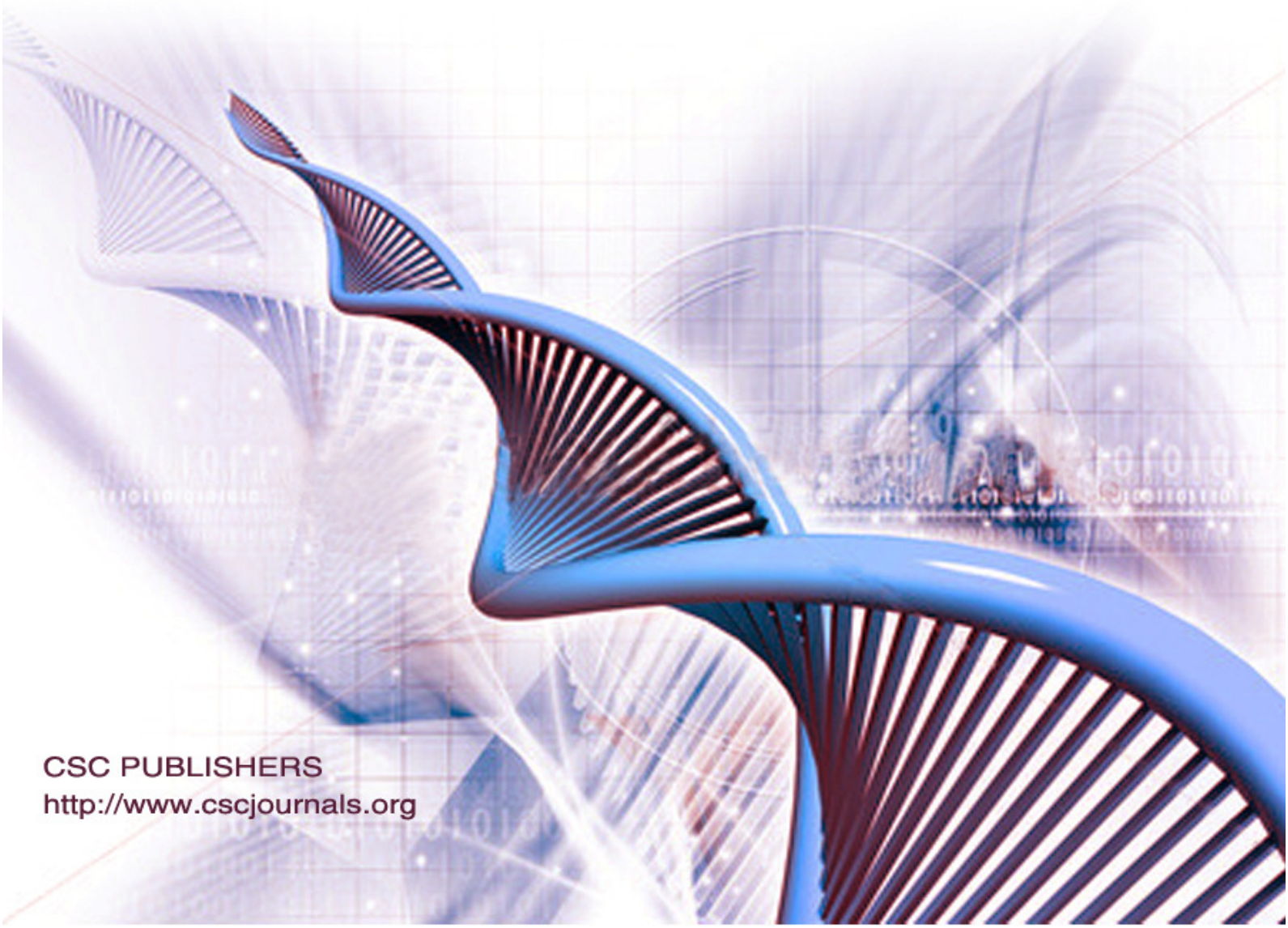
Volume 5 ▪ Issue 3 ▪ August 2011

INTERNATIONAL JOURNAL OF

BIOMETRICS AND BIOINFORMATICS (IJBB)

ISSN : 1985-2347
Publication Frequency: 6 Issues / Year



CSC PUBLISHERS
http://www.cscjournals.org

INTERNATIONAL JOURNAL OF BIOMETRICS AND BIOINFORMATICS (IJBB)

VOLUME 5, ISSUE 3, 2011

EDITED BY
DR. NABEEL TAHIR

ISSN (Online): 1985-2347

International Journal of Biometrics and Bioinformatics (IJBB) is published both in traditional paper form and in Internet. This journal is published at the website http://www.cscjournals.org, maintained by Computer Science Journals (CSC Journals), Malaysia.

IJBB Journal is a part of CSC Publishers

Computer Science Journals

http://www.cscjournals.org

INTERNATIONAL JOURNAL OF BIOMETRICS AND BIOINFORMATICS (IJBB)

Book: Volume 5, Issue 3, August 2011
Publishing Date: 31-08-2011
ISSN (Online): 1985-2347

IJBB Journal is a part of CSC Publishers
http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers, 2011

EDITORIAL PREFACE

This is the third issue of volume five of International Journal of Biometric and Bioinformatics (IJBB). The Journal is published bi-monthly, with papers being peer reviewed to high international standards. The International Journal of Biometric and Bioinformatics is not limited to a specific aspect of Biology but it is devoted to the publication of high quality papers on all division of Bio in general. IJBB intends to disseminate knowledge in the various disciplines of the Biometric field from theoretical, practical and analytical research to physical implications and theoretical or quantitative discussion intended for academic and industrial progress. In order to position IJBB as one of the good journal on Bio-sciences, a group of highly valuable scholars are serving on the editorial board. The International Editorial Board ensures that significant developments in Biometrics from around the world are reflected in the Journal. Some important topics covers by journal are Bio-grid, biomedical image processing (fusion), Computational structural biology, Molecular sequence analysis, Genetic algorithms etc.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with volume 5, 2011, IJBB appears in more focused issues. Besides normal publications, IJBB intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

The coverage of the journal includes all new theoretical and experimental findings in the fields of Biometrics which enhance the knowledge of scientist, industrials, researchers and all those persons who are coupled with Bioscience field. IJBB objective is to publish articles that are not only technically proficient but also contains information and ideas of fresh interest for International readership. IJBB aims to handle submissions courteously and promptly. IJBB objectives are to promote and extend the use of all methods in the principal disciplines of Bioscience.

IJBB editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJBB. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJBB provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
International Journal of Biometric and Bioinformatics (IJBB)

EDITORIAL BOARD

Dr Antonio Marco
University of Manchester
United Kingdom

Dr Peng Jiang
University of Iowa
United States of America

TABLE OF CONTENTS

Volume 5, Issue 3, August 2011

Pages

Life Expectancy Estimate With Bivariate Weibull Distribution Using Archimedean Copula

Eun-Joo Lee *elee@millikin.edu*
Department of Mathematics
Millikin University
Decatur, IL 62522

Chang-Hyun Kim *maraychk@gmail.com*
Illinois Natural History Survey
University of Illinois at Urbana-Champaign
Champaign, IL 61820

Seung-Hwan Lee *slee2@iwu.edu*
Department of Mathematics and Computer Science
Illinois Wesleyan University
Bloomington, IL 61701

Abstract

Archimedean copulas are used to construct bivariate Weibull distributions. Co-movement structures of variables are analyzed through the copulas, where the tail dependence between the variables is explored with more flexibility. Based on the distance between the copula distribution and its empirical version, a copula that may best fit data is selected. With extra computing costs, the adequacy of the copula chosen is then assessed. When multiple myeloma data are considered, it is found that relationship between survival time of a patient and the hemoglobin level is well described by the Clayton copula. The bivariate Weibull distribution constructed by the copula is used to estimate value at risk from which we investigate the anticipated longest life expectancy of a patient with the disease over the treatment period.

Keywords: Archimedean Copula, Dependence, Weibull Distribution, Value at Risk.

1. INTRODUCTION

Copulas are a useful tool used to model a joint distribution function of variables of interest. In particular, copulas have gained their importance as simple functions to describe the dependence structure of random variables in the joint distribution. As a model for the dependence structure, copulas have several advantages over other dependence measures such as the correlation coefficient (Sklar [28], Genest and Rivest [12], Nelson [26]). For example, using copulas, modeling both linear and non-linear dependencies of variables is possible, and the degree of dependence in the tail of the underlying distribution can be described (Embrechts et al. [7]). Many authors have studied the use of copula in applications, including risk management (Freez and Valdez [9]) and survival analysis (Zheng and Klein [30], Rivest and Wells [27]). In this work, we construct bivariate Weibull distributions using Archimedean copulas that reflect on the asymmetric dependence structure. These copulas are the Gumbel, Clayton, Frank and Independence copulas, each having different characteristics of tail dependence.

Copulas have varying amounts of tail dependence depending on the choice of copulas. Therefore, an important issue in using copulas is the choice of appropriate copulas. Poorly chosen copulas may lead to undesired results about the actual relationship between variables. The copula selection issue has been studied by many authors, including Melchiori [25], Durrelman et al. [6], Kumar and Shoukri [19], Frees and Valdez [9] and Genest and Rivest [12]. Similar to the procedures they have developed, we discuss the copula selection procedures

based on the distance of the copula distribution and its empirical version. With extra computing costs, we further examine the goodness of fit of the copula selected. The procedures are based on a process over the domain of the generator for Archimedean copulas. Under the null hypothesis of the no model misspecification, the distributions of the process from the distance measure can be easily approximated by the simulation technique. As a numerical measure for the assessment of the model adequacy, we consider the supremum of the process from which the empirical p -values are obtained.

Multiple myeloma is a progressive and invariably fatal disease caused by the accumulation of abnormal plasma cells in the bone marrow. The prognosis of the disease is often unpredictable and overall survival is ranged from a few months to more than 10 years (Kyle and Rajkumar [20]). Traditionally, multiple myeloma has been staged by the method developed by Durie and Salmon [5], although a newer staging method has been developed recently (Greipp et al. [13]). In the staging method by Durie and Salmon [5], it has been known that the level of hemoglobin (denoted by HB hereafter) in the blood of a multiple myeloma patient is strongly associated with the tumor mass and thus is a strong indicator of the disease progress (Durie and Salmon [5], Kyle and Rajkumar [20]). The objective of this paper is to demonstrate the benefits of using copulas to model dependencies in multiple myeloma data with a particular focus on potential survival time of a patient over the treatment period. This was carried out at the Medical Centre at the University of West Virginia. To simplify our discussion, the complete data points of survival time, in months, of male patients with the disease (denote by ST hereafter), i.e., the time from diagnosis until death from multiple myeloma, and the corresponding level of HB are considered in this paper, where the sample size is 22. See Krall et al. [18] and Collect [2] for details about the data. The effect of HB on the survival times of the patients is explored using the bivariate Weibull distribution constructed by copulas, where a measure of linear dependence is not so informative, as will be seen in Figure 2 and described in Section 4.2. We incorporate the copulas into the calculation of value at risk for the survival time. The value at risk is a risk measurement technique often used in the area of financial risk management (Jorion [17]). We use this method as a tool to estimate the anticipated maximum life span, i.e. maximum extension possible for a life, with reference to the level of hemoglobin that influences the survival time.

The layout of this paper is as follows. Section 2 presents Archimedean copula functions used in this work. Section 3 discusses the association parameter and the dependence measure of the copula functions. Section 4 constructs bivariate Weibull distributions using copula, checks the adequacy of the copula selected, and calculates value at risk associated with survival time. Concluding remarks are presented in Section 5.

2. COPULA MODEL

2.1 Copula Function

Estimating a multivariate distribution with correlations is not an easy process. A copula is a useful tool that accommodates this problem. It joins a multivariate distribution function to univariate marginal distribution functions, so a copula function is a multivariate distribution function. Specifically, a copula function, denoted by C, is a multivariate distribution function with uniform marginal distribution functions, $F_1, F_2, ..., F_p$, on the interval [0, 1], i.e., if for $x_1, x_2, ..., x_p$, $F(x_1, x_2, ..., x_p)$ is a multivariate probability distribution with marginals $F_1(x_1), F_2(x_2), ..., F_p(x_p)$, then $F(x_1, x_2, ..., x_p)$ can be written as

$$F(x_1, x_2, ..., x_p) = C(F_1(x_1), F_2(x_2), ..., F_p(x_p)).$$

For a bivariate case, the copula form is the easiest way to express and generate the joint distribution (Venter [29]). In this work, we primarily look at this bivariate copula. In the bivariate case, a copula is a function $C : [0,1]^2 \to [0,1]$ such that $C(u,0) = C(0,u) = 0$ for all u in $[0,1]$, $C(v,1) = C(1,v) = v$ for all v in $[0,1]$ and $C(u_2, v_2) - C(u_1, v_2) - C(u_2, v_1) + C(u_1, v_1) \geq 0$

for all $0 \le u_1 \le u_2 \le 1$ and $0 \le v_1 \le v_2 \le 1$. From this, when $F_1(x_1) = u$ and $F_2(x_2) = v$, a copula function $C(F_1(x_1), F_2(x_2))$ is a proper bivariate distribution function. Conversely, any bivariate distribution function $F(x_1, x_2)$ with continuous marginal distribution functions F_1 and F_2 can be uniquely expressed by a copula function

$$C(u, v) = F(F_1^{-1}(u), F_2^{-1}(v)).$$

The following theorem summarizes the above results (Sklar [28]).

Theorem (Sklar's Theorem) Let F be a bivariate joint distribution function of continuous random variables X and Y with corresponding marginal distribution functions F_1 and F_2. There exists a copula C (i.e., a bivariate distribution function on $[0,1]^2$ with uniform marginal distribution functions) such that, for $-\infty < x_1 < \infty, -\infty < x_2 < \infty$,

$$F(x_1, x_2) = P(X_1 \le x_1, X_2 \le x_2) = C(F_1(x_1), F_2(x_2)). \quad (1)$$

Note that Sklar's theorem simply implies that $C(u, v) = P(U \le u, V \le v)$ for uniform random variables U and V over $[0,1]$. We close this section by describing meaningful bounds for copula.

Theorem (Frechet-Hoeffding Bounds) For every copula C and every (u, v) in $[0,1]^2$,

$$\max(u + v - 1, 0) \le C(u, v) \le \min(u, v).$$

Note that by Sklar's theorem $F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$, where $C(u, v) = F(F_1^{-1}(u), F_2^{-1}(v))$ and $u = F_1(x_1), v = F_2(x_2)$. Thus,

$$\max(F_1(x_1) + F_2(x_2) - 1, 0) \le F(x_1, x_2) \le \min(F_1(x_1), F_2(x_2)).$$

2.2 Archimedean Copula

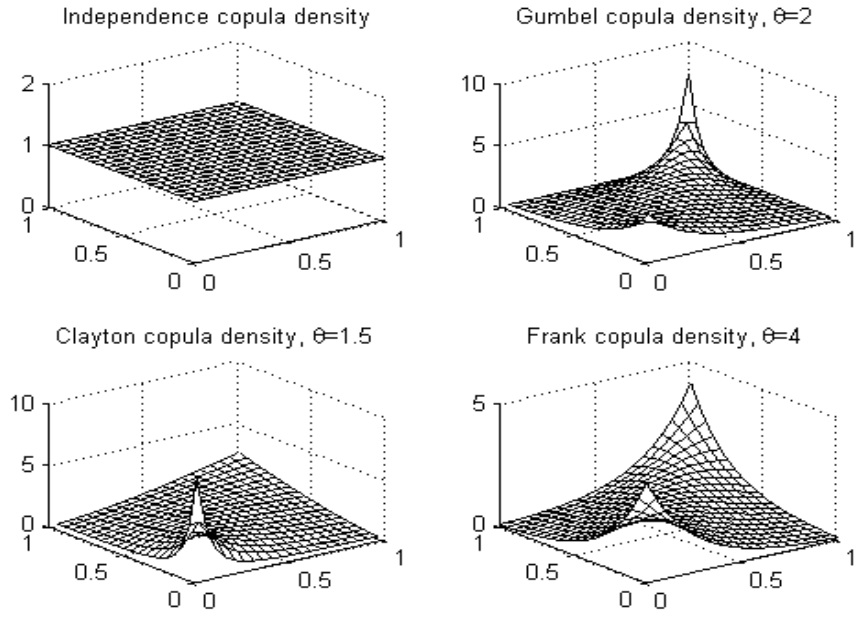


FIGURE1: Independence, Gumbel, Clayton, Frank copula density plots

The Archimedean copula is a convenient method to model a bivariate distribution due to its simple form and a variety of dependence structures. The use of Laplace transformations leads to the construction of Archimedean copulas. Specifically, let φ be a continuous monotonically decreasing function from $[0,1]$ to $[0,\infty)$ such that $\varphi(1)=0$ and $\varphi''(x)>0$. Define the pseudo inverse of φ as follows: $\varphi^{[-1]}(x)=\varphi^{-1}(x)$ for $0\le x\le\varphi(x)$ and zero for $\varphi(0)\le x\le\infty$. Note that if $\varphi(0)=\infty$, then $\varphi^{[-1]}=\varphi^{-1}$. Then, for real numbers, u and v, an Archimedean copula, C, of bivariate random variables U and V is given by

$$C(u,v)=\varphi^{[-1]}(\varphi(u)+\varphi(v)), \tag{2}$$

where $\varphi:[0,1]\to[0,\infty)$, and this φ is often called a generator of copula.

Archimedean copulas involve a tail dependence parameter, also referred as the association parameter. This describes the amount of dependence in the upper tail or lower tail of a multivariate distribution and can be used to analyze the dependence among extreme values. The generator φ contains all of the information about the dependence structure of the multivariate distribution of random variables in terms of the parameter of association. The association parameter is denoted by θ throughout this paper.

Based on the level of the tail dependence structure, we consider four families of Archimedean copulas in this paper. They are the Gumbel copula (Gumbel [14], Hougaard [15]), the Clayton copula (Clayton [1]) which is also referred to as Cook and Johnson's copula (Cook and Johnson [3]), the Frank copula (Frank [8]) and Independence (or Product) copula. Each family of the copulas is generated by the formula (2) through the generator φ. Specifically, Gumbel's copula is

$$C(u,v;\theta)=\exp\{-[(-\log u)^{\theta}+(-\log v)^{\theta}]^{1/\theta}\},$$

and the corresponding generator is $\varphi(t)=(-\log t)^{\theta}$ for $\theta\ge 1$. As a special case, the parameter $\theta=1$ implies independence between the distributions. With $\theta\to\infty$, the Gumbel copula attains the Frechet-Hoeffding lower bound, so the distribution is characterized by extreme values. This implies higher dependence in the upper tail. Clayton's copula is

$$C(u,v;\theta)=(u^{-\theta}+v^{-\theta}-1)^{-1/\theta}.$$

The Clayton copula generator is $\varphi(t)=\dfrac{t^{-\theta}-1}{\theta}$ for $\theta>0$. With $\theta\to\infty$, the Clayton copula attains the Frechet-Hoeffding upper bound, so higher dependence in the lower tail. As $\theta\to 0$, the Clayton copula implies independence between the distributions. The Frank copula is

$$C(u,v;\theta)=-\frac{1}{\theta}\log\left[1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right],$$

and its generator is $\varphi(t)=-\log\left[\dfrac{e^{-\theta t}-1}{e^{-\theta}-1}\right]$ $\theta\neq 0$. With $\theta\to\infty$, the Frank copula attains the Frechet-Hoeffding upper bound, and with $\theta\to-\infty$, it attains the Frechet-Hoeffding lower bound. When $\theta>0$ ($\theta<0$), it implies positive (negative) dependence between the distributions. When $\theta\to 0$, the copula implies independence between the distributions. Finally, the independence copula is

$$C(u,v)=u\cdot v,$$

with the generator $\varphi(t)=-\log t$. Note that there is no association parameter in the independence copula.

Tail dependence of copulas can be illustrated by their density function. The bivariate distribution function is defined in (1), and the corresponding density function is obtained by differentiation,

$$f(x_1, x_2) = c(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \cdot f_2(x_2)$$

where c is the density of C, and f_1 and f_2 are the marginals. Written this way, it is also possible to define Archimedean copulas in the multivariate case. See McNeil et al. [24] for details. Figure 1 displays the copula densities. As depicted from this figure, each copula has varying degrees of dependence according to values of θ. Note in this figure that the independence copula has unit everywhere. The estimation of parameters of the copulas is discussed in Section 3.1.

3. ASSOCIATION AND DEPENDENCE

3.1 Measuring Association

Two commonly used measures of association would be Spearman's ρ and Kendall's τ. They are based on the rank of data, so they have the invariance property under monotonic transformations. For Archimedean copulas, Kendall's τ has the copula representation, and so it captures perfect dependence. On the contrary, there seems to be no simple expression for Spearman's ρ. Kendall's τ is given by (Nelson [26], Joe [16], Genest and Mackay [10, 11])

$$\tau = 4\int_0^1 \int_0^1 C(u,v)dC(u,v) - 1. \quad (3)$$

From the expression in (3), Kendall's τ is calculated by a copula that contains the association parameter. Conversely, the association parameter can be measured by Kendall's τ obtained from data. For bivariate Archimedean copulas, where the two random variables are absolutely continuous, Kendall's τ can be readily calculated via the following identity (Genest and MacKay [10, 11])

$$\tau = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)} dt \ . \quad (4)$$

The copula generator contains the association parameter, and from (4) the generator can be expressed through Kendall's τ. Therefore, the association parameter, θ, is measured by solving the identity in (4). For example, it can be shown that for the Clayton copula,

$$1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)} dt = \frac{\theta}{\theta + 2},$$

and this yields $\tau = \frac{\theta}{\theta + 2}$. Similar algebra leads to $\tau = \frac{\theta - 1}{\theta}$ for the Gumbel copula. Unlike these two copulas, the Frank copula doesn't have a closed form of τ that can be directly expressed by θ. It is necessary to use numerical methods to solve the following identity

$$\tau = 1 - \frac{4}{\theta}\left(1 - \frac{1}{\theta}\int_0^\theta \frac{t}{e^t - 1} dt\right).$$

In this work, the random search method is utilized to estimate the Frank copula parameter, among other numerical methods.

The dependence structure of HB and ST is displayed in the scatter plot (Figure 2), where Kendall's τ is 0.2208. This and the procedures above result in θ = 1.2834, 0.5667 and 2.07 for the Gumbel, Clayton and Frank copulas, respectively.

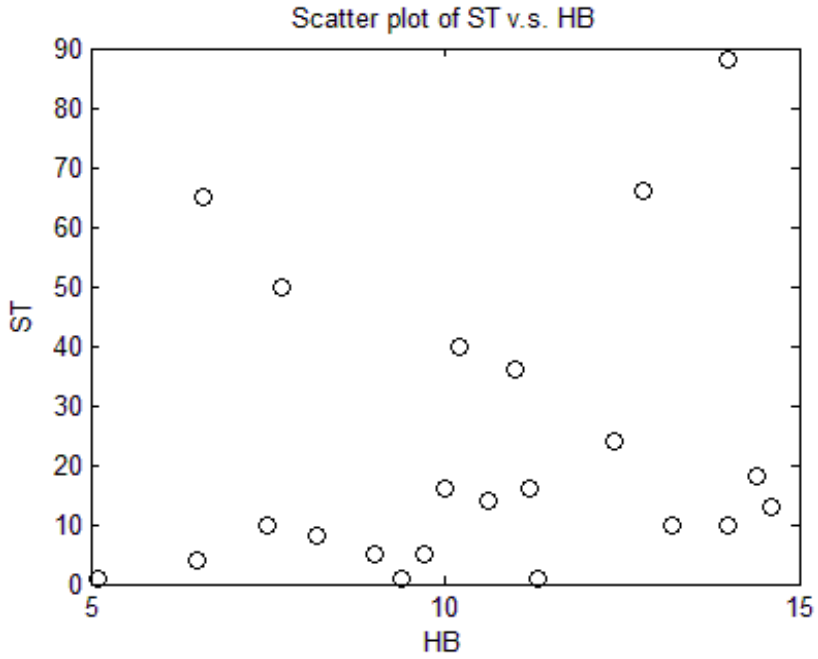


FIGURE 2: Scatter plot of HB vs ST

3.2 Dependence

Tail dependence deals with the degree of dependence in the tails of a bivariate distribution, so it describes the dependence structure of extreme events. For example, Figure 1 in Section 2.2 displays that different copulas show different behaviors in their tails. This implies that tail dependence may vary depending on the choice of copulas.

Let X_1 and X_2 be random variables with continuous distribution functions $F_1(x_1)$ and $F_2(x_2)$. The upper- and lower-tail dependence coefficients are defined as the limit of conditional probability, respectively,

$$\lambda_U = \lim_{u \to 1-} P(Y \geq F_2^{-1}(u) \mid X \geq F_1^{-1}(u)),$$

$$\lambda_L = \lim_{u \to 0+} P(Y \leq F_2^{-1}(u) \mid X \leq F_1^{-1}(u)),$$

for u in (0,1). If the value of the upper (lower) tail dependence coefficient is positive, then X_1 and X_2 have structure dependent at the upper (lower) tail. In contrast, zero tail dependence implies asymptotic independence. The tail dependences can be also expressed through copula, showing the fact that the tail dependence is a copula property,

$$\lambda_U = \lim_{u \to 1-} \frac{1 - 2u + C(u,u)}{1-u}, \quad \lambda_L = \lim_{u \to 0+} \frac{C(u,u)}{u}.$$

From this, the Gumbel, Clayton, Frank and independence copulas have $[0, 2-2^{\theta}]$, $[2^{-1/\theta}, 0]$, $[0,0]$ and $[0,0]$, respectively, where [a, b] = [lower, upper] tail dependence coefficients. By plugging in the values of θ found in Section 3.1, theoretical [lower, upper] tail dependence coefficients for the Gumbel, Clayton, Frank and independence copulas are [0,0.2838], [0.2943,0], [0,0] and [0,0], respectively. This indicates that the Gumbel copula has the upper tail dependence but does not have the lower tail dependence, the Clayton copula has the lower tail dependence but does not have the upper tail dependence, while the Frank and the independence copulas have neither. Numerical computations of the tail dependence using the limit formulas above are

reported in Table 1. It shows the same phenomena as in the theoretical analysis. For example, as u tends to 1 through values less than 1, λ_U for Gumbel tends to 0.2838.

$u \to 0+$	Gumbel λ_L	Clayton λ_L	Frank λ_L	Indep. λ_L.	$u \to 1-$	Gumbel λ_U	Clayton λ_U	Frank λ_U	Indep. λ_U
.10	0.1922	0.3806	0.1973	0.1000	.90	0.3459	0.1482	0.1973	0.1000
.05	0.1170	0.3486	0.1075	0.0500	.95	0.3147	0.0762	0.1075	0.0500
.005	0.0225	0.3076	0.0117	0.0050	.995	0.2869	0.0078	0.0117	0.0050
.001	0.0071	0.2995	0.0024	0.0010	.999	0.2844	0.0016	0.0024	0.0010
.00001	0.0003	0.2947	0.0000	0.0000	.99999	0.2838	0.0000	0.0000	0.0000

TABLE 1: Tail dependence coefficient for the copulas associated with data

4. THE PROCEDURES

4.1 Bivariate Weibull Distribution

In the parametric analysis of survival analysis, one of the commonly used models is the two-parameter Weibull distribution. We construct the bivariate Weibull distributions based on the four copula functions stated in Section 2.2. Specifically, given two marginal Weibull distributions

$$F_i(x_i) = 1 - e^{-(x_i/\beta_i)^{\alpha_i}}, i = 1,2,$$

it is possible to construct a bivariate distribution $F(x_1, x_2)$ such that $F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$. For example, choosing the Gumbel copula gives a bivariate Weibull distribution given by

$$F(x_1, x_2) = \exp\{-[(-\log F_1(x_1))^\theta + (-\log F_2(x_2))^\theta]^{1/\theta}\},$$

the Clayton copula yields a bivariate Weibull distribution given by

$$F(x_1, x_2) = (F_1(x_1)^{-\theta} + F_2(x_2)^{-\theta} - 1)^{-1/\theta},$$

the Frank copula leads to a bivariate Weibull distribution given by

$$F(x_1, x_2) = -\frac{1}{\theta}\log[1 + \frac{(e^{-\theta F_1(x_1)} - 1)(e^{-\theta F_2(x_2)} - 1)}{e^{-\theta} - 1}],$$

and the independence copula produces a bivariate Weibull distribution given by

$$F(x_1, x_2) = F_1(x_1)F_2(x_2).$$

For the multiple myeloma data, where X_1 and X_2 respectively represent HB and ST, it is found that Weibull distributions can be fit to them with α_1 = 11.307, β_1=3.859, and α_2=20.175, β_2=0.839. Associated with these, we use the four bivariate distribution functions above as the underlying distribution of returns to compute value at risk stated in Section 4.4.

4.2 Simulation

From the scatter plot of ST and HB in Figure 2 in Section 3.1, Pearson's correlation coefficient is not sufficiently informative on the dependence structure. It is problematic to identify the dependence (co-movement) of the variables at the tails. Computation of the linear correlation coefficient of ST and HB yields 0.1852. The corresponding $p-$value for testing the hypothesis of no correlation against the alternative that there is a non-zero correlation is 0.4094. This value

does not show pronounced evidence that the two variables are linearly dependent. So, the dependence structure of the variables is modeled via copula.
In Section 4.1, we created four bivariate Weibull distributions based on Gumbel, Clayton, Frank and the independence copulas. The first three copulas are respectively parameterized by θ=1.2834, 0.5667 and 2.07 as discovered in Section 3.1. Figure 3 shows 250 simulated values from the four bivariate Weibull distributions that use the Gumbel, Clayton, Frank and independence copulas. In this figure, it seems that the positive dependence is somewhat observed. A high level of hemoglobin (HB) tends to influence the survival time (ST) of patients in the Gumbel copula, indicating high tail dependence. Positive dependence between the variables is also observed in the Frank copula. However, it seems that there is no dominant copula among the Archimedean copulas considered here. The independence copula provides no distinct patterns due to the assumption of independence among the variables.

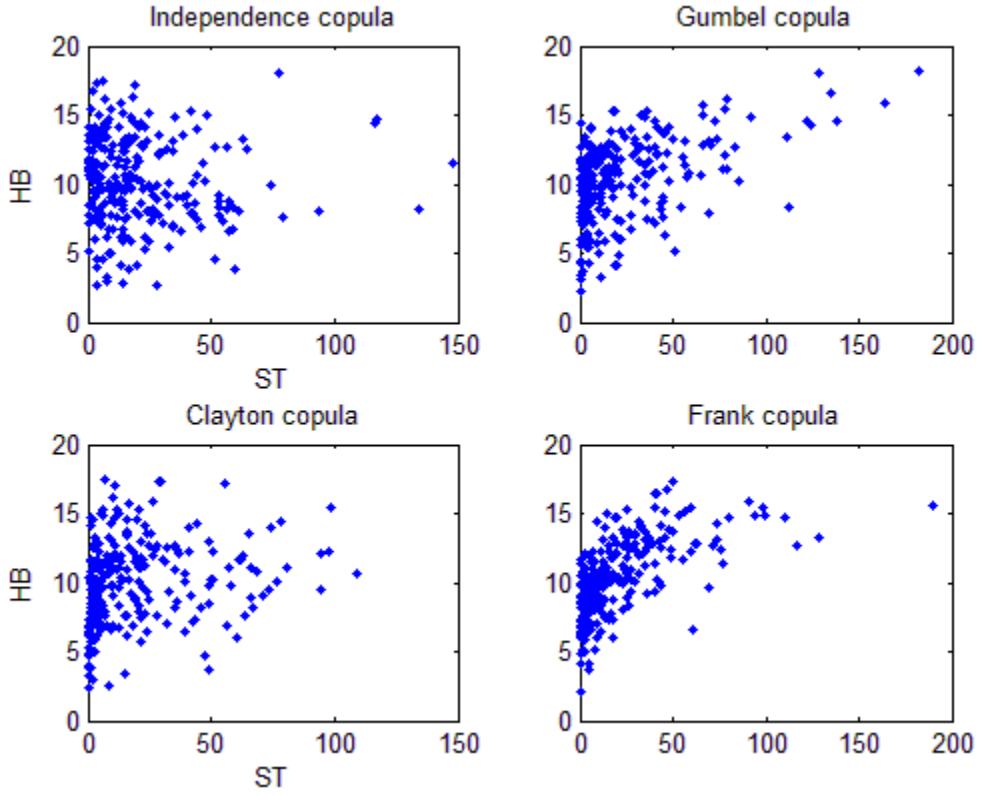


FIGURE 3: Indpendence, Gumbel, Clayton and Frank copulas, 1000 simulated samples

4.3 Copula Adequacy

A copula is the dependence structure of the data distribution. Since it has varying amounts of tail dependence depending on the choice of copulas, properly chosen copulas should be used in application. The adequacy of the copula selected also needs to be checked. These issues are discussed in this section. The procedures are based on the distance of the copula distribution and its empirical version.

Define a pseudo random variable T_i, for $i, j = 1,...,n$,

$$T_i = \{\text{the number of } (X_{1,i}, X_{2,j}), j = 1,...,n : X_{1,j} < X_{1,i}, X_{2,j} < X_{2,i}\}/(n-1).$$

Further, define $K(t) = P(T_i \le t)$ for t in [0,1]. Genest and Rivest [12] showed that the distribution of $C(u,v)$ is

$$K(t) = t - \frac{\varphi(t)}{\varphi'(t)}.$$

By plugging in, the following $K(t)$'s for the copulas are obtained: the Gumbel copula has $K(t) = t - \frac{t \log t}{\theta}$, the Clayton copula takes $K(t) = t - \frac{t^{1+\theta} - t}{\theta}$, the Frank copula gives $K(t) = t - \frac{e^{t\theta} - t}{\theta} \log \frac{e^{-t\theta} - 1}{e^{-\theta} - 1}$, and the independence copula yields $K(t) = t - t \log t$. Now, define an empirical distribution of $K(t)$,

$$\hat{K}(t) = \frac{1}{n} \sum_{i=1}^{n} I(T_i \le t),$$

where I is the indicator function. Then, we select the copula that best fits the data for which the distance of $K(t)$ and its estimate $\hat{K}(t)$ is minimized. Specifically, as usual in the literature, the best copula is selected as the one which minimizes the Kolmogorov-Smirnov type distance defined as

$$D(K, \hat{K}) = \int_0^1 \{K(t) - \hat{K}(t)\}^2 d\hat{K}(t). \quad (5)$$

For the Gumbel, Clayton, Frank and independence copulas associated with data, computation of $D(K, \hat{K})$ yields 0.0063, 0.0047, 0.0052 and 0.0162, respectively. This implies that the Clayton copula may be the best model for the data set considered. It appears that the independence copula is unlikely to be appropriate in this study.

We now check the validity of the copula chosen. The procedures are based on a process, derived from the distance measure in (5), over the domain of the copula generator. Define the process in t,

$$D(t) = \int_0^t \{K(t) - \hat{K}(t)\}^2 d\hat{K}(t)$$

for $0 < t \le 1$. Similar to Lin et al. [23], Lee and Yang [22] and Lee et al. [21], with the parameter of association, we generate data from the copula through simulation. From the simulated data, we obtain an estimate of the parameter, $\hat{\theta}$. Denote the resulting distribution of the copula by $K^*(t)$. The process associated with this and the parameter estimate is then given by

$$D^*(t, \hat{\theta}) = \int_0^t \{K^*(t) - \hat{K}(t)\}^2 d\hat{K}(t)$$

for $0 < t \le 1$, and this simulated process is an approximation to the observed process $D(t)$.

A large number of samples can be generated repeatedly from this simulated process. Since the null distribution of the copula is approximated by $D^*(t, \hat{\theta})$, there will be no distinguished behavior of $D(t)$ comparing to a large number of realizations produced from $D^*(t, \hat{\theta})$, if the copula fits the data. Under the null hypothesis that the copula model is valid, the process $D^*(t, \hat{\theta})$ will randomly fluctuate above, near zero. So, as a numerical measure for the assessment of the model adequacy, we consider the supremum of the process $D(t)$ over (0,1], $S = \sup_{0<t\le1} D(t)$. An unusually large value of S would indicate that the copula is not valid. Let $S^* = \sup_{0<t\le1} D^*(t, \hat{\theta})$. Then, the distribution of S is approximated by the conditional distribution of

S^* given the data. This implies that the p-value $P(S \geq s)$ can be approximated by $P(S^* \geq s)$, and $P(S^* \geq s)$ is estimated through the simulation technique. For the data considered here, using these procedures associated with the Clayton copula, the estimated p-value is 0.4945, which means the Clayton copula is appropriate. This estimated p-value is based on 1000 realizations of the simulated process as suggested by Lin et al. [23].
Using the same procedures, we also found that the results for Gumbel and Frank copulas are not significant (not shown). Therefore there is not sufficient evidence to reject the Gumbel, Clayton and Frank copulas. Thus all three classes of models may be applicable, although comparisons of the results from the individual models suggest that the Clayton copula may fit the data better. More data may be required to discriminate adequately between the three copulas.

4.4 Life Expectancy Estimate

Copula	VaR (90%)	ES (90%)	VaR (95%)	ES(95%)
Gumbel	54.6223	84.9359	74.8936	106.4190
Frank	54.5661	84.6025	74.7020	105.8340
Clayton	54.4268	84.3032	74.2875	105.5011
Independence	54.1984	83.6832	74.0877	104.4382

TABLE 2: Time estimates (in months) of VaR and ES

In this section, we employ the copulas to calculate value at risk. The value at risk (VaR) is a risk measurement technique often used in the area of Financial Risk Management (Jorion [17]). Consider a linear combination of X_1 and X_2, $Z = w_1 X_1 + w_2 X_2$, where X_1 and X_2 represent the same type of data, and w_1 and w_2 are the weights taken over the real number (R), for each variable, with distribution function F_Z. Let z be a realization of Z, and R be the set of real numbers. The Value at Risk of Z at probability level α is then defined as

$$\text{VaR}_\alpha(Z) = F_Z^{-1}(\alpha) = \inf\{z \in R \mid F(z) \geq \alpha\}.$$

Value at Risk is in fact an alternative notation for the quantile function of F_Z evaluated at α. In the area of Financial Risk Management, VaR is commonly used to estimate the largest potential loss that might be expected from holding a portfolio over a given period of time at a specified confidence level (Crouhy et al. [4]). For example, if a portfolio has a VaR of $1million at the 95 percent confidence level, then VaR is the cutoff loss such that the probability of losing at least $1million is less than 5 percent over a given time period. So VaR is a measure of risk that summarizes the distribution of returns into a single number. Similarly, in this work, we use this VaR as a tool to examine the anticipated life expectancy of a patient with multiple myeloma from diagnosis until death. As stated in Collet [2], multiple myeloma is a disease characterized by the accumulation of abnormal plasma cell in the bone marrow. Its proliferation within the bone causes pain and the destruction of bone tissue. The condition could be fatal unless treated.
To obtain the distribution of F_Z^{-1} under the setting above, we aggregate simulated returns of X_1 and X_2 associated with the weights, w_1 and w_2. In this work, where X_1 and X_2 respectively represent ST and HB, letting w_1=1 and w_2=0, based on the bivariate Weibull distribution, we get VaR for the survival time, and its procedures are based on a large number of simulated samples generated from copula. For our case, we simulated 5,000,000 samples from each copula to calculate VaR. Since our concern is with longest survival time, VaR is evaluated at the upper tail of the returns distribution of simulated values. Table 2 presents the estimated values of VaR at 90 and 95 percent confidence levels, i.e., 90% (longest) survival time and 95% (longest) survival

time, in months, from diagnosis until death from multiple myeloma. Expected shortfall (ES) is also used to examine the anticipated longest survival time. ES is the expectation in excess of VaR, indicating what we expect if an event occurs (Crouhy et al. [4]). ES averages data over all levels greater than or equal to VaR, and so it tells us the average size of the survival time in excess of VaR. In practice, ES is simply obtained by calculating the sample mean of the simulated values above the corresponding VaR. The estimates of ES are displayed in Table 2. For example, in the case of the Clayton copula, which is chosen as the most appropriate copula for data, VaR (95%) and ES (95%) show that the survival times of a patient under treatment could extend to 6.1 and 8.8 years, respectively.

5. CONCLUDING REMARKS

Using Archimedean copulas, bivariate Weibull distributions were constructed. Selecting a copula that may best fit data is important in applications. In an application for multiple myeloma data, it was shown that the Clayton copula best fits the data among the copulas considered. Four different copulas with the different tail dependencies were used to determine this outcome. With extra computing costs, the goodness-of-fit testing procedures of the copula chosen were evaluated. The tail dependence was identified and explained graphically. Based on the bivariate Weibull distribution, we calculated value at risk, where attention is confined to the upper tail of the distribution, to examine the anticipated longest life expectancy of a patient.

6. REFERENCES

[1] D.G. Clayton, “A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence”. Biometrika, 65:141-151, 1978

[2] D. Collect, “Modelling Survival Data in Medical Research”. Chapman, 1999

[3] R.D. Cook and M.E. Johnson, “A family of distributions for modeling non-elliptically symmetric multivariate data”. Journal of the Royal Statistical Society, Series B, 43, 210-218, 1981

[4] M. Crouhy, D. Galai, and R. Mark, “Risk Management”. McGraw-Hill, 2001

[5] B.G. Durie and S.E. Salmon, “A clinical staging system for multiple myeloma. Correlation of measured myeloma cell mass with presenting clinical features, response to treatment, and survival”. Cancer, 36(3), 842-854, 1975

[6] V. Durrleman, A. Nikeghbail and T. Roncalli, “Which copula is the right one?”. Credit Lyonnais, Available at SSRN: http://ssm.com/abstract=1032545, 2000

[7] P. Embrechts, A. McNeil and D. Straumann, “Correlation: Pitfall and Alternative”. Risk, 12, 69-71, 1999

[8] M.J. Frank, “On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$”. Aequationes Mathematicae, 19, 194-226, 1979

[9] M.J. Frees and E. Valdez, “Understanding relationships using copulas”. North American Actuarial Journal, 2, 1-25, 1998

[10] C. Genest and R.J. MacKay, “Copules Archimediennes et Failles de Lois Bidimensionnelles Don’t les Marges Sont Donnees”, The Canadian Journal of Statistics, 14, 145-159, 1986a

[11] C. Genest and R.J. MacKay, “The joy of copulas: Bivariate distributions with uniform marginals”. American Statistician, 40, 280-283, 1986b

[12] C. Genest, and L. Rivest, "Statistical inference procedures for bivariate Archimedean copulas". Journal of American Statistical Association, 88, 1034-1043, 1993

[13] P.R. Greipp, J. San Miguel, B.G. Durie, J.J. Crowley, B. Barlogie, J. Bladé, M. Boccadoro, J.A. Child, H. Avet-Loiseau, R.A. Kyle, J.J. Lahuerta, H. Ludwig, G. Morgan, R. Powles, K. Shimizu, C. Shustik, P. Sonneveld, P. Tosi, I. Turesson, and J. Westin, "International staging system for multiple myeloma". Journal of Clinical Oncology, 23, 3412-3420, 2005

[14] E.J. Gumbel, "Bivariate exponential distributions". Journal of American Statistical Association, 55, 698-707, 1960

[15] P. Hougaard, "A class of multivariate failure time distributions". Biometrika, 73, 671-678, 1986

[16] H. Joe, "Multivariate Models and Dependence Concepts". Chapman & Hall, London, 1997

[17] P. Jorion, "Value at Risk: The New Benchmark for Managing Financial Risk". McGraw-Hill Publication, 2007

[18] J.M. Krall, V.A. Uthoff and J.B. Harley, "A step-up procedure for selecting variables associated with survival". Biometrics, 31, 49-51, 1975

[19] P. Kumar and M.M. Shoukri, "Evaluating Aortic Stenosis using the Archimedean copula methodology". Journal of Data Science, 6, 173-187, 2008

[20] R.A. Kyle and S.V. Rajkumar, Multiple myeloma. Blood, 111(6), 2962-2972, 2008

[21] S. Lee, E.-J. Lee and B.O. Omolo, "Using integrated weighted survival difference for the two-sample censored data problem". Computational Statistics and Data Analysis, 52, 4410-4416, 2008

[22] S. Lee and S. Yang, "Checking the censored two-sample accelerated life model using integrated cumulative hazard difference". Lifetime Data Analysis, 13, 371-380, 2007

[23] D.Y. Lin, L.J. Wei and Z. Ying, "Checking the cox model with cumulative sums of martingale-based residuals". Biometrika, 80, 557-72, 1993

[24] A. McNeil, R. Frey and P. Embrechts, "Quantitative Risk Management: Concepts, Techniques and Tools". Princeton University Press, 2005

[25] M.R. Melchiori, "Which Archimedean copula is the right one?". Yield Curve, 37, 1-20, 2003

[26] R.B. Nelsen, "An introduction to copulas". Springer, 1999

[27] L. Rivest and M. Wells, "A Martingale Approach to the Copula-Graphic Estimator for the Survival Function under Dependent Censoring". Journal of Multivariate Analysis, 79, 138-155, 2001

[28] A. Sklar, "Functions de repartition a n dimensions et leurs merges". Publication of the Institute of Statistics, University of Paris, 8, 229-231, 1959

[29] G. Venter, "Tails of copulas". Proceedings of the Astin Colloquium, 2001.

[30] M. Zheng and J. Klein, “Estimates of marginal survival for dependent competing risks based on an assumed copula”. Biometrika, 82, 127-138, 1995

Extended Fuzzy Hyperline Segment Neural Network for Fingerprint Recognition

M. H. Kondekar *m_h_kondekar@yahoo.com*
College of Computer Science and Information Technology
Latur, Maharshtra, India.

U. V. Kulkarni *kulkarniuv@yahoo.com*
Professor, Dept. of C.S. & Engg.,
SGGS Institute of Engg. & Tech.
Nanded,Maharastra, India.

B. B. M. Krishna Kanth *bbkkanth@yahoo.com*
Research Scholar,
S.R.T.M. University,
Nanded, Maharashtra, India.

Abstract

In this paper we have proposed Extended Fuzzy Hyperline Segment Neural Network (EFHLSNN) and its learning algorithm which is an extension of Fuzzy Hyperline Segment Neural Network (FHLSNN). The fuzzy set hyperline segment is an n-dimensional hyperline segment defined by two end points with a corresponding extended membership function. The fingerprint feature extraction process is based on FingerCode feature extraction technique. The performance of EFHLSNN is verified using POLY U HRF fingerprint database. The EFHLSNN is found superior compared to FHLSNN in generalization, training and recall time.

Keywords: Biometrics, Fuzzy Neural Network, Hyperline Segment, Fingerprint Recognition, FingerCode

1. INTRODUCTION

The increased emphasis on secrecy and protection of information in databases, personal identification has become very important topic in today's network society. Biometric indicators have an advantage over traditional security identification methods, because these inherent attributes cannot be easily stolen. There are many biometric features that are used for people identification, like iris, face, retina, voice, gait, palm print and fingerprint. The convenience of current electronic applications has led to an explosive increase in their use. E-banking, electronic fund transfer, online shopping and virtual auctions are just some applications prevalently used by the public [1].

Fingerprints are widely used as personal identification technique around the world [2] due to its distinguished features from others. Fingerprints are fully formed at about seven months of fetus development and finger ridge configurations do not change throughout the life of an individual except due to accidents such as bruises and cuts on the fingertips [3]. This property makes fingerprints a very attractive biometric identifier. Fingerprint features are permanent and fingerprints of an individual are unique.

The fingerprint recognition algorithms can be broadly classified into minutiae-based and FilterBank-based algorithms. The minutiae-based matching algorithms first extract the local

minutiae such as ridge endings and ridge bifurcations from the thinning image [4] or the gray scale image, and then match their relative placement in a given fingerprint with the stored template. A number of matching techniques are available in the literature including point-based matching [4] and graph-based matching [5]. Although the minutiae-based matching is widely used in fingerprint verification, but it has problems in efficiently matching two fingerprint images containing different number of unregistered minutiae points. Further, it does not utilize a significant portion of the rich discriminatory information available in the fingerprints.

The FilterBank-based algorithm [6, 7, 8] uses a bank of Gabor filters to capture both local and global information in a fingerprint as a compact fixed-length FingerCode, which is suitable for matching and storage. Thus, it overcomes some of the problems with the minutiae-based matching algorithms. So, here we have used FilterBank-based algorithm Jain et al [8] for efficient and correct fingerprint feature extraction.

In this paper, we have applied EFHLSNN classifier which is an extension of Fuzzy Hyperline Segment Neural Network (FHLSNN) proposed by Kulkarni et al [9] to the problem of fingerprint recognition based on FingerCode feature data. The FHLSNN utilizes fuzzy sets as pattern classes in which each fuzzy set is an union of fuzzy set hyperline segments. The fuzzy set hyperline segment is an n-dimensional hyperline segment defined by two endpoints with a corresponding extended membership function.

The rest of the paper is structured as follows. The feature extraction method in our work is introduced in Section 2. Sections 3 give a brief introduction for the architecture of the EFHLSNN, followed by its learning algorithm in section 4. Section 5 demonstrates the testing results and performance comparison of the classifiers on fingerprint and Iris Fisher data set. Conclusions are made in Section 6.

2. FINGERPRINT FEATURE EXTRACTION

In this paper fingerprint feature extraction is done by using Poly U HRF Fingerprint database images of 320*240 sizes at 1200 dpi resolution. The feature extraction process is based on FilterBank-based FingerCode feature extraction algorithm which consists of following stages.

2.1 Reference Point Location

Fingerprints have many visible landmark structures and a combination of them could be used for establishing a reference point. Jain, Prabhakar, Hong, and Pankanti [8] had defined the reference point of a fingerprint as the point of maximum curvature of the concave ridges in the fingerprint image. The location of reference point is mainly dependent on good quality of image, for graceful handling of local noise in a poor quality fingerprint image; the detection should necessarily consider a large neighborhood in the fingerprint image. For locating a reference point of a fingerprint local ridge orientation is usually specified for a block rather than at every pixel; an image is divided into a set of non overlapping blocks and a single orientation is defined for each block.

2.2 Filtering and FingerCode Feature Extraction

Fingerprints have local parallel ridges and valleys, and well defined local frequency and orientation. Properly tuned Gabor filters [10, 11] can remove noise, preserve the true ridge and valley structures, and provide information contained in a particular orientation in the image. Before filtering the fingerprint image, it is normalized to the region of interest in each sector separately to a constant mean and variance. Normalization is performed to remove the effects of sensor noise and gray level distortion due to finger pressure differences.

An even symmetric Gabor filter has the following general form in the spatial domain:

$$G(x, y; f, \theta) = \exp\left\{\frac{-1}{2}\left[\frac{x'^2}{\delta_{x'}^2} + \frac{y'^2}{\delta_{y'}^2}\right]\right\} \cos(2\pi f x') \quad (1)$$

$$x' = x\sin\theta + y\cos\theta \quad (2)$$

$$y' = x\cos\theta - y\sin\theta \quad (3)$$

Where f is the frequency of the sinusoidal plane wave along the direction θ from the x-axis, and $\delta_{x'}$ and $\delta_{y'}$ are the space constants of the Gaussian envelope along x' and y' axes, respectively. Jain et al [9] had performed the filtering in the spatial domain with a mask size of 33 X 33. In this algorithm they have used eight different values for θ $(0^0,\ 22.5^0,\ 45^0,\ 67.5^0, 90^0,\ 122.5^0,\ 135^0, \text{and } 157.5^0)$ with respect to the x-axis. The normalized region of interest in a fingerprint image is convolved with each of these eight filters to produce a set of eight filtered features. These eight directional-sensitive filters capture most of the global ridge directionality information as well as the local ridge characteristics present in a fingerprint. The mean of each sector in each of the eight filtered features defines the components of FingerCode feature vector. The gray level in a sector in a disk represents the feature value for that sector in the corresponding filtered image.

3. TOPOLOGY OF PROPESED EFHLSNN

The architecture of the EFHLSNN consists of four layers as shown in Figure 1. In this architecture first, second, third and fourth layer is denoted as F_A, F_B, F_D and F_C respectively. The F_A layer accepts an input pattern and consists of n processing elements, one for each dimension of the pattern. The F_B layer consists of m processing nodes that are constructed during training. There are two connections from each F_A to F_B node; one connection represents one end point for that dimension and the other connection represents another end point of that dimension, for a particular hyperline segment as shown in Figure 2.

Each F_B node represents hyperline segment fuzzy set and is characterized by the transfer function. In Let $R_h = (R_{h1}, R_{h2}, R_{h3}, \ldots\ldots, R_{hn})$ represents the h_{th} input pattern, $V_j = (v_{j1}, v_{j2}, \ldots\ldots, v_{jn})$ is one end of the hyperline segment e_j and $W_j = (w_{j1}, w_{j2}, \ldots, w_{jn})$ is the other end point of e_j. Then the membership function of j_{th} F_B node is defined as

$$e_j(R_h, V_j, W_j) = 1 - f(x, y, l) \quad (4)$$

Which $x = l_1 + l_2$ and the distance l_1, l_2 and l are defined as

$$l_1 = \left(\sum_{i=1}^{n} \left| w_{ji} - r_{hi} \right|\right), \quad (5)$$

$$l_2 = \left(\sum_{i=1}^{n} \left| v_{ji} - r_{hi} \right| \right), \tag{6}$$

$$l = \left(\sum_{i=1}^{n} \left| w_{ji} - v_{ji} \right| \right), \tag{7}$$

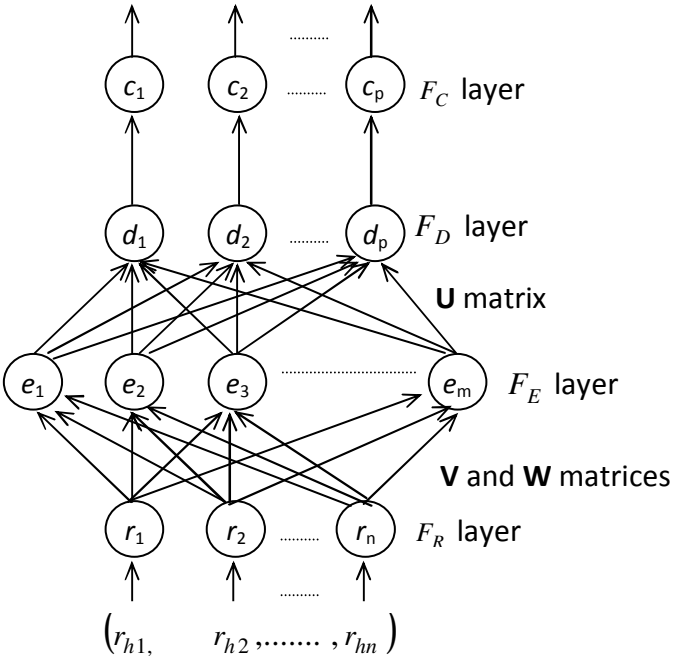


FIGURE 1: Extended Fuzzy hyperline segment neural network.

Here, in this paper we have used Manhattan distance for computing the values of l_1, l_2 and l as shown in equation (5), (6) and (*7).* The Manhattan distance has given best performance in terms of generalization, training and recall time in comparison with Euclidian distance [9] and distance between two position vectors as shown in equation (8), (9) and (10).

$$l_1 = \max \left(\sum_{i=1}^{n} \left| w_{ji} - r_{hi} \right| \right) \tag{8}$$

$$l_1 = \max \left(\sum_{i=1}^{n} \left| v_{ji} - r_{hi} \right| \right) \tag{9}$$

$$l_1 = \max \left(\sum_{i=1}^{n} \left| w_{ji} - v_{ji} \right| \right) \tag{10}$$

$f()$ is three parameter ramp threshold function defined as

$$f(x,\gamma,l)=0, \quad if \quad x = 1 \ otherwise$$

$$f(x,\gamma,l)=\begin{cases} x\gamma & if\ 0_j \le x\gamma \le 1 \\ 1 & if\ x\gamma > 1 \end{cases}$$

The fuzzy hyperline segment membership function for γ = 1, and with end points *w*=[0.5 0.3] and *v*=[0.5 0.7] is shown in Figure 3. This membership function returns highest membership value equal to one if the pattern $R_h$falls on the hyperline segment joined by two end points V_j and W_j. The membership value is governed by the sensitivity parameter γ, which regulates how fast the membership value decreases when the distance between R_h and e_j increases. For the given input pattern R_h, e_j 's output value is computed using equation (4).

Each node of F_C and F_D layer represents a class. The F_D layer gives soft decision and output of $k_{th}F_D$ node represents the degree to which the input pattern belongs to the class d_k. The weights assigned to the connections between F_E and F_D layers are binary values and stored in matrix *U*, and these values assigned to these connections are defined as

$$u_{jk}=\begin{cases} 1\ if \quad e_j is\ a\ hyperline\ of\ class\ d_k \\ 0 \quad otherwise \end{cases} \qquad (11)$$

for = 1, 2, ..., and =1, 2, ...,m.

Where e_j is the j_{th} F_E node and d_k is the k_{th} F_D node.

The transfer function of each F_D performs the union of appropriate (of same class) hyperline segment fuzzy values, which is described as

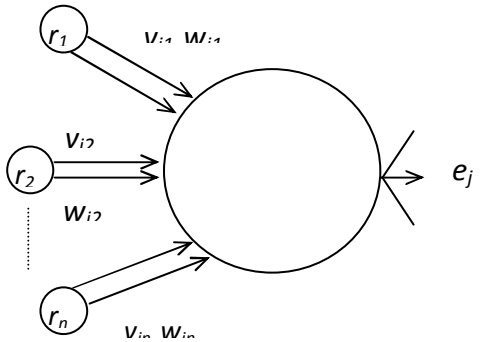


FIGURE 2: Implementation Extended Fuzzy Hyperline Segment

$$n_k = \max_{j=1}^{m} e_j u_{jk} \quad for\ k = 1,2,\ldots\ldots\ldots, p \qquad (12)$$

Each $F_C$node delivers non-fuzzy output, which is described as

$$C_k = \begin{cases} 0 \;\; if \;\; d_j > T \\ 1 \;\; if \;\; d_k == T \end{cases} \quad for\, T = \max(d_k) \quad for \;\; k = 1 \;\; to \;\; p. \tag{13}$$

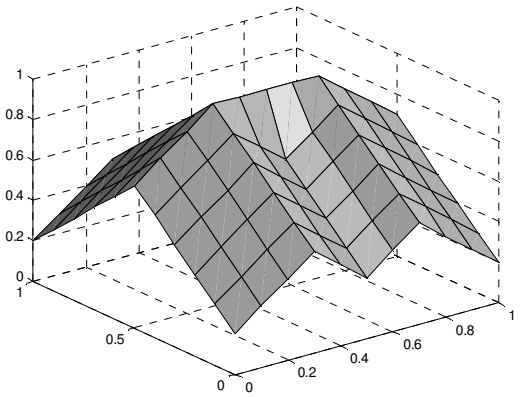


FIGURE 3: Extended Fuzzy Hyperline Segment membership function

4. EFHLSNN LEARNING ALGORITHM

The supervised FHLSNN learning algorithm for creating fuzzy hyperline segments in hyperspace consists of three steps, A: Creation of hyperline segments, B: Intersection test and C: Removing intersection. These steps are described below in detail.

4.1 Creation of Hyperline Segments

The length of hyperline segment is bounded by the parameter ζ, $0 \le \zeta \le \zeta_m$ and ζ_m depends on the dimension of feature vector. In the learning process appropriate values of ζ is selected and hyperline segment is extended only when the length of hyperline segment after extension is less than or equal to ζ. Assuming that the training set defined as $R \in \{R_h \mid h = 1,2,..., P\}$, the learning starts by applying the patterns one by one from the pattern set R. Given the h_{th} training pairs(R_h, d_h), find all the hyperline segments belonging to the class d_h. After this following four sub steps are carried out sequentially for possible inclusion of input patterns R_h.

Step 1: Determine whether the pattern $R_h$falls on any one of the hyperline segments. This can be verified by using fuzzy hyperline segment membership function described in equation (4). If R_h falls on any one of the hyperline segment then it is included, therefore in the training process all the remaining steps are skipped and training is continued with the next training pair.

Step 2: If the pattern R_h falls on any one of the hyperline passing through two end points of the hyperline segment, then extend the hyperline segment to include the pattern. Suppose e_j is that hyperline segment with end points V_j and W_j then $l1$, $l2$ *and* l are calculated using equations (5), (6) and (7). Where $l1$ is the distance of R_h from end pointW_j, $l2$ is the distance of R_h from end point V_j *and* l is the length of the hyperline segment.

2 (a): If $l1 > l2$ then test whether the point V_j falls on the hyperline segkent formed by the points W_j *and* R_h. This condition can be verified using equation (1) i.e. if $e_j(V_j, R_h, W_j) = 1$, then the

hyperline segment is extended by replacing end point V_j by R_h to include R_h, if extension criteria is satisfied. Hence

$$V_j^{new} = R_h \text{ and } W_j^{new} = W_j \quad (14)$$

2 (b): If $l2 > l1$ then test whether the point W_j, falls on the hyperline segment formed by the points V_j and R_h.If $e_j(V_j, R_h, W_j) = 1$, hyperline segment is extended by replacing end point W_j with R_h to include R_h, if extension criteria is satisfied. Hence

$$W_j^{new} = R_h \text{ and } V_j^{new} = V_j \quad (15)$$

Step 3: If hyperline segment is a point then extend it to include the pattern R_h, if extension criteria is satisfied as described by equation (11).

Step 4: If the pattern R_h is not included by any of the above sub-steps then new hyperline segment is created for that class, which is described as

$$W_{new} = R_h \text{ and } V_{new} = R_h \quad (16)$$

4.2 Intersection Test

The learning algorithm allows intersection of hyperline segments from the same class and eliminates the intersection between hyperline segments from separate classes. Intersection test is carried out as soon as the hyperline segment is extended either by sub-step 2 or sub-step 3 or created in sub-step 4.

Let $W_{ist} = [x_1, x_2, \ldots, x_n], V_{ist} = [y_1, y_2, \ldots, y_n]$represent two end points of extended or created hyperline segment and $W_n = [x'_1, x'_2, \ldots, x'_n], V_n = [y'_1, y'_2, \ldots, y'_n]$ are end points of the hyperline segment bf other class. First of all test whether the hyperlines passing through end points of two hyperline segments intersect. This is described by the following equations. The equation of hyperline passing through W_{ist} and V_{ist} is

$$\left[\frac{a_i - x_i}{y_i - x_i}\right] = r_1 \text{ for } i = 1,2, \ldots, n \quad (17)$$

and the equation of the hyperline passing through W_n and V_n is

$$\left[\frac{b_i - x'_i}{y_i - x'_i}\right] = r_2 \text{ for } i = 1,2, \ldots, n \quad (18)$$

where r_1, r_2 are the constant and a_i, b_ivariables. The equations (14) and (15) leads to set of n simultaneous equations which are described as

$$r_2(y_i - x_i) + x_i = r_2(y'_i - x'_i) + x'_i \text{ for } i = 1,2,\ldots,n \quad (19)$$

The values of r_1 and r_2 can be calculated by solving any two simultaneous equations. If remaining n-2 equations are satisfied with the calculated values of r_1 and r_2 thentwo hyperlines are intersecting and the point of intersection p_t is

$$P_c = (n_1(y_1 - x_1) + x_1, \ldots\ldots, n_i(y_n - x_n) + x_n) \quad (20)$$

4.3 Removing Intersection

If step 2(a) and step 3 has created intersection of hyperline segments from separate classes then intersection is removed by restoring the end point V_j as $V_j^{new} = V_j$, if sub-step 2(b) has created intersection then intersection is removed by restoring the end point W_j as $W_j^{new} = W_j$, and new hyperline segment is created to include the input pattern A_h, which is described by equation (13).

If the sub-step 4 creates intersection then it is removed by restoring the end points of previous hyperline segment of other class.

$$W_{new+1} = V_{new+1} = V_n \text{ and } V_n = W_n \quad (21)$$

5. SIMULATION RESULTS AND PERFORMANCE COMPARISON

The EFHLSNN is implemented using MATLAB 7.0. The results are obtained and compared with FHLSNN using fingerprint feature data set. The timing analysis of training and recall are depicted in Table 1. Table 2 gives performance comparison using recognition rates along with number of hyperline segments created.

Classifier	Training in seconds	Testing
FHLSNN using Euclidian Distance	0.1648	0.811566
EFHLSNN using Manhattan distance	0.1631	0.743732
EFHLSNN using Distance between two position vectors	0.1806	0.966453

TABLE 1: Timing Analysis with fingerprint data features

Classifier	Recognition Rate	Theta	Hyperline Segments
FHLSNN	100 %	1.4	200
EFHLSNN using Manhattan distance	100 %	0.2	259
EFHLSNN using Distance between two position vectors	100 %	0.2	259

TABLE 2: Percentage Recognition Rate with FHLSNN and EFHLSNN

As shown in Table 1 the training and testing time using EFHLSNN classifier takes less time compared to FHLSNN classifier using Euclidian distance and Distance between two position vectors.

The EFHLSNN classifier is also applied on standard Fisher Iris Database which also takes less time compared to FHLSNN classifier as depicted in Table 3.

Classifier	Training in seconds	Testing
FHLSNN using Euclidian Distance	0.5178	0.811566
EFHLSNN using Manhattan distance	0.4390	0.738011

TABLE 3: Timing Analysis with Fisher Iris dataset

Hence, the EFHLSNN classifier gives better recognition rates in comparison with FHLSNN in terms of less training and recall time along with 100 % recognition rate.

5. CONCLUSION

The EFHLSNN classifier using Manhattan distance has ability to train and recall patterns faster than FHLSNN classifier using Euclidian distance and Distance between two position vectors. Thus it can be used in real time applications for recognition purpose where less training and recall time is the prime demand. Generalization, training and recall time is also verified using Fisher Iris dataset , where almost similar performance is observed.

REFERENCES

[1] P.Meenen, and R.Adhami, "Fingerprinting for security", Proceedings of the IEEE Potentials, pp. 33-38, 2001.

[2] T. Song, Liang Huang, Che-Wei Liu, Jui-Peng Lin, Chien-Ying Li, and Ting-Yi Kuo, "A Novel Scheme for Fingerprint Identification", Proceedings of the Second Canadian Conference on Computer and Robot Vision (CRV'05) 0-7695-2319-6/05.

[3] W. J. Babler, "Embryologic Development of Epidermal Ridges and Their Configuration", Birth Defects Original Article Series, vol. 27, no. 2,1991.

[4] A.K. Jain, L. Hong, S. Pankanti, and R. Bolle, "An identity authentication system using fingerprints", Proceedings of the IEEE, Vol. 85, No. 9, pp. 1365-1388, 1997.

[5] A. K. Hrechak and I. A. Mchugh, "Automated fingerprint recognition using structural matching", Pattern Recognition, vol. 23, no. 8, pp. 893-904, 1990.

[6] A. K. Jain, S. Prabhakar, and L. Hong, "A multichannel approach to Fingerprint Classification", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 4, pp. 348-359, 1999.

[7] K. Jain, S. Prabhakar, L. Hong, and S, Pankanti, "FingerCode: a FilterBank for fingerprint representation and matching", Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR), vol. 2, pp. 187- 193, 1999.

[8] A. K. Jain, S. Prabhakar, L. Hong, and S. Pankanti, "FilterBank-based fingerprint matching", IEEE Transactions on Image Processing, vol. 9, no. 5, pp. 846-859, 2000.

[9] U.V. Kulkarni, T.R. Sontakke and G. D. Randale, "Fuzzy Hyperline Segment Neural Network for Rotation Invariant Handwritten Character Recognition", Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on Neural Network, Washington, DC, USA, 2001, pp. 2918 - 2923 vol.4.

[10] J. G. Daugman, "High confidence recognition of persons by a test of statistical independence", IEEE Transactions on Pattern Analysis and Machine Intelligance, vol. 15, no. 11, pp. 1148-1161, 1993.

[11] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters", J.Opt. Soc. Amer. A. vol. 2, pp. 1160-1169, 1985.

Electromyography Analysis for Person Identification

Suresh.M *sureshm_ap@yahoo.co.in*
Electronics and communication
Kalpataru Institute of Technology
Tiptur, 572202, India

Krishnamohan.P.G *pgkmohan@yahoo.com*
Electronics and communication
Jawaharlal Nehru Technological University
Hyderabad, 500085, India

Mallikarjun S Holi *msholi@yahoo.com*
Department of Biomedical Engineering
Bapuji Institute of Engineering and Technology
Davangere, 577004, India

Abstract

Physiological descriptions of the electromyography signal and other literature say that when we make a motion, the motor neurons of respective muscle get activated and all the innervated motor units in that zone produce motor unit action potential. These motor unit action potentials travel through the muscle fibers with conduction velocity and superimposed signal gets recorded at the electrode site. Here we have taken an analogy from the speech production system model as the excitation signal travels through vocal tract to produce speech; similarly, an impulse train of firing rate frequency goes through the system with impulse response of motor unit action potentials and travels along the muscle fiber of that person. As the vocal tract contains the speaker information, we can also separate the muscle fiber pattern part and motor unit discharge pattern through proper selection of features and its classification to identify the respective person. Cepstral and non uniform filter bank features models the variation in the spectrum of the signals. Vector quantization and Gaussian mixture model are the two techniques of pattern matching have been applied.

Keywords: Biometrics, Electromyogram, Gaussian mixture model (GMM), Identification, Vector Quantization.

1. INTRODUCTION

The EMG signal is the summation of the discharges of all the motor units within the pick-up range of the electrode. The nervous system always controls the muscle activity (contraction/relaxation). Hence, EMG signal is a complicated signal, which is controlled by the nervous system and is dependent on the anatomical and physiological properties of muscles. It is the study of muscle electrical signals. Muscle tissue conducts electrical potentials and the name given to these electrical signals is the muscle action potential. Surface EMG is a method of recording the information present in these muscle action potentials. When EMG is acquired from electrodes mounted directly on the skin, the signal is a composite of all the muscle fiber action potentials occurring in the muscles underlying the skin. These action potentials occur at random intervals. So at any one moment, the EMG signal may be either positive or negative voltage. Surface EMG signal consists of two major components. The First component is the firing frequency of the Motor Units. In the frequency domain, this component contributes spectral peaks which coincide with, and their quality depends on, the firing statistics of the individual MUs (Motor units). The spectral peaks are substantial at the mean frequency of the MUs firing frequencies, below 40 Hz, and have diminishing harmonics at the relatively higher frequencies [1], [2]. The second component is the resulting frequency spectra of the MUAP (motor unit action potential) shapes. This component

is influenced heavily by the recording arrangement, type of electrode, distance between the electrodes, and the distance to the recorded fibers, as well as fiber distribution inside the muscle(motor unit recruitments), fatigue, etc. [3], [4], [5], [6]. In surface recording due to the filtering through skin layers and of the recording arrangements [3], Surface EMG signal reaches only into the few hundred Hz regions. Surface EMG, is the result of the surface recording of the superposition of the many MUs at the same time. The interference pattern generated by such a recording does not exactly retain the shape of the individual MU's. But its spectral content depicts the information about MUAP shape and other characteristics.

Human ECG has unique wave shape, amplitude, due to anatomical structure of the heart and physiological Conditions [7].many researchers used ECG for person identification and verification [8]. [9]. [10]. In the past ten years, several studies have been proposed using brain waves, e.g. EEG as a biometric modality. [11]. [12]. [13]. [14]. [15]. [16]. [17]. [18].

We focus on biometric recognition task of authentication of person based on Electromyogram (EMG) signal as a biological trait. To design the more resembling devices one needs to extract useful information from EMG signal in order to generate enough discrimination among different persons. Since direct connection between intact muscle, intact central nervous system and brain is unique to an individual, EMG signals vary from individual to individual. The EMG signal is directly related to the physiology of each individual. These measurements are influenced by physiologic factors which include muscle fiber pattern, motor unit discharge pattern, changes in blood flow in the muscle, force generating capacity of each muscle, neural activity, and neurotransmitter activity in different areas within the muscle, skin conductivity, position, shape and size of the muscle. The EMG signals have different signatures depending on age, muscle development, motor unit paths, different density of bone, heat distribution of the muscle, skin-fat layer, and gesture style. The external appearances of two people gestures might look identical, but the characteristic of EMG signals are different. Regardless of what factors originate differences in the measurement, the fact that the EMG contains physiologic dependant singularities potentiates its application to person identification.

The goals of this work were to (1) build a state-of-the-art person identification system based on myoelectric signals, (2) to address major issues in the novel technology that have not yet been addressed in literature and (3) to demonstrate the practicability of EMG based person identification. One important goal of this work was to explore appropriate feature extraction and classification methods in order to develop state-of-the-art EMG based person identification system that achieves recognition results comparable to those that have so far been reported in literature.

2 ANATOMICAL AND PHYSIOLOGICAL BASICS OF SURFACE ELECTROMYOGRAPHY (SEMG)

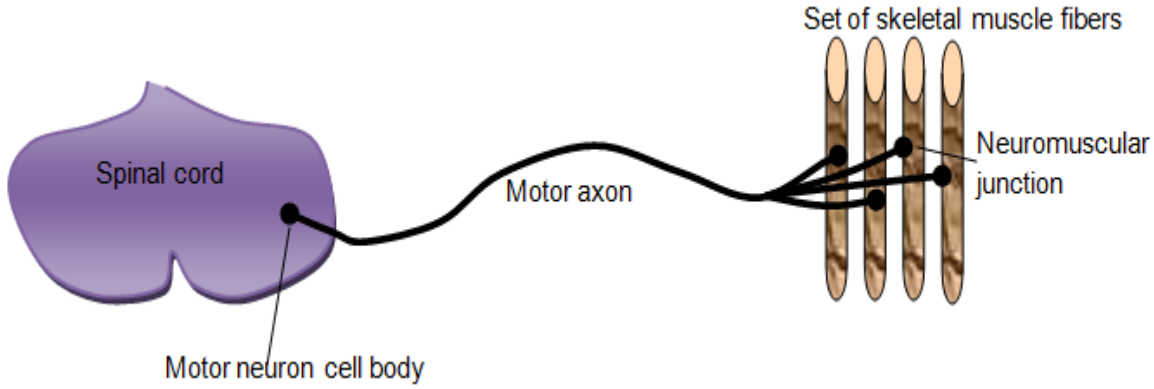
Skeletal muscle is the muscle attached to the skeleton. Its structural unit is the muscle fiber - an elongated cell ranging from 10 to 100 microns in diameter and from a few millimeters to 40cm in length [19]. Each muscle fiber is surrounded by a layer of connective tissue called endomysium. Groups of muscle fibers are wrapped by another layer of connective tissue called perimysium to form muscle bundles or fascicles. Skeletal muscles are composed of numerous fascicles. They are usually attached to bones via tendons composed of epimysium.

The contraction of skeletal muscle is controlled by the nervous system. Each muscle fiber can be Activated by one motor neuron (i.e. by one nerve) yet, one motor neuron can branch in up to several thousand branches, each one terminating in a different muscle fiber. A motor neuron and all the fibers it innervates are called a motor unit shown in fig.1. the term neuromuscular junction refers to the junction between a muscle fiber and the terminal of the motor neuron it is recruited by an ionic equilibrium between the inner and outer spaces of a muscle cell forms a resting potential at the muscle fiber membrane (approximately -80 to -90 mV when not contracted). The activation of an alpha-motor anterior horn cell (induced by the central nervous system or reflex)

results in the conduction of a excitation along the motor nerve. After the release of transmitter substances at the motor endplates, an endplate potential is formed at the muscle fiber innervated by this motor unit. The diffusion characteristics of the muscle fiber membrane are briefly modified and Na+ ions flow in. This causes a membrane Depolarization which is immediately restored by backward exchange of ions within the active ion pump mechanism, the Repolarization. If a certain threshold level is exceeded within the Na+ influx, the depolarization of the membrane causes an Action potential to quickly change from – 80 mV up to 30 mV. Mono-polar electrical burst is immediately restored by the repolarization phase and followed by an after Hyper polarization period of the membrane. Starting from the motor end plates, the action potential spreads along the muscle fiber in both directions and inside the muscle fiber through a tubular system.

FIGURE 1: Motor Unit

3 EQUIPMENT FOR SURFACE EMG RECORDING



To develop EMG data acquisition system we employed an instrumentation amplifier INA 101 to reject common mode signals. A single order high pass filter of 10 Hz is used in the feedback of instrumentation amplifier output stage to prevent the saturation of data due to Base line noise and motion Artifact. Then to remove other low frequency noise such as ECG and other biomedical signal interference, sixth order unity gain Chebyshev Active High Pass Filter (20hz) and to remove high frequency thermal noise a fourth order Butterworth Active Low pass filter is employed (600hz).to overcome the problem with non-ideal operation of op-amps at high frequency one single order passive low pass RC filter is also employed. To remove line frequency components Notch filter is implemented using an active filter module UAF042 from Texas instruments is used. We acquired the filtered and Amplified EMG data through Line-In port of sound card through MATLAB.

4 PERSON SPECIFIC INFORMATION OF EMG SIGNAL

The EMG signal is very rich in information as it contains task specific, site specific, muscle specific, subject specific and person specific information Traditionally, skeletal muscle were classified based on their color as red fiber and white fiber. Also, depending upon twitch capabilities, fibers are classified as fast twitch and slow twitch fibers. These two main categories of fibers become three when the white fiber is split into two sections as Type-I(Red fibers, - slow oxidative or slow twitch or fatigue resistant) fiber characterized with slow ATP splitting rate, Type-IIa (Red fibers, also called fast twitch A – fast oxidative) fiber characterized with high ATP splitting rate and Type – II b (White fiber –Fast Glycolytic , also, called fast twitch B or fatigable) fiber characterized with low myoglobin content, few mitochondria, few blood capillaries, large amount of glycogen, very high rate of ATP splitting and fatigue easily compared to Type-I and Type-IIa fibers. Fiber types, force generating capacity of muscle, Motor unit discharge pattern, muscle fiber pattern vary considerably from muscle to muscle, and person to person and thus we

have strong evidence that the EMG signal contains person specific information and is the objective of our work.

Motor unit discharge pattern is the source feature and the muscle fiber pattern is the system feature. By separating source and system feature similar to speaker identification using speech we can use EMG for person identification. However, appropriate EMG driven dynamic model is needed to be developed depending upon what information we need.

Any biologically assisted models commonly employ EMG as a means to monitor muscle activity. E.W.Theodo et all have developed an EMG assisted dynamic model which is unique in that it is person specific in terms of : 1) Anthrophometry (Muscle location and size), 2) Body mass characteristics, 3) Subject motion (inputs trunk as well as limb motion) and 4) Muscle activities[20]. Various EMG driven models to extract muscle specific information from EMG are discussed in [21]. Specific temporal change in the pattern of firing frequency that is depicted by spectral properties will be the model of person's muscle structure and this motivates to use EMG for person identification.

5 FEATURE EXTRACTION AND MODELING METHODS

5-1 Cepstral Features

Consider a given EMG waveform for a particular person. EMG frames of size 50 ms (100 samples for 2 kHz Signal) with a shift of 25 ms (50 samples for 2 kHz Signal) are taken. Therefore 400 frames for every ten second slot of EMG data obtained. In each session we have collected five slots of ten seconds each EMG data from all subjects. So for three sessions 6000 frames obtained.

5.2 Non Uniform Bank Features

Non uniform filter-bank: In this method, the signal spectrum will pass through a filter-bank set. The usage of these filter banks are motivated by the fact that, the EMG spectrum has some special shapes and are distributed by a non-linear scale in frequency domain. Using the filter-banks with spectral characteristics which are well-matched to those of the desired signal, the contribution of noise components in the frequency domain can be reduced. EMG spectrum is concentrated within the range of 20Hz-500Hz. In such a narrow band, full resolution is required to capture more information of the signal spectrum. In this work, EMG spectrum is simply processed by filtering out the frequency band outside the range of 20Hz-500Hz. non uniform filter bank extracts the corresponding bands of the input signal and has been widely used due to its consistent better performance.

5.3 Vector Quantization Modeling

The basic idea of using vector quantization is to compress a large number of cepstral vectors into a small set of code vectors. The VQ codebook is usually trained with the LBG algorithm to minimize the quantization error when replacing all feature vectors with their corresponding nearest code vectors. The Euclidean distance is often used as a quantization error measure.

5.4 Gaussians Mixture Modeling

GMMs are employed in various fields such as clustering and classification, but unlike k-means, they are able to build soft clustering boundaries i.e., points in space can belong to any class with a given probability and have the ability to represent smooth approximations for general probability density functions through the weighted sum of a finite number of Gaussian densities. The standard method used to fit a GMM to observe data is the expectation maximization (EM) algorithm [22], which converges to a maximum likelihood (ML) estimate. Plus several model order selection used to estimate the number of components of a GMM. GMMs are used to model the likelihoods of the features extracted from the EMG signal. GMMs are well-known flexible modeling tools able to approximate any probability density function.

6 EXPERIMENTAL RESULTS AND DISCUSSION

The EMG data is collected in three sessions with time gap of one day for a population of 49 (30 healthy male plus 19 female) subjects by placing the electrode at the same location on the Flexor carpi ulnaris muscle present in the forearm for all sessions. In each session, five slots of ten second each EMG data per subject are acquired. For EMG frame of 50 ms with an overlap of 25 ms 40 frames per second and 400 frames per 10 second slot are obtained. We have used first four slots (1600 frames) of all the subjects of randomly selected any two sessions data (that is 3200 frames) for training using vector quantization and Gaussian mixture model. Last fifth slot of data (400 frames) in all three sessions of 49 subjects was used for testing individually for different code book size in case of VQ and number of Gaussians in case of GMM. The following configurations using different combinations of features tested are

1) 39 dimensional base cepstra with VQ
2) 13 dimensional base cepstra plus 13 delta 13 delta with VQ
3) 13 dimensional base non uniform filter bank plus 13 delta and13 acceleration coefficients with VQ
4) 13 dimensional base non uniform filter bank plus 13 delta and 13 acceleration coefficients with GMM

A result of person identification using different combinations of features tested with VQ and GMM are tabulated in Table 1-4. As it can be seen in the table 1, we have explored that the VQ-based system achieved the best result of 93.8776% using the configuration with 39 dimensional base cepstra. Table 2 shows that, for the VQ-based system with 13 dimensional base cepstra plus 13 delta and 13 acceleration coefficients, the best score was 85.7143%, also using same set of features. Table 2-3 show that, In both cepstral based system and non uniform filter bank based system with VQ, the non uniform filter bank based system give higher person identification performance than cepstral based system. The reason is, in non uniform filter bank full resolution can be achieved to capture the more information present in the EMG signal. Based on the information present in EMG signal the non uniform filters bank extracts more hidden information. As it can be seen in the Table 4, GMM-based system achieved the best result of 97.9592% using 13 dimensional base non uniform filter bank plus 13 delta and 13 acceleration coefficients. For the VQ-based system with 39 dimensional non uniform filter banks, the best score was 93.8776%, also using same set of features. Further non uniform filter bank with GMM is exhibiting more effectiveness than the non uniform filter bank with VQ, because one of the powerful attributes of GMM is its ability to form smooth approximations to arbitrarily shaped densities and the individual component densities may model some underlying set of hidden classes. GMM also captures uncertainty in cluster assignments. In our experiment we have determined the optimal choice of code book size as 16, 32, 64 with VQ and Gaussians size as 2, 4, 8, 16, and 32 with GMM. For smaller number of Gaussian components in GMM good person identification performance obtained. This suggests that the non uniform filter bank coefficients with GMM are useful parametric measures from a biometric perspective and that they can be used to identify subjects.

EMG based person identification system using 39 dimensional base Cestrum-VQ approach provides average person identification performance of about 65.3050%, 13 dimensional base cepstra plus 13 delta 13 delta-VQ approach of about 58.9569%, filter bank - VQ approach of about 72.1088% including all untrained sessions of all code book sizes and Filter bank-GMM approach of about 73.3333% including all untrained sessions of all Gaussians mixtures. The results indicate that the EMG has significant biometric potential.

Training sessions	Code book size	Testing session		
		1	2	3
1,2	16	57.1429	93.8776	71.4286
	32	59.1837	93.8776	69.3878
	64	55.1020	91.8367	65.3061
2,3	16	57.1429	75.5102	91.8367
	32	53.0612	71.4286	87.7551
	64	51.0204	67.3469	89.7959
1,3	16	59.1837	77.55	91.8367
	32	55.1020	71.4286	87.7551
	64	51.204	71.4286	89.7959

TABLE 1: Results of Person Identification using 39 Dimensional Base Cepstra with VQ

Training sessions	Code book size	Testing session		
		1	2	3
1,2	16	55.1020	77.5510	63.2653
	32	59.1837	83.6735	61.2245
	64	55.1020	79.5918	57.1429
2,3	16	44.8980	69.3878	85.7143
	32	46.9388	69.3878	83.6735
	64	44.8980	69.3878	81.6327
1,3	16	44.8980	69.3878	85.7143
	32	46.9388	71.4286	83.6735
	64	44.8980	71.4286	81.6327

TABLE 2: Results of Person Identification using 13 Dimensional Base Cepstra plus 13 Delta 13 Delta with VQ

Training sessions	Code book size	Testing session		
		1	2	3
1,2	16	67.3469	93.8776	67.3469
	32	69.3878	93.8776	69.3878
	64	69.3878	93.8776	71.4286
2,3	16	63.2653	81.6327	91.8367
	32	63.2653	81.6327	91.8367
	64	67.3469	79.5918	91.8367
1,3	16	63.2653	83.6735	91.8367
	32	63.2653	81.6327	91.8367
	64	67.3469	81.6327	91.8367

TABLE 3: Results of Person Identification using 13 Dimensional Base Non uniform Filter Bank plus 13 Delta 13Delta with VQ

Training sessions	Number of Gaussians	Testing session		
		1	2	3
1,2	2	63.2653	97.9592	73.4694
	4	69.3878	97.9592	73.4694
	8	71.4286	95.9184	75.5102
	16	69.3878	97.9592	75.5102
	32	69.3878	97.9592	73.4694
2,3	2	57.1429	81.6300	91.8367
	4	63.2653	79.5918	97.9592
	8	65.3061	83.6735	97.9592
	16	65.3061	83.6735	97.9592
	32	65.3061	85.7143	95.9184
1,3	2	57.1429	81.6327	91.8367
	4	65.3061	77.5510	97.9592
	8	65.3061	83.6735	97.9592
	16	63.2653	83.6735	97.9592
	32	65.3061	85.7143	95.9184

TABLE 4: Results of Person Identification using 13 Dimensional Base Non uniform Filter Bank plus 13 Delta 13Delta with GMM

6 CONCLUSION

The proposed cepstral features, non uniform filter bank features and the associated modeling techniques have been shown to be suitable for the EMG in the human identification task, yielding a relatively good result in the experimental evaluation.

The results obtained in the present work corroborate the long existing line of research showing evidence that EMG carrying individual-specific information which can be successfully exploited for purpose of person identification.

7 REFERENCES

[1]. C. N. Christakos, "A population stochastic model of skeletal muscle and its use in the study of frequency characteristics of the muscle output activity with particular reference to tremor," Ph.D. dissertation, Chelsea Univ. College, London, England, 1980.

[2]. P.Lago and N.Jones, “Effect of motor unit firing statistics on EMG spectra” Med. Biol Eng. Comput., Vol.15, pp. 648-655, 1977.

[3]. L. Lindstrom and R. Magnusson, "Interpretation of myoelectric power spectra. A model and its applications” Proc. IEEE, vol. 64, pp. 653-662, 1977.

[4]. G. Agarwal and G. Gottlieb "An analysis of the electromyogram by Fourier simulation and experimental results” IEEE Trans Biomed. Eng., vol. BME-22, pp. 225-229, May 1975.

[5]. R. Le Fever and C. De Luca, "The contribution of individual motor units to the EMG power spectrum” in Proc. 29th Annu. Conf Eng., Med. Biol., Vol. 18, 1976, p.91.

[6]. G. Inbar, J. Allin, E. Golos, W. Koehler, and H. Kranz,"EMG spectral shift with muscle length, tension, and fatigue" in Proc. IEEE Melecon Conf., 1981.

[7]. N. S. Yogendra, and G. Phalguni, “ECG to Individual Identification”,IEEE Second International Conf on Biometrics: Theory, Applications and Systems (BTAS 2008), Washington DC, USA.

[8]. W. Yongjin, A. Foteini, H. Dimitrios, and N. P Konstatantinos,"Analysis of Human Electrocardiogram for Biometric Recognaition,"EURASIP Journal of advance in signal processing, 2008.doi:10.1155/2008/148658.

[9]. D. C. Chan, M. M Hamdy, A. Badre and V. Badee, "Person Identification using Electrocardiograms" CCECE apos;06. Canadian Conference on May 2006 pp:1 – 4.Y. Wang, K. N. Plataniotis, and D. Hatzinakos, "Integrating analytic and appearance attributes for human identification from ECG signal,"Proc of Biometrics symposiums, September 2006.

[10]. T. W. Shen, W. J. Tompkins, Y. H. Hu, "One-lead ECG for Identity Verification, Proc IEEE EMBS/BMES conf,62-63,2002.J. M. Irvine, K. B. Wiederhold, L. W. Gavshon, S. Israel, S. B. McGehee, R. Meyer, M. D. Wiederhold, "Heart rate variability: A new biometric for human identification", Proc International Conference on Artificial Intellegent, 1106-1111,2001.

[11]. D. Martin, D. Lodrova, "Liveness detection for biometric systems based on papillary lines," Int. J. Security and Its Applications, vol. 2, no. 4, Oct. 2008.

[12]. F. Su, L.-W. Xia, A. Cai, J.-S. Ma, Y.-B. Wu, "EEG-based personal identification: from proof-of-concept to a practical system," Proc. Int. Conf. Pattern Recognition, Istanbul, Turkey, Aug., 2010.

[13]. C. R. Hema, M. P. Paulraj, K. Harkirenjit, "Brain signatures: a modality for biometric authentication," Int. Conf. Electronic Design, 2008, pp. 1– 4.

[14]. N. Markus, W. Marc, S. Johannes, "Test–retest reliability of resting EEG spectra validates a statistical signature of persons," Clinical Neurophysiology, vol. 118, 2007, pp. 2519–2524.

[15]. M. Chisei, B. Sadanao, N. Isao, "Biometric person authentication using new spectral features of electroencephalogram (EEG)," Int. Symp. Intelligent Signal Processing and Communication Systems, 2008.

[16]. S. Sun, "Multitask learning for EEG-based biometrics," Int. Conf. Pattern Recognition, 2008, pp. 1–4.H.

[17]. Chen, X.-D. Lv, Z. J. Wang, "Hashing the mAR coefficients from EEG data for person authentication," IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2009, pp. 1445–1448.

[18]. A. Yazdani, A. Roodaki, S. H. Rezatofighi, K. Misaghian, S. K. Setarehdan, "Fisher linear discriminant based person identification using visual evoked potentials," IEEE Int. Conf. Signal Processing, 2008, pp. 1677–1680.

[19]. Recording techniques. In Selected Topics in Surface Electromyography for Use in the Occupational Setting: Expert Perspective. U.S. Department of Health and Human Services. DHHS (NIOSH) Publication No 91-100.

[20]. E.W.Theado, G.G.Knapik and W.S.Marras, "*Modification of an EMG –assisted biomechanical model for pulling and pushing*" in International Journal of Industrial Ergonomics, 37(2007)825-831.

[21]. David G.Loyd, Thor F.Beiser," *An EMG driven Musculoskeletal Model to estimate muscle forces and Knee joints moments in vivo"* Journal of Biomechanics 36(2003),765-776.

[22]. A. P. Dempster, N. M Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM algorithm," J. Roy. Statist. Soc. B, vol. 39, pp. 1-38, 1977.

Future Path Way to Biometric

C.N .Ravi Kumar *kumarcnr@yahoo.com*
Professor & Head of Department
Dept of Computer Science & Engg
S. J. College of Engineering
Mysore, India.

P. Girish Chandra *p.girishchandra@gmail.com*
Dept of Computer Science & Engg
S. J. College of Engineering
Mysore, India.

R. Narayana *narayanark17@gmail.com*
Dept of Computer Science & Engg
S. J. College of Engineering
Mysore, India.

Abstract

The nature of Pattern Recognition employing Biometric system causes the fact that not only the verification & identification system becomes practical but also enables recognition of the character and serves as a tool in diagnosing the disease of a person. One of the main interests for Biometric System is identification of person's psychological traits and personality types which can be accomplished by classifying the Biometric into different levels. Existing Biometric system can identify or verify the person but cannot declare the personality of a Person. Face and Hand provides Researchers and Psychologists with instrument of obtaining information about personality and psychological traits. Initially the paper describes the Different level of Biometric and need for classification. Later the paper mentions some of technologies to explore the entire field of personality through the skillful use of the computer data processing and biometric scanning. Applications of Personality Analysis include, but are not limited to, for use in daily living, private industry, civil service, education, law enforcement, military, medicine and psychology, throughout all aspects of human life.

Keywords: Personality, Biometric Levels, Gunas, Physiognomy, Psychometrics, Phase Facial Portrait, Ophthalmogeometry.

1. INTRODUCTION

In the field of biological sciences the terms "Biometrics" and "Biometry" are used from the beginning of the 20th century for the development of statistical and mathematical methods applicable to the problem of data analysis. Earlier the term "Biometrics" was used for Statistical methods for the analysis of data from agricultural field experiments to compare the yields of different varieties of wheat, for the analysis of data from human clinical trials evaluating the relative effectiveness of competing therapies for diseases, or for the analysis of data from environmental studies on the effects of air or water pollution on the appearance of human disease in a geographical area. Of late the terms are used to refer to identification technologies [10].

Each biometric characteristic has its own advantages and disadvantages, A number of biometric characteristics has been proposed for authentication purposes. Traditionally, they can be categorized into two major groups: physical or behavioral characteristics. Examples of physical characteristics include: DNA, Ear, Face, Fingerprint, Hand Geometry, Iris, and Retina. Behavioral characteristics include: Gait, Signature, and Voice. A summary of those characteristics can be found in [1] to [11].

Currently the hardware and software gadgets dedicated to human biometrics is mainly used for identification and verification purposes. This requires recording in some fashion an image of the subject that can later be utilized as a template to identify another image of the same part of that subject. These technologies are used to identify or authenticate persons based upon their physical characteristics stored as graphical information and/or digital data and templates about a person for identification or verification purposes [10][12].The application of biometrics can be found in the Medical, Convenience, and Security Biometrics.

2. HISTORICAL LAPSES AND CONTEMPORARY PERSPECTIVE IN BIOMETRICS

Till date the Biometric System was able to solve the significant problems in the field of identification & verification. How ever the research works taken up in these fields cannot determine the complete personality of a person using present Biometric system. The so many varied types of biometrics and also the depth that can be realized in each Biometric approach have motivated us to see the entire stretch of Biometric under various categories. The awareness of stretch has lead to the realization of the need for classification.

Looking at the vast potential application of biometrics, it is advisable to group the Biometrics based upon the purpose for which it is used. In the present context we feel the identification and verification using various Biometric systems can be classified as Biometric-I.

It is observed from Vedic time which is as old as 2000 BC the characters of an individual can be identified /read by analyzing the impressions of various human parts like Face, Hand, Ear ,etc. This analysis can be grouped as Biometric-II

Of late, a lot of research work is observed to identify various diseases that a person has by observing various parts of the body which can be collectively categorized as Biometric-III.

3. BIOMETRIC –II : JUDGING HUMAN CHARACTERS

Albert Einstein The great physicist said, "Time is the fourth Dimension". Vedas say that "Time is the first dimension". They say "in the beginning there was nothing"[17].Yes indeed Vedic literature contains knowledge about all fields. According to Vedas all material elements are infused with the modes of nature or gunas- satva, rajas, and tamas [18]. The description of human being in term of nine lotus petals and the three gunas or qualities occurs repeatedly in vedantic literature. Its first appearance is in atharva Veda. The summary of the same is as follows: “There is the nine-portalled lotus covered under three bands, in which Lives the Spirit with the atman within that the Veda-Knowers Know”. The three bands are three gunas or psychological qualities well known as satva , rajas ,and tamas , balanced dynamism , uncontrolled activity, inertia[19].

The each three gunas includes attributes in it. Satva guna is characterized by qualities such as cleanliness, truthfulness, gravity, dutifulness, detachment, discipline, mental equilibrium, respect for superiors, contentment, sharp intelligence, sense control, and staunch determination. Attributes of rajas guna include intense activity, desire for sense gratification, little interest in spiritual elevation, dissatisfaction with one’s position, envy of others, and a materialistic mentality. Qualities associated with tamas guna include mental imbalance, anger, ignorance, arrogance, depression, laziness, procrastination, and a feeling of helplessness [18]. The more explanation of gunas can be found in [18] [19].

The above mentioned gunas are refered in Big Five model ,the most popular model of psychological traits is a five-dimension personality model, as Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism .In contemporary psychology, the "Big Five" factors of personality are five broad domains or dimensions of personality which are used to describe human personality [46].

3.1. Methods for Characteristics Recognition

It is well known that "you can't manage what you don't measure", so measuring characteristic is also important. To measure human characteristics or identify personality types psychological researchers apply psychometric which is the field of study including the theory and technique of psychological measurement. Psychometric involves such research tasks as the construction of instruments and procedures for measurement; and the development of theoretical approaches to measurement [47].

Recent investigations have insisted on a difficulty inherent in all attempts to judge a person character [48]. The trait of a character which will be ascribed to a person depends on the situation in which the person's behavior is being observed. Therefore the outcome of psychometric methods may not be accurate, consistent and reliable.

The Biometric-II judges the characteristics of a person taking into account his body type, face, hand, sleeping style, walking gait and such other aspects. Thus the physical appearance comes into greater focus on the grounds that they are constant factors not governed by the intentions of an individual. If these two different approaches are fused, the forthcoming results will be accurate, reliable and consistent.

Of the various analysis adopted to judge the characteristics, we would like to confine ourselves to face analysis and hand analysis. The reason why these two have been chosen is that hand and face analysis themselves are very vast domains. Just the analysis of face involves study of as many aspects as position, shape and size of parts like ears, eyes, nose, mouth, chin, cheeks, forehead, temple, eyebrows and so on. Now coming to hand, again its shape, color, hand geometry, palm-lines, their texture, position, points of their beginning and end, style, fall into the focus of our study. The following sections delineate the role of face and hand in the judging the characteristics of persons.

3.2. Approaches to Character Recognition from Face Analysis

Two important factors help us get a sense of how face reading is 'wired' into the human experience. The role face plays in communication and the relationship between the body and the mind in terms of Vedic Psychology.

Western psychology considers distinctive universal expressions for anger, fear, disgust, sadness, enjoyment, contempt, surprise and interest for character recognition. The face is the index of the mind. It reveals the nature of childhood, past lives, and gives insight into the individual's propensities and aptitude. The basic idea behind Vedic face reading is that the gross physical body is lying on the subtle body which has been in development for many lifetimes In this regard the whole body can be a source of information. Some of this is described in the Vedic text dedicated to this subject (Samudrika Shastra). Although the whole body carries information the face is the most informative part of the body because it is the most muscularly complex.

There are three main approaches to Psychological characteristics recognition from face. Physiognomy, phase facial portrait and ophthalmogeometry, the first originally interprets different facial features; the second works with angles of facial features and facial asymmetry, and the third extracts and interprets eye region parameters. Physiognomy is a theory based upon the idea that the assessment of the person's outer appearance, primarily the face, facial features, skin texture and quality, may give insights into one's character or personality. According to Atharva Vedic "Muhasastra" by observing the face of a person as a whole, one can get a keen idea of what his characteristics are like starting from childhood till his old age.

The general study of regions such as Parents' region, Career and Success region, Life region, Middle Age region Love and Emotional region, Fertility and Children region, Old Age region give an idea about health of parents, career prospects of the individual. attitude towards life, earnings, longevity, negative qualities, poor health, conjugal problems, deep emotional involvement in

matters of love, its failure, procreation, energy, childlessness, early death, happiness or desolation in old age, dangers of drowning, poisoning and food or water-borne diseases.

Shapes of faces such as Round, Square, Rectangular, Triangular and also straight face, concave face, convex face represent laziness, dullness, creativity, humor, fastidiousness, toughness, aggressiveness, offensiveness, sociability, strong physique, honesty, quality for diplomacy and leadership, competence to work as executive and official, anger, being unskilled yet boastful and rough, temperamental cheerfulness and brilliance, hyperactiveness and sensitivity, disturbed married life, equipoise, pessimism, happily disposed optimism, refinement in nature, criminal tendency, self-reliance.

Forehead, Eyebrows, Eyes, Colour of Eyes, Nose, Lips, Mouth, Chin, Jaw help recognize intelligence, opulence one enjoys in life, impediments encountered in one's career, longevity, broad mindedness, perfection in work, want of seriousness, snobbery, prudence, calculative nature and insecure mind, instability and agitation, untrust worthiness, untimely death, cruelty and perversion, inability to hide affections and emotions, secretiveness, extroversion, cynicism , flirtation, highly impressionable and affectionate disposition, inventive and adaptable nature, passion, tenderness, egotisim, escapism, stability of mind, extravagance, cheerfulness and friendliness, oscillation and depression, sensuousness, graceful behavior, erudition, introversion, generosity, humor, diligence, steadiness, flexibility, will power and energy. In general it is observation of outer body in estimating the inner self of a person.

3.3. Approaches to Character Recognition from Hand Analysis

The ancient Indian sages gave the out knowledge of the hand, which they gathered through the means of seeing, touching and analyzing. The knowledge thus gain threw light on all the aspects of life. The Indian tradition holds that Anga(limbs) vidya (hand analysis) was first invented by the Sea God 'samudra' it was there after developed and handed down to humanity by the sages like Narada & Gargya. Lord Skanda is held to be the patron deity of this science .Further the study of the hand analysis spread in different country grew, flourished and found favour in the firmament of knowledge [28].

There are two key to successful works with the hands. Primarily the ability to be observant and second key is an ability to compare and contrast each feature. Another valuable tool in the world of hand analysis is a system to organize all the clues find in a hand. The order in which all the hand features dealt is also very important [29].

From the point of view of hand psychology, the hand is divided into three zones as shown in figure 1. The first zone, which includes the thumb and the index finger and the portion below them, gives knowledge of environment and conscious behavior and the will power. The second zone includes the ring finger and the little finger and the portion below them. This zone is a good indicator of the subconscious mind and the hidden characteristics of the person. The middle zone is the social zone and it indicates the social adaptability of the person [28].

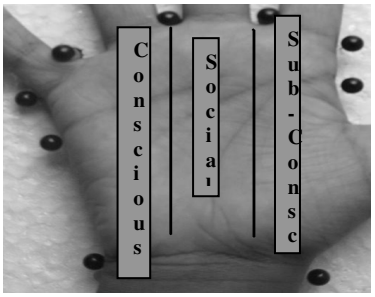


FIGURE 1: Three Zones on Hand.

The science of the hand is based on the shape of the hand. It can be pointed out here that the shape of the hand is so important that by its , one can know the type of mind is directing the subject's activities in life. By the formation of the Dog's foot one can tell for what particular kind of chase it is most suited; by the shape of the Horse's hoof one can tell what is the breed and what qualities particularly distinguished. In the same way, the shape of the hand sums up the whole of your mind and intelligence.

The raised portions of flesh are called Mounts in Hand analysis. They indicate the activity of various centers of our brain. They are the magnetic centers which deliver the message of the sub-conscious mind in the shape of lines. Each palm is divided into 9 parts known as MOUNTS. Each mount reveals mental as well as physical peculiarities of the individual such as the degree of egoism, extroversion, introversion sympathy, imaginative power, patience and endurance.

The study of the line is capable of explaining the finding out the good traits as well as the traits of various types of criminals. One can also judge with great accuracy the mental sickness and also the sickness due to psychological upset. The major lines are the heart, head, and life lines refer figure 2 and these lines are normally the deepest lines in the hand. They travel through every mount or zone of the hand and form a complete expression of an individual's energy and potential. To analyze the lines it is not just sufficient to concern only the individual lines but the other aspects like length, direction, origin and end position with respect to mounts, branches, islands and etc. should also be considered.

.

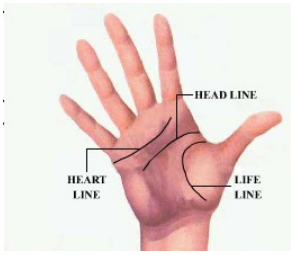


FIGURE 2: Major lines of the Hand

The line of heart is capable of tracing the more abstract, the more incomprehensible and the more subtle faculties of the person which ultimately decide the 'character 'and the 'personality'. The line of head is the most important line on the hand, as it indicates the brain power, mental forces and intellectual abilities. This line divides the hand into two parts .The upper parts show the mental development and the lower part shows the material development. The life line usually shows the vitality and energy of the person. The length, the depth and the continuity reveal the quantity of the vitality. This line also has great psychological aspect. The more characteristics that major lines represent can be found in [10, 11, 12, 14, 15, 16 project work chapter 3].

The major line in the palm contributes to the Major characterization of a person. The palm is a mirror of our Brain and mind. Dr Eugene Scheimann has referred to many experiments of his times to prove that the three main lines and the pattern on the skin of the palm are formed during the third and the fourth month of pregnancy [45 project work].

4. BIOMETRIC-III : AS A TOOL IN MEDICAL DIAGNOSIS

Biometrics-III extends the horizon (various precise landmarks) in identifying many health concerns. "Overall health is achieved through a combination of physical, mental and social well being which together commonly referred to as health triangle". [Wiki]The Indian concept of health can be understood from the available sources of the Vedas, Upanishads and Yoga.[Mart hum paper].These literature describes the balanced state of health conditions and any deviations from the balanced condition considered to be problematic. Good health means a perfect balance between the body, mind and soul, which is referred with the concept of 'Triguna'.

Among the four Vedas – Rig, Sama, Yajur, and Atharva, Atharvaveda deals with mental health and illness, i.e., the concept of 'Triguna' : 'vata', 'pitta', and 'kapha' : the motion, the liquid component, and the heat or energy component; as well as mental i.e., 'satva', 'rajas', and 'tamas'.The mental Triguna also contributes to the personality of a person as stated in the earlier section(Biometric II), while Biometrics III is more confined to the Physical Triguna i.e., 'vata', 'pitta', and 'kapha' that signifies health conditions.

Face is the index of mind. Our beauty is a reflection of our health, both physical as well as mental. Listen to your body; it is trying to tell you something! In the early days doctors diagnosed the diseases, by examining the tongue, the eyes, nails, various skin patterns (dermatoglyphics), and hair distribution. In general it is nothing but identifying the diseases through the body symptoms .

The health condition of a person is perceived by the Face and Hands. The Mishio Kushi Theory (Face Reading) contends that a weakness or toxicity in a certain organ or gland can cause acne on certain areas of the body. The radiance on the face represents the credibility of the body health. Similarly Hand interprets the thought impulses of the brain and hence these carvings and also all nerve endings happen to end here.Hence forth the Face and Hand can be made into a continuum of differential areas of face, which in turn represents the health condition of the body. The following sections delineate the role of Face and Hand in Serving as a Tool in Medical Diagnosis.

4.1. Medical assessment on the Basis of the Face Analysis

Face is broadly divided into: The fore head, The mid face, and The lower region, any variaition of health is reflected on any of the above region via acnes, pimples, wrinkles, lines, rashes, redness, and puffiness. Acne and pimples are the body reactions to the clogged pores i.e., toxins often cleanse through the skin, when the other eliminating systems are sluggish; hormonal imbalances over stimulate the oil glands and impurities in the blood will often affect the skin & these leads to clogging of pores.

In Oriental medicine the forehead-represents the nervous system, the gall bladder, liver and stomach. If there is no proper excretion of toxins by the gall bladder and intestine, results in pimples. The mid face-This is the area from the eyebrows to the bottom of the nose. The organs that dominate this region are the heart and the lungs. Kidneys, stomach and liver reflect in areas around the eyes .The lower region -This is the area from the nose down to the chin. It is related to the digestive track. The chin is related to the functioning of the kidneys and gall bladder and reproductive organs. Acne is often thought as a normal part of adolescence. This is not true, acne is merely a symptom that is accentuate by the rapid growth and change within an adolescent's body.
The table represents more clearly the various regions of face relating to internal organs and glands.

Regions of Face	Relating Organs/Glands
Forehead	Intestine
Above eye brows	liver
Between eyes	Spleen
Bridge of nose	Male reproductive
Cheeks	Lungs
Tip of nose or on ears	Heart
Creases at base of nostrils	Brain
Jaw line	Female reproductive
Upper lip	Stomach

TABLE 1: Various regions of face relating to internal organs and glands.

So listen to your body and do not jump to a conclusion and rush to the doctor just with the symptoms .Rather try to study the body changes and act accordingly.The oriental medicine challenges the existing ambiguous concepts, and provides solution.

4.2. Medical Assessment on the basis of the Hand Analysis

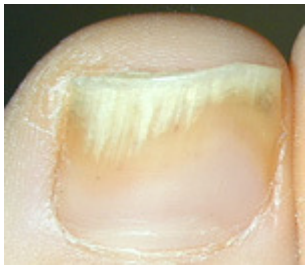
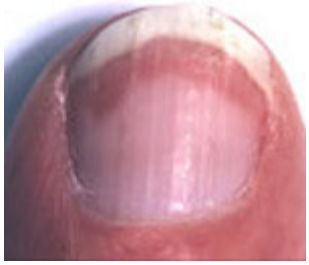
Information on the law and practice of hand reading has been found in Vedic scripts, the bible and early writings. Judging by the number of hands painted in prehistoric caves it would seem that palmistry interested humans since the Stone Age. Medical researchers studying skin patterns (dermatoglyphics) have discovered a correspondence between genetic abnormalities and unusual markings in the hand. Research has confirmed a link between specific fingerprint patterns and heart disease. Dermatoglyphics, an area of research which deals with the study of finger prints, and to a lesser extent, palmar creases, in terms of their development and their relationship to some birth defects and genetic conditions.

Supporting this view we have derived the ideas and research made by Martijn van Mensvoort, MSc. (Psychologist from The Netherlands), who has done more than 15 Years of work on Hand Analysis. The subsequent paragraph clusters together on how the different part of the Hand i.e., Nails, Fingers, dermatoglyphics, and Hand lines.

4.2.1 Assessment on the Basis of Nails

From the early 80's various works have been published which describe the clinical relevance of the nails. A classical example in this field is the work presented by Beaven & Brooks: Diagnosis. However, only a few years ago medical students were hardly aware of the clinical value of the nails. In order to fill this gap several dermatologists have combined their knowledge and created in 1997 'Nail-Tutor ': a visual personal computer program including 150 photo's which describe the anatomy and pathology of the nails.

In general one can say that only some diseases are frequently accompanied with nail abnormalities. The following table shows a sample overview of the most well-known diseases with description of the accessory nail abnormalities. More detailed description can be obtained from [30].

Onychomycosis 	Terry's nails 	Diabetes In a Atlas of Diseases of the Nail it is described that diabetes is relatively frequently accompanied with onychomycosis and 'Terry's nails' (half white, half pink nails).

Oil-drop nails	Beau lines	Psoriasis:
		Psoriasis nail diseases can be manifested in up to 55% of people who have psoriasis. Frequently observed nail abnormalities are: oil drop nails, Beau lines, onychomycosis, spots in the lunala, and splinter hemorrhages.

FIGURE 3: Assessment on the Basis of Nails

4.2.2 Assessment on the Basis of Dermatoglyphics

The famous magazine 'Nature' in 1963 presented an article published by L.S. Penrose who described observations which indicate that Down's syndrome is accompanied by specific characteristics in the dermatoglyphics. Stereotypical features in the dermatoglyphics in Down's syndrome are: Ulnar loops on all fingernails (+ possibly a radial loop on the ring finger), and in palm: a high positioned axial triradius, and a loop between the ring finger and the middle finger. The researchers have discovered that other genetic syndromes are accompanied with dermatoglyphics abnormalities as well. Another relevant factor appears to be the fact that the relative syndromes are usually accompanied with a high frequency of congenital heart disease.

4.2.3 Assessment on the basis of the Fingers

Of late the work of John T. Manning ('Digit Ratio') has reached the media worldwide. Manning discovered that the so-called 'Digit ratio' (the ratio of the index finger length and the ring finger length) might become a useful instrument in the diagnostic perspective of various medical and/or psychological problems.

4.2.4 Assessment on the basis of the Hand Lines

The scientific research has indicated that certain characteristics of the lines can indeed have some medical significance. However, in isolation these features have no value at all: only certain COMBINATIONS of features can provide a solid basis for a medical diagnosis. For instance various studies executed by medical researchers have shown that the so called simian crease is observed in about 60% of people who have Down's syndrome (mongolism).The diagnostic value of the other lines has not been established yet. However, it is premature to conclude from this observation that traditional hand analysts have gathered more insights on this matter. For, these 'alternative' insights have been constructed merely on the basis of anecdotal evidence.

5. CONCLUSION

The above paper provides a platform for future studies on Biometrics system. Although earlier studies have said that Biometrics can be used for verification and identification, we arrived at a conclusion that it has much more. For it can be used to identify the personality as well , as discussed in Biometric-II.These biometrics can also be used to analyze the health condition as discussed in Biometric-III.These levels can be extended depending upon other factors like to study of previous birth to Biometric IV. Some of the novel approaches used in finding the characters using the Hand Analysis confining to three major lines have obtained encouraging results. We hope this humble beginning will lead to a major leap in future for personality identification and for medical assessment of various disease diagnosis .

6. REFERENCES

[1] J.G.Daugman, "High confidence visual recognition of persons by a test of statistical Impedance", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11): 1148-1161, November 1993

[2] Davide Maltoni, Durio Maio, Anil K. Jain , Salil Prabhakar, "Handbook of Fingerprint Recognition", 2002

[3] Ruud M. Bolle, Jonathan H. Connell, Sharath Pankanti, Nalini K. Ratha, Andrew W. Senior, "Guide To Biometrics", 2003 Websites (accessed latest on January 13, 2005.)

[4] [online] http://csc.noctrl.edu

[5] [online] http://www.biometricsinfo.org

[6] [online] http://ctl.ncsc.dni.us/

[7] [online] http://www.gizmo.com.au

[8] [online] http://www.iirme.com

[9] Siddhesh Angle, Reema Bhagtani, Hemali Chheda , Thadomal Shahani , "Biometrics : A Further Echelon Of Security " 1993.

[10] D. Edward D. JD. Campbell, "III BIOMETRIC FUTURES", This is taken from a paper originally developed for Amida Biometrics, L.L.C., in 2002, by Edward D. Campbell, and edited for use on the IBMBS web site ins September, 2003.

[11][Online],http://www.griaulebiometrics.com/page/en-us/book/understanding-biometrics/introduction/authentication-technologies/comparison

[12] Nanavati S, Thieme M, Nanavati,R," Biometrics, Identity Verification in a Networked World, , Wiley Computer Publishing, 2002",[Online] Available: http://www.colby.edu/~jbrosenb/STS%20Project/web/; ,

[13] Lee Siow Mong, "The Chinese Art of Studying the Head, Face and Hands", *Eagle Trading Sdn Bhd, Malaysia,* 1989.

[14] Lad, Vasant, Ayurveda, "The Science of Self Healing", *Lotus Press*, pp 60-62, 1984.

[15] Santa Fe, N.M., Ros, Frank ,"The Lost Secrets of Ayuvedic Acupuncture", Lotus Press, Twin Lakes, WI. 108-112, 1994 .

[16] Ed. Sharles S, " Major Psychological Assessment Instruments", Newmark, 1996.

[17]Simon , "Vedic Astrology The Language of the Gods", Available http://www.soothsayers-india.com/vedic-astrology.

[18] Dr. Dhira Govinda Das , "The Vedic personality Inventory"

[19] Sri aurobindo kapila shastra , "Atharavabveda".

[20] Sri Selvam Siddhar, " Atharva Vedic "Muhasastra" Also known as FACE READING Commander Selvam, Available: http://forums.joeuser.com/366328.

[21] [online] http://www.howtoreadfaces.com/index.html

[22] [online] http://www.astrology.com.au/face/index.asp

[23] [online] http://physiognomyinfo.com/

[24] [online] http://www.dhyansanjivani.org/chin_face_read.asp

[25] "The naked face", Available: www.gladwell.com/2002/2002_08_05_a_face.htm

[26] Face Reading for astrology charts, horoscope reports, compatibility and star signs. Available: www.astrology.com.au/face/theface.asp

[27] The Art of Face Reading Available: www.chiofearth.com/article4.htm

[28] Dr .M.Katakkar, "Encyclopedia of Palm and Palm Reading", *UBSPD,* India, 1992.

[29] Judith Hipskind Collins, A to Z of Palmistry," The Essentials for reading Hands" Dallas, JAICO, 2008

[30]Caslon Analytics,Physiognomy, "Phenotypes ",Available: http:// www. caslon. com. Au physiognomynote4.htm

[31]CaslonAnalytics,Physiognomy,"Phenotypes",Available: http:// www. caslon. com. Au physiognomynote9.htm

[32]CaslonAnalytics, "Physiognomy:Character ", Available: http://www.caslon.com.au/idcrimeguide31.htm#b red Cheiro, Cheiro's Book of Fate and Fortune, America: New Lights

[33]Kenneth A. Lagerstrom, "Introduction to Cheirology", Available: http://www.humanhand.com/intro.html

[34]Martijn van Mensvoort, "Palm Reading Research - Unique Reports from the Fields of Hand Analysis" , Available: http://www.handresearch.com/research/palm-reading-research.htm

[35]Martijn van Mensvoort," Hand Research: Global News & Science about Hands", Available http://www.handresearch.com

[36]Martijn van Mensvoort, " Hand Analysis Research", Available http://www.handresearch.com/hand/Evolutie/10jaarEngels.htm

[37]Martijn van Mensvoort, "Discoveries about the Human Hand in the Fields of: Psychology, Psychiatry & Medics", Available http://www.handresearch.com/course/palm-reading-course.htm

[38]Martijn van Mensvoort," Hand Analysis: Psychology ", http://www.handresearch.com/hand/Evolutie/psychoEngels.htm

[39]Martijn van Mensvoort, "Hand Analysis Researcher Overview "http://www.handresearch.com/hand/Evolutie/overzichtEngels.htm

[40] Dr.Narayan Dutt Shrimali, "Practical Palmistry", Judhpur: Pustak Mahal, 1979

[41]Prof.O.P Verma," The principles of Palmistry", Vol-1, India: Ranjan Publications, 2005

[42]Prof.O.P Verma, "The principles of Palmistry", Vol-2, India: Ranjan Publications, 2007

[43][Online] http://www.marysherbs.com/heal/heal-f-rP.htm

[44] [Online] http://www.handresearch.com/medical-hand-analysis.htm

[45] [Online] http://www.handresearch.com/news/hands-on-cancer-hand-palm-cancers.htm

[46] [Online] http://en.wikipedia.org/wiki/Big_Five_personality_traits

[47] Ekaterina Kamenskaya1 , Georgy Kukharev2,"Recognition of Psychological Characteristics from Face".

[48] A Argelander, "The Personal Factor In Judging Human Character", Institute of Brain Research, Berlin, 2006.

[49] [Online] www.howtoreadfaces.com/intro.pdf

[50] [Online] http://www.indusladies.com/forums/

A Consistent and Efficient Graphical User Interface Design and Querying Organelle Genome "GUEDOS"

Hassan BADIR *hbadir@gmail.com*
Labtic, National School of applied sciences
University Abdelmalek essaadi
90020, Tangier, Morocco

Rachida FISSOUNE *fissoune@gmail.com*
National School of applied sciences
University Abdelmalek essaadi
Tetouan, Morocco

Amjad RATTROUT *ramjad@gmail.com*
University Claude Bernard
Lyon, France

Abstract

We propose a software layer called GUEDOS-DB upon Object-Relational Database Management System ORDMS. In this work we apply it in Molecular Biology, more precisely Organelle complete genome. We aim to offer biologists the possibility to access in a unified way information spread among heterogeneous genome databanks. In this paper, the goal is firstly, to provide a visual schema graph through a number of illustrative examples. The adopted, human-computer interaction technique in this visual designing and querying makes very easy for biologists to formulate database queries compared with linear textual query representation.

Keywords: Graphical User Interface, Complex Object, Bioinformatics, Gene, Genome, UML-BD

1. INTRODUCTION

Many biologists use freely and extensively a large number of databanks collecting considerable amount of information. Data from sequence databanks are daily submitted worldwide, usually by electronic mail, then manually checked and stored – generally-under a RDBMS. They are then broad casted to the main servers of the community – daily by Internet or quarterly by CDROM- under a flat file format (ASCII files).

There is no agreement on the format and, for the same data; many formats may b available, depending on the databank manager (Europe, US, Japan…). The number and the size of databanks are grouping rapidly (the size of major sequence databanks doubles each year).

Why Organelle Genome? Organelles (mitochondria and chloroplasts) are of interest for several reasons, including their:

- Possible bacterial origins,
- Relationship to the evolution of the nuclear genome,
- Central role in eukaryotic cell energy production
- Utility as population markers.

The most important advantage to genomics offered by organelles is the number of completely sequenced genomes already available and currently being sequenced (e.g. by OGMP: Organelle Genome Mega-sequencing Program). No larger collection of completely sequenced exists. Sequenced organelle genomes are information-rich datasets: firstly most of them belong to a highly analysed set including protein-coding, transfer RNA (tRNA) and ribosomal RNA (rRNA) genes, secondly the relationships between and among genomes can be determined (at both gene and genome levels), making them ideal for comparative genomic studies.

For many years, computer scientists have provided help to biologists for managing these data. As a result, a lot of retrieval systems, like SRS [11] or ACNUC, have been created which are based on indexing flat files. These simple tools obtain a wide success in the biologist community.
Web interfaces have also become very familiar and make the data very easy to access for any biologist. The ability to retrieve data from anywhere on the network has progressively leaded to a new problem: how to interconnect heterogeneous databanks? At present it is difficult for researchers to access all the relevant information associated with an sequence genome. Data are dispersed among a number of sources. In this present disorganized state, organelle genomic data constitute a major underexploited source of information. A wide discussion on this topic has started within the molecular biology community. To integrate such data warehouse is considered as one of the main problems to face up in bioinformatics.
Directly manipulating visual conceptual schemes of complex objects provide users with a clear and powerful mean of interaction. An end user of GUEDOS-DB is able to graphically build a database schema or modify an existing one, to load all the information initially provided by a data bank as GENBANK, to browse through the schema of the database in order to construct the queries about the data and to save them for later use. All of these activities are accomplished by using a unique graphical iconic representation.
As a consequence, different types of users, especially novices, can learn and use the query facilities in a more intuitive way, without necessity to remember the database scheme or the grammar of the query language.
In this paper we describe the visual query interface GUEDOS-DB (DBioMics) developed in the in object-oriented [2] environment for software engineering workshop development. A database schema based on an object-relational [1] data model using SQL3.
The user interface proposed in this paper:

- In descriptive: it provides a concise and complete visualization of the data scheme called Object Semantic Graph;
- Uses the same medium both for the description of the data scheme and for the representation of the formulated query on the semantic graph.
- Is interactive: the formulation of a query is made by simply designating the nodes and arcs of the displayed semantic graph and the syntactic units through the technique of direct manipulation.

The remainder of this paper is organized as follows. We first review the previously proposed biological databases in the next section. In section 3, we review the basic concepts of the object oriented data model. We then describe our proposed graphics user interfaces in section 4. Section 5 describes some graphical queries with a number of illustrative examples. Finally, we conclude in section 6.

2. RELATED WORK

In fact biological interfaces are developed upon one or more database management systems or one or more data file management systems. Database models can be relational, object-Oriented, object-relational or hierarchical. Some ad-hoc solutions have been proposed:

- The **Kabat antibody databank** [4] is one of the first realizations in this area. The system visualizes hierarchical schemes and includes computational tools. It is developed on P/FDM (Functional Data Model), an in-house OODBMS built at Aberdeen. Data are modelized in the FDM, and queries are made through DAPLEX language, which is based on PROLOG. Describing database schemes and querying are quite distinct.
- **GOBASE** [3] is implemented under SYBASE RDMS on a Sun SPARCstation. Custom software has been developed for make easy to populate and maintain the database. It uses the Perl language and Sybase's Open-Client development tools. The query interface is supported through WWW forms using the web/General gateway. The user interface allows the users to navigate in and to retrieve information from GOBASE, and includes the following entities: gene, intron, RNA, protein, organism, sequence, chromosome and signal. The interface also contains hypertext links to specific information in other internet-accessible databases. In future versions, the user interface will be expanded by adding new entities to the database and by offering analytical tools such as subsequence extraction and neighboring searches.

- **AGIS** (*agricultural Genome Information System)* [12] is a World Wide Web product of a cooperative effort between the department of Plant Biology, University of Maryland and College Park in USA. This databank consists of genome information for agricultural organisms. At present, it encompasses mostly crop and livestock or non-commodity animal species. Also included are a number of databases that have related information, e.g. germplasm and plant gene nomenclature data.
- **Genomic databases** can be viewed by ACDB [9] a widely used generic genome interface. It offers a specific visual interface. It is worth noting that this interface supports describing and querying schema but possible queries are too coarse.
- **Docking-D** [3], a prototype for managing ligands (PDB, HPSS…), is implemented with and OODBMS prototype VODAK [8]. According to its authors, the VODAK data model provides all standard features of object-oriented data model; the system includes SQL-Like query language with optimization and multi-user access modules. VODAK was initially created in response to the lack of a declarative query language in many object-oriented database systems like ObjectStore. It is still under development.

3. GUEDOS DESCRIBING

3.1 Architecture

3.1.1 Description

GUEDOS prototype is based on a graphical approach used to represent graphically database schemas. In addition, schema conception is realized by systemic and direct graphs manipulations by a set of available graphic operators. Also, GUEDOS is characterized by several criteria:

- **Flexible**: flexibility implies that the used graphical language should be adapted to possible uses.
- **Incremental**: practical interest is situated in the fact that, during schema conception phase, user can construct schema incrementally by defining more constraints.
- **Uniform**: uniformity imposes identical interaction modes for the different functionalities enhancing thus the tool ergonomic.

Figure 1 describes GUEDOS general architecture. It's composed mainly by two components: a graphical interface and an automatic transformation module.

GUEDOS Prototype was developed by java. It represents a semantic suite of LASCI-Complex toot developed earlier by in a thesis framework. This tool allows an automatic visualization of database structure, allowing thus an automatic transformation using algorithms developed by our team.

3.1.2 Functions

GUEDOS is a platform for editing and designing object and object relational database schemas. The systemic aspect in GUEDOS is the key solution for a guided construction of such a schema, because his objective is to facilitate requirements specifications of skilled designer or not, and to simplify him the design task. Its functions are to allow:

- To users to express their requirements by the means of one or several models like: universal relation with inclusion (URI), object attributes forest and UML DB-stereotyped classes diagram. The tool deals with these models by the means of a logical and syntax apprehension of those. User describe his requirements, and affine his constraints to obtain an initial conceptual schema. User can also normalize and transform a representation under the form of relations related by inclusions dependences into a normalized semantic graph. This normalization could be partial, leaving unchanged certain complex structures of fixed objects by user,
- To swap from one model -among those cited above- to another using transformation algorithms [6] and to generate SQL3 description or an XML schema,
- To personalize the obtained conceptual schema with respect to the foreseen processing, by introducing access methods, even denormalizing them.
- And to integrate many conceptual schemas into one without losing any information.

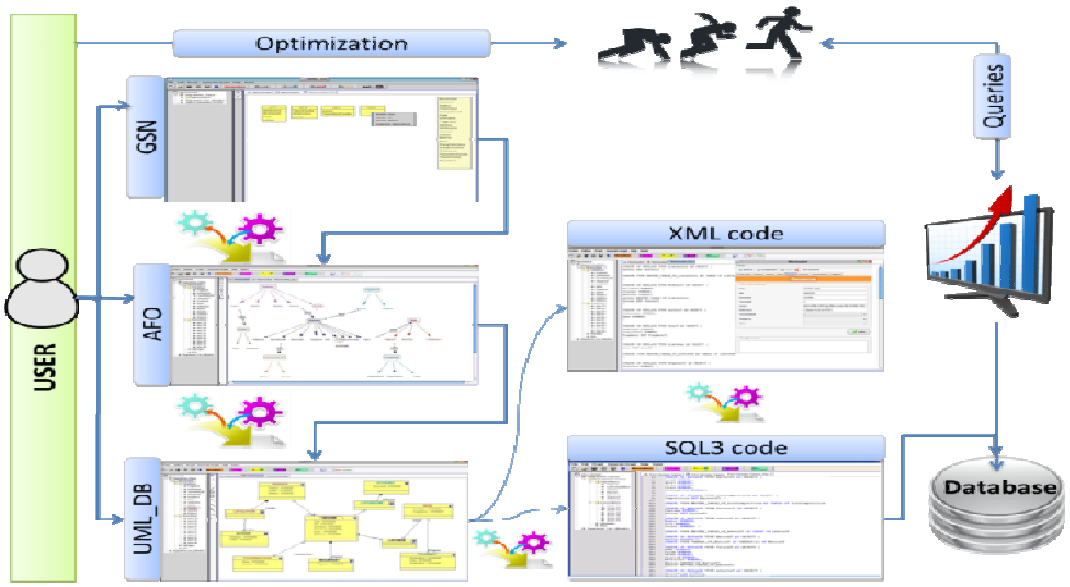


FIGURE 1: Architecture of GUEDOS

GUEDOS interface could be decomposed into three different screens, according to user need. In the first case, user looks to edit a set of attributes and functional dependencies to produce systemically NSG, using normalization algorithm; secondly, user constructs an OAF manually by specifying its constraints or automatically from the previously obtained NSG relying on a transformation algorithm developed in [6]. In the third case, user conceives a database stereotyped UML class diagram from an OAF previously obtained using transformation algorithm in [7]. GUEDOS has two other forms allowing to display SQL3 description and XML schema generated from classes diagram. Concretely, GUEDOS avoids waste of time generally noticed in redoing repetitive tasks, and allows then a more coherent information processing. User beneficing of supervision and control capacities on global tasks he accomplishes. It allows, more generally, a better comprehension of user conceptual framework.

3.2 Manipulation and Importing

The first part of this biological application has consisted in defining a data schema in the Graphical Object Data Model (GODM) [13] [6].
Figure 2 shows mitochondrial sequences graph. This schema respects the ontology and allows the indexation to the largest quantity of knowledge. The design of a GODM schema is not our object in this paper [6][14].
GUEDOS includes a schema editor, allowing designers to build such a GODM schema by picking graphical symbols from palette and positioning them into the workspace provided in an ad-hoc window. An analyst is able to work simultaneously on several schemas by using graphical window for each. Standard editing operations are available through pull-down menus.
The mitochondrial schema [10] represents seven fundamental classes corresponding to the types Genome, Gene, Fragment, *NotLeafTaxon*, *Reference*, *KeyWord* and *Author*. These are depicted using rectangle nodes, indicating that they correspond to abstract data types in the world. These classes are referencing each other using internal identifiers that are not visible to the user. In contrast with abstract types, attributes may be either atomic or complex.
An atomic attribute is depicted using oval such as *GeneType* of class Gene in Figure 3.

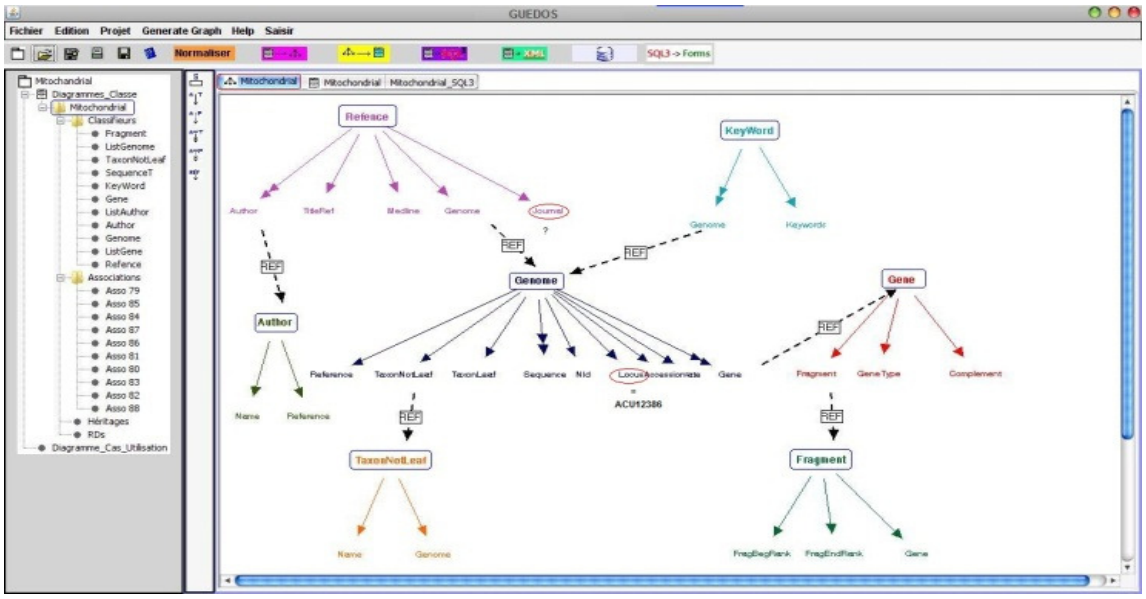


FIGURE 2: schema graph for mitochondrial sequences

A complex attribute is depicted using also oval, but decomposed into a tuple of attributes, which may be also either atomic or complex. Figure 2 also illustrates attribute-domain relationships. For example the *Genome* class has an attribute *NotLeafTaxon* referencing the *NotLeafTaxon* class. Then,

- The *NotLeafTaxon* class is the domain of attribute *NotLeafTaxon* of class **Genome**;
- An object instance of **Genome** has an object instance of *NotLeafTaxon* as the value of *NotLeafTaxon* attribute (single-valued attribute).

Figure 2 shows mitochondrial sequences schema graph. This schema respects the ontology and allows the indexation to the largest quantity of knowledge. In an Object Attributes Forest [7], each attributes forest has a color chosen randomly by GUEDOS or indicated by user. This color will be inherited by UML classes diagram during transformation. A complex class or an attributes tree could have one more keys. For example in figure 2, Genome has as keys *IdGenome* and Locus *TaxonNotLeaf* . Figure 3 represents DB-stereotyped classes diagram with derived keys from an OAF (figure 2) using UML-DB profile [6].

4. GUEDOS MANIPULATION AND QUERIES

There are two ways to extract data from a database. One way is extensional, by navigating through the database at the occurrence level. The other way is intentional, by formulating a query asserting which types in the database schema are relevant to given query and which attribute values and links among objects in order to define which subsets of those populations are relevant. A query also defines which data from the relevant subsets have to be put into the result (typically a projection operation) and how these data have to be structured for the end result (unless the result is by definition a flat or first normal form from a relation). In a classical query language as OQL/SQL, we can find these different specifications expressed as Select-From-Where blocks: the FROM clause restricts these classes to the relevant subsets, finally the Select clause specifies the projected data.

In a visual environment, the definition of the relevant object classes (the FROM clause) is performed by visualizing the database schema on the screen and by clicking on the desired class-nodes to lift them out from the schema into the query sub-schema (alternatively, by clicking on undesired types to reduce them from the schema, which gradually reorients it to the target query sub-schema). The query sub-schema is a sub-graph of the original schema graph.
The definition of the relevant subsets of the classes in the query (the WHERE clause) is expressed as conditions (predicates) upon desired attributes. Only these objects that match the

conditions will contribute to the result. Conditions in a query are divided in two categories: simple conditions which apply to object sets in a single class and composite conditions which apply to object sets in multiple classes.

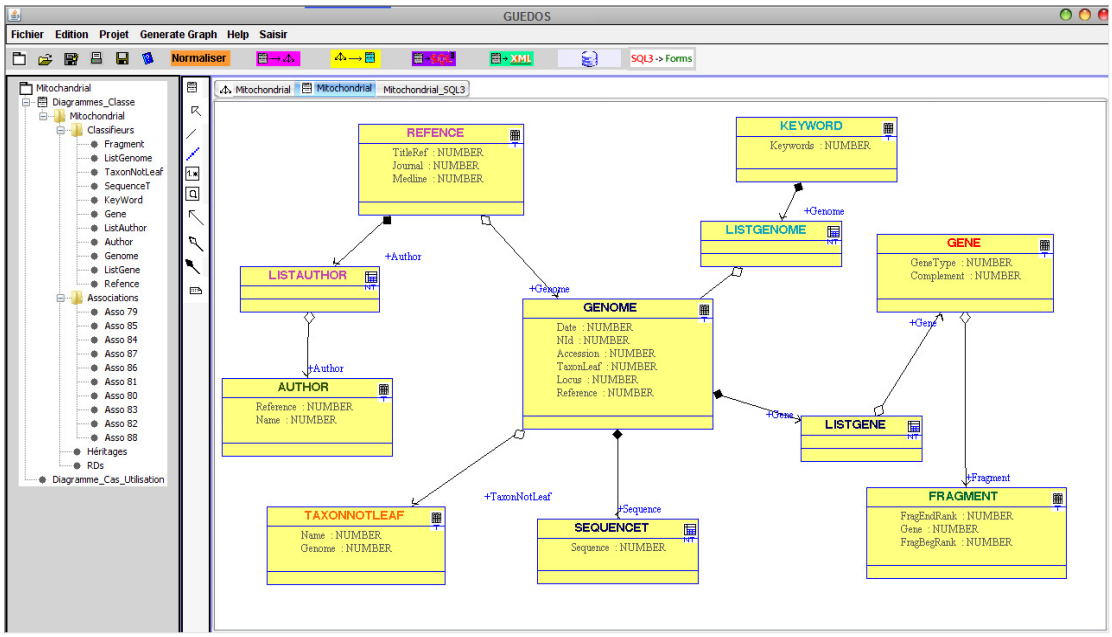


FIGURE 3: schema UML-DB for mitochondrial sequences

The definition of data to be presented to the user and the way by which they have to be presented (the SELECT clause) can be seen as the definition of a new, virtual object type. In the query sub-schema the projected attributes are represented by the symbol "?".

4.1 Complexes Queries in Object-relational Model

Object Relational Queries are the queries, which exploit the basic object relational model thus allowing the user to retrieve data from the new data structures that have been introduced in the model. Object Relational Queries can be broadly classified into: Queries having REF, Queries for Nested Table Structure (Aggregate Queries), Queries using index cluster, and Queries for Inheritance Relationship [16].

REF Queries

REF is incorporated in database by defining one attribute in the table, which holds the REF information of the attribute, which belongs to the other table. REF is used when there is an association relationship between two objects. REF queries are the type of queries, which involve REF either in projection or join or selection. REF is a logical pointer, which creates a link between two tables so that integrity in data can be maintained between the two tables. The attribute which holds the REF value is called as ref attribute (e.g. *Loginstaff_id in Login_t table*) and the attribute to which ref attribute points is called as referred attribute (e.g. *staffid in person_t* table). Ref attributes stores the pointer value of referred attribute as its real data value. Most important thing to keep in mind is that whenever we refer to REF we always give the alias of the table, which holds the referred attribute and not the table name. REF takes as its argument a correlation variable (table alias) for an object table. Generally, REF query consists of Projection, Joins and Selection.

```
SELECT <Projection List>
FROM <table1> <alias1>, <table2> <alias2>, … … …
WHERE <alias2>.<ref attribute> = REF(<alias1>) ;
```

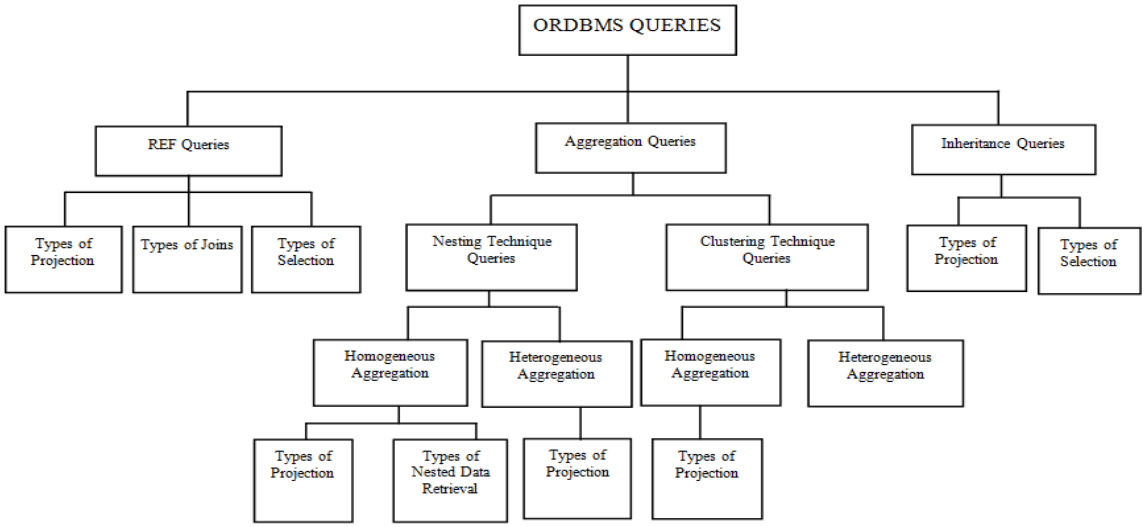


FIGURE 4: General Classification of the ORDBMS Queries

Projection refers to what we to get as a result of the execution of the query. Joins are the links, which are responsible for maintaining integrity in data, which is common to two tables. Selection is the condition based on which we do projection.

Aggregate Queries

Collection types give the flexibility of storing a series of data entries that are jointly associated to a corresponding row in the database. They can be further classified into two: VARRAYS, Nested Tables. We will not be discussing VARRAYS in this section, since at the moment we cannot write SQL statements to access the elements of the VARRAYS. The elements of the VARRAYS can only be accessed through PL/SQL block; hence it is out of the scope of this section.

Nested table is one of the ways for implementing aggregation. Nested table data is stored in a single table, which is then associated with the enclosing table or object type. In nesting technique, the relationship between "part" and "whole" is existence dependent type. If the data for the whole object is removed all of its part objects are removed as well. Nested Table is a user-defined datatype, which is linked to the main table in which it is nested. Generally nesting technique is of two types: Homogeneous and Heterogeneous, depending upon the number of the parts that the main table has.

Aggregation is an abstraction concept for the building composite objects from their component objects. Participating entities have "*Whole-Part*" type of relationship and the part is tightly coupled with whole. Aggregation can be done in the following two ways: Nesting Technique and Clustering Technique. Both the techniques store aggregate data efficiently but the degree of cohesiveness between whole and part data is more in nesting technique than in clustering technique. *Nesting* and *Clustering* technique can further be classified into *Homogeneous* and *Heterogeneous Aggregation* depending upon the number of parts they have,

```
SELECT <Nested table attribute1>,
       <Nested table attribute2>, … … …
FROM THE (SELECT <nested attribute>
          FROM <whole table> <alias1>
          WHERE <primary key condition>);
```

Clustering Technique Aggregation Queries

Clustering gives the flexibility of storing the "whole-part" type of information in one table. This technique is used when there is aggregation (i.e. the whole is composed of parts). This enforces

dependent type of relationship and each (i.e. either whole or part) has unique ID [17]. Clustering can be further classified into homogeneous and heterogeneous clustering aggregation.
Clustering technique implements the participating tables in "Whole–Part" type of relationship. The part information is tightly coupled with the corresponding whole record. For each whole info, we have many corresponding parts and this is achieved by creating cluster on the whole key. Index creation improves the performance of the whole clustering structure. Clustering can also be divided into two types depending upon the number of participating subtypes: Homogeneous Clustering and Heterogeneous Clustering.

```
SELECT CURSOR (SELECT <Projection List>
               FROM <main table name>
               WHERE Join AND[<condition>]),
       CURSOR (SELECT <Projection List>
               FROM <part table name>
               WHERE Join AND [<condition>])
FROM <whole table name> <alias1>,
     <part table name> <alias2>
WHERE <alias1>.<attribute name> = <alias2>.< attribute name >
AND [<condition>];
```

Inheritance Queries
Inheritance is a relationship between entities, which gives the flexibility to have the definition and implementation of one entity to be based on other existing entities. The entities are typically organized into hierarchy consisting of parent and child [17]. Child inherits all the characteristics of its parent and can also add new characteristic of its own. A parent can have many children but each child will have only one parent. There can be multilevel of inheritance (i.e. a parent child can have many other child's as well). Basically in Object Relational Database System, inheritance can be of three types: Union Inheritance, Mutual Exclusion Inheritance and Partition Inheritance. The basic difference between three types of inheritance is in implementation but for querying purposes they are basically the same. In this paper we have only taken Union Inheritance into consideration for writing SQL statements as the only difference between different types of inheritance is the way they are implemented in the database and not in terms of writing SQL. The general syntax for implementing inheritance relationship is as follows.
Generally inheritance queries consist of projection and selection. SQL for inheritance queries is same irrespective of type of inheritance or number of levels in the tree hierarchy unless mentioned specifically. The general syntax for inheritance queries is as follows.

```
SELECT VALUE(<alias>).<supertype attribute name1>,
       VALUE(<alias>).<supertype attribute name2>, … … …
FROM <table name> <alias>;
```

4.2 Query Editor
GUEDOS-ASCK includes an editor for graphical specification of queries, inserts and updates. In this section we present the main features of the editor. The discussion is limited to query formulation. The various steps which the process of query formulation comprises are:

a. Selecting the query sub-schema:

This is the initial step for all visual query languages. The portion relevant to the query is extracted from the database schema is reverse video.

b. Specifying predicates:

Predicates are stated here to be applied to database occurrences, so that only relevant data are selected. Predicates against complex objects may be rather clumsy. For the simplest ones (comparison of a mono-valued attributes with a constant) a graphical counter-part may easily be defined. A simple specification technique is to click on the attributes, select a comparison operator from a menu, and finally type the value or choose one from a list. For complex predicates (involving several quantifiers, for instance), there might be no simple way to express it graphically. Menus are sometimes used for syntactic editing of predicates. In GQL/ER [15], QBE-Like (Kari, 1990) forms are used to specify conditions on the selected nodes. In GUEDOS, textual

specification of Boolean or arithmetic expressions has been preferred to graphical representation: There are simpler for complex expressions.

c. Formatting the output:

The selection of projection of projected attributes defines the structure of the resulting entity type.

4.3 Query Examples

This section illustrates some examples of graphical queries in GUEDOS-Queries. Let us assume that the user wants to formulate a query for the schema show in Figure 1.
For each query we show:

- The **query window**, which displays twos superimposed graphs with different shades:
  - The object-relational schema which is bright and fixed in the background to guide the user,
  - The sub-graph under development, which is reverse video.

  The window has a selection bar which, in addition File and Edit menus, contains: the Query menu, used for choosing the type of query to be made on the query graph
- The **SQL3** code window for displaying the corresponding SQL3 code of the formulated visual query.
- The result window for displaying the result of the executed query.
- Consider the simple query, 'Print the NId and TaxonLeaf of Genome which the Locus is "ACU12386" 'in Figure 4.
  The user selects the Class boxes to which the projection and condition may be specified (Genome Class in this example). The user will proceed in this way.
- Designate the projected attributes ***NId*** and ***TaxonLeaf*** of Genome class by choosing the PROJECT option in the pull-down menu associated to this attributes.
- Establish the selection condition ***Locus*** = "*ACU12386*" of Genome class. The user carries out the following actions:
  - Double-click on the attribute *Locus* of *GENOME* class and the pull-down menu associated to this attribute is displayed,
  - Choose the predicate option in the menu
  - (implicit) choose the comparator "="
  - Implicit choose the option "value to be entered",
  - Enter the value "ACU12386" in the open box.
  - The answer to the request is a temporarily sub-class of the **Genome** class and is composed of appropriate objects. The sub-class name will be either the name of the query or a specific name given by the user. On the query sub-graph, the user selects the SHOW RESULT option in pull-down menu in the class node to visualize the temporarily objects.

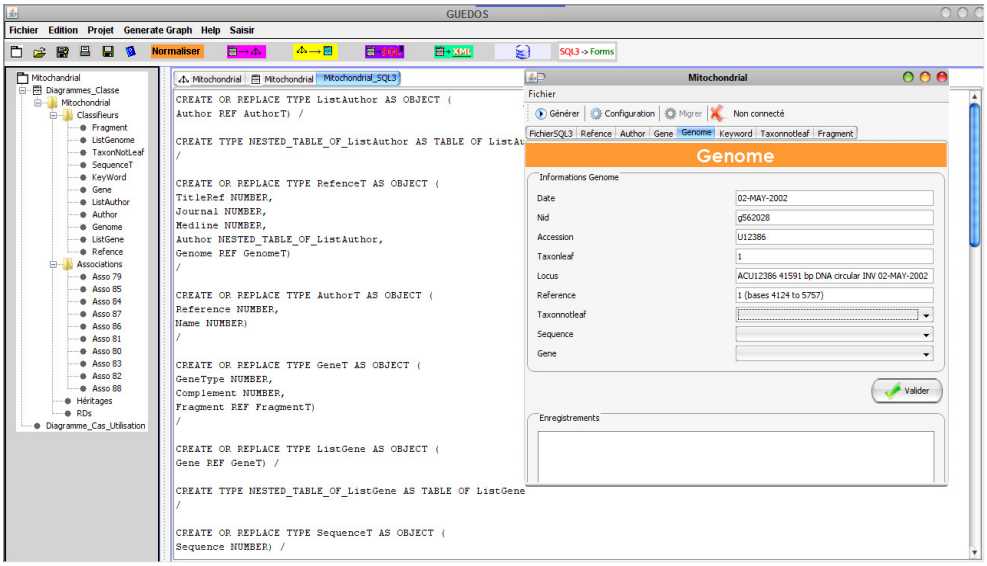


FIGURE 5: Example of query for mitochondrial sequences

5. CONSLUSION & FUTURE WORK

We have proposed a solution consisting on an integrated environment facilitating cohabitation of several models and techniques to sustain user when designing database schema. GUEDOS is specific to object-relational database schema design.

We have described the environment at whole, focusing at the same time on data static structure and dynamic process modeling. As perspectives, we tend to extend GUEDOS to characterize the obtained conceptual schema with respect to the previous processing, by introducing access methods, even denormalization, and to integrate several schemas conceptual schemas into on schema without lost of information. This work direction leads us to take more attention on optimization and design process using a self-optimization approach based on user preferences by selecting indexing methods, fragmentation and selection of views to materialize.

6. REFERENCES

[1] M. Stonebraker, P. Brown, 1999. Object-Relational DBMSs – Tracking the Next Great Ware, 2nd ed., Morgan Kaufmann, San Fransisco.

[2] Bertino E. and Martino L., 1993. Object Oriented Database Systems; Concepts and Architectures. Addison-Wesley Publishing Company Inc, (1993)

[3] Korab-Laskowska M., Pierre Rioux, Nicolas Brossard, Timothy G. Little-john1, Michael W. Gray2, B. Franz Lang, Gertraud Burger, 2001. The Organelle Genome Database Project (GOBASE). Nucleic Acids Research, volume 26, Issue 01: January 1 (2998) 138-144.

[4] Marie-Paule Lefranc, Véronique Giudicelli, Chantal Busin, Julia Bodmerl, Werner Müller, Ronald Bontrop, Marc Lemaitre, Ansar Malik, Denys Chaume, 1998. IMGT. The international ImMunoGeneTics database, Nucleic Acids Research, Volume 26, Issue 01: Jaunuary 1, 297-303.

[5] Moore, R., Lopes, J., 1999. Paper templates. In TEMPLATE'06, 1st International Conference on Template Production. SciTePress.

[6] Badir, H. and Pichat, E., 2005. An Interactive Tool for creative data modeling, in Database Technology and Applications for International Conference on Information Technology ITCC'05, IEEE, Las Vegas, Nevada

[7] Badir H., Tanacescu A., 2007. An efficient interface to handle complex structure for database design", ICEIS '07, Madeira, Potugal.

[8] Chang W., I.N. Shindyalov, C. Pu and P.E Bourne, 1994. Design and application of PDBlib, a C++ macromolecular class library. Computer Application in Biosciences, 10(6).

[9] Tateno Y., Kaoru Fukami-Kobayashi, Satoru Miyazaki Sugawara and Takashi Gojobori, 1999. DNA Data Bank of Japan at work on genome sequence data. Nucleic Acids Research 6 (19) 16-20.

[10] Lagesen K, et al. RN Ammer, 2007. Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35:3100–3108.

[11] Etzold T. and P. Argos. SRS, 1993. An indexing and retrieval tool for flat file data libraries. Computer Applications in the Biosciences. 9(1) 49-57.

[12] Stephen M. Beckstrom-Sternberg and D. Curtis Jamison, 1999. AGIS: Using the Agricultural Genome Information System, Bioinformatics: Databases and Systems, p 163-174

[13] Kim W., 1989. A model of queries for object-oriented databases. VLDB, pages 423-432.

[14] Kari S. And Rosenthal, A. G-WHIA, 1990. Conceptual Query Language-CQL: a visual user interface to application databases. IOS Press, pages 608-623.

[15] Lecluse, C. Richard, P. And Velez, F. O2, 1988. An Object-Oriented data model. EDBT, pages 556-562.

[16] D. Tania, Rahayu and Srivastava, , 2003, A Taxonomy for Object-Relational Queries, by IRM Press

[17] Loney, K. & Koch, G. (2002). *Oracle 9i: The Complete Reference.* Oracle Press.

INSTRUCTIONS TO CONTRIBUTORS

The *International Journal of Biometric and Bioinformatics (IJBB)* brings together both of these aspects of biology and creates a platform for exploration and progress of these, relatively new disciplines by facilitating the exchange of information in the fields of computational molecular biology and post-genome bioinformatics and the role of statistics and mathematics in the biological sciences. Bioinformatics and Biometrics are expected to have a substantial impact on the scientific, engineering and economic development of the world. Together they are a comprehensive application of mathematics, statistics, science and computer science with an aim to understand living systems.

We invite specialists, researchers and scientists from the fields of biology, computer science, mathematics, statistics, physics and such related sciences to share their understanding and contributions towards scientific applications that set scientific or policy objectives, motivate method development and demonstrate the operation of new methods in the fields of Biometrics and Bioinformatics.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJBB.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with volume 5, 2011, IJBB appears in more focused issues. Besides normal publications, IJBB intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

We are open to contributions, proposals for any topic as well as for editors and reviewers. We understand that it is through the effort of volunteers that CSC Journals continues to grow and flourish.

LIST OF TOPICS

The realm of International Journal of Biometrics and Bioinformatics (IJBB) extends, but not limited, to the following:

- Bio-grid
- Bioinformatic databases
- Biomedical image processing (registration)
- Biomedical modelling and computer simulation
- Computational intelligence
- Computational structural biology
- DNA assembly, clustering, and mapping
- Fuzzy logic
- Gene identification and annotation
- Hidden Markov models
- Molecular evolution and phylogeny
- Molecular sequence analysis
- Bio-ontology and data mining
- Biomedical image processing (fusion)
- Biomedical image processing (segmentation)
- Computational genomics
- Computational proteomics
- Data visualisation
- E-health
- Gene expression and microarrays
- Genetic algorithms
- High performance computing
- Molecular modelling and simulation
- Neural networks

CALL FOR PAPERS

Volume: 6 - **Issue:** 1 - February 2012

i. Paper Submission: November 30, 2011 **ii. Author Notification:** January 01, 2012

iii. Issue Publication: January / February 2012

CONTACT INFORMATION

Computer Science Journals Sdn BhD
B-5-8 Plaza Mont Kiara, Mont Kiara
50480, Kuala Lumpur, MALAYSIA

Phone: 006 03 6207 1607
006 03 2782 6991

Fax: 006 03 6207 1697

Email: cscpress@cscjournals.org

CSC PUBLISHERS © 2011
COMPUTER SCIENCE JOURNALS SDN BHD
M-3-19, PLAZA DAMAS
SRI HARTAMAS
50480, KUALA LUMPUR
MALAYSIA

PHONE: 006 03 6207 1607
006 03 2782 6991

FAX: 006 03 6207 1697
EMAIL: cscpress@cscjournals.org