Zusammenfassung

Nukleosome stellen die kleinste Untereinheit der kompaktesten Form eukaryotischer DNA, des Chromosoms, dar. Sie bestehen aus einem Protein, das über elektrostatische Wechselwirkungen an die DNA-Sequenz bindet. Eine bedeutende Rolle sowohl für die Transkription von Genen, als auch für die gezielte Positionierung des gesamten Nukleosom-Ensembles der Zelle spielen Regionen auf der DNA, die besonders nukleosomarm sind. In dieser Arbeit gehe ich der Frage nach, ob sich die thermodynamisch große Bedeutung dieser Regionen in einer evolutionären Selektion der molekularen Eigenschaft "Nukleosom-Ausschluss" niederschlägt.

Auf Basis eines biophysikalischen Modells für die Nukleosomen-Dichte erlauben es die verwendeten statistischen Verfahren, eine Fitness-Landschaft aus *S. cerevisiae*- und *S. paradoxus*-Sequenzdaten abzuleiten. Der dadurch nachgewiesene Selektionsdruck auf nukleosomarme Regionen beschreibt nicht nur die Evolution zwischen den Spezies korrekt, sondern zeigt sich auch auf dem Niveau einer gesamten Population. Das hierzu auf Polymorphismus-Daten angewandte Verfahren ermöglicht selbst im Falle eines geringen Selektionsdrucks auf einzelne Mutationen, wie hier, Aussagen über die Fitness-Landschaft.

Da die DNA-Sequenz zahlreichen Selektionsdrücken unterliegen kann, ist es von Interesse, jene davon herauszurechnen, die nicht auf die betrachtete Eigenschaft wirken. Es wird gezeigt, dass sich unter Verwendung von Polymorphismus-Daten eine solche Zerlegung erreichen lässt und damit die Möglichkeit, den Einfluss "scheinbarer" Selektion zu quantifizieren. Eine denkbare Quelle solcher scheinbarer Selektion sind Bindungsstellen von Transkriptionsfaktoren. Ihr kompetitives Binden an die DNA beeinflusst, neben der Sequenz selbst, die Bildung von Nukleosomen in der Zelle. Basierend auf der Anzahl der Transkriptionsfaktor-Bindungsstellen, wird hier ein vereinfachtes kollektives Modell zur genomweiten Beschreibung beider Prozesse dargestellt, welches eine gute Übereinstimmung mit experimentellen Daten in *S. cerevisiae* liefert.

Neben den nukleosomarmen Sequenzen existieren natürlich auch DNA-Zusammensetzungen, welche die Nukleosom-Bildung begünstigen. In Simulationen des thermodynamischen Bindungsprozesses zeigt sich hierbei eine charakteristische Verteilung von bindungsbegünstigenden Basenpaaren, welche in Experimenten wiederholt beobachtet, jedoch nie zufriedenstellend erklärt worden ist. Hier zeige ich, dass sie sich auf den zugrundeliegenden statistischen Prozess zurückführen lässt.

Abstract

Eukaryotic cells possess a cell nucleus which encloses the blueprint for the cell's construction and functioning, its DNA. In its most compact form, the DNA is condensed over several stages into chromosomes. At the very beginning of this compaction process lies the binding of so-called histone octamers − large protein complexes − to the DNA, forming structures called nucleosomes. Many of them assemble like beads on the DNA string, and the pattern thus formed is by no means arbitrary. Sequence composition has a large influence on this pattern, and it is nucleosome-depleted regions (NDRs) which play a particularly important role in this context. In fact, they serve to stably position the whole genomic ensemble of nucleosomes. In this thesis, I investigate the extent to which this function of NDRs has had a measurable impact on genome evolution. A biophysical modelling algorithm for the genome-wide nucleosome occupancy provides a genotype-to-phenotype map, associating a given DNA sequence with its nucleosome formation potential. It is then possible to infer a fitness landscape from the sequence data of the two yeast species *Saccharomyces cerevisiae* and *S. paradoxus*. This fitness landscape implies a selective pressure acting on the phenotype of histone repulsion for an ensemble of functional NDR sequences. It not only correctly describes the phenotypic divergence between the two species, but it is also recovered at the level of a population of individuals. The method applied in this context to polymorphism data affords the inference of selection even in the case of weakly selected individual mutations, like in this analysis.

Especially in regions between genes, DNA can be subject to several selective pressures. The majority of NDRs is indeed found in these regions, suggesting another important functional aspect in keeping gene promoters highly accessible to transcription factors, which are necessary for the initiation of gene expression. As a consequence, NDRs often overlap with transcription factor binding sites. It is therefore of interest to be able to disentangle such overlapping functions and clean the selection signal of sources of apparent selection that is unrelated to the phenotype under investigation. Here, a method to achieve this decomposition in polymorphism data is presented.

Given that histones and transcription factors jointly bind to promoter regions and the above-mentioned model for the binding of histones, a model description of joint binding is devised. This is based on the number of transcription factor binding sites present on the DNA sequence, which is taken to be an indicator of actual transcription factor binding in the living cell. It shows that even in an approximate limit, this approach can describe in vivo nucleosome positioning data of yeast. In addition, it permits the computation of the fraction of the total variance of these data caused by the two contributions of sequence histone affinity and transcription factor binding.

Apart from histone-repelling sequences, DNA can also exhibit nucleotide compositions that stably bind histones. Characteristic dinucleotide patterns found experimentally show an oscillating distribution of binding enhancers, yet also an additional modulation across the histone-binding region, which has not been explained satisfyingly by either experiment or theory. In this thesis, it is shown that thermodynamic modelling of the histone binding process yields the observed modulated dinucleotide pattern as a natural outcome of the underlying statistical process. Depending on the two parameters of the model, there exist two limiting shapes of the modulation, which were both observed in experiments.