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

Nicolas Parisot1,2, Jérémie Denonfoux1,2, Eric Dugat-Bony1,3, Pierre Peyret1,3 and Eric Peyretaillade1,3,*

1Clermont Université, Université d’Auvergne, EA 4678 CIDAM, BP 10448 and 2UMR CNRS 6023, Université Blaise Pascal, 63000 Clermont-Ferrand, France and 3Clermont 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@udamat.fr

Supplementary information: Supplementary data are available at Bioinformatics.

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. 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 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.
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**


