A LOCAL-GLOBAL GENE COMPARISON FOR ORTHOLOG DETECTION IN TWO CLOSELY RELATED EUKARYOTES SPECIES

Deborah Galpert Cañizares*, Michel Estopiñales Blay*, Reinier Millo Sánchez*, Claudia Companioni Brito*, Miguel Angel Fernández Marin**, and Carlos Morell Pérez**
*Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Cuba
**Universidad de Ciencias Informáticas y Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Cuba.

ABSTRACT
Ortholog detection has included the comparison of different gene features to build a phylogenetic tree or the initial genome correspondence graph. Many pre-processing procedures have been applied to prune graph structures before the clustering of potential orthologs. Then, some post-processing improvements have contributed in (>90%) of precision. Although, some algorithms yield high levels of precision, it is still the main target for comparative genomics community. In this paper, we present an ortholog detection algorithm which combines sequence homology, length and global genomes rearrangements into a novel local-global gene dissimilarity measure for the comparison of two closely related eukaryotes species. We use Locally Collinear Blocks reported by the "Multiple Alignment of Conserved Genomic Sequence with Rearrangements" software (MAUVE) to take into account global genome rearrangements. We build a weighted undirected complete bipartite graph representing the comparison of the two genomes with the global gene dissimilarity measure. The pre-processing step eliminates all edges with weight over 20% of the minimum weight. Next, we resolve ambiguities by keeping matches within synteny blocks. Finally, in the clustering process we search for Best Unambiguous Subsets representing homology groups and pairs of orthologs. We present an experiment with S. Cerevisiae and S. Bayanus with 98.45% of true classifications.

KEYWORDS: Ortholog Detection Algorithms, Similarity Measures, Bipartite Graph Partitioning

MSC: 68W25

RESUMEN
La detección de ortólogos ha incluido la comparación de diferentes rasgos de los genes para construir un árbol filogenético o un grafo de correspondencia entre genomas. Se han aplicado múltiples procedimientos de pre-procesamiento para podar las estructuras de grafos antes de agrupar los ortólogos potenciales. Además algunas mejoras de post-procesamiento han contribuido a (>90%) de precisión. A pesar de que algunos algoritmos arrojan altos niveles de precisión, ésta continúa siendo el principal objetivo de la comunidad científica que trabaja en genómica comparativa. En este trabajo presentamos un algoritmo de detección de ortólogos que combina la homología de las secuencias, la longitud y los reordenamientos globales en una nueva medida de disimilaridad entre genes local-global para la comparación de dos especies de eucariotas estrechamente relacionadas. Para tener en cuenta los reordenamientos globales de los genomas, utilizamos los Bloques Localmente Colineales reportados por el software de alineamiento múltiple de secuencias genómicas conservadas con reordenamientos "Multiple Alignment of Conserved Genomic Sequence with Rearrangements" (MAUVE). Construimos un grafo bipartito completo que representa la comparación entre los dos genomas con las medidas de disimilaridad globales entre los genes. El paso de pre-procesamiento elimina todos los arcos con peso por encima del 20% del mínimo peso. Luego resolvemos las ambigüedades conservando las correspondencias dentro de los bloques de orden conservado. Finalmente, en el paso de agrupamiento, buscamos los mejores subconjuntos no ambiguos que representan los grupos de homología y los pares de ortólogos. Presentamos un experimento con S. Cerevisiae y S. Bayanus con 98.45% de clasificaciones verdaderas.

1. INTRODUCTION

Genetic changes might be as subtle as mutations, insertions or deletions of individual nucleotides but as drastic as duplication or lost of chromosomal segments, entire chromosomes or complete genomes. Global genome rearrangements may be possible due to inversions, translocations, fusions and fissions. The resulting differences considering behaviour and chromosome organization may reduce some capabilities in sub-populations or emerging species.

On comparing various species we can observe some homologies in their genetic composition and in their gene organization. A lot of genes known as *orthologs* keep their sequence homology and function through different species starting from a common ancestor. Orthologs evolved by speciation while some other homolog genes known as *paralogs* evolved by duplication. *Inparalogs* are genes within species that duplicated after the

speciation event, while *outparalogs* duplicated before the speciation event. In genome comparison we can study groups of neighbour genes that preserve local organization (order and distance) throughout evolution. These groups are called *synteny blocks*.

Many studies have classified sets of orthologous sequences among annotated species, and therefore many multispecies eukaryotic databases of orthologous groups are available, for instance, NCBI euKaryotic Orthologous Group database (KOG) [54],ORTHOMCL_DB [6], INPARANOID [44], Eukaryotic Gene Orthologues database (EGO) [34], YOGY [48], ROUNDUP [10] and ORTHOINSPECTOR [36]. Some tools such as BLASTO [33] allow for a sequence search in a database set in order to assign this sequence to an orthologous group. These databases have been mainly developed through phylogeny-based orthology inference, all-against-all sequence comparison or hybrid methods [15], [29], [24] that include additional information such as protein interactions, synteny data and protein domains.

The phylogenetic approach [22], [61] includes the homolog clustering, the generation of correct multiple alignments for each group of homolog domain, the construction of a phylogenetic tree for each group, and finally, the extraction of orthologs from these trees. The tree methods typically reconcile gene and species trees in order to assign duplication and speciation nodes, as well as detect gene losses.

On the other hand, the all-against-all approach focuses on building ortholog groups by clustering pair-wise gene relationships mainly obtained from sequence similarity measures: asymmetric BLAST raw score [2], symmetric SW score [23], or symmetric BLAST E-value [1]. It is based on the fact that orthologs should be more similar than paralogs to each other. Thus, in the pre-processing step most algorithms use a cut-off value to prune the correspondence graph and an operational definition of orthology (best hit (BeT) [57], bi-directional best hit (BBH) [46], reciprocal best hit (RBH) [26], symmetrical best hit (SymBeT) [51] or reciprocal smallest distance (RSD) [60]. In section 2 we summarize some of these gene comparison techniques and their corresponding implementations in ortholog detection algorithms.

Other algorithms improve pre and post-processing steps taking duplication events, synteny data and global genome rearrangements into account. For example, SOAR [7] and MSOAR [18], [17] have used global rearrangement heuristics to estimate the evolution distance and, specifically, MSOAR incorporates a post-processing step to eliminate pairs of genes which most likely constitute inparalogs [17]. Its precision is (>90%). An algorithm based on Best Unambiguous Subsets (BUS) [29] starts from BLASTP [1] correspondences, weights graph edges by the amino acid sequence identity and the overall length of BLAST [1] matches and then eliminates all edges that are less than 80% of the maximum-weight edge. Before the single linkage graph clustering step, the BUS algorithm builds synteny blocks to eliminate some ambiguities between duplicated genes that are almost identical.

Despite the outstanding solutions in this field, precision was still a matter of discussion in the 'Quest for Orthologs' meeting at the Welcome Trust Conference Centre in Hinxton, UK in July 2009 [19]. Although new significant algorithms have emerged [53], [40], a recent benchmark [52] shows that those which integrate information from similarity searches, phylogenies, and synteny are more likely to be better choice for evolutionary genomics and functional studies. Hence, the need of gene comparing measures capable of merging different features while guaranteeing high levels of precision. Having this motivation, in this paper we present an all-against-all algorithm for ortholog detection between two closely related species (eukaryotes probably multi-chromosomal) where we combine the homology and length of the sequences with the global rearrangements information by using a novel local-global dissimilarity measure (in section 3).

In our algorithm (in section 4) we try to improve the pre-processing step considering a cut-off value and a synteny block membership criterion similar to the one in [29]. We follow a BUS-like clustering step [29]. We made some experiments with Saccharomyces Cerevisiae and Saccharomyces Bayanus (in section 5).

2. GENE COMPARISON BY USING AN OPERATIONAL DEFINITION OF ORTHOLOGY

Kuzniar in [32] uses the term "nearest neighbour" to collectively designate all ortholog detection algorithms that first calculate pair-wise sequence similarity and then apply an operational definition of orthology even though the approaches do not necessarily imply phylogenetic proximity [31].They are commonly used as first-pass approximations to find putative orthologs skimming the genome-wide matches between two species. For example, the construction of functionally annotated Groups of Ortholog Clusters (COGs) [57], [55], [56], [54] starts from the selection of best BLAST hits (BeTs) to multiple proteomes by using congruent "triangles" of BeTs from at least three different species. These minimal COGs are then merged by a single linkage into larger groups (protein families). Kuzniar refers to the disadvantages of the COG 'triangles' in the presence of gene losses. He points to deficiencies in the COG approach to differentiate between in- and out-paralogs

automatically. The user needs to investigate the pre-computed phylogenetic trees for duplication and speciation events. The automatic clustering procedure creates exclusive clusters, thus, multi-domain proteins must be handled manually.

The author in [46] presents the BBH approach for the detection of conserved clusters of genes based on the definition of a "run". A set of genes occurring on a chromosome is considered as a "run" if and only if all these genes occur on the same strand and the gaps between adjacent genes are 300 base pairs or less. Any pair of genes occurring within a single run is called a "close". Two genes x_a and x_b from two genomes *A* and *B*, x_a and x_b are called a BBH if and only if recognizable similarity exists between them, ie. FASTA scores [47] lower than $1.0 \text{ X } 10^{-5}$, and there is no gene z_b in *B* that is more similar than x_b to x_a and there is no gene z_a in *A* that is more similar than x_a to x_b. Genes (x_a,y_a) from *A* and (x_b,y_b) from *B* form a pair of "close" bidirectional best hits if and only if x_a and y_a are close, x_b and y_b are close, x_a and x_b are a BBH and y_a and y_b are a BBH.

The OFAM database of protein ortholog families is derived from BBH in the extensible data environment for computational genomics CoGenT++ [21].The authors use the BLASTP [1] bit score *bs* as an estimate of sequence similarity and calculate the E-value in a simplified yet uniform manner as follows: $E_{simpl} = L_{eff} X\ S_{eff} X\ 2^{(-bs)}$, where the effective database size S_{eff} is set to 10^8 residues and L_{eff} is the effective protein length (number of amino acid residues not masked by CAST [50]). They use an E-value cut-off of 10^{-5} on E_{simpl} and calculate the cut-off that would accept alignments covering 40% of the query protein length. They actually use the most permissive cut-off between this alternative cut-off and 10^{-5}.

A common procedure for identifying sequence pairs that are putatively orthologous, admissible for the estimation of relative evolutionary rate [60], is the identification of reciprocal best BLAST hits (RBH). Protein *x* in genome *A* is a reciprocal best hit of protein *y* in genome *B* if a forward search of genome *B* with protein *x* yields as the top hit protein *y*, and a reciprocal query of genome *A* with protein *y* yields as the top hit protein *x*. In [60] the author commented a potential pitfall of RBH in the sense that if the forward BLAST yields a paralog best hit, regardless of whether the reciprocal BLAST corrects the error by recovering an actual ortholog, both pairs will be excluded. Thus, while RBH will rightfully prevent admission of the paralog pair to the set of proteins for which relative evolutionary rates are estimated, it might also wrongly exclude an authentically ortholog pair from consideration. Despite of this potential limitation some of the ortholog detection tools reported in literature use RBH.

The ORTHOMCL algorithm [35] uses RBH with the application a normalization method introduced to cope with the fact that high similarity of inparalogs can bias the BLAST scores. First, the pairs with protein similarity scores under 100 are eliminated to get rid of false positives, so that only "most recent" paralogs (inparalogs) are included. Then, G_{AB}, representing the average score among all ortholog and inparalog pairs from genomes *A* and *B* (when *A*=*B*, G_{AB} means the average score among those paralog pairs with reciprocal best hits within a genome), and *G*, representing the average score among all pairs, are calculated. Finally, raw scores of pairs were divided by G_{AB}/G to obtain the final normalized scores.

In [8] an all-against-all FASTA search is conducted for all the proteins in one reference genome to identify the putative orthologs in other genomes. A subclass of putative orthologs is defined as RBH with additional two strict criteria: (1) FASTA expectation value [47] is $<10^{-10}$ and (2) the aligned region between two protein sequences is >80% of the protein length in the reference genome.

Recently, ORTHOINSPECTOR combines de use of RBH and BeT. Given a BLAST search result for a protein of organism *A*, all proteins of *A* with an E-value inferior to the E-value of the best hit in the organism *B* will define a potential group of inparalogs in *A* with respect to the internal node where species *A* and *B* coalesce (that is a group of inparalogs in *A* "with respect to *B*"). The putative list of inparalogs is then validated if the same minimal hypothesis of inparalogy is verified in the BLAST searches for each protein in the list. BLAST best hits are used to define the potential relationships existing between inparalog groups. A 1-to-1 relationship is described by a best hit between a protein of *A* and a protein of *B* complemented by a returning best hit from the protein of *B* to the protein of *A*, that is a reciprocal best hit. A 1-to-many relationship is described by a best hit from a given protein of *A* to any protein member of an inparalog group of *B* complemented by a returning best hit from any member of the inparalog group of *B* to the same protein of *A*. Finally, a many-to-many relationship is described by two best hits between proteins of two groups of inparalogs (a group in *A* and a group in *B*).

In the first INPARANOID program [51], pair-wise similarity scores are calculated with BLAST in four separate steps for organisms *A*, *B* and *C*: *A* versus *B*, *B* versus *A*, *A* versus *A* and *B* versus *B*. Sequence pairs with mutually best hits are detected. If an out-group species (dataset *C*) is used to detect cases of selective loss of orthologs, the similarity scores in bits between *A* versus *C* and dataset *B* versus *C* are calculated. To avoid problems of asymmetric scores between sequence pairs *x*-*y* and *y*-*x* all pair-wise scores are averaged. After the

all-against-all sequence comparison a threshold-pruning process has been applied in the algorithm. User adjustable cut-off values are applied to each pair-wise match: a score cut-off (50 bits); and an overlap cut-off. The *A-B* sequence pairs are eliminated if either sequence in *A* or sequence in *B* scores higher to out-group sequence than they score to each other. Additional orthologs (inparalogs) are clustered together with each remaining pair of potential orthologs. Overlapping clusters are resolved by a set of rules. Finally, they estimate the probability that a given pair of orthologs had mutual best score only by chance.

In release 7 of the algorithm [45], INPARANOID uses the SEG low-complexity filter to mask only during seeding but not during extension (soft masking). This more stringent low-complexity filtering permitted the procedure to lower the score threshold from 50 to 40 bits. However, matches accepted in the first pass are realigned using BLAST with SEG and compositional adjustment switched off, before the overlap criteria are applied to avoid the effect of shorter alignments. For both the query and the match sequence, the distance between the first and the last aligned residue must equal or exceed 50% of the length of the sequence. Furthermore, for both the query and the match sequence, the sum of the lengths of the aligned regions on that sequence must equal or exceed 25% of the length of the sequence. When there are multiple high-scoring segment pairs, INPARANOID requires that they maintain the same relative order on both sequences, and that they do not overlap by >5%.

In an effort to correct the mentioned source of error in RBH, Wall et al. developed the RSD algorithm [60], [10] that preserves the safeguard of reciprocal genome queries, but is less susceptible to exclude an ortholog if a paralog is returned as the top hit in either the forward or reverse steps of a reciprocal BLAST query. This approach has been shown to provide more comprehensive lists of orthologs than other methods that are based on BLAST alone. It is likely to be more accurate for identifying orthologs because it uses a phylogenetically-grounded measurement of similarity that matches certain assumptions about how orthologs in different species have evolved.

The RSD algorithm employs BLAST as a first step, starting with a subject genome *B*, and a protein query sequence *x*, belonging to genome *A*. A set of hits *H*, exceeding a predefined significance threshold (e.g. $E < 10^{-20}$) is obtained. Then, using the multiple alignment program CLUSTALW [58], each protein sequence in *H* is aligned separately with the original query sequence *x*. If the alignable region of the two sequences exceeds a threshold fraction of the alignment's total length (cut-off of 0.8), the program PAM [63] is used to obtain a maximum likelihood estimate of the number of amino acid substitutions separating the two protein sequences, given an empirical aminoacid substitution rate matrix [28]. The model under which a maximum likelihood estimate is obtained may include a variation in the evolutionary rate among protein sites, and for more distant comparisons they have generally assumed a gamma distribution with shape parameter $\alpha = 1.53$ [43]. Of all sequences in *H* for which an evolutionary distance is estimated, only *y*, the sequence yielding the shortest distance, is retained. This sequence *y* is then used for a reciprocal BLAST against genome *A*, retrieving a set of high scoring hits *L*. If any hit from *L* is the original query sequence *x*, the distance between *x* and *y* is retrieved from the set of smallest distances calculated previously. The remaining hits from *L* are then separately aligned with *y* and maximum likelihood distance estimates are calculated for these pairs. If the protein sequence from *L* producing the shortest distance to *y* is the original query sequence *x*, it is assumed that a true ortholog pair [60] has been found and their evolutionary distance is retained.

In general, the all-against-all detection of orthologs between two genomes inputs two FASTA format [47] protein sequence files *A* and *B*. The procedure starts with the calculation of all pair-wise similarity scores between all studied sequences. Then, it applies an operational definition of orthology with a pruning strategy and it follows with the application of some clustering algorithm such as the Markov Clustering algorithm (MCL) [59], the minimum common partition and the maximum cycle decomposition [7].

Instead of using an operational definition of orthology we are defining a similarity-based method with a local-global dissimilarity measure described in the next section. We compare genes measuring the dissimilarity of the DNA sequences, their length and their membership to truly homolog regions defined in [9] considering global genome rearrangements.

3. A NOVEL LOCAL-GLOBAL GENE DISSIMILARITY MEASURE

We model the dissimilarity between two genes *x* and *y* by dividing it into local functions for each feature and then combining them into a global function using the local-global principle [4]. In order to measure sequence similarity we select a global optimum alignment Needleman-Wunsch algorithm [42] implemented in the nwalign Matlab function [38] with default "Scalevalue" of 1, 'NUC44' [14] scoring matrix for nucleotides and default "*GapOpenValue"* and *"ExtendGapValue"* of 8. From the optimal global alignment score in bits [14] we define an association coefficient *ca* of gene sequences *x.s* and *y.s* for a local dissimilarity function d_1.

$$ca(x.s, y.s) = na(x.s, y.s) / na(x.s, x.s) \quad (1)$$

$$d_1(x.s, y.s) = \begin{cases} 1 - ca(x.s, y.s) \\ \quad if \;\; ca(x.s, y.s) > 0 \\ 1 \\ \quad if \;\; ca(x.s, y.s) \le 0 \end{cases} \quad (2)$$

Since recombination may produce genome rearrangements in evolution, ortholog regions may be reordered or inverted in relation with other genomes. Therefore, we use MAUVE multi-alignment software [9] to identify conserved segments that do not seem to be altered by genome rearrangements (Locally Collinear Blocks) (LCB) [9]. We consider that genes belonging to the same LCB will probably be orthologs. We define d_2 as a local distance function for each gene x_i, i = 1..n_1 + n_2 and j = 1.. *Number_of_LCBs* based on Jaccard Similarity Coefficient [27].

$$matLCB(i, j) = \begin{cases} 0 \quad if \;\; x_i \notin LCB_j \\ 1 \quad if \;\; x_i \in LCB_j \end{cases} \quad (3)$$

$$d_2(x, y) = 1 - CJaccard(x, y, matLCB) \quad (4)$$

The value of the Jaccard coefficient $CJaccard(x, y, matLCB)$ for a pair of gene vectors,*x* and *y*, ranges form 0 to 1 and is equal to the number of bits "on" in the corresponding pair of binary vectors in *matLCB*, divided by the number of bits "on" in either vector. The distance measure d_2, is called the Soergel distance andsatisfies the triangle inequality [37]. It represents the percentage of nonzero coordinates that differ in gene vectors *x* and *y* in *matLCB*.

Some other choices could be appropriate to compare binary vectors in the rows of *matLCB*: the Hamming metric [11], the Maryland Bridge coefficient [39] MB_{xy}, representing the average proportion of the overlap in these vectors, the similarity coefficient WA_{xy} [30] suggested by Korbel et al., the Simpson similarity index [49], [64] and the Dice similarity index [16] representing the arithmetic average cardinality of two sets.Dice and Jaccard's coefficients are monotonic in each other [62]. The standard correlation coefficient [11] and the Mutual Information [25] could also be used to measure the similarity between binary vectors with similar results [20].

Authors in [20] define the parametric family $A_{\lambda,xy} = \frac{X_{xy}}{B_\lambda}, -\infty < \lambda < \infty, (\lambda \neq 0)$ where $B_\lambda = \left(\frac{X_{xx}^{\lambda} + X_{yy}^{\lambda}}{2} \right)^{\frac{1}{\lambda}}$ and X_{xy} is the dot product of two vectors. They show that distance measures with $\lambda \geq 0$ attach more importance to shared presence of 1s, weighting shared vector coordinates by a factor of 2^{λ}. This may make two genes belonging to more LCBs more similar to each other, than any genes with different and low degrees of membership which is the generality in our problem. However, functions with negative λ values (as MB_{xy} and WA_{xy}) may tend to balance the shared 1's and the degree of LCB membership.

In [39], authors define the MB_{xy} coefficient to cope with the flaw in the Jaccard coefficient in that it systematically underestimates the similarity between genomes in some studies on gene content (presence-absence) trees when the sizes of comparing sets are about the same and their overlap is about half of the elements in each of them. They notice that the underestimate becomes even more striking when one set is much smaller than the other, as it frequently happens with genomes and the small value of the coefficient contradicts predictions on the evolutionary relationships between genomes. In our case it should be important to study the overlap ratio in order to detect possible errors in the similarity estimation of d_2. Such further study might change our measure choice for the LCB membership gene comparison. In fact, the Maryland Bridge coefficient not only has all the advantages and is co-monotone with the Jaccard coefficient but properly evaluates similarity in the cases where the latter fails [39].

We define the length of the sequence distance function d_3 as a renormalized difference [13] for an interval-scaled attribute where max_length and min_length are the minimum and maximum lengths of gene sequences:

$$d_3(x.s, y.s) = \frac{|length(x.s) - length(y.s)|}{\max\_length - \min\_length} \quad (5)$$

Each local functions has values ranging in [0..1] so we can classify them as interval-scaled attributes and select Minkowski Distance, Euclidean Distance, Manhattan Distance or Maximum Distance to compute the global gene dissimilarity [3]. Following the approach in [3] we can calculate the gene dissimilarity using an attribute-weighted function of mixed attributes or the weighted Euclidean Distance [13] with the corresponding study of the possible weight values with biological meaning. For now, the global dissimilarity measure d_g between gene x and gene y is defined as:

$$d_g(x,y) = \sqrt{\frac{d_1(x.s, y.s)^2 + d_2(x,y)^2 + d_3(x.s, y.s)^2}{3}} \quad (6)$$

4. ALGORITHM OVERVIEW

We construct $G(V,E)$ as a weighted bipartite undirected complete graph [12] describing the measure d_g between the two sets of genes A and B in the two species compared. The set of vertices V have order $n=n_1+n_2$ where n_1 are the total of genes in A and n_2 are the total of genes in B. Every edge $e=(x,y)$ in E connecting nodes $x \in A$ and $y \in B$ is weighted by $d_g(x,y)$ global measure (6).

$\boldsymbol{O(n_1 * n_2) * O(\max sequence\ length\ in\ X * \max sequence\ length\ in\ Y)}$ defines the algorithm time complexity.

Let $\boldsymbol{T}$ be the set of gene pairs $\boldsymbol{(x,y)}$ without ambiguities where $\boldsymbol{x \in X}$ and $\boldsymbol{y \in Y}$

Let ***BSyn*** be the set of synteny blocks.

$$\boldsymbol{marked(x,y) = \begin{cases} 0 & if\ (x,y) \notin \beta, \forall \beta \in BSyn \\ 1 & if\ \exists \beta \in BSyn, (x,y) \in \beta \end{cases}}$$

1) $\boldsymbol{\forall (x,y) \in T, marked(x,y) = 0}$
   a) Build a new synteny block $\boldsymbol{B \subseteq T}$ with $\boldsymbol{(x,y)}$, $\boldsymbol{marked(x,y) = 1}$
   b) $\boldsymbol{b1 = e1 = x, b2 = e2 = y}$
   c) $\boldsymbol{\forall (v,w) \in T, marked(v,w) == 0}$
      i) If $are_{physically_{near_{two_{genes(e1,v)}}}} \wedge keep_{gene_{order(b1,e1,v)}} \wedge are_{physically_{near_{two_{genes(e2,w)}}}} \wedge keep_{gene_{order(b2,e2,w)}}$ then
         (1) If $\boldsymbol{(v,w)}$ is physically near to the block $\boldsymbol{B}$ and keeps gene order, then $B = \mathrm{B} \cup \{(v,w)\}$
         (2) $\boldsymbol{marked(v,w) = 1}$
   d) If $\boldsymbol{B}$ contains more than two pairs, then $\boldsymbol{BSyn = BSyn \cup B}$
2) Return ***BSyn***.

Figure 1: Algorithm to build the synteny blocks.

In the pre-processing step we eliminate out-edges over 20% of the minimum-weight edge in the directed version of G. This initial pruning process creates numerous two-node sub-graphs representing one-to-one unambiguous matches between genes that are not closely-related paralogs. With these sub-graphs we build synteny blocks based on the physical distances between consecutive matched genes. Figure 1 shows the algorithm to build synteny blocks. We use the maximum distance for gene proximity defined in [29] as: two genes are near if they are 20kb apart, i.e. approximately 10 genes apart from each other. As in [29] in an additional pruning process we rather keep edges connecting additional genes within the synteny blocks of at least three genes since theses edges may represent ortholog relationships.

In the clustering step we separate BUS connected components in pruned graph G as in [29] such that the best match of any node within the subset of nodes is contained within the subset, and no node outside the subset has its best match within the subset. These two properties assure that the subsets will be both best and unambiguous and that the separation of subsets does not leave any orphan node or does not remove the strictly best match of any node [29]. Finally, each BUS represents a homology group, and especially each two-node BUS represents a pair of orthologs. The algorithm general schema is represented in Figure 2. The time complexity of the all-against-all calculation of the global gene dissimilarity measure (6)

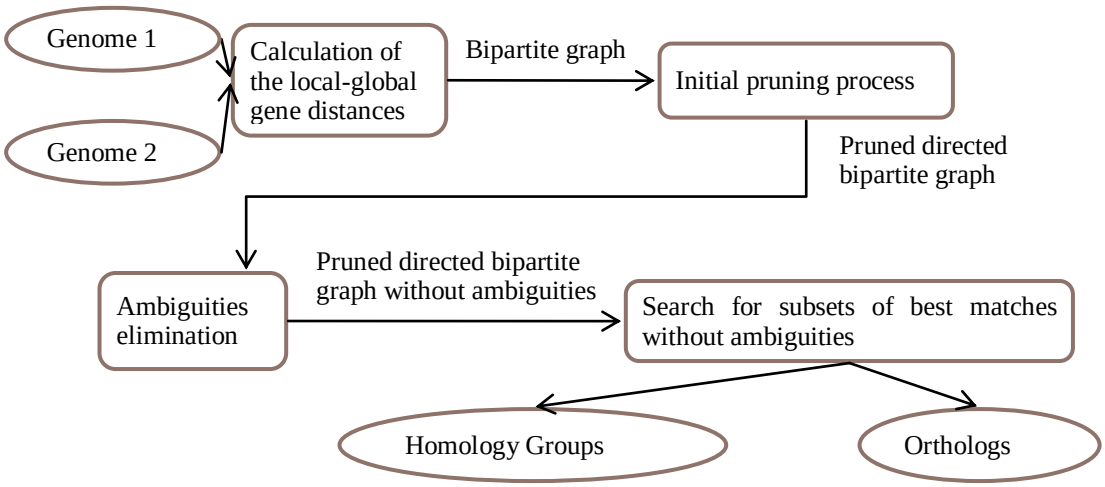

Figure 2: Overall schema of the algorithm.

5. EXPERIMENTS AND RESULTS

Saccharomyces Cerevisiae and Saccharomyces Bayanus are two closely related Saccharomyces species in the group Saccharomyces sensu stricto [29]. Due to processing capability limitations, we selected S. Cerevisiae chromosome 5 of the S288C annotation available in the NCBI Genomes Database [41] with 226 annotated genes. The complete S. Bayanus sequence was found in Biomax Database [5] with 4792 annotated genes.

First, we calculated 69 LCBs using MAUVE with its default parameters (see Figure 3). Then, we ran the all-against-all alignment. We found 28 unambiguous matches and 4 synteny blocks. We eliminated 10869 ambiguities. We found a total of 202 homology groups and 131 ortholog pairs. Figure 4 shows the plotting of the comparison between the two genomes.

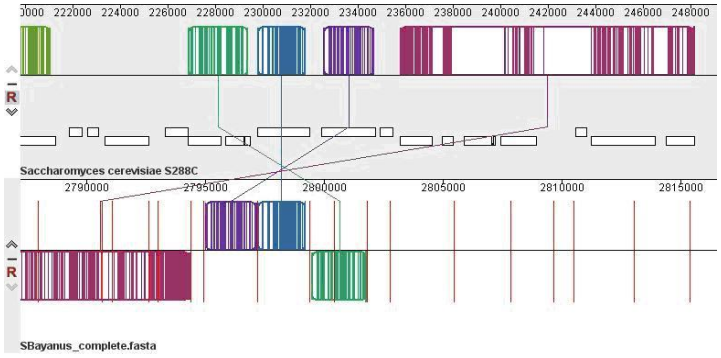

Figure 3: A region of the alignment in MAUVE with 4 LCBs.

In order to validate the results and because we could not find the counter part results in the available ortholog databases, we use the annotation of the two genomes. The annotated S. Bayanus genome has 194 S. Cerevisiae homolog genes. 191 genes were correctly classified so we achieved 98.45% of true classifications.

6. CONCLUSIONS/FUTURE WORK

Based on our novel local-global dissimilarity measure, our ortholog detection algorithm yields a promising performance. Further papers should present a validation performance comparison with whole genome datasets. The bipartite graph construction using our novel function should improve the one in [29] since it can incorporate new features to the gene comparison thus enhancing the complete prediction process. Our work pursues the inclusion of new gene features, some improvements in gene comparison and in different algorithm steps. We are also developing a parallel version of the algorithm.

.

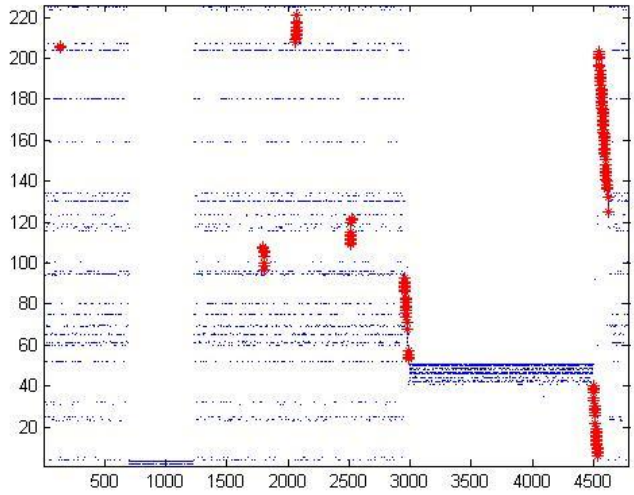

Figure 4: Undirected bipartite graph with the best matches and the ortholog pairs founded. Red asterisks represent the ortholog relationships.

REFERENCES

[1] ALTSCHUL, S. (1997): Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, **Nucleic Acids Research**, 25, 3389-3402.

[2] ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W. and LIPMAN, D.J. (1990): Basic local alignment search tool. **J. Mol. Biol.**, 215, 403-410.

[3] ANDRITSOS, P. (2002): **Data Clustering Techniques**. Tech Report CSRG-443, University of Toronto.

[4] BERGMANN, R. (2002): **Experience management: foundations, development methodology, and Internet-based applications**. Springer, Hildesheim.

[5] Biomax Informatics genome database (2009): Available in www.pedant.gsf.de. **Consulted**: 15-1, 2009.

[6] CHEN, F., MACKEY, A. J., STOECKERT, C. J. and ROOS, D. S. (2006): ORTHOMCL-DB: querying a comprehensive multi-species collection of ortholog groups. **Nucleic Acids Research**, 34(Database issue), D363-D368.

[7] CHEN, X., ZHENG, J., FU, Z., NAN, P. ZHONG, Y., LONARDI, S. and JIANG, T. (2005): Assignment of Orthologous Genes via Genome Rearrangement. **IEEE/ACM Trans. Comput. Biology Bioinform.**, 2, 302-315.

[8] CHEN, Y. and XU, D. (2005): Understanding protein dispensability through machine-learning analysis of high-throughput data. **Bioinformatics**, 21, 575-581.

[9] DARLING, A.C.E., MAU, B. and BLATTNER, F.R. (2004): MAUVE: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. **Genome Res.**, 14, 1394-1403.

[10] DELUCA, T.F., WU, I., PU, J., MONAGHAN, T., PESHKIN, L., SINGH, S. and WALL, D. P. (2006): Roundup: a multi-genome repository of orthologs and evolutionary distances. **Bioinformatics**, 22, 2044-2046.

[11] DEZA, E. (2006): Dictionary of Distances. Available in http://www.scribd.com/doc/53001767/Dictionary-of-Distances-M-Deza-E-Deza-Elsevier-2006-WW. **Consulted**: 15-1, 2009.

[12] DIESTEL, R. (2005): **Graph Theory**. Springer-Verlag, Heidelberg.

[13] DUCH, W. (2000): Similarity-based methods: a general framework for classification, approximation and association. **Control and Cybernetics**, 29, 1-30.

[14] DURBIN, R., EDDY, S., KROGH, A., and MITCHISON, G. (1998): **Biological Sequence Analysis**. Cambridge University Press, Cambridge.

[15] FADI TOWFIC, M.H.W.G., HONAVAR V. (2009): Detection of Gene Orthology Based On Protein-Protein Interaction Networks. **IEEE International Conference on Bioinformatics and Biomedicine, BIBM, Washington DC**.

[16] FRAKES, W. B. and BAEZA-YATES, R., (1992): **Information Retrieval, Data Structure and Algorithms**. Prentice Hall, New Jersey.

[17] FU, Z., CHEN, X., VACIC, V., NAN, P., ZHONG, Y. and JIANG, T. (2007): MSOAR: A High-Throughput Ortholog Assignment System Based on Genome Rearrangement. **Journal of Computational Biology**, 14, 1160-1175.

[18] FU, Z., CHEN, X., VACIC, V., NAN, P., ZHONG, Y., and JIANG, T. (2006): A Parsimony Approach to Genome-Wide Ortholog Assignment Assignment of Orthologous Genes via Genome Rearrangement. **IEEE/ACM Trans. Comput. Biology Bioinform., Springer**, 578-594.

[19] GABALDÓN, T.E.A. (2009): Joining forces in the quest for orthologs. **Genome Biology**. 10, 1-3.

[20] GLAZKO, G., GORDON and A., MUSHEGIAN, A. (2005): The choice of optimal distance measure in genome-wide datasets. **Bioinformatics**, 21(Suppl 3), 3-11.

[21] GOLDOVSKY, L., JANSSEN, P., AHRÉN, D., AUDIT, B., CASES, I., DARZENTAS, N., ENRIGHT, A. J., LÓPEZ-BIGAS, N., PEREGRIN-ALVAREZ, J. M., SMITH, M., TSOKA, S., KUNIN, V. and OUZOUNIS, C. A. (2005): CoGenT++: an extensive and extensible data environment for computational genomics, **Bioinformatics**, 21, 3806-3810.

[22] GOODSTADT, L. and PONTING, C.P. (2006): Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. **PLoS Comput Biol**, 2(9), 1134-1150.

[23] GOTOH, O. (1981): An Improved Algorithm for Matching Biological Sequences. **J. Mol. Biol.**, 162, 705-708.

[24] HATICE GULCIN OZER, J.C., FA ZHANG, B. Y. (2004): Clustering Of Eukaryotic Orthologs Based On Sequence And Domain Similarities Using The Markov Graph-Flow Algorithm. Available in http://www.biosci.ohio-state.edu/~ozer/pub/papers/icba04_hg_ozer.pdf. **Consulted**: 15-1, 2008.

[25] HINDLE, D. (1990): Noun classification from predicate-argument structures. **Proceedings of ACL-90, Pittsburg, Pennsylvania**, 268-275.

[26] HIRSH, A.E. and FRASER, H.B. (2001): Protein dispensability and rate of evolution. **Nature**, 411, 1046-1049.

[27] JACCARD, P. (1901): Étude comparative de la distribution florale dans une portion des Alpes et des Jura. **Bulletin del la Société Vaudoise des Sciences Naturelles**, 37, 547-579.

[28] JONES, D.T., TAYLOR, W.R. and THORNTON, J.M. (1992): The rapid generation of mutation data matrices from protein sequences. **Computer Application in Biosciences**, 8, 275-282.

[29] KAMVYSSELIS, M.K., **Computational comparative genomics: genes, regulation, evolution**. (2003): Department of Electrical Engineering and Computer Science. Massachusetts Institute of Technology, Massachusetts.

[30] KORBEL, J.O., SNEL, B., HUYNEN, M. A. and BORK, P. (2002): SHOT: a web server for the construction of genome phylogenies. **Trends Genet**, 18, 159-162.

[31] KOSKI, L.B. and GOLDING, G.B. (2001): The closest BLAST hit is often not the nearest neighbor. **J. Mol. Evol.**, 52, 540-542.

[32] KUZNIAR, A., VAN HAM, R.C., PONGOR, S. and LEUNISSEN, J. A. (2008): The quest for orthologs: finding the corresponding gene across genomes. **Trends Genet**, 24, 539-551.

[33] LANDWEBER, Y.Z.A.L.F. (2007): BLASTO: a tool for searching orthologous groups. **Nucleic Acids Research**, 35(Web Server issue), W678-W682.

[34] LEE, Y., SULTANA, R. PERTEA, G., CHO, J., KARAMYCHEVA, S., TSAI, J., PARVIZI, B., CHEUNG, F., ANTONESCU, V., WHITE, J., HOLT, I., LIANG, F. and QUACKENBUSH, J. (2002): Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). **Genome Research**, 12, 493-502.

[35] LI LI, C.J.S., JR. and DAVID, S. R. (2003): ORTHOMCL: Identification of Ortholog Groups for Eukaryotic Genomes. **Genome Research**, 13, 2178-2189.

[36] LINARD, B., THOMPSON, J. D., POCH, O. and LECOMPTE, O. (2001): OrthoInspector: comprehensive orthology analysis and visual exploration. **BMC Bioinformatics**, 12, 1471-2105.

[37] LIPKUS, A.H. (1999): A proof of the triangle inequality for the Tanimoto distance. **Journal of Mathematical Chemistry**, 26, 263-265.

[38] MATLAB R2007a Help. (2007): Available in www.mathworks.com. **Consulted**: 15-1, 2009.

[39] MIRKIN, B. and KOONIN E. (2003) A top-down method for building genome classification trees with linear binary hierarchies. Available in http://citeseerx.ist.psu.edu. **Consulted**: 15-1, 2009.

[40] MULLER, J. (2010): eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. **Nucleic Acids Res.**, 38(Database issue), D190-D195.

[41] National Center for Biotechnology Information. (2009): Available in http://www.ncbi.nlm.nih.gov/. **Consulted**: 15-1, 2009.

[42] NEEDLEMAN, S. B. and WUNSCH, C.D. (1970): A general method applicable to the search for similarities in the amino acid sequence of two proteins. **J. Mol. Biol.**, 48, 443-453.

[43] NEI, M., XU, P. and GLAZKO, G. (2001): Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. **Proc. Natl Acad. Sci. USA**, 98, 2497-2502.

[44] O'BRIEN K. P., REMM, M. and SONNHAMMER, ERIK L. L. (2005): INPARANOID: a comprehensive database of eukaryotic orthologs. **Nucleic Acids Research**. 33(Database issue), D476-D480.

[45] ÖSTLUND, G., SCHMITT, T., FORSLUND, K., FORSLUND, K., KÖSTLER, T., MESSINA, D. N., ROOPRA, S., FRINGS, O. and SONNHAMMER, E. L. L. (2010): INPARANOID 7: new algorithms and tools for eukaryotic orthology analysis. **Nucleic Acids Research**, 38(Database issue), D196-D203.

[46] OVERBEEK, R., FONSTEINZY, M., D'SOUZA, M., MALTSEV, N. and PUSCH, G. D. (1998): The use of gene clusters to infer functional coupling. **Proc. Natl. Acad. Sci. U. S. A.**, 96, 2896-2901.

[47] PEARSON W.R. (1990): Rapid and Sentive Sequence Comparison with PASTP and FASTA. Methods **Enzymol**, 183, 63-98.

[48] PENKETT, C. J., WOOD, V. and BA¨HLER, J. (2006): YOGY: a web-based, integrated database to retrieve protein orthologs and associated Gene Ontology terms. **Nucleic Acids Research**. 34(Web Server issue), W330-W334.

[49] PIELOU, E.C. (1975): **Ecological diversity**. Wiley, New York.

[50] PROMPONAS, V.J. (2000): CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. **Bioinformatics**, 16, 915-922.

[51] REMM, M., STORM, C. E. V. and SONNHAMMER, E. L. L. (2001): Automatic clustering of orthologs and inparalogs from pair-wise species comparisons. **J. Mol. Biol.**, 314, 1041-1052.

[52] SALICHOS, L. and ROKAS, A. (2011): Evaluating Ortholog Prediction Algorithms in a Yeast Model Clade, **PLoS ONE**, 6, 1-11.

[53] SHI, G., PENG, M. and JIANG, T. (2011): MultiMSOAR 2.0: An Accurate Tool to Identify Ortholog Groups among Multiple Genomes. **PLoS ONE**, 6, 1-9.

[54] TATUSOV, R. L., FEDOROVA, N. D., JACKSON, J. D., JACOBS, A. R., KIRYUTIN, B., KOONIN, E. V., KRYLOV, D. M., MAZUMDER, R., MEKHEDOV, S. L., NIKOLSKAYA, A. N., RAO, B. S., SMIRNOV, S., SVERDLOV, A. V., VASUDEVAN, S., WOLF, Y. I., YIN, J. J. and NATALE, D. A. (2003): The COG database: an updated version includes eukaryotes. **BMC Bioinformatics**, 4, 1-14.

[55] TATUSOV, R. L., FEDOROVA, N. D., JACKSON, J. D., JACOBS, A. R., KIRYUTIN, B., KOONIN, E.V., KRYLOV, D.M., MAZUMDER, R., MEKHEDOV, S. L., NIKOLSKAYA, A. N., RAO, B. S., SMIRNOV, S., SVERDLOV, A.V., VASUDEVAN, S., WOLF, Y. I., YIN, J. J and NATALE, D. A. (2000): The COG database: a tool for genome-scale analysis of protein functions and evolution. **Nucleic Acids Res**, 28, 33-36.

[56] TATUSOV, R. L., NATALE, D. A., GARKAVTSEV, I. V., TATUSOVA, T. A., SHANKAVARAM, U. T., RAO, B. S., KIRYUTIN, B.,GALPERIN, M. Y., FEDOROVA, N. D., and KOONIN, E. V. (2001): The COG database: new developments in phylogenetic classification of proteins from complete genomes. **Nucleic Acids Res**, 29, 22-28.

[57] TATUSOV, R.L., KOONIN, E. V. and LIPMAN, D. J. (1997): A genomic perspective on protein families. **Science**, 278(5338), 631-637.

[58] THOMPSON, J.D., PLEWNIAK, F., THIERRY, J.C. and POCH, O. (2000): DbClustal: Rapid and reliable global multiple alignments of protein sequences detected by database searches. **Nucleic Acids Res.**, 28, 2919-2926.

[59] VAN DONGEN, S., (2000): Graph clustering by flow simulation. Available in: http://Figitur-archive.library.uu.nl/dissertations/. **Consulted**: 15-1, 2009.

[60] WALL, D.P., FRASER, H. B. and HIRSH, A. E. (2003): Detecting putative orthologs. **Bioinformatics**, 19, 1710-1711.

[61] WATERHOUSE, R. M. (2011): OrthoDB: the hierarchical catalog of eukaryotic orthologs. **Nucleic Acids Res.**, 39(Database issue), D283-D288.

[62] WEEDS, J., WEIR, D. and MCCARTHY, D. (2004): Characterising Measures of Lexical Distributional Similarity. **COLING '04 Proceedings of the 20th international conference on Computational, Stroudsburg, PA, USA.**

[63] YANG, Z. (2000): **Phylogenetic Analysis by Maximum Likelihood (PAML)**. University College London, London.

[64] SNEATH, P.H.A. and SOKAL, R.R. (1973): **Numerical taxonomy. The principles and practice of numerical classification**. W H Freeman & Co (Sd), San Francisco.