BENCHMARKING DE MÉTODOS DE BINNING A PARTIR DE UN METAGENOMA CONCEPTUAL.

FRANCISCO LUCIANO ISSOTTA CONTARDO
INGENIERO EN BIOINFORMÁTICA

RESUMEN

La metagenómica implica el estudio de comunidades microbianas enteras mediante el muestreo directo y la secuenciación de los genomas presentes en un entorno determinado. En los últimos años, la disponibilidad de secuencias metagenomicas ha crecido de manera significativa, debido principalmente al enorme impacto de las tecnologías de secuenciación de nueva generación (NGS). A diferencia del análisis clásico de un genoma, que implica la manipulación y ensamble de un número moderado de lecturas (20,000 – 30,000) largas (650-800 pb), la metagenómica tiene el reto de analizar millones de lecturas cortas (35-250 pb) de múltiples microorganismos, muchos de los cuales se desconocen.

Uno de los problemas subyacentes del análisis metagenómico es la recuperación de los genomas individuales presentes en la muestra en estudio. La clasificación y el agrupamiento de lecturas o contigs de NGS en unidades taxonómicas operacionales putativos se conoce como binning. Varios métodos de bioinformática para binning se han descrito en los últimos años. Estos métodos se basan en principios diferentes (dependientes de la taxonomía e independientes de la taxonomía) y los softwares se ejecutan de forma local o a través de un servicio web. A pesar de la diversidad de métodos, no hay consenso sobre cuál es el mejor.

Con esto en mente, 11 de los métodos de agrupación disponibles han sido evaluados utilizando un metagenoma conceptual construido con el software Grinder (*). Este metagenoma incluye 9 genomas procariontes completamente secuenciados y 28 genomas virales de diversa filiación taxonómica, consta de 2,5 millones de lecturas. El rendimiento de los softwares de binning se evaluó a través de cuatro criterios (el porcentaje de bases clasificadas, el porcentaje de bases correctamente clasificadas, el puntaje taxonómico y la distancia de la variación de la información). Los métodos de clasificación se evaluaron para todo el conjunto de datos, y las fracciones microbianas y virales.

Se evaluaron cuarenta y siete programas de binning diferentes. De estos: 18 resultaron ser obsoletos, 16 fueron limitados en su aplicación y 13 estaban funcionales para realizar binning. De este último grupo, 2 no fueron evaluados por razones técnicas, dejando a 11 programas de binning que fueron evaluados en profundidad. Estos programas fueron clasificados de acuerdo a los criterios antes mencionados. Los tres programas de binning mejor evaluados fueron: DiScRIBinATE, MEGAN y Sort-ITEMS. El ranking completo se muestra en la Tabla 24.

ABSTRACT

Metagenomics entails the study of entire microbial communities via direct sampling and sequencing of genomes present in a given environment. In recent years, the availability of metagenomic sequences has grown significantly, mainly due to the enormous impact of Next Generation Sequencing (NGS) technologies. Unlike classical genome analysis involving the handling and assembly of a moderate number (20,000 – 30,000) of long reads (650-800 bp), metagenomics is challenged to analyze millions of short reads (35-250 bp) of multiple microorganisms, many of which are unknown.

One of the underlying problems of metagenomic analysis is the retrieval of the individual genomes present in the sample under study. Classification and grouping of NGS reads or contigs into putative operational taxonomic units is known as binning. Several bioinformatics methods for binning have been described in recent years. These methods are based on different principles (taxonomy dependent and taxonomy independent) and executable programs are run either locally or through a web service. Despite the diversity of methods there is no consensus on which is best.

With this in mind, 11 of the available binning methods have been benchmarked using a conceptual metagenome built with Grinder (*). This metagenome includes 9 completely sequenced prokaryotic and 28 viral genomes of diverse taxonomic affiliation and consists of 2.5 million reads. The binning performance of each softwares was evaluated and ranked through four criteria (percent of classified bases, percent of correctly classified bases, taxonomic score and distance of variation of information). Best scoring methods for the whole data set, and the microbial and viral fractions are ranked. 11

Fourty-seven different softwares programs for binning were evaluated. Of these: 18 were found to be obsolete, 16 were limited in their application and 13 were determined to be suitable for metagenomic binning. Of this latter group, 2 were not evaluated further for technical reasons, leaving 11 softwares programs that were evaluated in depth. These programs were ranked according to the aforementioned criteria. The top three programs according to correct binning were: DiScRIBinATE, MEGAN and SOrt-ITEMS. The complete ranking is shown in “Tabla 24”.