Diss. ETH No. 21361

Alignment of genomic sequences with intrinsic disorder and tandem repeats

A dissertation submitted to
ETH Zurich

for the degree of
Doctor of Sciences

presented by
Adam Mieczyslaw Szalkowski
Dipl. Informatik
Universität Karlsruhe (TH), Germany

born December 14, 1981
citizen of Germany

accepted on the recommendation of
Prof. Dr. Gaston H. Gonnet, examiner
Dr. Maria Anisimova, co-examiner
Prof. Erich Bornberg-Bauer, co-examiner
Prof. Alexandros Stamatakis, co-examiner

2013

Abstract

A prerequisite to most evolutionary analyses is a reliable prediction of homologous relationships in a given set of genomic sequences. Homologous genes or characters can be predicted computationally by either pairwise or multiple sequence alignment algorithms based on statistical models of sequence evolution. Although alignment methods have been being developed for multiple decades, for certain sequence features still improved algorithms are required. Here, we focus on new models and algorithms for sequences with intrinsic disorder and tandem repeats, which have been rarely dealt with explicitly in sequence alignment although they are present in $> 30\%$ of eukaryotic proteins.

Although intrinsically disordered protein regions (IDRs) do not have a well-defined tertiary structure, proteins with IDRs perform a multitude of functions, often relying on native disorder to achieve the necessary binding flexibility and strength. IDRs are frequently found in all three kingdoms of life, and may occur in short stretches or span whole proteins. IDRs are difficult to align as they usually evolve faster and have a biased amino acid composition. We build a database of homologous sequences containing intrinsic disorder and estimate a Markov model of amino acid evolution separately for ordered and disordered regions. This model can be used in e.g. sequence alignment and phylogenetic reconstruction and it gives insights into the evolution of IDRs, contrasting patterns specific to ordered protein regions and the corresponding IDRs. As an alternative to the DO model we evaluate the performance of context-specific profiles, a method originally proposed for pairwise homology detection, and extend its definition to multiple sequence alignment. As a result the multiple alignment of sequences with IDRs is improved with only linear time complexity with respect to sequence length.

Tandem repeats (TRs) are often present in proteins with crucial functions, responsible for resistance, pathogenicity, and associated with infectious or neurodegenerative diseases. This motivates numerous studies of TRs and their evolution, requiring accurate multiple sequence alignment. We develop a multiple sequence alignment algorithm which takes into account that TR units may be lost or inserted at any position of a TR region by replication slippage or recombination. By explicitly modeling and reconstructing these events we improve alignment quality and are able to study TR evolution e.g. by estimating TR indel frequencies in different clades of a phylogeny. The algorithm is implemented in a graph-based multiple sequence alignment framework, which produces phylogenetically sensible gap patterns while maintaining robustness by allowing alternative splicings and errors in the branching pattern of the guide tree.

Overall, this thesis contributes to faster and more accurate alignments of sequences with intrinsic disorder and tandem repeats, as we demonstrate on real as well as simulated data.

Zusammenfassung

Für die meisten evolutionären Analysen ist eine verlässliche Bestimmung von Homologieverhältnissen innerhalb der gegebenen Genomsequenzen vonnöten. Die Homologie von ganzen Genen oder auch einzelnen Zeichen kann rechnerisch mithilfe des paarweisen oder multiplen Sequenzalignments (MSA) vorhergesagt werden, welche auf statistischen Modellen der molekularen Evolution basieren. Obwohl Alignmentmethoden schon seit mehreren Jahrzehnten erforscht werden, besteht für bestimmte Sequenzmerkmale noch immer Bedarf nach spezifischen Algorithmen. In dieser Arbeit liegt der Schwerpunkt auf neuen Modellen und Algorithmen für Sequenzen mit intrinsischer Strukturlosigkeit und Tandemwiederholungen. Obwohl diese in $> 30\%$ eukaryotischer Proteine auftreten, wurden sie im Sequenzalignment nur selten explizit behandelt.

Trotz einer fehlenden eindeutigen tertiären Struktur, üben intrinsisch unstrukturierte Proteinregionen (IDRs) verschiedene Funktionen aus. Hierbei verleiht ihnen ihre natürliche Strukturlosigkeit oft die nötige Bindungsflexibilität und -stärke. IDRs treten häufig in allen drei Domänen (Bakterien, Archaeen und Eukaryoten) auf, und können kurze Abschnitte oder auch ganze Proteine umfassen. Desweiteren sind sie schwierig zu alignieren, da sie meist schneller mutieren und eine veränderte Aminosäuren-Zusammensetzung haben. Wir sammeln Gruppen homologer Sequenzen mit IDRs und bestimmen daraus separate Markov Modelle für strukturierte und unstrukturierte Regionen, welche die Evolution von Aminosäuren beschreiben. Diese Modelle können z.B. im Sequenzalignment oder für phylogenetische Rekonstruktion eingesetzt werden. Gleichzeitig gewähren sie Einblick in die Evolution von IDRs, indem evolutionäre Muster strukturierter Regionen mit denen intrinsisch unstrukturierter verglichen werden können. Als Alternative zu diesem DO Modell evaluieren wir kontext-spezifische Profile, eine Methode, die ursprünglich für das Alignment und die Homologieerkennung von entfernt verwandten Sequenzen entwickelt worden ist. Desweiteren passen wir die Methode für den Einsatz im multiplen Sequenzalignment an. Der dafür benötigte Mehraufwand hat nur lineare Komplexität bezogen auf die Sequenzlänge. Dank dieser Erweiterung konnte die Alignmentgenauigkeit von IDRs gesteigert werden.

Tandemwiederholungen (TRs) treten häufig in Proteinen mit lebenswichtigen Funktionen auf, welche verantwortlich sind für Resistenz und Pathogenität. Desweiteren werden sie mit Infektions- und neurodegenerativen Krankheiten in Zusammenhang gebracht. Dies begründet zahlreiche Studien, die sich mit TRs befassen, welche fehlerfreie MSAs benötigen. Wir entwickeln einen MSA Algorithmus, welcher Einfügungen und Verluste (Indels) ganzer TR Einheiten berücksichtigt, die an jeder Stelle von TR Regionen auftreten können, bedingt durch Rekombination oder das Verrutschen der DNA-Polymerase. Dadurch, dass diese Ereignisse explizit modelliert und rekonstruiert werden, verbessert sich die Alignmentgenauigkeit und wir sind in der Lage die Evolution von TRs zu studieren indem wir z.B. die Häufigkeiten von TR Indels in verschiedenen Stämmen eines Stammbaumes bestimmen. Der Algorithmus ist in einem Graph-basierten MSA-Programm implementiert, welches phylogenetisch sinnvolle Leerstellen (Gaps) einfügt. Trotzdem besitzt es die nötige Robustheit alternativ gespleißte Regionen und Fehler im gegebenen phylogenetischen Baum zu tolerieren.

Zusammenfassend trägt diese Arbeit zu schnellerem und genauerem Alignment von Sequenzen mit intrinsischer Strukturlosigkeit und Tandemwiederholungen bei, wie wir sowohl auf echten als auch auf simulierten Daten demonstrieren.