BIOINFORMATICS APPLICATIONS NOTE

Vol. 00 no. 00 2005
Pages 1–2

DNA-BAR: Distinguisher Selection for DNA Barcoding

B. DasGupta[a*] K.M. Konwar[b] I.I. Măndoiu[b†] and A.A. Shvartsman[b]

[a]Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607-7053, and
[b]Computer Science and Engineering Department, University of Connecticut, 371 Fairfield Rd., Unit 2155, Storrs, CT 06269-2155.

ABSTRACT

Summary: DNA-BAR is a software package for selecting DNA probes (henceforth referred to as *distinguishers*) that can be used in genomic-based identification of microorganisms. Given the genomic sequences of the microorganisms, DNA-BAR finds a near-minimum number of distinguishers yielding a distinct hybridization pattern for each microorganism. Selected distinguishers satisfy user specified bounds on length, melting temperature, and GC content, as well as redundancy and cross-hybridization constraints.

Availability: DNA-BAR can be used online through the web interface provided at http://dna.engr.uconn.edu/~software/DNA-BAR/. The open source C code, released under the GNU General Public License, is also available at the above address.

Contact: ion@engr.uconn.edu

INTRODUCTION

String barcoding is a recently introduced technique for genomic-based identification of microorganisms such as viruses or bacteria from among a set of previously sequenced microorganisms. Applications of this technique range from rapid pathogen identification in epidemic outbreaks to point-of-care medical diagnosis to monitoring of microbial communities in environmental studies (see (2; 7) and references therein). Microorganisms identification can be performed by spotting or synthesizing on a microarray the Watson-Crick complements of the distinguisher strings and then hybridizing to the array the fluorescently labeled DNA extracted from the unknown microorganism. Under the assumption of perfect hybridization stringency, the hybridization pattern can be viewed as a string of zeros and ones, referred to as the *barcode* of the microorganism. For unambiguous identification, distinguishers must be selected such that each microorganism has a distinct barcode.

Since it is difficult to ensure perfect hybridization stringency with current microarray technologies, a method for improving identification robustness is to use redundant distinguishability, e.g., to require that every two barcodes differ in at least r positions, where r is a given integer. Further improvements in identification robustness can be obtained by using a multi-step assay similar to those used for Single Nucleotide Polymorphism genotyping (5). First, primers complementing selected distinguishers are hybridized in solution with unlabeled DNA extracted from the unknown microorganism. Then, primer hybridizations are registered via a single-base extension reaction using the polymerase enzyme and fluorescently labeled dideoxynucleotides. Formed duplexes are separated by heating, and the resulting mixture is hybridized to a microarray containing the distinguishers. Finally, microarray fluorescence levels are used to learn the identity of extended primers and thus determine the barcode of the microorganism. The increased reliability of this multi-step assay comes from two sources. First, solution-based reactions are better understood and much easier to optimize compared to solid-phase hybridization. Second, the relevant oligonucleotides involved in the solid-phase hybridization step have much lower complexity compared to the whole genome of the microorganism, and are fully under the assay designer's control.

DNA-BAR is a tool for selecting sets of distinguishers to be used in this type of identification assays. The tool accepts as input genomic sequences, possibly containing degenerate bases, given either in Fasta format (http://ngfnblast.gbf.de/docs/fasta.html) or interactively entered by the user. Subject to the given barcode redundancy requirements, the tool attempts to minimize the number of distinguishers, since this reduces assay cost and enables higher effective primer concentration in the solution-based assay steps. The tool enforces user specified lower and upper bounds on distinguisher length, melting temperature, and GC content. The tool also enforces cross-hybridization constraints between extended primers and non-complementary distinguishers on the microarray using a hybridization model based on nucleation complex theory (1). According to this model, hybridization between two oligonucleotides can take place only if one contains as substring the reverse Watson-Crick complement of a substring of weight $> c$ of the other, where c is a given constant. The weight of a string is the number of *weak* bases (A and T) plus twice the number of *strong* bases (G and C).

ALGORITHM AND IMPLEMENTATION

We use a simple greedy distinguisher selection strategy – in every iteration we pick a substring that distinguishes the largest number of not-yet-distinguished pairs of genomic sequences. After selecting a distinguisher d, we discard all candidates that have in common with d a substring of weight $> c$. To achieve high scalability, we use an incremental algorithm for quickly generating a representative set of candidate distinguishers and collecting all their occurrences in the given genomic sequences, and employ a "lazy" strategy for updating coverage gains in the greedy selection phase of algorithm. Full implementation details can be found in (3).

RESULTS AND DISCUSSION

The results of a comprehensive set of experiments on both randomly generated and genomic datasets are reported in (3). Figure 1 gives the distinguishers selected by running DNA-BAR on a set

*Authors are listed in alphabetical order.

†To whom correspondence should be addressed.

Organism	Mb	Barcode								
Nanoarchaeum equitans Kin4-M	0.49	0	0	0	0	0	0	0	0	1
Mycobacterium tuberculosis CDC1551	4.40	0	0	0	0	0	0	1	0	0
Brucella suis 1330 chromosome 1	2.11	0	0	0	0	1	1	0	1	0
Leifsonia xyli subsp. xyli str. CTCB07	2.58	0	0	0	0	0	0	1	0	1
Mannheimia succiniciproducens MBEL55E	2.31	0	0	0	0	1	1	1	0	0
Geobacter sulfurreducens PCA	3.81	0	0	0	1	0	0	0	0	0
Rickettsia typhi str. Wilmington	1.11	0	0	0	0	0	1	1	0	1
Picrophilus torridus DSM 9790	1.55	0	1	0	0	0	0	0	0	1
Mesoplasma florum L1	0.79	0	0	0	0	0	0	0	1	1
Methylococcus capsulatus str. Bath	3.30	0	0	0	0	0	1	0	0	1
Propionibacterium acnes KPA171202	2.56	0	0	0	0	0	0	1	1	0
Mycoplasma mobile 163K	0.78	0	0	0	0	0	1	0	1	1
Mycoplasma hyopneumoniae 232	0.89	1	0	0	0	0	1	0	1	1
Bacillus licheniformis DSM 13	4.22	0	0	0	0	0	1	1	1	0
Legionella pneumophila subsp. pneumophila str. Philadelphia 1	3.40	0	0	0	0	0	1	1	0	0
Onion yellows phytoplasma OY-M DNA	0.86	0	0	0	0	1	1	1	1	0
Staphylococcus aureus subsp. Aureus strain MRSA252	2.90	0	0	1	0	0	1	1	1	1
Staphylococcus aureus strain MSSA476	2.80	0	0	0	0	0	1	1	1	1
Burkholderia pseudomallei strain K96243 chromosome 1	4.07	0	0	0	0	0	1	0	0	0
Bartonella henselae strain Houston-1	1.93	0	0	0	0	0	1	0	1	0
	GC (%)	60.0	45.5	60.0	50.0	57.1	50.0	52.6	42.9	40.0
	Tm (°C)	55.6	59.6	55.4	59.3	56.9	58.6	55.1	55.4	56.3

GATTCGAACCCCCGA
AACTGTCTCACGACGTTCTGAA
GTGGATGCCTTGGCA
AAGCGCGTCGCAAA
GGACTACCAGGGTATCTAATCCTG
CGGTTTTGTGCTTCATGG
CACAAGGAGTGAGTGTTGC
AAAGAAGATAGAGCAGCAGCT
CCATTGACAATTTCAACACC

Fig. 1. Distinguishers selected by running DNA-BAR on a set of 20 microbial genomic sequences with redundancy requirement of 1, distinguisher melting temperature range of 55-60°C, GC content range of 40-60%, and maximum common substring weight bound of 5.

Table 1. Number of distinguishers selected with distinguisher melting temperature range of 55-60°C, GC content range of 40-60%, and varying redundancy r and maximum common substring weight bound w.

r	#Fingerprints	#DNA-BAR distinguishers							
	$w = \infty$	$w = \infty$	12	10	9	8	7	6	5
1	20	7	7	7	7	7	7	7	9
2	40	11	11	12	12	12	12	13	12
5	100	25	25	25	26	26	32	–	–
10	200	48	52	49	55	65	–	–	–
20	400	99	107	114	127	150	–	–	–

of 20 microbial genomic sequences extracted from NCBI databases (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html) with redundancy requirement $r = 1$, distinguisher melting temperature range of 55-60°C, GC content range of 40-60%, and maximum common substring weight bound of 5. Table 1 gives the number of distinguishers obtained for the 20 microbial genomes using the same melting temperature and GC content bounds, redundancy varying between 1 and 20, and maximum common substring weight bound varying between 5 and 12. For comparison, we include in the table the number of DNA *fingerprints*, i.e., DNA substrings each appearing in a unique target sequence, required to achieve the same identification redundancy. DNA fingerprints are commonly used in genomic based identification (e.g., in the recent study of North American birds (4)). The results in Table 1 show that the number of non-unique distinguishers selected by DNA-BAR can be significantly smaller than the corresponding number of fingerprints (up to 4 times for the 20 microbial genomes in our experiment; experiments on simulated data suggest much higher reductions for larger number of sequences (3)). The reduced number of DNA-BAR distinguishers leads to lower assay cost, and, most importantly, makes it possible to enforce more stringent cross-hybridization constraints compared to the fingerprint approach. In future work we plan to experimentally validate our methods and extend them to the problem of simultaneously identifying a small number of microorganisms that may be present in the sample (6).

ACKNOWLEDGMENTS

BDG was supported in part by NSF grants CCR-0206795, CCR-0208749, and NSF CAREER grant IIS-0346973. KMK and AAS were supported in part by NSF ITR grant 0121277. IIM was supported in part by a Large Grant from the University of Connecticut's Research Foundation.

REFERENCES

[1]A. Ben-Dor, R. Karp, B. Schwikowski, and Z. Yakhini. Universal DNA tag systems: a combinatorial design scheme. *Journal of Computational Biology*, 7(3-4):503–519, 2000.

[2]J. Borneman, M. Chrobak, G.D. Vedova, A. Figueora, and T. Jiang. Probe selection algorithms with applications in the analysis of microbial communities. *Bioinformatics*, 1:1–9, 2001.

[3]B. DasGupta, K.M. Konwar, I.I. Măndoiu, and A.A. Shvartsman. Highly scalable algorithms for robust string barcoding. *Intl. J. of Bioinformatics Research and Applications*, to appear.

[4]P.D.N. Hebert, M.Y. Stoeckle, T.S. Zemlak, and C.M. Francis. Identification of birds through DNA barcodes. *Public Library of Science Biology*, 2:1657–1663, 2004.

[5]J.N. Hirschhorn et al. SBE-TAGS: An array-based method for efficient single-nucleotide polymorphism genotyping. *PNAS*, 97(22):12164–12169, 2000.

[6]G.W. Klau, S. Rahmann, A. Schliep, M. Vingron, and K. Reinert. Optimal robust non-unique probe selection using integer linear programming. *Bioinformatics*, 20 (Suppl. 1):i186–i193, 2004.

[7]S. Rash and D. Gusfield. String barcoding: Uncovering optimal virus signatures. In *Proc. 6th Annual International Conference on Computational Biology*, pages 254–261, 2002.