Towards Phylogenomic Reconstruction

Le Sy Vinh[1] & Andrés Varón[1,2] & Daniel Janies[3] & Ward C. Wheeler[1]

[1] *Division of Invertebrate Zoology, American Museum of Natural History, New York*
[2] *Computer Science Department, The City University of New York*
[3] *Department of Biomedical Informatics, Ohio State University*

Corresponding author:
Le Sy Vinh
Division of Invertebrate Zoology, American Museum of Natural History, Central Park West at 79th Street, 10024 New York, New York
Voice: 212-769-7582
Fax: 212-769-5277
Email: vle@amnh.org

Conference: Contributed to the 2007 International Conference on Bioinformatics & Computational Biology

Keyword: phylogenetic reconstruction, genome comparison, chromosome comparison, direct optimization

Abstract

Reconstructing phylogenies is one of the primary objectives in evolution studies. Efficient software to reconstruct phylogenies based on isolated genes has existed for decades, yet, phylogenetic reconstructions from whole genomes are only beginning. The diversification of genome sequencing projects has generated thousands of whole genomes making phylogenomic reconstruction a challenging research topic. In this paper, we present an approach for pairwise alignment construction which deploys both nucleotide and locus (a segment of nucleotides) operations to minimize the total edit cost between genomes. The cost is composed of three factors: nucleotide transformation costs between loci, indel costs of loci, and rearrangement costs between locus orders. This approach is embedded within a direct optimization scheme to reconstruct phylogenies from whole unaligned genomes. Performance of this approach is demonstrated in our software, POY4, to reconstruct phylogenies from Coronavirus and Poxvirus genomes.

1 Introduction

Understanding evolutionary relationships among species is one of the central objectives in biology. The evolutionary relationships of species can be presented by a tree (phylogeny) in which leaves represent observed taxa, internal nodes represent inferred ancestors. The rapid development of efficient sequencing technologies has resulted in a huge amount of genetic data. These data enable researchers to study evolution at the molecular level. To date, phylogenies are typically reconstructed based on isolated genes [1, and references therein]. A growing number of available genome sequences leads us to develop approaches to comparative analysis of whole genomes. Evolutionary change of genomes is complex and subject to both small and large scale variations. The small scale processes act at nucleotide level, *i.e,* nucleotide substitution and indel. The large scale processes act at locus level, *i.e.,* locus rearrangement, locus indel, and horizontal gene transfer. Figure 1 shows an example of locus rearrangement operations in genomes.

The typical approach to compare the evolutionary relationships among species is to analyze nucleotide transformations among isolated homologous sequences. Besides, locus orders of genomes also reveal phylogenetic signals, hence, they can be used for phylogeny reconstruction [2, and references therein]. However, the locus order-based approach is unlikely applicable to closely related genomes because their locus orders are usually identical. Moreover, current implementations that use locus order to infer phylogeny ignore all nucleotide transformations which might contain orders of magnitude more information [2, 3, 4]. Thus, combining phylogenetic signals from both nucleotide transformation and locus order is useful to understand the evolutionary relationships among species.

Approaches to align genomes have been developed previously [5, 6, 7, 8, 9]. The first class of approaches considers genomes merely as long sequences. Sequence partitioning strategies are often used to overcome limits of time and memory needed to align long sequences. However, this class is not applicable when locus rearrangements have occurred in genomes. The second class concentrates on detecting conserved areas among genomes, which are subsequently aligned using traditional sequence alignment techniques. Although this class allows locus rearrangement operations, it suffers from two principle shortcomings: large sequence areas might be excluded from alignments (especially for distantly related genomes) and one locus might be aligned with several loci. To our knowledge, there does not exist any approach to align two genomes such that: (1) all loci are determined automatically, (2) each locus is either aligned with only one hypothetically homologous locus or considered as a locus indel, (3) loci are allowed to rearrange, (4) minimizing the total cost to transform one genome into another genome comprising nucleotide transformation costs, locus indel costs, and locus order rearrangement costs. We call this alignment a *comprehensive genome pairwise alignment.*

To comprehend the evolutionary changes in genomes, we must simultaneously analyze multiple genomes linked by phylogeny. ‘Direct optimization’ (DO) simultaneously evaluates sequence homologies and tree topologies to reconstruct phylogenies from unaligned sequences [10]. A core task in DO is the determination of hypothetical ancestor sequences according to their descendent’s sequences. In other words, the task is the construction of the pairwise alignment between two sequences. For isolated sequences, this problem has been investigated [11, and references therein]. We are working on genome pairwise alignment at various levels of generality and difficulty. The first level considers each genome as a long annotated sequence (each genome is pre-divided into segments using annotations as boundaries) [4, 12]. Alignments between two annotated genomes can be constructed using pairwise alignment while allowing rearrangements in gene order [12]. However, this approach is not applicable when annotations are not well defined.

Original order of loci	g_1	g_2	g_3	g_4	g_5
Loci from position 3 to 5 are inverted	g_1	g_2	$-g_5$	$-g_4$	$-g_3$
Loci from position 3 to 5 are moved to position 1	g_3	g_4	g_5	g_1	g_2
Loci from position 3 to 5 are moved to position 1, then inverted	$-g_5$	$-g_4$	$-g_3$	g_1	g_2

Figure 1: Three types of locus rearrangements on a genome

To relax the assumptions required by the annotation approach, the second level of our approach considers each genome as a single chromosome-genome (SC-genome). This type of data covers a wide range of genomes, *e.g,* viral genomes. In this paper, we propose an algorithm for the construction of a *comprehensive SC-genome pairwise alignment.* In the third level, each genome is represented as a set of chromosomes in which locus operations inside chromosomes and between chromosomes are allowed. The proposed algorithm for comprehensive SC-genome pairwise alignment can be extended to reconstruct a comprehensive genome pairwise alignment.

The rest of paper is organized as follows: section 2 presents an overview of direct optimization scheme to reconstruct phylogenies from unaligned sequences. In section 3, we describe an approach to reconstruct the comprehensive SC-genome pairwise alignment which is subsequently embedded within the direct optimization scheme to reconstruct phylogenies from unaligned genomes. Experiments on viral genome data sets to demonstrate the ability of our approach are presented in section 4. Discussions and open problems are addressed in the last section.

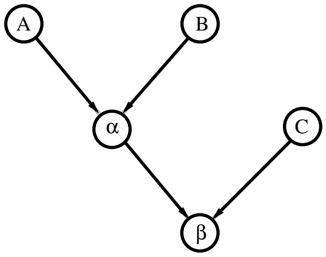

Figure 2: Topology and ancestor sequences are determined to minimize the total number of evolutionary events

2 Direct optimization

Direction optimization (DO) is a heuristic algorithm to the general NP-hard tree alignment optimization problem [13, 10]. DO searches tree topologies combined with reconstruction of ancestral sequences nodes in search of a global optimum of the the total number of evolutionary events such as mutations and insertion deletions but can also include locus level events. The key procedure is the determination of hypothetical ancestor sequences based on observed sequences following the post order travel. For example, ancestor sequences in Figure 2 is determined as follows. The pairwise alignment between A and B is constructed. As a result, the median sequence of A and B is inferred from the pairwise alignment and assigned to ancestor α. Following this, the median sequence of α and C is determined and assigned to ancestor β. The tree cost is calculated as the total number of evolutionary events over all branches. The phylogeny with minimum cost is considered as best. Since number of tree topologies increases combinatorially with number of sequences, heuristics must be applied to search for the best phylogeny in acceptable time [14].

3 Comprehensive SC-genome pairwise alignment

Here, we present an approach for construction of a comprehensive SC-genome pairwise alignment. To this end, first an algorithm is developed to detect conserved areas (blocks) between two genomes. It is not a novel algorithm but rather a consolidation of existing techniques to handle genomes from different levels of diversity [15, 6, 16, 9]. Conserved areas serve as anchors to divide each genome into a sequence of separated loci. Loci of the same block are initially considered homologous. To create the comprehensive SC-genome pairwise alignment, homologies (or indels) of other loci are determined by reconstructing pairwise alignment between two locus sequences when locus rearrangements are allowed.

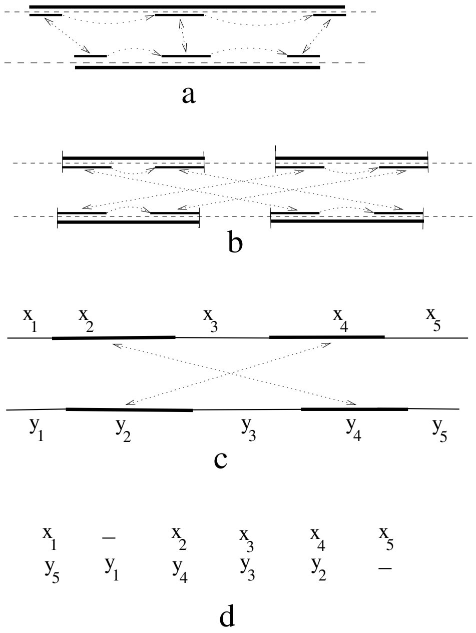
a
b
x_1 x_2 x_3 x_4 x_5
y_1 y_2 y_3 y_4 y_5
c
x_1 – x_2 x_3 x_4 x_5
y_5 y_1 y_4 y_3 y_2 –
d

Figure 3: a: non-rearranged seeds constitute a block, b: consecutive blocks are connected into a large block, c: blocks are used as anchors to divide genomes into loci, d: loci are aligned allowing order rearrangements

3.1 Detecting conserved areas

To detect conserved areas between two SC-genomes X and Y, three concepts, namely *segment, seed*, and *block*, are introduced to build up the algorithm progressively. To simplify the description of algorithm, genome orientations are not considered. However, this algorithm can be extended to integrate genome orientations into real implementations.

Let (s^X, e^X) denote a segment on X starting from nucleotide s^X and ending at nucleotide e^X ($s^X \leq e^X$). Intuitively, we say segment (s_1^X, e_1^X) is before segment (s_2^X, e_2^X) on X, denoted $(s_1^X, e_1^X) < (s_2^X, e_2^X)$, if $e_1^X < s_2^X$. The distance between two segments is $d_{12}^X = s_2^X - e_1^X$.

A pair of two identical segments (s^X, e^X) and (s^Y, e^Y) is called a *seed* between X and Y, denoted $S = (s^X, e^X, s^Y, e^Y)$. Seeds expose signals of conserved areas between X and Y. To eliminate spurious signals, seeds whose lengths are smaller than a predefined seed length threshold $\mathbf{l_s}$ are excluded from analysis ($\mathbf{l_s = 9}$ as default). All seeds between X and Y can be quickly identified using a suffix tree-based algorithm [17]. The number of seeds depends on two factors: the divergence between X and Y, and the seed length threshold $\mathbf{l_b}$. The score of a l_s long seed S is assigned by $(l_s \times c)$ where c is a predefined parameter ($c = 100$ as default). Obviously, a better seed is longer and have a higher score.

Consider two seeds $S_1 = (s_1^X, e_1^X, s_1^Y, e_1^Y)$ and $S_2 = (s_2^X, e_2^X, s_2^Y, e_2^Y)$, seed S_1 is before seed S_2 if segments of S_1 are before segments of S_2 on both X and Y. Distance d_{12} and shift g_{12} between S_1 and S_2 ($S_1 < S_2$) is measured as $d_{12} = \max\{d_{12}^X, d_{12}^Y\}$ and $g_{12} = |d_{12}^X - d_{12}^Y|$, respectively. The shift g_{12} indicates the minimum number of nucleotide indels needed to align two segments in between two seeds S_1 and S_2. The connecting score of S_1 and S_2 can be estimated approximately as $(o + (g_{12} - 1) \times e)$ where o and e are two predefined parameters ($o = -200, e = -10$ as default). The default values of c, o and e parameters are determined by experiments such that this score system can be used as a good optimal criterion for connecting seeds into larger areas.

Two seeds S_1 and S_2 are said to be *non-rearranged*, denoted $S_1 \leftrightarrow S_2$, if their distance d_{12} is not greater than a predefined non-rearranged threshold r. In other words, it is unlikely that rearrangement operations can be occurred in between non-rearranged seeds if they are connected. Experiments on real data shows that conserved areas between two moderately related genomes consist of about two seeds per thousand nucleotides. According to the observation, the default of r is 1000.

Experiments also show that seeds of over fifty nucleotides appears rarely, even between closely related genomes. That means individual seeds must be connected to construct larger conserved areas (blocks). Precisely, a list of seeds $(S_1, S_2, \ldots, S_n)$ can be connected into a block b if $S_i < S_{(i+1)}$ and $S_i \leftrightarrow S_{(i+1)}$ for $i = 1 \ldots (n-1)$ as illustrated in Figure 3a. Consider two block b_1 and b_2, block b_1 is before block b_2, denoted $b_1 < b_2$, if seeds of b_1 are before seeds of b_2. The block score is calculated as the sum of seed scores and connecting seed scores.

Due to the ordered property of seeds, the maximum score block can be constructed using dynamic programming. We are now ready to present an algorithm to detect conserved areas.

Detecting conserved areas algorithm:

1. Find the seed list L between X and Y. Set block list $B \leftarrow \emptyset$.

2. Find the maximum score block b. Add b into block list B. Remove seeds in block b from L. If L is not empty, repeat step 2.

3. Remove blocks in B whose lengths are smaller than a significant block length threshold $\mathbf{l_b}$ ($\mathbf{l_b = 100}$ as default). This guarantees that remained blocks indicate strong and reliable signals of conserved areas.

4. Resolve overlapped blocks. Specifically, if two blocks in B are overlapped, the smaller score block is removed.

5. Connect consecutive blocks in B to create larger blocks as illustrated in Figure 3b (blocks b_1 and b_2 are consecutive if $b_1 < b_2$ and $\nexists b \in B$ in between).

6. Output blocks in B as hypothetically conserved areas between X and Y.

3.2 Reconstructing pairwise alignment with rearrangements

Experiments with a large range of real data show that conserved areas often cover only a part of whole genomes. That means homologies (or indels) of other areas must be determined in order to gather additional phylogenetic signals. To this end, conserved blocks are deployed as anchors to partition X and Y into separated loci as demonstrated in Figure 3c. Thereafter, X and Y are represented as two sequences of loci $X = (x_1, x_2, \ldots, x_p)$ and $Y = (y_1, y_2, \ldots, y_q)$, respectively. Let λ denote a locus indel.

Here, we briefly define the pairwise alignment with rearrangements (PAR) between X and Y (the problem is fully described in [12]). Let $C(x_i, y_j) \in R^+$ be the edit cost to transform locus x_i into locus y_j for $i = 1 \ldots p,\ j = 1 \ldots q$. Note that $C(x_i, \lambda)$ and $C(\lambda, y_j)$ are respective indel costs of x_i and y_j. Since x_i and y_j are segments of nucleotides, the edit cost $C(x_i, y_j)$ can be calculated as the minimum number of nucleotide transformations between x_i and y_j. Consider two loci x_i and y_j of the same conserved block, they must be aligned together. To guarantee that, all edit costs $C(x_i, y'_j)$ and $C(x'_i, y_j)$ are set to infinity, $i' \neq i, j' \neq j$.

We denote Y_r a permutation of Y, that is, Y_r consists of the same set of loci in Y but in different orders. Let $R(Y, Y_r)$ be the rearrangement distance function between Y and its permutation Y_r. Typically, $R(Y, Y_r)$ is computed as the breakpoint distance or inversion distance [18, 19]. The rearrangement cost between Y and Y_r is measured as $R(Y, Y_r) \times c_r$ where c_r is the rearrangement operation cost.

Given edit cost matrix C and rearrangement distance function R, we construct the pairwise alignment A_r allowing locus rearrangements such that minimizing the total cost which is composed of three factors: edit costs between loci, indel costs of loci and rearrangement costs of locus orders. Figure 3d illustrates the A_r between two sequences $X = (x_1, x_2, x_3, x_4, x_5)$ and $Y = (y_1, y_2, y_3, y_4, y_5)$. Note that the PAR problem does not require genomes to have the same number of loci. Since an exact solution for PAR problem is likely intractable, heuristic approaches which compromise between time complexity and alignment quality have been proposed [12]. In a nutshell, the alignment A_r presents the comprehensive pairwise alignment between two SC-genomes X and Y.

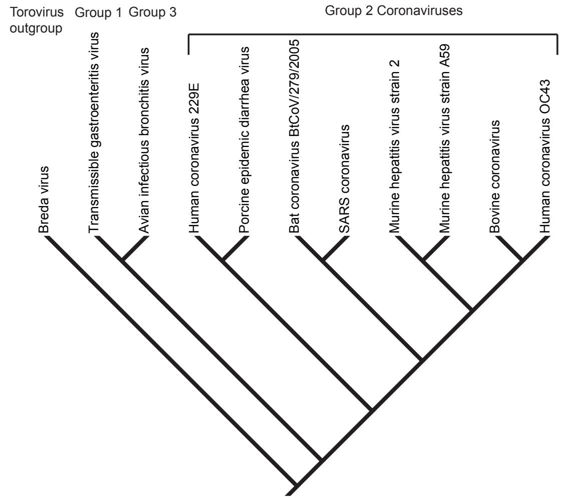
Torovirus outgroup
Group 1
Group 3
Group 2 Coronaviruses
Breda virus
Transmissible gastroenteritis virus
Avian infectious bronchitis virus
Human coronavirus 229E
Porcine epidemic diarrhea virus
Bat coronavirus BtCoV/279/2005
SARS coronavirus
Murine hepatitis virus strain 2
Murine hepatitis virus strain A59
Bovine coronavirus
Human coronavirus OC43

Figure 4: Genome-based phylogeny of 11 Coronaviruses where SARS-CoV shares a common ancestor with CoV from bats instead of small carnivors

4 Experiments

We are focusing on analyzing viral genomes to understand the origin and prediction of pathogenicity. To our knowledge, there does not exist any software to reconstruct phylogenies from unaligned genomes. To demonstrate the potential of our approach, we analyze Coronavirus genomes and Poxvirus genomes on a 3.2 GHz PC.

Severe Acute Respiratory Syndrome (SARS) is a novel human illness caused by a previously unrecognized Coronavirus, SARS CoV, of zoonotic origin [20]. Between November and August 2003, there were 8,422 cases and 916 deaths from SARS (WHO, 2003). Although more than 219 isolates of SARS CoV have been sequenced very little is known about genome diversity among the family Coronaviridae. Very few non-SARS Coronaviruses have been sequenced and the zoonotic potential of the Coronaviridae is not well understood [21]

There remain conflicting reports on the animal reservoir of SARS-CoV. Using small portions of the CoV genome, Guan et al. (2003) implicate small carnivores whereas Li et al. (2005) assert that bats are the animal reservoir of SARS-CoV [22]. Using the entire genome and a diverse sample of Coronaviruses, Janies et al. (2007) confirm that small carnivores are not the reservoir species and bats are the best candidates for the reservoir species [23].

The genome of Coronaviruses is comprised of a single-stranded, positive-sensed RNA molecule 27-31 kb in length [24]. To increase our understanding of Coronaviruses and their potential for reasssortment, we have examined a dataset of 11 Coronavirus genomes and a Torovirus outgroup. To search for the optimal phylogeny, five Wagner trees were built. Subsequently, the best of those was selected and improved further by the subtree pruning and re-grafting technique and finally considered as the optimal phylogeny (see Figure 4). The optimal tree found is in agreement with recently studies that SARS-CoV is related to group 2 Coronaviruses and shares a common ancestor with CoV from bats instead of small carnivores [22, 23]. Although rearrangements are not found among these 11 Coronaviruses, further experiments with more genomes must be investigated.

To examine our approach with larger genomes, we collected 13 Poxvirus genomes from NCBI (http://www.ncbi.nlm.nih.gov/). The genome sizes range from 184,900 to 228,250 bp. The pairwise alignment between two genomes is constructed in approximately ten seconds. Interestingly, conserved areas cover approximately 90% of the whole genomes. The average breakpoint distance between two genomes is four indicating a small number of locus rearrangements. To search for the optimal tree, five Wagner trees were built. Subsequently, the best of those was improved by nearest neighbor interchanges technique. Figure 5 presents the reconstructed phylogeny where five clades are constructed: Camelpox, Cowpox, Horsepox, Monkeypox and Variola. Diagnosing the tree shows 132 breakpoints. The rearrangements are not frequent between genomes inside the same group, *e.g.,* zero between Monkey_USA and Monkey_Sierra, two between Camelpox_M96 and Camelpox_CMS. However, a larger number of rearrangements occurs between ancestor sequences of different groups, *e.g.,* 24 breakpoints between Horsepox and Cowpox. The program took approximately ten hours indicating the potential of our approach for construction of phylogenies from large genomes and more taxa.

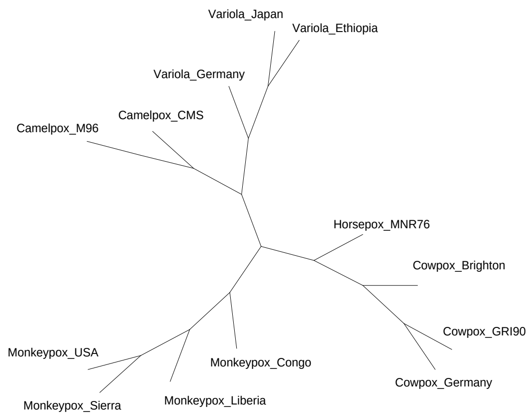

Figure 5: Genome-based phylogeny of 13 Poxviruses where five clades are reconstructed: Camelpox, Cowpox, Horsepox, Monkeypox and Variola.

5 Discussions

Reconstructing phylogenies from whole genomes requires a marriage between genome analysis techniques and phylogenetic reconstruction algorithms. We present an approach to align whole genomes which incorporates both nucleotide transformations and locus operations to minimize the cost to transform one genome into another genome. The approach is implemented in a direct optimization scheme to reconstruct phylogenies from unaligned genomes.

The performance of our approach is demonstrated on Coronavirus and Poxvirus genome data sets. The program reconstructs phylogenies in acceptable time. Default parameters are determined by experiments on real data rather than theoretical establishments. For example, the seed length threshold $\mathbf{l_b} = \mathbf{9}$ is experimentally considered as the best value for moderately related genomes. Although $\mathbf{l_b}$ can be adjusted, either largely increasing or decreasing $\mathbf{l_b}$ will decrease the efficiency of the approach.

Although the approach is described for single chromosome-genomes, we are working on an extension to cope with multiple chromosome-genomes in which loci operations within chromosomes and between chromosomes are allowed.

This approach can cope with locus indels and locus rearrangements, however, the evolution of genomes is certainly more complicated. Two other kinds of locus operations not considered are horizontal gene transfer and recombination [25, 26]. These processes typically result in additional homoplasy. Our ultimate goal is to design an approach incorporating all these operations to reconstruct phylogenies from genomes.

Acknowledgments

This paper is based upon work supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under grant “Novel analytical and empirical approaches to the origin and prediction of pathogenicity” (#W911NF-05-1-0271) and NSF-ITR grant “Information technology research: Phyloinformatics.”

References

[1] Felsenstein, J. *Infering Phylogenies*. Sinauer Associates, Sunderland, Massachusetts, 2004

[2] Sankoff, D., Nadeau, J. H. *Comparative Genome: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*. Kluwer, first edition, 2000

[3] Moret, B. M., Tangy, J., Wangz, L.-S., Tandy-Warnow. Steps toward accurate reconstructions of phylogenies from gene-order data. *J. Comput. Syst. Sci.*, 65:508–525, 2002

[4] Wheeler, W. C. Chromsomal character optimization. *Molecular Phylogenetics and Evolution*, 2007. In press

[5] Delcher, A. L., Kasif, S., Fleishman, R., Peterson, J., White, O., Salzberg, S. L. Alignment of whole genomes. *Nucleic Acids Research*, 27:2369–2376, 1999

[6] Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., DavidHaussler, Miller, W. Human mouse alignments with blastz. *Genome Research*, 13:103–107, 2003

[7] Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., Green, E. D., Sidow, A., Batzoglou, S. Lagan and multi-lagan: Efficient tools for large-scale multiple alignment of genomic dna. *Genome Research*, 13:721–731, 2003

[8] Brudno, M., Malde, S., Poliakov, A., Do, C. B., Couronne, O., Dubchak, I., Batzoglou, S. Glocal alignment: finding rearrangements during alignment. *Bioinformatics*, 19:i54–i62, 2003

[9] Darling, A. C., Mau, B., Blattner, F. R., Perna, N. T. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14:1394–1403, 2004

[10] Wheeler, W. C. Optimization alignment: The end of multiple sequence alignment in phylogenetics? *Cladistics*, 12(1):1–9, 1996

[11] Waterman, M. S. *Introduction to Computational Biology*. Chapman and Hall, London, UK, first crc press edition, 2000

[12] Vinh, L. S., Varon, A., Wheeler, W. C. Pairwise alignment with rearrangements. *Genome informatics*, 17(2):141–151, 2006

[13] Wang, L., Jiang, T. On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4):337–348, 1994

[14] Wheeler, W., Aagesen, L., Arango, C. P., Faivovich, J., Grant, T., Dhaese, C., Janies, D., Smith, W. L., Varon, A., Giribet, G. *Dynamic homology and phylogenetic systematics: a unified approach using poy*. American Museum of Natural History, New York, USA, 2006

[15] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990

[16] Brudno, M., Chapman, M., Gttgens, B., Batzoglou, S., Morgenstern, B. Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, 4:66, 2003

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

[18] Sankoff, D., Blanchette, M. Multiple genome rearrangement and breakpoint phylogeny. *J. Comput. Biol.*, 5:555–570, 1998

[19] Hannenhalli, Pevzner, P. A. Transforming cabbage into turnip. (polynominal algorithm for sorting signed permutations by reversals. *Proceedings of the 27th Annual ACM-SIAM Symposium on the Theory of Computing*, 178–189. 1995

[20] Ksiazek, et al. A novel coronavirus associated with severe acute respiratory syndrome. *The New England Journal of Medicine*, 348:1953–1966, 2003

[21] Narvas-Martin, SR., W. Sars: lessons learned from other coronaviruse. *Viral Immunolog*, 16:461–474, 2003

[22] Li, et al. Bats are natural reservoirs of sars-like coronaviruses. *Science*, 310:676–679, 2005

[23] Janies, D., Habib, F., Alexandrov, B., Pol, D. Evolution of genomes and host shifts among sars associated and related coronaviruses. *Cladistics*, 2007

[24] Lai, M. Coronavirus: organization, replication and expression of genome. *Ann. Rev. Microb.*, 44:303–333, 1990

[25] Maddison, W. P. Gene trees in species trees. *Syst. Biol.*, 46:523–536, 1997

[26] Posada, D., Crandall, K. A., Holmes, E. C. Recombination in evolutionary genomics. *Annual Review of Genetic*, 36:75–97, 2002