Medical Bioinformatics: Entwicklung einer Genexpressionsdatenbank für kolorektale Karzinome

Hintergrund:

Darmkrebs (kolorektales Karzinom, Colorectal Cancer, CRC) hat die zweithöchste Sterblichkeitsrate weltweit, was die Entwicklung fortschrittlicher Diagnostik und individueller Therapien erfordert. Für Darmkrebs konnten vier molekularen Subtypen (Consensus Molecular Subtypes, CMS) identifiziert werden, deren Genexpressionsprofile als wertvolle Grundlage für die Patientenstratifizierung und personalisierte Krebsprognose dienen. Diese Genexpressionsdaten werden daher als Eingabemerkmale für Klassifizierungsaufgaben verwendet, jedoch sind hierfür große Datenmengen erforderlich, die oftmals aus unterschiedlichen Quellen und Sammlungen zusammengestellt werden müssen. Über öffentlich zugängliche Datenbanken wie den Gene Expression Omnibus (GEO) kann nicht nur auf umfassende Sammlungen von molekulargenetischen Daten zugegriffen werden, sondern diese sind zudem ausführlich dokumentiert.

Expressionsdaten unterscheiden sich allerdings in mehreren Aspekten voneinander: Welche Methode wurde zur Erstellung der Daten verwendet (MicroArray, RNA-Seq,…)? Mit welcher Software wurden die Daten prozessiert (DESeq2, Limma,…)? Stammen die Daten aus unterschiedlichen Kohorten (Control-Treatment, Zeitreihen, Gesund-Krank,…)? Handelt es sich um Subtypen? Sind noch weitere Informationen zu den Proben verfügbar die einen Einfluss auf deren Interpretierung haben?

Für nachfolgende Analysen müssen die zusammengestellten Datensammlungen zudem normalisiert werden, damit die Werte innerhalb eines festgelegten Bereichs liegen. Dabei müssen auch die systematischen technischen Unterschiede berücksichtigt werden, welche entstehen wenn Proben in verschiedenen Chargen verarbeitet und gemessen werden und die nicht mit biologischen Variationen zusammenhängen. Diese Batch-Effekte (oder Chargeneffekte) müssen dabei korrigiert werden, um keinen Einfluss auf die Analysen zu haben.

Aufgabenstellung:

Ziel dieser Arbeit ist die Erstellung einer Datenbank zur Verwaltung und des Exportes von Genexpressionsdaten. Dabei sollen einzelne Datensätze oder Sammlungen direkt von GEO übernommen werden und automatisch mit sämtlichen Metadaten in die Datenbank integriert werden können. Bestehende Datensätze sollen sich nach Krebsart, Subtyp, Aufnahmemodalität und weiteren Attributen filtern lassen. Ausgewählte Datensätze sollen als Expressionstabelle mit Probenannotation exportiert werden können, wobei die Daten bereits normalisiert und der Batch-Effekt korrigiert ist.

Die Umsetzung des Projekts ist dabei weitestgehend frei gestaltet. Eine mögliche Herangehensweise ist die Umsetzung als Webanwendung: - Erstellung eines

Webservers mit python und flask - Anbindung einer MySQL oder PostgreSQL Datenbank zur Speicherung der Genexpresionsdaten und Metainformationen (zu den Proben, Sets und Quelle) - Implementierung des Datenimports von GEO für ausgewählte Beispielproben (nur benötigte Daten) - Tabellarische Übersicht über existierende Datensätze (mit Metainformationen wie Krebsart, Subtyp, Aufnahmemodalität, Datum, etc.) - Auswahlmöglichkeit für mehrere Datensätze und direkter Export der Expressionsdaten (nicht normalisiert) - Zusammenfassung über selektierte Sets (z.B. Verteilung der Expressionswerte in den einzelnen Sets) - Implementierung der Normalisierung (z.B. Quantile Normalisierung) - Implementierung der Batch-Effekt-Korrektur (z.B. LIGER)

Mögliche Erweiterungen:

- Harmonisierung der Expressionsdaten auf verschiedene Referenzgenome
- Implementierung verschiedener Normalisierungs- und Korrekturalgorithmen
- Vorauswahl des Exports durch Eingabe von Gensets
- Implementierung eines ansprechenden Frontends

Anforderungen:

- Studierende/r im Bereich Informatik, Medizininformatik, Bioinformatik oder einem verwandten Studiengang
- Fundierte Programmierkenntnisse in Python, Java, Groovy, Ruby, oder JavaScript/TypeScript sind erforderlich
- Kenntnisse in der Python/Flask oder einem anderen serverseitigem Webframework (node.js, Spring, Ruby on Rails, Grails) sind von Vorteil, jedoch nicht zwingend notwendig
- Kenntnisse in der Frontend/client-side Webentwicklung sind von Vorteil, jedoch nicht zwingend notwendig

Dauer und Betreuung:

Die Arbeit wird auf eine Dauer von 3 (Bachelorarbeit) bis 6 (Masterarbeit) Monaten ausgelegt. Während der gesamten Arbeit steht Ihnen ein Betreuer zur Seite. Der genaue Inhalt und die Aufgabenstellung können in gemeinsamer Absprache noch weiter verfeinert werden, um individuelle Interessen und Fähigkeiten zu berücksichtigen. Für etwaige Rückfragen stehen wir Ihnen gerne zur Verfügung.

Bei Interesse an diesem Thema melden Sie sich bitte bei florian.auer@uni-a.de

Weiterführende Informationen:

- Gene Expression Omnibus: https://www.ncbi.nlm.nih.gov/geo/
- Flask: https://flask.palletsprojects.com/en/3.0.x/
- MySQL: https://www.mysql.com/

- PostgreSQL: https://www.postgresql.org/
- Expressionsanalyse: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7771369
- Batch Effekt Korrektur: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1850-9