Approximability and Fixed-Parameter Tractability for the Exemplar Genomic Distance Problems *

Binhai Zhu

Department of Computer Science
Montana State University
Bozeman, MT 59717-3880
USA
Email: bhz@cs.montana.edu

Abstract. In this paper, we present a survey of the approximability and fixed-parameter tractability results for some Exemplar Genomic Distance problems. We mainly focus on three problems: the exemplar breakpoint distance problem and its complement (i.e., the exemplar non-breaking similarity or the exemplar adjacency number problem), and the maximal strip recovery (MSR) problem. The following results hold for the simplest case between only two genomes (genomic maps) $\mathcal{G}$ and $\mathcal{H}$, each containing only one sequence of genes (gene markers), possibly with repetitions.

1. For the general Exemplar Breakpoint Distance problem, it was shown that deciding if the optimal solution value of some given instance is zero is NP-hard. This implies that the problem does not admit any approximation, neither any FPT algorithm, unless P=NP. In fact, this result holds even when a gene appears in $\mathcal{G}$ ($\mathcal{H}$) at most two times.
2. For the Exemplar Non-breaking Similarity problem, it was shown that the problem is linearly reducible from Independent Set. Hence, it does not admit any factor-$O(n^{\epsilon})$ approximation unless P=NP and it is W[1]-complete (loosely speaking, there is no way to obtain an $O(n^{o(k)})$ time exact algorithm unless FPT=W[1], here k is the optimal solution value of the problem).
3. For the MSR problem, after quite a lot of struggle, we recently showed that the problem is NP-complete. On the other hand, the problem was previously known to have a factor-4 approximation and we showed recently that it admits a simple FPT algorithm which runs in $O(2^{3.61k}n + n^2)$ time, where k is the optimal solution value of the problem.

1 Introduction

In bioinformatics and computational biology, we constantly need to process various biological data to extract meaningful biological relation, like building a

* This research is partially supported by NSF, NSERC, Louisiana Board of Regents under contract number LEQSF(2004-07)-RD-A-35, and MSU-Bozeman's Short-Term Professional Development Leave Program.

phylogenetic tree. However, such a process usually involves solving hard combinatorial optimization problems which are typically NP-complete.

In the area of bioinformatics and computational biology, one would typically apply three methods to handle these NP-complete problems. One is to find an approximation solution, with the requirement being that the approximation factor is small (better close to one). The other is to look for an exact solution (FPT algorithm) when the solution size of the problem is small. The vast majority of practical solutions for bioinformatics and computational biology are heuristic ones, which are possibly based on some formal methods like integer linear programming, branch-and-bound, etc.

In this paper, we review the approximability and fixed-parameter tractability results for three problems related to exemplar genomic distance computation. In these problems, we are given some genomes or genomic maps and we try to optimize some solutions values by deleting some genes or gene markers. So these problem fit naturally for approximation and/or FPT solutions. Unfortunately, as we will review a bit later, some of these problems are very hard in both aspects. In other words, it might be impossible to design good approximation and/or FPT algorithms for them, unless P=NP or FPT=W[1]. On the other hand, many problems are still open along these lines.

The paper is organized as follows. In Section 2, we review the approximability and fixed-parameter tractability for the Exemplar Breakpoint Distance (EBD) problem. In Section 3, we review the approximability and fixed-parameter tractability for the Exemplar Non-breaking Similarity (ENbS) problem (which is the dual of EBD). In Section 4, we review the approximability and fixed-parameter tractability for the Maximal Strip Recovery (MSR) problem. In Section 5, we list a set of open problems to conclude this paper.

2 Approximability and Fixed-Parameter Tractability for EBD

In the genome comparison and rearrangement area, a standard problem is to compute the number (i.e., genetic distances) and the actual sequence of genetic operations needed to convert a source genome to a target genome. This problem is important in evolutionary molecular biology. Typical genetic distances include edit [23], signed reversal [26, 24, 6] and breakpoint [30], etc. (The idea of signed reversal and, implicitly, breakpoint, was initiated as early as in 1936 by Sturtevant and Dobzhansky [29].) In the past years, conserved interval distance was also proposed to measure the similarity of multiple sequences of genes [9]. Interested readers are referred to [21, 22] for a summary of the research performed in this area.

However, in genome rearrangement research, it is almost always assumed that each gene appears in a genome exactly once. Under this assumption, the genome rearrangement problem is in essence the problem of comparing and sorting signed permutations [21, 22]. However, this assumption is very restrictive and is only justified in several small virus genomes. For example, this assumption does not

hold on eukaryotic genomes where paralogous genes exist [25, 27]. On the one hand, it is important in practice to compute genomic distances, e.g., Hannenhalli and Pevzner's method [21], when no gene duplications arise; on the other hand, one might have to handle this gene duplication problem as well.

Sankoff first considered the problem of computing genomic distance with duplicated genes. About ten years ago, Sankoff proposed a way to select, from the duplicated copies of genes, the common ancestor gene such that the distance between the reduced genomes (*exemplar genomes*) is minimized [27]. A general branch-and-bound algorithm was also implemented in [27]. In [25], Nguyen, Tay and Zhang proposed to use a divide-and-conquer method to compute the exemplar breakpoint distance empirically.

For the theoretical part of research, it was shown that both of the problems of computing the signed reversal and breakpoint distances between exemplar genomes are NP-complete [7]. A few years ago, Blin and Rizzi further proved that computing the conserved interval distance between exemplar genomes is NP-complete [8]; moreover, it is NP-complete to compute the minimum conserved interval matching (i.e., without deleting the duplicated copies of genes). Recently we showed much stronger inapproximability result for the exemplar conserved interval distance problem (even under a weaker model of approximation) [12]. While various exemplar genomic distances have been researched before, in this survey we will focus on the exemplar breakpoint distance. In fact, all the inapproximability result for exemplar breakpoint distance under the normal model of approximation holds for any other genomic distance $d(-,-)$ satisfying $d(G,H)=0$ implies $G=H$ or $G=-H$.

2.1 Preliminaries

In the genome comparison and rearrangement problem, we are given a set of genomes, each of which is a signed sequence of genes. (In general a genome could contain a set of such sequences. The genomes we focus on are typically called *singletons*.) The order of the genes corresponds to the position of them on the linear chromosome and the signs correspond to which of the two DNA strands the genes are located. While most of the past research are under the assumption that each gene occurs in a genome once, this assumption is problematic in reality for eukaryotic genomes or the likes where duplications of genes exist [27]. Sankoff proposed a method to select an *exemplar genome*, by deleting redundant copies of a gene, such that in an exemplar genome any gene appears exactly once; moreover, the resulting exemplar genomes should have a property that certain genetic distance between them is minimized [27].

The following definitions are very much following those in [8]. Given n *gene families* (alphabet) $\mathcal{F}$, a genome $\mathcal{G}$ is a sequence of elements of $\mathcal{F}$ such that each element is with a sign ($+$ or $-$). In general, we allow the repetition of a gene family in any genome. Each occurrence of a gene family is called a *gene*, though we will not try to distinguish a gene and a gene family if the context is clear. Given a genome $G=g_1g_2...g_m$ with no repetition of any gene, we say that gene g_i *immediately precedes* g_j if $j=i+1$. Given genomes G,H, if gene

a immediately precedes b in G and neither a immediately precedes b nor $-b$ immediately precedes $-a$ in H, then they constitute a *breakpoint* in G. The *breakpoint distance* is the number of breakpoints in G (symmetrically, it is the number of breakpoints in H).

The number of a gene g appearing in a genome $\mathcal{G}$ is called the cardinality of g in $\mathcal{G}$, written as $card(g, \mathcal{G})$. A gene in $\mathcal{G}$ is called *trivial* if g has cardinality exactly 1; otherwise, it is called *non-trivial*. A genome $\mathcal{G}$ is called *r-repetitive*, if all the genes from the same gene family appear at most r times in $\mathcal{G}$. For example, $\mathcal{G} = c - adc - bdeb$ is 2-repetitive.

Given a genome $\mathcal{G}$ over $\mathcal{F}$, an *exemplar genome* of $\mathcal{G}$ is a genome G' obtained from $\mathcal{G}$ by deleting duplicating genes such that each gene family in $\mathcal{G}$ appears exactly once in G'. For example, let $\mathcal{G} = -bcaadag - e$, there are two exemplar genomes: $-bcadg - e$ and $-bcdag - e$.

The Exemplar Breakpoint Distance (EBD) problem is defined as follows:

Instance: Genomes $\mathcal{G}$ and $\mathcal{H}$, each is of length $O(m)$ and each covers n identical gene families (i.e., at least one gene from each of the n gene families appears in both $\mathcal{G}$ and $\mathcal{H}$); integer K.
Question: Are there two respective exemplar genomes of $\mathcal{G}$ and $\mathcal{H}$, G and H, such that the breakpoint distance between them is at most K?

In the next subsection, we present some hardness results on the approximability and fixed-parameter tractability for EBD, namely, the hardness to compute or approximate the minimum value K in the above formulation. Given a minimization (maximization) problem Π, let the optimal solution value of Π be OPT. We say that an approximation algorithm $\mathcal{A}$ provides a *performance guarantee* of α for Π if for every instance I of Π, the solution value returned by $\mathcal{A}$ is at most $\alpha \times \text{OPT}$ (at least OPT/α). Usually we say that $\mathcal{A}$ is a factor-α approximation for Π. For the obvious reason, we are only interested in polynomial time approximation algorithms. Readers are referred to [16, 19] for more details regarding the definitions related to approximation algorithms and NP-completeness.

As a well-known subject as well, an FPT algorithm for an optimization problem Π with optimal solution value $\text{OPT} = k$ is an algorithm which solves the problem in $O(f(k)n^c)$ time, where f is any function only on k and c is some fixed constant not related to k. More details on FPT algorithms can be found in [18].

2.2 Hardness Results

In [10], we presented the first set of inapproximability and approximation results for the Exemplar Breakpoint Distance problem, given two genomes each containing only one sequence of genes drawn from n identical gene families. We showed that even if a gene appears at most three times, deciding whether the optimal exemplar breakpoint distance is zero, i.e, whether $G = H$, is NP-complete. It was left as an open problem whether the result holds when each gene appears at most twice in each of the input genomes [10, 1]. This year, this open

question was finally answered, i.e., it remains NP-complete even when each gene appears at most two times [4]. Combining these results, we have the following inapproximability result.

Theorem 1. *If both $\mathcal{G}$ and $\mathcal{H}$ are 2-repetitive genomes, then the Exemplar Breakpoint Distance problem does not admit any polynomial time approximation (regardless of its approximation factor), unless P=NP.*

Proof. If we view the Exemplar Breakpoint Distance problem as a minimization problem, then the result in [4] implies that deciding whether OPT = 0 is NP-complete (even if the input genomes are 2-repetitive). Let $\mathcal{A}$ be any approximation algorithm for EBD with factor α. By definition, $\mathcal{A}$ returns an approximation solution value APP, with

$$\text{APP} \leq \alpha \times \text{OPT}.$$

When OPT = 0, clearly APP must also satisfy APP = 0. In other words, $\mathcal{A}$ would be able to solve the instance in [4] in polynomial time. This, however, contradicts with the corresponding NP-completeness result (unless P=NP). □

Regarding the fixed-parameter tractability for EBD, we have the following theorem.

Theorem 2. *If both $\mathcal{G}$ and $\mathcal{H}$ are 2-repetitive genomes, then the Exemplar Breakpoint Distance problem does not admit any FPT algorithm, unless P=NP.*

Proof. Again, if we view the Exemplar Breakpoint Distance problem as a minimization problem, then the result in [4] implies that deciding whether OPT = 0 is NP-complete (even if the input genomes are 2-repetitive). Let $\mathcal{B}$ be any FPT algorithm for EBD which runs in $O(f(k)n^c)$ time. When OPT = $k = 0$, $\mathcal{B}$ solves EBD in $O(f(0)n^c) = O(n^c)$ time. In other words, $\mathcal{B}$ would be able to solve the instance in [4] in polynomial time. This, again, contradicts with the corresponding NP-completeness result, unless P=NP. □

3 Approximability and Fixed-Parameter Tractability for ENbS

We comment that the negative results in Section 2.2 hold for any genomic distance $d(-,-)$ satisfying that $d(G,H) = 0$ implies $G = H$ or $G = -H$. This, of course, implies that all the exemplar genomic distance problems (like exemplar reversal, exemplar transposition, and exemplar conserved interval distances) do not admit any polynomial time approximation algorithms or any FPT algorithm, unless P=NP.

There have been two ways to handle this problem. One is to use a weak model of approximation, which will be covered as related to open problems in Section 5. The other, on the other hand, is to use a different similarity measure. In this case, one would try to maximize certain similarity measure. The most notably such measures include non-breaking similarity (or number of adjacencies) [13] and

the number of common intervals [3]. (A common interval is a pair of substrings appearing in the two genomes with the same genes, but possibly different orders. Example. $G = abced, H = deacb$. (abc, acb) is a length-3 common interval.) We will focus the non-breaking similarity, which is really the complement of the breakpoint distance.

For two exemplar genomes G and H over the same alphabet of size n, a breakpoint in G is a two-gene substring g_ig_{i+1} such that neither g_ig_{i+1} nor $-g_{i+1} - g_i$ is a substring in H. A *non-breaking point* (or an *adjacency*) is a common two-gene substring g_ig_{i+1} that appears either as g_ig_{i+1} or as $-g_{i+1} - g_i$ in G and H. The number of non-breaking points between G and H is also called the *non-breaking similarity* between G and H, denoted as $\text{nbs}(G, H)$. Clearly, we have $\text{nbs}(G, H) = n - 1 - \text{bd}(G, H)$. For two genomes $\mathcal{G}$ and $\mathcal{H}$, their *exemplar non-breaking similarity* $\text{enbs}(\mathcal{G}, \mathcal{H})$ is the maximum $\text{nbs}(G, H)$, where G and H are exemplar genomes derived from $\mathcal{G}$ and $\mathcal{H}$. Again we have $\text{enbs}(\mathcal{G}, \mathcal{H}) = n - 1 - \text{ebd}(\mathcal{G}, \mathcal{H})$.

The Exemplar Non-breaking Similarity (ENbS) problem is formally defined as follows:

Instance: Genomes $\mathcal{G}$ and $\mathcal{H}$, each is of length $O(m)$ and each covers n identical gene families (i.e., at least one gene from each of the n gene families appears in both $\mathcal{G}$ and $\mathcal{H}$); integer K.
Question: Are there two respective exemplar genomes of $\mathcal{G}$ and $\mathcal{H}$, G and H, such that the non-breaking similarity between them is at least K?

We have the following negative results which have been proved in [13].

Theorem 3. *If one of $\mathcal{G}$ and $\mathcal{H}$ is exemplar and the other is 2-repetitive, then the Exemplar Non-breaking Similarity problem does not admit any factor-n^ϵ polynomial time approximation unless P=NP.*

Proof. We give a sketch of proof from [13]. In [13], it was shown that Independent Set can be linearly reduced to ENbS; i.e., the input graph has an independent set of size k iff the constructed ENbS instance has a non-breaking similarity (or number of adjacencies) equal to k. As Independent Set cannot be approximated within a factor of n^ϵ unless P=NP [20], the theorem follows. □

Theorem 4. *If one of $\mathcal{G}$ and $\mathcal{H}$ is exemplar and the other is 2-repetitive, the Exemplar Non-breaking Similarity problem does not admit an FPT algorithm unless FPT=W[1].*

Proof. It is noted that the reduction from Independent Set to ENbS in [13] is in fact an FPT reduction. As Independent Set is W[1]-complete [18], the theorem simply follows. □

In fact, with the lower bound results proved in [15], Independent Set (hence ENbS) cannot be solved in $O(f(k)n^{o(k)})$ time even if k is bounded by an arbitrarily small function of n, unless ETH fails. (ETH — Exponential Time Hypothesis: 3SAT cannot be solved in subexponential time.)

4 Approximability and Fixed-Parameter Tractability for MSR

Given two genomic maps G and H represented by a sequence of n gene markers, a *strip* (syntenic block) is a sequence of distinct markers of length at least two which appear as subsequences in both of the input maps, either directly or in reversed and negated form. The problem *Maximal Strip Recovery* (MSR) is to find two subsequences G' and H' of G and H, respectively, such that the total length of disjoint strips in G' and H' is maximized An example is as follows: $G = abcde, H = cbdae$ and the optimal solution is $G' = H' = cde$.

The MSR problem was proposed to handle the elimination of noise and ambiguities in genomic maps. This is related to the well-known problem in comparative genomics — to decompose two given genomes into syntenic blocks, i.e., segments of chromosomes which are deemed to be homologous in the two input genomes. Two years ago, a heuristic method was proposed to handle the MSR problem [17, 32]. In [14], a factor-4 polynomial time approximation algorithm was proposed for the problem. This was done by applying the Maximum Weight Independent Set on 2-interval graphs, which admit a factor-4 approximation [5]. We also proved that several close variants of MSR, MSR-d (with $d > 2$ input maps), MSR-DU (with marker duplications), and MSR-WT (with markers weighted) are all NP-complete. It was left as an open problem whether the problem can be solved in polynomial time or is NP-complete [14].

Recently, in [31] we showed that MSR is in fact NP-complete, via a polynomial time reduction from One-in-Three 3SAT (which was shown to be NP-complete in [28, 19]). We summarize the results in [14, 31] as follows.

Theorem 5. *MSR is NP-complete, and it admits a factor-4 polynomial time approximation.*

As an effort to solve the MSR problem practically, we tried to solve MSR and its variants exactly with FPT algorithms, i.e., showing that MSR is fixed-parameter tractable [31]. Let k be the minimum number of markers deleted in various versions of MSR, the running time of our algorithms are $O(2^{3.61k}n + n^2)$ for MSR, $O(2^{3.61k}dn + dn^2)$ for MSR-d, and $O(2^{7.22k}n + n^2)$ for MSR-DU respectively. We summarize this result in [31] as follows.

Theorem 6. *Let k be the optimal number of gene markers deleted from the input genomic maps. MSR can be solved in $O(2^{3.61k}n + n^2)$ time; i.e., MSR is fixed-parameter tractable.*

Note that as k is typically greater than 50 in real datasets, our FPT algorithms are not yet practical.

5 Concluding Remarks and Open Problems

The negative results on EBD and ENbS do not mean that we have absolutely no way to tackle these problems. For instance, in [2], with integer linear pro-

gramming, very nice empirical results are obtained. Here, we try to present a different way to handle these problems formally.

In many biological problems, the optimal solution value OPT could be zero. (Besides EBD, in some minimum recombination haplotype reconstruction problems the optimal solution value could be zero.) As implied by Theorem 1, if computing such an optimal solution with zero solution value is NP-complete then the problem does not admit *any* polynomial time approximation (unless P=NP). However, in reality one would be satisfied to obtain a solution with value one or two. Due to this reason, we can relax the traditional definition of approximation to a *weak approximation*. Given a minimization problem Π, let the optimal solution of Π be OPT. We say that a weak approximation algorithm $\mathcal{W}$ provides a *performance guarantee* of α for Π if for every instance I of Π, the solution value returned by $\mathcal{W}$ is at most $\alpha \times (\text{OPT} + 1)$.

In [10–12] we showed that EBD and the exemplar conserved interval distance problems are both hard to approximate even under the weak approximation model. But for the exemplar reversal distance problem, no such result is known yet.

For the exemplar common interval number problem [3], the only negative result is its NP-hardness. It would also be interesting to know whether it admits an efficient polynomial time approximation. We conclude this paper with a list of open problems.

1. For the exemplar reversal distance problem, does there exist a good weak approximation?
2. For the exemplar common interval number problem, does there exist a good approximation?
3. For the MSR problem, does there exist a polynomial time approximation with factor better than 4?
4. For the MSR problem, does there exist a more efficient FPT algorithm?

Acknowledgments

I would like to thank my collaborators for this series of research: Zhixiang Chen, Richard Fowler, Bin Fu, Minghui Jiang, Lusheng Wang, Jinhui Xu, Boting Yang, and Zhiyu Zhao. Special thanks to Jianer Chen for answering many questions regarding FPT.

References

1. S. Angibaud, G. Fertin and I. Rusu. On the approximability of comparing genomes with duplicates. *Proc. 2nd Workshop on Algorithm and Computation (WALCOM'2008)*, LNCS 4921, pp. 34-45, 2008.
2. S. Angibaud, G. Fertin, I. Rusu, A. Thévenin and S. Vialette. Efficient tools for computing the number of breakpoints and the number of adjacencies between two genomes with duplicate genes. *J. Computational Biology*, 15:1093-1115, 2008.

3. G. Blin, C. Chauve, G. Fertin, R. Rizzi and S. Vialette. Comparing genomes with duplicates: a computational complexity point of view. *IEEE/ACM Trans. on Computational Biology and Bioinformatics)*, 4:523-534, 2007.
4. G. Blin, G. Fertin, F. Sikora and S. Vialette. The exemplar breakpoint distance for non-trivial genomes cannot be approximated. *Proc. 3nd Workshop on Algorithm and Computation (WALCOM'2009)*, to appear, 2009.
5. R. Bar-Yehuda, M.M. Halldórsson, J.(S.) Naor, H. Shachnai, and I. Shapira. Scheduling split intervals. *SIAM Journal on Computing*, 36:1–15, 2006.
6. V. Bafna and P. Pevzner, Sorting by reversals: Genome rearrangements in plant organelles and evolutionary history of X chromosome, *Mol. Bio. Evol.*, 12:239-246, 1995.
7. D. Bryant. The complexity of calculating exemplar distances. *In D. Sankoff and J. Nadeau, editors*, *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment, and the Evolution of Gene Families*, pp. 207-212. Kluwer Acad. Pub., 2000.
8. G. Blin and R. Rizzi. Conserved interval distance computation between non-trivial genomes. *Proc. 11th Intl. Ann. Comput. and Combinatorics (COCOON'05),* LNCS 3595, pp. 22-31, 2005.
9. A. Bergeron and J. Stoye. On the similarity of sets of permutations and its applications to genome comparison. *Proc. 9th Intl. Ann. Comput. and Combinatorics (COCOON'03),* LNCS 2697, pp. 68-79, 2003.
10. Z. Chen, B. Fu and B. Zhu. The approximability of the exemplar breakpoint distance problem. *Proc. 2nd Intl. Conf. on Algorithmic Aspects in Information and Management (AAIM'06)*, LNCS 4041, pp. 291-302, 2006.
11. Z. Chen, B. Fu, R. Fowler and B. Zhu. Lower bounds on the approximation of the exemplar conserved interval distance problem of genomes. *Proc. 12th Intl. Ann. Comput. and Combinatorics (COCOON'06),* LNCS 4112, pp. 245-254, 2006.
12. Z. Chen, B. Fu, R. Fowler and B. Zhu. On the inapproximability of the exemplar conserved interval distance problem of genomes. *J. Combinatorial Optimization*, **15**(2):201-221, 2008.
13. Z. Chen, B. Fu, B. Yang, J. Xu, Z. Zhao, and B. Zhu. Non-breaking similarity of genomes with gene repetitions. In *Proceedings of the 18th Annual Symposium on Combinatorial Pattern Matching (CPM'07)*, LNCS 4580, pages 119–130, 2007.
14. Z. Chen, B. Fu, M. Jiang, and B. Zhu. On recovering syntenic blocks from comparative maps. In *Proceedings of the 2nd Annual International Conference on Combinatorial Optimization and Applications (COCOA'08)*, LNCS 5165, pages 319–327, 2008.
15. J. Chen, X. Huang, I. Kanj, and G. Xia. Linear FPT reductions and computational lower bounds. In *Proceedings of the 36th ACM Symposium on Theory of Computing (STOC'04)*, pages 212–221, 2004.
16. T. Cormen, C. Leiserson, R. Rivest, C. Stein. *Introduction to Algorithms*, second edition, MIT Press, 2001.
17. V. Choi, C. Zheng, Q. Zhu, and D. Sankoff. Algorithms for the extraction of synteny blocks from comparative maps. In *Proceedings of the 7th International Workshop on Algorithms in Bioinformatics (WABI'07)*, pages 277–288, 2007.
18. R. Downey and M. Fellows. *Parameterized Complexity*, Springer-Verlag, 1999.
19. M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness.* Freeman, San Francisco, CA, 1979.
20. J. Hästad. Clique is hard to approximate within $n^{1-\epsilon}$. *Acta Mathematica,* **182**:105-142, 1999.

21. S. Hannenhalli and P. Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *J. ACM,* **46**(1):1-27, 1999.
22. O. Gascuel, editor. *Mathematics of Evolution and Phylogeny.*Oxford University Press, 2004.
23. M. Marron, K. Swenson and B. Moret. Genomic distances under deletions and insertions. *Theoretical Computer Science,* **325**(3):347-360, 2004.
24. C. Makaroff and J. Palmer. Mitochondrial DNA rearrangements and transcriptional alternatives in the male sterile cytoplasm of Ogura radish. *Mol. Cell. Biol.*, **8**:1474-1480, 1988.
25. C.T. Nguyen, Y.C. Tay and L. Zhang. Divide-and-conquer approach for the exemplar breakpoint distance. *Bioinformatics*, **21**(10):2171-2176, 2005.
26. J. Palmer and L. Herbon. Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *J. Mol. Evolut.*, **27**:87-97, 1988.
27. D. Sankoff. Genome rearrangement with gene families. *Bioinformatics*, **16**(11):909-917, 1999.
28. T. Schaefer. The complexity of satisfiability problem. In *Proceedings of the 10th ACM Symposium on Theory of Computing (STOC'78)*, pages 216–226, 1978.
29. A. Sturtevant and T. Dobzhansky. Inversions in the third chromosome of wild races of *drosophila pseudoobscura*, and their use in the study of the history of the species. *Proc. Nat. Acad. Sci. USA,* 22:448-450, 1936.
30. G. Watterson, W. Ewens, T. Hall and A. Morgan. The chromosome inversion problem. *J. Theoretical Biology*, **99**:1-7, 1982.
31. L. Wang and B. Zhu. On the tractability of maximal strip recovery. *This proceeding*, 2009.
32. C. Zheng, Q. Zhu, and D. Sankoff. Removing noise and ambiguities from comparative maps in rearrangement analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4:515–522, 2007.