Removing Noise and Ambiguities from Comparative Maps in Rearrangement Analysis

Chunfang Zheng, Qian Zhu, and David Sankoff

Abstract—Comparison of genomic maps is hampered by errors and ambiguities introduced by mapping technology, incorrectly resolved paralogy, small samples of markers, and extensive genome rearrangement. We design an analysis to remove or resolve most of these problems and to extract corrected data where markers occur in consecutive strips in both genomes. To do this, we introduce the notion of prestrip, an efficient way of generating these and a compatibility analysis culminating in a Maximum Weighted Clique (MWC) search. The output can be directly analyzed with genome rearrangement algorithms, allowing the restoration of some of the data not incorporated into the clique solution. We investigate the trade-off between criteria for discarding excessive prestrips to make MWC feasible in terms of retaining as many markers as possible in the solution and producing an economical rearrangement analysis. We explore these questions through simulation and through comparison of the rice and sorghum genomes.

Index Terms—Maximum Weight Clique, rice, sorghum, genome rearrangements, synteny blocks.

1 Introduction

The usual first step in comparative genomics is to decompose two genomes into synteny blocks, segments of chromosomes deemed to be homologous in the two genomes. The criteria for homology differ from study to study and allow for not only a degree of sequence divergence but also for some amount of insertion, deletion, or other local structural difference, depending on the scale of resolution of the analysis. The blocks are differently grouped into chromosomes and differently ordered and oriented in the two genomes being compared.

The construction of these synteny blocks based on traditional comparative maps is very vulnerable to errors and ambiguities in the position of the markers on a map. Errors are specific to the mapping technology producing the chromosomal marker positions. For example, statistical error in linkage disequilibrium calculations may result in ordering markers incorrectly. The same is true of mapping based on restriction enzymes, probe hybridization, and other techniques. Some of these may even occasionally map markers to the wrong chromosome. Another kind of problem involves ambiguous homology. Where there is nontandem duplication in one or both of the genomes being compared, whether it is a short fragment, a gene, or a larger segment, there is always the risk of matching up inappropriate pairs of markers as orthologs in the two genomes.

These problems are exacerbated when the number of genome rearrangements that differentiate one genome from the other is large relative to the number of markers available for their comparison. The general principle behind this is that, when two genomes contain a large number of consecutive markers in common, this lends confidence to the inference that the entire chromosomal segment containing the markers is a synteny block that was inherited intact from the common ancestor of the two genomes; at the other extreme, if a marker is the only one in common that is appearing on a chromosome in each of the genomes, it is more likely (though, of course, not necessary) that its position on these chromosomes is erroneous in at least one of the genomes. Thus, when many rearrangements have intervened since the common ancestor, the synteny blocks in common between the two genomes become more fragmented, that is, shorter, and more likely to contain only one marker, so that their status as bona fide synteny blocks is less certain. Similarly, if the genomes are densely sampled for markers, any particular synteny block is more likely to be confirmed by several consecutive markers in both genomes, whereas a sparse sample is more likely to have one per synteny block, reducing our confidence that these are genuine common inherited chromosomal segments.

These informal considerations suggest the principle whereby inferences that depend on the position of a single marker should not be given as much weight as inferences that are supported by two or more markers.

Even with this principle, a major source of difficulty in reconstructing synteny blocks is conflicting evidence for two incompatible blocks. For example, if two chromosomes in genome 1 contain $\cdots abc \cdots$ and $\cdots xyz \cdots$, respectively, it may happen that a chromosome in genome 2 contains $\cdots abxycz \cdots$. Then, we can reconstruct only one of abc or xyz as a synteny block common to both genomes. If we choose xyz, we either relegate c to the status of error or else infer that some rearrangements have occurred in this region to produce the interleaving pattern in genome 2.

- C. Zheng is with the Department of Biology, University of Ottawa, 160 Gendron Hall, 30 Marie Curie Street, Ottawa, Ontario, Canada, K1N 6N5. E-mail: czhen033@uottawa.ca.
- Q. Zhu is with the Biochemistry Program, University of Ottawa, 170 Gendron Hall, 30 Marie Curie Street, Ottawa, Ontario, Canada, K1N 6N5. E-mail: qzhu012@uottawa.ca.
- D. Sankoff is with the Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, Ontario, Canada, K1N 6N5. E-mail: sankoff@uottawa.ca.

Manuscript received 26 July 2006; revised 2 Feb. 2007; accepted 14 Feb. 2007; published online 22 Mar. 2007.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBBSI-0151-0706. Digital Object Identifier no. 10.1109/TCBBSI.2007.1075.

1545-5963/07/$25.00 © 2007 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

Whether these mapping difficulties are inherent in experimental methodology, paralogy, small marker sample, or elevated rearrangement rate, the effect on applications of comparative maps, for example, using knowledge about the genome of one organism to locate a marker in another, can be quite serious. Moreover, for the purposes of genome rearrangement inference, each mapping error typically introduces one or more spurious events into the rearrangement history so that even moderate rates of error can considerably inflate the inference of genomic distance.

We would like to construct a set of synteny blocks that are conflict-free, contain as much of the data as possible, and are credible from a genome rearrangement viewpoint. We may discard a limited amount of the data that conflicts with other data or whose inclusion would disproportionately inflate the distance. Our strategy is, first, to construct the set of all *prestrips*, which are certain common subsequences of two or more markers on single chromosomes in both genomes, second, to extract from this set a subset of mutually compatible prestrips containing a maximum number of markers, and, third, to restore to this subset any markers 1) not in any prestrips or 2) forming part of incompatible prestrips, whose inclusion does not cause the genomic distance based on the blocks to increase.

In Section 2, we formally define strips, prestrips, and other structures and quantities to be used in this paper. In Section 3, we present a new polynomial-time algorithm for generating all prestrips. Finding the optimal set of prestrips requires solving a maximum-weight clique (MWC) problem, as discussed in Section 4; this is the bottleneck in our procedure. We discuss the question of restoring additional compatible markers to the solution, in Section 5. We analyze the rice and sorghum comparative map in Section 6 and suggest constraints on the set of prestrips to reduce the input to MWC. We use simulations to assess the effects of these constraints on the solution.

2 Definitions: Strips, Prestrips, and Pure Strips

Let n be the number of distinct markers (not counting duplicates) in common in two genomes with χ_1 and χ_2 chromosomes and let $n_1 \geq n$ and $n_2 \geq n$ be the total numbers of markers, respectively, that is, counting each copy of a marker separately. In one genome, number all markers on any one of the chromosomes from left to right in increasing order, starting with marker 1. Continue the numbering sequence on a second chromosome and so on until finishing with the last marker on the χ_1st chromosome. If any subset of the markers is indistinguishable (duplicates, paralogs, gene family, and so forth), this set is called a *paralogy set* and is identified by a *paralogy set label* associated with each of its elements.

Then, each marker in the second genome receives the same label as its ortholog or as any of the paralogs of the latter in the first genome. Each paralogy set in one genome corresponds to a corresponding paralogy set in the other, even if it is only a trivial set containing one marker.

We now define *strips*, *prestrips*, and *pure strips*, as illustrated in Fig. 1. As mentioned in Section 1, during our analysis, we will be reducing the genomes by discarding some markers. *Strips* are defined relative to the current state of the two genomes before, during, or after reducing their size, but *prestrips* and *pure strips* are defined in terms of the original genome data only. Consider any $h \geq 2$ consecutive (contiguous) markers on a chromosome in the first genome. If the same h markers are consecutive on a chromosome in the second genome, with the same order and with each marker having the same orientation (DNA strand) in both genomes, they constitute a strip of *length* h. We call this a *forward* strip. Similarly, if the markers are consecutive on a chromosome in the second genome, but in the reverse order and with each marker having the opposite orientation in one genome to that which it has in the other, these markers constitute a *reverse* strip of length h.

```
ORIGINAL                  REDUCED

Genome 1                  Genome 1
abcdef                    abcd
lmnoprq                   lmoq
wdxyz                     wdyz

Genome 2                  Genome 2
lbcdpz                    lbcdz
-x-q-o-mbc                -q-o-m
wde-fry                   wdy
na                        a

Pre-strips                Strips
bcd,bc,cd,                bcd,moq,wdy
moq,mo,oq,
wdy,wd,dy,                Singletons not
lp,de,dz                  in pre-strips
                          but compatible
(Pure) strip              a,l,z
bcd,bc,de,wd
                          Discarded as noise
Common subsequences       e,f,n,p,r,x
not pre-strips
bd,mq,wy
```

Fig. 1. Strips and prestrips. The total number of distinct markers is $n = 17$. The total number of markers in the two genomes is $n_1 = 18$, $n_2 = 20$ since markers b, c, and d occur twice in one or both genomes. Minus signs indicate genes of different orientation (DNA strand) in the two genomes. Construction of reduced genomes is discussed in Sections 3, 4, and 5.

If a marker in a strip is a member of a paralogy set in one genome, it suffices that a member of the corresponding paralogy set occupies the corresponding position in the h markers in the other genome. Note that a strip is defined both by the markers in it and by its position in both genomes. Two or more strips may contain exactly the same markers but differ in where they appear, by virtue of the paralogy sets their markers belong to.

A prestrip satisfies the same definition as a strip except that the markers need not be contiguous. The markers in a prestrip must be in the same order on a chromosome in both original genomes and must conserve their orientations in the two genomes or else have the reverse order with all markers reversing their orientation. In addition, no marker not in a given forward or reverse prestrip may be located in both genomes between two successive markers that are in the prestrip if the marker has the same or reversed orientation, respectively, in the two genomes. As illustrated in Fig. 1, a prestrip P is a common subsequence, or a reverse common subsequence, of the markers on the two chromosomes which is *complete* in the sense that there is no other marker of appropriate orientation on both chromosomes that is between two successive markers in P. A prestrip that is a strip in the original genome data is called a *pure strip*.

We formulate our basic problem, Maximal Strip Recovery (MSR), as follows: *Given two genomes as described above, discard some subset of the markers, leaving only markers in disjoint strips $S_1, \cdots, S_r$ of lengths $h_1, \cdots, h_r$, respectively (estimates of the lengths of the synteny blocks that contain them), in the genomes thus reduced such that $\sum_{i=1}^{r} h_i$ is maximized.* The MSR problem corresponds to our previously stated goal of constructing a set of compatible strings containing as much of the data as possible. We postpone until Section 5 the question of restoring markers not in strips but compatible with the solution to the MSR problem.

3 The Prestrips

Though we are searching for strips, these are not generally visible in the original data and we have to construct them by discarding the markers disrupting their contiguity property. Thus, we search for prestrips—complete common subsequences or their reverse—in the two genomes, relying on the subsequent analyses to eliminate the disrupting markers and thus reveal the "underlying" strips. The justification for this is:

Proposition 1. *All possible strips that can be formed by the deletion of markers from two genomes and that can be part of a solution to the MSR problem are prestrips of these genomes.*

Proof. Consider any strip of length h, resulting from the deletion of m markers from the two genomes. We first restore all markers to the chromosomes containing the strip. Because the h strip markers are in the same order in the two chromosomes, they constitute a common subsequence P of the two chromosomes. Suppose P is not a complete common subsequence. Then, by definition, for some two successive markers in P, there is another marker in between them in both chromosomes. However, we can add this new marker to P to create a longer common subsequence Q so that deleting all of the remaining $m-1$ originally deleted markers creates a strip of length $h+1$, which would give a better value to the MSR problem. This contradicts the supposition that P is not complete and, so, P is a prestrip. □

The number of prestrips, however, may be exponentially large. Consider, for example, genome 1 to be the identity permutation of n^2 elements: $1, 2, 3, \cdots, n^2$ and genome 2 to be $n, n-1, \cdots, 2, 1, 2n, 2n-1, \cdots, n+2, n+1, 3n, \cdots$. The number of prestrips clearly grows exponentially with n.

Because the feasibility of our subsequent analysis is sensitive to how many prestrips we start with, we next prove that it suffices to constrain our search to certain small prestrips and obviate the necessity of generating all complete prestrips. The number of such small prestrips can only grow as a polynomial function of $n_1 + n_2$ and the algorithm for generating them runs in polynomial time.

We introduce new notation to describe prestrips. Every prestrip P has a unique representation R as a string of ps and 1s, where a p represents a pure strip and a 1 represents a marker not in a pure strip in P. This representation is easily constructed starting at either end of the prestrip and simply verifying for each marker in turn whether or not it is in a pure strip. For example, the prestrip $P = wdy$ in Fig. 1 can be represented by $p1$ because wd is a pure strip and y is in no pure strip in P.

Proposition 2. *The representation R of any prestrip can be decomposed into a sequence of terms of form p, 11, $1p$, $p1$, 111, and $1p1$.*

Proof. We use induction on the total number ν of ps and 1s in R, building R from left to right. A prestrip by definition must contain at least two markers and, so, for $\nu = 1$, the representation R must be simply p. If the Proposition is true for ν and R contains $\nu + 1$ terms, define R' to be the first ν terms of R. Consider the $\nu + 1\mathrm{st}$ term of R. If this is a p, then this can simply be added to any appropriate decomposition of R' to satisfy the proposition.

If the $\nu + 1\mathrm{st}$ term of R is a 1 and the $\nu\mathrm{th}$ term of R' is a p, then the last term of the decomposition of R' can only be a p or $1p$. This last term can be transformed to a $p1$ or a $1p1$ to satisfy the proposition for R.

Finally, if both the $\nu + 1\mathrm{st}$ term of R and the $\nu\mathrm{th}$ term of R' are 1s, then the last term of the decomposition of R' must be a 111, a $1p1$, a $p1$, or a 11. In the decomposition of R, these can be replaced by two 11s, a $1p$ plus a 11, a p plus a 11, or a 111, respectively. This completes the induction step. □

Note that we must allow for 111 and $1p1$ to be able to deal with representations R with an odd number of terms; otherwise, p, 11, $1p$, and $p1$ terms would suffice for the decomposition.

The implication of Proposition 2 is that the markers contained in any collection C of prestrips are also contained in a restricted collection of prestrips of form p, 11, $1p$, $p1$, 111, and $1p1$. Furthermore, the sum of the weights of the prestrips in the restricted collection is the same as that for C.

All prestrips of this form can be calculated by the following algorithm:

```
Algorithm Prestrips
Input: Genome 1 and genome 2 as a collection of $\chi_1$ and $\chi_2$ linear chromosomes, containing a total of $n_1$ and $n_2$ markers, each of the $n$ different markers appearing at least once in each genome.
For each of the $\chi_1\chi_2$ pairs of chromosomes, one from each genome,
    1. construct a dot plot of the markers in common;
    2. identify pure strips ($p$s) and singletons (1s) (note that
       every marker in a pure strip has an alternative
       identification as a singleton since it can be used in
       alternate prestrips);
    3. find which $p$s and 1s can be successive, taking into
       account that all markers in a prestrip must be in the
       same order in the two genomes or the reverse order;
       and
    4. construct prestrips of form $p$, 11, $1p$, $p1$, 111, and $1p1$.
Output: Prestrips
```

Let $n = \max(n_1, n_2)$. The construction of a dot plot requires $O(n^2)$ time, but, at each of the n dots found, the search for a possible preceding p or 1 also requires $O(n^2)$ time. Moreover, for the possibly n^2 combinations so detected, the search for a

TABLE 1
Prestrip Inclusion Criteria and Solution Characteristics

constraints	pre-	running	strips	markers	markers	total	
	strips	time	output	in output	restored	markers	distance
$G < 5$ no $11's$	441	2 minutes	97	286	17	303	49
$G < 4$	543	24 hours	124	309	13	322	69
$G < 3$	428	29 minutes	123	302	13	315	63
$G < 2$	257	6 seconds	115	282	36	318	65

possible third p or 1 also requires $O(n^2)$ time so that the whole algorithm might require $O(n^4)$ time. In practice, the sparseness of the dot plot, especially when there is little paralogy, assures that the running time is far less.

4 Maximum Weight Cliques

Once we have a set of prestrips, we next need to construct a matrix M of compatibilities among them. Two prestrips, P and Q, are incompatible, $M(P,Q) = 0$, if they share at least one marker or if one prestrip, say, P, contains a marker between two successive markers, in either genome, in the other prestrip, Q. Otherwise, P and Q are compatible and $M(P,Q) = 1$. This definition leads directly to:

Proposition 3. *Given any set C of pairwise compatible prestrips. Consider the reduced genomes produced by deleting all markers that are in none of the prestrips in C. In these reduced genomes, all of the markers in each prestrip in C appear as strips. The number of markers in each strip is the same as the number of markers in the corresponding prestrip.*

Proof. Consider any prestrip $P \in C$. By deleting all of the markers not in the prestrips, P is converted into a strip S; otherwise, one of the other prestrips would be incompatible with it, contrary to the hypothesis. Since none of its markers are deleted, S has the same markers as P. □

We next need to state the MWC problem. *Let G be a graph with positive weights associated to the vertices. Find a clique such that the sum of the vertex weight is maximum.*

From the two original n_1 and n_2-marker genomes, we wish to find a reduction, composed completely of strips $S_1, \cdots, S_r$, that maximizes $\sum_{i=1}^{r} h_i$. From the compatibility matrix, we construct a graph G with the prestrips as the vertices and with compatible vertices joined by an edge and vertex weights equal to the number of markers in the prestrip.

Proposition 4. *The solution C of the maximum weighted clique problem on G induces a reduction of the original genomes so that they are composed completely of disjoint strips and so that the total strip score is maximized.*

Proof. This follows from Propositions 2 and 3 and the definition of MWC. □

Kumlander's algorithm [6] for the MWC problem is based on a heuristic vertex coloring of sets of independent vertices, followed by two sorts of pruning of the clique search tree, one based on color classes and the other a backtrack search. Empirically, it has been demonstrated to work better than competing algorithms on denser graphs. In the case of genome comparison, we would expect G to become dense, in the sense of that in [1], as the number of chromosomes increases. In this case, any prestrip becomes less likely to involve the same two chromosomes as another prestrip and, hence, is more likely to be compatible with it.

As documented in Table 1, summarizing the map comparison discussed in Section 6, our current JAVA implementation of this algorithm generally requires a few minutes of computing time on an iBook G4 for 200 prestrips, but required up to 24 hours for some of our data sets containing 500 prestrips.

5 Restoration of Markers

The MWC solution is incompatible with any prestrip not in it, but it is not necessarily incompatible with all parts of such a prestrip. For example, it is possible that some prestrip of form $p1$ is not in the solution, but the singleton element in this prestrip does not intervene between any two successive markers of a prestrip in the solution and may thus be considered compatible. In addition, markers in no prestrip which play a role neither in the input nor in the output of the MWC could be similarly compatible with the solution.

The question arises of whether we want to reincorporate all such markers, some of them, or none of them into the solution. This answer depends on the biological context. Some of these markers may not have belonged to any prestrips, possibly due to sparse data, that is, low sampling density; were markers denser on the chromosomes, more of them might have been included in prestrips. Such markers are evolutionarily meaningful and, if we could identify them, should be tested for compatibility and possibly combined with the MWC solution. At the other extreme, we may think that the markers in our data in no prestrip or in no prestrip compatible with the MWC solution are primarily due to mapping error and, if we could identify them, should be discarded.

Whatever the policy toward restoring markers, with some data sets, we may have to formulate an optimal restoration algorithm since the restoration of one marker may affect whether or not another marker may be restored. With the stringent restoration criterion we will present next, however, and the small number of eligible markers in the real and simulated data to be discussed in Sections 6 and 7,

```
Genome 1                Genome 1 output
abcde                   ab cde
fghklm                  fgh kl

Genome 2                Genome 2 output
ab-l-k                  ab -l-k
fgh-mcde                fgh cde

MWC output strips       Distance = 2
ab,cde, fgh,kl          = 1 translocation
                        + 1 inversion

                        both before and after
                        restoration of marker m
```

Fig. 2. Restoration of marker that does not increase genomic distance.

the order in which markers are restored is a minor question and will not be discussed further here.

Since there is no way of identifying, in real data, exactly which markers excluded from the MWC solution are valid evidence of evolutionary relatedness or divergence of the two genomes and which are simply erroneous, we have recourse to genome rearrangement analysis. First, we use the strips output by the MWC to calculate the genomic distance between the two genomes [7], [9]. If we were to add a new marker at random ("noise") to both genomes, this would generally increase the distance by 1 or 2, even if it were compatible with all the strips in the solution. Thus, if we add a marker from among those not in the MWC and this does not increase the distance, this means that, when one genome is optimally transformed into the other, the new marker falls naturally into place with no extra effort and is fully consistent with the evolutionary history of all the markers in the solution.

Then, the final result in our analysis includes both the MWC solution plus all other markers that do not increase the genomic distance. An example of such a marker is **m** in Fig. 2.

6 A Comparison of the Rice and Sorghum Genomes

To illustrate our method, we compare data on the rice and sorghum genomes as accessible in the Gramene database [8]. The original references for the two maps we compare are [3] and [2].

In this comparison, the database reports 567 correspondences between the two genomes, involving $n_1 = 481$ rice markers and $n_2 = 567$ sorghum markers. The number of distinct markers was $n = 481$. A total of 69 of these were present in two or more copies in the sorghum data, with a maximum gene family size of 6.

We note that the marker data extracted from the Gramene files do not indicate the orientation or strand of the marker. Efficient genome rearrangement algorithms [7], [9] require this information. However, as conjectured in [5] and proven in [4], for all strips consisting of three or more consecutive markers in the same order in both genomes, same strandedness may be assumed for all of these markers and, for all strips in opposite orders, opposite strandedness can be assumed. In [4], it is also shown how to assign strandedness for strips containing only two markers. These same rules can all be used for prestrips since these will become strips if they are in the output of the MWC.

Our algorithm for generating prestrips produced 1,841 to enter as vertices into the MWC routine.

Our implementation of Kumlander's algorithm ran for a week without finding a solution for this large data set, so we reduced the problem by imposing two kinds of constraints on the prestrips. One constraint was to limit the number of "gap" markers G in either genome intervening between any two markers or pure strips in a prestrip on the grounds that a large gap lowers the credibility of the prestrip. Another constraint was to discard the prestrips with representation "11" on the grounds that these are the weakest evidence for synteny blocks aside from singletons. However, to ensure that only the weakest evidence was excluded in this case, we added prestrips of form $p11$ and $11p$ to the MWC input.

The results of these analyses are shown in Table 1. The first thing to note is that, even after all possible compatible markers, consistent with the output rearrangement distance, are restored, only 303-322 are present, meaning that 159-178 were discarded, of the maximum possible represented by $n_1 = 481$. This illustrates the importance of analyzing the marker data to remove noise and conflicts.

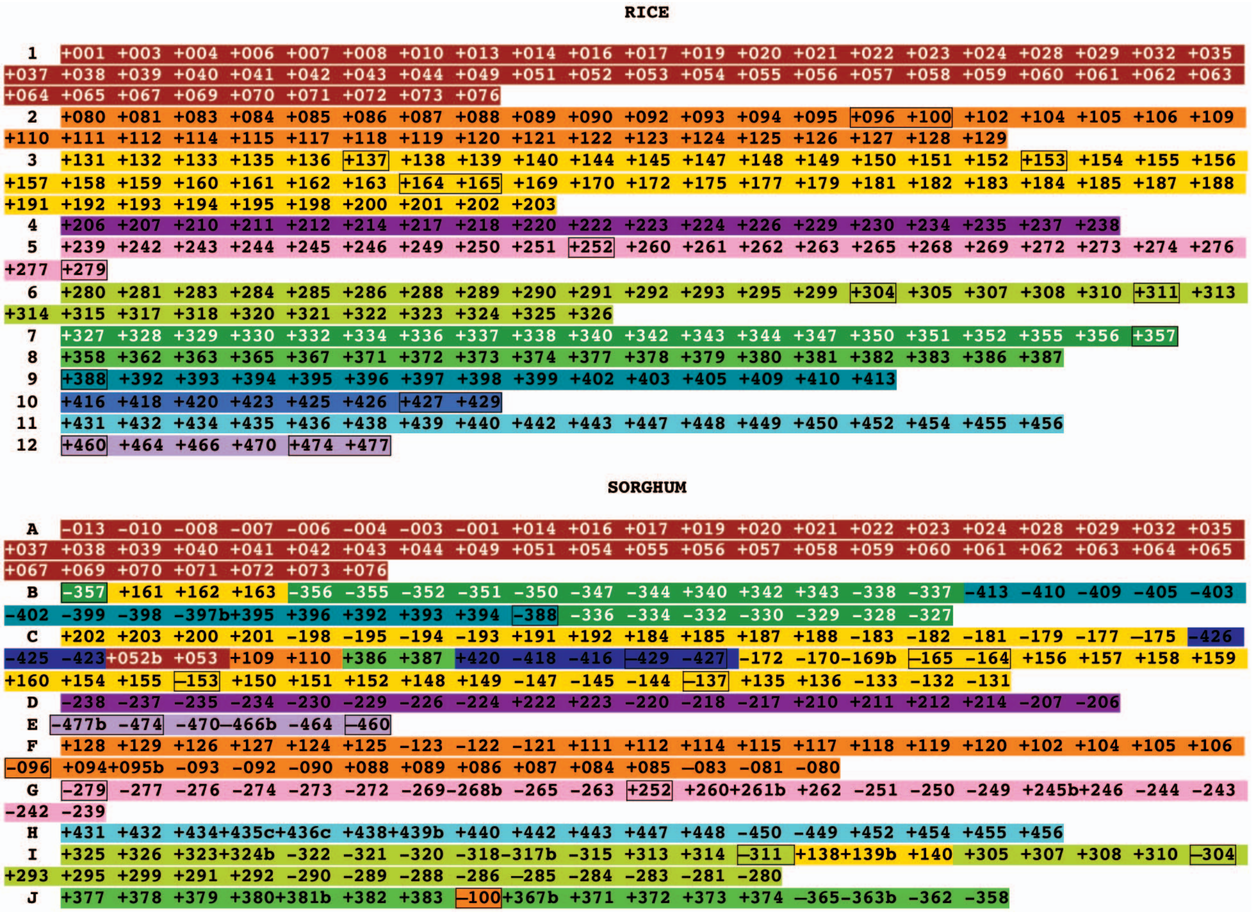
Another observation is the slight increase, if any, in the number of markers in the output as the gap size criterion is relaxed from $G < 2$ to $G < 4$, despite the great increase in the number of prestrips. Thus, the extra prestrips proved to be largely redundant.

Finally, we note the great drop in the genomic distance calculation as the "no 11s" constraint is added in the top row of the table. True, this comes at the cost of losing 19 markers from the output, but the fact that the distance lost is about equal to the number of markers lost suggests that these markers, coming largely from isolated "11" prestrips (that is, that could not be incorporated in $p11$ or $11p$ prestrips), do not carry authentic evolutionary information, using the same arguments about noise as in Section 5. The solution summarized in the top row of the table gives rise to the synteny blocks in Fig. 3.

7 Simulations

To further evaluate the results in Table 1, we carried out some simulations to create data sets similar to the rice-sorghum comparison. For the first genome, we defined 11 chromosomes with a total of 495 distinct genes, with a range of lengths (number of markers) similar to the rice genome. Based on the number of synteny blocks per chromosome in Fig. 3, we applied two inversions per chromosome, with breakpoints determined at random on the chromosome, and 23 reciprocal translocations, with breakpoints determined randomly, conditioned on not being on the same chromosome. The genome thus created was thus at a genomic distance of 45 from the initial one. We created 25 such data sets.

We then added noise by randomly changing the location of a random markers in the original genome, without changing them in the derived genome. This was done 200 times, independently in each of the 25 data sets, in order to simulate the same high level of conflict and noise observed in the rice-sorghum data.



RICE

1 +001 +003 +004 +006 +007 +008 +010 +013 +014 +016 +017 +019 +020 +021 +022 +023 +024 +028 +029 +032 +035
+037 +038 +039 +040 +041 +042 +043 +044 +049 +051 +052 +053 +054 +055 +056 +057 +058 +059 +060 +061 +062 +063
+064 +065 +067 +069 +070 +071 +072 +073 +076

2 +080 +081 +083 +084 +085 +086 +087 +088 +089 +090 +092 +093 +094 +095 +096 +100 +102 +104 +105 +106 +109
+110 +111 +112 +114 +115 +117 +118 +119 +120 +121 +122 +123 +124 +125 +126 +127 +128 +129

3 +131 +132 +133 +135 +136 +137 +138 +139 +140 +144 +145 +147 +148 +149 +150 +151 +152 +153 +154 +155 +156
+157 +158 +159 +160 +161 +162 +163 +164 +165 +169 +170 +172 +175 +177 +179 +181 +182 +183 +184 +185 +187 +188
+191 +192 +193 +194 +195 +198 +200 +201 +202 +203

4 +206 +207 +210 +211 +212 +214 +217 +218 +220 +222 +223 +224 +226 +229 +230 +234 +235 +237 +238

5 +239 +242 +243 +244 +245 +246 +249 +250 +251 +252 +260 +261 +262 +263 +265 +268 +269 +272 +273 +274 +276
+277 +279

6 +280 +281 +283 +284 +285 +286 +288 +289 +290 +291 +292 +293 +295 +299 +304 +305 +307 +308 +310 +311 +313
+314 +315 +317 +318 +320 +321 +322 +323 +324 +325 +326

7 +327 +328 +329 +330 +332 +334 +336 +337 +338 +340 +342 +343 +344 +347 +350 +351 +352 +355 +356 +357

8 +358 +362 +363 +365 +367 +371 +372 +373 +374 +377 +378 +379 +380 +381 +382 +383 +386 +387

9 +388 +392 +393 +394 +395 +396 +397 +398 +399 +402 +403 +405 +409 +410 +413

10 +416 +418 +420 +423 +425 +426 +427 +429

11 +431 +432 +434 +435 +436 +438 +439 +440 +442 +443 +447 +448 +449 +450 +452 +454 +455 +456

12 +460 +464 +466 +470 +474 +477

SORGHUM

A -013 -010 -008 -007 -006 -004 -003 -001 +014 +016 +017 +019 +020 +021 +022 +023 +024 +028 +029 +032 +035
+037 +038 +039 +040 +041 +042 +043 +044 +049 +051 +054 +055 +056 +057 +058 +059 +060 +061 +062 +063 +064 +065
+067 +069 +070 +071 +072 +073 +076

B -357 +161 +162 +163 -356 -355 -352 -351 -350 -347 -344 +340 +342 +343 -338 -337 -413 -410 -409 -405 -403
-402 -399 -398 -397b+395 +396 +392 +393 +394 -388 -336 -334 -332 -330 -329 -328 -327

C +202 +203 +200 +201 -198 -195 -194 -193 +191 +192 +184 +185 +187 +188 -183 -182 -181 -179 -177 -175 -426
-425 -423+052b +053 +109 +110 +386 +387 +420 -418 -416 -429 -427 -172 -170-169b -165 -164 +156 +157 +158 +159
+160 +154 +155 -153 +150 +151 +152 +148 +149 -147 -145 -144 -137 +135 +136 -133 -132 -131

D -238 -237 -235 -234 -230 -229 -226 -224 +222 +223 -220 -218 -217 +210 +211 +212 +214 -207 -206

E -477b -474 -470-466b -464 -460

F +128 +129 +126 +127 +124 +125 -123 -122 -121 +111 +112 +114 +115 +117 +118 +119 +120 +102 +104 +105 +106
-096 +094+095b -093 -092 -090 +088 +089 +086 +087 +084 +085 -083 -081 -080

G -279 -277 -276 -274 -273 -272 -269-268b -265 -263 +252 +260+261b +262 -251 -250 -249 +245b+246 -244 -243
-242 -239

H +431 +432 +434+435c+436c +438+439b +440 +442 +443 +447 +448 -450 -449 +452 +454 +455 +456

I +325 +326 +323+324b -322 -321 -320 -318-317b -315 +313 +314 -311 +138+139b +140 +305 +307 +308 +310 -304
+293 +295 +299 +291 +292 -290 -289 -288 -286 -285 -284 -283 -281 -280

J +377 +378 +379 +380+381b +382 +383 -100 +367b +371 +372 +373 +374 -365-363b -362 -358

Fig. 3. Syntenic blocks in sorghum genome color-keyed by rice chromosomes. Boundaries between small prestrips of the same color not shown. Markers restored post-MWC are framed in both genomes.

The average results of analyzing the data are presented in Table 2.

We note first that the number of prestrips is greatly reduced in these comparisons compared to the real data. A small part of this is due to the presence of some duplicate genes in the real data, leading to additional possibilities for constructing prestrips. However, the bulk of the extra prestrips in the original is due to the lack of strandedness in the marker data, whereas this information was conserved in the simulations. Suppose one genome had markers $\cdots a \cdots b \cdots c \cdots$ on a chromosome on the same strand, whereas the other genome had $\cdots a \cdots - b \cdots c \cdots$, where the minus sign indicates the opposite strand. Without the strandedness, there are three prestrips, ab, bc, and abc, while knowledge of the strandedness cuts the number of prestrips to one, namely, ac.

While there is still an excess of strips output in Table 1 over Table 2, this is negligible in comparison to the excess of prestrips.

Note that the number of markers output in the analyses of the real and simulated data are approximately the same —this was deliberate to make the data sets comparable and was determined by choosing 200 "noisy" events. Had we chosen 100 or 300, the results would have been very different.

TABLE 2
Prestrip Inclusion Criteria and Simulated Solutions

constraints	pre-	strips	markers	markers	total	noise	non-noise	dis-
	strips	output	in output	restored	markers	retained	cut	tance
$G < 5$ no $11's$	155.5	79.8	304.6	20.2	324.8	47.8	19.6	44.2
$G < 4$	191.0	97.9	306.4	16.8	323.2	58.5	13.0	55.9
$G < 3$	169.6	95.7	295.6	20.7	316.3	55.7	13.3	52.8
$G < 2$	136.8	90.4	276.0	22.2	298.2	51.4	16.0	47.9

The important result in Table 2 pertains to the distance calculation. Excluding the prestrips consisting of only two singletons produces an accurate estimate (44.2) of the original, prenoise, distance (45) between the two genomes. Allowing these two-marker strips results in the inclusion of more markers in the analysis, but at the expense of seriously inflating the distance estimate. The effect is not as great as with the real data, but it is clear that allowing these two-marker prestrips into the analysis does not add genuine evolutionary information.

Other points of interest in Table 2 are the numbers of "false positives and negatives" in the analysis, represented in the columns headed "noise retained" and "non-noise cut," respectively. Recall that a noise level of 200 means a total of 200 position changes occurred. Of these, in the final results, we note that only 47.8-58.5 were not excluded by the MWC or were restored after the MWC, which is not a very high rate of false positives for such noisy data. Of the markers that were never affected by noise, only 13-19.6 were excluded by the MWC and not restored, again an acceptable level of false negatives.

8 Conclusion

In this work, we have studied the conversion of the maximal weight strip problem to the MWC problem, based on the induced elimination of as few markers as possible from the genomes being compared. To increase the realism of our formulation, we have extended it to allow for paralogous markers.

We have shown how to reduce the prestrip computation to polynomial complexity by using only six types of small prestrip, though, with real data, this calculation is not the bottleneck. Applying the MWC algorithm to these small prestrips, followed by piecing all contiguous ones together, gives all of the same solutions as the previous method.

It is the MWC itself that is the bottleneck. Since our compatibility graph is dense, methods such as those introduced in [1] might speed up the MWC search. Vicky Choi (personal communication) has suggested that the availability of relatively fast programs for Minimum Weighted Vertex Cover, which is equivalent to MWC, could allow us to handle larger numbers of prestrips.

It might be thought that, with the advent of genome sequencing, the use of comparative mapping to study genome rearrangement would become obsolete. In fact, the trend toward low-coverage sequencing without finishing, leaving many gene order ambiguities, makes it likely that physical and genetic mapping methods will continue to predominate, aside from very few model organisms. This is especially relevant to comparative genomics applied to the phylogenetics of eukaryotic taxa, where we might wish to calculate the genomic distance among dozens or hundreds of organisms, but few, if any, of these will have sequenced genomes.

Acknowledgments

This research was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada. The authors thank the referees of this paper, as well as those of the conference version [10], for many helpful comments, suggestions, and questions. They particularly thank Vicky Choi for extensive advice during the revision of this paper.

References

[1] S. Arora, D. Karger, and M. Karpinski, "Polynomial Time Approximation Schemes for Dense Instances of NP-Hard Problems," *J. Computer and System Sciences,* vol. 58, pp. 193-210, 1999.

[2] J.E. Bowers, C. Abbey, S. Anderson, C. Chang, X. Draye, A.H. Hoppe, R. Jessup, C. Lemke, J. Lennington, Z. Li, Y.R. Lin, S.C. Liu, L. Luo, B.S. Marler, R. Ming, S.E. Mitchell, D. Qiang, K. Reischmann, S.R. Schulze, D.N. Skinner, Y.W. Wang, S. Kresovich, K.F. Schertz, and A.H. Paterson, "A High-Density Genetic Recombination Map of Sequence-Tagged Sites for Sorghum, as a Framework for Comparative Structural and Evolutionary Genomics of Tropical Grains and Grasses," *Genetics,* vol. 165, pp. 367-386, 2003.

[3] Rice-Gramene Annotated Nipponbare Sequence, http://www.gramene.org/Oryza_sativa/, 2006.

[4] S. Hannenhalli and P.A. Pevzner, "To Cut or Not to Cut (Applications of Comparative Physical Maps in Molecular Evolution)," *Proc. Seventh Ann. ACM-SIAM Symp. Discrete Algorithms,* pp. 304-313, 1996.

[5] J. Kececioglu and D. Sankoff, "Exact and Approximation Algorithms for the Inversion Distance between Two Permutations," *Proc. Fourth Combinatorial Pattern Matching Symp.,* A. Apostolico, M. Crochemore, Z. Galil, and U. Manber, eds., pp. 87-105, 1993, full version in *Algorithmica,* vol. 13, pp. 180-210, 1995.

[6] D. Kumlander, "A New Exact Algorithm for the Maximum-Weight Clique Problem Based on a Heuristic Vertex-Coloring and a Backtrack Search," *Proc. Fourth European Congress Math.,* 2005

[7] G. Tesler, "Efficient Algorithms for Multichromosomal Genome Rearrangements," *J. Computer and System Sciences,* vol. 65, pp. 587-609, 2002.

[8] D. Ware, P. Jaiswal, J. Ni, X. Pan, K. Chang, K. Clark, L. Teytelman, S. Schmidt, W. Zhao, S. Cartinhour, S. McCouch, and L. Stein, "Gramene: A Resource for Comparative Grass Genomics," *Nucleic Acids Research,* vol. 30, pp. 103-105, 2002.

[9] S. Yancopoulos, O. Attie, and R. Friedberg, "Efficient Sorting of Genomic Permutations by Translocation, Inversion and Block Interchange," *Bioinformatics,* vol. 21, pp. 3340-3346, 2005.

[10] C. Zheng and D. Sankoff, "Rearrangement of Noisy Genomes," *Proc. 2006 Int'l Workshop Bioinformatics Research and Applications.* Part II, V.N. Alexandrov et al., eds., pp. 791-798, 2006.

Chunfang Zheng received the bachelor's degree in biology from Beijing Sports University in 1989 and the bachelor's degree in computer science and the master's degree from the University of Ottawa in 2001 and 2005, respectively. She is a PhD candidate in the Biology Department at the University of Ottawa. She has published several computational biology papers on partially ordered genomes, the genome halving problem, and removing noise from comparative maps.

Qian Zhu is currently working toward the bachelor's degree, specializing in biochemistry and computer science, at the University of Ottawa. He has been a research intern in the Sankoff Lab during his undergraduate studies and has developed an expertise in genomic databases, notably, Gramene.

David Sankoff received the PhD degree in mathematics from McGill University under the direction of Donald Dawson and has been a member of the Centre de recherches mathématiques in Montreal for many years. He currently holds the Canada Research Chair in Mathematical Genomics in the Mathematics and Statistics Department at the University of Ottawa, and is cross-appointed to the Biology Department and the School of Information Technology and Engineering. His research interests include comparative genomics, particularly probability models, statistics, and algorithms for genome rearrangements. He is a fellow of the Royal Society of Canada and a fellow of the Evolutionary Biology Program of the Canadian Institute for Advanced Research.