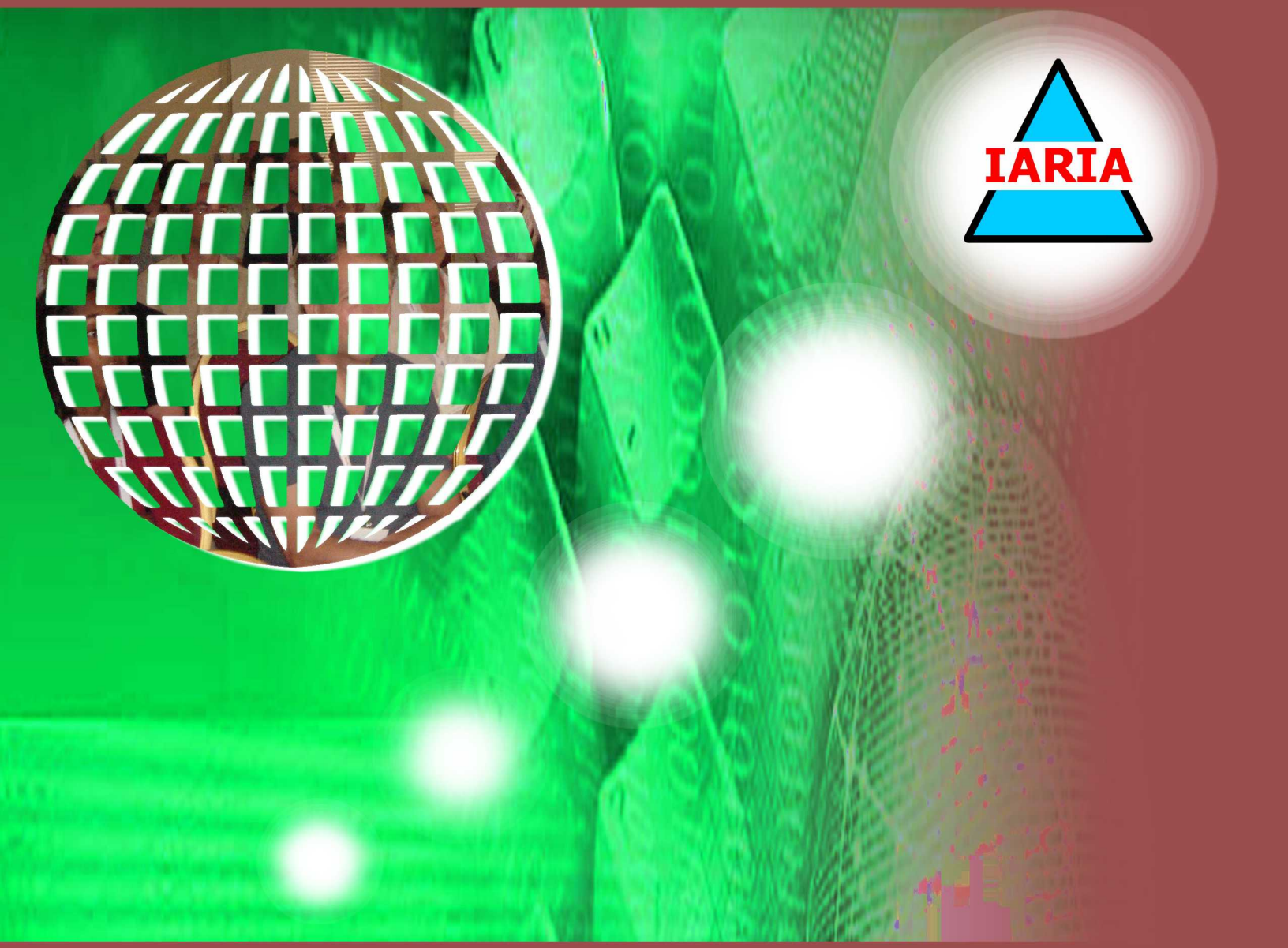
International Journal on Advances in Life Sciences



2012 vol. 4 nr. 1&2

The *International Journal on Advances in Life Sciences* is published by IARIA.
ISSN: 1942-2660
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Life Sciences
Volume 4, Number 1 & 2, 2012

International Journal on Advances in Life Sciences
Volume 4, Numbers 1 & 2, 2012

CONTENTS

Human Behaviour Analysis Using Data Collected from Mobile Devices

[1]Muhammad Awais Azam, [1]Jonathan Loo, [1]Sardar Kashif Ashraf Khan, [2]Muhammad Adeel, [3]Waleed Ejaz
[1]School of Engineering and Information sciences, Middlesex University, London, UK
[2]School of Electronic Engineering and Computer Science, Queen Mary University, London, UK
[3]Department of Information and Communication Engineering, Sejong University, Republic of Korea
*Corresponding Author {m.azam@mdx.ac.uk}

Abstract- **Human behaviours are multifarious and myriad in nature. It is a challenging task to envisage and learn the human behaviour from daily routine activities. The profusion of wireless enabled mobile devices in daily life routine and advancement in pervasive computing has opened new horizons to analyse and model the contextual information. The aim of this research work is to infer the behaviour of low entropy mobile people using contextual data collected from mobile devices such as GSM location patterns (cell tower ID data) and Bluetooth proximity data. Both the GSM and Bluetooth data itself do not reveal much information about the behaviour of the users. Therefore, the challenge is to find out whether such data can infer human behaviour to understand and aid the unusual activities and routines of low entropy people such as elderly people and early stages of dementia patients. In this paper, a framework is created to analyse the contextual data for behaviour detection. There are four different steps in this framework to achieve the objective of the research work. In the first step, the contextual data is first classifies into different locations to obtain the movement patterns of the users. In the second and third step, a probability matrix and training data is obtained respectively, depending upon the user's movement on daily and hourly basis. In the fourth step, a decision engine i.e. Neural Network (NN) and Decision Trees (DT) is used to detect the behaviour of the low entropy user. Results have shown that cell tower ID data gives behaviour of the user on high level scale for example movement patterns in GSM cells that does not help to identify any lower level activities such as attending the lecture, traveling in a bus. Whereas, Bluetooth data gives us more information about the lower level activities depending on the social relations and close proximity of other users.**

Keywords – Behaviour, Cell Tower ID, Bluetooth Proximity, Neural Networks, Jaccard Index, Decision Trees

I. INTRODUCTION

Detection and prediction of human behaviour from daily life activities is a challenging task. People can have both regular and varying daily life routines that make it a burning topic nowadays in social research circles. Modelling human behaviour such as individual routines from proximity data and social relations with gathered data of daily life activity patterns is an emerging realm in Ubiquitous Computing. Computers are becoming more and more pervasive and are embedded in everyday objects, such as cameras, music players, cars, clothing etc. There can be different sensing devices e.g., Radio Frequency Identification (RFID), motion sensors, GPS enabled tracking devices, and other context aware devices that can be used for real time proximity detection and daily life data gathering purposes. In particular, devices such as mobile phones provide a rich platform for various forms of data gathering by using its integrated sensors such as Bluetooth ID, digital camera, microphones and GPS transceivers. These sensors can give an individual's location, movement and proximity information for the whole period of cell phone usage. Specifically, Bluetooth radios are frequently incorporated into mobile devices [2].

This new generation of "smart devices" has created new ways to utilise the capability of computers and enhanced the area of Ubiquitous Computing by providing rich and detailed information about the context of the user. Context-aware computing, which is part of Ubiquitous Computing, uses sensors either in the environment or carried / worn by the users to extract and interpret the user's context, for example what resources are available, who is in close user's proximity. This contextual information can help to recognise different tasks and activities perform by the user.

Different researchers have worked on routine and activity classification using mobile phone data [2] [3] [13] [14]. They have tried to analyse the social relationships and daily life routine patterns of individuals using their cell phone data. They classified the cell tower ID data into different locations such as Home, Office, Elsewhere and No-signals and analyses the movement patterns. They have also used Bluetooth proximity data to differentiate between weekday or weekend activities. In our research work, we want to go a step further in behaviour analysis. As cell tower ID information can only give patterns of movement and location information. It cannot tell us about low level activities. For example, cell tower ID data can tell whether the user is at home or in office/campus, but it cannot tell in which activity such as attending the lecture or sitting in cafeteria, user is participating. On the other hand, proximity data can give information about the people and other devices that are in close vicinity of the user but it does not tell exact information about the user's location. If this proximity data can be classified into different locations, then proximity information can provide a good idea about the nature of activity that user is performing and this will help in analysing and understanding an individual's behaviour.

The aim and focus of this research work is on the detection of behaviour of the people who live low entropy lives that means they follow somewhat regular routines and exhibit less change in their behaviours as discussed in [3]. According to [3], if the user in his daily life, repeat the activities and routines with less change, it will be known as 'low entropy' behaviour. While a more change in daily routine patterns is considered as 'high entropy' behaviour. For example, a working person who follows the routine of going to the office and coming back home every day using the same means of the transport, or an elderly person with

regular routines [4] (e.g., an early stage of dementia patient) can be the examples of the people with more regular routines and hence less change in the behaviour. The motivation of this research work is to help elderly people and early stages of dementia patients to live their lives more independently by understanding their behaviour from wireless proximity data.

This work is an extension of [1] and [5], in which repeated patterns and behaviour of an individual was detected by using n-gram technique and considering only Bluetooth proximity data and the behaviour was detected by using only NN. The research work in [5] proves the concept that daily life traces of Bluetooth proximity data of a low entropy individual can give us enough repeated patterns in the data that can be further used for activity or behaviour detection. In [1], the unusual routines in the daily life of the user were detected by using the NN only.

In this research paper, we have used two different types of contextual data (GSM cell tower ID and wireless proximity data), that are rather collected independently, to analyse the behaviour of low entropy mobile people. Wireless proximity data that is used in this research work is of Bluetooth. The data set used in this paper is the reality mining dataset [3] collected at MIT for the year 2004-2005. Nokia 6600 cell phones were used to record the data of 100 users over the duration of 9 months. Different types of information were collected including phone status i.e. whether it is in use or charging or off, ID's of Bluetooth proximate devices, usage of mobile applications, cell tower ID data, call and SMS logs

The rest of the paper is as follows: Section-II contains related work on unusual activity detection and usage of Bluetooth as a sensing device. Section-III and Section-IV discusses the research objectives and the behaviour analyses framework respectively. Section-V discusses the behaviour analysis results using cell tower ID data and Section-VI contains the results of behaviour detection using Bluetooth proximity data. Summary of the work and notes on the direction planned for the future work is in Section-7.

II. Related Work

Detection of abnormality in human behaviour is very intricate and has been a challenging task in the past. Though, recent advancements in information technology had made it quite simpler. In last few years a lot of efforts have been made to observe the abnormal routines and daily life patterns of an individual [6] [7]. In [6], the author has presented a framework for the detection of unusual human behaviour inside an intelligent house. The author used motion sensors to detect the activities and unusual human behaviour patterns based on Markov Chain. Vector quantization is employed to reduce the sensor states and the transition between states is represented by probabilistic model. The above mentioned technique detects the unusual human behaviour either by computing the distance between the state transition probabilities or by the likelihood of human action. The distance between the state transition probabilities was calculated by using either Kullback-Leiber distance or Euclidian distance. Limitation of this work is that they only consider the indoor activities that can only happen inside the home. To analyse human behaviours and activities, some authors have also used devices other than motion sensors such as, accelerometers, digital cameras and microphones.

In literature some techniques has also been presented to analyse the accumulative behaviour of multiple individuals instead of one single individual. For example, in [8] the author proposed a framework based on identification of close proximity social behaviours. This work also focused on the movements inside a building. Similarly other multiple individual behaviour detection schemes such as group actions in meetings [9] and audio visual perception of a lecture in smart environment [10] are presented. However, majority of work in above mentioned studies have focused on indoor environment; as it is based on sensing devices which have several limitations such as short range of detection, less battery power and storage, or may not be very common that every person can use it without extra hardware, which is not feasible for outdoor environment.

The enormous penetration ability of Bluetooth technology have made it more suitable candidate to be used as a personal identifier. This capability can be exploited by using the mobile phone having Bluetooth technology as a sensing device. Nowadays the mobile phone is an indispensable part of our society with many types of embedded sensors. These sensors have been used in many worth mentioning applications such as social proximity sensing [11] [12], social behavioural modelling and routine classification [2] [3] [13] [14] and movement prediction [15] [16]. The significance of aforementioned studies is that these techniques have focused on how to recognize an individual's behavioural patterns and social routines but no one of them has classified the Bluetooth proximity data into different locations and predicts the behaviour by using the machine learning techniques.

In [13] and [14], researchers have presented a framework for daily life activity recognition based on the user's location and group affiliation. They used Author Topic Model (ATM) and hierarchical Bayesian topic models like Latent Dirichlet Analysis (LDA) for routine classification. The routines they classified are whether it is a weekday or a weekend depending upon the location of the user or the proximity information and whether the experimental subject is an engineering student or a business student. The proximity data is only classified depending upon the number of proximate devices. There classification of proximity data does not give any information about the location of the user.

In [15] and [16], NN are used to detect and predict user movement based only on cell tower IDs. They utilised the probability of user being at different locations. Our work is similar in one aspect with their work and that is; we have also utilized the probabilities of user being in different locations. Difference between our work and the work presented in [15] and [16] is that we have used real time data for our experiments and have used both cell tower ID and Bluetooth proximity data. In [17], researchers proposed a relaxing minimum description length (MDL) principle in order to build compatible decision trees that are suitable for novel behaviour detection. This relaxing MDL principle is to exploit additional tests/features in order to discriminate between normal and abnormal behaviours.

In [7], researchers detect abnormal event in solitary elder's daily life by mining the related data gained by sensors. They employ the association rules finding algorithm

with time cluster to analyse the elder's activities. In first step, they cluster each item of elder activity with time and then in the second step, all frequent item sets were found and strong association rules were created. Researchers in [18] work on the recognition of abnormal activities based on the Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM). They incorporate a Fisher Kernel into the One-Class Support Vector Machine (OCSVM) to filter out the most likely normal activities. Then from those normal activities, they derive a model to detect abnormal activities and tried to reduce false positives. In [19], researchers have presented a model for abnormal behaviour detection. That model considers user's location based on the cell tower ID and used Dynamic Bayesian Networks (DBN) to predict user's location. They proposed an X-Factor model, which is a DBN with a hidden variable. User's location according to this model not only depends on the hour of the day and day of the week but also this latent variable that represents the abnormal behaviour.

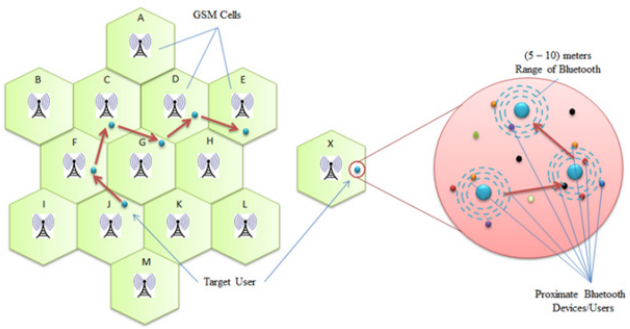
Most of the researchers as discussed above have focussed on routines and activity detection in closed and indoor environments and have used short range sensors that can work only in very close vicinity and have short battery life. This type of sensors cannot be used for outdoor environment. Our research work is not constrained of short range sensors and short battery lives. We explored the concept of mobile phone as a sensing device. Many other researchers as discussed above have also used mobile phones to get the proximity data and user's location from cell tower ID information. To the best of our knowledge, no one has classified the Bluetooth proximity data into different locations and obtained the user's movement patterns. In this paper, we address this concept and analysed the behaviour from cell tower ID and Bluetooth proximity data and found out that Bluetooth proximity data alone can be used to detect the behaviour of the low entropy mobile user. Results have shown that patterns in wireless proximity data can give us enough information about the routines of the user and unlikely cell tower ID data that can only give indications of user movement patterns at different locations, it can also give information about user's activities while staying at one particular location which is not possible to get from cell tower ID data.

III. Research Objectives

The primary aim of this research work is to find any anomalies in the behavioural patterns or routine activities of low entropy mobile people in order to aid in the detection of any unusual behaviours in elderly people or patients such as early stages of dementia patients. First objective is to utilize the contextual data (such as, cell tower ID and Bluetooth proximity data) available around us that can be obtained through different sensing devices, especially mobile phones, for behaviour detection. A framework is designed to analyse the behaviour of the low entropy users by using this contextual data.

The nature of Cell tower ID and Bluetooth proximity data is different from one another. Figure-1(a) shows the movement of a user in between different GSM cell towers. When a user is in the range of any GSM cell tower, ID of the cell tower is detected. This cell tower ID data only gives information about the user's movement in broad overview and cannot tell what type of activities user is performing within the range of detected cell towers. For example in Figure-1(a), user was in cell 'J', then moved to the cells 'F', 'C', 'G', 'D' and 'E' respectively. This information can only tell about the user's movement patterns and cannot give any idea about the activities that user is performing while at these locations. The purpose is to utilise this cell tower ID data to analyse behaviour of low entropy mobile people from the 'location data'. In order to detect the behaviour, two different machine learning algorithms have been used in the framework. The detection accuracy of both algorithms is also studied.

On the other hand, Figure-1(b) shows the detection of Bluetooth proximate devices. Cell tower ID data only can give user's location information, which in this case is cell 'X', whereas Bluetooth proximity data gives information about the people and other Bluetooth devices that are within the range of user's Bluetooth mobile device. Social relationship and group activities can be detected with this proximity data which is not possible to detect from the cell tower ID data. A weakness of Bluetooth proximity data is that it does not give any direct information about the location of the user. Location information is important to know in order to analyse the behaviour from daily routines and activities of the low entropy users. To obtain the location information from the Bluetooth proximity data is a challenging task and is also an objective of this research work. To get the location information from Bluetooth proximity data, we classify the Bluetooth detected devices into different groups that belong to locations such as Home, office and inferred the location of the user depending upon the detected devices. Another objective is to find out whether only Bluetooth proximity data can be used for behaviour and activity analyses and whether it can add more information about activities and daily routines of the user if we consider both cell tower ID and Bluetooth proximity data together.



(a) (b)

Figure 1. Scenario of GSM Cell Tower and Bluetooth Proximate Devices Detection

IV. Behaviour Analysis Framework

Figure-2 shows the overall framework which is going to be used to analyse the behaviour of low entropy mobile people from cell tower ID and Bluetooth proximity data. As aforementioned, real time traces of GSM cell tower ID and Bluetooth proximity data of low entropy people used for this

research work is obtained from the Reality Mining dataset. There are four steps in this framework that are followed to achieve the objective.

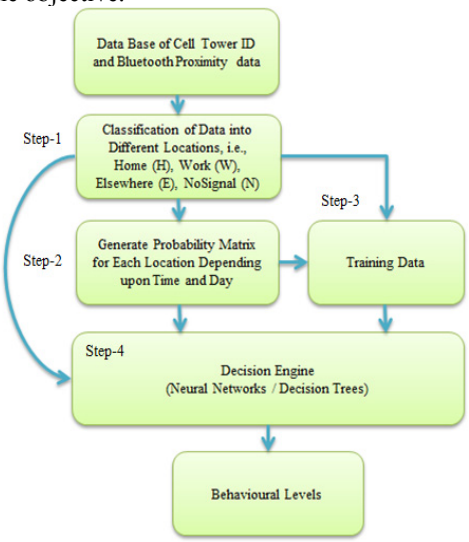


Figure 2. Behaviour Analysis Framework

Step-1 is to classify the cell tower ID and Bluetooth proximity data into different locations to find the activity and routine patterns of the user. The cell tower ID data which is obtained from the Reality Mining dataset is already classified into four different locations; i.e., Home (H), Work (W), Elsewhere (E) and NoSignal (N); whereas Bluetooth proximity data is in the form of list of detected proximate devices by the target user. Classification of Bluetooth proximity data into different locations is a great challenge and the classification procedure of Bluetooth proximate devices into different locations is explained in detail in Section-6.

Step-2 is to obtain a probability matrix predicting the location conditional on the hour of the day and day of the week from the classified information. This means, every entry of this matrix depends upon the specific hour of the day and whether it was a week day or a week end. Figure-3 shows the structure of the probability matrix for H, W, E and N for all twenty four hours. Each row in this matrix shows the hour of the day and each column shows the probability of H, W, E and N for that hour of the day. Depending upon this calculated probability; behaviour is divided into four different levels, shown in Table 1. Every entry of the probability matrix depends upon the specific hour of the day and whether it is a weekday or a weekend.

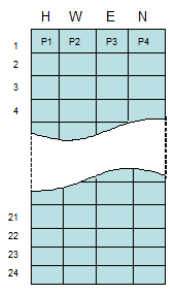


Figure 3. Probability Matrix for H, W, E and N

TABLE 1. BEHAVIOURAL LEVELS

Probability	Behaviour
0 < p < 0.25	Abnormal
0.25 < p < 0.5	Low Abnormal
0.5 < p < 0.75	Average Normal
0.75 < p <1	Normal

Step-3 is to utilise the probability matrix obtained in the setp-2 for preparing training data for the decision engine that is used for the detection of level of abnormality in the user's behaviour.

In step-4, the decision engine will use the training data obtained in the step-3, the probability matrix from step-2 and the classified data from setp-1 with a machine learning algorithm (NN or DT) to detect the behaviour of the user that deviates from the normal routines.

Next section discusses the behaviour analysis from cell tower ID data by using the above mentioned framework.

V. BEHAVIOUR ANALYSIS FROM CELL TOWER ID DATA

This section uses only the cell tower ID data of a low entropy user obtained from the reality mining dataset with the entropy level 23.06, calculated by using the Shannon's entropy equation shown in Equation-1, to find any anomalies in the daily life routines and behaviour of the user.

$$H(x) = -\sum_{i=1}^{n} p(i) \log_2 p(i) \qquad (1)$$

Cell tower ID gives information about the user's location and movement patterns. Step-1 is to classify the cell tower ID data into different locations to obtain the movement patterns of the user. As already discussed, the cell tower ID data that is used in this study is already classified into four different locations; i.e., H, W, E, N. This data is divided into twenty four time slots. Each time slot is represented by the associated presence information of the user (H, W, E, and N) during the one hour period as shown in Figure 4. The presence of user at specific location depends on the hour of the day and day of the week. For example, if the user has a regular routine of going to the office, then location of the user at 10a.m on Saturday morning cannot be the same at 10a.m on Monday morning. The daily life activities of an individual depend on the entropy level of the user as discussed in [3]. If the user is a low entropy user, his routines do not change much as compared to high entropy users, whose routines and activity patterns change continuously.

Step-2 is to obtain a probability matrix, which is generated depending on the hour of the day and day of the week from the classified information obtained in step-1 and then this probability matrix is used for the preparation of the training data in step-3.

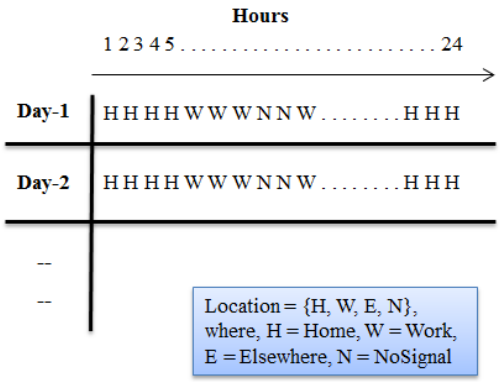


Figure 4. Format of cell tower ID Data

In step-4 decision engine detects the behaviour of the user. Two machine learning algorithms (NN and DT) are used for behaviour detection at this stage. The accuracy in terms of number of detections of both algorithms will be calculated and compared. The one, with the highest percentage accuracy will be used in the Section-6 with Bluetooth proximity data. Next section explains the behaviour analysis from cell tower ID data using NN.

A. Behaviour Analysis from Cell Tower ID Data Using Neural Networks

Figure-5 shows the basic architecture used to get the behaviour of an individual using NN. The neural network used here is Multi-layered Perceptron (MLP). Multi-layered Perceptrons have been created to try to solve the problem of non-linear classification of input instances by Rumelhart et al. [21]. A multi-layer neural network system consists of a large number of neurons connected with each other in a specific pattern. These neurons are normally divided into three classes; input layer neurons, hidden layer neurons and output layer neurons. The MLP used in this research work has four inputs and one output. Inputs are {Location, Hour, Day and Behavioural_Level}, where 'Location' gives the location of the user i.e., H, W, E, N, 'Hour' gives the hour of the day i.e., between 1 and 24, 'Day' gives the day of the week, i.e., between 1 and 7 and 'Behavioural_Level' gives the behavioural levels. Output of this neural network will give the level of abnormality of an individual for each hour of the day.

This gives twenty four samples of training data for one day. For each user, total training samples are (24 x numbers of days). 70% of these training data/samples are used for training the neural network whilst the remaining is used for cross validation and testing purposes. Training of the neural network is done till the cross validation error becomes less than 0.02, by using Mini-Batch training process [22]. The advantage of using Mini-Batch training is that it is a compromise between batch and incremental training. Back Propagation (BP) algorithm is used to estimate the weights of the neural network that includes the following steps:

- Provide a sample of training data to the NN.
- Calculate the error by comparing the desired output with the NN output.
- Adjust the weights of each neuron in order to lower the error value and again calculate the error.
- Repeat the steps unless reach the desired level.

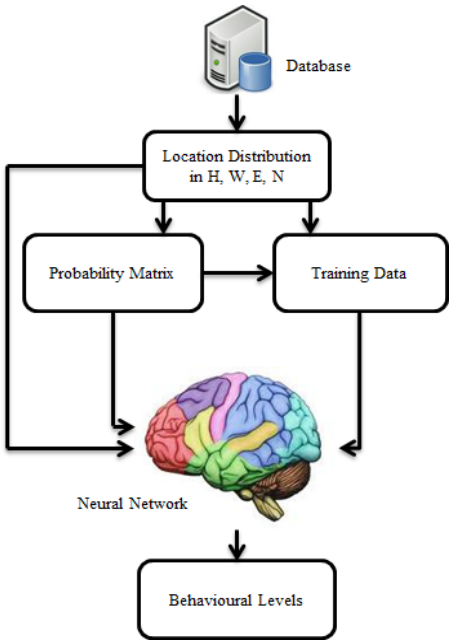


Figure 5. Behaviour Analysis from Cell Tower ID Data Using NN

Out of nine months of data available for this specific user; about 70% is used for training the neural network, one month data is used for behaviour detection and remaining is used for the cross validation purposes. Figure-6 shows the daily distributions of (H, W, E and N) transitions based on cell tower ID data of one month that is further used as a ground truth to detect the behaviour of the user and to calculate the accuracy of the NN in this specific scenario.

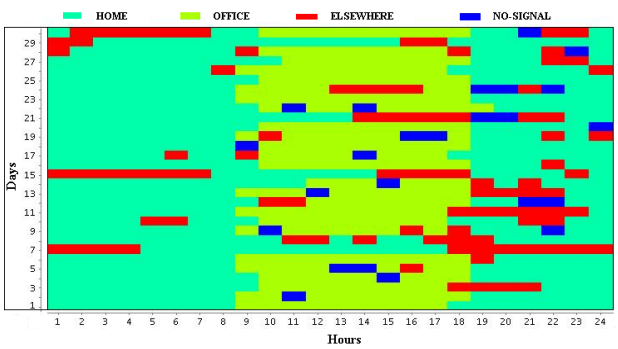


Figure 6. Distribution of (H, W, E and N) Transitions of Cell Tower ID Data

For understanding purposes, behaviour detection of the user for only two different days is discussed first. Figure-7 shows the comparison of behaviour of the user for Day-1 and Day-10. The trained neural network provides the behavioural levels for twenty four hours. First part of the figure shows the distribution of twenty four hours for Day-1 and Day-10 for the specific user in the form of H, W, E and N, whereas second part shows the inferred behaviour of the user. As the entropy level of the user is quite low, this figure shows that most of the time the behaviour of the user is average normal. Now if we look at day-10 in Figure-4, there is an unusual detection of 'Elsewhere' during 5-6am in the morning, which doesn't happen normally in usual daily routine of the user. Figure-6 also shows the detection of that unusual behaviour for day-10 in that specific time duration.

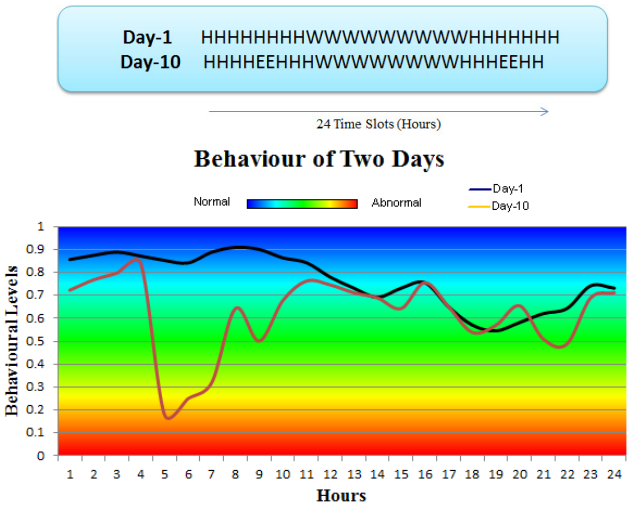


Figure 7. Comparison of Two Days of Behaviour Detected from Cell Tower ID Data Using NN

Figure-8 and Figure-9 show the behaviour of the user for one month time duration. Behaviour is divided into four levels as mentioned in Table-1. According to these levels, if the predicted behavioural value is near the '0', it means the users routine is more deviated from the normal routine activities and if it is near the '1', it is more normal. Figure-8 shows the first fifteen days and the Figure-9 shows the last fifteen days of the month. In Figure-8, the behaviour of the user for first nine days remains average normal as most of the predicted behavioural value lies between behavioural range of '0.5 - 0.7'. This can be verified from Figure-6 as well that shows the regularity in the distributions of 'Home' and 'Work' patterns and shows that user did not make any unusual movements. However, on 10th and 12th day of the month, between 5a.m – 7a.m and 10a.m – 12p.m respectively there is a change in behaviour when the user's (H, W, E and N) distributions in Figure-6 show an irregular routine activity. NN detects this behavioural change and is shown in the Figure-8 with two sharp low peaks on 10th and 12th day of the month.

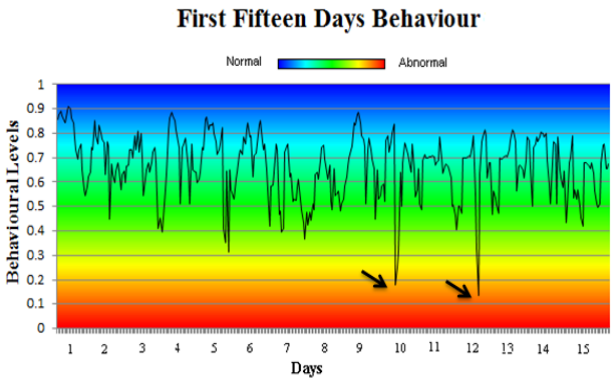


Figure 8. First Fifteen Days Behaviour Detected from Cell Tower ID Data Using NN

In Figure-9, last fifteen days of the month also show some routines that deviate from the normal behavior of the user. These routines are shown by sharp low peaks on 17th, 24th, 27th and 30th day of the month. These unusual routines are mostly detected on the week days in the morning before the office hours and some times during the office hours. As the user in these experiments belongs to academia, these results may show that, he or she most likely attending some seminar or a social function that is not part of the normal routine or due to health or traffic reasons, user sometimes comes late in the campus.

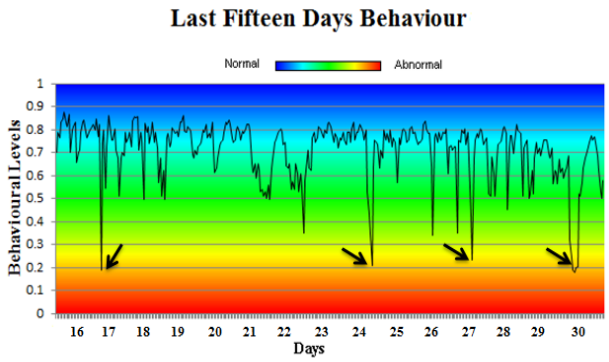


Figure 9. Last Fifteen Days Behaviour Detected from Cell Tower ID Data Using NN

B. *Behaviour Analysis from Cell Tower ID Data Using Decision Trees*

Figure-10 shows the basic architecture used to get the behaviour of an individual using DT algorithm. According to [20], DT classifies the instances by sorting them based on their feature values. Features are represented by different nodes in the DT's, and the value of the nodes is represented by the branches. Starting at the root node, each instance is classified and sorted depending upon the feature values. Root node is the feature value that best separates the data. The most well-known algorithm to build a DT is the C4.5 [23] and is used in this research work. The training and test data used for this algorithm is only the cell tower ID data of the same user as is considered in the previous section with NN. As already mentioned, this cell tower ID data is already divided into Home (H), Work (W), Elsewhere (E) and NoSignal (N) locations.

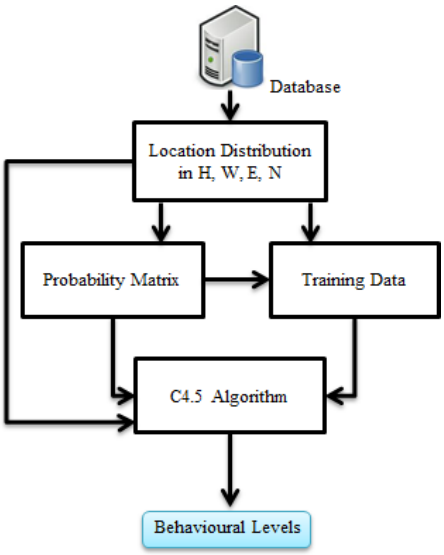


Figure 10. Data Processing Design Using C4.5 Algorithm

Figure-11 shows the behaviour detection of the user for the first fifteen days of the data using DT (C4.5 algorithm). These results show that the detections made by the DT are uniform as compared to NN, which are irregular. A DT consists of nodes and branches. Depending upon the four different behavioural levels, each node on DT represents a single behavioural level unlike NN that gives a predicted value that can be in between two different levels. The unusual routine activities are represented by the sharp low peaks in Figure-9. Another observation made is that all the unusual routines detected by DT lie in the range of 'Low_Abnormal' behavioural level and none of these are in 'Abnormal' level, unlike NN. A reason can be the biasing nature of the DT.

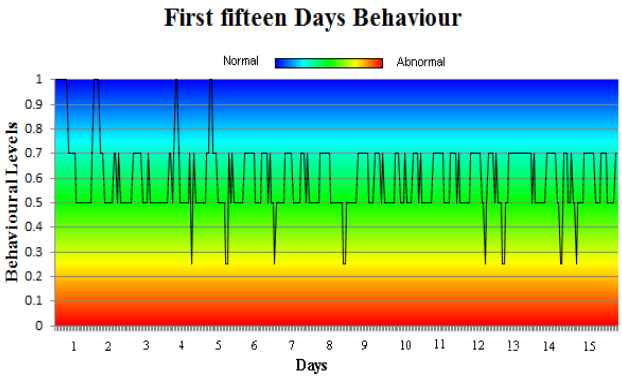


Figure 11. First Fifteen Days Behaviour Detected from Cell Tower ID Data using DT

C. *Accuracy Comparison between NN and DT in detecting behaviour from cell tower*

The comparison criteria for both NN and DT used here are the time required for learning and the percentage accuracy $P = N_c / N_t$, where 'N_c' is total number of correct detections and 'N_t' is the total number of detections. The results show that DT has some advantages over NN. The first advantage of induction of DT is its easy use and second advantage is that it requires fewer amounts of training data to train the classifier. The training time required for DT is also less as compared to the NN. For 3024 samples of training data and on a Pentium-IV 2.4GHz – processor with 2GB RAM, it takes only a few seconds until a decision tree has been trained. Whereas on the same system, NN takes about 13 minutes for complete training. However; the most important point for the specific problem here is, how accurate is the detection of behaviour by using DT as compared to the NN. For this purpose, percentage accuracy of both NN and DT is calculated. For 720 detections, the percentage accuracy of NN is 93% whereas DT gives the percentage accuracy of 86.3%. The bench data used for the calculation of percentage accuracy is the original data set of cell tower ID that was already classified into H, W, E and N. The reason behind the difference in the accuracy can be the biasing limitation of decision trees. As NN gives more accurate results in terms of percentage accuracy, for further processing and behaviour detection using Bluetooth proximity data, only NN will be used.

VI. Behaviour Analysis from Bluetooth Proximity Data

Bluetooth proximity data is available in the form of detected devices as a result of a scanning performed by the user's cell phone after every five minutes. Each scanning results a list of devices present within the range of 5-10m. The first aim is to classify this list of detected proximate Bluetooth devices into different locations, i.e. 'Home', 'Office', 'Other Devices' and 'No Devices Found'. List of Bluetooth proximate devices does not give any direct information about the location of the user. The reason for classification of Bluetooth devices is to obtain the user's movement patterns on daily and hourly basis. By doing so, the Bluetooth data format will become same as of cell tower ID data as shown in Figure-4 and the methodology applied on cell tower ID data can be used with the Bluetooth data as well.

Another reason is that the results obtained from cell tower ID data and the Bluetooth proximity data can be compared at the end to see if we could get some interesting anomalies in behaviour of the user.

After analysing the Bluetooth proximity data, user's home computer device was given the name 'Home' (H). That means all those time slots in which user detect his home computer device, considered as 'H' because it shows user's presence in the home. For office, there are many devices that user detects during office hours. To obtain a group of devices that belong to the office, we remove the weekends from one month data and use Jaccard index [24], to detect how similar the detected devices are throughout the office hours for all remaining weekdays. Jaccard similarity equation is:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (2)$$

where, 'A' and 'B' are sets of detected devices in two consecutive days. At the start, 'A' and 'B' represent day-1

and day-2, then day-2 and day-3 and so on up to the all remaining weekdays that left after removing the weekends from one month of Bluetooth data.

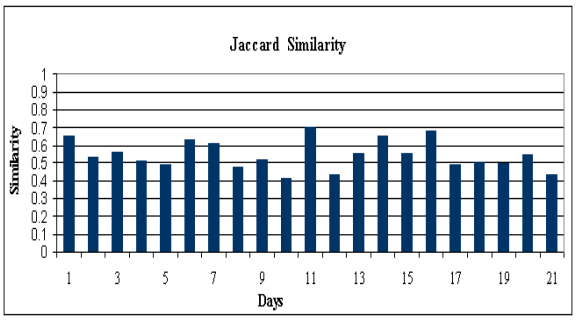


Figure 12. Jaccard Vertex Similarity

Figure-12 gives the similarity of detected Bluetooth proximate devices during the office hours between the pairs of consecutive days. The average similarity between the detected devices is above 0.5. This means there are many devices that user detects repeatedly during his office hours. All those devices that user detects for at least 70% of the days during office hours goes in 'Office' group. All other devices go in 'Other Devices' class.

After classifying the devices, a new data matrix is generated that contains twenty four time slots for each day as were in the case of cell tower ID data. Each time slot is assigned one of these classes (i.e., Home, Office, Other Devices, No Devices Found) depending upon the number of detections of the devices belonging to a specific class. Behaviour analysis frame work discussed in Section-4 is used to analyse the behaviour of low entropy user with Bluetooth proximity data using NN.

Figure-13 shows the Home/Work distribution of locations depending on the presence of user at different locations obtained from the Bluetooth data classes. The whole day is divided into twenty four time slots and each slot only represents one of the four classes depending upon the devices with which user spent most of his time.

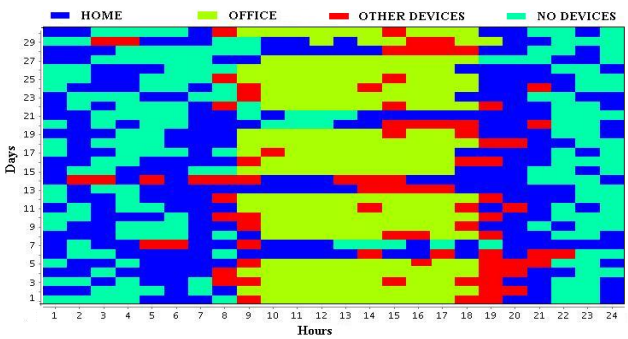


Figure 13. Distribution of Home/Work Transitions of Bluetooth Data

Figure-14 shows the fifteen days behaviour of the user detected from Bluetooth proximity data. An interesting observation can be made by analysing the behaviour detection results of both cell tower ID data in Figure-8 and Bluetooth ID data in Figure-14. It is observed that sometimes when behaviour detected from cell tower ID data is normal and no unusual pattern is detected, a change in behaviour or an unusual routine is detected from Bluetooth proximity data. For example on day-3, cell tower ID data shows normal behaviour in Figure-8, whereas an unusual routine is detected from Bluetooth proximity data on the same day shown in Figure-14. It can be said that it is more likely to be detecting unusual behaviour because during a regular routine of office hours of a weekday, user is supposed to detect 'Office Devices'. Cell tower ID data shows this as a normal behaviour because the user is in 'Office' where he should be normally. Whereas Bluetooth proximity data can be pointing towards some gathering or meeting of students or staff that is not part of the regular routine. Behaviour detected from Bluetooth proximity data can be pointing towards that activity.

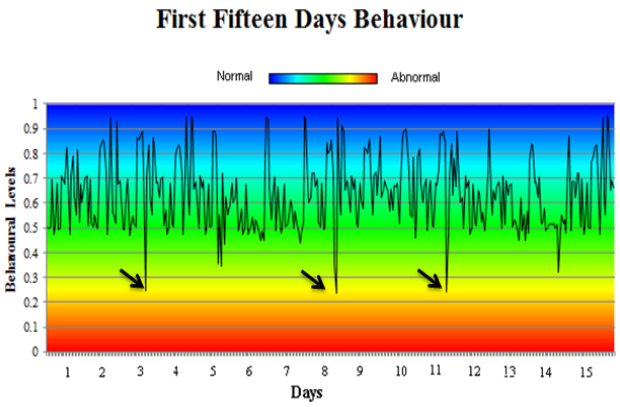


Figure 14. Fifteen Days Behaviour Using Bluetooth Proximity Data

VII. Further Analysis

So far, behaviour analysis obtained from contextual information of both GSM cell tower and Bluetooth proximity data have been presented. It is observed that the cell tower ID data gives high level behaviour of low entropy mobile people depending upon user's location and movement in different GSM cells. Unusual behaviour in these movement patterns can be obtained from the cell tower ID data. As GSM cells cover large physical area, it can only give user's location and does not give information about the low level activities such as attending the lecture, sitting in office with colleagues, going for shopping. On the other hand, Bluetooth proximity data gives information about other people and Bluetooth devices that are present in the close proximity of the user. It also gives information about the social relationships and most likely low level activities depending upon the detection of other proximate devices.

As the nature of contextual information obtained from cell tower and Bluetooth proximity data is different, it is interesting to analyse the difference of behaviour detected through this data statistically. For this purpose, we have used Kullback-Leibler (KL) Divergence [25], Kernel Density Estimation Function and Empirical Cumulative Distribution Function (ECDF). KL Divergence has been calculated and it gives the value of 0.4568. KL is a non-symmetric measure of the difference between two probability distributions as shown in Equation 3.

$$\mathrm{KL}(P||Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)} \qquad (3)$$

where $P(i)$ and $Q(i)$ is the value obtained by taking the histogram of cell tower ID data and Bluetooth proximity data respectively. The value obtained from KL shows that there is difference in the behaviour detected by cell tower ID data and Bluetooth proximity data. This supports the argument that it is possible to use only Bluetooth proximity data to detect some behaviour of low entropy people that deviates from the normal routine, although Bluetooth doesn't give strong information about the location of the user.

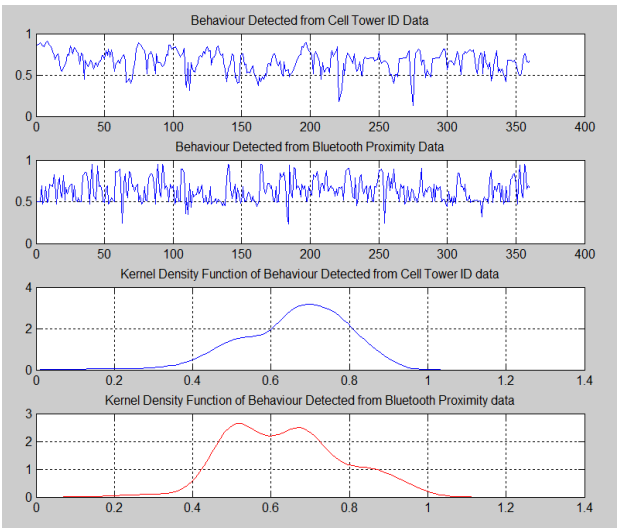


Figure 15. Kernel Density Estimation of Behaviour Detected from Cell Tower ID and Bluetooth Proximity Data

Figure-15 shows the Kernel Density Estimation [26] of two weeks of behaviour detected by both cell tower ID data and Bluetooth proximity data. Kernel Density Estimation is a non-parametric way of estimating the probability density function of a random variable. In our case, it estimates the probability density function of the behavioural values obtained by using both cell tower ID data and Bluetooth proximity data. The results in Figure-15 show that the Bluetooth proximity data show more unusual activities and routines. The peak shown around the behavioural value '0.5' in the Kernel Destiny Function of the Bluetooth data means that there are most likely many patterns in which users behaviour seems to be 'Low_Abnormal' means a little deviated from the normal routine. This shows that different routines and behaviour that deviate from the normal daily life routines of a low entropy user can be detected by using the Bluetooth proximity data.

In order to analyse difference of behaviour detected by using cell tower ID and Bluetooth proximity data in more detail, ECDF is also applied on the behavioural data. Empirical CDF is the cumulative distribution function associated with the empirical measure of the sample. Figure-16 shows the empirical distribution function applied on the behavioural data obtained using cell tower ID and Bluetooth proximity data. X-Axis shows the behavioural levels and Y-Axis gives the probability of exceeding the corresponding value on X-axis (Behavioural Levels). It shows the difference between the behaviour detected by cell tower ID and Bluetooth proximity data.

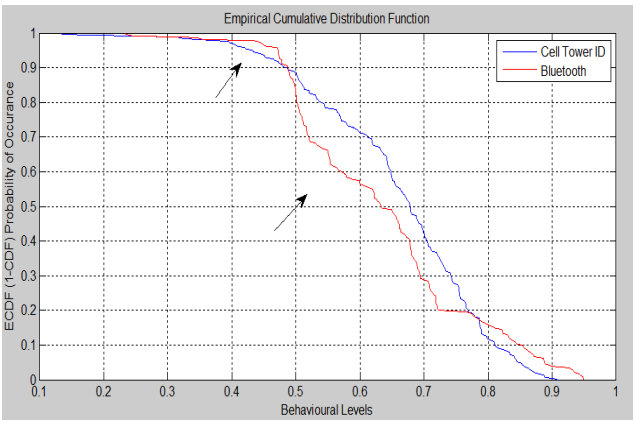


Figure 16. Empirical Cumulative Distribution Funtionc of Behaviour Detected from cell Tower ID and Bluetooth Proximity Data

Above mentioned results show that data collected from mobile devices such as cell tower ID and Bluetooth proximity data can be used for behaviour and routine activities detection. Bluetooth proximity data itself does not give much information about the location of the user but if this data is classified into different locations and user's movement patterns are obtained, then this data can give more insight into the behaviour of low entropy people. Proximity data gives information about the social relationship and activities that require social interaction of the users. If this data is used with the cell tower ID data, it can give extra information about the routines that deviates from the normal routine patterns. These results show that for low entropy users, the detection of unusual routines and behaviours by using only Bluetooth data is also possible. Low entropy users follow specific routines as compared to high entropy individuals, who live more diverse lives; therefor chances of detection of regular routines of low entropy users become more. This study aims to aid elderly people and patients to detect abnormal and unusual behaviours to avoid any accidents. Normally patients and elderly people have fixed and limited routines to follow that can likely be detected using Bluetooth devices by classifying the Bluetooth Proximity data into different activities or communities.

VIII. Summary and Future Work

In this paper, real time Bluetooth proximity and cell tower ID data is used to detect activities and routines of an individual that deviates from the normal daily life routines by using NN and DT. A low entropy user was selected for experiments due to the regularity and constancy in his routines. A successful detection of abnormal behaviour in this user's routines is done by using cell tower ID's and Bluetooth proximity data. NN and DT are used as decision engines to detect the behaviour of the user by using cell tower ID data. NN are found more accurate as compared to

the DT in detection. However; DT requires less data and takes less time for training.

Bluetooth proximity data is classified into four different categories by using Jaccard Index. To detect anomalies in more specific and lower level activities and routines, we need to classify the Bluetooth proximity data into temporal clusters. In future work, we will try to classify the Bluetooth proximity data on temporal scale to cover the minute details of the user's behaviour and will also try to predict the behaviour based on these classes and communities detection using pervasive computing`. This will provide one step further in the identification of unusual routines and activities by using only Bluetooth proximity data. This will help us to facilitate elderly people and patients who need more care and concern about their behaviour and unusual routines that can cause serious accidents.

ACKNOWLEDGMENT

This research work is funded by the Higher Education Commission of Pakistan under Faculty Development Program through University of Engineering & Technology Taxila, Pakistan.

REFERENCES

[1] Azam, M. A. Tokarchuk, L. Adeel, M. *"Human Behaviour detection Using GSM Location Patterns and Bluetooth Proximity Data"*. The Fourth International Conference on Mobile Ubiquitous Computing, Services and Technologies, pp. 428-433, Florence, Italy, 2010.

[2] Hermersdorf, M. Nyholm, H. Perkio, J. Tuulos, V. *"Sensing in Rich Bluetooth Environments"*- Workshop on WorldSensorWeb, in Proc. SenSys, 2006 - sensorplanet.org

[3] Eagle, N. Pentland, A. *"Reality mining: sensing complex social systems*". Personal and Ubiquitous Computing 2006 – Springer, Vol. 10, # 4, 255-268

[4] Gill, T. M. Desal, M.M. Evelyne, A. Holford, T. R. Williams, C. S. *"Restricted-Activity among Community-Living Older Persons: Incidence, Participants, and Health Care Utilization"*, Annals of Internal Medicine. http://www.annals.org/content/135/5/313.full.pdf+html

[5] Azam, M. A. Laurissa, T. (2009). *"Behaviour Detection Using Bluetooth Proximity Data";*. Proceedings of Networking & Electronic Commerce Research Conference, pp. 46-52 (NAEC 2009).

[6] Hara, K. Omori, T. Ueno, R. *"Detection of unusual human behaviour in intelligent house"*; Proceedings of the 2002 12th IEEE workshop on Neural Networks for Signal Processing, pp. 697-706, 2002.

[7] Yiping, T. Zhiying, Z. Hui, G. Huiqiang, L. Wei, W. Gang, X. *"Elder Abnormal Activity Detection by Data Mining"*, SICE Annual Conference in Sapporo, August 4-6, 2004, vol. 1, pp. 837–840 (2004) Japan

[8] Wren, C. Ivanov, Y. Kaur, I. Leigh, D. Westhues, J. "*SocialMotion: Measuring the Hidden Social Life of a Building"*. In: J. Hightower, B. Schiele, and T. Strang, (eds.) LoCA 2007. LNCS, vol. 4718, pp. 85–102. Springer, Heidelberg (2007)

[9] McCowan, I. Gatica-Perez, D. Bengio, S. Lathoud, G. *"Automatic Analysis of Multimodal Group Actions in Meetings"*. IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI) 27(3), 305–317 (2005)

[10] Stiefelhagen, R. Bernardin, K. Ekenel, H.K. McDonough, J. Nickel, K. Voit, M. Woelfel, M. *"Audio-Visual Perception of a Lecturer in a Smart Seminar Room*". In: Signal Processing - Special Issue on Multimodal Interfaces, vol. 86 (12). Elsevier, Amsterdam (2006)

[11] Eagle, N. "*Machine Perception and Learning of Complex Social Systems",* Ph.D. Thesis, Program in Media Arts and Sciences, MIT, June 2005

[12] Nicolai, T. Behrens, N. Yoneki, E. *"Wireless Rope: Experiment in social proximity sensing with Bluetooth"*. In fourth annual IEEE International Conference on Pervasive Computing, LNCS 4277, 2006

[13] Farrahi, K. Gatica-Perez, D. *"Daily Routine Classification from Mobile Phone Data*". In: Popescu-Belis, A., Stiefelhagen, R. (eds.) MLMI 2008. LNCS, vol. 5237, pp. 173–184. Springer, Heidelberg (2008)

[14] Farrahi, K. Gatica-Perez, D. *"What did you do today? Discovering daily routines from Large-Scale Mobile Data"*. In: MM 2008: Proceeding of the 16th ACM International Conference on Multimedia, pp. 849–852. ACM, New York (2008)

[15] Vukovic, M. Vujnovic, G. Grubisic, D. *"Adaptive User Movement Prediction for Advanced Location-aware Services"*, Proceedings of the 17th international conference on Software, Telecommunications and Computer Networks, pp. 343-347 (2009)

[16] Vukovic, M. Lovrek, I. Jevtic, D. "*Predicting user movement for advanced location-aware services"*. In 15th International Conference on Software, Telecommunications and Computer Networks, pp. 1–5. SoftCOM 2007, 2007.

[17] Tabia, K. Benferhat, S. *"On the Use of Decision Trees as behavioural Approaches in Intrusion Detection"*. In Seventh International Conference on Machine Learning and Applications, 2008. ICMLA'08.

[18] Hu, D. H. Zhang, X. Yin, J. Zheng, V.W. Yang, Q. *"Abnormal Activity Recognition Based on HDP-HMM Models",* AAAI Publications, 21st International Conference on Artificial Intelligence, pp. 1715–1720, 2009

[19] Eagle, N. Clauset, A. Quinn, J. A. *"Location Segmentaion, Inference and Prediction for Anticipatory Computing",* AAAI Spring Symposium, 2009

[20] Kotsiantis, S. *"Supervised Machine Learning: A Review of Classification Techniques*", Informatica Journal 31, pp. 249-268, 2007.

[21] Rumelhart, D. E. Hinton, G. E. Williams, R. J. *"Learning Internal Representations by Error Propapgation"*. In: Rumelhart D. E., McClelland J L et. al. (eds) Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT press, Cambridge, MA, Vol. 1, pp. 318-362.

[22] Linear Neural Networks, retrieved on 03-09-2011, retrieved from http://www.idsia.ch/NNcourse/linear2.html

[23] Quinlan, J. R. "*C4.5: Programs for Machine Learning*". Morgan Kaufmann San Francisco.

[24] Jaccard Index, retrieved on 03-09-2011, retrieved from http://www.statemaster.com/encyclopedia/Jaccard-index

[25] Kullback, S. Leibler, R. A. (1951). "*On Information and Sufficiency*". Annals of Methematical Statistics 22(1): 79-86.

[26] Rosenblatt, M. (1956). "*Remarks on some nonparametric estimates of a density function*". Annals of Mathematical Statistics 27: 832–837.

Monitoring Chronic Diseases Using Soft Computing Techniques and Rule-based Systems: The CHRONIOUS Case

Piero Giacomelli
Tesan S.p.A.,
email:giacomelli@tesan.it

Giulia Munaro
Tesan S.p.A.,
email:munaro@tesan.it

Roberto Rosso
Tesan S.p.A.,
email:rosso@tesan.it

***Abstract*—CHRONIOUS is an Open, Ubiquitous and Adaptive Chronic Disease Management Platform for Chronic Obstructive Pulmonary Disease(COPD) Chronic Kidney Disease (CKD) and Renal Insufficiency. It consists of several modules: an ontology based literature search engine, a rule based decision support system, remote sensors interacting with lifestyle interfaces (PDA, monitor touch screen) and a machine learning module. All these modules interact each other to allow the monitoring of two types of chronic diseases and to help clinician in taking decision for cure purpose. This paper illustrates how some machine learning algorithms and a rule based decision support system are used in the CHRONIOUS project, to monitor chronic patient. We will analyse how a set of machine learning algorithms can be used in smart devices to alert the clinician in case of a patient health condition worsening trend.**

***Keywords*-Telemedicine; chronic disease management; machine learning; soft computing techniques.**

I. INTRODUCTION

This notes extend the paper [1] that was presented at ETELEMED2011 [2]. Scientific advances over the past 150 years, particularly in the medical field, allowed the extension of life expectancy in western countries and this trend seems to increase in future years. Conservative estimates suggest that by 2030 in EU countries the proportion of people over 60 years regard the entire population will be around 50%; this means that we will see a gradual increase in the number of those subjects with chronic diseases (i.e., diseases not involving healing), that will therefore increase the cost and effort over health care facilities [3].

Such chronic diseases are slowing but constantly replacing malnutrition and infection as primary causes of mortality in the population [4]. The World Health Organization (WHO) has recently emphasized that chronic diseases are a global priority [5].It was calculated that, if governments where able to put in place public health policies that produce a 2% yearly reduction in mortality rates for chronic diseases, 36 million deaths could be prevented worldwide between 2005 and 2015 [6].

Reducing mortality rates caused by chronic diseases is also an economic priority, because it could save about 10% of the loss in income due to death and disability, which amounts to $8 billion in the developing countries only [7].

Chronic diseases are difficult to treat and, apart from deaths, have collateral social impact that are becoming an economic emergency both in western and developing countries. Considering the mean age growing in western countries population, chronic diseases will be a growing emergency in next years. As the number of patient with chronic diseases is rising there will be an increasing cost for hospitalization structure both public and private. Some specific diseases like Chronic Kidney Disease (CKD), sometimes there is, during the treatment of the disease, a non-return point from where the hospitalization is continuous as for dialysed people. The traditional approach that consists in periodic check-ups and periodic lab exams seems a model that won't be sustainable as the population gets older and the total number of patients with chronic diseases rises. At present, the physician deals with an increasing number of chronic patients that are lowering the periodic check-ups and so the reduced frequency is lowering the ability to prevent, if not death, worsening in patient's quality of life.

In the latest years, we have seen a tremendous growth in IT infrastructure, both from the hardware and communication capacity. Nowadays a common mobile phone is much more powerful in terms of hardware and software capacity than the first calculating machine that allowed the man to land on the moon forty years ago. The continuously growth of the World Wide Web (WWW) and, linked to this, the continuous growth in bandwidth capacity for data transmission allows to have cheaper and more widely available bandwidth, for larger portions of the population.

As a consequence of the exponential growth of hardware and software infrastructure, it is possible to rethink the whole approach to the treatment of complex chronic disease [8] [9] by limiting the hospitalization only to situation of severe worsening of patient condition. This was the original idea behind the EU funded CHRONIOUS project [10] [11]: constructing a generic platform to monitor, in an unobtrusive way, patient with chronic disease in two goals [12]:

- Improve the patients quality of life, by reducing as much as possible the hospitalizations.
- Allow the clinician a continuous monitoring the patients, both in standard and potential risk situations.

To gain this two goals, the CHRONIOUS platform has to integrate different technologies both hardware and software modules that need to interact among themselves. This paper is organized as follows: in the first section, the general structure of CHRONIOUS hardware and software modules are described. A deep analysis of the preprocessing algorithms covers the entire second section. Section three is dedicated to illustrate the machine learning algorithms. In last section, we will evaluate possible improvements to the Chronious intelligence system.

II. The CHRONIOUS system: an overview

CHRONIOUS system deals with (Chronic Obstructive Pulmonary Disease) COPD and (Chronic Kidney Disease) CKD. The two diseases were choosen mainly because they are ones the most difficult to treat. There is lot of literature that explain in detail why, but we can summarize in a brutal way by saying that this two diseases are greatly influenced by comorbidities. For example, a patient suffering from CKD tends to have also diabetes [13] and have an increased risk of cardiovascular disease. On the other side the connection between COPD and lung cancer is so deep that some author consider them two manifestation of the same pathology [14]. Even if it is true that not every COPD patient suffers of lung cancer or pulmonary neoplasia, respiratory diseases naturally lead to a fatigue of the cardiorespiratory system with obvious consequences on the health of the heart muscle, that clearly affect all the patient status.

Different comorbidities can lead to different treatments for the patients, and for sure a continuous monitoring of the patient conditions both in terms of measure taken from the patient himself (as, for example, glucose level for diabetic patients) and in terms monitoring general status could surely take advantage over traditional therapies. So the general ideas beside Chronious was to collect different physiological data for different diseases, in a way to allow a better evaluation on well agree indicators for the clinicians. Beside this data collection that is not a novel approach in remote patient monitoring a great attention was dedicated to create a system that is able to mimic the clinician period visits to alert the clinician in a case of a worsening trend. This means that Chronious is not a life saving system but one of the main focus was to create a system that is able to alert the clinician in a way to prevent an emergency hospitalization. The thing we were trying to avoid is to have as in traditional therapies a patient that is controlled 4/5 times a year directly by the doctor, and between two check-up, an emergency hospitalization is required because for example the patient change his diet without telling his clinician. For these different types of chronic diseases, according to the medical guidelines, it is important to monitor different data in a way to check patient health status and to activate suitable emergency alarms for the clinician (for the COPD: ECG, SPO2 and respiratory rate; for CKD: glucose level, body weight and blood pressure) in case of critical event.

The CHRONIUOS platform consists of many modules that act together

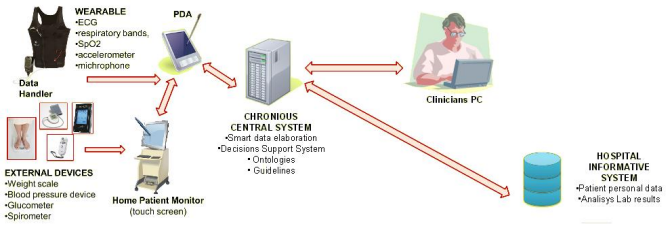


Figure 1. Chronious modules

We can organize them in three main frameworks

- Patient Sensor Framework
- Communication Framework.
- Monitor Framework.

The Patient Sensor Framework consists of hardware devices used to grab data from the patients. The hardware equipments, installed at patient's home, are

- Wearable and external devices.
- Touch screen Home Patient Monitor(HPM).
- Personal Digital Assistant (PDA).

The Wearable Device (WD) is a research project t-shirt equipped with the following sensors:

- a 3-lead Electrocardiogram (ECG).
- A microphone as a context-audio sensor.
- Two respiration bands (thorax and abdominal).
- An accelerometer.
- A sensor for measuring humidity as well as body and ambient temperature.

The external devices are a set of medical certified devices coupled with an Ambient Sensor (AS). The medical certified devices are:

- A weight scale.
- A blood pressure intake device.
- Two respiration bands (thorax and abdominal).
- A glucometer.
- A spirometer.

For the CKD patient, the devices provided are: weight scale, blood pressure device and glucometer. Every patient also interacts with a Touch screen medical PC for directly entering data for questionnaires about diet, psychological status and some question regarding the general patient status that mimic in a reduced way medical question posed by the clinician in a schedule examination(see Fig.2 and Fig.3).

All the data collected can be grouped in two types:

- Silent, when the data recording is automated and it does not involve the patient interaction, as respiratory frequency measurement for COPD patients from the wearable device

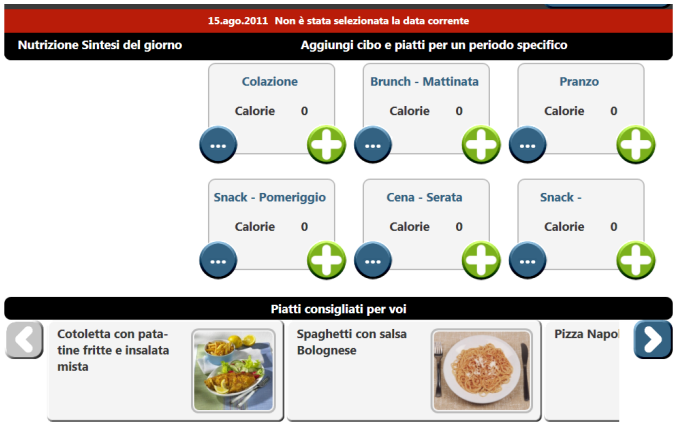


Figure 2. Diet questionnaire

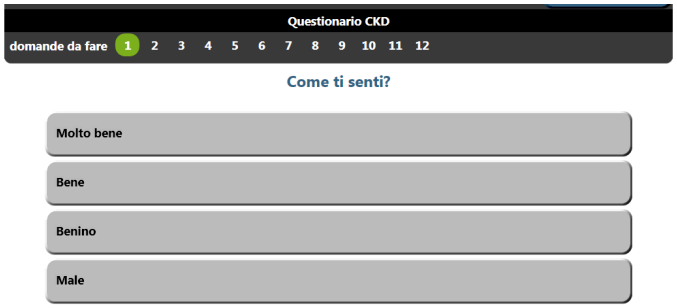


Figure 3. Doctor questionnaire

- no-silent, when it is requested a direct patient or caregiver interaction with the device, as for the blood pressure and questionnaires that are inserted on the HPM: for diet, activity and food intake monitoring. The no-silent data acquisition is particularly important for monitoring CDK patient lifestyle.

During the day there are several measurements, with different time intervals and different frequencies; only one data transmission is done, if there is no worsening in patient parameters. For COPD patient all the data are collected in a silent mode from the wearable t-shirt and transmitted to the PDA via Bluetooth; for CDK patient all the measurements, including a lifestyle questionnaire are stored in the HPM and transmitted in silent mode to the PDA. PDA is in charge of doing the following action

- Collect all the data from the devices.
- Use a set of machine learning algorithms to determinate if it is needed to force a non scheduled transmission to the Monitoring Framework.
- Transmit the collected and analysed information to the Monitoring Framework and receive back the changes, that will affect the interaction between the patient and the other devices.

The Communication Framework is in charge of transmitting data among devices and from PDA to Central DB. This transmission is done using messages in a predefined xml format. The device that is in charge of doing the transmission to the Monitoring Framework of the xmls is the PDA. Apart from the WD for which an ad-hoc proprietary binary transmission protocol has been developed, all other devices use a proprietary communication protocol. So the Communication Framework is also in charge to arrange all these difference in a uniform way. For example, for the weight scale, a proprietary software installed on the HPM is in charge of reading the measures from the weight scale. Once read the measure, the software stores it in a xml format that is different from the one used by the Communication Framework for data transmission to the Central System. Before communicating this measure, an xml transformation phase is necessary for standardizing the data an to tag it with a

device unique identifier that will be used by the Central System to associate measure to patient.

The Communication Framework is not only in charge of sending standardized xml to the central system but also once a day it connects to the Monitoring Framework to receive data from the central system. The data received are of two types:

- Frequency of measurements.
- Changes on drug intake for the patient.

Once the system is installed at patient's home, a standard set of measurements have been assigned to the patient (table I shows the frequency for CKD patient).

Table I
BASES FREQUENCY OF MEASUREMENTS FOR CKD PATIENT

Type	Day / Times per day
Glucometer	Monday once (8:00 am)
Blood pressure	Monday, Wednesday, Friday (8:00 am,5:00 pm)
Weight scale	Monday, Thursday (8:00 am)
Diet questionnaire	Monday (8:00 pm)
Clinical questionnaire	Monday (8:00 pm)

This schedule has been decided by the clinicians based on their experience in monitoring the chronic patient. But the clinician is free to change this schedule based on his experience and according to the suggestions provided by the Monitoring Framework. So, in case a patient needs to have an higher frequency of monitoring in some parameters, the clinician can decide to monitor them daily instead of the based three times per week. This allows a better monitoring of worsening trends in patient conditions. While the alerting system can transmit data from to the Central System without a schedule frequency, the data receiving phase is done once a day. As we underline above, Apart from measures the data

received cover also the drug intake. Being that the Central system store also the patient drug therapy, it is possible to change it to the patient. From the patient point of view all the data transfer is silent, the only interface is a software running on the patient HPM that every day display all the measurements that need to be done during the day and a drug reminder that display which pills need to be taken (see Fig.4).

Figure 4. HPM reminders

Move from the Communication Framework to the Monitoring framework consists we have four main parts

- Clinician interface.
- Decision Support System.
- Alerting system.
- Ontology Search Engine.
- Rule based editor.
- Interoperability engine.

Even if it this not possible in this paper to deeply describe every part we will give a brief description of each one, to allow to the reader to grab the complexity of the interaction involved in the Monitoring Framework.

Starting from the simple one, the clinician interface we can describe it as the main web interface for the clinician with the patient. Every clinician has a group of patient, determinated by the disease involved. For every patient, a detail health record is visualized. The clinician can see every alert the system have made for that patient, its history and how it was treated. Every measure that the central system received is visualized and the trend are plotted. Coupled with these information the clinician have the possibility as stated above to change frequency in measurements and drug intake. We can see the main screen in Figure 5.

The Decision Support System is a rule based engine that is able to receive as input nearly 100 different parameters about COPD/CDK patient condition and trends, parse every information and outcome a suggestion based on a set of if-then rules. The Decision Support System is coded as a web-service with different methods that is called by new

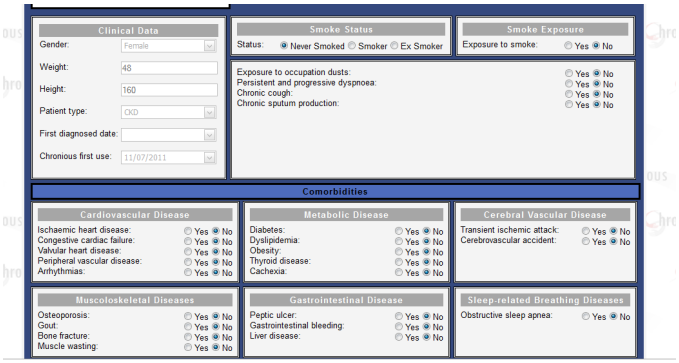


Figure 5. Clinician Interface

data insert into the Chronious Central DB. The web service use a set of JENA rules to infer a suggestion that will be displayed on the Clinician Interface. The mechanism is done using SPARQL [15] query language, on a set of xml based clinical treatment rules. As a basic example, one simple rule is the following one: if the body temperature of the patient is above 38 Celsius degrees the first aid department should be alerted for an hospitalization of the patient for monitoring purpose. The predefined set of rules codifies the clinician's expertise and the guidelines for the medical treatments of the COPD/CDK patients.

It is important to notice that even if the first triggering of an alert is done on the PDA, by the set of trained machine algorithm that will be described later, every alert is double checked by this engine that contains a larger set of rules. This because a new measure that comes to the Central System sent by the PDA could be somehow the first signal of a worsening trend in patient status. So if the PDA does the predefined one day transmission to the Monitoring Framework, even in this case the Decision Support System is called. This double check mechanism is particularly important for chronic patient because such patients can have a long period of stable conditions without any worsening but only with some fluctuations above the limits for an alert. In this situation the PDA will not reveal any worsening trend but in Decision Support System could output a suggestion on intensification of periodical check-up.

As we already pointed out, Chronious was not developed as a life saving system. However, the project faced the problem of a potential life risk situation. The alerting system is the answer to such problem. Every suggestion outputted from the Decision Support System is tagged with a level of severity. This level mimic the level used in first aid department of an hospital in case a chronic COPD / CKD request help.

These levels go from white to red, where red means that a patient is in danger of life. If the Decision Support System tags a suggestion as red the Alerting System sends an sms

to a number that fast reacts and eventually go to the patient home to see what is happening. Also, technical problems are arranged by the Alerting System, being that PDA use GPRS network to send data, we can have network problem so even if the data are already stored in the PDA the Central System cannot receive them. In this case, an email is sent to the technical team. This is possible because the Central Database contains the schedules of the measurements, so it is able to understand if a measure was missed. The Alerting System is based on a queue of alerts that are continuously monitored by a windows service and every new alerts is treated according to its severity and based on a set of configurable predefined rule.

Nowadays, the World Wide Web is the primary resource for medical knowledge. Nearly every clinician use it as a source for being up to date with the latest research information on how to treat diseases. Pubmed [16] is probably the first search engine used for such purpose. Lots of information are also stored in documentation that is present only in the medical structure, so we equipped Chronious with a an ontology search engine [17] that starting from raw document uses an personalized ontology to grab meanings for the two diseases covered by Chronious [18] [19] [19]. This with the goal of having a fast way for clinician, to find information related to COPD/CKD.

The ontology was created and fed with the standard guidelines for the treatment of COPD/CDK patient. Using an upload tool it is possible to enrich the original set of documents with new ones, to make more meaningful relationships between the various symptoms (see Fig. 6). It is interesting to notice that, at the time of writing, the first comparison between simple words search in Pubmed and Chronious Search Tool seems to indicate that the ontology search gives better result according to the clinician subjective feeling. This fact is in accordance with previous studies on the subject [20].

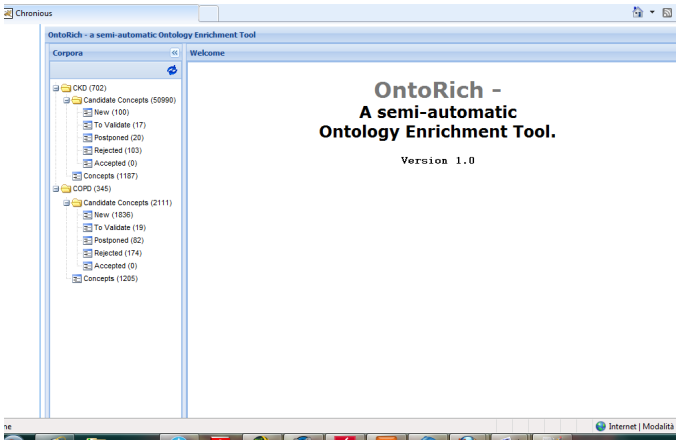


Figure 6. Ontology search

As we told the Decision Support System is a rule base engine based on a large set of if-then rules [21] [22] [23] [24]. The rules are codified using a JENA format [15] but the system gives the possibility to changes these rules according to the clinician intents. The interface that manages the rules, is the Rule Based editor that allow CRUD (Create, Read, Update, Delete) operations on the set of rules in a human readable format. The rule base editor is also equipped with a logical checker that controls if a new rule is in conflict with an existing one. Finally the Monitoring Framework is equipped with an interface written using HL7 messaging ANSI [25], to interface Chronious database with other medical structures. This was done because the medical structure use, in most cases, an RDBMS to store patient information and drug prescriptions. So this interface was needed to avoid duplicate entry by the doctor reducing the chance of errors. In the next section, we will analyse the preprocessing phase needed to activate the intelligence in the PDA that is the first decision system that activates the communication to the Monitoring Framework.

III. THE PDA CHRONIOUS PREPROCESSING PHASE

As we pointed out above, the Personal Digital Assistant is a smart phone equipped with WINDOWS MOBILE 6.5 Operating system, a SQL SERVER 2005 COMPACT database and a .NET FRAMEWORK 3.5. The PDA uses Bluetooth connection to send and receive data to and from HPM and Wearable t-shirt. It is equipped with a sim card for being able to do data transmission over GPRS network to the Central System. The data registered by the PDA are the following

- Data from the wearable jacket.
- Answers to questionnaires concerning dietary habits, drug intake and lifestyle of the patient.
- Data from the home patient monitor like blood pressure, glucometer measurements and body weight.

Once these data are collected, they are saved to the PDA database and a set of algorithms are triggered to analyse these data. Since the PDA analyses data for two different diseases, two sets of different algorithms are used. The fundamental data needed by COPD treatment are ECG signals and Respiration data, so in case of a COPD patient we have a first processing Electrocardiogram Pre-processing. After this, a Feature extraction phase is needed and at the end an Evaluation phase of the extracted features is done to determinate if an alert must be triggered to the Central System. For external devices used in particular by CDK patients, there is no need of a preprocessing phase because fundamental measures are the ones provided by the glucometer, the weight scale and the blood pressure measure; so, they are discrete time and directly used by the set of machine learning. Combined with these data, the answers to a set of queries concerning food intake, drug intake, lifestyle and mental status are passed to a set of machine learning algorithms to evaluate the whole patient condition. In the

next subsections, we will analyse first the COPD set of algorithms used.

A. Preprocessing of COPD signals

The aim of Preprocessing Phase is to improve the general quality of the ECG, for more accurate analysis and measurement, because there's the possibility to have some noises on the signals. Possible noises in the signal include

- Low frequency Base Line Wandering (BW) caused by respiration and body movements.
- High frequency random noises caused by mains interference (50 or 60Hz).
- Muscular activity and random shifts of the ECG signal amplitude caused by poor electrode contact and body movements.

The preprocessing comprises:

- Removal of base line wandering.
- Removal of high frequency noise.
- QRS detection.

The BW which is is an extrogenous low-frequency activity which may interfere with the signal analysis, rendering its clinical interpretation inaccurate and misleading. Two major techniques are employed for BW removal:

- Linear filtering [26]: involves the design of a LTI high pass filter with cut off in way that the clinical information in the ECG is preserved and the BW is removed as much as possible.
- Polynomial fitting [27]: includes the fitting of polynomials to representative points (knots) in the ECG, with one knot for each beat. Knots are selected from a silent segment, e.g., the PQ interval. A polynomial is fitted so that it passes through every knot in a smooth fashion.

The High Frequency Noise can be caused by the high frequency as well as power supply interference from the ECG signal. Its removal is done using:

- The Daubechies (DB4) wavelet employed on the basis of the resemblance and similar frequency response characteristics of the db4 basis function with the ECG waveform.
- Using wavelets to remove noise from a signal requires identifying which components contain the noise, using optimal methods to threshold them, and then reconstructing the signal using the thresholded coefficients.

The prepocessing phase finally deals with the QRS detection. The main features that should be calculated: the Inter-beat (RR) interval and the Heart Rate Variability (variation in the beat-to-beat interval). For the Inter-beat (RR) interval, two methods have been explored

- Filtering the ECG signal with continuous (CWT) and fast wavelet [28] transforms (FWT)[1].
- Following Pan-Tompkins [29], wavelets are used to remove noise from a signal requires identifying which component or components contain the noise, using optimal methods to threshold them, and then reconstructing the signal using the thresholded coefficients[2].

All the previous features are extracted from ECG signals. For COPD patient also the Respiratory Rate is a fundamental parameter that need to be analysed. In order to calculate the respiration rate using the reference respiration signal, a dominant frequency detection algorithm, based on short-time Fourier transform (STFT) [30], is applied.

The STFT is a localized Fourier transform, utilizing a Hamming window:

$$STFT(f(t)) = STFT(\omega, \tau) = \int_{\infty}^{\infty} f(t)w(t-\tau)e^{-j\omega t}dt \quad (1)$$

were $w(t)$ is the window function, commonly a Hann window or Gaussian hill centered around zero, and f(t) is the signal to be transformed. Because frequency components of the respiration signal are very low (2Hz), a window size of 60 seconds is selected. Every 60s, the hamming window is multiplied to the respiration signal, and the result is transformed to the frequency domain using Fourier transform. The dominant frequency is then detected by finding the maximum amplitude of the spectrums. When the dominant frequency components are found, inverse numbers are calculated in order to obtain the respiration rate. After this first preprocessing phase for COPD patients we wil now analyse the which kind of Features are extracted.

B. Features Extraction for COPD patients

From the Inter-beat (RR) interval and the Heart Rate Variability, several features can be extracted, either in time or in frequency domain.

Dealing with Time domain the values extracted are

1) SDNN(msec): Standard deviation of all normal RR intervals in the entire ECG recording using the following

$$sdnn = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(NN_i - m)^2} \quad (2)$$

[1]The reconstructed ECG signal after denoising contains only spikes with non-zero values at the location of QRS complexes. From this signal, the PQ junction and J point can be located as the boundaries of the spike. If the length of the spike is more or less than a predefined QRS length range it is annotated as noise and if the voltage is below a certain threshold, it is annotated as an artifact. The next stage is the detection of the T wave, and the P wave in the PQ interval. The peaks of Q, R and S waves are identified in the annotated part of the ECG signal from the PQ junction to J point.

[2]The algorithm includes a series of filters and methods that perform lowpass, high-pass, derivative, squaring, integration, adaptive thresholding and search procedures.

where NN_i is the duration of the i-th NN interval in the analysed ECG, n is the number of all NN intervals, and m is their mean duration.

2) SDANM(msec): Standard deviation of the mean of the normal RR intervals for each 5 minutes period of the ECG recording.
3) SDNNIDX (msec): Mean of the standard deviations of all normal RR intervals for all 5 minutes segments of the ECG recording.
4) pNN50 (intervals that are greater than 50 msec, computed over the entire ECG recording.
5) r-MSSD (msec): Square root of the mean of the sum of the squares of differences between adjacent normal RR intervals over the entire ECG recording the formula is

$$rMSSD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(NN_{i+1} - NN_i)^2} \quad (3)$$

where NNi is the duration of the i-th NN interval in the analysed ECG and n is the number of all NN intervals.

If we now move to the frequency domain the Feature Extraction on the PDA studies two bands:

1) The Low Frequency band (LF), which includes frequencies in the area $[0.030.15]$ Hz.
2) The High Frequency band (HF), which includes frequencies in the area $[0.150.40]$ Hz.

If we now move to the Respiration signal several features can be extracted either directly or indirectly we focused on:

1) Respiration Rate: The number of breaths per minute.
2) Tidal Volume (VT): The normal volume of the air inhaled after an exhalation.
3) Vital capacity (VC): The volume of a full expiration. This metric depends on the size of the lungs, elasticity, integrity of the airways and other parameters, therefore it is highly variable between subjects.
4) Residual volume (VR): The volume that remains in the lungs following maximum exhalation.

After all the preprocessing phase of the data gathered by the wearable devices, all these information are passed to a Classification System. The classification system is responsible for the analysis of the outcome from the preprocessing phase for the COPD patients and of the data gathered by the external devices and questionnaires entered by the patient himself. Below, we will see how the software is used to transform all these rich set of data in an information to be, in case, transmitted real-time or scheduled to the Monitoring Framework.

IV. The CHRONIOUS Classification System

After the collected data have been preprocessed for COPD patient and all the CKD patient input have been acquired, a set of machine learning algorithms are fired up to decide if a potentially risky situation is present. The aim of these tools is alerting the Central System that contains a rule based decision support system, for a better evaluation of the message triggered by the PDA. In case the message containing an alarm for life risk danger, the Central Decision Support System is able to alert the emergency staff or to suggest the clinician to modify the therapy approach. Most of these tools need a preprocessing phase for identifying the correct parameters that need to be validated by the clinicians. This means that in the first validation of the CHRONIOUS project a large amount of efforts has been dedicated to gather feedback from the clinicians about the correctness of the rules / parameters that have been inferred by the algorithms. In thise phase, another important effort has been dedicated by the technicians to evaluate some probability index for fuzzy data measures.

The CHRONIOUS Classification System is composed of the following part:

- A light rule based expert system.
- A supervised classification system.

The light rule based expert system is an xml parser that is able to extract from an xml a set of "'if then"' rules created an validated by the clinician. With these rules combined with the data collected, the rule base system is able to decide if a patient is in a potentially life-risk situation. For example the following rules will generate an immediate alarm to the central system:

- The Hearth Rate is above 120 bpm for both COPD and CKD patient.
- If the weight increase by 2% in the last 24 hour for CDK patient .

These rules are most for alarm triggering. It means that they aren't use for light monitor alerting. In the CHRONIOUS PDA system the Supervised Classification System is composed of the following machine algorithms:

- Support Vector Machines [31]
- Random Forest [32]
- Multi-Layer Perceptron [33]
- Decision Tree [34]
- Naïve Bayes [35]
- Partial decision Trees [36]
- Bayesian Network [37]

Apart from Bayesian Network, the other algorithms have been trained with a dataset to generate a set of rule that have been validated by clinician. The Bayesian Network have been used to identified possible rules about the mental and stress evaluator of the patient. This means that once trained, the rules generate can be used to identify is some stressful condition can alter some parameter and leading to a worsening of the general state of the patient. The use of these algorithms was needed because for the CKD patient the diet covers the most part of the medical treatment, so any factor that can influence a change on the diet intake,

would potentially and indirectly lead to a worsening of the patient condition. For example, if a female CKD patient feeld sad, these condition could lead her to eat a bigger piece of pie for satisfaction purpose. In general a bad feeling could lead chronic patients to be uncompliance with the medical treatment. However the kind of rules aren't liked the vital signs so their fuzzyness could be identified by these type of algorithms. Clearly while diet is important to avoid worsening on CDK patient conditions, on the other side the lifestyle could be and indirect cause of COKD worsening condition. Nevertheless for both of them we need static rules to have alarms sending, because for both for example having a body temperature above 38 could be a risky situation were an hospitalization is needed for both COPD and CKD patients. Considering the training dataset a set of 41 attributes have been identified. These data comes from the 2 sets of 2 hours health recordings (11 attributes), food input module (12 attributes), drug intake module (1 attribute), activity input module (2 attributes), questionnaire (13 attributes) and external device (2 attributes). The results of some classifier are shown in table II.

Table II
CLASSIFICATION SYSTEM: MAE: MEAN ABSOLUTE ERROR, RMSE: ROOT MEAN SQUARED ERROR, RAE : RELATIVE ABSOLUTE ERROR, CI: CORRECTLY/INSTANCES

Method	MAE	RMSE	RAE	CI
PART	0.1336	0.312	57.81 %	2.67
J48	0.1336	0.321	57.81 %	2.67
Forest	0.2	0.341	86.52 %	1.75
Naïve	0.127	0.343	54.95 %	2.67

Again we point out that even if there are some errors due to false positive matches, the PDA system in these case would only generate a rule that will send a message to the clinician that most of the times would only say that a light worsening is present. The core intelligence that deals with the central system would be the real suggestion system that would indicate to clinician a suggestion on how to act to possibly revert the trend. The Bayesian Network is the fundamental algorithm for the Mental support tool. It uses a set of attributes that affect a stress index and they weight based on the clinician's feedback as shown in table of Figure 2. When the total stress indicator is above a certain value a light alarm is triggered to the Central Database to inform clinician of a potentially worsening of patient conditions. In the same way a module in the PDA is in charge of the rules concerning the lifestyle od the patient: the lifestyle tool. It collects data entered by the patient or caregiver in a validation phase and using a Bayesian Network is able to compute an index of good or poor lifestyle of the patient also in this case if the poor lifestyle is find a light alarm is sent to the clinician for monitoring purpose.

During the first validation phase the training phase as been reprocessed several times as new data for the patient

Attribute	Different States	Probability of causing Stress (%)
Smoking	YES	90
	NO	10
Environmental Noise and rowded/Noisy places	High	70
	Medium	25
	Low	5
Hypoglycaemia	YES	85
	NO	15
Heart Rate	High	85
	Normal	15
Skin Temperature	Cool	35
	Sweat	65
Breathing asynchrony	YES	90
	NO	10
Sleep Disturbances (Questionnaires)	YES	62.5
	NO	37.5
Mood (Questionnaires)	Better	5
	Same	25
	Worst	70
Activity Comments	Feel sick, nausea	18
	Exhaustion, fatigue	23
	Discomfort in the chest, upper body, or jaw	23
	Irregular or extremely rapid heart beats	28
	None	8

Figure 7. Attributes

were collected. In fact, the knowledge that can be extracted using machine learning algorithms can be increased as new patients use the Chronious System.

V. CONCLUSIONS AN IMPROVEMENTS

In this paper, we presented a set of machine learning algorithms store in a smart phone that, combined with some external devices and patient specific data, can be used for a first monitor/alert system for treatment of patient affected by chronic diseases. Dealing with telemedicine application, these kinds of software, could help to improve patient quality of life and could be also a valid help for clinician to allow a more precise monitoring of patient conditions without need of the physical presence of the clinician. Apart these potentially advantages a PDA that equipped with these kinds of applications could suffer of some limitations. During developing phase we faced these problems:

- Heavy resource consumption of preprocessing algorithms.
- Updating a trained supervised algorithm.

The preprocessing algorithm for COPD parameter denoising is the most memory/CPU consumption. This can became a potentially problem when we deal with life risk situations, because while the algorithm denoise the ECG signal, the patient can loose consciousness and so precious time can be lost in these phase. Other time is lost due to the huge amount of signals transmitted from the wearable, this because we deal with an ECG signal that is composed of a mean of ten measurements per second so this means that 5 minute of ECG signal became nearly 3000 sql commands to a SQL SERVER 2005 COMPACT that is not so performing in this

case. Increasing hardware requirements of the PDA can be a first solution however it would be interesting to understand if a relational database is the best solution for storing these types of data or other structure would be more conformable for storing purpose.

PDA algorithms were developed using a Microsoft technology that at the time of writing have been exceeded by the cloud computing paradigm. For future purpose, having a compact RDBMS installed on the PDA, could became be an unnecessary requirement. The data collected could be transmitted directly to a Cloud storage and all the training algorithms could be moved on the cloud.

The other important issue is that once the supervised learning algorithms are trained any little change in some parameter will need to retrain the algorithms and most important, it would need a new validation of the outputs by the clinicians. This retrain can lead to difficulties when after this phase, the new trained algorithms need to be updated on the pda. The Chronious Communication Framework is only able to transmit values from the PDA to Monitor Framework and back, but cannot transmit back the trained configuration for the machine learning algorithms. In this case an interesting solution could be also to allow remote updating of the structure validated. For example a trained neural network could read the weights matrix from a structure upgradable by the Communication Framework. Apart from these improvements, and many others that could lead to a software system closer as much as possible to the clinician and patient needs, it is our opinion that with the smart devices that are closer to a normal PC, the algorithms presented in this paper could become an important part of the telemedicine platforms.

ACKNOWLEDGMENT

A reduced version of this paper was presented during [2] in the wonderful scenario of Le Gosier, Guadeloupe in the French Caribbean. The author that had the privilege to present his first research paper in such this, out of common world country, wishes to thank the IARIA organization for the unforgettable experience.

REFERENCES

[1] P. Giacomelli, G. Munaro, and R. Rosso, "Using soft computer techniques on smart devices for monitoring chronic diseases: the chronious case," February 23-28,2011, The Third International Conference on eHealth, Telemedicine, and Social Medicine, eTELEMED2011, Gosier, Guadeloupe, France, pp. 77-82, IARIA XPS Press, ISBN: 978-1-61208-119-9.

[2] "ETELEMED2011 web site," http://goo.gl/4HoLj, [retrieved: Dec, 2011].

[3] C. Zoccali, A. Kramer, and K. Jager, "Chronic kidney disease and end-stage renal disease a review produced to contribute to the report the status of health in the european union: towards a healthier europe," *NDT Plus*, vol. 3, no. 2, pp. 213–224, 2010.

[4] A. R. Omran, "The epidemiologic transition: A theory of the epidemiology of population change," *Milbank Quarterly*, vol. 83, pp. 731–757, 2005.

[5] W. H. Organization, "Preventing chronic diseases: a vital investment," http://goo.gl/umxCk, [retrieved: May, 2012].

[6] K. Strong, C. Mathers, S. Leeder, and R. Beaglehole, "Preventing chronic diseases: how many lives can we save?" *Lancet*, vol. 366, pp. 1578–1582, 2005.

[7] D. O. Abegunde, C. D. Mathers, T. Adam, M. Ortegon, and K. Strong, "The burden and costs of chronic diseases in low-income and middle-income countries," *The Lancet*, vol. 370, no. 9603, pp. 1929–1938.

[8] L. Bartoli, P. Zanaboni, C. Masella, and N. Ursini, "Systematic Review of Telemedicine Services for Patients Affected by Chronic Obstructive Pulmonary Disease (COPD)," *Telemedicine and e-Health*, vol. 15, no. 9, pp. 877–883, Nov. 2009.

[9] B. McKinstry, H. Pinnock, and A. Sheikh, "Telemedicine for management of patients with copd?" *Lancet*, vol. 374, no. 9691, pp. 672–673, 2009.

[10] "Chronious official web site," http://www.chronious.eu, [retrieved: Oct, 2011].

[11] M. van der Heijden, B. Lijnse, P. Lucas, Y. Heijdra, and T. Schermer, "Managing copd exacerbations with telemedicine," in *Artificial Intelligence in Medicine*. Springer Berlin / Heidelberg, 2011, vol. 6747, pp. 169–178.

[12] M. Vitacca, L. Bianchi, A. Guerra, C. Fracchia, A. Spanevello, B. Balbi, and S. Scalvini, "Tele-assistance in chronic respiratory failure patients: a randomised clinical trial," *European Respiratory Journal*, vol. 33, no. 2, pp. 411–418, 2009.

[13] A. J. Collins, J. A. Vassalotti, C. Wang, S. Li, D. T. Gilbertson, J. Liu, R. N. Foley, S.-C. Chen, and T. J. Arneson, "Who Should Be Targeted for CKD Screening? Impact of Diabetes, Hypertension, and Cardiovascular Disease," *American Journal of Kidney Diseases*, vol. 53, no. 3, pp. S71–S77, Mar. 2009.

[14] T. L. Petty, "Are copd and lung cancer two manifestations of the same disease?*," *Chest*, vol. 128, no. 4, pp. 1895–1897, 2005.

[15] "w3c sparql reference," http://www.w3.org/TR/rdf-sparql-query/, [retrieved: Aug, 2011].

[16] "Pubmed official site," http://www.pubmed.org, [retrieved: Sep, 2011].

[17] S. Kiefer, J. Rauch, R. Albertoni, M. Attene, F. Giannini, S. Marini, L. Schneider, C. Mesquita, and X. Xing, "An ontology-driven search module for accessing chronic pathology literature," in *On the Move to Meaningful Internet Systems: OTM 2011 Workshops*, P. H. R. Meersman, T. Dillon, Ed. Hersonissos, Crete, Greece: Springer Verlag, 2011, pp. 382–391.

[18] S. Kiefer, J. Rauch, R. Albertoni, M. Attene, F. Giannini, S. Marini, L. Schneider, C. Mesquita, X. Xing, and M. Lawo, "The chronious ontology-driven search tool: enabling access to focused and up-to-date healthcare literature," in *eChallenges e-2011 Conference*, M. C. P. Cunningham, Ed. IIMC International Information Management Corporation, 2011, pp. 1–8.

[19] P. Giacomelli, G. Munaro, and R. Rosso, "Can an ad-hoc ontology beat a medical search engine? the chronious search engine case." January 30- February 04,2012, The Fourth International Conference on eHealth, Telemedicine, and Social Medicine, eTELEMED2012, Valencia, Spain, pp. 215-220, IARIA XPS Press, ISBN: 978-1-61208-179-3.

[20] J. Z. Wang and F. Ali, "An efficient ontology comparison tool for semantic web applications," *Web Intelligence, IEEE / WIC / ACM International Conference on*, vol. 0, pp. 372–378, 2005.

[21] E. Seto, K. J. Leonard, J. A. Cafazzo, J. Barnsley, C. Masino, and H. J. Ross, "Developing healthcare rule-based expert systems: Case study of a heart failure telemonitoring system," *International Journal of Medical Informatics*.

[22] G. Kong, D.-L. Xu, R. Body, J.-B. Yang, K. Mackway-Jones, and S. Carley, "A belief rule-based decision support system for clinical risk assessment of cardiac chest pain," *European Journal of Operational Research*, vol. 219, no. 3, pp. 564–573, 2012.

[23] S. M. Maviglia, R. D. Zielstorff, M. Paterno, J. M. Teich, D. W. Bates, and G. J. Kuperman, "Automating complex guidelines for chronic disease: Lessons learned," vol. 10, no. 2, pp. 154–165, 2003.

[24] J. Chang, C. Ronco, and M. H. Rosner, "Computerized decision support systems: improving patient safety in nephrology." *Nature reviews. Nephrology*, vol. 7, no. 6, pp. 348–355, Jun. 2011.

[25] "HL7 official site," http://www.hl7.org, [retrieved: Jan, 2012].

[26] A. Kachenoura, F. Poree, G. Carrault, and A. I. Hernandez, "Comparison of four estimators of the 3d cardiac electrical activity for surface ecg synthesis from intracardiac recordings," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP '09, 2009, pp. 485–488.

[27] M. Kaur, B. Singh, and Seema, "Comparison of different approaches for removal of baseline wander from ecg signal," in *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*, ser. ICWET '11. ACM, 2011, pp. 1290–1294.

[28] Y. Chesnokov, D. Nerukh, and R. Glen, "Individually adaptable automatic qt detector," *Computers in Cardiology*, no. 2, pp. 337–340, 2006.

[29] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm." *IEEE Trans Biomed Eng*, vol. 32, pp. 230–236, 1985.

[30] P. S. Addison, "Wavelet transforms and the ecg: a review," *Physiological Measurement*, vol. 26, no. 5, p. R155, 2005.

[31] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel based learning methods*. Cambridge University Press, 2000.

[32] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[33] S. Haykin, *Neural Networks and Learning Machines (3rd Edition)*. Prentice Hall, 2008.

[34] J. R. Quinlan, "Learning decision tree classifiers," *ACM Comput. Surv.*, vol. 28, pp. 71–72, 1996.

[35] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, pp. 131–163, 1997.

[36] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," in *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998, pp. 144–151.

[37] D. Holmes and L. C. Jain, *Innovations in Bayesian Networks Theory and Applications*. Springer Netherlands, 2008.

Potential Antibacterial Targets in Bacterial Central Metabolism

Nichole Louise Haag, Kimberly Kay Velk, Chun Wu
Division of Natural Sciences
Mount Marty College
1105 W 8th St. Yankton, the United States
nichole.haag@mtmc.edu
kimberly.velk@mtmc.edu
cwu@mtmc.edu

***Abstract*—The emerging antibiotic resistant bacteria and their abilities for rapid evolution have pushed the need to explore alternative antibiotics less prone to drug resistance. In this study, we employed methicillin/multidrug-resistant *Staphylococcus aureus* (MRSA) as a model bacterial system to initiate novel antibiotic development. An *in silico* identification of drug targets in *MRSA* 252 strain and *MRSA* Mu50 strain respectively was described. The identified potential targets were classified according to their known or putative functions. We discovered that a class of essential non-human homologous, central metabolic enzymes falls into the scope of potential drug targets for two reasons: 1) the identified targets either do not have human counterparts or use alternative catalytic mechanisms. Based on major differences in active site structure and catalytic mechanism, an inhibitor of such a bacterial enzyme can be designed which will not inhibit its human cousin. 2) attacking bacterial energy-making machinery bypasses the usual drug resistance sites, paving the road to multi-faceted approaches to combat antibiotic resistance.**

Keywords-antibiotic resistance; Methicillin/multidrug-resistant Staphylococcus aureus; essential genes; drug targets; central metabolism.

I. INTRODUCTION

This paper is an extended version of the previously published conference paper [1]. The earlier paper detailed *in silico* identification of drug targets in MRSA 252 strain and MRSA Mu50 strain respectively and proposed that the development of a new class of antibiotics may be a potential solution to avoid bacterial antibiotic resistance.

One of the biggest medical breakthroughs of the twentieth century is the discovery of antibiotics [2], which was immediately followed by the unfortunate emergence of bacterial antibiotic resistance [3].The rapid rate of bacterial evolution to overcome the antibiotic action, the ability of a single pathogen strain to resist multiple drugs, as well as the stunning frequency of resistance occurring constitute a major challenge to the medical profession [3] and thus raised retrospective discussions of currently existing antibiotics [4-6]. Although there are hundreds of antibiotics on the market, it remains a fact that almost all existing antibiotics target only four cellular functions: protein synthesis, nucleic acid synthesis, cell walls synthesis or folate synthesis [6, 7, 8]. Bacterial resistance usually arises as the result of evolutionary adaptation of the target proteins that are subject to direct antibiotic attack [3]. Repetitively striking the same cellular sites leads to defensive bacterial gene mutation, which remains the primary cause of the prevalence of antibiotic-resistant bacteria [9], such as Methicillin/multidrug-resistant *Staphylococcus aureus* (MRSA) [10], Multidrug or Extensively Drug-Resistant Tuberculosis (MDR TB or XDR TB) [11,12], NDM-1 induced antibiotic -resistant *Escherichia coli* [13], *etc.* Hence, exploration of novel antibiotics with alternative modes of action is of great urgency [8]. The task falls on the shoulders of academia due to the fact that the pharmaceutical industry has ceased investing in antibiotic discovery owing to high cost, lengthening developing cycles, complexities and low profits along with failure of several recent investments in target-based approaches [14].

In this study, we employed MRSA as a model bacterial system because it is the most common bacterial pathogen isolated from humans [15, 16] with a significant morbidity and mortality [17]. The first MRSA case presented in the United Kingdom in 1961[18]. Shortly after, more variations were identified to be immune to β-lactam antibiotics (including penicillin, methicillin, oxacillin, and cephalosporins [19, 20]). Newly discovered MRSA strains have evolved to survive sulfa drugs, such as tetracyclines, and clindamycin [21]. Glycopeptide antibiotics, such as vancomycin and teicoplanin, considered drugs of "last resort", were used for the treatment of MRSA infections [22, 23]. However, recently discovered MRSA strains showed resistance even to vancomycin and teicoplanin [24, 25]. As of 2007, one variant found was resistant to six major kinds of antibiotics [26]. The beginning signs of MRSA infections are skin infections that resemble pimples, boils or spider bites. In immune-deficient patients, localized skin infections quickly spread through the bloodstream causing vital organ infections and possible death [27]. In a 2007 Centers for Disease Control and Prevention press release, there were about 94,000 cases of MRSA infections, contributing to around 19,000 deaths in the United States, which implies a mortality rate higher than that caused by HIV [28, 29]. The current treatment for MRSA infections is

still traditional broad-spectrum antibiotics such as lincosamides, sulfa drugs, glycopeptides [30-32], among which linezolids [33] daptomycin [34], Trimethoprim-sulfamethoxazole and MoxifloxacinHCl were considered relatively more effective [35, 36] though MRSA infections have become increasingly difficult to treat [31-33]. Thus, alternative treatments precisely targeting the root cause of MRSA infections needs to be established.

Novel antibiotic development starts with target screening [37]. In this paper, we reported the preliminary results of anti-MRSA drug development, *i.e.*, a systematic *in silico* approach for the identification of drug targets in two MRSA strains, MRSA 252 and MRSA Mu50 based on the following two criteria: essentiality to pathogen survival and absence from the human genome [38, 39]. A special list of enzymes targeting bacterial metabolism was identified, shedding light on a potentially new approach for antibiotic development.

II. METHODS

The objective of this study was to determine potential drug targets for alternative treatment of MRSA infections, to predict their enzymatic functions and to further shorten the list. We employed a reported *in silico* approach through a systematic and justified method [39, 40] for the identification of drug targets of MRSA infections following two criteria: essential to the survival MRSA and absent in the humans [38, 39].

MRSA genome National Center for Biotechnology Information (NCBI) gene bank contains at present complete genomic sequences of 13 MRSA strains. In this study, genomic sequences of MRSA 252 strain and MRSA Mu50 stain were studied respectively.

Sequence retrieval The genomic sequences of MRSA 252 and MRSA Mu50 were retrieved from the NCBI database respectively [41]. A total of 2656 genes from MRSA 252 strain and 2697 genes from MRSA Mu50 stain were purged at 90 % and 60% using CD-HIT [42] to remove paralogous or duplicate proteins.

Blasp against the database of essential genes (DEG) The resulting sequences were run through DEG [43**]** at an expectation (E-value) cutoff of 10^{-4}. The database of essential genes includes genes required for basic survival of *Staphylococcus aureus*, as well as more than 10 other bacteria, such as *E. coli*, *B. subtilis*, *H. pylori*, *S. pneumoniae*, *M. genitalium* and *H. influenzae,etc.*

Blasp against human genome The essential genes identified were subjected to BLASTP against the human genome (both refseq and nonrefseq) [44] to exclude any genes that have a significant match (E-value cutoff of 10^{-3} and lower) with human homologs. Genes having BLAST E-scores less than 10^{-3} were considered as having no close relatives in human.

Protein function assignment Information on the function of the identified proteins was derived from the annotated genome sequence through Integr8-Inquisitor [45] and/or EMBL/EBI/InterProScan [46].

Metabolic pathway study MRSA Metabolic pathways were obtained from KEGG database [47].

Amino Acid Alignment Analysis The interested protein sequences were submitted to SDSC Biology Workbench [48] for alignment in order to identify orthologs.

III. RESULTS AND DISCUSSION

The goal of this investigation was to determine potential drug targets for alternative treatment of MRSA infections and to classify and to analyze the identified targets. Out of the complete genomes of 13 MRSA strains that were sequenced and deposited in the NCBI gene bank, MRSA 252 and MRSA Mu50 were selected due to the fact that the former is a common strain in the USA [49] and the UK [50] and the latter, a methicillin and vancomycin resistant strain isolated in Japan [51] is commercially available for future molecular biological study (ATCC). The common method of drug target identification encompasses two steps: the identification of essential genes for bacterial viability [37] and the identification of genes absent in the human genome [38]. The former was performed by adopting the DEG database in our approach because this database compiles a list of all currently available essential genes in more than 10 prokaryotes including *Staphylococcus aureus* [41] and proved to be more accessible than conventional tools [39, 40]. On the other hand, the availability of the human genome sequence [52, 53] renders the latter step feasible. Following two newly published genomic analysis methods [39, 40], 2656 MRSA 250 and 2697 Mu50 genes were purged at 90 % and 60% using CD-HIT to remove

TABLE 1: GENOMICS ANALYSES OF MRSA 252 AND MRSA MU50 STRAINS REPECTIVELY.

Genes	**MRSA 252**	**MRSA Mu50**
Total number	2656	2697
Duplicates (>60% identical)	88	105
Non-paralogs	2568	2592
Essential genes [cut-off *E*-value < 10^{-4}]	499	496
Essential genes w/o human homologs[cut-off *E*-value < 10^{-3}]	133	134

TABLE 2. 133 ESSENTIAL, NON-HUMAN HOMOLOGOUS GENES IN BOTH MRSA 252 AND MRSA MU50 STRANS ENCODING DIFFERENT CLASSES OF PROTEINS AND THEIR KNOWN OR PUTATIVE FUNCTIONS

Categories	Classes	Groups	MRSA 252 *NCBI Gene Accession#*	MRSAMu50 *NCBI Gene Accession #*	Known or putative functions
Metabolism	Cellular respiration	Carbohydrate Catabolism	49482458	15923216	Formate acetyltransferase
			49482459	15923217	Formate acetyltransferase activating enzyme
			49482486	15923242	Xylitol dehydrogenase
			49483017	15923750	HPr kinase/phosphorylase
			49483247	15924074	Phosphoenolpyruvate-protein phosphatase ptsI
			49483033	15923765	Phosphoglyceromutase
			49483952	15924701	Acetate kinase
			49484267	15925031	Sucrose-6-phosphate hydrolase
			49484349	15925115	Fructose-bisphosphate aldolase
			49484367	15925133	Mannose-6-phosphate isomerase
			49484381	15925149	Mannitol-1-phosphate 5-dehydrogenase
			49484415	15925185	Galactose-6-phosphate isomerase subunit LacA
		Lipid Catabolism	49483384	15924216	Phosphatase/ dihydroxyacetone kinase
			49483425	15924288	Glycerol uptake operon antiterminator regulatory protein
		Amino acid catabolism	49482426	15923174	N-acetyl-γ-glutamyl-phosphate reductase
			49482779	15923539	N-acyl-L-amino acid amidohydrolase
			49483163	15923990	Thimet oligopeptidase homolog
			49483313	15924141	Glutamate racemase
			49483846	15924589	5'-methylthioadenosine nucleosidase/S-adenosylhomocysteine nucleosidase
			49484504	15925279	Urease subunit β
			49484120	15924869	Aminopeptidase ampS
			49484649	15925422	Glycerate kinase
			49484868	15925663	HisF cyclase-like protein
				15923177	Cystein Hydrolase
			49483520	15924318	Homoserine dehydrogenase
			49483584	15924384	Aspartate semialdehyde dehydrogenase
				15925319	Amino acid amidohydrolase
		Common metabolic pathway	49482818	15923578	Phosphotransacetylase
			49484161	15924909	Putative manganese-dependent inorganic pyrophosphatase
			49484002	57634637	Probable NAD(FAD)-utilizing dehydrogenase
	Bio-synthesis	Amino acid biosynthesis	49484873	15925668	Histidinol dehydrogenase
			49482425	15923173	Ornithine acetyltransferase
			49482586	15923346	5-methyltetrahydropter-oyltriglutamatehomo- cysteine methyltransferase
			49482696	15923462	Glutamate synthase, large subunit
			49483565	15924362	Tryptophan synthase β subunit
			49483583	15924383	Aspartokinase II
			49483655	15924456	Chorismate synthase
			49484279	15925043	dihydroxy acid dehydratase
			49484281	15925046	Ketol-acid reductoisomerase
			4948429	15925060	Alanine racease
			49484794	15925588	Pantoate--β-alanine ligase
		Fatty acid biosynthesis	49483392	15924219	Fatty acid/phospholipid synthesis protein
		Nucleotide biosynthesis	49482382	15923129	Phosphopentomutase
			49483421	15924248	Uridylate kinase
			49483664	15924468	Cytidylate kinase
		Cell wall biosynthesis	49484627	15925401	FemAB family protein
			49483567	15924364	FemA protein
			49482490	15923244	Teichoic acid biosynthesis protein (truncated TagF)
			49482939	15923673	Undecaprenyl Pyrophosphate Phosphatase
			49482995	15923728	UDP-N acetylenolpyruvoyl-glucosamine reductase
			49483182	15924008	UDP-N-acetylmuramoylalanyl-D-glutamate-2, 6-diaminopimelate ligase
			49484307	15925072	UDP-N-acetylmuramoylalanyl-D-glutamyl-2, 6-diaminopimelate-D-alanyl-D-alanyl ligase
			49484133	15924882	UDP-N-acetylmuramyl tripeptide synthetase

			49483346	15924173	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase
			49484348	15925114	UDP-N-acetylglucosamine 1-carboxyvinyltransferase
			49484309	15925074	Rod shape determining protein RodA
			49483587	15924387	Tetrahydrodipicolinate acetyltransferase
			49483980	15924730	UDP-N-acetyl-muramoyl-L-alanine synthetase
				57634647	UDP-N-acetylglucosamine 1-carboxyvinyltransferase
		Other biosynthesis	49482716	15923479	tetrapyrrole(corrin/porphy-rin) methylase
			49482722	15923485	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase
			49484013	15924759	Riboflavin biosynthesis
			49484795	15925589	3-methyl-2-oxobutanoate hydroxymethyltransferase
Transmission of genetic information	DNA replication, recombination and repair		49482254	15922991	Chromosomal replication initiation protein
			49482255	15922992	DNA polymerase III β subunit
			49482269	15923006	Replicative DNA helicase (DnaB-like)
			49483309	15924136	Excinuclease ABC subunit C
			49483633	15924434	Methyltransferase
			49483747	15924487	Integrase/recombinase
			49483811	15924552	DNA primase
			49483834	15924577	DNA polymerase III subunit delta
			49483926	15924674	Primosomal protein DnaI
			49483944	15924693	DNA polymerase III, β chain
			49484385	15925153	DisA bacterial checkpoint controller nucleotide binding
	Transcription and RNA processing		49483418	15924245	Transcriptional repressor CodY
			49483550	15924347	Transcription antiterminator
			49484097	15924845	SpoU rRNA methylase family protein
			49484908	15925703	Ribonuclease P
			49483433	15924260	Ribosome-binding factor A
			49483855	15924600	Transcription elongation factor
			49482590	15923350	Transcription terminator
			49483976	15924726	Catabolite control protein A
	Translation and posttranslational modifications		49483000	15923733	peptidase T
			49483039	15923772	SsrA-binding protein
			49483384	15924211	Hypothetical translation and posttranslational modifications
			49483609	15924409	Gcn5-related acetyltransferases
			49483778	15924518	Elongation factor P
Trans-membrane Proteins	Antibiotic Resistance		49482275	15923012	Metallo- lactamase
			49483344	15924171	Penicillin-binding protein
	Regulation		49483168	15923996	GTP pyrophosphokinase
			49483425	15924252	Zinc metalloprotease yluc
	Transport		49482431	15923179	Glucose-specific PTS, IIABC component
			49482476	15923232	PTS, IIBC component
			49482956	15923690	Gructose-specific PTS, IIABC component
			49483966	15924716	N-acetylglucosamine specific PTS, IIABC component
			49484378	15925146	Mannitol-specific PTS, IIBC component
			49484380	15925148	Mannitol specific PTS, IIA component
			49484538	15925313	PTS, arbutin-like, IIBC component
			49484739	15925528	Glucose-specific PTS, II ABC component
			49484838	15925631	PTS, IIABC component
			49483148	15923977	Oligopeptide transport system permease protein
			49484706	15925495	Gluconate permease
			49482866	15923628	Teichoic acid ABC transporter permease
			49484434	15925210	Cobalt transport protein
			49484516	15925291	Na^+/H^+ antiporter
			49484891	15925688	Nickel transport protein
			49484846	15925639	Bifunctional Preprotein translocase subunit SecA
			49483881	15924627	Bifunctional preprotein translocase subunit SecD/SecF
			49483265	15924092	Spermidine/putrescine-binding protein precursor homolog
			49482314	15923062	Potassium-transporting ATPase subunit A
			49482353	15923100	L-lactate permease homolog
			49484303	15925067	potassium-transporting ATPase subunit A
			49484446	15925220	Preprotein translocase subunit SecY
			49483071	15923829	ABC transporter substrate-binding protein
			49483075	15923833	ABC transporter-associated protein
			49483078	15923836	ABC transporter-associated protein

Other Proteins	Carrier proteins	49483175	15924003	Sodium/proton-dependent alanine carrier protein
		49482688	15923454	Lipoprotein
	Regulation	49482271	15923008	Response regulator protein
	Cell division	49482736	15923499	C ell division
		49483349	15924176	C ell division protein FtsZ
		49484905	15925700	Glucose-inhibited division protein B
	Other	49484374	15925142	Haloacid dehalogenase-like hydrolase
		49484612	15925386	Nitrate reductase β chain
		49484613	15925387	Respiratory nitrate reductase alpha chain
	Unknown function	49482472	15923228	Unknown
		49483005	15923738	Unknown
		49483022	15923755	Unknown
		49483024	15923757	Unknown
		49483035	15923767	Unknown
		49483546	15924343	Unknown
		49483928	15924676	Unknown
		49484792	15925584	Unknown

paralogues, respectively. The resulting 2568 MRSA 250 and 2592 Mu50 sequences were run through the database of essential genes (DEG) at an expectation cut-off of 10-4, yielding 499 and 496 essential genes respectively. Those 499 and 496 essential genes identified were subjected to BLASTP against the human genome [52, 53] to exclude any genes that have a significant match (E-value threshold of 10-3 and lower) with human homologs. Consensually, 133 MRSA 252 and 134 Mu50 genes respectively were considered as having no close relatives in humans. The results are summarized in table 1. Their known or putative functions annotated by Integr8-Inquisitor [46] and/or EMBL/EBI/InterProScan [47] are listed in table 2.

Among the 133 and 134 essential non-human homologous genes in MRSA 252 and Mu50 strains, respectively, 133 encode proteins that are well conserved between the two strains. Out of this conserved set, 63 are involved in metabolism, 24 participate in the transmission of genetic information, 29 represent transmembrane proteins, 9 have other functions such as regulation cell division and carrier proteins, *etc*., and 8 have unknown functions.

Our approach identified 14 genes in cell wall biosynthesis, most of which were validated by other research groups [54-56]. Among them, 6 are involved in the elongation of peptidoglycan, in agreement with previous studies [54, 55]. FemA family proteins are currently considered novel anti-staphylococcal targets due to the fact that they are involved in cell wall biosynthesis and expression of a methicillin resistance gene [56]. They are found to be essential in both MRSA 252 (NCBI Gene Accession#: 49484627 and 49483567) and Mu50 (NCBI Gene Accession#: 15925401 and 15924364) strains by our approach. Gene GI#49484133 in MRSA 252 and GI#15924882 in Mu50 respectively represent Staphylococcus aureus murE gene encoding UDP-N-acetylmuramyl tripeptide synthetase, which was demonstrated to be essential in *Staphylococcus aureus* through a method incorporating an IPTG controllable promoter [57].

Although the cell wall has long been considered an attractive target for antibiotic development because of its absence in humans, what should not be overlooked is that one of the most common antibiotic resistance mechanisms is the metamorphosis of cell-wall proteins, leading to antibiotic resistance. For example, β- lactam resistance was attributed to the expression of a group of cell wall penicillin-binding proteins (PBP-2') encoded by the mecA gene [58, 59]. Glycopeptide resistance is also considered to be caused by cell wall thickening resulting in binding vancomysin extracellularly [60, 61] and/or alteration of the drug-acting site in the cell wall from D-alanine-D-alanine to D-alanine-D-lactate owing to the expression of *vanA* resistance gene [62]. Hence, for novel antibiotic development, substances that anchor in sites other than the bacterial cell wall may have more potential because resistance usually arises as the result of gene mutation on the target proteins that are subject to direct antibiotic attack [63]. A 2006 review on mechanisms of bacterial antibiotic resistance suggested the exploration of novel antibiotics with alternative mechanisms of action [5].

Genes involved in transmission of genetic information including DNA replication, recombination and repair, transcription and RNA processing, translation, post-translational modification remain viable targets for antibacterial agent development [39, 40]. Our approach identified 24 of these candidate genes.

Our approach identified 29 membrane bound proteins. A recent review of anti-MRSA drug development indicated that agents anchoring in the bacterial membrane (*e.g.*, ceragenins and lipopeptides) showed great bactericidal effect and less prone to drug resistance due to the inability of bacteria to modify their targeted cellular sites in a way that is compatible with their survival [64]. Among this pool of proteins, 19 are involved in membrane transport, which represent valid drug targets

because pathogens usually lose their biosynthetic capabilities and rely on their hosts for the supply of essential nutrients [65, 66]. Thus, certain membrane transport proteins are of great importance in maintaining pathogen viability.

Our approach identified 30 energy metabolic (*i.e.* cellular respiration) genes in both MRSA 252 and MRSA Mu50, which are essential to staphylococcal survival with *E*-score < 10^{-4} but absent in human genome with *E*-score <10^{-3}. Currently there are limited numbers of commercially available antibiotics targeting energy metabolism. Those existing are mainly biological reagents such as oligomycin [67] and pesticides or piscicides such as antimycin A [68], not commonly used for humans because they affect both bacterial and human cells. Surprisingly, nature has provided us with a group of energy metabolic enzymes which are essential to pathogen survival while absent in humans. The differentiation lies in that those enzymes function through alternative mechanisms other than their counterpart enzymes in humans. Accumulating *in vitro* [69] and *in vivo* [70] evidence suggests that enzymes catalyzing bacterial cellular respiration with differentiated mechanisms of action are promising targets for novel antibiotic development. The inhibitors against those enzymes are able to hinder bacterial growth by inhibition of those enzymes without interfering with their human cousins. Most importantly, attacking bacterial energy-making machinery bypasses the usual bacterial mutation

TABLE 3 - POTENTIAL CENTRAL METABOLIC DRUG TARGETS FROM MRSA MU50 BASED ON DATA BASE OF ESSENTIAL GENES (DEG) HOSTED RECORDS OF CURRENTLY AVAILABLE ESSENTIAL GENES.

Class	MRSA 252	MRSA Mu50	Known or putative function	EC #	Identity with DEG genes				human homolog or ortholog
	Accession #	Accession #			Organism	E-Value	% Identity	% Similarity	
Carbohydrate Catabolism	SAR0217	SAV0226	Formate acetyltransferase	2.3.1.54	*H.influenzae* Rd KW20	0	62%	77%	No
	49482486 SAR0247	15923242 SAV0252	Xylitol dehydrogenase	1.1.1.137	*S. aureus* NCTC 8325	1e-163	80%	91%	No
	49483017 SAR0814	15923750 SAV0760	HPr kinase /phosphorylase	2.7.11.- 2.7.4.-	*S. aureus* NCTC 8325	1e-155	95%	95%	No
	49483247 SAR1057	15924074 SAV1084	Phosphoenolpyruvate-protein phosphatase	2.7.3.9	*S. aureus* N315	0	97%	97%	No
	49483033 SAR0831	15923765 SAV0775	Phospho-glyceromutase	5.4.2.1	*S. aureus* NCTC 8325	0	97%	97%	No
	49483952 SAR1789	15924701 SAV1711	Acetate kinase	2.7.2.1	*S. aureus* N315	0	92%	92%	No
	49484349 SAR2213	15925115 SAV2125	Fructose-bisphosphate aldolase	4.1.2.40	*S. aureus* N315	1e-163	100%	100%	No
	49484381 SAR2247	15925149 SAV2159	Mannitol-1-phosphate 5-dehydrogenase	1.1.1.17	*S. aureus* N315	0	100%	100%	No
	49484415 SAR2286	15925185 SAV2195	Galactose-6-phosphate isomerase subunit LacA	5.3.1.26	*S. aureus* N315	1e-76	100%	100%	No
	49482818 SAR0594	15923578 SAV0588	Phospho-transacetylase	2.3.1.8	*S. aureus* NCTC 8325	1e-169	93%	93%	No
Lipid Catabolism	49483425 SAR1238	15924288 SAV1298	Glycerol uptake operon antiterminator regulator	Un-classified	*S. aureus* N315	7e-98	100%	100%	No
Protein Catabolism	49483313 SAR1123	15924141 SAV1151	Glutamate racemase	5.1.1.3	*S. aureus* N315	1e-154	100%	100%	No
	49484504 SAR2373	15925279 SAV2289	Urease subunit β	3.2.2.16	*S. aureus* N315	2e-77	100%	100%	No
	49484120 SAR1969	15924869 SAV1879	Aminopeptidase ampS	3.4.11.-	*S. aureus* N315	0	100%	100%	No
Common metabolic pathway	49484161 SAR2012	15924909 SAV1919	manganese-dependent inorganic pyrophosphatase	3.6.1.1	*S. aureus* NCTC 8325	1e-172	100%	100%	No

sites for drug resistance [71-72]. Hence, exploration of antibiotics targeting alternative cellular functions such as central metabolic pathways may be a promising direction, and selective inhibition of targets specific to bacterial energy metabolism may be a potentially efficacious alternative in the treatment of MRSA infections. The enzymes on the higher priority list include *MRSA* fructose-bisphosphate aldolase, *MRSA* acetate kinase, *MRSA* phosphotransacetylase, *MRSA* formate acetyltransferase *and MRSA* xylitol dehydrogenase, *etc.* (table 3), which either do not have human homologues or adopt dramatically different catalytic mechanisms compared to their human cousins.

MRSA fructose-1, 6-diphosphate aldolase (NCBI Gene Accession#: 49483952 and 15924701 respectively) showed a 100 % match to both *Staphylococcus aureus* NCTC 8325 and *Staphylococcus aureus* N315 in Database of Essential Gene (DEG) with an identical expectation value of e^{-163} [73,74], suggesting the essential nature of this protein. It is well known that FBPA is one of the key enzymes in the glycolytic pathway that involves the breakdown of glucose [75]. FBPA is divided into two classes based on structural properties and catalytic mechanisms [75, 76]. Class I FBPA is mainly found in higher order organisms (*e.g.*, humans and animals). Catalysis in class I FBPA proceeds via a Schiff base intermediate formed by an active site lysine residue [75]. Class II FBPA is usually found in yeasts, bacteria, fungi, and parasites [76]. Catalysis in class II FBPA centers on the participation of a Zn (II) cofactor that coordinates to an enolate anion intermediate [76]. Based on major differences in active site structure and catalytic mechanism, an inhibitor of class II FBPA can be designed which will not inhibit class I FBPA. Thus, class II FBPA has long been considered as potential drug target in the development of antibiotics [77]. Multiple alignment of the sequence of MRSA FBPA with class II giardia FBPA and class I human FBPA was shown in Figure 1. *MRSA* FBPA (NCBI Gene Accession#: 49484349 and 15925115 respectively) exhibits 40.8% sequence identity to class II *giardia* FBPA while it exhibits only 18.8 % sequence

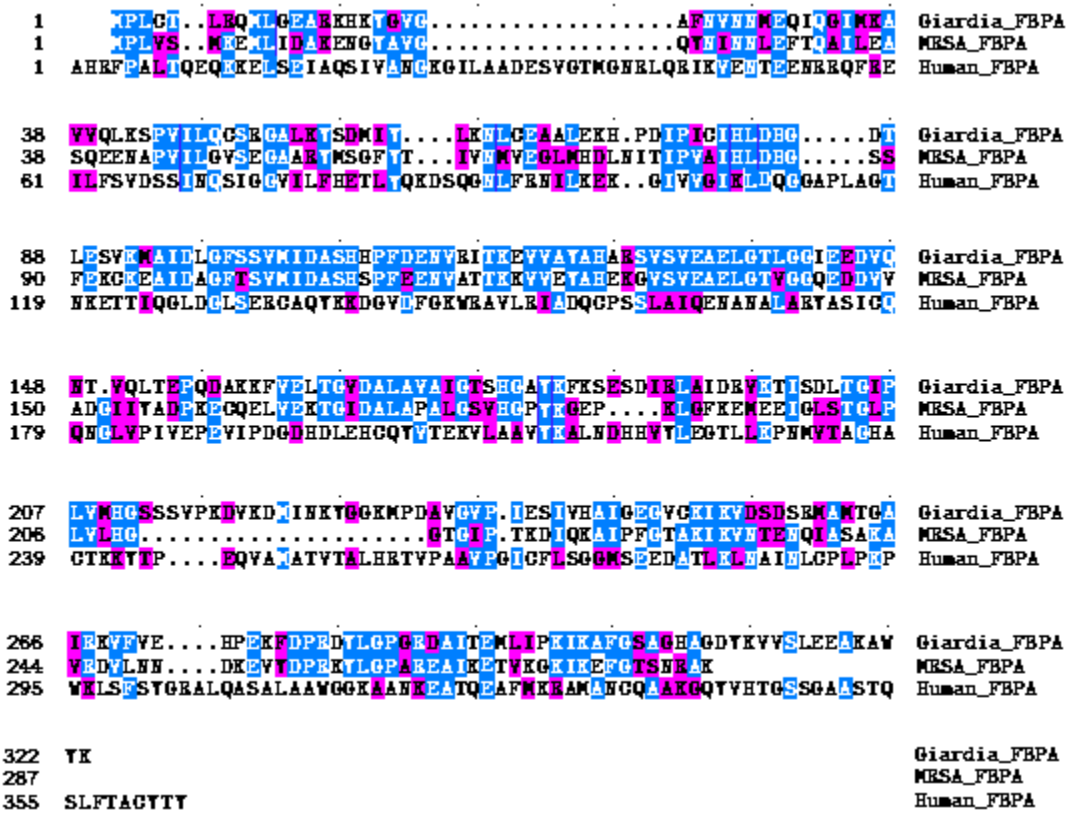


Figure.1 Alignment of the amino acid sequences of *MRSA* FBPA (NCBI GENE ACCESSION#:49484349 and 15925115 respectively) with class II *giardia* FBPA (2ISV) and class I *human* FBPA (1QO5). Numbering of the amino acids is indicated on the left. Identical amino acid residues in the alignment are indicated in light-blue shading and similar amino acid residues are indicated in purple shading. Gaps introduced during the alignment process are indicated as dots.

identity to class I *human* FBPA. Thus, *MRSA* FBPA should be putatively classified into class II FBPA, not class I FBPA. We have cloned and purified and characterized *MRSA* FBPA (unpublished result). Validation of the essential nature of class II *MRSA* FBPA through allelic replacement and inducible expression is underway in our research group.

MRSA acetate kinase (NCBI Gene Accession#: 49483952 and 15924701 respectively) demonstrated a 92 % match to *Staphylococcus aureus* N315 in Database of Essential Gene (DEG) with an expectation value of 0 [73, 74], suggesting the essential nature of this enzyme. Acetate kinase catalyzes the reversible phosphorylation of acetate to synthesize acetyl phosphate by transfer of a phosphoryl group from ATP. Acetate kinases are wildly distributed among prokaryotes [78] and some eukaryotes [79]. In aerobic conditions, this enzyme converts acetate to acetyl-CoA, a key intermediate in TCA cycle [80]. In anaerobic conditions, it plays a central role in synthesizing ATP from acetyl phosphate [81]. Prokaryotic acetate kinases are highly conservative. *MRSA* acetate kinase exhibits 44.6 % sequence identity to *E. coli*, 48.5 % sequence identity to *Salmonella typhimurium* acetate kinase, 51.3 % sequence identity to *Methanosarcina themophia* acetate kinase and 52.0 % sequence identity to *Lactobacillus sanfranciscensis* acetate kinase (Figure 2). Smith group has confirmed that it is a key enzyme in bacterial metabolism in a number of important fungal and protozoan pathogens. (*e.g.,* fungus *Cryptococcus neoformans* and protist *Entamoeba histolytica*). Its absence in humans suggests that it may also be a possible drug target [82]. We have cloned, purified and characterized *MRSA* acetate kinase (unpublished result). The development of crystal structures of *MRSA* acetate kinase is in progress at the laboratory of our collaborator Dr. Scott Lovell at the University of Kansas, which will allow us to preform structure-activity analysis as the basis of rational inhibitor design.

MRSA phosphotransacetylase (NCBI Gene Accession#: 49482818 and 15923578 respectively) demonstrated a 93 % match to *both Staphylococcus aureus* NCTC 8325 and *Staphylococcus aureus* N315 in Database of Essential Gene (DEG) with an identical expectation value of e^{-169} [73, 74], suggesting the essential nature of this enzyme. The gene encoding *MRSA* phosphotransacetylase has been cloned and the enzyme has been expressed in *E. coli* and purified. Kinetic assay of this enzyme is in progress.

Overall, we proposed that a class of essential, central metabolic enzymes, such as *MRSA* fructose-bisphosphate aldolase, *MRSA* acetate kinase, *MRSA* phosphotransacetyl-ase, *MRSA* formate acetyltransferase and *MRSA* xylitol dehydrogenase, *etc.* (table 3), which either do not have human homologues or functionally differentiate themselves from their human counterparts, are promising antibiotic drug targets. Because of the alterations in active site structure and mode of action of such a bacterial enzyme *v. s.* its human cousin (if there is one), through rational inhibitor design, an inhibitor of this enzyme can be designed which will not inhibit its human cousin. Nevertheless, this central metabolic inhibitor approach potentially decreases the risk of bacterial resistance against the antibacterial agents in that it bypasses the cellular sites where currently existing antibiotics regularly attack. In other words, since those cellular sites have not been repeatedly exposed to antibacterial agents, central metabolic inhibitors should be less prone to drug resistance induced by evolutionary adaptation.

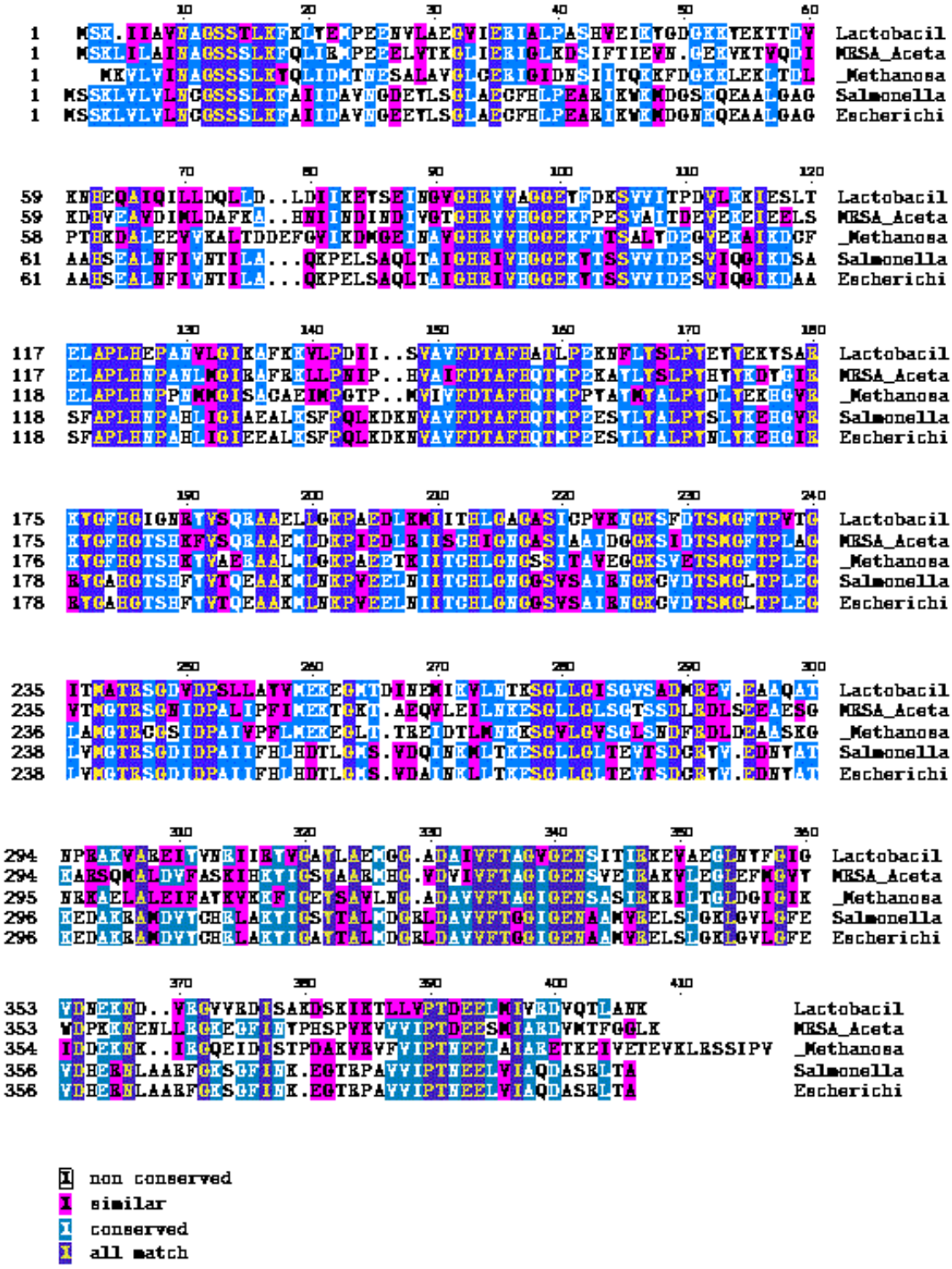


Figure.2 Alignment of the amino acid sequences of *MRSA* acetate kinase with *E. coli*, *Salmonella typhimurium* ,*Methanosarcina themophia* ,*Lactobacillus sanfranciscensis* acetate kinases. Numbering of the amino acids is indicated on the left. Identical amino acid residues in the alignment are indicated in light-blue shading and similar amino acid residues are indicated in purple shading. Gaps introduced during the alignment process are indicated as dote

IV. CONCLUSION AND FUTURE WORK

One of the crucial steps in narrow-spectrum antibiotics development is target identification. In this study, a putative set of candidate drug targets were elucidated by an *in silico* approach. The candidate genes are hypothetically required for survival of the candidate microorganisms and have no close human analogues. Many identified targets have been experimentally validated [56-59, 83-88]. By shortening the list of potential drug targets to a small pool of genes, the data presented in this paper facilitated our group and, may also aid other researchers in pursuing target validation and target characterization for alternative treatment of MRSA infections. Future directions include using a combination of kinetic assay and crystal structure development for enzyme characterization such as substrate recognition, catalytic site identification and reaction mechanism elucidation. Using ration drug design, tight-binding inhibitors will be designed followed by organic synthesis and *in vitro* evaluation. Once a nanomolar level inhibitor with high specificity is identified, development of X-ray crystal structures of enzyme-inhibitor complexes will be performed for further optimization. In principle, the premise is that the inhibitors of these targets should only be toxic to pathogens, but safe for use by humans. Proposed long-term work also includes extension of this approach to other bacterial systems to combat antibiotic resistance. It is even more crucial that this type of investigation is undertaken in academia than it would be if industry were still heavily investing in it.

This study sheds light on a potentially new class of MRSA antibiotics, which may pave the road to multi-faceted approaches to combat antibiotic resistance. From the broader perspective, blocking central metabolic pathways was usually considered as a forbidden area in drug development due to the possibility of affecting human central metabolism (*e.g.*, side effects of chemotherapies). If the assertion that certain central metabolic inhibitors are specific to pathogens not to humans is tested, it will reassure that we have moved in the right direction to tackle a major challenge.

ACKNOWLEDGMENT

We thank Dr. Adhar Manna (University of South Dakota) for the ongoing collaboration on target essentiality validation. We also appreciate Dr. Scott Lovell (University of Kansas) for the ongoing collaboration on crystal structure development. This publication was made possible by NIH Grant Number 2 P20 RR016479 from the INBRE Program of the National Center for Research Resources. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

REFERENCES

[1] N. L. Haag, K. K. Velk, and C. Wu, "*In silico* Identification of Drug Targets in Methicillin/Multidrug-Resistant *Staphylococcus aureus*," Proc. Third International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies, BIOTECHNO 2011, May 22-27, 2011, Venice, Italy, pp. 91-99, IARIA XPS Press, ISBN: 978-1-61208-007-9.

[2] C. T. Walsh, "Antibiotics: Actions, Origin, Resistance," ASM Press, Washington, DC, 2003.

[3] S. B. Levy, "Antibiotic and antiseptic resistance: impact on public health," Pediatr. Infect. Dis. J., Vol. 19, Oct. 2000, pp. S120–S122.

[4] S. B. Levy, "The antibiotic paradox: how miracle drugs are destroying the miracle," New York: Plenum, 1992.

[5] F. C. Tenover, "Mechanisms of antimicrobial resistance in bacteria," Am. J. Med., vol. 119 (6 Suppl 1), June 2006, pp. S3–S10, doi:10.1016/j.ajic.2006.05.219

[6] C. Nathan, "Antibiotics at the crossroads," Nature, vol. 431, Oct. 2004, pp. 899–902, doi :10.1038/431899a

[7] C. Barrett, and J. Barrett," Antibacterials: are the new entries enough to deal with emerging resistance problems?" Current Opinion in Biotechnology, 2003,14, pp. 621–626.

[8] C. Walsh, "Where will new antibiotics come from?" Nature Rev. Microbiol. Oct. 2003, Vol. 1, pp. 65–70.

[9] L. L. Marcusson, N. Frimodt-Møller, and D. Hughes, "Interplay in the Selection of Fluoroquinolone Resistance and Bacterial Fitness" PLoS Pathogens, August 2009, Vol. 5, pp. 1-8.

[10]Community-Associated MRSA Information for the Public. CDC. 11 August 2008

[11] Fact Sheet. Multidrug-Resistant Tuberculosis (MDR TB) . Division of Tuberculosis Elimination. CDC. 11 August 2008

[12] Fact Sheet. Extensively Drug-Resistant Tuberculosis (XDR TB) . Division of Tuberculosis Elimination. CDC. 11 August 2008

[13] K. K. Kumarasamy, M. A. Toleman, T. R. Walsh, J. Bagaria, F. Butt, R. Balakrishnan, and U. Chaudhary, *et al.* "Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: a molecular, biological, and epidemiological study". Lancet Infect Dis vol.10, August 2010, pp. 597–602, doi:10.1016/S1473-3099(10)70143-2.

[14] S. Projan, "Why is big Pharma getting out of antibacterial drug discovery?" Curr. Opin. Microbiol. vol. 6, Oct. 2003, pp. 427–430, doi:10.1016/j.mib.2003.08.003

[15] F. D. Lowy, " *Staphylococcus aureus* infections," N Engl J Med. 1998, Vol. 339, pp.520–32.

[16] V. M. Dukic, M. Z. David, and Lauderdale, D.S. "Internet queries and methicillin-resistant *Staphylococcus aureus* surveillance," Emerg. Infect. Dis. 2011 Jun http://www.cdc.gov/EID/content/17/6/1068.htm doi: 10.3201/eid1706.101451

[17] U. Lorenz, K. Ohlsen, H. Karch, A. Thiede, and J.Hacker, "Immunodominant Proteins in Human Sepsis Caused by Methicillin Resistant *Staphylococcus aureus,"* Advances in Experimental Medicine and Biology, 2002, Vol. 485, pp. 273-278, doi: 10.1007/0-306-46840-9_36

[18] M. Jevons, Celbenin"-resistant staphylococci, Br. Med. J., vol. 1, 1961, pp. 124-125.

[19] T. Foster, (1996). *Staphylococcus.* In: Barron's Medical Microbiology (Barron S *et al*, eds.), 4th ed. Galveston, TX.

[20] K. Okuma, K. Iwakawa, J. D. Turnidge, W. B. Grubb, J. M. Bell, F. G. O'Brien, G. W. Coombs, J. W. Pearman, F. C. Tenover, M. Kapi, C. Tiensasitorn, T. Ito, and K. Hiramatsu, "Dissemination of new methicillin-resistant *Staphylococcus aureus* clones in the community", J. Clin. Microbiol. vol. 40, Nov. 2002, pp. 4289–4294, doi: 10.1128/JCM.40.11.4289-4294.2002

[21] H. Huang, N. M. Flynn, J. H. King, C. Monchaud, M. Morita, and S. H. Cohen, "Comparisons of Community-Associated Methicillin-Resistant *Staphylococcus aureus* (MRSA) and Hospital-Associated MSRA

Infections in Sacramento, California," J. Clin. Microbiol. vol. 44, July 2006, pp. 2423-2427, doi:10.1128/JCM.00254-06
[22] R. C. Moellering, "Vancomycin: A 50-Year Reassessment," Clin. Infect. Dis. vol. 42, Suppl 1, Jan. 2006, pp. S3-4.
[23] P. L. Donald, "Vancomycin: A History," Clin. Infect. Dis. vol., 42, Suppl 1, Jan. 2006, pp. S5-S12.
[24] G. C. Schito, "The importance of the development of antibiotic resistance in *Staphylococcus aureus*," Clin. Microbiol. Infect., Suppl 1, Mar. 2006, pp. 16445718.
[25] K. Sieradzki, and A. Tomasz, "Inhibition of cell wall turnover and autolysis by vancomycin in a highly vancomycin-resistant mutant of *Staphylococcus aureus*," J. Bacteriol., vol. 179, April, 1997, pp. 2557-2566.
[26] C. Burlak, C. H. Hammer, M. Robinson, A. R. Whitney, M. J. McGavin, B. N. Kreiswirth, and F. R. DeLeo, "Global analysis of community-associated methicillin-resistant *Staphylococcus aureus* exoproteins reveals molecules produced *in vitro* and during infection Cell" Microbiol., vol. 9, Jan. 2007, pp. 1172–1190, doi:10.1111/j.1462-5822.2006.00858.x
[27] J. M. Voyich, M. Otto, B. Mathema, K. R. Braughton, A. R. Whitney, D. Welty, R. D. Long, D. W. Dorward, D. J. Gardner, G. Lina, B. N. Kreiswirth, and F. R. DeLeo, "Is panton-valentine leukocidin the major virulence determinant in community-associated methicillin-resistant *Staphylococcus aureus* disease?" J. Infect. Dis. vol.194, Dec. 2006, pp. 1761-1770.
[28] Centers for Disease Control and Prevention, (2007). MRSA: Methicillin-resistant *Staphylococcus aureus* in Healthcare Settings.
[29] R. M. Klevens, M. A. Morrison, J. Nadle, S. Petit, K. Gershman, S. Ray, L. H. Harrison, R. Lynfield, G. Dumyati, J. M. Townes, A. S. Craig, E. R. Zell, G. E. Fosheim, L. K. McDougal, R. B. Carey, and S. K. Fridkin, "Invasive methicillin-resistant *Staphylococcus aureus* infections in the United States," JAMA, vol. 298, Oct. 2007, pp. 1763-1771.
[30] J. D. Siegel, E. Rhinehart, M. Jackson, and L. Chiarello, (2008). Management of multi-drug resistant organisms in healthcare settings, 2006. US Centers for Disease Control and Prevention. Healthcare Infection Control Practices Advisory Committee. Accessed January 25.
[31] L. Nicolle, "Community-acquired MRSA: a practitioner's guide," CMAJ., vol. 175, June 2006, pp. 145, doi:10.1503/cmaj.060457
[32] Centers for Disease Control and Prevention. Epidemiology and ma1nagement of MRSA in the Community. October 26, 2007. Accessed January 25, 2008.
[33] J. A. Gorchynski, and J. K. Rose, "Complications of MRSA Treatment: Linezolid-induced Myelosuppression Presenting with Pancytopenia," Western Journal of Emergency Medicine, vol. 9, August 2008, pp. 177-178
[34] Cubist Pharmaceuticals, Inc. (2003). Daptomycin (Cubicin) package literature. Cubist Pharmaceuticals, Inc., Lexington, Mass.
[35] R. Moellering, Trends in New Drug Development; from Broad- to Narrow-Spectrum Antibiotics Program and abstracts from the 39th ICAAC Symposium 138, F 1363 .
[36] F. Tally, Trends in New Drug Development; from Broad- to Narrow-Spectrum Antibiotics Program and abstracts from the 39th ICAAC Symposium 138, F 1364 .
[37] S. D. Mills, "The role of genomics in antimicrobial discovery," J. Antimicrob. Chemother., vol. 51, Mar. 2003, pp. 749-752, doi: 10.1093/jac/dkg178
[38] P. F. Boreham, R. E. Phillips, and R. W. Shepherd, "Altered uptake of metronidazole in vitro by stocks of Giardia intestinalis with different drug sensitivities," Trans. R. Soc. Trop. Med. Hyg. vol. 82, May 1988, pp. 104-106.
[39] K. R. Sakharkar, M. K. Sakharkar, and V. T. Chow, "A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa*," *In Silico* Biology, vol. 4, 2004, pp. 355-360.
[40] V. Sharma, P. Gupta, and A. Dixit, "Identification of putative drug targets from different metabolic pathways of *Aeromonas hydrophila*," *In Silico*. Biology," vol 4, 2008, pp. 331-338.
[41] http://www.ncbi.nlm.nih.gov/genome/154 (06/12/2012)
[42] W. Li, L. Jaroszewski, and A. Godzik, "Clustering of highly homologous sequences to reduce the size of large protein databases," Bioinformatics, vol. 17, Mar. 2001, pp. 282-283, doi: 10.1093/bioinformatics/17.3.282
[43] R. Zhang, H. Ou, and C. Zhang, "DEG, a Database of Essential Genes," Nucleic Acids Res. vol. 32, (suppl 1), Jan. 2004, pp. D271-D272, doi: 10.1093/nar/gkh024
[44] http://www.ncbi.nlm.nih.gov/RefSeq/ (06/12/2012)
[45]http://www.ebi.ac.uk/integr8/InquisitorPage.do (06/12/2012)
[46] http://www.ebi.ac.uk/Tools/pfa/iprscan/(06/12/2012)
[47] http://www.genome.jp/(06/12/2012)
[48] http://workbench.sdsc.edu/ (06/12/2012)
[49] B. A. Diep, H. A. Carleton, R. F. Chang, G. F. Sensabaugh, and F. Perdreau-Remington, "Roles of 34 virulence genes in the evolution of hospital- and community-associated strains of methicillin-resistant *Staphylococcus aureus*," J. Infect. Dis. vol. 193, Apr. 2006, pp. 1495–1503, doi: 10.1086/503777
[50] A. P. Johnsona, H.M. Auckenb, S. Cavendishc, M. Gannerb, M. C. J. Walec, M. Warnera, D. M. Livermorea, and B. D. Cooksonb "Dominance of EMRSA-15 and -16 among MRSA causing nosocomial bacteraemia in the UK: analysis of isolates from the European Antimicrobial Resistance Surveillance System" (EARSS)". J. Antimicrob. Chemother., vol. 4, 2001, pp. 143–144, doi:10.1093/jac/48.1.143.
[51] K. Hiramatsu, N. Aritaka, and H. Hanaki, "Dissemination in Japanese hospitals of strains of *Staphylococcus aureus* heterogeneously resistant to vancomycin," Lancet, vol. 350, Dec. 1997, pp. 1670-1673, doi:10.1016/S0140-6736(97)07324-8
[52] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, *et al*., "Initial sequencing and analysis of the human genome," Nature vol. 409, Feb. 2001, pp. 860-921, doi:10.1038/35057062
[53] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, *et al*., "The sequence of the human genome" Science, vol. 291, Feb. 2001, pp. 1304-1351,doi: 10.1126/science.1058040.
[54] K. L. Longenecker, G. F. Stamper, P. J. Hajduk, E. H. Fry, C. G. Jakob, J. E. Harlan, R. Edalji, D. M. Bartley, K. A. Walter, L. R. Solomon, T. F. Holzman, Y. G. Gu, C. G. Lerner, B. A. Beutel, and V. S. Stoll, "Structure of MurF from *Streptococcus pneumoniae* co-crystallized with a small molecule inhibitor exhibits interdomain closure," Protein Sci., vol. 14, Dec. 2005, pp. 3039-3047, doi: 10.1110/ps.051604805
[55] A. Perdih, M. Kotnik, M. Hodoscek, and T. Solmajer, "Targeted molecular dynamics simulation studies of binding and conformational changes in *E. coli* MurD," Proteins, vol.68, Apr. 2007, pp. 243-254. doi: 10.1002/prot.21374
[56] A. Strandén, K. Ehlert, H. Labischinski, and B. Berger-Bächi, "Cell wall monoglycine cross-bridges and methicillin hypersusceptibility in a femAB null mutant of methicillin-resistant *Staphylococcus aureus*," J. Bacteriol. vol. 179, Jan. 1997, pp. 9–16.
[57] J. Malabendu, T. Luong, H. Komatsuzawa, M. Shigeta, and C. Y. Lee, "A method for demonstrating gene essentiality in *Staphylococcus aureus*," Plasmid, vol. 44, Mar. 2000. pp. 100-104, doi:10.1006/plas.2000.1473
[58] N. H. Georgopapadakou, and F. Y. Liu, "Binding of β-lactam antibiotics to penicillin-binding proteins of *Staphylococcus aureus* and *Streptococcus faecalis*: relation to antibacterial activity," Antimicrob. Agents Chemother. vol. 18, Nov. 1980, pp. 834-836.
[59] K. Ubukata, N. Yamashita, and M. d Konno, "Occurrence of a β-lactam-inducible penicillin-binding protein in methicillin resistant *staphylococci*," Antimicrob. Agents Chemother. vol. 27, May 1985, pp. 851-857.
[60] M. Kuroda, H. Kuroda, T. Oshima, F. Takeuchi, H. Mori, and K. Hiramatsu, "Two-component system VraSR positively modulates the regulation of cell-wall biosynthesis pathway in *Staphylococcus aureus*," Mol. Microbiol. vol.49, Aug. 2003, pp. 807–821, doi:10.1046/j.1365-2958.2003.03599.x
[61] L. Cui, H. Murakami, K. Kuwahara-Arai, H. Hanaki, and K. Hiramatsu, "Contribution of a thickened cell wall and its glutamine nonamidated component to the vancomycin resistance expressed by *Staphylococcus aureus* Mu50," Antimicrob. Agents Chemother. vol. 44, Sep. 2000, pp. 2276–2285.
[62] A. Severin, K. Tabei, F. Tenover, M. Chung, N. Clarke, and A. Tomasz, "High level oxacillin and vancomycin resistance and altered cell wall composition in *Staphylococcus aureus* carrying the staphylococcal mecA and the enterococcal vanA gene complex," J. Biol. Chem. vol. 279, Jan. 2004, pp. 3398-3407.

[63] A. A. Salyers, and D. D. Whitt, (2005). Revenge Of The Microbes: How Bacterial Resistance Is Undermining The Antibiotic Miracle, American Society for Microbiology Press, Washington, DC.
[64] F. Van Bambeke, M. P. Mingeot-Leclercq, M. J. Struelens, and P. M. Tulkens, "The bacterial envelope as a target for novel anti-MRSA antibiotics," Trends Pharmacol Sci. vol. 29, Mar. 2008, pp. 124-134.
[65] K. Lewis, "Multidrug resistance: versatile drug sensors of bacterial cells," Curr. Biol., vol. 9, Jun. 1999, pp. R403-R407, doi:10.1016/S0960-9822(99)80254-1
[66] C. M. Fraser, J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, R. D. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J. F. Tomb, B. A. Dougherty, K. F. Bott, P. C. Hu, T. S. Lucier, S. N. Peterson, H. O. Smith, C. A. Hutchison, 3rd and J. C. Venter, "The minimal gene complement of *Mycoplasma genitalium*," Science. vol. 270, Oct. 1995, pp. 397-403, DOI: 10.1126/science.270.5235.397
[67] Y. Kagawa, and E. Racker, "Partial resolution of the enzymes catalyzing oxidative phosphorylation. 8. Properties of a factor conferring oligomycin sensitivity on mitochondrial adenosine triphosphatase," J. Biol. Chem., vol.241, May 1966, pp. 2461-2466.
[68] D. Xia, C. H. Yu, H. Kim, J. Xia, A. M. Kachurin, L. Zhang, L. Yu, and J. Deisenhofer, "Structure of Antimycin A1, a Specific Electron Transfer Inhibitor of Ubiquinol-Cytochrome c Oxidoreductase" J. Am. Chem. Soc., vol. 121, Aug. 1999, pp. 4902–4903, DOI: 10.1002/chin.199933269
[69] A. Fredenhagen, S. Y. Tamura, P. T. M. Kenny, H. Komura, Y. Naya, K. Nakanishi, K. Nishiyama, M. Sugiura, and H. Kita, "Andrimid, a new peptide antibiotic produced by an intracellular bacterial symbiont isolated from a brown planthopper" J. Am. Chem. Soc. vol. 109, Jul. 1987, pp. 4409–4411, doi: 10.1021/ja00248a055
[70] C. Freiberg, J. Pohlmann, P. G. Nell, R. Endermann, J. Schuhmacher, B. Newton, M. Otteneder, T. Lampe, D. Häbich, and K. Ziegelbauer, "Novel bacterial acetyl coenzyme A carboxylase inhibitors with antibiotic efficacy in vivo," Antimicrob Agents Chemother, vol. 50, Aug. 2006, pp. 2707-2712.
[71] F. R. Stermitz, P. Lorenz, J. N. Tawara, L. A. Zenewicz, and K. Lewis, "Synergy in a medicinal plant: Antimicrobial action of berberine potentiated by 5'-methoxyhydnocarpin, a multidrug pump inhibitor," Proc. Natl. Acad. Sci. U.S. A. vol. 97, Feb. 2000, pp. 1433-1437.
[72] N. R. Guz, and F. R. Stermitz, "Synthesis and structures of regioisomeric hydnocarpin-type flavonolignans," J. Nat. Prod. Vol. 63, Aug. 2000, pp. 1140-1145, DOI: 10.1021/np000166d.
[73] S. F. Altschul, T. L. Madden, A. A. Schäffe, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Res. vol. 25, Sep. 1997, pp. 3389-3402.
[74] J.D. Thompson, D.G. Higgins, and T.J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," Nucleic Acids Res. vol. 22, Nov. 1994, pp. 4673-4680, doi: 10.1093/nar/22.22.4673.
[75] T. Gefflaut, C. Blonski, J. Perie, and M. Willson, "Class I aldolases: substrate specificity, mechanism, inhibitors and structural aspects," Prog. Biophys. Mol. Biol. vol.63, 1995, pp. 301-340.
[76] A. Galkin, L. Kulakova, E. Melamud, L. Li, C. Wu, P. Mariano, D. Dunaway-Mariano, T. E. Nash, and O. Herzberg, "Characterization, Kinetics, and Crystal Structures of Fructose-1,6-bisphosphate Aldolase from the Human Parasite *Giardia lamblia,"* J. Biol. Chem., vol. 282, Feb. 2007, pp. 4859–4867.
[77] D. Voet, and J. G. Voet, Biochemistry, 3rd ed., Chp. 17. Wiley: 2004, pp. 591
[78] A. Gorrell, and J. G. Ferry, "Investigation of the Methanosarcina thermophila acetate kinase mechanism by fluorescence quenching," Biochemistry, Dec 2007 Vol. 46, pp. 14170-14176.
[79] C. Ingram-Simth, S.R. Martin, and K.S. Smith,"Acetate kinase: not just a bacterial enzyme," Trends in Microbiology, June 2006, Vol. 14, pp. 249-253.
[80] A. J. Wolfe, "The acetate switch." Microbiol. Mol. Biol. 2005 Mar. Vol. 69, pp. 12-50.
[81] A. Gorrell, S. H. Lawrence, and J. G. Ferry, "Structural and Kinetic Analyses of Arginine Residues in the Active Site of the Acetate Kinase from *Methanosarcina thermophile*," The Journal of Biological Chemistry, 2005,March Vol. 280, pp. 10731-10742/doi: 10.1074/jbc.M412118200
[82]http://www.clemson.edu/cafls/spotlight_profiles/faculty/smith_kerry.html
[83] A. Boniface, A. Bouhss, D. Mengin-Lecreulx, and D. Blanot, "The MurE synthetase from Thermotoga maritima is endowed with an unusual D-lysine adding activity," J. Biol. Chem. vol. 281, Jun. 2006, pp. 15680-15686.
[84] T. Deva, E. N. Baker, C. J. Squire, and C. A. Smith, "Structure of *Escherichia coli* UDP-N-acetylmuramoyl:L-alanine ligase (MurC)," Acta. Crystallogr. D. Biol. Crystallogr. vol.62, Dec. 2006, pp. 1466-1474, doi:10.1107/S0907444906038376
[85] C. A. Smith, "Structure, function and dynamics in the mur family of bacterial cell wall ligases," J. Mol. Biol. vol. 362, Sep.2006, pp. 640-655, doi:10.1016/j.jmb.2006.07.066.
[86] K. Ehlert, "Methicillin-resistance in *Staphylococcus aureus* - molecular basis, novel targets and antibiotic therapy," Curr. Pharm. Des. vol.5, Feb.1999, pp. 45-55.

Identifying the Building Blocks of Protein Structures from Contact Maps Using Protein Sequence and Evolutionary Information

Hazem Radwan A. Ahmed
School of Computing and Information Science
Queen's University
Kingston, Ontario, Canada. K7L 3N6
Email: hazem@cs.queenu.ca

Janice I. Glasgow
School of Computing and Information Science
Queen's University
Kingston, Ontario, Canada. K7L 3N6
Email: janice@cs.queensu.ca

***Abstract*— 1D protein sequences, 2D contact maps and 3D structures are three different representational levels of detail for proteins. The problem of protein 3D structure prediction from 1D protein sequences remains one of the challenges of modern bioinformatics. The main issue here is that it is computationally complex to reliably predict the full 3D structure of a protein from its 1D sequence. A 2D contact map has, therefore, been used as an intermediate step in this problem. A contact map is a simpler, yet representative, alternative for the 3D protein structure. In this paper, we focus on the problem of identifying similar substructural patterns of protein contact maps (the building blocks of protein structures) using a structural pattern matching approach that incorporates protein sequence and evolutionary information. These substructural patterns are of particular interest to our research, because they could potentially be used as building blocks for a computational bottom-up approach towards the ultimate goal of protein structure prediction from contact maps. The results are benchmarked using a large standard protein dataset. We assess the consistency and the efficiency of identifying these similar substructural patterns by performing different statistical analyses (e.g., Harrell-Davis Quantiles and Bagplots) on different subsets of the experimental results. We further studied the effect of the local sequence information, global sequence information, and evolutionary information on the performance of the method. The results show that both local and global sequence information are more helpful in locating short-range contacts than long-range contacts. Moreover, incorporating evolutionary information has remarkably improved the performance of locating similar short-range contacts between contact map pairs.**

Keywords - protein structure prediction; contact map; case-based reasoning; evolutionary information; sequence information.

I. INTRODUCTION

Since the human genome sequence was revealed in April 2003, the need to predict protein structures from protein sequences has dramatically increased [2]. Proteins are complex macromolecules that are associated with several vital biological functions for any living cell. Such as, transporting oxygen, ions, and hormones; protecting the body from foreign invaders; and catalyzing almost all chemical reactions in the cell. Proteins are made of long sequences of amino acids that fold into three-dimensional structures. Because protein folding is not easily observable experimentally [3], protein structure prediction has been an active research field in bioinformatics as it can ultimately broaden our understanding of the structural and functional properties of proteins. Moreover, predicted structures can be used in structure-based drug design, which attempts to use the structure of proteins as a basis for designing new ligands by applying principles of molecular recognition [4].

In recent decades, many approaches have been proposed for understanding the structural and functional properties of proteins. These approaches vary from time-consuming and relatively expensive experimental determination methods (e.g., X-ray crystallography [5] and NMR spectroscopy [6]) to less-expensive computational protein modeling methods for protein structure prediction (e.g., ab-initio protein modeling [7], comparative protein modeling [8], and side-chain geometry prediction [9]).

Although the computational methods attempt to circumvent the complexity of the experimental methods with an approximation to the solution (predicted protein structures versus experimentally-determined structures), analyzing the three-dimensional structure of proteins computationally is not a straightforward task. Hence, two-dimensional representations of protein structures, such as distance and contact maps, have been widely used as a promising alternative that offers a good way to analyze the 3D structure using a 2D feature map [19]. This is because they are readily amenable to machine learning algorithms and can potentially be used to predict the three-dimensional structure, achieving a good compromise between simplicity and competency [45].

Despite the exhaustive research done in an effort to reliably predict the structure of proteins from their sequences, the gap between known protein sequences and computationally predicted protein structures is continuously growing because of the computational complexity associated with the problem [10]. The "Divide and Conquer" principle is applied in our research in an attempt to handle such a complex problem, by dividing it into two separate yet dependent subproblems, using a Case-Based Reasoning (CBR) approach [18]. Firstly, a contact map representing the contacts between amino acids is predicted using protein

sequence information. Secondly, protein structure is predicted using its predicted contact map [19].

Since contact map prediction offers a possible shortcut to predict protein tertiary structure, researchers have considered various approaches with encouraging results for predicting protein contact maps from sequence information and structural features. Approaches of protein contact map prediction vary from those that apply neural networks [11][12], to those that consider support vector machines [13] and association rules [14]. Various statistical approaches have also been attempted, including correlated mutations [15][16] and hidden Markov models [17].

Our research focuses on the problem of protein structure prediction from contact maps. In particular, we apply a CBR framework [18] to determine the alignment of secondary structures based on previous experiences stored in a case base, along with detailed knowledge of the chemical and physical properties of proteins [19].

The proposed CBR framework is based on the premise that similar problems have similar solutions. CBR solves new problems (e.g., protein structure prediction) by adapting the most similar retrieved solutions of previously solved problems. Several challenges arise here: firstly, how to retrieve the contact maps from the case-base with the most similar solutions (substructural patterns); secondly, how to adapt the new problem "query protein" to the retrieved solutions "template proteins"; thirdly, how to evaluate the adapted solution in an attempt to have a close solution to the native structure of the query protein, which will be saved as a new case in the repository of the CBR system for a later use. These three challenges correspond to the three main phases of our CBR system: *Retrieval*, *Adaptation*, and *Evaluation*, as shown in Figure 1.

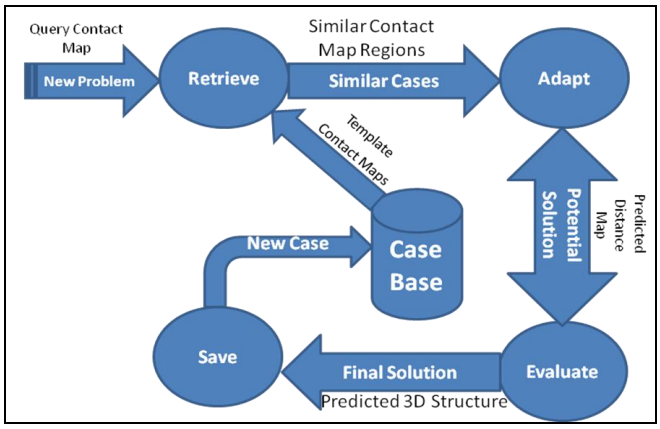


Figure 1. CBR framerwork for determining protein structure from contact maps.

The main challenge of the *Retrieval* phase for our problem domain is to find the template contact maps with the most similar solutions to the query contact map. Thus, it is necessary to have a robust similarity measure for contact maps to reliably compare each new contact map (i.e., a new case), with the contact maps in the case-base (i.e., previously solved cases). This measure is used so that the retrieved contact maps from the case-base have substructural patterns (e.g., secondary structures and structure motifs) that are in common with the new contact map.

Proposing an approach to determine and locate regions of similarity between contact map pairs help to identify these common or similar substructural patterns between the query contact map and every similar retrieved template contact map (known structure) from the case-base, which is crucial for the *Adaptation* phase of the CBR system. The main objective of this phase is to adapt the retrieved solutions to the new problem in the following way. All amino acid residues in the common substructural patterns for the current query contact map that have corresponding residues in the template structure are given the coordinate information from these residues [13]. This paper, an expansion of the work in [1], primarily focuses on the *Adaptation* phase, with the goal of identifying similar substructural patterns between protein contact map pairs using both sequence and evolutionary information.

The paper is organized as follows: Section II provides the reader with the background material required to understand the concepts used in our study. It describes distance and contact maps, gives examples of structural patterns of contact maps, and discusses protein similarity relationships at different representational levels of detail, as well as the structural classification of protein domains. Section III presents the experimental setup and the details of the multi-regional analysis of the contact map method used in our experiments. Section IV discusses the experimental benchmark dataset used in the study and shows the performance of the proposed method using statistical analyses, including a quantile-based analysis and a correlation analysis. The final section highlights the contributions, summarizes the main results of the study, and presents a set of potential directions for future research.

II. BACKGROUND MATERIAL

Contact and distance maps provide a compact 2D representation of the 3D conformation of a protein, and capture useful interaction information about the native structure of proteins. Contact maps can ideally be calculated from a given structure, or predicted from protein sequence. The predicted contact maps have received special attention in the problem of protein structure prediction, because they are rotation and translation invariant (unlike 3D structures). While it is not simple to transfer contact maps back to the 3D structure (unlike distance maps), it has shown some potential to reconstruct the 3D conformation of a protein from accurate and even predicted (noisy) contact maps [20].

A. Distance and Contact Maps

A distance map, *D*, for a protein of *n* amino acids is a two-dimensional *n* x *n* matrix that represents the distance between each pair of atoms of the protein. The distance may

be that between alpha-carbon atoms (Cα) [21], beta-carbon atoms (Cβ), or it may be the minimum distance between any pair of atoms belonging to the side chain or to the backbone of the two residues [22][23]. While the best definition for inter-residue distance is the minimum distance between side-chain or alpha-carbon atoms with a cut-off distance of around 1.0 Ångstrom (0.1 nm) [24], backbone atom-based definitions (e.g., Cα or Cβ distances) with longer distance cut-offs are more readily projected into three dimensions [26]. As shown in Figure 2(a), the darker the distance map region is, the closer the distance of its corresponding atom pairs is. The distance information can be further used to infer the interactions among residues of proteins by constructing another same-sized matrix called a contact map.

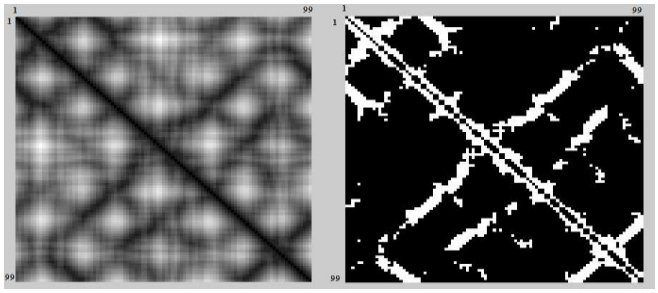
A contact map, *C*, is a two-dimensional binary symmetric matrix that represents pairs of amino acids that are in contact. A pair of residues is considered to be in contact if the distance between their alpha-carbon atoms is less than or equal a predefined threshold, i.e., their positions in the three-dimensional structure of the protein are within a given distance threshold (usually measured in Ångstroms), as shown in Figure 2(b).

An element of the i^{th} and j^{th} residues of a contact map, C*(i,j)*, can be defined as follows [19]:

$$\mathbf{C}(i,j) = \begin{cases} 1; & \text{if } D(i,j) \leq Threshold \\ 0; & otherwise \end{cases}$$

Where *D(i,j)* is the distance between amino acids i and j, ***1*** denotes *contacts* (white), and **0** denotes *no contacts* (black).

According to extensive experimental results presented in [25], contact map thresholds, ranging from 10 to 18 Å allow the reconstruction of 3D models from contact maps to be similar to the protein's native structure. In this paper, different contact map thresholds were applied in a series of experiments in order to find the contact map threshold that best suits our experiments.



(a) (b)

Figure 2. (a) Distance map for a protein of 99 amino acid residues. (b) contact map for the same protein of 99 amino acids after applying a distance threshold of 10 Ångstrom on its distance map. (local contacts < 3.8 Å are ignored – refer to Section III-C for details.)

As shown in Figure 3, a threshold of 8 Å missed some topological information about the protein, whereas a threshold of 12 Å added many contacts that are irrelevant to the topology of the protein. This suggests that a threshold of 10 Å is a good compromise; therefore, it was adopted in our experiments.

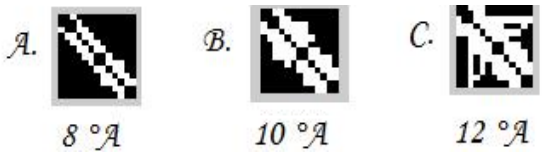


Figure 3. Applying different thresholds for a substructural pattern of contact map.

B. *Structural patterns of Contact Maps*

Different secondary structures of proteins have distinctive structural patterns in contact maps. In particular, an α-helix appears as an unbroken row of contacts between i, i ± 4 pairs along the main diagonal, while beta-sheets appear as an unbroken row of contacts in the off-diagonal areas. A row of contacts that is parallel to the main diagonal represents a pair of parallel β-sheets, while a row of contacts that is perpendicular to the main diagonal represents a pair of anti-parallel β-sheets [26]. Furthermore, contacts between secondary structure elements could also be recognized through a contact map. In general, contacts between α-helices and other secondary structure elements appear as broken rows or "tire tracks". If the two contacting elements are both helices, then the contacts appear in every 3 or 4 residues in both directions, following the periodicity of the helix. If one of the elements is a β-sheet, a periodicity of 2 in the contacts will appear, since the side chains in strands alternate between the two sides of the β-sheet [26].

C. *The Classification of Protein Domains*

The Structural Classification of Proteins (SCOP) database was designed by G. Murzin et al. [32] to provide an easy way to access and understand the information available for protein structures. The database contains a detailed and comprehensive description of the structural and evolutionary relationships of the proteins of known structure. Structurally and evolutionarily related proteins are classified into similar levels in the database hierarchy. Evolutionarily-related proteins are those that have similar functions and structures because of a common descent or ancestor. The main levels in the classification hierarchy of the SCOP database are as follows: 1) *Family* level that implies clear evolutionary relationship, 2) *Superfamily* level that implies probable common evolutionary origin, and 3) *Fold* level that implies major structural similarity.

D. *Protein Similarity Relationships*

Understanding protein similarity relationships is vital for the further understanding of protein functional similarity and evolutionary relationships. Although a protein with a

given sequence may exist in different conformations, the chances that two highly-similar sequences will fold into distinctly-different structures are so small that they are often neglected in research practice [30]. This suggests that sequence similarity could generally indicate structure similarity. Furthermore, a pair of proteins with similar structure has similar contact maps [31]. Therefore, as shown in Figure 4, by the transitivity relationship, a logical inference could be drawn regarding the association between sequence similarity and contact map similarity. The premise of the method of multi-regional analysis of contact maps in this paper is based on this transitive similarity relationship between contact map and protein sequence (via structure).

That being said, the counter relationships between contact map and sequence similarity, as well as structure and sequence similarity, are still questionable. This is due to the fact that protein structures are evolutionarily conserved better than protein sequences [33], since the protein sequences evolve rapidly compared to protein structures.

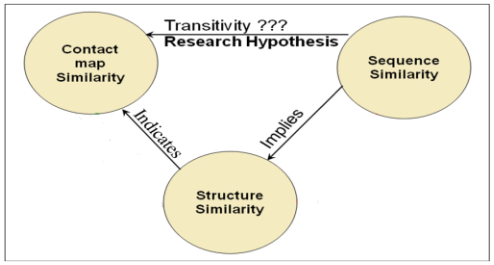


Figure 4. Protein similarity relationships at different levels of detail.

III. METHOD AND EXPERIMENTAL SETUP

This section describes the multi-regional analysis of the contact maps method used in the experiments. The method examines whether sequence similarity information helps in a pattern matching approach to locate regions of similarity in contact maps (the target substructural patterns) that correspond to local similarities in protein structures. The first stage of this method aims to align pairs of protein sequences for each combination pair of contact maps to find the most local similar subsequences. The next stage aims to quantify the similarity of contact maps regions that correspond to these similar subsequences found in the first stage. Finally, different statistical analyses were considered to evaluate the performance of the method, and to determine how well local protein sequence similarity leads to corresponding local contact map similarity.

A. Experimental Dataset

The benchmark Skolnick dataset is adopted for our experiments. The Skolnick dataset is a standard benchmark dataset of 40 large protein domains, divided into four categories as shown in Table I. It was originally suggested by J. Skolnick and described in [34]. The dataset has been used in several recent studies related to structural comparison of proteins [34][35][36]. The 40 protein domains are: 1b00A (1), 1dbwA (2), 1nat (3), 1ntr (4), 1qmpA (5), 1qmpB (6), 1qmpC (7), 1qmpD (8), 1rn1A (9), 1rn1B (10), 1rn1C (11), 3chy (12), 4tmyA (13), 4tmyB (14), 1bawA (15), 1byoA (16), 1byoB (17), 1kdi (18), 1nin (19), 1pla (20), 2b3iA (21), 2pcy (22), 2plt (23), 1amk (24), 1aw2A (25), 1b9bA (26), 1btmA (27), 1htiA (28), 1tmhA (29), 1treA (30), 1tri (31), 3ypiA (32), 8timA (33), 1ydvA (34), 1b71A (35), 1bcfA (36), 1dpsA (37), 1fha (38), 1ier (39), and 1rcd (40).

Each domain entry contains a name and its assigned index in parentheses. The domain name is the PDB code for the protein containing it. If multiple protein chains exist, the chain index is appended to the PDB code.

TABLE I. PROTEIN DOMAINS IN SKOLNICK DATASET

Categories	Global sequence similarity	Sequence length (residues)	Domain indices
1	15-30% (low)	124	1-14
2	7-70% (Med)	170	35-40
3	35-90% (High)	99 (Short)	15-23
4	30-90% (High)	250 (Long)	24-34

B. Sequence Analysis

For the sequence analysis stage, we align every combination pair of sequences. The SIM algorithm [37] is used for this purpose. This algorithm employs a dynamic programming technique to find user-defined, non-intersecting alignments that are the best (i.e., with the highest similarity score) between pairs of sequences. The results from the alignments are sorted descendingly according to their similarity score [38].

In this method, we are only interested in alignments of subsequence of at least 10 residues, and at most 20 residues. We are not interested in alignments of length less than 10 residues because these alignments would not form a complete substructural pattern (for example, the lengths of alpha helices vary from 4 or 5 residues to over 40 residues, with an average length of about 10 residues [39]). We are also not interested in long alignments because most methods for contact maps analysis are known to be far more accurate on local contacts (those contacts that are clustered around the main diagonal), than nonlocal (long-range) contacts [40]. Thus, to eliminate one source of uncertainty of the long-range contacts, alignments of a length greater than 20 residues are not considered.

In this experiment, a large penalty for opening a gap is used, since it is evident that affine gap scores [27][28] with a large penalty for opening a gap and a much smaller one for extending it, have generally proven to be effective. Opening gap penalty is a penalty for the first residue in a gap, and extended gap penalty is a penalty for every additional residue in a gap. Therefore, in our experiments, the open and extended gap penalties are set to 10 and 1 respectively. In an effort to analyze pairs of protein sequences, the best 100 local sequence alignments are generated from every pair of sequences. Then, a selection strategy is used to select the two alignments of 10-20

residues with the most and least similarity score (to check the performance in case of low and high similarity).

As for the substitution matrix, BLOSUM62 was adopted to score sequence alignment. The BLOSUM substitution matrix was developed by Henikoff [41] as a new approach for the Percent Accepted Mutation (PAM) scoring matrix that was developed earlier by Margaret Dayhoff who pioneered this approach in the 1970's [29]. Unlike PAM, the BLOSUM62 matrix did an excellent job in detecting similarities even for distant protein sequences.

C. Contact Map Analysis

The second stage of the method is to locate contact map regions that correspond to the most and least similar protein subsequences. In order to unbiasedly analyze the diagonal contact map regions, we ignored local contacts between each residue and itself on the main diagonal. Comparing the main diagonal of contact maps (protein backbone) will neither add meaningful information for their similarity nor dissimilarity (for example, even too distant contact maps will share a similar main diagonal). Furthermore, local contacts with distance less than 3.8 Å are ignored, based on the fact that the minimum distance between any pair of different residues cannot be less than 3.8 Å [40].

Two common similarity measures for binary data that can be used to measure contact map similarity are Simple Matching Coefficient (SMC) [43] and Jaccard's Coefficient (J) [44]. SMC is based on the Hamming distance. If two contact map regions have the same size, we can use SMC to count the number of elements in positions where they have similar values. SMC is useful when binary values hold equal information (i.e., symmetry).

$$SMC = \frac{C_{11} + C_{00}}{S} \quad (1)$$

where C_{11} is the count of nonzero elements (contacts) of both contact maps, C_{00} is the count of zero elements (no-contacts) of both contact maps, and S is the contact map size (i.e., the square of the sequence length for the contact map).

In contact maps, however, binary values do not hold equal information because of the fact that zero values hold no information (they mean there is no contact between protein residues) as opposed to non-zero values where contacts between protein residues occurred. Another drawback of SMC is that it considers counting zero values for both contact maps (C_{00}). These regions represent the "double absence" where there are no contacts for both contact maps, making them of less interest in this study.

On the other hand, Jaccard's Coefficient (J) is widely used in information retrieval as a measure of similarity. It is suitable for asymmetric information on binary (and non-binary) variables where binary values do not have to carry equal information, resolving the first issue of SMC. Furthermore, Jaccard's Coefficient (J) does not consider counting zero elements in the matrix (no contacts), removing the effect of the "double absence" that has neither meaningful contribution to the similarity, nor the dissimilarity, of contact maps. Therefore, Jaccard's Coefficient was chosen as the contact map similarity metric that best suits our experiments.

$$J = \frac{C_{11}}{S - C_{00}} \quad (2)$$

Where C_{11} is the count of nonzero elements (contacts) of both contact maps, C_{00} is the count of zero elements (no-contacts) of both contact maps, and S is the contact map size (i.e., the square of the sequence length for the contact map).

D. Sequence Gap and Region Displacement Problem

The displacement problem happens when a pair of aligned subsequences is very similar (greater than 70%), but their corresponding diagonal contact map regions are not as similar (less than 50-60%). This is noticed to occur as a result of a slight shift in the aligned subsequence pair either because of a gap in the alignment, or because of a slightly shifted alignment. In this case, if the right displacement is considered for one of the aligned subsequence in the correct direction with the correct number of residues, their corresponding diagonal contact map regions will perfectly overlay one another and their similarity can go up to 90%, as shown in Figure 5. The current experimental setup, however, (e.g., open gap penalty, extended gap penalty, etc.) are optimized to minimize the displacement problem. As shown in Figure 7, the proposed method was successful in locating the exact correct boundaries of contact map regions that perfectly overlay one another, in an effort to maximize their similarity. That is, if any boundary is shifted only by one or two residues, the local contact map similarity will be significantly dropped, as shown in Figure 5 and Figure 6.

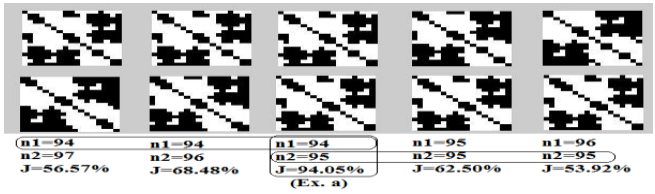


Figure 5. One example of the calculated region boundaries (n1 = 94 & n2 = 95) shows that the selected boundaries have the maximum Jaccard's coefficient (J = 94%) as opposed of 68% and 56% if the lower boundary is shifted by only one residue at a time, or 62% and 53% if the upper boundary is shifted by one residue at a time, instead.

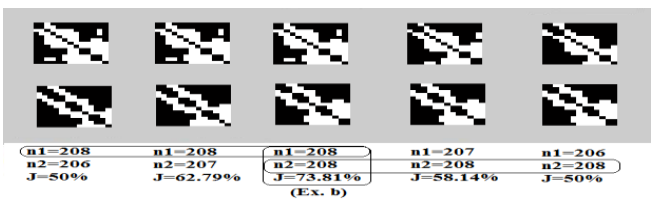


Figure 6. Another example of the calculated region boundaries of (Ex. b) also shows that the selected boundaries have the maximum Jaccard's coefficient (J = 73%) as opposed of 62% and 50% if the lower boundary is shifted by only one residue at a time, or 58% and 50% if the upper boundary is shifted by one residue at a time, instead.

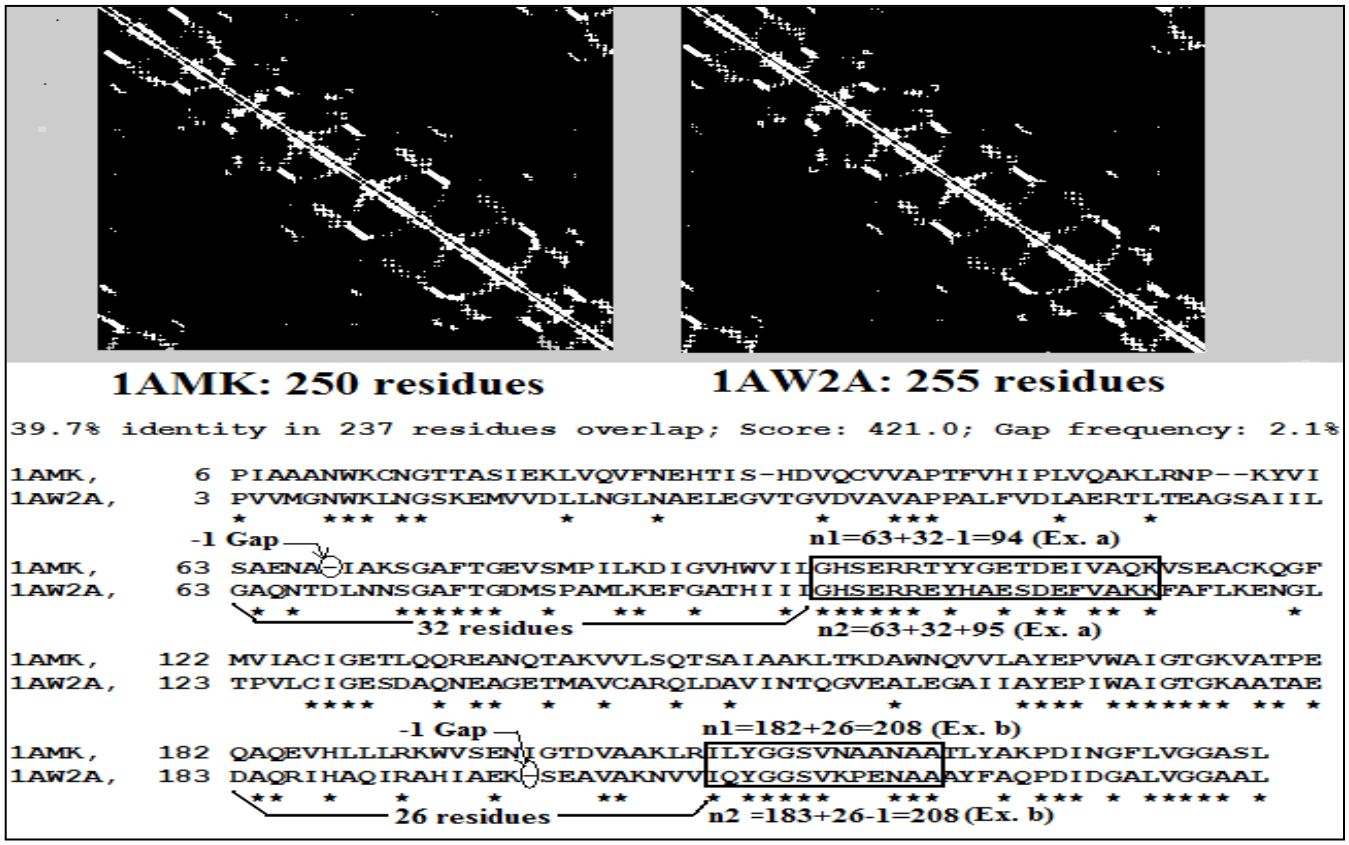


Figure 7. An illustration of the displacement problem between two highly-similar proteins (1AMK & 1AW2A). The gap length is subtracted from the start position of the upper boundary (n1 of Ex. a) and the lower boundary (n2 of Ex. b), since contact maps have no representation of gaps.

IV. RESULTS AND DISCUSSION

A. *The Big Picture*

To see the big picture of the problem, an all-against-all pair-wise analysis is performed on the benchmark Skolnick dataset, yielding several hundreds of pairwise alignment instances. The entire results of sequence and contact map similarity of each pairwise instance are presented as a 2D scatter plot to study the correlation between them, as shown in Figure 8. This figure draws a clear distinction between the correlation between sequence similarity and their contact map similarity in the diagonal area (short-range contacts), and the correlation between sequence similarity and their contact map similarity in the off-diagonal areas (long-range contacts).

Firstly, for long-range contacts, no matter how high the sequence similarity is the majority of the corresponding contact map similarity is very low (less than 20%). Thus, even high sequence similarity cannot help to suggest corresponding similarity for the long-range contacts. Secondly, for the short-range contacts, the plot reveals two different trends: 1) when sequence similarity is low (less than 60%), contact map similarity is indiscriminately dispersed between a very low similarity level (35%) and a very high one (90%), making it hard to reliably associate low sequence similarity to short-range contact map similarity. 2) When sequence similarity is high (greater than 60%), contact map similarity is apparently clustered in the upper-right corner of the plot (around 80%), suggesting a high correlation between local sequence similarity and short-range contact map similarity.

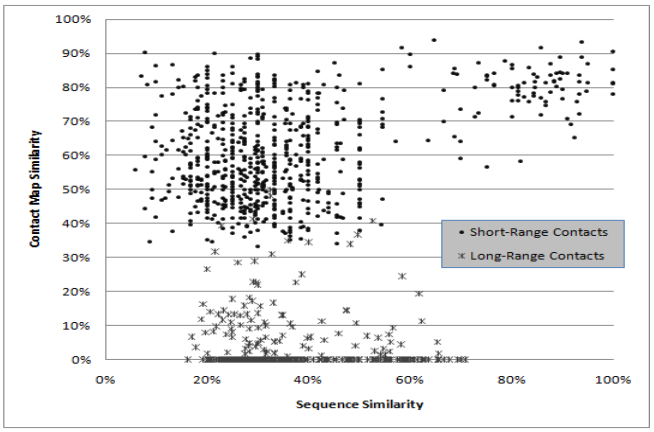


Figure 8. A 2D scatter plot showing the correlation between sequence similarities and their corresponding contact map similarities in the diagonal area (short-range contacts) and the off-diagonal areas (long-range contacts).

B. Harrell-Davis Quantiles

In an effort to improve performance in locating similar patterns in the diagonal regions of contact map pairs, evolutionary information (represented in SCOP family information) is proposed to be incorporated with the sequence information. As described in [36], the 40 protein domains of the Skolnick dataset are classified into five SCOP families. Based on SCOP family information, the results are distributed into four different groups: 1) the first group includes the results of pairs of protein subsequences that are most similar and of the same SCOP family. 2) The second group includes the results of pairs of protein subsequences that are most similar and of a different SCOP family. 3) The third group includes the results of pairs of protein subsequences that are least similar and of the same SCOP family. 4) The last group includes the results of pairs of protein subsequences that are least similar and of a different SCOP family.

Quantile-based analysis is performed to compare the four different groups. The q^{th} quantile of a dataset is defined as the value where the q-fraction of the data is below q and the (1- q) fraction of the data is above q. Some q-quantiles have special names: the 2-quantile (0.5 q) is called the median (or the 50^{th} percentile), the 4-quantiles (0.25 q) are called quartiles, the 10-quantiles (0.1 q) are called deciles, and the 100-quantiles are called percentiles. The 0.01 quantile = the 1^{st} percentile = the bottom 1% of the dataset, and the 0.99 quantile = the 99^{th} percentile = the top 1% of the dataset.

Using the online R statistics software in [46], the Harrell-Davis method for 100-quantile estimation is computed for this study. The Harrell-Davis method [48] is based on using a weighted linear combination of order statistics to estimate quantiles. The standard error associated with each estimated value of a quantile is also computed and plotted as error bars, as shown in Figure 10. Error bars are commonly used on graphs to indicate the uncertainty, or the confidence interval in a reported measurement. Figure 10(a) clearly shows that the results of contact map similarity of the same family are much better (higher) than those of a different family as in Figure 10(b). This supports the previous hypothesis that incorporating evolutionary information with sequence information improves the performance of locating remarkably better (highly-similar) diagonal contact map region. Comparing Figure 10(a) and Figure 10(c) reveals that low sequence information considerably deteriorates the method performance, even for the results of the same SCOP family. Whereas, comparing Figure 10(c) and Figure 10(d) demonstrates that with low sequence information, the performance is almost the same (poor), no matter if the protein pairs are of the same or of a different SCOP family.

C. Bagplots

A bagplot, initially proposed by Rousseeuwet et al. [49], is a bivariate generalization of the well-known boxplot [50]. In the bivariate case, the "box" of the boxplot changes to a convex polygon forming the "bag" of the bagplot. As shown in Figure [9], the bag includes 50% of all data points. The fence is the external boundary that separates points within the fence from points outside the fence (outliers), and is simply computed by increasing the bag by a given factor. Data points between the bag and fence are marked by a light-colored loop. The loop is defined as the convex hull containing all points inside the fence. The hull center is the center of gravity of the bag. It is either one center point (the median of the data) or a region of more than one center points, usually highlighted with a different color. Therefore, the classical boxplot can be considered as a special case of the bagplot, particularly when all points happen to be on a straight line. The bagplot provides a visualization of several characteristics of the data: its location (the median), spread (the size of the bag), correlation (the orientation of the bag), and skewness (the shape of the bag) [49].

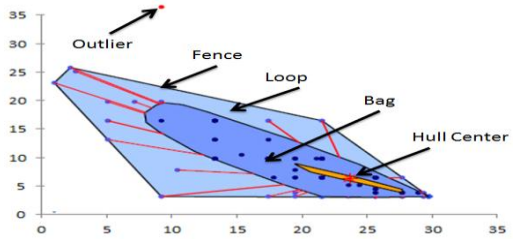


Figure 9. Basic elements of a generic bagplot.

In this statistical analysis, we study the effect of the global sequence similarity on the method performance. Thus, the factor that varies in this analysis is the global similarity information, while other factors will be fixed at their best settings obtained from Figure 10(a). In particular, 1) for the local similarity information, the subsequence pairs of the most local similarity will be used. 2) For the region of similarity, short-range contacts in the diagonal area will be considered. 3) For the evolutionary information, protein pairs will be of the same protein SCOP family. According to the global similarity information of the four categories of the Skolnick dataset (shown in Table I), the pair-wise results are further grouped into four clusters. Namely, 1) Low *vs*. Low, 2) Med *vs*. Med, 3) High *vs*. High (Short), and 4) High *vs*. High (Long). Using the online R statistics software in [47], the bagplots are computed for each cluster, in an effort to perform an in-depth correlation study of the experimental results between short-range contacts and most similar local subsequences at different ranges of global similarity. Although the available samples at the best settings are found to be considerably few, the global sequence information does appear to affect the method performance, as shown in Figure 11. For example, in Figure 11(a), even at the best settings, the center of gravity of the bag is fairly low (around ~62% for contact map similarity) in the case of low global similarity (15-30%). As for the rest of plots, the center of gravity is higher and remains almost the same (around 80% for contact map similarity), when global sequence similarity is medium and high. Samples of the retrieved contact map regions with highly-similar substructural patterns using the proposed pattern matching approach are shown in Figure 12.

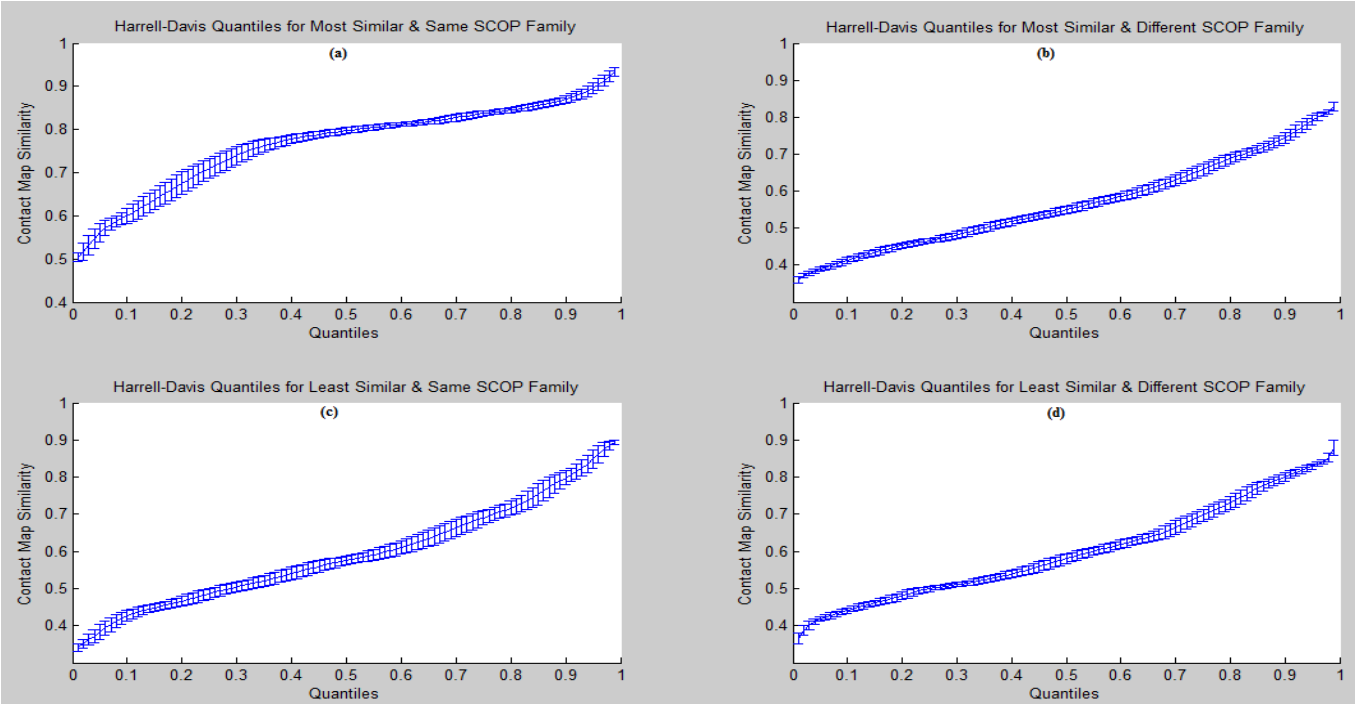


Figure 10. Harrell-Davis quantiles for different categories of the results, along with the error bars of the associated standard error for each reported quantile. (a) Shows the first category of the results of pairs of protein subsequences that are most similar and of the same protein class. (b) Shows category 2 of pairs of protein subsequences that are most similar and of the different protein class. (c) Shows category 3 for pairs of protein subsequences that are least similar and of the same protein class. (d) Shows the last category of pairs of protein subsequences that are least similar and of the different protein class.

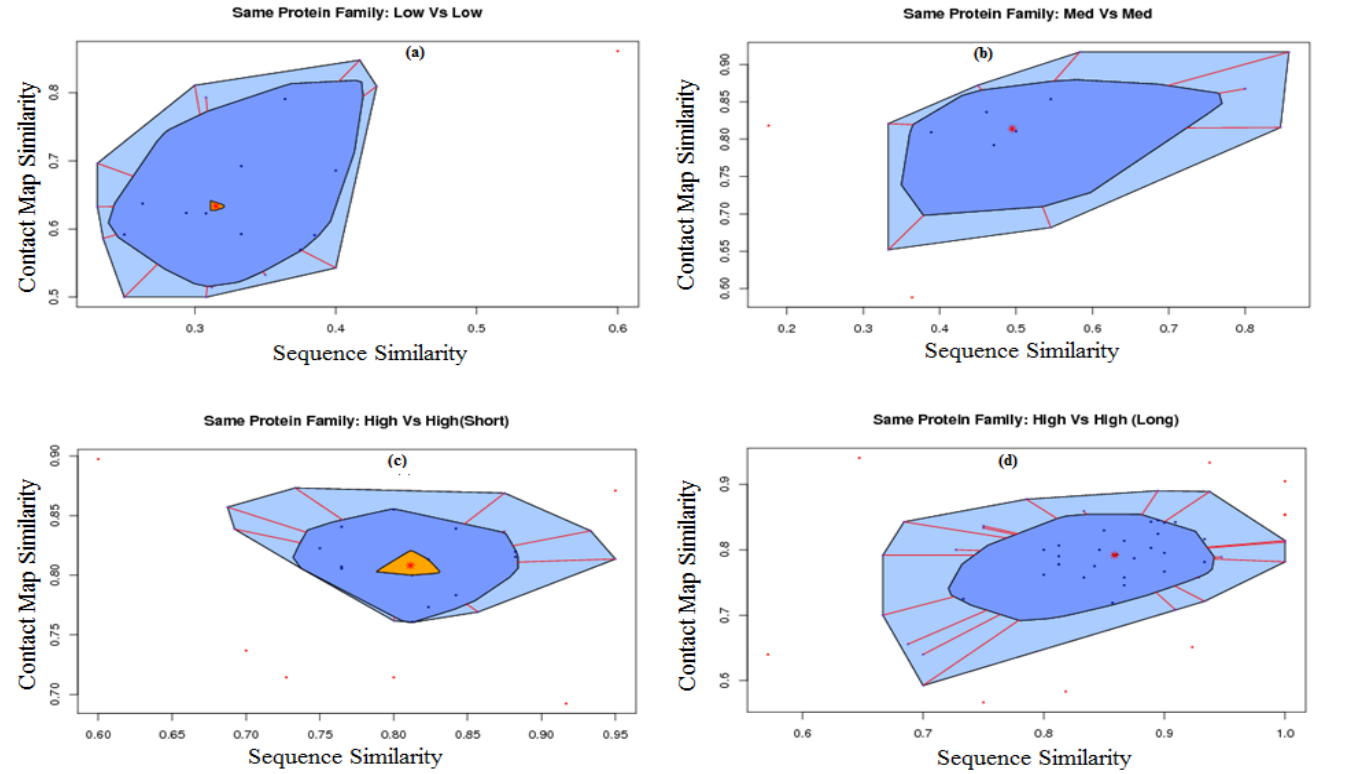


Figure 11. Bagplots for different clusters of the pair-wise results of most similar local subsequences and short-range contacts. (a) Shows the results of first cluster of pairs of protein sequences that are of low global sequence similarity (15-30%). (b) Shows the results of pairs of protein sequences that are of medium global sequence similarity (7 – 70%). (c) Shows the results of pairs of protein sequences that are of high global sequence similarity (35 – 90%) and short length (99 residues). (d) Shows the results of pairs of sequences that are of high global sequence similarity (30-90%) and long length (250 residues).

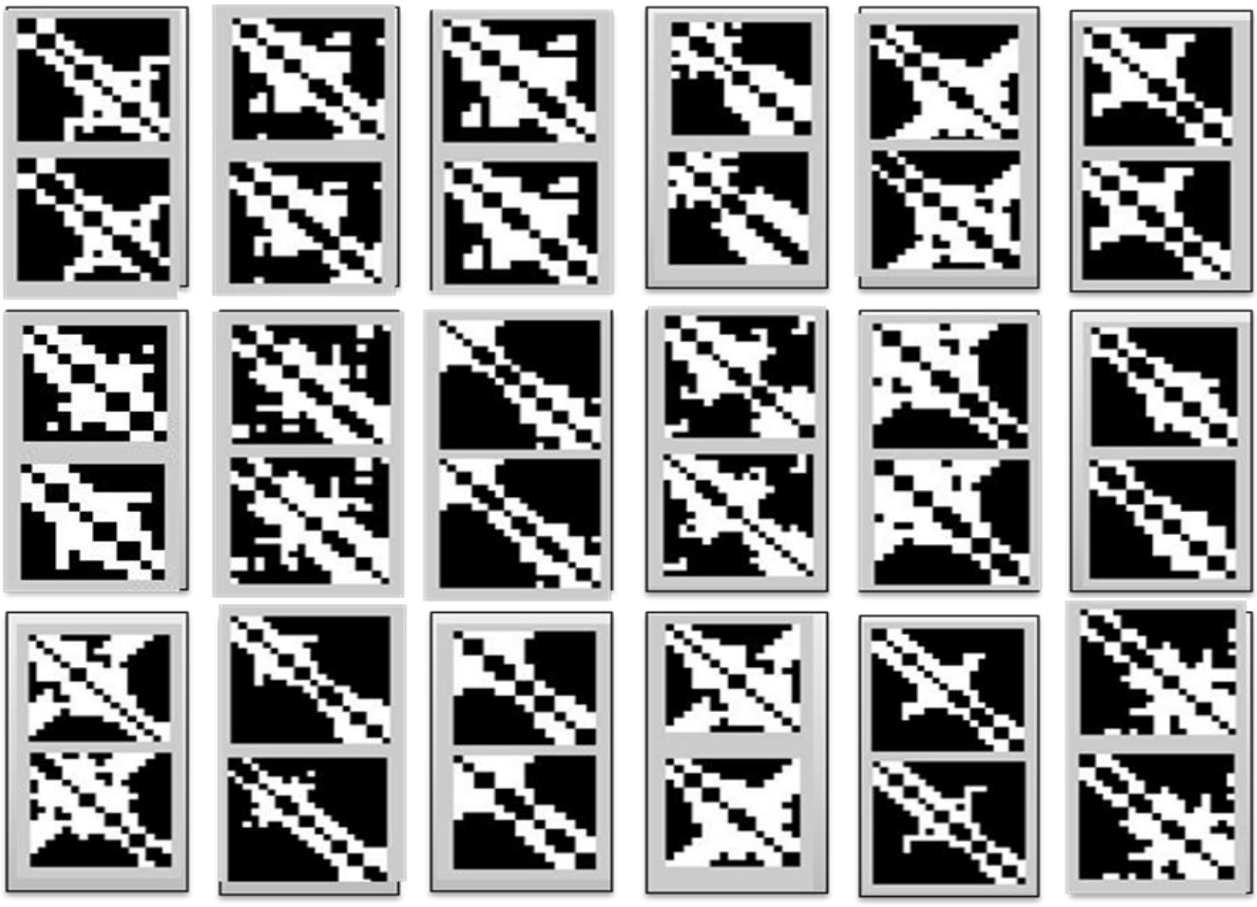


Figure 12. Samples of retrieved contact map regions with similar substructural patterns of 18 contact map pairs with similaty $\geqslant$ 75%.

V. CONCLUSION AND FUTURE WORK

This paper presented the case-based reasoning framework for protein structure prediction from predicted contact maps with the focus on locating similar substructural patterns between the query and template contact maps, as a necessary step for the *Adaptation* phase of the framework. In this paper, a structural pattern matching approach that incorporates both protein sequence and evolutionary information is proposed, with the ultimate goal of identifying the building blocks of proteins in a computational bottom-up approach to protein structure prediction from contact maps. A standard benchmark dataset of carefully-selected 40 large protein domains (Skolnick dataset) is adopted for this study as the experimental dataset.

To the best of our knowledge, this is the first-of-its-kind study to utilize sequence and evolutionary information in locating similar contact map patterns, with no comparable state-of-the-art results. The paper provides an extensive analysis for the three different factors believed to affect the performance of short-range pattern matching in the diagonal area, in particular, 1) local sequence information, 2) evolutionary information, and 3) global sequence information. Firstly, for local sequence information, high sequence similarity (above 60%) has demonstrated (using a scatter-plot analysis) to be a good indicator of a corresponding high diagonal contact map similarity (around 70-90%). This correlation, however, does not appear to be suitable when contacts are long-range (i.e., in the off-diagonal areas of contact maps), or when local sequence similarity is low (less than 60%). Secondly, for evolutionary information, the results proved (using a quantile-based analysis) to be considerably higher when protein pairs have a clear evolutionary relationship, i.e. when they are of the same SCOP family. Lastly, for global sequence information, the results are observed (using a bagplot analysis) to be superior when the global sequence similarity is not low (more than 30%).

Possible future work to improve pattern matching in the diagonal area would be to perform a dynamic expandable multi-regional analysis of contact maps to reduce any possibility of region displacement. That is, we may consider looking further in the neighborhood of the corresponding regions of similar local subsequences. As for the off-diagonal areas, alternative approaches could be employed instead of sequence and evolutionary information that both did not appear helpful in these areas due to the fuzzy nature of long-range contacts at the off-diagonal areas of contact maps [26]. We are currently looking into exploring *Swarm Intelligence* techniques [51] as a promising way to tackle the problem in the off-diagonal areas of contact maps, where the most uncertain, yet important, long-range contacts exist.

REFERENCES

[1] H. R. Ahmed and J. I. Glasgow, "Incorporating Protein Sequence and Evolutionary Information in a Structural Pattern Matching Approach for Contact Maps", The 3rd International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO'11), Venice, Italy, 2011.

[2] F. Collins, M. Morgan, and A. Patrinos, "The human genome project: lessons from large-scale biology," *Science*, vol. 300, 2003, pp. 286–290.

[3] R. D. Schaeffer and V. Daggett, "Protein folds and protein folding," *Protein Engineering, Design and Selection*, Vol. 24, no. 1-2, 2010, pp. 11-19.

[4] A. C. Anderson, "The process of structure-based drug design," *Chemistry and Biology*, vol. 10, 2003, pp. 787–797.

[5] J. Drenth, "Principles of protein X-ray crystallography," *Springer-Verlag*, New York, 1999, ISBN 0-387-98587-5.

[6] M. Schneider, X. R. Fu, and A. E. Keating, "X-ray versus NMR structures as templates for computational protein design," *Proteins*, vol. 77, no. 1, 2009, pp. 97–110.

[7] A. Kolinski (Ed.), "Multiscale approaches to protein modeling," 1st Edition, Chapter 10, *Springer*, 2011, ISBN 978-1-4419-6888-3.

[8] P. R. Daga, R. Y. Patel, and R. J. Doerksen, "Template-based protein modeling: recent methodological advances," *Current Topics in Medicinal Chemistry*, vol. 10, no. 1, 2010 , pp. 84-94.

[9] C. Yang et al., "Improved side-chain modeling by coupling clash-detection guided iterative search with rotamer relaxation," *Bioinformatics*, 2011, doi:10.1093/bioinformatics/btr00.

[10] F. Birzele and S. Kramer, "A new representation for protein secondary structure prediction based on frequent patterns," *Bioinformatics*, vol. 22, 2006, pp. 2628–2634.

[11] G. Pollastri and P. Baldi, "Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners," *Bioinformatics*, vol. 18, 2002, pp. 62–70.

[12] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio, "Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations," *Proteins*, vol. 5, 2001, pp. 157–62.

[13] Y. Zhao and G. Karypis, "Prediction of contact maps using support vector machines," *Proceedings of 3rd IEEE International Symposium on BioInformatics and BioEngineering (BIBE)*, Bethesda, MD, 2003, pp. 26–36.

[14] M. Zaki, J. Shan and C. Bystroff, "Mining residue contacts in proteins using local structure predictions," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 33, no. 5, 2003, pp. 789-801.

[15] D. Thomas, G. Casari and C. Sander, "The prediction of protein contacts from multiple sequence alignments," *Protein Eng*. vol. 9, 1996, pp. 941–948.

[16] M. Singer, G. Vriend and R. Bywater, "Prediction of protein residue contacts with a PDB-derived likelihood matrix," *Protein Eng.*, vol. 15, 2002, pp. 721–725.

[17] Y. Shao and C. Bystroff, "Predicting interresidue contacts using templates and pathways," *Proteins*, vol. 53, 2003, pp. 497–502.

[18] J. Kolodner, "An introduction to case-based reasoning," *Artificial Intelligence Review*, vol. 6, 1992, pp. 3-34.

[19] J. Glasgow, T. Kuo, and J. Davies, "Protein structure from contact maps: a case-based reasoning approach," *Information Systems Frontiers*, Special Issue on Knowledge Discovery in High-Throughput Biological Domains, Springer, vol. 8, no. 1, 2006, pp. 29-36.

[20] I. Walsh, A. Vullo, and G. Pollastri, "XXStout: improving the prediction of long range residue contacts," *ISMB 2006*, Fortaleza, Brazil.

[21] M. Vendruscolo, E. Kussell and E. Domany, "Recovery of protein structure from contact maps," *Folding and Design*, vol. 2, no. 5, 1997, pp. 295-306.

[22] P. Fariselli and R. Casadio, "A neural network based predictor of residue contacts in proteins," *Protein Eng.* vol. 12, 1999, pp.15–21.

[23] L. Mirny and E. Domany, "Protein fold recognition and dynamics in the space of contact maps," *Proteins*, vol. 26, 1996, pp. 391–410.

[24] M. Berrera, H. Molinari, and F. Fogolari, "Amino acid empirical contact energy definitions for fold recognition in the space of contact maps," *BMC Bioinformatics*, vol. 4, no. 8, 2003.

[25] M. Vassura et al., "Reconstruction of 3D structures from protein contact maps," *Proceedings of 3rd International Symposium on Bioinformatics Research and Applications*, Berlin, Springer, vol. 4463, 2007, pp. 578–589.

[26] X. Yuan and C. Bystroff, "Protein contact map prediction," *in Computational Methods for Protein Structure Prediction and Modeling*, Springer, 2007, pp. 255-277, doi:10.1007/978-0-387-68372-0_8.

[27] O. Gotoh, "An improved algorithm for matching biological sequences," *J. Mol. Biol.*, vol. 162, 1982, pp. 705-708.

[28] S. F. Altschul and B. W. Erickson, "Optimal sequence alignment using affine gap costs," *Bull. Math. Biol*., vol. 48, 1986, pp. 603-616.

[29] M. Dayhoff, "A model of evolutionary change in proteins, "*Atlas of Protein Sequence and Structure*, vol. 5, no. 3, 1978, pp. 345 – 352.

[30] E. Krissinel, "On the relationship between sequence and structure similarities in proteomics," *Bioinformatics*, vol. 23, 2007, pp. 717–723.

[31] Dictionary of secondary structure of proteins: available at http://swift.cmbi.ru.nl/gv/dssp/, 11.06.2012.

[32] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *J. Mol. Biol*., vol. 247, 1995, pp. 536–540.

[33] Gunnar W. Klau, "Comparing structural information in the life sciences: from RNA to metabolic networks," *Symposium on Bioinformatics and Biomathematics*, 2007.

[34] G. Lancia, R. Carr, B. Walenz, and S. Istrail, "101 Optimal PDB structure alignments: A branch-and-cut algorithm for the maximum contact map overlap problem," *Proceedings of Annual International Conference on Computational Biology (RECOMB)*, 2001, pp. 193-202.

[35] W. Xie and N. V. Sahinidis, "A branch-and-reduce algorithm for the contact map overlap problem," *Proceedings of RECOMB of Lecture Notes in Bioinformatics*, Springer, vol. 3909, 2006, pp. 516-529.

[36] P. Lena, P. Fariselli, L. Margara, M. Vassura, and R. Casadio, "Fast overlapping of protein contact maps by alignment of eigenvectors," *Bioinformatics*, vol. 26, no. 18, 2010, pp. 2250-2258. doi: 10.1093

[37] H. Xiaoquin and W. Miller, "A time-efficient, linear-space local similarity algorithm," *Advances in Applied Mathematics*, vol. 12, 1991, pp. 337-357.

[38] SIM: Alignment Tool for Protein Sequences, available at http://ca.expasy.org/tools/sim-prot.html, 11.06.2012.

[39] V. Arjunan, S. Nanda, S. Deris, and M. Illias, "Literature survey of protein secondary structure prediction," *Journal Teknologi*, vol. 34, 2001, pp. 63-72.

[40] Y. Xu, D. Xu, and J. Liang (Eds.), "Computational methods for protein structure and modeling," *Springer*, Berlin, 2007, ISBN: 978-1-4419-2206-9

[41] Henikoff and Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the national academy of sciences*, USA, vol. 89, 1992, pp. 10915-10919.

[42] S. F. Altschul and B. W. Erickson, "Optimal sequence alignment using affine gap costs," *Bull. Math. Biol*., vol. 48, 1986, pp. 603-616.

[43] K Teknomo, Similarity Measurement, available at: http:\\people.revoledu.com\kardi\ tutorial\Similarity, 11.06.2012.

[44] L. Lee, "Measures of distributional similarity," *Proceedings of the 37th annual meeting of ACL*, 1999, pp. 25–32.

[45] H. R. Ahmed and J. I. Glasgow, "Multi-regional analysis of contact maps towards locating common substructural patterns of proteins," *J Communications of SIWN*, vol 6, 2009, pp.90-98.

[46] P. Wessa, "Harrell-Davis quantile estimator", *in Free Statistics Software*, Office for Research Development and Education, 2007, URL: http://www.wessa.net/rwasp_harrell_davies.wasp/, 11.06.2012.

[47] P. Wessa, "Bagplot," *in Free Statistics Software*, Office for Research Development and Education, 2009, URL: http://www.wessa.net/ rwasp_bagplot.wasp/, 11.06.2012.

[48] F. E. Harrell and C. E. Davis, "A new distribution-free quantile estimator," *Biometrika*, vol. 69, 1982, pp. 635-640.

[49] P. J. Rousseeuw, I. Ruts, and J. W. Tukey, "The bagplot: A bivariate boxplot," *The American Statistician*, vol. 53, 1999, pp. 382–387.

[50] D. F. Williamson, R. A. Parker, and J. S. Kendrick, "The box plot: a simple visual method to interpret data," *Ann Intern Med*, vol. 110, 1989, pp. 916-921.

[51] S. Das, A. Abraham and A. Konar, "Swarm Intelligence Algorithms in Bioinformatics," *Studies in Computational Intelligence*. vol. 94, 2008, pp. 113–147.

Simulating Gene Expression Data To Estimate Sample Size For Class and Biomarker Discovery

Jiexin Zhang, Paul L. Roebuck, and Kevin R. Coombes
Department of Bioinformatics and Computational Biology
University of Texas M.D. Anderson Cancer Center
Houston, TX 77005, USA
Email: kcoombes@mdanderson.org
proebuck@mdanderson.org
jiexinzhang@mdanderson.org

***Abstract*—With modern advances in high-throughput technologies to measure gene expression profiles, researchers are eager to identify biomarkers that indicate pathogenic processes or pharmacologic responses. However, insufficient statistical power, often due to the limited sample sizes in real experiments, has hindered progress in this area. Realistic simulations can provide data to better estimate sample sizes and better evaluate analytical methods. Existing simulation tools have focused more on the technology and less on the biological complexity of patients and outcomes. In this paper, we describe an R package of gene expression simulation tools to address this problem. Our model incorporates both biological and technical noise on top of the true signal, transcriptional status, and block structures that mimic gene networks. More importantly, to simulate the multi-hit model of cancer development, our tool contains latent variables that link gene expression with binary outcome and survival data. We demonstrate the use of this R package by providing examples of simulated cancer subtype recovery and biomarker discovery.**

***Keywords*-gene expression; microarray; simulation; class prediction; multi-hit theory of cancer; biomarker**

I. Introduction

The "Ultimate Microarray Prediction, Inference, and Reality Engine" (Umpire) is an R package that allows researchers to simulate complex, realistic microarray data [1]. Simulations are useful for designing experiments and for evaluating proposed analytical methods. The simulation of microarray gene expression data sets has a long history: many of the earliest simulation tools focused on the simulation of microarray images, and were useful for developing better image processing algorithms [2]–[4]. Other simulation tools have attempted to explicitly model the steps in a microarray experiment, including printing, hybridization, dye effects, and scanning [5], [6]. As with many of the early statistical simulations [7]–[10], however, most tools use a model that simply compares two homogeneous populations of samples. Even more recent and more detailed simulations still assume that the data come from two homogenous populations [11]–[14]. Moreover, none of the existing simulation tools was designed to focus on the biological diversity related to such important outcomes as treatment response or survival.

To address this gap, we developed the `Umpire` package, which incorporates a heterogeneous model consistent with the multiple hit theory of carcinogenesis [15], [16]. Our package uses latent variables to simulate the connections between gene expression and either binary or time-to-event outcomes. Latent variables, also called hidden variables, are usually inferred from other variables rather than being observed directly [17]. For example, the latent variables in our simulation can be cancer subtypes that correspond to different survival rates, or biomarker expression levels that are linked with different treatment effects.

Advances in high-throughput technologies for gene expression measurement have spurred the development of analytical methods for dealing with the explosion of large amounts of biological data [18]–[21]. Three major questions addressed by these technologies are class comparison, class discovery, and class prediction [22]. The goal in class comparison is to find biological entities whose distributions differ among some pre-defined sample groups. Methods for class comparison include gene-by-gene t-tests or ANOVA coupled with multiple testing adjustments [23]. Class discovery involves performing unsupervised analyses to "learn" or "discover" subgroup structures in the data. The current state-of-the-art has evolved reasonable methods for class discovery, such as hierarchical clustering coupled with resampling techniques to assess robustness [24]. The goal of class prediction is to formulate gene signatures from a training data set, and then use the signatures to assign new samples to known classes [25]. The performance of class prediction methods is assessed with a rigorous approach involving independent testing data. There are some known pitfalls to building predictive models from microarray gene expression data that need special attention [26], [27]. Some studies have tried different strategies to boost the performance of class prediction [28], [29]. However, even though class prediction is the most important of the three problems, there is less agreement on the best (or even consistently good) methods for discovering complex models that can accurately predict biologically relevant outcomes such as treatment response or survival.

In spite of the difficulty in class prediction, there is an explosion of interest in biomarker research with the goal of incorporating biomarkers into drug development and leading to personalized medicine [30]–[37]. For example, about 30% of patients with breast cancer over-express the protein HER2, a member of the human epidermal growth factor receptor family. These patients do not respond to standard therapy, but benefit from Herceptin treatment in combination with chemotherapy [38]. This example illustrates the potential utility of biomarkers for patient selection. By selecting patients based on their biomarker profiles, we hope to enrich the pool of patients who have a greater probability of response to alternative treatment plans. If successful, this approach could lead to cheaper and faster clinical trials than the conventional ones.

Appropriate experimental designs are crucial to the biomarker discovery process. Sample size determination is a critical step in experimental design to ensure sufficient statistical power for making inferences about a population from a sample [39]. It is conceivable that the number of samples (typically between 100 and 300) included in most current studies is simply inadequate to learn effective predictive models. On one hand, the soundness of analytical tools cannot be evaluated accurately given the small sample size and the unknown "ground truth" of biology. On the other hand, biological changes can be masked by noise, which requires large number of samples in order to reveal the true signal. It is, however, extremely difficult to assess the possibility that more samples (and how many more) would convey sufficient predictive power. Although some progress has been made for binary classifiers [13], [40], [41], we do not have general theoretical methods to justify formal sample size computations that address the combination of feature selection and model building that goes into the discovery of predictive models from high-throughput biological data sets. Nor is it possible to collect gene expression data on $10,000$ patients in order to test empirically how many samples are really needed to learn good predictive models.

The obvious solution is to use simulation. If we can simulate many data sets, of different sizes, with realistic biological properties, then we can use those data sets to evaluate proposed methods for class prediction. Using the `Umpire` simulation package, we can generate realistic data to help answer the questions above. In the following sections, we first elaborate the design of `Umpire` and the parameters we implemented in the current version. We then discuss results from two sets of simulations to demonstrate the use of `Umpire` for cancer subtype recovery and biomarker discovery, respectively.

II. Homogeneous gene expression model

We begin by describing the underlying statistical model for simulating gene expression data that is implemented in the `Umpire` package. The fundamental object is a "random-vector generator" (RVG), which represents a specific multivariate distribution from which random vectors can be generated.

A. Additive and Multiplicative Noise

The observed signal, Y_{gi}, for gene g in sample i is:

$$Y_{gi} = \exp(H_{gi})S_{gi} + E_{gi}$$

where

$$S_{gi} = \text{true biological signal}$$

$$H_{gi} = \text{multiplicative noise}$$

$$E_{gi} = \text{additive noise.}$$

The noise model represents technical noise that is layered on top of any biological variability when measuring gene expression in a set of samples. Usually the microarray noise is considered a combination of additive and multiplicative components [42]. We modeled additive and multiplicative noise as normal distributions:

$$E_{gi} \sim \text{Normal}(\nu, \tau)$$

$$H_{gi} \sim \text{Normal}(0, \phi)$$

Note that we allow the additive noise to include a bias term (ν) that may represent, for example, a low level of cross-hybridization contributing some level of signal at all genes. The noise model is represented in the `Umpire` package by the `NoiseModel` class. The object-oriented and modular design makes it possible to add more elaborate noise models in the future, such as those described by Nykter and colleagues [5].

B. Active and Inactive Genes

We model the true biological signal S_{gi} as a mixture:

$$S_{gi} \sim (1 - z_g)\delta_0 + z_g T_{gi}$$

In this model, δ_0 is a point mass at zero, z_g defines the activity state (1 = active, 0 = inactive), and T_{gi} is the expression of a transcriptionally active gene. By allowing for some genes to be transcriptionally inactive, this design takes into account that the transcriptional activity of most genes is conditional on the biological context. Activity is modeled in `Umpire` using a binomial distribution, $z_g \sim Binom(p_0)$.

C. Expression Distributions

For most purposes, we assume that the expression, T_{gi}, of a transcriptionally active gene follows a log-normal distribution, $\log(T_g) \sim Normal(\mu_g, \sigma_g)$. In a class of samples, the mean expression of gene g on the log scale is denoted by μ_g and the standard deviation on the log scale is σ_g. Both μ_g and σ_g are properties of the gene itself and the sample class. Within a given simulation, we typically place hyperdistributions on the log-normal parameters μ_g and σ_g. We take $\mu_g \sim Normal(\mu_0, \sigma_0)$ to have a normal distribution with mean μ_0 and standard deviation σ_0. We take σ_g to have an inverse gamma distribution with *rate* and *shape* parameters. Reasonable values for the hyper-parameters can be estimated from real data. For instance, $\mu_0 = 6$ and $\sigma_0 = 1.5$ are typical values on the log scale of a

microarray experiment using the Affymetrix GeneChip® human arrays. The parameters for the inverse gamma distribution are determined by the method of moments from the desired mean and standard deviation; we have found that a mean of 0.65 and a standard deviation of 0.01 (for which $rate = 28.11$ and $shape = 44.25$) produce reasonable data.

D. Correlated blocks of genes

Biologically, genes are usually interconnected in networks and pathways. In fact, clustering methods are often used to group genes into correlated blocks. Thus, it is natural to simulate microarray experiments from this perspective. In our simulations, we usually allow the mean block size, ξ, to range from 1 to 1000, and the sizes of gene blocks to vary around the pre-defined mean block size. To be more specific, the block size follows a normal distribution with mean ξ and standard deviation $0.3\,\xi$. The case $\xi = 1$ is special, since we take the standard deviation of the block size to be zero so all genes are independent. At the other extreme, $\xi = 1000$ simulates large networks involving many genes.

The correlation matrix (Ω_b) for a block b, has 1's on the diagonal and ρ_b or $-\rho_b$ in the off-diagonal entries. We usually allow $\rho \sim Beta(pw, (1-p)w)$ to follow a beta distribution with parameters $p = 0.6$ and $w = 5$. We let θ denote the portion of negatively correlated genes within a block. In the simplest scenario, all genes in the same block have the same positive correlation ρ_b. In a more complicated scenario, $\theta = 0.5 - |x - 0.5|$ where x follows a beta distribution. Three types of x are considered: (1) $x \sim \text{Beta}(1,1)$, so θ is uniformly distributed between 0 and 0.5; (2) $x \sim \text{Beta}(5,5)$, so θ is likely to be close to 0.5; or (3) $x \sim \text{Beta}(0.5, 0.5)$, so θ is likely to be close to 0. Our pilot study showed that different θ's do not have a pronounced impact on the parameters of interest (data not shown). So, we only discuss the results obtained from $\theta = 0.5 - |x - 0.5|$ where $x \sim \text{Beta}(1,1)$.

The log expression values of genes within a block follow a multivariate normal (MVN) distribution. The mean vector is defined by μ_g as defined previously, and the covariance matrix Σ is defined as:

$$\Sigma_{i,j} = \Omega_{i,j} * \sigma_{g_i} * \sigma_{g_j}$$

where σ_{g_i} defines the standard deviation of gene i, which follows the inverse gamma distribution as described previously. More elaborate models can also be generated, by altering the variances or the correlation structure within the block.

We mentioned above that some genes would be transcriptionally inactive under certain biological conditions. Instead of simulating this active status for genes individually, we simulate whole blocks of genes being transcriptionally active or inactive. This models the idea that the entire pathway or network could be turned on or off under certain biological conditions.

III. The Multi-hit Model of Cancer

The multiple hit theory of cancer was first proposed by Carl Nordling in 1953 [15] and extended by Alfred Knudson in 1971 [16]. The basic idea is that cancer can only result after multiple insults (mutations; hits) to the DNA of a cell. We use the combinatorics of multiple hits to simulate heterogeneity in the population.

Let H be the number of possible hits (typically on the order of 10 to 20). We define a cancer subtype as a collection of hits (usually 5 or 6 out of those possible). Each subtype has a prevalence; by default, each subtype is equally likely to occur in the population. To simulate a set of patients, we start by assigning them to one of the cancer subtypes (with probabilities equal to the prevalences). We then use the individual hits as (unobserved) latent variables that influence gene expression, survival, and binary outcomes.

Specifically, let Z_h be a binary variable that indicates the presence ($Z_h = 1$) or absence ($Z_h = 0$) of a hit h. Then the probability p of an unfavorable (binary) outcome is simulated from a logistic model

$$log\left(\frac{p}{1-p}\right) = \sum_{h=1}^{H} \beta_i Z_i,$$

where the parameters $\beta_i \sim N(0, \sigma_B)$ are simulated from a normal distribution.

We simulate survival times from a Cox proportional hazards model [43], with

$$h(t) = h_0(t) \sum_{h=1}^{H} \alpha_i Z_i,$$

where $h_0(t)$ can be taken to be any desired survival model (usually exponential) and the coefficients $\alpha_i \sim N(0, \sigma_A)$ can be taken to be either independent of or related to the β_i depending on the goal of the simulation.

Finally, each hit is assumed to affect the expression of one correlated block of genes (representing the effect on a single biologically pathway) by altering the mean expression of the genes in that block. The absolute change of the mean expression values on log scale for a block of genes is given by $\Delta_g \sim \text{Gamma}(\alpha, \beta)$. Both parameters for this gamma distribution are set to 10 so that the absolute fold change on the log2 scale is 1, and the long tail on the right hand side of the distribution allows a few genes to have large fold changes. A gene in the changed block is randomly assigned to be up-regulated or down-regulated in cancer patients.

IV. Implementation

The statistical model that we have just described is implemented using S4 classes in the R statistical software programming environment. Version 1.2.3 of the `Umpire` package is available from the R repository at http://bioinformatics.mdanderson.org/OOMPA; detailed instructions on how to install the package can be found at http://bioinformatics.mdanderson.org/Software/

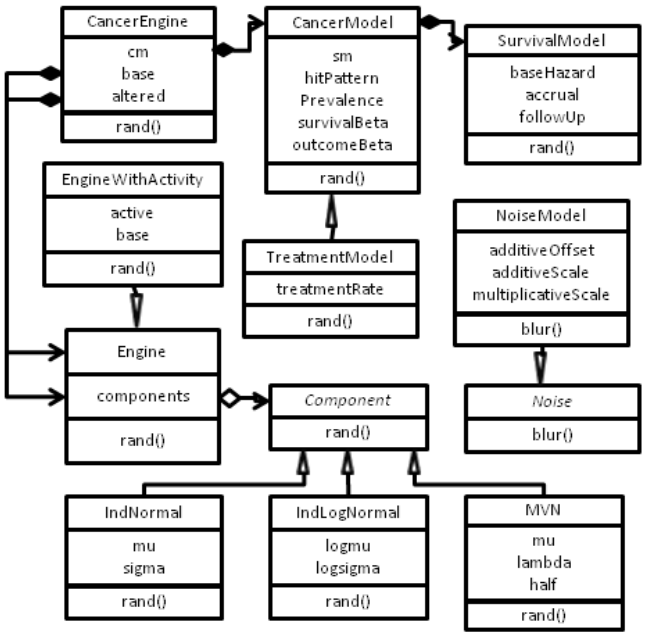


Fig. 1. UML diagram of classes in the Umpire package.

OOMPA. Figure 1 presents a diagram of the class structure using the Unified Modeling Language (UML). The main class, `CancerEngine`, contains one `CancerModel` and two `Engine` objects. The `CancerModel` object is used to simulate clinical data, including cancer subtypes, binary outcomes, and survival times. This object contains a matrix of hit patterns and a vector of prevalences that characterize the cancer subtypes being simulated.

The basic survival model assumes that the survival times follow an exponential distribution; other survival distributions can be simulated by deriving a subclass of `SurvivalModel`. The `rand()` method for `SurvivalModel`s takes an optional extra parameter, β, that represents a vector of logarithmic hazard ratios to modify the survival distributions for individual patients depending on the latent pattern of hits and, possibly, the treatment they receive.

Each `Engine` is used to simulate vectors of gene expression data. An `Engine` is a list of components; for the simulations described in this paper, we use a combination of `IndependentNormal` and `MVN` (multivariate normal) components. Additional components can be derived from the abstract `Component` class to simulate data from other distributions. For example, one might use Poisson distributions or negative binomial distributions to simulate the kinds of count-based gene expression data that are produced by next generation sequencing technologies. The pair of `Engine` objects in a `CancerEngine` represent the baseline gene expression (with no hit) and the altered gene expression that occurs in the presence of a hit; which expression pattern is used for any simulated sample depends on the subtype and hit pattern generated by the associated `CancerModel`.

Noise is applied to simulated gene expression data, using the `blur()` method, after the "true" signal is simulated. In the simulation presented here, we use a straightforward model of additive and multiplicative white noise. The general design, however, allows for the incorporation of more elaborate noise models by deriving additonal subclasses of the abstract `Noise` class.

The block structure is only indirectly specified by the class structure. For the simulations presented here, we implement it by constructing `Engine` objects consisting of `MVN` components with block sizes drawn from an appropriate distribution.

TABLE I
Number of significant genes, by sample size and FDR.

	N = 100	N = 300	N = 500
FDR = 0.01	12	86	144
FDR = 0.05	22	135	209
FDR = 0.1	37	169	253
FDR = 0.2	74	249	354
FDR = 0.3	127	346	446

V. Simulation Results

To illustrate the usage of the `Umpire` package, we performed two sets of simulation of microarray data with associated survival data.

A. Cancer Subtype Recovery

In the first simulation, we assumed that there are 20 possible hits (H1 to H20), and that 5 hits at a time define a cancer subtype. We also assumed that there were 6 distinct, equally likely, cancer subtypes. As above, each of the 20 hits corresponds to a correlated block of gene expression and also affects survival. We assumed that there were 100 additional correlated blocks of genes that were unrelated to cancer or to survival. Blocks were simulated to contain a mean of 100 genes with a standard deviation of 30. Gene means, standard deviations, and correlation structures were simulated using the distributions and hyperparameters described above. We simulated survival by assuming an exponential baseline hazard function.

We analyzed the simulated data using an approach that is common in the field. Specifically, we fit gene-by-gene univariate Cox proportional hazards models. We recorded the p values for a log-rank test of the significance of each gene. We then fit a beta-uniform mixture (BUM) model [44] to the set of p-values, and used the BUM model to estimate the false discovery rate (FDR). Table I shows the number of genes called significant as a function of the FDR and the sample size. For an FDR of 20%, Table II separates these results into groups depending on the membership of genes in different correlated blocks. Recall that 20 correlated blocks of genes were associated with cancer-related hits; the blocks of "irrelevant" genes are collected in the row of the table labeled "FP" to denote obvious false positive findings. The first column of Table II shows the number of cancer subtypes (patterns) that included each hit; the second column shows the coefficient of that (latent) hit

TABLE II
NUMBER OF SIGNIFICANT GENES AS A FUNCTION OF THE SAMPLE SIZE AND THE TRUE HIT STATUS.

	Patterns	Alpha	N = 100	N = 300	N = 500
H1	4	0.291	0	8	10
H2	2	0.366	0	5	11
H3	1	0.090	0	3	11
H4	0	0.278	0	1	0
H5	1	1.428	0	2	2
H6	3	0.313	0	1	2
H7	0	0.496	0	0	0
H8	1	-0.428	1	5	13
H9	3	-2.135	6	34	40
H10	0	0.631	2	1	0
H11	1	0.047	17	38	44
H12	2	0.422	0	13	27
H13	2	1.062	1	7	12
H14	0	1.433	0	2	0
H15	2	2.514	0	6	15
H16	1	-0.384	0	3	3
H17	1	-0.841	1	10	14
H18	2	0.299	0	13	16
H19	2	1.358	10	25	32
H20	2	-1.674	6	35	41
FP	0	0.000	30	37	61

in the simulated survival model. Note that even though there were 20 possible hits, four of them (G4, G7, G10, and G14) were not actually included in the patterns of 5 hits that defined the 6 cancer subtypes in this simulation. Using 100 samples, we only discovered multiple genes that represented 5 of the cancer-related gene blocks. Using 500 samples, we discovered multiple genes representing all 16 "active" cancer-related gene blocks.

Figure 2 displays heatmaps of the genes selected as significant at the 20% FDR level using either 100 or 500 samples. The color bar along the top reflects the true cancer subtype for each patient. The color bar along the side displays the gene memberships in cancer-related gene blocks, with white representing genes belonging to non-cancer-related blocks, which are false positives. When using 100 samples, not all patients with different cancer subtypes are well separated. We observe distinct gene expression patterns in patients with subtype 1 and 5, but not in other patients. On the gene level, the 74 significant genes come from 8 cancer-related gene blocks. With 500 samples, all six cancer subtypes are well separated by clustering, and their distinct gene expression patterns are visible in the heatmap. On the gene level, the 354 significant genes cover 16 out of 20 cancer-related gene blocks. In both heatmaps, the false positive genes, represented by the white color bar, are recognizable by their lack of correlation with other selected genes.

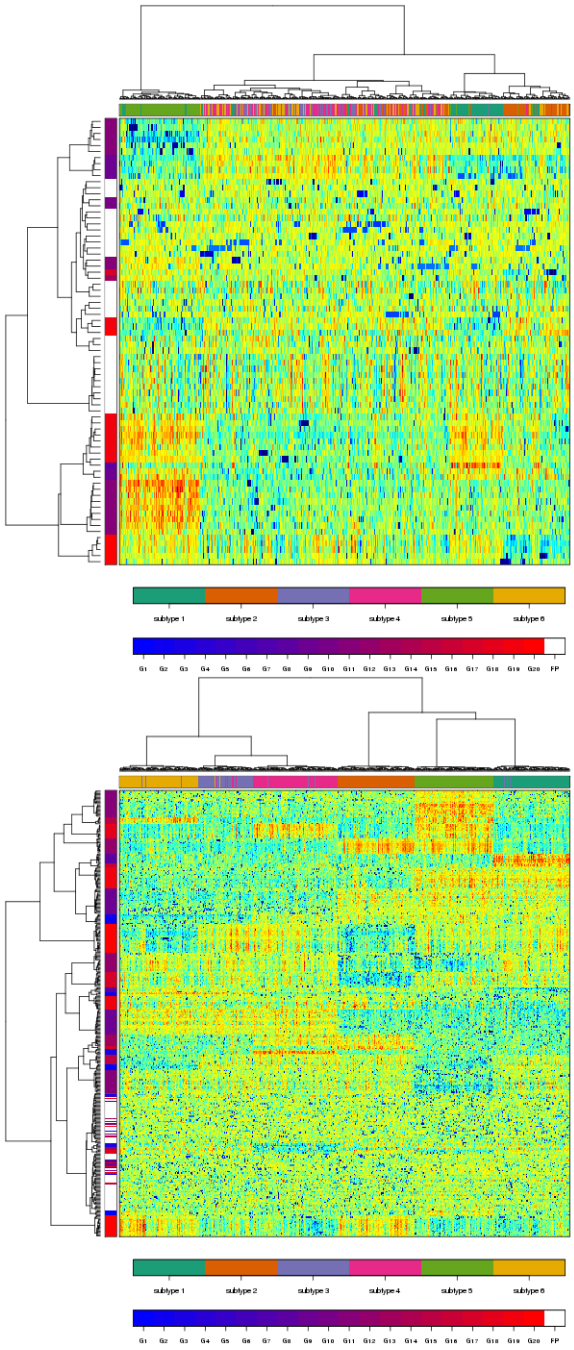


Fig. 2. Heatmaps of the significant genes at FDR = 20% using 100 (top) or 500 (bottom) samples.

B. Patient Selection In Clinical Trials

The second set of simulations involves biomarker identification and patient selection during clinical trials. We assume that a randomized clinical trial is conducted with two arms with equal probability to compare the performance of some standard therapy with a potentially better alternative therapy. The hazard ratio between the alternative treatment and the standard treatment in the full population is called HR_{trt}. We simulated time-to-event outcome, which might represent overall survival, progression-free survival, or other similar clinically relevant endpoints, between the two arms. Note that other types of endpoints can be easily added to the `Umpire` package. A latent variable L indicates whether each patient is marker-positive ($M+$) or marker-negative ($M-$). The time-to-event outcome is linked with the treatment and the latent variable. Only $M+$ patients will benefit from the alternative treatment.

Genes in five correlated blocks out of 100 total blocks are

differentially expressed between $M+$ and $M-$ groups. The goal is to identify some complex (probably multivariate) marker that separates the initial patient population into two groups ($M+$ and $M-$), such that the hazard ratio between treatment arms in the $M+$ group, HR_{M+}, is a substantial improvement over the hazard ratio HR_{Trt} in the full population.

We simulated survival using an exponential baseline hazard function with a median progression-free survival time of 18 weeks. The true benefit in the patients who have the marker is simulated as $HR_{M+} = 0.55$. Assuming that 30% of patients contain this marker, as in the example of HER2 described above, we simulated different sizes of patient cohorts ranging from 100 to 1500. Each scenatio was simulated 10 times for variance estimation. We also simulated independent testing data sets of size 200.

For each training data set, we performed K-means clustering [45] on each gene with $K = 2$. To select potential biomarkers, we searched for genes whose two groups corresponded to different hazard ratios between the two treatment arms. We fit gene-by-gene univariate Cox proportional hazards models. The p-values corresponding to the interaction term between treatment and the gene grouping are further modeled using the BUM model to estimate the FDR. With FDR cutoff 20%, we selected significant genes for each set of training data. Similarly, K-means clustering was performed on each gene in the test data sets. We then calculated the percentage of significant genes voting for $M+$ as a multivariate predictor that a patient is $M+$. Figure 3 shows receiver operating characteristic (ROC) curves [46] of the predictions in the testing data sets for different size training data sets. We observe that more training samples yield more accurate predictions. In this simulation, the area under the ROC curve (AUC) is larger than 0.9 when the number of patients is at least 500.

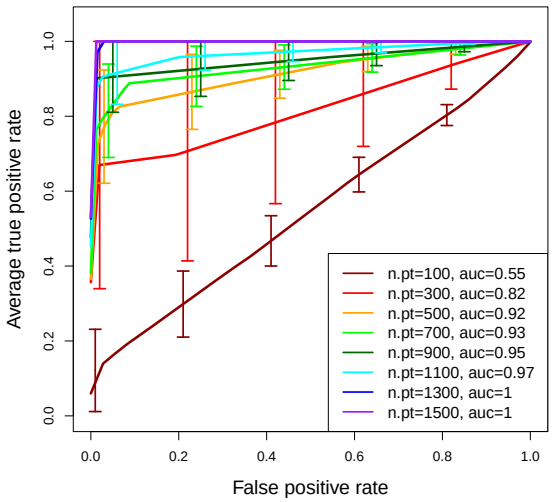


Fig. 3. ROC curves for patient selection with markers identified from different sized patient cohorts. The vertical bars correspond to standard error from 10 simulation, and the AUCs are shown in the legend.

VI. Conclusion

We have described the `Umpire` R package and shown that it can be used to simulate microarray data that is related to survival outcomes in complex ways. In our simulation, many assumptions are based on our extensive experience derived from working with real Affymetrix GeneChip® data sets. We recognize that some of the modeling assumptions that we used might seem simplified considering the complex biology. However, one advantage of implementing `Umpire` with S4 classes in R is that the package is flexible enough to allow easy addition of components representing alternative models of gene expression.

The two sets of simulations that we have presented, which use a plausible set of biologically meaningful parameters, suggest that both class discovery studies and biomarker discovery studies looking for signatures to predict time-to-event outcomes may need more than the 100–300 samples that have frequently been used in practice. In order to elucidate the true subgroup structure, our first simulation required about 500 samples. In order to discover biomarker signatures that could identify a subgroup of patients more likely to respond to an alternative treatment, our second simulation also0 required at least 500 patients. In this context, it is interesting to note that the ongoing effort of The Cancer Genome Atlas (TCGA) to apply comprehensive high-throughput molecular biology techniques to a variety of different cancers intends to study about 500 samples of each type [47].

The results of the simulation also suggest that we may need better methods for combining gene expression values into predictive signatures. First, the common statistical approach that tries to optimize the coefficients of all 354 selected genes using 500 samples is unlikely to succeed. Moreover, since we know "ground truth" for this particular simulation, we know that there are 16 independent factors that influence survival. From the heatmap on the bottom of Figure 2, we would also estimate that there are many distinct expression patterns that contribute to survival. This observation suggests two possible approaches. On the one hand, we could group correlated genes together into simpler factors that can be included in predictive models. For example, we could perform a principal components analysis and use the first few principal components (PCs) as predictors. For our simulated data, there are approximately five non-random PCs; the appropriate number of PCs in a real data set could potentially be estimated from a scree plot of the amount of variance explained by each PC. The selected PCs could then be used as predictors in a Cox proportional hazards model. On the other hand, the same heatmap indicates the presence of six subtypes of cancer. An alternative approach would be to use those six subtypes as a categorical predictor; these could also be tested in a Cox model. In this case, the obvious next step would be to develop a robust multi-category classifier.

We do not pursue these approaches further in the

current paper. However, the `Umpire` package provides the tools that are necessary to evaluate a range of analytical methods on data sets with different sizes and properties. The availability of this tool should contribute to the development of better methods to learn useful predictors of biologically relevant outcomes.

Acknowledgment

This research was supported by grants P30 CA016672, R01 CA123252, P50 CA070907, and P50 CA140388 from the National Cancer Institute of the United States National Institutes of Health.

This document was prepared using Sweave, a literate programming tool for the R statistical software environment. Complete source code, including all code necessary to run the simulations and generate the figures and tables, is available upon request.

References

[1] J. Zhang and K. R. Coombes, "UMPIRE: Ultimate microarray prediction, inference, and reality engine," in *BIOTECHNO 2011, The Third International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies*, 2011, pp. 121–125.

[2] C. K. Wierling, M. Steinfath, T. Elge, S. Schulze-Kremer, P. Aanstad, M. Clark, H. Lehrach, and R. Herwig, "Simulation of DNA array hybridization experiments and evaluation of critical parameters during subsequent image and data analysis," *BMC Bioinformatics*, vol. 3, p. 29, 2002.

[3] Y. Balagurunathan, E. R. Dougherty, Y. Chen, M. L. Bittner, and J. M. Trent, "Simulation of cDNA microarrays via a parameterized random signal model," *J Biomed Opt*, vol. 7, no. 3, pp. 507–23, 2002.

[4] D. S. Lalush, "Characterization, modeling, and simulation of mouse microarray data," in *Methods of Microarray Data Analysis III*, S. M. Lin and K. F. Johnson, Eds. Boston: Kluwer Academic Publishers, 2003, pp. 75–92.

[5] M. Nykter, T. Aho, M. Ahdesmaki, P. Ruusuvuori, A. Lehmussola, and O. Yli-Harja, "Simulation of microarray data with realistic characteristics," *BMC Bioinformatics*, vol. 7, p. 349, 2006.

[6] C. J. Albers, R. C. Jansen, J. Kok, O. P. Kuipers, and S. A. van Hijum, "SIMAGE: simulation of DNA-microarray gene expression data," *BMC Bioinformatics*, vol. 7, p. 205, 2006.

[7] K. Dobbin and R. Simon, "Comparison of microarray designs for class comparison and class discovery," *Bioinformatics*, vol. 18, no. 11, pp. 1438–45, 2002.

[8] A. Szabo, K. Boucher, W. L. Carroll, L. B. Klebanov, A. D. Tsodikov, and A. Y. Yakovlev, "Variable selection and pattern recognition with gene expression data generated by the microarray technology," *Math Biosci*, vol. 176, no. 1, pp. 71–98, 2002.

[9] I. Lonnstedt and T. Speed, "Replicated microarray data," *Statistica Sinica*, vol. 12, pp. 31–46, 2002.

[10] M. S. Pepe, G. Longton, G. L. Anderson, and M. Schummer, "Selecting differentially expressed genes from microarray experiments," *Biometrics*, vol. 59, no. 1, pp. 133–42, 2003.

[11] P. de Valpine, H. M. Bitter, M. P. Brown, and J. Heller, "A simulation-approximation approach to sample size planning for high-dimensional classification studies," *Biostatistics*, vol. 10, no. 3, pp. 424–35, 2009.

[12] R. S. Parrish, H. J. Spencer III, and P. Xu, "Distribution modeling and simulation of gene expression data," *Computational Statistics and Data Analysis*, vol. 53, pp. 1650–1660, 2009.

[13] C. F. Aliferis, A. Statnikov, I. Tsamardinos, J. S. Schildcrout, B. E. Shepherd, and F. E. Harrell Jr., "Factors influencing the statistical power of complex data analysis protocols for molecular signature development from microarray data," *PLoS One*, vol. 4, no. 3, p. e4922, 2009.

[14] Y. Guo, A. Graber, R. N. McBurney, and R. Balasubramanian, "Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms," *BMC Bioinformatics*, vol. 11, p. 447, 2010.

[15] C. O. Nordling, "A new theory on cancer-inducing mechanism," *Br J Cancer*, vol. 7, no. 1, pp. 68–72, 1953.

[16] J. Knudson, A. G., "Mutation and cancer: statistical study of retinoblastoma," *Proc Natl Acad Sci U S A*, vol. 68, no. 4, pp. 820–3, 1971.

[17] D. Borsboom, G. Mellenbergh, and J. van Heerden, "The theoretical status of latent variables," *Psychological Review*, vol. 10, no. 2, pp. 203–219, 2003.

[18] N. Belacel, Q. Wang, and M. Cuperlovic-Culf, "Clustering methods for microarray gene expression data." *OMICS.*, vol. 10, no. 4, pp. 507–31, 2006.

[19] W. Kong, C. Vanderburg, H. Gunshin, J. Rogers, and X. Huang, "A review of independent component analysis application to microarray gene expression data." *Biotechniques*, vol. 45, no. 5, pp. 501–20, 2008.

[20] J. Chen, "Key aspects of analyzing microarray gene-expression data," *Pharmacogenomics*, vol. 8, no. 5, pp. 473–82, 2007.

[21] G. Hatfield, S. Hung, and P. Baldi, "Differential analysis of dna microarray gene expression data," *Mol Microbiol.*, vol. 47, no. 4, pp. 871–7, 2003.

[22] R. M. Simon, E. L. Korn, L. M. McShane, M. D. Radmacher, G. W. Wright, and Y. Zhao, *Design and Analysis of DNA Microarray Investigations*, ser. Statistics for Biology and Health. New York, NY: Springer-Verlag, 2003.

[23] S. Dudoit and M. van der Laan, *Multiple Testing Procedures with Applications to Genomics*, ser. Springer Series in Statistics. New York, NY: Springer, 2008.

[24] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nat Rev Genet*, vol. 7, no. 1, pp. 55–65, 2006.

[25] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–7, 1999.

[26] A. Dupuy and R. Simon, "Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting." *J Natl Cancer Inst.*, vol. 99, no. 2, pp. 147–57, 2007.

[27] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy." *Lancet*, vol. 365, no. 9458, pp. 488–92, 2005.

[28] J. Deutsch, "Evolutionary algorithms for finding optimal gene sets in microarray prediction." *Bioinformatics*, vol. 19, no. 1, pp. 45–52, 2003.

[29] R. Simon, M. Radmacher, K. Dobbin, and L. McShane, "Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification." *J Natl Cancer Inst.*, vol. 95, no. 1, pp. 14–8, 2003.

[30] Q. Ye, L. Qin, M. Forgues, P. He, J. Kim, A. Peng, R. Simon, Y. Li, A. Robles, Y. Chen, Z. Ma, Z. Wu, S. Ye, Y. Liu, Z. Tang, and X. Wang, "Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning." *Nat Med.*, vol. 9, no. 4, pp. 416–23, 2003.

[31] D. Danila, K. Pantel, M. Fleisher, and H. Scher, "Circulating tumors cells as biomarkers: progress toward biomarker qualification," *Cancer J.*, vol. 17, no. 6, pp. 438–50, 2011.

[32] T. Sigdel and M. Sarwal, "Recent advances in biomarker discovery in solid organ transplant by proteomics." *Expert Rev Proteomics.*, vol. 8, no. 6, pp. 705–15, 2011.

[33] J. Kalinina, J. Peng, J. Ritchie, and E. Van Meir, "Proteomics of gliomas: initial biomarker discovery and evolution of technology." *Neuro Oncol.*, vol. 13, no. 9, pp. 926–42, 2011.

[34] P. Rakowska and M. Ryadnov, "Nano-enabled biomarker discovery and detection." *Biomark Med.*, vol. 5, no. 3, pp. 387–96, 2011.

[35] J. Ross, "Biomarker-based selection of therapy for colorectal cancer." *Biomark Med.*, vol. 5, no. 3, pp. 319–32, 2011.

[36] S. Dupouy, N. Mourra, V. Doan, A. Gompel, M. Alifano, and P. Forgez, "The potential use of the neurotensin high affinity receptor 1 as a biomarker for cancer progression and as a component of personalized medicine in selective cancers." *Biochimie.*, vol. 93, no. 9, pp. 1369–78, 2011.

[37] E. Galanis, W. Wu, J. Sarkaria, S. Chang, H. Colman, D. Sargent, and D. Reardon, "Incorporation of biomarker assessment in novel clinical trial designs: personalizing brain tumor treatments." *Curr Oncol Rep.*, vol. 13, no. 1, pp. 42–9, 2011.

[38] J. Ross, E. Slodkowska, W. Symmans, L. Pusztai, P. Ravdin, and G. Hortobagyi, "The HER-2 receptor and breast cancer: ten years of targeted anti-HER-2 therapy and personalized medicine." *Oncologist*, vol. 14, pp. 320–368, 2009.

[39] R. Fisher, *The Design of Experiments*. Macmillan, 1971.

[40] K. K. Dobbin and R. M. Simon, "Sample size planning for developing classifiers using high-dimensional DNA microarray data," *Biostatistics*, vol. 8, no. 1, pp. 101–17, 2007.

[41] K. K. Dobbin, Y. Zhao, and R. M. Simon, "How large a training set is needed to develop a classifier for microarray data?" *Clin Cancer Res*, vol. 14, no. 1, pp. 108–14, 2008.

[42] J. Kim, D. Shin, and Y. Lee, "Effect of local background intensities in the normalization of cDNA microarray data with a skewed expression profiles." *Exp Mol Med.*, vol. 34, no. 3, pp. 224–32, 2002.

[43] D. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society*, vol. 34, no. 2, pp. 187–220, 1972.

[44] S. Pounds and S. Morris, "Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values." *Bioinformatics*, vol. 19, pp. 1236–42, 2003.

[45] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press*, 1967, pp. 281–297.

[46] M. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine." *Clin Chem.*, vol. 39, no. 4, pp. 561–77, 1993.

[47] L. Chin, W. Hahn, G. Getz, and M. Meyerson, "Making sense of cancer genomic data," *Genes and Development*, vol. 25, pp. 534–555, 2011.

Facilitating Bioinformatics Research through a Mobile Cloud with Trusted Data Provenance

Jinhui Yao[1,2], Jingyu Zhang[2], Shiping Chen[1], Chen Wang[1], David Levy[2], Qing Liu[3]

[1]*Information Engineering Laboratory, CSIRO ICT Centre, Australia*
{Firstname.Lastname}@csiro.au
[2]*School of Electrical and Information Engineeering, University of Sydney*
{jinhui, jyzhang, dlevy}@ee.usyd.edu.au
[3]*CSIRO Plant Industry, Canberra, Australia*
q.liu@csiro.au

***Abstract*—Cloud provides a cheap yet reliable outsourcing model for anyone who needs scalable computing resources. Together with the Cloud, Service Oriented Architecture (SOA) allows the construction of scientific workflows to bring together various scientific computing tools offered as services in the Cloud, to answer complex research questions. In those scientific workflows, certain critical steps need the participation of research personnel or experts. It is highly desirable that scientists have easy access, such as mobile devices, to the workflows running in the Cloud. Furthermore, since the participants in this cross-domain collaboration barely trust each other, achieving reliable data provenance becomes a challenging task. In this paper, we propose a concept of mobile-cloud by combining mobile and cloud together in a bioinformatics research application scenario. A mobile-cloud framework is developed, which facilitates the use of mobile devices to manipulate and interact with the scientific workflows running in the Cloud. The Mobile Cloud system acts as a trusted third party to record provenance data submitted by the participating services during the workflow execution. We have implemented a prototype which allows the bioinformatics workflow design and participation using mobile devices. We prove the concept of mobile-cloud with the prototype and conducted performance evaluation for the significant points of the bioinformatics workflow.**

Keywords -Cloud Computing,Accountability,Service Oriented Architecture,Mobile Cloud,Data Provenance

I. INTRODUCTION

The emergence of computing resource provisioning known as the Cloud has revolutionized classical computing. It provides a cheap yet reliable outsourcing model for anyone who needs scalable computing resources. Given the fact that many scientific breakthroughs need to be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets [2], Cloud computing offers the promise of "democratizing" research, as a single researcher or small team can have access to the same large-scale computing resources as well-funded research organizations without the need to invest in purchasing or hosting their own physical IT infrastructures.

On the other hand, the concept of Service Oriented Architecture (SOA) allows flexible and dynamic collaborations among different service providers. A service can either directly be used for its mere functions or be composed with other services to form new value-added workflows [3]. Through SOA, scientific workflows can be used to bring together various scientific computing tools and resources offered as services in the Cloud to answer complex research questions. Workflows describe the relationship of individual computational components and their input and output data in a declarative way. In astronomy, scientists are using workflows to generate science-grade mosaics of the sky [4], to examine the structure of galaxies [5]. In bioinformatics, researchers are using workflows to understand the underpinnings of complex diseases [6].

In scientific workflows, certain critical steps need the participation of respective research personnel or experts. For example, how the workflow should be designed and which scientific tools need to be involved must be decided by the experts in the area. And some complex patterns generated from the experiments need to be visually inspected by the scientists, who will determine the next a few steps for further analysis. In this regard, it is highly desirable that scientists can have an easy access to the services in the Cloud so that they can design and participate in the workflows efficiently.

Furthermore, data provenance has been widely acknowledged as an important issue for scientific experiments [28-30], for the provenance data collected during the experiment can be used to understand, reproduce the experiments conducted; identify the way data are derived. However, within this service-oriented collaboration, each service provider or individual researcher is from different organizations. The cross-domain collaboration intuitively suggests that the participants should be unnecessary to fully trust each other even though they need to collaborate. This implies they will question each other, e.g., 1) if a particular participant has employed proper data provenance mechanisms during the experiment; 2) if the recorded provenance data have been or will be tampered with; and 3) if the issuer and the integrity of the provenance data are somehow verifiable. These doubts caused by the lack of trustworthiness makes achieving reliable data provenance a challenging task, hence reduce the incentives of individuals to participant in such cross-domain collaborative scientific workflow and are harmful for the wide adaptation of this computing paradigm. Therefore, a means to

record the provenance data during the experiments in a trustworthy way is needed.

To address the above needs, with the impressive advances in the technology, we believe using mobile devices can be an ideal solution. The processes in a workflow can be thoroughly integrated with portable devices. All activities are decided and monitored on time from the way that fit the human environment instead of forcing users to passively accept the computing results from cloud service.

In this paper, by extending our previous work [1] we propose a novel design which facilitates the use of mobile devices to manipulate and interact with the scientific workflows running in the Cloud. In our system, the users can choose the services in the Cloud to form the workflows via their mobile devices, and each mobile device can serve as one service node to be involved in the workflows designed. A significant aspect of the Mobile Cloud is its ability to provide trusted data provenance. Mobile Cloud serves as a trusted third party to record provenance data submitted by the participating services during the workflow execution. By enforcing strong accountability via the use of cryptographic techniques, the provenance data submitted by the participants in the workflow are undeniably linked to the submitter, which means its issuer and integrity can be cryptographically verified. With these verifiable provenance data recorded by a trusted platform, the collaborating entities can have a much better sense of trust in the validity of the provenance information they need to use.

The main contributions of this article are: 1) we design a Mobile Cloud system as a middleware layer to facilitate the use of mobile devices to design and interact with the scientific workflows running in the Cloud; 2) we define and illustrate the concept of strong accountability and the way it can be applied to record activity traces with provability; 3) we propose a novel approach to obtain activity traces from the execution of workflows and use them to construct data provenance graph to illustrate provenance information; and 4) we evaluate the performance of the Mobile Cloud system in the Cloud with real services.

II. THE APPLICATION SCENARIO

In the area of gene research, the recent development of the microarray technology [7] have led to rapid increase in the variety of available data and analytical tools. Some recent surveys published in Nucleic Acids Research describes 1037 databases [8] and over 1200 tools [9]. The analysis of microarray data commonly requires the biologist to query various online databases and perform a set of analysis using both local and online tools.

To illustrate with an example, here we explain the research study of the genetic cause of colorectal cancer, i.e., identify the genetic variation in human DNA that makes people susceptible to colorectal cancer. The rat azoxymethane (AOM) model of CRC is often used in dietary intervention studies as it induces mutations in genes which are also found to be mutated in human adenomas and adenocarcinomas. To define the baseline variation in global gene expression, the biologists extract RNA from mucosa scraped from colon and analyze the global gene expression using the Affymetrix Gene Chip. Data is normalized and then analyzed for differential expression.

By contrasting the results from normal and cancer mice, biologists can identify candidate genes through statistical analysis. Further analysis—such as searching for the functions known to these genes — are commonly performed to examine whether and how the candidate genes relate to the colorectal cancer. The followings are the data acquisition and analysis steps to perform the study of microarray experiment:

Quality Control. The raw microarray result data are processed, visualized and inspected by an expert, who can identify errors and discard the experiment.

Normalization. Microarray results from different samples need to be normalized before any meaningful comparison can be conducted.

Gene Differentiation. By contrasting the results from cancerous and healthy tissues, differentially expressed genes—candidate genes that are active in cancer are identified by applying some statistical methods (e.g. LIMMA).

Gene Study. Most differentially expressed genes are further studied to understand the biological functions of the disease. There are various resources available for study. For example, gene symbols and descriptions could be retrieved from the Rat Genome Database and/or BioMart. Gene Ontology (GO) and KEGG databases could provide gene functions and molecular pathways information respectively. Experts need to be involved to make good decision as which study to conduct and which database to use.

We can see that the four standard analysis procedures we listed above not only can be extremely computing intensive but also require some decision making from the research scientists or experts at certain critical steps (e.g. quality control). It easily follows that, a viable approach to conduct such researches must utilize certain computing platform that has enormous computing capacity, yet research scientists can easily interact with the platform and the computing process conducted. This is essentially the reason for which we promote the "Mobile Cloud" - a composition of the Cloud, and the mobile devices – to be a suitable paradigm for complicated bioinfomatics researches.

III. A MOBILE-CLOUD SYSTEM FOR BIOINFORMATICS REARCH

As we have established in previous sections, we propose to compose the Cloud and the mobile devices to conduct complex bioinformatics researches. The bioinformatics research scenario we chose is the study for the cause of colorectal cancer. Figure 1 shows our proposed system with this research scenario.

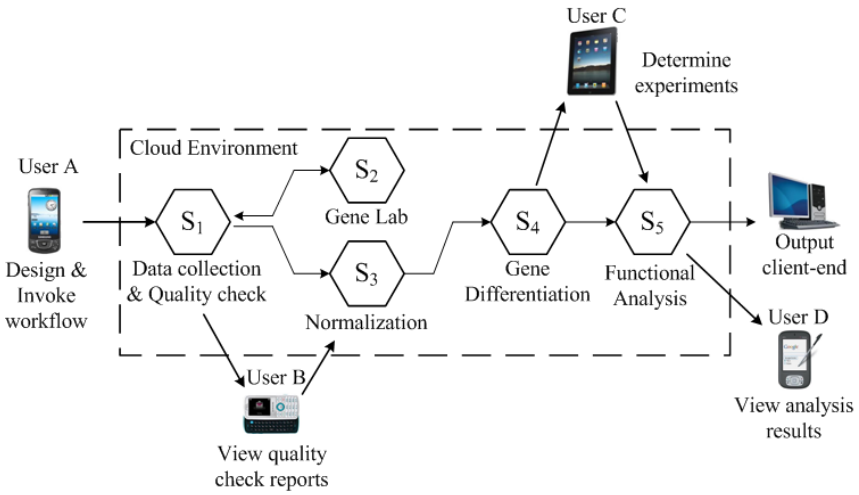


Figure 1. Overview of the proposed Moble-Cloud system

In the Cloud, different computing intensive gene research tools are deployed by different research bodies and provided as services. Outside the Cloud, research scientists or gene analysts locate the desired services in the Cloud, and use them to compose a workflow for studying the cancer. In a gene research lab, we assume the gene data in the subject microarray chips are scanned and archived in some digital database, which can be reached from the Cloud or itself could be a Cloud storage service [31] such as Amazon S3. The Mobile Cloud operates as this: a researcher (user A) designs the scientific workflow and composes the needed services in the Cloud, then he invokes the first service – "Data collection and Quality Check", which retrieves the gene data from the nominated "Gene Lab" where the gene subjects are stored, then conducts quality checks on the gene data. Once finished, the data is sent to the next service – "Normalization" and a quality report is sent to user B for confirmation. If user B confirms the data quality, the normalization service will normalize the data and send the results to "Gene Differentiation". Another report is sent to user C After the differentiation, to choose the suitable experiment for the functional analysis. When the workflow is complete, the results are sent to a client end and a final report is sent to user D. We can see in the workflow multiple research scientists are involved. They participate in the workflow by using portable or desktop devices to invoke or receive output from the services.

Our argument for using mobile devices to design and participate in the workflows is intuitive. As mentioned, in the workflow there are "critical steps" that require decision making by experts in the respective area, in order to continue the process. For example, after the quality check, an important decision needs to be made about whether the quality of the raw data suffices the requirements of the experiment. The experiment should be paused before the expert in charge has reviewed the quality check reports and confirmed the usability of the raw data. Therefore, mobile devices are indeed ideal for this task for its outstanding mobility compared to desktop computers or even laptop computers, i.e. one can freely use his mobile devices while waiting in a queue, on a bus, or even walking. Further, given the recent impressive advances in the mobile technology, the computing capability of mobile devices - however limited compared to desktops or laptops - is more than enough to run basic UI or display data sets and processing reports. Therefore, we believe mobile devices such as smart phones or tablet computers are indeed ideal to be used as light client-end to drive the heavy bioinformatics research workflows in the Cloud.

A. Overall architecture of Mobile Cloud

To enable mobile devices to construct and participate in the workflows running the Cloud, we have developed the Mobile Cloud middleware layer (MC-layer) to facilitate these. This middleware shall be deployed or even provided by the cloud environment provider in that environment to facilitate efficient interactions with the services and the clients. Figure 2 provides an overview of the architecture, which consists of a user interface (residing on mobile devices), a Cloud environment containing various services and a middleware layer consists of three function units. Their respective functionalities are summarized as follows:

- **Cloud Environment** provides various services deployed by respective providers. The services have registered their access end point with the MC-layer.
- **Service Repository/Composition** stores the information about the services in the Cloud that have registered with it. It helps the user to search for the services that best satisfy the requirements specified, and compose them into workflows.
- **Workflow Execution** conducts two jobs: (a) orchestrating workflows during the operation; (b) invoking Web services according to the workflow defined.
- **Trusted Data Provenance Unit** (TPU) records cryptographically signed provenance data submitted by the participating services during the execution of the workflows. Using the recorded data, it monitors the status of the execution and allows the clients to query data

provenance traces (this part will be elaborated in detail in section 4).

- **User Interface** allows users to register, design workflows and participate in a running workflow.

For mobile devices to construct workflows, they first need to send a search request to the Service Repository in order to get a list of the services/workflows they are looking for. A convenient UI has been implemented on the mobile devices to allow the users to easily design the workflows using the services listed by the Service Repository (the UI will be elaborated in the evaluations). Once the workflow have been designed, a representative XML based description script is generated to be submitted to the Service Composition unit. The Service Composition unit thus according to the script, composes the services to form the desired workflows. The services can be composed in two ways: i) centrally composed, where the MC-layer invokes the services in the sequence designed by the user; and ii) remotely orchestrated, where certain orchestration scripts such as BPEL will be generated and distributed to all the services involved for deployment.

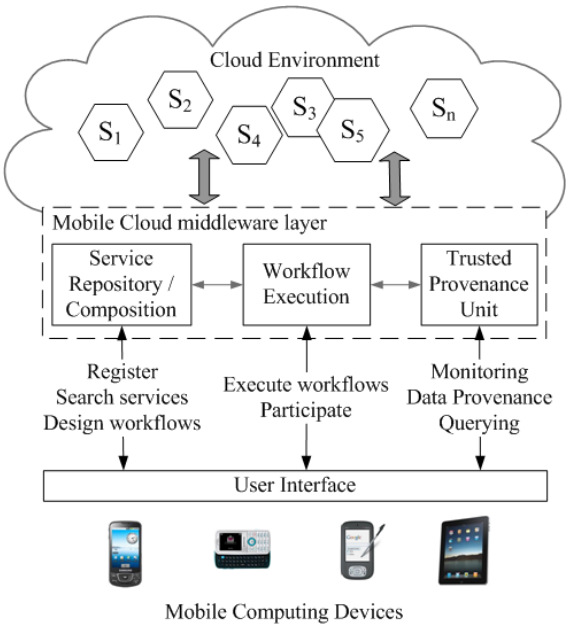


Figure 2. Overview of Mobile-Cloud architecture

B. *Workflow design through abstract description script*

In our system, the workflow designed by the users is an abstract workflow, that is, the users only need to specify the type of service needed, and the MC-layer will search its service reporsitory and select the best suited ones according to the user's specifications. Table 1 gives a sample of the workflow description script. As it is developed based on the BPEL, "sequences" and "flows" are used to specify serial and parallel composition, and "Actions" are used define the invocation operations. The sample describes the first half of the gene analysis workflow in Figure 1. In some actions, the endpoint is set to be "OPTIMAL". This is to tell the Service Composition unit to choose the best suited services.

TABLE I. SAMPLE WORKFLOW DESCRIPTION SCRIPT

```
<sequence name="main">
   <Action operation="start" invoker="client"
   endpoint="QualityCheck" type="send&forget".../>
   <Action operation="fetchGene" invoker="QualityCheck"
   endpoint="GeneLab" type ="send&receive".../>
   <flow>
      <Action operation="sendForApproval"
      invoker="QualityCheck" endpoint="user B" type
      ="send&forget".../>
      <Action operation="normalization"
      invoker="QualityCheck" endpoint="OPTIMAL" type
      ="send&forget".../>
   </flow>
   …
</sequence>
```

As we have established in our system design, mobile devices will be involved in the workflows as web services. To facilitate this, we created a customized web service engine to run on the mobile devices. Using this engine, mobile devices can both send and receive service requests, as well as interpreting the workflow description scripts delivered by the MC-layer. Once a user has designed and submitted a workflow, the workflow description script will be forwarded to the research personnel that are involved. The mobile devices they are using will interpret the workflow script and save the workflow logic. When a service request is received during the execution of the workflow, the UI will allow the user to view the content (e.g. quality check reports) and provide the list of the services that the user should send output request to according to the workflow logic (e.g. normalization services). For the technical details of the MC-layer, please refer to our previous publications about the Web Service Management System (WSMS) [13].

IV. ACCOUNTABILITY FOR COMPLIANCE AND PROVENANCE

The workflows in the Cloud are constructed using services provided by different parties who barely know each other. The correctness of the resultant workflow relies on the individual correctness of all participators. That is, if the service is compliant to the pre-defined workflow logic, or Service Level Agreement (SLA). The scientific integrity of the gene analysis results will be highly questionable if the services involved can act willy-nilly and get away with processing errors.

On the other hand, for scientific experiments not only the resultant data are considered, the steps of how these data are derived along the process can also be very valuable. It has been widely realized that data provenance plays an important role in the scientific researches [14]. It follows that, trusted data provenance mechanisms are necessary in such systems with participants from different administrative domains. Provenance data should be preserved in a trustworthy way that, the contributors of the data are committed to their truthfulness. This naturally leads us to the issue of accountability. In this

section, we illustrate our design to incorporate strong accountability into the "Mobile Cloud" to address these issues.

A. Accountability for trustworthiness

Accountability can be interpreted as the ability to have an entity account for its behaviors to some authorities [10]. This is achieved by binding each activity conducted to the identity of its actor with proper evidence [11]. Such binding should be achieved under the circumstance that all actors within the system are semi-trusted. That is, each identified actor may lie according to their own interest. Therefore, accountability should entail a certain level of stringency in order to maintain a system's trustworthiness. Below, we identify several desirable properties of a fully accountable system:

- **Verifiable:** The correctness of the conducted process can be verified according to the actions and their bindings recorded.
- **Non-repudiable:** Actions are bound to the actors through evidence, and this binding is provable and undeniable.
- **Tamper-evident:** Any attempt to corrupt to recorded evidence inevitably involves the high risk of being detected.

We illustrate our proposed approach in Figure 3. In our approach, accountability can be incorporated into activity-based workflow by requiring the entity conducting the process to log non-disputable evidence about the activities in a separate entity. In the figure, after incorporating accountability into an ordinary process, entity A is now required to perform logging operations before and after conducting the activity in its process. The evidence is logged in a separate entity - entity B - so that entity A cannot access the logged evidence. The evidence needed to be logged should contain enough information to describe the conducting activity. In our simple example, which is intuitive enough, the evidence should include the states of the factors concerning the start of the activity (e.g. the input variables) and the factors concerning its completion (e.g. the output value).

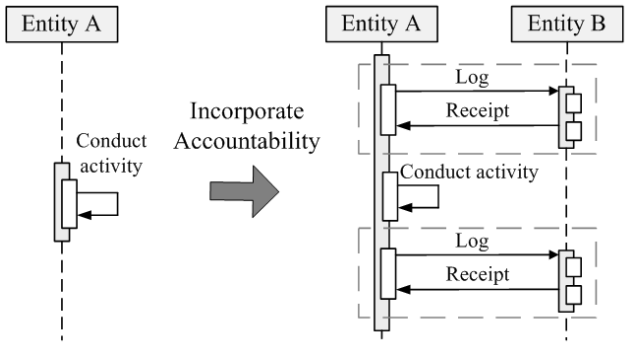


Figure 3. Example of incorporating accountability into processes

The logging operations require the employment of PKI in all involved service entities. Each of them has its own associated public-private key pair issued by certificated authorities. The logging operations are as follows:

1. The logger (entity A) signs the evidence (E) by its private key (K_{A-}) to create a digital signature of the evidence (S_A).
2. The evidence and its signature are then logged in a separate entity (entity B).
3. When received, entity B creates a receipt by signing entity A's signature with entity B's private key (K_{B-}).
4. Lastly, the receipt (S_B) is sent back to the logger (entity A) in the reply.

Assuming the digital signature is un-forgeable, the signed evidence in entity B can be used to verify entity A's compliance; and yet any corruption or deletion applied to the evidence will be discovered using the receipt received by entity A. Under the circumstance that neither of the service entities is trusted; and assume they will not conspire to cheat, this structure manages to ensure the proper preservation of evidence associated with the process conducted.

B. Logging provenance data a Trusted Provenance Unit

In Mobile Cloud, the *Trusted Provenance Unit* (TPU) acts as the separate entity B, dedicated to provide accountability to all underlying services involved in the workflow. Figure 4 shows the structure. All the mobile devices, service nodes in the Cloud as well as local computing nodes that are involved in the workflow, register with TPU and submit provenance data during the execution of the workflow.

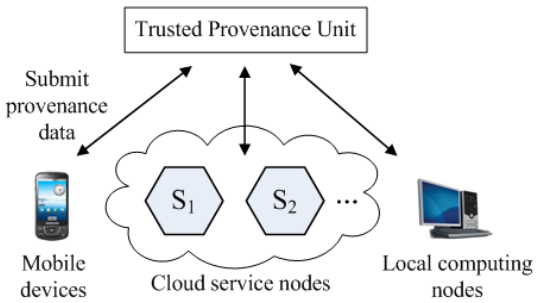


Figure 4. TPU records provenance data from various sources

The provenance data can be recorded in various ways, for instance, if the service invocations are all relayed by the MC-layer, they can be simply archived when received. Here we illustrate a generic approach to incorporate the data logging into the workflows by transforming the workflow descriptive scripts. Business process or workflows are often defined through process descriptive languages, which will be interpreted by orchestration engines (e.g. Apache ODE) to conduct the process accordingly. A good example of the process descriptive language is Business Process Execution Language (BPEL) [34]. BPEL models the business activities into several basic activity types, and then composes those types to describe the whole process. The core activity types include:

1. *Receive*, receiving the request from a requestor. This activity type will specify the variable to which the input data is to be assigned.

2. *Invoke*, invocation to an endpoint (service). Invoke activity type will specify the variable used as the input and the variable used to store the output data for this invocation.

3. *Reply*, replying the invocation. A variable will be specified to be returned to the requestor as the result.

To add logging activities into the workflow, we can insert *invoke* activity types into the BPEL script to invoke a certain endpoint (e.g. logging service) with the provenance data to be logged. And due to the distinct natures of *receive*, *invoke* and *reply* activity types, the rules used to decide the insertion locations are in fact quite straightforward. For the *receive* activity, an *invoke* should be inserted right after it, to log the input data received. For the *invoke* activity, one *invoke* should be inserted before this activity and another to be inserted after, to log the input data and the reply data of the invocation respectively. And finally for the *reply* activity, an *invoke* needs to be inserted just before it to log the result data that is about to be returned to the requester. The invocation endpoint for the *invoke* activities inserted (i.e. logging service) should either be a service in the same domain of the logger, or a trusted party nominated by the logger, which in turn signs the evidence on the logger's behalf and forward the signed evidence to the TPU.

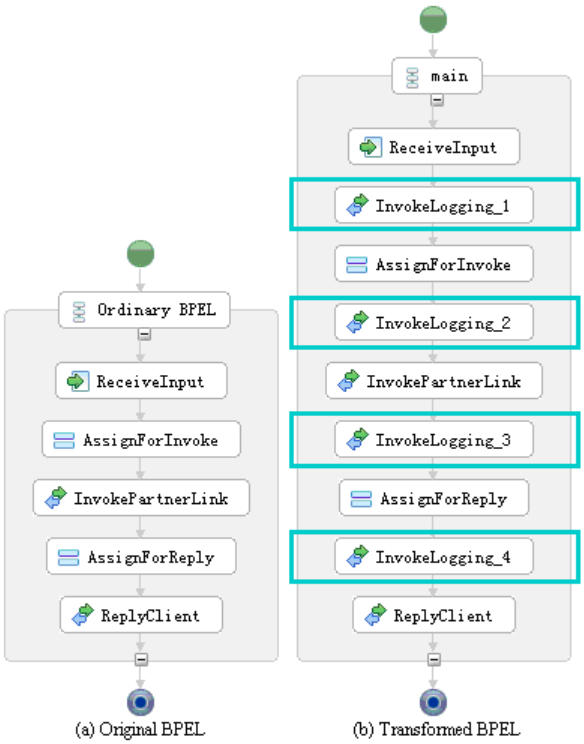


Figure 5. Transformation of BPEL

To further illustrate this transformation process, we have presented an example in Figure 5. Figure 5(a) shows the graphical view of an ordinary sample BPEL. This simple process is started by receiving an input (ReceiveInput); then a partner link (collaborating service) is invoked in turn (InvokePartnerLink), and finally, replies the result to the client (ReplyClient). Figure 5(b) is the BPEL after the transformation. We can see in Figure 5b that four logging invoke activities (the InvokeLogging) have been inserted, one after the "ReceiveInput"; one before and one after "InvokePartnerLink"; and one before "ReplyClient". Because BPEL is entirely based on xml schema, any xml schema parser will be capable of analyzing and inserting activities into it. The implementation details of the incorporation of accountability have been elaborated in our previous work [12].

Here the evidence can be any intermediate gene analysis data generated by the tools in the Cloud, or the decisions made by research personnel participated. With the evidence data logged, the core functionalities provided by the TPU are:

- Compliance verification. Through the analysis of the evidence data, the correctness of the behaviors of the underlying services is continuously validated.
- Data provenance. The evidence recorded capture the evolution path of the data as well as the entities responsible for each step.
- Workflow status monitoring. A global view over the workflow is maintained by the TPU. Such information can be used to assist the functioning of the MC-layer and the underlying services.

C. *Architectural design of Trusted Provenance Unit*

TPU is responsible of recording the provenance data from all the participants of the workflow. As accountability requires the submitter of the data to sign the data before submission to commit its truthfulness, services and the entities involved in the workflow are needed to register their identity documents (e.g. X.509) at MC-layer. When a new abstract workflow is proposed by a researcher, the Service Repository/Composition unit first find the services that best suit the specified requirements, then the filled workflow script is transformed by TPU to have logging activities (refer to [12] for details). Meanwhile, TPU uses the knowledge obtained from the documents registered to generate analysis logics to process the incoming data during the execution of the workflow. The resultant data provenance information will be delivered to the user through querying and visual displays.

The internal architectural design of TPU is shown in Figure 6. In the initialization phase, registered information about the services, like WSDL, X.509 certificate etc.; and information about the workflows, like BPEL scripts are transmitted to TPU from Service Repository/Composition unit. TPU first transform the workflow script to incorporate logging activities and send the transformed script for redeployment; then it uses the registration information received to generate two components: "Monitoring Logic" and "Provenance Logic". In the monitoring phase, the provenance data will be submitted

from the participants in the workflow. These data will first be analysed by the monitoring logic to find obvious compliance violations (e.g. QoS service level agreement); then be processed by the provenance logic to generate data provenance information to be stored in the data warehouse.

When the provenance data are received, the provenance logic first labels the provenance data with information regarding the "four Ws": who, when, where and what. In general, the provenance logic will add labels explaining what this provenance data is about, in which workflow (where) it is generated, at what time and by which participant (who). Then, based on the knowledge obtained from the documentation registered, the provenance logic links the different provenance data with Open Provenance Model [32] edges to form a provenance graph. An example of such graph is displayed in Figure 7. We can see from the example, the provenance information of the data pieces (circles marked with numbers) are expressed in terms of their links to the activities (round rectangles) that used or/and generate them. The figure is a visual display of the provenance graph, it is not necessarily an actual graph when stored in the data warehouse. The provenance logic simply needs to label the data so they are linked with each other.

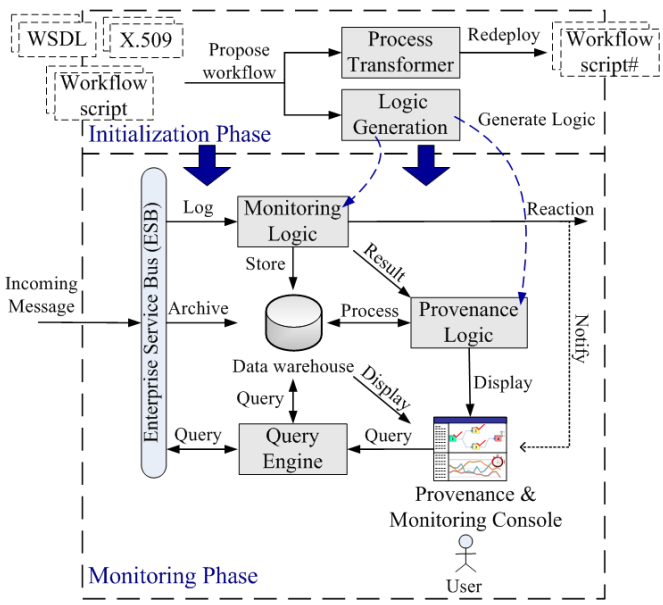


Figure 6. Internal architecture of Trusted Provenance Unit

The query engine provides an interface for the users to fetch the provenance information about specific data. In order to enable simple and efficient querying, a query language in XML is developed, called SWQL (Simple Workflow Query Language). SWQL allows the user to specify the information regarding the "four Ws" to fetch the desired provenance data. An example is shown in Listing 2. The example is a query to fetch all the differentiated gene (what) recorded from the colorectal cancer workflow (where), submitted by service A and B (who) from 9am to 5pm on 20 July 2011 (when).

The provenance and monitoring console is a graphical user interface to display provenance and monitoring information as well as let users query the data warehouse. During the execution of the workflow, the evolution of the data will be displayed in terms of the provenance graph generated by the provenance logic, and the status of the workflow will been shown. More details about the console will be discussed in the evaluation section.

TABLE II. AN EXAMPLE OF SWQL QUERY

```
<SWQL>
    <Action>Find</Action>
    <DataIdentifier>
     <Type>Differentiated gene</Type>
    </DataIdentifier>
    <EntityIdentifier>
     <Entity>Differentiation service A</Entity>
     <Entity>Differentiation service B</Entity>
    </EntityIdentifier>
    <TimeInterval>
     <From>9AM-20JUL2011</From>
          <To>5PM-20JUL2011</To>
    </TimeInterval>
    <WorkflowIdentifier>Colorectal cancer
    </WorkflowIdentifier>
     ...
</SWQL>
```

V. EVALUATION

We developed a prototype system to showcase our mobile-cloud concept. Our system consists of three parts: i) a client UI deployed in the mobile device; ii) an MC-layer for composing workflows and provenance; and iii) a number of demonstrating service nodes in Amazon EC2. We implemented five services nodes in EC2 to represent the gene research tools provided by different organizations. The services are linearly composed (one node finishes its job then invokes the next) to form a workflow using BPEL. The information about the services as well as the workflow are registered at the MC-layer, which is deployed in another computing instance in EC2. A remote user designs and invokes the workflow using the client UI locally deployed in the mobile device. With this setting, in this section, we will elaborate the implementation of the client UI; examine the communication overhead introduced when provenance data are logged at TPU during the execution; and we show some processing latency when a real gene database (KEGG) is involved in a workflow.

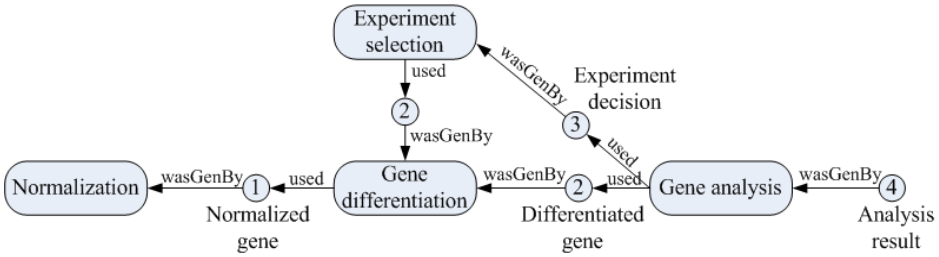
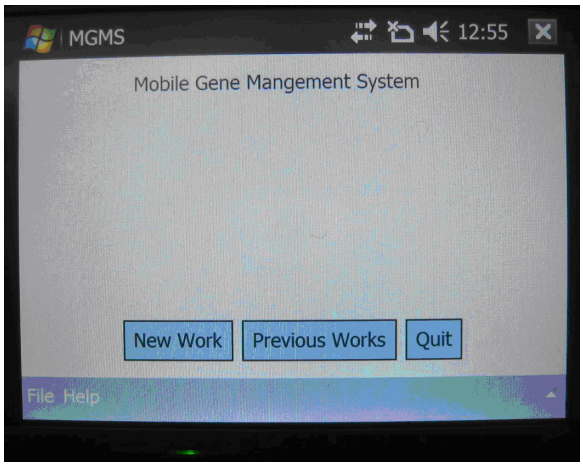


Figure 7. An example of provenance graph

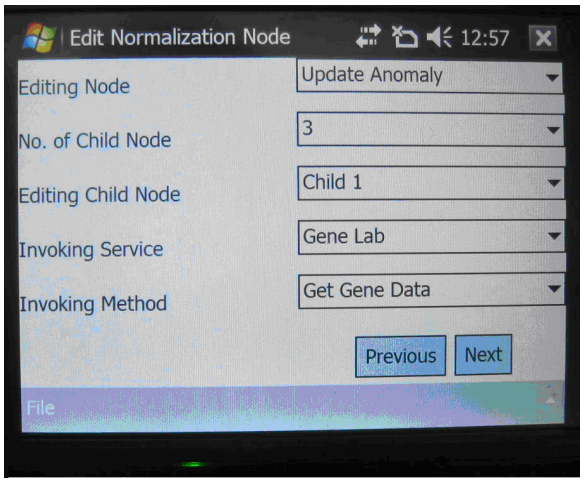
The UI on mobile device is developed using Java platform, micro edition (J2ME). The mobile web service feature is deployed and runs on a HTC 9500 mobile phone, which is running on IBM WebSphere Everyplace Micro Environment that supports a connected device configuration (CDC1.1). Figure 8 (a) and (b) show two screen shots of the Mobile Gene Management System (MGMS) - a scientific workflows design and surveillance tools. A user can define or edit a scientific process from the "New Work" button or "Previous Work" button as shown in Figure 8 (a). Then, the user can select into process items and specify their detail information as shown in Figure 8 (b). System users define the steps from four aspects, what services carry out these tasks; the number of child nodes; which methods/services are invoked; and what are the inputs and outputs of each step. Finally, an abstract workflow in BPEL will be generated and uploaded to the WSMS in Cloud, which will instantiate the abstract workflow by filling up the endpoints in the BPEL with the best concrete services URLs.

We have conducted testing to evaluate the latency introduced by incorporating the logging actions into the workflow. Figure 9(a) shows the overall latency to finish the process with untransformed BPEL scripts and with transformed ones. We have tested the workflow with request message size from 0.1KB (equivalent to a sentence) to 50KB (equivalent to a medium size document). For the process with transformed BPEL scripts to log the entire input/output messages (the series marked with "circles"), the latency introduced compared to the untransformed one (the series marked with "squares") grows as the request message becomes larger. In percentage terms, on average we observed a 30% increase in the overall process latency. Intuitively, this latency is significant to the business process; however it can be improved through the use of hash functions.

The hash of the message computed using collision-resistant hash functions (e.g., SHA-1), which is a very small digest (160 bits for SHA-1), can be logged as a substitute. Because the hashes computed are collision-resistant, which means it is theoretically impossible to have two different items with the same hash, so the hash can be logged to represent the data. We can see in the graph, the extra latency is significantly reduced if the BPEL scripts are transformed only to log the hash of the evidence (the series marked with "triangles"). In fact, since the size of the hash is fixed regardless of the data size, the extra latency almost remains constant regardless of the size of the request message. The overhead introduced will

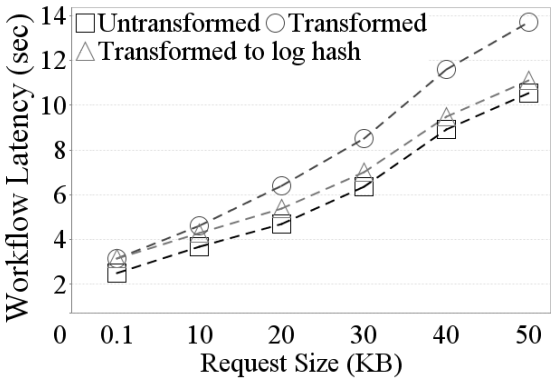


(a.) Main menu

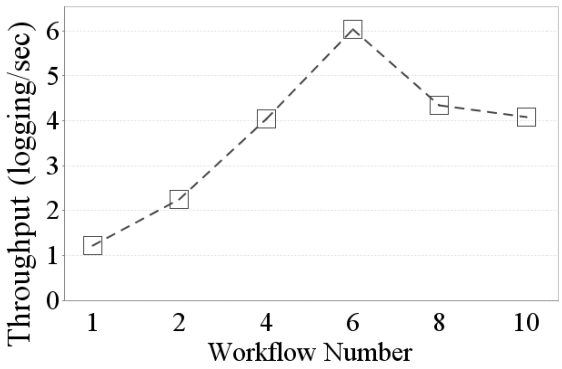


(b.) Designing a workflow

Figure 8. Screen shots of Mobile Cloud client end UI



(a.) Overall execution latency of the workflow



(b). Throughput of TPU under different loads

Figure 9. Performance evaluation

become more and more negligible when the size of the messages transferred increases. In practice, it is not often that the provenance data is urgently needed to be logged at runtime. When the system is idle, the provenance data can be eventually logged and verified according to the hash values. This eventual-logged strategy can further improve the performance and reduce the overhead.

As the MC-layer will be managing a number of workflows, naturally, it is interesting to find out the processing capability of the TPU. To evaluate this, we replicated the workflow we have implemented (the colorectal cancer workflow), and execute multiple workflows replicated concurrently. As such, multiple service nodes will be submitting provenance data to the TPU deployed in a computing instance simultaneously. With this setting, we evaluate the processing throughput of the TPU when it is under different loads (in terms of logging received per unit time). Figure 9(b) shows the testing results. In the figure we can see that, the processing throughput of TPU improves as the number of workflows increments, it reaches its peak when TPU is monitoring 6 workflows, and then it decays gradually if more workflows are involved in the monitoring. We tested this with messages of size 50KB, the processing operations conducted by AS involves both SLA monitoring and provenance data processing, which may need to fetch history data from the data warehouse to make conclusions. Since the computing power of a computing instance is fixed, an decrease in message size or processing complexity will shift the peak towards right to occur when more workflows are involved, and vice versa.

To evaluate the performance of gene retrieving from gene bank services, we selected 6 example genes which are the genetic causes of colorectal cancer and retrieve their genetic neighbors from KEGG disease Database [22]. We test the response time from 0 neighbors to 50 neighbors. As shown in Figure 10, it is clear that the latency is slowly increasing while we are increasing the number of neighbors. The has-581 continually yields the best performance at all stages from the 1427msec for retrieving 0 gene neighbor to 2746.8msec for getting 50 neighbors. However, has-10297 spent 2078msec to search 0 neighbors and it cost 2912.6msec for finding 50 neighbors.

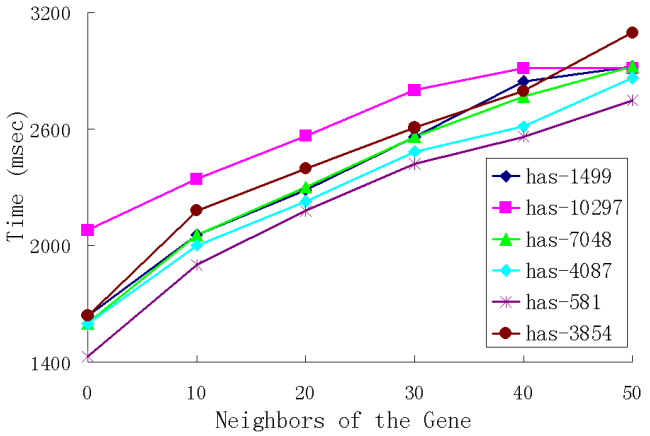


Figure 10. Gene retrieval experiment with KEGG

VI. Related Work

Mobile computing provides a luggable computation model for users. Its portability makes it very ideal for many application scenarios. To extend its limited computing power, research communities have proposed novel designs to leverage the Cloud. [23] proposed a virtual cloud system, and [24] detailed a distributed computing platform using mobile phones. They improve the capacities of mobile phones in the purpose of storage and computation. In the literature, [25-27] presented some computation offloading studies that move some parts of the applications to run on the Cloud. Executing parts of application remotely can save battery lifetimes and significantly extend computing resources. However, these

solutions do not support platform-independent cooperative interaction over an open network. In addition, after moving some parts of applications from stand-alone handheld devices to the cloud, several issues need to be considered in advance such as privacy, trustworthy or provenance.

The importance of provenance for scientific workflows has been widely acknowledged by various research communities. Many approaches have been proposed to record the derivations of the data during the scientific process. Approaches like [15][16] allow the designer to capture the intermediate data forms generated by the experiments at different granularities. In our work, we introduced the concept of accountability which not only provides data provenance but can enforce compliance among the service providers. Compliance assurance has been studied decently in recent years, some remarkable works include [18-21]. Our work differs from them at the point that we consider a more hostile environment where all service entities are expected to behave in any possible manner and deceive for their own benefit. Cryptographic techniques are deployed in our system to ensure the evidence (provenance data) are undeniable.

VII. CONCLUSION

Cloud computing has emerged as a way to provide a cost effective computing infrastructure for anyone with large needs for computing resources. Together with the Service Oriented Architecture, research scientists can construct scientific workflows composed of various scientific computing tools offered as services in the Cloud to answer complex research questions.

In this paper, we have described a Mobile Cloud system which enables mobile devices to design and participate in the scientific workflows running in the Cloud. The scientific researchers can use mobile devices to sketch an abstract workflow design to be submitted to the mobile cloud middleware layer, which will recommend and compose the optimal services according to the designer's requirements. On top of that, we further incorporated accountability mechanisms to provide trusted data provenance during the execution of the scientific workflows. Trusted data provenance implies that the recorded provenance data about a certain workflow is cryptographically verifiable to be attributed to the responsible services who, issued them. The provenance data thus can be used with confidence that its source is verifiable and its integrity has been preserved.

In the future development, it will be interesting to explore the utilization of the trusted provenance data collected, to improve the service recommendation for workflow design. The applicability of a particular service in a certain workflow and its performances in the past executions can provide much information to the research scientists and the recommendation system about characteristics of this service and its eligibility for the workflow under design. Another direction of development is to utilize existing workflow platforms or service repositories (e.g. BioCatalogue [33]) to construct workflows and provide trusted data provenance. In this way we can testify the concept of Mobile Cloud and trusted data provenance in the practice, improve our methodology so as to offer more value and insights to the community.

REFERENCES

[1] J. Yao, J. Zhang, S. Chen, C. Wang, D. Levy. Facilitating Bioinformatics Research with Mobile Cloud. In proc. International Conference on Cloud Computing, GRIDs, and Virtualization, pp.161-166, 2011

[2] W. Lu, J. Jackson and R. Barga. AzureBlast: A Case Study of Developing Science. In proc. Workshop on Scientific Cloud Computing, pp. 413-420, 2010.

[3] O. Moser, F. Rosenberg, S. Dustdar. Non-Intrusive monitoring and service adaptation for WS-BPEL. In proc. international conference on World Wide Web, pp. 815-824, 2008.

[4] Montage. http://montage.ipac.caltech.edu.

[5] I, Taylor, M, Shields, I. Wang, and R. Philp. Distributed P2P computing within Triana: A galaxy visualization test case. In proc. IEEE International Parallel and Distributed Processings Symposium, 2003.

[6] T. Oinn, P. Li, D.B. Kell, C. Goble, A. Goderis, M. Greenwood, D. Hull, R. Stevens, D. Turi, and J. Zhao. . Taverna/myGrid: Aligning a workflow system with the life sciences community. In *Workflows in e-Science*, Springer, 2006.

[7] M. Schena, D. Shalon, R.W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, October 1995.

[8] M. Y. Galperin. The molecular biology database collection: 2008 update. *Nucleic Acids Research*, November 2007.

[9] M. D. Brazas, J. A. Fox, T. Brown, S. McMillan, and B. F. F. Ouellette. Keeping pace with the data: 2008 update on the bioinformatics links directory. *Nucleic acids research*, 36, July 2008.

[10] R. Mulgan. Accountability: An ever-expanding concept? In: Public Administration, pp 555–573, 2000.

[11] A. R. Yumerefendi, J. S. Chase. Trust but verify: accountability for network services. In proc. ACM SIGOPS European workshop, article No. 37, 2004.

[12] J. Yao, S. Chen, C. Wang, D. Levy, and J. Zic. Accountability as a service for the cloud, in proc. IEEE International Conference on Services Computing, pp. 81–90, 2010.

[13] Q. Yu, X. Liu, A. Bouguettaya, and B. Medjahed. Deploying and managing web services: issues, solutions, and directions. *VLDB J.*, 17(3):537–572, 2008.

[14] Y. L. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science, in trans. ACM SIGMOD Record, volume 34, issue 3, pp. 31-36, September 2005.

[15] I. T. Foster, J.-S. Vöckler, M. Wilde, and Y. Zhao. Chimera: A Virtual Data System for Representing, Querying, and Automating Data Derivation, in *SSDBM*, 2002.

[16] J. Zhao, C. A. Goble, R. Stevens, and S. Bechhofer. Semantically Linking and Browsing Provenance Logs for Escience, in *ICSNW*, 2004.

[17] J. Myers, C. Pancerella, C. Lansing, K. Schuchardt, and B. Didier. Multi-Scale Science, Supporting Emerging Practice with Semantically Derived Provenance, in ISWC workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, 2003.

[18] E. Mulo, S. Dustdar, U. Zdun. Monitoring Web Service Event Trails for Business Compliance. In Proc. International Conference on Service-Oriented Computing and Applications , pp. 1-8, 2009.

[19] M. Huang, L. Peterson, A. Bavier. PlanetFlow:maintaining accountability for network services. In Proc. ACM SIGOPS Operating Systems Review, pp. 89-94, 2006.

[20] Y. Zhang, K. Lin, J.Y.J. Hsu. Accountability monitoring and reasoning in service-oriented architectures. In Trans. Service Oriented Computing and Applications, Volume 1, Number 1, pp. 35-50, 2007.

[21] A. C. Squicciarini, W. Lee, B. Thuraisingham, E. Bertino. End-to-end accountability in grid computing systems for coalition information

sharing. In Proc. Workshop on Cyber Security and Information Intelligence Research , 2008.

[22] M. Kanehisa, S. Goto, M. Furumichi, M. Tanable and M. Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. In trans. Nucleic Acids Research, volume 38, Database issue, pp. 355-360, 2010.

[23] G. Huerta-Canepa and D. Lee. A virtual cloud computing provider for mobile devices. presented at the Proceedings of the 1st ACM Workshop on Mobile Cloud Computing Services: Social Networks and Beyond, San Francisco, California, pp. 61-65, 2010.

[24] J. Zhang, et al.. mBOSSS+: A Mobile Web Services Framework. in Services Computing Conference (APSCC), 2010 IEEE Asia-Pacific, pp. 91-96, 2010.

[25] I. Giurgiu, et al.. Calling the cloud: enabling mobile phones as interfaces to cloud applications. the 10th ACM/IFIP/USENIX International Conference on Middleware, pp. 83-102, May,2009

[26] K. Kumar and Y. Lu. Cloud Computing for Mobile Users. Computer, vol. 18, issue 99, pp. 51-56, 2010.

[27] R. Kemp, N. Palmer, T. Kielmann, and H. Bal. Cuckoo: a Computation Offloading Framework for Smartphones. In MobiCASE '10: Proceedings of The Second International Conference on Mobile Computing, Applications, and Services, pp. 62-81, 2010.

[28] C. Scheidegger, D. Koop, E. Santos, H. Callahan, J. Freire, C. Silva. Tackling the provenance challenge one layer at a time. Concurrency and Computation: Practice and Experience, 20(5), pp. 473–483, 2008.

[29] Y. Simmhan, B. Plale, D. Gannon. Karma2: Provenance management for data driven workflow. International Journal of Web Services Research 5(2), pp. 1–22, 2008.

[30] J. Zhao, C. Goble, R. Stevens, D. Turi. Mining taverna's semantic web of provenance. Concurrency and Computation: Practice and Experience 20(5), pp. 463–472, 2008.

[31] M. Palankar, A. Iamnitchi, M. Ripeanu, S. Garfinkel. Amazon s3 for science grids: a viable solution. In: Workshop on Data-aware distributed Computing, pp. 55–64, 2008.

[32] L. Moreau, J. Freire, J. Futrelle, R. McGrath, J. Myers, P. Paulson. The open provenance model. (2007). URL: http://openprovenance.org/

[33] BioCatalogue, URL: http://www.biocatalogue.org/

[34] T. Andrews, F. Curbera, H. Dholakia, et al.: Business process execution language for web services (BPEL4WS) specifications (2003). URL http://download.boulder.ibm.com/ibmdl/pub/software/dw/specs/ws-bpel/ws-bpel.pdf

www.iariajournals.org

International Journal On Advances in Intelligent Systems

ICAS, ACHI, ICCGI, UBICOMM, ADVCOMP, CENTRIC, GEOProcessing, SEMAPRO, BIOSYSCOM, BIOINFO, BIOTECHNO, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS, ENERGY, COLLA, IMMM, INTELLI, SMART, DATA ANALYTICS

issn: 1942-2679

International Journal On Advances in Internet Technology

ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING, MOBILITY, WEB

issn: 1942-2652

International Journal On Advances in Life Sciences

eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO, SOTICS, GLOBAL HEALTH

issn: 1942-2660

International Journal On Advances in Networks and Services

ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION, VEHICULAR, INNOV

issn: 1942-2644

International Journal On Advances in Security

ICQNM, SECURWARE, MESH, DEPEND, INTERNET, CYBERLAWS

issn: 1942-2636

International Journal On Advances in Software

ICSEA, ICCGI, ADVCOMP, GEOProcessing, DBKDA, INTENSIVE, VALID, SIMUL, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS, IMMM, MOBILITY, VEHICULAR, DATA ANALYTICS

issn: 1942-2628

International Journal On Advances in Systems and Measurements

ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL, INFOCOMP

issn: 1942-261x

International Journal On Advances in Telecommunications

AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA, COCORA, PESARO, INNOV

issn: 1942-2601