ISSN 1870-4069

A distance measure for building phylogenetic trees: a first approach

Eunice Ponce-de-Leon-Senti[1], Elva Diaz[1], Hector Guardado-Muro[3],
Daniel Cuellar-Garrido[1], Juan José Martinez-Guerra[2], Aurora Torres-Soto[1],
Dolores Torres-Soto[4], Arturo Hernandez-Aguirre[5]

[1] Autonomous University of Aguascalientes, Computer Sciences Department, Ags., Mexico

[2] Autonomous University of Aguascalientes, Chemistry Department, Ags., Mexico

[3] Universidad Juárez Autónoma de Tabasco,
División Académica de Informática y Sistemas, Tabasco,
Mexico

[4] Autonomous University of Aguascalientes, Information Systems Department, Ags., Mexico

[5] CIMAT, A.C., Departamento de Ciencias Computacionales, Guanajuato, Gto., Mexico

eponce@correo.uaa.mx, ediazd@correo.uaa.mx, cuellar_garrido@correo.uaa.mx,
atorres@correo.uaa.mx, jjmartin@correo.uaa.mx,
hector_guardado_muro@outlook.com,
mdtorres@correo.uaa.mx, artha@cimat.mx

Abstract. We propose a distance measure for building phylogenetic trees in comparative genomics. The measure is based on the Bidirectional Best Hit (BBH) concept. The idea behind the measure is, insofar as that two organisms share more Bidirectional Best Hits (BBHs) they are more similar. Although, in general, the sizes of the genomes are different; a similarity measure between two organisms is defined having in the numerator the number of BBHs that exist between them, and in the denominator the semi-sum of the sizes of the genomes of both organisms in order to define a proportion. A distance measure is defined based on the similarity measure. Some restrictions on the number of BBHs between organisms are needed for fulfilling the triangle inequality. We apply different algorithms for building phylogenetic trees reported in literature, using the distance matrix as input, and we obtain suitable phylogenetic trees for a study case of 33 whole proteomes of fungi.

Keywords: distance measure, phylogenetic trees, bidirectional best hits, whole genome or proteome, comparative genomics.

pp. 149–162; rec. 2017-03-14; acc. 2017-05-06

1 Introduction

The many methods for reconstruction phylogenetic trees can be classified into three main categories: parsimony [7], [10], distance [23], [25], [24], and likelihood [8] methods. Our approach is based on the distance methods. The distance methods use the evolutionary distance between operational taxonomic units (in our case, species). Two organisms sharing a recent common ancestor be more similar to each other than two organisms whose last common ancestor was farther. By this reason, it should be possible to infer evolutionary relationships from similarities found between organisms. This is the principle that underlies the distance methods of phylogenetic reconstruction. A distance matrix is generated at each step. The distances represent the dissimilarity between each pair of taxa. The resultant matrix is used to generate a phylogenetic tree. The minimum evolution method [23] use the least-squares method (LSD) in order to give each generated tree a score. Then the tree with the lowest LSD can be found. The unweighted pair group method with arithmetic mean (UPGMA) [25] constructs a tree by identifying the shortest distance in the matrix, clustering those two taxa into a single Operational Taxonomic Unit for use in all subsequent calculations, calculating a new distance matrix, and then repeating these steps. The biggest disadvantage of the UPGMA method is that assumes equal rates of change in each lineage since they diverged from a common ancestor. The neighbor joining method [24] resembles the UPGMA method, but the most important difference is to allow unequal rates of evolution in different branches of the tree.

The concept of bidirectional best hit (BBH) in comparative genomics has had great attention in the last years, that is the case of Wolf and Koonin's paper [27]. They find out experimentally a tight link between orthologs and bidirectional best hits in the case of bacterial and archaeal genomes. Overbeek and his colleagues [19] defined and explained BBHs and used them to detect conserved clusters of genes in order to show that they give evolutionary advantages on individuals and populations. Wolf and Koonin [27] in 2012 said that the ortholog conjecture is the cornerstone of all functional annotation of sequenced genomes and introduced another conjecture, the BBHO (bidirectional best hits (BBH)-orthology equivalence conjecture).

The objective of this paper is to propose a distance, in order to build phylogenetic trees in comparative genomics. To attain this objective the first step is to define an indicator to link the evolutionary theory with the data at hand. In our case we use the bidirectional best hits (BBHs) concept [19], which induces on a set of genes, a reflexive and symmetric binary relation [22], but can not establishes a transitive relation between them. That is, if $\{x_r^i, x_s^j\}$ and $\{x_s^j, x_k^t\}$ are BBHs where x_r^i is a gene that belongs to genome G^i, x_s^j is a gene that belongs to genome G^j and x_k^t is a gene that belongs to genome G^t, then we can not affirm that $\{x_r^i, x_k^t\}$ is a BBH, in general, i.e., there is no guaranty of the existence of this latter pair of genes as a bidirectional best hit. However, we can define a metric based on the BBH concept in order to measure how many genes share two genomes under the BBH's relation. A distance matrix will be used then to construct phylogenetic trees from knowing distances based methods reported in literature [9].

The case of study, in this paper is a database of 33 whole proteomes of fungi, i.e., completely sequencing proteomes. The fungi are selected because they are very important organisms, that won their place as an object of fundamental research because they affect our daily lives as causative agents of disease, as sources of food, as agents for recycling of biomass, as key ingredients in industrial processes, as essentials tools in medicine (penicillin- that changed the population growth pattern), and as models to study properties of evolution.

Until 2006 most fungal phylogenies had been derived from single gene comparison, or from concatenated alignment of a small number of genes. After 2006 the availability of greater number of data laid the basis to reconstruct phylogenetic trees from whole genomes. Fitzpatrick and his colleagues in 2006 used 42 whole genomes and agglomerative methods to construct a phylogenetic tree [11]. Wang and his colleagues in 2009 [26] had constructed a kingdom-wide fungal phylogenetic tree for 82 sequenced genomes using an alignment free composition method (CV) previously successfully applied to prokaryotik and viral phylogenics [21]. The method is based on whole genomes [20].

The next section discusses the fundamental concepts and notation that establish a relationship between a biological problem, i.e. the building of phylogenies and the mathematical formalism. In the section 3 the properties of the distance measure are proven. The section 4 explains the steps to follow, in order to apply the genomic distance measure. The section 5 gives the results and discussion in the case of study, and finally, the section 6 gives the conclusions and future work.

2 Basic Concepts and Notation

In order to formulate mathematically the biological problem of finding a phylogenetic tree in comparative genomics we have defined some involved concepts from molecular biology. The comparative genomics is a very powerful tool in Molecular Biology [14] and in particular, the phylogenetic reconstruction has become one of the main tools of comparative genomics [16], especially, in the case of the enormous amounts of data generated by several molecular biology methods. A phylogeny is, in general, the evolutionary history of a group of related entities. A genome is the entire genetic constitution of a living organism. We define here the concepts of a whole genome, the genome's size, a BBH, the set of all BBHs between two genomes and the genomic distances matrix. In literature, whole genome is a genome that is sequenced completely. We define a whole genome as a n-tuple of ordered genes, and genome's size as the number of genes of the genome, i.e., n. Henceforth, we will give all definitions using the genome concept, but these can be reformulated mathematically for the proteome concept.

Definition 1. *A whole genome is defined as a n-vector of genes, $G = (x_1, ..., x_n)$. The number of genes in the whole genome represents the genome's size, i.e., n. Let $G^i = (x^i_1, ..., x^i_{n_i})$ and $G^j = (x^j_1, ..., x^j_{n_j})$ be genomes, n_i and n_j denote the size of genomes G^i and G^j respectively.*

Henceforward, when we refer to a genome, we are supposing it is a whole-genome.

Definition 2. *For all genomes G^i the number of genes n_i in G^i satisfies that*

$$n_i \in \mathbf{Z}^+ - \{0\} \; . \tag{1}$$

This is trivial for modelling the biological problem because there is no reason to consider a genome G^i, if this genome do not have genes. The values that the number of genes n_i can take, are positives and integers.

Definition 3. *The BBH definition [19] is as follows: Let G^i and G^j be genomes. Two genes, x_r^i , the r-th element of the n_i-tuple G^i, and x_s^j, the s-th element of the n_j-tuple G^j, are called a Bidirectional Best Hit (BBH), if and only if recognizable similarity exists between them (in our case, we use the BLAST similarity score with a E-value [12] lower than* 1.0 x 10^{-5}*), and there is no gene x_k^j of the n_j-tuple G^j that is more similar than x_s^j is to x_r^i, and there is no gene x_k^i of the n_i-tuple G^i that is more similar than x_r^i is to x_s^j. We denote it as, $x_r^i \leftrightarrow_{BBH} x_s^j$.*

Definition 4. *The set of all BBHs between two genomes G^i and G^j is the set*

$$B_{i,j} = \{\{x_r^i, x_s^j\} : x_r^i \leftrightarrow_{BBH} x_s^j\}, \tag{2}$$

and the cardinality of $B_{i,j}$ is the number of BBHs that exist between G^i and G^j, and is denoted by $|B_{i,j}|$.

Definition 5. *For all pair of genomes G^i and G^j the number of BBHs, that is, $|B_{i,j}|$ satisfies that $|B_{i,j}| \in \mathbf{Z}^+$ and it can not exceed the genes number of the shortest genome between both, i.e.,*

$$|B_{i,j}| \leq min\{n_i, n_j\} \; , \tag{3}$$

where n_i and n_j are the genomes size of G^i and G^j respectively.

In the first part of this definition, we have considered that the number of BBHs between two genomes G^i and G^j is a positive number that can be zero. This refers to that do not exist any pair of genes taking one of G^i, and the other of G^j that they are a BBH. Of course, $B_{i,j}$ is a positive number because it is related with the number of genes of G^i and G^j. The second part of this definition about the upper bound of BBHs number between two genomes means that the upper bound of BBHs number is the size of the smallest genome.

Definition 6. *The genomic distances matrix between a set of genomes is a symmetric matrix constructed for all organisms of the study, from a distance measure between all pair of genomes of these organisms.*

3 Proposed Distance Measure

In this part of the chapter, the distance measure for building phylogenetic trees from genomic information is introduced. This distance measure is made from a similarity measure, based on bidirectional best hits [19]. The idea behinds of this is to consider the following premise: two organisms that share more BBHs than other two, should be more similar. To model the real biological problem, the different sizes of the genomes of the involved organisms should be considered in the definition of the measure. The measures defined here are measures between whole genomes.

Definition 7. *A similarity measure δ between G^i and G^j is defined as follows:*

$$\delta(G^i, G^j) = \frac{2|B_{i,j}|}{n_i + n_j}, \tag{4}$$

$|B_{i,j}|$ is the total number of BBHs between G^i and G^j. n_i is the total number of genes from genome G^i, and n_j is the total number of genes from genome G^j.

The similarity measure is standardized with respect to the size of their genomes in order to eliminate the effect of the difference of genomes sizes. The dissimilarity measure is defined as additive inverse of the similarity measure [5] as follows:

Definition 8. *A dissimilarity measure d between G^i and G^j is given by*

$$d(G^i, G^j) = 1 - \delta(G^i, G^j) = 1 - \frac{2|B_{i,j}|}{n_i + n_j}. \tag{5}$$

In the following we demonstrate that $d(G^i, G^j)$ obeys the following four properties, i. e., $d(G^i, G^j) \geq 0$, reflexivity, symmetry and triangle inequality.

Property 1. The dissimilarity measure between two genomes G^i and G^j is a positive number or is equal to zero.

Proof.

$$d(G^i, G^j) = 1 - \delta(G^i, G^j) = 1 - \frac{2|B_{i,j}|}{n_i + n_j} \tag{6}$$

$d(G^i, G^j)$ is a positive number or is equal to zero, if and only if,

$$1 \geq \delta(G^i, G^j) \geq 0, \tag{7}$$

That is

$$1 \geq \frac{2|B_{i,j}|}{n_i + n_j} \geq 0. \tag{8}$$

Since Definition 5

$$|B_{i,j}| \leq min\{n_i, n_j\}, \tag{9}$$

because $|B_{i,j}|$ is the number of BBHs between G^i and G^j, and this number can not overtake the number of genes of the smallest genome because of BBH's definition. Suppose $n_i \leq n_j$, that is, $|B_{i,j}| \leq n_i$ by Definition 5. The following steps are true:

$2|B_{i,j}| \leq 2n_i = n_i + n_i$ substitute the second n_i by n_j knowing $n_i \leq n_j$,
$2|B_{i,j}| \leq n_i + n_j$ dividing both sides by $n_i + n_j$, where

$$\frac{2|B_{i,j}|}{n_i + n_j} \leq 1. \tag{10}$$

An analogous demonstration is possible as $n_j \leq n_i$.

Demonstrate that $\delta(G^i, G^j) \geq 0$, i.e., $\frac{2|B_{i,j}|}{n_i+n_j} \geq 0$. $|B_{i,j}| \in \mathbf{Z}^+$ by BBH's definition and also $2|B_{i,j}| \in \mathbf{Z}^+$. n_i and $n_j \in \mathbf{Z}^+$. The quotient of positive integer numbers belong to $\mathbf{R}^+$. Finally, $n_i \neq 0$ and $n_j \neq 0$ by Definition 2. Finally

$$0 \leq \frac{2|B_{i,j}|}{n_i + n_j} \leq 1. \tag{11}$$

Property 2. Reflexivity property: For every genome G^i, $d(G^i, G^i) = 0$.

Proof.

$$d(G^i, G^i) = 1 - \delta(G^i, G^i) = 1 - \frac{2|B_{i,i}|}{n_i + n_i} = 1 - \frac{2n_i}{2n_i} = 1 - 1 = 0. \tag{12}$$

$|B_{i,i}| = n_i$ because the number of BBHs of a genome G^i with itself is the same as the number of genes the genome G^i has.

Property 3. Symmetry property: For all pair of genomes, G^i and G^j, $d(G^i, G^j) = d(G^j, G^i)$.

Proof.

$$d(G^i, G^j) = 1 - \delta(G^i, G^j) = 1 - \frac{2|B_{i,j}|}{n_i + n_j} = 1 - \frac{2|B_{j,i}|}{n_j + n_i} = 1 - \delta(G^j, G^i) = d(G^j, G^i). \tag{13}$$

$|B_{i,j}| = |B_{j,i}|$ because the number of BBHs of G^i with G^j is the same as the number of BBHs of G^j with G^i, due the BBH's definition.

Property 4. Triangle inequality property: Let G^i, G^j and G^k be the genomes in a database, and let n_i, n_j, n_k be the genome sizes respectively. Since $n_i, n_j, n_k \in \mathbf{Z}^+ - \{0\}$ and $|B_{i,k}|, |B_{k,j}|, |B_{i,j}| \in \mathbf{Z}^+$ and suppose that the following expressions are fulfilled,

$$|B_{ik}| = \frac{n_i + n_k}{2}\alpha \leq min\{n_i, n_k\}, \tag{14}$$

$$|B_{kj}| = \frac{n_k + n_j}{2}\beta \leq min\{n_k, n_j\}, \tag{15}$$

$$0 \leq \alpha + \beta \leq 1\ , \tag{16}$$

and then the triangle inequality $d(G^i, G^j) \leq d(G^i, G^k) + d(G^k, G^j)$ is fulfilled. Notice that $\alpha \geq 0$ and $\beta \geq 0$ because in the suppositions (14) and (15), the left side of both equations is a positive number, that is, $|B_{ik}| \in \mathbf{Z}^+$ and $|B_{kj}| \in \mathbf{Z}^+$ as we show in definition 5.

Proof. Since the suppositions in (14) and (15) we have

$$2|B_{ik}| = (n_i + n_k)\ \alpha, \tag{17}$$

$$2|B_{kj}| = (n_k + n_j)\ \beta, \tag{18}$$

that is

$$1 - \frac{2|B_{ik}|}{n_i + n_k} = 1 - \alpha, \tag{19}$$

$$1 - \frac{2|B_{kj}|}{n_k + n_j} = 1 - \beta. \tag{20}$$

Adding (19) and (20),

$$1 - \frac{2|B_{ik}|}{n_i + n_k} + 1 - \frac{2|B_{kj}|}{n_k + n_j} = 2 - \alpha - \beta. \tag{21}$$

Since the supposition (16) we have

$$2 - (\alpha + \beta) \geq 1. \tag{22}$$

Finally, using inequality (11), we obtain that the triangle inequality

$$1 - \frac{2|B_{ik}|}{n_i + n_k} + 1 - \frac{2|B_{kj}|}{n_k + n_j} \geq 1 \geq 1 - \frac{2|B_{ij}|}{n_i + n_j}, \tag{23}$$

i.e.

$$d(G^i, G^j) \leq d(G^i, G^k) + d(G^k, G^j) \tag{24}$$

is fulfilled for all $n_i, n_j, n_k \in \mathbf{Z}^+ - \{0\}$ and $|B_{i,k}|, |B_{k,j}|, |B_{i,j}| \in \mathbf{Z}^+$ and for all G^i, G^j and G^k that satisfy (14), (15) and (16). The property has been demonstrated.

We have demonstrated that the dissimilarity measure d defined in Definition 8 is a distance measure. In the following we refer d as a distance measure.

The suppositions (14), (15) and (16) made for the proof of triangle inequality property means that the number of BBHs for any two pairs (i, k) and (k, j) of organisms will be always less than the size of the shortest genome, in a factor α and a factor β respectively, and $0 \leq \alpha + \beta \leq 1$.

4 Materials and Methods

In this section we describe the method for applying the proposed distance measure for whole genomes in order to build phylogenetic trees using the proposed distance as the input of several distance based algorithms reported in literature [9]. The first step is to define the set of organisms in the study, and to establish a research hypothesis to test. The second step is to obtain the whole genomes or whole proteomes of these organisms from an appropriated database in the web. The third step is to obtain the BBHs for all pair of organisms using BLAST [1]. The fourth step is to make the genomic distance matrix using Definition 5 and finally, to run the algorithms that build the phylogenetic trees and to analyse the results with respect to the research hypothesis.

4.1 Data Description

In this study, the fungal phylogenetic trees are made with whole proteomes of 33 representative fungi from the following phyla: one organism of Zygomycota phylum, two of Chytridiomycota phylum, four of Basidiomycota phylum, and 26 of Ascomycota phylum. The proteomes used (see Table 1) are obtained from the Broad Laboratory [3], the Bordeaux Bioinformatics Center (CBiB) [2], and The European Bioinformatics Institute [6]. Each one is selected because it is a whole proteome.

Table 1 contains the list of the 33 fungal organisms for this study detailing its genus, species, variety, proteome size, and and the institute from where the data was obtained. Table 2 contains information important for reading and interpreting the obtained phylogenetic tree, such as the taxonomy in terms of the phylum, subphylum, class and order for all fungal proteomes in study.

4.2 Preprocessing of Data

A part of the data analysis is realized using BLAST (Basic Local Alignment Search Tool) [1] a free program from the National Center for Biotechnology Information (NCBI) [17]. This program performs different types of analysis. In this case the program is used to compare proteins that belongs to each pair of fungi. This program is used to obtain for each protein of the first fungus the one of the second fungus that best resembles it. This correspondence is denoted as a “best hit”. The analysis is performed then in reversed order, and if the same protein that was a best hit is given back as a best hit too, it is said that a “bidirectional best hit” has occurred (see Definition 3) if the BLAST similarity scores in both directions are lower than 1.0 x 10^{-5}. The fundamental basis to the construction of the phylogenetic tree in our approach is to obtain all the BBHs existing between every pair of fungal proteomes.

The cut off point for determining which a best hit is and what is not, is determined by their expectation value. The expectation value (or E) in BLAST is a statistical significance threshold for reporting matches against database sequences. The typical value of E found in the literature for obtaining the BBHs is of 1.0 x 10^{-5} due the necessity of assessing the best resemblance between proteins of the species, not allowing alignments that appear very similar at first sight but, in closer examination they are not because they were well aligned by chance [4].

Table 1. Fungal organisms used in this analysis are listed.

No.	Identifier	Genus	Species	Variety	Proteins	Citation
1.	ASHBYA	*Ashbya*	*gossypii*	Q	4718	SP
2.	ASPERF	*Aspergillus*	*fumigatus*	Afu	9888	BROAD
3.	ASPERN	*Aspergillus*	*nidulans*	AN	10665	BROAD
4.	ASPERT	*Aspergillus*	*Terreus*	ATEG	10406	BROAD
5.	BATRAC	*Batrachochytrium*	*dendrobatidis*	BDEG	8818	BROAD
6.	BOTRYT	*Botrytis*	*cinerea*	BC1G	16389	BROAD
7.	CANDAL	*Candida*	*albican*	CAWG	6157	BROAD
8.	CANDGL	*Candida*	*glabrata*	CAGR	5215	GNL
9.	CANDGU	*Candida*	*guilliermondii*	PGUG	5920	BROAD
10.	CANDLU	*Candida*	*lusitaniae*	CLUG	5936	BROAD
11.	CANDTR	*Candida*	tropicalis	CTRG	6258	BROAD
12.	CHAETO	*Chaetomium*	globosum	CHGG	11124	BROAD
13.	COCCID	*Coccidiodes*	*immitis*	CIMG	10457	BROAD
14.	COPRIN	*Coprinus*	*cinereus*	CC1G	13544	BROAD
15.	CRYPTO	*Cryptococcus*	*neoformans*	CNAG	7302	BROAD
16.	DEBARY	*Debaryomyces*	*hansenii*	DEHA	6319	GNL
17.	FUSAGR	*Fusarium*	*graminearum*	FGSG	13321	BROAD
18.	FUSAOX	*Fusarium*	*oxysporum*	FOXG	17608	BROAD
19.	FUSAVE	*Fusarium*	*verticilloides*	FVEG	14195	BROAD
20.	HISTOP	*Histoplasma*	*capsulatum*	HGAC	9349	BROAD
21.	KLUYVE	*Kluyveromyces*	*lactis*	KLLA	5327	GNL
22.	LODDER	*Lodderomyces*	*elongisporus*	LELG	5796	BROAD
23.	MAGNAP	*Magnaphorte*	*grisea*	MGG	12832	BROAD
24.	NEUROS	*Neurospora*	*crassa*	NCU	9823	BROAD
25.	PUCCIN	*Puccinia*	*graminis*	PGTG	20567	BROAD
26.	RHIZOP	*Rhizopus*	*oryzae*	RO3G	17467	BROAD
27.	SACCHA	*Saccharomyces*	*cerevisiae*	SCRG	5388	BROAD
28.	SCHIZO	*Schizosaccharomyces*	*japonicus*	SJAG	5168	BROAD
29.	SCLERO	*Sclerotinia*	*sclerotiorum*	SS1G	14522	BROAD
30.	STAGON	*Stagonospora*	*nodorum*	SNU	16597	BROAD
31.	UNCINO	*Uncinocarpus*	*reesii*	UREG	7798	BROAD
32.	USTILA	*Ustilago*	*maydis*	UM	6522	BROAD
33.	YARROW	*Yarrowia*	*lipolytica*	YALI	6436	GNL

We take each pair combination P^i, P^j for every 33 whole proteomes in BLAST being the proteins of the fungus P^i the query and the proteins of the fungus P^j the database in order to obtain 1056 files of best hits. The next step is to take every pair of files in which we stored the best hits between P^i and P^j from $i \rightarrow j$ and from $j \rightarrow i$ to obtain the bidirectional best hits. After preprocessing the BBHs between all pairs of fungi, 528 independent files are obtained. This number corresponds to the number of necessary comparisons in order to obtain the BBH's number for each pair of fungi. When this magnitude is obtained for all fungi pair, it is possible to obtain the genomic distances matrix (see Definition 6) using the distance measure (Definition 5).

A computer program tests the accomplishment of restrictions made in the proof of triangle inequality, that is, the restrictions (14), (15) and (16) for all pair of organisms in our database, i.e., for any three fungi, the triangle inequality for the distance measure is fulfilled.

5 Results and Discussion

MEGA7 software [13] is used for running different phylogenies with its distance based algorithms. The genomic distances matrix with the distance values for

Table 2. The usual fungal taxomomy. The zygomycota and chytridiomycota phyla, have only one fungus respectively. The basidiomycota phylum has 4 fungi. The rest of the fungi belongs to ascomycota phylum, and are classified in three subphyla: pezizomycotina, saccharomycotina, and taphrinomycotina. This taxonomy corresponds to [11], and [26].

Identifier	Phylum	Subphylum	Class	Order
ASHBYA	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
ASPERF	Ascomycota	Pezizomycotina	Eurotiomycetes	Eurotiales
ASPERN	Ascomycota	Pezizomycotina	Eurotiomycetes	Eurotiales
ASPERT	Ascomycota	Pezizomycotina	Eurotiomycetes	Eurotiales
BATRAC	Chytridiomycota	Chytridiomycotina	Chytridiomycetes	Chytridiales
BOTRYT	Ascomycota	Pezizomycotina	Leotiomycetes	Helotiales
CANDAL	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
CANDGL	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
CANDGU	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
CANDLU	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
CANDTR	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
CHAETO	Ascomycota	Pezizomycotina	Sordariomycetes	Sordariales
COCCID	Ascomycota	Pezizomycotina	Eurotiomycetes	Onygales
COPRIN	Basidiomycota	Agaricomycotina	Hymenomycetes	Agarigales
CRYPTO	Basidiomycota	Agaricomycotina	Hymenomycetes	Tremellales
DEBARY	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
FUSAGR	Ascomycota	Pezizomycotina	Sordariomycetes	Hypocreales
FUSAOX	Ascomycota	Pezizomycotina	Sordariomycetes	Hypocreales
FUSAVE	Ascomycota	Pezizomycotina	Sordariomycetes	Hypocreales
HISTOP	Ascomycota	Pezizomycotina	Eurotiomycetes	Onygales
KLUYVE	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
LODDER	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
MAGNAP	Ascomycota	Pezizomycotina	Sordariomycetes	Magnaporthales
NEUROS	Ascomycota	Pezizomycotina	Sordariomycetes	Sordariales
PUCCIN	Basidiomycota	Pucciniomycotina	Pucciniomycetes	Pucciniales
RHIZOP	Zygomycota	Mucoromycotina	Zygomycetes	Mucorales
SACCHA	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
SCHIZO	Ascomycota	Taphrinomycotina	Schizosaccharomycetes	Schizosaccharomycetales
SCLERO	Ascomycota	Pezizomycotina	Leotiomycetes	Helotiales
STAGON	Ascomycota	Pezizomycotina	Dothideomycetes	Pleosporales
UNCINO	Ascomycota	Pezizomycotina	Eurotiomycetes	Onygales
USTILA	Basidiomycota	Ustiloginomycotina	Ustilaginomycetes	Ustilaginales
YARROW	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales

each pair of proteomes of fungi is the input to MEGA 7 software. We use three distance based methods for building phylogenetic trees: the UPGMA, the Neighbor-Joining and the Minimum Evolution.

In Fig. 1 the phylogenetic tree using the UPGMA Method is presented [25]. The evolutionary history was inferred from the UPGMA method. The optimal tree with the sum of branch length = 7.06527874 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances are calculated from Definition 5.

In Fig. 2 the phylogenetic tree using the Neighbor-Joining method is presented. The evolutionary history is inferred using the Neighbor-Joining [24] method. The optimal tree with the sum of branch length = 7.09372415 is shown. The tree is drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances are calculated from Definition 5.

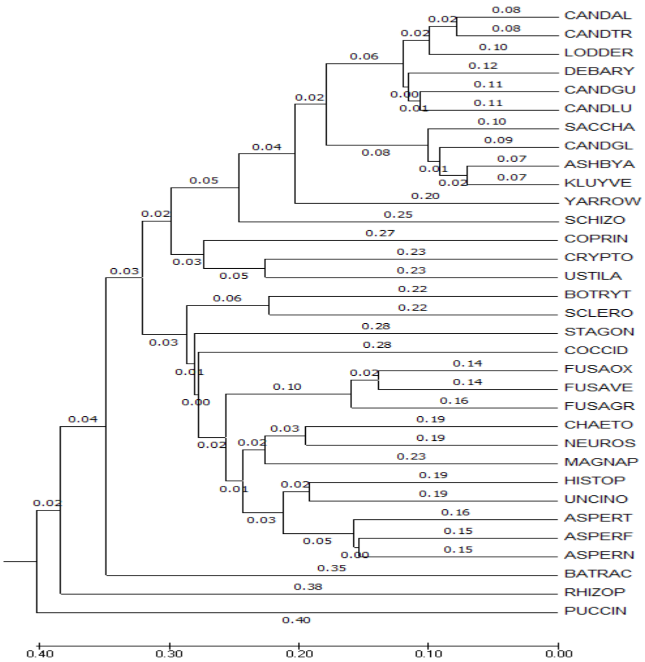

Fig. 1. Phylogenetic tree using UPGMA method.

In Fig. 3 the phylogenetic tree using the Minimum Evolution method is presented. The evolutionary history is inferred using the Minimum Evolution method [23]. The optimal tree with the sum of branch length = 7.09372415 is shown. The tree is drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances are calculated from Definition 5. The Minimum Evolution tree is searched using the Close-Neighbor-Interchange (CNI) algorithm [18] at a search level of 1. The Neighbor-joining algorithm [24] is used to generate the initial tree.

For all methods we can observe that the Saccharomycotina subphylum has been well identified respect to the main clades of subphylum, CTG and WDG in all phylogenetic trees. All organisms in our database that belong to Saccharomycotina subphylum are of the same order, that is, saccaromycetales. The members of CTG clade in our database are CANDLU, CANDGU, DEBARY, LODDER, CANDTR, and CANDAL. The members of WGD clade in our database are SACCHA, CANDGL, KLUYVE, and ASHBYA.

The Pezizomycotina subphylum has five different orders, Eurotiales, Helotiales, Sordariales, Onygales and Hypocreales. In all mentioned orders for all phylogenetic trees obtained, the topologies of the order's subtrees are the same as reported in [26].

In our database four fungi belong to the Basidiomycota phylum, they have

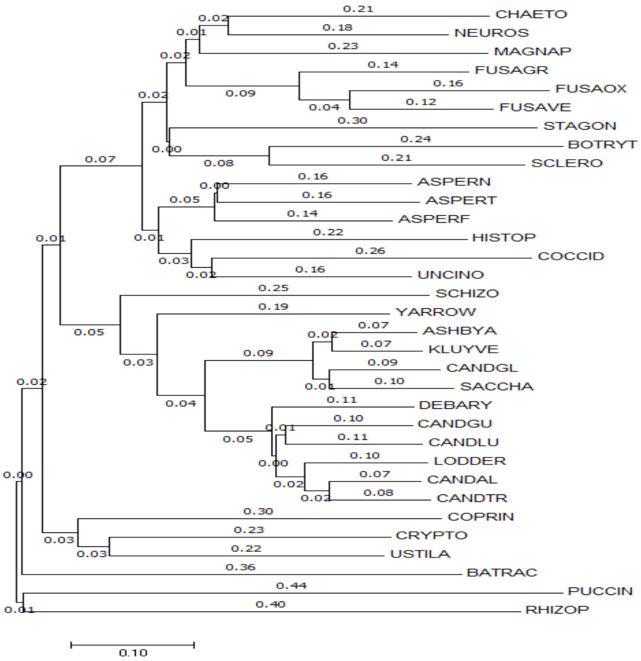

Fig. 2. Phylogenetic tree using Neighbor-Joining method.

been classified well in the Neighbor-Joining and Minimum Evolution methods. Only in the UPGMA method three fungi belonging to the Basidiomycota phylum (USTILA, CRIPTO and COPRIN see Table 2) have been classified wrong. On the other hand, the RHIZOP fungus belonging to the Zygomyccota phylum obtains an adequate position in the topology of the phylogenetic tree. The same situation occurs in the case of BATRAC fungus, which is classified as belonging to the Chytridiomycota phylum. In the case of the Taphrinomycotina subphylum exits the discussion about two different places where to assign it [15]. We obtain that the SCHIZO fungus, single representative of Taphrinomycotina subphylum in our study, is branching as a sister group to Saccharomycotina.

6 Conclusions and Future Work

The dissimilarity measure d between whole genomes has been demonstrated as a distance measure and it satisfies the four properties of a distance, i.e., the measure is a positive number or equal to zero, it is reflexive, symmetric and fulfils the triangle inequality if the suppositions (14), (15) and (16) are supported. Specially, the distance measure d is less than or equal to 1.

The resulting phylogenetic trees are in agreement with the most part of topologies and groups reported in the literature, for example, the fungal phylogenies

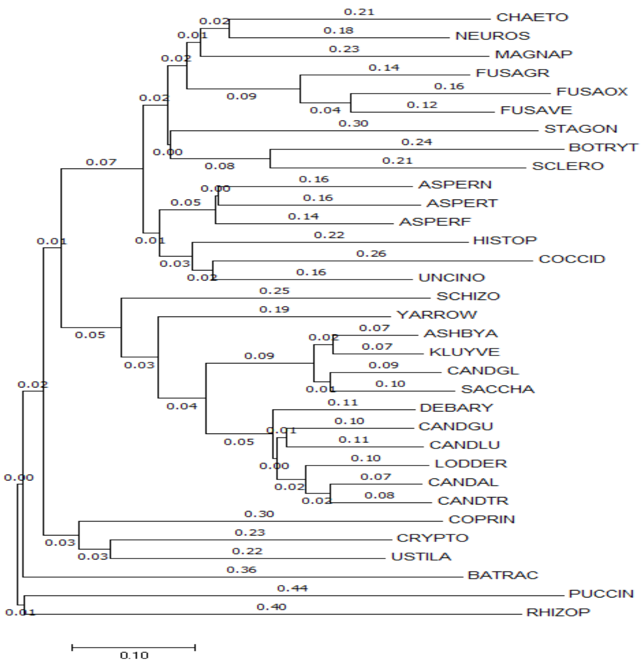

Fig. 3. Phylogenetic tree using Minimum Evolution method.

obtained in [11], and [26]. As seen in Table 2 where appears the taxonomy in terms of the phylum, subphylum, class and order for all fungal proteomes in study, our resulting phylogenetic trees classify correctly the Pezizomycotina subphylum until the order level. In the case of Saccharomycotina subphylum two main clades are identified. The Neighbor-Joining and Minimum Evolution methods obtain very similar phylogenies and the best results. They are also very similar to phylogenies reported in [11], and [26].

The future work will be to test if incorporating BBH - based phylogenetic tree structural information contributes to study the different functional groups of proteins for a set of organisms in study.

Acknowledgments. We would like to acknowledge the support for this project (PIINF15-1) given by the Autonomous University of Aguascalientes, Aguascalientes, Mexico.

References

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J.: Basic local alignment search tool. Journal of Molecular Biology 215(3):403–410 (1990)
2. Bordeaux Bioinformatics Center (CBiB), http://proteome.cgfb.u-bordeaux.fr

3. Broad Laboratory, http://archive.broadinstitute.org/ftp/pub/annotation/fungi
4. Brenner, S.E., Chothia, C., Hubbard, T. J.: Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc. Natl. Acad. Sci. USA, 95(11):6073–6078, (1998)
5. Deza, M. M., Deza,E.: Encyclopedia of Distances. Springer-Verlag (2013)
6. European Bioinformatics Institute,
ftp://ftp.ebi.ac.uk/pub/databases/integr8/last_release/fasta/proteomes
7. Farris, J. S.: Methods for computing Wagner trees. Systematic Zoology 19:83–92 (1970)
8. Felsenstein, J.: Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. Journal of Molecular Evolution, 17(6):368–376 (1981)
9. Felsenstein, J.: Inferring Phylogenies. Sinauer Associates, Inc. (2004)
10. Fitch, W.M.: Toward defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology 20(4):406–416 (1971)
11. Fitzpatrick, D. A., Logue, M.E., Stajich, J. E., Butler, G.: A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evolutionary Biology 6(99):1–15 (2006)
12. Gertz, E. M.: BLAST Scoring Parameters. Available at:ftp://ftp.ncbi.nlm.nih.gov/blast/documents/developer/scoring.pdf (2005)
13. Kumar, S., Stecher, G., Tamura, K.: MEGA7 Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Molecular Biology and Evolution 33:1870–1874 (2016)
14. Lackie, J. M.: The dictionary of cell and Molecular Biology. Academic Press Elsevier (2007)
15. Liu, Y., Leigh, J. W., Brinkmann, H., Cushion, M. T.: Phylogenomic Analyses Support the Monophyly of Taphrinomycotina, including Schizosaccharomyces Fission Yeast. Mol. Biol. Evol. 26(1):27–34 (2009)
16. Mushegian, A. R.: Foundations of Comparative Genomics. Elsevier Academic Press (2007)
17. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov
18. Nei, M., Kumar, S.: Molecular Evolution and Phylogenetics. Oxford University Press, New York (2000)
19. Overbeek, R., Fonstein, M., D'Souza, M., Push, G. D., Maltsev, N.: The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. 96(6):2896–2901 (1999)
20. Qi, J., Luo, H., Hao, B.: CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res. 32:W451–W47 (2004)
21. Qi, J., Wang, B., Hao, B.: Whole Proteome Prokaryote Phylogeny without Sequence Alignment: A K-String Composition Approach. J. Mol. Evol. 58:1–11 (2004)
22. Rosen, K. H.: Discrete Mathematics and Its Applications. McGraw-Hill (1999)
23. Rzhetsky, A., Nei, M.: A simple method for estimating and testing minimum evolution trees. Molecular Biology and Evolution 9:945–967 (1992)
24. Saitou, N., Nei, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406–425 (1987)
25. Sneath, P.H.A., Sokal, R.R.: Numerical Taxonomy. Freeman, San Francisco (1973)
26. Wang, H., Xu, Z., Gao, L., Hao, B.: A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evolutionary Biology 9(195):1–13 (2009)
27. Wolf, Y. I., Koonin, E. V.: A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol Evol. 4(12):1286–1294 (2012)