Genome-scale DNA sequence recognition by hybridization to short oligomers

Aleksandar Milosavljević[1], Suzana Savković[2],
Radomir Crkvenjakov[3], David Salbego, Hope Serrato,
Heidi Kreuzer, Anne Gemmell, Shawna Batus[2],
Danica Grujić[4], Susan Carnahan, Jovanka Tepavčević

Center for Mechanistic Biology and Biotechnology
Argonne National Laboratory
Argonne, Illinois 60439-4833

Abstract

Recently developed hybridization technology (Drmanac *et al.* 1994) enables economical large-scale detection of short oligomers within DNA fragments. The newly developed recognition method (Milosavljević 1995b) enables comparison of lists of oligomers detected within DNA fragments against known DNA sequences. We here describe an experiment involving a set of 4,513 distinct genomic E.coli clones of average length 2kb, each hybridized with 636 randomly selected short oligomer probes. High hybridization signal with a particular probe was used as an indication of the presence of a complementary oligomer in the particular clone. For each clone, a list of oligomers with highest hybridization signals was compiled. The database consisting of 4,513 oligomer lists was then searched using known E.coli sequences as queries in an attempt to identify the clones that match the query sequence. Out of a total of 11 clones that were recognized at highest significance level by our method, 8 were single-pass sequenced from both ends. The single-pass sequenced ends were then compared against the query sequences. The sequence comparisons confirmed 7 out of the total of 8 examined recognitions. This experiment represents the first successful example of genome-scale sequence recognition based on hybridization data.

Introduction

How much information about the sequence of a particular DNA fragment is necessary in order to recognize similarity of the fragment to a known sequence in a DNA sequence database? What is the fastest and most economical way of obtaining partial sequence information that suffices for recognition? As the amount of sequenced DNA grows, these questions will become ever more relevant for the process of discovery in molecular biology and genetics.

Much of the practice of molecular genetics is currently based on three general types of DNA sequence comparisons: "wet" comparisons where pairs of long single-stranded DNA fragments are hybridized in a chemical experiment; "dry" comparisons in the computer where a sequenced fragment is compared against a database of known sequences; and "restriction fingerprint" comparsons where a particular DNA sequence fragment is cut by restriction endonucleases to obtain a specific set of fragment lengths, which can then be compared against other sets of fragment lengths that are either obtained experimentally or computed based on known sequences.

In addition to these currently most well-established methods, a method for reliable mutual comparison of clones based on their *hybridization signatures* has recently been proven in practice (Milosavljević *et al.* 1995). It has also been proposed (Lennon & Lehrach 1991; Drmanac *et al.* 1991) that a hybridization signature of a particular DNA fragment may be compared against a known DNA sequence. We here present the first large-scale verification of a newly proposed method (Milosavljević 1995b) of this kind.

A hybridization signature consists of real-valued hybridization scores of a set of short oligomer probes against a DNA fragment. The oligomers with the highest hybridization scores are taken as most likely to occur within the fragment. A list of highest-scoring oligomers is subsequently compared against known DNA sequences in order to determine significant similarities.

Two kinds of comparison methods of this kind have been proposed earlier by other authors: the first

[1] Present address: CuraGen Corporation
322 E. Main St., Branford, Connecticut 06405
email: amilosav@curagen.com
[2] Present address: University of Illinois at Chicago,
840 Southwood St., Chicago, Illinois, 60612
[3] Present address: HySeq Inc.
670 Almanor Ave., Sunnyvale, California 94086
[4] Present address: Department of Medicine,
Beth Israel Hospital, Boston, Massachusetts 02215

method is to count common occurrences of oligomers in the fragment and the candidate matching sequence (Lennon & Lehrach 1991; Drmanac *et al.* 1991); the second method is to reconstruct the sequence (completely or partially) based on oligomer overlaps and then to compare the reconstructed sequence against a candidate matching sequence (Drmanac *et al.* 1991). The first method ignores shared subwords between the oligomers altogether, while the second method utilizes shared subwords (oligomer overlaps) for the purpose of sequence reconstruction, which may be thought of as an intermediate step in the process of recognition of similarity. Most importantly, the second method considers shared subwords only prior to the moment when a candidate matching sequence is available.

In contrast to these previously proposed methods, the newly developed method that we apply in this paper (Milosavljević 1995b) simultaneously considers both shared subwords and the structure of the candidate matching sequence. The novelty of the recognition method applied here stems from the fact that it utilizes oligomer lists directly in a similarity search, thus avoiding sequence reconstruction as an intermediate step. It has been theoretically proven (Milosavljević 1995b) that a direct recognition method achieves superior performance in terms of specificity of recognition.

The significance of the hybridization-based recognition method stems from the ease with which large-scale hybridization experiments can be performed. In contrast to gel-based sequencing and restriction analysis, which are essentially one-dimensional separation experiments, the hybridization experiments do not require one-dimensional separation and thus can be economically conducted on a much larger scale by utilizing high-density two-dimensional arrays of immobilized DNA fragments (format 1 hybridization experiments) or oligomer probes (sequencing chip, or format 2 experiments (Fodor *et al.* 1993; Mirzabekov 1994; Southern, Maskos, & Elder 1992)). The data collection for our experiment was performed on a format 1 hybridization production line that was originally developed for the purpose of complete (Drmanac *et al.* 1993) and partial (Drmanac *et al.* 1991) sequencing by hybridization. Data collection throughput of over one million probe/target hybridization scores per day can be achieved in a laboratory of small size by utilizing current hybridization technology (Drmanac *et al.* 1994).

In this paper we present first experimental genome-scale verification of the recognition method. A set of 4,513 distinct genomic E.coli clones of length 2kb was selected out of a total pool of 15,328 clones based on hybridization signatures with 636 short oligomer probes. (The experiment was based on a total of $15,328 * 636 = 9,748,608$ individual clone/probe hybridization scores.) For each selected clone, a list of oligomers that are putatively identified in it was selected. A database consisting of the 4,513 oligomer lists was used for comparisons against known E.coli sequences. The queries used for database searches were overlapping windows covering 40kb of known E.coli genomic sequence. A total of 8 clones that exhibited outstanding similarity scores to particular query sequences were single-pass gel-sequenced. Out of 8 putative identifications, 7 were confirmed by comparing the single-pass sequence against the query sequence.

Methods

The sequence recognition method applied in our experiment has been described in detail in (Milosavljević 1995b). We here only summarize the most relevant features.

In order to compare a particular oligomer list against a particular sequence, a scoring function must be designed. Our method is based on a scoring function that utilizes potential oligomer overlaps: oligomers that significantly overlap may contribute more toward a score, provided their occurrrences in the candidate matching sequence also overlap. The recognition algorithm efficiently constructs contiguous stretches of overlapping oligomers that most resemble the candidate matching sequence, ignoring overlaps that do not resemble the sequence; the final score depends on the degree of achieved resemblance. The algorithm tolerates imperfect oligomer overlaps and imperfect similarity with the candidate matching sequence.

There is a large number of DNA sequence comparison methods (Hide, Burke, & Davison 1994; Pietrokovski, Hirshon, & Trifonov 1990; Pizzi *et al.* 1991; Blaisdell 1986; Quentin 1994; Pearson & Lipman 1988) that are also based on subword composition. Such methods may be viewed as completely opposite of our recognition method: in such methods sequences are *broken* into subwords for the purpose of efficient sequence comparison, while in our method the set of oligomers is implicitly treated as a sequence.

The first step of the recognition method is basically the same as the first step in sequencing by hybridization (Drmanac & Crkvenjakov 1993): clones that are few hundred to few thousand bases long are hybridized with oligomer probes of short length under conditions that enable approximate discrimination of those probes whose complement is present in the sequence of the clone from the probes whose complement is not present. The oligomers that are detected to occur in the sequence by this method can then be used to reconstruct the sequence. However, if we were to use erroneously reconstructed sequences (even partially reconstructed) for similarity searches, we may not identify matching sequences.

The error is decreased in our method due to the elimination of the sequence reconstruction step: lists of oligomers are used directly, without any *a priori* commitment to a particular sequential arrangement. This point is illustrated in Figure 1: an optimally recon-

structed sequence (optimal in the sense that it is the shortest superstring with up to 2 mismatches) achieves a score of 14 matches, a total of 4 matches less than the recognition score of 18. The difference is due to the fact that in case of recognition the oligomers can be assembled so that they most resemble the original sequence, while the sequence reconstruction process does not utilize the information about the candidate matching sequence.

The intuitively appealing argument to the effect that oligomer lists should be used directly in sequence similarity searches can be made more formal (Milosavljević 1995b). The formal argument is based on the fact that sequence reconstruction is an algorithmic step, and thus can only destroy information about the original sequence.

While the information-theoretic argument about using the oligomer lists directly is both intuitively and theoretically clear, it is not obvious that it is practical. Fortunately, simple and efficient algorithms for direct comparison of oligomer lists against known sequences exist. One type of algorithm (described in (Milosavljević 1995b)) is based on a variant of the algorithm that was used for sequence comparisons (Milosavljević 1995a). In the following, we verify the algorithm in a genome-scale recognition experiment.

Experimental results

The E.coli library was prepared by partial SauIIIa digestion followed by size selection for 2kb fragments. The fragments were subcloned in plasmids, which were then used for transfection of E.coli hosts. The E.coli hosts were plated, manually picked and grown in 96-well plates. The plasmid inserts were then amplified by PCR. Individual PCR products were arrayed using a robotic spotting technique at a density of 31,104 spots per 16 × 24 cm nylon filter as described in (Drmanac *et al.*, 1994).

The data used in this presentation is based on a single type of filter. A total of 7,664 distinct PCR products were spotted in duplicate, occupying a total of 15,328 spots on the filter. The remaining spots on the filter contained similar PCR products from a number of related bacteria plus a small fraction of control clones of known sequence; these spots were not considered in the present study. A total of 20 physical copies of this type of filter were made in order to be able to perform hybridizations with different probes in parallel.

A total of 636 short oligomer probes were consecutively hybridized with the filters as described in (Drmanac *et al.*, 1993). Most of the probes consisted of mixtures of 33P-labeled oligomers of general formula $(N)_{0-2}(B)_7(N)_{0-2}$, where B denotes a specific base and N denotes a degenerate position. Hybridization was conducted under conditions that enable approximate detection of single-basepair mismatches. Phosphorimaging technology was applied to scan the hybridized filters for radioactivity. Specialized image-analysis programs developed by J. Jarvis and R. Drmanac (unpublished) were used for automated hybridization scoring and tracking of individual spots. The complete record of the experiment, including the hybridization scores, was then semi-automatically stored in a Sybase relational database. The data were analyzed by a suite of C, C++, and UNIX C-shell programs with a direct access to the database.

In order to control redundancy of coverage of the E.coli genome, a set of distinct clones was selected based on a clustering of hybridization signatures. A hybridization signature, consisting of hybridization intensities with each of 636 probes was compiled for each spot. The hybridization intensities within a signature were obtained by applying mass scaling and rank scaling, as described in (Milosavljević *et al.* 1995). Due to duplicate spotting, each clone was represented by two signatures. Approximately 20% of signatures of poor quality were eliminated; the remaining signatures were grouped into 4,513 clusters of highly overlapping or identical clones by applying the clustering algorithm described in (Milosavljević *et al.* 1995). A representative member was then selected from each cluster for inclusion in the database.

For each selected signature a list consisting of 140 oligomer probes that exhibit the highest intensities was compiled; such lists were augmented by additional 140 reverse complementary oligomers due to the fact that both strands of PCR products were spotted and oligomer orientation could not be resolved. The 4,513 lists consisting of 280 oligomers each were used as a database of E.coli genome. Known E.coli genomic DNA sequences were then used to query the database.

A 40kb segment from the entry ECD077.15 in the E.Coli Database Collection (Wahl & Kroeger 1994) (depicted in in Figure 2) was split into overlapping windows of length 2kb. Overlap between consecutive windows was 1.8kb. Each of the windows was in turn used as a query sequence for the database of 4,513 oligomer lists.

Each query sequence was compared against the oligomer lists using the method described in (Milosavljević 1995b). The degree of match between an oligomer list s and a target sequence t was measured using the scoring function $I(s;t)$, which is an estimate of algorithmic mutual information between the oligomer list and the candidate matching sequence. Two parameters were considered for each query sequence: the top $I(s;t)$ score with a particular oligomer list and the difference between the top score and the second highest score, termed absolute and relative scores, respectively. A list of all the clones that resulted in a relative score of 10 bits or more were further considered. This resulted in a list of 11 putative recognitions, as listed in Table 1.

In order to test the accuracy of recognition, the putatively recognized clones were single-pass gel-sequenced from both ends on an ABI sequencer (average sequenc-

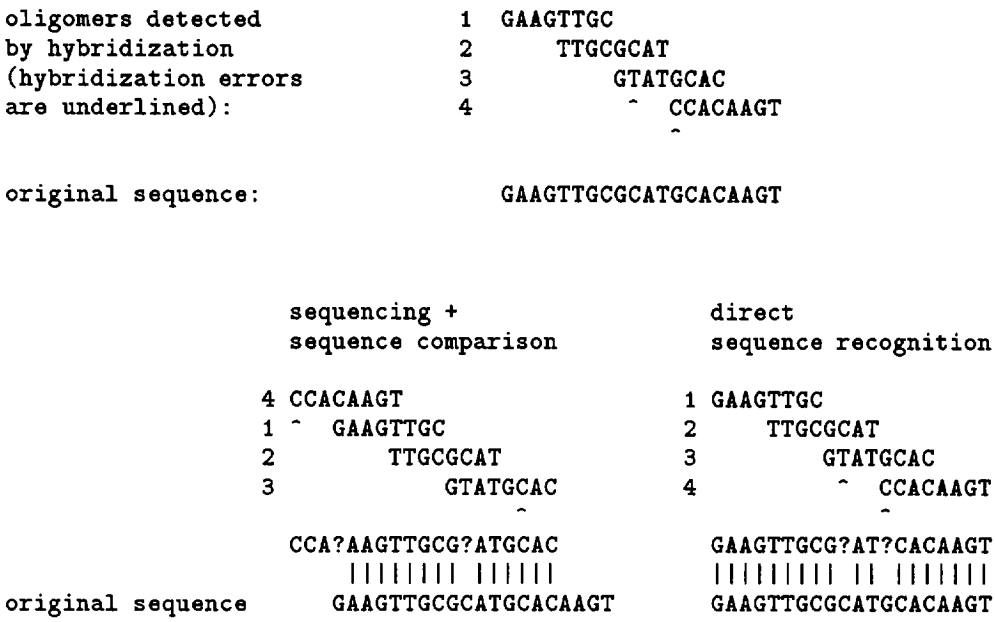

Figure 1: Sequencing vs. sequence recognition. In this hypothetical example, four oligomers are (imperfectly) detected by hybridization (top). The oligomers can be used to first optimally reconstruct sequence fragment and then compare the reconstructed sequence fragment against the original sequence (bottom left) or to directly compare oligomer lists against the original sequence (bottom right). Note that the optimally reconstructed sequence (bottom left) has only 14 matching letters with the true sequence, while the same list of oligomers can be assembled so that it achieves 18 matches with the true sequence (bottom right).

ing run was around 300bp). Three clones could not be sequenced due to technical reasons. Correct recognition was confirmed in 7 out of a total of 8 sequenced clones, as summarized in Table 1.

Discussion

We have presented a genome-scale verification of a newly proposed method (Milosavljević 1995b) for sequence recognition by comparison of hybridization signatures against known sequences. The present study shows that the highest-scoring clones can be recognized based on hybridizations with 636 probes if the database contains an equivalent of multiple bacterial genomes.

While the current study does not address the problem of false negatives, we were able to recognize clones covering about one half of the 40kb query fragment. While the missing clones may simply not be present in our database, we feel that additional hybridizations would enable a significant number of additional recognitions. An interesting open question, which is to be resolved experimentally, is how many more probes must be hybridized in order to recognize almost all the clones covering a particular sequenced region.

The verification experiment presented here included as a preliminary step an application of the previously established method for clone clustering based on hybridization signatures (Milosavljević *et al.* 1995). The combination of clone clustering and sequence recognition exemplifyed in our experiment provides a basis for novel approaches to the study of genome structure and function.

Clone clustering and sequence recognition can be jointly applied in a novel strategy for studying gene expression (Milosavljević *et al.* 1996). A currently very popular method to study patterns of gene expression is to partially sequence randomly selected clones from a particular cDNA library (Adams *et al.* 1991): the abundance of a particular clone (indicating the expression level of a particular gene) is measured by the number of times the clone is resequenced; the sequence is also used for homology searches against DNA sequence databases. While the current large-throughput sequencing methodology enables sequencing of hundreds of thousands of cDNA fragments per year, a much more rapid and economical method is needed in order to quantitatively study the expression of genes across different cell types, developmental stages, and physiological conditions. Large-throughput hybridization experiments (Drmanac & Drmanac 1994) combined with the clustering method (Milosavljević *et al.* 1995) and the sequence recognition method demonstrated in this paper provide a much higher throughput

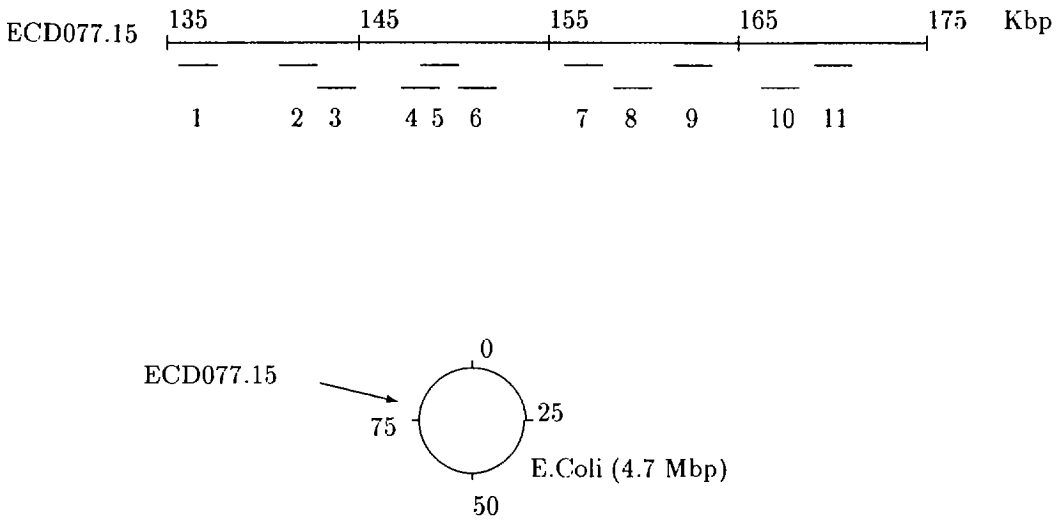

Figure 2: Positions of 11 windows (numbered 1-11) that exhibited highest recognition scores agaist particular clones. The windows are positioned within the 40kb fragment from the locus ECD077.15. The window numbers can be used for cross-reference with Table 1, which gives more detailed information.

and are much more economical. A recent study of gene expression in infant brain (Milosavljević *et al.* 1996) confirms the viability of this new strategy.

On a more general level, as the amount of sequenced DNA grows, a typical sequencing experiment will in the near future typically result in the *resequencing* of an already sequenced fragment. Thus, instead of obtaining the information about a completely new sequence, a sequencing experiment will result in the recognition of similarity to an already known sequence. The critical observation here is that the amount of information necessary for recognition is much less than for determination of an unknown sequence: assuming a 300-basepair DNA sequence fragment, and a database containing the complete $3,000,000,000$-basepair Human genome, the amount of recognition information is $log_2 3,000,000,000 \approx 32$ bits (enough to store a pointer to the occurrence fo the fragment in the database) whereas the sequencing experiment gives a total of $300 * 2 = 600$ bits of information, a nearly 20-fold redundancy over 32 bits. In contrast, sequence recognition by hybridization to a set of oligomers may give just the right amount of information in a much more economical way. For example, an individual clone may be simultaneously hybridized to a set of oligomers immobilized on a sequencing chip (format 2 hybridization experiment). The number of immobilized oligomers may be chosen so as to give just the right amount of information that is necessary for recognition.

Acknowledgements

This work was supported by the U.S. Department of Energy, Office of Health and Environmental Research, under Contract W-31-109-ENG-38.

References

Adams, M.; Kelley, J.; Gocayne, J.; Dubnick, M.; Polymeropoulos, M.; Xiao, H.; Merril, C.; Wu, A.; Olde, B.; Moreno, R.; Kerlavage, A.; McCombie, W.; and J.C., V. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* 252:1651–1656.

Blaisdell, B. 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences U.S.A.* 83:5155–5159.

Drmanac, R., and Crkvenjakov, R. 1993. Method of sequencing of genomes by hybridization of oligonucleotide probes. US patent No. 5,202,231.

ECD077.15 135-175 Kbp					
win. num.	1	2	3	4	5
win. pos. (bp)	135601-137600	140801-142800	142801-144800	147201-149200	148201-150200
rel. score (bits)	14	27	13	15	10
true pos. (bp)	-	140749-142751	142851-144893	147228-149230	148686-150370
6	7	8	9	10	11
154201-156200	155801-157800	158401-160400	161601-163600	166201-168200	169001-171000
21	13	36	15	10	18
154229-156408	155327-157609	–	?-?	–	169250-171262

Table 1: Detailed information for individual recognitions. Windows within the 135-175 Kbp fragment from the E. Coli locus ECD077.15 are numbered for cross-reference with Figure 2. For each individual recognition, three items are listed: position of the window, relative score (in bits) of the best-matching clone, and the exact position of the clone, as determined by sequencing from both ends. Question marks denote sequenced fragments that do not resemble sequence close to the matching window. In three cases, denoted by bars, clones could not be sequenced due to technical reasons.

Drmanac, S., and Drmanac, R. 1994. Processing of cDNA and genomic kilobase-sized clones for massive screening, mapping and sequencing by hybridization. *BioTechniques* 7:328–336.

Drmanac, R.; Lennon, G.; Drmanac, S.; Labat, I.; Crkvenjakov, R.; and Lehrach, H. 1991. Partial sequencing by hybridization: Concept and applications in genome analysis. In *The First International Conference on Electrophoresis, Supercomputing and the Human Genome*, 60–74. World Scientific, Singapore.

Drmanac, R.; Drmanac, S.; Strezoska, Z.; Paunesku, T.; Labat, I.; Zeremski, M.; Snoddy, J.; Funkhouser, W.; B., K.; Hood, L.; and Crkvenjakov, R. 1993. DNA sequence determination by hybridization: A strategy for efficient large-scale sequencing. *Science* 260:1649–1652.

Drmanac, R.; Drmanac, S.; Jarvis, J.; and Labat, I. 1994. Sequencing by hybridization. In Venter, J., ed., *Automated DNA Sequencing and Analysis Techniques*. New York: Harcourt Brace Jovanovich. 29–36.

Fodor, S.; Rava, R.; Huang, X.; Pease, A.; Holmes, C.; and Adams, C. 1993. Multiplexed biochemical assays with biological chips. *Nature* 364:555–556.

Hide, W.; Burke, J.; and Davison, D. 1994. Biological evaluation of d^2, and algorithm for high-performance sequence comparison. *Journal of Computational Biology* 1(3):199–215.

Lennon, G., and Lehrach, H. 1991. Hybridization analyses of arrayed cDNA libraries. *Trends in Genetics* 7(10):314–317.

Milosavljević, A.; Strezoska, Z.; Zeremski, M.; Grujic, D.; Paunesku, T.; and Crkvenjakov, R. 1995. Clone clustering by hybridization. *Genomics* 27:83–89.

Milosavljević, A.; Zeremski, M.; Strezoska, v.; G̃rujić, D.; Dyanov, H.; Gemmell, A.; Batus, S.; Salbego, D.; P̃aunesku, T.; Soares, B.; and Crkvenjakov, R. 1996. Discovering distinct genes represented in 29,570 clones from infant brain cDNA libraries by applying sequencing by hybridization methodology. *Genome Research* 6:132–141.

Milosavljević, A. 1995a. Discovering dependencies via algorithmic mutual information: a case study in DNA sequence comparisons. *Machine Learning Journal* 21:35–50.

Milosavljević, A. 1995b. DNA sequence recognition by hybridization to short oligomers. *Journal of Computational Biology* 2(2):355–370.

Mirzabekov, A. 1994. DNA sequencing by hybridization – a megasequencing method and a diagnostic tool? *Trends in Biotechnology* 12:27–32.

Pearson, W., and Lipman, D. J. 1988. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences U.S.A.* 85:2444–2448.

Pietrokovski, S.; Hirshon, J.; and Trifonov, E. 1990. Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. *Journal of Biomolecular Structure and Dynamics* 7(6):1251–1268.

Pizzi, E.; Attimonelli, M.; Liuni, S.; Frontali, C.; and Saccone, C. 1991. A simple method for global sequence comparison. *Nucleic Acids Research* 20:131–136.

Quentin, Y. 1994. Fast identification of repetitive elements in biological sequences. *Journal of Theoretical Biology* 166:51–61.

Southern, E.; Maskos, U.; and Elder, J. 1992. Analyzing and comparing nucleic acid sequences by hybridization to arrays of iligonucleotides: evaluation using experimental models. *Genomics* 13:1008–1017.

Wahl, R., and Kroeger, M. 1994. Echerichia coli database collection, release 20.