Updated CIBM research plan - Nathanael Fillmore

In this document, I describe two projects that I plan to work on, one with Prof. Colin Dewey and one with Prof. Michael Newton. For each project, I describe **(a)** the biological problem, **(b)** motivation for a new approach to the problem, and **(c)** my proposed contribution.

1 Transcriptome assembly from RNA-seq data

With Colin Dewey, I plan to work on de novo transcriptome assembly from RNA-seq data. Prof. Dewey's biological collaborator on the project is Prof. James Thompson.

(a) The biological problem is as follows. A transcriptome is the collection of all RNA molecules produced in a cell or in a collection of cells. One can think of a transcriptome as (i) a collection of transcripts together with (ii) the abundance, or level of expression, of each transcript within the cell. RNA-seq is a new protocol that uses high-throughput sequencing technology in order to measure gene expression [13]. In RNA-seq, (i) RNA is isolated from a sample and (ii) converted to cDNA fragments; then (iii) a high-throughput sequencer is used to generate millions of short reads from the cDNA fragments [5]. Finally, (iv) since each read corresponds to only a small fragment of a transcript, a computer program is used to infer the transcriptome from the collection of reads. I plan to work on a new approach to step (iv).

(b) Existing approaches to transcriptome assembly from RNA-seq data have important shortcomings. (i) Reference-based approaches to transcriptome assembly (e.g., Scripture [2] and Cufflinks [12]) have been highly successful - but reference-based approaches require a reference genome, and we are interested in organisms for which it is not yet practical to sequence the genome. (ii) Numerous methods for de novo transcriptome assembly, where no reference genome is used, also exist (e.g., SOAPdenovo [6], Rnnotator [8], Trans-ABySS [10], STM and Multiple-*k* [11], Trinity [1], Genovo [4], IsoLasso [7], and T-IBDA [9]). Most of these approaches adapt tools originally designed for de novo genome assembly, for example overlap graphs and de Bruijn graphs, to de novo transcriptome assembly. However, important differences between transcriptomes and genomes reduce the usefulness of these tools for transcriptome assembly. For example, in a genome each chromosome occurs exactly once or twice, but in a transcriptome, different genes have widely varying levels of expression. Also, a single gene in the genome may be transcribed into several different isoforms in the transcriptome. A separate problem with most existing approaches to de novo transcriptome assembly is lack of a clear objective function that allows the user to understand and trust the program's output on new RNA-seq data.

(c) In order to address these issues, Prof. Dewey's group has been working on a probabilistic generative model that is specifically designed for transcriptome assembly from RNA-seq data. So far, they have used the model to estimate expression levels of isoforms from RNA-seq data, when the isoforms are known in advance [5]. I plan to work on an algorithm based on the model to learn the transcriptome when even the isoforms are unknown. To get started, I will work on an evaluation of the overall approach by analyzing the model on RNA-seq data against a known transcriptome.

2 Progression and gene expression in cervical cancer

With Michael Newton, I plan to work on an aspect of the NIH-funded SUCCEED (Study to Understand Cervical Cancer Early Endpoints and Determinants) project. Prof. Newton's biological collaborators on this project are Prof. Paul Ahlquist and Prof. Paul Lambert.

(a) The biological problem is as follows. We have whole genome expression profiles for $n = 128$ tissue samples, divided into four pathologic groups: putatively normal samples, early stage lesions (cervical intraepithelial neoplasia [CIN] 1 and 2), later stage lesions (CIN 3) and frank cancer. Roughly an equal number of tissue samples are in each group. Each tissue sample was measured by an Affymetrix whole genome microarray, which contains about 54,000 probe sets.

(b) The SUCCEED members have already carried out several analyses of these data, including analyses that aim to identify genes showing various patterns of differential expression among the four pathological groupings. However, an alternative analysis can potentially be useful in helping understand these preliminary findings. Motivation for the alternative analysis is given by the following biological and technical facts. Each expression profile is measured from a collection of around 1000 cells - so it represents a mixture of theoretical profiles. The stages of cancer progression of cervical tissue are characterized in part by changes in the proportion of cells of particular types. E.g., normal tissue is organized in layers with more well-differentiated cells at the surface and with less differentiated, but more actively dividing cells further inside the tissue. Neoplastic lesions shift the balance of types, at least partly by having relatively more of the less differentiated types and having fewer of the well-differentiated types. Note that the different types will have different gene-expression profiles.

(c) Prof. Newton has outlined at a high level a probabilistic generative model of cervical cancer progression based on the facts mentioned above. My plan is as follows. (i) First, I am currently working out several versions of this model in detail. One version given in the appendix. (ii) After details are worked out, possibly in several different ways, I will develop an algorithm to find the maximum-likelihood parameters of the model. (iii) After we have an algorithm to find the MLE, we will investigate aspects of the type-specific gene-expression profiles and also mixing proportions of different types at each stage according to the MLE. (iv) Depending on the results of this investigation, we will either look for a different model, or we will link in other relevant data into the above model, in order to get a clearer picture of cervical cancer progression.

References

[1] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn a Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, (May), May 2011.

[2] Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, John L Rinn, Eric S Lander, and Aviv Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature biotechnology*, 28(5), May 2010.

[3] C.M. Kendziorski, M.A. Newton, H. Lan, and M.N. Gould. On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, 22:3899–3914, 2003.

[4] Jonathan Laserson, Vladimir Jojic, and Daphne Koller. Genovo: de novo assembly for metagenomes. *Journal of Computational Biology*, 18(3):429–443, 2011.

[5] Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, February 2010.

[6] Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics (Oxford, England)*, 25(15):1966–7, August 2009.

[7] Wei Li, Jianxing Feng, and Tao Jiang. IsoLasso: A LASSO Regression Approach to RNA-Seq Based Transcriptome Assembly. In *Research in Computational Molecular Biology*, pages 168–188. Springer, 2011.

[8] Jeffrey Martin, Vincent M Bruno, Zhide Fang, Xiandong Meng, Matthew Blow, Tao Zhang, Gavin Sherlock, Michael Snyder, and Zhong Wang. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics*, 11(1):663, 2010.

[9] Yu Peng, Henry C M Leung, S M Yiu, and Francis Y L Chin. T-IDBA : A de novo Iterative de Bruijn Graph Assembler (Extended Abstract). (1):337–338, 2011.

[10] Gordon Robertson, Jacqueline Schein, Readman Chiu, Richard Corbett, Matthew Field, Shaun D Jackman, Karen Mungall, Sam Lee, Hisanaga Mark Okada, Jenny Q Qian, Malachi Griffith, Anthony Raymond, Nina Thiessen, Timothee Cezard, Yaron S Butterfield, Richard Newsome, Simon K Chan, Rong She, Richard Varhol, Baljit Kamoh, Anna-Liisa Prabhu, Angela Tam, Yongjun Zhao, Richard A Moore, Martin Hirst, Marco A Marra, Steven J M Jones, Pamela A Hoodless, and Inanc Birol. De novo assembly and analysis of RNA-seq data. *Nature Methods*, 7(11):909–12, October 2010.

[11] Y. Surget-Groba and J. Montoya-Burgos. Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Research*, 20(10):1432–1440, August 2010.

[12] Cole Trapnell, Brian a Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):516–520, May 2010.

[13] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, January 2009.

3 Appendix

We have gene-expression profiles for $n = 128$ tissue samples. Each gene-expression profile $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,G})$ consists of the (raw) expression levels for $G \approx 54000$ genes. We also know the stage s_i in the progression that each tissue sample was taken from. We think of these $x_{i,g}$ and s_i as realizations of random variables $X_{i,g}$ and S_i.

Since each tissue sample is in reality a mixture of T ($T > 1$ but not too large) different cell types, we model each expression profile X_i as a mixture of type-conditional profiles. In particular, we represent "the type of a particular cell from tissue sample i being of some type" by a (hidden) random variable T_i, so each T_i takes values in $\{1, \ldots, T\}$, and $P(T_i = t)$ is the fraction of cells of type t in the tissue sample i.

The tissue samples' stages s_i are fixed, just based on what tissues the experimenters chose to look at, and there does not seem to be much to be gained by modeling them. For this reason, we are interested in the conditional joint distribution of the $X_{i,g}$ given that $S_i = s_i$ for $i = 1, \ldots, n$. In other words, we want to specify some parametric form (and ultimately learn the parameters) for

$$P(\cap_{g=1}^{G} \cap_{i=1}^{n} \{X_{i,g} \in A_{i,g}\} | \cap_{i=1}^{n} \{S_i = s_i\})$$

We specify the parametric form in several steps.

Step 1. First, we assume that the expression levels are independent by genes, i.e., $X_{i,g}$ and $X_{i',g'}$ are independent for any $g \neq g'$. (But we do not assume that the profiles are independent by tissue sample, i.e., for a fixed gene g, $X_{i,g}$ and $X_{i',g}$ are not necessarily independent.) Thus:

$$\begin{aligned} & P(\cap_{g=1}^{G} \cap_{i=1}^{n} \{X_{i,g} \in A_{i,g}\} | \cap_{i=1}^{n} \{S_i = s_i\}) \\ & = \prod_{g=1}^{G} P(\cap_{i=1}^{n} \{X_{i,g} \in A_{i,g}\} | \cap_{i=1}^{n} \{S_i = s_i\}) \end{aligned}$$

Step 2. We decompose the profiles into mixtures based on type. I.e., for fixed g, we stipulate that

$$\begin{aligned} & P(\cap_{i=1}^{n} \{X_{i,g} \in A_{i,g}\} | \cap_{i=1}^{n} \{S_i = s_i\}) \\ & = \sum_{(t_1, \ldots, t_n)} P(\cap_{i=1}^{n} \{T_i = t_i\} | \cap_{i=1}^{n} \{S_i = s_i\}) P(\cap_{i=1}^{n} \{X_{i,g} \in A_{i,g}\} | \cap_{i=1}^{n} \{S_i = s_i\}, \cap_{i=1}^{n} \{T_i = t_i\}) \\ & = \sum_{(t_1, \ldots, t_n)} P(\cap_{i=1}^{n} \{T_i = t_i\} | \cap_{i=1}^{n} \{S_i = s_i\}) P(\cap_{i=1}^{n} \{X_{i,g} \in A_{i,g}\} | \cap_{i=1}^{n} \{T_i = t_i\}) \end{aligned}$$

where the sums over $(t_1, \ldots, t_n)$ are over all combinations of t_i in $\{1, \ldots, T\}$. (If T is large, there will be a lot of terms here, but we have already required that T is not too large.)

Step 3. We define the mixing proportions. We assume that (i) T_i are conditionally independent given the S_i, and (ii) $P(T_i = t_i | \cap_{i=1}^{n} \{S_i = s_i\}) = P(T_i = t_i | S_i = s_i)$. These are reasonable assumptions, because biologically we think that the proportion of cells of various types in a tissue depends (only or at least primarily) on the tissue's type. So:

$$P(\cap_{i=1}^{n} \{T_i = t_i\} | \cap_{i=1}^{n} \{S_i = s_i\}) = \prod_{i=1}^{n} P(T_i = t_i | S_i = s_i) = \prod_{i=1}^{n} p_{s_i, t_i}$$

where $(p_{1,t},\ldots,p_{4,t})_{t=1}^{T}$ are parameters that we want to find.

Step 4. We stipulate that the gene expression levels (for a fixed gene g) for tissue samples i and i' are conditionally independent given that $T_i \neq T_{i'}$. (This assumption follows [3], Section 3.) This assumption is reasonable, since if a cell from tissue i is of a different type than a cell from tissue i', we shouldn't think that their expression levels of gene g are related. (In contrast, if the cells are of the same type, then their levels would be related.)

In order to express the assumption precisely, we first define some index sets:

$$\mathscr{I}_t = \mathscr{I}_t(t_1,\ldots,t_n) = \{i : t_i = t\}$$

Note that the $\mathscr{I}_t$ are disjoint sets. Also note that the $\mathscr{I}_t$ depend on $(t_1,\ldots,t_n)$, though we will not always write this explicitly below, in order to save space.

Using this notation, we express the stipulation stated above:

$$\begin{aligned}
&P(\cap_{i=1}^{n}\{X_{i,g} \in A_{i,g}\}|\cap_{i=1}^{n}\{T_i = t_i\})\\
&= \prod_{t=1}^{T} P(\cap_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}\{X_i \in A_i\}|\cap_{i=1}^{n}\{T_i = t_i\})\\
&= \prod_{t=1}^{T} P(\cap_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}\{X_i \in A_i\}|\cap_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}\{T_i = t\})
\end{aligned}$$

Step 5. We plug in the Gamma-Gamma model from [3]. In particular, following [3], we stipulate that for each fixed tissue type t the measure on Borel$(\mathbb{R}^{|\mathscr{I}_t(t_1,\ldots,t_n)|})$ which is induced by

$$(\times_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}A_i) \mapsto P(\cap_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}\{X_i \in A_i\}|\cap_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}\{T_i = t\})$$

has the following density (wrt Lebesgue):

$$f_0(x_1,\ldots,x_{|\mathscr{I}_t|};\theta_t') = \int_{\mathbb{R}} \Big(\prod_{i=1}^{|\mathscr{I}_t|} f_{obs}(x_i|\mu;\alpha_t)\Big)\pi(\mu;\theta_t')\,d\mu$$

which is parameterized by $\theta_t' = \{\alpha_t, \alpha_{0,t}, \nu_t\}$. Here, f_{obs} is a gamma density and π is an inverse-gamma density, with specific functional forms given in [3], Section 4.

Step 6. Finally, putting all the pieces above together,

$$\begin{aligned}
&P(\cap_{g=1}^{G}\cap_{i=1}^{n}\{X_{i,g} \in A_{i,g}\}|\cap_{i=1}^{n}\{S_i = s_i\})\\
&= \prod_{g=1}^{G}\sum_{(t_1,\ldots,t_n)}\Big[\prod_{i=1}^{n} p_{s_i,t_i}\Big]\Big[\prod_{t=1}^{T} P(\cap_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}\{X_i \in A_i\}|\cap_{i\in\mathscr{I}_t(t_1,\ldots,t_n)}\{T_i = t\})\Big]
\end{aligned}$$

and the measure induced by

$$(\times_{g=1}^{G}\times_{i=1}^{n}A_{i,g}) \mapsto P(\cap_{g=1}^{G}\cap_{i=1}^{n}\{X_{i,g} \in A_{i,g}\}|\cap_{i=1}^{n}\{S_i = s_i\})$$

has density

$$f(x_{1,1},\ldots,x_{n,G};\boldsymbol{\theta}) = \prod_{g=1}^{G} f_g(x_{1,g},\ldots,x_{n,g})$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_t', p_{1,t},\ldots,p_{4,t})_{t=1}^{T}$ and

$$f_g(x_{1,g},\ldots,x_{n,g};\boldsymbol{\theta}) = \sum_{(t_1,\ldots,t_n)} \Big[\prod_{i=1}^{n} p_{s_i,t_i}\Big]\Big[\prod_{t=1}^{T} f_0(\mathbf{x}_{\mathscr{I}_t(t_1,\ldots,t_n),g};\boldsymbol{\theta}_t')\Big]$$

and $\mathbf{x}_{\mathscr{I}_t(t_1,\ldots,t_n),g} = (x_{i,g} : i \in \mathscr{I}_t(t_1,\ldots,t_n))$.