RIFLE: Rapid Identification of Microorganisms by Fragment Length Evaluation

Henning Hermjakob
GBF Braunschweig
Mascheroder Weg 1b
D - 38124 Braunschweig
hhe@gbf-braunschweig.de

Dr. Robert Giegerich
Technische Fakultät
University of Bielefeld
D - 33501 Bielefeld
robert@techfak.uni-bielefeld.de

Dr. Walter Arnold
Department of Biology
University of Bielefeld
D - 33501 Bielefeld
warnold@post.uni-bielefeld.de

Abstract

Biological macromolecules represent a valuable source of information for the identification and phylogenetic classification of microorganisms. One of the most commonly used macromolecules for this task is the 16S rDNA. The WWW-based RIFLE system presented here supports large-scale identification tasks by comparing 16S rDNA restriction patterns to a database of restriction patterns derived from sequence databases. Computing efficency and robustness against experimental errors are gained by emploing a new distance measure for restriction patterns, the fragment length distance. Results from the application of the system to the identification of uncultured microorganisms associated with the seagrass *halophila stipulacea* show the reliability of the method.

Introduction

Restriction fragment patterns of DNA and RNA molecules are easily obtained by enzymatic digestion and gel electrophoresis. Such patterns have a wide variety of uses, such as phylogenetic identification, food control, or plausibility checks subsequent to a PCR run (Gurtler, Wilson, & Mayall 1991; Weidner, Arnold, & Pühler 1995). The general setting is that a particular macromolecule, the query, is to be identified in a family of candidates of known origin via comparison of their restriction fragment patterns. This technique can be applied in an ad-hoc way as long as the family of candidates is rather small, data noise is low, and identification tasks arise only occasionally.

A recent computational attempt at this problem was presented at ISMB96 (Kim *et al.* 1996). However, due to an (as we show below) unfortunate notion of "closeness" between patterns, that approach runs into problems of algorithmic complexity, and does not appear to be a viable basis for a workable tool.

The RIFLE approach was designed to support large-scale identification tasks. Restriction patterns of the query are obtained by laboratory techniques. Patterns of known candidates are obtained from a sequence database. At the heart of RIFLE there is a new method of evaluating restriction pattern similarity, called *fragment length distance*, which is intuitively simple and has a number of pleasing mathematical properties. Great care is given to the correct handling of all kinds of error and uncertainty in the query and the database. We show that identification results are generally very good, and by using several digests with varying enzymes, the power of the method seems to be limited only by the quality of the reference databases. This approach is embedded in a state-of-the-art WWW interface. RIFLE can be accessed at URL http://bibiserv.techfak.uni-bielefeld.de/RIFLE/.

Overview of the RIFLE Method

Laboratory and computational procedures

Figure 1 shows an overview of the RIFLE method. We base the following discussion on the use of 16S rDNA, which is frequently used for the genetic identification of bacteria. In general, the process can be adapted to an arbitrary marker gene for which a sufficient number of related sequences is available.

In a first step, the DNA of the samples is isolated. As the subsequent PCR step specifically amplifies the 16S rDNA, the DNA isolation may be rather crude, even cell lysates may suffice. In the next step, a specific subsequence of the 16S rDNA is amplified using a universal 16S rDNA primer pair. The isolated PCR product is digested by a small number of restriction enzymes in seperate digests. Multiple digests are not used because they yield too many small fragments. Finally, the fragment lengths are determined using gel electrophoresis and subsequent silver staining. For each sample and each restriction enzyme used, a sorted list of fragment lengths is obtained. For a detailed description of the experimental method see (Weidner, Arnold, & Pühler 1995).

The generation of a database of theoretical restriction patterns faithfully models the laboratory procedure. The starting point is one of the publicly avail-

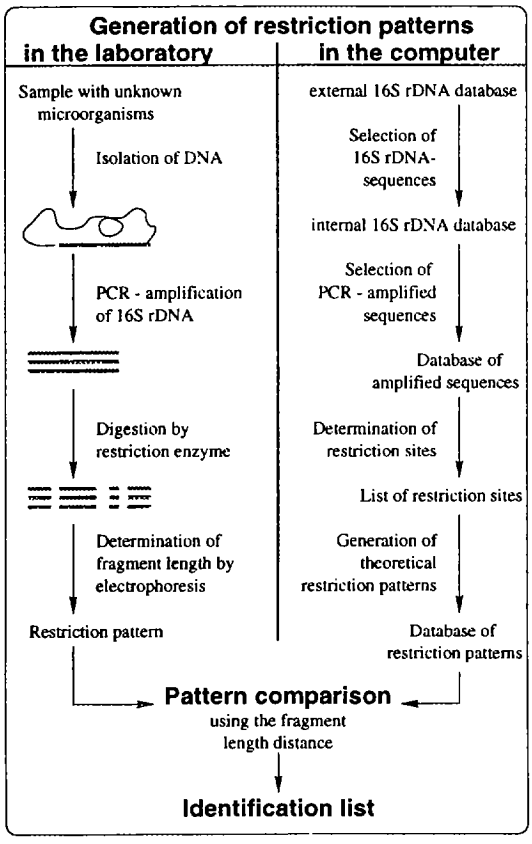

Figure 1: Overview of the RIFLE method

able 16S rDNA databases, e. g. the RDP database (Maidak *et al.* 1996). Only sequences satisfying certain quality criteria are entered in the internal 16S rDNA database. Next, the PCR amplification of a specific subsequence of the 16S rDNA is modeled. The subsequences between the binding sites of the primers (e. g. universal 16S rDNA primers) used in the laboratory are extracted and stored in the database of amplified sequences. For each of these sequences and the restriction enzymes defined in the RIFLE system, a list of restriction sites is determined. From these lists the theoretical restriction patterns are calculated and stored in a restriction pattern database.

To identify the query patterns obtained in the laboratory, they are compared against the theoretical restriction patterns of the database. The comparison is based on a new way of evaluating restriction pattern similarity, the *fragment length distance* described in detail below. For each query pattern, an identification list is displayed. This is a list of organisms whose theoretical restriction patterns are similar to the query pattern. Organism names are sorted in the order of decreasing similarity to the query. The smaller the distance between query pattern and theoretical pattern, the higher the probability of a correct identification of the query.

Sources of uncertainty and reliability of results

In the identification task described here, we must distinguish four sources of uncertainty:

1. Fragment patterns are more abstract than sequence data, and even a perfect match of fragment patterns does not necessarily imply sequence identity.

2. The experimentally obtained fragment lengths contain an experimental error of up to 10%, due to the methodical limitations of the fragment length determination by gel electrophoresis.

3. The sequence data in the database may contain errors. They also may contain ambiguous base symbols.

4. The query sequence may not be represented in the database at all, or it may be excluded from consideration due to to sequence incompleteness, high ambiguity, or PCR primer mismatch.

In the context of this uncertainty, it is particularly important that a measure of pattern similarity can be evaluated precisely (rather than by a heuristic approximation only). This ensures that there are *no false negatives* in the following sense: If the query does not achieve a close match to any of the sequences in the *restriction pattern* database, we can be sure that the species at hand is not among the family of candidates represented therein.

On the other hand, a close match of the query to one of the candidates can very well happen by chance. Let p_1 be the probability of such a match by chance for the restriction pattern obtained with enzyme 1, p_2 for enzyme 2, and so forth. The probability that k enzymes independently achieve close matches is therefore $p_1 * ... * p_k$, or conversely, the reliability of identification by close matches in k runs grows with $1 - p_1 * ... * p_k$. This increase in reliability is confirmed by the results shown below.

Application and Results

For the validation of the RIFLE approach, experiments were performed with several microorganisms currently investigated by Dr. Stefan Weidner, for which the sequence data were also obtained. Only with respect to this standard of truth, the quality of identification can be evaluated, and the experimental error that has to be handled can be measured precisely.

The 16S rDNA of 15 microorganisms associated with the seagrass *Halophila stipulacea* were amplified with the primer pair *R1n-U2*, which amplifies the 16S rDNA of *E. coli* from position 22 to 1085. Restriction patterns of the PCR products were generated by digestion with the enzymes *Hpa*II, *Hinf*I, *Hha*I and *Rsa*I (Weidner, Arnold, & Pühler 1995). The average error in the patterns so produced was 5%.

The nucleotide sequences of the 16S rDNAs were added to the 16S rDNA database of the RDP project (Maidak *et al.* 1996). RIFLE's resulting restriction pattern database contained entries for 1212 sequences. The RIFLE system was used to compare each restriction pattern generated in the laboratory against all theoretical restriction patterns generated from the nucleotide sequence database.

For each organism, Table 1 shows the ranking of the correct sequence in the RIFLE identification list. An organism has been correctly identified if its ranking equals 1. Rank '–' means that the organism was not contained in the first 20 items of the identification list. Together with the ranking, the fragment length distance between query and its correct counterpart is indicated in each column. Where several enzymes are used, the average distance is given.

The first two columns show the individual results obtained with the enzymes *Hpa*II and *Hinf*I. The next rows show the combined results obtained with two, three and four enzymes.

Due to the large number of candidates, a run with a single enzyme will typically produce many hits by chance, witnessed by the '–'-entries in the columns for *Hpa*II and *Hinf*I: The true candidate often receives a ranking greater than 20. With two restriction enzymes, 6 of 15 organisms are correctly identified. With three restriction enzymes, 11 organisms are correctly identified, and with four enzymes, 14 of 15 organisms are correctly identified.

These results show that microorganisms can be reliably identified by their 16S rDNA restriction patterns when separate digests with several restriction enzymes are evaluated. Three to four restriction enzymes per sample already yield reliable results.

Computation of Theoretical Restriction Patterns from the Database

The generation of theoretical restriction patterns has been designed following the example of the laboratory procedure. While the latter has already been described in detail in other works, the former procedure is described in more detail below.

Selection of 16S rDNA sequences

The publicly available 16S rDNA databases contain numerous partial sequences or sequences with many ambiguous base symbols of the IUPAC code. As such sequences are unsuitable for the generation of theoretical restriction patterns, only those sequences fulfilling certain quality criteria are selected in a preprocessing step and stored in the internal 16S rDNA database. Currently, a maximum of eight ambiguous base symbols per sequence is accepted. As RIFLE can process several input database formats, the sequences are stored in a standardised internal format to faciliate subsequent data handling.

Selection of PCR-amplified sequences

The next processing step models the PCR amplification of a 16S rDNA subsequence. For each sequence the subsequence between the binding sites of the PCR primers used in the laboratory is extracted. Parameters allow the adaption to laboratory conditions, e.g. the number of allowed mismatches between primer and binding site. As RIFLE can manage several input databases and several primers, for each database/primer combination a separate database of amplified sequences is built.

Determination of restriction sites

The determination of theoretical restriction sites corresponds to the digestion of the PCR product with restriction enzymes. To locate the position of restriction sites in the sequences, a standard approach to efficient pattern search in a text has been used, the suffix tree method (Giegerich & Kurtz 1995). Once the database has been converted into a suffix tree, restriction sites for arbitrary enzymes can be determined with an effort that is independent of the size of the database. The suffix tree algorithm had to be generalized due to the ambiguous base symbols of the IUPAC code. Two cases have to be distinguished:

- The recognition site of the restriction enzyme contains ambiguous base symbols of the IUPAC code.

  In this case, one restriction enzyme posesses multiple recognition sites. E.g. the enzyme *Ban*I with the recognition site G/GYRCC cuts the four oligonucleotides G/GTACC, G/GTGCC, G/GCACC, and G/GCGCC.

- The DNA sequence contains ambiguous base symbols of the IUPAC code.

  In this case, *potential* restriction sites arise. E.g. when using the restriction enzyme *Hpa*II with the recognition sequence C/CGG, the DNA sequence

Organism	Rank and distance of the source organisms in the identification list using the restriction enzyme(s)									
	*Hpa*II		*Hinf*I		*Hpa*II, *Hinf*I		*Hpa*II, *Hinf*I, *Rsa*I		*Hpa*II, *Hinf*I, *Hha*I, *Rsa*I	
	rank	dist.	rank	dist.	rank	dist.	rank	dist.	rank	dist.
L1	-	24	-	61	2	42	2	44	2	33
L2	7	12	-	41	1	26	1	25	1	51
L3	3	34	11	23	1	28	1	28	1	34
L8	-	167	4	97	8	132	2	100	1	79
L11	-	230	1	20	2	125	1	104	1	92
L15	7	80	-	49	2	64	1	59	1	92
L17	20	102	2	22	1	62	1	53	1	50
L18	-	88	5	61	3	74	1	74	1	76
L20	8	61	-	145	10	103	1	84	1	72
L21	-	71	-	69	1	70	2	59	1	53
L24	-	90	16	112	1	101	1	77	1	85
L26	20	40	-	36	2	38	1	54	1	57
L28	1	14	-	87	14	50	9	54	1	52
L30	-	46	13	83	2	64	1	73	1	67
L31	4	47	-	80	1	63	1	61	1	56

Table 1: Identification of 15 microorganisms using up to four enzymes

C/CRG constitutes a potential restriction site which will be treated appropriately in the calculation of theoretical restriction patterns.

Taking this problem into account, for each sequence and each restriction enzyme, a sorted list of secure and potential restriction sites is generated.

Calculation of theoretical restriction patterns

The analog step to the determination of fragment lengths in the laboratory is the calculation of theoretical restriction patterns from the list of restriction site positions. Be F_i the length of restriction fragment i. In the case where P_{i+1} is a secure restriction site, $F_i = P_{i+1} - P_i$.

If P_{i+1} is a potential restriction site, two possibilities have to be considered:

- P_{i+1} *is* a restriction site. Then two fragments $F_i = P_{i+1} - P_i$ and (subsequently) $F_{i+1} = P_{i+2} - P_{i+1}$ result.
- P_{i+1} is *not* a restriction site. Then only one fragment $F_i = P_{i+2} - P_i$ results.

Of course, this case distinction applies recursively at site P_{i+2}.

Therefore, one potential restriction site leads to two different theoretical restriction patterns which are termed *restriction pattern variants*. As restriction sites are independent of each other, each potential restriction site leads to a duplication of the number of variants; p possible restriction sites lead to 2^p variants. As a high number of variants results in a high probability of random similarities between one of the pattern variants and laboratory patterns as well as in a high load on system resources, sequences which yield too many pattern variants are not entered in the restriction pattern database.

The resulting restriction pattern database is used as a basis for the comparison between theoretical and laboratory restriction patterns.

The Fragment Length Distance

The classic notion of sequence distance

The edit distance of two sequences and its dual, similarity, are well-established concepts in bioinformatics (Waterman 1995). We recall the basic definitions: Let $\mathcal{A}$ be an alphabet, and '-' a gap character not in $\mathcal{A}$. An alignment of two sequences s and t with characters from $\mathcal{A}$ is an arrangement of s and t as a 2 by n matrix, such that the first row contains s, possibly interspersed with gap symbols, and the same holds for t in the second row. No column must contain gap symbols only. We denote such an alignment $(\bar{s}, \bar{t})$. A score function w has type $\mathcal{A} \cup \{'-'\} \times \mathcal{A} \cup \{'-'\} \to \Re$. For this paper, we require w to satisfy the axioms of a metric on

$\mathcal{A}$. The score of an alignment under score function w is $w(\overline{s},\overline{t}) = \sum_{i=1}^{n} w(\overline{s_i},\overline{t_i})$. An optimal alignment of s and t is one with minimal score, and the edit distance $D_w(s,t)$ of s and t is the score of an optimal alignment. If w is a metric on $\mathcal{A}$, then D_w is a metric on sequence space. $D_w(s,t)$ can be calculated in $O(n^2)$ time by the well-known dynamic programming scheme.

Typically, $\mathcal{A}$ is the DNA alphabet and w is the unit cost function, or $\mathcal{A}$ is the amino acid alphabet, and w is derived from the famous PAM matrices. In our application, the sequences at hand are restriction patterns. The alphabet here is the set of positive integers, with 0 taking the role of the gap symbol.

Distance of restriction maps and restriction profiles

So far we have used the term restriction pattern in an informal way. For the sake of a precise mathematical treatment, we now have to distinguish restriction maps and restriction profiles. We will, however, continue to use the term restriction pattern to refer to either maps or profiles.

Definition 1 (restriction map and profile)
A restriction map *is a sequence of nonnegative integers. It specifies the length of the restriction fragments of an underlying DNA sequence* in their original order.

A restriction profile *is a sequence of nonnegative integers in decreasing order. It specifies the length of the restriction fragments of an underlying DNA sequence,* irrespective of their original order.

Clearly, profiles are abstractions[1] from maps by ignoring the native fragment ordering. Having fragments in the profile ordered by size is a convention as well as major convenience, as we shall see.

Let σ_s be the abstraction function that transforms a restriction map s into a restriction profile by sorting its entries. $\sigma^{-1}(p)$ assigns to profile p the set of maps obtained by arbitrary permutations of p.

Definition 2 (fragment length score)
The score function fl is defined by $fl(x,y) = |x-y|$. *It is called the fragment length score.*

For comparing fragment lengths, this seems a natural, albeit simplistic definition. Note that $fl(x,0) = x$, i.e. a missing fragment is scored according to its length.

Definition 3 (fragment length distance)
Let s and t be both either restriction maps or restriction profiles. The fragment length distance *of s and t is* $D_{fl}(s,t)$, *i. e. their edit distance under the fragment length score.*

Example 1 illustrates this for two sequences that can be considered either as restriction maps or restriction profiles. It shows the dynamic programming matrix calculated for $s = (23,17,5)$ and $t = (24,17,10,5)$. $D_{fl}(s,t) = 11$, and the paths in the matrix indicate two alternative optimal alignments.

Example 1 (fragment length distance)

			t					
			24		17		10	5
		0	24		41		51	56
		↘						
	23	23	1		18		28	33
			↘					
s	17	40	18		1	→	11	16
					↘		↘	
	5	45	23		6		6	→ **11**

It is illustrative to see how this score function balances two sources of error, measurement error and lost fragments. For example, given the restriction maps 100, 20 and 100, 50, 10, our distance measure yields a score of 40, based on the optimal alignment 100-100, 50-20, delete 10. Given the similar length of the two smallest segments, one might vote to have those matched, leading to a (suboptimal) alignment 100-100, delete 50, 20-10, with a score of 60. But then, this alignment implies that a fragment of length 50 has been lost, while two much smaller fragments have been detected. Of course, there is no way to tell from our data what really has happened. In the first case, only an equivalent of the smallest segment is considered lost, and from this, the better score of the first alignment seems justified. Finally, note if we modify the example and give equal size to the smallest fragments in each map, both alignments come out even.

Our problem at hand is to compare restriction profiles (from gel electrophoresis) to restriction maps (from the sequence database).

Definition 4 (distance of profile and map)
The fragment length distance between a restriction profile p and a restriction map t is defined as $D_{fl}(p,t) = \min_s\{D_{fl}(s,t) | s \in \sigma^{-1}(p)\}$.

In order to exclude false negatives, this definition requires that the best possible reordering of the profile wrto the map must be considered. Note that $|\sigma^{-1}(p)| = n!$ when n is the length of p. Minimization over $\sigma^{-1}(p)$ in the setting of (Kim *et al.* 1996) leads to NP-completeness (claimed but not proved in (Kim *et al.* 1996)). The fragment length distance, however, has very convenient properties:

[1] (Kim *et al.* 1996) reserves the term restriction pattern for (unordered) multisets of fragment lengths. We found no use for this intermediate level of abstraction.

Theorem 1 (reordering Theorem)
$D_{fl}(p,t) = D_{fl}(p,\sigma_t(t))$

Proof: See Appendix A.

This settles an argument raised at ISMB96: Since gel electrophesis yields no ordering information anyway, all ordering information can be abstracted from in the algorithm. $\sigma_t(t)$ can be obtained in $O(n*log(n))$ time or even in $O(n)$ if we impose an uppper bound on the length of fragments, and with standard dynamic programming, $D_{fl}(p,\sigma_t(t))$ is obtained in $O(n^2)$.

In fact, we can do even better in the case when at least one of the two partners is a restriction profile rather than a restriction map:

Theorem 2 (profile evaluation Theorem) *Let p, q be two restriction profiles. Without loss of generality, let q be the shorter one, and let q' be q extended (if necessary) by 0s to achieve the same length n as p. Then,* $D_{fl}(p,q) = \sum_{i=1}^{n} fl(p_i,q_i)$

Proof: See Appendix A.

Hence $D_{fl}(p,q)$ can be calculated in linear time for two restriction profiles, as well as, due to Theorem 1, $D_{fl}(p,t)$ for a profile and a map. Note that this does not hold for two maps.

Comparing the fragment length distance to previous approaches

Previous approaches are based on a notion of fragment "identity": Two fragments are considered "identical" if their lengths differ by an amount less than a certain threshold value or a certain percentage of the fragment length. This holds for the widely used Sørensen-Dice coefficient (Jackson, Somers, & Harvey 1989) as well as the method of (Kim *et al.* 1996) (where the latter uses a percentage threshold and does not even achieve symmetry). The crux of these methods is that they behave in a chaotic way with data close to the threshold value. For example, for the two patterns $p = [400,300,200,100]$ and $q = [404,295,204,95]$, one obtains 100%, 50%, or 0% identity, depending on whether a threshold of 5, 4 or 3 is chosen. Conversely, if we fix the threshold value and vary the data lightly, anything — from a perfect match to a total mismatch — can happen. Such examples may seem contrived and might be recognized by an experienced person doing a singular experiment. However, in large scale identification tasks, no matter how the threshold value is chosen, such cases are bound to occur, and are likely to go unnoticed. $D_{fl}(p,t)$, by contrast, behaves in a continous way. No threshold value is necessary. In the above case, $D_{fl}(p,t) = 18$, and can change only slightly with slight changes in the data.

Another drawback of measures based on counting "identical" fragments is that they achieve only a small number of values (0, 0.25, 0.5, 0.75, and 1.0 are possible in the above example). The fragment length distance can differentiate in a much more fine-grained way. This is important for ranking closest neighbours when the family of candidates is rather large.

The fragment length distance is quite intuitive as it represents the sum of fragment differences. It would be desirable to give the fragment length score a probabilistic interpretation, as suggested by (Platt & Dix 1995) in the context of physical mapping, but is not clear how this could be achieved here. The restriction patterns are derived from a database of closely related sequences and therefore cannot be considered random.

The RIFLE Software

The RIFLE WWW interface is freely available on *bibiserv*, the recently established bioinformatics server located at Bielefeld, Germany. Just have a look at http://bibiserv.techfak.uni-bielefeld.de/RIFLE/.

Not all parameters of the RIFLE software can be described here. RIFLE features include

- individual adaption to laboratory processes, e.g. exactness of fragment length determination,
- screening against multiple restriction pattern databases in the same query,
- combination of multiple restriction enzymes to attain a high reliability of identifications,
- and a response time in the order of minutes.

Any additional sequence databases, primers and restriction enzymes can be integrated on demand. Appendix B gives an abbreviated example of a RIFLE identification list.

Conclusion

Our immediate goal is to further evaluate RIFLE by obtaining feedback from its users. A data management model must be devised, as each application creates its individual internal database, which must persist for some time. One possible extension is anticipated: According to Definition 3, RIFLE can also be used to evaluate two restriction maps rather than profiles against maps. Theorem 2 does not apply, and the computational effort goes up to $O(n^2)$. The laboratory work for obtaining (ordered!) restriction maps is much higher. But then, the information content in these data increases with $n!$, so there may be contexts where the extra effort is well spent.

Let us close with a little contemplative remark: While we are involved with the development of various tools for bioinformatics (e.g. (Dress, Perrey, & Füllen 1995), (Giegerich, Meyer, & Schleiermacher 1996)), the RIFLE approach has a particular charme, by which it has become our favourite example when it comes to explaining the nature of bioinformatics to outsiders or newcomers of this emerging field. It lies in the perfect analogy between biochemical laboratory procedures and the computational procedures, requiring a true integration of knowledge from molecular biology and computer science. A little mathematics is required to come up with a suitable method of fragment pattern comparison with good computational characteristics. Finally, a state-of-the-art WWW interface makes the resulting tool immediately accessible to the research community. This is what bioinformatics is all about: interdisciplinary, cooperative, distributed (and fun).

References

Dress, A.; Perrey, S.; and Füllen, G. 1995. A divide and conquer approach to multiple alignment. In Rawlings, C.; Clark, D.; Altman, R.; Hunter, L.; Lengauer, T.; and Wodak, S., eds., *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, 107–113. AAAI press.

Giegerich, R., and Kurtz, S. 1995. A comparison of imperative and purely functional suffix tree constructions. *Science of Computer Programming* 25(2-3):187–218.

Giegerich, R.; Meyer, F.; and Schleiermacher, C. 1996. Genefisher–software support for the detection of postulated genes. In States, D. J.; Agarwal, P.; Gaasterland, T.; Hunter, L.; and Smith, R. F., eds., *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, 68–77. AAAI press.

Gurtler, V.; Wilson, V. A.; and Mayall, B. C. 1991. Classification of medically important clostridia using restriction endonuclease site differences of pcr-amplified 16s rdna. *Journal of General Microbiology* 137:2673–2679.

Jackson, D. A.; Somers, K. M.; and Harvey, H. H. 1989. Similarity coefficients: Measures of co-occurrence and association or simply measures of occurrence? *The American Naturalist* 133:436–453.

Kim, J.; Cole, J. R.; Torng, E.; and Pramanik, S. 1996. Inferring relatedness of a macromolecule to a sequence database without sequencing. In States, D. J.; Agarwal, P.; Gaasterland, T.; Hunter, L.; and Smith, R. F., eds., *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, 125–133. AAAI press.

Maidak, B. L.; Olsen, G. J.; Larsen, N.; Overbeek, R.; McCaughey, M. J.; and Woese, C. R. 1996. The ribosomal database project (rdp). *Nucleic Acids Research* 24:82–85.

Platt, D., and Dix, T. 1995. A model for comparing genomic restriction maps. In *The 28th Hawaii International Conference on Systems Sciences*, 24–31.

Waterman, M. S. 1995. *Introduction to Computational Biology. Maps, Sequences and Genomes.* Chapman & Hall, London, UK.

Weidner, S.; Arnold, W.; and Pühler, A. 1995. Diversity of uncultured microorganisms associated with the seagrass halophila stipulacea estimated by restriction fragment length polymorphism analysis of pcr-amplified 16s rrna genes. *Applied and Environmental Microbiology* 62(3):766–771.

Appendix A: Proofs

Lemma 1 (shuffle and re-align)
Let s, t be sequences and π an arbitrary permutation. There is a permutation π' such that $D_w(\pi'(s), \pi(t)) \leq D_w(s,t)$.

Proof:
Let $w(\overline{s}, \overline{t}) = D_w(s,t)$. Choose a permutation π_1 such that its restriction to non-gap positions in $\overline{t}$ equals π. Note that $w(\pi_1(\overline{s}), \pi_1(\overline{t}))$ is merly a permutation of columns in $(\overline{s}, \overline{t})$. Let π' be π_1 restricted to the non-gaps in $\overline{s}$. Then, $(\pi_1(\overline{s}), \pi_1(\overline{t}))$ is an alignment of $(\pi'(s), \pi(t))$, but not necessarily an optimal one. Hence, $D_w(\pi'(s), \pi(t)) \leq D_w(s,t)$.

Lemma 1 holds fo an arbitraty score function w with additive gap costs. The subsequent lemmas rely on particular properties of the fragment length score fl.

Lemma 2 (decrease score by sorting step)
Let $(\overline{p}, \overline{t})$ be an alignment of profile p and map t. Let $j = i + 1$ and $t_j > t_i$. Transform $(\overline{p}, \overline{t})$:

1. *If $\overline{p_i}$ or $\overline{p_j}$ is a gap, exchange columns i and j.*
2. *Otherwise, exchange $\overline{t_i}$ and $\overline{t_j}$ (leaving $\overline{p}$ unchanged).*

Claim:
By this transformation the score of the alignment cannot increase.

Proof:
Case 1 is trivial, as the score is not affected. Note that the order of non-gaps in $\overline{p}$ is unchanged.

Case 2 has $\overline{p_i} \geq \overline{p_j} > 0$ and $t_j > t_i$. The score contribution of columns i and j is

- $c_{before} = |\overline{p_i} - \overline{t_i}| + |\overline{p_j} - \overline{t_j}|$
- $c_{after} = |\overline{p_i} - \overline{t_j}| + |\overline{p_j} - \overline{t_i}|$

Case analysis shows $c_{after} \leq c_{before}$:

- $\overline{p_i} \geq \overline{p_j} \geq \overline{t_j} \geq \overline{t_i}$:
  $c_{before} = c_{after}$
- $\overline{p_i} \geq \overline{t_j} \geq \overline{p_j} \geq \overline{t_i}$:
  $c_{before} = \overline{p_i} - \overline{t_i} + \overline{t_j} - \overline{p_j} \geq \overline{p_i} - \overline{t_i} + \overline{p_j} - \overline{t_i} = c_{after}$
- $\overline{p_i} \geq \overline{t_j} \geq \overline{t_i} \geq \overline{p_j}$:
  $c_{before} = \overline{p_i} - \overline{t_i} + \overline{t_j} - \overline{p_j} \geq \overline{p_i} - \overline{t_j} + \overline{t_i} - \overline{p_j} = c_{after}$

Other cases by symmetry.

Lemma 3 (sort t to profile)
Given an alignment $(\overline{p}, \overline{t})$, there is an alignment $(\overline{p}, \sigma_t(\overline{t}))$ with $fl(\overline{p}, \sigma(\overline{t})) \leq fl(\overline{p}, \overline{t})$.

Proof:
By successive application of Lemma 2, non-gaps in $\overline{t}$ can be sorted to $\sigma_t(t)$, while the order of non-gaps in $\overline{p}$ is not changed. Hence the resulting alignment is one of $(p, \sigma_t(t))$. The score cannot increase.

Lemma 4 (overall proof of Theorem 1)
$D_{fl}(p,t) := \min_s\{D_{fl}(s,t) | s \in \sigma^{-1}(p)\} = D_{fl}(p, \sigma_t(t))$

Proof:
Let $s \in \sigma^{-1}(p)$ such that $D_{fl}(s,t)$ is minimal. Let $A_1 = (\overline{s}, \overline{t})$ be an optimal alignment. Let $A_2 = (\pi(\overline{s}), \pi(\overline{t}))$ such that non-gaps in $\overline{s}$ are sorted in decreasing order. $fl(A_2) \leq fl(A_1)$ by Lemma 1. Now Lemma 2 applies to A_2. Let $A_3 = (\rho\pi(\overline{s}), \rho\pi(\overline{t}))$ such that non-gaps in $\overline{s}$ remain in order, and $\rho\pi(\overline{t}))$ is sorted. $fl(A_3) \leq fl(A_2)$. Note that A_3 is an alignment of $(p, \sigma_t(t))$. Hence $D_{fl}(p, \sigma_t(t)) \leq fl(A_3) \leq D_{fl}(s,t)$.

We now show by contradiction that inequality cannot hold. Assume $D_{fl}(p, \sigma_t(t)) = fl(\overline{p}, \sigma_t(t)) < D_{fl}(s,t)$.

Choose τ such that the original order of non-gaps in $\overline{t}$ is reestablished in $\tau(\sigma_t(t)$. Let $A_4 = (\tau(\overline{p}), \tau(\sigma_t(t)))$. By Lemma 1, $fl(A_4) < D_{fl}(s,t)$. Since A_4 aligns a permutation of p with t, this contradicts the minimality of s.

Lemma 5 (gaps go right)
There is an optimal alignment of two profiles (p,q) where gaps occur only at the end of $\overline{p}$ or $\overline{q}$.

Proof:
Let $A_1 = (\overline{p}, \overline{q})$ be an optimal alignment. Assume an internal gap, i.e. columns i and $j = i + 1$ with $\overline{p_i} \geq \overline{p_j}$ and $0 = \overline{q_i} < \overline{q_j}$. Their score contribution is $c_{before} = \overline{p_i} + |\overline{p_j} - \overline{q_j}|$. Let us construct A_2 by exchanging the gap with $\overline{q_j}$. We obtain $c_{after} = \overline{p_i} - \overline{q_j} + \overline{p_j}$. Case analysis shows that $c_{after} \leq c_{before}$:

- $\overline{p_i} \geq \overline{p_j} \geq \overline{q_j}$:
  $c_{before} = \overline{p_i} + \overline{p_j} - \overline{q_j} = \overline{p_i} - \overline{q_j} + \overline{p_j} = c_{after}$
- $\overline{p_i} \geq \overline{q_j} \geq \overline{p_j}$:
  $c_{before} = \overline{p_i} + \overline{q_j} - \overline{p_j} \geq \overline{p_i} - \overline{q_j} + \overline{p_j} = c_{after}$
- $\overline{q_j} \geq \overline{p_i} \geq \overline{p_j}$:
  $c_{before} = \overline{p_i} + \overline{q_j} - \overline{p_j} \geq \overline{q_j} - \overline{p_i} + \overline{p_j} = c_{after}$

Since the order of non-gaps is not changed, A_2 is an alignment of (p,q). Since A_1 is optimal, A_2 is optimal, and the interior gap has moved one position to the right. By iterating this construction, we obtain an optimal alignment of (p,q) with all gaps moved to the right.

The existence of an optimal alignment in this form immediately proves Theorem 2.

Appendix B: Example Output from the RIFLE Application

```
** Parameters **

* Databases *
Halophila stipulacea-associated microorganisms
RDP unaligned sequences, June 1995

* Primer *
R1n-U2

* Experimental artefacts *
20 base pairs before the forward primer
20 base pairs after the reverse primer

* Laboratory resolution *
Fragments shorter than
80 base pairs will be ignored in the theoretical patterns.
Subsequent theoretical fragments whose length difference is smaller than
4 % of the length of the shorter fragment will be considered as one fragment.

* Filtering of results *
Only organisms with an edit distance of at most
200 base pairs to the laboratory pattern will be displayed.
Only the 20 organisms with the smallest edit distances will be displayed.
Organisms for which restriction pattern data is missing for some of
the selected enzymes are skipped.
[...]

* Restriction pattern 0.3: L8 [477 346 207 86], digested with HpaII *
Accession Number       Distance   Variants Strain
M58818 (Release 77.0         59          4 Lactobacillus fructivorans.
L10942                       72          1 str. agg8.
M99704                       78          1 Lactobacillus acidophilus
M58820 (Release 77.0         78          2 Lactobacillus gasseri.
U02521                       80          1 Ehrlichia sp.
M26634 (Release 77.0         81          4 Desulfuromonas acetoxidans.
U03775                       81          1 Ehrlichia bovis.
M58744 (Release 77.0         82          2 Gardnerella vaginalis.
M75038                       88          4 Haemophilus actinomycetemcomitans.
M75035                       88          2 Haemophilus actinomycetemcomitans str. FDC Y4.
[...]

** Combined results **
* Pattern group 3: L8 *
Accession Number        Average      HpaII      HinfI       HhaI       RsaI Strain
-                            79        167         97         18         36 Weidner L8
L10942                       94         72         96         17        193 str. agg8.
M75053                      128        260        158         21         73 Pasteurella langaa.
M35017                      151        114        106        319         67 Actinobacillus
                                                                             lignieresii str. CM 2
L34628                      157        135        211         10        274 Eubacterium xylanophilum.
L34421                      188        204        150         79        322 Eubacterium ventriosum.
M75051                      189        160        160        261        175 Pasteurella dagmatis.
M59123 (Release 77.0        196        254        235         27        269 Haloanaerobium praevalens.
-                           200        102        328        257        113 Oceanospirillum pusillum.
```