Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

ZBH Zentrum für Bioinformatik

Prof. Dr. M. Rarey, K. Schomburg **24.10.2013**

2. Übung zur Vorlesung Grundlagen der Chemieinformatik

Abgabe als pdf bis 01.11.2013 8:00h an gci-uebung@zbh.uni-hamburg.de

Aufgabe 1: Patente (A) (4 Punkte)

Unter `www.freepatentsonline.com` kann man nach Patenten weltweit suchen. Geben Sie hier "Acyclovir" ein.

- Wofür genau ist das Patent, welches als oberstes in der Liste erscheint (Patentnummer US20100069410)?
- Wann wurde es eingereicht und wann veröffentlicht?

Schauen Sie sich das abgelegte pdf an.

- Was wird mit den Diagrammen gezeigt?
- Auf Seite 53 des pdfs finden Sie eine Markush-Struktur des Purins. Schauen Sie sich die variablen Substituenten/Positionen an. Glauben Sie, dass alle Strukturen welche über die Markush-Struktur beschrieben werden, die gleiche antivirale Aktivität haben und warum?
- Welche Informationen aus dem Patent würden Sie interessieren, wenn Sie selbst ein Patent zu einem antiviralen Wirkstoff einreichen wollten?

Aufgabe 2: Workflowtool Orange (A) (8 Punkte)

Die Software Orange Canvas ist ein Open-Source-Workflow Tool, welches durch eine interaktive GUI erlaubt, statistische Auswertungen, Visualisierungen und Maschinelle Lernverfahren über Widgets zu kombinieren. Die Steuerung ist ebenfalls über die Skriptsprache Python möglich, in dieser Übung können Sie allerdings alles über die GUI steuern, wenn Sie nicht so vertraut mit Python sind. Durch die Einteilung in Widgets

sind die Module einfach miteinander kombinier- und austauschbar. Die Software ist sowohl für Experten als auch für Anfänger geeignet. Heute werden Sie die Software je nach Ihrem Wissensstand eher als Blackbox benutzen. Im Laufe der Vorlesung werden Sie aber die verwendeten Methoden und Verfahren kennen lernen.

Sie können sich die Software selber von `http://orange.biolab.si/` auf ihren eigenen Rechner herunterladen, oder eine vorinstallierte Version im ZBH benutzen. Dafür starten Sie Orange Canvas mit dem Befehl:

```
sh /usr/local/zbhtools/orange/Orange-2.7.2/bin/start_orange-canvas.sh
```

Wenn Sie die Version aus dem ZBH nutzen, ist die Auswahl an Visualisierungs-Widgets begrenzt. Bitte geben Sie auf Ihrer Abgabe an, ob sie die ZBH-Variante oder eine eigens installierte Version genutzt haben.

Schauen Sie sich nun als erstes die Tutorials an. Unter
`http://orange.biolab.si/start-using/`
finden Sie beschrieben, wie Sie Daten in die Workflows der drei Tutorien laden. Machen Sie sich mit dem Tool vertraut. Auf der Webseite finden Sie weitere Informationen zu den einzelnen Widgets.

Sie sollen nun analysieren, inwiefern Moleküleigenschaften mit Molekülklassen und Aktivitäten gegenüber bestimmten Targets korrelieren. Als erstes sollen Sie sich Eigenschaften von Molekülen anschauen, die der ACE (Acetylcholinesterase) gegenüber Aktivität zeigen. Dafür nutzen Sie die Datei `ace.csv` aus dem zip-Archiv `data.zip`.
Bauen Sie sich einen Workflow, der eine Datei einlesen kann, den Inhalt der Datei anzeigen kann und Eigenschaften visualisiert. Schauen Sie sich die unterschiedlichen Widgets an, mit denen sie Eigenschaften visualisieren können und bauen Sie in den Workflow eine oder mehrere Visualisierungen ein, die nützlich sein können, um die Eigenschaften zu beurteilen. Machen Sie einen Screenshot von ihrem Workflow, nachdem Sie die einzelnen Widgets mit den Kommentarmöglichkeiten über die Pfeile und den Textboxen mit einer kurzen Beschreibung versehen haben. Machen Sie ebenfalls einen Screenshot von einem Beispiel einer von Ihnen gewählten Visualisierung und begründen Sie, warum Sie diese als sinnvoll betrachten. Fügen Sie die Screenshots in Ihr Abgabedokument ein.

Als nächstes wollen Sie Moleküle miteinander vergleichen, die für unterschiedliche Proteine Aktivität zeigen. Dafür verwenden Sie die Datei `ace_cox1_cox2.csv` aus dem zip-Archiv `data.zip`. In dieser Datei finden Sie Eigenschaften von Molekülen welche für die ACE und für zwei verschiedene Formen der Cyclooxygenase, COX1 und COX2, Aktivitäten zeigen.
Wählen Sie nun erneut eine Visualisierung, die Ihnen sinnvoll erscheint und machen Sie einen exemplarischen Screenshot der Visualisierung. Versuchen Sie nun, die Daten zu klassifizieren. Verwenden Sie das Distanz-Maß “Example Distance“ zum Clustern und

lassen Sie sich das Ergebnis mit “Hierarchical Clustering“ anzeigen (wählen Sie hier als Annotation das Attribut “type“ aus). Was ist der Unterschied, wenn sie unterschiedliche Linkage-Verfahren benutzen? Finden Sie, dass das Clustering gut in der Lage ist, zwischen den Molekülen, welche an die ACE, die COX1 oder die COX2 binden, zu unterscheiden? Machen Sie einen Screenshot ihres Workflows für Ihre Abgabe.

Als letztes wollen Sie evaluieren, ob maschinelle Lernverfahren in der Lage wären, die Moleküle anhand ihrer Eigenschaften richtig nach ihrer Aktivität zu klassifizieren. Oft werden maschinelle Lernverfahren als Black Box genutzt. Vorher muss getestet werden, ob die Verfahren in der Lage sind, das vorliegende Problem zu klassifizieren. Dazu reduzieren Sie zunächst die Variablen, indem Sie die “Rank“ Funktionalität nutzen. Hier stellen Sie bitte beim Scoring “Information Gain“ ein. Verknüpfen Sie ein “Data Table“-Widget mit dem “Rank“-Widget. Hier sehen Sie nun nur noch die übrig gebliebenen Attribute. Auf diesen sollen Sie nun die Lernverfahren SVM, Naive Bayes und Random Forest testen. Hierbei können Sie ähnlich wie im Tutorial die drei Widgets mit einem “Test Learners”’ -Widget verbinden und von Ihrem gerade erzeugtem “Data Table“-Widget ebenfalls eine Verbindung zu dem “Test Learners“-Widget ziehen. Vergleichen Sie nun die drei Lernverfahren, indem Sie sich die “Classification accuracy“, “Sensitivity“, “Specificity“ und die “Area under ROC curve“ anschauen. Recherchieren Sie die Begriffe und erläutern Sie diese in ein bis zwei Sätzen. Welches maschinelle Lernverfahren würden Sie wählen und warum? Machen Sie ebenfalls einen Screenshot von ihrem Workflow für die Abgabe.