Anhang

A Zusammenfassung

Mit Einführung der *Microarray*-Technologie begann ein neues Kapitel in der statistischen Bioinformatik: Mit Hilfe dieser Mikrochips lässt sich der Momentanzustand einer Zelle auf Transkriptionsebene festhalten. Das bedeutet, dass man mit einem einzigen Experiment die Aktivität – die Expression – von mehreren Zehntausend Genen beobachten kann. Entnimmt man nun Proben aus verschiedenen Geweben und vergleicht die Expression innerhalb der Proben miteinander, so stellen sich zwei Fragen. Erstens, gibt es Gene, deren Expression sich zwischen den Proben unterscheidet? Und zweitens, wenn wir Unterschiede finden, sind diese auch signifikant?

Die erste Frage lässt sich beantworten, indem man beispielsweise die Differenz der Expressionsmittelwerte in den unterschiedlichen Proben berechnet. Die Signifikanz drückt sich dann im p-Wert aus. Er gibt an, wie wahrscheinlich es ist, rein zufällig eine gleiche oder höhere Differenz zu beobachten. Je kleiner der p-Wert, desto eher können wir uns darauf verlassen, dass sich die Genexpression tatsächlich unterscheidet. Die entsprechenden Gene nennen wir *differenziell exprimiert*. Allerdings kann man auch bei kleinen p-Werten nicht ausschließen, dass es sich um einen zufälligen Befund handelt. Erschwerend kommt hinzu, dass biologische Daten häufig starken Schwankungen unterliegen. Um den Kreis der Kandidaten einzuschränken, wendet man so genannte *p-Wert-Filter* an. Der Filter legt den maximalen p-Wert fest, bis zu dem die Unterschiede als signifikant betrachtet werden. Stellt man sich die Verteilung der beobachteten p-Werte als Histogramm vor, so definiert der Filter eine vertikale Trennung, nämlich in die signifikanten Gene mit zulässigem kleinen p-Wert und in die nicht-signifikanten Gene mit zu großem p-Wert.

P-Wert-Filter eignen sich für Experimente mit wenigen Genen. Je höher die Anzahl der Gene jedoch ist, desto stärker fällt neben den zufälligen Schwankungen eine weitere Eigenart der Daten ins Gewicht: Ein großer Teil der Gene zeigt keine

Unterschiede, sondern wird gleichmäßig in allen Proben exprimiert. Die Theorie besagt, dass die p-Werte solcher konsistenten Gene gleichverteilt sind. Die Gleichverteilung definiert eine horizontale Trennung des p-Wert-Histogramms. Die Höhe der Trennlinie liefert einen natürlichen Schätzer für den Anteil der konsistenten Gene im Experiment. Die Verteilung der p-Werte *über* der Linie ermöglicht Aussagen über Anzahl und vor allem Verteilung der differenziell exprimierten Gene.

Es liegt in der Definition eines p-Wert-Filters, dass die Liste der signifikanten Gene auch konsistente Gene enthalten wird. Man wird immer einen gleichverteilten Anteil in die Liste mit einschließen. In einem gewöhnlichen *Microarray*-Experiment beobachtet man jedoch häufiger kleine p-Werte als große. Die Dichtefunktion der p-Werte fällt bis auf die Höhe der Gleichverteilung ab. Da eben der Anteil der gleichverteilten p-Werte immer gleich bleibt, haben nun Gene mit kleinem p-Wert eine höhere Wahrscheinlichkeit differenziell exprimiert zu sein als Gene mit höheren p-Werten. Schätzt man die Wahrscheinlichkeit differenzieller Expression für jedes Gen, so beantwortet dies obige zweite Frage. Diese Wahrscheinlichkeit nennt man *local false discovery rate.* Die lokale *false discovery rate* eignet sich hervorragend für die großen Datensätze, die bei *Microarray*-Experimenten anfallen. Je mehr p-Werte, desto verlässlicher gelingt die Schätzung.

In der vorliegenden Arbeit schlagen wir zwei neue Konzepte vor, um die Schätzung der lokalen *false discovery rate* zu verbessern. In Kapitel 5 stellen wir eine iterative Schätzmethode vor. Verglichen mit dreizehn anderen Verfahren lieferte unser Schätzer sehr gute Ergebnisse in einer Simulationsstudie. Kapitel 6 widmet sich einer Beobachtung, die in der Literatur unseres Wissens nach noch nicht behandelt wurde: Da die Gene und somit auch ihre gemessenen Expressionswerte untereinander stark korrelieren, benutzt man zur Berechnung der p-Werte häufig ein unzureichendes Hintergrundmodell. Wir stellen einen Algorithmus vor, der die Schätzung des Hintergrundmodells verbessert. Diese Verbesserung trägt gleichzeitig dazu bei, dass die lokale *false discovery rate* akkurater geschätzt werden kann.

Publikationen:

1. Scheid S and Spang R (2007). Compensating for unknown confounders in microarray data analysis using filtered permutations. Accepted for *Journal of Computational Biology*.
2. Scheid S and Spang R (2006). Permutation filtering: A novel concept for significance analysis of large-scale genomic data. In: Apostolico A, Guerra C, Istrail S, Pevzner P, and Waterman M (Eds.), *Research in Computational Molecular Biology: 10th Annual International Conference, Proceedings of RECOMB 2006, Venice, Italy, April 2-5, 2006*, Lecture Notes in Computer Science vol. 3909, Springer, Heidelberg, pp. 338–347. ISSN: 03029743, ISBN: 3540332952.
3. Lottaz C, Yang X, Scheid S, and Spang R (2006). OrderedList - a Bioconductor package for detecting similarity in ordered gene lists. *Bioinformatics* **22**(18):2315–2316.
4. Yang X, Bentink S, Scheid S, and Spang R (2006). Similarities of ordered gene lists. *Journal of Bioinformatics and Computational Biology* **4**(3):693–708.
5. Scheid S, Lottaz C, Yang X, and Spang R (2006). Similarities of ordered gene lists - User's guide to the Bioconductor package OrderedList. CompDiag Technical Report Nr. 2006/01.
6. Scheid S and Spang R (2005). Chapter 18 - Microarray data analysis: Differential gene expression. In: Nuber U (Ed.), *DNA Microarrays*, Taylor & Francis Group, UK. ISBN: 0415358663.
7. Westhoff TH, Scheid S, Tölle M, Kaynak B, Schmidt S, Zidek W, Sperling S, and van der Giet M (2005). A physiogenomic approach to study the regulation of blood pressure. *Physiological Genomics* **23**(1):46–53.
8. Scheid S and Spang R (2005). twilight; a Bioconductor package for estimating the local false discovery rate. *Bioinformatics* **21**(12):2921–2922.
9. Scheid S (2005). Correspondence clustering of Dortmund city districts. In: Weihs C and Gaul W (Eds.), *Classification - The Ubiquitous Challenge. Proceedings of the 28th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Dortmund, March 9-11, 2004*, Springer, Heidelberg. ISBN: 3540256776.
10. Scheid S and Spang R (2004). A stochastic downhill search algorithm for estimating the local false discovery rate. *IEEE Transactions on Computational Biology and Bioinformatics* **1**(3):98–108.

11. Scheid S and Spang R (2004). Estimation of local false discovery rate - User's guide to the Bioconductor package twilight. CompDiag Technical Report Nr. 2004/01.
12. Kunz M, Ibrahim SM, Koczan D, Scheid S, Thiesen HJ, and Gross G (2004). DNA microarray technology and its applications in dermatology. *Experimental Dermatology* **13**(10):593–606.
13. Grzeskowiak R, Witt H, Drungowski M, Thermann R, Hennig S, Perrot A, Osterziel KJ, Klingbiel D, Scheid S, Spang R, Lehrach H, and Ruiz P (2003). Expression profiling of human idiopathic dilated cardiomyopathy. *Cardiovascular Research* **59**(2):400–11.
14. Scheid S and Spang R (2003). A false discovery rate approach to separate the score distributions of induced and non-induced genes. In: Hornik K, Leisch F, and Zeileis A (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Technical University of Vienna, Austria. ISSN: 1609-395X.
15. Scheid S (2001). Die verallgemeinerte Lognormalverteilung (The generalized lognormal distribution). Master Thesis, University of Dortmund.

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Darmstadt, im Dezember 2006 Stefanie Scheid