Candidate Ortholog Clusters in Human, Mouse and Chicken genomes

Akshay Vashist, Ilya Muchnik, Casimir Kulikowski
Department of Computer Science
Rutgers, The State University of New Jersey
Piscataway, NJ-08854 U.S.A.
Phone: +732-445-2006, Fax: +1-732-445-0537, E-mail: vashisht@cs.rutgers.edu,
muchnik@dimacs.rutgers.edu, kulikows@cs.rutgers.edu

Abstract - The recently completed chicken genome, and previously available human and mouse genomes provide us an opportunity to understand the evolutionary relationship between mammals and aves at the molecular level by finding groups of orthologs in these genomes. Using a recently developed tool for automatic large scale screening of candidate orthologs in a multi-genome data, we extracted candidate ortholog clusters in these complete genomes.

We obtained 14,254 candidate ortholog clusters that cover 81% of all genes in the three complete genomes. There are 9,733 candidate ortholog clusters that contain genes from all the three organisms. They cover about 70% of all genes in these organisms. Based on the Pfam annotations of genes we found that 95% of clusters are consistently annotated, since genes within each of these clusters are annotated by a single Pfam annotation. A comparison with manually curated 565 known ortholog triplets in these genomes shows that candidate ortholog clusters related to these ortholog triplets are the extensions of the ortholog triplets. Among the 565 known ortholog triplets, 549 were preserved in our results, demonstrating that our procedure was able to capture the essence of stringent criteria used by experts. Additionally, we were able to estimate the stability of the ortholog triplets and found that 562 of these are stable.

I. Introduction

One of the fundamental problems in comparative genomics is to find orthologous genes, genes in different organisms that are related through vertical inheritance [FIT 70]. The set of orthologous sequences serves to functionally annotate genes, select targets for various experimental studies, provide anchors to determine related genetic elements such as promoters, and are essential for molecular phylogenetics.

The classical method to find orthologous genes involves a detailed multicriteria approach relying primarily on comparison of genes from two genomes [MAK 98]. Such detailed analysis by experts includes known information about gene function, domain structure for corresponding proteins, evolutionary distance between genomes [OUY 03]. With the availability of complete genome data for several genomes the problem of finding ortholog clusters on large data must be addressed computationally. Currently, there are a few servers [TAT 97] that contain database ortholog clusters from many genomes. The methods employed by these databases can be characterized as semi-automatic because an initial grouping of genes is performed by automatic clustering procedures, followed by a manual curation of clustering results. Due to the requirement of manual curation, such semi-automatic approaches do not scale up to address the current needs. In our work to find candidate ortholog clusters on chicken, human and mouse data, we use an extension of the completely automatic tool for large scale screening of ortholog clusters in multi-genome data. The basic method along with its validation tests is described in [VAS 04].

We obtained 14,254 candidate ortholog clusters which cover approximately 80% of all genes in the three genomes. There are 9,733 clusters that contain genes from all the three genomes and genes in these clusters make up 70% of all genes in the genomes and 85% of all genes in the candidate ortholog clusters. About a 52% of all candidate orthologs are consistently annotated by Pfam [BAT 04] i.e, all genes in each of these clusters have the same Pfam annotation. Only 5% of the candidate ortholog clusters have genes which are annotated by different Pfam families. However, the functional descriptions for these Pfam families are similar. Genes from 34% of the clusters do not have any hit in the Pfam database. These are likely to be genes which are unrelated to annotated data in Pfam.

The candidate ortholog clusters are also compared with the known 565 ortholog triplets found in the same organisms by a multi-criteria approach through extensive manual curation [OUY 03]. The comparison showed that 549 out of the 565 known ortholog clusters have a strong correspondence with our candidate ortholog clusters. The tool also allowed to estimate the 'stability' of the known 565 ortholog clusters. The 549 ortholog triplets that are related to our clusters demonstrate high stability. From this perspective the corresponding clusters can be considered us extensions of known ortholog triplets.

The outline of the paper is as follows. In section 2, we present the description of data used in this study and section 3 presents the results of the clustering and their validation using Pfam annotations. The comparison of candidate ortholog cluster with the known manually curated ortholog triplets is presented in section 4. The method to extract candidate ortholog clusters is described in section 5. Section 6 is the discussion.

II. Data

There are three different data used in this study. The first of these data are the complete set of proteins in chicken (*Gallus gallus*), human (*Homo sapiens*) and mouse (*Mus musculus*) for extracting candidate ortholog clusters. Secondly, the Protein families database, Pfam [BAT 04], was used for annotating the protein sequences these complete genomes. Finally, the expert curated 565 triplet ortholog clusters [OUY 03] were used to validate the candidate ortholog clusters.

The human genome [LAN 01] was completed in 2001, mouse genome [CHI 03] was completed in 2003, and recently the chicken genome [HIL 04] was completed in 2004. The complete proteomes for these organisms were downloaded from Ensembl (updated Dec 08, 2004) [AL. 04] in which the size of the human, mouse and chicken proteome is 33860, 32442, and 28416 respectively. We also use the species tree for chicken, human and mouse as the input data.

We used the sequences in the protein families database, Pfam [BAT 04] (rel. 12) to associate annotations with the sequences from chicken, human and mouse. This version of Pfam contains 865,065 sequences classified into 7,677 protein functional families. The association was performed using the sequence search tool blastp [ALT 97] with 10^{-4} as the e-value cutoff. A sequence was annotated with the set of Pfam families for its blast hits, but if the e-value for the best hit was worse than 10^{-4}, no annotation was associated with it. The Pfam annotation was used to measure the consistency of sequences within a candidate ortholog cluster from a functional family perspective.

A strong means to validate ortholog clustering results is a set of manually curated ortholog clusters. The third source of data is the set of 565 ortholog triplets in the chicken, human and mouse genomes found by experts [OUY 03] [1]. This database of triplets was constructed when complete genomes from mouse and chicken were not available. The ortholog triplets were produced through extensive filtering of human, mouse and chicken sequences in the NCBI "nr" database in 2002. Only those sequences from the three genomes which had the same description were considered for finding candidates of orthologs. The orthologs were required to maintain a stringent level of sequence identity, both at the nucleotide and the amino-acid level, over the entire length of the sequence. In general, 90% sequence identity was required. This was followed by removal of close paralogs and extensive manual curation of results using various criteria such as complete analysis of coding sequences and usage of codon bias. An exception to this procedure was for the triplet annotated as *interleukin-2 (IL-2)*; Blast search of chicken *IL-2* does not find any mammalian *IL-2* sequence and characterization of this triplet is based on expert knowledge of curators [OUY 03].

TABLE I

Size/Orgs.	C+H+M	H+M	C+H	C+M	Total
2		1,792	484	479	2,755
3	2,693	809	189	146	3,837
4	2,020	219	60	36	2,335
5	1,258	109	22	6	1,395
6-10	2,654	102	23	12	2,791
11-15	631	16			648
16-25	333	7			340
26-50	104	4			108
51-100	26	4			30
>100	14	1			15
Total	9,733	3,063	778	680	14,254

Joint distribution of size and organisms in clusters. The first column represent the size of the clusters. In the first row contains the names of the organism in a clusters; the organisms are represented by their first alphabet character, so C+H+M means a cluster contains genes from chicken, human and mouse.

To compare the candidate ortholog clusters with the manually curated 565 triplet ortholog clusters, we identified 1,695 amino-acid sequences from the triplets in our ortholog clusters using the Genbank [BEN 04] accession numbers provided in [OUY 03]. Additionally, we also validated the automatically produced ortholog clusters using their Pfam annotations.

III. Clustering Results

The input to clustering procedure was 94,718 sequences from the complete chicken, human and mouse genomes, of which 18,053 were classified as genes novel to these organisms as they could not be associated with any orthologs in the given genomes. The method produced 14,254 clusters that contain sequences from at least two organisms. We call these clusters the candidate ortholog clusters. The candidate ortholog clusters include 2,755 clusters of size 2, 3,837 clusters of size 3, and 2,335 of size 4. The size of most clusters is less than 10, although the largest cluster has 1,079 genes (closely followed by clusters with size 944 and 526), and there are 15 clusters that contain more than 100 genes. There are 9733 candidate ortholog clusters which have sequences from all the three genomes, whereas 3,063 have sequences only from human-mouse genomes, 778 have sequences from human-chicken genomes and 680 contains only mouse-chicken genomes.

Table 1 summarizes the distribution of different groups of organisms in various size clusters. The first column indicates the size (or, the range of size) for the corresponding rows. The first row indicates the group of organisms for which the distribution is shown in the corresponding column. For instance, the second column corresponds

[1] This set of orthologs was compiled by experts to study the evolutionary distance of various proteins, the synonymous and nonsynonymous rates of substitution, and evolutionary distances of different proteins in the chicken, human and mouse genome

to clusters that contain genes from all the three organisms: chicken (C), human (H) and mouse (M). There are 2,693 clusters which contain exactly one gene from each of the three organisms. We find that the number of clusters containing genes from all three organisms is significantly higher than the clusters that contain genes from only two organisms. This domination is also observed for various groups of organisms: the number of clusters with human-mouse genes is greater than clusters chicken-human genes which is greater than clusters with chicken-mouse genes. Also, the number of clusters with genes from human and mouse is considerably larger than the chicken-human or chicken-mouse clusters. One would expect this based on the close evolutionary distance between human and mouse. Likewise, another interesting fact related to the composition of clusters containing genes from three organisms is that these clusters contain similar number of human and mouse genes but different number of chicken genes.

A simple test for validating the candidate ortholog clusters is by estimating the consistency of functional annotation for sequences within the clusters. An identical functional annotation for genes from different genomes does not imply an orthologous relationship between them, on the other hand, sequences from a candidate ortholog must have identical or very similar functional annotation. We analyzed the consistency of candidate ortholog clusters using the Blast search of genes in a cluster with the genes in the Pfam database, as described in section 2. We were able to annotate 59,950 sequences of the 94,718 protein sequences in these three genomes. Among these 59,950 sequences, a total of 55,065 sequences belong to clusters and cover 72% of all sequences in the clusters. It is interesting to note that when a gene in a cluster is annotated by Pfam, all other genes in the cluster are also likely to be annotated by Pfam. Similarly, if there is a gene in a cluster that is not annotated by Pfam, it is likely that none of the genes in the cluster are annotated. In all, 86% of the clusters can be divided into two groups: clusters all of whose genes are annotated with the same set of Pfam families, and clusters none of whose genes are annotated with Pfam families.

The results of assessing homogeneity of clusters using Pfam annotations are given in Table 2. This table presents the joint distribution of groups of organisms in a cluster and the number of Pfam families which are associated with genes in the cluster. The first column represents the number of Pfam families using which genes in a cluster are annotated. The first row represents the groups of organisms whose genes are members of a cluster. For instance, there are 2,815 clusters that contain genes from three organisms but none of the genes could be annotated using Pfam. Additionally, there 1,625 clusters, containing human and mouse genes, whose genes are annotated by a single Pfam family; and there are no chicken-human clusters whose genes are associate with 3 different Pfam families.

Based on Pfam annotation for sequences within clusters, it is possible to classify the clusters into 4 different classes.

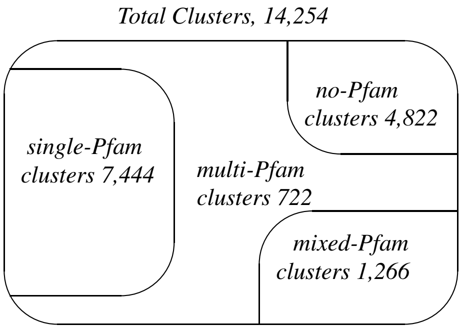

Fig. 1. Consistency of clusters according to Pfam annotations.

The *no-Pfam clusters* are those that do not contain any sequence that could be annotated using Pfam. Clusters in which all sequences are annotated by a single Pfam family are the *single-Pfam clusters*, while clusters whose sequences are annotated by different Pfam families are called the *multi-Pfam clusters*. The final class of clusters is the *mixed-Pfam clusters* and is related to clusters that contain some sequences which cannot be annotated using Pfam but those that can be annotated have a single Pfam annotation. The results for these classes are summarized in Fig. 1. The 4,822 clusters in the no-Pfam class contain 18516 sequences (24% of all sequences in clusters). These clusters cannot be validated using Pfam. Among the remaining 9,432 clusters, 7,444 clusters (52% of candidate ortholog cluster) belong to the single-Pfam class and these contain 43,321 sequences (57% of all sequences in clusters). There are 1,266 mixed-Pfam clusters comprising 7554 sequences (10% of all sequences in clusters). In summary, there are only 722 (5% of all clusters) clusters whose sequences are annotated with different Pfam families, therefore, our candidate cluster are very consistent based on Pfam annotation. When a cluster contains sequences which are annotated using different Pfam families, the cluster can be divided into "homogeneous subclusters" such that all sequences in the subclusters have the same annotation. The same annotation means that the sequences from the subcluster have hits which are members from the same Pfam family. Based on this, the entire cluster is assigned the annotation of the largest subcluster. Let us consider a cluster whose cardinality is N and suppose, its largest homogeneous subcluster has cardinality n, then the ratio n/N can be considered as a degree of consistency for the annotation of a cluster. As described earlier, there are 722 clusters in which sequences are associated with different Pfam families. The average of the degree of consistency, over clusters that are associated with more than one Pfam family, is 0.68 [2]. When this average is considered over 14,254 clusters, the value is 0.95. We analyzed the functional descriptions of different Pfam

[2] This number is a raw mean and is biased towards the contribution from the small size clusters, for instance a cluster containing 3 sequences associated with two different Pfam families contributes a maximum value of 0.67 to this average. So 0.68 implies a very degree of consistency among large sized clusters in multi-Pfam category.

TABLE II

Pfam/Organisms	C+H+M	H+M	C+H	C+M
0 pfam	2815	1382	340	285
1 pfam	6335	1625	396	354
2 pfam	537	52	42	40
3 pfam	36	4		1
Others	10			
Total	9733	3063	778	680

Annotation results for clusters using Pfam. This joint distribution shows number of Pfam families associated with a cluster, and also the organisms whose sequences are included in a cluster. The row corresponding to *0 pfam* means there are 2815 clusters containing sequences from human, mouse and chicken in which no sequences could be annotated using Pfam; similarly there are 1382, 340, 285 clusters containing sequences from human-mouse, chicken-human and chicken-mouse genomes, respectively whose sequences could not be annotated using Pfam.

TABLE III

Cluster Description	No. of clusters	Triplets in clusters
clusters with 1 triplet	437	437
clusters with 2 triplets	38	76
clusters with 3 triplets	6	18
clusters with 4 triplets	3	12
clusters with 6 triplets	1	6
Total	485	549

Relationship between candidate ortholog clusters and manually curated ortholog triplets.

families associated with the 722 clusters in the multi-Pfam class. This analysis shows that the functional description of different Pfam families associated with a multi-Pfam cluster is highly similar. For instance, cluster with id *7692* has 24 sequences of which 15 are annotated as Pfam *PF02932* and 9 sequences annotated as Pfam *PF02931*, but the functional description for Pfam *PF02932* is *Neurotransmitter-gated ion-channel transmembrane region* and the functional description for Pfam *PF02931* is *Neurotransmitter-gated ion-channel ligand binding domain*. So, despite their association with more than one Pfam family, sequences in cluster with id *7692*, are consistently annotated. This case is representative of results in general, and sequences in candidate ortholog clusters are highly consistent with Pfam annotations.

IV. COMPARISON WITH THE KNOWN 565 ORTHOLOG TRIPLETS AND THEIR ANALYSIS

When automated approaches are used to find orthologs, the methods usually find clusters of orthologous sequences. Although the sequences in a cluster of orthologous sequences are mostly related through orthologous relationship, these clusters may contain multiple sequences from an organism. These multiple sequences from an organism in an ortholog cluster are highly similar to each other and to other orthologous sequences in an ortholog cluster. Such genes are called in-paralogs [3] [REM 01] and despite high sequence similarity, these may be involved in slightly different functions from those of the main orthologs. Methods to detect orthologs whether completely automatic [VAS 04] or semi-automatic [TAT 97], [REM 01] produce clusters of orthologous genes (orthologs along with in-paralogs) that contain in-paralogs along with orthologs. On the other hand, extensive manual curation based methods [OUY 03] use a multicriteria approach to find exact orthologs in the given set of genomes. Our automatic method [VAS 04] finds clusters of orthologous genes which we call candidate ortholog clusters where as the 565 triplets [OUY 03] produced through manual curation are likely to be exact orthologs in the chicken, human, and mouse genomes.

The candidate ortholog clusters produced by the automatic method on complete genomes were compared with the set of manually curated 565 ortholog triplets. This comparison is limited to our clusters that contain genes from the ortholog triplets. We also used the Pfam annotation of genes in the ortholog triplets and the candidate ortholog clusters towards this comparison.

Since ortholog triplets are considered as benchmark in this comparison, we studied how the ortholog triplets are represented in automatically produced candidate ortholog clusters. Among the 565 ortholog triplet clusters, 549 are subsets of single candidate ortholog clusters produced by our method, while the remaining 16 ortholog triplets are destroyed in our clustering results. Of these 16 ortholog triplets, 14 are fragmented such that a pair of genes belongs to a single cluster while 2 ortholog triplets are completely destroyed, i.e, each of the genes in these triplets belongs to three different candidate ortholog clusters. There are 6 triplets that are fragmented in such a way that human-mouse genes belong to a single cluster and the chicken genes are classified as singletons. Also, there are 6 triplets that are broken such that mouse-chicken genes are in a single cluster and 2 triplets are fragmented such that human-chicken genes are in a single cluster.

The 549 triplets that are subsets of single clusters correspond to 485 different candidate ortholog clusters produced by our method (see Table 3). There are 437 triplets that are included in as many candidate ortholog clusters. Although, in most cases our clusters are extensions of triplets, there are 25 ortholog triplets that match our clusters exactly. There are 48 candidate ortholog clusters which contain more than one ortholog triplet (see Table 3).

The clusters containing more than one ortholog triplet represent a small number of cases and it was possible to analyze these clusters in detail. Among 48 such clusters, 41 are associated with single Pfam families. In the remaining 7, there are 5 instances in which, genes in the ortholog

[3] Paralogs in a genome that arise through recent duplication events.

triplets themselves are annotated by different Pfam families. For example, cluster with id *12087* contains 10 genes and the two ortholog triplets contained in this cluster are annotated as *osteoglycin* and *dermatan sulphate proteoglycan*. While all genes in the first ortholog triplet are annotated with *PF00560*, in the second triplet, only the human gene is annotated with *PF00560* whereas mouse and chicken genes are annotated with *PF01462*. The functional description for *PF00560* is *Leucine Rich Repeat* and that for *PF01462* is *Leucine rich repeat N-terminal domain* implying that the genes in the ortholog triplets are related despite the different annotation for triplets.

From the perspective of annotation for the ortholog triplets, we analyzed the candidate clusters which contained more than one ortholog triplets. We found that 34 out of 48 clusters contain triplets that are related to different subunits of the same gene. An example of this case is the cluster with id *6896*. This cluster contains 15 genes and houses 3 ortholog triplets whose annotations are *histone deacetylase 1*, *histone deacetylase 2* and *histone deacetylase 3*. All these three ortholog triplets are annotated with single Pfam family.

Thus candidate ortholog clusters compare well with existing the manually curated ortholog triplets and our candidate clusters capture and preserve the manually curated orthologs. In cases where candidate orthologs contain more than one ortholog triplets, the ortholog triplets often correspond to different domains of the same gene and would belong to the same cluster of orthologous genes (groups of orthologs along with in-paralogs) and hence should be members of the same candidate ortholog cluster.

V. Methods

In this section, we shortly describe the method to find candidate orthologs developed and validated in [VAS 04]. More precisely, we describe here a modification of the method presented in [VAS 04]. Our tests show that this modification models the ortholog problem better and thus is an improvement over the original method. Our study of orthologous genes in chicken, human and mouse genomes was conducted using the modified approach.

Let V be the set of all genes, $V =< V_1, V_2, \dots, V_k, \dots, V_n >$, where n is the number of genomes and $V_k,\ k = 1, 2, \dots, n$ is the subset of V containing genes from the genome k. An arbitrary subset H ($H \subseteq V$) can be denoted as $H =< H_{i_1}, H_{i_2}, \dots, H_{i_l} >$ where H_{i_t} is the subset of H containing genes from the genome i_t; and l is the number of different genomes in H. Below we consider only such H for which $l \geq 2$. Using this notation we can introduce a coefficient $\pi(i, H)$ to estimate the degree of membership (or, degree of orthologous membership) of the gene i to subset of genes H ($i \in H$):

$$\pi(i,H) = \sum_{\substack{t=1\\ t\neq s(i)}}^{l} p(s(i),t)\left(\sum_{j\in H_t} m_{ij} - \sum_{j\in V_t\setminus H_t} m_{ij}\right) \quad (1)$$

where m_{ij} is the similarity between the genes i and j; $s(i)$ is the genome to which i belongs and $p(s(i), t)$ is the distance between the genome $s(i)$ and t on the species tree (see fig. 2). Using the coefficient of similarity in (1) between a gene and a subset of genes from another genome, any arbitrary subset H of genes from multiple genomes is associated with a score, $F(H)$, that quantifies the strength of orthologous relationship among genes in the subset H. The score $F(H)$ is defined as:

$$F(H) = \min_{i\in H} \pi(i,H) \quad (2)$$

Then, an candidate ortholog cluster H^* is defined as the subset that has maximum score over all possible subsets of genes from the set of all genes, V: of genomes:

$$H^* = \arg\max_{H\subseteq V} F(H) \quad (3)$$

The definition (3) of an ortholog cluster requires us to solve a combinatorial optimization problem. This problem can be efficiently solved if the score function satisfies certain properties. These properties along with the algorithm are given in [VAS 04].

In an earlier formulation, we had used the $\pi(i, H) = \sum_{j\in H_t} m_{ij}$ to define the score function in (2). However, the definition in (1) models the ortholog extraction problem better. The current modified definition (1) has two components, the first component is related similarity values between genes and the second component , $p(s(i), t)$, is the distance between genomes. In the first part, the first term, $\sum_{j\in H_t} m_{ij}$, has the role to accumulates all pairwise similarity values between a gene i and all genes in the subset H_t, while the second term, $\sum_{j\in V_t\setminus H_t} m_{ij}$, estimates how this gene is related to genes from genome t that are not included in H_t. A large positive difference between these two terms ensures that the gene i_s is highly similar to genes in H_t and at the same time very dissimilar from genes not included H_t. So, a large positive value of the first component in (1) measures the similarity of the gene i_s to H_t in contrast to its similarity to genes outside H_t in the genome V_t. The first component in (1) is based on observed values of sequence similarity, but sequence similarity between orthologs from distantly related organisms is not always very high. The purpose of the second component in (1) is to bias the clustering method to include similar genes from distantly related organisms in a cluster. The function $p(s, t)$ may be defined using the time since divergence between organisms, or using the topology of the species tree for the given set of organisms.

In this paper, we define $p(s, t)$ as the height of the subtree rooted at the most recent ancestor of organisms s and t. We

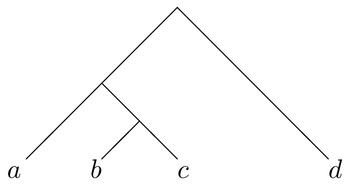

Fig. 2. An example species tree to illustrate the calculation of the genome distance.

illustrate the calculation of this distance for the hypothetical organisms a, b, c and d whose evolutionary relationship is represented by the species tree in Fig. 2. The distance $p(a, b)$ between organisms b and c is 1 because the height of subtree with leaves b and c as leaves is 1. The distance between a and b (or c) is 2 because the subtree rooted at the most recent ancestor of a and b (or c) has height 2. Similarly, the common ancestor of a and d is indicated by the root of the species tree, and so the subtree of interest in this case is the entire tree whose height is 3. Alternately, if the time-scales were represented as the branch lengths of the tree, one could define $p(s, t)$ as a function of the actual time-scales.

As described earlier, the clustering procedure requires pairwise similarity values between genes. The sequences similarity searches among the genomic sequences were performed using Blast [ALT 97]. We used the Blast in with default parameters, and the bit-score values was used to measure similarity between any two sequences. We considered only a subset of the top hits for any given gene. To be precise, if a given gene i_s in genome s has the gene j_t as its best-hit in genome t with score m_{ij}^{st}, we considered all those hits for i_s from genome t which had bit-score larger than $m_{ij}^{st}/2$. The idea behind such selection is to avoid low-scoring spurious hits for a given gene. The values sequence similarity scores between genes within an ortholog family vary across ortholog families and one can not use a constant threshold across all ortholog families. So, to avoid problems related to spurious matches without filtering out potential orthologs, and in this direction we used a top fraction of hits for a gene from another genome.

VI. Discussion

An ideal method for identifying groups of orthologous genes would involve a measure of similarity between a sequence and a subset of sequences, possibly based on a multiple alignment of a group of sequences. Further, true orthologs can be obtained by resolving the gene tree for genes in the ortholog cluster with the species tree. Since multiple alignment is computationally expensive and gene-tree species-tree resolution cannot be solved efficiently either, such approaches are in principle difficult. Although, approaches to use multiple alignments followed by resolving gene-tree with species trees are known in principle, there are few methods that actually do it [ABA 02]. The method, we have used, addresses the first of these issues by using a measure of similarity between a gene and a group of genes. The method ignores the similarity between genes within a genome, and this enables it focus on detection of orthologs rather than paralogs.

In this paper, we also show a way to incorporate the phylogenetic information through the use of (1) to bias the procedure to recognize related genes in distantly related organisms. Such bias can be introduced into the procedure using various different formulations and we present a simple but effective means to achieve the desired effect. One of the issues with introducing a bias has to do with the strength of the bias; if one introduces a large bias, the procedure may become specific for producing ortholog clusters that contain genes only from distantly related organisms. However, in our experiments using the proposed formulation, we did not encounter such a phenomenon.

We found candidates of ortholog clusters in the complete genomes from chicken, human and mouse, using an automatic method to screen for ortholog clusters. About 20% of genes are those that appear to be specific to a single organism, on the other hand, 70% of all genes belong to clusters that contain genes from all the three organisms. Although same Pfam annotation for genes does not indicate orthologous relationship, such annotation can be used to measure consistency of genes within clusters. We found that 52% clusters contain genes that share a single Pfam annotation while 34% of clusters contain genes that could not be annotated with Pfam families.

We compared the candidate ortholog clusters with manually curated known 565 ortholog triplets on the same genomes. These triplets were obtained through multicriteria approach using extensive manual curation when mouse and chicken genomes were incomplete. Comparison with manually curated known ortholog triplets show that our method recapitulates the stringency of these methods, in other words, most often the manually extracted ortholog triplets are parts of our candidate ortholog clusters. In some cases, our clusters include more than one manually curated ortholog cluster but annotations of such clusters indicates that sequences in these ortholog clusters are highly similar.

The 565 ortholog triplets were extracted when the chicken and mouse genomes were only partially complete and all genes from these organisms were not available for analysis. We analyzed these triplets to assess their stability in the presence of complete genome data. Our candidate clusters show that 549 of these 565 triplets can not be destroyed even the presence of complete genome data, and 2 among remaining 16 are completely destroyed such that each of the genes in these triplets belongs to different clusters. This shows that manually extracted clusters are mostly stable even though they were extracted from an incomplete data.

An assessment of the stability of manually extracted ortholog triplets was performed using the score function we have proposed (2). This test determines if a given ortholog cluster can be fragmented into smaller subclusters. We call a cluster *compact*, if the given cluster does not contain any

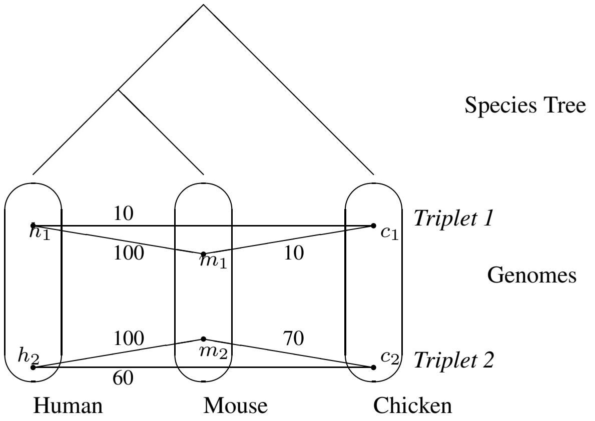

Fig. 3. The figure illustrates evolutionary relationship between human, mouse and chicken using the species tree. It also shows two example ortholog clusters in these genomes found by experts. The first ortholog triplet *Triplet 1* contains genes h_1, m_1 and c_1. The strength of similarity between these genes is represented as weights on the respective edges. Similarity, the second ortholog triplet *Triplet 2* contains genes h_2, m_2 and c_2.

subset whose score value is larger than the score value of the original cluster. To illustrate compactness of a triplet consider the two example triplets shown in Fig. 3. According to the species tree for chicken, human and mouse shown in the figure, the distance between human and mouse is 2, while that between chicken and human or chicken and mouse is 2. Then, using (2), it is clear that the score value for the entire *Triplet 1* is 40 while the score value for the subset $\{h_1, m_1\}$, containing human and mouse genes, is 60. So, *Triplet 1* is not compact as it contains a subset with a larger score value. When we consider the *Triplet 2*, the highest scoring subset is the entire triplet with a score of 220, and it is therefore compact.

We found that 562 of the 565 triplet ortholog clusters do not contain any subset whose score value is higher than the score value for the entire triplet. There are three triplets that are not compact. All of these triplets are related to genes that are involved in cell and organism defense. One of these triplets annotated as IL-2 was found solely based on the functional properties. The chicken *IL-2* gene has no sequence similarity to the human and mouse *IL-2* genes. The other two triplets, annotated as *interferon alpha* and *CD4*, are also cases in which the similarity value between human and mouse is much higher than similarity value between chicken-human or chicken-mouse genes. In fact, for the triplet related to *CD4*, we found another gene in the complete chicken genome whose similarity to the human and mouse genes is much higher compared to the gene in triplet. This case demonstrates that extracting ortholog clusters in incomplete genomes, even through rigorous analysis can occasionally be error prone.

The proposed method provides not only finds candidate orthologs but also provides an infrastructure for various other comparative genomic studies through the use of different similarity functions. The ortholog clusters found in this study are useful for a deeper understanding of the relationship between mammals and aves. It would be interesting to study various rates of evolution and substitution bias using ortholog clusters built on complete genome data. We expect that this would provide similar results as shown by using the manually curated ortholog clusters. Our method can find candidate ortholog clusters in a large set eukaryotic genomes and thus enables us to study variations and similarities in evolutionary rates of different organisms and different protein families.

REFERENCES

[ABA 02] Abascal F., Valencia A., *Clustering of proximal sequence space for identification of protein families*, *Bioinformatics*, vol. 18, p. 908-921, 2002.

[AL. 04] Birney E. B. , et al. , *An overview of Ensembl*, *Genome Res*, vol. 14, p. 925-928, 2004.

[ALT 97] Altschul S., Madden T., Schaffer A., Zhang J., Zhang Z., Miller W., Lipman D., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, *Nucleic Acids Res*, vol. 25, p. 3389-3402, 1997.

[BAT 04] Bateman A., Coin L., Durbin R., Finn R. D., Hollich V., Griffiths-Jones S., Khanna A., Marshall M., Moxon S., Sonnhammer E. L. L., Studholme D. J., Yeats C., Eddy S. R., *The Pfam protein families database*, *Nucleic Acids Res*, vol. 32, p. 138-141, 2004.

[BEN 04] Benson D., Karsch-Mizrachi I., Lipman D., Ostell J., Wheeler D., *GenBank: update*, *Nucleic Acids Res*, vol. 32, p. D23-26, 2004.

[CHI 03] Chinwalla A. T., et al., *Initial sequencing and comparative analysis of the mouse genome*, *Nature*, vol. 420, p. 520-562, 2003.

[FIT 70] Fitch W. M., *Distinguishing homologous from analogous proteins*, *Systematic Zoology*, vol. 19, p. 99-113, 1970.

[HIL 04] Hillier L. W., et al., *Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution*, *Nature*, vol. 432, p. 695 - 716, 2004.

[LAN 01] Lander E. S., et al., *Initial sequencing and analysis of the human genome*, *Nature*, vol. 409, p. 860-921, 2001.

[MAK 98] Makalowski W., Boguski M., *Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences*, *Proc Natl Acad Sci USA*, vol. 95, p. 9407-9412, 1998.

[OUY 03] Ouyang M., Case J., Tirunagaru V., Burnside J., *Five hundred sixty-five triplets of Chicken, Human and Mouse candidate orthologs*, *J Mol Evol*, vol. 57, p. 271-278, 2003.

[REM 01] Remm M., Strom C. E., Sonnhammer E. L., *Automatics clustering of Orthologs and In-paralogs from pairwise species comparisons*, *J Mol Biol*, vol. 314, p. 1041-1052, 2001.

[TAT 97] Tatusov R., Koonin E., Lipmann D., *A genomic perspective on protein families*, *Science*, vol. 278, p. 631-637, 1997.

[VAS 04] Vashist A., Kulikowski C. A., Muchnik I. B., Automatic screening for groups of orthologous genes in comparative genomics using multiple-component clustering, Technical Report 2004-33, DIMACS, 2004.