How to piece together life's cookbook?

Andries Willem Boers

In my thesis project, I have been working on reconstructing genomes. Genomes contain information about how different molecules in an organism can be put together, just as a cookbook contains information on how different meals can be prepared. In this summary, I will focus on the technological challenge of reconstructing genomes, using a cookbook as a metaphor.

Genomes are built up out of DNA, as cookbooks are built up out of letters. A challenge with reading genomes is that current sequencing technology can only read a few hundred to thousand DNA units at a time, while genomes consists of millions of these units. So to read life's cookbook, we only get to read a few sentences at a time.

A way to handle this challenge is to keep on opening the cookbook at random places, read some sentences and note them down. After a while, we can piece sections of the book together based upon the overlap between different stretches of sentences. Fortunately, I have been spared the effort of reading the thousands of sentences that would be required here. Computers did the heavy lifting for me.

During the reconstruction of the human genome, pure DNA samples could be used. The field in which I'm working has a special challenge associated with it, however. Lokiarchaeum, the microorganism around which my thesis project revolved, does not grow in the laboratory. Its genome was therefore not available in pure form, but we had to read its genome content directly from a sample taken from the Danish sea floor, where it was part of an entire microbial community. Continuing the cookbook metaphor, imagine that we're now interested in finding a specific book within a swimming pool filled with books, and we are still only able to read a few sentences at one time.

How to handle this challenge? Well, we simply start by piecing book sections together again based upon overlap. It is only the scale of the effort that has increased here; we'll get there by reading millions (!) of sentences and billions (!!) of letters. After we've reconstructed cookbook sections in this way, we can group sections based upon two types of information. One type is based upon 'word preference'; some cookbooks prefer certain combinations of letters over others, and these patterns correspond between different section. The other type is 'section coverage', sections belonging to books of which more copies were present in the swimming pool will have more read sentences covering them. Sections can be sorted into different groups, based upon their degree of coverage.

Based upon this methodology, I have managed to piece together some genomes that are likely to be closely related to Lokiarchaeum, a microbe whose genome was previously reconstructed by the Ettema lab. Studying Lokiarchaeum and its relatives could unveil more information about the origin of the eukaryotes. The eukaryotes are a group of organisms that includes plants, animals, you and me. Please read my report if you want to find out more about why Lokiarchaeum is interesting and which methods to use if you want to reconstruct your book of choice from a swimming pool filled with cookbooks.

Degree project in Biology, Master of Science (2 years), 2016
Examensarbete i biologi 30 hp till magisterexamen, 2016
Biology Education Center and Department of Cell and Molecular Biology, Uppsala University
Supervisors: Eva Fernández Cáceres and Thijs J.G. Ettema
External opponent: Katarzyna Zaremba-Niedzwiedzka