Java Applikation zur Klassifikation von Infektionserregern aus Microarraydaten

Adl C[1], Vierlinger K[2], Wiesinger-Mayr H[2], Nöhammer C[2], Schreier G[1]
[1]*ARC Seibersdorf research GmbH, Medizintechnik, eHealth Systems, Wiener Neustadt, Österreich*
[2]*ARC Seibersdorf research GmbH, Life Sciences, Molekulare Diagnostik, Seibersdorf, Österreich*
christoph.adl@arcsmed.at

Einleitung und Fragestellung Die korrekte und schnelle Klassifikation eines Infektionserregers ist für eine erfolgreiche und gezielte Therapie von großer Bedeutung. Neben der Identifikation des Erregers selbst, ist vor allem die frühzeitige Erkennung allfälliger Antibiotika-Resistenzen des Erregers ein wichtiger Schritt zur effizienten und erfolgreichen Behandlung. Für die Keimidentifikation, sowie zur Bestimmung seines Antibiotika-Resistenz-Spektrums, muss bislang eine Reinkultur des Erregers angelegt werden. Dies nimmt viel Zeit in Anspruch und erfordert somit eine sofortige – lange vor der Diagnose erfolgende – Verabreichung von Antibiotika. Diese birgt, falls der Erreger letztlich eine Resistenz gegen das verabreichte Antibiotikum aufweist, die Gefahr der Behandlung des Patienten mit einem unwirksamen Antibiotikum in sich und trägt zudem zur weiteren Resistenz-Ausbreitung bei. Durch den Einsatz von diagnostischen Microarrays kann sowohl die Klassifikation des Erregers, als auch die Bestimmung dessen Resistenz gegen Antibiotika innerhalb weniger Stunden erfolgen. Wie dabei vorgegangen wird, zeigt
Abb. 1.

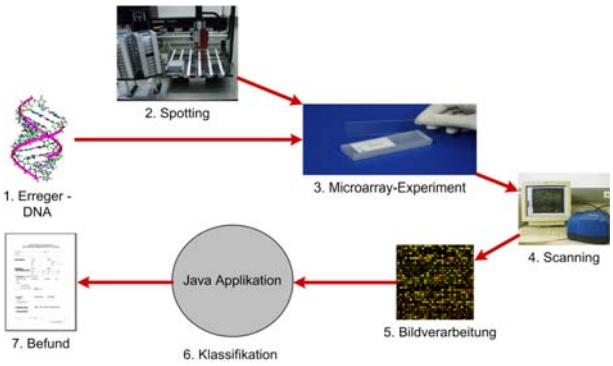

Abb. 1: Prozess der Erreger-Detektion bzw. Antibiotikaresitenzbestimmung mit diagnostischen Microarrays.

Der Vorteil in der Verwendung von diagnostischen Microarrays liegt, im Vergleich zu den derzeit verwendeten Verfahren, im wesentlich schnelleren Ablauf der Klassifikation und Antibiotikaresistenzbestimmung des Erregers. Das bedeutet, dass die Zeit, die bis zu einer effektiven Behandlung vergeht, wesentlich verkürzt wird und die Überlebenschancen des Patienten steigen.
Die hier vorgestellte Anwendung führt die Klassifikation von Infektionserregern, sowie die Erkennung von Antibiotikaresistenzen durch. Dabei kommen Methoden des Machine-Learning zum Einsatz, wobei die Applikation mit einer repräsentativen Auswahl von bekannten Infektionserreger- bzw. Antibiotikaresistenz-Microarrays trainiert wird. Der erstellte Klassifikator kann abgespeichert werden und danach bei der Klassifikation von neu hybridisierten – also nicht zum Training verwendeten – Microarrays zum Einsatz kommen. Dieser Prozess wird in [1] näher beschrieben.
Material und Methoden Implementiert wurde die vorliegende Anwendung als stand-alone Java Applikation. Durch die Verwendung von Java ist eine einfache Interaktion mit dem Statistik-Paket R [2] und dem Machine-Learning-Framework WEKA [3] möglich. R bietet über die Bioconductor-Pakete (http://www.bioconductor.org/, [4]) umfangreiche Möglichkeiten der Verarbeitung von diversen, bei Microarray-Scannern eingesetzten Formaten (z.B. GenePix Result Format – GPR). Um Funktionalitäten aus R in der Anwendung zu nutzen, kommen Rserve und JRclient (http://rosuda.org/Rserve/) zum Einsatz. WEKA ist ein Open Source Machine-Learning-Framework und ebenfalls in Java implementiert, wodurch eine Integration sehr vereinfacht wird.
Die Funktionen der Applikation gliedern sich in vier Assistenten:

1. Assistent zum Trainieren eines neuen Infektionserreger-Klassifikators
2. Assistent zum Klassifizieren neuer Infektionserreger-Microarrays mit einem gelernten Klassifikator (aus Punkt 1.)
3. Assistent zum Trainieren eines neuen Infektionserreger-Antibiotika-Resistenz-Klassifikators
4. Assistent zum Klassifizieren neuer Microarrays hinsichtlich Antibiotikaresistenz mit einem gelernten Klassifikator (aus Punkt 3.)

Der Ablauf des Assistenten zum Trainieren eines neuen Infektionserreger-Klassifikators verläuft dabei in den folgenden fünf Schritten:

Schritt 1 Auswahl einer Anzahl von Microarray-Dateien (z.B. im GenePix Results Format GPR). Jede dieser Dateien enthält das Ergebnis des Scans eines diagnostischen Microarrays (siehe Abb. 1, Schritt 4 und 5) – also für jedes einzelne DNS Stück (Sonde), das am Microarray aufgetragen wurde, Signal-Intensitätswerte, die von einem, in die bakterielle DNA für den Assay eingebauten, Fluoreszenzfarbstoff (z.B. dCTP-Cy3) herrühren. Die Verarbeitung der Microarray-Daten geschieht mit dem limma-Paket (http://bioinf.wehi.edu.au/limma/) von R [2].

Schritt 2 Auswahl von bestimmten Infektionserregern. Hier kann die Menge der verwendeten Microarray-Dateien durch die Auswahl einzelner Erreger noch reduziert werden. Da es bei diesem Assistenten um das Trainieren eines neuen Klassifikators geht, muss für jede Microarray-Datei der zugehörige Erreger bekannt sein. Es wurden bisher nur Testdaten mit genau einem Erreger verwendet, da dies auch in der klinischen Realität überwiegend der Fall ist.

Schritt 3 Auswahl von Sonden des Microarrays, die für die Klassifikation verwendet werden sollen. In diesem Schritt können unspezifische Sonden von der Klassifikation ausgeschlossen werden. Falls bei dieser Auswahl Sonden selektiert werden, die nicht auf allen ausgewählten Microarrays vorhanden sind, werden diese als „missing values“ behandelt. Trotzdem kann eine Klassifikation stattfinden, da solche Werte in WEKA toleriert werden.

Schritt 4 Auswahl des Klassifikators und seiner Parameter, sowie Festlegung der Anzahl der Teilmengen bei der n-fachen stratifizierten Kreuzvalidierung. Dabei kann man aus dem gesamten Repertoire des Machine-Learning-Frameworks WEKA [3] schöpfen.

Schritt 5 Präsentation der Ergebnisse der Kreuzvalidierung, Übersicht über den erzeugten Klassifikator und den Datensatz an Microarray-Dateien (siehe

Schritt 6 Abb. 2). Schließlich kann der Klassifikator noch mit einem Namen gespeichert werden, um ihn danach für die Klassifikation neuer Microarray-Dateien verwenden zu können.

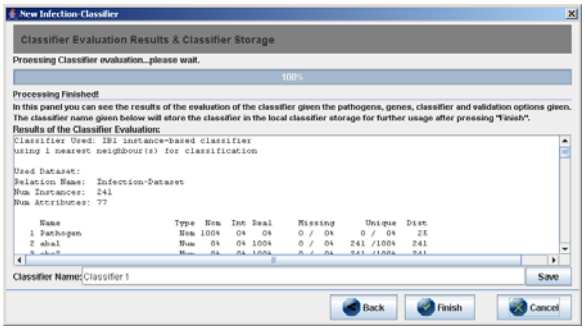

Abb. 2: Screenshot des letzten Schritts des Assistenten zum Trainieren eines neuen Infektionserreger-Klassifikators. Angezeigt werden Informationen über den trainierten Klassifikator, den verwendeten Microarray-Datensatz, sowie das Ergebnis einer n-fachen Kreuzvalidierung.

Der Assistent zum Trainieren eines neuen Antibiotikaresistenz-Klassifikators wird ähnlich funktionieren. Anstatt erregerspezifischer Sonden werden auf den Microarrays Antibiotikaresistenz-Sonden eingesetzt. Bei den Assistenten zur Klassifikation neuer Microarrays mit bestehenden Klassifikatoren (Punkte 2 und 4) werden eine Menge von Microarray-Dateien und ein bereits gelernter Klassifikator ausgewählt. Als Ergebnis erhält man für jeden zu klassifizierenden Microarray den wahrscheinlichsten Erreger bzw. eine Menge von Antibiotika, gegen die dieser Erreger resistent ist – gemeinsam mit einer Abschätzung über die Richtigkeit der Klassifikation.
Ergebnisse Das Ergebnis der Implementierung (
Abb. 2) zeigt, dass sich durch den kombinierten Einsatz von R und WEKA die Vorteile beider Systeme optimal nutzen lassen. Die gute Unterstützung diverser, in der molekularen Diagnostik verwendeten Dateiformate durch R, das breite Spektrum an Klassifikatoren, Validierungsverfahren und Visualisierungen von WEKA und die einfache Interaktion beider Tools mit Java, waren die Basis für eine schnelle und einfache Implementierung der

Applikation. Bereits fertig umgesetzt sind die Assistenten zum Trainieren eines neuen Infektionserreger-Klassifikator (Punkt 1) und zum Klassifizieren neuer Infektionserreger-Microarrays mit einem gelernten Klassifikator (Punkt 2).
Die besten Ergebnisse mit einem Test-Datensatz von 241 Microarrays konnten mit einem Nearest-Neighbour Klassifikator erzielt werden (vgl. [1]). Über 96% der 25 verschiedenen getesteten Erreger-Arten konnten richtig zugeordnet werden. Bei der Beschränkung auf die Gattung – was für eine Behandlung bereits ausreichend ist – waren sogar alle Klassifikationen des Test-Datensatzes richtig (Evaluierungsmethode: Leave-One-Out Kreuzvalidierung).
Diskussion und Ausblick Die viel versprechenden Ergebnisse müssen noch im klinischen Umfeld bestätigt werden. Durch die offene Architektur der Applikation ist es möglich, diese auch in anderen Bereichen einzusetzen, in denen diagnostische Microarrays Anwendung finden – beispielsweise in der Tumordiagnostik. Die vorgestellte Applikation zeigt, dass die Kombination von Machine-Learning-Verfahren und modernen molekularbiologischen Verfahren für eine schnelle und effiziente Infektionsdiagnostik wesentliche Vorteile gegenüber bisherigen Verfahren bringen kann.

Literatur

[1] Wiesinger-Mayr H, Vierlinger K, Pichler R, Presterl E, Hirschl A, Bodrossy L, Nöhammer C, Human pathogen identification using DNA chips including statistical evaluation of microarray data sets, 2006, in preparation.
[2] R Development Core Team, R: A Language and Environment for Statistical Computing, Vienna, Austria, 2005, http://www.R-project.org.
[3] Witten I H, Frank E. Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
[4] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology 2004;5:R80.