Bastien Chevreux
Dr. sc. hum.

MIRA: An Automated Genome and EST Assembler

Geboren am 05.07.1972 in Duisburg
Diplom der Fachrichtung Medizinische Informatik am 1. März 1997 an der Universität Heidelberg

Promotionsfach: Biophysik (DKFZ)
Doktorvater: Prof. Dr. S. Suhai

Shotgun sequencing genomic sequences for subsequent reconstruction is comparable to assembling a jigsaw puzzle. These puzzles, of course, are much more complex than the average jigsaw puzzle: they tend to be about 500 to 5 million pieces, printed on both sides, with many vital pieces possibly missing. Some of the pieces are dirty or unrecognisable, and several pieces from another puzzle might have been mixed in. Additionally, a few pieces themselves appear to have been cut and reassembled by a very impatient two-year-old with a pair of scissors and a bottle of glue.

The extensively studied reconstruction of the unknown, correct contiguous nucleic acid sequence by inferring it through the help of a number of representations1 is called the assembly problem. The devil is in the details, however. If the collected sequences were 100% error free, then many problems would not occur. In reality, the extraction of data by electrophoresis is a physical process in which errors due to biochemical phenomena show up quite often. Current laboratory technologies total an error rate that might be anywhere between 0.1% – for good parts in the middle of a sequence – and more than 10% in bad parts at the very beginning and at the end. This error rate, combined with the sometimes exacerbating fact that both DNA and RNA tend to contain highly repetitive stretches with only very few bases differing across different repeat locations, impedes the assembly process in a daunting way.

A new strategy for assembling genomic shotgun and EST sequence data was developed and worked out in this thesis. It combines novel enhancements like repeat detection and on-the-fly automatic editing with strengths of existing assemblers. The strategy also provides the assembler with the ability to use and – more importantly – to acquire by itself additional knowledge present in the assembly data. Furthermore, the knowledge acquisition was combined with the ability to resolve potential conflicts – like long term repeats in genome sequencing projects or different mRNA transcripts in EST projects – during the assembly by falling back to trace signal analysis routines.

Especially the possibility to discriminate alternative solutions – due to previously unknown short and long term repeats – during the assembly process constitutes a systematic improvement in quality of assembly algorithms that produce sequences as accurate as possible.

The main aim achieved in this thesis was to reduce assembly errors caused by repetitive sequences as well as to increase the reliability of consensus sequences derived from automatically assembled projects. The results presented demonstrate that the combination of the methods and algorithms devised for this thesis leads to a system that achieves this task. It reliably accomplishes the given task of reconstructing genomic or transcriptomic sequences from DNA or RNA fragments. This is done through the detection, analysis and classification of repetitive elements or single nucleotide polymorphisms which in turn prevents grave misassemblies that occur in other systems.

In most analysed assembly comparisons, the quality of the resulting consensus sequences was improved and the number of errors per kilobase consensus sequence was decreased. The improved strategy described here therefore permits to use resulting sequences almost directly for the design of further investigative studies with high quality and precision requirements like, e.g., the design of oligo probes in clinical micro-array hybridisation screening experiments.