Title:

Oligo_kmer: K-mer-based tool for the design of oligonucleotide probes for Rapid Microbial Detection

Abstract/Purpose:

Rapid detection of pathogenic microbes is crucial for biosurveillance, disease diagnostics, and epidemiological studies. Oligonucleotide-based hybridization methods provide cost-effective and highly specific detection of pathogens even at the strain level. The Oligo_kmer tool addresses the limitations of existing oligonucleotide design methods, providing a user-friendly GUI interface for retrieving DNA probes, efficient memory usage, and customization options for DNA probe design. Moreover, the tool utilizes an SQL database for storing the unique oligonucleotides which could help in better data management and data sharing. The features implemented in the tool help in designing high-quality DNA probes for hybridization techniques, thereby enabling faster disease diagnostics in the clinics and better biosurveillance strategies.

Project Objectives/Goals:

1. Develop a memory-efficient tool that can identify unique kmers/oligonucleotides specific to each microbial species and is suitable for use as a DNA probe in a hybridization chip.
    - Generate kmers for both the strands of the microbial genome.
    - Filter kmers based on criteria like GC content, secondary structure, Tm, etc.
    - Compare the kmers of each microbial species to kmers of all other microbial species using hamming distance in order to identify unique kmers
2. Develop an SQL database to store the unique kmers/oligos for the microbial species analyzed.
3. Develop a GUI interface for querying the SQL database for easy retrieval of oligonucleotides for microbial species

Background:

The rapid detection of pathogenic microbes is crucial for biosurveillance, disease diagnostics, and epidemiological studies. Culture-dependent microbiological techniques have been widely adopted for the detection and identification of bacterial pathogens. However, these methods are often time-consuming and require extensive manual effort, presenting significant limitations. The PCR-based method and Hybridization based method are the two most commonly used molecular biology methods for the rapid detection of pathogens. The hybridization method offers some advantages in comparison to PCR based method as listed below.

1. **High throughput:**

The hybridization technique allows the detection of multiple microbes simultaneously, although PCR-based can be multiplexed, the optimization of PCR based method is more challenging (Cleven et al., 2006).

2. **Simultaneous identification of multiple genes:**

Identification of multiple genes of a pathogen provides some advantages like the ability to detect virulence factors, antibiotic resistance genes, etc. (Cleven et al., 2006).

3. **Limited chances of cross-reaction:**

Since the DNA probes are physically separated from each other in a DNA chip, the chances of cross-reaction is reduced in comparison to multi-plex PCR (Strauss et al., 2015)

4. **Cost-effective and on-site detection capabilities:** colorimetric sensor array-based DNA chips do not require specialized equipment for signal detection and can be used for on-site detection assays (Kim et al., 2022).

However, the design of high-quality and highly specific DNA probes presents a significant challenge in the application of these techniques. In addition to being highly unique to the pathogen in question, which helps in preventing cross-hybridization, the probes should have to fulfill other important criteria such as common Tm for all the probes used in the chip which is crucial for uniform hybridization, absence of sequence repeats and secondary structures which otherwise could interfere with the hybridization and GC content in the range of 40%-60% (Hu et al., 2007, Rouillard et al., 2003, Li et al., 2005). The Oligo_kmer tool was developed for the custom design of oligonucleotides providing an easy user interface to the users. PairwiseAligner class in Biopython's Bio.Aligner module (Bio.Aligner, 2023) was used for implementing Smith-Waterman algorithm for identifying secondary structures in the kmers (Cock et al., 2009). Distance module (Distance, 2023) was used for calculating the Hamming distance between the kmers. Bio.SeqUtils.MeltingTemp module (MeltingTemp, 2023) was used for calculating the Tm of the kmers. MySQLdb(MySQL-python 1.2.5) was used as the API for providing the interface between SQL database and Python.

Biological or health motivation:

Given the rising threat of the emergence of antibiotic resistance, epidemics, and pandemics there is an increasing need for efficient and rapid pathogen detection methods both in the clinical setting and for better biosurveillance. Oligonucleotide-based hybridization methods provide cost-effective and highly specific detection of pathogens even at the strain level. Moreover, DNA chips employing the hybridization method could be used for on-site pathogen detection. The Oligo_kmer tool helps in designing DNA probes for the hybridization chip, facilitating faster diagnosis and effective biosurveillance, which could be used for public health decision-making and formulating disease management strategies.

Significance:

The oligo_kmer tool can be used for designing high-quality DNA probes which could help in better biosurveillance, faster disease diagnostics, and faster detection of antibiotic resistance /virulence genes. The Oligo kmer tool could be a valuable tool for basic researchers as well as epidemiologists.

Novelty:

While existing tools like ArrayOligoSelector (Caudy et al., 2011) and UPS 2.0 (Chen et al., 2010) offer some similar functionalities, they have limitations regarding customization options, user accessibility, outdated Python version as well as lack of SQL support. The Oligo_kmer tool addresses these limitations and introduces additional features such as a SQL database for collaborative research and a GUI interface for broader user accessibility, setting it apart from existing solutions.

Project Components:

1. Python script: Oligo_kmer script is used for identifying unique kmer for each of the records in the MULT-FASTA file as well as for storing the unique kmer in a SQL database
2. MySQL database: A MySQL database has been implemented for storing the unique kmers
3. Web page: A web page has been designed for accessing the unique kmers of the processed records

Data provenance:

The data for the file mini.fasta is a subset of data obtained from the public database HMP Human Microbiome Project (https://www.hmpdacc.org/)

USERS:

Basic researchers, epidemiologists, health care providers for disease diagnostics, and Biotech companies who design DNA chips.

IMPLEMENTATION CONSTRAINTS:

Lack of knowledge in implementing multi-threading/multi-processing limited making the code more efficient.

PRIVACY:

Nil

References:

1. Cleven BE, Palka-Santini M, Gielen J, Meembor S, Krönke M, Krut O. Identification and characterization of bacterial pathogens causing bloodstream infections by DNA microarray. Journal of clinical microbiology. 2006 Jul;44(7):2389-97.

2. Strauss C, Endimiani A, Perreten V. A novel universal DNA labeling and amplification system for rapid microarray-based detection of 117 antibiotic resistance genes in Gram-positive bacteria. Journal of microbiological methods. 2015 Jan 1;108:25-30.
3. Kim TY, Lim MC, Lim JA, Choi SW, Woo MA. Microarray detection method for pathogen genes by on-chip signal amplification using terminal deoxynucleotidyl transferase. Micro and Nano Systems Letters. 2022 Dec;10(1):1-8.
4. Hu G, Llinás M, Li J, Preiser PR, Bozdech Z. Selection of long oligonucleotides for gene expression microarrays using weighted rank-sum strategy. BMC bioinformatics. 2007 Dec;8:1-3.
5. Rouillard JM, Zuker M, Gulari E. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic acids research. 2003 Jun 15;31(12):3057-62.
6. Li X, He Z, Zhou J. Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. Nucleic acids research. 2005 Jan 1;33(19):6114-23.
7. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, De Hoon MJ. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 Jun 1;25(11):1422-3.
8. Caudy AA. Design of Custom Oligonucleotide Microarrays for Single Species or Interspecies Hybrids Using Array Oligo Selector. Molecular Methods for Evolutionary Genetics. 2011:233-41.
9. Chen SH, Lo CZ, Su SY, Kuo BH, Hsiung CA, Lin CY. UPS 2.0: unique probe selector for probe design and oligonucleotide microarrays at the pangenomic/genomic level. InBMC genomics 2010 Dec (Vol. 11, No. 4, pp. 1-7). BioMed Central.
10. Bio.Aligner(2023) https://biopython.org/docs/1.76/api/Bio.Align.html.
11. Distance(2023) https://pypi.org/project/Distance/.
12. MeltingTemp(2023)https://biopython.org/docs/1.76/api/Bio.SeqUtils.MeltingTemp.html
13. MySQL-python 1.2.5 https://pypi.org/project/MySQL-python/