2012 23rd International Workshop on Database and Expert Sytems Applications

Whole-Genome Phylogeny by virtue of Unic Subwords

Matteo Comin, Davide Verzotto
Department of Information Engineering
University of Padova
Padova, Italy
Email: comin,verzotto@dei.unipd.it

***Abstract*—With the progress of modern sequencing technologies a number of complete genomes is now available. Traditional motif discovery tools cannot handle this massive amount of data, therefore the comparison of complete genomes can be carried out only with ad hoc methods. In this work we propose a distance function based on subword compositions, which extends the Average Common Subword approach (ACS) of Ulitsky *et al.* [15]. ACS is closely related to the cross entropy estimated between two entire genome sequences, and thus to some set of "independent" subwords, namely the irredundant common subwords. Then, we filter the irredundant common subwords by means of underlying-paired motifs, which relate to each other regions of two genome sequences. This set of motifs is, by construction, linear in the size of input and without overlap; we call the selected motifs, underlying-paired irredundant common subwords, or simply *unic subwords*. Preliminary results show the validity of our method, and suggest novel computational approaches for analyzing the evolution of genomes.**

I. Whole-Genome Sequence Analysis: Background

The global spread of low-cost high-throughput sequencing technologies has made publicly available a number of complete genomes, and this number is still growing quite rapidly day by day. In contrast, only few computational methods can really handle as input entire chromosomes, or entire genomes. Similarly, the global alignment of large genomes has become a prohibitive task even for supercomputers, hence simply infeasible. To overcome this recent obstacle, in the last ten years a variety of alignment-free methods have been proposed. In principle they are all based on counting procedures that characterize a sequence based on its constituents, e.g., k-mers [3], [11].

For example, Sims *et al.* recently applied the Feature Frequency Profiles method (FFP) presented in [11] to compute a whole-genome phylogeny of mammals [10] —i.e., large eukaryotic genomes, including the human genome— and of bacteria.

In brief, in this k-mer based approach, they first estimate the parameter k in order to compute a feature vector for each sequence; this vector is composed by the frequency of each possible k-mer. Each feature vector is then normalized by the total number of k-mers found (i.e., by the sequence length), obtaining a probability distribution vector, or feature frequency profile, for each genome. FFP finally computes the distance matrix between all pairs of genomes by applying the Jensen-Shannon divergence to their frequency profiles. For completeness, we notice that, in large eukaryotes, they filter out high-frequency and low-complexity features among all the k-mers found.

When comparing genomes it is well known that different evolutionary mechanisms can take place. In this framework, two closely related species are expected to share larger portions of DNA than two distant ones, whereby also other large complements and reverse-complements, or inversions, may occur [13]. In this work we will take into account all these symmetries, in order to define a measure of comparison between genomes.

In this sense, an important fact is that most methods in the literature use only a portion of complete genomes [15]. For instance, there are approaches that use only the genic regions [3], [14] or the mitochondria [8]; in other cases, methods filter out regions that are highly repetitive or with low complexity, as for [11]. Recently, it has been shown that the evolutionary information is also carried by the non-genic regions [10]. For several families of viruses, we are not even able to estimate a complete phylogeny by analyzing their genes, since these organisms may share a very limited genetic material [15].

A. Average Common Subword Approach

Among the many distance measures proposed in the literature, which in most cases are dealing with k-mers, as seen above, an effective and particularly elegant method is the Average Common Subword approach (ACS), introduced by Ulitsky *et al.* [15]. In short, given two sequences s_1 and s_2, where s_1 is the reference sequence, it counts the length $l[i]$ of the longest subword starting at position i of s_1 that is also a subword of s_2, for every possible position i of s_1 (see Table I). This count is then averaged over the length of s_1. The general form of ACS follows:

$$ACS(s_1, s_2) = \frac{\sum_{i=1}^{|s_1|} l[i]}{|s_1|}.$$

The ACS measure is intrinsically asymmetric, but with simple operations can be reduced to a distance-like measure. We can notice the similarity with the cross entropy

1529-4188/12 $26.00 © 2012 IEEE
DOI 10.1109/DEXA.2012.10

of two probability distributions P and Q: $H(P,Q) = -\sum_x p(x)\log q(x)$, where $p(x)\log q(x)$ measures the number of bits needed to code an event x from P if a different coding scheme based on Q is used, averaged over all the possible events x.

$s_1[i]$	A	C	A	C	G	T	A	C
$l_1[i]$	2	1	4	3	3	3	2	1

$s_2[j]$	T	A	C	G	T	G	T	A
$l_2[j]$	3	4	3	2	1	3	2	1

Table I
EXAMPLE OF COUNTERS $l[i]$ FOR THE ACS APPROACH. COUNTERS $l_1[i]$ AND $l_2[j]$ FOR THE COMPUTATION OF $ACS(s_1, s_2)$ AND $ACS(s_2, s_1)$, RESPECTIVELY, WHERE $s_1 =$ ACACGTAC, $s_2 =$ TACGTGTA, AND $i, j = 1, \ldots, 8$.

The beauty of the ACS measure is that it is not based on fixed length subwords, but it can capture also variable length matches, in contrast to most methods that are based on fixed sets of k-mers. In fact, with the latter the choice of the parameter k is critical, and every method needs to estimate k from the data under examination, typically using empirical measurements [11]. From the theoretical prospective Burstein *et al.* [15] showed that the ACS approach mimics the cross entropy estimated between two large sequences generated by a finite-state Markov process.

ACS proved to be useful for reconstructing whole-genome phylogenies of viruses, bacteria, and eukaryotes, outperforming in most cases the state-of-the-art methods [15]. Here we aim to characterize and improve the ACS method, filtering out motifs that might be not useful for a whole-genome phylogeny of different organisms. In particular, we want to discard common motifs occurring in regions covered by other more significant motifs.

II. MATERIALS AND METHODS

In this section we propose a distance measure between entire genomes based on UNderlying-paired Irredundant Common subwords, or *unic subwords* (pronounced as "unique subwords") .

A. Irredundant Common Subwords

In the literature, the values $l[i]$ captured by the ACS approach are called the *matching statistics*, as described in detail in Gusfield *et al.* [7]. Here we aim to characterize the matching statistics with associated motifs, in order to identify which motifs are essential for the ACS measure.

The notion of irredundancy has been introduced in [2] and later modified for the problem of protein comparison [4], [5]. In this paper we consider this version, also called irredundant common motifs, but we restricted the domain only to subwords (i.e., without mismatches/don't cares). This ensures that there exists a close correspondence between the irredundant common subwords and the matching statistics.

Definition 1: (Irredundant/Redundant common subword) A common subword w is *irredundant* if and only if at least an occurrence of w in s_1 or s_2 is not covered by other common subwords. A common subword that does not satisfy this condition is called a *redundant common subword*.

As in the case with don't cares, we note that every irredundant common subword w is the result of some intersection of the two entire sequences, where each meet, in this case, corresponds to a particular a set of subwords. We further observe that the set of all irredundant common subwords $\mathcal{I}_{s_1,s_2}$ is a subset of the well-known linear set of *maximal common subwords*, defined as the common subwords for which the list of occurrences cannot be deduced by the list of a longer subword, possibly adding an offset d (see [1] for a more complete treatment of this topic). Therefore, the number of irredundant common subwords is bounded by $m + n$, where $|s_1| = n$ and $|s_2| = m$.

In summary, the notion of irredundant common subwords is useful to decompose the information given by ACS into several patterns, and then perform an additional filtering on the most representative common motifs for each region of the sequences s_1 and s_2.

B. Unic Subwords

When comparing entire genomes we want to avoid that large non-coding regions, which by nature tend to be highly repetitive, may overcount the same subwords a multiple number of times, misleading the final similarity score. In fact, while analyzing massive genomes, the number of repeated motifs is very high, particularly in the non-genic regions. For instance, in our experiments the number of irredundant common subwords can easily reach $2(m + n)/\log_4(m + n)$ elements in many pairwise comparisons, where m and n are the lengths of s_1 and s_2; and a very large number of overlaps between these subwords is present. Therefore we need to filter out part of this information, and select for each region of the sequences the "best" common subword by some measure.

In this regard, we must recall the definition of motif priority and of underlying motif, adapted from [6] to the case of pairwise sequence comparison. We will take as input the irredundant common subwords and the underlying quorum $u = 2$. Let now w and w' be two distinct subwords. We say that w has priority over w', or $w \to w'$, if and only if either $|w| \geq |w'|$, or $|w| = |w'|$ and w rank lower than w' in the lexicographic order. In this case, every subword can be defined just by its length and one of its starting positions in the sequences, meaning that any set of subwords is totally ordered with respect to the priority rule. Moreover, we say that an occurrence l of w is *tied* to an occurrence l' of a subword w', if $(E_{l,k} \cap E_{l',k'}) \neq \emptyset$ and $w' \to w$, where k and k' are, respectively, the lengths of w and w'. Otherwise, we say that l is *untied* from l'. Now, let $s = s_1 s_2$ be the string obtained through the concatenation of s_1 with s_2, and

let $\mathcal{I}_{s_1,s_2}$ be the set of irredundant common subwords that lie on s.

Definition 2: (Underlying-paired representative set, Unic subword) A set of subwords $\mathcal{U}_{s_1,s_2} \subseteq \mathcal{I}_{s_1,s_2}$ is said to be the *underlying-paired representative set* of s if and only if:

(i) every subword w in $\mathcal{U}_{s_1,s_2}$, called *unic subword*, has at least two occurrences, one in s_1 and the other in s_2, that are untied from all the untied occurrences of other subwords in $\mathcal{U}_{s_1,s_2} \setminus w$, and

(ii) there does not exist a subword $w \in \mathcal{I}_{s_1,s_2} \setminus \mathcal{U}_{s_1,s_2}$ such that w has at least two untied occurrences, one per sequence, from all the untied occurrences of subwords in $\mathcal{U}_{s_1,s_2}$.

As for the underlying motifs [6], it is easy to see that this set of unic subwords exists, and is unique for a concatenation s. A direct procedure to discover the whole set $\mathcal{U}_{s_1,s_2}$ can be obtained from the algorithm in [6]. As a corollary we know that the untied occurrences of the unic subwords can be mapped into the sequences s_1 and s_2 without overlaps in case of distinct subwords, resulting in a total length linear in the size of the sequences.

Thus, following the interesting experimental results obtained with the ACS approach, here we aim to select the irredundant common subwords that best fit each region of s_1 and s_2, employing a technique that we call *Unic Subword Approach* or, in short, USA. This technique is based on a simple pipeline. It first selects the irredundant common subwords and subsequently filters out the subwords that are not underlying motifs.

C. Efficient Computation of the Unic Subwords

Unlike the ACS method that can efficiently compute the matching statistics, the algorithm we will describe in the following requires a little more computation due to the filtering of the underlying-paired motifs. We first show how to compute the irredundant common subwords from the matching statistics, and then we present an approach for the selection of the unic subwords among these motifs by exploiting some algorithmic techniques.

1) Discovery of the Irredundant Common Subwords: One can use the matching statistics to compute the irredundant common subwords in a simple way, thus exploiting the fast algorithms proposed in [7], [15]. These algorithms use two different data structures, either the suffix tree or the suffix array, to find all possible right-maximal occurrences of common subwords between s_1 and s_2.

Here we use instead the generalized suffix tree. The first step consists in making a depth-first traversal of all nodes of T_{s_1,s_2}, and coloring each internal node with the colors of its leaves (each color corresponds to an input sequence). In this traversal, for each leaf i of T_{s_1,s_2}, we capture the closest ancestor of i having both the colors c_1 and c_2, say the node $\bar{w}$. Then, w is a common subword, and i is one of its right-maximal occurrences (in s_1 or in s_2); we select all subwords having at least one right-maximal occurrence. The resulting set of subwords, that is linear in the size of the sequences $O(m+n)$, represents a superset of the irredundant common subwords, since their right-maximal occurrences could be not left-maximal. The second part that keeps only the left-maximal occurrences is omitted due to space limitations.

2) Selection of the Unic Subwords: Once acquired the irredundant common subwords and their tree $T^{\mathcal{I}}_{s_1,s_2}$, composed by at most $m+n$ nodes, we filter out the subwords that are not underlying-paired for the case $s = s_1 s_2$, obtaining the set of unic subwords $\mathcal{U}_{s_1,s_2}$. As in [6], this process first requires to sort the subwords. Then, other two steps are required for each subword w: checking for the untied occurrences of w, and storing these occurrences.

In conclusion, our approach requires $O((m+n)\log\min\{m,n\}\log\log\min\{m,n\})$ time and $O(m+n)$ space to discover the set of all unic subwords $\mathcal{U}_{s_1,s_2}$ by employing a generalized suffix tree for s_1 and s_2.

D. A Distance-like Measure based on Unic Subwords

In the following we report the basic steps of our distance-like measure, similarly to ACS.

Let us assume that we have computed $\mathcal{U}_{s_1,s_2}$.For every subword $w \in \mathcal{U}_{s_1,s_2}$ of length k we sum up the score $h^{s_1}_w \sum_{i=1}^{k} i = h^{s_1}_w k(k+1)/2$ in $USA(s_1,s_2)$, where $h^{s_1}_w$ is the number of its untied occurrences in s_1 with respect to $\mathcal{U}_{s_1,s_2}$. Then, we average $USA(s_1,s_2)$ over the length of the first sequence, s_1, yielding

$$USA(s_1,s_2) = \frac{\sum_{w\in\mathcal{U}_{s_1,s_2}} h^{s_1}_w |w|(|w|+1)}{2|s_1|}.$$

Similarly to the method ACS we can compute a symmetrical distance-like measure $d_{USA}(s_1,s_2)$ between the sequences s_1 and s_2:

$$\overline{USA}(s_1,s_2) = \frac{\log_4(|s_2|)}{USA(s_1,s_2)} - \frac{2\log_4(|s_1|)}{(|s_1|+1)},$$

$$d_{USA}(s_1,s_2) = \frac{\overline{USA}(s_1,s_2) + \overline{USA}(s_2,s_1)}{2}.$$

We can easily see that the correction term rapidly converges to zero as $|s_1|$ increases; moreover we notice that $d_{USA}(s_1,s_2)$ grows as the two sequences s_1 and s_2 diverge. From now we will simply refer to the measure $d_{USA}(s_1,s_2)$ as the Unic Subword Approach measure, or USA.

III. Experimental Results

A. Genome Datasets and Reference Taxonomies

We assess the effectiveness of the Unic Subword Approach on the estimation of whole-genome phylogenies of different organisms. We test our distance function on three types of datasets that consider complete genomes among viruses, prokaryotes, and unicellular eukaryotes.

In the first dataset we selected 54 virus isolates of the 2009 human pandemic Influenza A – subtype H1N1, also called the "Swine Flu". The influenza A virion has eight segments of viral RNA with different functions. We concatenate these segments by means of a symbol not in Σ, e.g., '$' or 'N', according to their natural order. To compute a reference taxonomic tree, we perform an extensive multiple sequence alignment using the ClustalW2 tool as suggested by many scientific articles on the 2009 Swine Flu [12]. Then, we compute the tree using the DNAML tool from the PHYLIP software package[1], which implements the maximum likelihood method for DNA sequences.

In the second dataset we selected 18 prokaryotic organisms among the species used in [15] for a prokaryotic DNA genome phylogenomic inference. We chose the species whose complete genome has been sequenced and published, and whose phylogenetic tree structure can be inferred by well-established methods in the literature. The organisms come from both major prokaryotic domains: Bacteria, 10 organisms in total, and Archaea, 8 organisms in total. We compute their tree-of-life by using genes that code for the 16S ribosomal RNA, a small ribosomal subunit characterizing prokaryotes and widely used to reconstruct their phylogeny. Then we perform a maximum likelihood estimation on the aligned set of sequences, and use DNAML from PHYLIP in order to compute a reference tree based on the resulting estimations.

In the third dataset we selected 5 eukaryotic taxa of the protozoan genus *Plasmodium* whose genomes have been completely sequenced. Plasmodium are unicellular eukaryotic parasites best known as the etiological agents of malaria infectious disease. The sequences have lengths ranging from 18 Mbp to 24 Mbp, accounting for a total 106 Mbp. We used as reference tree the taxonomy computed by Martinsen *et al.* [9], as suggested by the Tree of Life Web Project (ToL).[2]

B. Whole-Genome Phylogeny Reconstruction

We exploited the above datasets to compare our method, the Unic Subword Approach (USA), with other efficient state-of-the-art approaches in the whole-genome phylogeny reconstruction challenge: ACS, FFP, and FFP$_{RY}$. The FFP$_{RY}$ method, instead of FFP, employs a reduced alphabet, the Purine-Pyrimidine alphabet (RY), which is composed by two character classes: [A,G] (both purine bases, denoted by R) [C,T] (both pyrimidines, denoted by Y). We did not perform tuning operations on the methods or a preliminary filtering of the sequences.

We reconstruct the phylogenomic trees from the distance matrices using the Neighbor-Joining algorithm (NJ) as implemented by the NEIGHBOR tool in the PHYLIP package.

We compute the symmetric difference of Robinson and Foulds (R-F) to compare the resulting topologies, assuming all edges of unit length, to the respective reference trees.

[1]PHYLIP (phylogenetic inference package) is a free computational phylogenetics software package available at http://evolution.genetics.washington.edu/phylip.

[2]The Tree of Life web project is hosted by the University of Arizona and available at http://www.tolweb.org.

Species	Group	USA	ACS	FFP	FFP$_{RY}$
Influenza A	Viruses	80/102	84/102	100/102	96/102
Archaea	Prokaryotes	4/10	4/10	6/10	6/10
Bacteria	Prokaryotes	6/14	10/14	6/14	10/14
Arch. & Bact.	Prokaryotes	20/30	22/30	20/30	22/30
Plasmodium	Eukaryotes	0/4	0/4	4/4	0/4

Table II
COMPARISON OF WHOLE-GENOME PHYLOGENY RECONSTRUCTIONS. NORMALIZED ROBINSON-FOULDS SCORES WITH THE CORRESPONDING REFERENCE TREE.

C. Performance Comparison and Statistics

Table II compares the phylogenomic reconstruction of our method with that of the other state-of-the-art approaches, by showing the R-F difference with respect to the reference taxonomy of each species. We ran FFP and FFP$_{RY}$ for different values of k (the fixed subword length) as suggested by [11], retaining the best results in agreement with the reference trees.

Our method, USA, achieves good performance in every test considering the R-F difference with the reference taxonomy, and very good performance if we further analyze the resulting phylogenies, as in Figure 1. We achieve in all cases at least the score of the best performing method, outperforming the other methods for sequences that share large parts, as in the case of viruses.

More in detail, Figure 1 shows that our approach can distinguish the two main clades of the 2009 Swine Flu (in green and red), that have been outlined in [12]. The origin of the flu could reside in the Mexican isolate of early April 2009 (Mexico/4108, in green), to which all other green isolates may ensue, from California/06 to the European isolates. Two sub-clades for the U.S. states of California and Texas are highlighted within the red clade, most probably corresponding to the first major evolutions of the viral disease.

Similar results are obtained for the second dataset. USA can easily distinguish the Archaea domain, from the Bacteria domain, and also other sub-clades with respect to the reference tree (figure not shown).

For the third dataset, the whole-genome phylogeny of the genus Plasmodium generated by USA (figure not shown) corresponds exactly to the taxonomy found in the literature.

In Table III we present some statistics for the unic subwords. We can see that only a few subwords are selected on average among the irredundant common subwords. Removing the high-frequency subwords (which were very few), we notice that the unic subwords typically have lengths

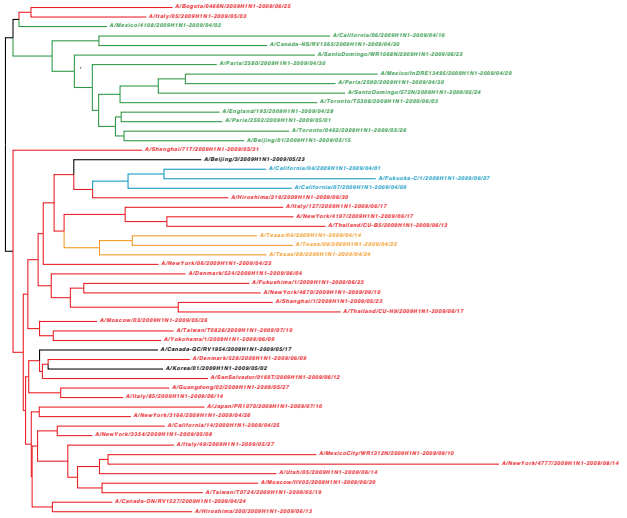

Figure 1. Whole-genome phylogeny of the 2009 world pandemic Influenza A (H1N1) generated by USA. In green and red we point out the two main clades, where the green Mexico/4108 is probably the closest isolate to the origin of the flu. In blue and orange are two of the possible early evolutions of the viral disease. In black, the organisms which in the literature do not fall into one of the two main clades.

$\geq log_4 m$, and in the case of viruses they can also be very large, capturing more information than FFP. Furthermore, each unic subword has attested on average only few occurrences per sequence, in general only one occurrence per sequence.

Count	Influenza A	Arch. & Bact.	Plasmodium
Min genome size	12,976 bp	650 kbp	18,524 kbp
Max genome size	13,611 bp	8,350 kbp	23,730 kbp
Average genome size	13,230 bp	2,700 kbp	21,380 kbp
Irredundants $\lvert\mathcal{I}_{s_1,s_2}\rvert$	3,722	3,167 k	16,354 k
Unic subwords $\lvert\mathcal{U}_{s_1,s_2}\rvert$	60	112 k	706 k
Min $\lvert w\rvert$ in $\mathcal{U}_{s_1,s_2}$	6	10	12
Max $\lvert w\rvert$ in $\mathcal{U}_{s_1,s_2}$	1,615	25	266
Average $\lvert w\rvert$ in $\mathcal{U}_{s_1,s_2}$	264	14	20
Untied inversions	28 %	31 %	33 %
Untied complements	22 %	20 %	19 %

Table III
MAIN STATISTICS FOR THE UNIC SUBWORD APPROACH AVERAGED OVER ALL EXPERIMENTS.

We can further analyze the average number of inversions and complements, where the increasing size of sequences seems to attest their values to 33 % and 19-20 %, respectively, out of the total number of unic subwords. However, this fact may be relegated to the nature of the sequences considered.

In conclusion we have shown that the unic subwords can be used for the reconstruction of phylogenetic trees. Preliminary experiments have shown very good performance in the identification of major clusters for viruses, prokaryotes, and unicellular eukaryotes. In the future we plan to investigate the presence of other genomic subtle signals among the unic subwords selected during the reconstruction.

REFERENCES

[1] Maximal words in sequence comparisons based on subword composition. A. Apostolico. *Algorithms and Applications*, T. Elomaa, H. Mannila, P. Orponen (Eds.), vol. 6060 of Lecture Notes in Computer Science. Springer, 2010, pp. 34–44.

[2] Incremental paradigms of motif discovery. A. Apostolico, L. Parida. *J. Comput. Biol.*, 2004, 11(1): 15–25.

[3] Genomic DNA k-mer spectra: models and modalities. B. Chor, D. Horn, N. Goldman, Y. Levy, T. Massingham. *Genome Biol.*, 2009, 10: R108.

[4] Classification of protein sequences by means of irredundant patterns. M. Comin, D. Verzotto. Proceedings of the 8th Asia-Pacific Bioinformatics Conference (APBC). *BMC Bioinformatics*, 2010, 11(Suppl.1): S16.

[5] The Irredundant Class method for remote homology detection of protein sequences. M. Comin, D. Verzotto. *J. Comput. Biol.*, 2011, 18(12): 1819–29.

[6] Comparing, ranking and filtering motifs with character classes: Application to biological sequences analysis. M. Comin, D. Verzotto. To appear in *Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data*, M. Elloumi, A.Y. Zomaya (Eds.). Wiley, 2012.

[7] *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.* D. Gusfield. Cambridge University Press, 1997.

[8] An information-based sequence distance and its application to whole mitochondrial genome phylogeny. M. Li, J.H. Badger, X. Chen, *et al. Bioinformatics*, 2001, 17(2): 149–54.

[9] A three-genome phylogeny of malaria parasites (*Plasmodium* and closely related genera): Evolution of life-history traits and host switches. E.S. Martinsen, S.L. Perkins, J.J. Schall. *Mol. Phylogenet. Evol.*, 2008, 47: 261–73.

[10] Whole-genome phylogeny of mammals: Evolutionary information in genic and nongenic regions. G.E. Sims, S.-R. Jun, G.A. Wu, S.-H. Kim. *PNAS*, 2009, 106(40): 17077–82.

[11] Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. G.E. Sims, S.-R. Jun, G.A. Wu, S.-H. Kim. *PNAS*, 2009, 106(8): 2677–82.

[12] Origins and evolutionary genomics of the 2009 swine-origin H1N1 Influenza A epidemic. G.J.D. Smith, D. Vijaykrishna, J. Bahl, *et al. Nature*, 2009, 459(7250): 1122–25.

[13] Inverse symmetry in complete genomes and whole-genome inverse duplication. S.-G. Kong, W.-L. Fan, H.-D. Chen, *et al. PLoS ONE*, 2009, 4(11): e7553.

[14] Proper distance metrics for phylogenetic analysis using complete genomes without sequence alignment. Z.-G. Yu, X.-W. Zhan, G.-S. Han, *et al. Int. J. Mol. Sci.*, 2010, 11(3): 1141–54.

[15] The average common substring approach to phylogenomic reconstruction. I. Ulitsky, D. Burstein, T. Tuller, B. Chor. *J. Comput. Biol.*, 2006, 13(2): 336–50.