BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 23 2012, pages 3161–3162
doi:10.1093/bioinformatics/bts597

Data and text mining

Advance Access publication October 9, 2012

KASpOD—a web service for highly specific and explorative oligonucleotide design

Nicolas Parisot[1,2], Jérémie Denonfoux[1,2], Eric Dugat-Bony[1,3], Pierre Peyret[1,3] and Eric Peyretaillade[1,3,*]

[1]Clermont Université, Université d'Auvergne, EA 4678 CIDAM, BP 10448 and [2]UMR CNRS 6023, Université Blaise Pascal, 63000 Clermont-Ferrand, France and [3]Clermont Université, Université d'Auvergne, UFR Pharmacie, 63000 Clermont-Ferrand, France

Associate Editor: David Posada

ABSTRACT

Summary: KASpOD is a web service dedicated to the design of signature sequences using a *k*-mer–based algorithm. Such highly specific and explorative oligonucleotides are then suitable for various goals, including Phylogenetic Oligonucleotide Arrays.

Availability: http://g2im.u-clermont1.fr/kaspod.

Contact: eric.peyretaillade@udamail.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 13, 2012; revised on September 25, 2012; accepted on September 28, 2012

1 INTRODUCTION

Environmental DNA microarrays, including Phylogenetic Oligonucleotide Arrays (POAs), are key technologies that are well adapted to profiling environmental communities (Dugat-Bony *et al*., 2012b). The extreme diversity of microorganisms, however, means that molecular community exploration or specific analysis of microbial groups are faced with a new challenge: designing group-specific probe sets that must harbour a high coverage (i.e. being able to hybridize with all the target sequences) and a high specificity, showing no cross-hybridizations with non-target sequences (Loy *et al*., 2008). Sensitivity (i.e. being able to detect even low abundance targets) and uniformity (i.e. uniform thermodynamic behaviours for all the probes) are also main criteria in the selection of the best probe set (Wagner *et al*., 2007).

The development of comprehensive POAs requires integrating large datasets produced by metagenomics projects to assess the coverage and specificity of the probe set. Unfortunately, many available probe design programmes are not suitable to deal with such data (Dugat-Bony *et al*., 2012b). To overcome this limitation, two recent strategies have been implemented (Bader *et al*., 2011; Hysom *et al*., 2012). Despite major speed improvements, both strategies are still not able to define explorative probes. They only define regular oligonucleotides found uniquely in the target group, whereas explorative probes take into account the sequence variability within the target group to define new combinations not yet deposited in public databases but potentially present in the environment.

In spite of large amounts of data, our current vision of the microbial diversity is, indeed, still incomplete. This is partially explained by the tremendous diversity of microbial species, ecological niches and technological limits: detecting 90% of the richness in some complex environments could require tens of thousands of times the current sequencing effort (Quince *et al*., 2008). Microarrays coupled with explorative probe design strategies are, therefore, well suited to survey complete microbial communities, including microorganisms with uncharacterized sequences (Dugat-Bony *et al*., 2012a; Terrat *et al*., 2010).

Currently, the only software dedicated to POAs that allows the design of explorative probes is the PhylArray programme (Militon *et al*., 2007), which relies on group-specific alignments before the probe design step to identify conserved probe-length regions. Building large multiple sequence alignments, however, represents a time-consuming task that is not compatible with high-throughput data.

Here we propose KASpOD, a fast and alignment-free algorithm to detect group-covering signature sequences allowing the design of explorative probes.

2 METHODS

2.1 Usage

KASpOD takes as input a target sequence set and a database of non-target sequences. The web interface accepts two parameters to design signatures: the oligonucleotide length (18–31-mer), and the edit distance between signatures and full-length sequences to perform specificity and coverage evaluation steps. The edit distance is defined as the total number of differences, gaps and/or mismatches allowed between the probe and its target.

2.2 Algorithm

KASpOD consists of three computational stages (Fig. 1).

2.2.1 Search for group-specific k*-mers* The first stage is the extraction of every *k*-mer from both the target and the non-target groups by using Jellyfish version 1.1.4 (Marcais and Kingsford, 2011). For large target groups (>100 sequences), a noise-reduction step is performed to remove *k*-mers occurring only once. Every *k*-mer found in both groups is then removed from the signature candidates, as it occurs exactly in the non-target group.

*To whom correspondence should be addressed.

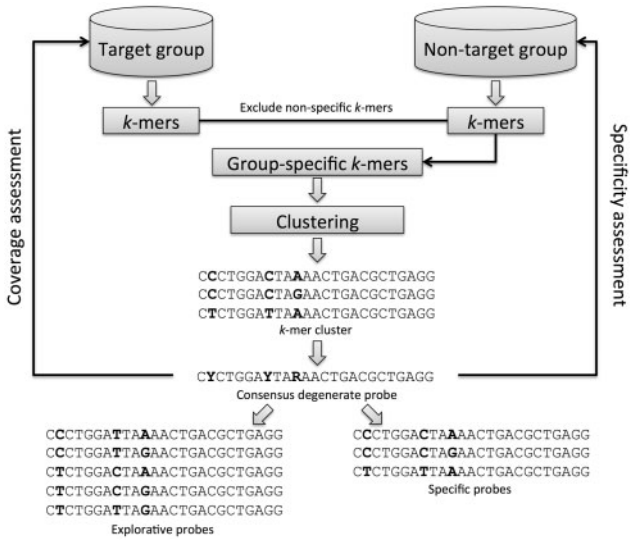

Fig. 1. The KASpOD programme workflow

2.2.2 Consensus signature sequences building The second stage consists of clustering fully overlapping *k*-mers using CD-HIT version 4.5.4 (Li and Godzik, 2006) at an 88% identity clustering threshold. For each cluster, a degenerate consensus signature is built taking into account sequence variability at each position.

2.2.3 Coverage and specificity evaluation The last stage performs a coverage assessment of each degenerate consensus *k*-mer against the target group, by using PatMaN version 1.2.2 (Prüfer *et al.*, 2008). Coverage is computed using the number of exact or non-exact (with at most the edit distance) matches in the target group. Specificity is assessed in the same way by comparing degenerate probes against the non-target group sequences.

3 RESULTS

We used KASpOD to design 25-mer probes for 1295 prokaryotic genera based on the recently published Greengenes taxonomy (McDonald *et al.*, 2012) (see Supplementary Data 1 for complete procedure). Finally, 22 613 group-specific signatures were designed (Supplementary Table 2) and are freely available on the KASpOD website (http://g2im.u-clermont1.fr/kaspod/about.php). This high-quality probe set could be used to build a POA to allow monitoring of complete prokaryotic communities in complex environmental samples. The probe set was not filtered using thermodynamic calculations, to let the users select the entire probe set, or subset, for their own applications, such as Polymerase Chain Reaction (PCR), Fluorescence In Situ Hybridization (FISH), gene capture or *in silico* for rapid sequence identification.

A runtime performance analysis of the web service has been performed and results are available in the Supplementary Data 3.

As KASpOD does not allow the generation of probes longer than 31 nucleotides, an interesting strategy would be to combine KASpOD and GoArrays (Rimour *et al.*, 2005) to concatenate two short probes with a random linker. This approach produces oligonucleotide probes as specific as short probes and as sensitive as long ones. KASpOD could, therefore, be used for applications such as Functional Genes Arrays, offering the opportunity to generate group-specific and explorative probes, allowing a broad coverage of multiple sequence variants for a given gene family.

ACKNOWLEDGEMENTS

The authors thank S. Terrat and A. Mahul for their help.

Funding: This work was supported by Direction Générale de l'Armement (DGA).

Conflict of Interest: none declared.

REFERENCES

Bader,K.C. *et al.* (2011) Comprehensive and relaxed search for oligonucleotide signatures in hierarchically clustered sequence datasets. *Bioinformatics*, **27**, 1546–1554.

Dugat-Bony,E. *et al.* (2012a) In situ TCE degradation mediated by complex dehalorespiring communities during biostimulation processes. *Microb. Biotechnol.*, **5**, 642–653.

Dugat-Bony,E. *et al.* (2012b) Detecting unknown sequences with DNA microarrays: explorative probe design strategies. *Environ. Microbiol.*, **14**, 356–371.

Hysom,D.A. *et al.* (2012) Skip the alignment: degenerate, multiplex primer and probe design using K-mer matching instead of alignments. *PLoS One*, **7**, e34560.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Loy,A. *et al.* (2008) probeCheck—a central resource for evaluating oligonucleotide probe coverage and specificity. *Environ. Microbiol.*, **10**, 2894–2898.

Marcais,G. and Kingsford,C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

McDonald,D. *et al.* (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.*, **6**, 610–618.

Militon,C. *et al.* (2007) PhylArray: phylogenetic probe design algorithm for microarray. *Bioinformatics*, **23**, 2550–2557.

Prüfer,K. *et al.* (2008) PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics*, **24**, 1530–1531.

Quince,C. *et al.* (2008) The rational exploration of microbial diversity. *ISME J.*, **2**, 997–1006.

Rimour,S. *et al.* (2005) GoArrays: highly dynamic and efficient microarray probe design. *Bioinformatics*, **21**, 1094–1103.

Terrat,S. *et al.* (2010) Detecting variants with Metabolic Design, a new software tool to design probes for explorative functional DNA microarray development. *BMC Bioinformatics*, **11**, 478.

Wagner,M. *et al.* (2007) Unravelling microbial communities with DNA-microarrays: challenges and future directions. *Microb. Ecol.*, **53**, 498–506.