

Größenbestimmung bei Microarrayexperimenten Klassenvergleiche und Classifier

Inhaltsverzeichnis:

0.1	Summary	Seite 2
0.2	Microarrays	↓
1.0	Einführung und Hintergrund	
2.0	Definition und Notation	Seite 3
3.0	Single Label Microarrays	↓
4.0	Dual Label Microarrays	Seite 4
	4.1 Simple Reference Design	↓
	4.1.1 Technische Replikate und Dyeswap	
	4.2 Balanced Block Design	Seite 5
	4.3 Single paired Design und dye-swap paired Design	Seite 6
5.0	Versuchsgröße bei der Prognostic Marker Bestimmung	↓
	5.1 Effekt von Pooling	
6.0	Wahl des Signifikanzlevels	
7.0	Auswahl der Varianzparameter	Seite 7
8.0	Auswirkung von kleinen n	↓
9.0	Ergebnisse der Formeln	Seite 8
10.0	Training von Classifiern	
11.0	Fazit	Seite 9
12.0	Anhang	↓
	12.1 Versuchsgrößenberechnung für Single Channel Arrays	Seite 10
	12.2 Bijektive Funktion zur Überführung von Single- in Dual-Label-Design	
	12.3 Versuchsgröße bei Pooling	↓
	12.4 Überblick über Formeln zur Versuchsgrößenberechnung	Seite 11
13.0	Literaturangabe	↓

0.1 || Summary:

Die Versuchsgrößenbestimmung bei Microarray Experimenten ist sehr wichtig, da die Versuchsgröße maßgeblich die Kosten beeinflusst. Allerdings macht die Komplexität und die zu analysierende Datenmenge dies sehr schwierig, so dass viele Forscher verleitet sind, die Versuchsgröße nur zu schätzen, anstatt sie genau zu berechnen.

Hier wird nun die Herleitung von Formeln auf Basis des Papers von Kevin Dobbin und Richard Simon¹ gezeigt, um die Versuchsgröße bei Klassenvergleichen und der Entwicklung von Prognostic Markern zu bestimmen. Dabei wird auch auf den Einfluss von „Pooling“, technischen Replikaten und „Dye Swap“ eingegangen. Dabei werden eine Vielzahl von typischen Experiment-Situationen mit Single Layer und double Layer Microarrays bedacht.

Die Berechnungen basieren auf einfachen stochastischen Modellen und kommen zu einfach anzuwendenden Formeln für die Versuchsgrößenbestimmung.

Im zweiten Teil wird ein Modell zum Trainieren eines Classifiers vorgestellt. Auch hier geht es um die Berechnung der benötigten Versuchsgröße, um den Classifier ausreichend zu trainieren. In unserem Fall handelt es sich um ein Sequentielles Modell.

0.2 || Microarrays:

Microarrays dienen zur Erkennung von Genexpressionsprodukten. Sie bestehen neben anderen Materialien aus Glas- oder Keramikplatten, auf denen Gensequenzen eines Organismus aufgebracht sind. Der Ort jedes Gens ist bekannt. Bei sogenannten Oligonukleotidchips der Firma „Affymetrix“ sind Oligonukleotidsequenzen der Gene auf der Platte fixiert, die aus ca 20-25 Oligonukleotiden bestehen. Die Chips besitzen eine wesentlich höhere Spotdichte von bis zu 40.000/cm² im Gegensatz zu normalen cDNA-Chips, deren Spotdichte bei 100-1000 bei einer Größe von 8*12cm liegt.

Die Genprodukte der zu untersuchenden Zellen werden auf den Chip gegeben. Sie sind mit Fluoreszenzmarkern versehen. Diese Genprodukte, auch „Targets“, binden nun mit den komplementären Strängen auf dem Chip. Durch eine anschließende Waschung werden noch nicht-, oder unzureichend gebundene Targets entfernt.

Durch einen Laser können nun die Fluoreszenzfarbstoffe zum Leuchten gebracht

werden. Von einer Kamera wird ein Bild des Chips festgehalten. Dabei werden ein Laser und ein Bild pro Farbe benötigt.

Das Bild wird anschließend noch bearbeitet. Erst wird es normalisiert, hierbei wird eine Skalierung der Leuchtstärke vorgenommen, um die Ergebnisse mit denen anderer Experimente vergleichen zu können.

Nun werden noch vorhandene Fehler im Bild, die zum Beispiel durch Salzkristalle entstanden sind, beseitigt.

Die gewonnenen Daten werden letztendlich in der Genexpressionsmatrix gespeichert.

1.0 || Einführung und Hintergrund

Microarrayexperimente sind oft komplex, generieren eine große Masse von Daten und bedürfen sorgfältiger Planung. Die hier präsentierten Formeln können genutzt werden, um die Versuchsgröße zu bestimmen, sowie ein breites Spektrum an experimentellen Zielen, einschließlich des Klassen – Vergleiches und der Identifikation von auf Gen -Expression basierenden Markern zu erreichen. Laut der Microarray Literatur gehört der Klassen –Vergleich zu Experimenten, die zum Ziel haben, zwei verschiedene Klassen von Proben zu vergleichen (z.B. Krebsgewebe im Gegensatz zu normalen Gewebe des selben Organs oder histologisch verschiedene Typen von Krebs-Exemplaren), normalerweise um die unterschiedliche Genexpressionen in den zwei Typen zu identifizieren

Es wurden relativ wenig Arbeiten in der Microarray Literatur veröffentlicht, die sich mit der Bestimmung der benötigten Versuche für Klassen – Vergleichsexperimente oder für die Entwicklung von Prognostic Markern beschäftigen.

Wir konzentrieren uns hier auf zwei statistische Ansätze: (1) Klassen-Vergleich und (2) Prognostic –Marker Entwicklung, bei denen das Ziel ist, einen Multigen-Vorhersagemarker zu konstruieren. Das sind übliche Ziele in Microarrayexperimenten in der Krebsforschung. Die Gleichungen für die Samplegröße bieten Einsicht in die Auswirkungen, die Varianzparameter und ExperimentDesign auf die benötigte Versuchsgröße haben, welche „Black Box“ Algorithmen und Simulationen manchmal nicht bieten. In der Praxis, vor allem bei kleinen Versuchsgrößen, sind allerdings statistisch arbeitende Computerprogramme zur Abschätzung der Versuchsgröße vorzuziehen.

Beim Vergleich von Genklassen ist man an der unterschiedlichen Expression der einzelnen Gene interessiert.

Wenn man einen Prognostic Marker entwickeln möchte startet man, auch wenn das Ziel ein

¹ Kevin Dobbin & Richard Simon (2005)

Multigen Marker ist damit, im ersten Schritt die Gene zu identifizieren, die mit dem Ausbrechen der Krankheit unmittelbar in Verbindung stehen.

Auf diesen Ergebnissen werden dann alle weiteren Entwicklungen basieren, weshalb es auch in diesem Paper um Einzelgenvergleiche geht.

Alle Sample-Größen Formeln, die hier präsentiert werden, basieren auf der Annahme eines normal linearen Modells für jedes Gen. Diese Annahme ist angemessen nahe an der Wahrheit für log-Intensive Daten, weil viele Microarray Experimente, die diesen Ansatz verwenden, biologisch bedeutende Differential-Ausdrücke entwickelten. Diese konnten durch andere unabhängige Verfahren verifiziert werden. In Kapitel 9.0 werden dazu noch einige empirische Untersuchungen gezeigt.

2.0 || Definitionen und Notationen

Wir nehmen an, dass alle Daten hintergrundkorrigiert und normalisiert wurden. Im Falle von Affymetrixdaten simulieren wir die zusammengefasste Intensität für ein Gen und nicht die für die individuellen PM und MM Scores. Wir benutzen eine allgemeine Notation, sowohl für Single- Layer, als auch für Double- Layer Microarray Experimente. Y_{gadvfs} repräsentiert eine Fluoreszenzintensität.

- g: Der Index $g=1,2,\dots,G$ steht für das jeweilige Gen
- a: Der Index $a=1,\dots,n$ für das Array oder die Glasscheibe.
- d: Der Index d für die benutzte Farbe oder Farben, $d=1$ für Single Label Microarrays und $d=1,2$ für Dual Label Microarrays.
- v: Der Index $v=1,2$ steht für die verschiedenen Phänotypen oder die vorhandenen Varianten. Der Einfachheit halber gehen wir generell davon aus, dass es zwei unterschiedliche Typen gibt.
- f: Der Index $f=1,\dots,F$ steht für die biologisch verschiedenen Proben innerhalb eines Phänotyps (z.B. die unterschiedlichen Menschen innerhalb eines Tumor Experimentes oder Mäuse in einem Mäusemodellexperiment). Wir nehmen an, dass jeder Phänotyp von der gleichen Anzahl an Individuen repräsentiert wird.
- s: Letztendlich steht der Index $s=1,\dots,m$ für die Unterproben die aus der gleichen biologischen Quelle entnommen wurden (z.B. eine Gewebeprobe von einem Individuum, die in verschiedene Stücke geteilt wurde, von denen jedes separat in einem Microarray verwendet wird).

Mit jedem Gen g werden zwei verschiedene Level von Variation verbunden:

- τ_g^2 : Biologische Variation aufgrund der Heterogenizität von Individuum zu Individuum der Genexpression innerhalb eines Phänotyps wird dargestellt durch τ_g^2
- σ_g^2 : Die Variation die experimentelle Fehler aufgrund von technischen Ungenauigkeiten wiedergibt, wird durch σ_g^2 dargestellt

In der Formel für die Versuchsgröße bezieht sich m auf die Anzahl der unterschiedlichen Subsamples von jedem Sample (z.B. die Anzahl von technischen Kopien des jeweiligen Samples) und n auf die totale Anzahl im Experiment verwendeter Microarrays.

$z_{\alpha/2}$: $z_{\alpha/2}$ ist die Wahrscheinlichkeit für fälschlicherweise als positiv eingestufte Genexpressionen (false positive).

z_{β} : z_{β} ist die Wahrscheinlichkeit für fälschlicherweise nicht entdeckte Genexpressionen (false negative).

δ : Abschließend ist δ die Distanz zwischen den Klassen-Mitteln.

Für die Variablen τ_g^2 , σ_g^2 , $z_{\alpha/2}$, z_{β} und δ werden dabei Werte gewählt die aus vorherigen Experimenten bekannt sind.

3.0 || Single Label Microarrays

Einige Microarraysysteme, wie das „Affymetrix Array“ nutzen ein Single Label Design. Dabei werden nur die Genexpressionsprodukte eines Organismus auf eine Platte gegeben. Es wurde viel Arbeit geleistet, um ein adäquates Maß der Genexpression für diese Typen von Arrays zu finden [2], sowie eine Methode zur Normalisierung eines Datensets zum Vergleich zweier Experimente zu entwickeln. In diesem Paper wird davon ausgegangen, dass dies bereits geschehen ist, und dass ein hintergrund-korrigiertes, normalisiertes Genexpressionsmaß Y_{gadvfs} für die Gene des Arrays verfügbar ist, die mit folgendem Modell beschrieben werden:

$$\text{Log}(Y_{\text{gadvfs}}) = G_g + GV + (GF)_{\text{gf}(v)} + \epsilon_{\text{gadvfs}}$$

Wobei G_g das durchschnittliche Genexpressionslevel des Gens g in der gesamten Population ist. GV_{gv} ist der Effekt der Klasse oder des Typen und GF_{gf} ist der individuelle Sample Effekt (von einem bestimmten biologischen Individuum). Dieser zufällige Effekt hat die Varianz τ_g^2 . ϵ_{gadvfs} repräsentiert den unabhängigen, normalverteilten Fehler mit der genspezifischen Varianz σ_g^2 .

Im Anhang wird genau darauf eingegangen, wie die Formel für die Anzahl an benötigten Single-Label

Microarrays zustande kommt, wenn die Varianzen bekannt sind. Die Formel lautet:

$$n = 4m \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta} \right]^2 \left(\tau_g^2 + \frac{\sigma_g^2}{m} \right)$$

wobei m die Anzahl an technischen Replikaten pro Sample ist und n die absolute Anzahl an benötigten Microarrays. Die absolute Anzahl an benötigten biologisch verschiedenen Samples ist demnach:

$$n/m = 4 \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta} \right]^2 \left(\tau_g^2 + \frac{\sigma_g^2}{m} \right)$$

Eine Abschätzung der Menge $\tau_g^2 + \frac{\sigma_g^2}{m}$ ist nun

notwendig. Wenn keine technischen Replikationen geplant sind, so dass m=1, entspricht die Menge $\tau_g^2 + \sigma_g^2$ der Varianz der log- Intensität für dieses Gen über die Samples der gleichen Klasse. Wenn technische Vervielfältigungen geplant sind, so dass

m > 1 entspricht, ist die Menge $\tau_g^2 + \frac{\sigma_g^2}{m}$ die

Varianz der durchschnittlichen log Intensität für dieses Gen über die Samples der gleichen Klasse. Wenn für jedes Individuum die Genexpression über den Durchschnitt der m log-Intensität von zu diesem Individuum zugehörigen Arrays abgeschätzt wird, erhalten wir die gesuchte Varianz. In Laboren, die routinemäßig mehrere technische Replikate fertigen, sind meist solche Einschätzungen aufgrund vieler tausend vorheriger Versuche vorhanden.

4.0 || Dual-Label Microarrays

Bei Dual Label Versuchen sind die Expressionsprodukte zweier Organismen auf eine Platte aufgebracht. Beide Regionen sind allerdings streng von einander getrennt.

Durch dieses Verfahren lassen sich beide Proben besser miteinander vergleichen.

Wir nehmen an, dass das normalisierte, hintergrundberichtigte Intensitätsmaß für diese cDNA Microarrayexperimente mit folgendem Modell beschrieben werden kann:

$$\text{Log}(Y_{\text{gadvfs}}) = G_g + GA_{\text{ga}} + GD_{\text{gd}} + GV_{\text{gv}} + (GF)_{\text{gf(v)}} + \epsilon_{\text{gadvfs}}$$

Ein Spot pro Gen pro Array wird angenommen, so dass GA_{ga} einen bestimmten Spot auf dem jeweiligen Array beschreibt. Der GF Term betrachtet einen zufälligen Sample Effekt, der mit dem Klasseneffekt verflochten ist. Dieser Term entspricht einer Normalverteilung mit 0 im Mittelwert und einer Varianz von τ_g^2 . Wir nehmen

an, dass das Maß für Fehler genspezifisch sein könnte, so dass ϵ_{gadvfs} eine unabhängige zufällige

Variable mit der Varianz σ_g^2 ist. Der GD_{gd} Farbeffekt-Term ist nur enthalten, wenn er abschätzbar ist. Im einfachen Referenz Design (siehe 4.1) ist er Z.B. vom G_g Gen Haupteffekt absorbiert.

Dieses Modell für Dual-layer Microarrays ist äquivalent mit dem Model, das für Single-Layer Microarrays präsentiert wurde. Dies wird im Anhang in Sektion 12.2 durch eine bijektive Abbildung gezeigt. Unsere generelle Notation ist durchwegs τ_g^2 für die biologische Variation und

σ_g^2 für die technische Fehler Variation der Log-Intensitäten. Die Interpretation dieser Populations-Parameter wird sich mit dem Kontext verändern. Z.B. würde man nicht den gleichen technischen Fehler für olinukleotid Daten wie für cDNA Daten erwarten. Genauso beeinflussen verschiedene Experiment Designs die Varianz. Daher muss der Kontext (Single Label, Dual Label, Referenz Design, Block Design, Paired Samples) beachtet werden. Wir werden hierbei nicht den Design Auswahl Aspekt betrachten. Zu diesem kann man mehr in ² lesen.

4.1 || Simple Reference Design

Eine Art der Dual Label Microarrays sind die Simple Reference Designs. Hierbei werden auf einer Seite des Chips die Genprodukte des Testorganismus, auf der anderen Seite die einer Referenzprobe betrachtet. So ist immer ein Bezugspunkt vorhanden, um die Ergebnisse zu bewerten und mit Ergebnissen anderer Experimente zu vergleichen.

Eine Formel für die erwartete Anzahl an benötigten Microarrays beim Vergleich zweier Klassen im Simple Reference Design ohne technische Replikate (m) ist :

$$n = 4 \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta} \right]^2 \left(\tau_g^2 + 2\sigma_g^2 \right)$$

wobei jeweils die Hälfte der Arrays jeder Gruppe zugewiesen ist.

Hierbei ist $\tau_g^2 + 2\sigma_g^2$ die Varianz des log-Verhältnisses für Gen g in einer der Typen oder Klassen des RNA Samples.

Im Gegensatz zum Singel Label Design beträgt die technische Varianz $2\sigma_g^2$ da 2 Label genutzt sind.

Die Abschätzung aus vorherigen Daten wird benötigt.

Wenn mehr als 2 Klassen verglichen werden, steigen die Vergleichsmöglichkeiten und die Versuchsgrößenbestimmung wird von den Fehlerraten, die man kontrollieren möchte abhängen.

² Dobin and Simon (2002)
Kendzierski et al. (2003).

4.1.1 || Technische Replikate und „Dye swaps“ im Reference Design

Technische Replikate werden verwendet, wenn nicht genügend Proben zur Verfügung stehen. Jede Probe wird dann mehrfach verwendet. Dabei ist darauf zu achten, dass alle Proben gleich häufig verwendet werden.

Die Formel für die benötigten Samples lautet :

$$n = 4m \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta} \right]^2 \left(\tau_g^2 + \frac{2\sigma_g^2}{m} \right)$$

Die absolute Anzahl an benötigten biologisch unterschiedlichen Samples ist n/m.

Anders als bei der Größenberechnung bei der keine technischen Replikate benutzt werden, wird eine geschätzte Varianz für die Berechnung nicht genügen. Eine separate Schätzung sowohl für die biologische Variation, als auch die Varianz der Experimentfehler wird nötig sein. Da gleiche Proben mehrfach benutzt werden, steigt ihr relativer Einfluss auf das Gesamtergebnis, im Verhältnis zur technischen Variation. Abbildung 1 zeigt die Auswirkung von technischen Replikaten sowohl auf die benötigte Anzahl an Arrays, als auch auf die Anzahl von benötigten Samples für einige typische Microarray Design Situationen. Das Varianzverhältnis ist das Verhältnis von biologischen Varianz zu experimentellen Fehler Varianz $\tau_g^2 / 2\sigma_g^2$ und liegt normalerweise im Bereich zwischen 2 und 10^3 .

Bei sogenannten Dye-Swap Arrays wird die Farbgebung nach einem Versuch vertauscht und der Versuch wiederholt. Damit sollen Fehler, die durch den Farbstoff entstehen, kompensiert werden.

Für sie sollten die gleichen Berechnungen größtenteils anwendbar sein,

obwohl manche sagen, dass Dye Swapping die Erwartungen mehr verbessert als einfache technische Replikationen ohne Dye Swapping. Wie aus der Tabelle (Abb. 1) abgelesen werden kann, gibt es eine

Reduktion bei der Menge der benötigten Samples, wenn jedes Sample mehr als einmal untersucht wurde, aber dieser Vorteil hat auch Kosten, meist in Form von signifikant höheren Anzahlen von Arrays die durchlaufen werden müssen.

Abb.1:

Varianz Verhältnis	Technische Replikate / Sample	Anzahl benötigter Arrays	Anzahl benötigter Samples
2	1	49	49
	2	74	37
	3	99	33
	4	124	31
4	1	49	49
	2	82	41
	3	114	38
	4	148	37

Hier ist der Effekt von technischen Replikaten auf die benötigte Anzahl an Arrays zu sehen. Gewählte Werte: $\alpha=0,001$; $\beta=0,05$; $\delta=1$

Ist n_m die Anzahl an benötigten Arrays, wenn m technische Replikate von jedem Sample auszuführen sind, dann ist die Beziehung zwischen den benötigten Arrays, um eine äquivalente Aussage bezüglich der Zuverlässigkeit des Microarray Experiments mit einem technischen Replikat zu treffen, gegenüber m technischen Replikaten pro Sample :

$$n_m = n_1 \frac{m(\tau_g^2 / \sigma_g^2) + 2}{(\tau_g^2 / \sigma_g^2) + 2}$$

, Zum Beispiel, wenn

$$\frac{\tau_g^2}{\sigma_g^2} = 2 \text{ ist dann ist } n_2 = 1,5n_1 \text{ und } n_3 = 2n_1.$$

4.2 || Balanced Block Design

Das Balanced Block Design ist auch eine Form der Dual Label Experimente. Bei ihm werden also auch auf jeder Seite des Arrays die Expressionsprodukte von verschiedenen Zelltypen aufgetragen. Allerdings sind es im Gegensatz zum Single Reference Design beides Proben, die es zu untersuchen gilt. Hier kann man die zwei benutzten Proben leicht und mit wenig Aufwand miteinander vergleichen. Da aber ein Referenzpunkt fehlt, ist der Vergleich mit anderen Proben ohne weiteren Versuch nicht möglich. Durch diese Spezialisierung ist das Design also nur eingeschränkt zu verwenden.

Verwendet man diese Art von Design, kann die benötigte Anzahl von Arrays in einem Klassenvergleich sinken.

³ Dobin and Simon (2002)
Kendziorski et al. (2003).

Die Anzahl an benötigten Arrays beim Balanced Block Design ist:

$$n = \frac{(z_{\alpha/2} + z_{\beta})^2}{\delta^2} (\tau_{g,1}^2 + \tau_{g,2}^2 + 2\sigma_g^2)$$

wobei $\tau_{g,1}^2$ und $\tau_{g,2}^2$ den Varianzen in den log Intensitäten der biologischen Variationen in der Genexpression der jeweiligen Population entsprechen. (Das Balanced Block Design nutzt per Definition keine technischen Replikate, deshalb gibt es auch keinen Parameter m dafür in der Gleichung.)

Die Varianz in den log-Verhältnissen in einem Balanced Block Design wird generell größer sein, als in einem Reference Design, da keine Referenz verfügbar ist.

4.3 || Simple paired Design und dye-swap paired Design

Bei einem Paired Design wird eine natürliche Paarung der Samples untersucht. Dies sind z.B. Samples vom gleichen Individuum vor und nach einer Behandlung, oder Samples von Tumorzellen und gleichzeitig existierendem gesunden Gewebe des gleichen Organs. Damit ist nicht das physische Paaren von cDNA Samples auf dem gleichen Array bezeichnet, das üblich für alle Dual Label Microarray Experimente ist. Es werden zwei Paired Designs behandelt: Zum einen das Simple Paired Design und zum anderen ein Paired Design, das mit Dye Swap arbeitet, also die Versuche noch einmal mit vertauschten Farben wiederholt. Die Anzahl an benötigten Arrays für balanced paired Design ohne dye swap ist:

$$n_{balanced} = \frac{(z_{\alpha/2} + z_{\beta})^2}{\delta^2} (2\sigma_g^2 + \eta_g^2)$$

wobei $2\sigma_g^2 + \eta_g^2$ die Varianz der log-Verhältnisse für die gepaarten Proben ist. Wir haben die Notation von unserem vorherigen τ_g^2 auf η_g^2 für die biologische Komponente geändert, da die biologische Varianz konzeptionell anders ist bei gepaarten Daten. η_g^2 repräsentiert nicht die Variation im Expressionslevel des Gens in der Population, sondern den Effekt, den der Krebs (bzw. ein anderer Unterschied zwischen den Proben) auf die Genexpression in der Population hat. Z.B. könnte Gen g in Krebs Gewebe im Gegensatz zu normalem Gewebe hoch reguliert sein, aber der Betrag der erhöhten Expressivität mag von Individuum zu Individuum schwanken. Wenn jedes Sample doppelt durchlaufen wird, einmal mit jedem Farbstoff (wird oft komplettes Dye Swap genannt), dann ist die Anzahl der benötigten Arrays für dieses Dye Swap Paired Design:

$$n_{dyeswap} = \frac{(z_{\alpha/2} + z_{\beta})^2}{\delta^2} (2\sigma_g^2 + \eta_g^2)$$

Man beachte das

$$1 \leq \frac{n_{dyeswap}}{n_{balanced}} \leq 2$$

wobei $n_{dyeswap}$ und $n_{balanced}$ die jeweilige Anzahl an benötigten Arrays wiedergibt. Der Wert des Verhältnisses ist meiner Meinung nach sehr nahe an 2, da man für das Dye Swap-Verfahren jedes Array 2 mal mit jeweils anderen Farben betrachten muss. Der Grund warum man nicht genau doppelt so viele Arrays benötigt ist lediglich die geringere technische Fehlerrate, aufgrund der herausgefilterten Störung, durch die einzelnen Farbstoffe. Diese Störung ist normalerweise sehr gering. Hinzu kommt, dass man technische Replikate benötigt, um den Versuch ein zweites mal mit vertauschten Farben durchzuführen. Das hebt die technische Fehlerrate wieder an.

5.0 || Versuchsgröße bei der Prognostic Marker Bestimmung

Bis jetzt ging es um reine Klassenvergleiche, bei denen man die unterschiedlichen Expressionslevel unterschiedlicher Gene herausfinden wollte. Ein anderes übliches Ziel ist die Entwicklung sogenannter Prognostic Marker. Dies sind bestimmte Schlüsselgene oder ganze Klassen von Genen, die auf den Ausbruch einer Krankheit hindeuten.

5.1 || Der Effekt von Pooling

Beim Pooling werden die Expressionsprodukte mehrerer Zellen aus verschiedenen Probanden zusammengemischt. Aus diesem Gemisch werden Proben auf die Arrays gegeben, welche zufällig entnommen werden.

Die Formel die bei k biologisch unabhängigen Proben gilt ist:

$$n = 4m \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta} \right]^2 \left(\frac{\tau_g^2}{k} + \frac{2\sigma_g^2}{m} \right)$$

Zur Herleitung der Formel siehe Anhang Punkt 12.3.

Man kann sofort sehen, dass eine hohe Anzahl an Samples die Versuchsgröße senkt.

6.0 || Wahl des Signifikanzlevels

In den bisher vorgestellten Formeln muss man neben den Varianzparametern auch α und β wählen. α ist die Wahrscheinlichkeit ein Gen als „anders exprimiert“ zu klassifizieren, obwohl das nicht der Fall ist (false positiv) und β ist die Wahrscheinlichkeit ein ungewöhnlich exprimiertes Gen nicht zu finden (false negativ). Also $1-\beta$ ist die Wahrscheinlichkeit dafür ein positiv, also

außergewöhnlich exprimiertes Gen zu finden (true positiv). Wenn man z.B. 10.000 biologisch unabhängige Gene hat von denen keins außergewöhnlich exprimiert wird, dann würde ein α von 0.001 bedeuten, dass fälschlicherweise 10 Gene als außergewöhnlich exprimiert erkannt werden.

Wählt man $1-\beta$, also die Entdeckungsrate, mit 95%, so heißt das, dass 95% der anders exprimierten Gene entdeckt werden. Das heißt, wenn wir z.B. 20 anders exprimierte Gene haben, die Wahrscheinlichkeit alle 20 zu finden nur bei ca. 36% liegt.

Eine Möglichkeit die false-positiven Gene gering zu halten ist, α möglichst klein zu wählen. Das hat aber mehrere Nachteile, zum einen könnte die Schranke so streng werden, dass in der Praxis eine sehr große Anzahl an Arrays benötigt würde, was Kosten- und Zeitaufwand enorm in die Höhe treibt. Außerdem fällt bei einer sehr kleinen erwarteten Anzahl die Varianz stärker ins Gewicht.

Statt dessen kontrolliert man eher die false discovery rate, also die Rate der Gene die fälschlicher Weise als anders expriemiert eingestuft wurden. Eine solche Falscheinstufung wird im folgenden als FD (false discovery) bezeichnet. Dem gegenüber stehen die als anders exprimiert eingestuftene Gene, auf die dies auch wirklich zutrifft (true discoveries = TD).

Die false discovery rate FDR wird durch folgende Formel von Benjamini und Hochberg (1995) ausgedrückt:

$$FDR = E \left[\frac{\#FD}{\#FD + \#TD} \right]$$

(FDR = 0 falls $\#FD + \#TD = 0$)

Nehmen wir an, dass π der Anteil an anders exprimierten Genen ist. Weiterhin nehmen wir an, dass wir alle Gene in zwei Klassen aufteilen können. Entweder sind sie (1) normal exprimiert oder (2) anders exprimiert um einen festen Betrag δ .

Die erwartete Anzahl an false discoveries ist also $E[\#FD] = \alpha(1-\pi)G$, wobei G der Anzahl an Genen entspricht. Die erwartete Anzahl an true discoveries ist $E[\#TD] = (1-\beta)\pi G$. Daraus folgt dann:

$$E[FDR] \approx \frac{\alpha(1-\pi)}{\alpha(1-\pi) + (1-\beta)\pi}$$

$$= \left\{ 1 + \left(\frac{1-\beta}{\alpha} \right) * \left(\frac{\pi}{1-\pi} \right) \right\}^{-1}$$

Obwohl diese Abschätzung eine kleine Abweichung hat⁴, wird sie verwendet.

Diese Abschätzung wurde mit einer Monte Carlo Methode getestet und die Ergebnisse waren für die üblicherweise benutzten Parameter sehr gut.

Wie man in Abb.2 sieht, hängt die FDR hauptsächlich von π und α ab. Der Parameter $1-\beta$ hat nur einen sehr geringen Einfluss.

α mit 0.001 zu wählen ist ein sehr einfacher Weg, die erwartete FDR niedrig zu halten. Andere Möglichkeiten kann man z.B. in den Quellen⁵ oder ⁶ nachlesen.

7.0 || Auswahl der Varianzparameter

Wie schon angemerkt, muss man die biologischen Varianzen schätzen. Dazu werden Erfahrungswerte aus vorherigen Experimenten benötigt. Diese sollten dem Experiment für das man die Varianzen benötigt möglichst ähnlich sein, sowohl von den äußeren Bedingungen her, als auch von den Proben. Die konservativste Schätzung hierbei ist es, die schlechteste, also größte Varianz der vorherigen Versuche zu nehmen. Üblicherweise schätzt man die Varianz aber ab, indem man den Mittelwert aller Proben des oberen Viertels verwendet, oder der Spanne in denen 90% der Proben lagen. Bei Reference Designs mit cDNA Daten ist ein Wert von 0,5 bei humanen Proben und 0,15 bei Mäuse Proben üblich.

8.0 || Auswirkung von kleinen n

Alle hier vorgestellten Formeln gelten nicht für kleine Versuchsgrößen. Dies ist ein wichtiger Punkt, da viele Experimente aus zeit- und kostentechnischen Gründen in kleinem Rahmen ablaufen.

Dies hat verschiedene Gründe, z.B. wird davon ausgegangen, dass die Varianz bekannt ist, aber gerade in kleinen Versuchen, hat eine einzelne Abweichung der Varianz vom erwarteten Wert eine starke Auswirkung.

Dieses Problem ist schon lange bekannt und es gibt eine Reihe von Software die diese statistischen Fehler korrigiert.

⁴ Billingsley, 1995, S.80

⁵ Efron et al. (2001)

⁶ Reiner et al (2003)

π	α	$1-\beta$	$\hat{E}[\text{FDR}]$	#FD in 10000 Genen	# wirklich anders exprimierter Gene	Erwartete Anzahl an nicht entdeckten anders exprimierten Genen
0,005	0,001	0,95	0,17	8,5	50	2,5
0,005	0,001	0,9	0,18	9	50	5
0,005	0,001	0,8	0,2	10	50	10
0,05	0,001	0,95	0,02	10	500	25
0,05	0,001	0,9	0,02	10	500	50
0,05	0,001	0,8	0,02	10	500	100
0,2	0,001	0,95	0,004	8	2000	100
0,2	0,001	0,9	0,004	8	2000	200
0,2	0,001	0,8	0,004	10	2000	400
0,005	0,01	0,95	0,68	34	50	2,5
0,005	0,01	0,9	0,69	35	50	5
0,005	0,01	0,8	0,71	36	50	10
0,05	0,01	0,95	0,17	85	500	25
0,05	0,01	0,9	0,17	170	500	50
0,05	0,01	0,8	0,19	340	500	100
0,2	0,01	0,95	0,04	20	2000	100
0,2	0,01	0,9	0,04	40	2000	200
0,2	0,01	0,8	0,05	100	2000	400
0,005	0,005	0,95	0,51	25,5	50	2,5
0,005	0,005	0,9	0,53	26,5	50	5
0,005	0,005	0,8	0,55	27,5	50	10
0,05	0,005	0,95	0,09	45	500	25
0,05	0,005	0,9	0,1	50	500	50
0,05	0,005	0,8	0,11	55	500	100
0,2	0,005	0,95	0,02	40	2000	100
0,2	0,005	0,9	0,02	40	2000	200
0,2	0,005	0,8	0,02	40	2000	400

Abb.2:

9.0 || Ergebnisse der Formeln

Wenn man ein lineares Modell zu Grunde legt, arbeiten die Formeln sehr zuverlässig. Es wurden 7399 Gene verwendet und jeweils 1000 mal getestet.

Wie wir in Abbildung 3 sehen, ist sowohl das beobachtete α Level, als auch die beobachtete Power jeweils etwas größer als der errechnete Wert. Allerdings nähert sich die Power bei größer werdenden Versuchsgrößen dem erwarteten Wert an. Die Abschätzungen für α und β liegen also immer innerhalb der wahren, experimentell bestimmten Werte.

10.0 || Training von Classifiern

Ein Classifier soll Gene als Prognostic Marker erkennen. Um dies zu gewährleisten, muss er zuerst trainiert werden. Wie immer bei Microarrayexperimenten, soll das mit möglichst wenig Versuchsaufwand gelingen. Hierfür wird nun ein sequentieller, also iterativer Ansatz vorgestellt. Dieser bietet viele Vorteile gegenüber einer statischen Berechnung der Versuchsgröße. Zum einen wird das Stopp-Kriterium in jeder Iteration überprüft. Sobald es erreicht ist, kann das Training beendet werden. So werden auf keinen Fall mehr Versuche als nötig durchgeführt. Gleichzeitig ist durch diese Methode garantiert, dass das erwünschte Ziel erreicht wird,

Abb.3:

Mean Shift	Fold change	α -Level	Observed Level	1- β Power	Observed Power	Mean sample size
0,5849625	1,5	0,001	0,0009	0,9	0,903	133,3
0,5849625	1,5	0,052	0,052	0,9	0,911	68,9
1	2	0,001	0,001	0,9	0,911	50,6
1	2	0,05	0,06	0,9	0,928	25,7
2	4	0,001	0,001	0,9	0,921	16,8
2	4	0,05	0,075	0,9	0,953	8,8

wenn nicht eine vorher definierte Grenze maximaler Versuche erreicht ist. Durch das sequentielle Verfahren „lernt“ der Algorithmus aus eigenen Erfahrungen der Vorrunde

Die verwendete Formel sieht wie folgt aus:

$$N \geq \min \left[\left(\frac{z_{1-\alpha} * \hat{k}_N}{\varepsilon - N^{-1} \sum_{i=1}^N Q_i} \right)^2, N_0 \right]$$

wobei gilt $\hat{k}_N > 0$ und $0 < N^{-1} \sum_{i=1}^N Q_i < \varepsilon$.

ε ist die zu unterschreitende Wahrscheinlichkeit dafür, dass der Classifier eine Fehleinschätzung trifft. Sie wird gegeben. $Z_{1-\alpha}$ ist hierbei die Wahrscheinlichkeit, dass der Classifier unter ε sinkt.

$N^{-1} \sum_{i=1}^N Q_i$ ist der Durchschnittswert, der bisherigen Ergebnisse. Q_i ist der jeweilige Lernalgorithmus wie z.B. K-nearest neighbour, der trainiert werden soll. \hat{k}_N ist der Schätzer für den Durchschnitt der Varianz der Richtigklassifizierung. N_0 dient als Zähler für die korrekten Einschätzungen die hintereinander getroffen wurden.

Der Algorithmus funktioniert folgendermaßen:

1. Starte bei $N_0=0$
2. Trainiere Classifier aufgrund der vorangegangenen Daten.
3. Ziehe neue Probe und klassifiziere diese.
4. Setze $Q_i=0$ und $N_0=N_0+1$ falls Klassifikation korrekt.
5. Setze $Q_i=0$ und $N_0=0$ falls Klassifikation falsch
6. Berechne \hat{k}_N
7. Beende Algorithmus, falls Stopp-Kriterium erreicht, oder $N=M$. Ansonsten springe zu Punkt 2.

In unserem Fall ist die Abbruchsbedingung, das Unterschreiten von ε ohne vorher eine Falschklassifizierung vorgenommen zu haben.

Eine mögliche Verbesserung wäre es, wenn nur eine bestimmte Anzahl von Klassifizierungen in Reihenfolge richtig sein müsste.

11.0 || Schlussfolgerungen und Ausblick

Die Methoden zur Berechnung von Versuchsgrößen für verschiedene Zwecke sind bereits sehr gut. Bei den hier vorgestellten Formeln ist aber eine hohe Anzahl an Versuchen nötig, um ein zuverlässiges Ergebnis zu erhalten.

Es ist ebenfalls nötig, Ergebnisse aus vorherigen, vergleichbaren Versuchen zu besitzen.

Diese gibt es heute aber für viele Probentypen. Ist der Austausch dieser Daten gut, so ist diese Einschränkung nur relativ gering.

12.0 || Anhang

12.1 || Versuchsgrößenberechnung für Single Channel Arrays.

$$Y_{g1fs} = z_{fs}, Y_{g2fs} = w_{fs}$$

$$z_{ij} = \mu_x + x_i + \varepsilon_{ij}; i = 1..n; j = 1..m$$

$$x_i \sim Normal(0, \tau_x^2)$$

$$\varepsilon_{ij} \sim Normal(0, \sigma^2)$$

$$w_{kl} = \mu_y + y_k + \varepsilon_{kl}; k = 1..n; l = 1..m$$

$$y_k \sim Normal(0, \tau_x^2)$$

$$\varepsilon_{kl} \sim Normal(0, \sigma^2)$$

Y_{g1fs} bezeichnet die Leuchtintensität. g ist dabei der Index des Gens, 1 bzw. 2 gibt den Phänotyp an. f steht für die unterschiedlichen Individuen und s für die Unterproben eines Individuums.

z_{ij} und w_{kl} repräsentieren die normalisierten, hintergrund-korrigierten log-Intensitäten.

Der Einfachheit halber werden hier alle Beweise nur für z_{ij} gezeigt. Die Beweise für w_{kl} sind identisch.

μ_x Ist der über die x_i gemittelte Erwartungswert. x_i ist der Erwartungswert für das Gen i mit einer Varianz von $Normal(0, \tau_x^2)$ und ε_{ij} ist der Erwartungswert für die jeweiligen Unterproben mit der Varianz $Normal(0, \sigma^2)$.

Wir wollen nun wissen ob $\mu_x = \mu_y$ ist. Der log Likelihood ist:

$$\begin{aligned}
 & -\frac{n}{2} \log \left(\left\| \sum_x \right\|^{-m/2} \right) \\
 & -\frac{1}{2} \sum_i (Z_i - \mu_x J_{m,1})^T \sum_x^{-1} (Z_i - \mu_x J_{m,1}) \\
 & -\frac{n}{2} \log \left(\left\| \sum_y \right\|^{-m/2} \right) \\
 & -\frac{1}{2} \sum_i (W_i - \mu_y J_{m,1})^T \sum_y^{-1} (W_i - \mu_y J_{m,1}) \\
 & \sum_x \text{ bzw. } \sum_y \text{ sind Schätzer für } x \text{ bzw. } y.
 \end{aligned}$$

$J_{m,1}$ ist ein mit Einsen gefüllter Vektor der Länge m . Z_i ist der Vektor $(z_{i1}, \dots, z_{im})^T$. \sum_x ist die korrespondierende symmetrische Covarianzmatrix zu den Z 's

Wenn man annimmt, dass die Varianzparameter bekannt sind, dann sind auch die Covarianzmatrizen bekannt und es kann ihre Symmetrie gezeigt werden. Dann gilt auch die

Transformation $Z_i^* = \sum_x^{-1/2} Z_i$, also die inverse Wurzel aus Z_i^* zu bilden aus der die Covarianzmatrix $\text{cov}(Z_i^*) = I$ resultiert, wobei I die Einheitsmatrix repräsentiert.

Durch algebraische Umformungen erhält man:

$$I = E \left[\sum_x^{-1/2} Z_i \right] = J_{m,1} \frac{\mu_x}{\sqrt{\sigma^2 + m\tau_x^2}}. \text{ Nun ist}$$

also $\tau_x = \tau_y = \tau$. Mit ziehen von Stichproben folgt nun:

$$\begin{aligned}
 \bar{Z}_i^* & \sim \text{Normal} \left(\frac{\mu_x}{\sqrt{\sigma^2 + m\tau_x^2}}, \frac{1}{m} \right) \\
 \sqrt{\sigma^2 + m\tau_x^2} \bar{Z}_i^* & \sim \text{Normal} \left(\mu_x, \tau^2 + \frac{\sigma^2}{m} \right)
 \end{aligned}$$

Wenn die Null-Hypothese $H_0: \mu_x = \mu_y$ lautet, ergibt sich die Formel unter Einbeziehung von $1-\beta$ und α bei Distanz δ :

$$n = 4 \left[\frac{z_{\alpha/2} + z_\beta}{\delta} \right]^2 \left(\tau_g^2 + \frac{\sigma_g^2}{m} \right).$$

$4 \left[\frac{z_{\alpha/2} + z_\beta}{\delta} \right]^2$ ergibt sich aus der Formel zur Berechnung der Standartabweichung

$$\frac{z_{\alpha/2} + z_\beta}{\sqrt{\text{Versuchsgröße}}} = \text{Standartabweichung,}$$

die 4 ist gleich 2^2 da bei den Versuchen im Reference Design immer auch Referenzwerte in den Versuchen gemacht werden.

12.2 || Bijektive Funktion zur Überführung von Single- und Dual-Label-Designs

Es soll gezeigt werden, dass man für Dual Label Designs von den selben Voraussetzungen wie bei Single Label Designs ausgehen kann. Ist dies gezeigt, kann man auch die Formel für Single Label Designs auf Double Label Arrays anwenden.

Die Funktionen für die beiden Designs waren:

-Single Label:

$$\log(Y_{gvf}^{\{1\}}) = G_g^{\{1\}} + GV_{gv}^{\{1\}} + (GF)_{gf(v)}^{\{1\}} + \varepsilon_{gvf}^{\{1\}}$$

-Dual Label:

$$\begin{aligned}
 \log(Y_{gadvf}^{\{2\}}) &= G_g^{\{2\}} + GA_{ga}^{\{2\}} + GD_{gd}^{\{2\}} \\
 &+ GV_{gv}^{\{2\}} + (GF)_{gf(v)}^{\{2\}} + \varepsilon_{gadvf}^{\{2\}}
 \end{aligned}$$

Die einzigen Unterschiede sind der GA_{ga} - und der GD_{gd} - Term. GA ist unerheblich für die Versuchsgrößenberechnung, da nur die jeweilige Seite des Arrays bezeichnet wird. GD ist der Effekt der Farbe, der bei Single-Label Arrays im G_g -Effekt enthalten ist, da es nur eine Farbe gibt und auf diesen abgebildet werden kann.

12.3 || Versuchsgröße bei Pooling

Der Ausdruck für Dual-Label Designs war:

$$\begin{aligned}
 \log(Y_{gadvf}^{\{2\}}) &= G_g^{\{2\}} + GA_{ga}^{\{2\}} + GD_{gd}^{\{2\}} \\
 &+ GV_{gv}^{\{2\}} + (GF)_{gf(v)}^{\{2\}} + \varepsilon_{gadvf}^{\{2\}}
 \end{aligned}$$

Wenn wir nun davon ausgehen, dass wir k Proben mischen, so müssen wir die Formel nur in einem Punkt abändern zu:

$$\begin{aligned}
 \log(Y_{gadvf}^{\{2\}}) &= G_g^{\{2\}} + GA_{ga}^{\{2\}} + GD_{gd}^{\{2\}} \\
 &+ GV_{gv}^{\{2\}} + \frac{1}{k} \sum_{f=1}^k (GF)_{gf(v)}^{\{2\}} + \varepsilon_{gadvf}^{\{2\}}
 \end{aligned}$$

Es wird also ein Schnitt über die Proben gebildet.

$$\text{var} \left[\frac{1}{k} \sum_{f=1}^k (GF)_{gf(v)}^{\{2\}} \right] = \frac{\tau^2}{k}$$

12.4 || Überblick über Formeln zur Versuchsgrößenberechnung

$$n = 4m \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta} \right]^2 \left(\tau_g^2 + \frac{\sigma_g^2}{m} \right)$$

Simple Reference Design:

$$n = 4m \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta} \right]^2 \left(\tau_g^2 + \frac{2\sigma_g^2}{m} \right)$$

Balanced Block Design:

$$n = \frac{(z_{\alpha/2} + z_{\beta})^2}{\delta^2} (\tau_{g,1}^2 + \tau_{g,2}^2 + 2\sigma_g^2)$$

Single paired Design und dye-swap paired Design:

$$n = \frac{(z_{\alpha/2} + z_{\beta})^2}{\delta^2} (2\sigma_g^2 + \eta_g^2)$$

Pooling:

$$n = 4m \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta} \right]^2 \left(\frac{\tau_g^2}{k} + \frac{2\sigma_g^2}{m} \right)$$

13.0 || Literaturverzeichnis

- *Billingsley*, (1995), Probability and Measure, 3rd Edition. New York: Wiley
- *Kevin Dobbin & Richard Simon (2005)* : Sample size determination in microarray experiments for class comparison and prognostic classification, Biostatistics 2005, Seiten 27-38
Supplementary Material for: Sample Size Determination in Microarray Experiments for Class Comparison and Prognostic Classification
- *Dobin and Simon (2002)*. Comparison of microarray designs for class comparison and class discovery. Bioinformatics 18, 1438-1445
- *Efron et al. (2001)* Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association 96, 1151-1160
- *Kendzioriski et al. (2003)*. The efficiency of pooling in mRNA microarray experiments. Biostatistics 4, 465-477
- *Reiner et al (2003)*, Identifying differential expressed genes using false discovery rate controlling procedures, Bioinformatics 19, 368-375
- *Weijang J. Fu et al. (2004)* : How many samples are needed to build a classifier: a general sequential approach, Bioinformatics vol 21 no1 2005, Oxford Press, Seiten 63-70.