ALGORITHMS FOR CONSTRUCTING COMPARATIVE MAPS

Debra S. Goldberg

Susan McCouch

Jon Kleinberg

Comparative maps are a powerful tool for aggregating genetic information about related organisms, for inferring phylogenetic relationships, and for examining hypotheses about the evolution of gene families and the functional significance of orthologous genes. Construction of any genetic map is laborious, but compiling comparative maps across multiple species requires a large investment of manual effort on the part of biologists. In this paper we present efficient algorithms that help in automating this effort and offer an explicit set of principles on which to base the construction of such maps. We compare the results of three approaches: manual expert analysis, a simple linear algorithm, and a more complex stack-based algorithm. All three methods produce remarkably similar results, with the stack algorithm more closely approximating the manual expert analysis.

1 Background

Comparative mapping. Comparative mapping is based on the principle that the order of homologous genes along the chromosomes of different eukaryotic species is often conserved both within and between families. Conservation of gene order (synteny) in defined chromosomal regions of different species suggests that those chromosomal segments are likely to be derived from a common ancestral linkage group (i.e. that the segments are homeologous). The construction of comparative maps between the genomes of different species using a common set of gene probes allows us to exploit the collective research accumulated for each of the species under consideration, and it suggests hypotheses about the origin and phylogenetic relationships among species as well as fundamental principles of genome evolution.

Comparative maps allow us to use structural and functional information about one genome to make predictions about another genome. In particular, these maps aid in the efficient localization of genes of interest and contribute to the isolation and characterization of those genes. The isolation of genes from organisms with large genomes and/or low or uneven levels of recombination along the chromosomes may be facilitated by first isolating a homologous gene in a well characterized genome. In particular, for plant species, we would like to localize, clone, and characterize genes controlling functions such as predisposition to diseases, crop yield potential, nutritional quality, and response to environmental stress, including pollutants or toxins. Comparative maps also offer opportunities to gain new insights into the evolution of multi-gene families, the interaction of specific genes in complex metabolic and physiological pathways, and the distinctive nature of certain developmental patterns and adaptations that have occurred in individual taxa over the course of evolution.

Rice and Maize. The rice genome provides a good basis for comparative mapping efforts because it has a small, diploid genome (approximately 430 Mb [2]) with relatively few internal duplications, well-developed genetic maps (containing about 3000 RFLP, SSLP, and morphological markers), a physical map covering more than 80% of the genome, and an international initiative to sequence the entire genome [20, 36]. The maize genome offers a marked contrast to rice, with a genome size six times larger (approximately 2600 Mb [2], or nearly the same size as the human genome), a high proportion (60-80%) of rapidly evolving repetitive DNA, and a recent polyploidization event (followed by a subsequent return to disomic inheritance) resulting in global gene duplication. Because of its genome complexity, there is little chance that maize will be fully sequenced soon, and it stands to benefit from comparative studies with maps of other grass species having less complex genomes, such as rice. The numerous, well-characterized mutants available in maize, in addition to its well-developed molecular map, are invaluable for the characterization of plant gene function and plant development, so rice may also benefit from rice-maize comparative map studies.

Based on the use of homologous gene probes (cDNA markers which hybridize to both the rice and maize genomes (see [1, 51])), conserved linkage segments can be defined in rice and maize. Most of these homeologous segments are duplicated within the maize genome, reflecting the polyploidization event that distinguishes the maize lineage. The extensive segmental similarity of the genomes coupled with the complexity offered by the recent polyploidization event make the rice-maize comparative map an excellent starting point for the development of a robust algorithmic model that can handle the complications that arise in comparative mapping studies.

Related work on comparative mapping and genome comparison. In order to construct a comparative map, one must first decide how homeologous portions of different genomes will be identified and defined. This can be done in several fundamentally different ways. One way is to hybridize a portion of one genome (e.g. a chromosome) with the entirety of another genome, as can be accomplished via chromosome painting using FISH (see, for example, [3, 50]). Another possibility is to identify orthologous loci using conserved genes, RFLPs, or microsatellite markers which can be reciprocally mapped onto the genomes of multiple organisms. This has been done for many groups of plants and animals. Using both of these approaches, many comparative maps have been constructed by biologists; for example, see work on Solanaceae [25, 43], Brassica [23], grasses [1, 9, 12, 47, 51], rat-mouse-human [41], dog-human [6], cat-human [29], and mouse-human [10], as well as more general reviews of comparative studies of mammals [8, 15, 32, 33]. In all of these studies, it has become apparent that some lineages are marked by a slow rate of chromosomal evolution, with few breakpoints and minimal rearrangements while other lineages show evidence of a more rapid rate of evolutionary change, marked by global reorganization and consequently, much smaller regions of synteny. In the case of prokaryotes, Koonin and Galperin comment that although protein sequences are frequently conserved, there is little conservation of genome organization [21]. This illustrates that very different mechanisms of genome evolution operate in eukaryotic and prokaryotic genomes.

As the resolution of comparative mapping improves, smaller regions of synteny can be increasingly recognized. Increased marker density makes it possible to document chromosomal relationships among rapidly evolving genomes and among ever more distantly related organisms. It also provides insights into the mechanisms that drive genome evolution in both rapidly changing and highly conserved genomes. Chromosomal evolution is marked by duplication, deletion, fusion, fission, translocation, transposition, and inversion events. In some eukaryotes, there can be extensive genome rearrangements [23, 42] that add significantly to the difficulty of constructing and interpreting comparative maps. With this in mind, we have attempted to construct algorithms that are general enough to be widely applicable in a variety of situations.

It is beneficial to consider the genomes of more than two organisms when building a comparative map. Although in this work we will develop algorithms that work on pairs of genomes, it is helpful to begin with a model system where there are many related comparative maps defined by experts. The grass family is indeed such a system. In addition to the maize-rice maps, there are also comparative maps for millet-rice [13], saccharum-sorghum [26], sorghum-maize [34], rice-wheat [22], rice-triticeae-oat [44], wheat against rice, maize and oat [45], and oat against wheat, rice, and maize [46]. There is also a low resolution map which puts grass species on a single circular comparative map [12, 27].

Work in the area of discrete algorithms has developed methods for inferring evolutionary "distance" between genomes, based on finding parsimonious sequences of genome rearrangements — such analysis typically asks, What is the minimum number of rearrangement events needed to convert one given genome into another? One can view this as a generalization of the standard *edit distance* used for sequence alignment [16, 49], viewed at the scale of whole genomes and allowing highly non-local transformations [37, 38]. For surveys on this approach, see Pevzner and Waterman [35], Hannenhalli [17], and Nadeau and Sankoff [31]. An algorithmic approach in a different spirit can be found in Sankoff, Ferretti, and Nadeau [39]; the authors describe rules for identifying corresponding regions in two genomes, without explicitly trying to minimize the length of a sequence of genome rearrangements. This is much more closely related to our approach here, and we compare our work with that of [39] below.

2 The Chromosome Labeling Problem

From a computational point of view, comparative mapping involves a spectrum of activities centered around modeling the evolutionary relationships between genomes of different species. When dealing with a pair of species, as we will do here, some of the basic concerns are

(i) identifying a large number of loci on the two genomes that can be put into correspondence, through markers that can be comparatively mapped to locations in both genomes;

(ii) inferring larger segments in the two genomes exhibiting extensive homeology; and

(iii) hypothesizing a sequence of evolutionary events by which the two species have diverged from a common ancestor.

Note that these activities build on each other in order; (ii) makes use of the data from (i), and higher-level results as in (iii) can be expressed in terms of the homeologous segments found in (ii).

This hierarchy of issues can be seen in the comparative analysis of rice and maize performed by Wilson et al. [51], the initial point of departure for our work here. In [51], a large collection of loci comparatively mapped between rice and maize was used to define a relatively small collection of homeologous segments; and it was from the structure of these segments that hypotheses could be made about the genomes of ancestral grass species. Consider Figure 2, a schematic representation of data collected by Wilson et al. for maize chromosomes 1 and 6.[1] In parentheses following each marker name is the rice linkage group in which a corresponding marker is found — a linkage group in the present context is specified by a rice chromosome number 1-12, and the symbol 'L' or 'S' to indicate the long or short arm of the chromosome. To the left of each list of markers is a *labeling* of the maize chromosome defined by Wilson et al.: it partitions (most of) each maize chromosome into a small number of segments defined by a preponderance of markers homologous to one arm of a single chromosome in rice, labeling this maize segment with the appropriate rice chromosome arm.

The following principles underly the construction of this labeling. First, it should represent a high-level *global view* of the relationships between the rice and maize chromosomes; in other words, there should be relatively few distinct segments in the labeling, so that a large volume of marker data can be distilled into a concise representation from which further hypotheses can be made at a global

[1] Our representation here differs from that of [51] in the following ways: we list only markers that were comparatively mapped in rice; we do not indicate distance between markers on the chromosomes; and for certain subsets of the markers whose relative orders could not be resolved (in other words, they were at distance 0 from each other), we have fixed a linear order according to the most statistically probable linkage relationships.

level. Second, consistent with providing a representation at a global scale of resolution, the labeling need not "explain" the presence of every marker. While maize contains, on average, twice as many copies of each locus as does rice (due to the global polyploidization event that gave rise to the modern maize lineage), local gene duplication and transposition events in both genomes have resulted in a situation where it is frequently impossible to distinguish which copy is the true ortholog. Thus, cDNAs mapping to positions that do not conform to predicted synteny relationships are allowed in a global labeling, and positively identified as small deviations. The rationale for this is that gene markers that are "out-of-place" at the level of resolution of the current map provide potentially valuable information as "seeds" for future, higher resolution mapping studies where additional information may provide the missing links necessary to identify new, smaller regions of homeology.

Here is a simple, if not entirely apposite, analogy for this approach to labeling. If we consider the partition of the earth's surface into continents and oceans (a type of labeling), the representation at a global scale of resolution may not seek to model the fact that small bodies of water (lakes) are contained in the large land masses, and small patches of land (islands) are contained in the large water masses. Such a high-level representation is valuable for reasoning at this scale; in order to consider finer scales, one must take these more detailed features into account.

The Present Work: A Computational Approach. The construction of global chromosome labelings as in Figure 2 has essentially been a manual process, performed by domain experts using underlying knowledge about the species being compared. In this work, we model chromosome labeling as a computational problem; we ask: Is there a simple algorithmic rule that can generate labelings similar to those built by hand in Figure 2? Such an algorithm would not only be useful in automating the process of constructing such labelings; it would also be useful for making explicit the assumptions that underly such labelings, so that we can reason about their consequences more directly. Further it would provide a rational basis for moving to higher-level comparisons in the future — for example, the comparison of different pairwise comparative maps to each other, particularly maps constructed independently by different research groups.

Note that while we wish to model chromosome labeling computationally, it is not *a priori* a precisely defined computational problem. Indeed, our goal will be to design an underlying model together with efficient algorithms for producing labelings. In this way, we can bring a formal problem definition to bear on issue (ii) in the hierarchy of problems above, much the way that the formalization of genome rearrangement problems brought a mathematical concreteness to issue (iii). Guided by the motivation above, we favor simpler models and algorithms, with few tunable parameters, as these impart a greater conceptual robustness to the labelings obtained.

To discuss these issues more concretely, we settle on the following general terminology. We begin with two genomes, the *base* and the *target*. We wish to *label* segments of the target using names of linkage groups from the base. In our case, maize is the target, and the labels will be the chromosome arms of rice. (Thus, we consider a set of 24 labels: the long and short arm of each of the 12 rice chromosomes.) A consequence of using this coarse-grained set of labels is that we do not address the question of whether relative order has been preserved in the segments being labeled. This is based on the assumption that a significant cluster of markers in the target genome which all belong to the same linkage group in the base genome provides strong evidence of significant synteny. Our approach does not seek to identify inversions or other intrachromosomal rearrangements, leaving this to a more fine-grained level of resolution.

A simplification in our model is that the markers on the target chromosome are assumed to be fully and correctly ordered. This will not be strictly the case in practice, since it may be impossible to distinguish the order of nearby markers with a finite mapping population, and such markers will be mapped to the same location. For other markers, the order inferred through mapping experiments may be in error. Nevertheless, it is possible for us to obtain an order that represents a good approximation to a correct, total ordering; we leave more detailed concerns about this issue for future work.

The construction of chromosome labelings is a natural setting in which to formalize the trade-off between parsimony and accuracy. We seek to partition each chromosome in the target genome into a sequence of contiguous *segments*, each with a given label. We seek to do this in a way that minimizes a *penalty function* consisting of the following two types of terms:

(a) A penalty that increases with the number of segments we use in the partition. (A larger number of segments constitutes a less parsimonious labeling.)

(b) A penalty for each marker that does not belong to the linkage group used to label its segment. (Such "out-of-place" markers are not well explained by the labeling.)

In effect, such a model seeks to interpolate between the following two extremes: a labeling consisting of a single segment (which minimizes penalties of type (a), but incurs a lot of penalty of type (b)), and a labeling in which each marker belongs to its own segment (which can minimize penalty terms (b), but incurs a large penalty of type (a)). Moreover, our basic models will turn out to have a single parameter, essentially the relative values of the penalty terms of types (a) and (b).

We introduce the models and algorithms formally in the next two sections. We begin with a simple *linear* model that can be viewed as a type of hidden Markov model. (See e.g. [14] for an overview of hidden Markov models and some of their applications.) This approach turns out to have some shortcomings — in effect, it is too "local" in its behavior — and we modify it to a *stack-based* model, in which the penalty terms of type (a) are derived from a stack-like relation among the segments in the partition. We find that our stack model produces labelings that closely correspond to those of Wilson et al.; the reader can see an example of this for maize chromosomes 1 and 6 in Figure 3, and we discuss the comparison among the methods in the final sections of the paper. This transition from local Markov models to stack-based models for the purpose of capturing long-range dependencies has a long history of analogues in the study of programming languages and natural language [7, 19]

Thus we see that the chromosome labeling problem addresses issues that lie naturally between the low-level identification of corresponding loci between genomes, and the bulk of the algorithmic work on finding short sequences of genome rearrangements to explain evolutionary divergence. The work of Sankoff, Ferretti, and Nadeau on conserved segment identification [39] can be viewed as proceeding from similar motivation, and addressing a similar type of issue in comparative mapping: given a pair of genomes, they wish to find corresponding pairs of *conserved regions* that show a high degree of synteny. (See also [30].) There are several fundamental differences between their work and ours. First, they seek a model that "explains" the presence of every marker; to keep the number of regions small despite this, they allow for regions to be non-contiguous. Second, while they formulate their problem in terms of a penalty function to be minimized, their function is more complex: it contains three tunable parameters (capturing the extents to which each region is short, dense, and not interrupted by other regions), and it is not known how to efficiently find the optimal partition under this function. In contrast, the objective functions underlying our models can be solved to optimality by efficient algorithms.

3 The Linear Model

Our most basic model is a direct adaption of the principles discussed above. We fix a chromosome in the target genome, and let $M = \langle 1, 2, \ldots, n \rangle$ denote the sequence of markers in order on this chromosome. For each marker i, we assume it has been comparatively mapped to a single linkage group ℓ_i in the base genome. (Markers that have not been comparatively mapped in the base genome are not informative for our purposes; below, we will mention an extension to markers with more than one associated linkage group in the base genome.) The label set L consists of all linkage groups in the base genome; let k denote the number of labels in L. We define a simple comparison function $\delta(\cdot, \cdot)$ on pairs of labels as follows: $\delta(a, b) = 0$ if $a = b$; and $\delta(a, b) = 1$ if $a \neq b$.

A *labeling* of the chromosome is a function $f : M \to L$; in other words, it assigns a label to each marker. We encode penalties of types (a) and (b) from the previous section as follows. For a constant s, we impose a penalty of s for each consecutive pair of markers i and $i+1$ such that $f(i) \neq f(i+1)$; this is a boundary between adjacent segments, and we are charged a *segment opening penalty* of s for introducing the new segment. For a constant m, we impose a penalty of m for each marker i such that $f(i) \neq \ell_i$; this is a marker that is not "explained" by the labeling f. The sum of all these penalties defines the objective function; formally, we can write it as

$$Q(f) = s\left(|\{i : f(i) \neq f(i+1)\}|\right) + m\left(|\{i : f(i) \neq \ell_i\}|\right).$$

For our objective function to yield meaningful labelings, we must have $0 < m < s$; indeed, we may assume with no loss of generality that $m = 1$, so that s (or, more properly, the ratio s/m) is the single parameter of the model.

We now describe an efficient algorithm, based on dynamic programming, that computes a labeling f of minimum total penalty. For any value of i between 1 and n, and any $a \in L$, we let $S[i, a]$ denote the optimal (minimum) penalty of a labeling of the prefix of M of length i which ends in label a, and let f^*_{ia} be such an optimal labeling. Now, let f' denote the labeling of the first $i-1$ markers in f^*_{ia}. f^*_{ia} (and hence f') assigns some label b to marker $i-1$, where possibly $b = a$; if f' does not have penalty $S[i-1, b]$, we could replace it with a better labeling of the first $i-1$ markers ending in b, resulting in a labeling better than f^*_{ia}. But this is not possible, so f' achieves the penalty $S[i-1, b]$.

This justifies the following recurrence relation.

$$S[i, a] = m \cdot \delta(\ell_i, a) \quad + \quad \min_{b \in L} (\ S[i-1, b] \quad + \quad s \cdot \delta(b, a)\ )$$

Beginning with the initialization $S[0, a] = 0$ for each label a, we can build up the values $S[i, a]$ in order of increasing i.

We can then determine an optimal labeling for all of M: it is one that achieves the minimum value of $S[n, a]$, over all labels $a \in L$. The recurrence takes $O(k)$ time to invoke for each value of $S[\cdot, \cdot]$; and there are kn such values to compute. Thus the total running time is $O(k^2 n)$. Since we view the label set as having fixed constant size, this is a running time linear in the number of markers.

4 The Stack Model

We now describe a more sophisticated model that provides labelings on rice-maize data closer to that of Wilson et al. [51]; it is designed to take into account certain long-range correlations in the sequence, and in the process corrects some counter-intuitive behavior exhibited by the linear model.

To begin with, we consider an informative example, an instance of the labeling problem in which $s = 2t$, for a number t, and M is a sequence of markers of length $9t$. For three distinct labels a, b, and c, the first $3t$ markers in M have $\ell_i = a$; the next $3t$ have $\ell_i = b$; and the final $3t$ have $\ell_i = c$. Then one can check that the unique optimal solution under the linear model is the obvious labeling that produces three segments labeled **a**, **b**, and **c**. Now consider the same instance, except that the final $3t$ markers have $\ell_i = a$. In this case, the unique optimal solution under the linear model is a single segment labeled **a**. The point is that in this latter instance, changing to a segment labeled **b** and then back to one labeled **a** would cost $2s = 4t$, and so it is worth paying for $3t$ out-of-place markers in the middle in order to have a single segment labeled **a**. In the first instance, on the other hand, there was still a cost of $2s = 4t$ for two new segments; but there, the alternative was to pay $6t$ for out-of-place markers.

Thus, somewhat surprisingly, the linear model treats labelings of the form **a-b-c** and **a-b-a** differently; to capture the types of analysis described by Wilson et al., we would like a model that treats such labelings comparably. Intuitively, this requires a way to handle long-range correlations in a labeling more accurately.

To accomplish this, we add a push-down stack to the model, where segments we wish to remember are saved in a last-in-first-out (LIFO) manner. Thus, at all times there will not just be a current segment label, but also an auxiliary stack of labels that have been seen earlier in the labeling. There will now be several ways to switch from a current label **a** to a new one: (i) we can *replace* **a** with a new label **b**, as in the linear model; (ii) we can *push* a new label **b** on top of **a**, so that **a** will be saved beneath **b**; or (iii) we can *pop* **a** off the top of the stack, revealing whichever label is lying just below **a**. The key point is that while operations (i) and (ii) incur the usual segment opening penalty s, the *pop* operation (iii) will incur zero cost.

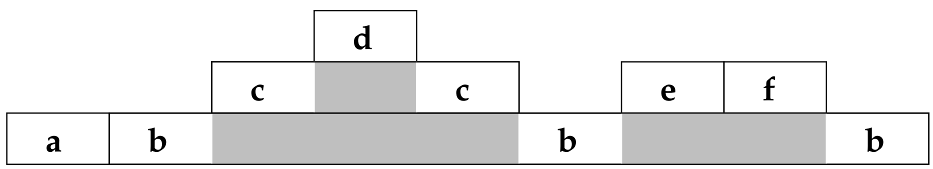

Figure 1: Intuitive notion of stacking segments.

As an example of this process, consider Figure 1. We see that the segment labeled **c** has been pushed on top of the segment labeled **b**, and that the segment labeled **d** has been pushed on top of that. Then **d** is popped off, leaving **c** visible, and then **c** is popped off, leaving **b** visible. Subsequently **e** is pushed on top of the segment labeled **b**, **f** replaces **e** linearly, and then is popped off once again leaving label **b** visible. The labeling described in this figure is **a-b-c-d-c-b-e-f-b**.

This stack model suggests a correspondence with the genome rearrangement process of insertion. When a segment is pushed on top of another segment, it effectively splits the existing segment into two pieces, one on either side of the new segment, as if the new segment is inserting itself into the old. While this is a suggestive and potentially useful connection, we do not mean to overemphasize the role of insertions in the construction of comparative maps; rather, the stack model is fundamentally designed for algorithmic reasons, to improve on the behavior of the linear model, and to better capture the type of analysis in Wilson et al.'s notion of labeling.

An Optimal Algorithm. We thus seek a labeling f that minimizes a penalty function composed of the following terms: a cost of m for each marker i such that $f(i) \neq \ell_i$; a cost of s for each segment that is pushed onto the stack of current labels; and a cost of s for each segment that replaces a previous segment as in the linear model; each *pop* will have no cost.

We can find such a labeling of minimum cost, together with the associated sequence of pushes, pops, and replacements, by an efficient algorithm based on dynamic programming. The algorithm will be similar in spirit to dynamic programming algorithms for parsing strings with respect to a context-free grammar [19], and for predicting RNA secondary structures [14]. This is natural, since both of these latter applications involve using stack structures to model long-range dependencies; at the same time, our algorithm exhibits some differences at a technical level.

We keep much of the same notation as in the previous section. For any values i and j such that $1 \leq i \leq j \leq n$, and any $a \in L$, let $M[i,j]$ denote the subsequence of M which starts at position i and ends at position j. We let $S[i,j,a]$ denote the optimal (minimum) cost of a labeling f of $M[i,j]$ with the property that $f(i) = a$. Let f^*_{ija} be such a labeling with optimal cost $S[i,j,a]$.

Clearly $S[i,i,a] = m \cdot \delta(\ell_i, a)$. For $j > i$, we claim that $S[\cdot,\cdot,\cdot]$ satisfies the following recurrence relation.

$$S[i,j,a] = \min \left(\begin{array}{l} \min\limits_{b \in L} \ S[i{+}1,j,b] \ + m \cdot \delta(\ell_i, a) \ + s \cdot \delta(b, a) \ , \\ \min\limits_{i<k<j} \ S[i,k,a] + S[k{+}1,j,a] \end{array} \right)$$

We can prove this as follows. Consider the label b assigned to marker $i+1$ by f^*_{ija}. If $b = a$, then we can apply the first line of the recurrence. If $b \neq a$, and the stack in f^*_{ija} never returns to the current copy of a in $M[i{+}1,j]$, we can again apply the first line of the recurrence. Finally, suppose $b \neq a$, and the stack in f^*_{ija} returns to the current copy of a for at least one marker in $M[i{+}1,j]$. Then $j > i+1$; let $k+1 > i+1$ be the minimum index at which the stack returns to the current copy of a. The transition from the label at k to the label a at $k+1$ has zero cost, since it is achieved by a *pop*; thus we can apply the second line of the recurrence.

Using this recurrence relation, we can build up all values of $S[i,j,a]$ iteratively as follows. We initialize $S[i,i,a] = m \cdot \delta(\ell_i, a)$ for all i and all $a \in L$. We then compute all $S[i,j,a]$ using the recurrence in order of increasing $j - i$. Finally, an optimal labeling is one that achieves the minimum value of $S[1,n,a]$, over all labels $a \in L$.

Each invocation of the recurrence relation involves the examination of $O(k+n)$ quantities, and takes $O(k+n)$ time; since we assume $k \leq n$, this can be written as $O(n)$. There are $O(kn^2)$ values $S[i,j,a]$ that must be computed, so the total running time is $O(kn^3)$, or cubic in the number of markers.

Extensions. In order to have our model more closely match the results achieved by the biologists, and to better model actual biological data, several small extensions were made to the algorithm described above.

Although in eukaryotes the two arms of a chromosome are considered to be different linkage groups, they are connected. Beginning a segment labeled with the other arm of the previously labeled segment should therefore be cheaper than beginning an unrelated segment. To accomplish this without adding new parameters, we modified the algorithm so that beginning a related segment (i.e. the opposite arm of the segment we are replacing or popping) costs $s/2$, half as much as beginning an unrelated segment. Also, for markers comparatively mapped to a centromeric region of a rice chromosome, or to an unknown arm of the chromosome, we imposed a penalty of $m/2$ to label them with either arm of the appropriate chromosome. For markers that were comparatively mapped to multiple locations in rice, we allowed the algorithm to choose the better location in computing an optimal labeling.

We made a few modifications to further favor labelings with fewer segments. In cases where there were multiple optima, we reported a primary labeling based on a tie-breaking rule in which segments were extended for as long as possible. Also, we increased the cost for a *pop* operation from 0 to a very small positive quantity $\varepsilon > 0$; this causes the algorithm to favor labelings that do not perform *pop*'s when they do not strictly improve the objective function.

5 Results and Discussion

The Wilson map was the initial yardstick with which the success of the algorithms was measured. We have since undertaken preliminary tests on data from other species. It is important to note that the Wilson map was constructed as a single map aligning the entire maize genome with the rice genome, and information from one part of the maize genome was leveraged against decisions for other portions of the genome. In contrast, the maps produced by our algorithms are constructed separately for each maize chromosome.

The stack-based algorithm (including the extensions discussed above) was run on the data set used by Wilson et al. [51], with the following modifications. The Wilson data set is not fully ordered, since several markers may be mapped to the same location (due to the limited number of recombination events available for interpreting order in the small mapping populations used in these studies), and other markers are mapped with lower confidence (such that they can be positioned within an interval, but not to a predicted point on the map). Because our algorithms require that the markers be presented in a linear order, markers were ordered according to the most statistically probable linkage relationships. Our algorithms also require that all markers under consideration in the target genome (in this case, maize) are labeled with a corresponding putative region of homeology in the base genome (rice). The manual expert analysis did not impose this requirement, and in fact there are portions of the maize genome which, at the current level of resolution, show no synteny corresponding to any segment in rice. Ignoring this, an evaluation of the comparative maps generated by the stack algorithm side-by-side with the comparative maps in [51] show few significant differences for low values of the segment opening penalty s.

In two of the ten maize chromosomes (maize 6 and maize 7), all markers included in a syntenic segment in [51] were included in the identically-labeled segment by our algorithm. The results of the stack model for maize chromosome 6 are shown in Figure 3. The linear model produced only two segments for this chromosome, labeled 6S and 5L; the fact that it did not produce a segment labeled 6L is a direct consequence of the linear model's difficulty in handling labelings of the form **a-b-a**, as discussed above.

The differences in the other eight chromosomal maps were of several types. Biologists consider the short arms of rice chromosomes 11 and 12 to be largely syntenic to each other, indicative of a duplication in the rice lineage. As such, tracts of maize markers that are homologous to similarly ordered markers on rice 11S are considered to give evidence for synteny with both rice chromosomes 11S and 12S, and vice versa. This complication has not been incorporated into our model, accounting for differences in the constructed maps for maize chromosomes 3 and 10.

Our algorithms did not detect small homeologous segments identified in the Wilson maps on maize chromosomes 2, 3, and 9 because there were not more than 2 maize markers providing evidence of synteny. In constructing the Wilson maps, genomes from other members of the grass family were taken into consideration, as well as other domain-specific knowledge to corroborate these inferences about synteny; but this kind of information was not available to our algorithms.

Our algorithm performed *pop* operations on maize chromosomes 4, 5, and 8 for the purpose of matching a single additional homologous marker. It is unclear if these differences represent possible improvements to the Wilson map (see discussion for chromosome 1 below for an example of how this is possible) or not.

In each of maize chromosomes 1 and 5 the stack algorithm produced a syntenic segment which was not included in [51], but which looks suggestive, and will be investigated further by biologists. The results for maize chromosome 1 are shown in Figure 3. This same map was generated with the segment opening penalty s set at 2, 3, or 4. The addition of the first segment labeled **10L** was caused by the ability to pop a label, since otherwise a segment with just two markers would not be created, even with $s = 2$ due to our tie-breaking rule. This alternate comparative map suggests that the segment syntenic to **8L**, which is determined to have been inserted into a **3S-10L-3L** composite chromosome at the boundary between the **3S** and **10L** segments in [51], was in fact inserted a short distance away from this boundary in the middle of the **10L** segment. The linear model did not find this segment, instead producing the same result as the Wilson map.

The output comparative maps for the stack model which were most similar to the Wilson maps as described above were obtained with the segment opening penalty as shown in Table 1. Where multiple penalty values are shown, they all produce the same labeling.

6 Further Directions

We are currently pursuing further extensions of the algorithms described here. One direction is to incorporate a whole-genome perspective for optimizing the labeling. We are also investigating richer frameworks for labeling that more fully integrate ordering and distance information among markers.

Maize chromosome	segment opening penalty (s)
1	2,3,4
2	2,3
3	2,3,4,5,6,7
4	2,3
5	2*
6	2,3,4,5
7	2,3,4,5,6
8	2,3,4,5,6,7
9	2,3,4
10	2,3

* $s = 3$ produces map more like Wilson map without the suggestive new segment

Table 1: Values of segment opening penalty s for which the resulting labeling approximates the Wilson map.

We have begun preliminary investigations of the performance of the algorithms on data from other species, including mouse-human data obtained from the Web site of The Jackson Laboratory [28]. Our analysis indicates that, given their efficiency, the algorithms described here will be able to scale up to input sizes significantly larger than what we have dealt with in the rice-maize comparison. In addition, because the algorithms are based on a general model, involving a small number of clearly delineated assumptions, they are applicable to a range of settings, and a variety of datasets.

Acknowledgments. The authors are grateful for the help of Sandra Harrington, who provided the rice-maize data used in our analysis. We also thank Sam Cartinhour and David Schneider for many discussions while formulating the problem.

The work of the first author was supported in part by NSF Training Grant DEB-9602229 and by the Packard Foundation Fellowship of the third author. The work of the second author was supported in part by USDA National Research Initiative grant 94-37310-0661 and Cooperative State Research Education and Extension Service NYC 149-401. The work of the third author was supported in part by a David and Lucile Packard Foundation Fellowship, an Alfred P. Sloan Research Fellowship, an ONR Young Investigator Award, and NSF Faculty Early Career Development Award CCR-9701399.

References

[1] SN Ahn and SD Tanksley. Comparative linkage maps of the rice and maize genomes. *Proc. Natl. Acad. Sci. USA* 90:7980-7984, 1993.

[2] K Argumuganathan, ED Earle. Nuclear DNA content of some important plant species. *Plant Mol Biol Rep* 9:208-218, 1991.

[3] A Auch, J Wienberg, R Stanyon, N Arnold, S Tofanelli, T Ishida, T Cremer. Reconstruction of genomic rearrangements in great apes and gibbons by chromosome painting. *Proc. Natl. Acad. Sci. USA* 90:7980-7984, 1993. 89: 8611-8615, 1992.

[4] AK Bansal. An automated comparative analysis of 17 complete microbial genomes. *Bioinformatics* 15(11):900-908 1999.

[5] M Bonierbale, R Plaisted, and SD Tanksley. RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. *Genetics* 120:1095-1103, 1988.

[6] M Breen, R Thomas, MM Binns, NP Carter, and CF Langford. Reciprocal chromosome painting reveals detailed regions of conserved synteny between the karyotypes of domestic dog (Canis familiaris) and human. *Genomics* 61:145-155, 1999.

[7] N Chomsky. Three models for the description of language. *IRE Transactions on Information Theory* 2(3):113-124, 1956.

[8] BP Chowdhary, T Raudsepp, L Fronicke, H Scherthan. Emerging patterns of comparative genome organization in some mammalian species as revealed by Zoo-FISH. *Genome Research* 8: (6) 577-589, 1998.

[9] GL Davis, MD McMullen, C Baysdorfer, T Musket, D Grant, M Staebell, G Xu, M Polacco, L Koster, S Melia-Hancock, K Houchins, S Chao, EH Coe Jr. A Maize Map Standard With Sequenced Core Markers, Grass Genome Reference Points and 932 Expressed Sequence Tagged Sites (ESTs) in a 1736-Locus Map. *Genetics* 152:1137-1172, 1999.

[10] RW DeBry and MF Seldin. Human/mouse homology relationships. *Genomics* 33(3):337-351, 1996.

[11] L Delcher, S Kasif, RD Fleischmann, J Peterson, O White and SL Salzberg. Alignment of whole genomes. *Nucleic Acids Research* 27(11):2369-2376, 1999.

[12] KM Devos and MD Gale. Comparative genetics in the grasses. *Plant Mol. Biol.* 35:3-15, 1997.

[13] KM Devos, ZM Wang, J Beales, T Sasaki, MD Gale. Comparative genetic maps of foxtail millet (Setaria italica) and rice (Oryza sativa). *Theor. Appl. Genet* 96:63-68, 1998.

[14] R Durbin, S Eddy, A Krogh, G Mitchison. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids.* Cambridge University Press, 1997.

[15] J Gellin, S Brown, JAM Graves, M Rothschild, L Schook, J Womack, M Yerle. Comparative gene mapping workshop: progress in agriculturally important animals. *Mammalian Genome* Vol 11, Iss 2, pp 140-144, 2000.

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

[17] S Hannenhalli. *Transforming Mice into Men (A Computational Theory of Genome Rearrangements).* Ph.D. thesis, Pennsylvania State University, 1995.

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

[19] JE Hopcroft and JD Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, 1979.

[20] An Intenational Collaboration to Sequence the Rice Genome. http://demeter.bio.bnl.gov/rice.html.

[21] EV Koonin and MY Galperin. Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Current Opinion in Genetics & Development* 7:757-763, 1997.

[22] N Kurata, G Moore, Y Nagamura, T Foote, M Yano, Y Minobe, M Gale Conservation of genome structure between rice and wheat. *Bio/Technology* 12:276-278, 1994.

[23] U Lagercrantz. Comparative Mapping Between *Arabidopsis thaliana* and *Brassica nigra* indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. *Genetics* 150:1217-1228, 1998.

[24] S Liu and KE Sanderson. Highly plastic chromosomal organization in *Salmonella typhi* *Proc. Natl. Acad. Sci. USA*, 93:10303-10308, 1996.

[25] KD Livingstone, VK Lackney, JR Blauth, R van Wijk, MK Jahn. Genome Mapping in Capsicum and the Evolution of Genome Structure in the Solanaceae. *Genetics* 152:1183-1202 1999.

[26] R Ming, SC Liu, YR Lin, J daSilva, W Wilson, D Braga, A vanDeynze, TF Wenslaff, KK Wu, PH Moore, W Burnquist, ME Sorrells, JE Irvine, AH Paterson . Detailed Alignment of Saccharum and Sorghum Chromosomes: Comparative Organization of Closely Related Diploid and Polyploid Genomes. *Genetics* 150:1663-1682, 1998.

[27] G Moore, KM Devos, Z Wang, and MD Gale. Grasses, line up and form a circle. *Curr. Biol.* 5:737-739, 1995.

[28] Mouse Genome Database (MGD), Mouse Genome Informatics Web Site, The Jackson Laboratory, Bar Harbor, Maine. World Wide Web (URL: http://www.informatics.jax.org/). (March, 2000).

[29] WJ Murphy, M Menotti-Raymond, LA Lyons, MA Thompson, and SJ O'Brien. Development of a feline whole genome radiation hybrid panel and comparative mapping of human chromosome 12 and 22 loci. *Genomics* 57:1-8, 1998.

[30] J Nadeau and B Taylor. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. USA*, 81(1984), pp. 814-818.

[31] J Nadeau and D Sankoff. Counting on comparative maps. *Trends in Genetics*, 14(12):495-501, 1998.

[32] SJ O'Brien, JE Womack, LA Lyons, KJ Moore, NA Jenkins, and NG Copeland. Anchored reference loci for comparative genome mapping in mammals. *Nature Genetics* 3:103, 1998.

[33] SJ O'Brien, M Menotti-Raymond, WJ Murphy, WG Nash, J Wienberg, R Stanyon, NG Copeland, NA Jenkins, JE Womack, and JA Marshall Graves. The promise of comparative genomics in mammals. *Science* 286:458-481, 1999.

[34] MG Pereira, M Lee, P Bramel-Cox, W Woodman, J Doebley, R Whitkus Construction of an RFLP map in sorghum and comparative mapping in maize. *Genome* 37:236-243, 1994.

[35] PA Pevzner and MS Waterman, "Open Problems in Computational Molecular Biology," *Proc. Israel Symposium on Theory of Computing and Systems*, 1995, pp. 158–173.

[36] Rice Genome Research Program. http://www.dna.affrc.go.jp:82/.

[37] D Sankoff. Edit distance for genome comparison based on non-local operations. In *3rd Annual Symposium on Combinatorial Pattern Matching*, pages 121–135, 1992.

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

[39] D Sankoff, V Ferretti, and JH Nadeau. Conserved Segment Identification. *Journal of Computational Biology* 4(4):559-565, 1997.

[40] D Sankoff. Genome rearrangement with gene families. *Bioinformatics* 15(11):909-917, 1999.

[41] T Serikawa, ZH Cui, N Yokoi, T Kuramoto, Y Kondo, K Kitada, JL Guenet. A comparative genetic map of rat, mouse and human genomes. *Experimental Animals* 47(1):1-9, 1998.

[42] SD Tanksley, R Bernatzky, NL Lapitan, JP Prince. Conservation of gene repertoire but not gene order in pepper and tomato. *Proc. Natl. Acad. Sci. USA*, 85:6419-6423, 1988.

[43] SD Tanksley, MW Ganal, JP Prince, MC de Vicente, MW Bonierbale, P Broun, TM Fulton, JJ Giovannoni, S Grandillo, GB Martin, R Messeguer, JC Miller, L Miller, AH Paterson, O Pineda, MS Roder, RA Wing, W Wu, and ND Young. High density molecular linkage maps of the tomato and potato genomes. *Genetics* 132: 1141-1160, 1992.

[44] AE VanDeynze, J Dubcovsky, KS Gill, JC Nelson, ME Sorrells, J Dvorak, BS Gill, ES Lagudah, SR McCouch, R Appels. Molecular-genetic maps for group 1 chromosomes of Triticeae species and their relation to chromosomes in rice and oat. *Genome* 38:45-59, (1995)

[45] AE VanDeynze, JC Nelson, ES Yglesias, SE Harrington, DP Braga, SR McCouch, ME Sorrells. Comparative mapping in grasses. Wheat relationships. *Mol. Gen.. Genet.* 248: 744-754, (1995)

[46] AE VanDeynze, JC Nelson, LS O'Donoughue, SN Ahn, W Siripoonwiwat, SE Harrington, ES Yglesias, DP Braga, SR McCouch, ME Sorrells. Comparative mapping in grasses. Oat relationships. *Mol. Gen.. Genet.* 249:349-356, (1995)

[47] AE VanDeynze, ME Sorrells, WD Park, NM Ayres, H Fu, SW Cartinhour, E Paul, SR McCouch. Anchor Probes for comparative mapping of grass genera. *Theor. Appl. Genet.*, 97:356-369, 1997.

[48] P Vincens, L Buffat, C André, J Chevrolat, J Boisvieux and S Hazout. A strategy for finding regions of similarity in complete genome sequences. *Bioinformatics* 14(8):715-725, 1998.

[49] MS Waterman. *Introduction to Computational Biology*. Chapman-Hall, 1995.

[50] J Wienberg, A Jauch, R Stanyon, T Cremer. Molecular cytotaxonomy of primates by chromosomal in situ suppression hybridization. *Genomics* 8: (2) 347-350, 1990.

[51] WA Wilson, SE Harrington, WL Woodman, M Lee, ME Sorrells, SR McCouch. Inferences on the genome structure of progenitor maize through comparative analysis of rice, maize, and the domesticated panicoids. *Genetics* 153: (1) 453-473, 1999.

Center for Applied Mathematics
Cornell University
Ithaca, NY 14853 USA
E-mail: debra@cam.cornell.edu

Department of Plant Breeding
Cornell University
Ithaca, NY 14853 USA
E-mail: srm4@cornell.edu

Department of Computer Science
Cornell University
Ithaca, NY 14853 USA
E-mail: kleinber@cs.cornell.edu

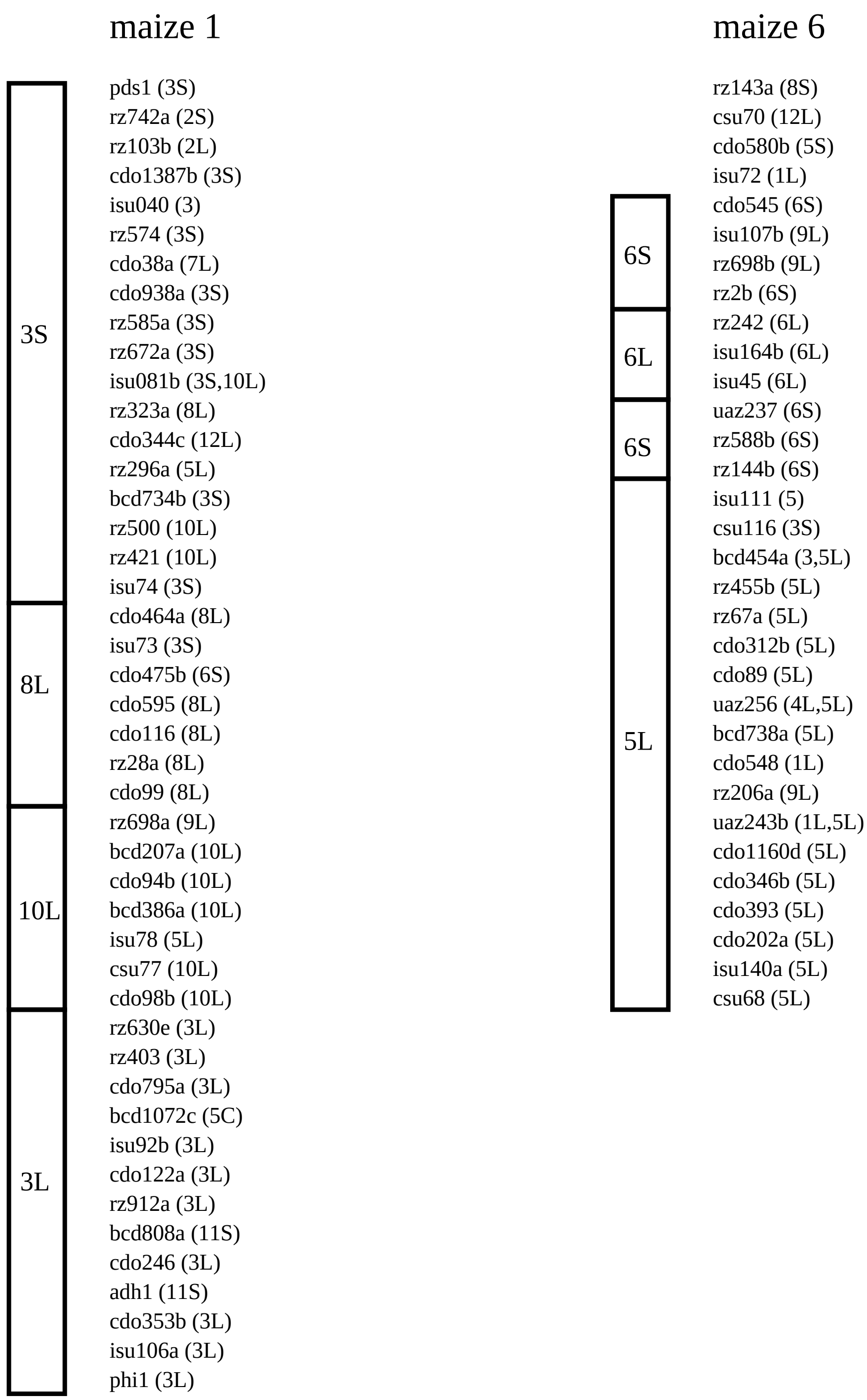

Figure 2: Markers comparatively mapped in rice for maize chromosomes 1 and 6.

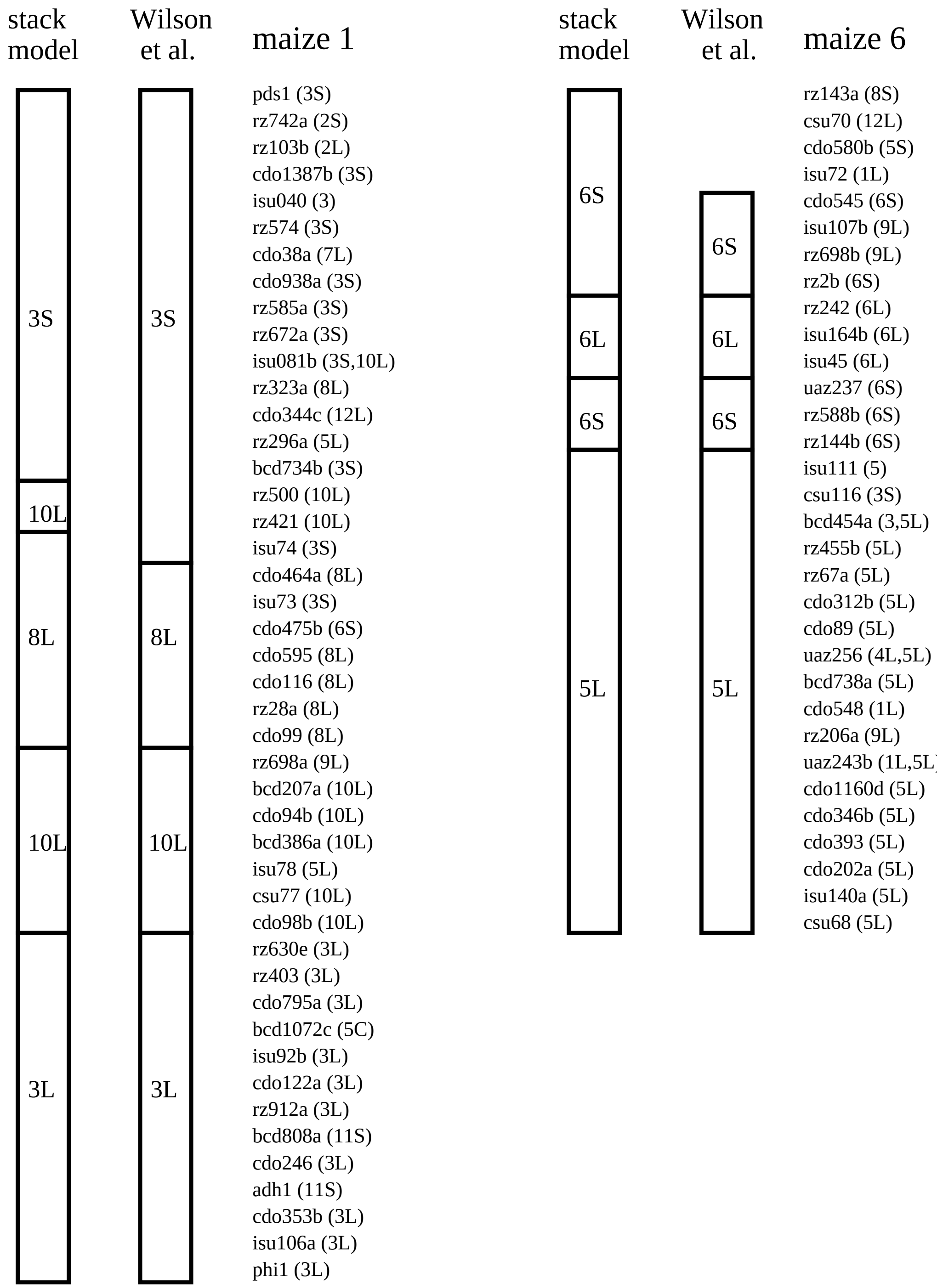

Figure 3: Results of manual and automated chromosome labeling.