MODEC to perform practical motif discovery on the model plant *Arabidopsis thaliana*. The main contributions in this paper include: i) a variation of RMMS, namely MAR-G is developed by combining both the background probability and GC-content score to measure

the motif model quality; ii) we extend the mining regions from promoters to UTRs, which provides an initial insight into the comparative study of regulatory elements over non-coding regions; iii) with the support of well-annotated genome structure information , a relatively large-scale motif mining process is carried out in its non-coding regions (promoters and UTRs).

As the results shown in the last section, the expected motif patterns are returned by both MODEC and the state-of-the-art approach MEME with the low degree of dissimilarities. Top five most enriched words predicted by both algorithms demonstrate the pattern diversities across different non-coding districts in *Arabidopsis thaliana*, which give an important clew that different functional elements may be found in different non-coding regions. Based on the comparison with MEME, MODEC also proves the capability of performing motif prediction in practice.

A detailed analysis will be carried out which mainly focuses on the location distributions of the predicted motifs and their possible functional categories. Also, as the future extension of this study, the same discovery process is going to apply to introns in *Arabidopsis thaliana* which are another important district for gene regulation.

References

1. Galas, D.J., Schmitz, A.: DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 5, 3157–3170 (1978)
2. van Helden, J., André, B., Collado-Vides, J.: Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J. Mol. Biol. 281, 827–842 (1998)
3. Bailey, T.L., Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28–36. AAAI Press, Menlo Park (1994)
4. Tompa, M., Li, N., Bailey, T.L., et al.: Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotechnology 23, 137–144 (2005)
5. Hu, J., Li, B., Kihara, D.: Limitations and potentials of current motif discovery algorithms. Nucleic Acids Res. 33, 4899–4913 (2005)
6. Chan, T.-M., Leung, K.-S., Lee, K.-H.: TFBS identification based on genetic algorithm with combined representations and adaptive post-processing. Bioinformatics 24, 341–349 (2008)
7. Li, L.P., Liang, Y., Bass, R.L.L.: GAPWM: a genetic algorithm method for optimizing a position weight matrix. Bioinformatics 23, 1188–1194 (2007)
8. Wei, Z., Jensen, S.T.: GAME: detecting cis-regulatory elements using a genetic algorithm. Bioinformatics 22, 1577–1584 (2006)
9. Li, X., Wang, D.H.: Computational Discovery of Regulatory DNA Motifs Using Evolutionary Computation. In: CEC-IEEE 2010: IEEE Congress on Evolutionary Computation (accepted 2010)
10. Fiume, E., Christou, P., Giani, S., Breviario, D.: Introns are key regulatory elements of rice tubulin expression. Planta 218, 693–704 (2004)
11. Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., Kellis, M.: Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434, 338–345 (2005)

12. Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D., Koornneef, M.: *Arabidopsis thaliana*: a model plant for genome analysis. Science 282, 662–682 (1998)
13. Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., et al.: The arabidopsis information resource (TAIR): gene structure and function annotation. Nucleic Acids Res. 36, D1009–D1014 (2008)
14. Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L., Van de Peer, Y.: Unraveling transcriptional control in arabidopsis using cis-regulatory elements and coexpression networks. Plant Physiol. 150, 535–546 (2009)
15. Wang, D.H., Lee, N.K.: MISCORE: mismatch-based matrix similarity scores for DNA motif detection. In: Köppen, M., Kasabov, N., Coghill, G. (eds.) ICONIP 2008. LNCS, vol. 5506, pp. 478–485. Springer, Heidelberg (2008)
16. Benos, P.V., Bulyk, M.L., Stormo, G.D.: Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res. 30, 4442–4451 (2002)
17. Wang, D.H.: Characterization of regulatory motif models. Technical Report, La Trobe University, Australia (October 2009)
18. Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouzé, P., Moreau, Y.: A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17, 1113–1122 (2001)
19. Galtier, N., Piganeau, G., Mouchiroud, D., Duret, L.: GC content evolution in mammalian genomes, the biased gene conversion hypothesis. Genetics 159, 907–911 (2001)
20. Mahony, S., Hendrix, D., Golden, A., Smith, T.J., Rokhsar, D.S.: Transcription factor binding site identification using the Self-Organizing Map. Bioinformatics 21, 1807–1814 (2005)

Prediction of Non-coding RNA and Their Targets in *Spirulina platensis* Genome

Tanawut Srisuk[1,2], Natapol Pornputtapong[2], Supapon Cheevadhanarak[2,3], and Chinae Thammarongtham[4]

[1] Bioinformatics and Systems Biology Program, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
[2] Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi, Bangkhuntien, Bangkok, 10150, Thailand
[3] School of Bioresources and Technology, King Mongkut's University of Technology Thonburi, Bangkhuntien, Bangkok, 10150, Thailand
[4] Biochemical Engineering and Pilot Plant Research and Development Unit, National Center for Genetic Engineering and Biotechnology at King Mongkut's University of Technology Thonburi, Bangkhuntien, Bangkok, 10150, Thailand
tanawut.cool@gmail.com

Abstract. Non-coding RNAs (ncRNAs), transcripts that have function without being translated to protein, have a number of roles in the cell including important regulatory roles. Efforts to identify the whole set of ncRNAs and then to elucidate their functions would gain better biological understanding. Although ncRNA is another type of genome constituent, most of the genes for ncRNA are overlooked by standard genome annotation of genome sequencing projects. This also happens in *Spirulina platensis* genome sequencing project. It is because gene finding tools generally are able to identify only protein-coding genes but not non-protein-coding ones. In this study, *S. platensis* ncRNAs were detected by comparative genomics approach using computational tools, together with RNA secondary structure prediction. It was found that more than 100 predicted ncRNA loci matched with known ncRNAs for example cobalamin riboswitch, RNaseP, Signal Recognition Particle RNA, Group II intron RNA and Yfr1. It has been reported that Yfr1 has been found in most cyanobacterial genomes sequenced. The result showed that more than 70 putative loci were similar to Group II intron RNAs. In addition, approximately 100 predicted ncRNA loci were not matched with any known ncRNAs. The predicted targets for some putative ncRNAs are also proposed.

Keywords: non-coding RNA prediction, non-coding RNA target prediction.

1 Introduction

Besides protein-coding genes, the genes coded for these RNAs have also been recognized as genome constituents since a large fraction of the transcriptome consists of non-protein-coding RNAs [1]. They are involved in many biological processes such

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 106–117, 2010.

as transcriptional regulation, chromosome replication, RNA stability and translational regulation, and even proteins stability and translocation [2]. Some of them act as catalytic molecules. Consequently, efforts to identify the whole set of ncRNAs and then to elucidate their functions for better biological understanding are more and more prominent. Availabilities of complete genome sequence data have made it possible to computationally identify ncRNAs in sequenced genomes using bioinformatics approach. Although experimental verification is necessary, it has been recognized that computational identification may be an effective approach to first detect ncRNAs candidates, including novel ncRNA species, followed by biochemical assessment. RNA classification is a step involving in RNA gene annotation. Since ncRNAs are conserved in structures rather than their primary sequences, using a simple homology search is not efficient enough in RNA classification. Comparing structures of unknown ncRNAs, along with similarity at primary sequence level, to known ncRNAs is more powerful. To perform this strategy, a statistical model so-called covariance-model (CM) is a method of choice since CMs are integrated with both primary sequence and secondary structure information of known ncRNAs. CMs will be trained and then be used to search for the region in given sequences which are similar to the feature it has learned [3]. Infernal package [4] is a suite of tools for these tasks ranging from CM construction to searching for the region which is similar to CM in given sequences.

ncRNAs are a heterogeneous group of functional RNAs and showing up in all kingdoms, including prokaryotic domain. In bacteria, ncRNAs mostly function as coordinators of adaptation processes in response to environmental changes, integrating environmental signals and controlling target gene expression. For cyanobacteria, including *Spirulina,* regulatory circuits involving ncRNAs can be expected as well. Identification of ncRNAs will facilitate investigation of another level of controls in *Spirulina,* in addition to regulation by protein mediators, and providing new insights into growth and adaptation to stresses of this cyanobacterium. *Spirulina* genome sequencing project has been established by Thai research consortium. Availabilities of such cyanobacterial genome information together with bioinformatics approach make it possible to computationally identify putative ncRNA genes. In this study, the *Spirulina* genome were computationally analyzed and screened for ncRNAs. Their putative targets were also predicted.

2 Materials and Methods

Sequences of intergenic regions (IG) of *S. platensis* were extracted from genome sequence data and compare with sequences of 34 related species genomes, including *Nostoc punctiforme* ATCC 29133 [5], *Prochlorococcus marinus* subsp. marinus str. SS120 [6], *Synechococcus* sp. JA-3-3Ab, *S.* sp. JA-2-3B'a(2-13), *S.* sp. WH 8102, *S.* sp. CC9902, *S.* sp. CC9605, *S.* sp. CC9311, *S.* sp. WH 7803, *S.* sp. RCC307, *S.* sp. PCC 7002, *S. elongatus* PCC 6301, *S. elongatus* PCC 7942, *Synechocystis* sp. PCC 6803, *Chlorobium tepidum* TLS, *N.* sp., *Thermosynechococcus elongatus* BP-1,

P. marinus str. MIT 9211, *P. marinus* str. MIT 9215, *P. marinus* str. MIT 9301, *P. marinus* str. MIT 9303, *P. marinus* str. MIT 9312, *P. marinus* str. MIT 9313, *P. marinus* str. MIT 9515, *P. marinus* str. NATL1A, *P. marinus* str. NATL2A, *P. marinus* str. AS9601, *P. marinus* subsp. pastoris str. CCMP1986, *Gloeobacter violaceus* PCC 7421, *Anabaena variabilis* ATCC 29413, *Trichodesmium erythraeum* IMS101, *Acaryochloris marina* MBIC11017, *Microcystis aeruginosa* NIES-843, and *Cyanothece* sp. ATCC 51142 [7]. Multiple sequence alignments between *S. platensis* IGs and corresponding regions from other genomes were constructed. These alignments were scores, by ncRNA prediction tool, for possibility of being ncRNAs. This procedure was also applied to *Arthrospira maxima* CS-328 and *Lyngbya* sp. PCC 8106 [8]. The methodology of this work is outlined in the work flow shown on Fig. 1.

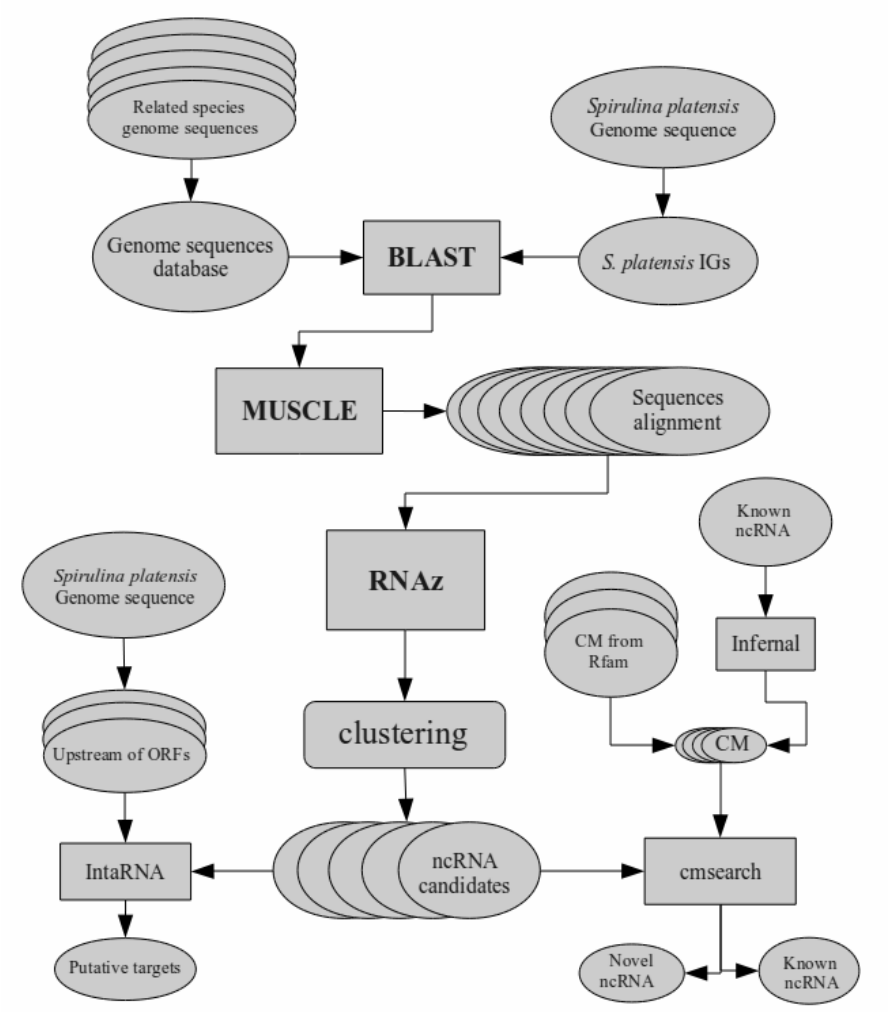


Fig. 1. Work flow of the proposed methodology

2.1 BLAST Based Alignment Generating

To prepare multiple sequence alignment as input for ncRNA prediction program, IG sequences (include 50 base from preceding and consequence ORFs) were searched against cyanobacteria genome database for similar regions in other species using ncbi BLASTN [9] with relaxed parameter (-q -1 -r 1 -F F -e 1e-10) because ncRNAs tend to conserved in their structures rather than primary sequences. Similar regions of each *S. platensis* IGs were grouped by their lengths and positions on *S. platensis* genome then aligned with each IG by MUSCLE [10] with increased gap penalty to avoid gap insertion in distantly related sequences.

2.2 RNAz Scoring and Result Clustering

RNAz [11] package, an ncRNA prediction tool, was used to calculate probability of being ncRNA for each alignment input. Before scoring, each alignment was pre-processed, by scripts in RNAz package, into appropriate form. After each input was scored, results with ncRNA probability score (RNAz P-value) > 0.9 (the closer to 1 the more likely to be ncRNA) were kept. To recover some part of long ncRNA which may lost in prediction due to sliding windows of 40 nt (nucleotides) in preprocessing step, they were also joined into a single ncRNA locus if their locations were over-lapped or located within 40 nt from each other.

2.3 ncRNA Candidate Classification

In order to classify which individual ncRNAs belong to known ncRNAs, each candidate was compared with covariance model (CM) from Rfam database [12] and several our-owned constructed CMs of known ncRNA using cmsearch from INFERNAL package. The resulting matches with E-value lower than 1e-5 were reported as true known ncRNA homologues.

2.4 ncRNA Targets Prediction

To predict targets of each ncRNA candidate, in term of translational regulation, an interaction score between each ncRNA candidate and 5' upstream region (250 nt before start codon to 150 nt after start codon) of predicted ORFs were calculated by IntaRNA [13] then top rank scores from each ncRNA candidate were reported.

3 Results and Discussion

3.1 Predicted ncRNA in *S. Platensis* Genome

A set of 3,976 IGs were searched by BLAST against 34 related species genomes then 10,639 of BLAST hits were selected and grouped before aligning into 2003 alignment. Using RNAz, 334 putative ncRNA loci were computationally identified from *S. platensis* genome. The lengths of these loci are varied between 52 and 1482 nt. Some of particular loci predicted may be only partial, not full length, ncRNAs. Some of them can be merged together into a single larger locus for example RNaseP RNAs and Group II intron RNAs. Based on the method used in this work, the transcription direction of the predicted loci were not determined. Localization of promoter and

terminator of each locus will reveal transcription direction and also increase reliability of putative ncRNA candidates. However experimental identification of promoters is relatively complicated and computational prediction is still challenging. Notably this work primarily focused on ncRNAs located on IGs. Antisense transcripts were not detected by the method used. Such anti-sense ncRNAs can be predicted by CM searching or other methods. The statistical value of the predicted loci is RNAz P-value which is between 0 and 1. The default P-value for a locus to be reported as RNA is 0.5 or higher. The higher P-value the more significant the locus is, inferring a plausible secondary structure forming locus. For all 334 predicted loci, their P-values are 0.9 or above. Therefore they are considered to be putative ncRNA candidates.

3.2 Classification of ncRNA Candidates

By CM searching, 129 ncRNA candidates were matched with CM obtained from Rfam and classified into 12 known RNA families including of 1 5.8S rRNA, 2 Cobalamin riboswitches, 1 CRISPR-DR57, 79 group II introns, 1 mir-598, 2 PK-G12 rRNAs (23S rRNA pseudoknot), 1 bacterial RNaseP type A, 1 bacterial signal recognition particle (SRP), 3 SSU-5s, 37 tRNAs, 1 Yfr1 and 1 Yfr2b. Several long length RNA loci matched to more than 1 family. Interestingly, RNA candidates in a large group containing 79 loci are classified as Group II intron RNA.

3.2.1 Yfr1

Yfr1 is an ncRNA which exists in many cyanobacteria and locates between *trxA* and *guaB*. It is approximately 50-70 nt in length. It has been found that Yfr1 consists of approximately 10-nt conserved unpaired region (5'-ACUCCUCACAC-3') between two stem loop structures [14]. It has been reported that Yfr1 is required for growth under stress condition and target to SbtA mRNA which plays an important role for sodium-dependent bicarbonate transport. In addition, estimated abundance of Yfr1 is about 18,000 molecules per cell [15]. Furthermore, Yfr1 inhibit translation of two outer membrane protein genes by direct base pairing mechanism at ribosome binding site [16]. In *S. platensis* and *A. maxima* genomes, the putative Yfr1 was predicted as a locus resided between thioredoxin and IMP dehydrogenase gene in the genome. This agrees with the report when the Yfr1 was first identified. Fig. 2 represents cmsearch result which indicates matched ncRNA candidate of Yfr1 CM.

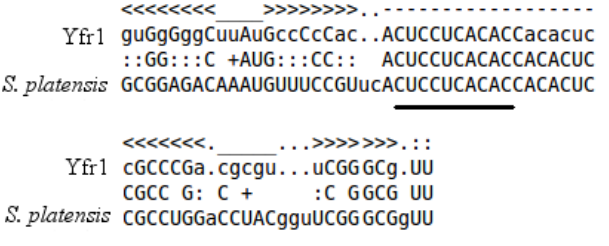


Fig. 2. Alignment from "cmsearch" represents matching between Yfr1 CM from Rfam and ncRNA candidate in *S. platensis*. Middle line between Yfr1 and ncRNA represent matching, either in primary sequence or secondary structure as described in [17]. Conserved region according to previous report is indicated by underlining.

3.2.2 Group-II intron

Bacterial group II intron RNAs are mobile retro-element and catalytic unit which is spliced by lariat intermediate mechanism. These introns transferred to target site by ribonucleprotein complexes assembled from intron-encoded proteins and excised intron RNA lariat. Group II intron RNAs are a large biomolecule consisting of highly structured RNAs with six distinct double-helical domains and reverse transcriptase ORF in fourth domain [18]. There are 79 and 89 ncRNA candidates, which are identified as group II intron RNA, in *S. platensis* and *A. maxima*, respectively. Many loci of group II intron RNAs in *S. platensis* are located near reverse transcriptase or transposase coding sequences. Since the CM for group II intron RNAs is constructed from partial group II intron RNA sequences which is conserved within this RNA family, the hits returned from CM searching are reported as partial sequences matched with the CM. The full length of Group II intron RNA can be traced by investigated neighboring loci predicted as ncRNAs and shared similarity with other regions of group II intron RNA. Fig. 3 represents alignment of group II intron CM to match locus on *S. platensis* genome.

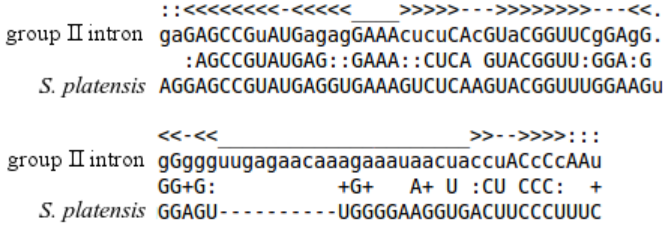


Fig. 3. Alignment from"cmsearch" result represents matching between an ncRNA candidate in *S. platensis* and CM of a group II intron partial structure from Rfam.

3.2.3 Cobalamin riboswitch

Cobalamin riboswitch is a conserved regulatory element located at 5' untranslated region (UTR) of vitamin B_{12} related genes [19]. In *S. platensis* and *A. maxima* genomes, there are three loci of putative cobalamin riboswitchs. A locus of *S. platensis* cobalamin riboswitch locates on 5' UTR of 5-methyltetrahydropteroyltriglutamate-homocysteine S-methyltransferase (MetE) homologous gene. In *Mycobacterium tuberculosis*, cobalamin riboswitch resided on 5' UTR of MetE gene and involved in transcriptional control of the gene [20]. Another predicted locus in *S. platensis* genome matched with cobalamin riboswitch located upstream of cobalamin biosynthesis protein CobW gene [21]. In addition, a possible cobalamin riboswitch was predicted in *S. platensis* genome as it matched with CM of the one reported to be located on 5' UTR of cobalamin biosynthesis protein CbiM gene [22]. However, this locus matched to the CM with E-value higher than 1e-5 which was lower than the cut-off for this work. Fig. 4 represents cmsearch result, indicated how the cobalamin riboswitch CM matched to ncRNA candidate.

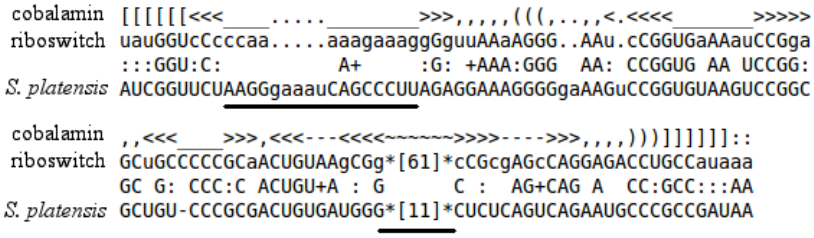


Fig. 4. Alignment from cmsearch result represents matching between Cobalamin riboswitch CM and ncRNA candidate in S. platensis, miss-match regions are indicated by underlining.

3.2.4 Signal Recognition Particle RNA (SRP)

SRP is a ribonuclecoprotein complex, composed of an RNA component and one or more protein domain(s). SRP is required for secretion or integration to membrane for proteins those consisted of signal peptide [23]. Single copy of SRP RNA was classified in each genome of *S. platensis*, *A. maxima, Nostco* sp. PCC 7120, and *Lyngbya* sp. PCC 8160. In *S. platensis* and *A. maxima*, SRP RNA candidate are adjacent to downstream of YCII-related gene and these loci from both species are identical. According to genome annotation, homologues of signal recognition particle protein subunit Ffh and signal recognition particle-docking protein FtsY which are involved in proteins translocation machinery exist in *S. platensis* genome. Fig. 5 represents "cmsearch" result, indicates how ncRNA candidate matches to SRP CM.

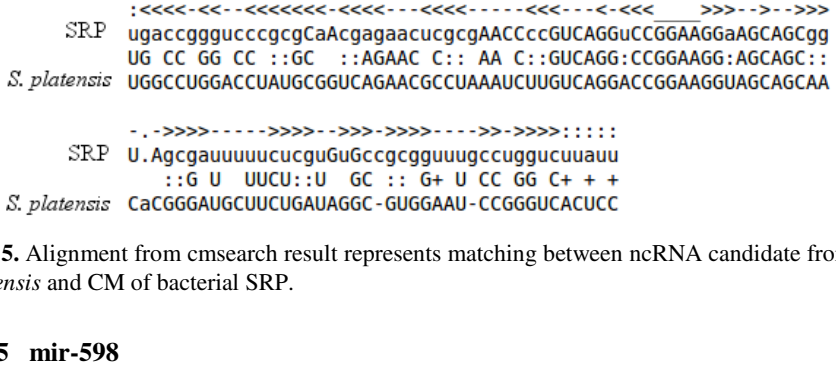


Fig. 5. Alignment from cmsearch result represents matching between ncRNA candidate from *S. platensis* and CM of bacterial SRP.

3.2.5 mir-598

mir-598 is a type of microRNA which is widely observed in animals and plants. Characteristic of RNAs in this family is approximately 20-nt long stretch single stranded RNA which is processed from approximately 100 nt-stem loop RNA by proteins complex. These microRNAs regulate translation by base-pairing mechanism to mRNA after incorporated with an argonaute protein complex [24]. In *S. platensis*, mir-598 candidate locates between predicted hydrogenase accessory protein HypB gene and predicted hydrogenase nickel insertion protein HypA gene. Although microRNAs have been reported only in mammalian and plants but argonaute protein homologues have been reported in several strains of cyanobacteria [25]. This may be considered as an RNA interference system in cyanobacteria. Fig. 6 shows "cmsearch" results which indicate how an ncRNA candidate is matched to mir-598 CM.

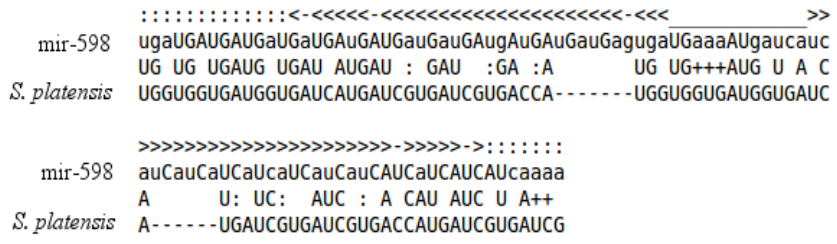


Fig. 6. Alignment from cmsearch result represents matching between ncRNA candidate in *S. platensis* and mir-598 CM.

3.2.6 Bacterial Rnase P Type A

A predicted locus, located on the upstream region of predicted adenylate/guanylate cyclase coding gene, matched with the CM of RNaseP constructed from conserved region of RNase P RNA. Bacterial endoribonuclease RNase P which composed of catalytic RNA domain and protein subunits is essential to generate mature 5' region of tRNA [26]. RNase P cleaves 5' element of pre-tRNA through hydrolysis at specific phosphodiester bond. In addition, other RNase P substrates have been reported also which are polycistronic mRNA [27], tmRNA [28], bacteriophage C4 antisense RNA precursor [29] and SRP RNA [30]. Fig. 7 represents "cmsearch" result which shows how ncRNA candidate is matched to CM of Bacterial RNaseP type A.

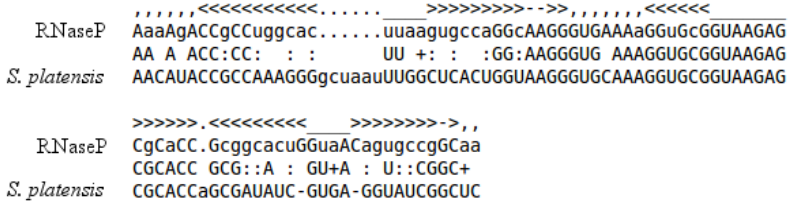


Fig. 7. Alignment from cmsearch result represents matching between ncRNA candidate in *S. platensis* and CM of bacterial RNaseP type A.

3.3 Non-coding RNA Target Prediction

After excluding putative tRNA loci, interaction scores between each ncRNA candidate and all protein-coding sequences in *S. platensis* were calculated by IntaRNA. Candidates for ncRNA targets were extracted from upstream regions of predicted protein-coding genes of *S. platensis* (250 bases upstream of each start codon and 150 bases downstream of each start codon for each protein-coding sequence).

3.3.1 Yfr1 Predicted Targets

Prediction result for Yfr1 indicates that Yfr1 interacts with its target by conserved region (5'-UCACUCCUCACACCACACU-3') located between two stem loops. The conserved region forms base pairing to putative 5' UTR of predicted ORFs listed in Table 1. Many interactions were predicted to occur around -20 to -6 nt (refer to ORF start codon) and some were predicted to occur within ORF region. Yfr1 may regulate

Table 1. Top rank of predicted targets for Yfr1 homolog

gene name	target on mRNA	target on ncRNA	Interaction score
DegT/DnrJ/EryC1/StrS aminotransferase	-20 .. -6	39 -- 52	-16.84
Undecaprenyl-phosphate galactose phosphotransferase	-100 .. -89	41 -- 52	-15.35
phosphoenolpyruvate synthase	32 .. 43	38 -- 50	-15.35
glutamate racemase	42 .. 56	35 -- 50	-15.33
sulfotransferase	-80 .. -65	38 -- 52	-15.18
chromosome partitioning protein, ParB family	-226 .. -216	39 -- 50	-14.45
DNA-cytosine methyltransferase	73 .. 84	36 -- 48	-14.19
Amine oxidase	-117 .. -102	38 -- 52	-14
ATP-dependent metalloprotease FtsH	-178 .. -164	38 -- 52	-13.93
catalytic domain of components of various dehydrogenase complexes	-224 .. -209	38 -- 52	-13.7

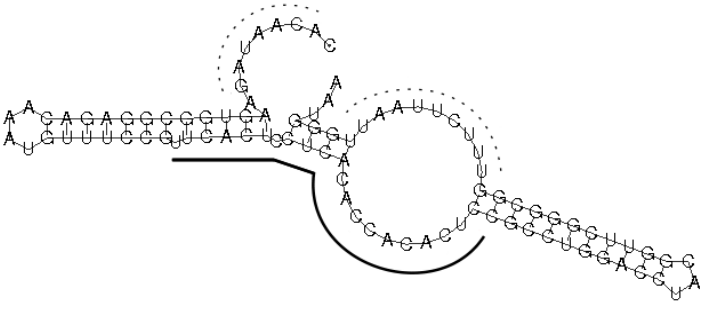


Fig. 8. Predicted structure of putative Yfr1 in *S. platensis* using RNAfold [32]. Thick line represents interaction region and dot line represents A-U rich regions which are Hfq binding motif.

translation of these genes since the predicted interaction regions are at putative ribosome binding sites. Furthermore, predicted Yfr1 structure at interaction site was closed by complementary base pairing from 5' and 3' tails which were adjacent to Hfq binding motif, indicated by A-U rich region and stem loop (Fig. 8) [31]. This suggests that Hfq may involve in Yfr1 target regulation by binding to the motifs to prevent an interaction site from pairing with those 5' and 3' tails.

3.3.2 mir-598 Homolog Predicted Targets

Predicted targets of mir-598 homologue are varied in term of functions and interaction sites while predicted interacting site on mir-598 homolog are likely to be at around the 1st nt to the 50th nt and the 1st nt to the 14th nt as shown in Table 2. Considering to predicting interactions which occur at a half of stem loop on ncRNA (Fig. 9)

which is similar to interaction of micro RNA regulation in higher organisms, this may be RNA interference in *S. platensis*.

Table 2. Top rank of predicted targets for mir-598 homolog

gene name	target on mRNA	target on ncRNA	Interaction score
putative transposase	-244 .. -197	1 -- 50	-24.24
Hemolysin-type calcium-binding region	-30 .. 24	3 -- 50	-22.09
4-hydroxyphenylpyruvate dioxygenase	96 .. 140	28 -- 71	-19.89
short-chain dehydrogenase/reductase SDR	-136 .. -121	1 -- 16	-19.42
SCP-like extracellular	-21 .. 22	4 -- 50	-19.40
glycosyl transferase family 2	-193 .. -132	1 -- 50	-19.37
Polypeptide-transport-associated domain protein ShlB-type	134 .. 147	1 -- 14	-17.95
ribose-phosphate pyrophosphokinase	-193 .. -180	3 -- 14	-17.47
anaerobic ribonucleoside-triphosphate reductase activating	-182 .. -133	3 -- 50	-17.28
restriction endonuclease	134 .. 149	1 -- 14	-17.21
homoserine kinase	-19 .. 34	1 -- 50	-17

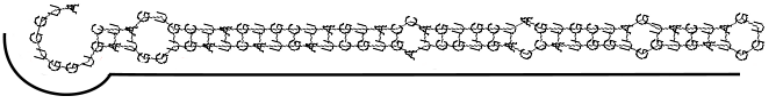


Fig. 9. Predicted structure of mir-598 homolog using RNAfold. Thick line represent interaction region which is predicted to interact with targets.

4 Conclusion and Further Works

This work provided a set of 334 putative ncRNAs including of 129 known ncRNAs and 205 unkown loci for verification and further analysis. In addition, a list of predicted targets for these ncRNAs was also acquired. To investigate reliability and role(s) of predicted ncRNAs in *S. platensis*, an integrating information from "cmsearch" result, ncRNA target prediction, location of transcription regulatory site around ncRNA loci and ncRNA-protein(s) interaction will be performed before analyzing by experimental approach.

Acknowledgments

Srisuk, T. would like to thank National Center for Genetic Engineering and Biotechnology, Thailand (BIOTEC), and King Mongkut's University of Technology Thonburi for the scholarship of Bioinformatics program.

References

1. Washietl, S., Hofacker, I.L., Lukasser, M., Hüttenhofer, A., Stadler, P.F.: Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat. Biotechnol. 23, 1383–1389 (2005)
2. Storz, G.: An expanding universe of noncoding RNAs. Science 296(5571), 1260–1263 (2002)
3. Meyer, I.M.: A practical guide to the art of RNA gene prediction. Brief Bioinform. 8(6), 396–414 (2007)
4. Nawrocki, E.P., Kolbe, D.L., Eddy, S.R.: Infernal 1.0: Inference of RNA alignments. Bioinformatics 25(10), 1335–1337 (2009)
5. http://genome.jgi-psf.org/nospu/nospu.download.html
6. http://www.sb-roscoff.fr/Phyto/Genome_Cyanos/ProSS120/
7. ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/
8. http://www.ncbi.nlm.nih.gov
9. Alteschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic Local alignment search tool. J. Mol. Biol. 215, 403–410 (1990)
10. Edgar, R.C.: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5), 1792–1797 (2004)
11. Washietl, S., Hofacker, I.L., Stadler, P.F.: Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. USA 102(7), 2454–2459 (2005)
12. Griffith-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., Bateman, A.: Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. 33, 121–124 (2005)
13. Busch, A., Richer, A.S., Backofen, R.: IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. Bioinformatics 24(24), 2849–2856 (2008)
14. Voß, B., Gierga, G., Axmann, I.M., Hess, W.R.: A motif-based search in bacterial genomes identifies the ortholog of the small RNA Yfr1 in all lineages of cyanobacteria. BMC Genomics 8, 375 (2007)
15. Nakamura, T., Naito, K., Yokota, N., Sugita, C., Sugita, M.: A Cyanobacterial Non-coding RNA, Yfr1, is required for Growth Under Multiple Stress Conditions. Plant Cell Physiol. 48(9), 1309–1318 (2007)
16. Richter, A.S., Schleberger, C., Backofen, R., Steglich, C.: Seed-based INTARNA prediction combined with GFP-reporter system identifies mRNA targets of the small RNA Yfr1. Bioinformatics 26(1), 1–5 (2010)
17. Eddy, S.R.: The Infernal user's guide, August 15 (2009), http://infernal.janelia.org/
18. Toro, N., Jimenez-Zurdo, J.I., Gracia-Rodriguez, F.M.: Bacterial group II introns: not just splicing. FEMS Microbiol. Rev. 31, 342–358 (2007)
19. Rodionov, D.A., Vitreschak, A.G., Mironov, A.A., Gelfand, M.S.: Comparative genomics of the vitamin B_{12} metabolism and reglation in prokaryotes. J. Biol. Chem. 278(42), 41148–41159 (2003)
20. Warner, D.F., Savvi, S., Mizrahi, V., Dawes, S.S.: A riboswitch regulates expression of the Coenzyme B_{12}-Independent methionine synthase in *Mycobacterium tuberculosis*: Implications for differential methionine synthase function in strain H37Rv and CDC1551. J. Bacteriol. 189(9), 3655–3659 (2007)
21. Kazanov, M.D., Vitreschak, A.G., Gelfand, M.S.: Abundance and functional diversity of riboswitches in microbial communities. BMC Genomics 8, 347 (2007)

22. Rodinov, D.A., Hebbeln, P., Gelfand, M.S., Eitinger, T.: Comparative and functional genomics analysis of prokaryotic Nickel and Cobalt uptake transporters: Evidence for a novel group of ATP-binding cassette transporters. J. Bacteriol. 188(1), S1 (2006)
23. Nagai, K., Oubridge, C., Kuglstatter, A., Menichelli, E., Isel, C., Jovine, L.: Structure, function and evolution of the signal recognition particle. EMBO J. 22(14), 3479–3485 (2003)
24. Bartel, D.P.: MicroRNAs: target Recognition and Regulatory Functions. Cell 136(2), 215–233 (2009)
25. Makarova, K.S., Wolf, Y.I., van der Oost, J., Koonin, E.V.: Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. Biol. Direct. 4, 29 (2009)
26. Frank, D., Pace, N.: Ribonuclease P: Unity and diversity in a tRNA processing ribozyme. Annu. Rev. Biochem. 67, 153–180 (1998)
27. Alifano, P., Rivellini, F., Piscitelli, C., Arraiano, C.M., Bruni, C.B., Carlomagno, M.S.: Ribonuclease E provides substrates for ribonuclease P-dependeent processing of a ploycistronic mRNA. Genes Dev. 8(24), 3021–3031 (1994)
28. Komine, Y., Kitabatake, M., Yokogawa, T., Nishikawa, K., Inokuchi, H.: A tRNA-like structure is present in 10Sa RNA, a small stable RNA from *Escherichia coli*. Proc. Natl. Acad. Sci. USA 91(20), 9223–9227 (1994)
29. Hartmann, R.K., Heinrich, J., Schlegl, J., Schuster, H.: Precursor of C4 antisense RNA of bacteriophages P1 and P7 is a substrate for RNase P of *Escherichia coli*. Proc. Natl. Acad. Sci. USA 92(13), 5822–5826 (1995)
30. Peck-Miller, K.A., Altman, S.: Kinetics of the processing of the precursor to 4·5 S RNA, a naturally occurring substrate for RNase P from *Escherichia coli*. J. Mol. Biol. 221(1), 1–5 (1990)
31. Moll, I., Afonyushkin, T., Vytvytska, O., Kaberdin, V., And Blasi, U.: Coincident Hfq binding and RNase E cleavage sites on mRNA and small regulatory RNAs. RNA 9(11), 1308–1314 (2003)
32. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., Schuster, P.: Fast Folding and Comparison of RNA Secondary Structures. Monatshefte für Chemie. 125(2), 167–188 (1994)

Reconstruction of Starch Biosynthesis Pathway in Cassava Using Comparative Genomic Approach

Oratai Rongsirikul[1], Treenut Saithong[1,2,4], Saowalak Kalapanulak[1,2,4], Asawin Meechai[1,3,4], Supapon Cheevadhanarak[1,2,4], Supatcharee Netrphan[5], and Malinee Suksangpanomrung[5]

[1] Bioinformatics and Systems Biology Program
[2] School of Bioresources and Technology
[3] Department of Chemical Engineering
[4] Systems Biology and Bioinformatics Lab (SBI),
Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
[5] National Center for Genetic Engineering and Biotechnology, Pathumthani, Thailand
o.r.n@hotmail.com, treenut.sai@kmutt.ac.th,
saowalak.kal@kmutt.ac.th, asawin.mee@kmutt.ac.th,
supaponche@yahoo.com, supatchareen@biotec.or.th,
malineec@biotec.or.th

Abstract. Cassava is one of the most attractive crops nowadays because it can produce and accumulate large amount of starch in its roots. Cassava starch is widely used as food, feed and raw materials for biochemical industries. Due to the increasing demand of cassava starch, the starch biosynthesis pathway is thus of interest for metabolic engineering, aiming at strain improvement. However, the uncertainties in the metabolic pathway of starch biosynthesis in cassava retard the rate of achievement. Availability of recently released cassava genome motivates us to reconstruct the starch biosynthesis pathway in cassava using comparative genomic approach. Here, nucleotide sequences of the template plants (*i.e.* Arabidopsis and potato) were compared with the sequence of cassava collected from three sources: Phytozome (genomic sequence), Cassava full-length cDNA and Cassava genome (ESTs) databases. The metabolic pathway of approximately 34 enzymes was constructed, including pathway from sucrose metabolism to amylose and amylopectin synthesis. The resulting pathway is a good initial point toward the complete pathway reconstruction.

Keywords: Metabolic pathway reconstruction, Starch biosynthesis, Cassava, Comparative genomic approach.

1 Introduction

Cassava (*Manihot esculenta*) is one of the leading plants of the world for serving as human food and animal feed [1], and it is considered as one of the most important economic crops in Thailand. Cassava starch is also used in various nonfood applications such as paper, textile, plywood, glue, alcohol, cosmetic, and pharmaceutical

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 118–129, 2010.

industries. The applications of cassava starch are usually dictated by the physico-chemical characteristics of the produced starch, which are related to the proportion of amylopectin and amylose [2]. Therefore, the value of cassava is not only determined by the yield, but also by the properties of the produced starch.

Starch is a heterogeneous mixture of highly branched amylopectin and less branched amylose, which are complex structure polymers comprised of α-glucan monomer units. Starch is synthesized and stored in the chloroplast during the day as a transient product in leaf, called *transitory starch*, before being degraded and re-synthesized during the night to restore in the amyloplast storage organ as *permanent starch* [2]. Starch is accumulated in plants in both short-term and long-term time frames depending on the purpose. Short-term storage is required in leaves to provide a source of carbohydrate that can be used to maintain metabolic functions when photosynthesis is inactive. Long-term storage of carbohydrate in tubers or seeds is necessary to support reproductive tissue development. In vascular plants, starch is synthesized within the plastid by a complex metabolic pathway containing four basic steps: substrate activation (catalyzed by ADP-glucose pyrophosphorylase), polymer elongation (starch synthase), polymer branching (branching enzymes) and debranching (debranching enzymes) [3].

Various breeding programs have been developed to improve plant starch production both in terms of yield and properties (*e.g.* solubility and swelling power) to serve industrial requirements [4]. To attain this goal, genes in starch biosynthesis pathway have been characterized in various model plants. As with cassava, a number of starch-biosynthesis related genes was studied, including α-amylase (MEamy2) [5], Sucrose transporter1 (MeSUT1) [6], ADP-glucose pyrophosphorylase (AGPase) [7], Granule bound starch synthase II (GBSSII) [8], and Branching enzyme [7]. This knowledge provides more insight into the starch biosynthesis process in cassava, at least in terms of pathway components. Though to date the whole pathway of the starch biosynthesis in cassava is not yet fully understood.

Employing comparative genomic approach, the pathway of starch and sucrose metabolism in cassava was first constructed by Sakurai *et al* (2007) [9]. Under the limitation of cassava genome data, the first draft pathway was reconstructed using full-length cDNA EST of cassava under various conditions. Also, the authors employed Arabidopsis, which is evolutionarily far from cassava, as a single template. The published pathway may thus not fully describe starch and sucrose metabolism in cassava. Recently, genome sequence of cassava was released to the public, increasing an opportunity to acquire more complete metabolic map of the starch and sucrose metabolism in cassava.

The availability of the novel data motivates us to start revising the previous reconstructed pathway of starch biosynthesis in cassava. Therefore, the objective of this study is to reconstruct the metabolic pathway of starch biosynthesis in cassava by using a comparative genomic approach. Here, multiple template plants were employed to increase the span of template; yet we presented the result of using only two template plants. We focused on the starch biosynthesis pathway starting from sucrose metabolism and ending at amylose and amylopectin synthesis. Only nucleotide

sequence alignments (BLASTn) between the template plants and cassava were performed due to the restriction of data availability of cassava. The nucleotide sequences of Arabidopsis and potato were collected from the KEGG database. For cassava, genome data were gathered from three main sources: (i) genomic sequence (http://www.phytozome.net), (ii) full-length cDNA (http://amber.gsc.riken.jp/cassava) [9], and (iii) ESTs (http://cassava.igs.umaryland.edu).

At the present, the first assembly of cassava genome is not yet completed and it is not fully annotated. We thus addressed an inverse similarity search approach, which is simply using cassava genome as a database for the search instead of being a query as in a normal similarity search process. Through this simple method, we can reconstruct more detailed pathway of the starch biosynthesis in cassava.

2 Method

2.1 Data Resources

Template plants. Starch biosynthesis pathways in template plants were used as an outline for building such a pathway in cassava. The multiple templates were employed in this study to capture cassava genes that are absent in a single template plant. Template plants were selected according to the data availability and the evolutionary distance (between template and cassava). Here, Arabidopsis and potato were selected for the following reasons. Arabidopsis is a well-studied plant that most of all genes in this pathway were annotated and characterized, while potato is a starch storage plant evolutionarily closely related to cassava. Due to data restriction, nucleotide sequences of potato used in this study were obtained from ESTs experiment available in the KEGG database, instead of gene sequences as in case of Arabidopsis.

Cassava. Nucleotide sequences of cassava were collected from three sources of data: genomic sequence, full-length cDNA ESTs, and partial-sequenced ESTs. First, genomic sequence is obtained from genome sequencing of Cassava Genome Project 2009. Nucleotide sequence of the first cassava genome assembly, containing 416 Mb, is available before scientific publication in Phytozome database (http://www.phytozome.net). The data is provided in a form of scaffolds (~11,243) with no annotation, and can be downloaded in a multi-fasta format. Although cassava genome size is 760 Mb, it is believed that the 416 Mb of the first genome assembly nearly covers all of the genic regions in the cassava genome. The current genome sequence is estimated to cover 95 percent of known cassava genes. Second, full-length cDNA sequence is obtained from full-length cDNA ESTs from leaves and roots (http://amber.gsc.riken.jp/cassava) that were constructed under normal, heat, drought, aluminum, and post harvest physiological deterioration conditions by Sakurai *et al* (2007) [9]. Third, partial sequence is obtained from partial ESTs that were performed in cassava. This data is available in GenBank.

The overall processes used in this study are shown in Fig. 1.

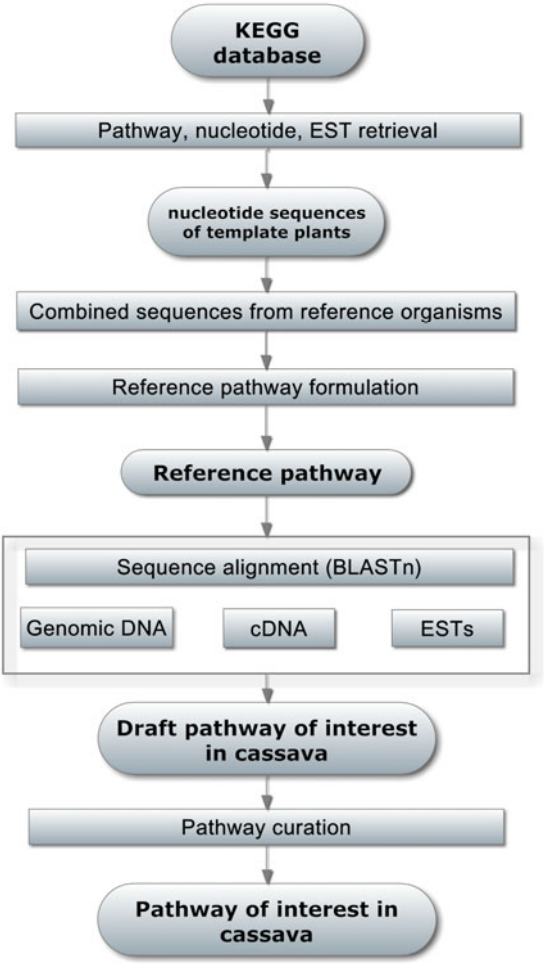


Fig. 1. The overview of methodology. Pathways as well as nucleotide sequences of template plants were retrieved from the KEGG database to formulate the reference pathway. Such nucleotide sequences (i.e. genes and ESTs) were subsequently used as a query to search for their homologue in cassava. Overlaying the search results of the homologous genes in cassava on the reference pathway can outline the pathway of interest in cassava. The resulting pathway was then curated with experimental evidence in literature once again at the end of the reconstruction process.

2.2 Reference Pathway Formulation

The reference pathway was formulated under the consideration of at least three features: boundary of the pathway, constituent components in the pathway, and the reactions between components in the pathway. The boundary of the reference pathway of interest (*i.e.* related to starch biosynthesis process) was defined according to the

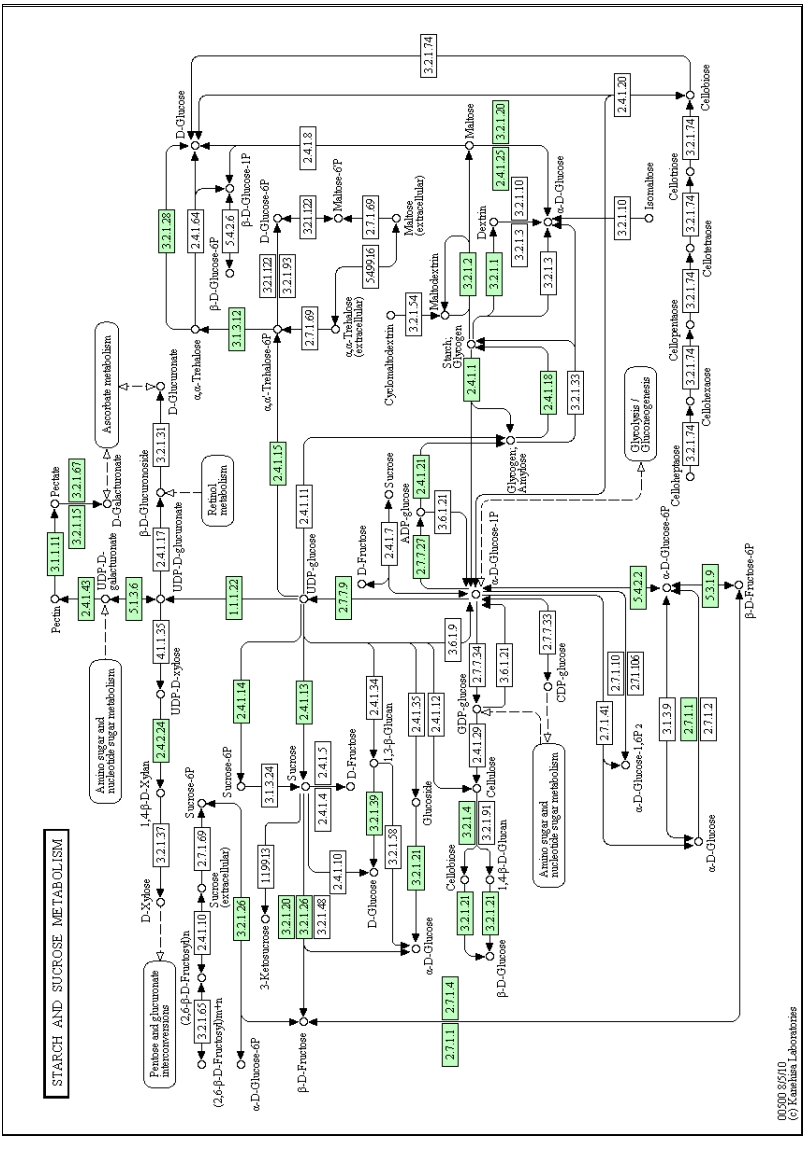


Fig. 2. Starch and sucrose metabolism in KEGG mapped with Arabidopsis genes (green boxes). The boxes represent enzymes with EC number. The circles represent metabolites.

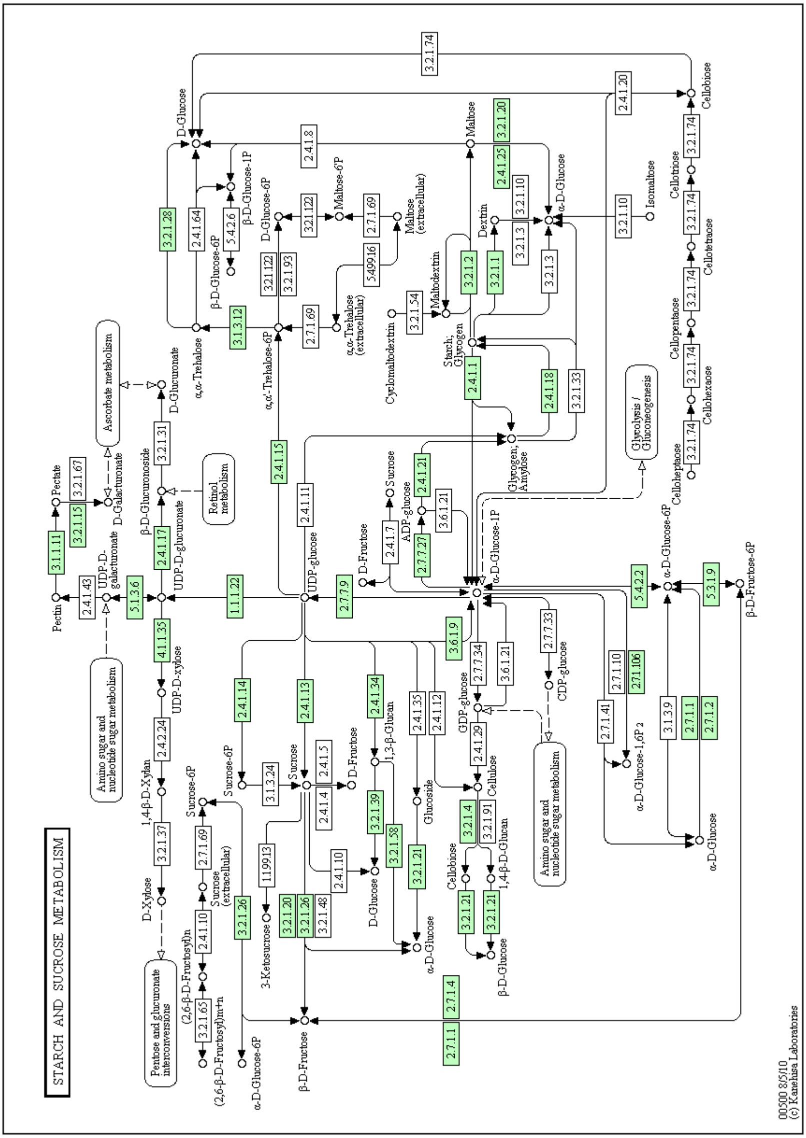


Fig. 3. Starch and sucrose metabolism in KEGG mapped with potato ESTs. (green boxes). The boxes represent enzymes with EC number. The circles represent metabolites.

KEGG metabolic network in the KEGG database, from which much of the data used in this study was obtained. Starch and sucrose metabolism pathways in the KEGG metabolic network (Fig. 2 and Fig. 3) demonstrate the scope of metabolic pathway under study, covering 78 biochemical reactions, 71 enzymes (EC numbers) and 50 chemical reactions. However, the KEGG pathway presents a non-species-specific network that includes all biochemical reactions of metabolite derivation occurring in living organisms (*i.e.* animals, plants and microorganisms). To formulate the plant-specific reference pathway, Arabidopsis (genes) and potato (ESTs) data were mapped onto the KEGG pathway, and only metabolites, enzymes and reactions existing in Arabidopsis and potato were kept in the formulated reference pathway. The formulated reference pathway was redrawn as shown in Fig. 4.

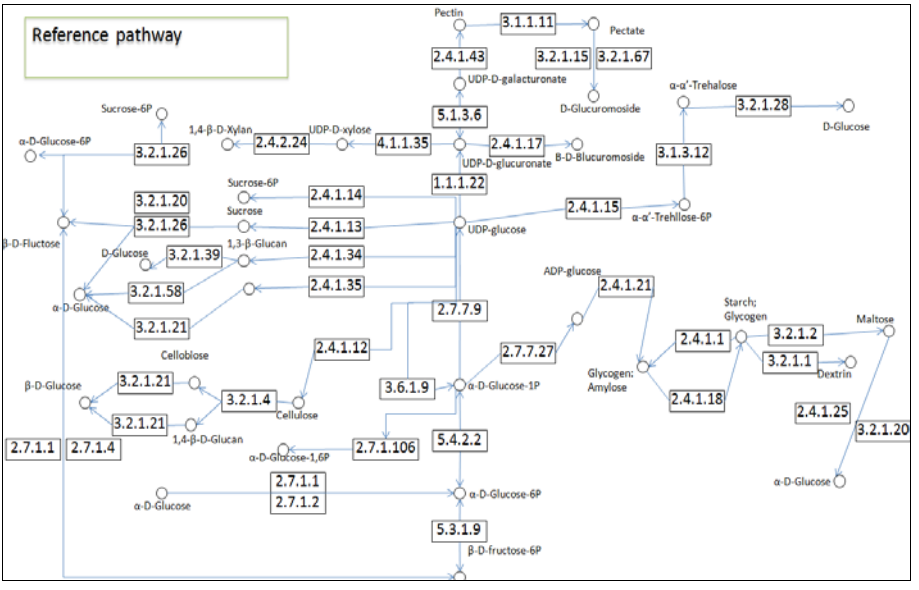


Fig. 4. Reference pathway of sucrose metabolism and starch biosynthesis developed by using Arabidopsis and potato as templates. Boxes represent enzymes with specific EC numbers, and circles represent metabolites.

2.3 Sequence Similarity Search (BLASTn)

Sequence similarity search was performed in all sources of cassava genome data, in order to computationally identify cassava genes involved in starch biosynthesis process. Stand-alone BLAST 2.2.23 was employed for searching into the downloaded genomic sequence of cassava (from Phytozome database), while web-based BLAST provided in the websites was employed for searching into cDNA and EST sequence libraries: cassava full-length cDNA database (http://amber.gsc.riken.jp/cassava) and cassava genome database (http://cassava.igs.umaryland.edu). Sequence alignment through both stand-alone BLAST 2.2.23 and web-based BLAST was performed under default setting. Arabidopsis genes and potato ESTs were used as a query for searching their homologue in cassava genome based on the data available in the three databases. Sequences with

high similarity (percent identity ≥75% and e-value ≤ 1e-10) were defined as a homologue of the query sequence in cassava, sharing a similar enzymatic function.

2.4 Pathway Reconstruction

The availability of starch biosynthesis genes in cassava were predicted by sequence similarity search (BLASTn), whereby the function (including EC number) of the homologue sequence was assumed to be the same as that of the query. Pathway of starch biosynthesis in cassava was, thus, reconstructed by mapping these predicted components onto the reference pathway (Fig. 4).

3 Results and Discussion

3.1 Reference Pathway of Sucrose Metabolism and Starch Biosynthesis

The reference pathway of sucrose metabolism and starch biosynthesis was redrawn in Fig. 4. Basically, it is a combined network of such pathway in Arabidopsis (31 enzymes) and potato (34 enzymes) template plants. The resulting reference pathway composes of 35 biochemical reactions, 37 enzymes (EC numbers) and 33 metabolites. The larger reference pathway developed from multiple template plants may increase the possibility to identify starch-biosynthesis related genes in cassava.

3.2 Identification of the Enzymes in the Sucrose Metabolism and Starch Biosynthesis Pathway in Cassava via Similarity Search (BLASTn)

Based on starch and sucrose metabolism pathway in the KEGG database, nucleotide sequences of 116 genes (31 EC numbers) of Arabidopsis and 231 ESTs (34 EC numbers) of potato were retrieved and used as a BLAST query. The results of similarity search allow us to identify at least 34 enzymes (EC numbers) in cassava metabolic pathway. The number of predicted enzymes in cassava in this study is significantly increased from the previous study that relied only on a single template plant [9]. BLAST results are summarized in Table 1.

Table 1. Summary of the similarity search (BLASTn)

Templates	Arabidopsis	Potato	Arabidopsis+potato
Number of nucleotide sequences retrieved from KEGG database	116 (genes)	231 (ESTs)	116 genes 321 ESTs
Number of the enzymes (EC numbers) corresponding to the retrieved nucleotide sequences	31	34	37
Number of nucleotide sequences in cassava matched with the template sequences	112 (genes)	187 (ESTs)	112 genes 187 ESTs
Number of the enzymes (EC numbers) in cassava predicted from similarity search (corresponding to the matched sequences)	25 (see Fig. 5)	30 (see Fig. 6)	34 (see Fig. 7)

3.3 Sucrose Metabolism and Starch Biosynthesis Pathway in Cassava

Cassava sucrose metabolism and starch biosynthesis pathway reconstructed based on similarity search against Arabidopsis genes (Fig. 5) comprises of 25 EC numbers (marked as blue boxes). The pathway contains 25 reactions and 3 gap reactions (marked as white boxes). In the same manner, the reconstructed pathway based on potato ESTs composes of 30 EC numbers which lie in 29 reactions. Though more EC numbers and reactions can be identified in the latter case, the pathway still contains 2 gaps (Fig. 6).

When combining all EC numbers that were identified from both template plants, the putative metabolic network of sucrose metabolism and starch biosynthesis in cassava covers 34 EC numbers in 33 reactions *with no* gap. (Fig. 7). Interestingly, the use of multiple template plants not only allows more cassava enzymatic genes to be identified, but it also helps in network gap closure. Our finding contributes to increase information regarding the process of sucrose metabolism and starch biosynthesis in cassava as follows.

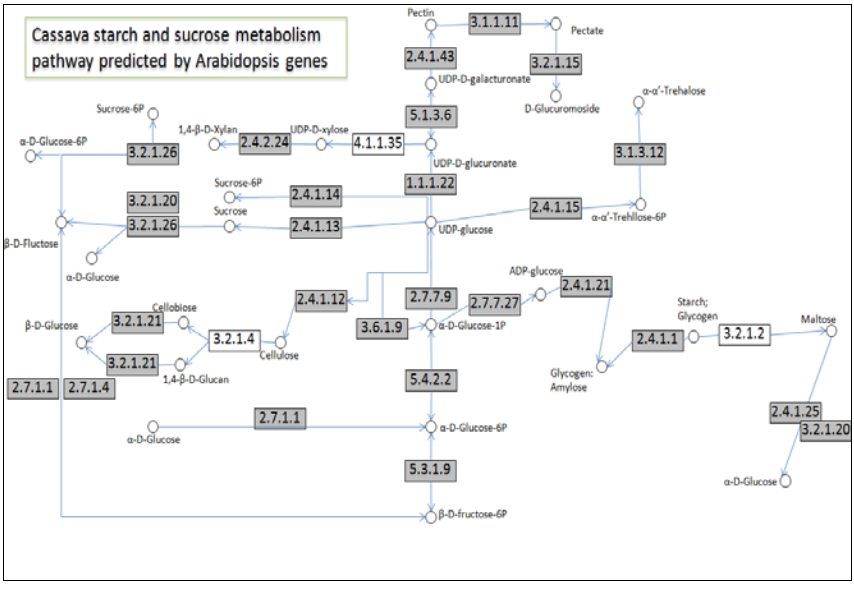


Fig. 5. Sucrose metabolism and starch biosynthesis pathway in cassava using Arabidopsis as a template. The gray boxes represent enzymes existing in cassava, predicted through similarity search, and the white boxes represent the gap between the existing enzymes.

Starch biosynthesis pathway in cassava is likely to be closer to that in potato than in Arabidopsis. This is suggested by the similarity search results using a single template plant. In total of 34 enzymes (EC numbers) identified from the multiple template plants, nine enzymes (EC numbers) in cassava were identified only when using potato as a template, while only four enzymes (EC numbers) were found as a result of using Arabidopsis as a template. The nine enzymes identified by potato are (1) cellulose

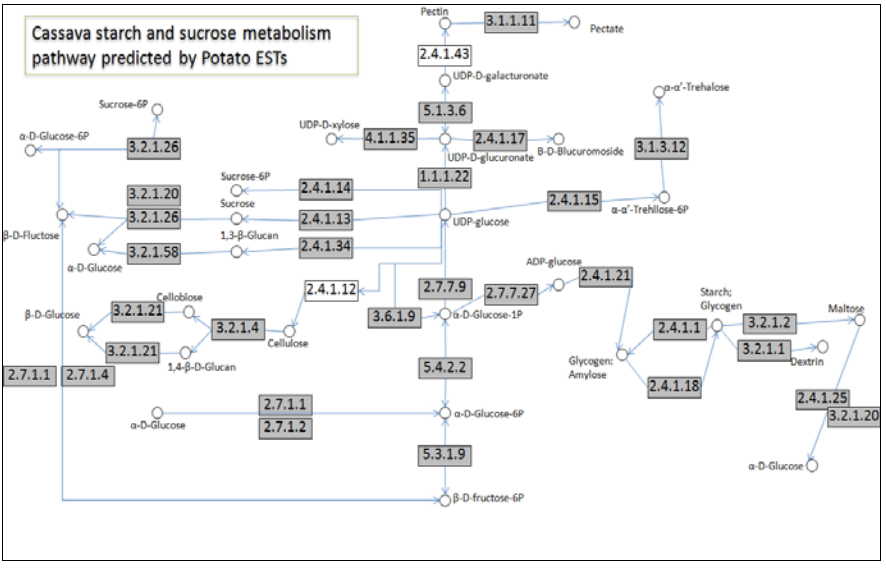


Fig. 6. Sucrose metabolism and starch biosynthesis pathway in cassava using potato as a template. The gray boxes represent enzymes existing in cassava, predicted through similarity search, and the white boxes represent the gap between the existing enzymes.

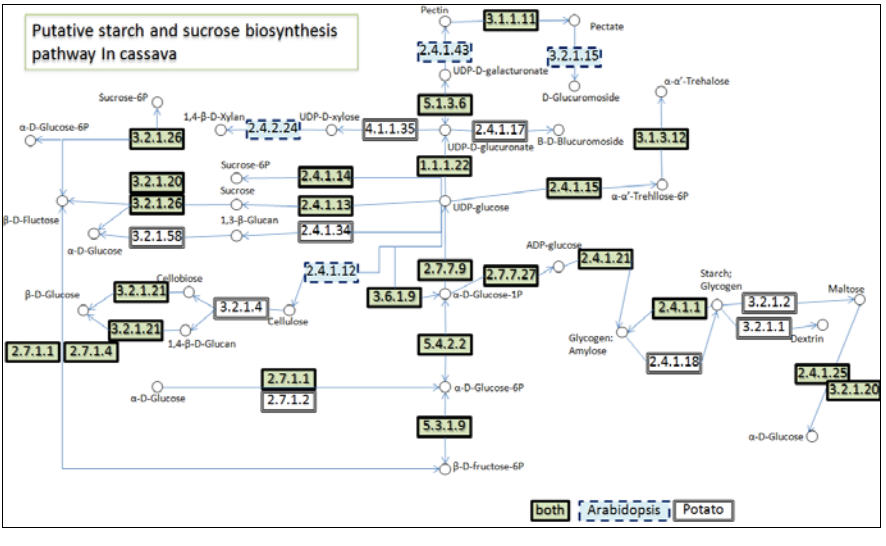


Fig. 7. Sucrose metabolism and starch biosynthesis pathway in cassava. The green boxes with thick border line represent existing enzymes in cassava that were identified from both templates. The blue boxes with dash border line and white boxes represent existing enzymes in cassava predicted from using Arabidopsis or potato as template respectively, and the white boxes represent the gap between existing enzymes.

(EC 3.2.1.4), (2) 1,3-β-glucan synthase (EC 2.4.1.34), (3) glucokinase (EC 2.7.1.2), (4) UDP-glucuronate decarboxylase (EC 4.1.1.35), (5) glucuronosyltransferase (EC 2.4.1.17), (6) α-amylase (EC 3.2.1.1), (7) β-amylase (EC 3.2.1.2), (8) starch branching enzyme (SBE; EC 2.4.1.18), and (9) glucan 1,3-beta glucosidase (EC 3.2.1.58); and the four enzymes identified by Arabidopsis are (1) cellulose synthase (EC 2.4.1.12), (2) 1,4-β-D-xylan synthase (EC 2.4.2.2), (3) polygalacturonate 4-α-galacturonosyltransferase (EC 2.4.1.43), and (4) polygalacturonase (EC 3.2.1.15).

Besides the numbers of identified genes, at least three key enzymes in starch biosynthesis pathway were identified by using potato as a template: starch synthase (SS; EC 2.4.1.21), ADP-Glucose pyrophosphorylase (AGPase; EC 2.7.7.27) and starch branching enzyme (SBE; EC 2.4.1.18). For Arabidopsis single template, only SS and AGPase, but not SBE, were predicted to exist in cassava. This result corresponds well to the work published by Sakurai *et al.* (2007) [9].

Moreover, the result of more genes identified from the multiple template plants allows us to learn the advantage of such a template to capture more components of the network, and also to pinpoint the weakness of using a single template plant. Compared to the previously published pathway [9], the novel reconstruction can cover more EC numbers and reactions, which only 26 reactions were identified from matching the fulllength cDNA with the Arabidopsis genes.

Though we succeeded to reconstruct more complete pathway of sucrose metabolism and starch biosynthesis in cassava, this putative pathway might be only an initial step toward the reconstruction of the whole network. Thus, in this stage and the early of next stage, we will give the priority to the improvement in the power of computational prediction rather than the sophisticated experiment for validation. For example tBLASTn will be employed for similarity search of the evolutionarily closely related organisms. Furthermore, for more complete pathway, more template plants may be required to cover all possible starch biosynthesis genes in cassava. In addition, this pathway may require further curation and analysis, such as tracing each enzyme back to the experimental evidence or curated databases. These suggested points will be revisited in our next reconstruction.

4 Conclusion

Nucleotide sequence alignment allows us to find genes encoding enzymes to reconstruct cassava starch biosynthesis pathway, involving 33 metabolic reactions. The results showed the high conservation of starch biosynthesis pathway among the studied plants; however, more template plants might increase the coverage of starch biosynthesis genes in cassava. Although, this putative pathway is not yet complete, it brings better understanding into the metabolic process of cassava. Our reconstructed pathway might be a seed for further development of the complete pathway that will have a big impact to metabolic engineering both in terms of design and analysis. A good design and accurate modification of metabolic pathway would accelerate the achievement in strain improvement.

References

1. Nassar, N.M.: Cassava, Manihot esculenta Crantz, genetic resources: origin of the crop, its evolution and relationships with wild relatives 1, 298–305 (2002)
2. Beyene, D., Baguma, Y., Mukasa, S.B., Sun, C., Jansson, C.: Characterisation and role of isomaylase1 (MEISA1) gene in cassava. Af. Crop Sci. J. 18, 1–8 (2009)
3. William, C.P., Michael, T.M.: Control of primary metabolism in plants, p. 285. Blackwell Publishing, Malden (2006)
4. Nuwamanya, E., Baguma, Y.: Quantification of starch physicochemical characteristics in a cassa segregating population. Af. Crop Sci. J. 16, 191–202 (2009)
5. Tangphatsornruang, S., Naconsie, M., Thammarongtham, C., Narangajavana, J.: Isolation and characterization of an [alpha]-amylase gene in cassava (Manihot esculenta). Plant Physiology and Biochemistry 43, 821–827 (2005)
6. Worawut, Y., Suksangpanomrung, M., Limapaseni, T.: Expression of cassava Manihot exculenta Crantz. In: Sucrose transporter1 (MeSUT1) gene in yeast Saccharomyces cerevisiae and characterization of the protein. 33rd Congress on Science and Technology of Thailand (2002)
7. Munyikwa, T.R.I.: Isolation and characterisation of starch biosynthesis genes from cassava (Manihot esculenta Crantz). Munyikwa, [S.l.] (1997)
8. Munyikwa, T.R.I., Langeveld, S., Salehuzzaman, S.N.I.M., Jacobsen, E., Visser, R.G.F.: Cassava starch biosynthesis: new avenues for modifying starch quantity and quality. Euphytica 96, 65–75 (1997)
9. Sakurai, T., Plata, G., Rodriguez-Zapata, F., Seki, M., Salcedo, A., Toyoda, A., Ishiwata, A., Tohme, J., Sakaki, Y., Shinozaki, K., Ishitani, M.: Sequencing analysis of 20,000 full-length cDNA clones from cassava reveals lineage specific expansions in gene families related to stress response 7, 66 (2007)

Catalog of Genetic Variations (SNPs and CNVs) and Analysis Tools for Thai Genetic Studies

Sattara Hattirat[1], Chumpol Ngamphiw[2], Anunchai Assawamakin[2], Jonathan Chan[3], and Sissades Tongsima[2,*]

[1] Bioinformatics and Systems Biology Program, King Mongkut's University of Technology Thonburi, Bangkok, Thailand 10140
[2] Biostatistics and Informatics Laboratory, Genome Institute, National Center for Genetic Engineering and Biotechnology, Thailand Science Park, Klong Luang, Patumthani 12120
[3] School of Information and Technology, King Mongkut University of Technology Thonburi, Bangkok, Thailand 10140
sissades@biotec.or.th

Abstract. The Thailand SNP database (ThaiSNPdb) initiative is the first attempt to catalog both Single Nucleotide Polymorphisms (SNPs) and Copy Number Variations (CNVs) from 32 healthy individuals in the central region of Thailand using the 5th generation Affymetrix SNP genotyping arrays. The aim of this initiative is to facilitate genetic studies of Thais by systematically cataloging large-scale population genetic polymorphism data from Thais combining with data from other different populations. Comparative views of both SNPs and CNVs were made possible with standard comprehensive and interactive graphic technology, called GBrowse. The database allows easy browsing and comparisons of genetic polymorphism data from Thai populations as well as others, which were retrieved from several public variation databases including NCBI dbSNP, HapMap3, JSNP and Database of Genomic Variant (DGV). As a result, this database can be considered as one of the largest collections of SNPs and CNVs. Furthermore, to enable genetic analysis, ThaiSNPdb offers three common genetic tools including linkage disequilibrium (LD), haplotype blocks and tagging SNPs. In conclusion, ThaiSNPdb is an invaluable platform to support the studies in personalized medicine, forensic sciences and even cytogenetic studies in the case of CNV analyses. ThaiSNPdb is available on the Internet and can be publicly accessed at http://www.biotec.or.th/thaisnp.

Keywords: Thai genetic variation, SNP, CNV, database, genetic analysis tools.

1 Introduction

The completions of the human genome sequence [1] and the advances in sequencing and genotyping technologies over the last decade [2, 3] have accelerated the studies of human genetic polymorphisms. Many polymorphisms at the DNA level have biological implications and thus are studied extensively in the field of molecular medicine [4, 5],

* Corresponding author.

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 130–140, 2010.

contributing to the understanding in molecular pathogenesis, gene-disease association and predisposing genes.

The types of genetic polymorphisms that are widely studied include microsatellites (short tandem repeats or STRs), Single Nucleotide Polymorphisms (SNPs) and Copy Number Variations (CNVs). Currently, due to the robustness in SNP genotyping [6, 7], the use of SNPs as molecular markers is becoming very popular, replacing the more polymorphic form, yet less robust, microsatellites. In other words, SNPs, by their nature of variation among the four possible genetic bases, do not contain as much polymorphic information as STRs, which could exist in many forms ranging up to many hundreds possible copies in one genome. Furthermore, SNPs could be studied in very high density and are suitable for studies of multifactorial diseases [4, 5]. Another form of genetic variation is CNV, which alter the copy number of genetic material (encompassing one or more genes) [8]. CNV is gaining more popularity owing to its involvement in quantitative gene expressions [9, 10]. Studies on population genetics and molecular pathogenesis during the past five years have been shifted toward the use of SNPs as molecular markers while the use of CNVs is steadily increasing.

Proliferation of the number of genetic variations makes it very difficult to manage without well-designed database infrastructures. Nonetheless, only few genetic variation databases provide large-scale variation data from multiple populations including HapMap3 [11] and NCBI dbSNP [12]. Other public variation databases are dedicated to specific applications or focus specifically on particular ethnics, such as JSNP [13] and the YH Database [14]. The proposed ThaiSNP database was first designed to be an ethnic specific genetic variation database. However, to facilitate genetic studies of other populations when comparing with Thais, we extended the database features by including a graphical interface, which can display genetic information from multiple populations. This was accomplished by incorporating SNP data from HapMap3 and dbSNP. This population-wide comparative feature among SNP data from different sources could greatly assist many technological-driven studies, e.g., pharmacogenomics and other genome-wide association studies. In this paper, we present the construction of ThaiSNP database and its multi-ethnic comparative visualization of Thai people and other genetic variation of other populations.

2 Methodologies

2.1 Data Sources

The SNP and CNV information were obtained from DNA samples of 32 healthy Thai people who were recruited according to the criteria shown in Table 1. Genotyping was performed using the 5th generation Affymetrix SNP genotyping 500K array, which comprises common SNPs reported by Affymetrix. The SNPs from this platform are compatible with other SNP data reported in both dbSNP and HapMap3 (i.e., having large number of common SNPs). Copy number interpretation was analyzed using SNP genotyping from 32 Thais as input. We utilized the R package module [15] in this analysis.

Table 1. Criteria for selecting 32 healthy Thai DNA samples used in the inference of SNPs and CNVs discovery and genotyping

	Criteria	Remarks
1	Being 50-60 years of age.	*Genetic disease should be detectable at this range*
2	Being in the unbroken lineage of at least 3 Thai generations.	
3	Being in good health	- *No serious chronic illnesses* - *No hospitalization due to chronic illness* - *No history of monogenic diseases* - *Laboratory tests with no serious or chronic diseases*

Apart from the new SNP and CNV data gathered in this study, ThaiSNPdb also hosts genetic polymorphism information from other studies on Thai populations, including 3864 SNPs distributing on 368 candidate genes (unpublished data, only on the gene body) and 1536 SNPs from 228 genes (drug metabolizing enzyme), which were reported earlier [16]. Furthermore, ThaiSNPdb stores SNP and CNV information from other public databases, including dbSNP [12], HapMap3 project [11], JSNP [13] and Database of Genome Variance (DGV) [18]. The database also provides gene and disease-gene information retrieved from RefSeq genome build 36.3. Detailed summary of the genetic polymorphism data stored in ThaiSNPdb is shown in Table 2.

Table 2. Third-party genetic polymorphism data stored in ThaiSNP database

	Sources	Types	Numbers	Remarks
1	Thailand SNP Discovery project	SNP	3,864	*On 368 genes with focus on gene body only*
2	Drug metabolizing enzyme SNPs	SNP	1,536	*On 228 drug metabolizing enzyme genes*
3	dbSNP build 129	SNP	14,735,067	
4	HapMap public release 27	SNP	4,165,577	*1,207 individuals from 11 populations*
5	JSNP release 35	SNP	184,081	
6	Database of Genomic variants	CNV	21,107	

2.2 System Design

ThaiSNP database employs MySQL as the main database engine. Fig. 1 illustrates the overall integration of data into ThaiSNP database. The data representation in ThaiSNPdb can be grouped into 2 parts: (1) the detailed information of chromosomes, genes, SNPs and CNVs in the HTML format; (2) the graphical interface in the form of Generic Feature Format (GFF), rendered by GBrowse [19]. The SNP and CNV information can be queried through web interface rendered by Python Webware 1.0.2 [20].

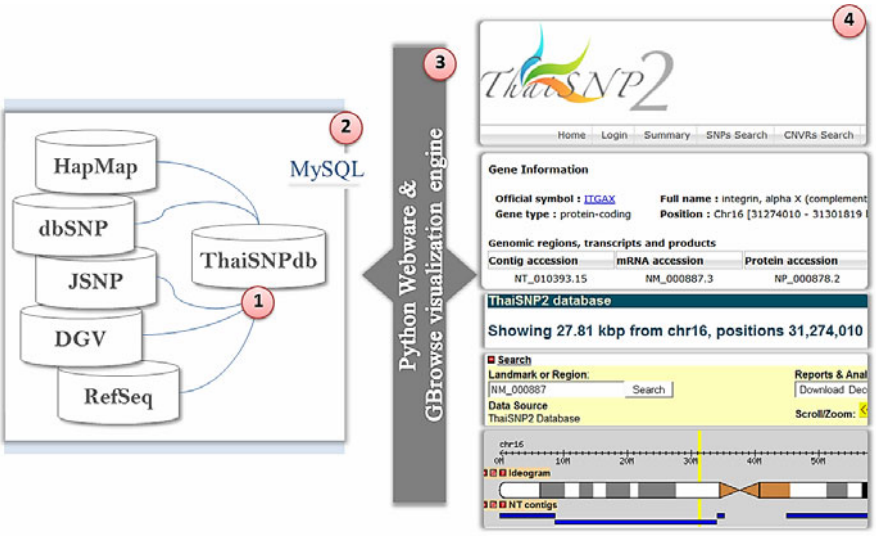


Fig. 1. Implementation of ThaiSNP database system: 1) Storage of third-party genetic polymorphism data (see Table 2 for details); 2) MySQL serving as database management; 3) Python Webware providing web interface and GBrowse providing graphical representation of cytogenetic information; 4) snapshots of homepage, HTML query results and graphical cytogenetic results.

3 Results and Discussions

3.1 Data Analysis

A total of 440,333 SNPs were identified in the SNP genotyping process. The 3,230 CNVs identified were distributed throughout all chromosomes. All CNVs span more than 5,000 bases long, accounting for 1,889 distinct CNV regions. Several analyses including 1) expected SNP heterozygosity, 2) individual neighbor joining tree from all populations and 3) histogram of CNV distribution over the entire genome.

Heterozygosity. Since this is the first genome-wide collection of SNPs from central Thai populations, we want to see how different the Thais are comparing to other populations in terms of genetic diversity. We compared the expected heterozygosity from 299,837 overlapping SNPs with genotyping data from HapMap3 samples. Fig. 2 shows the bar chart of expected heterozygosity of each population reported in our database. The Caucasian descendants appeared to have highest values of expected heterozygosity while African and Asian descendants revealed lower values. The heterozygosity values do not tell us much about genetic relatedness among the populations derived from similar origins; for example, descendants from Africa do not share the similar expected heterozygosity values. A better view on population genetic relatedness is from the following neighbor joining analysis.

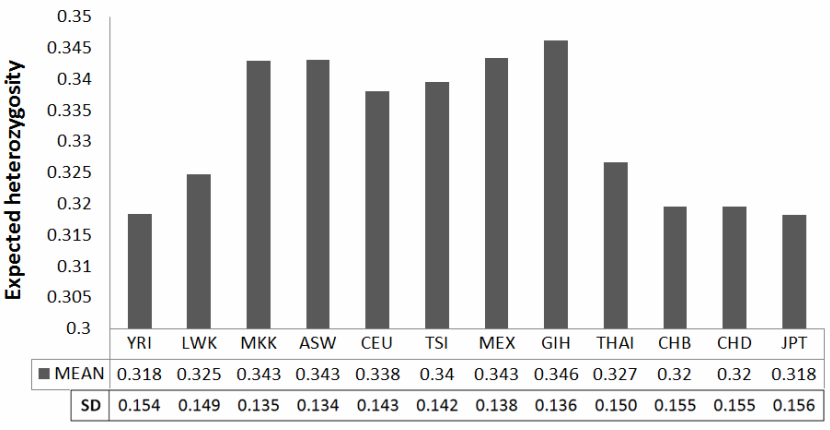


Fig. 2. Expected heterozygosity calculated from genotype data from each population reported by ThaiSNPdb including 32 Thai data (THAI) and the data from 11 HapMap populations. (Population descriptors: ASW: African ancestry in Southwest USA, CEU: Utah residents with Northern and Western European ancestry from the CEPH collection, CHB: Han Chinese in Beijing, China, CHD: Chinese in Metropolitan Denver, Colorado, GIH: Gujarati Indians in Houston, Texas, JPT: Japanese in Tokyo, Japan, LWK: Luhya in Webuye, Kenya, MEX: Mexican ancestry in Los Angeles, California, MKK: Maasai in Kinyawa, Kenya, TSI: Toscans in Italy, YRI: Yoruba in Ibadan, Nigeria).

Phylogenetic. Phylogenetic analysis was performed on the individual allele sharing distance (ASD) matrix of all populations [21]. This ASD matrix captured the underlying genetic differences among all reported individuals. The software Mega4 [22] was used to generate the neighbor-joining tree [23]. The phylogenetic result is shown in Fig 3. It could be seen from the tree that many clusters are classified according to their geographical or ethnic groups, for example JPT, CHB and CHD are grouped together in one major branch. Particularly, the Thai individuals formed a dependent branch in the tree indicating a unique genetic pattern. The Thai genetic variants can be used to study the genetic diversity of Asian peopling and fill the variant spectrum of genetic pattern in Asia.

CNV distribution. The predicted CNV results from 32 Thais were compared with those reported in the database of genomic variants (DGV). The dense SNP markers from 5th generation Affymetrix SNP chip enable the CNV calling program "Aroma.affymetrix" [15] to report more precise CNVs that can suggest possible boundaries of CNV regions (CNVRs) reported in DGV if Thai CNVs are entirely encapsulated within the DGV CNVRs. Moreover, we observed that there were some new CNVs, which were present in Thais only (Fig. 4). This suggested us that CNVs might play a major role in making the Thais genetically unique and standing out from the rest of their peers. More CNV differential comparisons are required to further investigate this issue between every pair of populations.

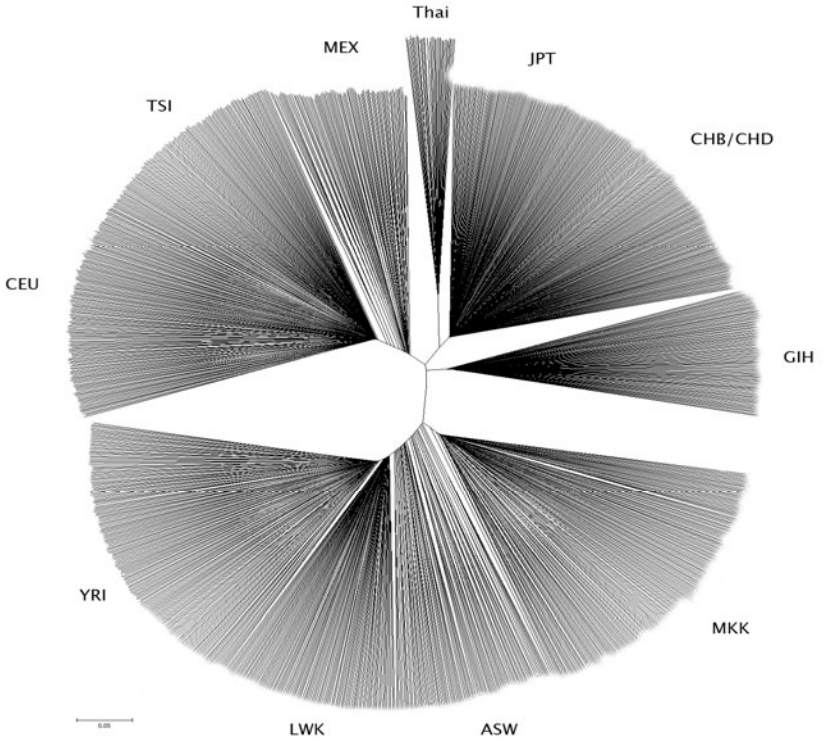


Fig. 3. Phylogenetic tree using neighbor joining analysis approach with allele sharing distance calculated from all individuals.

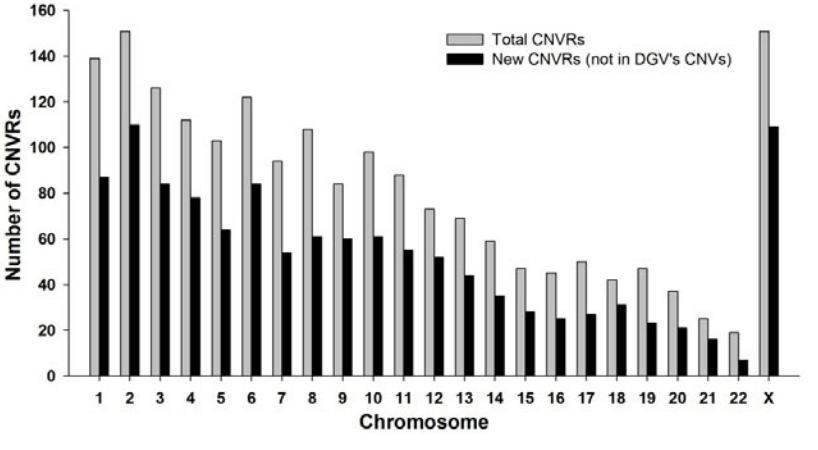


Fig. 4. Number of CNVRs in each chromosome, the light gray columns represent number of total CNVRs. While the dark shades indicate the new CNVRs that have not been reported by the DGV database.

3.2 Database Query

ThaiSNPdb provides 3 main search portals to access the database: SNP search, CNV search and genotype and LD block calculation. Querying results include SNP locations, flanking sequences, alternative SNP identifiers (SNP ids) from various databases, allele frequencies in different populations and SNPs' functions referred by their genomic locations. The map viewer to facilitate cross-database comparisons also displays SNPs and CNVs from other databases. Fig. 5 shows the query results in HTML (Fig. 5a) and graphical display format (Fig. 5b).

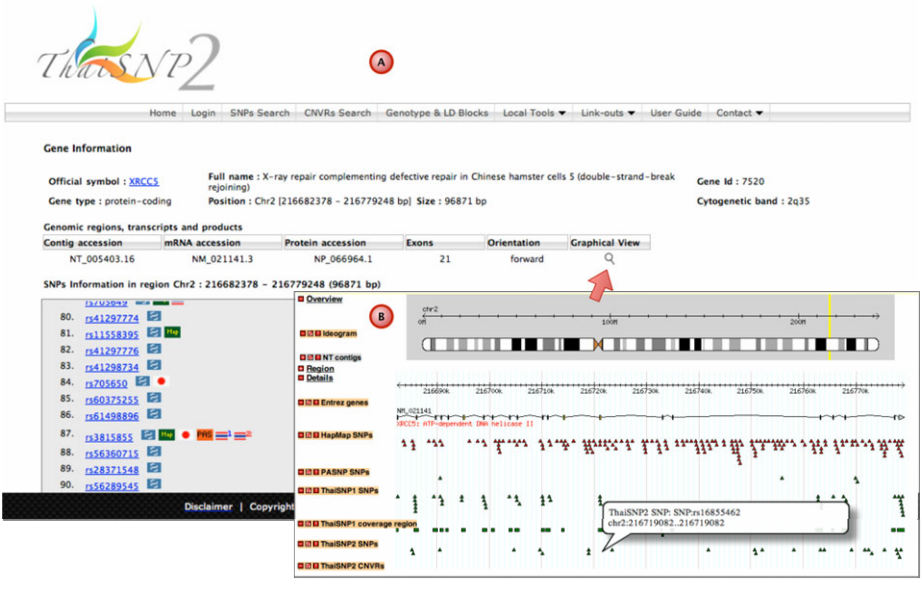


Fig. 5. SNP results from the query of the SNP search page (A) Query results, including gene information and labels indicating SNPs from various databases, in HTML format (B) GBrowse graphical display of SNP locations on a cytogenetic map

SNP Search

This is for accessing SNP and gene information from chromosomal locations, diseases of interest or SNP ids. The genes' upstream and downstream regions can be specified to display all the SNPs in that range.

CNVR Search

The query results from this search feature will be displayed as all CNV regions (CNVRs) on the specified chromosome. All of the CNVRs are larger than 5,000 bases. Sample ids, CNV sizes, CNV types and SNPs present on each CNRV will also be shown. CNVs are categorized as *0: homozygous deletion*, *1: hemizygous deletion*, *2: normal*, *3: single copy gain* and *4-6: multiple copy gain*. Location of each CNVR can be viewed on interactive graphical chromosome charts rendered by GBrowse.

Genotypes and LD Blocks
The genotypes of the Thai samples from the ThaiSNP projects can be queried either by specifying chromosome regions or RefSeq gene name/gene IDs. The query results can be downloaded in Comma Separated Value (CSV) and HapMap style format, the latter of which allows further analyses of LD blocks with Haploview program [24].

GBrowse Graphical Representation
Graphical representation in the form of cytogenetic maps (rendered by GBrowse) accompanied with SNP and CNV indicators is also provided to facilitate cross-database comparisons (see the list of databases included in Table 2). The cytogenetic map marked with SNPs and CNVs of interest can be accessed by choosing a graphical view option on the query result webpage. Similarly, from the graphical view, users can click on the markers to see the details of each polymorphism in that region in normal HTML format. The graphical maps can be downloaded as decorated FASTA files or HapMap GFF files.

3.3 Database Tools

ThaiSNPdb provides a number of interactive analyses including calculation of allele frequencies, Hardy-Weinberg equilibrium, heterozygosity and functional SNP prediction. The parameters, such as population numbers and expected SNPs on genes or genomes, can be adjusted. In order to manage the large information on ThaiSNPdb, the requested information is managed on demand. Users have to log in and, under the menu Genotype and LD blocks, input their queries by gene names or location on chromosomes. The system will search and return all SNPs in the specified areas.

Furthermore, ThaiSNPdb provides a tool for calculating haplotypes, using embedded Haploview which utilizes HapMap format file as input data. The genotype information in the database can also be used to analyze linkage disequilibrium. The HapMap file can be automatically created and analyzed by using ThaiSNP LD Block function. Fig. 6 illustrates LD plot results from Haploview.

3.4 Possible Uses of ThaiSNPdb

ThaiSNPdb can be used to compare many aspects in human genetic polymorphisms, for example, allele frequencies and heterozygosity in different populations and chromosomal regions with many polymorphisms and gene functions in relation to regions of polymorphism. The integration of such knowledge is crucial in the study of genetic epidemiology and disease risk across populations. ThaiSNPdb can also be used to study monogenic disease, which requires the use of genetic markers to compare linkage. Moreover, the Thai specific CNV information can be used in molecular pharmacology and the study of drug responses that are related to unevenly distributed CNVs among individuals.

Information from ThaiSNPdb can be used in population genetics, particularly in the studies of genetic relatedness and population structure of Thai people. Such studies could provide better understanding of admixture ratio in Thai population. Furthermore, it could classify individuals into populations or subpopulations. This is of great importance when considering other applications that utilize genetic polymorphism information in genome-wide association studies.

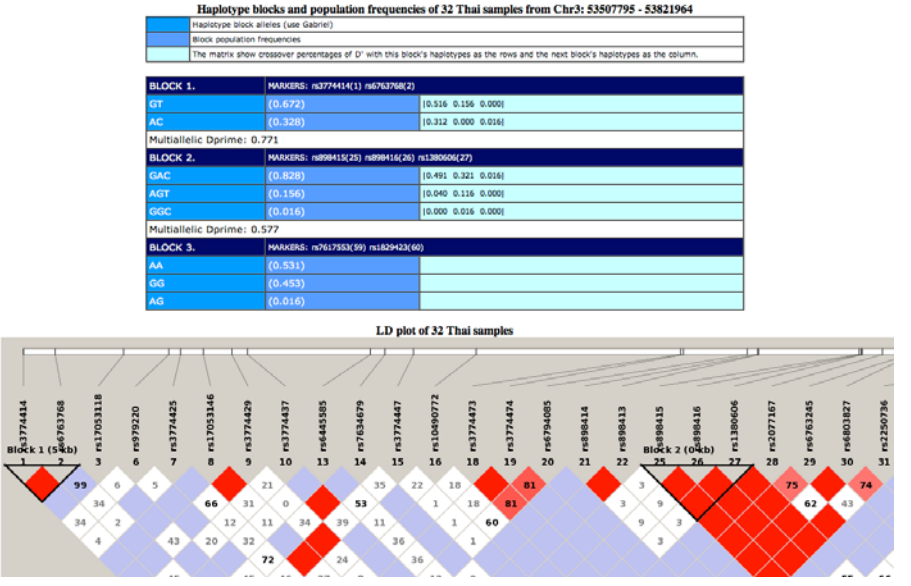


Fig. 6. LD plot result analyzed and rendered graphically by Haploview

4 Conclusion

With ever-increasing amounts of genotype data, ThaiSNPdb serves as a central public genetic database for genetic studies in Thailand. It not only serves as a local reference for genetic makeups, but also provides comparative view of the genetic variances across many populations. The database is equipped with many flexible search features and easy-to-use graphical user interface, offering rich information on both SNP and CNV information. The exporting features are provided for researchers to extract the selected variants to be used in their researches. Data organization of such nature shall lead to a more effective use of genetic polymorphisms of Thai population.

Acknowledgments. This work was supported by the National Center for Genetic Engineering and Biotechnology of Thailand, School of Information Technology and School of Bioresources and Technology King Mongkut's University of Technology Thonburi. Anunchai Assawamakin was supported by BIOTEC postdoctoral grant.

References

1. Collins, F.S., Morgan, M., Patrinos, A.: The Human Genome Project: Lessons from Large-Scale Biology. Science 11, 286 (2003)
2. Shendure, J., Ji, H.: Next-Generation DNA Sequencing. Nat. Biotechnol. 26(10), 1135–1145 (2008)
3. Hawkins, R.D., Hon, G.C., Ren, B.: Next-Generation Genomics: an Integrative Approach. Nat. Rev. Genet. 11, 476–486 (2010)

4. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., Hirschhorn, J.N.: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. 9, 356–369 (2008)
5. Flintoft, L.: Human Disease: Joining the Dots from SNPs to Proteins. Nat. Rev. Genet. 9, 496 (2008)
6. Vignal, A., Milan, D., SanCristobal, M., Eggen, A.: A Review on SNP and other Types of Molecular Markers and their Use in Animal Genetics. Genet. Sel. E 34, 275–305 (2002)
7. Thomson, R.C., Wang, I.J., Johnson, J.R.: Genome-enabled development of DNA markers for ecology, evolution and conservation. Mol. Ecol. 19(11), 2184–2195 (2010)
8. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., Cho, E.K., Dallaire, S., Freeman, J.L., Gonzalez, J.R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J.R., Marshall, C.R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M.J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D.F., Estivill, X., Tyler-Smith, C., Carter, N., Aburatani, H., Lee, C., Jones, K.W., Scherer, S.W., Hurles, M.E.: Global Variation in Copu Number in the Human Genome. Nature 444(23), 444–454 (2006)
9. Estivill, X., Armengol, L.: Copy Number Variants and Common Disorders: Filling the Gaps and Exploring Complexity in Genome-Wide Association Studies. PLoS Genet. 3(10), 1787–1799 (2007)
10. Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y.H., Hicks, J., Spence, S.J., Lee, A.T., Puura, K., Lehtimäki, T., Ledbetter, D., Gregersen, K., Bregman, J., Sutcliffe, J.S., Jobanputra, V., Chung, W., Warburton, D., King, M.C., Skuse, D., Geschwind, D.H., Gilliam, T.C., Ye, K., Wigler, M.: Strong association of de novo copy number mutations with autism. Science 316(5823), 445–449 (2007)
11. International HapMap Consortium: A Second Generation Human Haplotype Map of Over 3.1 Million SNPs. Nature 449(7164), 851–861 (2007)
12. Sherry, S.T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Sirotkin, K.: dbSNP: the NCBI Database of Genetic Variation. Nucleic Acids Research 29(1), 308–311 (2001)
13. Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T., Nakamura, Y.: JSNP: a Database of Common Gene Variations in the Japanese Population. Nucleic Acids Res. 30(1), 158–162 (2002)
14. Li, G., Ma, L., Song, C., Yang, Z., Wang, X., Huang, H., Li, Y., Li, R., Zhang, X., Yang, H., Wang, J., Wang, J.: The YH Database: the First Asian Diploid Genome Database. Nucleic Acids Res., D1025–D1028 (2009)
15. Bengtsson, H., Simpson, K., Bullard, J., Hansen, K.: Aroma.Affymetrix: A Generic Framework in R for Analyzing Small to Very Large Affymetrix Data Sets in Bounded Memory. Tech Report 745, Department of Statistics, University of California, Berkeley (2008)
16. Mahasirimongkol, S., Chantratita, W., Promso, S., Pasomsab, E., Jinawath, N., Jongjaroenprasert, W., Lulitanond, V., Krittayapoositpot, P., Tongsima, S., Sawanpanyalert, P., Kamatani, N., Nakamura, Y., Sura, T.: Similarity of the Allele Frequency and Linkage Disequilibrium Pattern of Single Nucleotide Polymorphisms in Drug-Related Gene Loci between Thai and Northern East Asian Populations: Implications for Tagging SNP Selection in Thais. J. Hum. Genet. 51(10), 896–904 (2006)
17. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., Lee, C.: Detection of Large-Scale Variation in the Human Genome. Nat. Genet. 36(9), 949–951 (2004)

18. Pruitt, K.D., Tatusova, T., Maglott, D.R.: NCBI Reference Sequence (RefSeq): a Curated Non-redundant Sequence Database of Genomes, Transcripts and Proteins. Nucleic Acids Res. 1(35) (Database issue), D61–D65 (2007)
19. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., Lewis, S.: The Generic Genome Browser: a Building Block for a Model Organism System Database. Genome Res. 12(10), 1599–1610 (2002)
20. Webware for Python, `http://www.webwareforpython.org/`
21. Mountain, J.L., Cavalli-Sforza, L.L.: Multilocus Genotypes, a Tree of Individuals, and Human Evolutionary History. Am. J. Hum. Genet. 61(3), 705–718 (1997)
22. Smith, S., Beaulieu, J., Donoghue, J.: Mega-Phylogeny Approach for Comparative Biology: an Alternative to Supertree and Supermatrix Approaches. BMC Evol. Biol. 9(1), 37 (2009)
23. Saitou, N., Nei, M.: The Neighbor-Joining Method: a New Method for Reconstructing Phylogenetic Trees. Mol. Biol. E 4(4), 406–425 (1987)
24. Barrett, J.C., Fry, B., Maller, J., Daly, M.J.: Haploview: Analysis and Visualization of LD and Haplotype Maps. Bioinfo. 21(2), 263–265 (2005)

The Genome Atlas Resource

Matloob Qureshi, Eva Rotenberg, Hans-Henrik Stærfeldt,
Lena Hansson, and David W. Ussery

Center for Biological Sequence Analysis, Department of Systems Biology,
The Technical University of Denmark, 2800 Lyngby, Denmark

Abstract. The Genome Atlas is a resource for addressing the challenges of synchronising prokaryotic genomic sequence data from multiple public repositories. This resource can integrate bioinformatic analyses in various data format and quality. Existing open source tools have been used together with scripts and algorithms developed in a variety of programming languages at the Centre for Biological Sequence Analysis in order to create a three-tier software application for genome analysis. The results are made available via a web interface developed in Java, PHP and Perl CGI. User-configurable and dynamic views of Chromosomal maps are made possible through an updated GeneWiz browser (version 0.94) which uses Java to allow rapid zooming in and out of the atlases.

Keywords: Genome atlas, web interface, chromosomal maps, genome analysis.

1 Introduction

There are, at the time of writing, over 1200 completed Archaeal and Bacterial genome sequences available in the major public repositories of sequence data. However, for a number of reasons these repositories are not in complete synchronisation with each other [1]. Furthermore, the number of genomes published per year has been rising rapidly since the first genome published in 1995 [2] and the advent of "next generation" sequencing technologies which allow a bacterial genome to be sequenced, assembled and annotated in a day [3,4] serve to highlight the need for tools to assist with comparative genomic analysis.

Most bacterial genomes would be thousands of pages long, if viewed as text. Therefore there is a need to collate these projects and present them together in an integrated site, along with links to a graphical overview of the chromosomal sequences. We use asynchronous client-server communication to develop zoomable atlases, which can go from a full chromosomal view, down to the level of individual nucleotides, smoothly and quickly. Although there are other web-based services such as Entrez NCBI genomes [5] and EnsEMBL genomes [6], only the CBS Genome Atlas application focuses on collecting prokaryotic sequence data from multiple sources together with the results of detailed genomic and structural analyses in a user-configurable and dynamic manner [7, 8].

In this work, open source tools such as BioPerl and eHive [9,10] have been used together with Perl scripts and algorithms in other programming languages to develop a three-tier software application for genome analysis [11]. This resource is available

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 141–150, 2010.

at http://www.cbs.dtu.dk/services/GenomeAtlas/. The structure of the rest of this paper is as follows: section 2 covers the three tier architecture of the genome atlas, section 3 describes the GeneWiz chromosome atlas browser, section 4 presents the genome atlas data model, Some examples of scientific uses of the web pages are described in section 5 and finally conclusions are made in the last section.

2 Three Tier Architecture

The Genome Atlas has been redeveloped as three-tier application. Multi-tier applications are concerned with partitioning components of an application into layers, each concerned with a different aspect of the system in order to facilitate flexibility and reuse. Multi-tier architectures are often used in client server applications [11]. The three tiers of Genome Atlas system are described diagrammatically in **Fig. 1**. The data tier is concerned with storage and access to the information used by the application. A data access API developed using BioPerl presents an interface to this information that can be accessed from the logic tier. This tier is involved with the coordination and processing of data and is where the analysis pipeline resides. The final tier of the application is the presentation tier where Genome Atlas data is presented via a web server using a combination of PHP, Perl web pages and the GeneWiz Java application.

2.1 The Data Tier

Completed prokaryotic genome sequence data are downloaded from the three main public sequence repositories, GenBank, EMBL and DDBJ. Although they regularly synchronise data amongst themselves, a small number of projects can be missing from one or more resource. Therefore the Genome Atlas database draws data from all three resources and also from projects referenced in the Genomes On-Line Database (GOLD). The MD5 sum of the DNA sequence is used to minimise redundancy of information stored in the Genome Atlas. Project accession numbers and identifiers are used to map data between the different sources. The GOLD database provides references to genome projects that are not yet in the main repositories and these are also downloaded where possible.

The DNA sequence data are stored as files in GenBank and EMBL format in a directory and file structure as described previously [7]. A MySQL database is used to collate all the information associated with a project, including the location of the all sequence files and the associated taxonomic data. The results of any computationally intensive analyses are stored in the database where appropriate or in as a set of files in a similar directory structure to the raw sequence data files. A data access layer written in Perl with BioPerl provides an integrated object-oriented interface to all the stored information. This can be accessed, by the analysis pipeline in the logic tier and by any other scripts written by users with accounts at the institute.

2.2 The Logic Tier

The logic tier consists of a set of bioinformatic applications written in a variety of programming languages including C, Perl, Python and shell scripts. The execution of these applications in sequence to create an analysis pipeline is managed by the eHive

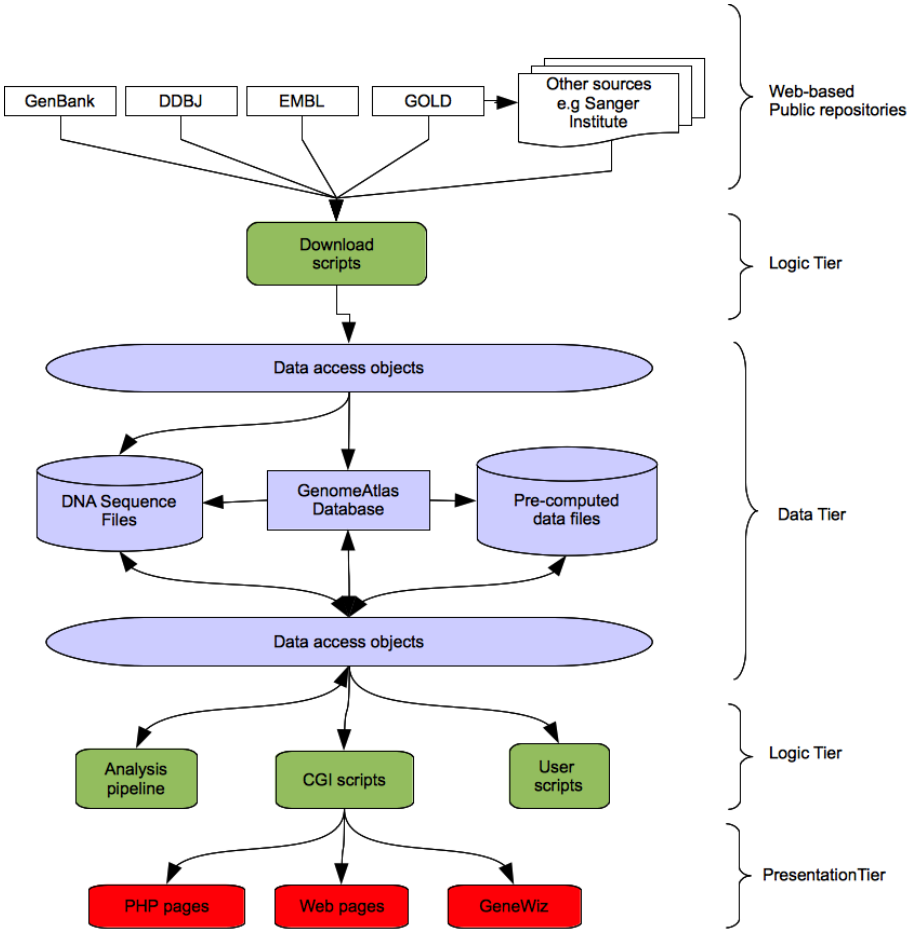


Fig. 1. Genome Atlas Application Diagram. The application is divided into three-tiers. The Data tier is concerned with the storage and retrieval of information. Data access objects are used to present a simple common application interface to the underlying data. The Logic tier performs more complex processing. The download scripts for acquisition of Sequence data from public repositories and the analysis pipeline reside in this tier. The third tier is the presentation tier, where data is reformatted and presented in web forms, static and dynamic chromosome atlases, and genome summaries.

workflow management system developed for the EnsEMBL project [10]. The executable of interest is called from a Perl module that controls the input, output and execution environment of the software. The module together with the pre- and post-conditions are stored in the eHive database. Thus the eHive system allows analyses to be grouped into a set of independent work packages which make efficient use of our computing cluster. It also manages the flow of data from one set of analyses to the next and provides a fault-tolerant system for handling situations when problems occur.

2.3 The Presentation Tier

The Genome Atlas web pages are written in PHP and are updated regularly. The basis for displaying data and analyses in the Genome Atlas is the chromosomal map [12]. Many kinds of chromosomal maps can be displayed together, so that the chromosome can be visualised together with genes, repeat sequences and structural data. Some of these are generated by the analysis pipeline and stored as vector graphics, compressed bit maps or binary files whilst others are created on request in the logic tier.

3 The GeneWiz Chromosome Atlas Browser

The GeneWiz browser provides a dynamic scalable and zoomable view of the chromosomal maps. It is written in Java, using the AWT and Swing libraries for the GUI components. The data can be taken from existing genome projects stored in Genome Atlas in which case the browser makes use of predefined settings and binary files. Alternatively, user defined data can be uploaded in a number of formats. The properties of the chromosome are calculated by the CBS computing cluster and displayed as tracks in the chromosome map display. Asynchronous requests are made to the server whilst the display is rendered so when the atlas is clicked or a section is selected there is a smooth zoom in or out from one level of magnification to another, all the way through to the individual bases in the sequence. Pre-calculated binary files are requested by the client and the server returns the data necessary for the display [8]. Some of the tracks, for example, global inverted repeats are displayed with a colour scheme based on global properties, whilst the colours and statistics for other tracks, such as GC skew, are recalculated according to the region of the chromosome being viewed. Details for open reading frames are displayed from the genome annotation when the pointer is moved over the CDS on the appropriate track. The GeneWiz browser also allows views and settings to be stored as a session in a user defined directory. Browsing may then be resumed at any time by loading the session file. The browser also exports data in fasta, support vector graphic, postscript or pdf format.

4 The Genome Atlas Data Model

The information stored in the Genome Atlas data tier is coordinated via a set of tables in a MySQL relational database. The data model for this database is shown in **Fig. 2**. The main table concerned with storing data for coordinating the Genome Atlas with external repositories is the Projects table. This table links an internal project identifier and version with data related to the Organism, such as NCBI organism taxonomy. It also references information about the project from public repositories stored in the Sources table. A project is linked to one or more DNA sequences through the Segments table. This table contains information about to the actual location of the DNA sequence and other related data such as length and MD5 check-sum. The MD5 values are used to generate incremental version numbers whenever the DNA sequences referenced by Segments are updated. The results of a given bioinformatic analysis are

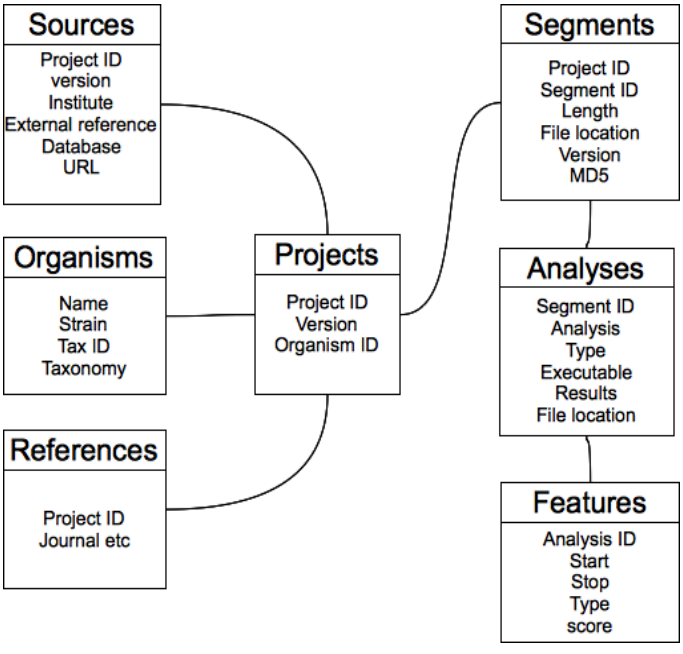


Fig. 2. Genome Atlas Database Data model. The main tables in the Genome Atlas are shown. The Projects table links to the Sources, Organisms and References tables. This provides a way of querying external meta-data linked to the project. The Segments table links the Projects to DNA Sequence files stored in a file system, whilst the Analyses table connects the DNA Sequence to the results produced by running the analysis pipeline on a project.

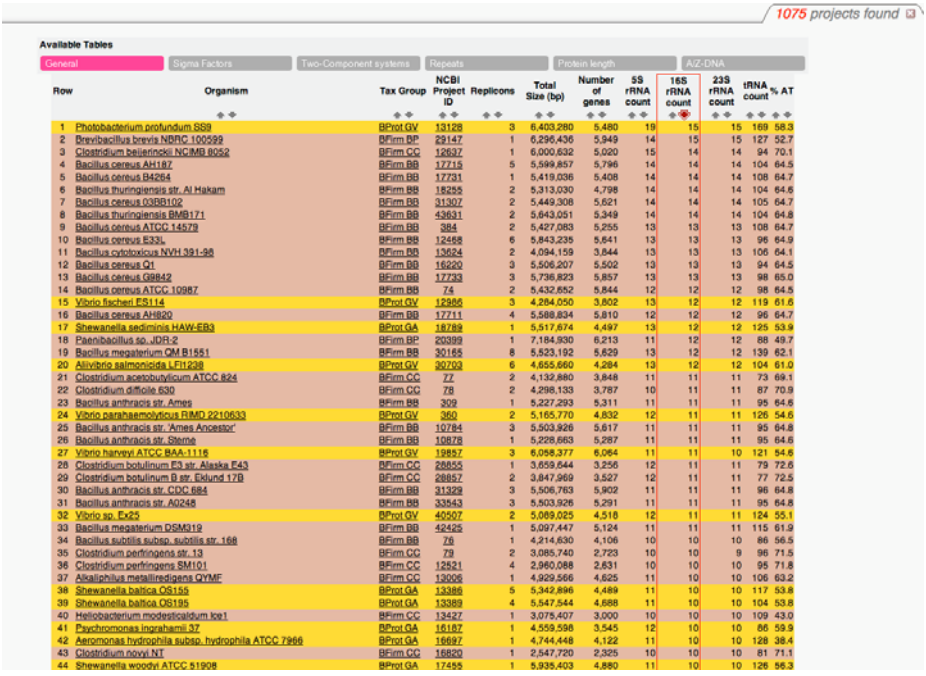
stored in the Analyses table. This table records the analysis module, the executable and version together with the results and any file locations or sequence features generated by the analysis pipeline. This allows the logic tier to select attributes and features based created by specific versions of analysis executables.

5 Using the Genome Atlas for Research

The Genome Atlas can be used to analyse the properties of genomes in order to test hypotheses. Here we present two short investigations conducted through the Genome Atlas application.

5.1 Correlation between Growth Rate and the Properties of Chromosomes

The Genome Atlas analysis pipeline includes analysis by RNAmmer [13], therefore we can display the number of predicted 5S, 16S and 23S ribsomal RNA genes in each genome. This can found on the General Genomes web page. The genomes table displayed here can be sorted on any column in ascending or descending order by clicking on the arrows underneath the column heading (**Fig. 3**).



1075 projects found

Available Tables

General | Sigma Factors | Two-Component systems | Repeats | Protein length | AIZ-DNA

Row	Organism	Tax Group	NCBI Project ID	Replicons	Total Size (bp)	Number of genes	5S rRNA count	16S rRNA count	23S rRNA count	tRNA count	% AT
1	Photobacterium profundum SS9	BProt GV	13128	3	6,403,280	5,480	19	15	15	169	58.3
2	Brevibacillus brevis NBRC 100599	BFirm BP	29147	1	6,296,436	5,949	14	15	15	127	52.7
3	Clostridium beijerinckii NCIMB 8052	BFirm CC	12637	1	6,000,632	5,020	15	14	14	94	70.1
4	Bacillus cereus AH187	BFirm BB	17715	5	5,599,857	5,796	14	14	14	104	64.5
5	Bacillus cereus B4264	BFirm BB	17731	1	5,419,036	5,408	14	14	14	108	64.7
6	Bacillus thuringiensis str. Al Hakam	BFirm BB	18295	2	5,313,030	4,798	14	14	14	104	64.6
7	Bacillus cereus 03BB102	BFirm BB	31307	2	5,449,308	5,621	14	14	14	105	64.7
8	Bacillus thuringiensis BMB171	BFirm BB	43631	2	5,643,051	5,349	14	14	14	104	64.8
9	Bacillus cereus ATCC 14579	BFirm BB	384	2	5,427,083	5,255	13	13	13	108	64.7
10	Bacillus cereus E33L	BFirm BB	12468	6	5,843,235	5,641	13	13	13	96	64.9
11	Bacillus cytotoxicus NVH 391-98	BFirm BB	13624	2	4,094,159	3,844	13	13	13	106	64.1
12	Bacillus cereus Q1	BFirm BB	16220	3	5,506,207	5,502	13	13	13	94	64.5
13	Bacillus cereus G9842	BFirm BB	17733	3	5,736,823	5,857	13	13	13	98	65.0
14	Bacillus cereus ATCC 10987	BFirm BB	74	2	5,432,552	5,844	12	12	12	98	64.5
15	Vibrio fischeri ES114	BProt GV	12886	3	4,284,050	3,802	13	12	12	119	61.6
16	Bacillus cereus AH820	BFirm BB	17711	4	5,588,834	5,810	12	12	12	96	64.7
17	Shewanella sediminis HAW-EB3	BProt GA	18789	1	5,517,674	4,497	13	12	12	125	53.9
18	Paenibacillus sp. JDR-2	BFirm BP	20399	1	7,184,930	6,213	11	12	12	88	49.7
19	Bacillus megaterium QM B1551	BFirm BB	30165	8	5,523,192	5,629	13	12	12	139	62.1
20	Aliivibrio salmonicida LFI1238	BProt GV	30703	6	4,655,660	4,284	13	12	12	104	61.0
21	Clostridium acetobutylicum ATCC 824	BFirm CC	77	2	4,132,880	3,848	11	11	11	73	69.1
22	Clostridium difficile 630	BFirm CC	78	2	4,298,133	3,787	10	11	11	87	70.9
23	Bacillus anthracis str. Ames	BFirm BB	309	1	5,227,293	5,311	11	11	11	95	64.6
24	Vibrio parahaemolyticus RIMD 2210633	BProt GV	360	2	5,165,770	4,832	12	11	11	126	54.6
25	Bacillus anthracis str. 'Ames Ancestor'	BFirm BB	10784	3	5,503,926	5,817	11	11	11	95	64.8
26	Bacillus anthracis str. Sterne	BFirm BB	10878	1	5,228,663	5,287	11	11	11	95	64.6
27	Vibrio harveyi ATCC BAA-1116	BProt GV	19857	3	6,058,377	6,064	11	11	10	121	54.6
28	Clostridium botulinum E3 str. Alaska E43	BFirm CC	28855	1	3,659,644	3,256	12	11	11	79	72.6
29	Clostridium botulinum B str. Eklund 17B	BFirm CC	28857	2	3,847,969	3,527	12	11	11	77	72.5
30	Bacillus anthracis str. CDC 684	BFirm BB	31329	3	5,506,763	5,902	11	11	11	96	64.8
31	Bacillus anthracis str. A0248	BFirm BB	33543	3	5,503,926	5,291	11	11	11	95	64.8
32	Vibrio sp. Ex25	BProt GV	40507	2	5,069,025	4,516	12	11	11	124	55.1
33	Bacillus megaterium DSM319	BFirm BB	42425	1	5,097,447	5,124	11	11	11	115	61.9
34	Bacillus subtilis subsp. subtilis str. 168	BFirm BB	76	1	4,214,630	4,106	10	10	10	86	56.5
35	Clostridium perfringens str. 13	BFirm CC	79	2	3,085,740	2,723	10	10	9	96	71.5
36	Clostridium perfringens SM101	BFirm CC	12521	4	2,960,088	2,631	10	10	10	95	71.8
37	Alkaliphilus metalliredigens QYMF	BFirm CC	13006	1	4,929,566	4,625	11	10	10	106	63.2
38	Shewanella baltica OS155	BProt GA	13386	5	5,342,896	4,489	11	10	10	117	53.8
39	Shewanella baltica OS195	BProt GA	13389	4	5,547,544	4,688	11	10	10	104	53.8
40	Heliobacterium modesticaldum Ice1	BFirm CC	13427	1	3,075,407	3,000	10	10	10	109	43.0
41	Psychromonas ingrahamii 37	BProt GA	16187	1	4,559,598	3,545	12	10	10	86	59.9
42	Aeromonas hydrophila subsp. hydrophila ATCC 7966	BProt GA	16697	1	4,744,448	4,122	11	10	10	128	38.4
43	Clostridium novyi NT	BFirm CC	16820	1	2,547,720	2,325	10	10	10	81	71.1
44	Shewanella woodyi ATCC 51908	BProt GA	17455	1	5,935,403	4,880	11	10	10	126	56.3

Fig. 3. The General Genome Atlas table here shows more than a thousand bacterial genomes sorted by number of 16S Ribsomal RNA in descending order

The first entry in the table is *Photobacterium profundum* (GenBank project ID 13128) which is a member of Vibrionales; known for their short doubling time [14]. The other group heavily represented at the top of the table are Firmacutes, notably Clostridia. The second entry, *Brevibacillus brevis* (Genbank project ID 29147), is a member of this group. It has been observed that the leading and lagging strand of bacterial chromosomes have differing "skews" of nucleotides present in one of the two stands. The A/T and G/C biases of short oligomers can be plotted along the chromosome to allow visualisation of any such biases [15]. **Fig. 4** shows these plots of the oligomer bias towards the leading strand for a number of different organisms.

Fig. 4a shows the strand bias of *Photobacterium profundum* (a Gram negative Gammaproteobacteria); it can be seen that the Guanines are over-represented on the leading strand whilst the Adenines are on the lagging strand. In contrast, **Fig. 4b** shows *Brevibacillus brevis* (a Gram positive Firmacute). The green and the blue lines follow the same direction with respect to replication origin and terminus. The Adenine bias in **Fig. 4a** is much weaker and in the opposite direction. Thus, although both are highly streamlined, they show the opposite strand bias. This difference in the bias probably reflects a historical contingency as the *polC* gene encodes a proofreading subunit in DNA polymerase which is present in *Brevibacillus brevis* but absent from *Photobacterium profundum* [14].

Looking back at the table in **Fig. 3**, the third organism on the list, *Clostridium beijerinckii* (GenBank project ID 12637) has an AT content of 70%, which is considerably higher than many bacterial genomes. Many of the Firmicute genomes in this list

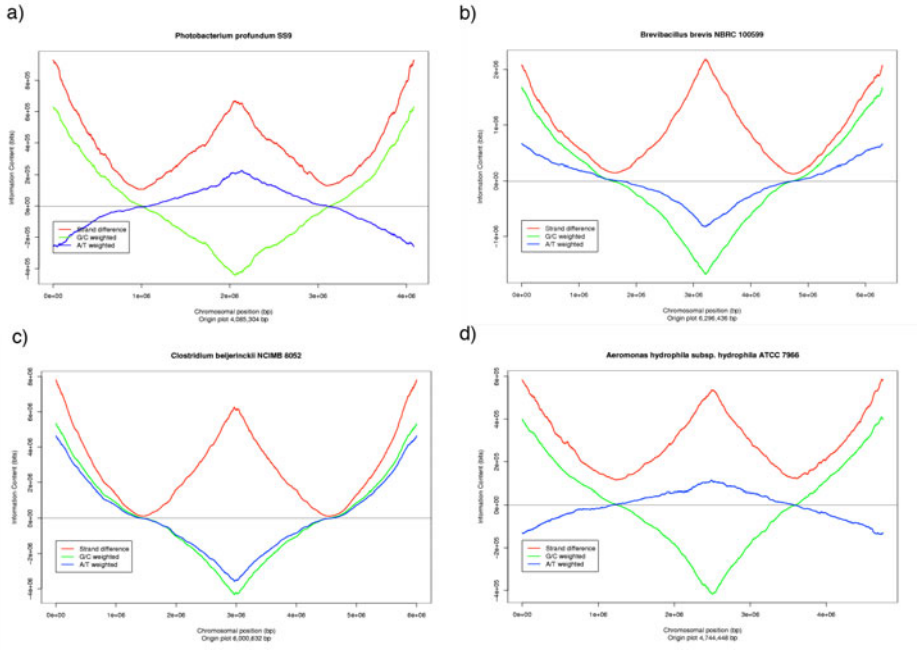


Fig. 4. Plots of the G/C skew, A/T skew on the leading strand and the strand difference of oligomers from the chromosome. (a) shows *Photobacterium profundum* SS9, (b) is *Brevibacillus brevis* NBRC 100599, (c) is *Clostridium beijerinckii* NCIMB 8052 and (d) is *Aeromonas hydrophila* subsp. hydrophila ATCC 7966.

are AT rich. We can see a similar oligomer skew in **Fig. 4c**. For comparison, the skew diagram for a GC rich organism from the table in **Fig. 3.** is also shown (*Aeromonas hydrophila*: GenBank project 16697). Notice there is still a strong strand bias, although, like *Photobacterium profundum*, the Guanines and Adenines are biased on the opposite strand (i.e the blue and green lines are opposite).

5.2 Zooming into Regions of Chromosomes

We chose to visualise the chromosome of *Brevibacillus brevis* (the second organism shown in **Fig. 3**) via the GeneWiz browser (**Fig. 5**). There are many repeat regions visible in the chromosome map (shown in the red and blue lanes in the inner circles). There are also some regions of high intensity in the 'position preference' lane (shown as dark green). These areas of the chromosome are likely to exclude chromatin proteins, and as such are often the location for highly expressed genes [16].

Fig. 5a Shows the whole chromosome zoomed out whilst **Fig. 5b**, shows a magnified view of the area near the top of the circle. This is in a region closer to the replication origin, containing genes coding for the beta/beta prime subunit of RNA polymerase, which are known to be highly expressed. At the bottom of the chromosome, close to the replication terminus, is another region, containing the *lrg* operon, which controls synthesis on an antibiotic peptide linear gramicidin. This operon can

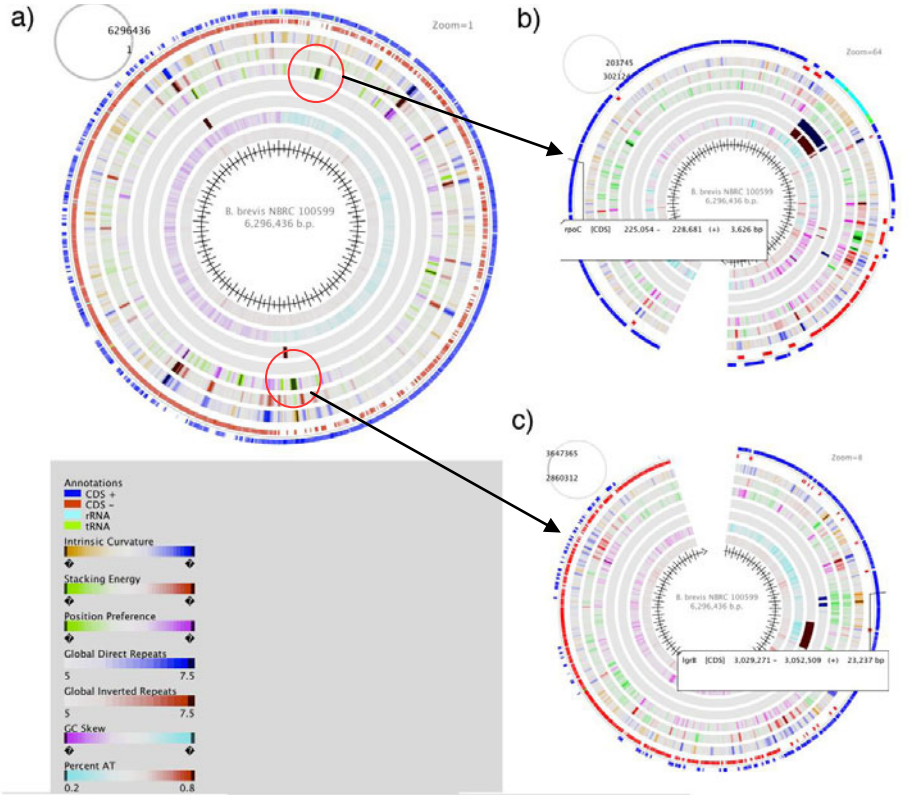


Fig. 5. Zoomable atlases. (a) shows the whole chromosome, and (b) and (c)a zoom of two regions.

also be highly expressed under the right conditions such as sporulation [17]. This ability to view the physical properties of the chromosome in order to identify regions likely to contain highly expressed genes followed by closer analysis by magnification of the region of interest represents a very useful tool for studying bacterial chromosomes.

6 Conclusion

We hope that the Genome Atlas application will result in the development of an application capable of scaling to provide a useful resource for analysis of completed prokaryotic genomes as the number of such projects continues to increase. The use of widely accepted software development patterns such as multi-tier architecture and of existing open source projects wherever possible should provide a robust platform for dealing with the explosion of various new types of high-throughput sequencing data [18], as well as emerging single-molecule methods or the so-called ‘third generation’ sequencing technologies [19].

We have also demonstrated how some of the unique features of the Genome Atlas database, such as the ability to view and sort by the number of predicted ribosomal RNA genes and the zoomable atlases can be used to assist biological investigations.

The only method of interaction with the Genome Atlas currently is a web browser, but future work on the Genome Atlas could include making the application available as a web service based on the Representational State Transfer (REST) architecture [20]. This would allow users to make use of resources provided by the application for via automated tools and incorporated in services of their own.

References

1. Lagesen, K., Ussery, D.W., Wassenaar, T.W.: The One Thousandth Genome - A Cautionary Tale. Microbiology 156, 603–608 (2010)
2. Fleischmann, R.D., et al.: Whole-genome random sequencing and assembly of *Haemophilus influenzae*. Rd. Science 269, 496–512 (1995)
3. Flicek, P., Birney, E.: Sense from sequence reads: methods for alignment and assembly. Nature Methods 6, S6–S12 (2009)
4. Reeves, G.A., Talavera, D., Thornton, J.M.: Genome and proteome annotation- organization, interpretation and integration. J. R. Soc. Interface 6, 129–147 (2009)
5. Wheeler, D.L., et al.: Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 36, D13–D21 (2008)
6. Kersey, P.J., et al.: Ensembl Genomes: Extending ensembl across the taxonomic space. Nucleic Acids Research 38, D563–D569 (2010)
7. Hallin, P.F., Ussery, D.W.: CBS Genome Atlas Database: A dynamic storage for bioinformatic results and sequence data. Bioinformatics 20, 3682–3686 (2004)
8. Hallin, P., Stærfeldt, H., Rotenberg, E., Binnewies, T., Benham, C., Ussery, D.: GeneWiz browser: An Interactive Tool for Visualizing Sequenced Chromosomes. Standards in Genomic Sciences 1(2), October 14 (2009), `http://standardsingenomics.org/index.php/sigen/article/view/sigs.28177` (Date accessed: August 26, 2010)
9. Stajich, J.E., et al.: The Bioperl toolkit: Perl modules for the life sciences. Genome Research 12, 1611–1618 (2002)
10. Severin, J., Beal, K., Vilella, A., Fitzgerald, S., Schuster, M., Gordon, L., Ureta-Vidal, A., Flicek, P., Herrero, J.: eHive: An Artificial Intelligence workflow system for genomic analysis. BMC Bioinformatics 11, 240 (2010)
11. Ramirez, A.O.: Three-Tier Architecture. Linux Journal (75), July 01 (2000), `http://www.linuxjournal.com/article/3508`
12. Jensen, L.J., Carsten, F., Ussery, D.W.: Three Views of Microbial Genomes. Research in Microbiology 150, 773–777 (1999)
13. Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., Ussery, D.W.: RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35, 3100–3108 (2007)
14. Reen, F., Almagro-Moreno, S., Ussery, D., Boyd, E.: The genomic code: inferring Vibrionaceae niche specialization. Nature Reviews: Microbiology 4, 697–704 (2006)
15. Worning, P., Jensen, L.J., Hallin, P.F., Stearfeldt, H.H., Ussery, D.W.: Origin of Replication in Circular Prokaryotic Chromosomes. Environmental Microbiology 8, 353–361 (2006)

16. Willenbrock, H., Ussery, D.W.: Prediction of highly expressed genes in microbes based on chromatin accessibility. BMC Mol. Biol. 13, 8–11 (2007)
17. Nakai, T., Yamauchi, D., Kubota, K.: Enhancement of linear gramicidin expression from Bacillus brevis ATCC 8185 by casein peptide. Biosci. Biotechnol. Biochem. 69(4), 700–704 (2005)
18. Hawkins, R.D., Hon, G.C., Ren, B.: Next-generation genomics: an integrative approach. Nature Reviews Genetics 11, 476–486 (2010)
19. Munroe, D.J., Harris, T.J.: Third-generation sequencing fireworks at Marco Island". Nature Biotechnology 28, 426–428 (2010)
20. Fielding, R.T., Taylor, R.N.: Principled Design of the Modern Web Architecture. ACM Transactions on Internet Technology Association for Computing Machinery 2(2), 115–150 (2002)

INVERTER: INtegrated Variable numbER Tandem rEpeat findeR

Adrianto Wirawan[1,2], Chee Keong Kwoh[1], Li Yang Hsu[2], and Tse Hsien Koh[2]

[1] School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore
{adri0004,asckkwoh}@ntu.edu.sg
[2] Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, 5 Lower Kent Ridge Road, Singapore 119074, Singapore
li_yang_hsu@nuhs.edu.sg, koh.tse.hsien@sgh.com.sg

Abstract. A tandem repeat in DNA is a sequence of two or more contiguous, approximate copies of a pattern of nucleotides. Tandem repeats occur in the genomes of both eukaryotic and prokaryotic organisms. They are important in numerous fields including disease diagnosis, mapping studies, human identity testing (DNA fingerprinting), sequence homology and population studies. Although tandem repeats have been used by biologists for many years, there are few tools available for performing an exhaustive search for all tandem repeats in a given sequence. In this paper, we present INVERTER, a *de novo* tandem repeat finder without the need to specify either the pattern or a particular pattern size, integrated with a data visualization tool. INVERTER is implemented in Java and has a built-in user-friendly Graphical User Interface. A standalone version of the program can be downloaded from http://bmserver.sce.ntu.edu.sg/INVERTER. Comparison search result of INVERTER with an existing software tool is presented. The use of INVERTER will assist biologists in discovering new ways of understanding both the structure and function of DNA and protein.

Keywords: Variable Number Tandem Repeat, exact match search, non-exact match search, data visualization.

1 Introduction

Most genomes have a high content of repetitive DNA. Fifty percent of the human genome, for example, consists of repeated sequences[1]. A tandem repeat (TR) in DNA is a sequence of two or more contiguous, approximate copies of a pattern of nucleotides. Variable Number Tandem Repeat (VNTR) is a location in a genome where a short nucleotide sequence is organized as a tandem repeat.

VNTRs appear in biological sequences with a wide variety and occur in the genomes of both eukaryotic and prokaryotic organisms. They are found in both coding and non-coding regions of DNA. Expansions of repeats found in the protein-coding portions of genes can affect the function of the gene by causing synthesis of malfunctioning

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 151–164, 2010.

proteins. Repeats in non-coding regions have been shown to affect biological processes by affecting gene expression, transcription and translation.

VNTRs are essential in genetics and biology research. They are important as genetic markers[2] as well as responsible for over 30 inherited diseases in humans. Expansions of simple DNA repeats have been linked to hereditary disorders in humans, including Fragile X Syndrome[3], Myotonic Ddystrophy[4], Huntington's Disease[5], spinal and bulbar muscular atrophy[6], various Spinocerebellar Ataxias and Friedreich's Ataxia[7]. These diseases are sometimes called the *repeat expansion diseases* since they are caused by long and highly polymorphic VNTRs [8, 9]. Tetra- or pentanucleotide VNTRs in the human genome are the genetic markers used in DNA forensics[10]. Since the number of adjacent repeated units varies from individual to individual, the copy number of a tandem repeat can be used to identify an individual, and relations such as parent or grandparent. VNTRs are also used in population studies[11], conservation biology[12] as well as multiple sequence alignments[13].

Although VNTRs have been used by biologists for many years, there are few tools available for performing an exhaustive search for all VNTRs in a given sequence. Popular existing software tools for finding VNTRs in a sequence include: Tandem Repeats Finder (TRF)[14], mreps[15], ATRHunter[16], STRING[17], and T-REKS[18].

One of the difficulties involved in locating VNTRs is in the precision of finding a tandem repeat given its loose definition. Exact repeats, i.e. repeats that do not allow any errors, are clearly defined. Once we introduce errors, such as insertions and deletions of single or multiple bases, we have to define what constitutes a tandem repeat. Each of the tools for locating VNTRs relies on certain assumptions and definitions. Thus, the output of the different tools differs, each offering different insights into the presence of repeated sequences. Furthermore, none of the above software tools provide data visualization of the resulting VNTRs to facilitate users to correlate annotations, observe visual patterns, and view useful statistics of the data.

In this paper, we present INVERTER, a *de novo* tandem repeat finder without the need to specify either the pattern or a particular pattern size, integrated with a data visualization tool. INVERTER is aimed to identify both exact match and non-exact match VNTRs and provide data visualization which would allow users to correlate annotations, observe visual patterns, and view useful statistics of the data. INVERTER is implemented in Java and has a built-in user-friendly Graphical User Interface. The use of INVERTER will assist biologists in discovering new ways of understanding both the structure and function of DNA and protein.

The remainder of the paper is organized as follows. Section 2 explains the problem definition and implementation details of INVERTER. Experimental works on 6 completed projects of *Acinetobacter baumanii* genomes using the INVERTER are presented in Section 3. Section 4 concludes the paper.

2 Method

2.1 Problem Definition

There are two principal families of VNTRs: microsatellites and minisatellites. Microsatellites are short tandemly repeated DNA sequences of 1-6 base pairs in

Table 1. Commonly employed terms for VNTRs

Biological definition	Mathematical/Computational description	Features	Example
Perfect	Exact match	100% identical copies	$(A)_n$, $(ATC)_n$
Imperfect	Approximate-Hamming Distance	Substitutions (=mismatches)	$(AC)_nAT(AC)_m$
Interrupted*	Approximate Edit-Distance	substitutions, insertions, deletions (=interruptions)	$(ACG)_nT(ACG)_m$, $(AT)_nCGAG(AT)_m$
Compound/ complex	'Fuzzy'	multiple motifs, periods, substitutions	$(ACG)_nT(TC)_m$

* Interrupted repeats are often included in imperfect repeats

length[19]. Minisatellites, on the other hand, consists of moderately 10-100 base pairs of DNA sequences[20].

It is well known that sequences are subject to many kinds of modifications, such as point mutations, i.e. substitutions, insertions and deletions (indels for short), and expansions, i.e. exact tandem replications of some tracts. Table 1 shows the commonly employed terms for VNTRs based on their types of repeats[21].

Given a sequence S of length l, $S = \{s_1, s_2, \ldots s_l\}$. S contains exact match VNTR r of length k at position p if r can be partitioned into consecutive n sub-sequences of pattern q, as defined in the following equation

$$S = wrw',$$

where $r = \{s_p,\ldots,s_{p+k}\} = (q)_n$, $n > 1$, $w = \{s_1,\ldots,s_{p-1}\}$ and $w' = \{s_{p+k+1},\ldots,s_l\}$.

INVERTER is aimed to identify both exact match and non-exact match VNTRs and provide data visualization which would allow users to correlate annotations, observe visual patterns, and view useful statistics of the data. More specifically, INVERTER is more optimized towards minisatellite VNTRs, although it can be used to identify and visualize microsatellite VNTRS. The reason is that due to the size of microsatellite VNTRs, visual patterns and statistical result provided may not be as significant as in minisatellite VNTRs.

2.2 Implementation

INVERTER is designed to be portable and therefore it is implemented in Java due to its platform-independent characteristics. Java supports four popular Operating Systems, i.e. Windows, Linux, Solaris and MacOS.

INVERTER also has a built-in GUI to allow user to set parameters needed for the identification and visualization of VNTRs in DNA sequences. A standalone version can be downloaded from http://bmserver.sce.ntu.edu.sg/INVERTER. Fig. 1 illustrates the workflow diagram of INVERTER. Currently, there are three features that are available to users, i.e. identification of exact match VNTRs and non-exact match VNTRs in DNA sequences as well as visualization of the resulting tandem repeat data.

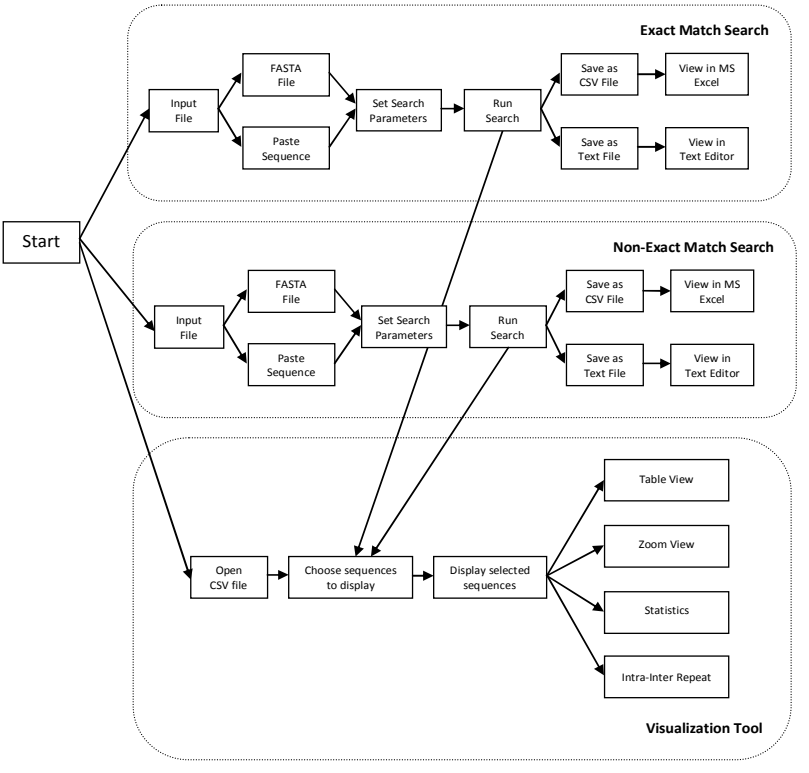


Fig. 1. Workflow diagram of INVERTER

2.2.1 Identification of Exact Match Tandem Repeats

The first feature allows users to identify exact match VNTRs in the input nucleotide sequences. Users can open an input file containing one or more sequences in FASTA format or copy and paste a sequence directly in FASTA format to INVERTER. Several parameters and options of the program can be defined by the users, as seen in Fig. 2.

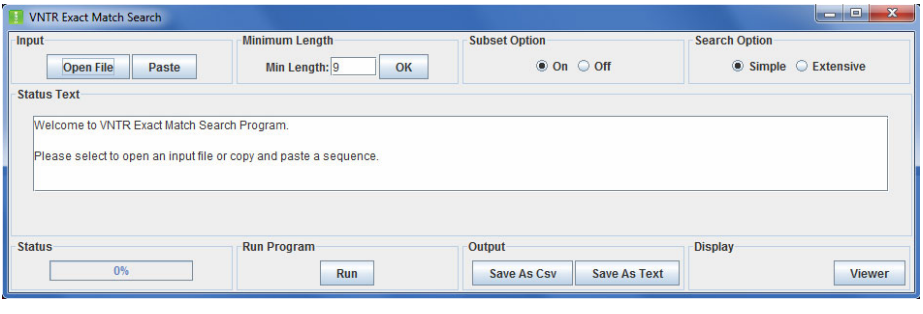


Fig. 2. VNTR Exact Match Search

Minimum length l_{min} defines the minimum length of pattern to be reported in the result in terms of base pairs. The default value l_{min} is 9 bps and is chosen based on the analysis of known minisatellites of biological importance. Setting the value of l_{min} to a smaller value will generate more VNTR results but also increase the probability of inclusion of irrelevant repeats (noise). Given a sequence S of length l, the theoretical maximum of l_{max} is $l/2$. However, although it is possible to increase the maximum length value, it is not feasible to verify it using wet lab experiment due to the limitation of current gel-based technology verification technique, e.g. pulsed-field gel electrophoresis (PFGE). Therefore, the maximum length of searched pattern l_{max} is currently limited to 200.

The subset option indicates whether subset VNTRs should be included in the search result. The *on* option will include subset VNTRs in the result, while *off* will exclude them. The default option is *on*. Given a VNTR r of length k, $r = \{r_1, r_2, \ldots, r_k\}$, a subset VNTR $z = \{z_1, z_2, \ldots, z_k\}$ of length k_z containing m consecutive pattern q_z can be defined as follows:

$$r = vzv',$$

where $z = (q_z)_m$, $1 < m < n$, $1 \leq k_z \leq k$, $v = \{r_1,\ldots,r_{c-1}\}$, $v' = \{r_{c+kz+1},\ldots,r_k\}$ and $1 < c \leq k_z$. For example, a VNTR r = ACGTACGTACGT contains subset VNTRs CGTACGTA, GTACGTAC and TACGTACG.

The search option indicates whether search should be a *simple* or *extensive search*. The *simple search* denotes that each VNTR is searched only in its origin sequence. On the other hand, *extensive search* searches for the occurrence of each VNTR in every sequence in the input file and is suitable to find inter-relationships of VNTRs among the input sequences. The default option is *simple search*.

Given a set of input sequences S and input parameters l_{min} and l_{max}, the exact match search will generate a list of VNTR V, which is then filtered for the occurence of VNTRs consisting of only N-nucleotides. N-nucleotide stands for an unknown nucleotide in some databases. It can be either of the four nucleotides (A, C, G or T). Therefore, it is logical to remove the N-VNTRs in order to reduce redundancy and improve the efficiency of the result. Subsequently, INVERTER checks for the subset and extensive search flag and then executes the necessary processing accordingly. The pseudocode of the exact match search is illustrated in Fig. 3.

After the search has finished, users can opt to use the visualization tool to view the result or save the result in a comma-separated value (csv) or text format. The saved search result contains the exact match VNTRs with their respective relevant data needed for the visualization tool, i.e. their origin sequence, the length of the origin sequence, the pattern, the initial position of the VNTR in the sequence, the length of the VNTR and the number of count.

2.2.2 Identification of Non-exact Match Tandem Repeats

The second feature of INVERTER allows users to identify non-exact match VNTRs. For this purpose, we implement a heuristic algorithm based on the unsupervised classification K-means algorithm[22]. Other VNTR finder tools using K-means algorithm include T-REKS[18].

```
Start
Get input data and relevant input parameters l_min and l_max;
While there are still unprocessed input sequence in |S|
     For i:0 to l_max
          Search for potential VNTR candidate with length >
            l_min using sliding window;
          If a VNTR is found
               Process VNTR data;
               Add VNTR to V;
          End If
     End For
End While
Filter for N-VNTRs;
If Subset Flag is TRUE
     Do subset processing
End If
If Extensive Search Flag is TRUE
     For i:1 to |V|
          For j:1 to |S|
               If the origin of current VNTR is S_j
                    Get the data from
               Else
                    Search the VNTR in other sequences
                     in the dataset;
          End For
     End For
End If
End
```

Fig. 3. Exact Match Search Pseudocode

We use short strings (SS) to detect for the presence of VNTRs and use the lengths between identical neighbouring SSs as datapoints of the K-means algorithm. Hence, all datapoints are partitioned into k clusters for user-defined k. For each partition a centroid is defined. K initial centroids are selected from the dataset. Distances between each datapoint and the centroids are then calculated to assign the datapoint to the cluster which has the nearest centroid. This procedure repeated iteratively until convergence is met. Statistically, the longer is the sequence the higher is the number of occurrences of a given SS. The increase of the occurrences will amplify a background noise and decrease the quality of detection at the clustering steps. Therefore, we split the input sequences to smaller chunks of length 1500 or less to avoid the reduction of detection quality in long sequences and then concatenate them after the search. This strategy is also used in T-REKS. For our implementation, we choose the length of SS l_{SS} to be 4 and k to be 20. These values are obtained empirically.

The candidate VNTR lengths is determined first by selecting the most frequent length *mfl* within each cluster generated. This step is applied to each type of SS found. If a cluster has several most frequent lengths which occur the same number of times, the shortest length is chosen.

Not all *mfl*s may correspond to the VNTR lengths, because a given short string may occur more than one time within a repeat. Hence, we filter the *mfl*s. First, we consider only SSs which are separated by lengths that are equal or close to the *mfl*s. The threshold of closeness of the length to the *mfl*s is proportional to the length, so it

takes into account the variability of the lengths in biological tandem repeats. Second, we scan the sequence and do not consider a downstream SS of the neighbouring SSs except for those which length correspond to one chosen *mfl*s. The lengths are then re-calculated and the scanning is repeated for each of the *mfl*s and leads to k new sets of re-calculated lengths.

K-means algorithm is simultaneous applied to all *mfl*s of all type of SS which will provide k most frequent *mfl*s that can be considered as candidate VNTRs. The level of sequence similarity between the putative repeats of each run is evaluated using Multiple Sequence Alignment (MSA) center-star approach. Based on the obtained MSA of the repeats constituting the runs, we obtain a consensus sequence and subsequently use it as a reference for similarity calculation. Given an alignment made by m repeats of length l. Hamming distance d_i between the consensus sequence and a repeat r_i with $1 \leq i \leq m$ are calculated. A similarity coefficient for the whole alignment as $sim = (m*l - \sum \mathrm{Di})/m*l$ where $0 \leq sim \leq 1$. The pseudocode of the non-exact match search is illustrated in Fig. 4.

```
Start
Get input data and relevant input parameter sim;
While there are still unprocessed input sequence in |S|
      If l_Si >= 1500
            Split S_i;
            For all the split sequences
                  For all SS
                        Apply k-Means and assign datapoints;
                        Filter mfls;
                        MSA string alignment;
                        Do similarity calculation;
                        Bridge the split sequences;
                  End For
                  If candidate fulfills criteria
                        Add VNTR to list;
                  End If;
            End For
      Else
            For all SS
                  Apply k-Means and assign datapoints;
                  Filter mfls;
                  MSA string alignment;
                  Do similarity calculation;
            End For
            If candidate fulfills criteria
                  Add VNTR to list;
            End If
      End If
End While
```

Fig. 4. Non-Exact Match Search Pseudocode

2.2.3 Visualization Tool

To our knowledge, INVERTER is the first tandem repeat discovery tool with an integrated visualization tool. The visualization tool is aimed to provide data visualization of the resulting VNTRs, which allow users to correlate annotations, observe visual patterns, and view useful statistics of the data, which are not available in other tools.

The availability of this novel tool is aimed to help biologists visualize VNTRs in raw genomic data as well as overall view of the entire data in form of useful statistics, hence facilitating new ways of understanding both the structure and function of DNA and protein. An example of this is frequency analysis of DNA sequences, in which some of tandem repeat patterns have been associated with known genetic factors e.g. a 3bp repeats has been found to be characteristic of exonic regions[23, 24] and a 10-11bp repeats has been found to be characteristic of DNA prone to supercoiling[25].

Users can directly use the visualization tool on the latest search result or open a csv file containing the result of a previous search. The visualization tool partitions the VNTRs based on their respective origin sequences in tab form. Users can display all the sequences or choose selectively which ones they are interested for comparison purposes. The visualization tool includes four features, i.e. table view, zoom view, statistics and intra-inter repeat.

The zoom view provides an overall view of the positions VNTRs in the sequence. The x axis shows the relative position while the y axis shows the id of the VNTRs, respectively. Users can zoom in and zoom out of a particular area in the sequence to get a more detailed visualization of the VNTRs in the sequence. VNTRs marked with repeat reference are color coded in accordance to the table view while VNTRs that are not marked with repeat reference are given grayscale colors, ranging from light gray to black. For the latter, the higher the number of count of a particular VNTR, the lighter the color associated to it. Tooltip showing the VNTR sequence and the starting position of the VNTR will appear if users mouseover a particular VNTR. Furthermore, users can choose to display only selected VNTRs marked with repeat reference, with or without the non-repeat reference VNTRs for isolation purposes. In addition, users can also save the visualization as a PNG image file or print it directly from the user interface. Fig. 5 shows an example of the zoom view of the *Acinetobacter baumannii* AYE complete genome.

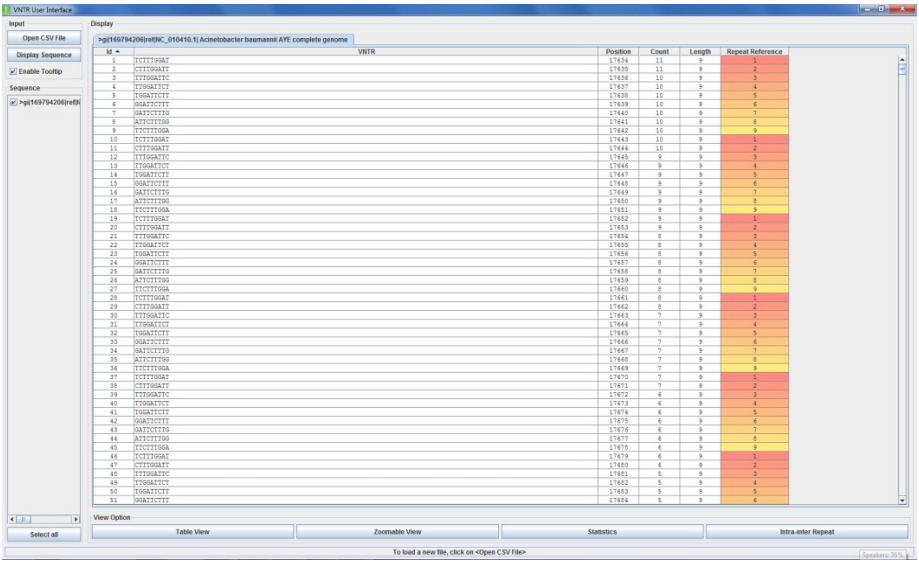


Fig. 5. Table view of the *Acinetobacter baumannii* AYE complete genome

Table view displays the search result in spreadsheet form. It shows the identification number (id), the VNTR, the initial position of the VNTR in the sequence, the length of the VNTR, the number of count and whether the VNTR is a repeat reference. If a VNTR is marked with repeat reference, it indicates that the VNTR appears in either another position in the origin sequence or in another sequence in the one of selected sequences. The repeat reference cell is color coded, up to 1024 unique colors and the value inside the cell indicates the reference number that particular VNTR. The spreadsheet can be sorted according a particular column (the default is id). In addition, the columns can be expanded and interchanged to allow greater flexibility for users. Example of the table view of the *Acinetobacter baumannii* AYE complete genome is illustrated in Fig. 6.

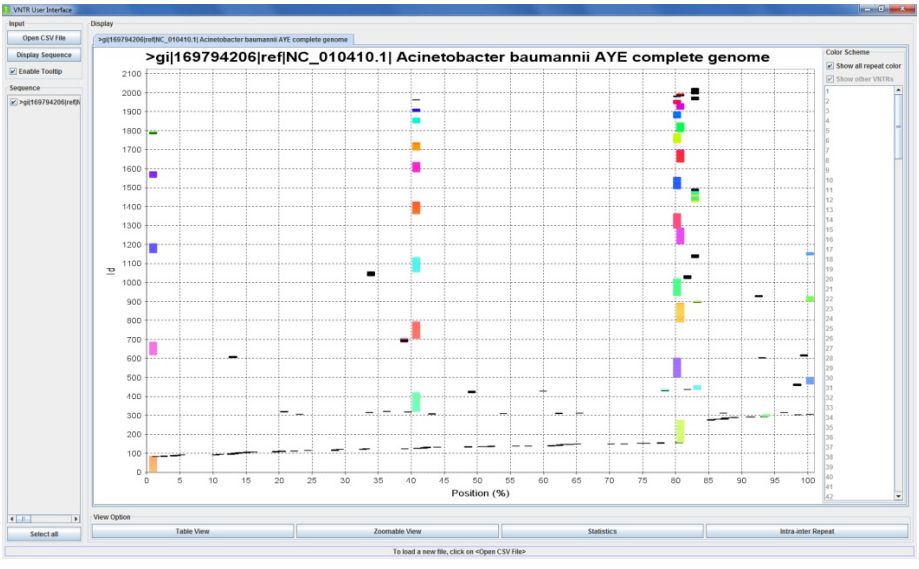


Fig. 6. Zoom view of the *Acinetobacter baumannii* AYE complete genome

The main idea of the statistics feature is for users to be able to make a quick decision on whether or not the VNTRs in the sequence are worth pursuing further. This feature will be very useful especially when users are dealing with numerous sequences, as it summarizes how VNTR is distributed over the sequences. The statistics include the maximum value, minimum value, average, median, mode and histogram of the length and count of the VNTRs in the search result, respectively. In addition, repeat reference statistics are also included, namely the total number of repeat reference in the sequence and their respective references, total number of *unique* repeat reference in the sequence and their respective references as well as the total number of multiple occurrence of repeat reference in the sequence. Example of the statistics of the *Acinetobacter baumannii* AYE complete genome is shown in Fig. 7.

Intra-inter repeat shows the intra-relationship and inter-relationship of VNTRs marked with repeat reference. Intra-relationship means that the VNTR appears in

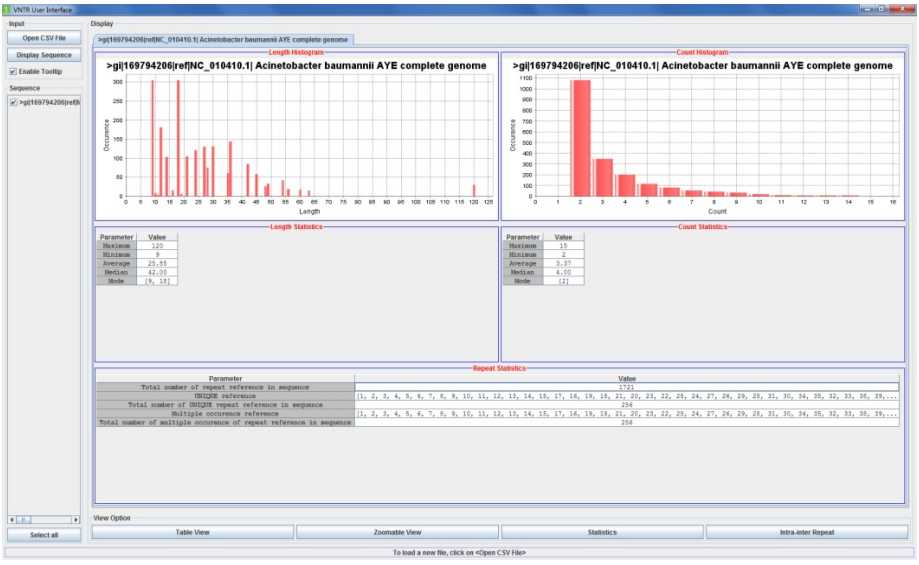


Fig. 7. Statistics of the *Acinetobacter baumannii* AYE complete genome

another position in the origin sequence while inter-relationship means that the VNTR appears in another sequence. The cell is color coded in accordance to the table view and the value inside the cell of row *r* and column *c* indicates the number of occurrence of that a VNTR that is marked with a repeat reference *r* in sequence *c*. Fig. 8 illustrates an example of relationship of the *Acinetobacter baumannii* AYE and *Acinetobacter baumannii* AB307-0294 complete genomes.

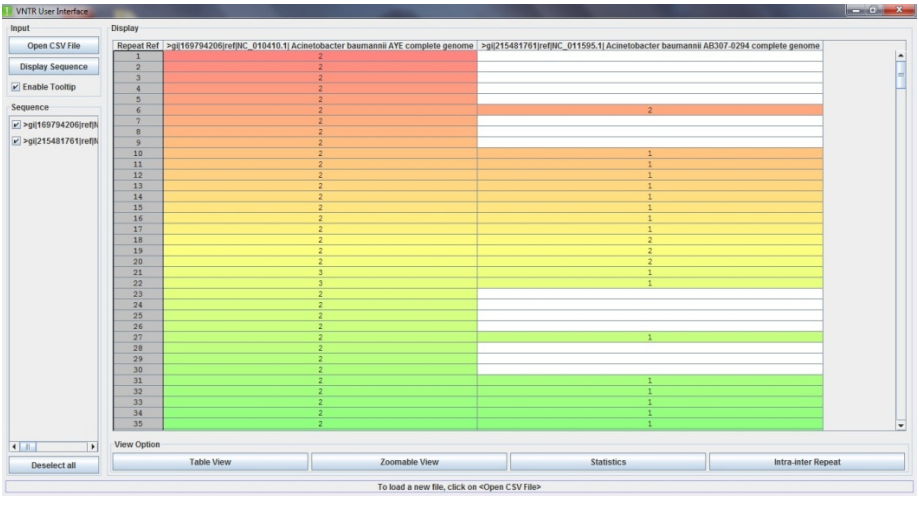


Fig. 8. Intra-inter repeat relationship of the *Acinetobacter baumannii* AYE and *Acinetobacter baumannii* AB307-0294 complete genomes

3 Result

Acinetobacter baumannii is a major nosocomial pathogen with many clones resistant to most available antibiotics. It affects primarily immunocompromised patients, often in intensive care and burns units[26]. The reported antibiotic resistance includes the *carbapenems*, which have been the antibiotics of choice against this organism. Infections can therefore be difficult to treat, and are associated with increased morbidity and mortality[27, 28].

We have examined 6 completed genome projects of *Acinetobacter baumannii* obtained from the NCBI website (http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj&cmd=Retrieve&list_uids=21111,30993,17827,17477,28921,13001), i.e. AB0057, AB307-0294, ACICU, ATCC17978, AYE and SDF. The parameter values of l_{min} and l_{max} chosen for the experiment are 9 and 200, respectively. Our experiment is benchmarked on an Intel Core i5-540M 2.40 GHz CPU, 4 GB RAM running Windows 7 64 bit. The runtime is measured in seconds. The result of the experiment is shown in Table 2.

Table 2. Exact match result on 6 complete *Acinetobacter baumanii* genomes

***A. baumanii* genome**	**RefSeq**	**Length (Mbps)**	**Subset option off**		**Subset option on**	
			#VNTRs found	**Runtime (s)**	**#VNTRs found**	**Runtime (s)**
AB0057	NC_011586	4.050513	570	42	1164	43
AB307-0294	NC_011595	3.760981	821	38	2720	39
ACICU	NC_010611	3.904116	508	39	1201	40
ATCC17978	NC_009085	3.976747	394	40	838	41
AYE	NC_010410	3.936291	787	39	2023	39
SDF	NC_010400	3.421954	1524	35	4600	35

We compare INVERTER result with T-REKS[18]. The similarity parameter P_{sim} of T-REKS is set to 1.0 to ensure that it search only for exact match VNTRs and we set the type to DNA sequence.

In general, all of the VNTRs reported by T-REKS are found by INVERTER as well. Furthermore, INVERTER found several more VNTRs that are not reported in T-REKS. However, there are a total of 4 VNTRs that are reported in the T-REKS but are not included in the INVERTER result with subset option off. By modifying the subset option parameter to on, these 4 VNTRs are included in the INVERTER result. The reason is because INVERTER found a more significant VNTR compared to the T-REKS result, and therefore it relegates that particular VNTR as a subset VNTR. In all 4 cases, the more significant VNTR has a higher count value compare to the subset counterpart. The difference of INVERTER result with T-REKS is highlighted in Table 3.

Table 4 shows non-exact match result of INVERTER on the 6 completed genome projects of *Acinetobacter baumannii*. The similarity threshold value *sim* chosen for the experiment is 0.85. The runtime is measured in seconds. As a comparison, T-REKS needs 17 min to analyze a medium size genome of *Drosophila melanogaster* using a Pentium 4 3.0 GHz and 2 GB of RAM[18]. In some genomes, INVERTER found less non-exact match compared to the exact match. One possible reason for this

Table 3. Difference of INVERTER result with T-REKS

A.baumanii genome	T-REKS	INVERTER (Subset option off)
AB307-0294	ATTCTTTGG	CTTTGGATT
ATCC17978	GGATTCTTT	TTCTTTGGA
AYE	GGATTCTTT	TTCTTTGGA
SDF	GACAGCGATTCGGATTCTGACTCA	TGACTCAGACAGCGATTCGGATTC

Table 4. Non-exact match result on 6 complete *Acinetobacter baumanii* genomes

A. baumanii genome	RefSeq	Length (Mbps)	#VNTRs (non-overlapping)	Runtime (s)
AB0057	NC_011586	4.050513	450	802
AB307-0294	NC_011595	3.760981	260	743
ACICU	NC_010611	3.904116	355	771
ATCC17978	NC_009085	3.976747	318	786
AYE	NC_010410	3.936291	416	779
SDF	NC_010400	3.421954	372	677

anomaly is that the non-exact match result excludes overlapping VNTRs of different tandem repeats that can be detected in the same region. Another possibility is that DNA sequences of length 1500 may contain more than the determined maximum number of clusters *k* different lengths of VNTRs (20 in this case) and that the splitting step may lead to the failure to detect some VNTRs. Our future work includes the optimization of parameters for the non-exact matches as well as to compare the time complexity with other tools.

4 Conclusions

In this paper, we present INVERTER, a *de novo* exact match tandem repeat finder which main advantage is that there is no need to specify either a pattern or a particular pattern size, integrated with a data visualization tool and a built-in user-friendly Graphical User Interface. INVERTER is designed to be portable; hence it is written in Java. It is therefore usable without problems on any CPUs with Windows, Linux, Solaris and MacOS operating systems. A standalone version of the program can be downloaded from http://bmserver.sce.ntu.edu.sg/INVERTER. Three features are currently available to users, i.e. identification of exact match VNTRs and non-exact match VNTRs in DNA sequences as well as visualization of the resulting tandem repeat data.

INVERTER is aimed to identify both exact match and non-exact match VNTRs and provide data visualization which would allow users to correlate annotations, observe visual patterns, and view useful statistics of the data. More specifically, INVERTER is more optimized towards minisatellite VNTRs. The visualization tool is aimed to provide data visualization of the resulting VNTRs, which would allow users to correlate annotations, observe visual patterns, and view useful statistics of the data. The visualization tool includes four features, i.e. table view, zoom view, statistics and intra-inter repeat.

Acknowledgments. The work described in this paper was supported by the National Medical Research Council - Exploratory / Developmental Grant NMRC EDG-08nov025.

References

1. Collins, F.S., Morgan, M., Patrinos, A.: The Human Genome Project: Lessons from large-scale biology. Science 300, 286–290 (2003)
2. Kannan, S.K., Myers, E.W.: An algorithm for locating nonoverlapping regions of maximum alignment score. SIAM Journal on Computing 25, 648–662 (1996)
3. Verkerk, A.J.M.H., Pieretti, M., Sutcliffe, J.S., Fu, Y.H., Kuhl, D.P.A., Pizzuti, A., Reiner, O., Richards, S., Victoria, M.F., Zhang, F., Eussen, B.E., Van Ommen, G.J.B., Blonden, L.A.J., Riggins, G.J., Chastain, J.L., Kunst, C.B., Galjaard, H., Caskey, C.T., Nelson, D.L., Oostra, B.A., Warren, S.T.: Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell 65, 905–914 (1991)
4. Fu, Y.H., Pizzuti, A., Fenwick Jr, R.G., King, J., Rajnarayan, S., Dunne, P.W., Dubel, J., Nasser, G.A., Ashizawa, T., De Jong, P., Wieringa, B., Korneluk, R., Perryman, M.B., Epstein, H.F., Caskey, C.T.: An unstable triplet repeat in a gene related to myotonic muscular dystrophy. Science 255, 1256–1258 (1992)
5. MacDonald, M.E., Ambrose, C.M., Duyao, M.P., Myers, R.H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S.A., James, M., Groot, N., MacFarlane, H., Jenkins, B., Anderson, M.A., Wexler, N.S., Gusella, J.F., Bates, G.P., Baxendale, S., Hummerich, H., Kirby, S.: A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72, 971–983 (1993)
6. La Spada, A.R., Wilson, E.M., Lubahn, D.B., Harding, A.E., Fischbeck, K.H.: Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature 352, 77–79 (1991)
7. Campuzano, V., Montermini, L., Molto, M.D., Pianese, L., Cossee, M., Cavalcanti, F., Monros, E., Rodius, F., Duclos, F., Monticelli, A., Zara, F., Canizares, J., Koutnikova, H., Bidichandani, S.I., Gellera, C., Brice, A., Trouillas, P., De Michele, G., Filla, A., De Frutos, R., Palau, F., Patel, P.I., Di Donato, S., Mandel, J.L., Cocozza, S., Koenig, M., Pandolfo, M.: Friedreich's ataxia: Autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science 271, 1423–1427 (1996)
8. Mirkin, S.M.: DNA structures, repeat expansions and human hereditary disorders. Current Opinion in Structural Biology 16, 351–358 (2006)
9. Usdin, K.: The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. Genome Research 18, 1011–1019 (2008)
10. Jeffreys, A.J.: DNA typing: Approaches and applications. Journal of the Forensic Science Society 33, 204–211 (1993)
11. Bruford, M.W., Wayne, R.K.: Microsatellites and their application to population genetic studies. Current Opinion in Genetics and Development 3, 939–943 (1993)
12. Spong, G., Hellborg, L.: A near-extinction event in lynx: Do microsatellite data tell the tale? Conservation Ecology 6 (2002)
13. Benson, G.: Sequence alignment with tandem duplication. In: Conference Sequence alignment with tandem duplication, pp. 27–36 (1997)
14. Benson, G.: Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Research 27, 573–580 (1999)

15. Kolpakov, R., Bana, G., Kucherov, G.: mreps: Efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Research 31, 3672–3678 (2003)
16. Wexler, Y., Yakhini, Z., Kashi, Y., Geiger, D.: Finding approximate tandem repeats in genomic sequences. Journal of Computational Biology 12, 928–942 (2005)
17. Parisi, V., De Fonzo, V., Aluffi-Pentini, F.: STRING: Finding tandem repeats in DNA sequences. Bioinformatics 19, 1733–1738 (2003)
18. Jorda, J., Kajava, A.V.: T-REKS: Identification of Tandem REpeats in sequences with a K-meanS based algorithm. Bioinformatics 25, 2632–2638 (2009)
19. Turnpenny, P., Ellard, S.: Emery's Elements of Medical Genetics. Elsevier, London (2005)
20. Jeffreys, A.J., Wilson, V., Thein, S.L.: Hypervariable 'minisatellite' regions in human DNA. Nature 314, 67–73 (1985)
21. Merkel, A., Gemmell, N.: Detecting short tandem repeats from genome data: Opening the software black box. Briefings in Bioinformatics 9, 355–366 (2008)
22. MacQueen, J.B.: Some Methods for Classification and Analysis of MultiVariate Observations. In: Cam, L.M.L., Neyman, J. (eds.) Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)
23. Sanchez, J., Lopez-Villasenor, I.: A simple model to explain three-base periodicity in coding DNA. FEBS Lett. 580, 6413–6422 (2006)
24. Lopez-Villasenor, I., Jose, M., Sanchez, J.: Three-base periodicity patterns and self-similarity in whole bacterial chromosomes. Biochem. Biophys. Res. Commun. 325, 467–478 (2004)
25. Schieg, P., Herzel, H.: Periodicities of 10-11 bp as indicators of the supercoiled state of genomic DNA. J. Mol. Biol. 343, 891–901 (2004)
26. Turton, J.F., Matos, J., Kaufmann, M.E., Pitt, T.L.: Variable number tandem repeat loci providing discrimination within widespread genotypes of acinetobacter baumannii. European Journal of Clinical Microbiology and Infectious Diseases 28, 499–507 (2009)
27. Wareham, D.W., Bean, D.C., Khanna, P., Hennessy, E.M., Krahe, D., Ely, A., Millar, M.: Bloodstream infection due to Acinetobacter spp: Epidemiology, risk factors and impact of multi-drug resistance. European Journal of Clinical Microbiology and Infectious Diseases 27, 607–612 (2008)
28. Dijkshoorn, L., Nemec, A., Seifert, H.: An increasing threat in hospitals: Multidrug-resistant Acinetobacter baumannii. Nature Reviews Microbiology 5, 939–951 (2007)

Design of an *Enterobacteriaceae* Pan-Genome Microarray Chip

Oksana Lukjancenko and David W. Ussery

Center for Biological Sequence Analysis, Department of Systems Biology,
The Technical University of Denmark, 2800 Kongens Lyngby, Denmark

Abstract. Microarrays are a common method for evaluating genomic content of bacterial species and comparing unsequenced bacterial genomes. This technology allows for quick scans of characteristic genes and chromosomal regions, and to search for indications of horizontal transfer. A high-density microarray chip has been designed, using 116 *Enterobacteriaceae* genome sequences, taking into account the enteric pan-genome. Probes for the microarray were checked *in silico* and performance of the chip, based on experimental strains from four different genera, demonstrate a relatively high ability to distinguish those strains on genus, species, and pathotype/serovar levels. Additionally, the microarray performed well when investigating which genes were found in a given strain of interest. The *Enterobacteriaceae* pan-genome microarray, based on 116 genomes, provides a valuable tool for determination of the genetic makeup of unknown strains within this bacterial family and can introduce insights into phylogenetic relationships.

Keywords: *Enterobacteriaceae*, Pan-genome, DNA microarray analysis, gene, *Escherichia coli.*

1 Introduction

The risk of dying from disease caused by a bacterial infection is greater than that associated with any other type of disease, including cancer or heart attacks [1, 2]. Epidemic infectious diseases are the most serious causes of mortality and morbidity worldwide, more than all other diseases combined. Infections contribute to significant economic loss in most parts of the world, including first world countries that have high income and developed surveillance and control systems [3, 4]. Every year thousands of people are infected by bacterial pathogens, most of which are transmitted through food [5]. The outcome from food-borne human infections can range from mild self-limiting diarrhea to severe illness that requires hospitalization. In rare cases, food-borne illnesses are even fatal [5, 6]. Enteric bacteria, particularly *Salmonella enterica* subsp. *enterica*, are among the leading food-borne pathogens [6, 7]. In light of this, the detailed and rapid investigation of enteric pathogens is essential in modern epidemiology and clinical diagnostics.

Enterobacteriaceae are pervasive. They are widespread in the environment, existing in water, soil, food, and plants, as well as in the normal intestinal flora of many animals and humans [8-12]. Pathogens within this group have developed a diversity

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 165–179, 2010.

of strategies to overcome protective host barriers in order to invade the host, resist innate immune response, multiply in specific and normally sterile body sites, and damage cells in order to establish and maintain a successful infection [13, 14]. Genera within *Enterobacteriaceae* family are of interest, as well, because of problems from food spoilage and for that reason are of considerable economic importance [15].

Bacterial genomes vary in size, even among the strains of the same species. Bacterial species can be characterized by its pan-genome. As defined by Tettelin *et al.*, the microbial pan-genome is a complete collection of various genes located within populations at a particular taxonomic level, commonly within a species. The pan-genome concept can of course be expanded to higher levels, such as genus or even a bacterial family. The pan-genome includes a core-genome, which is a minor fraction of the entire gene pool that is shared between all the given strains. Furthermore, there is a much larger, dispensable portion of bacterial genes, that are missing in one or more strains. Also there are some genes that appear to be unique to each strain [16, 17]. Strain-specific genes can, even among a particular species, make up a notably large portion of the pan-genome [18].

Many methods have been developed for characterizing genetic variation. Use of DNA microarrays is becoming a standard procedure for evaluating genotyping – that is, looking at the genetic content of a bacterial species. The price for microarrays used for genotyping was historically expensive, but now is becoming competitive with the cost of other commonly used typing methods, such as previously widely used multilocus sequence typing (MLST). Moreover, it is becoming increasingly popular, quick, and cost-effective to define the presence and absence of each of the assigned genes in the pan-genome of a species. Thus, microarrays, imprinted with all the genes from species' pan-genome can be used to compare and characterize the genomic content of unknown bacterial isolates and to achieve accurate typing information, that can be useful in epidemiological investigations and clinical diagnostics [1, 19]. For instance, array comparative genomic hybridization (aCGH) is frequently used in human cancer studies to genotype cell lines by determination of gene loss and copy number variations [20] or to detect single nucleotide polymorphisms at target loci [21]. Additionally, microarrays have been widely used in human screenings for the determination and genotyping of bacterial species. Microarrays have changed considerably since they were first introduced. Early microarrays for the *E. coli* genome consisted of long fragments of chromosomal DNA (~1000 to 2000 base-pairs), attached to a microscope slide. Later, Affymetrix made an array covering the entire *E. coli* K-12 genome using a set of 10 to 15 probes (synthetic 25mers) for each gene [22], followed shortly by an array which contained 4 *E. coli* genomes [23, 24]. Custom-designed NimbleGen chips have been made including 7 and then 32 *E. coli* genomes [25, 26].

This study describes the design and use of a high-density oligonucleotide microarray covering the pan-genome of 116 genomes within the *Enterobacteriaceae* family. Probes are designed to distinguish among organisms at the level of genera, species, and even single strains. Moreover, probes for determination of particular gene families, comprising *Enterobacteriaceae* pan-genome, are defined. The performance of this microarray is evaluated both *in silico* and experimentally. Its utility is illustrated for the hybridization of genomic DNA in order to compare uncharacterized isolates which have not been sequenced with the 116 known, sequenced strains. A microarray chip approximating the complete pan-genome of *Enterobacteriaceae*

provides optimal sensitivity to characterize isolates. Gene family microarray analysis is useful for medical and environmental diagnoses and will provide an alternative to costly genome libraries, as well as to the sequencing of environmental samples.

2 Materials and Methods

2.1 Bacterial Strains

In this study, one hundred and twelve complete *Enterobacteriaceae* genome sequences and four in progress, which were publically available in GenBank database at the time of analysis (February, 2010), were used for custom microarray design. An overview of the used strains is shown in Table 1 and the complete collection of the strains is described in supplementary Table S1[1].

Table 1. *Enterobacteriaceae* genera used in the design of the microarray chip

Genus	Number of strains	Genus	Number of strains
Buchnera	6	*Photorhabdus*	2
Citrobacter	3	*Salmonella*	18
Cronobacter	2	*Serratia*	1
Dickeya	3	*Shigella*	8
Edwardsiella	2	*Sodalis*	1
Enterobacter	2	*Wigglesworthia*	1
Escherichia	35	*Xenorhabdus*	1
Klebsiella	4	*Yersinia*	14
Pectobacterium	3	*Erwinia*	4
Proteus	3	*Candidatus**	3

* *Candidatus* is not a genus; however some strains were included as they were classified as *Enterobacteriaceae* at the time of study.

Twelve bacterial strains included in experimental evaluation of the chip are listed in Table 3 (Results section).

2.2 Pan-Genomics

The pan-genome was estimated, as described by Snipen *et al* [27]. Briefly, all protein sequences were compared by BLASTP [28]. Two proteins were attributed to a single gene family if they satisfied the 50/50 rule, meaning that when they could produce a pairwise BLASTP alignment covering at least 50% amino of the length of the longest protein with at least 50% of amino acid identity. Each genome was compared successively: for each n additional genome, that genome was compared to any combinations of n-1 genomes and the number of identical 'core genes' and 'genome specific genes' (specific for genome n) were counted for each n. All cumulative BLASTP hits found in the whole set of genomes were plotted as a running total and were considered as pan-genome, which increases as more genomes are added. The number of gene families with at least one representative in every genome was plotted for the core-genome.

[1] Available at http://www.cbs.dtu.dk/~dave/Supplementary_TableS1.pdf

2.3 The Custom-Microarray Design

The custom probe set for the microarrays was designed around 78 different groups of genomes (the list of groups is presented in the Results section, Table 2) including a collection of generic probes for the entire enteric core (97 genes), as well as for the probes that differentiate each genus within *Enterobacteriaceae*. The custom probe set was followed by more specialized probe sets for species-specific classification within *Klebsiella*, *Salmonella*, *Escherichia*, *Shigella*, and *Yersinia* genera and further probe groups were specific for strain and pathotype for *Escherichia coli* genus. Additionally, sets of probes for all the gene families, comprising pan-genome, were included. The custom microarrays, manufactured by NimbleGen, were based on the NimbleGen 12-plex platform.

2.4 Constructing Target Gene Sets

The genome sequences in this study (Table S1) were searched for genes using the Prodigal gene-finding approach [29] in order to standardize gene finding. All protein-coding sequences were aligned all-against-all using BLASTP [28], and similarity was decided according to 50/50 rule. Proteins that satisfy this rule were assigned to one protein family. 'Group specific gene families' (as described above) were found using batch Perl script, which outputs a list of gene families that are either common to or complementary to the genomes included in pan- and core-genome plots (depending on whether unique or core genes are extracted). Representative sequences from each gene network were selected by choosing the organism from which the genes should be extracted. Unique genes were considered to be those that appeared to be conserved only among the strains belonging to a particular group.

2.5 Probe Selection for Target Genes

Probes for target genes were selected using the OligoWiz program, previously described by Wernersson *et al.* [30][31]. At each position along all the input sequence, the suitability of placing a probe was evaluated according to several criteria: melting temperature (ΔT_m), cross-hybridization, folding (self-annealing), position (within the transcript), and 'low-complexity' (absence of subsequences that occur very commonly in the genome/transcriptome). The weighting scores for these criteria are as follow: cross-hybridization, 39%; ΔT_m, 26%; folding, 13%; position, 13%; and low-complexity, 9%. No probes were accepted unless an overall score of at least 0.3 was obtained, and all probes were required to have a length in the range of 42 bp to 50 bp. OligoWiz was originally designed for single genome use, and thus, the program was modified in order to make the mechanisms screening for cross-hybridization less strict as described by Vejborg *et al.* [32]. A new modified scheme included a log-transformation in the underlying calculations. The net effect is insignificant near the upper boundary of the score, but next to the lower boundary it increases the discriminatory power of the tool.

$$\text{BLAST max score} = 1 - \sum_{n}^{i=1} \log(1 + \sum_{m}^{m=1} \frac{hm,i}{100} \quad (1)$$

2.6 Probe Evaluation *in silico*

Probes were aligned against a database consisting of all possible gene sequences in the total data set using BLASTN. The affinity of each probe for every gene was determined and expressed as the number of identical base pairs and by the E-value. Sequences for which the E-value was lower than 0 were extracted using a batch Perl script. Probes that matched strains not expected to belong to particular group were excluded from the further analysis. If more than ten probes per gene remained available after filtering, only not-overlapping ones were used for subsequent analysis. This resulted in the reduction of candidate probes from 106,657 to 53,644. Consequently, the number of probes targeting each gene ranged from 3 to 14 with a median coverage of about 7 probes per gene.

2.7 DNA Preparation and Hybridization

All the experimental isolates were kindly provided by the laboratory of Frank Møller Aarestrup (DTU Food, The Technical University of Denmark). All test strains were grown overnight on blood agar and genomic DNA was isolated as described in the protocol for the Easy-DNA kit from Invitrogen [33]. The method used is briefly described here: the lysis of the cells was performed by the addition of solution A and subsequent incubation at 65°C. Proteins and lipids were precipitated and extracted by the addition of solution B and chloroform. The solution was then centrifuged to separate the solution into two phases. The DNA was in the upper, clear aqueous phase, the proteins and lipids were in the solid interface, and the chloroform formed the lower phase. The DNA was then removed, precipitated with ethanol, and re-suspended in TE buffer.

The genomic DNA was labeled with cy3 dye and hybridized to NimbleGen custom arrays according to Arrays User's Guide for CGH analysis as provided by the manufacturer of the arrays (Roche NimbleGen, Madison, Wisconsin, USA).

2.8 Analysis Methods

In the initial step, the raw data from multiple microarrays was extracted using NimbleScan software, developed by Roche NimbleGen, and combined as a single input. Data analysis was performed in R (a statistical software program), using the 'oligo' package for analyzing oligonucleotide arrays at the probe level. The package was obtained from Bioconductor [34]. The probes were mapped to each gene group, including position, according to the design. Chip analysis workflow then continued as follows:

1. Performance of probe-level normalization using robust multi-array average (RMA) algorithm. RMA method had a three-step procedure consisting of background correction, normalization, and summarization to obtain gene-level relative intensity measures from probe-level intensities [35].
2. Estimation of gene 'on/off' status based on the summarized gene relative intensities and the median of these intensities for each of the 78 groups.

Supporting microarray chip design information is publicly available[2].

[2] http://www.cbs.dtu.dk/~dave/Microarray_Chip_Design_Lukjancenko_2010.ndf

3 Results

3.1 Pan-Genome and Core-Genome Estimation

For each of the considered bacterial strains listed in Table S1 (Supplementary data), the genome sequence was downloaded from NCBI/GenBank. Genes were predicted by Prodigal [29], and translated into proteins. This resulted in a dataset of 887,184 entries with considerable redundancy due to the presence of the same gene in multiple genomes. To reduce the homology, proteins were grouped into the gene families. Proteins were considered conserved (belonging to the same gene group) if they showed at least 50% amino acid identity in a BLASTP alignment covering at least 50% of the length of the longest protein. The combined pan-genome of 116 genomes within *Enterobacteriaceae* was estimated and appeared to contain 44,838 gene families. The core-genome, that is, the number of conserved genes present in all 116 genomes, was estimated to be comprised of 97 conserved gene families.

3.2 Probe and Microarray Design

In the presented *Enterobacteriaceae* pan-genome microarray design strategy, the probe set was designed around 78 different groups of genomes. The microarray was made up of a collection of probes for each genus within *Enterobacteriaceae,* being species-specific for *Klebsiella*, *Salmonella*, *Escherichia*, *Shigella*, and *Yersinia* genera; strain and pathotype specific for *Escherichia coli* genus; core genes; and all protein families, comprising pan-genome. Using the data from the pan- and core-genome estimation step, the number of 'group-specific' genes and probes was determined and are shown in Table 2. Genes were considered to be 'group-unique' if they were found only within genomes, belonging to a particular group, and were absent in all of the rest genomes among a set of 116 genomes.

The final result was a set of 52,356 *Enterobacteriaceae* target sequences, representing genes of both specific groups and pan-genome gene families. The oligos were then selected using OligoWiz [31] based on several criteria, including their specificity, self-annealing, presence of low-complexity sequences, and their lengths adjusted so as to standardize the hybridization strength. Probes were filtered in order to avoid complimentarity with unwanted targets. In the end a set of 130,540 non-overlapping probes with an average length of 49 bp were obtained. The average number of probes per target gene was about 7, although the actual number for any given target depended on the length of the sequence, since shorter sequences have space for fewer non-overlapping probes. For set of probes that represent gene families an average of 3 probes per family was used.

3.3 Validation of the Custom Arrays

The chip design was evaluated by analyzing and comparing hybridization data from twelve control strains, shown in Table 3. Microarray data can have noise, coming from multiple variations which can occur during the array manufacturing process, the preparation of the biological sample for the hybridization, the hybridization of the samples to the array itself, and the quantification of the spot intensities [35]. To remove such variation, which obviously will affect the measured gene intensity levels,

Table 2. Number of 'group specific' gene families and probes before and after *in silico* validation

Probe group	Number of genes before validation	Number of probes before validation	Number of genes after validation	Number of probes after validation
Buchnera genus	14	200	14	123
Candidatus strains	41	584	41	373
Citrobacter genus	20	171	15	95
Cronobacter genus	271	3224	270	2002
Dickeya genus	155	2129	155	1398
Edwardsiella genus	318	3803	317	2447
Enterobacter genus	40	511	40	318
Erwinia genus	217	2919	217	1840
Escherichia genus	1	15	1	10
Escherichia coli 042	106	1047	79	450
Escherichia coli 536	142	1207	95	436
Escherichia coli 55989	72	646	45	272
Escherichia coli APEC	116	1287	14	83
Escherichia coli APEC O1	116	1287	14	83
Escherichia coli Avirulent	69	508	39	241
Escherichia coli B phylogroup	14	175	14	100
Escherichia coli CFT073	292	2251	115	393
Escherichia coli E24377A	249	1700	90	511
Escherichia coli EAEC	72	646	45	272
Escherichia coli ED1a	159	1545	146	823
Escherichia coli EHEC	21	173	13	27
Escherichia coli EPEC	142	1685	126	893
Escherichia coli ETEC	249	1700	90	511
Escherichia coli ExPEC	52	392	17	131
Escherichia coli HS	90	642	44	313
Escherichia coli IAI1	67	499	39	238
Escherichia coli IAI39	77	609	48	262
Escherichia coli K-12	11	159	11	113
Escherichia coli O103:H2	65	693	50	377
Escherichia coli O111:H-	148	1536	54	250
Escherichia coli O127:H6	142	1685	126	893
Escherichia coli O157:H7	68	709	52	379
Escherichia coli O26:H11	74	690	48	280
Escherichia coli S88	52	392	17	131
Escherichia coli SE11	178	1692	70	360
Escherichia coli SE15	58	609	49	328
Escherichia coli SMS-3-5	145	1064	106	501
Escherichia coli UMN026	113	1026	85	505
Escherichia coli UPEC	121	983	49	179
Escherichia coli UTI89	85	754	35	192
Escherichia/Shigella genera	15	184	15	113
Klebsiella genus	242	3296	242	2090
Klebsiella pneumoniae 342	11	93	8	50
Klebsiella pneumoniae MGH 78578	21	237	14	49
Klebsiella pneumoniae NTUH-K2044	339	2636	233	863

Table 2. (*Continued*)

Klebsiella variicola At-22	115	1282	110	758
Pectobacterium genus	166	2287	166	1422
Proteus genus	355	4782	355	3006
Photorhadbus genus	318	4392	318	2728
Salmonella genus	69	933	69	575
Salmonella enterica Agona	136	1151	111	568
Salmonella arizonae	477	3828	474	2245
Salmonella enterica Choleraesuis	92	804	44	87
Salmonella enterica Dublin	101	526	22	77
Salmonella enterica Enteritidis	20	217	9	55
Salmonella enterica Gallinarum	10	88	5	14
Salmonella enterica Heidelberg	91	608	51	249
Salmonella enterica Newport	189	1967	111	351
Salmonella enterica Paratyphi A	10	80	7	10
Salmonella enterica Paratyphi B	436	1982	175	547
Salmonella enterica Paratyphi C	54	266	20	47
Salmonella enterica Schwarzengrund	139	1025	122	498
Salmonella enterica Typhi	69	759	63	326
Salmonella enterica Typhimurium	9	113	3	30
Serratia genus	780	10393	780	6777
Shigella boydii	19	164	16	52
Shigella dysenteriae	113	1216	98	348
Shigella flexneri	17	218	17	123
Shigella genus	28	401	25	178
Shigella sonnei	48	531	32	152
Sodalis genus	420	5697	420	3464
Wigglesworthia genus	212	3029	212	1789
Xenorhabdus genus	82	855	82	527
Yersinia genus	97	4189	97	809
Yersinia enterocolitica	336	1312	336	2655
Yersinia pestis	7	26	5	5
Yersinia pseudotuberculosis	23	165	13	24
Core genes	97	1378	97	850
Gene families	42151	180219	27536	76896

normalization was performed. A set of twelve arrays (one 12plex array) used in the experiment was printed at the same time, so background noise effects were expected to be reasonably similar across all arrays. Only one out of the twelve the results were not as anticipated. The single exception being for the *Salmonella enterica* serovar Choleraesuis isolate, which shows variation. Thus it was decided to exclude hybridization data of this isolate from further analysis. RMA normalization, performed for microarray data of the remaining eleven samples, made the distribution of probe intensities for each array in a set of arrays nearly the same.

In the workflow of further microarray data analysis, the evaluation of which genus, species, pathotype/serovar or strain, the experimental isolate is most likely to be similar to. For each of the seventy-eight gene sets, the median of signal intensities were calculated. The analysis was performed based on both distribution of probe log intensities and the signal median. The examples are shown in Figures 1-3, which visualize

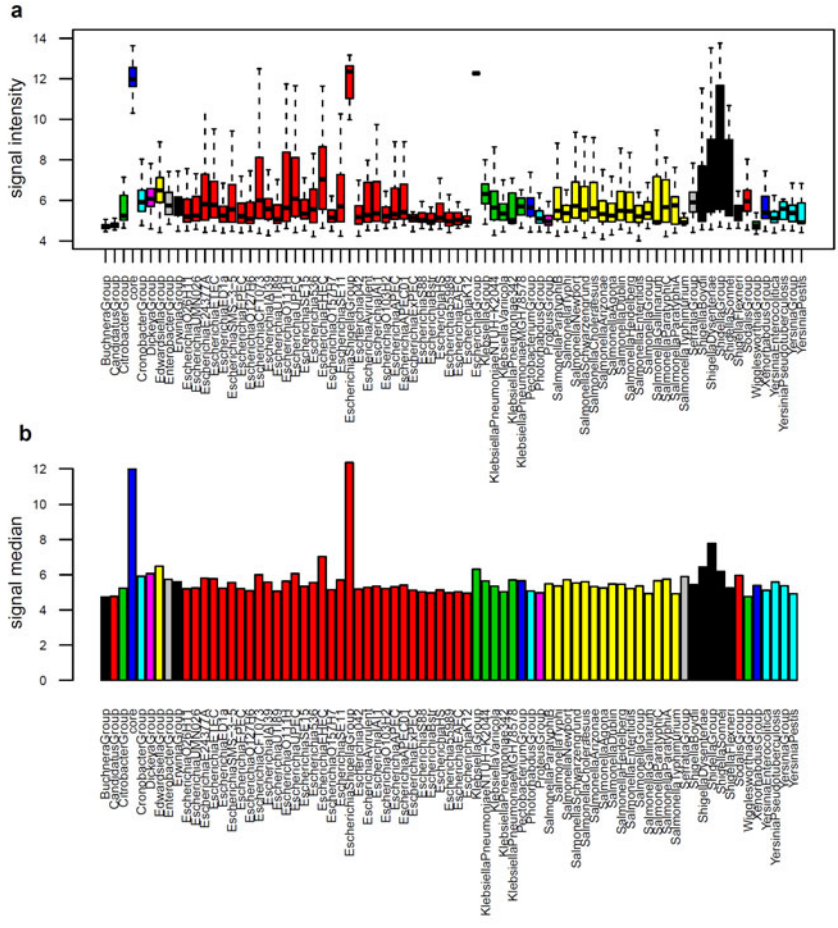


Fig. 1. Distribution of signal intensity and signal median for Escherichia coli ECOR20 strain among the set of seventy-eight groups, mentioned previously in Table 2. a. Box-and-whisker plot, showing signal intensity distribution. b. Bar plot, showing expression signal median distribution. X-axis elements are sorted by genus, based on the order showed in Table 2. Colour code is based on the genera, where 12-colour palette represents 20 genera.

the resulting plots for single representative of three chosen genera *Escherichia*, *Salmonella* and *Yersinia*. Those were *Escherichia coli* ECOR20, *Salmonella enterica* serovar Dublin and *Yersinia frederiksii*, respectively. Table 3 overviews the results for all the eleven isolates, used in the study.

Both box-and-whisker and bar plots for *Escherichia coli* ECOR20, represented in Fig. 1, show high signal intensity among the genes comprising core and *Escherichia*-and-*Shigella* groups. Additionally, results show high similarity to several pathogenic *E. coli* strains, such as *Escherichia coli* CFT073, and strains of O111:H-, UPEC and EHEC pathotypes. Apart from being highly expressed among the genes belonging to *Escherichia* genus, microarray data show relatively high signal level to *Shigella* genus

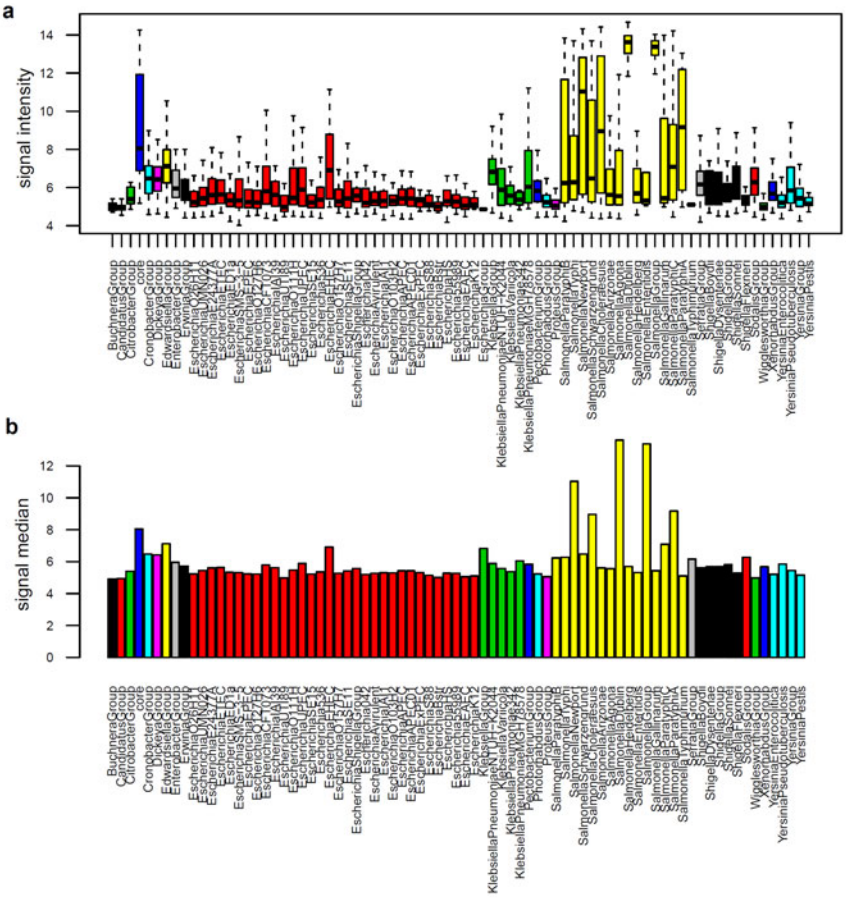


Fig. 2. Distribution of signal intensity and signal median for Salmonella enterica serovar Dublin strain among the set of seventy-eight groups, mentioned previously in Table 2. a. Box-and-whisker plot, showing signal intensity distribution. b. Bar plot, showing expression signal median distribution. X-axis elements are sorted by genus, based on the order showed in Table 2. Colour code is based on the genera, where 12-colour palette represents 20 genera.

strains, thus, resulting in another proof of *Escherichia* and *Shigella* genera strains being very similar.

Fig. 2 visualizes the comparison of data for *Salmonella enterica* serovar Dublin isolate. Genes have high intensity values within strains belonging to *Salmonella* genus and core group. The highest similarity is shown to be Dublin serovar; however, DNA sequences appeared to hybridize with the high strength to Newport, Choleraesuis and Paratyphi A serovar representing probes as well.

In the case of the chosen representative for *Yersinia* genus, *Yersinia frederiksi*, results, shown in Fig. 3, are not that positive, since any obvious high intensity signal cannot be seen. This might occur as a consequence of impropriate isolation of genomic DNA, low concentration of labeled DNA, which was obviously not enough for proper hybridization to target genes, or cross-hybridization effect.

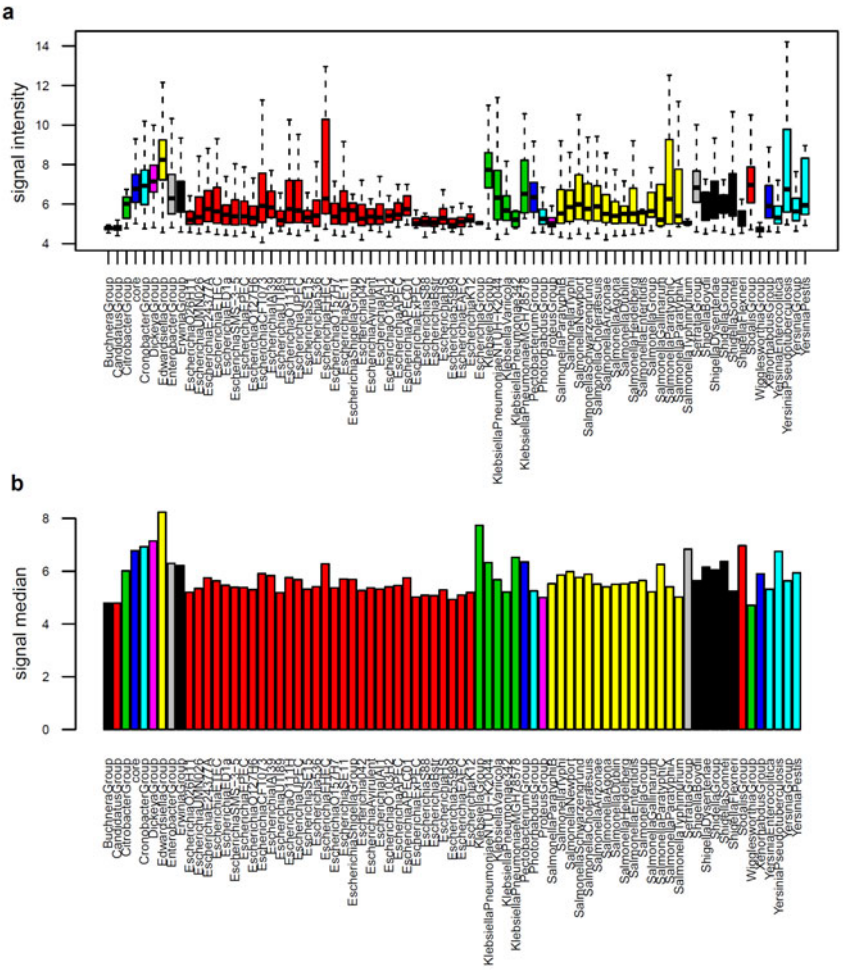


Fig. 3. Distribution of signal intensity and signal median for Yersinia frederiksii strain among the set of seventy-eight groups, mentioned previously in Table 2. a. Box-and-whisker plot, showing signal intensity distribution. b. Bar plot, showing expression signal median distribution. X-axis elements are sorted by genus, based on the order showed in Table 2. Colour code is based on the genera, where 12-colour palette represents 20 genera.

Isolates, results for which are presented in Table 3, show different chip performances. Several of them can be easily proved to belong to a particular genus, specific species and be most likely similar to a particular genus, species or serovar/serotype.

However, some samples, likewise *Yersinia frederiksii*, do not show obvious results. This can consider the presence of uncertainties included in genomic DNA purification and sample preparation for the hybridization.

Table 3. Overview of experimental validation results

Isolate / Distinguishing level	Genera	Species	Pathotype/Serovar
Escherichia coli ECOR20	+	+	-
Salmonella enterica serovar Dublin D6	+	+	+
Salmonella enterica serovar Paratyphi B var Java b	+	+	+
Salmonella enterica serovar Isangi 2005-60-2087-1	+	+	
Salmonella enterica Typhimurium HN-GSS-2007-016	+	+	+
Salmonella enterica serovar Choleraesuis 2870/08			
Shigella sonnei phase 12006-077	-	-	
Shigella flexneri 4 2006-054	+	+	
Shigella boydii 9S	-	-	-
Yersinia entericolitica O3 98-30624-5	-	-	-
Yersinia ruckerii NCTC 10476	-	-	-
Yersinia frederiksii P963	-	-	-

'+' is a positive result, '-' is a negative result and absence of any mark means no analysis with this purpose was made or results are not analysed

4 Discussion and Perspective

The design of a microarray chip covering 116 bacterial genomes has proven to be a considerable challenge. Multiple aspects had to be examined, such as the number of possible sequences to be included in the database, various criteria to select the unique set of genes to particular groups of genomes, and to design probes for them. The greatest difficulty was to optimize these criteria and to filter out the false positive representative sequences for each sequence of interest. Some genera within *Enterobacteriaceae*, such as *Escherichia* and *Shigella*, are quite similar, thus it was difficult to find genus-specific genes. For example, the *Escherichia* genus appeared to have only a single gene family conserved among all the strains belonging to this genus, and being absent in the other enterics. Thus it was an obvious decision to design probes for *Escherichia*-and-*Shigella* genera-specific genes.

Along with choosing representative sequence for each of unique gene family, a problem of selecting the right organism to extract representative sequences for core-genome set became evident. In this study, core-genome genes were extracted from type species of the type genus *Escherichia coli* K-12 MG1655 strain. The unique sets of genes were selected on protein level, that is, similarity/dissimilarity was based on alignment using BLASTP, and gene family members were considered based on the 50/50 rule, described above. Thus this might be an explanation of why some probes did not show high intensity levels at the DNA level as was predicted.

Selecting the probes is indeed a challenging aspect. On the one hand, probes should cover all versions of the same gene, however, at the same time they should be able to distinguish between different genera, species, pathotypes/serovars, and strains. Furthermore, the array should allow various numbers of probes per gene in order to acquire the sufficient coverage of genes. Longer sequences require higher numbers of probes, whereas design of the same number of probes for short genes would result in low quality probes [36]. Therefore, the challenge is to find the best possible solution, with least time, money, and personal energy consumption.

Several improvements and suggestions could be considered for the design of an *Enterobacteriaceae* pan-genome microarray chip. To obtain more sufficient unique gene finding, searchs should be done on DNA level with an appropriate cut-off value. Alignment using the BLASTN algorithm would be able to efficiently identify homologous nucleotide sequences based on similarity and would be helpful in avoiding non-specific probes.

Furthermore, for the validation of the chip step, sample preparations, such as genomic DNA isolation, labeling, and preparation to hybridize an array should be done according to protocols. Purity of DNA should be checked before the DNA labeling step to avoid small quantities of labeled DNA, which hybridizes to wrong sequences and fails to recognize the expected target sequence.

5 Conclusion

In this study, an *Enterobacteriaceae* pan-genome microarray chip was developed based on 116 genomes within this bacterial family. The typical genome size (with the exception of the reduced endosymbiont genomes of *Buchnera*, *Wigglesworthia* and *Sodalis* genera) contained between 3500 and 5500 genes. This made it possible to find at least 10 genus-, species- and pathotype/serovar-genes among all the analysed genomes. This resulted in 53644 unique probes, which were expected to hybridize to particular target sequence. High-density pan-genome microarrays can be very useful in both characterizing DNA content and monitoring expression levels for thousands of genes simultaneously. The comparison of two or more arrays can display the distinct patterns of gene expression or signal intensity level that are useful in the definition of unknown strains or genes included in these genomes. Using some experimental tests the ability of the microarray to determine bacterial strains within *Escherichia* spp., *Shigella* spp., *Salmonella* spp. and *Yersinia* spp. was demonstrated. Most of the results showed discriminative power, although some samples did not show a clear connection to the bacterial strain they are most likely to be similar to. This could be due to low quality DNA from the experiment.

It can be concluded that a *Enterobacteriaceae* pan-genome microarray, based on 116 genomes provides a perfect tool for determination of the genetic makeup of unknown strains within this bacterial family and can introduce insights into phylogenetic relationships.

Acknowledgments. This work is supported by grants from the Danish Center for Scientific Computing and the Danish Research Council. The authors would like to thank Colleen Ussery for help in editing the manuscript.

References

1. Hall, B.G., Ehrlich, G.D., Hu, F.Z.: Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. Microbiology 156, 1060–1068 (2010)
2. Sørensen, T.I., Nielsen, G.G., Andersen, P.K., Teasdale, T.W.: Genetic and environmental influences on premature death in adult adoptees. N. Engl. J. Med. 318, 727–732 (1988)

3. Helms, M., Vastrup, P., Gerner-Smidt, P., Mølbak, K.: Short and long term mortality associated with foodborne bacterial gastrointestinal infections: registry based study. BMJ 326, 357 (2003)
4. Ternhag, A., Törner, A., Svensson, A., Ekdahl, K., Giesecke, J.: Short- and long-term effects of bacterial gastrointestinal infections. Emerging Infect. Dis. 14, 143–148 (2008)
5. Mead, P.S., Slutsker, L., Dietz, V., McCaig, L.F., Bresee, J.S., Shapiro, C., Griffin, P.M., Tauxe, R.V.: Food-related illness and death in the United States. Emerging Infect. Dis. 5, 607–625 (1999)
6. Litrup, E., Torpdahl, M., Malorny, B., Huehn, S., Helms, M., Christensen, H., Nielsen, E.M.: DNA microarray analysis of *Salmonella* serotype Typhimurium strains causing different symptoms of disease. BMC Microbiol. 10, 96 (2010)
7. Laupland, K.B., Schønheyder, H.C., Kennedy, K.J., Lyytikäinen, O., Valiquette, L., Galbraith, J., Collignon, P.: *Salmonella enterica* bacteraemia: a multi-national population-based cohort study. BMC Infect. Dis. 10, 95 (2010)
8. Cheng, S., Hu, Y., Zhang, M., Sun, L.: Analysis of the vaccine potential of a natural avirulent Edwardsiella tarda isolate. Vaccine 28, 2716–2721 (2010)
9. Lindberg, A.M., Ljungh, A., Ahrné, S., Löfdahl, S., Molin, G.: *Enterobacteriaceae* found in high numbers in fish, minced meat and pasteurised milk or cream and the presence of toxin encoding genes. Int. J. Food Microbiol. 39, 11–17 (1998)
10. Musgrove, M.T., Northcutt, J.K., Jones, D.R., Cox, N.A., Harrison, M.A.: *Enterobacteriaceae* and related organisms isolated from shell eggs collected during commercial processing. Poult. Sci. 87, 1211–1218 (2008)
11. Stiles, M.E., Ng, L.K.: *Enterobacteriaceae* associated with meats and meat handling. Appl. Environ. Microbiol. 41, 867–872 (1981)
12. Wright, C., Kominos, S.D., Yee, R.B.: *Enterobacteriaceae* and *Pseudomonas* aeruginosa recovered from vegetable salads. Appl. Environ. Microbiol. 31, 453–454 (1976)
13. Cossart, P., Sansonetti, P.J.: Bacterial invasion: the paradigms of enteroinvasive pathogens. Science 304, 242–248 (2004)
14. Hornef, M.W., Wick, M.J., Rhen, M., Normark, S.: Bacterial strategies for overcoming host innate and adaptive immune responses. Nat. Immunol. 3, 1033–1040 (2002)
15. Olsson, C., Ahrné, S., Pettersson, B., Molin, G.: DNA based classification of food associated *Enterobacteriaceae* previously identified by biology microplates. Syst. Appl. Microbiol. 27, 219–228 (2004)
16. Glasner, J.D., Marquez-Villavicencio, M., Kim, H., Jahn, C.E., Ma, B., Biehl, B.S., Rissman, A.I., Mole, B., Yi, X., Yang, C., Dangl, J.L., Grant, S.R., Perna, N.T., Charkowski, A.O.: Niche-specificity and the variable fraction of the *Pectobacterium* pan-genome. Mol. Plant Microbe Interact 21, 1549–1560 (2008)
17. Tettelin, H., et al.: Genome analysis of multiple pathogenic isolates of *Streptococcus* agalactiae: implications for the microbial ”pan-genome”. Proc. Natl. Acad. Sci. U.S.A. 102, 13950–13955 (2005)
18. Lefébure, T., Stanhope, M.J.: Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol. 8, 71 (2007)
19. Phillippy, A.M., Deng, X., Zhang, W., Salzberg, S.L.: Efficient oligonucleotide probe selection for pan-genomic tiling arrays. BMC Bioinformatics 10, 293 (2009)
20. Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., Dairkee, S.H., Ljung, B.M., Gray, J.W., Albertson, D.G.: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat. Genet. 20, 207–211 (1998)

21. Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lipshutz, R., Chee, M., Lander, E.S.: Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280, 1077–1082 (1998)
22. Khodursky, A.B., Peter, B.J., Cozzarelli, N.R., Botstein, D., Brown, P.O., Yanofsky, C.: DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in Escherichia coli. Proc. Natl. Acad. Sci. U.S.A. 97, 12170–12175 (2000)
23. Wei, Y., Lee, J.M., Richmond, C., Blattner, F.R., Rafalski, J.A., LaRossa, R.A.: High-density microarray-mediated gene expression profiling of Escherichia coli. J. Bacteriol. 183, 545–556 (2001)
24. Jacobsen, L., Durso, L., Conway, T., Nickerson, K.W.: *Escherichia coli* O157:H7 and other E. coli strains share physiological properties associated with intestinal colonization. Appl. Environ. Microbiol. 75, 4633–4635 (2009)
25. Willenbrock, H., Fridlyand, J.: A comparison study: applying segmentation to array cgh data for downstream analyses. Bioinformatics 21, 4084–4091 (2005)
26. Willenbrock, H., Petersen, A., Sekse, C., Kiil, K., Wasteson, Y., Ussery, D.W.: Design of a seven-genome *Escherichia coli* microarray for comparative genomic profiling. J. Bacteriol. 188, 7713–7721 (2006)
27. Snipen, L., Almøy, T., Ussery, D.W.: Microbial comparative pan-genomics using binomial mixture models. BMC Genomics 10, 385 (2009)
28. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and psi-BLAST: a new generation of protein database search programs. Nucleic. Acids. Res. 25, 3389–3402 (1997)
29. Hyatt, D., Chen, G., Locascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J.: Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119 (2010)
30. Wernersson, R., Nielsen, H.B.: OligoWiz 2.0–integrating sequence feature annotation into the design of microarray probes. Nucleic Acids Res. 33, W611–W615 (2005)
31. Wernersson, R., Juncker, A.S., Nielsen, H.B.: Probe selection for DNA microarrays using OligoWiz. Nat. Protoc. 2, 2677–2691 (2007)
32. Vejborg, R.M., Bernbom, N., Gram, L., Klemm, P.: Anti-adhesive properties of fish tropomyosins. J. Appl. Microbiol. 105, 141–150 (2008)
33. Easy-DNA kit (2010), `http://tools.invitrogen.com/content/sfs/manuals/easydna_man.pdf`
34. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H., Zhang, J.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80 (2004)
35. Do, J.H., Choi, D.: Normalization of microarray data: single-labeled and dual-labeled arrays. Mol. Cells 22, 254–261 (2006)
36. Willenbrock, H., Hallin, P.F., Wassenaar, T.M., Ussery, D.W.: Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray. Genome Biol. 8, 267 (2007)

Multi-objective Particle Swarm Optimisation for Phase Specific Cancer Drug Scheduling

Mohammad S. Alam, Saleh Algoul, M. Alamgir Hossain, and M.A. Azim Majumder

University of Bradford, Bradford, UK
{M.Alam1,S.K.A.Algoul,M.A.Hossain1}@Bradford.ac.uk
A.A.Majumder@Bradford.ac.uk

Abstract. An effective chemotherapy drug scheduling requires adequate balancing of administration of anti-cancer drugs to reduce the tumour size as well as toxic side effects. Conventional clinical methods very often fail to balance between these two parameters due to their inherent conflicting nature. This paper presents a method of phase specific drug scheduling using a close-loop control method and multi-objective particle swarm optimisation algorithm (MOPSO) that can provide solutions for trading-off between the cell killing and toxic side effects. A close-loop control method, namely Integral-Proportional-Derivative (I-PD) is designed to control the drug to be infused to the patient's body and MOPSO is used to find suitable parameters of the controller. A phase specific cancer tumour model is used for this work to show the effects of drug on tumour. Results show that the proposed method can generate very efficient drug scheduling that trade-off between cell killing and toxic side effects and satisfy associated design goals, for example lower drug doses and lower drug concentration. Moreover, our approach can reduce the number of proliferating and quiescent cells up to 72% and 60% respectively; maximum reduction with phase-specific model compared to reported work available so far.

Keywords: Phase specific scheduling, Cancer chemotherapy, Cell compartment, Feedback control, Multi-objective optimisation, Particle Swarm Algorithm.

1 Introduction

Cancer refers to a set of disease where normal cells of the body lose their mechanisms which are responsible for controlling their growth and motility. Chemotherapy is one of the essential treatment methods for cancer. The main aim of chemotherapy is to minimise the number of cancer cells after a number of fixed treatment cycles with minimum toxic side effects. The efficiency of the dosages of the treatment is often measured as the interval of time from the start of therapy, until the end of treatment. The most important challenge of cancer treatment is to maintain the normal physiological states of the patient's body system during the course of different treatment schedules. This can be achieved by optimising chemotherapy treatment in such a way as to reduce tumour burden to a minimum level with minimum/acceptable toxic side effects. The simplest mathematical models which are commonly used in

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 180–192, 2010.

research for optimal control of cancer chemotherapy assume the entire cell cycle as one compartment [1]. In many cases, these single compartment models prove to be inadequate due to the over-simplified nature of the model. The cell cycle is modelled in the form of multiple compartments which describe different cell phases or combine phases of the cell cycle into clusters. In general, the cycle comprises of five stages as show in Figure 1. A brief description of different stages is given below [2], [3]:

First stage of the cell cycle is called Post mitotic gap which is indicated by G1. The cell prepares for DNA synthesis in this stage. In the second stage, denoted by S, DNA synthesis takes place in preparation for cell division. The third stage, G2 is called Pre-mitotic gap when specialised proteins and RNA are synthesised in preparation for cell division. In the fourth phase, called Mitotic phase (M), cell division takes place to produce two identical daughter cells and the last phase is Resting phase, indicated by G0. The cells become quiescent in this phase, i.e., viable but unable to divide.

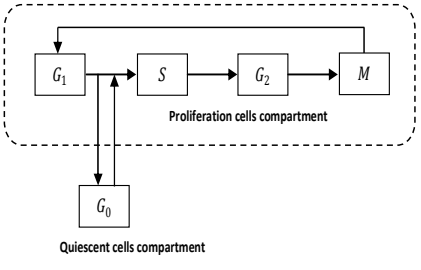


Fig. 1. Schematic diagram of different phases of cell cycle

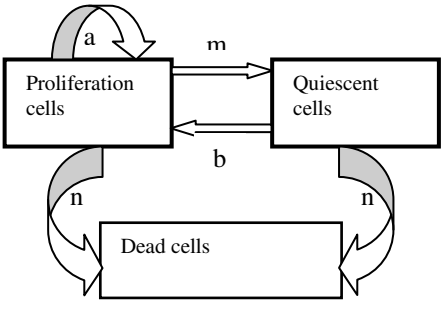


Fig. 2. Two compartments functional within tumour tissue

Of the multi-compartment models, the simplest and at the same time most natural ones, are two/three compartment models; which divide the cell cycle into two/three compartments. Figure 2 shows a two-compartment model where proliferating part contains actively dividing cells whereas quiescent part is inactive cells, but capable of dividing if a certain stimulus is given. The dead cells are unable to divide because they have completed their life cycle. In model, P (Proliferating) present the combination of the first four stages of the cell cycle as mentioned earlier ($\mathrm{G}_1, \mathrm{S}, \mathrm{G}_2$ and M) and Q (Quiescent cells) indicates stage G_0. The parameters m and b express the immigrants between the proliferating cells and quiescent cells respectively. Here a indicates to the growth rate of cycling cells and n is the natural decay of the cycling cells [4].

A number of models have been developed and used to characterise the evolution and effects of treatment on cancer by dividing the tumour into number of compartments (phase-specific) as considered in [5], [6]. Petrovski and co-workers in [7] used a relatively new bio-inspired algorithm, called particle swarm optimisation (PSO) to design chemotherapy drug scheduling using aforementioned design objective and constraints. The same authors also utilised multi-objective evolutionary algorithms in [8] to design chemotherapy drug scheduling where drug doses and toxic side effect were set as constraints.

This paper presents a novel method of chemotherapy drug scheduling using feedback control strategy and Multi-Objective Particle Swarm Optimisation (MOPSO). A close-loop control, namely Integral-Proportional-Derivative (I-PD) is used to control the drug doses to be infused to the patient's body. MOPSO optimisation process is employed to find parameters of the controller that trade-off between two conflicting objectives; reducing cancerous cells and toxic side effects simultaneously. The research aims at scheduling of a single drug, so, a two compartment phase specific cancer tumour model is developed and used for this work.

2 Mathematical Model of Two Compartment Model

A number of differential equations used to build a two compartment model are stated below [3]:

$$\frac{dP}{dt} = (a - m - n)P(t) + bQ(t) - g(t)P(t), \quad P(0) = P_0 \tag{1}$$

$$\frac{dQ}{dt} = mP(t) - bQ(t), \quad Q(0) = Q_0 \tag{2}$$

$$\frac{dY}{dt} = \delta y(t)\left(1 - \frac{Y(t)}{K}\right) - g(t)Y(t), \qquad Y(0) = Y_0 \tag{3}$$

$$\frac{dD}{dt} = u(t) - \gamma D(t), \quad D(t) = D_0 \tag{4}$$

$$g(t) = k_1 D(t) \tag{5}$$

$$\frac{dT}{dt} = D(t) - \eta T(t) \tag{6}$$

Where P and Q represent population of proliferating and quiescent cells. Here parameters $a, m, b \; and \; n$ indicate the rate of growth of proliferation cells, immigrant from cycling to quiescent cells, immigrant from quiescent cells to cycling cells and natural death of cycling cells respectively. Parameter $g(t)$ indicates the effects of drug on tumour cell which is the rate of cell killing per unit drug. Equations (1) and (2) show the rate of change of cell population in the proliferating and quiescent compartments of the tumour site during the period of treatment and equation (3) indicates the effect of chemotherapy where $Y(t)$ indicates the normal cells population, δ and K present the growth rate of the normal cells and the carrying capacity of normal cells respectively. $Y(0)$ is the initial value of normal cell population at the beginning of the treatment. Equation (4) shows the rate of change of drug concentration at the tumour site during the treatment cycle. Here $u(t)$ is the amount of drug doses to be infused to patient's body and γ is drug decay which is related to the metabolism of drug inside patient's body. Equation (5) shows the relationship between drug concentration at the tumour site and cell killing rate k_1 Equation (6) shows the relationship between level of toxicity $T(t)$ and drug concentration $D(t)$

where parameter η indicates the rate of elimination of toxicity. Using the above equations, a Matlab/Simulink [9], [10] model was developed with parameters and values as illustrated in Table 1.

Table 1. Parameters of Patient Model [3]

Symbol	Parameters	Values
a	The rate of growth Proliferating cells	0.5 day^{-1}
m	The mutation rate of proliferating cells to quiescent cells	0.218 day^{-1}
n	The natural end of the cycling cells	0.477 day^{-1}
b	The mutation rate of quiescent cells to proliferating cells	0.05 day^{-1}
δ	The rate of normal cell growth	0.1 day^{-1}
K	The carrying capacity of normal cell	10^9 cells
P	The proliferating cells population	$2x10^{11}$
Q	The quiescent cells population	$8x10^{11}$
Y	The normal cells population	10^9
$\mathbf{Y_{min}}$	The limitation of normal cells	10^8

3 Proposed Control Schema

A schematic diagram of chemotherapy drug scheduling scheme for cancer treatment is shown in Figure 3. A feedback control method is developed in order to maintain a predefined level of drug concentration at tumour sites. A close-loop control; Integral-Proportional-Derivative (I-PD) is used to control the drug to be infused to the patient's body. The proposed I-PD controller involves three parameters, the proportional gain k_p, integral gain k_i, and derivative gain k_d, Drug concentration at the tumour is used as the feedback signal to the controller which is compared with a predefined reference level. The difference between reference input and drug concentration at tumour site, output- $D(t)$, of the model is called the error, $e(t)$ which is used as input to the controller. The output of the controller, $u(t)$ as:

$$u(t) = K_i \int_0^t e(t)dt - [K_d \frac{d}{dt} D(t) + K_p D(t)] \quad (7)$$

$$e(t) = (X_D - D(t)) \quad (8)$$

It is noted that X_D indicates reference signal to the controller which can be depicted as the desired drug concentration to be maintained at the tumour site during the whole period of treatment. The reference input, both magnitude and pattern, is very crucial in the proposed drug scheduling scheme since reduction of cancerous cells largely depends on drug concentration developed in plasma and at the tumour site. It is noted that when the $e(t)$ is zero, the drug concentration at tumour site will be equal to the desired drug concentration. In such case, the cell killing will be maximum. The efficacy of the drug doses depends on three parameters k_i, k_p and k_d of I-PD controller. In this work, step input signal is chosen as the reference input to the close-loop control system that will ensure approximately a constant level of drug

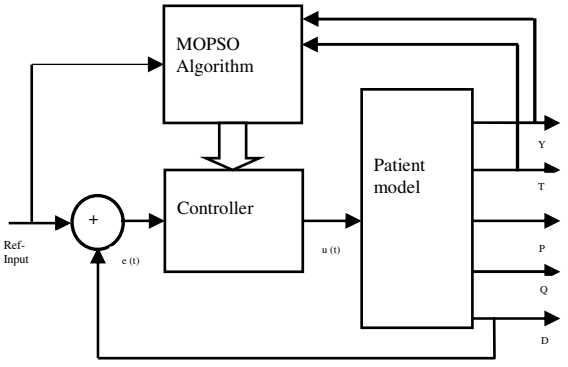


Fig. 3. Schematic diagram of the proposed drug scheduling scheme

concentration for most of the time of the treatment cycle. In this work, a MOPSO is used to find three parameters of the I-PD controller. It is noted that the chemotherapy drug scheduling is design for a period of 84 days [2].

3.1 Design Objectives and Goal Values

The design objectives and goal values of different performance measures of chemotherapy drug scheduling are as follows:

a) The number of proliferating cells should be reduced to a minimum or acceptable level at the end of the treatment. Dua et al., in [3] reported a reduction of approximately 70% for phase specific treatment. In this work, the acceptable goal value for reduction of proliferating cell is approximately set at 65% in the multi-objective optimisation process.
b) The number of quiescent cells is also required to be minimum at the end of the treatment. With Chemotherapy treatment, the number will reduce and the reduction is set at 55% as minimum acceptable value in this work.
c) The number of normal cells is inversely proportional to the toxicity developed in patient's body. So this is required to be as high as possible throughout the whole treatment period.
d) The level of toxicity should be as low as possible during the whole period of treatment. The maximum toxicity should not exceed 100.
e) Due to risk of toxic side effects, the drug doses infused to the patient's body should be low but effective in the treatment.
f) Since a close-loop control strategy has been used to control the drug scheduling, stability of the whole system is very crucial. As mentioned earlier, drug concentration is used as the feedback in the control scheme, settling of this parameter within a range of ±2% of the reference signal (desired drug concentration) has been used as a constraint. Design objectives and goal values of phase specific drug scheduling are listed in Table 2.

Table 2. Design objectives and goal values

Notations	Design objectives	Accepted goal values
RC_P	% of Reduction of proliferating cells	$RC_P > 65\%$
RC_Q	% of Reduction of quiescent cells	$RC_Q > 55\%$
Y(t)	Number of normal cells	$Y(t) > 1x10^8$
T(t)	Toxicity	$T(t) \leq 100$
D(t)	Drug concentration	$10 < D(t) \leq 50$

4 Multi-objective Particle Swarm Optimisation

PSO was first designed to simulate birds seeking food, defined as a "cornfield vector" [11]. PSO is a population-based search algorithm and is initialised with a population of random solutions, called particles and each particle in the PSO has an associated velocity. Particles fly through the search space with velocities which are dynamically adjusted according to their historical behaviours. The dynamic equation of basic PSO [12], with inertia coefficient is given as:

$$v_{id} = \omega \times v_{id} + c_1 \times r_1 \times (p_{id} - x_{id}) + c_2 \times r_2 \times (p_{gd} - x_{id}) \tag{9}$$

$$x_{id} = x_{id} + v_{id} \tag{10}$$

where c_1 and c_2 are positive constants, r_1 and r_2 are two random functions in the range [0,1], $X_i = (x_{i1}, x_{i2}, ..., x_{id})$ represents the i-th particle, $P_i = (p_{i1}, p_{i2}, ..., p_{id})$ represents the best previous position (the position giving the best fitness value) of the i-th particle, the symbol g represents the index of the best particle among all the particles in the population, and $V_i = (v_{i1}, v_{i2}, ..., v_{id})$ represents the rate of the position change (velocity) for particle i. Here p_{id} represents the best previous position of the i-th particle and p_{gd} represents the best particle among all the particles in the population. These two terms, p_{id} and p_{gd}, are usually known as local guide (or pbest)' and 'global guide' (or gbest). This algorithm deals with only one objective. In order to handle more objectives, some changes are needed in the operation of single objective PSO. In case of single objective optimisation problem, p_{id} and p_{gd} are selected based on the objective function either minimum or maximum value as far as minimisation or maximisation problem is concerned. For a multi-objective optimisation a wide range of solutions is obtained. So the main challenge, in designing a MOPSO algorithm, is to select pbest and gbest for each particle so as to obtain a wide range of solutions that trade-off among the conflicting objectives.

4.1 Selection Method of Global Guide and Local Guide

In the proposed algorithms, a new technique is introduced that combines external archive and non-dominated fronts of the current population in order to select gbest for

each particle. An external archive and associated control mechanism, as used in [13], is also employed here. For a two-objective optimisation problem, Figure 4 shows the state of the external archive and solutions of the current particles in the objective domain. The dark circles inside a 2-D grid structure indicate the non-dominated solutions found so far while circles on the right represent solutions of current particles in a 2-D objective domain and the number associated with them indicate index of the particles in the initial population. The current solutions are sorted based on Pareto dominance and several non-dominated fronts (ND fronts) are formed as shown in Figure 4.

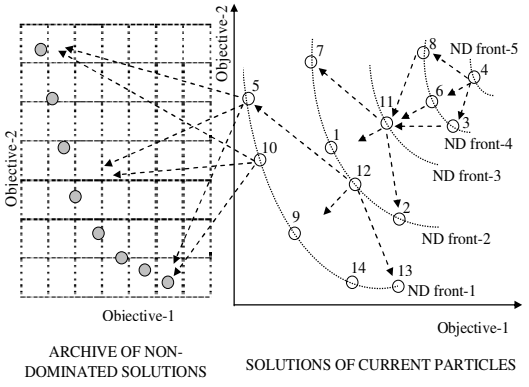


Fig. 4. Schematic diagram for finding gbest guide for particle in MOPSO

For each particle on ND front-1, the corresponding gbest is selected from the external archive based on fitness sharing and roulette wheel selection method (see Figure 4). Details of this process can be found in [14]. For particles on the remaining fronts, i.e., ND front-2, 3, 4 and 5: gbest of each particle is selected in the following way:

At first, shared fitness of each particle in the current population is calculated based on the exact non-dominated sorting GA (NSGA) fitness assignment scheme which was adopted by Srinivas and Deb [15]. Then, for each particle on ND fron-2, the corresponding gbest is selected from particles lying on the immediate lower front (better solutions), i.e., ND front-1, based on shared fitness and roulette wheel selection method (see Figure 4). This process continues for particles residing on the remaining ND fronts. Local guide or pbest for each particle is selected based on fitness sharing technique as explained in [16].

5 Implementation

The whole simulation is carried out in the Simulink environment with some .m-files of Matlab® [9], [10]. The Simulink model is chosen because it allows for simple construction of the model and control system with simple built-in components. At first, MOPSO has been used to design chemotherapy drug scheduling which find trade-off between two competing objectives; (i) number of proliferating cells at the end of the treatment and (ii) average level of toxicity over the whole period of treatment. The objectives are formulated as follows:

$$f_1(x) = P(t_f) \tag{11}$$

$$f_2(x) = \frac{1}{t_f}\int_o^{t_f} T(t)dt \tag{12}$$

where $P(t_f)$ is number of proliferating cells at the end of the treatment, $T(t)$ is the toxicity and t_f is the total period of chemotherapy treatment, which is 84 days (12 weeks). Stability of the close-loop system and Design objectives, as listed in Table 2 and are used as constraints in the optimisation process. A swarm of 20 particles having 3 elements each, i.e., 20 × 3 is created randomly within the range of [0, 2]. Each particle represents a solution where the three elements are assigned to controller parameters; proportional gain k_p, integral gain k_i, and derivative gain k_d respectively. The acceleration coefficients of MOPSO algorithm are set as $c_1 = c_2 = 1.5$, and inertia coefficient ω is gradually decreased from 1.4 to 0.1 with generation. Particles (solutions) not satisfying aforementioned design constraints are penalised with very large numbers, called penalty function. This penalty function will guide the particles towards better search region. The optimisation process is run for a maximum generation of 200. The maximum number of solutions that the external archive can keep is limited to 100. The Pareto optimal set at generation 200 is shown in Figure 5.

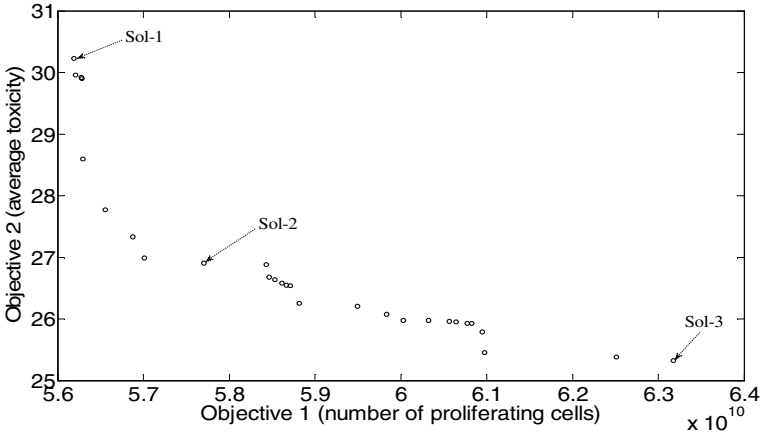


Fig. 5. Pareto optimal set of MOPSO optimisation at generation 200

5.1 Validation and Results

In order to evaluate the effectiveness of MOPSO in chemotherapy drug scheduling, several representative solutions are further assessed. To validate the solution set, three solutions are selected on the Pareto front, one from each region. Solutions are selected in such a way that two fall on either extreme points of the two objectives, the other is at approximately in the middle of objective domain. Three selected solutions, as

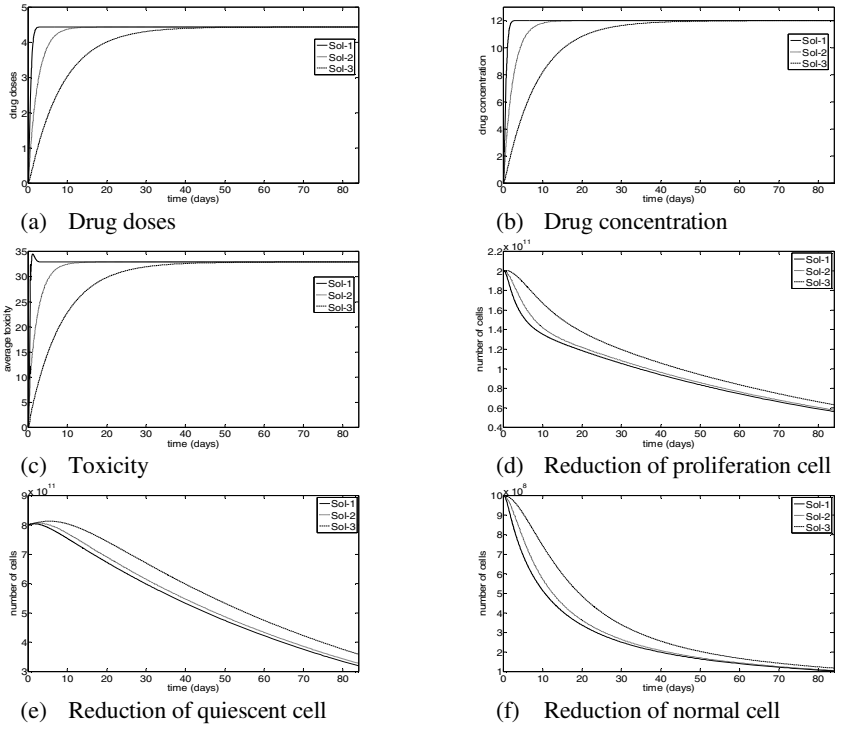
shown in Figure 5 will be denoted as Sol-1, Sol-2 and Sol-3 for further discussion. To make the chemotherapy drugs effective, the drug concentration at the tumour site should be maintained at a desired level for the whole period of treatment and this is implemented by using a fixed level of signal, called step input. In this work, the reference to the controller is selected by trial and error so that the maximum toxicity always remains below the maximum allowable value as indicted in design objective in Table 2. The fixed reference value is set at 12 in this work.

To obtain different performance measures in relation to chemotherapy treatment, decision variables, k_p, k_i and k_d corresponding to solutions; Sol-1, Sol-2 and Sol-3 are feed to the I-PD controller and the whole system along with the patient model is simulated for 84 days. Then the output of the I-PD controller, $\mathrm{u(t)}$, which is the chemotherapy drug scheduling is recorded in each case. Moreover, outputs of the patient model, such as, drug concentration at tumour site, toxicity, reduction of proliferating and quiescent cells and changes in normal cells are also recorded due to the infusion of the chemotherapy doses. Figure 6(a) shows the chemotherapy drug scheduling for Sol-1, Sol-2 and Sol-3. The response of the patient model due to the infusion of these drug scheduling are shown in Figures 6(b)-6(f).

Moreover, several performance measures of chemotherapy treatment, such as maximum and average levels of drug doses, toxicity and drug concentrations for all three solutions are recorded and presented in Table 3. Furthermore, percentage of reductions in proliferating and quiescent cells at the end of chemotherapy treatment are also determined and showed in Table 3. The number of normal cells remaining at the end of treatment gives an indication about the physiological state of the patient. So this number is also calculated and displayed in same Table.

***a) Drug Scheduling*:** Figure 6(a) shows the chemotherapy drug scheduling for Sol-1, Sol-2 and Sol-3. In all cases, the drug doses increase from zero and finally become stable at a level of 4.4. It is noted that the rate of increase is different for different solutions. For Sol-1, the doses reach maximum value of 4.4 within the first week of treatment and for the remaining periods it becomes stable at that value. For Sol-2, the chemotherapy drug scheduling takes slightly more than two weeks to be to reach the maximum and stable level whereas for Sol-3, it takes nearly seven weeks. Although in all three cases the maximum chemotherapy drug doses are same but the average drug doses over the whole period of treatment are different. For Sol-1, the average drug dose is maximum, which is 3.9 whereas this value is minimum (=3.4) for Sol-3. For Sol-2, the average drug dose is 3.5. It is noted that, the drug doses are much lower compared to conventional doses during 84 days of treatment [3], [6]. It is important to mention that, phase specific chemotherapy drugs, such as Vinca alkaloids, Hydroxyurea, Cytosine arabinoside, Methotrexate, 6-Mercaptopurin, 6-Thioguanine, Procarbazin, VM-26 and VP16-213 [6] are, in general, toxic agents and lower doses of these drugs may reduce the toxic side effects during the treatment cycle and thereby improve the quality of life of the patient [2].

Toxicity: The toxicities, for Sol-1, Sol-2 and Sol-3, developed due to the corresponding chemotherapy drug scheduling are shown in Figure 6(c). For all three solutions, the toxicities gradually increase from the first day of treatment and finally settle to a steady value after few weeks in a similar manner as observed in case of drug scheduling and drug concentration. The maximum level of toxicity is observed



(a) Drug doses (b) Drug concentration

(c) Toxicity (d) Reduction of proliferation cell

(e) Reduction of quiescent cell (f) Reduction of normal cell

Fig. 6. Performance measures of chemotherapy treatment for Sol-1, Sol-2 and Sol-3

Table 3. Performance measures of drug scheduling techniques

Example Solutions	For the whole period of treatment						At the end of 84 days treatment		
	Drug doses		Drug concentration		Toxicity		Reduction of Proliferating Cells	Reduction of Quiescent cells	No. of Normal cells
	Max	Avg	Max	Avg	Max	Avg			
Sol-1	4.4	3.9	12	10.6	34.4	30.2	72%	60%	$1.03x10^8$
Sol-2	4.4	3.5	12	9.5	32.9	26.9	71%	59%	$1.05x10^8$
Sol-3	4.4	3.4	12	9.1	32.8	25.3	68%	55%	$1.17x10^8$

with the drug scheduling obtained with Sol-1 and the value is 34.4 whereas the minimum toxicity is caused by Sol-3. The average toxicities for Sol-1, Sol-2 and Sol-3 are 30.2, 26.9 and 25.3 respectively.

c) Reduction of cells: The main aim of chemotherapy treatment is to reduce proliferating and quiescent cells without affecting normal cells much during the treatment. Before the treatment starts, the numbers of all cells are listed in Table 1. Figure 7(d) shows the reduction of proliferating cells during the whole period of

treatment. For Sol-1, Sol-2 and Sol-3, the percentage of reductions obtained using the drug scheduling shown in Figure 6(a) are 72%, 71% and 68% respectively. During the treatment period, the number gradually decreases depending on chemotherapy drug doses and this is observed for all three solutions in Figure 6(d). Similar trend is observed in case of quiescent cells and it is shown in Figure 6(e). It is important to note that the reduction for Sol-1, Sol-2 and Sol-3 are 60%, 59% and 55% respectively. Figure 6(f) shows the changes of normal cells during the whole period of treatment for all cases. For Sol-1, Sol-2 and Sol-3, the number of normal cells remaining at the end of 84 days treatment are 1.03×10^8, 1.05×10^8 and 1.175×10^8 respectively. It is mentioned that in all solutions, the number is higher that the threshold value as indicated in Table 2. Moreover, these higher values of remaining normal cells are attributed to lower toxic side effects and better physiological conditions of patients.

Remark 1: *Dua and co-workers in [3] designed chemotherapy drug scheduling for cell cycle specific model, as used in the present work, and reported reductions of 70% and 50% for proliferating and quiescent cells at the end of treatment. In present work, Sol-1 and Sol-2 result a reduction of 72% and 71% for proliferating cells which are slightly more than the reported one. More importantly to note that, example solutions; Sol-1, Sol-2 and Sol-3 of present work can reduce the quiescent cells up to 60%, 59% and 55% respectively which are significantly higher than the reported result. It is clearly evident that cell reductions for both proliferating (except Sol-3) and quiescent cells are better in case of proposed model.*

Remark 2: *Considering the physiological state of the patient and state of the cancer, the oncologist can choose a suitable solution from the objective space suitable for the patient. For patients, having better physiological conditions and requiring faster response, chemotherapy drug scheduling resulting from or around Sol-1 can be chosen. On the other hand, patients having relatively poor physiological conditions and vulnerable to toxic side effects may be given chemotherapy doses based on solutions residing around Sol-3. In general, chemotherapy drug scheduling resulting from solutions close to Sol-2 may be preferred unless there are some specific reasons.*

6 Comparative Performances

This section presents a comparative performance analysis of the proposed drug scheduling pattern with some reported works using similar cancer tumour models. The outputs of the proposed drug scheduling scheme is compared with the results reported by Dua et al in [3]. Figure 7 compares of the percentage of reduction of proliferating and quiescent cells at the end of treatment cycles with proposed model and the reported one in [3]. It is noted that the reduction of proliferating cells in case of proposed model is 72% compared to 70% in reported one. It is also noted that the reduction of quiescent cells is 60% whereas the reported model yields only 50%. It is clearly evident that cell reductions for both proliferating and quiescent cells are better in case of proposed model.

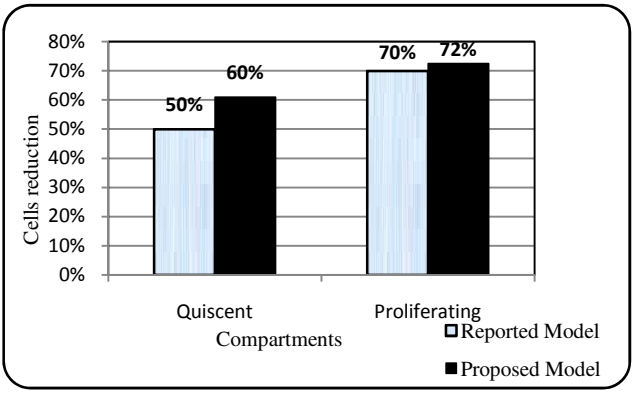


Fig. 7. Comparative performance for reported (Dua et al, [3]) and proposed model in Sol-1

7 Conclusions

This paper has presented a method of chemotherapy drug scheduling using a close-loop control method and multi-objective particle swarm optimisation (MOPSO). Two main objectives of chemotherapy treatment; reducing cancerous cells and reducing toxic side effects are always found in conflict. MOPSO optimisation process is used to design the drug scheduling that trade-off between these two. A close-loop control method, namely I-PD is designed to control the drug to be infused to the patient's body for a cell cycle specific treatment and MOPSO is used to find acceptable/suitable parameters. In the proposed method, several design objectives, constraints and associated goal values are defined prior to the optimisation process and a wide range of non-dominated solutions have been obtained satisfying all design goals, known as Pareto-optimal set, which trade-off among competing objectives. It is interesting to note that the design approach can offer flexibility in decision making and suitable solution can be picked under different trade-off interventions for cancer treatment. It is noted that the drug scheduling pattern of the proposed model offers better performance as compared to the reported models available till date. The same control strategy and optimisation technique can be extended for multidrug or combination chemotherapy regimen. Future work will include verification of the proposed scheduling with clinical data and experiments.

Acknowledgement

This research has been supported by the EU Erasmus Mundus Project - eLINK (east-west Link for Innovation, Networking and Knowledge exchange) under External Cooperation Window – Asia Regional Call (EM ECW – ref. 149674-EM-1-2008-1-UK-ERAMUNDUS).

References

[1] Martin, R.: Optimal control drug scheduling of cancer chemotherapy. Automatica, 1113–1122 (1992)

[2] Martin, R., Teo, K.: Optimal control of drug administration in chemotherapy tumour Growth. World Scientific, Singapore (1994)

[3] Dua, P., Dua, V., Pistikopoulos, N.: Optimal delivery of chemotherapeutic agents in cancer. Computer and Chemical Engineering 32, 99–107 (2008)
[4] Swierniak, A., Ledzewicz, U., Schättler, H.: Optimal control for a class of compartmental models in cancer chemotherapy. Int. J. Appl. Math. Comput. Sci. 13(3), 357–368 (2003)
[5] Ochoa, M., Burke, E.: An evolutionary approach to cancer chemotherapy scheduling. Springer science 8, 301–318 (2007)
[6] Liang, Y., Leung, K., Mok, T.: Evolutionary drug scheduling models with different toxicity metabolism in cancer chemotherapy. Applied soft computing 8, 140–149 (2008)
[7] Petrovski, A., Sudha, B., McCall, J.: Optimising cancer chemotherapy using particle swarm optimization and genetic algorithms. In: Yao, X., Burke, E.K., Lozano, J.A., Smith, J., Merelo-Guervós, J.J., Bullinaria, J.A., Rowe, J.E., Tiňo, P., Kabán, A., Schwefel, H.-P. (eds.) PPSN 2004. LNCS, vol. 3242, pp. 633–641. Springer, Heidelberg (2004)
[8] McCall, J., Petrovski, A., Shakya, A.: Evolutionary Algorithms for Cancer Chemotherapy Optimization. Computational Intelligence in Bioinformatics, 265–296 (2008)
[9] The Mathworks, Inc.: Simulink Control Design User's Guide (2008)
[10] The Mathworks, Inc.: MATLAB Reference Guide (2010)
[11] Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proc. of IEEE International Conference on Neural Networks (ICNN), Perth, Australia, vol. IV, pp. 1942–1948 (1995)
[12] Eberhart, R., Shi, Y.: Particle swarm optimization: developments, applications and resources. In: Proc. Congress on Evolutionary Computation 2001, Seoul, Korea. IEEE Service Centre, Piscataway (2001)
[13] Coello Coello, C.A., Toscano Pulido, G., Salazar Lechuga, M.: Handling multiple objectives with particle swarm optimization. IEEE Transactions on Evolutionary Computation 8, 256–279 (2004)
[14] Tokhi, M.O., Alam, M.S.: Particle Swarm Optimisation Algorithms and Their Application to Controller Design for Flexible Structure Systems. IST Transactions of Control Engineering-Theory and Applications 1(3(9)), 12–25 (2010) ISSN 1913-8849
[15] Srinivas, N., Deb, K.: Multiobjective Optimization Using Nondominated Sorting in Genetic Algorithms. Evolutionary Computation 2(3), 221–248 (1994)
[16] Deb, K.: Multi-objective optimization using evolutionary algorithms. Wiley, Chichester (2001)

A Vaccine Strategy for Plant Allergy by RNA Interference – An *in Silico* Approach

Ramya Ramadoss and Chee Keong Kwoh

School of Computer Engineering, Nanyang Technological University,
Nanyang Avenue, Singapore 639798
ASCKKWOH@ntu.edu.sg
http://www.ntu.edu.sg

Abstract. Worldwide population affected by allergic rhinitis and asthma are estimated to 400 million and 300 million respectively, and the medical costs for treatment are estimated to exceed that of tuberculosis and AIDS allied. The main objective of this research is to propose a vaccine design strategy for the management of allergy through siRNA vaccination in silencing IgE VH region. The allergen *Che a 3* was chosen to demonstrate our approach. Docking interactions between *Che a 3* and modeled structures of heavy chain variable region of 31 Immunoglobulin E clones were analyzed in AutoDock. Concurrently, small interference RNA sequences targeting the Immunoglobulin E clone with least binding energy were designed in siDRM.

Keywords: Allergy, Asthma, Immunoglobulin E, Vaccine, Immunotherapy, Small interference RNA, AutoDock, Bioinformatics, Docking, *In silico*.

1 Introduction

Allergy is defined as the acute immune reaction induced by allergic compounds. The main objective of this study is to mediate the molecular mechanisms behind allergy and propose a novel strategy to reduce this adverse reaction. It has been demonstrated by Zhang *et al.*[1] that bioinformatics can serve as a catalyst to drive the wet lab experiments into a cost-effective and time-effective paradigm to formulate epitope-based vaccines. Various remedies for allergy in current practice are reviewed by Holgate and Polosa [2]. Traditional treatments include Corticosteroids, β2- adrenoceptor agonists, Mediator antagonists and synthesis inhibitors and phosphodiesterase inhibitors. However, they fail to eradicate natural history of the disease [3],[4], virus-induced exacerbations [30] and ineffective in smoking asthma patients. And their side effects like anaphylaxis [6], central nervous system incitement and cerebral appraisals of sleep and early morning behavior [7] were inevitable. This led to the discovery of cetirizine, levocetirizine, loratadine and desloratadine [8]. The side effects and non-uniform effectiveness among patients instigated evolvement of Allergen-specific immunotherapy (SIT).

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 193–207, 2010.

Where T regulatory (T REG) cells boost protective immunotolerance against allergens and maintain balance between TH1 and TH2 cells populations to subdue allergic reactions [9]. SIT entails injection of allergic protein(s) in incremental doses to suppress immune responses mediated by mast cells, basophils and eosinophils and allergen-specific immunoglobulin A (IgA) and immunoglobulin G4 (IgG4) antibodies on sensitization. Vaccines based on allergen extracts, allergoids, peptides, recombinant allergens, epitope modified allergens or allergen-CpG fusion molecules [10],[11],[12] and RNA interference (RNAi) mediated therapies are currently being widely researched. Functionally, mRNA molecules are translated into a protein. So, targeting mRNA than a protein is potentially an efficient approach [12]. It has been shown that RNAi is highly competent in suppression of specific genes when compared to the traditional antisense approaches [13],[14],[15] .

Suzuki *et al* [16] have established that siRNA dependent silencing of CD40 mediated immune responses [17],[18] can be effective therapy against allergy. Amidst various allergic reactions that can be suppressed by siRNA, IgE seems to be an efficient candidate as it was evolved in mammals as the first line of defense against pathogens [19]. Furthermore, therapies arresting IgE mediated responses have been highly successful. Antibody specific to low-affinity IgE receptor FcξRII, Lumiliximab has passed Phase 1 trial against asthma [20] and the IgE-specific antibody, Omalizumab (Xolair; Novartis Pharmaceuticals Ltd) has been effective in treatment of asthma and other allergic diseases [21].

IgE is associated with many other mechanisms apart from aggravating allergic responses like restraining malaria parasites [22], helminthes infection [23], *Trichinella spiralis* infection [24] and ovarian tumor cells [25] making it vulnerable to block expression of complete IgE molecule to combat allergy. Alternatively, gene expression of heavy chain variable (VH) region of IgE specific to allergies can be silenced. Steering towards VH region is quiet convincing as a potent IgE is assembled only on successful rearrangement of VH region by V(D)J recombination [26],[27] . Also, it is the variable domain H3 [28] (heavy chain region) that determine specificity of antigen binding sites in an Ig [28],[29] . The abovementioned discussions affirm an effective therapy for allergy by targeting VH region of IgE while preserving its role in other immune reactions.

The main objective of this study is to design siRNA based vaccination against allergy *in silico*. siRNA sequences are designed using the online tool, siDRM [30] which is available at `http://sirecords.umn.edu/siDRM/` pertaining to its high Positive Predictive value compared to other tools [30] . The vaccine is involved in silencing gene expression of IgE VH region. siRNA is designed to target the IgE specific to a particular allergen. Furthermore, Allergen *Che a 3* and IgE VH regions were docked in AutoDock [31],[32],[33] to identify the genetic variant capable of recognizing the epitopes in the allergen where binding sites of IgE are predicted in the online prediction tool, Q-SiteFinder (`http://www.modelling.leeds.ac.uk/qsitefinder/`) [34] prior to docking analysis.

2 Methods

2.1 IgE Sequence Retrieval from NCBI

Coker et al [35] have published 112 protein sequences for IgE VH region which were retrieved from NCBI. Only 31 sequences among them could be successfully modeled in Swiss-Model Workspace.

2.2 Protein Modeling

Protein structures of IgE were modeled in Swiss-Model Workspace (URL: http://swissmodel.expasy.org/workspace/) using the Protocol devised by Bordoli et al. [36] which consists of the following steps (See Fig 1): Step 1: The target sequence (IgE VH region mRNA) was first examined by submitting its FASTA format or UniProt Accession Code in Sequence Features Scan session (with default settings), found under Tools. Step 2: Suitable template (s) for building homology model(s) was identified in Template Identification session (with default options) under Tools. Step 3: Most identical template spanning one or more domains of the target with least e-value was selected from the resulting hit list (a condensed graphical overview of template coverage with respect to domain boundaries with underlying target-template alignment and SWISS-MODEL template library (SMTL)). Step 4: Target-template sequence identity was considered for choice of modeling modes. (Automated mode : >50% identity, Alignment mode: 50%-30% identity, Project mode: <30% identity)

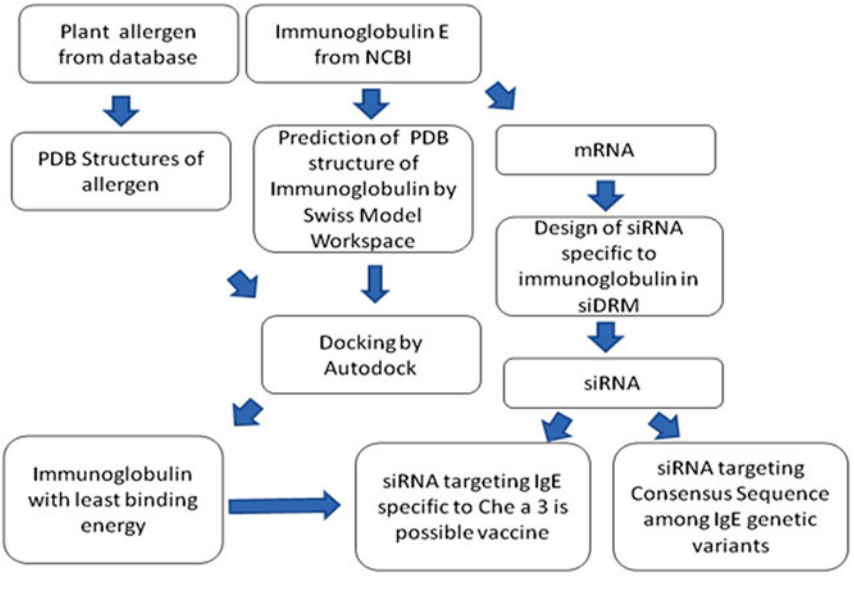


Fig. 1. The chronology of methods is summarized in this figure

Step 5: Modeling was done in the respective mode and model coordinates were output in PDB file format. Step 6: Quality of the modeled structure was estimated in 'Structure Assessment session' under 'Tools,' to identify incorrect regions.

2.3 Collection of Allergen Sequences

mRNA sequences of plant food and air borne allergens were retrieved from the public database Food Allergy Research and Resource Program (FARRP) (`http://www.allergenonline.org/`) [37] (See Supplementary1) by filtering based on the text terms Aero plant and Food plant for type of allergens. PDB structures were obtained from Swiss-Prot [38] (See Fig. 2; Supplementary2).

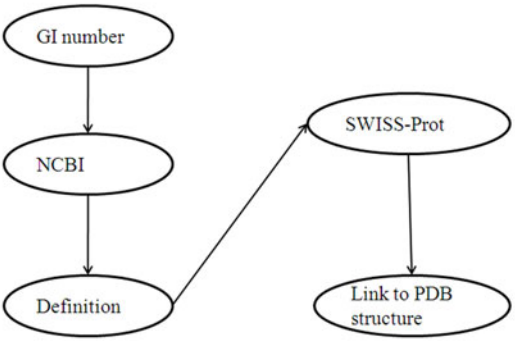


Fig. 2. Steps involved in fetching structural information of allergens

2.4 Docking in AutoDock

The PDB structures of the ligand (*Che a 3*) and receptor (IgE VH region) proteins were used after the addition of polar hydrogen atoms. The ligand protein of *Che a 3* is composed of 4 chains (See Fig. 3) with 2068 non polar hydrogen bonds, 1594 rotatable bonds and a torsion degree of freedom of 1224. The protein was split into individual chains and each chain was docked to each receptor of interest. Each chain was made rigid by making all rotatable bonds non-rotatable. A binding site detection program, QSiteFinder was used to identify the binding sites in receptor protein which were ranked based on binding energy. Residues in the top ranked binding site were made flexible in receptor. Mass-centered grid maps were generated with 0.75Å spacing [39] in AutoGrid program for the whole protein receptor. The best fitted conformation of the ligand was identified in AutoDock [31],[32],[33] where each chain of *Che a 3* (ligand) was docked with a total of 31 immunoglobulins (receptors) to find the IgE specific to that allergen. Each docking was performed in 50 trials (runs) (See Table 2 and Fig. 4).

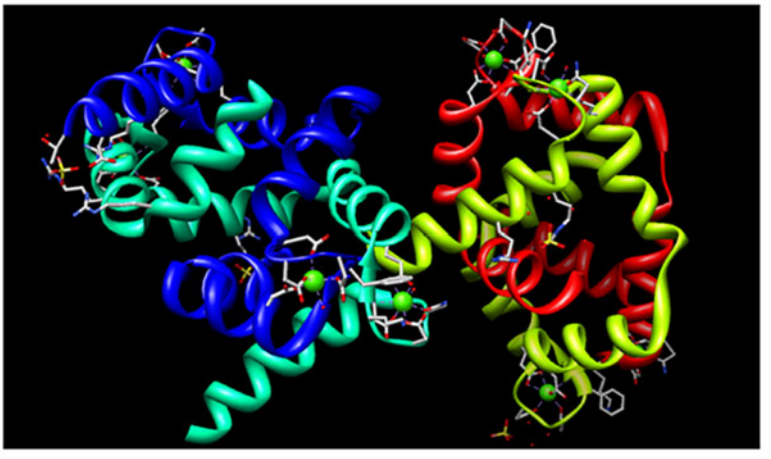


Fig. 3. *Che a 3* with each of its chain depicted in different colors: Chain A-Blue, Chain B- Cyan, Chain C- Green, Chain D-Red

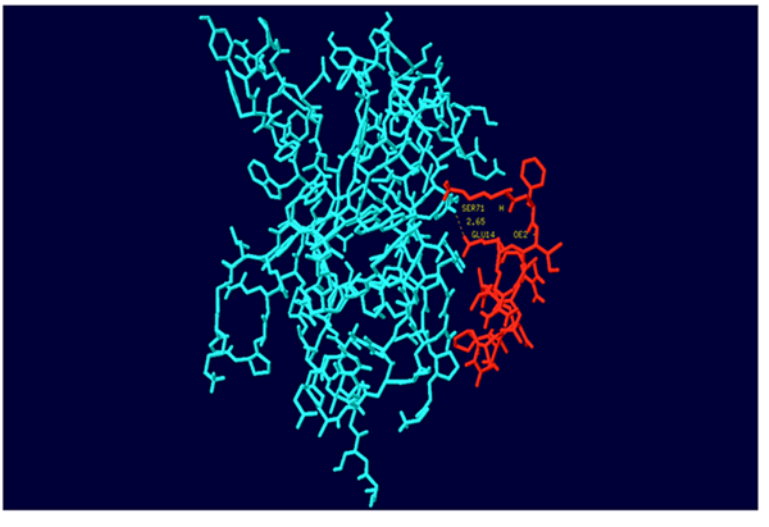


Fig. 4. The IgE, RH (represented in Cyan) bound to Che a 3 A chain (represented in Red) with a hydrogen bond of length 2.65Å between Serine 71 and Glutamine 14 respectively

2.5 siRNA Design Specific to IgE

Each immunoglobulin mRNA sequence was input into the online siRNA designing tool, siDRM [30] (URL: `http://sirecords.umn.edu/siDRM/`) with all default parameters to design a siRNA specific to each immunoglobulin. siRNA for all the genetic variants of IgE as a whole was designed based on the protocol devised by Birmingham *et al.*[40].

3 Results

3.1 Allergen Database

Complete record of 432 aero and 335 food allergens were fetched from FARRP (Food Allergy Research and Resource Program) database dated 4 November 2008. Each record is defined by the fields of species name and common name of the plant source, name of the allergen assigned by International Union of Immunological Societies (IUIS), Gene Identification Number and sequence length. The PDB structures thus obtained were validated for their accuracy as suggested by Kosloff and Kolodny [41] that only the structures with >70% sequence identity are similar. There were a total of 24 aero and 38 food allergens meeting this requirement (See Supplementary2).

3.2 Immunoglobulin E heavy Chain Variable Region

Search for IgE heavy chain variable region in Map Viewer [42] database of NCBI shed light on its respective chromosomal regions in Human genome. The region of interest was centered at Chromosomes 14q, 16p and 21p.

3.3 Modeling Protein Structures

The structures of IgE VH region were modeled using SWISS-MODEL workspace [43] as specified by Bordoli *et al.* [36]. The result from Sequence Feature annotation session was contributed by three individual tools, InterPro, domain scan tool, PsiPred, secondary structure prediction tool and DisoPred, disorder (Flexible, dynamic regions that can be partially or completely extended in solution) prediction tool to analyze features of the target sequence (See Supplementary3). The results of InterPro domain scan revealed that all the proteins were composed of immunoglobulin-like Domain and immunoglobulin V-set domain. And, the results of DisoPred indicated prevalence of disorderliness in 4-10 residues in the protein segment ranging from 8th to 15th position (See Supplemntary4). Also, the results from PsiPred reveals overrepresentation of extended β-sheets among secondary structures in 14 IgEs but 15 of them had an equal contribution from coil and sheet structures. And, four IgEs were more coiled in structure. Template Identification session facilitated selection of suitable templates for protein structure modeling (See Supplementary5).

The template with highest identity and least e-value was chosen for modeling. The template identities varied from 60-65% and Automated mode was chosen for modeling. The result page of SWISS-Model workspace [36],[43],[44] include energy profiles from ANOLEA statistical potential [45], GROMOS force field [46] along with percentage sequence identity between target and template (to build the model). ANOLEA statistical potential is to analyze the packing quality of predicted models which is graphically presented with y-axis representing energy for each amino acid in the protein chain (See Supplementary6). The green spikes (negative values) represent favorable energy environment whereas the red spikes

(positive values) are unfavorable energy environment for a given amino acid. While, GROMOS is used for analysis of conformations obtained by computer simulations and the results are similar to ANOLEA with green spikes (negative values) represent favorable energy environment and the red spikes (positive values) represent unfavorable energy environment for a given amino acid.

3.4 AutoDock

Each chain of the aero plant allergen *Che a 3* was docked to each of 31 immunoglobulins of interest. It is clear from the results that the IgE clone RH has least binding energy when docked to chain A of *Che a 3* (See Table 1; Supplementary7).

Table 1. Immunoglobulin clones with least six binding energies

Ligand Protein (Chain)	Receptor protein	Overall Binding Energy (Kcal/mol)	Ref R.M.S (Å)	Running time (Hours)
A	Homo sapiens clone BG immunoglobulin E variable region	8.77	109.77	6hrs 25min 15.43 sec
A	Homo sapiens clone PO immunoglobulin E variable region	8.76	127.74	7hrs 03min 05.04sec
A	Homo sapiens clone PT immunoglobulin E variable region	3.97	43.65	8hrs 23min 17.76sec
A	Homo sapiens clone RC immunoglobulin E variable region	5.9	77.72	9hrs 33min 49.39sec
A	Homo sapiens clone RD immunoglobulin E variable region	6.32	122.35	15hrs 46min 41.87sec
A	Homo sapiens clone RH immunoglobulin E variable region	0.14	42.21	10hrs 35min 52.77sec

Table 2. siRNAs designed (in siDRM [12]) to target IgE heavy chain variable region

Immunoglobulin	siRNA	GC content	Location	Off targets
Homo sapiens Clone RH immunoglobulin E variable region	CCGGUUCACCAUCUCCAGA	57%	198-216	Yes
Consensus Sequence among variable regions of Immunoglobulin E clones	CCGAUUCACCAUCUCCAGA	52%	197-215	No

3.5 siRNA Specific to Immunoglobulins

siRNA against the immunoglobulin specific to *Che a 3* was designed with the tool siDRM [30]. (See Table 2). Possibility for inhibition of non targeted transcripts by the designed siRNA was monitored by checking the homology of the respective siRNA with other transcripts. For each siRNA being designed, it was checked for the following possibilities in siDRM [30] : a) Homology to the whole transcript (5'UTR or CDS or 3'UTR), b) Homology of its subsequence by excluding last two positions to the 3'UTR region of another transcript, c) Homology of position 2-8 (seed region) to the 3'UTR region of another transcript, or d) Homology of position 2-8 (seed region) to the 3'UTR region of another transcript and the homologous region is followed by four consecutive mismatches.

4 Discussion

The variable region in IgE is most often specific for an allergen and is different in each repertoire produced by B cells as proved by Xu and Davis [47] in their experiments on transgenic mouse with diverse variable region in heavy chain and restricted lambda light chains, which concluded that most of the antibody specificity was generated by the molecular diversity specified by heavy chain.

This study focuses on the approach of designing a siRNA based vaccine (targeting IgE heavy chain variable region) *in silico*. Though, siRNA based vaccines in allergen immunotherapy are prevalent [16],[48] it will be a new strategy to focus on IgE heavy chain variable region as a vaccine target. A very recent research about development of IgE-based and allergen specific gene vaccine for food allergy pioneered by Behnecke *et al.* [49] further supports this viewpoint. siRNA is designed for the IgE clone specific to a particular allergen after docking analysis. In this study, the IgE specific for aero allergen from *Chenopodium album, Che a 3,* is analyzed based on their least binding energy obtained by docking. *Che a 3* is a tetramer with 86 residues in each chain (See Fig. 3), 2068 non-polar hydrogen bonds, 163 aromatic bonds, 1594 rotatable bonds and 1224 torsion degrees of freedom. But since AutoDock [31],[32],[33] docking tool is optimized only for 2048 molecules, there was a need to perform chain-wise docking to IgE.

The interaction between two proteins typically involves binding between specific domains [50]. Hence, the domain based docking can give a greater insight into the interaction between allergen and immunoglobulin. Since each domain is enclosed by individual chains, *Che a 3* was split into distinct chains in SWISS PDB viewer [51]. While proteins can be docked by any of the two approaches, blind and focused dockings. In blind docking, the receptor protein is docked without knowledge of its binding site with Mass-centered grid box. In focused docking, grid box covering the predicted ligand binding sites on receptor protein is defined. Comparing both, blind docking has greater efficacy than focused [52],[53].

In this research, the protein is made rigid except for the residues in top ranking binding site predicted by QSiteFinder [30] and mass-centered grid box. The ligand protein was made rigid and receptor protein was docked in AutoDock

[31],[32],[33] docking tool. As stated earlier, each of the 4 chains of allergen were docked with individual immunoglobulin. From the docking results, (See Table 1; Supplementary6) Chain A and RH IgE VH region exhibited least binding energy of 0.14 Kcal/mol. In RH-Chain A complex, residue GLU 14 in *Che a 3* and SER 71 in IgE form a hydrogen bond between them with the latter as donor and former as acceptor of electrons (See Fig. 4). To explain this phenomenon, we explore the modeled structure of IgE, RH clone. The disorderliness of RH was the least (See Suplementary3). The disordered regions in a protein, are capable of binding their partners with both high specificity and low affinity [54], hence, though the interactions are specific, they are prone to be unstable, which explains the increase in binding energies with increase in disorderliness. On the other hand, the immunoglobulin RM exhibited the third least disorderliness after RG and RH but requires highest energy for binding.

This annotates contribution of other factors apart from disorderliness for increase in binding energy. Structural and energetic analysis of changes during binding process construes that factors like small perturbations on protein structure, hydrogen bonding, buried surface area; shape complementarity and cooperativity between proteins, each have an effect on binding [55]. Adequacy of binding is also determined by protein dynamics, solvation potential, amino acid composition, conservation, electrostatics and hydrophobicity [56]. Residue SER 71 of the IgE RH lies within the sequence containing heterodimer interface (See Supplementary8). While, on the other hand, residue GLU 14 in *Che a 3* was significant too, since, it was nearly conserved among other Polcalcin containing plant allergens and occasionally subjected to homologous exchange with aspartic acid [57]. Hence, it is a good site to induce cross reactivity among Polcalcins. As deduced earlier, the IgE clone RH can be specific to *Che a 3* and siRNA designed against it can be formulated as a vaccine. According to the protocol [40], a siRNA is efficient only if its GC content lies within the optimal range of 30-64%.

Fortunately, both the siRNA targeting RH and consensus sequence of IgE repository had an optimal GC content of 57% and 52% respectively. Also, siRNA candidates must lack the motif GTCCTTCAA correlated with Interferon induction leading to non specificity [58]. There were no siRNAs with more than 6 consecutive Gs or Cs which might have rendered them ineffective [39] (See Supplementary9). siRNAs must be unconserved across multiple organisms. Hence were cross-checked among 62 seed siRNA regions identified by Lewis et al [59].

The resulting siRNA targeting RH heavy chain variable region is predicted to exhibit Off-targeting in siDRM [30]. Furthermore, siRNAs complementary to each of 31 genetic variants were designed (See Supplementary9). Surprisingly, 18 out of 31 variants exhibited Off-targeting while 3 had no siRNA complementary to them. Unique siRNAs and those capable of Off-targeting were examined using YMF [60],[61],[62] and FindExplanators [63] (URL: http://bio.cs.washington .edu/software.html) for motif discovery and selection of significant motifs respectively. No significant motifs were deduced from Off-targeting siRNAs while unique siRNAs were overrepresented with the motif, CGAUUCAC (Z-Score:

268.95, Position: 2-10) which may be significant for their uniqueness. Moreover, a consensus sequence,

[GAGGTGCAGCTGCTGGAGTCTGGGGGAGGCGTGGTGAAGCCTGGGG
GGTCCCTGAGACTCTCCTGTGCAGCCTCTGGATTCACCTTCAGTAGTT
ATTACATGAACTGGGTCCGCCAGGCTCCAGGGAAGGGGCTGGAGTGG
GTGG-AGTTATTATTAATAATGGTAGTAGTAAAAACTACGCCGACTCCG
TGAAGGGCCGATTCACCATCTCCAGAGACAACTCCAAGAACACCCTGT
ACCTGCAGATGAACAGCCTGAGAGCCGAGGACACGGCCGTCTATTAC
TGTGCGAGAGGAGG-GGTGG-GATGCTTTTGACTACTGC-ACTAGGGC
GGCGGCGGCGGCGTCGC]

was derived from 294 different clones [64],[65],[66],[67],[68],[69] of IgE VH region with the help of the alignment tool CINEMA [70] and a targeting siRNA was designed (See Table 2) which is found to be unique. The gene sequences were retrieved from NCBI with key words Immunoglobulin E variable region and Immunoglobulin heavy chain variable region and the duplicates were removed.

5 Conclusion

The siRNA can be injected in B-Cells to validate their silencing activity prior development of allergy vaccines for maximized vigor. In silico screening strategy preceding laboratory investigations can lead to great breakthroughs and is an economical method to accelerate the pace of researches in developing effectual immunotherapies for allergy. The siRNA, CCG**A**UUCACCAUCUCCAGA targets the consensus sequence at a region with 51-89% conservation among the genetic variants. Furthermore, it has no off-targeting and can serve as a vaccine for any allergy irrespective of the source of allergen. On the contrast, the siRNA, CCG**G**UUCACCAUCUCCAGA complimentary to the IgE (Clone RH) specific to *Che a 3* exhibits 'Off-targeting'. Both of the RNAs are complimentary to similar positions in their target sequences (See Table 2) and differ only by a nucleotide at their 4th position. Thus, proves the reliability of the siRNA against IgE clone RH as a possible vaccine.

Supplementary Documents

Supplementary 1: Databases with Information about allergens:
`https://docs.google.com/document/edit?id=1Fbq3B-luz1YbrxZ8lxd_RWfzjjNWvx55WRSmK7kVhC4\&hl=en#`

Supplementary 2: List of Allergens with PDB Structures:
`https://docs.google.com/document/edit?id=11s1vQ-SwMru52sc--3IqcZ9WHA8_mWMhWJ3rC_o4zHE\&hl=en#`

Supplementary 3: Sequence Feature annotation in SWISS-MODEL workspace:
`https://docs.google.com/present/edit?id=0AeEKk700XoYVZGNrYnRtM3JfMTY2ZGdtYnpnZzI\&hl=en`

Supplementary 4: Disorderliness of predicted Model given by DisoPred:
`https://docs.google.com/document/edit?id=1JhGF86z2xG15b-pmze6BqXtXB62PvNKuQxYNxi5UnF0\&hl=en#`

Supplementary 5: Results of Template identification session SWISS-Model Workspace:
`https://docs.google.com/document/edit?id=1K84rljE2fhIYT6YMOVvo31sbzLzkAtLZ4wxnKOWqCbo\&hl=en#`

Supplementary 6: ANOLEA statistical potential:
`https://docs.google.com/present/edit?id=0AeEKk700XoYVZGNrYnRtM3JfMTYwY2dmdDNxamM\&hl=en`

Supplementary 7: Autodock results:
`https://docs.google.com/document/edit?id=1gEJWL5MjRdnMEqQmmVvpZSGP_3tQj8mqix6okmTfXF0\&hl=en#`

Supplementary 8: Result page of Conserved Domain Search in NCBI:
`https://docs.google.com/present/edit?id=0AeEKk700XoYVZGNrYnRtM3JfMTYzZzU2dmdtY3o\&hl=en`

Supplementary 9: siRNA targeting IgE heavy chain variable region:
`https://docs.google.com/document/edit?id=1PycmdhiX_195sKf-wuO-P6WU7n-_iHfK83KATT3z6Ww\&hl=en#`

References

1. Zhang, G., Khan, A., Srinivasan, K., Heiny, A., Lee, K., Kwoh, C., et al.: Hotspot Hunter: a computational system for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes. BMC Bioinformatics 9 (suppl. 1:S19) (2008)
2. Holgate, S.T., Polosa, R.: Treatment strategies for allergy and asthma. Nat. Rev. Immunol. 8(3), 218–230 (2008)
3. Bisgaard, H., Hermansen, M.N., Loland, L., Halkjaer, L.B., Buchvald, F.: Intermittent inhaled corticosteroids in infants with episodic wheezing. The New England Journal of Medicine 354, 1998–2005 (2006)
4. Guilbert, T., Morgan, W., Zeiger, R., Mauger, D., Boehmer, S., Szefler, S., et al.: Long-term inhaled corticosteroids in preschool children at high risk for asthma. The New England Journal of Medicine 354, 1985–1997 (2006)
5. Harrison, T.W., Oborne, J., Newton, S., Tattersfield, A.E.: Doubling the dose of inhaled corticosteroid to prevent asthma exacerbations: randomised controlled trial. Lancet. 363, 271–275 (2004)
6. MacLennan, S., Barbara, J.A.J.: Risks and side effects of therapy with plasma and plasma fractions. Best Practice and Research Clinical Haematology 19(1), 169–189 (2006)
7. Hindmarch, I., Parrott, A.: A repeated dose comparison of the side effects of five antihistamines on objective assessments of psychomotor performance, central nervous system arousal and subjective appraisals of sleep and early morning behaviour. Arzneimittelforschung 28(3), 483–486 (1978)

8. Cuvillo, A., Mullol, J., Bartra, J., Dvila, I., Juregui, I., Montoro, J., et al.: Comparative pharmacology of the H1 antihistamines. Journal of Investigational Allergology and Clinical Immunology 16(1), 3–12 (2006)
9. Kindt, T.J., Goldsby, R.A., Osborne, B.A., Kuby, J.: Kuby immunology, 6th edn. W.H. Freeman, New York (2006)
10. Gawchik, S., Saccar, C.: Pollinex Quattro Tree: allergy vaccine. Expert Opinion on Biological Therapy 9(3), 377–382 (2009)
11. Carns, J., Robinson, D.: New strategies for allergen immunotherapy. Recent Patents on Inflammation and Allergy Drug Discovery 2(2), 92–101 (2008)
12. Popescu, F.-D.: Antisense- and RNA interference-based therapeutic strategies in allergy. Journal of Cellular and Molecular Medicine 9(4), 840–853 (2005)
13. Kurreck, J.: Antisense technologies. Improvement through novel chemical modifications. European Journal of Biochemistry 270(8), 1628–1644 (2003)
14. Wadhwa, R.K.S., Miyagishi, M., Taira, K.: Know-how of RNA interference and its applications in research and therapy. Mutatation Research 567(1), 71–84 (2004)
15. Bagasra, O., Prilliman, K.R.: RNA interference: The molecular immune system. Journal of Molecular Histology 35(6), 545–553 (2004)
16. Suzuki, M., Zheng, X., Zhang, X., Li, M., Vladau, C., Ichim, T.E., et al.: Novel Vaccination for Allergy through Gene Silencing of CD40 Using Small Interfering RNA. The Journal of Immunology 180, 8461–8469 (2008)
17. Taylor, P.A., Friedman, T.M., Korngold, R., Noelle, R.J., Blazar, B.R.: Tolerance induction of alloreactive T cells via ex vivo blockade of the CD40: CD40L costimulatory pathway results in the generation of a potent immune regulatory cell. Blood 99, 4601–4609 (2002)
18. Taylor, J.J., Mohrs, M., Pearce, E.J.: Regulatory T cell responses develop in parallel to Th responses and control the magnitude and phenotype of the Th effector population. Journal of Immunology 176, 5839–5847 (2006)
19. Thornton, C., Holloway, J., Popplewell, E., Shute, J., Boughton, J., Warner, J.: Fetal exposure to intact immunoglobulin E occurs via the gastrointestinal tract. Clinical and Experimental Allergy 33, 306–311 (2003)
20. Poole, J.A., Meng, J., Reff, M., Spellman, M.C., Rosenwasser, L.J.: Anti-CD23 monoclonal antibody, lumiliximab, inhibited allergen-induced responses in antigen-presenting cells and T cells from atopic subjects. Journal of Allergy and Clinical Immunology 116, 780–788 (2005)
21. Holgate, S., Casale, T., Wenzel, S., Bousquet, J., Deniz, Y., Reisner, C.: The anti-inflammatory effects of omalizumab confirm the central role of IgE in allergic inflammation. Journal of Allergy and Clinical Immunology 115, 459–465 (2005)
22. Duarte, J., Deshpande, P., Guiyedi, V., Mcheri, S., Fesel, C., Cazenave, P.-A., et al.: Total and functional parasite specific IgE responses in Plasmodium falciparum-infected patients exhibiting different clinical status. Malaria Journal 6(1) (2007)
23. Erb, K.J.: Helminths, allergic disorders and IgE-mediated immune responses: Where do we stand? European Journal of Immunology 37, 1170–1173 (2007)
24. Naohiro Watanabe, F.B., Korenaga, M.: IgE: a question of protective immunity in Trichinella spiralis infection. Trends in Parasitology 21(4), 175–178 (2005)
25. Karagiannis, S.N., Wang, Q., East, N., Burke, F., Riffard, S., Bracher, M.G., et al.: Activity of human monocytes in IgE antibodydependent surveillance and killing of ovarian tumor cells. European Journal of Immunology 33, 1030–1040 (2003)
26. Alt, F.W., Yancopoulos, G.D., Blackwell, T.K.: Ordered rearrangement of immunoglobulin heavy chain variable region segments. EMBO Journal 3, 1209–1219 (1984)

27. Alt, F., Oltz, E., Young, F., Gorman, J., Taccioli, G., Chen, J.: VDJ recombination. Immunology Today 13, 306–314 (1992)
28. Stephanie Culler, T.R.H., Glassy, M., Chau, P.C.: Canonical Structure Repertoire of the Antigen-binding Site of Immunoglobulins Suggests Strong Geometrical Restrictions Associated to the Mechanism of Immune Recognition. Journal of Molecular Biology 254(3-1), 497–504 (1995)
29. Chothia, C., Lesk, A.M.: Canonical structures for the hypervariable regions of immunoglobulins. Journal of Molecular Biology 196, 901–917 (1987)
30. Gong, W., Ren, Y., Zhou, H., Wang, Y., Kang, S., Li, T.: siDRM: an effective and generally applicable online siRNA design tool. Bioinformatics 24(20), 2405–2406 (2008)
31. Goodsell, D., Olson, A.: Automated docking of substrates to proteins by simulated annealing. Proteins 8, 195–202 (1990)
32. Morris, G., Goodsell, D., Huey, R., Olson, A.: Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. Journal of Computer-Aided Molecular Design 10, 293–304 (1996)
33. Morris, G., Goodsell, D., Halliday, R., Huey, R., Hart, W., Belew, R., et al.: Automated docking using a Lamarckian genetic algorithm and empirical binding free energy function. Journal of Computational Chemistry 19, 1639–1662 (1998)
34. Alasdair, T.R.L., Richard, M.J.: Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 29(9), 1908–1916 (2005)
35. Coker, H.A., Harries, H.E., Banfield, G.K., Carr, V.A., Durham, S.R., Chevretton, E., et al.: Biased use of VH5 IgE-positive B cells in the nasal mucosa in allergic rhinitis. Journal of Allergy and Clinical Immunology 116(2), 445–452 (2005)
36. Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., Schwede, T.: Protein structure homology modeling using SWISS-MODEL workspace. Nature Protocols 4(1-3) (2009)
37. Hilemana, R.E., Silvanovicha, A., Goodmana, R.E., Ricea, E.A., Holleschaka, G., Astwooda, J.D., et al.: Bioinformatic Methods for Allergenicity Assessment Using a Comprehensive Allergen Database. International Archives of Allergy and Immunology 128(4) (2002)
38. Bairoch, A., Apweiler, R.: The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Research 25(1), 31–36 (1997)
39. Kroemer, R.T., Hecht, P., Guessregen, S., Liedl, K.R.: Improving the Predictive Quality of CoMFA Models. In: 3D QSAR in Drug Design, pp. 41–56. Springer, Netherlands (2006)
40. Birmingham, A., Anderson, E., Sullivan, K., Reynolds, A., Boese, Q., Leake, D., et al.: A protocol for designing siRNAs with high functionality and specificity. Nature Protocols 2(9), 2068–2078 (2007)
41. Kosloff, M., Kolodny, R.: Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins 71(2), 891–902 (2008)
42. Arnold, K., Bordoli, L., Kopp, J., Schwede, T.: The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22, 195–201 (2006)
43. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., et al.: Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 35(D5-D12) (2007)
44. Schwede, T., Kopp, J., Guex, N., Peitsch, M.: SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Research 31, 3381–3385 (2003)
45. Guex, N., Peitsch, M.C.: SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling. Electrophoresis 18, 2714–2723 (1997)

46. Melo, F., Feytmans, E.: Assessing protein structures with a non-local atomic interaction energy. Journal of Molecular Biology 277, 1141–1152 (1998)
47. Xu, J., Davis, M.: Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. Immunity 13, 37–45 (2000)
48. Ichim, T., Popov, I., Riordan, N., Izadi, H., Zhong, Z., Yijian, L., et al.: A novel method of modifying immune responses by vaccination with lipiodol-siRNA mixtures. Journal of Translational Medicine 3(4) (2006)
49. Behnecke, A., Chen, L., Saxon, A., Zhang, K.: Development of an IgE-based Allergen Gene Vaccine for Severe Food Allergy. Journal of Allergy and Clinical Immunology 121(2), S212 (2008)
50. Katia, S.G., Raja, J., Elena, Z., Teresa, M.P.: Predicting domain-domain interactions using a parsimony approach. Genome Biology 7, R104 (2006)
51. Kaplan, W., Littlejohn, T.G.: Swiss-PDB Viewer (Deep View). Briefings in Bioinformatics 2(2), 195–197 (2001)
52. Huang, B., Schroeder, M.: Using protein binding site prediction to improve protein docking. Gene. 422(1-2), 14–21 (2008)
53. Ghersi, D., Sanchez, R.: Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites. Proteins: Structure, Function, and Bioinformatics 74(2), 417–424 (2009)
54. Dunker, A.K., Obradovic, Z.: The protein trinitylinking function and disorder. Nature Biotechnology 19, 805–806 (2001)
55. Reichmann, D., Rahata, O., Cohena, M., Neuvirtha, H., Schreiber, G.: The molecular architecture of protein-protein binding sites. Current Opinion in Structural Biology 17(1), 67–76 (2007)
56. Horn, J.R., Kraybill, B., Petro, E.J., Coales, S.J., Morrow, J.A., Hamuro, Y., et al.: The Role of Protein Dynamics in Increasing Binding Affinity for an Engineered ProteinProtein Interaction Established by H/D Exchange Mass Spectrometry. Biochemistry 45(58), 8488–8498 (2006)
57. Verdino, P., Barderas, R., Villalba, M., Westritschnig, K., Valenta, R., Rodriguez, R., et al.: Three-Dimensional Structure of the Cross-Reactive Pollen Allergen Che a 3: Visualizing Cross-Reactivity on the molecular Surfaces of Weed, Grass, and Tree Pollen Alergens. The Journal of Immunology (2007)
58. Hornung, V., Guenthner-Biller, M., Bourquin, C., Ablasser, A., Schlee, M., Uematsu, S., et al.: Sequence-specific potent induction of IFN-alpha by short interfering RNA in plasmacytoid dendritic cells through TLR7. Nature Medicine 2005(3), 263–270 (2005)
59. Lewis, B., Burge, C., Bartel, D.: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120(1), 15–20 (2005)
60. Sinha, S., Tompa, M.: YMF: a Program for Discovery of Novel Transcription Factor Binding Sites by Statistical Overrepresentation. Nucleic Acids Research 31(13), 3586–3588 (2003)
61. Sinha, S., Tompa, M.: Discovery of Novel Transcription Factor Binding Sites by Statistical Overrepresentation. Nucleic Acids Research 30(24), 5549–5560 (2002)
62. Sinha, S., Tompa, M.: A Statistical Method for Finding Transcription Factor Binding Sites. In: Eighth International Conference on Intelligent Systems for Molecular Biology, San Diego, CA, pp. 344–354 (August 2000)
63. Blanchette, M., Sinha, S.: Separating real motifs from their artifacts. Bioinformatics 17, S30–S38 (2001)

64. Andrasson, U., Flicker, S., Lindstedt, M., Valenta, R., Greiff, L., Korsgren, M., et al.: The human IgE-encoding transcriptome to assess antibody repertoires and repertoire evolution. Journal of Molecular Biology 362(2), 212–227 (2006)
65. Dahlke, I., Nott, D., Ruhno, J., Sewell, W., Collins, A.: Antigen selection in the IgE response of allergic and nonallergic individuals. Journal of Allergy and Clinical Immunology 117(6), 1477–1483 (2006)
66. Galibert, L., van Dooren, J., Durand, I., Rousset, F., Jefferis, R., Banchereau, J., et al.: Anti-CD40 plus interleukin-4-activated human naive B cell lines express unmutated immunoglobulin genes with intraclonal heavy chain isotype variability. European Journal of Immunology 25(3), 733–777 (1995)
67. Davies, J.M., O'Hehir, R.E.: Immunogenetic characteristics of immunoglobulin E in allergic disease. Clinical & Experimental Allergy 38(4), 566–578 (2008)
68. Snow, R., Djukanovic, R., Stevenson, F.: Analysis of immunoglobulin E VH transcripts in a bronchial biopsy of an asthmatic patient confirms bias towards VH5, and indicates local clonal expansion, somatic mutation and isotype switch events. Immunology 98(4), 646–651 (1999)
69. van der Stoep, N., van der Linden, J., Logtenberg, T.: Molecular evolution of the human immunoglobulin E response: high incidence of shared mutations and clonal relatedness among epsilon VH5 transcripts from three unrelated patients with atopic dermatitis. Journal of Experimental Medicine 177(1), 99–107 (1993)
70. Pettifer, S.R., Sinnott, J.R., Attwood, T.K.: UTOPIAUser-Friendly Tools for Operating Informatics Applications. Comparative and Functional Genomics 5(1), 56–60 (2004)

Unsupervised Algorithms for Population Classification and Ancestry Informative Marker Selection

Apaporn Rodpan[1], Pongsakorn Wangkumhang[2], Anunchai Assawamakin[2], Santitham Prom-on[3], and Sissades Tongsima[2,*]

[1] Bioinformatics and Systems Biology Program, King Mongkut University of Technology Thonburi, Bangkok, Thailand
[2] Biostatistics and Informatics Laboratory, Genome Institute, National Center for Genetic Engineering and Biotechnology, Thailand Science Park, Pathumthani
[3] Computer Engineering, King Mongkut University of Technology Thonburi, Bangkok, Thailand
sissades@biotec.or.th

Abstract. Single Nucleotide Polymorphisms (SNPs) can be used to identify the differences among populations. However, for high-level organisms, there are numerous number of SNPs distributed throughout entire of the genomes. Animal breeders can make use of these genetic markers to different subpopulations. For economical purpose, finding a minimum number of SNPs that can accurately identify different breeds is needed. In this paper, given a set of SNP genotyping samples, without knowing what breed a sample belong to (unlabeled samples), we developed a framework to classify these samples into different animal groups (breeds) based on their genotyping profiles. The proposed framework further identifies a small set of SNPs, called ancestry informative markers (AIMs) that can accurately classify these samples to these groups. The proposed framework adopted the Principal Component Analysis (PCA) technique, and Student's t-test, to cluster unlabeled genotype data and determine AIMs, respectively. This unsupervised approach can avoid potential ascertainment biases due to mistakenly label some samples or having unlabeled data to be classified.

Keywords: AIMs, ancestry informative markers, SNPs, Principal Component Analysis, Student's t-test, population structure.

1 Introduction

Single Nucleotide Polymorphisms (SNPs) are the most common form of genetic variation in all organisms. SNPs have been adopted in various fields of studies including forensics [1], disease association studies [2], assignment of individuals to populations [3,4], and studies of dispersal, wildlife management and livestock breeding management. Since collection of all SNPs forms a unique genetic profile of each individual, they can be used to group different individuals at a population level, for

* Corresponding author.

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 208–216, 2010.

which not all SNPs are needed in the classification process. For this purpose, it is desirable to discover only small number of informative markers, called ancestry informative markers (AIMs), to be used in the population classification process.

In order to determine AIMs, SNPs that are specific to a certain population but not others have to be determined. To identify these informative SNPs, statistical techniques such as Student's t-tests [5,6] and F statistics [6] have been deployed. Screening of candidate SNPs was carried out with a greedy approach [7] or a ranking method [6]. A classification task, which involves supervised machine learning [6] and the use of parametric genetic model [8] was proposed to computationally verify these selected SNPs. However, as the accuracy of supervised classification techniques have to be assessed by the pre-defined sample labels (can also be called classes, population names or breeds in some contexts), the supervised machine learning techniques cannot address the problems of classifying unlabeled data. Moreover, sample labels are usually derived from phenotypes or traits, e.g., colors, height. This can be problematic, especially when performing AIM selection, if the wrong phenotypic assumption is made.

In this study, we propose a framework, which employs an unsupervised exploratory data analysis technique, called Principal Component Analysis (PCA) and a statistical hypothesis Student's t-test to identify AIMs. The proposed technique does not require genetic model and/or parameter, which are very difficult to estimate. The proposed combined framework can efficiently determine AIMs from assorted genotypic samples. Such AIMs can be used to determine the corresponding subpopulations with high accuracy.

2 Methods

The proposed combination technique involves two main processes: (1) informative marker selection (2) assignment of individuals to the corresponding subpopulations.

2.1 Informative Marker Selection

Genotype Data Conversion. Genotype information is categorical data and cannot to be analyzed by PCA directly; hence such information must be converted into numerical before performing PCA. Homozygous wild types, the homozygous alleles with the highest frequencies, are represented by '0'. Heterozygous alleles are represented by '1'. Finally, the homozygous variant types, the homozygous alleles with lower frequencies, are represented by '2'. We encode the missing values with the value '-1'. The zero mean input data matrix to PCA can be generated by, at each SNP locus, the values 0, 1, 2 and -1 will be subtracted with the locus mean value [9].

Determination of Population Structure and Number of Ancestry (K). Population structure and inferred ancestry numbers (K) were analyzed without the use of predefined population labels. PCA was employed in this process to reduce the dimensionality of the SNP data. Each sample containing large number of SNPs is transformed into a vector, which appears as a single data point in the PCA space. A group of samples with similar genotypic profiles will be shown as a conglomerate in this space. Since SNP genotype matrix can be very large owing to the number of input SNPs, using

traditional PCA covariance analysis can be computationally and memory intensive. Generally, the number of samples is much smaller than the number of SNP markers. Hence, Singular Value Decomposition (SVD) technique was employed to reduce the aforementioned complexities. This was done by finding a covariance matrix, XX^T, whose rank is as large as the number of samples. This is much smaller in size comparing to the SNP genotype matrix.

$$X = USV^T \tag{1}$$

When calculating covariance matrix,

$$XX^T = US^2U^T. \tag{2}$$

where U is a left unitary matrix,
S is a diagonal matrix with non-negative real numbers on diagonal,
V is a right unitary matrix.

Therefore, we can calculate the left unitary matrix using Equation 3.

$$V = S(1:r,1:r)^{-1} U(:,1:r)^T X \tag{3}$$

where r is the rank of matrix S
":" represents the continuity of the row or column indices of the matrix (Matlab notation).

The data can then be transformed further using Equation 4.

$$X_t = XV^T \tag{4}$$

where X_t is the transformed SNP data matrix.

Matrix S is used to calculate the degree of group heterogeneity (diversity of samples within the group), which was used in the termination step. It was found that the differences of eigenvalues in matrix S are related to the existence of population structure (potentially containing sub-populations). The rates of change of eigenvalues could thus be used as a termination condition. A constant or linear rate of change of eigenvalues indicates an absence of sub-populations in the dataset, while the non-linear indicates the existence of sub-populations.

The diagonal entries of matrix S are eigenvalues (λ_1, λ_2,..., λ_r) and they are ranked in a descending order and plotted (**Fig. 1**). A hundred percent coverage is defined by λ_1 (**a**) to λ (**b**). The value λ_J (**c**) is designated as the point where 90% coverage was reached. In order to determine the linearity of the rate of change of eigenvalues, λ_1 to λ_J are used. The included eigenvalues are sufficient to warrant the original data structure while eliminating the insignificant portions. A line (**d**) connecting λ_1 and λ_J can be drawn. Then, we calculated the mean of the distances from the line to each eigenvalue.

$$e = \frac{1}{J}\sqrt{\sum_{i=1}^{J}\left(\lambda_i - \hat{\lambda}_i\right)^2} \tag{5}$$

where e is a degree of heterogeneity and
$\hat{\lambda}_i$ is a point on the line corresponding to λ_1.

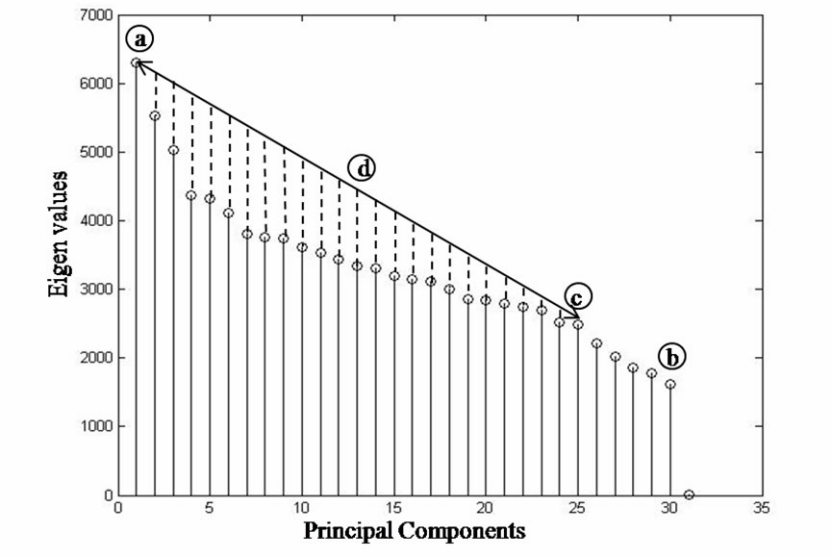


Fig. 1. Illustrates how heterogeneity (*e* value) is obtained

The threshold value of 0.15 was obtained from observations and could be applied to every set of genotype data with similar eigenvalues' rates of change as in **Fig. 1**. In other words, the *e* value that was smaller than 0.15 indicates the observed group has less chance of having sub-populations (i.e., they belong to the same population). If *e* is larger than or equal to 0.15, there should be two or more sup-populations in the group. Applying *k*-means clustering algorithm (k=2) on X_t always yields two groups. As a result, X_t will be clustered into X_1 and X_2. Equations 1 to 5 will be repeated until the termination condition is met. The processed results can be presented in an unbalanced binary tree with the resulting populations at the termination nodes. The numbers of leaf nodes indicate the inferred population number (K). **Fig. 2.** represents the flow chart of the aforementioned steps.

Identifying Ancestry Informative Markers. SNPs that are significantly correlated with one population, but not the others, can be used as the ancestry informative markers (AIMs) for that population. In this process, we chose informative markers using Student's t-test. Generally, t-test is used to investigate if the mean of the two classes are different (two-sample t-test). This can be achieved by calculating t statistic ratio between the difference of the mean value of each group and the variance of these two groups combined. In this case, t-test is used to indicate whether a SNP at a specific location is statistically different among the all populations. From K populations, we can identify a distinct polymorphism by calculating Equation 6.

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{S_{\bar{x}_i - \bar{x}_j}} \tag{6}$$

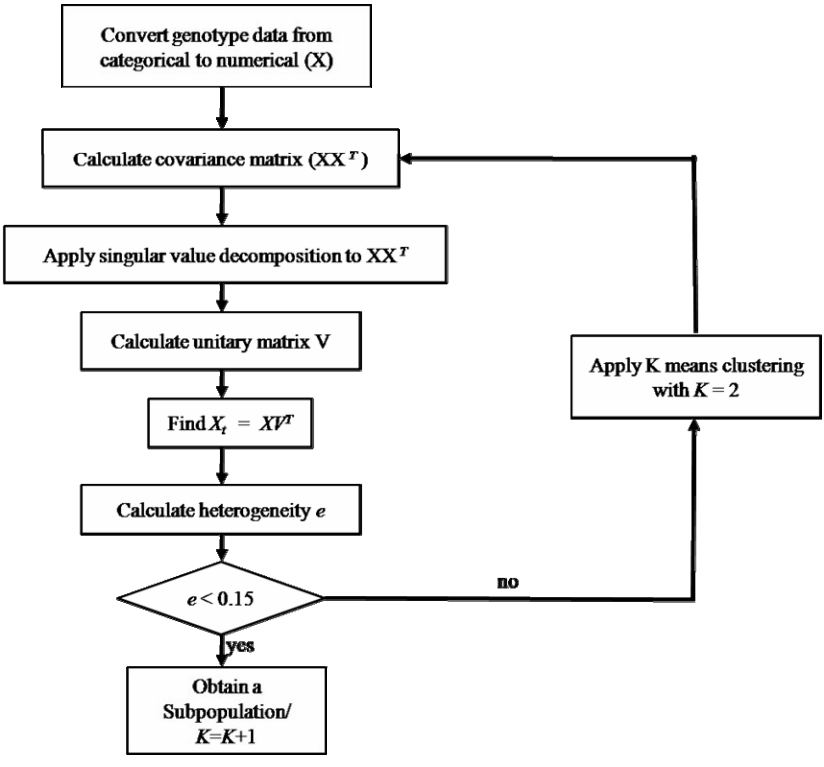


Fig. 2. Flow diagram of determining population structure and number of ancestry

$$\text{where } S_{\bar{x}_i - \bar{x}_j} = \sqrt{\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}} \tag{7}$$

Let $\mathbf{X_i}$ and $\mathbf{X_j}$ be SNP matrices containing samples that belong to populations *i* and *j* respectively, where $\boldsymbol{i} \neq \boldsymbol{j}$ and *i, j* = {1, 2,…, K}. Student's t-test can be calculated using Equation 6. The *p*-values obtained in this manner signify the differences of the mean values for each SNP. The *p*-values will be sorted in an ascending order. SNPs with *p*-values smaller than 10^{-5} will be selected as the informative markers that can differentiate populations *i* from *j*. The same selection process using t-test is performed over all possible pairs of populations *i* and *j*. At the end, the top rank informative markers from each pair of populations are union, resulting in the AIM set.

2.2 Assigning Samples to Sub-populations

The algorithm described in **Fig. 2.** can be used to classify individuals to different sub-populations because of the embedded clustering procedure used to split samples into two different groups. The clustering is done only when the observed group contain significant structure (does not meet the stopping criterion). In order to verify if AIMs are unique and have the power to classify each sample, we repeat the process in **Fig. 2.** and check if 1) number of ancestry (K) and 2) individual assignment results are the

same as performing this algorithm on the full dataset. In other words, these informative SNPs are expected to have the same power to discriminate the subpopulations.

Dataset

To verify the proposed algorithm, we performed the experiments on a large genotyping dataset containing multiple breeds of cattle. The bovine genotype data previously published in [10] were downloaded from http://www.animalgenome.org/bioinfo/resources/util/q_bovsnp.html. These samples are the major beef cattle and dairy cattle in many countries, including Thailand. The data composed of 9,239 SNPs obtained from Affymetrix GeneChip Bovine Mapping 10K SNP Kit from 230 individual bovines, accounting for 9 species (Table 1).

Table 1. Number of bovine samples from each breed. The breed name is abbreviated using three letters shown in the parentheses.

Breeds	Origins	Samples #
Charolais (CHL)	France	22
Santa Gertrudis (SGT)	USA	24
Jersey (JER)	England	28
Holstein (HOL)	Netherlands	33
Brahman (BRM)	India	25
Norwegian Red (NRC)	Norway	24
Hereford (HFD)	England	27
Limousin (LMS)	France	21
Angus (ANG)	Scotland	26

3 Results and Discussion

To determine AIMs for breed identification, Student's t-test was used. The t-test identifies the SNPs, which significantly varied between breeds but containing high allele frequency within the same breed. Our proposed model produced 44 AIMs (see Supplement at http://www4a.biotec.or.th/GI/tools/aims). We validated the effectiveness of 44 AIMs by repeating the classification process described in the method section to obtain the individual assignment accuracy. The result presented in **Table 2** shows that the individual assignment accuracy is 97.39% (224 out of 230 individuals were correctly classified).

We tested the effectiveness of this method by comparing with the results on the same bovine dataset published by DeNise et al. [8]. DeNise's approach was based on the difference of allele frequencies between cattle breeds. Their algorithm made use of predefined breed labels as the training datasets. The statistical tests were used to determine the difference in each allele frequency between a species and other populations, resulting in 45 AIMs. In their work, STRUCTURE [11], was used to infer the number of population K. STRUCTURE find the optimal K by estimating the posterior

Table 2. Individual assignment results from our proposed method using 44 AIMs

Breeds	Identify population by PCA and using 44 AIMs								
	1	2	3	4	5	6	7	8	9
CHL	-	-	-	22	-	-	-	-	-
SGT	-	20	-	4	-	-	-	-	-
JER	28	-	-	-	-	-	-	-	-
HOL	-	-	-	-	31	2	-	-	-
BRM	-	-	-	-	-	-	25	-	-
NRC	-	-	-	-	-	24	-	-	-
HFD	-	-	27	-	-	-	-	-	-
LMS	-	-	-	-	-	-	-	-	21
ANG	-	-	-	-	-	-	-	26	-

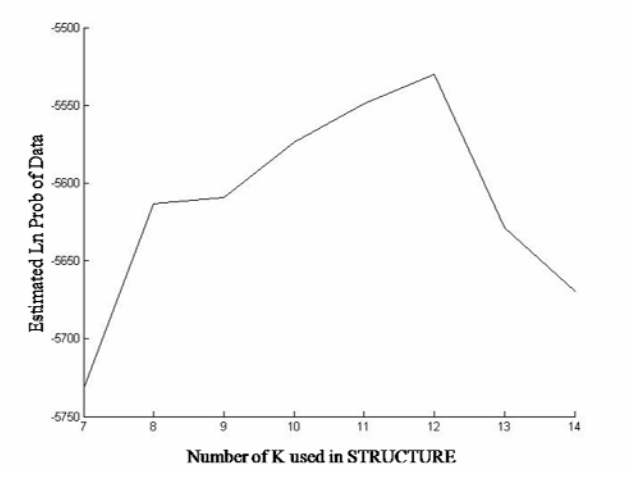


Fig. 3. DeNise's average *Estimated Ln Prob of data* for 45 markers. The inferred *K* value was identified as the highest value (K=12) of *Estimated Ln Prob of Data.*

probability of each K called *Estimated Ln Prob of data.* This value indicates the probability that a specific breed can be clustered into K ancestral groups, as shown in **Fig 3**. The highest value of *Estimated Ln Prob of Data* was -5529.94, obtained when K = 12.

We followed DeNise's method by using the program STRUCTURE and a set of parameters consisting of 45 markers, 230 samples, 20,000 burn-in periods, and 10,000 repeats. We also used Markov Chain in STRUCTURE to cluster by estimating the admixture coefficients (*Q* matrix) for each sample. Average *Q* values were then used as input in the place of sample genomes. High average *Q* values indicate high

probability that the samples pertained to the ancestry. We used this method to identify the sample strains. The results can be downloaded from our Supplementary data link (http://www4a.biotec.or.th/GI/tools/aims). The identification results of DeNise *et al*'s supervised method are summarized in **Table 3**. The accuracy identification (85.65%) was calculated from the number of individuals that were correctly classified.

Table 3. Individuals assignment to sub-population from DeNise et al.

Assignment of individuals to populations, using AIMs from DeNise et al.'s approach												
Breeds	1	2	3	4	5	6	7	8	9	10	11	12
CHL	1	-	-	-	9	2	-	-	10	-	-	-
SGT	-	-	-	-	-	1	2	-	-	-	-	21
JER	1	-	-	-	-	1	-	-	-	-	26	-
HOL	3	1	-	-	-	-	-	29	-	-	-	-
BRM	-	-	-	-	-	-	25	-	-	-	-	-
NRC	5	18	-	-	-	-	-	1	-	-	-	-
HFC	-	-	-	25	1	-	-	-	-	-	1	-
LMS	-	-	-	-	1	-	-	-	-	20	-	-
ANG	-	-	23	-	1	2	-	-	-	-	-	-

4 Conclusion

Our framework offers three main analyses: 1) find the population structure without predefined labels, 2) accurately determine the numbers of populations or sub-populations (K) within sample pool, and 3) assign individuals to their correct populations or breeds. The accuracy of assignment of individual bovine to their breeds was calculated from the major group of individuals that are accurately assigned, accounting for 97.39% or 224 individuals out of 230. Our approach can identify AIMs from unlabeled populations with high accuracy of population classification. However, application of this method is not restricted only to bovine breeds classification, it can also be applied to the classification of other species including ethnic groups of human.

Acknowledgments. This work was supported by the National Center for Genetic Engineering and Biotechnology of Thailand, School of Information Technology and School of Bioresources and Technology King Mongkut's University of Technology Thonburi. Anunchai Assawamakin was supported by BIOTEC postdoctoral grant.

References

1. Budowle, B., Van Daal, A.: Forensically Relevant SNP Classes. BioTechniques 44, 603–610 (2008)
2. Moore, J.H., Gilbert, J.C., Tsai, C.-T., Chiang, F.-T., Holden, T., Barney, N., White, B.C.: A flexible computational framework for detecting, characterizing and interpreting patterns of epistasis in genetic studies of disease susceptibility. J. Theor. Biol. 241, 252–261 (2006)
3. Pritchard, J.K., Donnelly, P.: Case-Control Studies of Association in Structured or Admixed Populations. Theor. Popul. Biol. 60, 227–237 (2001)
4. Guinand, B., Topchy, A., Page, K.S., Burnham-Curtis, M.K., Punch, W.F., Scribner, K.T.: Comparisons of Likelihood and Machine Learning Methods of Individual Classification. J. Hered. 93, 260–269 (2002)
5. Park, J., Hwang, S., Lee, Y.S., Kim, S.C., Lee, D.: SNP@Ethnos: a Database of Ethnically Variant Single-Nucleotide Polymorphisms. Nucleic Acids Res. 35 (Database issue), D711–D715 (2007)
6. Zhou, N., Wang, L.: Effective Selection of Informative SNPs and Classification on the HapMap Genotype Data. BMC Bioinformatics 8 (2007)
7. Rosenberg, N.A.: Algorithms for Selecting Informative Marker Panels for Population Assignment. J. Comput. Biol. 12, 1183–1201 (2005)
8. DeNise, S.K., Charteris, P., Rosenfeld, D.: Compositions, Methods and Systems for Inferring Bovine Breed, PatentStorm: Application: 10750622 (2005)
9. Patterson, N., Price, A.L., Reich, D.: Population structure and eigenanalysis. PLoS Genet. 2(12), e190 (2006)
10. Bovine Genome Project, `http://www.hgsc.bcm.tmc.edu/projects/bovine/index.html`
11. Pritchard, J.K., Stephens, M., Donnelly, P.: Inference of population structure using multilocus genotype data. Genetics 155(2), 945–959 (2000)

Genome-Based Screening for Drug Targets Identification: Application to Typhoid Fever

Arporn Juntrapirom[1], Saowalak Kalapanulak[1,2], and Treenut Saithong[1,2]

[1] Bioinformatics and Systems Biology Program
[2] School of Bioresources and Technology,
King Mongkut's University of Technology Thonburi, Bangkok, Thailand
arj_5104@hotmail.com, saowalak.kal@kmutt.ac.th,
treenut.sai@kmutt.ac.th

Abstract. *Salmonella enterica serovar Typhi CT18* (*S. Typhi*) is the causative agent of typhoid fever in human beings. Currently, most of the drugs used to treat this sickness have adverse side-effects. Moreover, drug-resistant strains are emerging as a serious threat for the disease. Therefore, the most effective drug targets are urgently demanded for the development of new faster-acting antibacterial agents. In this paper, a published method for drug targets identification in *Mycobacterium tuberculosis* metabolism by Kalapanulak was applied to typhoid fever. The whole genome of *S. Typhi* was investigated and 282 genes were proposed as new drug targets. Interestingly, 34 drug–affected and essential genes from the three current antibiotics are all found in our proposed drug targets.

Keywords: *Salmonella Typhi,* typhoid fever, pathogenic bacteria, genome-scale, drug targets.

1 Introduction

Typhoid fever, an infectious disease, can be called by various names, such as gastric fever, abdominal typhus, infantile remittant fever, slow fever, nervous fever, pythogenic fever, etc. In 1829, Louis gave the name of "typhoid" as a derivative from typhus [1]. Typhoid fever remains a common disease in the developing world, where it affects about 12.5 million people each year. Around 10% of them will develop severe or complicated disease. Annually, more than 600,000 people die from the disease [2] .

Salmonella enterica serovar Typhi CT18 (S. Typhi), a gram negative bacterium, causes the vast majority of typhoid fever, systemic infection, in humans. It is transmitted from a patient to a normal person by the ingestion of food or water contaminated with the feces of an infected person. The pathogen usually attacks the surface of the intestine in humans but it can develop and adapt to grow into the deeper tissues of the spleen, liver, and the bone marrow. The most common symptoms characterized by the disease often include a sudden onset of a high fever, a headache, and nausea [3]. Some patients who are infected with *S. Typhi* become life-long carriers that serve as the reservoir for the pathogen. Moreover, the causative agent has an endotoxin (which is typical of gram negative organisms), as well as the Vi antigen, which increases its

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 217–225, 2010.

virulence. More importantly, it is a strong pathogen for humans due to its resistance to the innate immune response system [4].

At present, patients are treated with some antibiotics that kill the *Salmonella* bacteria. Before using antibiotics, the fatality rate was 20% and deaths occurred from overwhelming infection, pneumonia, intestinal bleeding, or intestinal perforation. With antibiotics and supportive care, mortality was reduced to 1-2%. Because of appropriate antibiotic therapy, patients are usually better within one to two days and recovery within seven to ten days. Several antibiotics are used to cure the disease, for example chloramphenicol, the original drug of choice for many years. Unfortunately, because of its serious side effects, chloramphenicol has been replaced by other antibiotics. Currently, the patients have to take multiple drugs i.e., chloramphenicol, ciprofloxacin, ceftriaxone, cefexime for the treatment but most of them still have adverse effects. Moreover, drug-resistant strains emerge existing serious problem. Besides medicine, a vaccine was developed during World War II by Ralph Walter Graystone Wyckoff for prevention [5]. However it is no longer recommended for use, it has a high rate of side effects (mainly pain and inflammation at the site of the injection). Therefore, the new effective drug targets are urgently demanded for developing new faster-acting antibacterial agents.

Nowadays, biological information of *S. Typhi* is available in various resources, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) database [6] collecting the gene functions of all annotated genes in any living organisms such as biochemical reactions of enzymatic genes, transport reactions of transporter genes. Additionally, Universal Protein Resource (UniProt) [7], and InterPro databases [8] are useful data resources for obtaining protein and protein signature information of living organisms, including *S. Typhi*. For InterPro database, protein signatures describing the same family or domain in terms of sequence position and protein coverage from other 10 protein signature databases are integrated into single InterPro entries. It brings us the occasion to analyze protein signatures of individual organism systematically.

Thanks to the availability of useful biological information mentioned above, it provides us an opportunity to investigate *S. Typhi* as a whole system by using systems biology approach and apply for drug targets identification. In 2009, Barh *et al.* applied comparative genomics approach for prediction of essential genes in *Neisseria gonorrhoeae* [9]. Like previous work, Dutta *et al*. applied the same *in silico* strategies for identifying drug targets of *Helicobacter pylori* [10]. Furthermore, Doyle *et al.* applied orthology-based approach for predicting essential genes in the pathogenic nematodes [11]. Additionally, in 2009 Kalapanulak proposed a novel method for identifying drug targets against *Mycobacterium tuberculosis* (*Mtb*), the causative agent of tuberculosis in humans [12]. Not only was genomic data integrated, but the wet experimental results (transposon site hybridization) were also included as a raw data for identifying drug targets in *Mtb* metabolism. Interestingly, 13 *Mtb* metabolic genes from all 42 proposed drug targets were found in a list of 70 current validated drug targets. The *in silico* approaches for identifying drugs targets have much benefit for biologists to screen for possible drug targets before doing their single-gene knockout experiment, one of the conventional methods for identifying drug targets.

In this work, a published novel method for drug targets identification by Kalapanulak in 2009 was applied to typhoid fever by integrating biological information of *S. Typhi* from various resources such as protein signatures from Interpro database, and

genome information from KEGG database. We further compared protein signatures between human host and the pathogenic organism. The proposed drug targets have been identified based on two criteria. First, their protein signatures do not match with any human protein signatures in order to reduce side effects. Second, they are essential genes required by *S. Typhi* for optimal growth from TraDis (transposon directed insertion-site sequencing) experiment.

2 Methodology

A published approach for identifying drug targets in *Mycobacterium tuberculosis* metabolism has been applied to *S. Typhi*. The whole genome of *S. Typhi* has been investigated by comparing its all protein signatures corresponding to genes in the genome between *S. Typhi* and human through InterPro accession numbers (InterPro Ac.) from InterPro database. Not only was the metabolic network of *S. Typhi* investigated, but we also did the analysis for the whole genome. Every gene in the *S. Typhi* genome has been considered as possible drug targets. The methodology for drug targets identification in *S. Typhi* is illustrated in Fig. 1.

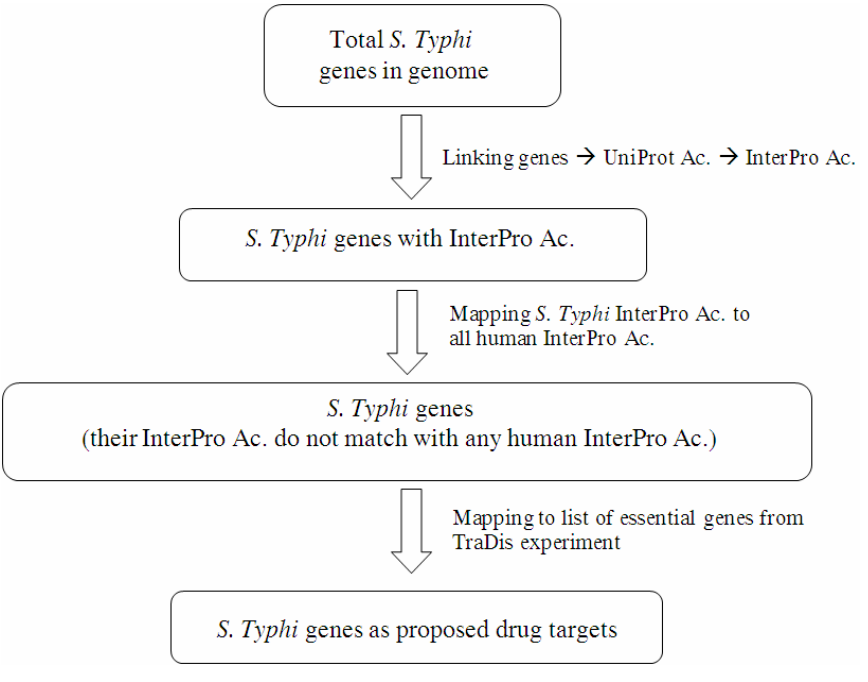


Fig. 1. The methodology for drug targets identification in *S. Typhi*

In the first step, linking genes to their protein signatures in terms of InterPro accession numbers, all genes and their UniProt accession numbers (UniProt Ac.) belonging to human (*H. Sapiens*) and *S. Typhi* genome were retrieved from KEGG database.

Next, the relations between UniProt accession numbers and InterPro accession numbers were extracted from UniProt database. Finally, the relations between genes and their InterPro accession numbers were created through UniProt accession numbers since we cannot link genes to their InterPro accession numbers directly. Hence, all genes with InterPro accession numbers were obtained and prompted for doing the protein signatures comparison between *S. Typhi* and human.

In the second step, a comparison of InterPro accession numbers between human and *S. Typhi* was made. A Visual Basic code was written for doing systematic comparison between the two groups of InterPro accession numbers. The unique InterPro accession numbers of each *S. Typhi* gene were compared with the whole set of human InterPro accession numbers. Eventually, the *S. Typhi* genes, of which the number of unmatched InterPro accession numbers are the same as the number of all InterPro accession numbers, are proposed as preliminary proposed drug targets.

In the last step, the preliminary proposed drug targets were mapped with the list of 356 essential genes required by *S. Typhi* for optimal growth from TraDis experiment (transposon directed insertion-site sequencing). Consequently, the drug targets were proposed based on two criteria: A) their protein signatures are unique in *S. Typhi*; B) they are essential genes reported from the TraDis experiment [13].

3 Results

3.1 Linking *S. Typhi* and Human Genes to Their Protein Signatures in Term of InterPro Accession Numbers

All gene names and their UniProt accession numbers of human and *S. Typhi* were downloaded from KEGG database by using the following URLs:

- ftp://ftp.genome/pub/keg/genes/organism/hsa/hsa_uniprot.list
- ftp://ftp.genome/pub/keg/genes/organism/sty/sty_uniprot.list

The results show that all genes in both of the genomes obtain UniProt accession numbers as illustrated in Table 1.

Table 1. Characteristics of Gene-UniProt-InterPro accession number relationship for *S. Typhi* and human

Organisms	Numbers of genes in genomes	Numbers of genes with UniProt Ac.	Numbers of genes with InterPro Ac.
S. Typhi CT18	4,679	4,679	3,938(84.16%)
H. Sapiens	22,339	22,339	16,915(75.72%)

The relationship between UniProt accession numbers and InterPro accession numbers of *S. Typhi* and human was extracted from the UniProt database. All data were downloaded by specifying the URLs because directly downloading from their web

interface resulted in inconsistent data. The URLs, http://www.uniprot.org/uniprot?query=organism:taxID(220341)+AND+database:interpro&format=tab&compress=yes&columns=id,database(interpro) and http://www.uniprot.org/uniprot?query=organism:taxID(9606)+AND+database:interpro&format=tab&compress=yes&columns=id,database(interpro), were used to download the data of *S. Typhi* and human, respectively. Consequently, we obtained 4,679 UniProt entries for *S. Typhi*. However, 3,938 of them provided at least one InterPro accession number per entry. On the other hand, 22,339 UniProt entries were retrieved for human but only 16,915 of them have InterPro accession numbers. It should be mentioned that not all UniProt entries in both *S. Typhi* and human have InterPro accession numbers. This is because the InterPro database covers only around 84.9% of UniProtKB, for the latest public release version, Release 2010_08, in 2010 [7]. Moreover, one protein may have one or more InterPro accession numbers depending on the number of identified protein signatures. Finally, 3,938 genes of *S. Typhi* had 10,666 unique InterPro accession numbers and 72,631 unique InterPro accession numbers were obtained from 16,915 human genes.

3.2 Comparing Protein Signatures between *S. Typhi* and *H. Sapiens* via InterPro accession Numbers

The InterPro accession numbers of 3,938 *S. Typhi* genes in Table 1 were compared to 72,613 InterPro accession numbers of 16,915 human genes. We wrote the Visual Basic code for doing the systematic comparison. Eventually, we found 2,419 *S. Typhi* genes, of which all InterPro accession numbers do not match any human InterPro accession numbers, from the total of 3,938 genes (Fig. 2). These 2,419 genes are proposed as preliminary drug targets against *S. Typhi.*

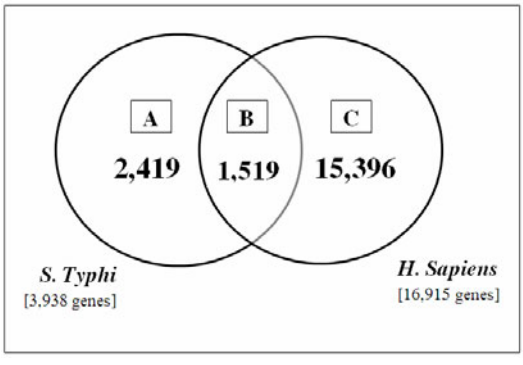


Fig. 2. Protein signature comparison between *S. Typhi* and human (*H. Sapiens*) via InterPro accession numbers; A) number of proposed preliminary drug targets against *S. Typhi*, B) number of genes that their protein signatures match between *S. Typhi* and *H. Sapiens* and C) number of *H. Sapiens* genes that their InterPro accession numbers do not match with any InterPro accession numbers of *S. Typhi*.

3.3 Mapping the List of Preliminary Drug Targets to Essential Genes from TraDis Experiment

The 2,419 genes in Fig. 2 were further compared with the list of 356 essential genes and 4,162 non-essential genes reported by Langridge *et al.* using transposon directed insertion-site sequencing (TraDis) experiment [13]. Consequently, the 2,419 genes that have been proposed as preliminary drug targets against *S. Typhi* were classified into two groups: A) 282 genes are essential, B) 2,137 genes are non-essential as shown in Fig. 3. Finally, all 282 genes in region A have been proposed as new drug targets. We further classified all proposed drug targets into seven categories based on their protein functions as shown in Fig. 4. Half of them are metabolic genes, more attractive as drug targets because of their potential for assayability and good druggability precedents. However, further analysis such as 3D-structure of proteins is still necessary for prioritizing the proposed drug targets.

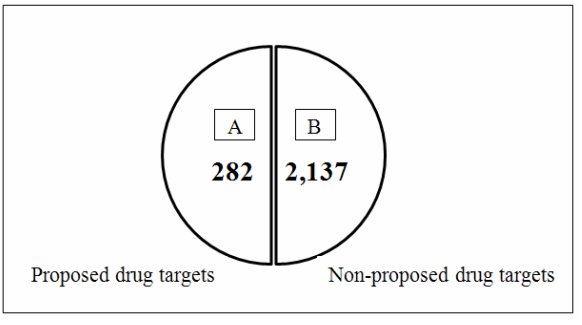


Fig. 3. Comparison between the preliminary drug targets and essential genes from TraDis experiment; A) number of proposed drug targets and B) number of non-proposed drug targets

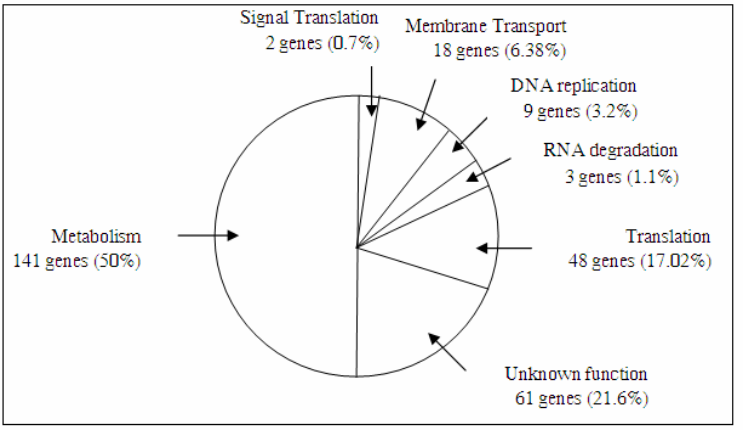


Fig. 4. Classification of 282 proposed drug targets into seven categories including Metabolism, Signal Transduction, Membrane Transport, DNA replication, RNA degradation, Translation, and Unknown function.

4 Discussion

The 282 genes from the whole genome of *S. Typhi* have been proposed as new drug targets based on two criteria: A) all of their protein signatures are different from human protein signatures; B) they are essential genes from TraDis experiment. In order to state the confidence of the novel approach for genome-scale identification of drug targets, we compared the proposed drug targets with 44 drug-affected genes reported by Becker *et al*. in 2006 [14]. The results show that 34 drug-affected genes have been found in our proposed drug targets.

In term of statistics for clarification of the confidence of these 282 proposed drug targets from the total of 3,938 investigated genes, we compared the proposed and non-proposed drug targets against 43 drug-affected genes from literature. One drug-affected gene reported by Becker *et al*. has not been included for statistical calculation because it did not have any protein signatures. Therefore, it is not include in the 3,938 investigated genes. Of the 282 proposed drug targets, 34 (12%) were current drug-affected genes. Of the 3,656 non-proposed drug targets, 9 (0.3%) were current drug-affected genes. To demonstrate the probability for obtaining 34 or more drug-affected genes from randomly selecting 282 *S. Typhi* genes without replacement from the total of 3,938 investigated *S. Typhi* genes, we calculated the hypergeometric probability for obtaining 34 or greater drug-affected genes using the below formulas [15].

$$p(x \geq 34) \quad = \quad 1 - p(x \leq 33) \tag{1}$$

$$p(x \leq 33) \; = \; \sum_{i=0}^{33} \frac{\binom{3,895}{249+i}\binom{43}{33-i}}{\binom{3,895}{282}} \; = \; 0.999999999998596 \tag{2}$$

With a population size = 3,938, sample size = 282, number of successes in the population = 43, number of successes in the sample (x) = 33.

From the probability calculation we found that the probability for receiving 34 or more drug-affected genes from randomly selecting 282 genes of 3,938 investigated *S. Typhi* genes is closed to 1×10^{-12}. It means that it is uncommon to receive 34 or more drug-affected genes when randomly sampling 282 genes from the total of 3,938 *S. Typhi* genes. Therefore, the identification of 34 drug-affected genes from 282 proposed drug targets (12%) is significant enough to confirm the quality of the published novel method in term of genome-scale analysis.

Moreover, we further compared 44 drug-affected genes with essential genes based on TraDis experiment. Interestingly, all 34 drug-affected and essential genes are all found in our proposed drug targets (Fig. 5). Among 34 drug-affected genes, 32 of them function as metabolic genes, whereas the other two are unknown genes. Therefore, we may conclude that the published novel method has high accuracy for predicting drug-affected genes that are essential for the pathogen and most of them are metabolic genes.

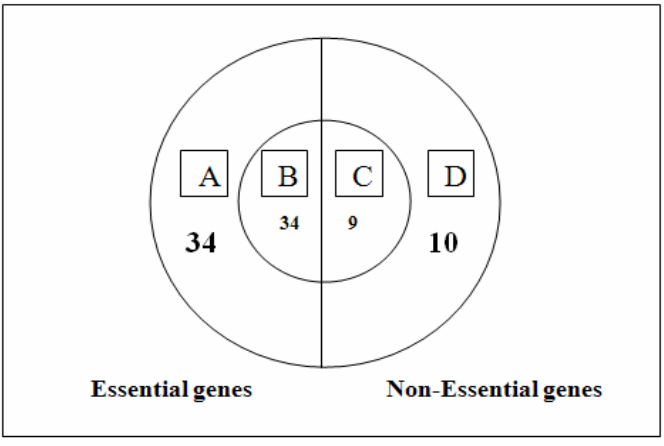


Fig. 5. Classification of 44 drug-affected genes on essential genes and non-essential genes from TraDis experiment. A) number of drug-affected genes from three current antibiotics (Becker *et al*., 2006) and essential genes based on TraDis experiment, B) number of proposed drug targets, all of them found in drug-affected and they are essential genes based on TraDis experiment, C) number of non-proposed drug targets, all of them found in drug-affected genes and they are non-essential genes based on TraDis experiment, D) number of drug-affected genes from three current antibiotics (Becker *et al*., 2006) and non-essential genes based on TraDis experiment.

5 Conclusions and Future Work

Two hundred and eighty-two genes of *S. Typhi* have been proposed as novel drug targets based on the *in silico* screening approach. Drug targets are proposed based on two criteria: A) no protein signature matching with any human protein signatures and B) essential genes from TraDis experiment. Interestingly, 34 drug–affected and essential genes from the three current antibiotics are all found in our proposed drug targets. It brings to the achievement of applying the published novel method for identifying new drug targets against typhoid fever. For future work, we will further improve the published method in order to overcome the limitation of the method. Genes without protein signature information were not included in the analysis, even though they can be interesting drug targets. Moreover, the modified method will be implemented for drug targets identification as a web application tool. It will facilitate biologists to do computational screening via our user-friendly tool before doing their wet experiments.

References

1. Kaye, K.S., Kaye, D.: Salmonella infections (including typhoid fever). In: Goldman, L., Ansiello, D. (eds.) Cecil Medicine, 23th edn. Saunders Elsevier, Philadelphia (2007)
2. Ivanoff, B., Levine, M.M.: Typhoid fever: continuing challenges from a resilient bacterial foe. Bulletin de l'Institut Pasteur 95, 129–142 (1997)

3. Shanahan, P.M., Jesudason, M.V., Thomson, C.J., Amyes, S.G.: Molecular analysis of and identification of antibiotic resistance genes in clinical isolates of *Salmonella typhi* from India. J. Clin. Microbiol. 36, 1595–1600 (1998)
4. Everest, P., et al.: The molecular mechanisms of severe typhoid fever. Trends in Microbiology 9, 316–320 (2001)
5. International Union of CRYSTALLOGRAPHY, http://ww1.iucr.org/people/wyckoff.htm
6. Ogata, H., et al.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 27, 29–34 (1999)
7. Apweiler, R., et al.: The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. 38, D142–D148 (2010)
8. Hunter, S., et al.: InterPro: the integrative protein signature database. Nucleic Acids Res. 37, D211–D215 (2009)
9. Barh, D., Kumar, A.: In silico identification of candidate drug and vaccine targets from various pathways in Neisseria gonorrhoeae. Silico. Biol. 9, 225–231 (2009)
10. Dutta, A., et al.: In silico identification of potential therapeutic targets in the human pathogen *Helicobacter pylori*. In Silico. Biol. 6, 43–47 (2006)
11. Doyle, M.A., et al.: Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. BMC Genomics 11, 1–14 (2010)
12. Kalapanulak, S.: High Quality Genome-Scale Metabolic Network Reconstruction of Mycobacterium tuberculosis and Comparison with Human Metabolic Network: Application for Drug Targets Identification. PhD Thesis. School of Informatics, The University of Edinburgh (2009)
13. Langridge, G.C., et al.: Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Res. 19, 2308–2316 (2009)
14. Becker, D., et al.: Robust *Salmonella* metabolism limits possibilities for new antimicrobials. Nature 440, 303–307 (2006)
15. StatTrek Teach yourself statistics, http://www.StatTrek.com

Author Index