Modeling Sequencing Errors
by Combining Hidden Markov Models

Claudio Lottaz\textsuperscript{1,2,*}, Christian Iseli\textsuperscript{1,3},
C. Victor Jongeneel\textsuperscript{1,3} and Philipp Bucher\textsuperscript{1,2}

\textsuperscript{1}Swiss Institute of Bioinformatics (SIB)
\textsuperscript{2}Swiss Institute of Experimental Cancer Research (ISREC)
\textsuperscript{3}Office of Information Technology, Ludwig Institute for Cancer Research (LICR)
155, ch. des Boveresses, CH-1066 Epalinges s/Lausanne (Switzerland)

Abstract
Among the largest resources for biological sequence data is the large amount of expressed sequence tags (ESTs) available in public and proprietary databases. ESTs provide information on transcripts but for technical reasons they often contain sequencing errors. Therefore, when analyzing EST sequences computationally, such errors must be taken into account. Earlier attempts to model error prone coding regions have shown good performance in detecting and predicting these while correcting sequencing errors using codon usage frequencies. In the research presented here, we improve the detection of translation start and stop sites by integrating a more complex mRNA model with codon usage bias based error correction into one hidden Markov model (HMM), thus generalizing this error correction approach to more complex HMMs. We show that our method maintains the performance in detecting coding sequences.

Keywords coding region prediction, sequencing errors, expressed sequence tags, hidden Markov models

Contact Claudio Lottaz\textsuperscript{1} Claudio.Lottaz@molgen.mpg.de

\textsuperscript{*}Current adress: Max-Planck-Institute for Molecular Genetics, Ihnestr. 73, D-14195 Berlin, phone +49 30 8413 1177, fax +49 30 8413 1176, Claudio.Lottaz@molgen.mpg.de

\textsuperscript{1}Current adress: Max-Planck-Institute for Molecular Genetics, Ihnestr. 73, D-14195 Berlin (Germany)