Population Sequencing from Chromatogram Data

D. Dueck[1], C. Abha[2], C. Moore[2], J. Sarich[2], D. Goodridge[3], D. Heckerman[1], D. Sayer[3], S. Mallal[2], N. Jojic[1]
[1]Microsoft Research, One Microsoft Way, Redmond, WA. [2]Centre for Clinical Immunology and Biomedical Statistics, Royal Perth Hospital, Wellington Street, Perth, Western Australia. [3]Conexio 4, East Fremantle 6158 Western Australia.

One of the key components of sequencing technologies [1]is proper separation of a single species/strain/allele of the target sequence (*e.g.*, gene) from a sample. Traditionally, this is achieved chemically, for example through the use of specific primer sequences. However, it is possible that multiple related species are picked up with the same primer. This is especially problematic in sequencing RNA or proviral DNA, when the virus in question is highly variable and each individual is infected with a different swarm of viral strains. In the case of HIV, when dominant sequences in the population differ by one or more insertions/deletions, standard sequencing techniques fail to recover any component strains satisfactorily. For example, regions of HIV such as the envelope proteins (which are subjected to strong selective pressures from host defenses and result in multiple mutations, insertions, and deletions), most chromatograms obtained by DNA sequencers have unusable sections.

We show that chromatograms of mixed sequences, such as the segment shown in Figure 1, can be used to accurately infer the individual strains thus eliminating the need for additional sequencing steps (*e.g.*, new primer synthesis, cloning of individual viral variants). To this end, we have developed a statistical generative model of raw chromatogram data and an appropriate inference algorithm for maximizing the likelihood of an observed chromatogram. To illustrate this technique, we employ an automated ABI 3730XL sequencer to capture mixed samples of proviral DNA, however, this could be used in other applications where the need for sequencing populations, rather than individual strains or alleles, arises.

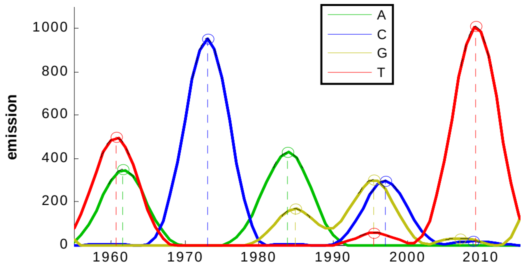

Figure 1: Mixed-sequence chromatogram

To experimentally test our algorithms performance, we focused on sequencing Human Immunodeficiency Virus (HIV). For that purpose, we first obtained samples from HIV-infected individuals in the WA cohort [2]. We cloned 6KB PCR amplified fragments corresponding to the region 3060bp – 9500bp of the reference strain (HXB2) into TOPO TA cloning vector (Invitrogen) according to the manufacturer's instructions.

Clones were randomly selected and sequenced using internal primers and the BigDye terminator kit. The products were run on an automated DNA sequencer 3730XL. The two major species were then selected and the corresponding clones were mixed in varying proportions (1:0, 1:1, 5:3, 5:1 and 0:1) and sequenced as detailed above. Chromatograms of the mixed samples were analyzed by the presented algorithm, providing as output the inferred individual strains for each mixture which were then compared with the sequences of the original clones. In many cases, the separated components had fewer than 1% differences to the ground truth. This compares favourably to the output of the basic sequencer, where error rates were typically an order of magnitude worse (ranging up to 40%). In addition, the algorithm recovered the mixing ratios fairly well, even though the amplitude variance across the chromatograms was fairly high. For example, the estimated relative concentrations for two 1:1 mixtures were 48:42 and 49:51, and the inferred ratio for the 5:3 mixtures was 60:40. Finally, we find that an important cue in separating the mixtures is the phase offset of the strains (a slight shift in the chromatogram components corresponding to the two strains), which the algorithm infers at sub-sample accuracy.

[1] Sanger, F., Nicklen, S and Coulson, A. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.*, **74**:5463-5467.

[2] Mallal, S. (1998) The Western Australian HIV Cohort Study, Perth, Australia. *Journal of Acquired Immune Defficiency Syndromes and Human Retrovirology*; **17** (Suppl 1): S23-S27.