AN $O(N^2 \log N)$ RESTRICTION MAP COMPARISON AND SEARCH ALGORITHM

EUGENE W. MYERS*
Department of Computer Science,
University of Arizona,
Tucson, AZ 85721, U.S.A.

XIAOQIU HUANG†
Department of Computer Science,
The Pennsylvania State University,
University Park, PA 16802, U.S.A.

We present an $O(R \log P)$ time, $O(M + P^2)$ space algorithm for searching a restriction map with $M$ sites for the best matches to a shorter map with $P$ sites, where $R$, the number of matching site pairs, is bounded by $MP$. As first proposed by Waterman et al. (1984, Nucl. Acids Res. 12, 237–242) the objective function used to score matches is additive in the number of unaligned sites and the discrepancies in the distances between adjacent aligned sites. Our algorithm is basically a sparse dynamic programming computation in which “candidate lists” are used to model the future contribution of all previously computed entries to those yet to be computed. A simple modification to the algorithm computes the distance between two restriction maps with $M$ and $N$ sites, respectively, in $O(MN (\log M + \log N))$ time.

1. Introduction. A restriction map of a DNA strand is an ubiquitous tool of molecular biology. For each of a finite and small number of restriction enzymes, a “map” gives the position along the strand of the recognition sites at which each enzyme cleaves the DNA. For example, Kohara et al. (1987) recently produced a map of the DNA of the bacteria E. coli with respect to eight restriction enzymes. The map of the first 25 kilobases of this 4.72 million nucleotide sequence is shown in Fig. 1 as originally presented by Kohara. There is an open bar for each restriction enzyme (its name is to the left of the bar) and each vertical line within a bar indicates the location of a recognition site for the given enzyme. Experimental errors inherent in the techniques used to produce such maps result in inaccurate site locations, and missing and spurious sites. Consequently one cannot expect perfect alignment and correlation between maps arising from different experimental sources. This fact alone necessitates software that can compute optimal alignments between two potentially similar maps. But there are many other uses for such a comparison algorithm, among them: assembling complete physical maps from partial ones, locating potentially conserved regions between and within species, finding the

* This author's work was supported in part by National Library of Medicine Grant R01-LM4960.
† This author's work was supported in part by National Library of Medicine Grant R01-LM5110.
physical location of sequenced segments of DNA, and diagnosing genetic disorders by analysing polymorphisms.

![Figure 1. The first 25 kb of the Kohara map.](image)

This paper focuses on the problem first considered by Waterman et al. (1984), of aligning two restriction maps under a measure of similarity that is additive in the number of unaligned sites and the discrepancies in the distances between adjacent aligned sites. This similarity measure admits a dynamic programming formulation that leads directly to an \(O(M^2P^2)\) worst-case algorithm where \(M\) and \(P\) are the number of sites in the two maps. Recently, Miller and Huang (1988) considered the variation of finding subregions of a long map of \(M\) sites that are highly similar to a short probe map of \(P\) sites, where \(M\) is orders of magnitude larger than \(P\). For this problem they designed an algorithm whose worst-case performance is \(O(MP^3)\) but empirical evidence suggests that its expected case performance is \(o(M^P)\). Concurrently, Huang (1988) gave an \(O(MP^2)\) worst-case algorithm.

This paper presents an \(O(MP \log P)\) algorithm for comparing restriction maps or searching a long map for an approximate match to a short probe. The basic idea for the algorithm is an outgrowth of recent work on fast sequence comparison algorithms in the case where gap costs are concave (Hirschberg and Larmore, 1987; Miller and Myers, 1988; Eppstein et al., 1988). While the objective function for aligning restriction maps appears at first to be affine, its dependence on map distances implies a dynamic programming recurrence for which multiple indels or gaps must be treated as a unit. Eppstein et al. (1990), have considered such recurrences for the problems of determining first-order RNA secondary structures and sparse sequence comparisons. Indeed, we discovered that their work encompasses that of this paper when slightly altered and specialized appropriate to the problem of comparing maps. Nonetheless, we proceed in this paper to give an algorithm development concretely tied to the map comparison problem. Our approach also differs in its conception, being based on the candidate-list paradigm introduced in our earlier work.

The remainder of the paper is organized as follows. In the next section preliminary concepts and notations are introduced. In Section 3, we give the outline of our algorithm based on a candidate-list paradigm. In Section 4,