Poster H-81

TOFI: A Software Tool for Oligonucleotide Fingerprint Identification

Authors:
Waibhav Tembe (*Biotechnology HPC Institute, US Army Medical Research and Materiel Command*)
Nela Zavaljevski (*Biotechnology HPC Institute, US Army Medical Research and Materiel Command*)
Elizabeth Bode (*Diagnostic Systems Division, US Army Medical Research Institute of Infectious Diseases*)
Catherine Chase (*Diagnostic Systems Division, US Army Medical Research Institute of Infectious Diseases*)
Jeanne Geyer (*Diagnostic Systems Division, US Army Medical Research Institute of Infectious Diseases*)
Leonard Wasieloski (*Diagnostic Systems Division, US Army Medical Research Institute of Infectious Diseases*)
Gary Benson (*Departments of Biology and Computer Science, Boston University*)
Jaques Reifman (*Biotechnology HPC Institute, US Army Medical Research and Materiel Command*)

Short Abstract: TOFI implements an automated computational approach to identify from a target genomic se-quence its in silico DNA fingerprints satisfying microarray experimental constraints. The GUI-driven software pipeline integrates various bioinformatics algorithms on a high performance computing platform. We employed TOFI to identify in silico DNA fingerprints for several patho-gens, which were then experimentally evaluated.

Long Abstract:
Introduction

Advances in DNA microarray technology and computational methods have unlocked new opportunities to identify “DNA fingerprints,” i.e., short DNA sequences that uniquely identify a specific genome. We have developed a bioinformatics tool, termed TOFI, which supports the identification of DNA fingerprints for the design of microarray-based pathogen diagnostic assays.

The challenges in the computational identification of DNA fingerprints involve: (a) large search space due to the large number and size of genomic sequences, (b) accurate computer modeling of DNA-DNA hybridization on microarray, and (c) the evaluation of cross-hybridization of candidate fingerprints with non-target genomic sequences.

We present an integrated approach for the computational identification of DNA fingerprints. The key contributions of our work are: (a) the provision of a quantifiable definition of a DNA fingerprint stated both from a computational as well as an experimental point of view, and (b) the analytical proof that all in silico fingerprints satisfying the stated definition are found using our approach.

From a pure computer science standpoint, a DNA fingerprint for a target genome gt could be defined as a classic string comparison problem to identify any subsequence of gt that is not a subsequence of any other genome gn from a reference genomic database, such as GenBank. However, from an experimental standpoint, DNA microarrays impose on DNA subsequences a set of design constraints, such as length, GC content, and the secondary structure of the duplex. We refer to all subsequences of gt that satisfy DNA microarray-based constraints as “DNA probes.” DNA fingerprints are those DNA probes that discriminate, in a hybridization experiment, between target and non-target sequences. We estimate the discriminatory power (specificity) of a DNA probe from the number of mismatches, gaps, and insertions/deletions in the sequence alignment and compare it with a threshold T. Hence we define an “in silico” DNA fingerprint for a target genome as follows: A DNA probe p of length L is considered an in silico DNA fingerprint of gt if and only if an optimal sequence alignment between p and any other sequence gn has at most L-T matches.

Computational Approach

The algorithm implemented in TOFI consists of three steps:
Step 1. Comparison of the target genome with a known biological near-neighbor genome: The search space is reduced by discarding the common sequences between the target and a near-neighbor genome longer that a specified minimum length M. Candidate sequences, i.e., those subsequences of target genome that are not present in the near-neighbor genome, are obtained using the suffix tree-based open source software MUMmer.
Step 2. DNA probe design: DNA probes that satisfy the experimental constraints, such as length, melting temperature, GC content, are extracted from candidate sequences using the commercial Oligonucleotide Modeling Platform software.
Step 3. Specificity determination: Specificity of each DNA probe is quantified by examining the best alignment with the reference NCBI RefSeq database, produced by BLAST running in parallel computing environment.

It can be analytically proven that the identification of all DNA probes of length L between Lmin and Lmax having less than or equal to L-T exact matches with any non-target genome implies that the parameters in these steps are related. Specifically, the minimum length of the exact matches M in Step 1 is related to the length constraints on the DNA probes in Step 2 and the specificity threshold T in step 3 as follow: M=Lmax-T+1.

Results

TOFI was used to identify DNA fingerprints for the plague-causing pathogen Yersinia pestis by using Yersinia pseudotuberculosis as the near-neighbor genome. The empirical specificity criterion T=15 based on the BLAST alignments retained 146 in silico DNA fingerprints that underwent further screening (e.g., restriction enzyme cleavage sites), leaving only 99 in silico fingerprints for testing on custom microarrays.

Ten DNA microarray chips (six for Y. pestis and four for Y. pseudotuberculosis), each containing a few replicates of the 99 in silico DNA fingerprints and a number of control sequences were fabricated.

Diagnostic assays demand that a fingerprint has a very low hybridization response for non-targets and a high hybridization response for the target. We analyzed hybridization responses for Y. pestis and Y. pseudotuberculosis to identify valid fingerprints based on alternate rules. For example, 20 probes could be selected by using a minimum threshold value of 2.0 for Y. pestis responses and a maximum threshold value of 0.5 for Y. pseudotuberculosis responses, while about 25 probes could be selected using a minimum threshold of 2.0 for Y. pestis and allowing a maximum threshold of 1.0 for Y. pseudotuberculosis. In each case, a sufficiently large number of probes would allow for detection redundancy.

It is important to note that the in silico fingerprints are only valid with respect to a reference genomic sequence database. This requires that the identified fingerprints be continually tested against newly sequenced genomes.

TOFI is available as a GUI-driven, standalone application running in a U.S. Department of Defense high performance computing environment. DNA fingerprints for Y. pestis have been obtained in 48 hours using 32 CPUs of a Linux cluster.

Conclusions

TOFI is a bioinformatics tool that allows biologists to identify in silico genomic fingerprints for the design of microarray-based diagnostic assays. This work differs from previous ones in that a precise, formal definition of a DNA fingerprint is provided. More importantly, given the desired length of a fingerprint and its required number of non-matching base pairs, we provide an algorithm that guarantees that all in silico fingerprints are identified. TOFI successfully identified DNA fingerprints for a number of pathogens. The fingerprints have been preliminarily validated through experimental tests with related organisms. Further testing, with a standard panel of non-target genomes, is underway.

Acknowledgements

The authors thank the staff of the Advanced Biomedical Computing Center, National Cancer Institute, Frederick, MD, for the computational support.

This work was sponsored by the U.S. Department of Defense High Performance Computing Modernization Program (HPCMP), under the High Performance Computing Software Applications Institutes (HSAI) initiative, and the U. S. Defense Threat Reduction Agency.

Disclaimer

The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the
views of the U.S. Army or the U.S. Department of Defense.