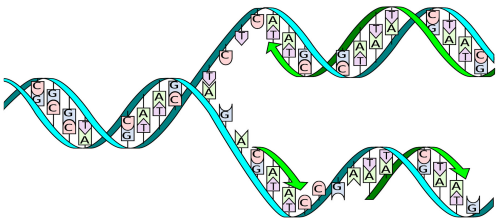
Hannes Diethelm

Graduate Candidate	Hannes Diethelm
Examiner	Prof. Dr. Guido Schuster
Co-Examiner	Gabriel Sidler
Subject Area	Sensor, Actuator and Communication Systems

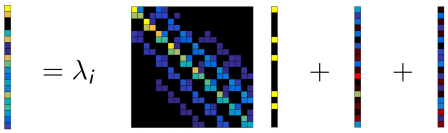
All Your Base revisited: DHBC

A fast and robust base caller for next generation DNA sequencing



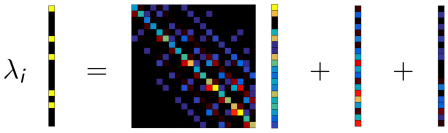
DNA replication process which is used in next generation sequencing

$$\operatorname{vec} I_i = \lambda_i A \operatorname{vec} S_i + \operatorname{vec} N + \operatorname{vec} \varepsilon_i$$



Forward model according to All Your Base

$$\lambda_i \operatorname{vec} S_i = \tilde{A} \operatorname{vec} I_i + \operatorname{vec} \tilde{N} - \operatorname{vec} \tilde{\varepsilon}_i$$



Inverse model firstly used in DHBC

Introduction: In the year 2012, Tim Massingham and Nick Goldman published the article "All Your Base: a fast and accurate probabilistic approach to base calling". They proposed a new statistical model of the DNA sequencing process as a foundation for their base calling algorithm. According to their paper, it is the best performing algorithm for this task at that time. They used data from the Illumina Genome Analyzer, but it should also work for similar next generation sequencers. Based on this work, Lukas Schmid and Roman Koller reimplemented the algorithm in Matlab. After a tutorial about sparse and low rank recovery problems, the idea came up to use these methods to recover the interaction matrix used by AYB because this matrix should be sparse according to the model. At the beginning, this was the main target of this thesis.

Proceeding: In a first step, the Matlab code received was reworked and optimized. A speedup of ten times was already possible reducing the base calling time for the used data from five minutes to 30 seconds. Unfortunately, the sparse modeling took about 25 times longer than the Tikhonov regularization used in AYB, without significantly improving the quality, hence this approach was temporary suspended. Other ways of improving the quality where researched. By inverting the model, and so minimizing the base calling error instead of the intensity error, the performance of the original AYB on Illumina data is nearly reached without using the time consuming Viterbi algorithm. On alterantive data, the quality is even significantly better than the original AYB. Nearly at the end of the thesis, a trick was discovered which made the sparse model approach computationally equivalent to the Tikhonov regularization. Unfortunately, the sparse model does not lead to better results, hence Tikhonov regularization is still used.

Result: The new algorithm is now called DHBC: Diethelm Hannes Base Caller. It is also implemented in C which gives a two times speedup compared to Matlab. By heavy verification on many data sets, the robustness has been proven. It is 4 to 15 times faster than the original AYB in C code and 300 times faster than the previous Matlab implementation, while it gives 20 to 55% more perfect matches on alternative data. However, on Illumina data, there are up to 4% less perfect matches.