

Genome Annotation

Project Assignment – *Bioinformatics* module 2009

23rd of March, 2009

The experimental side of bioinformatics is spewing out genomes at an ever increasing rate. A blank genome is of little use, especially in higher organisms where some estimates of the ‘junk’ percentage are well into the nineties. Once a genome has been completed, the first task of bioinformatics is to annotate the genome. This consists of inferring all the ‘functional units’ of the genome. What should be considered a functional unit can be discussed at length, but it is suggested that you focus on protein coding gene finding, non-coding RNA gene finding, and alternative splicing. Other project groups will be looking into regulatory annotation, functional annotation, and patterns of molecular evolution on the same genomic region. You are encouraged to discuss and collaborate with the other groups.

The first thing needed to annotate a genome is a genome. Rather than working on a full genome, which can be a rather time consuming process, we will restrict your attention to a small region containing known genes believed to be associated with lung cancer. So the first thing you should do is to retrieve the part of the human genome containing the three genes *CHRNA5*, *CHRNA3*, and *CHRNB4*. Include about 50,000 bases upstream and downstream of this region. This should be your core data set for this project.

The second thing needed to annotate a genome is annotators. Given the limited time available, we do not expect you to develop novel state-of-the-art annotators, but rather to investigate existing annotation tools, how they work and their performance on your data set. These instructions should give some pointers of where to begin, but if you come across other methods you find interesting, do pursue these. You may also at your discretion decide to ignore one or more leads in these instructions. The important thing is that you carry out an in-depth annotation study of your data. At the end of the module, that is, on Friday the 3rd of April, you will be expected to give a presentation of about 45 minutes describing your findings and the concepts underlying the methods you have been applying.

One of the reference tools for gene finding based on a single genome is Genscan [2]. It applies a hidden Markov model describing the normal structure of a eukaryotic gene, complete with exons, introns, start and stop codons, etc. Applying this to your core data set you should get a small set of gene predictions. How does this compare to the annotation in the human genome browser? Are the scores reported by Genscan in any way informative about which predictions that can actually be trusted?

In the decades prior to its publication, estimates of the number of genes in the human genome were somewhat higher than the number now generally accepted. This is not just a thing of the distant past [1]. Did you observe a similar outcome with your Genscan annotation? One way of improving the specificity of an annotation is to require corroborating evidence. With more and more genomes becoming available, a very natural choice for such evidence is conservation across genomes. As is the case with single sequence gene finding, several choices exist for so-called comparative gene finding. A good place to start it by going to the software section of <http://www.geneprediction.org/>. If you are up for installing software, we of course recommend EvoGene [9], but you may want to explore other options that mention comparative, phylogenetic, pair, or evolutionary in their description. For a comparative method to work, you will of course also have to locate auxiliary data from other species that is homologous to your human core data set. The more distant the species are, the stronger the signal left by conservation will be. However, at too large a distance, genes may have been lost or gained, or it may not even be possible to find homologous regions. Does comparative methods significantly improve the predictions in your core data?

So far we have only been conserved with finding genes encoding proteins. The core data was chosen based on a set of genes known to encode proteins, but for annotation purposes it is still interesting to explore the presence of non-coding RNA genes, i.e. DNA that encodes a functional RNA molecule. It is the consensus that *ab initio* single genome RNA gene finding is not feasible [10]. A couple of comparative RNA gene finders [8, 11] are publicly available. Is there any evidence for conserved RNA genes in your core data? What are the basic signals that the two gene finders are attempting to detect? What are the estimated error rates of these gene finders? Would it be sensible to try to incorporate more subtle signals like [6, 7], and if so, how much do you believe could be gained? It is a lot easier to find something when you know what you are looking for. A popular example of RNAs to look for are μ RNAs (or miRNAs). Mature μ RNA binds to the 3' end of mRNA which marks it for degradation. Are there any known μ RNAs in your core data? Can you find any evidence for μ RNA targets in the genes in your core data?

One final thing you may want to explore is presence of alternative splicing in the genes we have been looking at. Though the number of protein coding genes is believed to be in the low 20,000s, the number of resulting protein variants is much higher. The reason for this is that a gene may encode many spliceforms, i.e. the exons from the pre-mRNA spliced together to form the mature mRNA that is translated to a protein may be subject to context dependent or random variation. A few good places to start an investigation into alternative splicing are the alternative splicing gallery [5], BIPASS [4], and ASPicDB [3].

References

- [1] Jeffrey L Bennetzen, Craig Coleman, Renyi Liu, Jianxin Ma, and Wusirika Ramakrishna. Consistent over-estimation of gene number in complex plant

- genomes. *Current Opinion in Plant Biology*, 7:732–736, 2004.
- [2] Chris Burge and Samuel Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268:78–94, 1997.
- [3] T. Castrignano, M. D’Antonio, A. Anselmo, D. Carrabino, A. D’Onorio De Meo, A.M. D’Erchia, F. Licciulli, M. Mangiulli, F. Mignone, G. Pavesi, E. Picardi, A. Riva, R. Rizzi, P. Bonizzoni, and G. Pesole. ASPicDB: a database resource for alternative splicing analysis. *Bioinformatics*, 24(10):1300–1304, 2008.
- [4] Z. Lacroix, C. Legendre, L. Raschid, and B. Snyder. BIPASS: Bioinformatics Pipeline Alternative Splicing Services. *Nucleic Acids Research*, 35:W292–W296, 2007.
- [5] J. Leipzig, P. Pevzner, and S. Heber. The alternative splicing gallery (ASG): bridging the gap between genome and transcriptome. *Nucleic Acids Research*, 32(13):3977–3983, 2004.
- [6] Irmtraud M. Meyer and István Miklós. Co-transcriptional folding is encoded within RNA genes. *BMC Molecular Biology*, 5:10, 2004.
- [7] Naila K. Mimouni, Rune B. Lyngsø, Sam Griffiths-Jones, and Jotun Hein. An analysis of structural influences on selection in RNA genes. *Molecular Biology and Evolution*, 26:209–216, 2009.
- [8] Jakob Skou Pedersen, Gill Bejerano, Adam Siepel, Kate Rosenbloom, Kerstin Lindblad-Toh, Eric S. Lander, Jim Kent, Webb Miller, and David Haussler. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Computational Biology*, 2(4):e33, 2006.
- [9] Jakob Skou Pedersen and Jotun Hein. Gene finding with a hidden markov model of genome structure and evolution. *Bioinformatics*, 19(2):219–227, 2003.
- [10] Elena Rivas and Sean R. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16:583–605, 2000.
- [11] Stefan Washietl, Ivo L. Hofacker, and Peter F. Stadler. Fast and reliable prediction of noncoding RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2454–59, 2005.