Digitale Plattform zum Citizen Science Projekt „Schwarm-Genomik“

1 Antragsteller/in

Arbeitsgruppe Computational Systems Biology
Prof. Dr. Toni Goßmann

2 Kurzbeschreibung des Projektes

Wir befinden uns im Zeitalter der BigData, in dem in verschiedensten Disziplinen riesige Datenmengen angehäuft und ausgewertet werden können. Leider tritt aber dabei in den Hintergrund, ob die verwendeten Daten überhaupt repräsentativ sind und über ausreichende Qualität verfügen. In dem Projekt Schwarm-Genomik soll innovative Lehre mit Forschung verbunden werden und Laien und Experten gleichermaßen für bioinformatische Analysen gewonnen werden. Durch den Einsatz von möglichst vielen Teilnehmenden in diesem Citizen-Science Projekt soll ermittelt werden, inwieweit biologische Genomdatenbanken wirkliche biologische Diversität widerspiegeln. Die Ergebnisse werden aufbereitet und öffentlich in einem Webportal zur Verfügung gestellt.

3 Details zum Projekt

3.1 Istzustand vor Beantragung

Jüngste Fortschritte von Sequenzierungstechnologien haben zu einer Blüte genomischer Ressourcen geführt, was zum Beispiel durch großangelegte Genomprojekte für Wirbeltiere (z.B. Bird 10K (OBrien, Haussler, and Ryder 2014) oder Bat 1K (Teeling et al. 2018)) veranschaulicht wird. In naher Zukunft werden Referenzgenome fast aller Arten verfügbar sein. Allerdings unterscheiden sich die Genome auch zwischen Individuen einer Art. Daher ist die geeignete Auswahl eines Referenzindividuums für ein Referenzgenom von allerhöchster Bedeutung. Leider wurde und wird diesem Aspekt wenig Aufmerksamkeit geschenkt, was dazu führt, dass die Qualität von Referenzgenomen stark schwankt.

Bereits im Jahr 2001 wurde die Veröffentlichung des humanen Erbguts gefeiert. Tatsächlich war dieses aber äußerst lückenhaft und erst im letzten Jahr wurde das vorläufig vollständige humane Genom publiziert (Nurk et al. 2022). Dies ist nur ein prominentes Beispiel dafür, dass die Qualität von Referenzgenomen stark variiert. Aus biologischer Sicht stellt sich daher eine entscheidende Frage: Welche Qualität haben Referenzgenome überhaupt und können wir aus den Genomen einzelner Individuen wirklich etwas über die genetische Diversität der entsprechenden Spezies lernen? Um das zu beantworten, ist Ziel des Projekts Schwarm-Genomik, zu verstehen, welche Faktoren die genetische Vielfalt in biologischen Datenbanken beeinflussen.

Anhand bioinformatischer Analysen ist es möglich mit Hilfe des Genoms eines einzelnen Individuums Rückschlüsse auf die gesamte Spezies zu ziehen. Dies betrifft zum Beispiel die generelle genetische Vielfalt einer Art, die demografische Entwicklung oder aber auch parasitären Befall oder Durchmischungen mit anderen Arten. Aber auch die Qualität eines Referenzgenoms kann ermittelt werden. Durch das Zusammentragen verschiedener Kenngrößen von möglichst vielen Spezies können biologische Muster und technische Artefakte ermitteln werden. Die größte Herausforderung ist hierbei Genomdaten so gleichartig zwischen den einzelnen Spezies aufzubereiten, dass sie auch vergleichbar sind. Dies bedarf manueller Aufbereitung und Analyse der genomischen Daten. Ziel ist es daher, interessierte Laien (Citizen-Science) in die Analyse mit einzubeziehen.

Daher wird bei Schwarmgenomik folgender Ansatz verfolgt: Teilnehmende (z.B. Studierende während des Kurses Schwarmgenomik oder externe Interessierte) werten einzelne, selbstgewählte Speziesgenome mit Hilfe von vorher entwickelten bioinformatischen Analysen selbstständig aus. Diese werden als kleine Anleitung, z.B. anhand eines WiKis und kurzer Videotutorials, zur Verfügung gestellt. Ziel ist es zudem auch, dass sich Teilnehmende über einen moderierten Blog austauschen und gegenseitig helfen. Anschließend werden die einzelnen Ergebnisse über eine Plattform zusammengetragen. Diese Plattform soll dann dazu dienen, die komplexen Daten zu visualisieren und dadurch eine gemeinsame Auswertung des Schwarms zu erreichen. Ziel ist es dabei möglichst viele ("Schwarm") Genome in die Analyse aufzunehmen, um ein umfangreiches Bild der vorhandenen Referenzgenome zu bekommen.

3.2 Projektziel/Projektbeschreibung

Ziel des Projektes ist es, komplexe bioinformatische Analysen so zu vereinfachen und aufzubereiten, dass auch Laien unter Anleitung Teile von oder vollständige Genomanalysen durchführen können, um sich somit mit der Thematik besser vertraut zu machen. Zum anderen werden die erzeugten Daten auch eine unschätzbare Quelle für biologische Fragestellungen darstellen und sollen der Allgemeinheit und insbesondere der Wissenschaft für Folgeanalysen zur Verfügung gestellt werden. Zudem sollen auch Rohdaten der Allgemeinheit zur Verfügung gestellt werden.

Generelle Idee: Das Projekt ist vollständig computerbasiert und kann sowohl online als auch im Rahmen eines Kurses durchgeführt werden. Inhalte decken dabei Bereiche der Populationsgenetik, molekularen Evolution und des Data Mining von biologischen Daten ab. Zudem wird eine allgemeine Einleitung mit einer Einführung in die Genomsequenzierung und relevante populationsgenetische Methoden angeboten. Anschließend werden die Teilnehmenden unter Anleitung das Genom einer selbstgewählten Spezies analysieren. Erzielte Ergebnisse der einzelnen Spezies werden dann auf einer Plattform zusammengeführt und dadurch ausgewertet. Alle Ergebnisse werden online verfügbar gestellt, um somit auch Folgeanalysen zu ermöglichen.

Wer wird angesprochen? Ein grundlegendes Interesse an computerbasierter Analyse und Auswertung von biologischen Daten ist Voraussetzung, wobei vorhandene Programmiererfahrungen vorteilhaft, jedoch nicht notwendig, sind. Gegebenenfalls nicht vorhandene Programmierkenntnisse können durch angebotene Lerneinheiten (z.B. unterstützt durch Videotutorials oder Lernmaterialien) erworben werden.

3.3 Einzelmaßnahmen, Schritte etc.

Das Projekt wird im Rahmen des Kurses Schwarmgenomik aufgesetzt. Dort werden die Übungsmaterialien mithilfe der teilnehmenden Studierenden angefertigt und verbessert. Eine Hilfskraft wird diese dann entsprechend so aufarbeiten, dass Schwarmgenomik auch für Studierende außerhalb des Kurses und sogar außerhalb der Universität (z.B. durch Laien) durchführt werden kann. Konkret ergeben sich folgende Arbeitsschritte:

1. Entwicklung einer ansprechenden Weboberfläche auf einem Server, um das Interesse an bioinformatischer Analyse zu steigern. Ankündigung des Projektes mithilfe einer Publikation in einem einschlägigen Journal (zum Beispiel Genome Biology and Evolution oder G3).
2. Durchführung bioinformatischer Analysen. Innerhalb der Weboberfläche wird eine automatisierte Einrichtung einer de.NBI Workstation eingebettet, dies wird mit Hilfe des Supports vor Ort (RBG), des ITMCs und des de.NBI Supports erreicht.

3. Entwicklung des Analyse-Workflows mit modularem Aufbau, damit ein Genom etappenweise und nicht komplett von einer einzelnen Person analysiert werden muss. So werden einzelne Aspekte bereits Lernerfolge ermöglichen und dadurch die Motivation erhöhen. Zudem wäre es damit möglich, dass sich eine Gruppe von Teilnehmenden intensiv mit einem Referenzgenom beschäftigt und die Aufgaben untereinander verteilt. Außerdem soll ein Wiki bzw. Blog eingerichtet werden, der es ermöglicht Fragestellungen unter den Teilnehmenden zu kommunizieren. Dies wird dann von der Hilfskraft moderiert. Dazu sollen modular aufgebaute Scripte zur Verfügung gestellt werden. Diese sind zum Teil bereits vorhanden, und sollen durch kleine Videos ergänzt werden. Es besteht daher die Möglichkeit sowohl textbasierte als auch videobasierte Hilfe anzubieten. Dabei soll insbesondere Unterstützung und Infrastruktur der Medienabteilung der Universität genutzt werden.
4. Einrichten einer Schnittstelle zum Hochladen der erzielten Daten sowie Qualitätsmanagement. Geplant ist für jedes erfolgreich analysierte Referenzgenom ein Zertifikat (virtuelle Urkunde) an die Beteiligten auszustellen. Datenqualität soll dadurch gewährleistet werden, dass die Genome mehrfach und unabhängig analysiert werden und bei Diskrepanzen ein Expertenteam eine Reanalyse durchführt.
5. Einrichten einer Schnittstelle zur Visualisierung der Daten, die auf der Weboberfläche integriert und erweitert wird, insbesondere zur generellen Nutzung und Downloadmöglichkeit. Dies könnte zum Beispiel mit R oder Javabasierten Tools (Vue JS, Vite) realisiert werden.

3.4 Geplante Laufzeit

Generell handelt es sich um ein Langzeitprojekt, dass laufend aktualisiert werden soll. Für diese spezifische QVM-Maßnahme wird der Kern die Veranstaltung “Schwarmgenomik” bilden, welche ich im Sommersemester anbieten werde, sowie deren Nachbereitung Anfang des WS 2023/24.

3.5 Indikatoren zur Evaluation des Projektes

a) Einrichtung eines Webservers und Vorhandensein einer Weboberfläche
b) Anzahl aufbereiteter Genome
c) Anzahl vorhandener Lehrmaterialien („Lehrpakete“)
d) Blognutzung
e) Anzahl aktiver Nutzer*innen

3.6 Nachhaltigkeit/Verstetigung

Es handelt sich um ein Langzeitprojekt, das laufend aktualisiert werden soll. Grundlage bildet dazu das jährliche Seminar “Schwarmgenomik”, welches ich an der Universität Dortmund anbiete und insbesondere biologisch interessierte Studierende anspricht. Ziel ist es, forschungsorientierte Lehre so anzubieten, dass sowohl Lernerfolge ermöglicht werden und gleichzeitig neues Wissen generiert wird. Durch die jährliche Erweiterung des Datenbestandes wird es zudem auch möglich sein, neue und größere Fragestellungen, die z.B. neue Tiergruppen betreffen, zu bearbeiten. Geplant ist, den aktuellen Fortschritt regelmäßig (z.B. alle 2 Jahre) zu publizieren.

Es ist bereits jetzt absehbar, dass die Anzahl an Referenzgenomen in den nächsten Jahren in die Höhe schnellen wird. Eine Plattform, die die einheitliche Aufbereitung und Analyse dieser Genome ermöglich, wird dann sehr relevant und auch langfristig nutzbar bleiben.