Abstract

Similarity searching tools are generally divided to two distinct groups: Alignment-based and alignment-free methods. There is no standard comparison between alignment-based and alignment-free methods` performance. In this project, I compared CLUSTALW, T-COFFEE and MUSCLE algorithms with four alignment-free methods, Kr, D2, D2z and MplusD to find out which of these methods are more efficient with respect to time and accuracy. In both alignment-based and alignment-free, results are represented by phylogenetic trees. Here, all phylogenetic trees are compared against the reference tree to determine how accurate the methods are.

I found that alignment-based methods were the most accurate comparing to alignment-free methods. On the other hand, Kr and D2z gave approximately correct results with substantial gain in speed.

I conclude that for inferring phylogenetic trees containing taxa that have diverged for a long time, alignment-based methods should be used. However if many closely related species are studied, then an alignment-free method should worth well.

Abstrak

Perkakas perbandingan keserupaan boleh dibahagikan secara am kepada dua kumpulan: kaedah berasaskan penjujukan dan kaedah bebas penjujukan. Walaupun kaedah berasaskan penjujukan lebih diutamakan kerana ia berkisar pada konsep homologi, ia biasanya mengambil masa yang lama apabila kita ingin menjujukan turutan-turutan yang sangat panjang daripada banyak taxa. Sementara itu, kaedah bebas penjujukan yang berasaskan frekuensi perkataan dalam suatu jujukan lebih cepat secara relatif dan mungkin memberikan topologi yang lebih kurang tepat. Dalam projek ini, saya telah membandingkan tiga kaedah berasaskan penjujukan: CLUSTALW, T-COFFEE dan MUSCLE terhadap empat kaedah bebas penjujukan: Kr, D2, D2z dan MplusD untuk mengetahui yang manakah kaedah ini lebih cekap untuk membina pepohon filogenetik dalam aspek kelajuan dan ketepatan, dengan menggunakan algoritma *Neighbour Joining*. Dengan menggunakan suatu pohon rujukan yang mengandungi *clade* primat dan burung yang dibina daripada jujukan genom mitokondria dan dijustifikasikan secara biologi, saya mendapati bahawa kaedah berasaskan penjujukan adalah paling tepat berbanding kaedah bebas penjujukan – mereka semua memberikan topologi yang sama seperti pohon rujukan. Sementara itu, Kr dan D2z memberikan keputusan yang lebih kurang tepat dengan kelebihan kelajuan. Saya merumuskan bahawa untuk menganggarkan pepohon filogenetik yang mengandungi taxa yang sudah berpisah untuk masa yang lama, kaedah berasaskan penjujukan harus dipakai. Walaubagaimanapun, jika banyak spesis yang berkait rapat dikaji, maka barangkali kaedah bebas penjujukan agak berguna.

Acknowledgment

I am really graceful to my supervisor Dr. Tsung Fei Khang for giving me the opportunity to involve in this project and devoting his valuable time and unlimited support for directing me all the way throughout this project. I want to take this opportunity to express my gratitude to him for not only guiding me in this project but also helping me to open my mind into new area of feature technology.

I would like to thank my family members, especially my husband, Masoud, for supporting and encouraging me to pursue my project.

Last but not least, my thanks and appreciations go to my friend Hoda for helping me every time and everywhere, when I was facing difficulty.

Table of Contents

List of Figures

List of Tables