Praktikum zur Vorlesung Algorithmische Bioinformatik WS2011/2012

Roland Krause, Matthias Winkelmann, Patrick Pett, Knut Reinert

5. Dezember 2011

1 Aufgabe T3

Bestimmen Sie den Stammbaum der Bakterien.

2 Ziele und Ablauf

Das Pflichtpraktikum zur Vorlesung *Algorithmische Bioinformatik* umfasst die Anwendung von Algorithmen, die in der Vorlesung besprochen wurden, um Hypothesen für Experimente zu generieren. Ausgehend von vollständig sequenzierten Genomen wurden um die Jahrtausendwende eine Reihe von Bioinformatik-Analysen durchgeführt, die im Rahmen des Praktikums nachgearbeitet werden.

Ziel des Praktikums sind saubere, reproduzierbare Implementierungen der Algorithmen, besonders aber ihre Anwendung und die Interpretation der Ergebnisse.

2.1 Zeit und Ort

Die Teilnahme am Praktikum ist verpflichtend. Die Anwesenheit ist nur beim Einführungstermin am 5.12. nach der Vorlesung erforderlich. Das Praktikum läuft bis zum 16.12., Abgabetermin für den Bericht ist der 23.12.2011. Die Präsentationen und Nachbesprechungen erfolgen während der Tutorien am 3.1.2012.

Wer wegen Klausuren oder Kursen teilweise fernbleiben muss, schreibe eine E-Mail mit Begründung an `mailto:krause@molgen.mpg.de`.

Am 6. Dezember 2011 findet kein Tutorium statt. Am 13. wird der aktuelle Übungszettel besprochen und jede Gruppe berichtet über ihren Stand.

2.2 Ablauf

Zwischen den Aufgabengruppen dürfen Sie sich helfen. Tipps gibt es umsonst, ansonsten erwähnen Sie in ihrem Bericht, wenn sie Programme anderer Gruppen, zum Beispiel zum Vergleich oder wegen ungelöster Probleme angewandt haben.

Am MPI besteht nach Absprache die Möglichkeit der Nutzung des PC Pools.

Da der Platz am MPI auf 20 Personen beschränkt ist, sollten Sie sich überlegen, ob Sie am MPI arbeiten wollen.

Fragen und Terminanfragen richten Sie bitte an die Mailing Group *AlBi*[1]. Wir werden die Gruppe nutzen, um weitere Informationen während des Praktikums weiterzugeben und Ihre Anfragen zu beantworten.

[1] `https://lists.fu-berlin.de/listinfo/AlBi`

2.3 Bericht und Vortrag

Zum Praktikum ist ein Bericht anzufertigen, der etwa 10 Seiten reinen Text umfassen soll (1,5 zeilig, 12 Punkt). Orientieren Sie sich an wissenschaftlichen Arbeiten. Er besteht verbindlich aus

1. Einleitung
   (a) Hintergrund
   (b) Aufgabenstellung (Ihre Zusammenfassung)
   (c) Überblick der Ergebnisse
2. Ergebnisse und Diskussion
   mit Kapiteln zu den einzelnen Aufgabenblöcken
3. Abschluss und Bewertung
4. Literaturverzeichnis mit Referenzen, ordentlich formatiert und idealerweise kommentiert
5. Aufstellung der Beiträge der Gruppenmitglieder. Beschreiben Sie kurz wer in Ihrer Arbeitsgruppe welchen Teil der Aufgaben bearbeitet hat. Sie können sich an den *authors' contributions* in Artikeln in BMC Bioinformatics[2] orientieren.
6. Verzeichnissen der Abbildungen und Tabellen
7. Eventuell weitere Anhänge, etwa umfangreichere Abbildungen oder Tabellen

Code gehört nicht in der Bericht, es sei denn, sie wollen auf Besonderheiten in der Implementierung verweisen. Dann geben Sie nur die entscheidenen Zeilen an. Reichen Sie den Code, aber nicht die Ausgangsdaten, BLAST-Ergebnisse und andere einfach zu erzeugende Zwischendaten ein.

Der Bericht ist als zusammenhängendes PDF per E-Mail an `mailto:krause@molgen.mpg.de` einzureichen. Beachten Sie bei der Erstellung von Grafiken die Größe. Stichtag für die Abgabe ist Freitag, der 23. Dezember 2011, 18 Uhr.

Berichte, die nicht akzeptiert werden, müssen überarbeitet und erneut eingereicht werden.

Beginnen Sie den Aufschrieb am besten direkt mit den ersten Arbeiten und schreiben den Bericht zeitgleich als Logbuch, das Sie abschließend überarbeiten.

Sie werden Ihre Arbeit vortragen. Dazu sollte jeder Ihrer Gruppe in der Lage sein. Sie brauchen keinen ausgefeilten Vortrag vorbereiten sondern sollten sich anhand von Abbildungen aus dem Bericht und Stichpunkten durch Ihre Arbeit gehen. Diese Präsentation sollte etwa zehn Minuten dauern.

2.4 Grafiken

Der Bericht soll 3-4 aufbereitete Darstellungen der Ergebnisse in Publikationsqualität enthalten. Nach Möglichkeit ist die erste eine Übersicht der Vorgehensweise und der Ergebnisse (*Figure 1*) Alle Abbildungen sollen aussagekräftig, kompakt und übersichtlich sein. Achsenbeschriftungen, Legenden und klare Zuordnung von Datensätzen sind ebenso gefordert. Umfangreichere Abbildungen können Sie im Anhang unterbringen. Beachten Sie die Dateigröße.

[2] `http://www.biomedcentral.com/bmcbioinformatics/`

2.5 Code

Die Implementierung ist so zu strukturieren, dass sie vollständig wiederholt werden kann. Versuchen Sie Eingabedateien, Rechnerstrukturen und Umgebungen flexibel zu halten, sowie die Verwendung externer Programmpakete auf ein vernünftiges Minimum zu beschränken. Nach Möglichkeit läuft Ihr Code auf den Linux-Maschinen im Rechnerpool des MPI.

Strukturieren Sie Ihren Code in sinnvoller Weise. Verwenden Sie geeignete Datenstrukturen, sowie Funktionen oder Methoden, die 30 Zeilen nicht überschreiten.

Sie können Sprachen nach Wahl einsetzen, bevorzugt sind Python und R. BioPython und ähnlich Bibliotheken können eingesetzt werden; sie sind jedoch nicht erforderlich und wegen der Kürze der Zeit und den häufig unvollständigen Implementierungen unter Umständen hinderlich.

Code, der übermäßig repetitiv oder schlecht strukturiert ist (Copy&Paste-Programmierung, Spaghetti-Code) kann abgelehnt werden und muss anschließend refaktoriert werden.

Jedes Mitglied Ihrer Gruppe sollte einen Teil der Programmieraufgaben übernehmen. Dazu tragen Sie den Namen desjenigen im Kopf der Datei ein, die er oder sie geschrieben hat. Sie können gerne einander helfen. Als Autor gilt (im Rahmen des Praktikums) derjenige, der den Code eingegeben hat.

2.6 Hinweise für Literatursuchen

Für die Literatursuche sollten Sie sich mit den folgenden Stellen vertraut machen. Für Ihren Bericht sollten Sie weitere Informationen sammeln. Wenn Sie einen Artikel zitieren, sollten Sie wenigstens den Abstract gelesen haben.

1. PubMed `http://www.pubmed.com`
2. Google Scholar `http://scholar.google.com`
3. Zur Verfolgung von Zitierungen bietet sich der Thompson citation index `http://isiknowledge.com` an. Er ist aus dem MPI für Molekulare Genetik zu erreichen, eventuell auch über die FU.
4. Zu vielen Themen finden sich gute Einträge in der Wikipedia, insbesondere der englisch-sprachigen `http://en.wikipedia.org`. Beachten Sie die richtige Zitierweise aus der Wikipedia mit Angabe der Version[3].

2.7 Datenquellen

Um die Laufzeiten der Programme, besonders der Homologiesuchen, überschaubar zu halten, werden nur 20 Genome verwendet. Damit kann man bereits gute und verlässliche Aussagen machen. Im einzeln sind dies:

E. coli K12, *Bacillus subtilis*, *Bacteriodes thethaiotaomicron* VPI-5482, Candidatus *Pelagibacter ubique*, *Chromobacterium violaceum*, *Clostridium perfringens* 13, *Corynebacterium glutamicum*, *Deinococcus radiodurans*, *Helicobacter pylori*, *Listeria monocytogenes*, *Mycobacterium tuberculosis* H37Rv, *Methylacidiphilum infernorum* , *Porphyromonas gingivalis*, *Pseudomonas aeruginosa PAO1*, *Salmonella typhimurium* LT2, *Staphylococcus aureus*, *Streptococcus pneumoniae* R6, *Synechocystis*, *Thermotoga maritima*, *Treponema pallidum*.

Diese Genome repräsentieren für dieses Praktikum den gesamten Hintergrund an Sequenzen. Sie können bei Interesse weitere Genome zu dem Datensatz hinzufügen. Ihr Referenzgenom ist *Bacteriodes thethaiotaomicron* VPI-5482.

Zur Erstellung eines Stammbaumes ist es hilfreich, eine *outgroup* zu Verfügung zu haben. Dazu bieten sich die Archaeon *Haloquadratum walsbyi* an.

[3] `http://en.wikipedia.org/wiki/Citing_Wikipedia`

Bei der Genomauswahl empfiehlt es sich häufig, den ältesten sequenzierten Stamm zu wählen, weil es häufig der Laborstamm ist, für den die meisten experimentellen Daten bestehen.
Außerdem muss man immer überprüfen, ob der Stamm vollständig sequenziert ist, oder ob das Projekt noch läuft. Diese Informationen erhält man in der GenBank-Datei.

3 Aufgaben

3.1 Literaturarbeit

- Ermitteln und lesen Sie die Publikation, in der die Sequenzierung Ihres Referenzgenoms beschrieben wurde. Geben Sie einen kurzen Überblick über den Organismus.
- Überlegen Sie, wie aktuell die Informationen aus diesem Artikel heute sind.
- Ermitteln Sie, wie häufig der Artikel zitiert wurde.

3.2 Genomvergleiche mittels BLAST

Für Ihre Hauptaufgabe benötigen Sie einen Abgleich der Ähnlichkeiten der Proteine untereinander, meistens um Orthologe bestimmen zu können. In einer erste Analyse reicht es allerdings aus, nur das Referenzgenom zu betrachten. Sie können dazu eventuell auf die Skripte aus den Übungen zurückgreifen. Für die vergleichende Genomik werden typischerweise vollständige Homologie-Vergleiche aller Proteine durchgeführt.

1. **BLAST** Ihre Ausgangsdaten ermitteln Sie durch Vergleich mittels BLAST der Protein aller 20 Genome gegen die Proteine Ihres Referenzgenoms (*Bacteriodes thethaiotaomicron* VPI-5482).
   (a) Genomdaten laden Sie am besten vom NCBI[4]. Automatisieren Sie dies reproduzierbar mittels `wget` o. ä.. Beachten Sie, dass Bakterien teils mehrere Plasmide und Chromosomen besitzen und dann mehrere Dateien pro Spezies verwendet werden müssen. Für Information über Lage im Genom und Annotation arbeiten Sie mit der GenBank-Dateien (`.gbk`), die entsprechende AA-Sequenzen finden Sie im FASTA-Format (`.faa`). Wofür stehen die anderen Dateiendungen?
   (b) Für BLAST benötigen Sie die Sequenzdaten im FASTA-Format. Vor den Suchen müssen Sie die Datenbank indizieren.
   (c) Untersuchen Sie die Möglichkeiten, die die lokale Installation von `blast` bietet, um es nachher mit dem Parsen der Daten einfacher zu haben, z.B. die Option, die Ergebnisse als Tabelle zu erhalten.
   (d) Erstellen Sie sich ein kleines Testset und probieren Sie verschiedene Optionen, bevor Sie die gesamten Daten laufen lassen.
   (e) Schreiben Sie einen geeigneten Parser, um sich einen Überblick über die Anzahl der Treffer eines Gens im gesamten Datensatz zu verschaffen. Erstellen Sie dazu Abbildungen.
2. **Orthologe** Ermitteln Sie die *Bidirectional Best Hits* von jedem Protein des untersuchtem Genoms (*Bacteriodes thethaiotaomicron* VPI-5482) zu jedem anderen Genom des Datensatzes. Bereiten Sie die Daten anschaulich auf.
3. **Inparaloge und Koorthologe** Bestimmen Sie, welche Gene Ihres Referenzgenoms Inparaloge sind. Dazu müssen Sie einfach die Proteine ermitteln, deren bester Treffer im eigenen Genom liegt. Für die Bestimmung des *Bidirectional Best Hits* (s.o) sind diese Proteine nämlich störend. Da die Duplikation der Inparalogen per Definition nach dem letzten betrachteten Speziationsereignis stattgefunden hat, sollten diese zwei eng verwandten Gene für die Analyse als eines betrachtet werden (Koorthologie). Führen Sie die Orthologie-Untersuchung erneut durch, indem sie entweder das Protein aus dem Datensatz entfernen und die BLAST-Suchen wiederholen (Zeitaufwand) oder indem Ihr Programm zur Bewertung der Orthologen diese Information berücksichtigt (Programmieraufwand).
   Werten Sie aus, wie viele Proteine als Inparaloge betrachtet werden. Für sehr kleine Genome kann es passieren, dass auf diese Weise keine Inparalogen gefunden werden. Testen Sie in diesem Fall die Funktionalität ihrer Implementierung an einem geeigneten Beispiel.

[4] `ftp://ftp.ncbi.nih.gov/genomes/Bacteria`

3.2.1 Gene und Proteine

In der Praktikumsanleitung und in der Literatur werden die Begriffe *Gen* und *Protein* austauschbar genutzt. Prinzipiell sind wir an den Eigenschaften aller Gene interessiert, da aber der Großteil der Gene Protein-kodierend sind, werden nur diese untersucht.

3.2.2 Frage

1. Warum führt man den Vergleich mittels der Protein-Sequenzen der Gene durch und nicht der DNA-Sequenzen?

3.3 Phylogenetischer Referenzbaum

Erstellen Sie einen phylogenetischen Baum aus den Ihnen zu Verfügung stehenden Proteinen.

1. **Genauswahl für Referenz-Stammbaum** Um einen möglichst verlässlichen Stammbaum erstellen zu können, benötigen Sie konservierte Proteine. Suchen Sie Gene, die in möglichst allen betrachteten Organismen einmal vorhanden sind, d.h. in jedem Genom nur genau einen Treffer mit BLAST haben.

   Sollten Sie keine finden, können Sie die Kriterien leicht relaxieren und Gene zulassen, die in einigen Spezies dupliziert wurde oder die in ein oder zwei Spezies fehlen.
   Bakterielle Gene mit vielen Paralogen sind wahrscheinlicher in ihrer Evolution horizontal weitergegeben worden als solche, die nur einmal auftreten. Daher sollte für den Stammbaum möglichst Gene mit wenigen Paralogen verwendet werden. Beachten Sie, dass Treffer mit e-values > 1 ruhig zu berücksichtigen sind. Tatsächlich gibt es genügend Gene, die die Kriterien mit kleinen e-values erfüllen.

2. **Verbesserte Homologiesuchen** Suchen Sie die PFAM-Domänen der für Proteinfamilien, bei denen wenige Spezies keine BLAST-Treffer haben und suchen Sie mit `hmmsearch` aus dem HMMER-Paket[5].

3. **Referenz-Stammbaum** Erstellen Sie aus den Genen einen phylogenetischen Baum nach einem geeigneten Alignment mit `muscle`. Inspizieren Sie das Alignment kritisch, besonders in Hinblick auf die Alignierung von Proteinen, die eventuell keine Homologen sind. Verwenden Sie für die erste Analyse Neighbor Joining. Erfüllt der Baum Ihre Erwartungen? Sie können Ihr Ergebnis mit dem von Ciccarelli *et al.* erstellten Baum vergleichen[1].

4. **Verbesserter Referenzbaum** Erstellen Sie den Baum mit PhyML[6] oder TreePuzzle[7]. Suchen Sie eine geeignete Anzahl von Bootstraps.

5. ***concatenated alignment*** Fügen Sie weitere Proteine, die ihre Bedindungen der Konservierung erfüllen zu einem sogenannten *concatenated alignment* zusammen. Erzeugen Sie einen weiteren Baum und vergleichen Sie die Topologie sowie die Bootstrap-Werte. Hierzu ist es eventuell nötig, die Protein-Sequenzen händisch zu bearbeiten. Besser ist die Verwendung eines Parsers für konservierte Regionen (gap-freie Spalten) oder die Verwendung von `gblocks`[8].

6. **Inkongruente Genbäume** Wenn es frühzeitig zu einer Duplikation eines Proteins und nachfolgendem Verlust in einzelnen Spezies kam, könnte der erstellte Baum nicht der Phylogenie entsprechen. Wiederholen Sie die Prozedur für verschiedene beliebige Proteine und überprüfen Sie, welche Gene mit Ihrem Referenzbaum übereinstimmen.

[5] `http://hmmer.janelia.org/`
[6] `http://www.atgc-montpellier.fr/phyml/binaries.php`
[7] `http://www.tree-puzzle.de/`
[8] `http://molevol.cmima.csic.es/castresana/Gblocks.html`

3.3.1 Fragen

- Welche Distanz in Jahren liegt zwischen den Spezies Ihres Stammbaumes? Schlagen Sie nach.
- Welche Ansätze verfolgen TreePuzzle und PhyML? Wie unterscheiden Sie sich?

4 Anhang

4.1 Versionierung

Bei gemeinsamem Arbeiten ist es zweckmäßig, ein Versionierung mittels Subversion, Git oder Bazaar durchzuführen.

4.2 Weitere nützliche Programme

- **Artemis** Visualisierung von Genomsequenzen. Funktioniert gut für Bakterien[9].
- **Artemis Comparison Tool** Zum Vergleich von zwei Genomen
- **iTOL** Eigentlich *interactive tree of life* Darstellung von phlyogenetischen Bäumen im Webbrowser[10].
- FigTree

4.3 Server am MPI

Für größere Berechnung steht am MPI Molgen in 40-Knoten-Cluster zu Verfügung, der sich auch ohne Rechenerlaubnis am MPI nutzen lässt. Dazu muss man sich lokal auf den Linux-Rechnern des PC pools anmelden. Der Benutzername ist ist "l" + *Computername*, das Passwort "gast". Der Zugang ist stark beschränkt, kann aber die meisten Programme der Infrastruktur am MPI nutzen.

Zur Nutzung mache man sich mit den Programmen `qhost`, `qsub` und `qstat` vertraut (`man qsub`). Man lege ein Skript, etwa namens `blastgr12.sh` an, dass die entsprechenden Befehle enthält und schicke es mit `qsub -q students.q blastgr12.sh` ab. Es empfiehlt sich, alle Pfade zu Daten und Programmen voll zu qualifizieren.

Das Verzeichnis für alle Berechnung sollte `/scratch/hal64_hpc/pcpool/` + *Gruppenname* sein. Berechnungen im Home-Verzeichnis schränken die Netzwerkverbindung des Instituts stark ein und werden mit Besuchen vom Systemadministrator und sowie Kuchenbacken geahndet.

4.4 Literaturverzeichnis

Literatur

[1] F. Ciccarelli, T. Doerks, C. von Mering, C. Creevey, B. Snel, and P. Bork. Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science*, 311(5765):1283–1287, 2006.

[9] `http://www.sanger.ac.uk/resources/software/artemis/`
[10] `http://itol.embl.de`