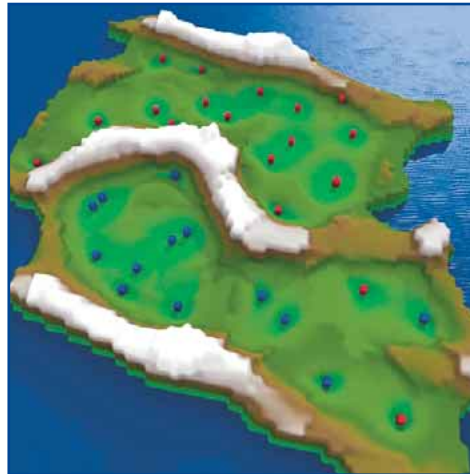
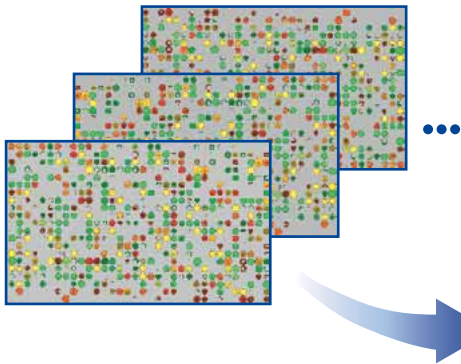


# Emergente Data Mining Methoden

für die Diagnose von Tumoren mit DNA Microarrays

Philipps-Universität Marburg



## Neuroblastome

Neuroblastome sind bösartige Krebserkrankungen die im Kindesalter auftreten. Sie sind die dritthäufigste bösartige Krebserkrankung im Kindesalter. Bundesweit erkranken jährlich ca. 150-200 Kinder an diesem Tumor. Der Tumor entwickelt sich aus Vorläuferzellen des autonomen Nervensystems, welches die unwillkürlichen Funktionen, wie Herz- und Kreislauf, Darm- und Blasenaktivität, steuert. Es sterben insgesamt ca. 40 % der erkrankten Kinder innerhalb der ersten fünf Jahre.

Tritt der Tumor im Säuglingsalter auf, so kann es jedoch sein, dass er sich ohne jede Behandlung zurückbildet. Ein wichtiges Forschungsziel ist es daher, diese harmlose Form des Neuroblastoms von Tumoren abzugrenzen, die schlechte Überlebenschancen bieten.

## DNA-Microarrays

DNA-Microarrays sind ein modernes, gentechnisches Verfahren, bei dem die Genaktivität, die sog. Expression, von tausenden verschiedener Gene in einer Zelle simultan gemessen wird. Die Aktivität bestimmter Gene in einer Zelle ist ein Indikator für den Zellzustand und die Aufgaben die gerade von der Zelle wahrgenommen werden. Insbesondere in der Tumor-Forschung erhofft man sich durch dieses Verfahren einen Einblick in die Mechanismen des Tumorgeschehens. DNA-Microarrays stellen eine Anordnung von tausenden verschiedener DNA Fragmente dar. An diesen können die Gene ankoppeln, die in der zu messenden Zelle vorkommen.

Die Expressionsrate, d. h. die Aktivität von Genen wird durch ein Lumineszenzverfahren gemessen. Hierzu werden die einzelnen Gen-Messpunkte zum Aussenden von Licht angeregt. Das Helligkeitsmuster dient als Messgröße für die Expressionsrate von Genen. Kennzeichnend für diese Technik ist, dass sehr viele Variablen (die Gene) zugleich an einer Probe gemessen werden.

Hergestellt werden DNA-Microarrays mit Verfahren, die auch bei der Herstellung von Computerchips üblich sind. DNA-Microarrays werden daher auch „Gen-Chips“ genannt.

In der hier beschriebenen Anwendung wurden von der Arbeitsgruppe Molekulare Biologie und Tumorforschung (Prof. Eilers) 4608 Gen-Expressionen für jeden Patienten in verschiedenen Stadien von Neuroblastom-Tumoren gemessen.

Prof. Dr. Alfred Ultsch, Neuroinformatik und Künstliche Intelligenz  
ultsch@informatik.uni-marburg.de

Prof. M. Eilers, Institut für Molekulare Biologie und Tumorforschung  
Dipl.-Inform. Ulrich Penndorf  
penndorf@mathematik.uni-marburg.de

Philipps-Universität Marburg  
Hans-Meerwein-Straße  
D-35032 Marburg

Telefon +49 (0) 64 21 / 2 82-21 85  
Telefax +49 (0) 64 21 / 2 82-89 02

### Emergente Data Mining Methoden

Die Arbeitsgruppe Neuroinformatik (Prof. Ultsch) hat Verfahren entwickelt, mit denen wichtige Eigenschaften hochdimensionaler Daten entdeckt und klassifiziert werden können. Das sog. U-Matrix Verfahren nutzt die Fähigkeit der Selbstorganisation von Neuronalen Netzen oder künstlichen Lebensformen (Artificial Life) um hochdimensionale Daten sinnvoll arrangieren zu können. Die dabei entstehenden Strukturen werden als Landschaft sichtbar gemacht. Die Struktur solcher Datenlandschaften bieten einen direkten Einblick in die Zusammenhänge im ansonsten nicht anschaulichen Raum der Gen-Expressionen. Mit diesem Verfahren können die bei DNA-Microarrays entstehenden Daten, im vorliegenden Fall 4608 Expressionsdaten pro Microarray, direkt visualisiert werden.

Ein hoher Berg zwischen den Bildern von Microarrays auf der Landschaftsdarstellung der U-Matrix bedeutet einen deutlichen Unterschied im Vorhandensein von Genen in einer Zelle. Aus der Struktur dieser Landschaft kann somit auf die Gleichheit oder Verschiedenheit von Gen-Expressionen geschlossen werden.

Die hier gezeigte U-Matrix wurde aus DNA-Microarrays von Patienten mit Neuroblastom-Tumoren konstruiert. Die roten und blauen Objekte stellen dabei jeweils ein Microarray dar. Die Struktur der Landschaft zeigt eine deutliche Auftrennung der Microarrays in zwei Gruppen. Die Farbgebung der Objekte stellt eine Eigenschaft dar, die nicht zur Konstruktion der U-Matrix verwendet wurde. Rot dargestellt sind Tumore die eine bekannte genetische Eigenart, das mehrfache Vorkommen des sog. N-myc Genes, besitzen. In der Regel bedeutet dies eine schlechte Prognose für den Patienten. Die Blau dargestellten Microarrays besitzen nur ein einfaches Vorkommen dieses Gens. Die U-Matrix zeigt also, dass dieses Unterscheidungsmerkmal durch eine große Anzahl von Expressionsdaten auf den DNA-Microarrays repräsentiert wird.



### Differenzierung von DNA Arrays mittels sig\*

Die Strukturen dieser Landschaften werden durch das sig\* Verfahren erschlossen. Dieses Verfahren erklärt die gesehene Strukturen in möglichst prägnanter und verständlicher Form als Entscheidungsregeln. Bei der Analyse von DNA-Microarrays dient das sig\* Verfahren zur Identifikation von Genen die für eine Diagnose von entscheidender Bedeutung sind.

Als eine Anwendung des sig\* Verfahrens wurden die bedeutsamen Gene für zwei verschiedene Tumorstadien identifiziert. Hierzu wurden Daten von Patienten mit Tumorstadium 1 und 4 betrachtet. Tumorstadium 1 bedeutet ein relativ leichtes, Tumorstadium 4 ein relativ schweres Stadium der Erkrankung.

Das sig\* Verfahren hat 17 Gene identifiziert, mit denen eine Unterscheidung der beiden Tumorstadien zuverlässig und sicher gefällt werden kann.

### Identifikation bedeutsamer Gene

Für die Entwicklung von praktisch einsetzbaren gen-diagnostischen Verfahren ist es wichtig die für eine Diagnose entscheidenden Gene zu kennen. Auch aus Kostengründen verbietet sich eine Messung von Tausenden von Genen. Es galt also eine Methode zu entwickeln, welche die kleinstmögliche Menge von Gene identifiziert, mit denen Krankheiten zuverlässig diagnostiziert werden können.

Die Kombination der in der Arbeitsgruppe Neuroinformatik (Prof. Ultsch) entwickelten Verfahren U-Matrix und sig\* ermöglicht es, die Datenflut von DNA-Microarrays auf eine überschaubare Menge von diagnostisch bedeutsamen Genen zu reduzieren. Mit diesen Verfahren ist es möglich, mit einer relativ kleinen Menge von Genen, in der Regel zwischen 20 und 80 Genen, sichere diagnostische Entscheidungen zu treffen.

