Algorithms for analyzing signals in DNA Applications to transcription and translation

Author(s):
Friberg, Markus

Publication Date:
2007

Permanent Link:

https://doi.org/10.3929/ethz-a-005378601 →

Diss. ETH No. 17096

Algorithms for analyzing signals in DNA: applications to transcription and translation

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

for the degree of
Doctor of Sciences

presented by
MARKUS FRIBERG
M.Sc.
born January 24, 1978
citizen of Sweden

accepted on the recommendation of
Prof. Dr. Gaston H. Gonnet, examiner
Prof. Dr. Hauke Hennecke, co-examiner

2007

Abstract

Hundreds of genomes have been sequenced, resulting in huge amounts of DNA data. Still, our understanding of how organisms work remains limited. The interpretation of DNA sequences is far from trivial, and sophisticated algorithms are required to make sense out of long strings of four seemingly simple nucleotide bases: A, C, G, and T.

This thesis contains algorithms that can be used to analyze signals in DNA. I have focused on two problems: The first is prediction of transcription factor binding sites, which relates to the study of non-coding DNA. This is a central problem in bioinformatics, because it is of great interest to biologists, and it is challenging from a computational and statistical point of view.

The second problem is the analysis of codon bias, which involves the study of coding DNA. There is a redundancy in the 61 codons which code for the 20 amino acids. We examine the extent to which protein expression levels can be predicted from DNA-based features. Furthermore, we present and evaluate a model of tRNA reusage, which is measured with a tRNA pairing index.

The analysis of microarray and other high-throughput experimental data is helpful for our understanding of gene transcription. The analysis of those data sources is complicated (it can even be a thesis in itself), but through collaborations with other researchers I have had the opportunity to use this data both as input to my algorithms (in the case of transcription factor binding site prediction) and as a validation of proposed models (in the case of the tRNA pairing index).

Needless to say, my work is by no means a comprehensive deciphering of all transcriptional and translational mechanisms. However, I believe to have at least partly contributed to some new biological information, and to some new methods for the study of genomic data.

Zusammenfassung

Das Sequenzieren von mittlerweile mehreren hundert Genomen hat zu riesigen Datenmengen von DNA geführt. Unser Verständnis wie Organismen funktionieren ist aber immer noch beschränkt. Die Interpretation von DNA-Sequenzen ist alles andere als trivial, und ausgeklügelte Algorithmen sind notwendig, um lange Ketten von vier scheinbar einfachen Nukleinbasen (A, C, G, und T) verstehen zu können.

Diese Dissertation behandelt Algorithmen für die Analyse von Signalen in DNA. Ich habe mich auf zwei Probleme fokussiert. Das erste ist die Vorhersage von Bindungsstellen für Transkriptionsfaktoren, die sich im nicht-kodieren Teil der DNA befinden. Dies ist ein wichtiges Problem in der Bioinformatik, weil es für biologische Anwendungen wichtig und ausserdem von einem algorithmischen und statistischen Standpunkt her anspruchsvoll ist.

Das zweite Problem ist die Analyse vom "Codon Bias" im kodierenden Teil der DNA. Der genetische Code ist redundant, weil die 20 Aminosäuren von 61 verschiedenen Codons kodiert werden können, die allerdings nicht gleich häufig verwendet werden. Wir untersuchen, in welchem Ausmass die Genexpression anhand von der DNA-Zusammensetzung vorhergesagt werden kann. Ausserdem präsentieren und untersuchen wir ein Modell für die tRNA-Wiederverwendung, gemessen durch einen "tRNA Pairing Index".

Die Analyse von Microarray- und anderen experimentellen Daten ist von Nutzen für unser Verständnis der Gentranskription. Die Analyse dieser Daten ist kompliziert (und kann sogar eine eigene Dissertation sein), aber durch Zusammenarbeit mit anderen Forschern habe ich die Möglichkeit gehabt, diese Daten sowohl als Input für meinen Algorithmus (im Falle von der Vorhersage von Bindungsstellen für Transkriptionsfaktoren) als auch zur Validierung vorgeschlagener Modelle (im Falle vom tRNA Pairing Index) zu verwenden.

Diese Arbeit entschlüsselt natürlich nicht alle Transkriptions- und Translationsmecha-

nismen. Ich glaube aber, mindestens teilweise zu neuen biologischen Informationen und zu neuen Methoden für die Analyse von Genomdaten beigetragen zu haben.