SNP Selektion für Assoziationsstudien – Exposé einer Studienarbeit

Betreuer: Ulf Leser[1], Jochen Hampe[2], Michael Krawczak[3]
Bearbeiter: Andreas Wollstein (März 2004 bis August 2004)

[1]WBI - Institut für Informatik Humboldt Universität zu Berlin
[2]Mucosa Research Group - 1. Medizinische Klinik Christian Albrechts Universität zu Kiel
[3]Institut für Medizinische Informatik und Statistik - 1. Med. Klinik Christian Albrechts Universität Kiel

Problemstellung

Durch Assoziationsstudien wird versucht, Krankheitsgene auf dem Genom zu lokalisieren. Die Lokalisation beruht auf der Detektion von Krankheitsassoziation mit bestimmten „Single Nukleotid Polymorphismen“. Für die effiziente Durchführung einer Assoziationsstudie ist die Auswahl der SNPs entscheidend. Ziel ist es, eine möglichst informative Menge für die Studie zu verwenden. Wenn im Verlauf der Studie keine ausreichende Assoziation der betrachteten Loci gefunden wurde ist man gezwungen, die vorhandene Auswahl von typisierten SNPs zu erweitern. Nicht jeder beliebige, zusätzlich typisierte SNP kann die Assoziationswahrscheinlichkeit verbessern. Der Informationsgewinn eines neuen, zu einer vorhandenen Menge hinzugefügten SNPs, ist abhängig von seiner Entfernung von, und seinem Kopplungsungleichgewicht zu den anderen SNPs. (Hampe et al. 2003 [1]). Im Labor ist man daran interessiert, den Typisierungsaufwand möglichst gering zu halten. Kostenintensive, zusätzliche Typisierungen sind davon abhängig, wie gut man seine Assoziationsstudie dadurch noch erhärten kann. Hampe & Krawczak beschreiben eine Strategie die Auswahl der SNPs zu optimieren, damit die Assoziationswahrscheinlichkeit maximal wird. Hierzu wird die sogenannte Mapping Utility eingeführt [1], ein Maß für den Informationsgewinn eines neuen SNPs basierend auf Shannons Entropie. Mit dem $\kappa_{\min}$ -Wert [1], der den maximal möglichen Informationsgewinn eines zusätzlichen SNPs bewertet, erhält man eine Entscheidungshilfe für den Abbruch bzw. Fortführung einer Studie. Durch das statistische Auswahlverfahren kann man bis zu 30% an Genotypisierungsaufwand und Zeit sparen.

Ziel

Im Rahmen dieser Studienarbeit soll eine öffentlich zugängliche Webdatenbank erstellt werden, die es dem Forscher über ein Frontend ermöglicht, die informativsten SNPs aus einem selbst gewählten Intervall für sein Experiment zu berechnen. Es bieten sich ihm zwei Optionen:

1. Er kann sich vor seiner Studie die informativsten SNPs in einem gewählten Chromosomabschnitt bestimmen lassen.
2. Zu einer gegebenen, typisierten Menge SNPs, kann er sich weitere, maximal aussichtsreiche SNPs, zur weiteren Typisierung vorschlagen lassen.

Herangehensweise

Es wird eine Datenbank modelliert, welche populationsspezifische Genotyp Daten aus dem Hapmap Projekt [3], der Affymetrix [4]-und ABI Datenbank [5] vereint. Diese Genotypdaten und die dazugehörenden SNPs bilden die Grundlage für die Berechnung der SNP-Auswahl. Hierzu müssen alle in der Datenbank vorhandenen SNPs in dem gewählten Intervall und der gewählten Population paarweise in ihren

gegebenen Haplotypfrequenzen und Entropien evaluiert werden. Diese Berechnung hat auf chromosomweiten Intervallen quadratische Komplexität. Die paarweisen Haplotypfrequenzen werden populationsspezifisch, *a priori* in einer sogenannten Haplotypmatrix ausgerechnet und in der Datenbank gespeichert. Zur Reduzierung der quadratischen Komplexität wird der Swept Radius, ein Maß für die Rekombinationshäufigkeit (Morton et al. 2002 [2]) verwendet, der pro SNP berechnet werden muß. Der Swept Radius ist eine biologische Konstante und gibt zu einem SNP an, ab welchem Abstand keine Marker mehr betrachtet werden müssen und somit aus der Haplotypmatrix ausgeschlossen werden können. Der Nutzer erhält nicht die Möglichkeit die Datenbasis zu verändern. Ein Export der Genotypen bleibt ihm ebenfalls verwehrt. In einer Standalone Java-Applikation werden administrative Teilaufgaben der Datenbank, wie das Importieren der SNPs und Genotypdaten und die Berechnung des Sweptradius und der Haplotypmatrix, implementiert. Wenn die wichtigsten Teilgebiete dieses Projektes funktionieren, wird zur Datenbank ein einfaches Webfrontend auf Basis dynamisch generierter HTML-Seiten programmiert.

Kooperation

Diese Studienarbeit wird in Kooperation mit der Mucosa Research Group durchgeführt, einem Hochdurchsatz SNP-Genotypisierungszentrum in dem die SNP Selektion in einem praxisnahen Umfeld erprobt wird.

Literatur

[1] Jochen Hampe, Stefan Schreiber, Michael Krawczak. *Entropy-based SNP selection for genetic Association studies*. Human Genetic 2003 Vol. 114: 36-43

[2] N.E. Morton, W. Zhang, P. Taillon-Miller, S. Ennis, P.-Y. Kwok, A. Collins. *The optimal measure of allelic association.* PNAS 2001 Vol. 98 No. 9: 5217-5221

[3] Hapmap Projekt Seite: http://www.hapmap.org

[4] Affymetrix: https://www.affymetrix.com

[5] ABI: http://www.appliedbiosystems.com/