Die Transkriptionsregulation beschreibt die zeitliche und räumliche Expression der Gene. Mit Hilfe der Identifikation von transregulatorischen Elementen, wie beispielsweise Transkriptionsfaktorbindestellen, können regulatorische Netzwerke besser verstanden werden. Regulatorische Netzwerke beschreiben zelluläre Prozesse wie zum Beispiel die Zellentwicklung und das Entstehen von Krankheiten.

Beim herkömmlichen rechnergestützten Ansatz zur Identifikation von Transkriptionsfaktorbindestellen wird auf Sequenzierungsmethoden zurückgegriffen, um die DNA des Genoms nach Sequenzen mit unterschiedlichen Bindungsneigungen zu Transkriptionsfaktoren (TF) zu durchsuchen. Mit diesem Ansatz ist es jedoch nicht möglich aktive Bindestellen vorherzusagen. Eine aktive Bindestelle ist beispielsweise dann gegeben, wenn an der DNA-Sequenz ein TF bindet. Dieser auf Sequenzierungstechniken beruhende Ansatz nimmt keinen Bezug darauf, dass der Zustand des Chromatins dynamisch zwischen offen (so dass ein TF binden kann) und geschlossen (so dass kein TF binden kann) wechseln kann.

Mit Sequenzierungsmethoden der nächsten Generation (next generation sequencing) kann offenes Chromatin genomweit identifiziert werden. Beispiele hierfür sind die Kombination von Chromatin ImmunoPrecipitation (ChIP-seq) oder DNase I Verarbeitung (DNase-seq) mit der Sequenzierungstechnik. Aktuelle Studien haben belegt, dass die Verwendung von ChIP-seq und DNase-seq zur Bestimmung von offenem Chromatin einen positiven Einfluss auf die Identifikation von aktiven TFBS haben. Dabei wird die Suche nach charakteristischen DNA-Sequenzen auf die Bereiche eingeschränkt, an denen das Chromatin offen ist und die TF somit binden können.

Wir führen in dieser Arbeit ein rechnergestütztes Rahmenwerk ein, das DNase-seq und ChIP-seq Daten kombiniert, um aktive TFBS vorherzusagen. Wir und andere Forscher haben beobachtet, dass es bei aktiven TFBS ein ausgeprägtes Muster in DNase-seq und ChIP-seq Daten gibt. Unser Rahmenwerk führt zunächst eine Normalisierung des Signals aus und sucht dann in den Daten nach diesen Mustern, den sogenannten Fußabdrücken. Dabei wird das Genom mit einem Hidden Markov Modell segmentiert. Unsere Methode mit dem Namen HINT (HMM-basierte Identifikation von TF Fußabdrücken) ist als „rechnergestützte Fußabdruck Methode“ klassifiziert.

In unserer Evaluierungsstudie haben wir die vorhergesagten Fußabdrücke von HINT mit bereits validierten Fußabdrücken verglichen. Dabei haben wir Statistiken erzeugt, um unsere Methode mit anderen zu vergleichen. Unsere Experimente sind mit insgesamt 14 verglichenen Methoden und 233 TF die umfangreichsten.

Zudem haben wir HINT erfolgreich bei zwei biologischen Studien angewandt, um regulatorische Elemente, die bei bestimmten biologischen Bedingungen vorkommen, zu identifizieren. HINT ist ein nützliches rechnergestütztes Rahmenwerk für biologische Studien in der regulatorischen Genomik.