Nachhaltigkeit der Forschungssoftware AUGUSTUS zur Genomannotation

Fabian Gumz, Steffen Herbold, Henry Mehlan and Mario Stanke
herbold@cs.uni-goettingen.de, mario.stanke@uni-greifswald.de

UNIVERSITÄT GREIFSWALD
Wissen lockt. Seit 1456

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Zusammenfassung

AUGUSTUS [1–4] ist ein in C++ entwickeltes Bioinformatik-Werkzeug für die Genomannotation [5]. Obwohl es oft verwendet wurde, gab es erhebliche Defizite bezüglich der Benutzbarkeit. Im Programm "Nachhaltigkeit Forschungssoftware" wurden bisher sowohl der Entwicklungsprozess als auch die Usability verbessert. Viele weitere Verbesserungen sind geplant.

Web Service

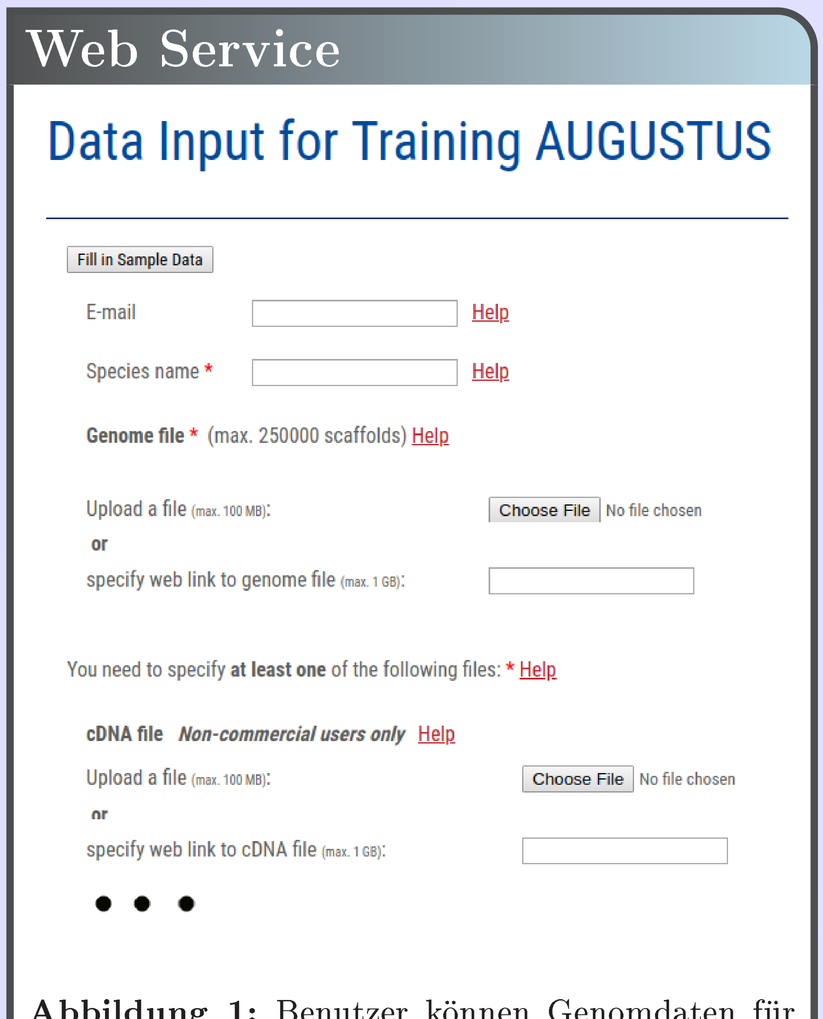

Abbildung 1: Benutzer können Genomdaten für Jobs hochladen, die mehrere Tage laufen, und so eine eigene Installation vermeiden.

Nachhaltige Entwicklung

- Umzug der Entwicklung von privatem Subversion Server nach GitHub
- Weiterenwicklung durch Featurebranches und Pullrequests.
- Continuous Integration mit Hilfe von TRAVIS-CI
- Automatisch erzeugte API Dokumentation mit DoxyGen

Usability Verbesserungen

- Bereitsstellung eines Dockerimages zur Virtualisierung und zur Benutzung unter Windows
- Extern bereitgestellte Debian-Med & Ubuntu-Pakete
- Bessere Dokumentation der Installation

Danksagung

Gefördert durch
DFG Deutsche Forschungsgemeinschaft
im Rahmen der Ausschreibung
"Nachhaltigkeit von Forschungssoftware"

Genvorhersage

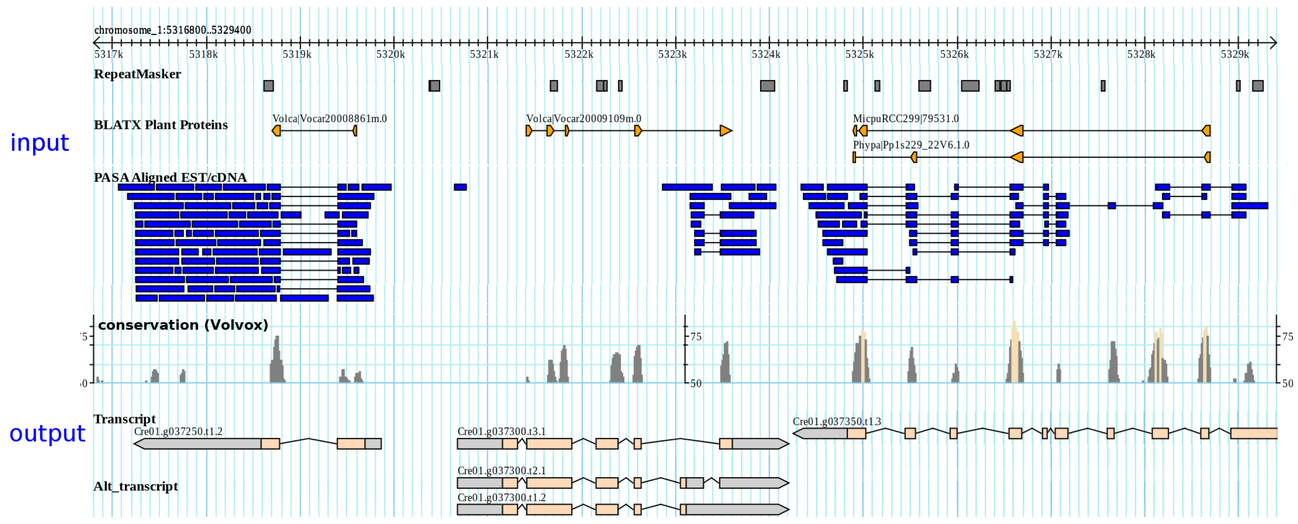

Abbildung 2: AUGUSTUS sagt die Lage und Struktur von Genen in DNA-Sequenzen vorher unter Verwendung einer Methode maschinellen Lernens (Conditional Random Field, ein probabilistisches graphisches Modell) für die Segmentierung und Klassifizierung eines Eingabegenoms.

```
chrX   AUGUSTUS   gene          445   1848   1   +   .   g2
chrX   AUGUSTUS   transcript    445   1848   .   +   .   g2.t1
chrX   AUGUSTUS   start_codon   445   447    .   +   0   transcript_id "g2.t1"; gene_id "g2";
chrX   AUGUSTUS   CDS           445   582    .   +   0   transcript_id "g2.t1"; gene_id "g2";
```

Abbildung 3: Ausschnitt einer Beispielausgabe.

Benutzerrepositorium für baumstrukturierte Daten

- Spezifisch für eine Spezies gelernte Parameter (ca. 1MB) sollen von Anderen wiederverwendet werden können
- Parametertraining ist aufwändig, aber die Parameter von nah verwandten Spezies sind geeignet (z.B. menschliche Parameter für das Genom einer Maus)
- *Reproduzierbarkeit* nur bei gleicher Programmversion, gleichen Parametern und gleichen Daten
- Repositorium für Parameter soll *taxonomische Suchen* erlauben, etwa "benutze eine Version der Parameter vom nächsten Verwandten im Repositorium von *Musca domestica*" (Stubenfliege)

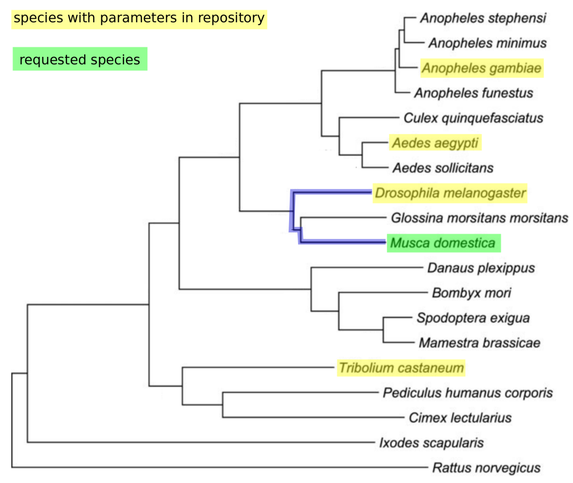

Abbildung 4: Unter vorher von Benutzern hochgeladenen Parameterdateien soll eine zu einer Anfragespezies taxonomisch nahe gewählt werden.

Regressionstests für Vorhersagegenauigkeit

Gene Sensitivity	40.74%
Gene Specificity	37.97%
Exon Sensitivity	77.16%
Exon Specificity	77.70%
Nucleotide Sensitivity	94.45%
Nucleotide Specificity	89.84%

Tabelle 1: Sensitivität und Spezifität zum Messen der Performanz

- Die durchschnittliche Vorhersagegenauigkeit kann auf Testdaten geschätzt werden (links)
- Änderungen die ungewollt die Vorhersagen verschlechtern müssen vermieden werden.
- Regressionstests können signifikante Verschlechterungen automatisch erkennen und verhindern.

Weitere geplante Verbesserungen

- Bessere Integration mit Gendatenbanken
- Neues Pipeline/Workflow-System
- Umfangreiche automatische Tests
- Die Parallelisierung mittels Multithreading

Literatur

[1] M. Stanke, M. Diekhans, R. Baertsch, and D. Haussler. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics*, 24(5):637–644, 2008.

[2] K.J. Hoff and M. Stanke. WebAUGUSTUS – a web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucleic Acids Research*, 2013.

[3] M. Stanke and S. Waack. Gene prediction with a hidden Markov model and new intron submodel. *Bioinformatics*, 19 Suppl. 2:ii215–ii225, 2003.

[4] M. Stanke, R. Steinkamp, S. Waack, and B. Morgenstern. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.*, 32:W309–W312, 2004.

[5] Katharina Hoff and Mario Stanke. Current methods for automated annotation of protein-coding genes. *Current Opinion in Insect Science*, 7:8–14, 2015.