Inferring Relatedness of a Macromolecule to a Sequence Database Without Sequencing

Jin Kim[1], James R Cole[2], Eric Torng[1] and Sakti Pramanik[1]
Department of Computer Science[1]
and
Center for Microbial Ecology[2]
Michigan State University
East Lansing, MI 48824
{kimj@cps, colej@pilot, torng@cps, pramanik@cps}.msu.edu

Abstract

Derivation of biological information of a macromolecule isolate based on sequence similarity is playing a significant role in numerous areas of biological research. However, it is often the case that a researcher obtains more macromolecule isolates than can be sequenced practically, due either to the high cost of sequencing or lack of specialized equipment and personnel. To overcome this difficulty, we study the problem of obtaining biological information (such as sequence information) about a macromolecule isolate using only (i) the fragmentation pattern of that isolate obtained from digestion with enzymes and (ii) a database D of sequences. We investigate a three phase approach to solving this problem. In the first phase, we obtain a restriction pattern of the isolate while analytically deriving the corresponding restriction maps of the sequences in the database. In the second phase, we identify a set $S \subseteq D$ of sequences which have restriction maps that are most similar to the unknown isolate's restriction pattern. This task is complicated by the fact that we have only approximate fragment lengths for the unknown isolate and that we do not know the actual ordering of the unknown isolate's fragments. Despite these difficulties, we derive experimental results which indicate maximum matching techniques are effective in identifying the correct set most of the time. In the third phase, we use the set S to infer biological information (such as sequence information or hierarchical classification information) about the unknown isolate. We demonstrate experimentally that the closeness of the sequences in the set S to each other can be used to infer the relatedness of the unknown isolate to the sequences of the set S. Furthermore, the confidence of this inferred information is strongly correlated to the minimum pairwise relatedness of any two elements in S.

Key words. algorithms, sequence databases, phylogenies, search, sequence similarity, restriction mapping

Introduction

Sequence database searches have become a normal first step in the identification and characterization of new macromolecule isolates. The information obtained from such database searches and comparisons often yields novel insights into isolate origin, function, and evolution. A number of very good tools exist for searching sequence databases, for example BLAST (Altschul *et al.* 1990) and FASTA (Pearson & Lipman 1988). These tools can be very powerful in rapidly discovering even weak similarities to a query. However, all of these database search tools require that the molecular sequence of the query be known.

Even though improvements in methods for nucleic acid and protein sequencing have substantially reduced the cost of obtaining sequences, there are many situations in which sequencing costs are still too high or the required specialized equipment is just not available. Often, a researcher is able to obtain more macromolecule isolates than it is practical to sequence. Many of these isolates may be identical, and some rapid method is needed to obtain a set of unique isolates.

One simple method of categorizing multiple isolates is to use the restriction pattern obtained from digestion with certain enzymes (proteases and restriction endonucleases). These enzymes typically cleave the macromolecular substrate at specific subsequences. By comparing the pattern of fragment lengths produced by the isolates, sets of non-identical patterns can be selected for further study.

At a further level of sophistication, the number of matching (same size) fragments between patterns may be used to calculate a measure of (phylogenetic) relatedness between isolates (Nei & Li 1979). If the positions of cleavage sites along the molecule (cleavage site map) are known as well as the fragment sizes, then several more accurate methods exist for estimate the relatedness between isolates (Nei & Li 1979; DeBry & Slade 1985; Felsenstein 1992; Holsinger & Jansen 1993). Some work has been done on constructing cleav-

age site maps from digestion patterns (Pearson 1982; Fitch, Smith, & Ralph 1983; Durand & Bregegere 1984; Bellon 1988; Grigorjev & Mironov 1990; Wright *et al.* 1994), but this work typically assumes that there are overlapping fragments. In our case, there are no overlapping fragments to guide us in constructing a cleavage site map.

Cleavage site maps can be computed easily from molecular sequences. We are interested in how such maps generated from a sequence database might be used to help characterize a molecular isolate. Assume a cleavage pattern for a query molecular isolate (measured experimentally to within some expected limit of accuracy) and a database containing sequences with varying degree of sequence similarity to the query isolate. We would like to be able to extract the set of sequences with cleavage site maps most similar to the (unknown) map of the isolate. In addition, we would like some estimate of how closely related these sequences are to our query molecule in terms of sequence similarity or other biological measures. Related work has also been done in identifying the location of restriction maps of short query probes in longer restriction maps (Miller, Ostell, & Rudd 1990; Miller, Barr, & Rudd 1991). However, this problem differs from ours in that the ordering of the query fragments is known.

Grouping isolates by cleavage site maps may not be useful in itself, unless this grouping implies some more standard measure of relationship, such as primary sequence similarity. Although primary similarity between two sequences can be used to directly estimate the expected fraction of shared enzyme sites, sequence similarity can not be accurately estimated from the fraction of matching sites alone. Any attempt at performing the reverse calculation requires some knowledge of the underlying distribution of similarity in the pool from which the two sequences were drawn. This leads us to use experimental results for evaluating the effectiveness of different methods of inferring information about a query isolate.

Similar work has been done in classifying unknown peptides by using mass profiles where the unknown peptide is digested by enzymatic or chemical means and the masses of the resulting fragments are determined by mass spectrometry (James *et al.* 1993; Shaw 1993; Cottrell 1994). However, when dealing with peptides, several problem parameters are quite different than when dealing with nucleotides. First, the masses of the resulting fragments can be obtained with essentially no error. Second, the masses reveal significant information about the amino acid composition of each fragment.

We perform our experiments using the Ribosomal Database Project (RDP) database of bacterial 16S rRNA gene sequences (Maidak *et al.* 1994). This database is a member of a class of databases containing sequences derived from that of an unknown common evolutionary ancestor. In addition, the phylogenetic relationships of the sequences in the database have been estimated and are available from the RDP, as is a biological classification scheme based on these relationships. Also, comparison of restriction fragment patterns of rRNA genes (rDNA) is an accepted method of discerning relatedness between isolates in microbiological studies (Moyer, Dobbs, & Karl 1993).

In this work, we present a method for finding the set of database sequences with the most sites in common with a query restriction fragment pattern. This method uses sequences in the database as templates to assemble the fragments into putative restriction site maps. The results of this method are shown to be similar to results obtained when the exact fragment order is known for both the query isolate and the database sequences. Also, a method to estimate the sequence similarity between the query and database result set is presented. In addition, we demonstrate that the results can be used to place the query in an existing biological classification scheme.

Overview of Problem Solution

Our basic problem is to determine biological information such as primary sequence information about an unknown query isolate q using a database D sequences. Furthermore, we do not allow biological sequencing of the unknown query isolate q.

We investigate a three phase approach to solve this problem.

1. First, we use enzymes to obtain a *restriction pattern* of the query isolate q. Simultaneously, we analytically compute *restriction maps* of all the sequences $s \in D$.

2. We then compute the *closeness* of the restriction map of each sequence $s \in D$ to the restriction pattern of q.

3. This closeness information is then used to infer biological information about q.

We describe each of these steps in more detail in the following sections.

We note here that our methods for completing each step, particularly the third step, are dependent on the characteristics of the database D. We first show that analytic methods which try to infer the primary sequence similarity between the query isolate q and any

sequence $s \in D$ require a priori knowledge of the relationship of the underlying probability distribution of the primary sequence similarity of any sequence $s \in D$ to the sequence of isolate q. Straightforward analytic methods are further handicapped by the assumption that for closely related sequences $s, s' \in D$, the relationship of the restriction map of s to the restriction pattern of q is independent of the relationship of the restriction map of s' to the restriction pattern of q. As a result, we experimentally evaluate different methods for inferring information about the query isolate q. Our experimental results demonstrate that effective inference techniques do exist which utilize the set $S \subseteq D$ of sequences which have restriction maps closest to the restriction pattern of q. This result has good consequences for step two of our approach where we show that computing how close the restriction map of any sequence $s \in D$ to the restriction pattern of q may be computationally difficult but that determining if the restriction map of a sequence $s \in D$ is extremely close to the restriction pattern of q is computationally tractable.

Restriction Patterns and Restriction Maps

The basic biological processing we perform on the isolate q is digestion by enzymes which produces a restriction pattern of q. At the same time, we analytically compute the restriction maps of all sequences $s \in D$ that would have been formed if we had digested these sequences with the same enzymes. To facilitate later comparison of the restriction pattern of q to the restriction maps of $s \in D$, we assume that the database sequences represent molecules with ends at positions homologous to the ends of the query molecule. In practice, the end points of query molecules may be known if, for example, they are produced by the PCR reaction using primers to conserved regions. The primers define the query endpoints. For the specific RDP database, PCR is the method of choice for isolating rRNA genes.

The two processes of obtaining a restriction pattern of q and computing restriction maps for $s \in D$ differ in their precision. Because computing the restriction maps is analytical, we can compute the restriction maps exactly. However, while it is technically possible to measure the exact length of the restriction fragments of q, the methods required are not suitable for a rapid, inexpensive screen. A more realistic assumption is that fragment sizes would be estimated by simple gel or capillary electrophoresis; this leads us to assume that fragments f which have actual length $|f|$ will be measured to have length between $(1-\epsilon)|f|$ and $(1+\epsilon)|f|$. In our experiments, we estimate ϵ to be 5%.

Definitions

We now formally define the terms restriction map, restriction pattern, the closeness of two restriction maps, and the closeness of a restriction map to a restriction pattern. We begin by defining restriction maps, restriction patterns, and the equivalence relation that restriction patterns induce on restriction maps.

In the following definitions, s is a nucleotide sequence, i is a nucleotide isolate, $s(i)$ is the underlying (and unknown) sequence of isolate i, and z is an enzyme with a nucleotide recognition sequence of length $2l$. We abuse notation and use z to refer to both the enzyme z and z's nucleotide recognition sequence.

Definition 1 Let $last(s, j)$ denote the last j nucleotides in s. Let $first(s, j)$ denote the first j nucleotides in s.

Definition 2 The *restriction map* $RM_z(s)$ formed by digesting sequence s by enzyme z is the ordered tuple $(s_1, s_2, \ldots, s_n)$ such that

- $s = s_1 s_2 \cdots s_n$.
- For $1 \leq i \leq n-1$, the $last(s_i, l) first(s_{i+1}, l) = \overline{z}$ (we assume the enzyme cuts in the middle of its recognition sequence).
- For $1 \leq i \leq n$, $\overline{z}$ is not a subsequence of s_i (we assume that enzyme recognition sequences do not overlap in s).

We will abuse notation and often refer to the ordered tuple of fragment lengths, $(|s_1|, |s_2|, \ldots, |s_n|)$, as a restriction map as well.

Definition 3 Let $RM_z(s) = (s_1, s_2, \ldots, s_n)$. We define the *restriction pattern* $RP_z(s)$ to be the *unordered* multiset of fragment lengths $\{|s_1|, |s_2|, \ldots, |s_n|\}$. We say that $RM_z(s)$ *yields* $RP_z(s)$.

Note the previous two definitions apply only to sequences s and not isolates i. We will typically use $RM_z(i)$ and $RP_z(i)$ as shorthand to represent $RM_z(s(i))$ and $RP_z(s(i))$.

Note many different restriction maps yield the same restriction pattern. For example, if there are n different fragment lengths in $RP_z(s)$, then any ordering of these n fragments is a restriction map which yields $RP_z(s)$ as a restriction pattern.

Definition 4 We define two restriction patterns to be *isomorphic* if they yield the same restriction pattern. We denote the set of isomorphic restriction maps which yield the restriction pattern $RP_z(s)$ with the notation $[RP_z(s)]$.

For example, the restriction map $(100, 200, 300)$ and $(300, 200, 100)$ are isomorphic but neither is isomorphic

to the restriction map $(100, 200, 100, 200)$. It is not hard to see that this isomorphic relation defines an equivalence relation on the set of restriction maps.

We now define closeness metrics for restriction maps and restriction patterns. First, we need some notation. Let $sites(RM_z(s))$ denote the number of cleavage sites in the restriction map $RM_z(s)$. Let $sites([RP_z(s)])$ denote the number of cleavage sites in any of the restriction maps in $[RP_z(s)]$. Note $sites(RM_z(s))$ and $sites([RP_z(s)])$ is one less than the number of fragments in $RM_z(s)$ and $RP_z(s)$.

Definition 5 We define a *common* site in sequences s and s' to be a cleavage site that appears in the same nucleotide position in both s and s'. Let $common(RM_z(s), RM_z(s'))$ denote the number of common cleavage sites in the restriction maps $RM_z(s)$ and $RM_z(s')$. Let $maxcommon(RM_z(s), RP_z(s'))$ denote the maximum number of common cleavage sites in any restriction map in $[RP_z(s')]$ and $RM_z(s)$.

Definition 6 We define the closeness of the two restriction maps, denoted $closeness(RM_z(s), RM_z(s'))$, to be $\max(sites(RM_z(s)), sites(RM_z(s'))) - common(RM_z(s), RM_z(s'))$. We define the closeness of a restriction map to a restriction pattern, denoted $closeness(RM_z(s), RP_z(s'))$, to be $\max(sites(RM_z(s), sites(RP_z(s'))) - maxcommon(RM_z(s), RP_z(s'))$.

Definition 7 We define the set of sequences $CLOSE(q, D, j) \subseteq D$ to be the sequences $s \in D$ such that $closeness(RM_z(s), RP_z(q)) \leq j$.

Computing Closeness Information

Step two of our approach is computing $closeness(RM_z(s), RP_z(q))$ for each sequence $s \in D$. While we show that this problem appears to be a computationally intractable, we describe how we can efficiently determine if $closeness(RM_z(s), RP_z(q)) = 0$. This allows us to compute the set $CLOSE(q, D, 0)$ which is sufficient for step three of our approach.

We first observe that the problem of computing $closeness(RM_z(s_1), RM_z(s_2))$ reduces to the problem of computing $common(RM_z(s_1), RM_z(s_2))$ and the problem of computing $closeness(RM_z(s_1), RP_z(s_2))$ reduces to that of computing $maxcommon(RM_z(s_1), [RP_z(s_2)])$. We next note that the second problem, computing $maxcommon(RM_z(s_1), [RP_z(s_2)])$, is NP-complete via a reduction from the 3-Partition problem (we omit this simple proof from this writeup).

As a result, we have focused on a restricted version of this problem which can be solved in polynomial time. In particular, we give a simple polynomial time algorithm which solves the decision problem, "Is $closeness(RM_z(s_1), [RP_z(q)]) = 0$?" We can apply this algorithm to all sequences $s \in D$ to efficiently compute the set $CLOSE(q, D, 0)$.

We first simplify the problem by observing the answer is no unless $sites(RM_z(s)) = sites([RP_z(q)])$. Thus, we need only consider sequences $s \in D$ which satisfy the constraint $sites(RM_z(s)) = sites([RP_z(q)])$. Next, we observe that for these sequences s, the answer is yes if and only if $RP_z(s)$ is identical to $RP_z(q)$ assuming that the fragment lengths are exact. However, while the set of fragment lengths for $RP_z(s)$ where $s \in D$ may be computed exactly, the fragment lengths for $RP_z(q)$ are only approximate. We only know that the actual length of fragment f ranges between $(1 - \epsilon)|f|$ and $(1 + \epsilon)|f|$. Thus, we must relax our yes condition. In particular, we must consider the length of a query fragment f_q to be identical to the length of any database fragment f_s if $(1 - \epsilon)|f_q| \leq f_s \leq (1 + \epsilon)|f_q|$. However, we must still insure that a query fragment f_q matches only one database fragment f_s and that every query fragment is matched. Suppose there are n fragment lengths in both $RP_z(s)$ and $RP_z(q)$. This problem is easily solved in $O(n \log n)$ time by sorting both multisets of fragment lengths and verifying that the i^{th} shortest database fragment length matches the i^{th} shortest query fragment length for $1 \leq i \leq n$.

Ideally, we would like to find the sequences $s \in D$ such that $RM_z(s) = RM_z(q)$. However, our data is imprecise in two ways which complicates this task. First, we have only $RP_z(q)$, not $RM_z(q)$. Second, we have only the approximate lengths, not the exact lengths, of each query fragment. The natural question to ask is, how much do these data imprecisions affect the correctness of our algorithm? We perform some experiments with the RDP database to to show that these imprecisions do not seriously affect the accuracy of our algorithm. These experiments and results are presented in section 5.

Inferring Information About the Query Isolate

We now consider the problem of inferring information about the isolate q from the closeness information we have computed. We first describe analytic methods for inferring information about q given $closeness(RM_z(s), RP_z(q))$ for any $s \in D$. Unfortunately, these techniques have three drawbacks which limit their applicability to biological databases such as the RDP database. First, the analytic methods require some a priori knowledge on the relationship between q and D such as the distribution of similarities between elements of D and q. Second, the analytic meth-

ods typically assume sequences evolve and change only through mutation or substitution; that is, insertions and deletions of nucleotides are typically not modeled. Third, the analytic methods typically assume that two sequences $s, s' \in D$ vary from $s(q)$ independently, even if we know s and s' are biologically similar. For these reasons, we are forced to evaluate our methods of inference experimentally. We describe a basic heuristic approach and experimentally verify the applicability of this approach.

Mathematical Models and Their Drawbacks

In this section, we describe two methods for using $closeness(RM_z(s), RP_z(q))$ to infer primary sequence information about q. In both methods, we assume that we actually have $closeness(RM_z(s), RM_z(q))$ instead of $closeness(RM_z(s), RP_z(q))$.

The first method is based upon the work of Nei and Li (Nei & Li 1979). The basic assumption in this model is that all the sequences in D and the underlying sequence $s(q)$ of isolate q are derived from a common ancestor. Sequences diverge over time as nucleotides mutate via a Poisson process. As time goes on, the number of shared sites between q and s typically decreases as more and more nucleotides mutate. Nei and Li develop methods for using the number of common sites present in s and $s(q)$ to determine the amount of time that has progressed since s and $s(q)$ shared a common ancestor. Using the Poisson process, it is then possible to compute the number of changes that have occurred in each nucleotide position from the common ancestor to both s and $s(q)$. This can then be used to estimate the primary sequence similarity between s and $s(q)$.

The second method is based upon Bayesian analysis. Given the primary sequence similarity α between s and $s(q)$ and the assumption that each nucleotide position in s is identical to the same nucleotide position in $s(q)$ with probability α, we can compute the probability distribution on the random variable that represents the number of sites common to both restriction maps. However, in our setting, we actually have the number of common sites shared by the restriction maps of s and $s(q)$, and we desire to compute the primary sequence similarity α between s and $s(q)$. If we are also given the fact that the primary sequence similarity between s and $s(q)$ is drawn from a known probability distribution X, we can compute a probability distribution on α using a Bayesian analysis.

Unfortunately, both of these methods have several flaws which limit their applicability to biological databases such as RDP. First, the assumptions on evolution are too restrictive. In both cases, only mutations are considered; insertions and deletions are not allowed. Furthermore, both models assume that nucleotides mutate with a given probability distribution which we have access to. This is an extremely strong assumption which does not seem to be justifiable in a general setting. Finally, the Bayesian analysis method does little to account for the fact that biologically similar sequences s and s' are likely to either both share a site with $s(q)$ or both not share a site with $s(q)$. For these reasons, it seems unlikely either method can be used in general database settings reliably. Indeed one fundamental assumption that is shared by both models is that the number of common sites shared by sequences s and $s(q)$ is a binomial random variable with p based upon the primary sequence similarity between s and $s(q)$. However, our experiments with the RDP database indicate the number of common sites does not seem to follow this binomial distribution (we omit this data from this writeup). As a result, we explore other methods for inferring information about $s(q)$ and use experimental means to evaluate these methods.

Closest Sequence Methods

In this section, we focus on using the set of sequences $CLOSE(q, D, 0)$ to infer information about the query isolate q. The basic premise behind this method is that sequences which have primary sequences that are most identical to $s(q)$ are the ones most likely to be in the set $CLOSE(q, D, 0)$. The extended premise is that the similarity of $s(q)$ to any sequence in $CLOSE(q, D, 0)$ is, with high probability, lower bounded by the maximum pairwise dissimilarity between any two sequences in $CLOSE(q, D, 0)$.

We use this basic premise with two different biological metrics of similarity. The first metric is primary sequence similarity. We say that two sequences s and s' have primary sequence similarity $sim(s, s') = \alpha$ for $0 \leq \alpha \leq 1$ if $100\alpha\%$ of the corresponding nucleotide positions in s and s' are identical. Define $sim(CLOSE(q, D, 0)) = \min_{s_i, s_j \in CLOSE(q,D,0)} sim(s_i, s_j)$. Finally, define $sim(q, CLOSE(q, D, 0))$ to be the quantity $\min_{s \in CLOSE(q,D,0)} sim(q, s)$. The extended premise implies that $sim(CLOSE(q, D, 0))$ lower bounds $sim(q, CLOSE(q, D, 0))$ with high probability.

The second metric is a biological family similarity that is based on a biological classification hierarchy for the database D. Each sequence in D is classified by an x digit number for $1 \leq x \leq 5$ (note each digit may have a different range of values). We say that two sequences s and s' have level similarity $level(s, s') = i$ if their classifications are identi-

cal to the i^{th} digit. Define $level(CLOSE(q, D, 0)) = \min_{s_i, s_j \in CLOSE(q,D,0)} level(s_i, s_j)$. Finally, we define $level(q, CLOSE(q, D, 0))$ to be the quantity $\min_{s \in CLOSE(q,D,0)} level(q, s)$. The extended premise implies that $level(CLOSE(q, D, 0))$ lower bounds $level(q, CLOSE(q, D, 0))$ with high probability.

If the extended premise is true and if both $sim(CLOSE(q, D, 0))$ and $level(CLOSE(q, D, 0))$ are high, we can, with high confidence, infer fairly precise primary sequence information or classification information about q.

In general, it seems that $sim(q, CLOSE(q, D, 0))$ and $level(q, CLOSE(q, D, 0))$ should increase if the number of cleavage sites in $RP_z(q)$ increases. That is, it seems unlikely that dissimilar sequences will have restriction maps that are close to that of q. On the other hand, it also seems likely that $sim(q, CLOSE(q, D, 0))$ and $level(q, CLOSE(q, D, 0))$ will decrease in size as the number of cleavage sites in q increases. These relationships, however, are difficult to analyze mathematically, and thus it is difficult to generate conditions where our technique will be effective. As a result, we experimentally evaluate this basic procedure. As we see in the next section, our experiments indicate this method can be used to effectively infer useful information about roughly half the sequences in RDP and that the error rate is quite small.

Experimental Results

Experimental Procedure

The sequence data used in this study was obtained from the Ribosomal Database Project (RDP) (release number 5 of May 17, 1995 (Maidak *et al.* 1994)). The RDP provides curated databases of ribosomal RNA related information and analysis services. The database used in this study was a subset of the bacterial 16S rRNA database distributed by the RDP. This database contains 16S rRNA sequence information from about 3000 different bacterial isolates and environmental samples. These sequences are distributed by the RDP as pre-aligned sequences with alignment gaps inserted so that homologous residues appear at the same position in all sequences. This alignment was produced by the RDP curators using, in addition to primary sequence similarity, secondary structure and other higher-order information. Inspection of the alignment indicates that highly diverged regions are often aligned to conserve putative secondary structures without regard to primary sequence. Most of the sequences in this database are incomplete, usually at the two ends. To construct the subset used here, incomplete sequences were removed after first selecting the region corresponding to positions 46 through 1406 of E. coli (Brosius *et al.* 1978). The resulting database contained 1575 sequences. Sequence similarity values were calculated for all pairwise combinations of the 1575 sequences. The similarity values clustered around the mean value of 72% identity, with a tail stretching toward higher similarity values (Figure 1).

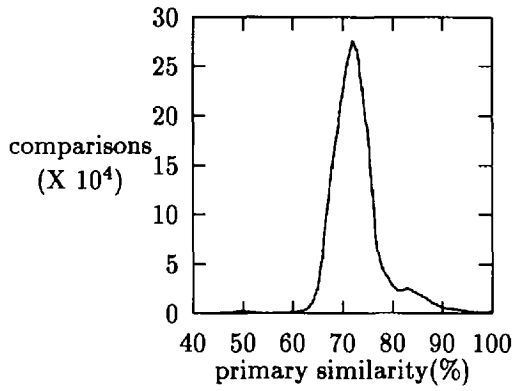

Figure 1: Pairwise sequence similarity. The pairwise sequence similarity was determined for all pairwise combinations of 1575 sequences in test database. Only sequence regions considered in further analysis were included. The similarity of the aligned pairs was determined using the pre-aligned sequences supplied by RDP.

Five separate restriction map data sets were computer generated from the 1575 sequences. For each data set, positions matching recognition sites from two commercial restriction enzymes were identified and the length between sites calculated (after removing alignment gaps). Ambiguity codons in the database sequences were treated as not matching any recognition site. The recognition sites (enzymes) chosen to generate the five sets were: AGCT + CAGT; CTAG + GATC; CUGC + ACGT; GTAC + TCGA; and TTAA + ATAT. The average number of sites in the combined 7875 digests was 10.66 with a median of 10 (Table 1).

The maps in a data set were chosen one at a time as queries. To simulate experimental data, the fragments sizes of the query were assumed to be accurate to only +/- 5%. The result set of database sequences where all sites could be matched was determined for each query, assuming the fragment order for the query was known (ordered query). These result sets were compared with the expected results calculated using the exact site positions from the aligned sequences. These ordered query result sets were missing one or more sequences found using pre-aligned sequences in 749 of the 7875 trials (9.5%). In addition to base changes (point mutations), rRNA genes have accumu-

site	seq.no	site	seq.no	site	seq.no	site	seq.no
1		11	904	21	4	31	
2	6	12	737	22	1	32	
3	19	13	714	23	5	33	1
4	86	14	605	24	1	34	
5	233	15	418	25		35	
6	367	16	284	26	1	36	
7	585	17	152	27		37	
8	802	18	57	28		38	
9	891	19	35	29		39	
10	955	20	11	30		40	

Table 1: Sequences by number of sites

lated insertions and deletions over the course of evolution. These insertions and deletions may have caused homologous fragments to no longer have sizes within the 5% error bound. The result of these size changes is apparently to cause pattern mismatches over shorter evolutionary distances than would be predicted from site conservation alone.

Another 312 of the 7875 trials (4.0%) produced result sets with extra sequences. Some of these extra matches may be due to peculiarities in the RDP alignment causing occasional site mismatch when comparing aligned sequences; however some extra matches are probably due to matching of non-homologous sites within the 5% error bound for the ordered query tests. Even if all of the additional matches are due to incorrectly pairing non-homologous sites, the percentage of query results affected is still relatively small.

The trials were repeated without assuming the order of query fragments was known (unordered query). Any differences in result sets between ordered and unordered queries represent incorrect pairing of non-homologous regions between query and database sequences (incorrect ordering). Only 266 of the 7875 trials (3.4%) produced result sets with extra sequences not in the ordered query result sets. However, if mismatches were allowed in the site matching, the results deteriorated. When up to one query and/or database site mismatch was allowed, 69.6% of the unordered result sets contained sequences not in the ordered query result sets. When up to two query and/or database site mismatches were allowed the percentage of result sets with incorrect matches increased to 78.4%. Because of this rapid increase in incorrect matches, only perfect matching data will be considered in the following experiments.

Similarity Between Query and Result Set

The average primary sequence similarity between unordered query and result was 94%. This was dependent, as expected, on the number of sites in the

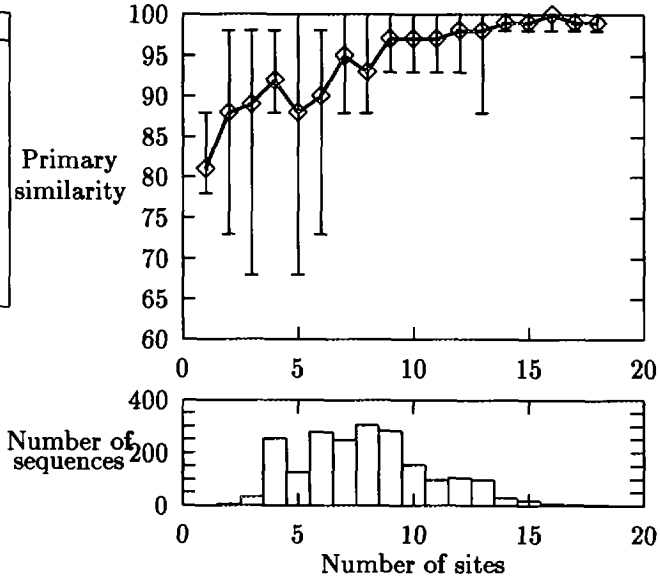

Figure 2: Primary similarity of selected sequences (unordered). The sequence similarity was calculated as described for figure 1 for every query versus each sequence in its result set. The data is shown by number of query sites as mean and range after discarding the 5% highest and lowest values. Bar chart at bottom indicates the number of queries with non-empty result sets.

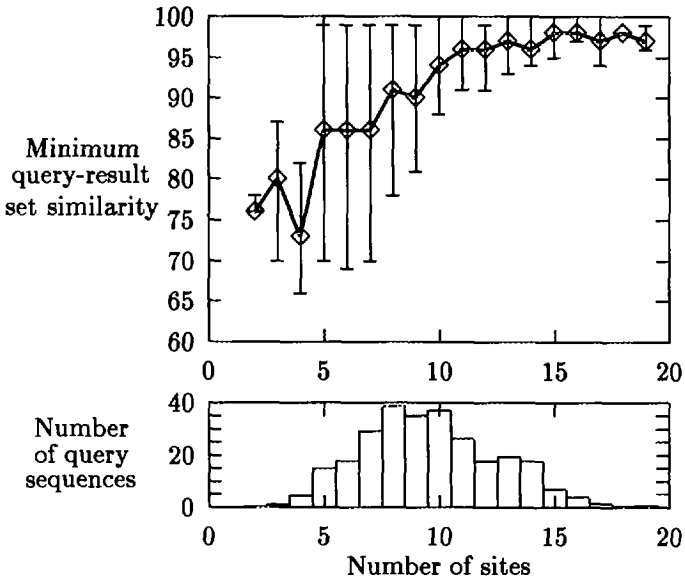

Figure 3: Minimum query - result set similarity. The minimum sequence similarity value between query and result set was calculated as in figure 1. These results are shown by number of query sites as mean and range after discarding the 5% highest and lowest values. Bar chart at bottom indicates the number of queries with result sets with size greater than one.

query (Figure 2). Seven or more sites were required to produce similarity values above 80% for more than 95% of the matches and 10 sites or more were sufficient to produce similarity values above 90% for more than 95% of such matches.

For result set containing several sequences, the sequence similarity between the query sequence and the least similar result sequence $sim(CLOSE(q, D, 0))$ provides a measure of how closely the query was identified. The average minimum similarity was 89.8%. The distribution of these minimum similarity values increases with number of sites in a manner similar to that seen in figure 2 (Figure 3). Although the average similarity was above 90% for queries with seven or more sites, nine or more sites were required to produce minimum similarity values above 80% for more than 95% of result sets and eleven sites were sufficient to produce minimum similarity values above 90% for more than 95% of the result sets.

Primary Sequence Similarity

In actual practice, the minimum similarity between query and result set could not be calculated since the query sequence would not be known. However, for result sets with more than one sequence, the minimum pairwise similarity between results can be calculated. This minimum result appeared to be a good predictor of the minimum query - database similarity (Table 2). A minimum query - result similarity of $> 80\%$ was observed for 99% (2456/2476) of result sets with $\geq 80\%$ minimum result - result similarity. For results sets with a minimum query - result similarity of $\geq 90\%$, 98% (1940/1988) also had minimum query - result similarities $\geq 90\%$.

sim(CLOSE(q,D,0))	*sim(q,CLOSE(q,D,0))*			
	90-100	80-89	70-79	60-69
90-100	1940	38	6	4
80-89	79	399	8	2
70-79		40	132	9
60-69			32	51

Table 2: Relationship between $sim(CLOSE(q, D, 0))$ and $sim(q, CLOSE(q, D, 0))$

Biological Classification Similarity

In addition to aligned rRNA sequence databases, the RDP also distributes phylogenetic information inferred from these sequences. This includes a hierarchical classification scheme consistent with the inferred phylogeny. At the most basic level, all sequences tested here are members of category 2, the Bacteria. There are 15 categories at the next level (2.1 through 2.15), 24 at the third level, 94 at the fourth, and 99 at the fifth and highest level. Not all sequences are categorized to all five levels, we assumed an implied category of 0 for all undefined levels.

level(q,CLOSE(q,D,0))	*level(CLOSE(q,D,0))*				
	1	2	3	4	5
1	149	1	3	2	22
2		46	3		20
3			319	42	29
4				374	41
5					2853

Table 3: Relationship between $level(q, CLOSE(q, D, 0))$ and $level(CLOSE(q, D, 0))$

By this classification scheme, 95% of these result sets contained sequences with identical classification for at least three levels (Table 3). Of these, 99% matched the query classification for the first three levels. 77% of these result sets were identical at all five levels; of these 95% match the query classification at all five levels. These results demonstrate the feasibility of typing bacterial 16S genes by restriction pattern. Although result sets were found for only about half the queries, in practice obtaining additional digestion patterns from an isolate is relatively simple. Of the 1575 separate query sequences considered here, only 287 produced empty result sets with all five digests. Many of these may not have close neighbors in the current database. In practice, isolates with no matches in several digests may be candidates for further characterization.

Acknowledgments

This work was partially supported by NSF grant BIR 912006 to the Center for Microbial Ecology.

References

Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. M.; and Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

Bellon, B. 1988. Construction of restriction maps. *CABIOS* 4:111–115.

Brosius, J.; Palmer, M. L.; Kennedy, P. J.; and Noller, H. F. 1978. Complete nucleotide sequence of a 16s ribosomal rna gene. *Proc. Natl. Acad. Sci. U.S.A* 75:4801–4805.

Cottrell, J. 1994. Protein identification by peptide mass fingerprinting. *Peptide Research* 7(3):115–124.

DeBry, R., and Slade, N. A. 1985. Cladistic analysis of restriction endonuclease cleavage maps within a maximum-likelihood framework. *Syst. Zool.* 34(1):21–34.

Durand, R., and Bregegere, F. 1984. An efficient program to construct restriction maps from experimental data with realistic error levels. *Nucleic Acids Res.* 12:703–716.

Felsenstein, J. 1992. Phylogenies from restriction sites: a maximum-likelihood approach. *Evolution* 46(1):159–173.

Fitch, W. M.; Smith, T. F.; and Ralph, W. W. 1983. Mapping the order of dna restriction fragments. *Gene* 22:19–29.

Grigorjev, A. V., and Mironov, A. A. 1990. Mapping dna by stochastic relaxation: a new approach to fragment sizes. *CABIOS* 6:107–111.

Holsinger, K. E., and Jansen, R. E. 1993. Phylogenetic analysis of restriction site data. *Methods in Enzymology* 224:439–455.

James, P.; Quadroni, M.; Carafoli, E.; and Gonnet, G. 1993. Protein identification by mass profile fingerprinting. *Biochemical and Physical Research Communications* 195(1):58–64.

Maidak, B. L.; Larsen, N.; McCaughey, M. J.; Overbeek, R.; Olsen, G.; Fogel, K.; Blandy, J.; and Woese, C. R. 1994. The ribosomal database project. *Nucl. Acids Res.* 22:3485–3487.

Miller, W.; Barr, J.; and Rudd, K. 1991. Improved algorithms for searching restriction maps. *CABIOS* 7(4):447–456.

Miller, W.; Ostell, J.; and Rudd, K. 1990. An algorithm for searching restriction maps. *CABIOS* 6(3):247–252.

Moyer, C. R.; Dobbs, F. C.; and Karl, D. M. 1993. Estimation of diversity and community structure through restriction fragment length polymorphism distribution analysis of bacterial 16s rrna genes from a microbial mat at an active, hydrothermal vent system. *Appl. Environ. Microbiol.* 60:871–879.

Nei, M., and Li, W. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Genetics* 76(10):5269–5273.

Pearson, W. R., and Lipman, D. J. 1988. Improved tools for biological sequence comparison. basic local alignment search tool. *Proc. Natl. Acad. Sci. U.S.A* 85:2444–2448.

Pearson, W. 1982. Automatic construction of restriction site maps. *Nucleic Acids Res.* 10:217–227.

Shaw, G. 1993. Rapid identification of proteins. *Proc. Natl. Acad. Sci. USA* 90:5138–5142.

Wright, L. W.; Lichter, J. B.; Reintz, J.; Shifman, M. A.; Kidd, K. K.; and Miller, P. L. 1994. Computer-assisted restriction mapping: an integrated approach to handling experimental uncertainty. *CABIOS* 10:443–450.