



Das Max-Planck-Institut für Marine Mikrobiologie sucht zum nächstmöglichen Zeitpunkt einen

Studenten (m / w) der Fachrichtung Informatik

zur Anfertigung einer
Masterarbeit
mit dem Themenvorschlag:

„Kompressionsbäume für ARB-Datenbanken“

Hintergrund

Das SILVA-Projekt (www.arb-silva.de) stellt umfassende, qualitätsgeprüfte rRNA-Datenbanken u.a. im ARB-Format zur Verfügung. ARB (<http://www.arb-home.de>) bietet einen grafischen Editor für Sequenzdaten und deren Metadaten sowie eine Sammlung von Tools für die phylogenetische Analyse der Daten. Für den Zugriff auf die Daten verwendet ARB eine eigene In-Memory-Datenbank. Um den Speicherverbrauch und die Größe der erzeugten Dateien zu reduzieren, kann die ARB-Datenbank mittels eines Baumes optimiert werden. Aktuell werden dafür phylogenetische Bäume verwendet, mit denen sich allerdings keine optimalen Kompressionsraten erzeugen lassen.

Ziel

Entwicklung eines Programms zur Berechnung eines möglichst optimalen Kompressionsbaumes für eine ARB-Datenbank sowie die Evaluation der erzielten Kompressionsraten im Vergleich zu Kompressionsraten von Bäumen, die mit Programmen zur Berechnung von phylogenetischen Bäumen erstellt wurden.

Aufgaben

- Anforderungsanalyse in Zusammenarbeit mit dem SILVA-Team und dem ARB-Entwickler
- Implementation des Programms gemäß der Anforderungsanalyse
- Dokumentation der Entwicklung im Ticket-System (JIRA)
- Erstellung einer Projektdokumentation (Build-System, Installation, etc.) und einer Benutzerdokumentation
- Berechnung von phylogenetischen Bäumen mindestens mit RAxML und FastTree
- Evaluation der implementierten Lösung
 - Vergleich der Kompressionsbäume und der phylogenetischen Bäume hinsichtlich ihrer Kompressionsraten
 - Vergleich von Programmlaufzeiten und Speicherverbrauch
- Veröffentlichung des Programms unter einer Open-Source-Lizenz

Anforderungen

- Das Programm muss in C/C++ als Kommandozeilentool erstellt werden und es muss auf dem institutseigenen Rechencluster lauffähig sein (HPC-Cluster mit Sun Grid Engine). Der Cluster kann auch für das Berechnen der Vergleichsbäume verwendet werden.
- Sämtliche Dokumentationen müssen in englischer Sprache verfasst werden.

Kenntnisse in Biologie oder Bioinformatik sind nicht notwendig. Die benötigten Fachbegriffe, Konzepte und Formate werden zu Beginn der Arbeit erläutert. Um Zugriff auf die Institutsressourcen zubekommen, muss aus rechtlichen Gründen ein unbezahlter Gastvertrag mit dem Max-Planck-Institut geschlossen werden.

Die Betreuung der Arbeit erfolgt gemeinsam durch das SILVA-Team, im Wesentlichen jedoch durch Dr. Jan Gerken (Dipl. Inf.). Für die Dauer der Masterarbeit wird ein Büroplatz zur Verfügung gestellt. Auf Wunsch steht der Leiter des SILVA-Projekts, Prof. Dr. Frank Oliver Glöckner, als Gutachter für die Masterarbeit zur Verfügung – bitte beachten Sie hierzu auch die Studienordnung Ihres Fachbereichs.

Bei Interesse schicken Sie bitte einen kurzen Lebenslauf sowie einen aktuellen Notenspiegel per E-mail an:

Prof. Dr. Frank Oliver Glöckner
Leiter der Forschungsgruppe mikrobielle Genomik und Bioinformatik
Max Planck Institut für Marine Mikrobiologie
Jacobs University Bremen gGmbH

E-mail: jobs@arb-silva.de

Bremen, den 22.01.2018