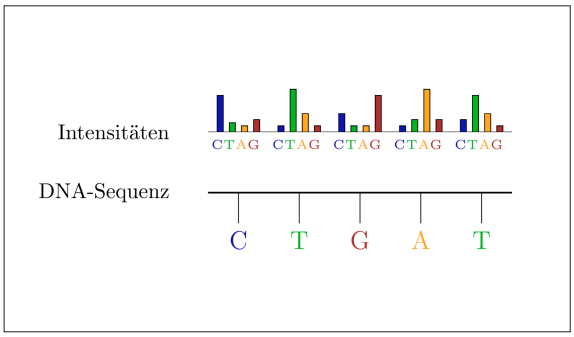
Hannes Badertscher

Tabea Méndez

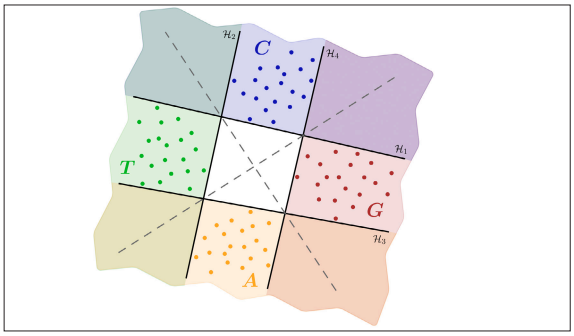
Diplomanden	Hannes Badertscher, Tabea Méndez
Examinator	Prof. Dr. Guido Schuster
Experte	Gabriel Sidler, Eivycom GmbH, Uster, ZH
Themengebiet	Digitale Signalverarbeitung

Support Vector Machines for Basecalling

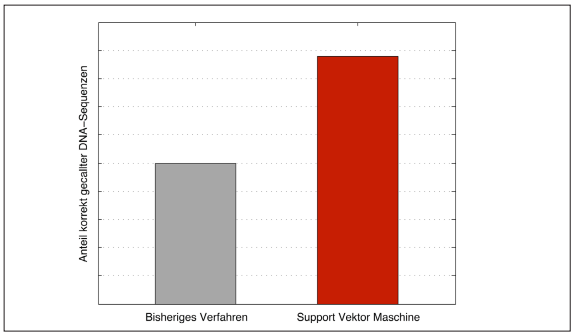
Bestimmung der Basenabfolge in einer DNA-Sequenz mithilfe einer Supervised Learning Strategie



Prozess des Basecalling: Detektion der korrekten Base anhand der gegebenen Intensitäten



Trennebenen der Support-Vektor-Maschinen, welche die Messpunkte (Vektoren der Intensitäten) in die vier Gruppen (C, T, A, G) aufteilt



Ergebnisse der Support-Vektor-Maschine, verglichen mit dem bisherigen Verfahren

Ausgangslage: In den letzten Jahren wurden verschiedene Next-Generation-Sequencing-Verfahren entwickelt, welche eine schnelle und günstige Bestimmung von DNA-Sequenzen erlauben. Bei solchen Verfahren wird jede Position in der DNA-Sequenz mit einem Laser belichtet, wobei gleichzeitig die vier Intensitäten, welche den möglichen Basen (C, T, A, G) entsprechen, gemessen werden. Aus den gemessenen Intensitäten muss anschliessend die entsprechende Base bestimmt werden, was als Basecalling bezeichnet wird. Verschiedene Fehlereinflüsse, z.B. ein Übersprechen (Crosstalk) zwischen den Intensitäten oder eine Verschiebung innerhalb der DNA-Sequenz (Phasing), erschweren das Basecalling und begrenzen die Möglichkeiten heutiger Verfahren. Um Fehlereinflüsse zu korrigieren, nutzen bestehende Verfahren statistische Modelle oder Supervised-Learning-Strategien (SLS).

Vorgehen: In der vorliegenden Arbeit wird das Basecalling mit Support-Vektor-Maschinen (SVM), einer SL-Strategie, gelöst. In einem ersten Schritt wurde dazu ein bestehender SVM-Basecaller analysiert, welcher bereits gute Ergebnisse liefert. Dabei lernt die SVM in einer Trainingsphase die Muster der Fehlereinflüsse, indem sie Messpunkte (Vektoren der Intensitäten) bekannter DNA-Sequenzen der jeweiligen Base zuordnet und die vier Gruppen mittels Trennebenen voneinander separiert. Neue Messdaten können anschliessend mittels der gefundenen Trennebenen den Basen zugeordnet werden. In einem zweiten Schritt wurde ein neuer Basecaller entwickelt. Dieser wurde anschliessend mit systematischen Tests optimiert und mit bestehenden Verfahren verglichen.

Ergebnis: In dieser Arbeit wurde gezeigt, dass das Basecalling mit Support-Vektor-Maschinen deutlich bessere Resultate erreicht als bisher eingesetzte Verfahren. So konnte mit dem neu entwickelten Basecaller der Anteil korrekt gecallter DNA-Sequenzen gegenüber dem bisherigen Verfahren um 76% gesteigert werden. Dadurch sinkt die Fehlerrate (Anteil falsch gecallter Basen) auf unter 10%. Die Erhöhung der für das Basecalling benötigten Zeit von wenigen Minuten auf insgesamt 1716 Minuten (Training: 1152 min, eff. Basecalling: 546 min) könnte durch eine zukünftige Parallelisierung der Software noch stark reduziert werden. Somit bildet diese Arbeit eine solide Grundlage für mögliche Weiterentwicklungen.