ISSN: 2153-0602

OPEN ACCESS Freely available online

Journal of Data Mining in Genomics & Proteomics

Research Article

Whole-Proteome Tree of Insects: An Information-Theory-Based “Alignment-Free” Phylogeny and Grouping of “Proteome Books”

JaeJin Choi[1,2], Byung-Ju Kim[1,3,4], Sung-Hou Kim[1,3,4*]

[1]Department of Chemistry and Center for Computational Biology, University of California, Berkeley, USA;[2]Division of Biological Systems and Engineering; [3]Department of Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, USA;[4]Human Genome Research Center, Incheon National University, Incheon, Korea

ABSTRACT

Background: An “organism tree” of insects, the largest and most species-diverse group of all living animals, can be considered as a metaphorical and conceptual tree to capture a simplified narrative of the complex and unpredictable evolutionary courses of the extant insects. Currently, the most common approach has been to construct a “gene tree”, as a surrogate for the organism tree, by selecting a group of highly alignable regions of each of the select genes/proteins to represent each organism. However, such selected regions account for a small fraction of all genes/proteins and even smaller fraction of whole genome of an organism. During last decades, whole-genome sequences of many extant insects became available, providing an opportunity to construct a “whole-genome or whole-proteome tree” of insects using Information Theory without sequence alignment (alignment-free method).

Results: A whole-proteome tree of the insects shows that (a) the demographic grouping-pattern is similar to those in the gene trees, but there are notable differences in the branching orders of the groups, thus, the sisterhood relationships between pairs of the groups; and (b) all the founders of the major groups have emerged in an “explosive burst” near the root of the tree.

Conclusion: Since the whole-proteome sequence of an organism can be considered as a “book” of amino-acid alphabets, a tree of the books can be constructed, without alignment of sequences, using a text analysis method of Information Theory. Such tree provides an alternative view-point of constructing a narrative of evolution and kinship among the extant insects.

Keywords: Organism tree; Whole-genome tree; Feature Frequency Profile; Jensen-Shannon divergence; Evolutionary progression scale; Cumulative genomic divergence; Arthropodal burst

INTRODUCTION

Sequence-alignment-based “gene trees”

An “organism tree” of insects can be considered as a practically useful narrative to convey a simplified evolutionary relationship among the insects. However, it is a metaphorical and conceptual tree that cannot be experimentally validated. Thus, it is expected that the effort will be continuing to find one or more “surrogate trees” derived from various descriptors of the characteristics associated with each insect and to find improved methods to estimate evolutionary distances from the divergence of the descriptors under as few subjective assumptions as possible at the time of investigation. At present, the best descriptor of an insect is its whole-genome sequence information. However, for several decades, due to the technical difficulties and high cost of whole genome sequencing, the most practically feasible and common approach to construct a surrogate tree has been to construct a “gene tree”. To construct a gene tree a group of genes common among the insects of study are selected, align the sequence regions that are highly homologous for each gene family under an assumption that such regions represent the characteristics of each insect organism. Then, the divergence of certain characteristics, most commonly, point substitution rates within each paired aligned-regions, is calculate to represent the evolutionary distances among the insect organisms.

Such “alignment-based” gene trees, strictly speaking, represent the phylogeny of the regions of the selected genes, but not full aspects all genes, because the aligned regions account, in general, for a very small fraction of the all genes, let alone of the whole genome.

Correspondence to: Sung-Hou Kim, Department of Chemistry, University of California, Berkeley, CA. USA, Tel: 510-708-2564, E-mail:sunghou@berkeley.edu

Received: September 21, 2021; **Accepted:** October 5, 2021; **Published:** October 12, 2021

Citation: Choi JJ, Kim BJ, Kim SH (2021) Whole-Proteome Tree of Insects: An Information-Theory-Based “Alignment-Free” Phylogeny and Grouping of “Proteome Books”. J Data Mining Genomics Proteomics. 12:002.

Furthermore, this approach makes a set of debatable assumptions about evolutionary process such as: (a) the collection of such highly homologous sequence-regions contains enough information to represent each organism, (b) non-homologous genes do not contribute to the evolution of individual organisms and (c) one type of mutation, e.g., point substitutions, account for the most important mutational events of evolution. Thus, strictly speaking, such gene trees may reveal the evolutionary narrative of the selected genes, at best, but not of the insect organisms [1].

Information-theory-based ("alignment-free") "whole-genome trees"

This situation has since changed significantly in two important aspects: (a) During last decades, a large number (over 134 species, mostly insects, as of 2020) of whole-genome sequences of extant insect species have been accumulating in public databases, and (b) Information Theory, developed to analyze linear electronic signals, was found to be adaptable to analyze other linear information, such as natural languages and genomic information, without sequence alignment ("alignment-free"). In this approach, whole content of a digitized linear information, such as simple binary electronic signals as well as complex natural languages and whole genome sequence, can be described by "n-Gram" or "k-mers". N-Gram of a linear sequence is the collection of all overlapping subsequences of length n, and it contains all information necessary to reconstruct the original sequence. Furthermore, the information divergence (difference) between two n-Grams can be estimated by, for example, Jensen-Shannon divergence without alignment, alignment-free, of the sequences. Such approach has been widely tested and validated for comparing texts and books of natural languages since 1990s and gene sequences consisting of coding and non-coding regions since 1986 [2-5].

Such validated methods have been adapted and optimized to handle whole-genome/proteome sequences in Feature Frequency Profile (FFP) method by representing, first, the whole-genome sequence or whole-proteome sequence of an organism as a "book" of 4-base alphabets or 20-amino acid alphabets, respectively, without space or delimiters within the book or within each chapters [6]; second, each "book" is formulated into an n-Gram, which is named here as Feature Frequency Profile (FFP), where a Feature is a unique component of n-Gram, which is a string of linked alphabets of an optimal length. Thus, each FFP has all the information to reconstruct the "book" of a given whole genome/proteome. The analogy here is that Feature Frequency Profile represents a "genome/proteome book", just as "Word Frequency Profile" represents a regular book in text analysis method of Information Theory.

OBJECTIVE

In FFP method, whole-genome information is used under the assumptions very different from the gene trees: (a) whole genome or proteome sequence of an organism represents the organism better than those of short regions of highly homologous sequence from a set of selected genes used in the gene trees and (b) a combination of all types of mutations, such as point substitution, insertion/deletion of various length, recombination, duplication, transfer or swapping of gene etc., contribute to the evolutionary processes of the organisms, rather than only point substitution rates in the gene trees. Thus, whole-genome/proteome tree may provide an independent view of the evolutionary relationship among the insects.

Earlier examples of whole-proteome trees

In the last decade, such genome trees have been constructed for species of Bacteria and Archaea Domain, of Fungi Kingdom and, most recently, of samples of all domains and kingdoms of all living organisms at Order level using over 4,000 species with complete or near-complete genome sequences in public databases [7-9]. These studies have shown that: (a) among three types of genome trees (whole-genome DNA tree, whole-transcriptome RNA tree, and whole-proteome amino-acid tree) the whole-proteome tree produces the most topologically stable trees; and (b) the grouping of extant organisms based on whole-proteome sequence agrees well, in general, with those of the corresponding gene trees, but there are notable differences in the evolutionary branching order among the groups and significant differences in relative branch lengths.

In this study we present a view of the whole-proteome Tree of Insects (ToIn), based on whole-proteome sequences of 134 diverse arthropod species (123 insects plus 11 non-insect arthropods), available in the NCBI database [10].

RESULTS

To compare the gene trees with our whole-proteome ToIn we chose two recent and very comprehensively-analyzed gene trees: The first one is the recent "alignment-based" gene tree of 144 insect taxa based on 1,478 single-copy protein-coding nuclear genes [11]. The second is the gene tree for 76 arthropod taxa [12] based on up to 4,097 single copy protein-coding genes. In both cases, the number of alignable genes used is a small fraction of about 10,000 to 31,000 genes in each whole genome of extant arthropods. Both gene trees agree with each other on the branching order of the Order groups of the study populations, and on similar time spread of the emergence of the founders of the groups in chronological time scale estimated based on available fossils and calibration methods under various assumptions.

In comparing our whole-proteome ToIn to these two gene trees, we focus on two aspects separately: grouping patterns and phylogeny of the groups. For the former, we cluster our study population by several unsupervised clustering algorithms using only the evolutionary distances estimated from the "divergence" distances among whole-proteome FFPs, as whole proteome characteristics (see Construction of whole-proteome Tree of Insects in Materials and Methods) with no assumption of specific evolutionary models. We then ask whether the "clustering pattern" is similar to the "clading pattern" in the gene trees and in our whole-proteome ToIn, recognizing that the tree constructions assume respective evolutionary models in addition to the evolutionary distance estimations. For the phylogeny of the groups, we compare the order of branching of the groups and their emergence point on the evolutionary progression scale (see "Cumulative Genomic Divergence (CGD)" and "Evolutionary Progression Scale" in Materials and Methods).

Robust demographic grouping pattern by clustering and clading

Clustering: We have tested the grouping pattern of the insects by several unsupervised clustering algorithms, such as Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS), and t-Distributed Stochastic Neighbor Embedding (t-SNE) [13,14], all of which are based only on the evolutionary distances, estimated by the divergence of whole-proteome sequences among all pairs of the study organisms (see Construction of whole-proteome Tree of Insects in Materials and Methods). Of these, t-SNE and classical PCA showed the clustering pattern most compatible with the current grouping of arthropods with common and scientific names, mostly based on morphological characteristics. Figure 1, a t-SNE clustering, shows about 17 clusters of the organisms in this study that are most stable under a wide range of "perplexity", the adjustable parameter of t-SNE, which is roughly related to the "looseness" of each cluster. Another clustering by PCA (see Supplemental information) also shows a similar number of clusters, although some of them are not as well resolved as in t-SNE. Such approximate agreement is unexpected, because the assumptions and algorithms in t-SNE and PCA are completely different (Figure 1).

Clading: As a second method of the grouping, we use the clading pattern of the organisms in our whole-proteome ToIn. Figure 2 shows the topology of the ToIn, constructed using Neighbor-Joining method implemented in BIONJ [15,16]. In this study, we use the divergence distances of whole-proteome sequences as the estimates for the evolutionary distances between the organisms (see Construction of whole-proteome Tree of Insects in Materials and Methods). We also assume an evolutionary model of Maximum Parsimony (minimum evolution) in a way that the "chosen neighbors" to be joined are those that minimize the total sum of branch-lengths at each stage of step-wise joining of neighbors starting from a star-like tree [15]. The tree shows that all 17 clusters in Figure1 can be identified among the clades in the ToIn. (Figure 2)

Robustness of grouping: The grouping pattern by clustering and clading based on whole-proteome information also agrees well with those of the gene trees except for Hemiptera group (see Notable differences in grouping and phylogenic positions in Discussions). Thus, it is surprising that the demographic grouping pattern is robust regardless of not only the information type (morphology, select protein-characteristics or whole-proteome characteristics), but also the methods (clustering or clading) used in grouping the extant organisms (see Similarities in grouping patterns in Discussion). However, not surprisingly, there are significant differences from the gene trees in branch-length and branching order of the groups, which are used to predict the position, in the evolutionary progression scale, when the "founders" of each group emerged (see below).

Emergence of the "Founders" of all major groups in a staged "burst"

For the following results we define "Cumulative Genomic Divergence (CGD)" for an internal node of the ToIn as the cumulative scaled-branch-length from the tree root to the node (see Cumulative Genomic Divergence (CGD) in Evolutionary Progression Scale in Materials and Methods) to represent the extent of the "evolutionary progression" of the node. The progression has a scale such that the root node of ToIn is at CGD=0 (see Outgroup in Discussions) and the leaf nodes of the extant organisms at CGD 100, on average.

"Arthropodal burst" near the root of ToIn: Figures 3 and 4 show the whole-proteome tree with CGD values. It reveals that the "founders" of all major groups of insects as well as non-insect arthropods (at Subphyla and Order levels) emerged in a staged burst within a short evolutionary progression span between CGD of 1.6 and 5.8 (marked by a small red arc), near the root of the tree in Figures 3-5. This burst is reminiscent of the "deep burst" of the founders of all Kingdoms of Life in the whole-proteome Tree of Life (Figure 2 of reference 9).

Order of emergence of the "founders" for all major groups: Figures 4 and 5 shows a series of staged emergence of the founders of all major groups within the burst. The first groups to emerge (branch out) in a narrow range of CGD values of 1.6 and 1.9 are the founders of Subphylum Crustacea and Chelicerata (arachnids and Atlantic horseshoe crab). Then, the founders of members of Class Insecta emerge: the founders of Hemiptera-A group (aphids and a psyllid) at CGD of 3.7, the founders of Diptera group at CGD of 4.1, and the founders of the remaining five Order-level groups (Lepidoptera, Hemiptera-B (bugs, a planthopper and a whitefly), Coleoptera, Blattodea+a thrips, and Hymenoptera groups) at CGD of 4.4, 4.8, 5.2, 5.8, and 5.8, respectively (Figures 4 and 5).

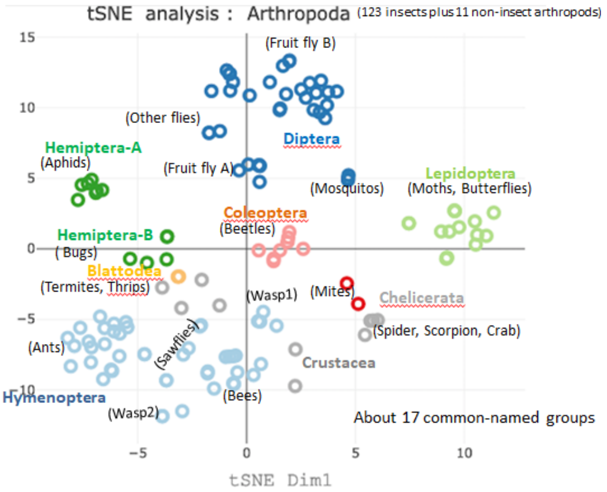

Figure 1: Unsupervised clustering (grouping) of 134 extant arthropods (123 insects plus 11 non-insect arthropods) by t-SNE. t-SNE [14] is a machine learning algorithm of clustering and reducing a high-dimensional data into a two or three dimensional space for easy visualization by emphasizing resolution of clusters, but de-emphasizing the distances between the clusters. There are about 17 clusters with their common names in parentheses. These clusters can be assorted into six Order groups and two Subphylum groups in colored bold-letters, correspond to Hemiptera, Lepidoptera, Diptera (blue), Coleoptera (pink), Blattodea (yellow), Hymenoptera (green), Chelicerata (gray and red) and Crustacea (gray). Some clusters are loose, and there are a few unclustered organisms.

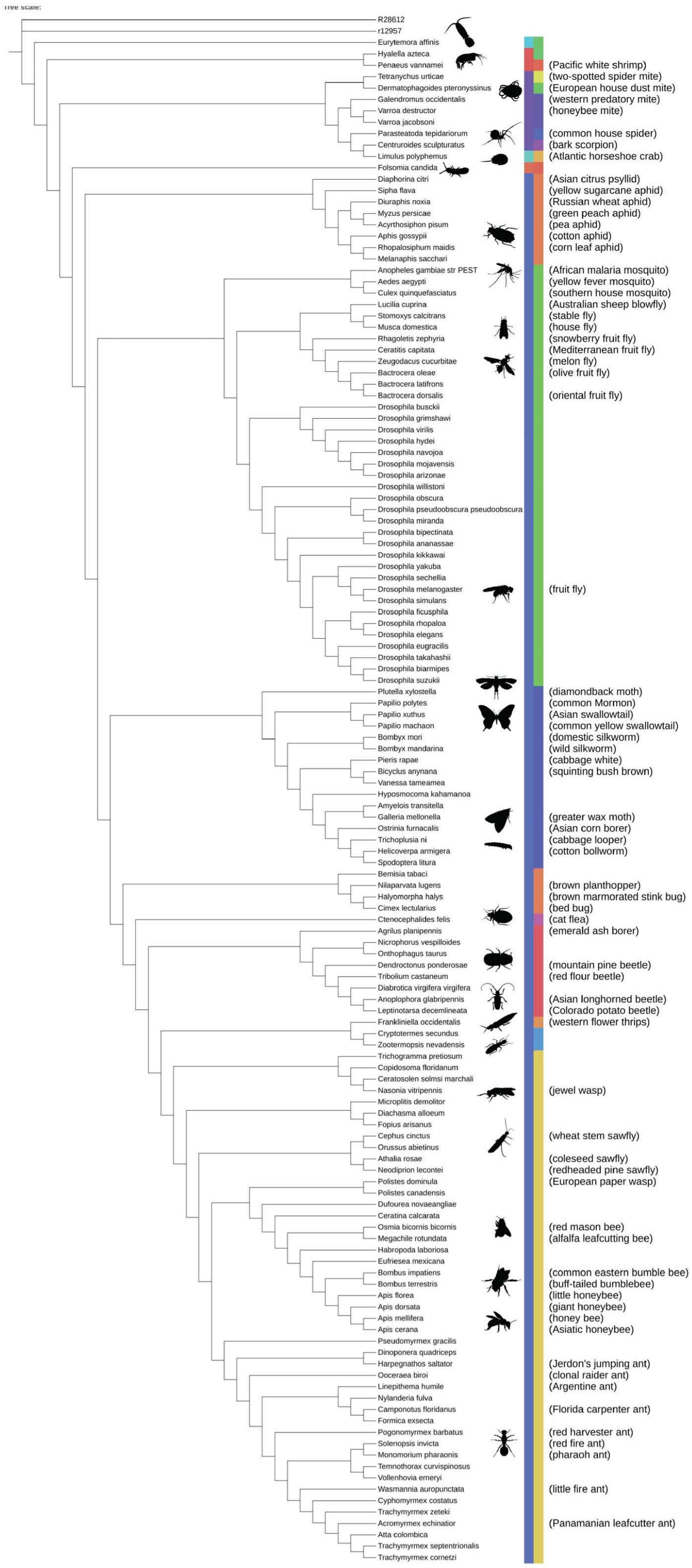

Figure 2: Topology of the linear representation of whole-proteome ToIn. The colors of the first (inner) band distinguish organisms in different Classes, and those of the second (outer) band among different Orders (the names of different color-bands are shown in Fig. 3). Scientific names and common names, when available, of each organism are also listed. The silhouettes of sampled organisms are shown next to their names. To emphasize the clading pattern, all branch-lengths are ignored. The first two items refer to two members of the outgroup (see Outgroup in Discussions) constructed by shuffling [17, 18] the whole-proteome sequences of two arthropods. The visualization of the ToIn was made using iTOL [19-21].

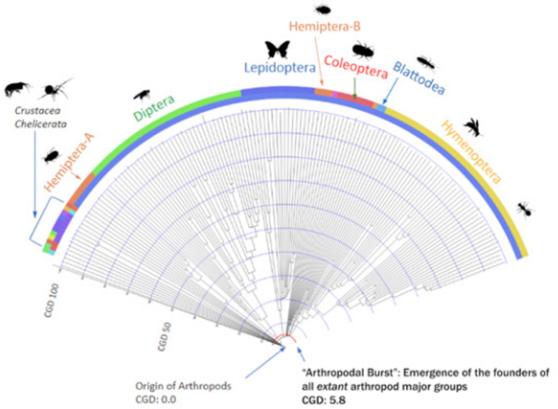

Figure 3: "Pie" representation of whole-proteome ToIn with cumulative branch-lengths scale. This view of the whole-proteome ToIn shows all branch-lengths to emphasizes the progression of evolution of each member in the study population from the origin at CGD=0 to the extant forms of the members at CGD= near100. The small red arc near the root is at CGD=5.8, by which point of the evolutionary progression, the founders of all major groups (consisting of 7 Order groups and 2 Subphylum groups shown in Fig. 1) have emerged in a staged "burst", suggesting that the remaining 94.2 on CGD scale corresponds to further diversification and gradual evolution of the founders and common ancestors within each major group toward their extant forms. The visualization of the ToIn was made using iTOL [21].

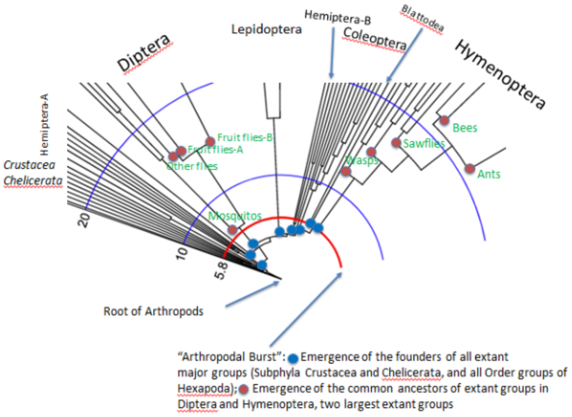
Hemiptera-B
Blattodea
Lepidoptera
Coleoptera
Diptera
Hymenoptera
Hemiptera-A
Crustacea
Chelicerata
Fruit flies-B
Fruit flies-A
Other flies
Bees
Sawflies
Wasps
Ants
Mosquitos
20
10
5.8
Root of Arthropods
"Arthropodal Burst": ● Emergence of the founders of all extant major groups (Subphyla Crustacea and Chelicerata, and all Order groups of Hexapoda); ● Emergence of the common ancestors of extant groups in Diptera and Hymenoptera, two largest extant groups

Figure 4: Expanded view of Fig. 3 near the root of the whole-proteome ToIn. Examples of the founders of all major groups are shown as blue dots, and the common ancestors of extant groups within two major groups, Diptera and Hymenoptera, as red dots. The visualization of the ToIn was made using iTOL [21].

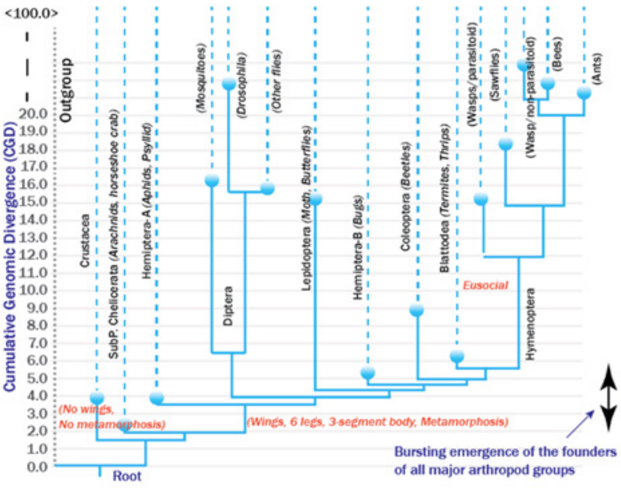
<100.0>
Cumulative Genomic Divergence (CGD)
20.0
19.0
18.0
17.0
16.0
15.0
14.0
13.0
12.0
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
Outgroup
Crustacea
SubP. Chelicerata (Arachnids, horseshoe crab)
Hemiptera-A (Aphids, Psyllid)
(Mosquitoes)
(Drosophila)
(Other flies)
Diptera
Lepidoptera (Moth, Butterflies)
Hemiptera-B (Bugs)
Coleoptera (Beetles)
Blattodea (Termites, Thrips)
(Wasps/parasitoid)
(Sawflies)
(Wasp/non-parasitoid)
(Bees)
(Ants)
Eusocial
Hymenoptera
(No wings, No metamorphosis)
(Wings, 6 legs, 3-segment body, Metamorphosis)
Bursting emergence of the founders of all major arthropod groups
Root

Figure 5: Simplified whole-proteome ToIn. The vertical axis shows cumulative genomic divergence (CGD) values, which ranges from zero to around 100, and they correspond to the extent of evolutionary progression from the root of the ToIn to the extant leaves. For simplicity, "singletons" (that do not belong to any named groups) are not shown, and all the leaf nodes and their branches of a common-named group (in parenthesis) are combined into a single dotted line coming out from their common ancestor node of the extant group shown as a sphere. Each internal node represents a "pool of founding ancestors" (see Supplementary Information Fig. S3). Dotted vertical lines are to indicate that they are arbitrarily shortened to accommodate large jumps of CGD values within a limited space of the figure. The double-headed arrow at bottom right indicates the short range of the CGD values, within which the founders of all the major groups of the extant organisms in this study have emerged in a "burst". For our interpretation of horizontal lines and vertical lines, see Supplementary Information Fig. S3.

DISCUSSION

Similarities in grouping patterns

As mentioned earlier, it is surprising that the grouping patterns at Order level between the gene trees and our whole-proteome ToIn are very similar (see below for one notable exception of Hemiptera) despite the facts that the types of input data (multiple-aligned regions of selected proteins vs. whole-proteome) and grouping methods used (clustering vs. clading) are very different. A possible implication is that each group members evolved, during the most of the evolutionary progression, largely "isolated" within the group without significant genomic intermixing between the groups, thus, resulting in each group having much smaller genomic variation within the group than between the groups.

Dissimilarities in branching orders and branch-lengths

It is not surprising that the branching orders and branch-lengths are not similar, because the assumptions under which the estimations for evolutionary distances among the organisms are very different: in the gene trees, the distances are calculated for the aligned portions of the selected genes for, e.g., point-substitutional mutations, while in our tree they are done for all proteins going through all types of mutations.

Evolutionary progression scale vs. chronological time scale

Chronological time scale is linear, but the "evolutionary progression scale", which is proportional to the degree of whole-genomic divergence, is most likely not strictly linear, because any significant geological events may accelerate or decelerate the evolutionary progression globally for all organisms, as well as differently for each subgroup of organisms. However, the direction of arrows in both scales are the same, suggesting that the two scales may be calibrated when sufficient fossils, other independent records, and improved calibration methods become available (see "Cumulative Genomic Divergence (CGD)" in "Evolutionary Progression Scale" in Materials and Methods). Meanwhile, we use the evolutionary progression scale to compare the order of emergence of the founders of various major groups under the assumption that the whole-genome divergence among organisms increases as evolution progresses, similar to the physical entropy of universe increases as the universe evolve.

"Burst" vs. gradual emergence of the founders of major groups

While cognizant of the difference and similarity of the two scales, the most dramatic difference is observed in the span of the scales within which the founders of all major groups at Order level emerged in the gene trees and in our whole-proteome ToIn. In the gene trees, the founders of all the groups at Order level emerged gradually during a long chronological time span of about 350 Million years (Myrs) corresponding roughly 60% of about 570 Myrs between the tree root to the extant arthropods (Figure 1 of reference 11), or about 210 Myrs corresponding to roughly 37% of the same full chronological scale Figure 2 of reference 12). In drastic contrast, the founders of all the major groups in our tree emerged within about 5.8% of the full evolutionary progression scale in a sudden burst ("Arthropodal burst") near the root of our whole-proteome tree, followed by a long gradual evolution spanning the most of the evolutionary progression scale (CGD of 5.8 to 100 on average) in Figures 3,4. Despite the quantitative difference between the chronological time scale and the evolutionary progression scale mentioned above, this drastic contrasting observations between the two types of trees may have an important qualitative evolutionary implication in constructing the narrative for the birth of the insects.

Notable differences in grouping and phylogenic positions

Despite the drastic difference in the emergence pattern of the founders (burst vs. gradual) mentioned above, the order of emergence of the major groups at Order and Subphylum levels agree between the gene trees and our whole-proteome tree except some notable differences in Hemiptera and Blattodea as discussed below. These differences may get resolved once the whole genome sequences of more diverse members of the two groups become available.

Hemiptera: In the gene trees, the bugs, cicadas, and aphids form Hemiptera clade [11,12]. But, in our whole-proteome ToIn (Figure 3) as well as in t-SNE (Figure 1) and PCA clustering plots (Supplemental Information), Hemiptera is divided into two separate clusters, which we call Hemiptera-A (aphids and a psyllid) and Hemiptera-B (planthopper, whitefly, stink bug, bed bug, cat flea, ash borer). The former consists of the most "primitive" Hemiptera, which feed passively on mostly one plant species or its closely related plants, is at the basal position of all Hexapoda. But, the latter is at basal or sister position to a larger clade consisting of Coleoptera, Blattodea and Hymenoptera. The separation into two clusters (which are based solely on the whole proteome sequences only, and not constrained by the assumptions and related confounding issues associated with algorithm of tree building by Neighbor-Joining) indicates that the difference in the whole-proteome information contents between the two separate clusters is greater than those within each cluster. The separation is also observed in our ToIn despite the algorithmic constraints of Neighbor-Joining method.

Blattodea: Two termites (Blattodea) and one thrips (Thysanoptera), both eusocial and hemimetabolous, form a clade in our whole-proteome ToIn and the clade is sister (or basal) to Hymenoptera group, which is also eusocial but holometabolous. However, in one gene tree [11], Blattodea group (cockroaches and termites, which are eusocial and hemimetabolous) is a member of a larger clade Polyneoptera and placed at the basal position to all other Order groups of Insecta, which are largely non-social and hemi- or holo-metabolous, while, in the other gene tree [12], Blattodea group forms a separate clade, and is placed near the basal position of all other Order groups of Insecta. This is in contrast to what we observe in our ToIn, where non-social Hemiptera-A, not Blattodea, is the basal group of Insecta, consistent with the assumption that eusociality of Blattodea is likely to be a trait acquired late in the insect evolution.

Out-group

Since our method does not require multiple sequence alignment, we constructed, as was described in our earlier works on whole-proteome trees, the proteome sequence of an "artificial (faux) arthropod" by "shuffling" the alphabets of the whole proteome sequence of an organism in the study group [17,18]. We used two such artificial arthropods (named R28612 and r12957) to form the out-group for the study. Each has the same size and amino acid composition of corresponding protein of an extant arthropod, but does not have gene sequences information for the organism's survival.

SUMMARY

A whole-proteome tree constructed using "alignment-free" FFP method is based on a set of assumptions and algorithms that are fundamentally different from those used in the most-commonly accepted "alignment-based" gene trees based on homologous proteins. Our tree provides not only another view-point to consider in constructing the narrative of kinship among the extant insects (plus a few non-insect arthropods) in evolutionary progression scale, but also highlights some of the significant differences between the two types of trees that may have important implications in understanding the evolution of insect diversity. Our tree raises the following debatable issues

1. The founders of all order-level groups of the extant insects emerged in a "burst" rather than through a long stretch of chronological time as suggested in the gene trees.

2. Hemiptera-A, consisting of some of the most "primitive" Hemiptera, may be the basal group of all members of Insecta, rather than eusocial Blattodea as suggested in the gene trees.

3. Blattodea may be sister to Hymenoptera, a eusocial group.

MATERIALS AND METHODS

Design concept

Construct a "whole-proteome tree" by the following design concepts:

- Treat each whole-proteome sequence of an insect as a "book" of the amino acid alphabets
- Compare the information content among all "books" of the study population by calculating the divergence of the information content using a text comparison method of Information Theory [19].
- Construct the whole-proteome tree of the study population to estimate a cumulative genomic divergence as the evolutionary distance for each internal node of the tree
- Compare the whole proteome tree with recent gene trees of insects/arthropods in grouping and the kinship among the groups

Sources and selection of proteome sequences

We downloaded the proteome sequences for 134 arthropods from NCBI RefSeq DB using NCBI FTP depository [10] (as of December 2019). Protein sequences derived from all organelles were excluded from this study. Also excluded from our study are those derived from whole genome sequences assembled with "low" completeness based on two criteria

a) The genome assembly level indicated by NCBI as "contig" or lower (i.e. we selected those with the assembly levels of 'scaffold', 'chromosome' or 'complete genome')

b) The proteome size smaller than the smallest proteome size among highly assembled arthropod genomes (Anopheles gambiae str. PEST with 14,089 proteins at "chromosome" assembly level; TaxID 180454).

All taxonomic names and their Taxon Identifiers (TaxIDs) of the organisms in this study are from NCBI taxonomy database, and listed in Supplementary Information, Dataset S1, where "N/A" indicates an unassigned taxonomic order.

Construction of whole-proteome tree of insects

Based on our earlier experiences of constructing whole-proteome trees of prokaryotes, fungi and all life forms by Feature Frequency Profile (FFP) method, following choices have been made to obtain a topologically stable whole-proteome ToIn of maximum parsimony (minimum evolution) by BIONJ:

a) Among three types of genomic information (DNA sequence of the whole genome, RNA sequence of whole transcriptome and amino acid sequence of whole proteome) whole-proteome trees are most "topologically stable" as estimated by Robinson-Foulds metric at respective "optimal Feature-length" [20].

b) For FFP as the "descriptor" of the whole proteome of each organism, the optimal Feature-length is about 10 amino-acid string (Figure S2).

c) Jensen-Shannon Divergence (JSD) is a convenient measure of "divergence of information content" bound between zero and 1, as the distance of dissimilarity between two whole-proteome descriptors, for constructing the distance matrix of BIONJ. It is important to note that such FFP of a whole-proteome sequence of an organism has all the information necessary to reconstruct the original whole proteome sequence.

"Cumulative Genomic Divergence (CGD)" and "Evolutionary progression scale"

In Information Theory, the Jensen-Shannon Divergence (JSD), bound between zero and one, is commonly used as a measure of the dissimilarity between two probability distribution of informational features. The FFP as the descriptor for a linear sequence information of the whole proteome of an organism is such a probability distribution. Thus, a JSD value of two FFPs, used as a measure of the divergent distance between two proteome sequences, is also bound between 0 and 1, corresponding to the JSD value between two FFPs of identical whole proteome sequences and two completely different whole proteome sequences, respectively. Any whole proteome-sequence "dissimilarity" between two extant organisms accumulated during the evolution can be considered as caused by changes of, ultimately, genomic sequences due to all types of mutational events, such as point substitutions, indels, inversion, recombination, loss/gain of genes, etc. as well as other unknown mechanisms, and they will bring JSD somewhere between 0.0 and 1.0 depending on the degree of the sequence divergence.

In this study the collection of the JSDs for all pairs of the study organisms plus 2 out-group members constitutes the "distance matrix" for BIONJ. Since all the branch-lengths are derived from the JSD values, the cumulative branch-length of an internal node, which we call "cumulative genomic divergence (CGD)" of the node along the presumed evolutionary lineage, can be considered as the point of evolutionary stage reached by the node on an "evolutionary progression scale". For convenience of assigning the nodes on the progression scale, CGDs are scaled such that the CGD value at the root node of the ToIn is at zero and the leaf nodes of the extant organisms around 100 on average, corresponding to the current genomic states of the organisms, which we define as the end point of the "evolutionary progression scale" for the organisms at present.

Clustering methods

We use two unsupervised methods to observe the clustering patterns based solely on whole-proteome sequences: Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor embedding (t-SNE). Both are dimensional reduction methods, but with strength and weakness for our purposes, which help to visualize any clustering pattern in the data distribution. Both are based only on the evolutionary distances, estimated by the divergence of whole-proteome sequences among all pairs of the study arthropods. In PCA, the distances within a cluster as well as between two clusters are quantitative, thus, two close clusters nearby may not resolve well. In t-SNE, which applies Machine Learning to emphasize the resolution of nearby clusters, but the inter-cluster distances are de-emphasized, thus, not quantitative.

DECLARATIONS

Acknowledgement

We gratefully acknowledge the comments and advices from the arthropod experts at University of California, Berkeley, CA., Profs. Kipling Will and Peter Oboyski, and the constructive suggestions on our application of Information Theory and Unsupervised Clustering methods from Prof. Se-Ran Jun of College of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR, and Dr. Chao Zhang, Chief Scientific Officer, of Plexxikon, Inc. All the silhouettes are generated by Rigel Sison.

Funding

This research was partly supported by a grant (to SHK) from World Class University Project, Ministry of Education, Science and Technology, Republic of Korea and a gift grant to University of California, Berkeley, CA. (to support JJC). SHK acknowledges having appointments as Visiting Professorships at Yonsei University, Korea Advanced Institute of Science and Technology, and Incheon National University in South Korea during the manuscript preparation.

Competing financial interests

Authors declare that there are no competing financial interests in connection with this paper.

Author contribution

Conceptual design of the study and speculative interpretations and implications of the results by SHK; filtering and curation of genomic and proteomic sequence data from NCBI database, computational-algorithm design, programming and execution by JJC and BJK; interpretation of computational results by SHK, JJC and BJK; unsupervised clustering by various algorithms were performed by BJK; manuscript preparation by SHK with extensive discussions with JJC and BJK; all figures are designed by SHK, JJC and BJK.

COMPUTER CODE AVAILABILITY

The FFP programs for this study (2v.3.0) written in GCC(g++) is available in Github: https://github.com/jaejinchoi/FFP.

REFERENCES

1. Pace NR. Mapping the Tree of Life: Progress and Prospects. Microbiol Mol Biol Rev. 2009;73:565-576.
2. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol. 2017;18:1-17.
3. Blaisdell BE. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci .1986;83:5155-51599.
4. Lin J. Divergence measures based on the Shannon entropy. IEEE Trans Inf Theory. 1991;37:145-151.
5. Deerwester S, Dumais ST, Furnas GW, Landauer TK , Indexing by Latent Semantic Analysis. J Am Soc Inf Sci. 1990; 41: 391-407
6. Sims GE, Jun S, Wu GA, Kim S. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc Natl Acad Sci. 2009;106:2677-2682.
7. Jun SR, Sims GE, Wu GA, Kim SH. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. Proc Natl Acad Sci. 2010;107:133-138.
8. Choi JJ, Kim SH. A genome Tree of Life for the Fungi kingdom. Proc Natl Acad Sci. 2017;114:201711939.
9. Choi JJ, Kim S-H. Whole-proteome tree of life suggests a deep burst of organism diversity. Proc Natl Acad Sci. 2019;117:756155.
10. OLeary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:733-745.
11. Misof B, Liu S, Meusemann K, Peters RSR, Donath A, Mayer C, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science. 2014;346:763-768.
12. Thomas GWC, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K, et al. Gene content evolution in the arthropods. Genome Biol. 2020;21:15.
13. R Core Team. R: A language and environment for statistical computing. 2016.
14. Van Der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579-2605.
15. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4:406-425.
16. Gascuel O. BION J: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 1997;14:685-695.
17. Knuth DE. Semi-numerical algorithms, in The art of computer programming. 3rd edition. Boston: Addison-Wesley; 1973.
18. Fisher RA, Yates F. Statistical tables for biological, agricultural and medical research. London: Oliver and Boyd; 1948;27:378.
19. Shannon, C. A Mathematical Theory of Communication. Bell Syst Tech J. 1948;27:379-423.
20. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci.1981;53:131–147.
21. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47: 256-259.