Diss. ETH N° 22150

COMPUTATIONAL EPIGENOMICS AND PROGRESSION OF CANCER

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

THOMAS SAKOPARNIG

Master of Science, King's College London

born on 10. 05. 1986

citizen of Austria

accepted on the recommendation of

Prof. Dr. Niko Beerenwinkel, ETH Zürich
Prof. Dr. Mark Robinson, University of Zurich
Dr. Florian Markowetz, Cancer Research UK Cambridge Institute
Prof. Dr. Tanja Stadler, ETH Zürich

2014

Abstract

Although recent technological advancements give rise to unprecedented amounts of biological data, our understanding of fundamental biological processes such as transcription regulation, cell differentiation or carcinogenesis did not keep pace with the growth of available data. In this thesis I present four computational studies which I conducted as a PhD student in the Computational Biology Group at the Department of Biosystems Science and Engineering at ETH Zürich. These studies focus on modeling epigenomic regulation of transcription and progression of cancer on inter and intra-tumor levels.

We integrated epigenomic marker data – ChIP-chip (chromatin immunoprecipitation followed by microarray analysis) and ChIP-seq (chromatin immunoprecipitation followed by next generation sequencing) data with transcriptomics data – RNA sequencing and short RNA sequencing (Chapter 2), in order to assess how well binding of chromatin modifiers predicts transcriptional activity. The study was performed on *Drosophila melanogaster* S2 cells. We compared the predictive models for transcription as measured by RNA sequencing with the predictive models for stalling as measured by short RNA sequencing. This study reveals complex interactions of transcriptional regulators in the context of gene expression.

Throughout my work on cancer development, I was particularly interested on modeling cancer progression. First, introduced a Bayesian version of Conjunctive Bayesian networks, a probabilistic model which identifies dependencies between mutations, together with an efficient algorithm for inference (Chapter 3). Therefore, we developed five MCMC structure move types which allow for significant speed-ups when sampling the structure of CBNs compared to state-of-the-art Bayesian network MCMC sampling techniques. The method was applied to comparative genome hybridization (CGH) data from renal cell carcinoma and sequencing data glioblastoma. Furthermore, we established a method for discriminating dependent cancer events from independent cancer events (Chapter 4). This method is based on the observation that dependent events display a punctuated pattern of occurrence in time. We applied this method to copy number aberration (CNA) data from ovarian and breast cancer and single nucleotide variant (SNV) data from breast cancer. We found already known as well as novel driver candidates highly ranked in the list of predicted dependent events. Finally, we developed BitPhylogeny, a probabilistic framework for reconstructing intra-tumor phylogenies (Chapter 5). In contrast to competing methods which are based on SNV frequency or CNA data, we focused on CpG methylation data from bisulfite sequencing. Application of BitPhylogeny to 32 samples from four different colon cancers revealed the differentiation hierarchies present in these tumors.

Kurzfassung

Technischer Fortschritt ermöglicht die Generierung von biologischen Daten in zuvor niegesehenem Ausmass. Das Verständnis wie grundlegende biologische Prozesse wie Transkriptionsregulation, Zelldifferenzierung oder Krebsentstehung funktionieren, hält aber nicht Schritt mit dem Wachstum von Biodatenbanken. In dieser Dissertation präsentiere ich vier rechnerbasierte Studien, welche ich als Doktorrand in der Computional Biology Gruppe am Department für Biosysteme der ETH Zürich durchgeführt habe. Die Studien konzentrieren sich auf die Modelierung von epigenomischer Regulierung von Transkription und auf Tumorprogression.

Es wurden epigenomische Markerdaten – ChIP-chip (Chromatin Immunoprecipitation mit anschliessender Microarray Analyse) und ChIP-seq (Chromatin Immunoprecipitation mit anschliessender Sequenzierung) Daten mit transkriptomischen Daten – RNA Sequenzierung und short RNA Sequenzierung - integriert. Das Ziel war die Beurteilung wie gut Bindung von Chromatinmodifizierern die transkriptionelle Aktivität vorhersagen kann. Als Modelsystem wurden *Drosophila melanogaster* S2 Zellen verwendet. Wir vergleichen prediktive Modelle für Transkription welche mit RNA Sequenzierung gemessen wurde mit prediktiven Modellen für stalling welche mit short RNA Sequenzierung gemessen wurde. Diese Studie enthüllt komplexe Interaktionen von Transkriptionsregulatoren im Kontext von Genexpression.

Im Bereich der Tumorentstehung konzentrierte ich mich auf den Bereich der Tumorprogression. Es wird eine Bayesianische Version der Conjunctive Bayesian Networks inklusive eines effizienten Algorithmus für Inferenz beschrieben. Wir entwickelten fünf MCMC Strukturänderungen welche signifikante Geschwindigkeitserhöhungen beim Samplen der Struktur von CBNs verglichen mit modernen Bayesianischen Netzwerk MCMC Methoden erlauben. Diese Methode wurde auf comparative genome hybridization (CGH) Daten von Nierenzellkarzinomen und Sequenzierungsdaten von Glioblastomaproben angewandt. Weiters, präsentiere ich eine Diskriminierungsmethode um abhängige Krebsereignisse von unabhängigen zu unterscheiden (Kapitel 4). Diese Methode basiert auf der Beobachtung, dass abhängige Krebsereignisse konzentriert in der Zeit auftreten. Wir wenden diese Methode auf Kopienzahlvariationsdaten von Ovarialkarzinomen und Brustkarzinomen und Punktmutationsdaten von Brustkariznomen an. Schon bekannt und auch neue Krebstreiberkandidaten wurden dabei in der Liste der prognostizierten abhängigen Krebstreiber identifiziert. Wir haben ein Wahrscheinlichkeits-basiertes Rahmenwerk, genannt BitPhylogeney, entwickelt, welches auf die Rekonstruktion von intra-Tumor Phylogenien abzielt. Im Gegensatz zu konkurierenden Methoden, welche auf SNV Frequenzdaten oder CNA Daten ausgerichtet sind, ist unsere Methode auf die Analyse von CpG Methylierungsdaten ausgerichtet. Die Anwendung von BitPhylogeny auf 32 Proben von vier verschiedenen Darmkrebstumoren deckt die Tumorzellhierarchien dieser Tumore auf.