ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2006
Seria: AUTOMATYKA z. 145 Nr kol. 1728

Maria ŁUSZCZKIEWICZ, Andrzej POLAŃSKI
Politechnika Śląska

OCENA SKOKOWYCH ZMIAN PROCESU LOSOWEGO Z WYKORZYSTANIEM ALGORYTMU EXPECTATION-MAXIMIZATION (EM) NA PRZYKŁADZIE ZMIENNYCH W CZASIE PROFILI EKSPRESJI GENÓW BAKTERII DEINOCOCCUS RADIODURANS

Streszczenie. Bardzo często w naukach biomedycznych przeprowadzane eksperymenty polegają na obserwacji zjawisk przy jednoczesnym mierzeniu odpowiedzi systemu w pewnym horyzoncie czasowym. Istnieje tylko kilka metod analizy zmiennych w czasie profili genów. W artykule zaproponowano metodę grupowania genów w poszczególnych chwilach czasu, których poziom ekspresji może być indukowany tym samym/tymi samymi sygnałami, co w efekcie powoduje, iż geny te wykazują podobny poziom ekspresji. W celu budowy modelu wykorzystano złożenia rozkładów normalnych Gaussa (Gaussian Mixture Models) wraz z algorytmem Expectation-Maximization (Dempster, Liard, Rubin 1977, Bilmes 1998). W przeciwieństwie do wielu innych metod, których wyniki silnie zależą od zastosowanych miar i parametrów, zaproponowany model nie posiada żadnych predefiniowalnych parametrów, tak wiec wyniki można uznać za wiarygodniejsze. Z uwagi na swoją konstrukcję może on zostać łatwo zaadaptowany do badania danych o innej strukturze.

PARAMETER ESTIMATION OF RANDOM PROCESS CHANGES USING EXPECTATION-MAXIMIZATION ALGORITHM FOR IDENTIFYING GROUPS OF GENES OF DEINOCOCCUS RADIODURANS IN TIME-COURSE MICROARRAY EXPERIMENTS

Summary. Observation of a biological phenomenon over a certain period of time simultaneously measuring object responses is a common practice in biomedical research. On the other hand there are only several methods of time-course data analysis. We propose method of grouping genes which expression's levels were measured in several time points, assuming that grouped gene's expression levels are induced by the same factor/factors. Due to built a model, we used Gaussian Mixture Model combined with

Expectation-Maximization algorithm. In opposition to many other methods which results are strongly correlated to choice of parameters and methods, proposed approach's results are more stable due to lack of predefined parameters. Moreover, our model can be easily adapted to different experiment structure (several patients with same disease but treated with different drugs).

1. Wprowadzenie

Bardzo często w naukach biomedycznych przeprowadzane eksperymenty polegają na obserwacji zjawisk przy jednoczesnym mierzeniu odpowiedzi systemu w pewnym horyzoncie czasowym. Takie badania sa uważane za kluczowe w zrozumieniu wielu procesów biologicznych i interakcji między nimi. Rozwój technologii mikromacierzy DNA w ostatnich latach pozwala na jednoczesne monitorowanie poziomów ekspresji tysięcy transkryptów. Eksperymenty uwzględniające obserwacje prowadzone na przestrzeni pewnego okresu czasu są bardzo wartościowymi i wnoszącymi wiele w dotychczasową wiedzę o procesach zachodzących w żywych organizmach i środowisku je otaczającym oraz ich wzajemnych interakcjach. W badaniach takich można jednocześnie monitorować wyodrębnione markery w wielu próbkach w pewnym horyzoncie czasowym, co w efekcie pozwala uzyskać pełniejszy obraz zmian zachodzących w komórce, równocześnie uzyskując informacje o zależnościach pomiędzy poszczególnymi genami.

Głównym celem takich badań jest wyodrębnienie genów, które charakteryzują się odmiennymi profilami ekspresji, gdy reprezentują różne warunki (np. różne rodzaje raka lub przebieg terapii na przestrzeni czasu). Niestety, podstawowym problemem w badaniu takich zjawisk jest brak odpowiednich narzędzi do analizy otrzymanych wyników. W przeciwieństwie do bogatej literatury dotyczącej analizy ekspresji w poszczególnych chwilach czasu (bez nawiązania do poprzednich i przyszłych), istnieje niewiele metod służących do analizy przebiegów czasowych. Wczesne badania profili zmiennych w czasie wykorzystywały regresję liniową i analizę wariancji ANOVA (Guo i in. 2003, Xu, Olson i Zhao, 2002). Yuan i Kendziorski (2006) zastosowali Ukryte Modele Markowa (HMM), by wyodrębnić geny różnicujące w każdej chwili czasu.

W wielu przypadkach oprócz analizy stanu w zdefiniowanych momentach, badaczy interesuje to, co dzieje się w komórce pomiędzy chwilami, w których są wykonywane pomiary. Hong i Li (2004) zaproponowali metodę modelowania profili wykorzystującą model liniowy kombinacji kilku krzywych (B-spline functions). Stosowano też modelowanie za pomocą złożeń rozkładów normalnych (np: Ouyang 2004).

W nieniejszej pracy przedstawiamy metodę obliczeniową, pozwalającą na analizę przebiegów czasowych ekspresji genów, przy założeniu że poziomy ekspresji mogą przyjmować wartości z pewnego dyskretnego zbioru, natomiast ich

pomiary są obarczone szumem gaussowskim. Podejście to jest użyteczne dla automatycznego grupowania genów o podobnych właściwościach biologicznych bez uprzedniego definiowania parametrów, które silnie wpływają na uzyskiwane wyniki, a tym samym na wiarygodność i ich biologiczną interpretacje. Zaproponowany algorytm obliczeniowy wykorzystuje odpowiednie złożenie normalnych rozkładów gęstości Gaussa, a iteracyjna estymacja parametrów modelu następuje przy wykorzystaniu algorytmu Expectation-Maximization (Dempster, Liard, Rubin (1977), Bilmes (1998)).

W kolejnych rozdziałach pokazano szczegółowe wyprowadzenie metody oraz wskazano pierwsze uzyskane za jej pomocą wyniki w eksperymentach przeprowadzonych na bakterii Deinoccocus radiodurans (Liu 2003).

2. Założenia metody

Badany eksperyment przedstawia biologiczny proces ciągły,obserwowany w pewnym horyzoncie czasowym, którego dyskretne stany – będące realizacją zmiennej losowej $\mathcal{X}$ – są obserwowalne i mierzalne w chwilach czasu t=1,...,T. Zbiór obserwacji (profil ekspresji genów) jest NxT - elementowy. Zakłada się, że wektory danych są niezależne. W praktyce, $\mathcal{X}_i$ zawiera zmienne odpowiadające t pomiarom poziomu ekspresji i-tego genu w kolejnych chwilach czasu. W dalszych rozważaniach realizacje wektora $\mathcal{X}_t$ (zawierającego poziomy ekspresji wszystkich badanych genów w pewnej chwili czasu t) będą zapisywane jako

$$\chi_t = \begin{pmatrix} x_{1,t} \\ \vdots \\ x_{N,t} \end{pmatrix}.$$

Zakłada sie, że w każdej chwili t eksperymentu występowały te same czynniki stymulujące, lecz o różnym, zmiennym w czasie, natężeniu (ekspresje grupy genów, które potencjalnie "odpowiedziały", są przedstawione jako rozkłady normalne Gaussa) o stałych w czasie parametrach μ oraz σ, lecz ich udziały α_t (interpretowane jako udziały poszczególnych procesów w ogólnym profilu ekspresji badanych genów) są zmienne i zależne od odpowiedniego zestawu danych χ_t.

Celem obliczeń jest wyprowadzenie ocen parametrów modelu opisującego skokową zmianę stanu badanego procesu.

Funkcja największej wiarygodności dla poszczególnych chwil czasu t jest definiowana następująco

$$p(\chi_t|\Theta) = p(x_{1,t}, x_{2,t}, ..., x_{N,t}|\Theta) = \prod_{i=1}^{N} p(x_{i,t}|\Theta) = \mathcal{L}_t(\Theta|\chi_t); \qquad (1)$$

gdzie: $x_{i,t}$ oznacza i-ty pomiar w chwili t, a $p(x_{i,t}|\Theta)$ to funkcje gęstości prawdopodobieństwa. Funkcja $p(\chi_t|\Theta)$ jest funkcją łączną gęstości prawdopodobieństwa dla pomiarów $x_{1,t}, x_{2,t}, ..., x_{N,t}$.

Funkcję największej wiarygodności łączną dla wszystkich chwil czasu można przedstawić jako

$$\mathcal{L}(\Theta|\chi) = \prod_{t=1}^{T} \prod_{i=1}^{N} p(x_{i,t}|\Theta). \tag{2}$$

Celem jest znalezienie zestawu parametrów Θ, które maksymalizują funkcję wiarygodności. Poszukiwany zestaw parametrów może być zdefiniowany jako

$$\Theta_t^* = \text{argmax}_\Theta \mathcal{L}(\Theta|\chi_t). \tag{3}$$

Przyjmuje się istnienie pewnych niezmierzonych obserwacji. Zakłada się, że zdefiniowane i zebrane pomiary χ_t zostały wygenerowane przez funkcje gęstości prawdopodobieństwa o nieznanych parametrach reprezentujące poszczególne czynniki stymulujące. Zatem, χ_t może być traktowane jako zbiór niekompletnych danych. Jeśli przyjmie się, że kompletny zestaw danych dla chwili t może być zdefiniowany jako $\mathcal{Z}_t = (\chi_t, \mathcal{Y}_t)$, wtedy łączna funkcja gęstości dla wektora danych odpowiadającemu chwili t może być zdefiniowana jako

$$p(\mathcal{Z}_t|\Theta) = p(\chi_t, \mathcal{Y}_t|\Theta) = p(\mathcal{Y}_t|\chi_t, \Theta)p(\chi_t|\Theta). \tag{4}$$

Dla powyższego przedstawiania funkcji gęstości można zauważyć, że

$$\mathcal{L}(\Theta|\mathcal{Z}_t) = \mathcal{L}(\Theta|\chi_t, \mathcal{Y}_t) = p(\chi_t, \mathcal{Y}_t|\Theta). \tag{5}$$

Algorytm EM pozwala na wyznaczenie spodziewanych wartości logarytmicznej funkcji największej wiarygodności $\log p(\chi_t, \mathcal{Y}_t|\Theta)$ dla kompletnego zestawu danych w chwili t , co znane jest jako E-step

$$\mathcal{Q}_t(\Theta, \Theta^{(i-1)}) = E[\log p(\chi_t, \mathcal{Y}_t|\Theta)|\chi_t, \Theta^{(i-1)})], \tag{6}$$

gdzie $\Theta^{(i-1)}$ jest aktualnym zestawem estymat parametrów (wyliczonych w poprzedniej iteracji), który jest wykorzystywany w celu aktualizacji parametrów Θ optymalizowanych dla zwiększenia wartości funkcji $\mathcal{Q}_t$. Po wyznaczeniu zestawu parametrów Θ następnym krokiem jest ich maksymalizacja (M step):

$$\Theta^{(i)} = \text{argmax}_\Theta \quad \mathcal{Q}_t(\Theta, \Theta^{(i-1)}). \tag{7}$$

3. Wyznaczanie parametrów modelu

Model probabilistyczny opisujący zjawisko może być zdefiniowany jako

$$p(\chi_t|\Theta) = \sum_{j=1}^{M} \alpha_{j,t} p_{j,t}(\chi_t|\theta_j), \tag{8}$$

gdzie parametry są zdefiniowane jako $\Theta_t = (\alpha_{1,t}, ...\alpha_{M,t}, \theta_1, ..., \theta_M)$, przy założeniu że $\sum_{j=1}^{M} \alpha_{j,t} = 1$.

Co więcej, każda funkcja $p_{j,t}$ jest funkcją gęstości prawdopodobieństwa dla pomiarów w chwili t parametryzowaną przez θ_j. Biorąc to pod uwagę, można przyjąć, że przedstawiony model składa się z M normalnych funkcji gęstości prawdopodobieństwa (reprezentujących poszczególne czynniki stymulujące) w każdej z chwil t, którym są przyporządkowane współczynniki skali $\alpha_{j,t}$. Dla przedstawionego modelu logarytmiczna funkcja największej wiarygodności dla niekompletnego zestawu danych może być przedstawiona jako

$$\log(\mathcal{L}(\Theta|\chi)) = \log \prod_{t=1}^{T} \prod_{i=1}^{N} p(x_{i,t}|\Theta) = \sum_{t=1}^{T} \sum_{i=1}^{N} \log(\sum_{j=1}^{M} \alpha_{j,t} p_{j,t}(x_{i,t}|\theta_j)). \tag{9}$$

Zdefiniowane nieobserwowane pomiary $\mathcal{Y}_t = (y_{i,t})_{i=1}^{N}$ dostarczają informację o tym, która funkcja gęstości wygenerowała w pewnej chwili t poszczególny pomiar. Zakłada się także, że $y_{i,t} \in 1, ..., M$ dla każdej pary (i,t) oraz $y_{i,t} = k$, jeśli i-ta próbka w chwili t została wygenerowana przez k-tą komponentę złożenia. Kompletna (przy znajomości zestawu pomiarów $\mathcal{Y}_t$) logarytmiczna funkcja największej wiarygodności dla pomiarów χ_t może być przedstawiona (korzystając z zależności (1),(4),(5))jako

$$log(\mathcal{L}(\Theta|\chi, \mathcal{Y})) = \sum_{t=1}^{T} \log(p(\chi_t, \mathcal{Y}_t|\Theta) = \sum_{t=1}^{T} \log(p(\mathcal{Y}_t|\chi_t, \Theta)p(\chi_t|\Theta)).$$

Korzystając z reguły Bayesa

$$p(\mathcal{Y}_t|\chi_t, \Theta) = \frac{p(\chi_t|\mathcal{Y}_t, \Theta)p(\mathcal{Y}_t)}{p(\chi_t|\Theta)},$$

otrzymujemy

$$log(\mathcal{L}(\Theta|\chi, \mathcal{Y})) = \sum_{t=1}^{T} \sum_{i=1}^{N} \log(p(x_{i,t}|y_{i,t}, \theta_{(y_i)})p(y_{i,t})) = \sum_{t=1}^{T} \sum_{i=1}^{N} log(\alpha_{(y_i,t)} \cdot p_{(y_i,t)}(x_{i,t}|\theta_{(y_i)}) \tag{10}$$

Przy założeniu że zbiór pomiarów $\mathcal{Y}_t$ w chwili t oraz zbiór parametrów dla tej chwili jest znany (początkowo wybrano wartości losowe) wykorzystując regułę Bayesa, można zapisać

$$p(y_{i,t}|x_{i,t}, \Theta_) = \frac{\alpha_{(y_i,t)} p_{(y_i,t)}(x_{i,t}|\theta_{(y_i)})}{p(x_{i,t}|\Theta)} = \frac{\alpha_{(y_i,t)} p_{(y_i,t)}(x_{i,t}|\theta_{(y_i)})}{\sum_{k=1}^{M} \alpha_{k,t} p_{k,t}(x_{i,t}|\theta_k)} = \frac{\alpha_{j,t} p_{j,t}(x_{i,t}|\theta_j)}{\sum_{k=1}^{M} \alpha_{k,t} p_{k,t}(x_{i,t}|\theta_k)} \tag{11}$$

gdzie α_j oznacza udział składowej j-tej.

W poniższych obliczeniach zakłada się, że znany jest zbiór Y_t - początkowe wartości określające przynależność do poszczególnych składowych zostały wygenerowane losowo.

3.1. Wyznaczenie oceny parametru μ

Wykorzystano zależności (9) oraz (10)

$$\frac{\partial \log \mathcal{L}}{\partial \mu_j} = \sum_{t=1}^{T}\sum_{i=1}^{N} \frac{1}{p(x_{i,t}|\Theta)} \cdot \frac{\partial}{\partial \mu_j} p(x_{i,t}|\Theta) = \sum_{t=1}^{T}\sum_{i=1}^{N} \frac{1}{p(x_{i,t}|\Theta)} \cdot \frac{\partial}{\partial \mu_j} p(x_{i,t}|y_{i,t},\Theta) p(y_{i,t}) =$$

$$\sum_{t=1}^{T}\sum_{i=1}^{N} \frac{p(y_{i,t})}{p(x_{i,t}|\Theta)} \cdot \frac{\partial}{\partial \mu_j} p(x_{i,t}|y_{i,t},\Theta) = \sum_{t=1}^{T}\sum_{i=1}^{N} \frac{p(y_{i,t})}{p(x_{i,t}|\Theta)} \cdot p(x_{i,t}|y_{i,t},\Theta) \cdot \frac{\partial}{\partial \mu_j} \log p(x_{i,t}|y_{i,t},\Theta)$$

Korzystając z reguły Bayesa można zapisać

$$p(y_{i,t}|x_{i,t},\Theta) = \frac{p(x_{i,t}|y_{i,t},\Theta)p(y_{i,t})}{p(x_{i,t}|\Theta)}.$$

Wstawiając do powyższego wzoru, otrzymuje się:

$$\frac{\partial \log \mathcal{L}}{\partial \mu_j} = \sum_{t=1}^{T}\sum_{i=1}^{N} p(y_{i,t}|x_{i,t},\Theta) \cdot \frac{\partial}{\partial \mu_j} \log p(x_{i,t}|y_{i,t},\Theta) =$$

Należy zauważyć, że

$$p(x_{i,t}|y_{i,t},\Theta) = \frac{1}{\sigma_j\sqrt{(2\pi)}} \cdot e^{-\frac{1}{2}\frac{(x_{i,t}-\mu_j)^2}{\sigma_j^2}},$$

stąd wyznaczenie

$$\frac{\partial}{\partial \mu_j} \log[\frac{1}{\sigma_j\sqrt{(2\pi)}} \cdot e^{-\frac{1}{2}\frac{(x_{i,t}-\mu_j)^2}{\sigma_j^2}}] = \frac{\partial}{\partial \mu_j}[\log \frac{1}{\sigma_j\sqrt{2\pi}}] - \frac{\partial}{\partial \mu_j}[\frac{(x_{i,t}-\mu_j)^2}{2\sigma_j^2}] = \frac{1}{\sigma_j^2}(x_{i,t}-\mu_j)$$

prowadzi do

$$\frac{1}{\sigma_j^2}\sum_{t=1}^{T}\sum_{i=1}^{N}(p(y_{i,t}|x_{i,t},\Theta)\cdot(x_{i,t}-\mu_j)) =$$

$$\frac{1}{\sigma_j^2}\sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t}\cdot p(y_{i,t}|x_{i,t},\Theta) - \frac{1}{\sigma_j^2}\sum_{t=1}^{T}\sum_{i=1}^{N}\mu_j \cdot p(y_{i,t}|x_{i,t},\Theta) = 0$$

a stąd ostatecznie

$$\mu_j^{new} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t}\cdot p(y_{i,t}|x_{i,t},\Theta)}{\sum_{t=1}^{T}\sum_{i=1}^{N} p(y_{i,t}|x_{i,t},\Theta)}. \tag{12}$$

3.2. Wyznaczenie oceny parametru σ

Analogicznie do poprzedniego wyprowadzenia otrzymuje się estymatę

$$\sigma_j^{2new} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{N} p(y_{i,t}|x_{i,t},\Theta)(x_{i,t}-\mu_j)^2}{\sum_{t=1}^{T}\sum_{i=1}^{N} p(y_{i,t}|x_{i,t},\Theta)}. \tag{13}$$

3.3. Wyznaczenie oceny parametru α

W celu wyznaczenia ocen udziałów $\alpha_{j,t}$ dla poszczególnych chwil czasu t korzysta się z logarytmicznej funkcji największej wiarygodności dla pomiarów χ_t, a nie kompletnego zbioru danych χ, gdyż w chwilach $t=1...T$ parametry odpowiednich składowych normalnych nie ulegają zmianie:

$$(\mu_{k,t=1} = \mu_{k,t=2} = ... = \mu_{k,t=T}) \text{ oraz } (\sigma_{k,t=1} = \sigma_{k,t=2} = ... = \sigma_{k,t=T}),$$
$$\text{ale } (\alpha_{k,t=1} \neq \alpha_{k,t=2} \neq ... \neq \alpha_{k,t=T}).$$

Wartości poszczególnych udziałów $\alpha_{j,t}$ są różne w kolejnych chwilach czasu - zależą od wektora danych wyłącznie w chwili t.
Analogicznie do poprzednich wywodów (8-11)

$$log(\mathcal{L}_t(\Theta|\chi_t, \mathcal{Y}_t)) = \log(p(\chi_t, \mathcal{Y}_t|\Theta)) = \log(p(\mathcal{Y}_t|\chi_t, \Theta)p(\chi_t|\Theta)) =$$

$$= \sum_{i=1}^{N} \log((p(y_{i,t}|x_{i,t}, \Theta)p(x_{i,t}|\Theta)) = \sum_{i=1}^{N} \log(p(y_{i,t}|x_{i,t}, \Theta) \cdot \alpha_{j,t}p(x_{i,t}|\theta_i)).$$

By wyznaczyć $\alpha_{j,t}$, wprowadza się mnożnik Lagrange'a λ, przy założeniu że $\sum_j^M \alpha_{j,t} = 1(*)$.

$$\frac{\partial}{\partial \alpha_{j,t}}[\sum_{i=1}^{N} \log(\alpha_{j,t} \cdot p(y_{i,t}|x_{i,t}, \Theta)) + \sum_{i=1}^{N} \log(p(x_{i,t}|\theta_i)p(y_{i,t}|x_{i,t}, \Theta)) + \lambda(\sum_{j=1}^{M} \alpha_{j,t} - 1)] = 0.$$

$$\sum_{i=1}^{N} \frac{1}{\alpha_{j,t}} p(y_{i,t}|x_{i,t}, \Theta)) + \lambda = 0$$

Pamiętając o (*) oraz sumując po wszystkich udziałach $\alpha_{j,t}$, otrzymuje się

$$\lambda = -N,$$

stąd

$$\alpha_{j,t}^{new} = \frac{1}{N} \sum_{i=1}^{N} p(y_{i,t}|x_{i,t}, \Theta)).$$

4. Błędy metody

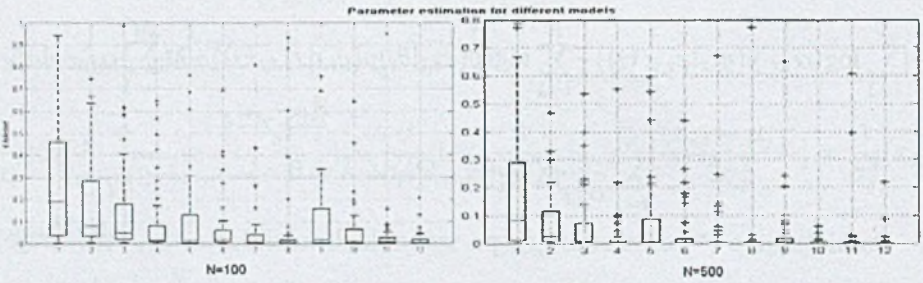
Przeprowadzono analizę błędów estymacji metody (rys. 1). W tym celu wykonano symulacyjne badanie dla sztucznych danych w T=10 chwilach czasu, składających się z I=3 modeli, po N=100 oraz N=500 próbek w każdym modelu. Dodatkowo, zbadano, jak błędy estymacji zależą od struktury modelu. W tym celu przyjęto J=4 wzorce modeli. Każdy z nich (1-4) składa się z K=3 składowych, będących rozkładami normalnymi Gaussa. Modele budowane są wg (8), gdzie dla

każdej chwili t:
– poszczególne prawdopodobieństwa $p_{j,k}$ pochodzą z rozkładów Gaussa $\mathcal{N}_{j,k} \sim (\mu_{j,k}, \sigma_{j,k})$ o właściwościach:
1) Wartość wariancji (odchylenia standardowego)jest niezmienna $\forall_{j,k} \sigma_{j,k} = \sigma$.
2) Wartość średnia dla poszczególnych rozkładów jest definiowana jako $\mu_{(j,1)} = \mu$ oraz $\mu_{(j,k>1)} = U \sim [(\mu_{j,1} - \gamma[j] \cdot \sigma), (\mu_{j,1} + \gamma[j] \cdot \sigma)]$.
3) Współczynnik skali ma własność $\alpha_{(j,k)} = U \sim [0, 1]$ oraz $\sum_{k=1}^{K=3} \alpha_{j,k,t} = 1$;
gdzie: U- oznacza rozkład równomierny, a $\gamma = [1, 2, 3, 5]$ jest parametrem definiującym strukturę modelu. Wraz ze wzrostem wartości parametru γ rośnie rozróżnialność poszczególnych składowych modelu. Dla każdego modelu wygenerowano r=50 realizacji. Oczywiście, ze wzrostem liczby obserwacji znacząco wzrasta dokładność estymacji parametrów. Analizę modelu przedstawiono dla N=100 oraz N=500 obserwacji.

Błąd estymacji poszczególnych parametrów określony jest wskaźnikiem:

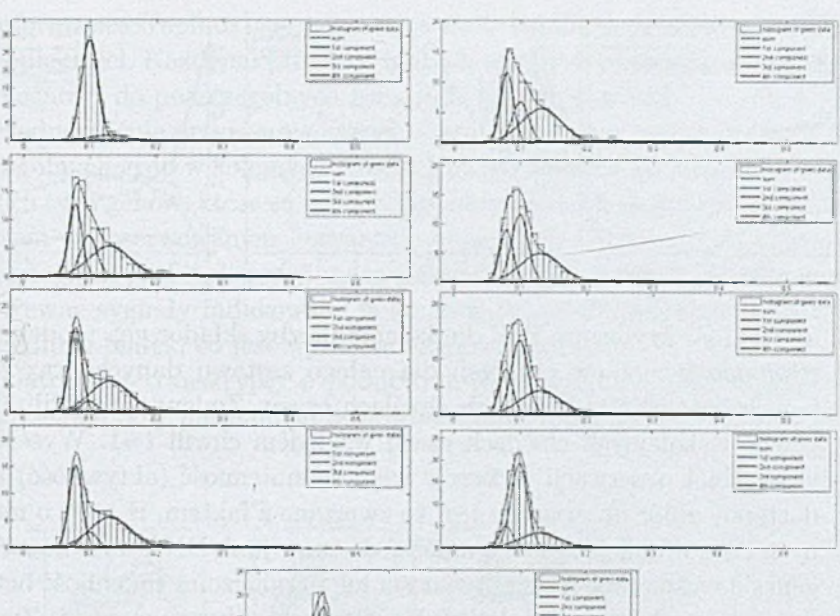
$$I = \sum_{i=1}^{M} \sum_{t=1}^{T} \sum_{l=1}^{L} \frac{1}{M \cdot T \cdot L} \left(\frac{\vartheta_{i,t,l} - \bar{\vartheta}_{i,t,l}}{\vartheta_{i,t,l}}\right)^2, \tag{14}$$

gdzie:
i- oznacza kolejne składowe, t- oznacza kolejne chwile czasu, l- oznacza kolejne parametry.



Rys. 1. Błędy estymacji poszczególnych parametrów. Każdy model (1-4) reprezentowany jest przez 3 tzw. "pudełka" dla 3 estymowanych parametrów: α, μ, σ. Odpowiednio "pudełka" 1,4,7,10 odpowiadają estymacji parametru α, "pudełka" 2,5,8,11 odpowiadają estymacji parametru μ, natomiast "pudełka" 3,6,9,12 odpowiadają estymacji parametru σ

Obserwując, jakim błędem obarczona jest estymacja poszczególnych parametrów, okazuje się, że o ile parametry rozkładów (μ, σ) są dość dokładnie estymowane, to nawet niewielkie odchyłki tych parametrów od wartości zadanych powodują znaczne błędy w wyznaczeniu udziałów α, co spowodowane jest konstrukcją algorytmu, który stara się zapewnić poprzez współczynnik skali najlepsze dopasowanie między danymi pomiarowymi a modelem o wyestymowanych parametrach. Poza tym można zauważyć, że największe błędy estymacji są generowane



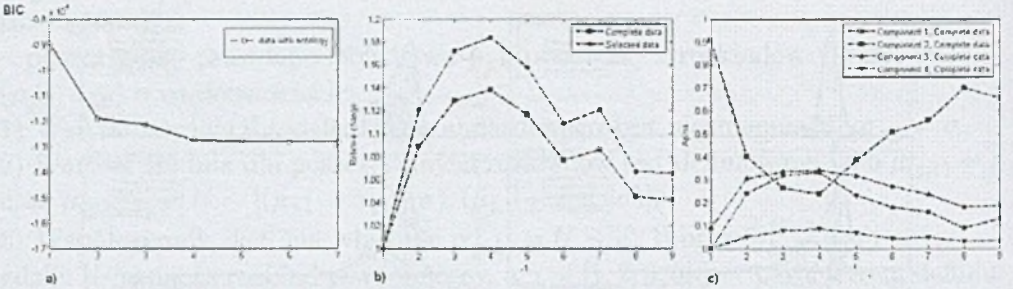
Rys. 2. Stany procesu w kolejnych chwilach czasu oraz rozkład na składowe normalne Gaussa

dla modelu *I*. Ma to swoje uzasadnienie, wynikające z faktu, iż konstrukcja tego modelu zakłada, iż jego składowe są trudno rozróżnialne, stąd istnieje wiele kombinacji parametrów dość wiernie odzwierciedlających histogram obserwacji.

5. Ocena skokowych zmian stanu procesu dla danych rzeczywistych Deinococcus Radiodurans

Przedstawiony model wykorzystano do analizy danych rzeczywistych (Liu 2003). Eksperyment polegał na badaniu poziomu ekspresji 3103 genów w 9 chwilach czasu (t=0h, 0.5h, 1.5h, 3h, 5h, 9h, 12h, 24h) bakterii Deinococcus Radiodurans. Z dalszych analiz wykluczono geny, dla których nie posiadano pełnej informacji o poziomie ekspresji w każdej z chwil czasu t. W kolejnym kroku wykluczono geny nie posiadające ontologii, czyli informacji o pełnionej roli (tab. 1). Na potrzeby poniższych badań wartości poziomów ekspresji zostały przeskalowane do przedziału $0 \div 1$ (rys. 2). Dla danych posiadających ontologię zbadano jaki model najlepiej oddaje ich strukturę. Jako kryterium wyboru ilości składowych modelu wykorzystano kryterium Bayesian Information Critierion (Schwarz 1978)

$$BIC = -2\log(\hat{L}) + k\log(N), \tag{15}$$



Rys. 3. a) Wartość kryterium BIC dla kolejnej liczby składowych modelu. b) Porównanie poziomów ekspresji dla całego zestawu danych oraz posiadających ontologię w kolejnych chwilach czasu. Zmienność profili ekspresji genów w kolejnych chwilach czasu względem chwili t=1. Wyselekcjonowany zbiór obserwacji wykazuje większą zmienność (aktywność) niż cały dostępny zbiór obserwacji. Jest to związane z faktem, iż geny o niepoznanych dotąd funkcjach mogą należeć do tzw. 'junk DNA' lub 'housekeeping genes' i wykazywać brak aktywności lub ograniczoną zmienność bez względu na warunki, w jakich był przeprowadzany eksperyment. c) Zmienność współczynnika skali w czasie dla danych posiadających ontologię dla 4 składowych normalnych o niezmiennych w czasie parametrach

Tabela 1

Parametry dla modelu składającego sie z T=9 chwil czasu oraz M=4 składowych

$\begin{pmatrix} \mu_{j=1}, \sigma_{j=1} \\ \vdots \\ \mu_{j=4}, \sigma_{j=4} \end{pmatrix}$	$\begin{pmatrix} 0.0792, 0.0982 \\ 0.0992, 0.1110 \\ 0.1302, 0.1734 \\ 0.2470, 0.3346 \end{pmatrix}$	$\begin{pmatrix} \alpha_{j=1,t=1}, \ldots, \alpha_{j=4,t=1} \\ \vdots \ddots \vdots \\ \alpha_{j=1,t=T}, \ldots, \alpha_{j=4,t=T} \end{pmatrix}$
		$\begin{pmatrix} 0.0000, 0.9091, 0.0897, 0.0012 \\ 0.2409, 0.4034, 0.3042, 0.0514 \\ 0.3200, 0.2654, 0.3359, 0.0786 \\ 0.3311, 0.2396, 0.3417, 0.0875 \\ 0.2313, 0.3837, 0.3173, 0.0676 \\ 0.1809, 0.5061, 0.2715, 0.0414 \\ 0.1592, 0.5563, 0.2359, 0.0486 \\ 0.0905, 0.6999, 0.1788, 0.0308 \\ 0.1312, 0.6521, 0.1858, 0.0310 \end{pmatrix}$

gdzie:
- $\hat{L}$ to "najlepsza" logarytmiczna funkcja największej wiarygodności,
- N to liczba obserwacji (dla każdej chwili czasu t),
- k to liczba niezależnych parametrów (stopni swobody).

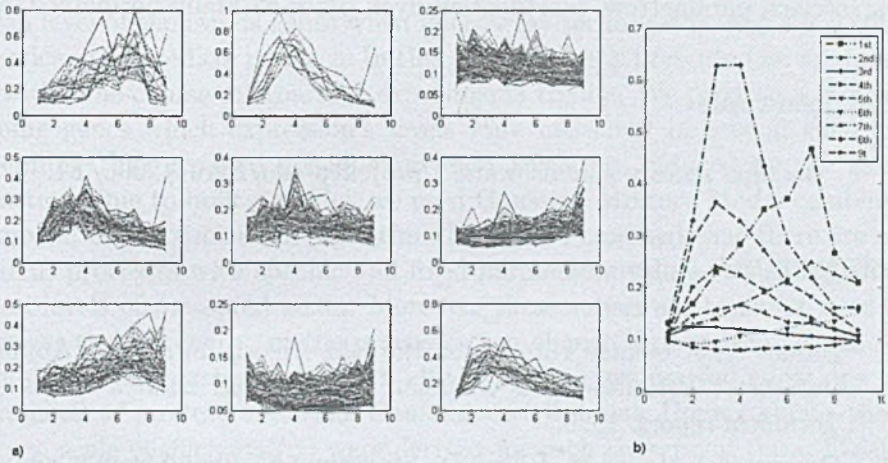
Analizując badany zbiór obserwacji (rys. 3) zauważa się, że przytłaczająca większość badanych genów nie zmieniała komponenty, do której została zakwalifikowana (ewentualne odstępstwa występowały w pierwszej i ostatniej chwili czasu). Największe zainteresowanie budzą te geny, których profile ekspresji wykazują dużą zmienność.

W celu zdefiniowania grup transkryptów o podobnych profilach ekspresji przeprowadzono klasteryzację hierarchiczną metodą "Complete linkage" (Matlab

- Bioinformatics Toolbox) przy zastosowaniu korelacji pomiędzy profilami jako miary odległości. Każdemu profilowi nadano wagę odpowiadającą częstości zmian przynależności do poszczególnych grup w kolejnych chwilach.

Jednocześnie przypisanie każdemu profilowi miary zmian przynależności do poszczególnych grup w kolejnych chwilach czasu pozwala na wyodrębnienie z całego zbioru tych genów, które są najbardziej aktywne i ich ekspresja ulega znacznym wahaniom w obserwowanym horyzoncie czasowym. Uważa sie, że geny o podobnych kształtach profili ekspresji mogą należeć do tych samych szlaków regulatorowych. Pewne sygnały indukowane w komórkach indukują/wstrzymują produkcję odpowiednich białek, co jest wyrazem ekspresji genów.

Natomiast transkrypty o podobnych poziomach ekspresji mogą być sterowane przez wspólny mechanizm odpowiedzialny za produkcję białek w określonej ilości.



Rys. 4. Wynik klasteryzacji hierarchicznej danych. a) Dla wszystkich genów. b) Uzyskane kształty profili ekspresji

6. Wnioski

Zaproponowana metoda analizy danych zmiennych w czasie pozwala na grupowanie w każdej z chwil czasu genów ze względu na ich poziom ekspresji, który może być indukowany takim samym sygnałem. Metoda ta jest niezależna od zadanych predefiniowalnych parametrów. Zakładamy, że wspomniane czynniki występowały o różnym, zmiennym w czasie natężeniu (ekspresje grupy genów, które potencjalnie "odpowiedziały", są przedstawione jako rozkłady normalne Gaussa) o stałych w czasie parametrach μ oraz σ, lecz ich udziały α_t (interpretowane jako udziały poszczególnych procesów w ogólnym profilu ekspresji badanych genów) są zmienne. W rezultacie, typowane są w każdej dyskretnej chwili czasu transkrypty

o zbliżonym poziomie ekspresji.

Przedstawiona metoda pozwala na wyznaczenie transkryptów o największej aktywności, czyli tych, które są przyporządkowane w kolejnych chwilach do różnych grup definiowanych przez rozkłady normalne. Potencjalnie geny te mogą być różnicującymi z uwagi na swoją ekspresję w kolejnych fazach eksperymentu.

Metodę łatwo zmodyfikować w celu wyszukiwania genów różnicujących (np. z uwagi na zastosowaną terapię, różną postać badanej choroby, płeć, wiek itd.). W tym celu każdą obserwację traktujemy nie jako zmierzoną w kolejnej chwili czasu, lecz jako odrębny obiekt (np. reprezentujący jeden ze sposobów leczenia). Zakładamy, na przykład, że dla poszczególnych wariantów choroby ekspresja poszukiwanych genów różni się. Stąd wniosek, iż najbardziej interesujące są te transkrypty, które przynależą do różnych grup o pewnych stałych parametrach dla poszczególnych obserwacji. Jeśli gen potencjalnie nie różnicuje pomiędzy przypadkami, to dla nich wszystkich powinien należeć do tej samej grupy lub do grup o zbliżonych wartościach parametrów, reprezentowanych przez rozkłady normalne Gaussa.

Podziękowanie

Badania zostały sfinansowane z projektu 409/RAu1/2006/T1.

LITERATURA

1. Bilmes J.: A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical report. 1998.
2. Dempster A., Laird N., Rubin D.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, Series B,vol.39, no.1, 1977, p. 1–38.
3. Hong F., Li H.: Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics, vol 19 no. 4, 2003, p. 474–482.
4. Guo W., Dai M., Ombao H.C., von Sachs R.: Smoothing Spline ANOVA for Time-Dependent Spectral Analysis. Journal of the American Statistical Association, 98, 2003, p. 643–652.
5. Liu Y. et al.: Transcriptome dynamics of Deinococcus radiodurans recovering from ionizing radiation. Proc.Natl. Acad Sci. Apr.1;100(7), 2003, p. 4191–6.
6. Ouyang M., Welsh W., Georgopoulos P.: Gaussian mixture clustering and imputation of microarray data. Bioinformatics, 20, 2004, p. 917–923.
7. Schwarz, G.: Estimating the dimension of a model. Ann. Stat.6, 1978, p. 461-–464.
8. Xu X.L., Olson J.M., Zhao L.P.: A regression-based method to identify differentially expressed genes in microarray time course studies and its application

in an inducible Huntington's disease transgenic model Hum. Mol. Genet., August 15, 11(17), 2002, p. 1977–1985.

9. Yuan M., Kendziorski C.: Hidden Markov Models for Microarray Time Course Data in Multiple Biological Conditions. Journal of the American Statistical Association, to appear, 2006.
10. Ouyang M., Welsh W.J., Georgopoulos P.: Gaussian mixture clustering and imputation of microarray data. Bioinformatics, 20, 2004, p. 917–923.

Recenzent: Prof. dr hab. inż. Mariusz Ziółko

Abstract

Observation of a biological phenomenon over a certain period of time (i.e. expression level of thousands genes when microarray technique is used) is a common practice in biomedical research. In this article there is presented an approach to deal with time-course microarrays experiments results. We propose a method of grouping genes which expression's levels were measured in several time points, assuming that grouped gene's expression levels are induced by the same factor/factors. Due to built a model, we used Gaussian Mixture Model combined with Expectation-Maximization algorithm. Firstly we assumed that there are some (hidden) processes with specific and fixed parameters values, which influence expression levels of measured genes. Moreover these values are constant over time. However the "strength" of the processes can change during time. Secondly, taking into account expression levels in all time points, we derived equations for each parameter of mixture of normal Gaussian distributions. Having known these parameters, scale coefficients (α) were derived for each time point data. Finally we applied presented method to Deinococcus Radiodurans data [5]. In opposition to many other methods which results are strongly correlated to choice of specific parameters and methods, proposed approach's results are more stable due to lack of predefined values. Moreover, our model can be easily adapted to different experiment structure (several patients with same disease but treated with different drugs).