SORTING BY SIGNED REVERSALS

PHILLIP COMPEAU

At the turn of the 20th Century, biologists surmised that new traits must be introduced by chromosomal mutations. In 1921, Alfred Sturtevant examined the genetic linkage maps from two species of fruit fly (*Drosophila*) and observed that an interval of genes located on chromosome 3 had been inverted in one of the species. His discovery provided the first substantial evidence of a specific mutation occurring at some point during the evolutionary history of an entire species.

A decade later, Theophilus Painter suggested the investigation of polytene chromosomes to analyze mutations directly. These chromosomes, found in cells taken from the salivary glands of certain fruit flies, are made huge as a result of unchecked gene replication without mitosis. When dyed, areas of a chromosome undergoing a greater amount of gene transcription will appear lighter, dividing the gigantic polytene chromosomes into clear alternating black and white bands of varying widths that biologists can view with a light microscope (Fig. 1). Sturtevant worked alongside Theodosius Dobzhansky to compare banding patterns from the polytene chromosomes of different fly species, and they noted chromosomal segments in which segments of bands had clearly been inverted (Fig. 2).

A chromosomal segment inversion, or *reversal*, occurs when an interval of DNA coils into a loop-de-loop and the endpoints of the interval trade bonds (Fig. 3). Reversals are now known to be one of a variety of structural genomic mutations, most of which are very harmful to the mutated cell, but some of which may equip the organism with a beneficial trait that can be positively reinforced by the action of natural selection in further generations.

If chromosomes C_1 and C_2 from two related species share (roughly) the same n genes, then we may consider a model in which the two chromosomes have been separated solely as a result of reversals. Therefore, we would like to calculate the minimum number of reversals required to change C_1 into C_2 to provide a function representing evolutionary distance between chromosomes. We call this minimum number of reversals the *reversal distance* between C_1 and C_2, denoted $d(C_1, C_2)$.

For the sake of simplicity, we will label the genes of C_2 in order with $1, 2, \ldots, n$. We then represent C_1 by a *signed permutation*, or an ordering of $\{1, 2, \ldots, n\}$ indexing the gene ordering of C_1, and for which an element is positive (negative) if the gene it represents has the same (opposite) orientation in C_2. For example, in Fig. 4a, we have that $C_1 = (-4, 1, 3, -5, 2)$ (and of course $C_2 = (1, 2, 3, 4, 5)$). A reversal applied to C_1 simply inverts the corresponding interval of indices and changes their sign; for instance, reversing the interval formed of the middle three indices of C_1 yields the new signed permutation $(-4, 5, -3, -1, 2)$ (Fig. 4b). Note that reversals of a single index are allowed.

We may also represent the relationship between C_1 and C_2 graphically as follows. Assign a "head" (h) and "tail" (t) to each gene, representing the gene's endpoints. A gene's orientation is now determined by the ordering of its two ends, so that for instance -2 may be encoded as $(2\text{h}, 2\text{t})$. Adding two null gene ends to represent the beginning and ending of the chromosome, C_1 may be represented as $(0\text{h}, 4\text{h}, 4\text{t}, 1\text{t}, 1\text{h}, 3\text{t}, 3\text{h}, 5\text{h}, 5\text{t}, 2\text{t}, 2\text{h}, 6\text{t})$. We now form a graph whose vertices are defined by a linear embedding of the gene ends from C_1, and in which adjacent ends in C_1 are joined by blue edges; join adjacent ends from C_2 by red edges in the same graph. Notice that the edges $\{1\text{t}, 1\text{h}\}, \{2\text{t}, 2\text{h}\}$, and so on will be both red and blue; we may remove these edges, as they do not provide us with any relevant information regarding the relationship between C_1 and C_2, which yields the *breakpoint graph* of C_1 and C_2, denoted $\text{B}(C_1, C_2)$ (Fig. 5a).

We may view the effect of a reversal on C_1 by how it changes the breakpoint graph with respect to C_2. Notice two facts. Firstly, observe (Fig. 5b) that a reversal inverts the vertices contained in the affected interval of C_2 but only changes the overall graph structure by rearranging two blue edges. Secondly, we will have transformed C_1 into C_2 when the breakpoint graph consists of $n+1$ cycles of length 2. We are now ready to state our problem.

PROBLEM: For the signed permutation C_1 on 500 elements given below (see "Problem Instance"), calculate $d(C_1, C_2)$, where $C_2 = (1, 2, \ldots, 500)$. Also, provide a minimum sequence of reversals that will yield this distance.

Solution Checking

A user will submit a value d for the reversal distance as well as a $(d+1) \times 500$ matrix A. The solution will be considered correct iff:

1. $d = 495$
2. The first row of A encodes the signed permutation representation of C_1: $[-371\ -119-245\ \ldots\ 221]$.
3. The final row of A encodes the signed permutation representation of C_2: $[1\ 2\ 3\ \ldots\ 500]$.
4. For each j, $1 \leq j \leq d$, the permutation in the $(j+1)$th row represents some reversal of the permutation in the jth row.

Draft Solution

The fact that any reversal applied to C_1 rearranges two blue edges of $\mathrm{B}(C_1, C_2)$ implies that the number of cycles in the breakpoint graph changes by at most 1 with any reversal of C_1, which immediately yields the following lemma.

Lemma 1. *$d(C_1, C_2) \geq n + 1 - c$, where c is the number of cycles in $B(C_1, C_2)$.*

For our particular instance of C_1, $\mathrm{B}(C_1, C_2)$ will have 6 cycles, implying that $d(C_1, C_2) \geq 495$. We can obtain an upper bound on $d(C_1, C_2)$ of $2n-1$ (i.e., 999) via a greedy algorithm that inductively places the largest element at the end of the current permutation and then flips it if necessary. The lower bound of Lemma **??** seems to be more realistic, and so we may wonder if it is in fact tight. Thus, we would like to devise an algorithm that always "splits" a cycle so as to increase the total number of cycles in the breakpoint graph by 1; even better, we would like to convert a $(2m+2)$-cycle $(m \geq 1)$ into a $2m$-cycle and a new 2-cycle at each step. Note that 2-cycles correspond precisely to blue edges of the form $\{k\mathrm{h}, (k+1)\mathrm{t}\}$ in $\mathrm{B}(C_1, C_2)$.

We can create a new 2-cycle via a reversal iff some elements k and $k+1$ in C_2 have opposite signs, or equivalently iff in $\mathrm{B}(C_1, C_2)$, both $k\mathrm{h}$ and $(k+1)\mathrm{t}$ (always joined by a red edge) appear on the same side of blue edges $\{k\mathrm{h}, x\}$ and $\{(k+1)\mathrm{t}, y\}$. This scenario occurs in turn iff the number of vertices separating $k\mathrm{h}$ and $(k+1)\mathrm{t}$ is odd. In this case, we call $\{k\mathrm{h}, (k+1)\mathrm{t}\}$ a *good edge* (or *good pair*) and the reversal that replaces blue edges $\{k\mathrm{h}, x\}$ and $\{(k+1)\mathrm{t}, y\}$ with blue edges $\{k\mathrm{h}, (k+1)\mathrm{t}\}$ and $\{x, y\}$ a *good reversal.*

Define the *score* of a reversal as the number of good pairs resulting from that reversal, and define a reversal to be *optimal* if it is a good reversal having maximum score over all good reversals. This definition implies the following heuristic for calculating reversal distance:

Heuristic 2. *At each step, select an optimal reversal.*

Though far from obvious, it turns out that this heuristic is exact for almost every signed permutation as $n \to \infty$; in particular, it can be applied to calculate $d(C_1, C_2)$ along with a minimum collection of reversals for our given instance. Though it is not necessary for the solution of this problem, in what follows we will illustrate why this heuristic is often exact.

Form a new graph called the *interleaving graph* of C_1 and C_2 (denoted $\mathrm{I}(C_1, C_2)$), whose vertices are the red edges of $\mathrm{B}(C_1, C_2)$, and for which two vertices v, w are joined by an edge if the edges corresponding to v and w cross in $\mathrm{B}(C_1, C_2)$. This crossing occurs precisely when the intervals between those red edges intersect without inclusion. Predictably, we will call a vertex of $\mathrm{I}(C_1, C_2)$ *good* if it encodes a good edge from $\mathrm{B}(C_1, C_2)$ and *bad* otherwise. Good vertices can be determined directly from the interleaving graph: the fact that the number of vertices separating a good pair in $\mathrm{B}(C_1, C_2)$ must be odd implies that a vertex of $\mathrm{I}(C_1, C_2)$ is good iff its degree is odd.

In [1], Bergeron noted that applying the (good) reversal encoded by a good vertex v acts on the interleaving graph by complementing the subgraph induced by v. Thus v will become isolated, and by a parity argument, each vertex incident to v will change from good to bad or vice-versa. Vertices not originally incident to v will of course have the same degree in the new interleaving graph. This proves the following lemma.

Lemma 3. *The score of the reversal encoded by a good vertex v is:*

$$T + b(v) - g(v) - 1$$

where T is the total number of good vertices in $I(C_1, C_2)$, $b(v)$ is the number of bad vertices adjacent to v, and $g(v)$ is the number of good vertices adjacent to v.

We now have an easily calculable formula for the score of a good reversal. For that matter, it will give us our desired result. Define a component of $\mathrm{I}(C_1, C_2)$ containing at least two vertices to be *good* if it contains a good vertex and *bad* otherwise; because reversals corresponding to good vertices only influence the interleaving graph locally, our heuristic will not work if we create any bad components. The following result, which can be proven relatively quickly by contradiction using Lemma 3, demonstrates why optimal reversals are so important.

Theorem 4. *An optimal reversal does not create new bad components in the interleaving* *graph.*

Thus, as long as $\mathrm{I}(C_1, C_2)$ does not contain any bad components to begin with, we will ensure that our heuristic is exact. It turns out that most random signed permutations do not contain bad components; it can easily be confirmed that this is the case for our chosen instance as well.

Additional Comments

A student could guess the correct distance from the lower bound, which is why we also demand the matrix component of the solution. More importantly, we reiterate that a heuristic will solve this problem. Using a heuristic seems to be acceptable, as the heuristic is nontrivial, exact in the majority of cases, and requires a reasonable amount of insight; for that matter, heuristics are so fundamental to bioinformatics (and computer science in general) that it would be both difficult and imprudent not to include them in this project.

Two objections are immediately relevant. First, a student could randomly select a good reversal that is not necessarily optimal at each step (or even just choose a random reversal increasing the number of cycles in the breakpoint graph) and get lucky enough to never create a bad component. We do not see any way in avoiding this contingency, and to alleviate it we propose suggesting further resources to students submitting correct responses. Second, a student could attempt to produce an exact algorithm that works on all possible datasets, which would require the student to recreate the expansive work of Hannenhalli and Pevzner in [2], which took a decade to be presented somewhat concisely in [1]. We have tried to avoid the possibility of a student getting bogged down in generality by providing the hint at the end of the problem statement urging users not to solve the problem for all possible inputs.

References

[1] Anne Bergeron. A very elementary presentation of the hannenhalli-pevzner theory. *Discrete Applied Mathematics*, 146(2):134–145, 2005.

[2] Sridhar Hannenhalli and Pavel Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *Journal of the ACM*, pages 178–189. ACM Press, 1995.

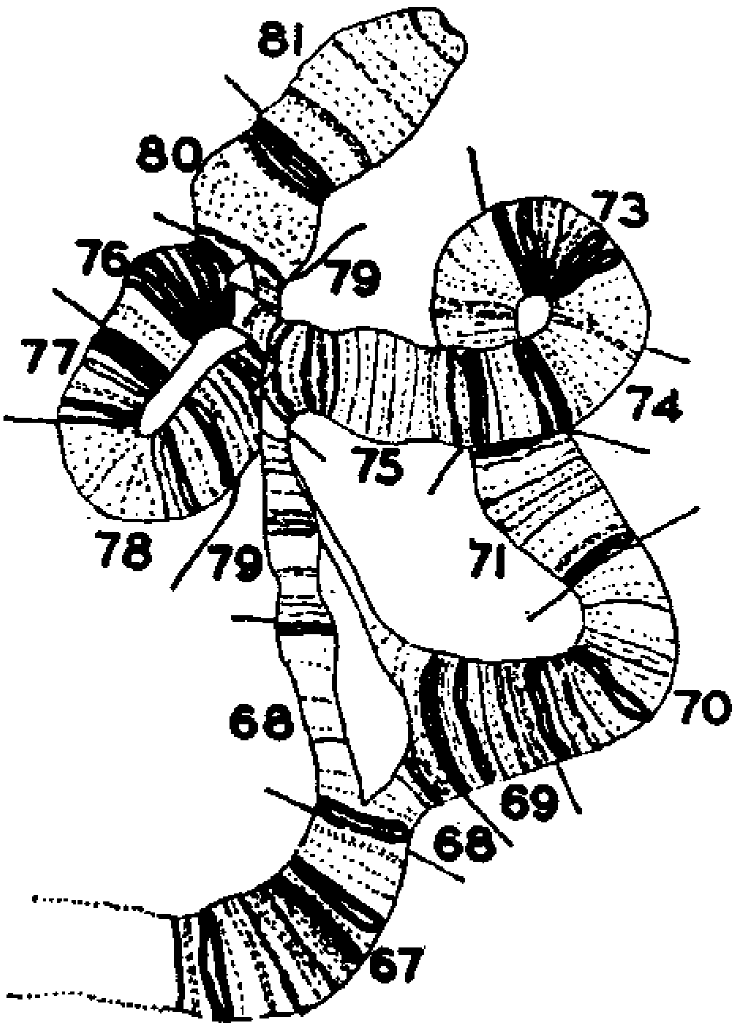

Fig. 1: Chromosome staining reveals banding patterns of polytene chromosomes of *Drosophila*. Reproduced from the work of Dobzhansky and Sturtevant ("Inversions in the Chromosomes of *Drosophila obscura*," *Genetics*, 1938).

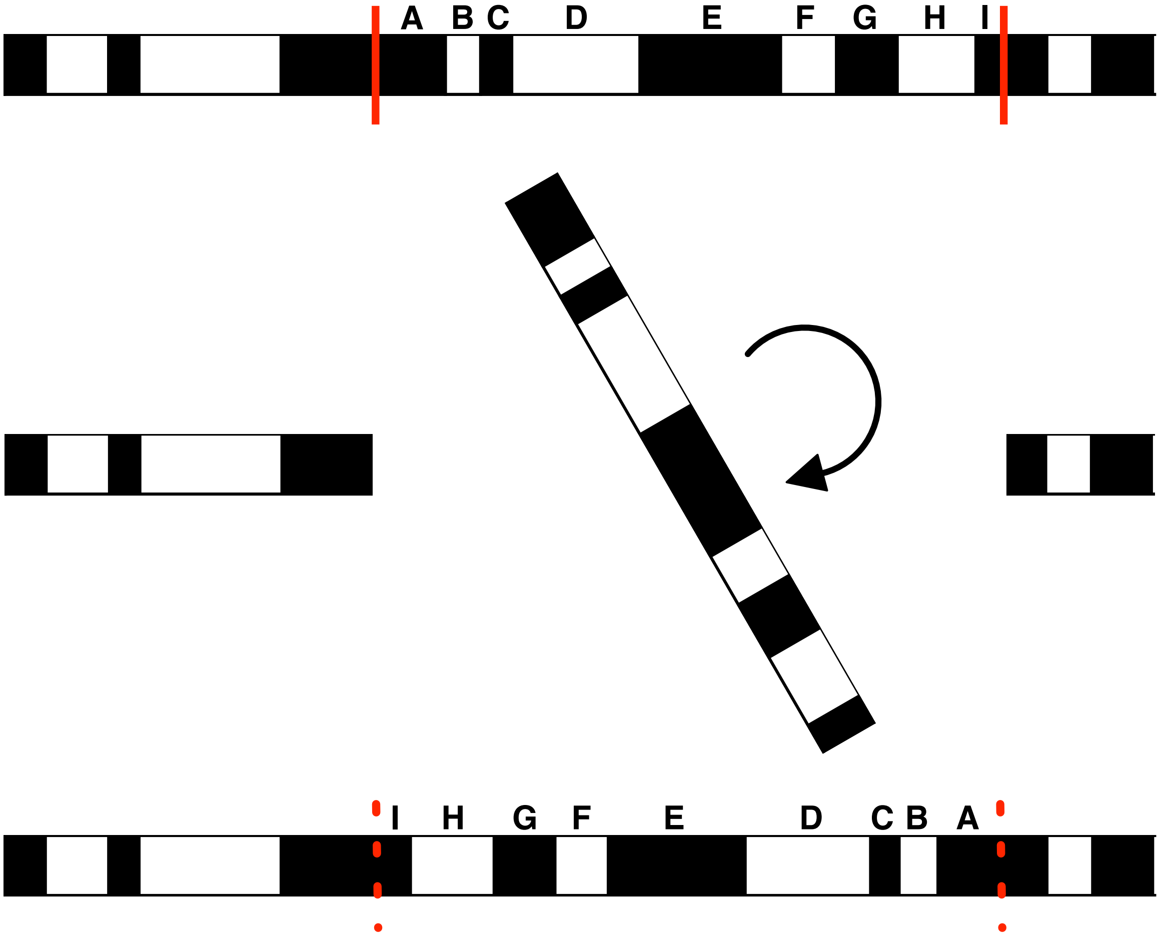

Fig. 2: A simplified visualization of how chromosome inversions may be inferred from banding patterns. The bottom chromosome may be produced from the top chromosome by inverting the detached interval. Dobzhansky and Sturtevant noticed this fact by comparing the widths of different bands; our simplified example shows how 9 bands (labeled A-I) have been inverted.

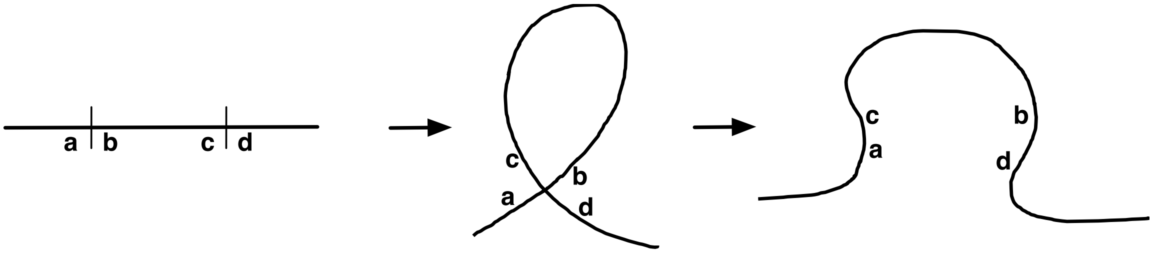

Fig. 3: Simple illustration of how a reversal may occur when a chromosome becomes coiled on itself and forms new bonds. Notice that **a** and **b** were adjacent in the original chromosome, as were **c** and **d**, and that the reversal has created new adjacencies **a** and **c** as well as **b** and **d**.

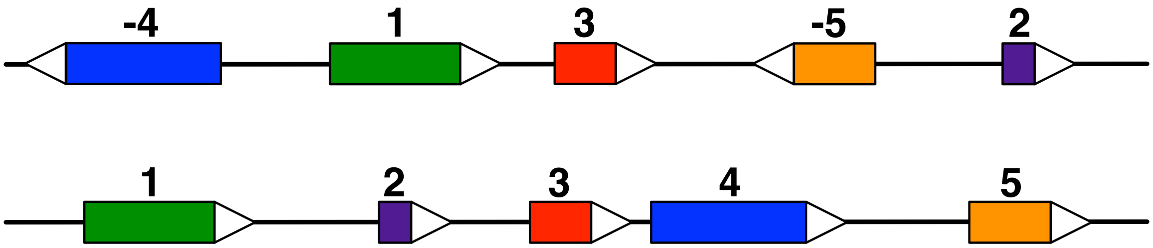

Fig. 4: Two simplified hypothetical chromosomes C_1 (top) and C_2 (bottom) containing the same genes. We may number the genes of C_2 in order, which yields the signed permutation (1, 2, 3, 4, 5); the gene ordering of C_1 therefore implies the signed permutation (-4, 1, 3, -5, 2).

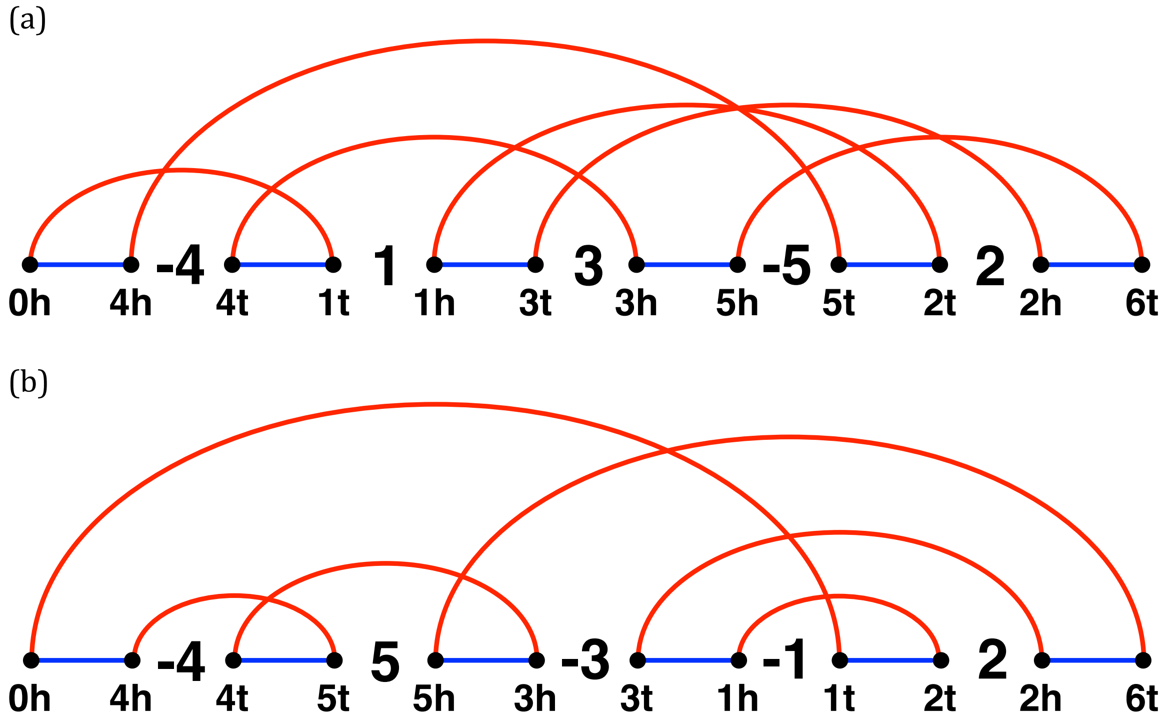

Fig. 5: (a) The breakpoint graph of chromosomes C_1 and C_2 from Fig. 4. (b) Applying a reversal on the interval containing the middle three indices of C_2 gives a new chromosome whose breakpoint graph with C_1 is depicted. Note that the only change in the edges is that blue edges {4t, 1t} and {5t, 2t} have been replaced by {4t, 5t} and {1t, 2t}.