Ordenamento Pseudo Cronológico do Ciclo Celular

João M. Dinis, Rita M. C. de Almeida

Instituto de Física - UFRGS

Introdução

O ciclo celular é consequência de uma dinâmica metabólica extensiva, em nível de genoma completo, responsável pela proliferação celular, de tecidos e outras funções essenciais para sobrevivência dos organismos. O ciclo celular pode ser dividido em três grandes fases *G1*, *S* e *G2M*.

Ao longo deste trabalho, usaremos medidas de expressão gênica por *RNASeq* de células únicas *T* de *Mus musculus*, disponibilizados no banco de dados *Array Express* sob o código *E-MTAB-2805*[2]. Estas células foram cultivadas, marcadas e classificadas em *G1*, *S* ou *G2M* por *FACS* (*Fluorescence-activated cell sorting*, método de classificação celular por sinal fluorescente).

Objetivos

1. Aplicar um controle de qualidade para filtrar erros usuais da técnica;
2. Normalização dos dados pelo comprimento total de cada gene para cada amostra;
3. Efetuar um *transcriptograma*[3] destes dados normalizados;
4. Efetuar uma análise de componentes principais (*PCA*);
5. Construção e validação do modelo cronológico para o ciclo celular.

Ordenamento

A partir de informações sobre associação entre proteínas, retiradas do banco de dados *STRING*, construímos uma matriz de adjacência (M_{ij}) onde:

$$M_{ij} = \begin{cases} 1 & \text{se as proteínas } i \text{ e } j \text{ estão associadas} \\ 0 & \text{se as proteínas } i \text{ e } j \textbf{ não} \text{ estão associadas} \end{cases}$$

Em seguida, definimos uma função custo E para a matriz de adjacência:

$$E = \sum_{i=1}^{13314} \sum_{j=i+1}^{13314} \mid i-j \mid [\mid M_{ij} - M_{ij+1} \mid + \mid M_{ij} - M_{ij-1} \mid + \mid M_{ij} - M_{i+1j} \mid + \mid M_{ij} - M_{i-1j} \mid]$$

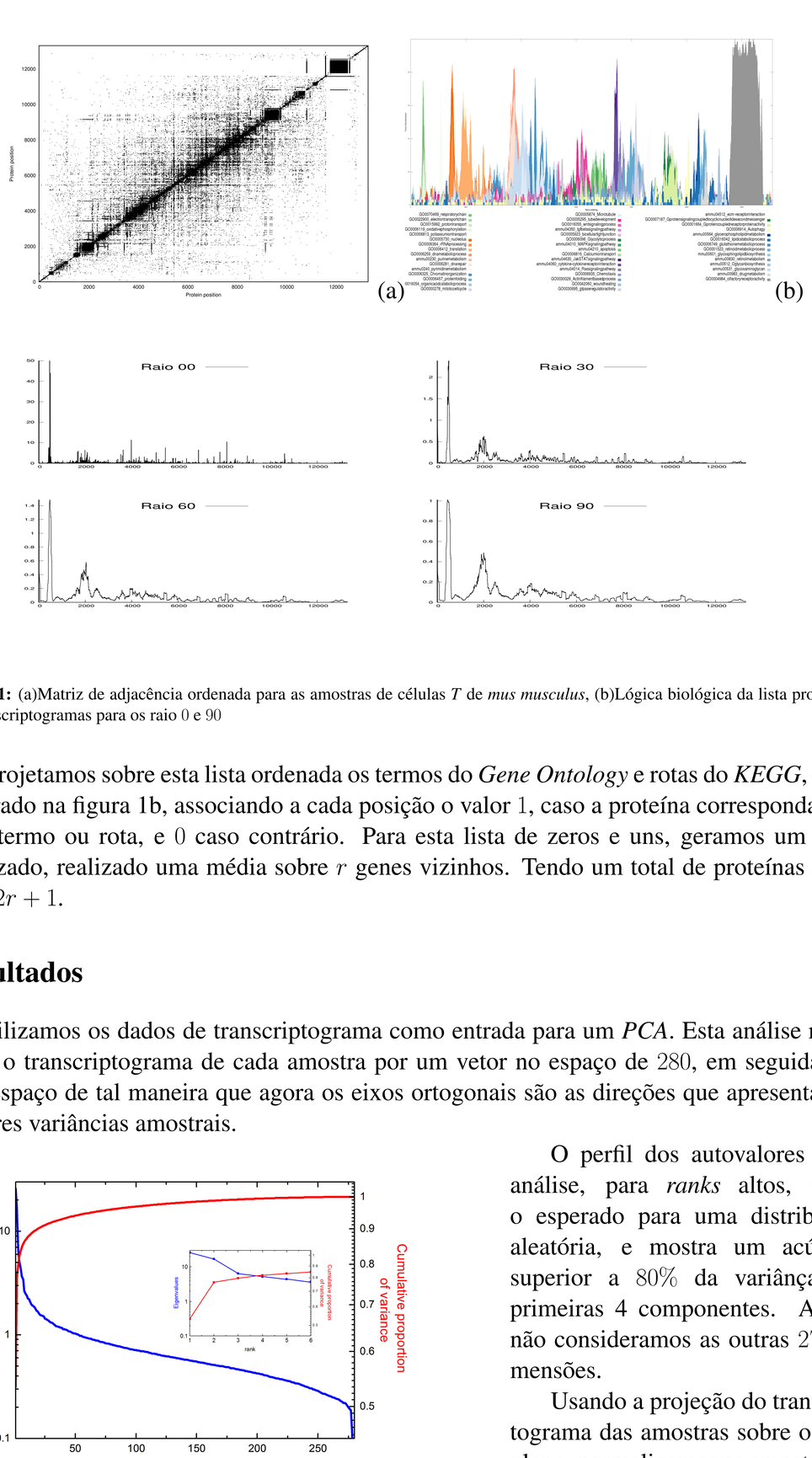

Figure 1: (a)Matriz de adjacência ordenada para as amostras de células *T* de *mus musculus*, (b)Lógica biológica da lista proteínas e (c)Transcriptogramas para os raio 0 e 90

Projetamos sobre esta lista ordenada os termos do *Gene Ontology* e rotas do *KEGG*, como mostrado na figura 1b, associando a cada posição o valor 1, caso a proteína corresponda a algum termo ou rota, e 0 caso contrário. Para esta lista de zeros e uns, geramos um perfil suavizado, realizado uma média sobre r genes vizinhos. Tendo um total de proteínas avaliadas $2r+1$.

Resultados

Utilizamos os dados de transcriptograma como entrada para um *PCA*. Esta análise representa o transcriptograma de cada amostra por um vetor no espaço de 280, em seguida rota este espaço de tal maneira que agora os eixos ortogonais são as direções que apresentam as maiores variâncias amostrais.

Figure 2: Autovalores da análise de *PCA*.

O perfil dos autovalores desta análise, para *ranks* altos, segue o esperado para uma distribuição aleatória, e mostra um acúmulo superior a 80% da variança nas primeiras 4 componentes. Assim, não consideramos as outras 276 dimensões.

Usando a projeção do transcriptograma das amostras sobre o novo plano, normalizamos as amostras tal que:

$$\sum_{i=1}^{4} (c_i^a)^2 = 1$$

Onde c_i^a é o coeficiente da amostra a na direção da i-ésima componente principal. Esta normalização faz com que a maior parte das amostras fiquem dispostas em uma circunferencia de raio 1 sobre o plano $P1$ x $P2$.

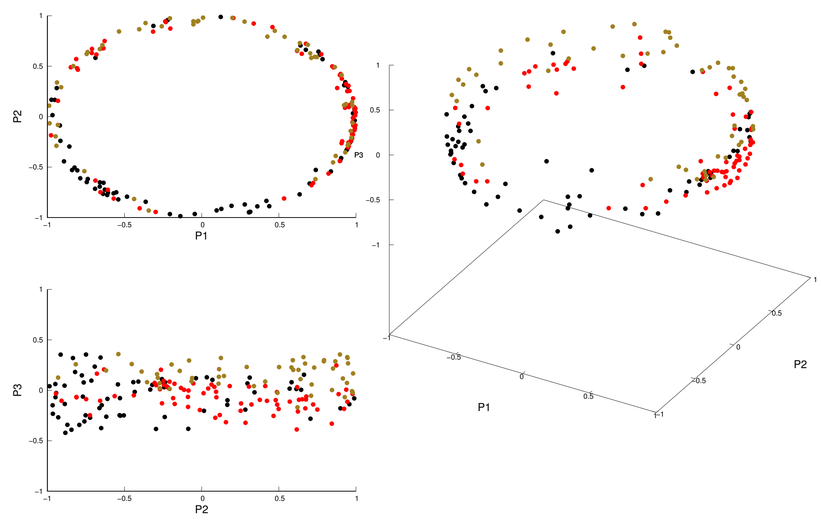

Figure 3: Proejção das amostras normalizadas pelas 4 primeiras componentes com a condição adicional que $((c_1^a)^2 + (c_2^a)^2) \geq 0.81$

Como mostrado na figura 3 , não podemos identificar todas as fases do ciclo em um plano bidimenssional, portanto, uma possível solução seria projetar as amostras sobre o plano $P1$ x $P2$ x $P3$ onde podemos separar as amostras em três diferentes grupos no intervalo $-2\pi \leq \delta \leq 2\pi$:

a) $c_2^a < 0$: Amostras classificadas como *G1*

b) $c_2^a > 0$ e $c_3^a > 0$: Amostras classificadas como *S*

c) $c_2^a > 0$ e $c_3^a < 0$: Amostras classificadas como *G2M*

Com base nessa separação, consideramos que um ordenamento pseudo cronológico para ciclo celular pode ser obtido calculando:

$$\delta^a = \begin{cases} 2\pi - \theta_0 + \tan^{-1}(c_2^a/c_1^a) & \text{se } c_3^a > 0 \\ -2\pi + \theta_0 - \tan^{-1}(c_2^a/c_1^a) & \text{se } c_3^a < 0 \\ \tan^{-1}(c_2^a/c_1^a) - \theta_0 & \text{se } \tan^{-1}(c_2^a/c_1^a) > \theta_0 \\ -\tan^{-1}(c_2^a/c_1^a) + \theta_0 & \text{se } \tan^{-1}(c_2^a/c_1^a) < \theta_0 \end{cases}$$

Sendo θ_0 o ponto onde a componente c_3^a troca de sinal.

Validação do modelo e o comportamentos das ciclinas, CDK's e CDKN's

Ciclinas, *CDK's* e *CDKN's* são enzimas conhecidas por regular processos do ciclo celular. A validação deste modelo pode ser feita através dos seus comportamentos:

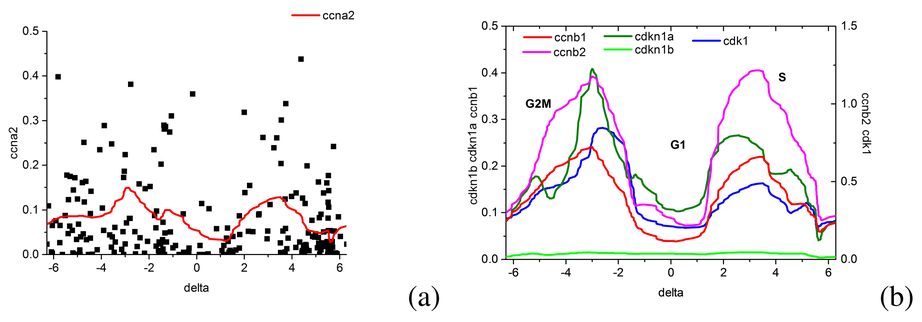

Figure 4: (a)Expressão da ccan2 ao longo da trajetória, a linha vermelha é resultado da aplicação de um filtro *FFT* (*fast Fourier transform)*, considerando um grupo de 12 pontos. (b)Comportamento das expressões de ccns e cdkn's ao longo da trajetória δ.

Conclusões e Perspectivas

Tendo este modelo como validado, efetuaremos uma inferência estatística para, então, construir um ciclo ideal, ou seja, construir um modelo suavizado onde podemos mostrar como as expressões dos genes devem normalmente se comportar.

References

[1] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. HTSeq – a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2014.

[2] Florian Buettner, Kedar N. Natarajan, F. Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J. Theis, Sarah A. Teichmann, John C. Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *nature biotechnology*, 33(2):155–160, 2015.

[3] Samoel R.M. da Silva, Gabriel C. Perrone, João M. Dinis, and Rita M.C. de Almeida. Reproducibility enhancement and differential expression of non predefined functional gene sets in human genome. *BMC Genomics*, 15(1181), 2014.