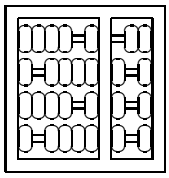

INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

A Rearrangement-Based Approach to Compare Whole Genomes of *Vibrionaceae* Strains

P. P. Côgo *J. Meidanis*

Technical Report - IC-08-005 - Relatório Técnico

February - 2008 - Fevereiro

A Rearrangement-Based Approach to Compare Whole Genomes of *Vibrionaceae* Strains*

Patrícia P. Côgo† João Meidanis‡

Abstract

The *Vibrionaceae* family comprises a wide variety of organisms, including some severe human pathogens. Currently, 8 completely sequenced genomes are public available[1], and a natural challenge that arises from this scenario is to employ this information to establish a modern phylogeny.

Toward this goal, we propose an approach based on genomic rearrangements to compare complete genomes of *Vibrionaceae* strains, which can in principle be applied, with adaptations, to other species. In our approach, we employ a profile-based methodology to identify homologous genes and model evolutionary events such as gene losses, lateral gene transfers, and chromosomal rearrangements to estimate evolutive distance, which we believe can substantially complement analyzes based on gene markers or phenotypic characteristics.

1 Introduction

Vibrionaceae is a heterogeneous family, comprising organisms from ten different genera. Vibrios, i. e., *Vibrionaceae* strains, are abundantly found in aquatic environments or in association with eukaryotes. Although some associations could be symbiotic, severe pathogens of humans and other species belong to this group — including *Vibrio cholerae*, the etiologic agent of cholera, a severe disease in developing countries [24].

Nowadays, the analysis by gene markers has almost completely substituted phenotypic analysis to define vibrio taxonomy. Although the genomic method presents advantages, it has some weaknesses too and, in some cases, it could be not conclusive enough [24].

An outstanding drawback of this method is the attempt to extend to the entire genome the evolutionary history constructed from a single gene, or from a small set of genes. There is no assurance that this inference is true in general, especially for bacteria — where a variety of mutation events, in addition to nucleotide substitutions, have been shown to occur. In particular, it does not take into consideration evolutionary events affecting gene content and gene order, such as gene losses, gene duplications, lateral gene transfers, and

*This research was partially supported by grants by FAPESP (05/53279-6) and CNPq (470420/2004-9).

†Institute of Computing, University of Campinas, Campinas, Brazil

‡Institute of Computing, University of Campinas, and Scylla Bioinformatics, Campinas, Brazil

[1]Source: NCBI, 11/11/2007

chromosomal rearrangements — that occur very frequently in bacteria and are responsible for the large diversity of this group [5, 18, 19].

To overcome these problems, complementing the results obtained from phylogenetic markers, we began using genome rearrangement theories and tools to help analyze complete genomes of vibrio species, accounting for variations in gene content and chromosomal rearrangements to estimate evolutionary distances.

Hannenhalli and Pevzner [12] report that in 1936 Sturtevant and Dobzhansky implicitly discussed issues related to genome rearrangement distances. The groundbreaking work by Nadeau and Taylor in 1984 provided a statistical estimate of the distance between human and mouse [17]. The combinatorial aspects of the problem were put forward by Sankoff *et al.* in the 1990's [21]. Applications to analyzing actual biological cases, to name but a few, include the work by Bafna and Pevzner [2], where the evolution of the X chromosome in mammals was analyzed, Hannenhalli et al. [11], where whole genomes of herpes viruses were analyzed, Belda et al. [3], where the class of γ-proteobacteria was analyzed by various methods, including genome rearrangements, and Lin et al. [15], analyzing vibrio species.

Nevertheless, while rearrangement theory researches usually consider that the problems of determining homology have been solved or pay little attention to them [20], here we focused in establishing a reliable methodology to determine families of orthologous genes, following some criteria usually not accomplished by traditional methods based on sequence similarity. Our method uses information provided by HAMAP project [8] and in our own complementary set of profiles [10].

This HAMAP/profile based method was combined in this work with rearrangement algorithms in order to estimate evolutive distances between six completely sequenced vibrio genomes, namely: *Vibrio cholerae*, *Vibrio parahaemolyticus*, *Vibrio fisheri*, *Vibrio vulnificus* strains CMCP6 and YJ016, and *Photobacterium profundum*. The phylogenetic trees obtained when each chromosome was analysed individually coincide with the trees obtained using different genomic methods.

The rest of the paper is organized as follows. Discussion about homology and a description of the method employed in this work are presented in section 2. Section 3 presents the model used in this work to estimate evolution based on the rearrangement theory. Results obtained and discussion are presented in section 4.

2 Homology

Homology is defined as origin in a common ancestor. Homology detection is crucial in various biological problems, including comparative genomics, taxonomic analysis, inference of protein structures, and others. For this reason, many researchers have developed different methods to identify homologous genes or proteins. These methods are based on widely different techniques, ranging from simple sequence comparison to very complex clustering algorithms.

Despite the large amount of studies in this area, there is no consensus about the best approach to detect homology. While simple methods such as pairwise sequence alignment are efficient when analyzing conserved sequences [23], to find an effective method to analyze

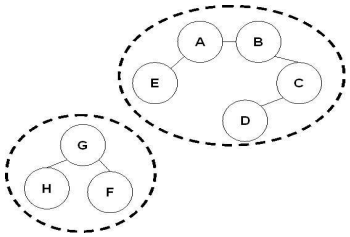

Figure 1: An example of homology identification through similarity. Dashed curves delimit gene families created by a single linkage method.

distantly related sequences is still an open problem.

Based on our observations, we defined a few very important properties that the method must possess, described in the next section. Not all existing methods satisfy these properties.

2.1 Criteria for Homology

The precision of the evolutive estimation based on gene content and chromosomal rearrangement is strongly dependent on the homology detection accuracy. In order to define a consistent and accurate method, we delineated two criteria that it must observe.

The first of them is *transitivity*, that is, if A is homologous to B, and B is homologous to C, then A must be homologous to C. This is an intrinsic rule of homology relationships. Indeed, if A and B have a common ancestor, and B and C also do, then A and C must have a common ancestor as well.

Sequence similarity is a popular method used to detect homology. A score above a certain threshold is indicative of homology. This method may violate transitivity, as illustrated in Figure 1. In this picture, letters represent genes and edges connect genes whose similarity is larger than a certain threshold. If similarity is the criterion employed to identify homology, A and B would be homologs, as would B and C. In spite of this, the homology between A and C is not detected, because they are not sufficiently similar.

Usually, this problem is solved grouping homologous genes by sequence clustering [9]. For example, the application of a single linkage method (i.e, where each member of a group is, at least, similar to one other member) to the case depicted in Figure 1 results in two gene families {A, B, C, D, E} and {F, G, H}.

Despite solving the transitivity problem, this solution brings in another problem. In our example, A, B, and C are members of the same family. However, if B is taken out of the analysis, A and C would be no longer considered homologous. This leads to an undesirable property, namely, the homology between A and C depends on the presence of a third gene B. To avoid this problem, we postulate that our homology detection method should be independent of the set of genes being analyzed. According to this rule, homology can not be based on similarity between two sequences, but in a external criterion to assign a sequence in a gene family, independently of the other genes.

In summary, our homology identification method must observe the following criteria:

1. Transitivity;

2. Independence of the genes being analyzed.

Complying with these rules, we adopted a homology identification method based on information of protein families provided by the HAMAP project. Unfortunately, a large fraction of vibrio proteins was not covered by HAMAP yet. So, we proceed a collection of our own set of profiles to describe the remain families. Our methodology is detailed in the next sections.

2.2 Profiles

Profiles are weighted matrices that describe families of proteins. They are constructed from multiple sequence alignments and constitute an efficient method to identify even distantly related sequences. We say that a sequence belongs to a family when its alignment to the profile results in a "match", that is, a score higher than a threshold [10, 22]. If a sequence belongs to more than one family, we take the family with highest score.

This method may produce false positives or false negatives, as any method will, but at least it satisfies the criteria of transitivity and independence of the genes being analyzed, defined in Section 2.1. Since all sequences that "match" the same profile will receive the same family "name", transitivity is guaranteed. In addition, the alignment between a sequence and a profile does not depend on any other sequence being analyzed.

2.3 Determining Homology

The first step in our method was to send the genomic data, in the form of 27282 proteins from all six vibrio strains, to the HAMAP server[2] The query resulted in the classification of 3217 genes, which represent only 12% of the complete set. In order to determine homology for the remaining proteins, we proceed to augment HAMAP with our own profiles.

The first step in the profile development procedure is to choose the multiple sequence alignments. Here, we selected these sequences from GeneBank using the program BLASTP [1] applying the following criteria:

- If the protein length is greater than or equal to 120 aa:
  - Select the 50 best hits with *e-value* $\leq 1e^{-50}$ and whose length does not differs more than 30% from the query length;
- If the protein length is less than 120 aa:
  - Select the 50 best hits with *e-value* $\leq 1e^{-3}$.

After this step, the selected sequences were aligned using ClustalW [6]. Then, the program PFMAKE, provided by the PROSITE project [22], was used to construct a profile from this multiple alignment.

[2]Submission date: 08/31/2006.

The profile constructed is called an "unscaled profile", *i.e.*, the scores assigned from this profile do not have any biological interpretation [22]. In order to accomplish meaningful scores, we assigned to each profile a normalization function. This is a linear function of the form:

$$f(x) = \frac{10x}{a}$$

where x is the unscaled score and a is the result of the alignment between the original protein and its own unscaled profile. Therefore, if an alignment between a protein and a profile presents a normalized score close to 10, we assign this protein to the family described by the profile. In this work, we assume 8.5 as the lowest and 11.75 as the highest score to affirm that a protein is assigned by a profile.

Naturally, before a profile is added to our database, we check whether it really describes a new family or is equivalent to another, previously constructed, profile, in which case the repeated profile is discarded.

This procedure is repeated for each protein that was not classified by the HAMAP project, until all of them have been assigned to a family described by a profile or have been classified as a singleton.

3 Estimating Evolution

In this work we apply a set of rearrangement operations to represent biological events relevant to vibrio evolution. Our model is based on the occurrence of four basic operations: *fusions*, *fissions*, *reversals*, and *gene losses* (modeled as a special case of fission).

Translocations, block-interchanges, and transpositions (which can be seen as special cases of block-interchanges), are not fundamental events, but rather special cases or results of two successive operations, namely, a fission followed by a fusion. Genes losses and lateral gene transfers can be viewed as special cases of fissions and fusions, respectively, combined with a "chromosome" loss/gain event.

In this section we describe our model in more detail, relating our treatment for gene losses, duplications, and finally describing the rearrangement distances estimation.

3.1 Gene Losses

The hypotheses of equality of gene content usually adopted in theoretical works about genomic rearrangement is far from true in vibrio genomes, that are in continuous evolution through gene losses and lateral gene transfers in order to accomplish a better environmental adaptation [24]. To take into account variation in gene content, we include a special operation:

- Gene Loss.

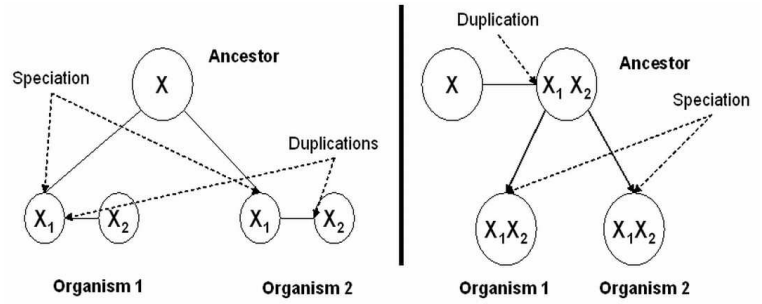

Figure 2: Comparison between two genomes, each having two copies of gene X. In case (a), the paralog genes come from a duplication process; in case (b), the two copies are inherited from the ancestral genome.

This operation is treated as a case of gene fission and will also have unitary weigth. After application of this operation on all genes that belong to just one of the organisms, the genomes being compared will have the same gene content — satisfying the assumptions of the rearrangement algorithms.

3.2 Duplications

Gevers and Van de Peer, in their study about duplications in vibrio genomes, emphasize the role played by these events in adaptation to changing environments and exploitation of new niches. The fraction of duplicated genes in vibrio genomes reflects their importance: the analysis of complete genomes have shown rates between 28,19% (*Vibrio choleare*) and 36,66% (*Photobacterium profundum*) [9].

Nevertheless, it is not an easy task to identify and correctly handle duplications, as illustrated in Figure 2. In this example, two genomes are being compared and each of them has two copies of gene X. There are two ways of treating them: (i) keeping a single copy in each genome and handling paralogs as gene gain/loss or (ii) associate each copy in one of the genomes with one in the other genome. The correct treatment depends on the closest ancestral genome: if it had just one copy of X, as illustrated in Figure 2 item (a), the first solution would be more appropriate. On the other hand, if the ancestor has had also two copies of X, as exemplified in item (b), the second solution would be more correct.

Since we usually one do not have enough information about the closest common ancestor, we adopt a compromise here. For each gene family with more than one member in one or both genomes, a ultrametric tree with all family members is built. Then, this tree is split into subtrees corresponding to subfamilies.

We can associate the degree of strictness with a certain height at which to cut the phylogeny. The method consist in pruning the phylogenetic tree as close as possible to the root, in order to create subtrees without paralog genes. Each subtree will constitute a new family, as exemplified in Figure 3.

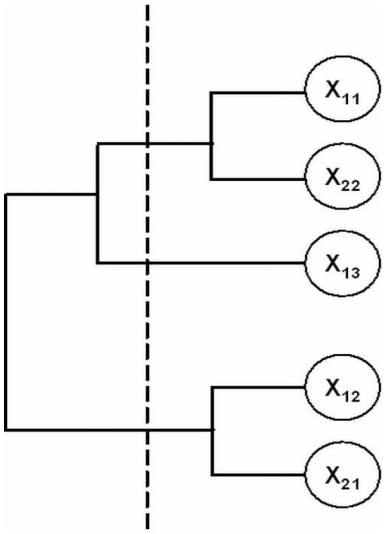

Figure 3: Two genomes are being compared and a family composed by five genes is identified: genes $\{X_{11}, X_{12}, X_{13}\}$ belonging to organism 1, and $\{X_{21}, X_{22}\}$ belonging to organism 2 (the first subscript identifies the organism to which the gene belongs). After adequate pruning, we will have three new gene families, namely $\{X_{11}, X_{22}\}$, $\{X_{13}\}$, and $\{X_{12}, X_{21}\}$, with no paralogs in any of the families.

After this procedure, there will be no duplicated genes. Therefore, we could apply rearrangement algorithms to the resulting genomes.

3.3 Rearrangement Distances

After we have determined homolog structures and have treated genes gain/loss, and duplications, the final problem is to estimate rearrangement distances, *i.e.*, to find the minimal number of mutational operations necessary to transform a genome into another.

While most classical genome rearrangement works describe methods and algorithms to calculate distances using a single operation [7, 14], our approach is to employ a set of basic operations and combinations of them in order to accomplish more realistic and meaningful distances.

In accordance with these principles, Yancopoulos *et al.* introduce a universal operation called *double-cut-and-join* that models fissions, fusions, translocations, reversals and block-interchanges [25].

Briefly, a *double-cut-and-join* operation consists in cutting the genome in two points, resulting in 4 unconnected vertices, and rejoining them in a different way. Depending on the way this reconnection is done, it could result in a translocation (including fission and fusion as special cases) or in a reversal — each of these operations has weight 1. Because a block-interchange can only be accomplished with two consecutive operations, it receives weight 2.

We employ in this work the method described by Bergeron *et al.* [4] to apply the *double-cut-and-join* operation in order to compute rearrangement distances in multichromossomal genomes.

4 Experimental Results

We have applied our method to compare complete genomes of six vibrio complete genomes. The steps necessary to estimate the evolutive distance between each pair of vibrio genomes are briefly described below.

1. Classify each vibrio protein as either a HAMAP family member, or as covered by one of our additional profiles, or as singleton;

2. Identify families with multiple members in one or both genomes being compared. These families are subdivided, as described in Section 3.2. At the end of this step, there are no more duplicated genes;

3. Eliminate sequences that do not have a homolog in the other genome, as discussed in Section 3.1. They will be counted as a gene loss/gain operation;

4. Finally, apply a rearrangement algorithm to the genomes with same gene content and no duplications, according to Section 3.3.

Here, two different approaches were employed to estimate evolution among the six vibrio species: analysing the two chromosomes combined and separately. The results obtained are described in the next sections.

4.1 Analysing the two vibrio chromosomes combined

When the two chromosomes are combined to estimate the evolution between the six vibrio species, the resulting matrix distance is represented in table 1. The phylogenetic tree showed in figure 4 is a graphical representation of these distances, built with MEGA 4 software [13].

	V. c.	*V. p.*	*V. f.*	*V. v.* C.	*V. v.* Y.	*P. p.*
V. choleare	0	6292	6277	6296	5969	7351
V. parahaemolyticus		0	7154	6973	6614	7728
V. fisheri			0	7165	6745	8111
V. vulnificus CMCP				0	5835	7682
V. vulnificus YJ016					0	8114
P. profundum						0

Table 1: Evolutive distances among six vibrio species.

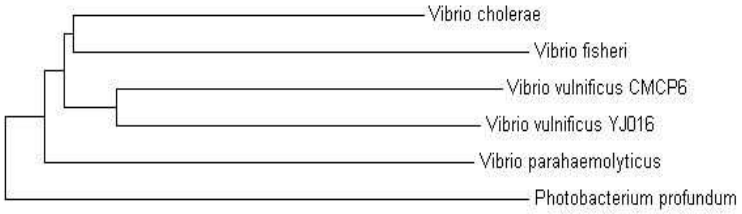

Figure 4: Phylogenetic tree of the six vibrio strains built from our distance matrix using the Neighbor-Joining method.

This phylogenetic tree agrees in some aspects with the accepted tree of the *Vibrionacea* family built by traditional analysis of gene markers [24]. Nevertheless, *V. choleare* and *V. fisheri* are wrongly grouped. In order to correct our results, we analysed each chromosome individually.

4.2 Analysing the two choromosomes individualy

When the bigger and the smaller chromosomes are analysed separately, we have obtained the distance matrices show in tables 2 and 3, respectively. Figures 5 and 6 present the corresponding phylogenetic trees.

	V. c.	*V. p.*	*V. f.*	*V. v.* C.	*V. v.* Y.	*P. p.*
V. choleare	0	3906	4403	3906	4150	4645
V. parahaemolyticus		0	4701	4031	4288	4929
V. fisheri			0	4548	4848	5062
V. vulnificus CMCP				0	3831	4757
V. vulnificus YJ016					0	5052
P. profundum						0

Table 2: Evolutive distances between the six vibrio strains analysing only the bigger chromosome.

The phylogenetic trees showed in figures 5 and 6 have the same topology and they agree with the accepted phylogenetic tree of the *Vibrionacea* family built by traditional analysis of gene markers [24]. This result validates of our method to compare whole genomes, and also provides an indication that rearrangement events have played an important role in the evolution of vibrios.

Our method also allows the computation of the last universal common ancestor, much in the same way as Koonin and colleagues did with gene loss and gain, but here involving more rearrangement operations [16].

	V. c.	*V. p.*	*V. f.*	*V. v.* C.	*V. v.* Y.	*P. p.*
V. choleare	0	2427	2061	2226	2336	2926
V. parahaemolyticus		0	2697	2667	2779	3512
V. fisheri			0	2416	2561	2942
V. vulnificus CMCP				0	2006	3266
V. vulnificus YJ016					0	3407
P. profundum						0

Table 3: Evolutive distances between the six vibrio strains analysing only the smaller chromosome.

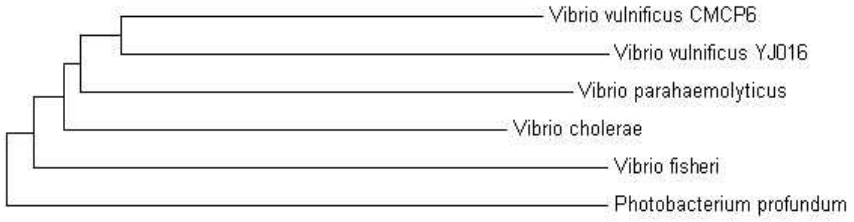

Figure 5: Phylogenetic tree from six vibrio strains obtained from the analysis of their bigger chromosome.

Although, the divergences between the trees obtained from the analysis of each chromosome and the analysis of their combination can attest that the rearrangement operations, specially translocations, must be better weighting.

5 Conclusions

We propose in this work a method based on genome rearrangements to analyze evolution of completely sequenced vibrio species.

Our approach covers completely the genomic comparison process, including treatments to genes gain/loss, that can represents mutational events such lateral gene transfers, and duplications. In order to obtain more realistic and meaningful rearrangement distances, we use in this work the *double-cut-and-join* operation, which models almost all rearrangement events, such as reversals, translocations (including fission and fusion), and block-interchanges.

The estimated phylogenetic trees analysing each chromosome individually coincide with the trees of the *Vibrionacea* family obtained using other genomic methods. This result validates our method and asserts to the adequacy of the genome rearrangements theory in large scale genomes comparison. However, the divergences between these trees and the tree obtained from the analysis of the two chromosomes combined demonstrate that the rearrangement opperations, specially translocations, must be better weighting.

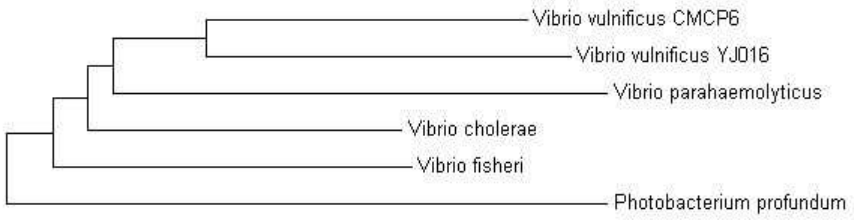

Figure 6: Phylogenetic tree from six vibrio strains obtained from the analysis of their smaller chromosome.

Our methodology can also be applied to estimate the last universal common ancestor, in the same way as Koonin and colleagues, but involving more complex rearrangement operations.

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Computacional Biology*, 215:1403–1410, 1990.

[2] B. Bafna and P. Pevzner. Sorting by transpositions. *SIAM J. Discrete Mathematics*, 11(2):224–240, 1998.

[3] E. Belda, A. Moya, and F. J. Silva. Genome rearrangement distances and gene order phylogeny in gamma-Proteobacteria. *Mol. Biol. Evol.*, 22(6):1456–1467, 2005.

[4] Anne Bergeron, Julia Mixtacki, and Jens Stoye. A unifying view of genome rearrangements. In *proocedings of WABI 2006*, volume 4175 of *Lecture Notes in Computer Science*, pages 163–173, 2006.

[5] C. Chen, K. Wu, Chang Y., C. Chang, H. Tai, T. Liao, Y. Liu, H. Chen, A. Shen, J. Li, C. Shao, C. Lee, L. Hor, and S. Tsai. Comparative genome analysis of *Vibrio vulnificus*, a marine pathogen. *Genome Research*, 13:2577–2587, 2003.

[6] R. Chenna, T. Sugawara, H.and Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. son. Multiple sequence alignment with the CLUSTAL series of programs. *Nucleic Acids Res*, 31:3497–3500, 2003.

[7] D. A. Christie. Sorting permutation by block-interchanges. *Information Processing Letters*, 60(4):165–169, 1996.

[8] A. Gattiker, K. Michoud, C. Rivoire, A. H. Auchincloss, E. Coudert, T. Lima, P. Kersey, M. Pagni, C. J. Sigrist, C. Lachaize, A. L. Veuthey, E. Gasteiger, and A. Bairoch. Automatic annotation of microbial proteomes in Swiss-Prot. *Comput. Biol. Chem.*, 27:49–58, 2003.

[9] D. Gevers and Y. Van de Peer. Gene duplicates in vibrio genomes. In F. L. Thompson, B. Austin, and Swings, editors, *Invited book chapter in The Biology of Vibrios*. J. ASM Press, 2006.

[10] M. Gribskov, A. D. Mclachlan, and D. Eisenberg. Profile analysis: detection of distantly related proteins. *Proc. Natl. Aca. Sci. USA*, 84:4355–4358, 1987.

[11] S. Hannenhalli, C. Chappey, E. Koonin, and P. A. Pevzner. Genome sequence comparison and scenarios for gene rearrangements: a test case. *Genomics*, 30:299–311, 1995.

[12] S. Hannenhalli and P.A. Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *Proceedings of the 27th Annual Symposium on the Theory of Computing (STOC 95)*, pages 178–189, 1995.

[13] Tamura K., Dudley J., Nei M., and Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, 24:1596–1599, 2007.

[14] H. Kaplan, R. Shamir, and R. E. Tarjan. Faster and simpler algorithm for sorting signed permutations by reversal. *SIAM Journal of Computing*, 29(3):880–892, 2000.

[15] C. Lin, C. Lu, H. Chang, and C. Tang. An efficient algorithm for sorting by block-interchanges and its application to the evolution of vibrio species. *Journal of Computacional Biology*, 12(1), 2005.

[16] Boris G. Mirkin, Trevor Fenner, Michael Galperin, and Eugene Koonin. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology*, 2(3), 2003.

[17] J. H. Nadeau and B. A. Taylor. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Nat. Acad. Sci. USA*, 81:814–818, 1984.

[18] H. Ochman, J. G. Lawrence, and E. A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405:299–304, 2000.

[19] H. Ochman and N. A. Moran. Genes lost and genes found: Evolution of bacterial pathogenesis and symbiosis. *Science*, 292(5519):1096–1098, 2001.

[20] D. Sankoff and N. El-Mabrouk. Genome rearrangement. *In Current Topics in Computational Biology*, 2001.

[21] D. Sankoff, G. Leduc, N. Antoine, B. Paquin, B. F. Lang, and R. Cedergren. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc. Nat. Acad. Sci. USA*, 89:6575–6579, 1992.

[22] C.J.A. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinformatics*, 3:265–274, 2002.

[23] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1):33–36, 2000.

[24] F. L. Thompson, T. Iida, and J. Swings. Biodiversity of Vibrios. *Microbiology and Molecular Biology Reviews*, 68(3):403–431, 2004.

[25] S. Yancopoulos, O. Attie, and R. Friedberg. Efficient sorting genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346, 2005.