Algorithms and hardness results for DNA physical mapping, protein identification, and related combinatorial problems

Author(s):
Cieliebak, Mark

Publication Date:
2003

Permanent Link:

https://doi.org/10.3929/ethz-a-004651579 →

Diss. ETH No. 15258, 2003

Algorithms and Hardness Results for DNA Physical Mapping, Protein Identification, and Related Combinatorial Problems

A dissertation submitted to the
Swiss Federal Institute of Technology, ETH Zurich
for the degree of Doctor of Technical Sciences

presented by
Mark Cieliebak
Dipl. Informatiker, University of Dortmund
born February 28, 1970 in Hagen, Germany

accepted on the recommendation of
Prof. Dr. Peter Widmayer, examiner
Prof. Dr. Thomas Erlebach, co-examiner

Abstract

In this thesis, we focus on two applications of digestion experiments, namely *physical mapping of DNA* and *protein identification*, and study the computational complexity of combinatorial problems that arise in this context.

Digestion experiments play an important role in molecular biology. In such experiments, enzymes are used to cleave DNA molecules or proteins at specific sequence patterns, the *restriction sites*. The resulting fragments are used in many different ways to study the structure of DNA and proteins, respectively.

In the Double Digest problem, we are given the lengths of DNA fragments arising from digestion experiments with two enzymes, and we want to find a physical map of the DNA, i.e., the positions of the restriction sites of the enzymes along the DNA sequence. Double Digest is known to be NP-hard. We show that the problem is even strongly NP-hard, even if the two enzymes always cut at disjoint restriction sites. Moreover, we show that for partial cleavage errors the problem to find solutions with a minimum number of errors is hard to approximate.

In the Partial Digest problem, we are given DNA fragment lengths arising from digestion experiments with only one enzyme, and we again ask for a physical map of the DNA. Neither a proof of NP-hardness nor a polynomial–time algorithm is known for Partial Digest. We study variations of Partial Digest that model missing fragments, additional fragments, and erroneous fragment lengths, and show that these variations are NP-hard, hard to approximate, and strongly NP-hard, respectively.

The Equal Sum Subsets problem, where we are given a set of positive integers and we ask for two subsets such that their elements add up to the same total, is known to be NP-hard. Equal Sum Subsets can be used to prove NP-hardness for Partial Digest variants. Motivated by this, we study variations of Equal Sum Subsets, where we, for instance, allow any positive rational factor between the sums of the two subsets. We give (pseudo–)polynomial algorithms or (strong) NP-hardness proofs,

respectively, for several natural variations of Equal Sum Subsets.

In the second part of this thesis, we address the problem of protein identification. The *mass fingerprint* of a protein contains the masses of fragments that emerge when digesting the protein. Mass fingerprints are used, for instance, to search for proteins in large protein databases, without sequencing them. The Mass Finding problem arises in this context. Here, we are given a mass M and a protein sequence, and we ask whether there is a fragment of the protein that has mass M. Mass Finding can be solved easily in time linear in n, the length of the protein sequence. We present an algorithm that solves the problem even in sublinear time $O(\frac{n}{\log n})$. This algorithm uses a data structure that is generated in a preprocessing step, and that requires only linear storage space.

A different approach to identifying a protein is to establish its amino acid sequence (*de novo sequencing*). Here, a fragment of the protein (*peptide*) is dissociated, and the masses of the resulting pieces are measured using tandem mass spectrometry. This yields an *MS/MS spectrum* of the peptide. For the case of error–free data, algorithms exist that construct the amino acid sequence of a peptide from its MS/MS spectrum. We have implemented a software tool (Audens) that allows for de novo peptide sequencing even in the case of erronoeous data, and evaluated its performance on real–life spectra.

One problem that arises in the context of Audens is the Decomposition problem, where we ask whether a given mass can be represented as a sum of amino acid masses. This problem is known to be NP-hard. We show that Decomposition can be solved in polynomial time if the number of different amino acid masses is constant, or if the masses of all but a constant number of amino acids are polynomially bounded. On the other hand, we show that if we ask for the *minimum* or *maximum* number of amino acids whose masses add up to the given mass, then no polynomial–time algorithm can guarantee any constant approximation ratio (unless $\mathsf{P} = \mathsf{NP}$).

Zusammenfassung

In dieser Arbeit betrachten wir zwei Anwendungen von Verdau-Experimenten (*digestion experiments*): *physikalische Kartierung von DNS–Molekülen* und Identifikation von Proteinen. Wir untersuchen die algorithmische Komplexität von verschiedenen kombinatorischen Problemen, die in diesem Zusammenhang auftreten.

Verdau-Experimente spielen eine wichtige Rolle in der Molekularbiologie. In diesen Experimenten werden Enzyme verwendet, um DNS–Moleküle oder Proteine an bestimmten Sequenzmustern, den *Restriktionsmustern*, aufzuspalten. Die entstehenden Fragmente werden verwendet, um die Struktur der DNS-Moleküle bzw. Proteine zu untersuchen.

Beim Double Digest Problem sind die Längen von DNS–Fragmenten aus Verdau-Experimenten mit zwei Enzymen gegeben. Hieraus soll eine physikalische Karte der DNS berechnet werden, die die Positionen in der DNS–Sequenz angibt, an der die Restriktionsmuster der Enzyme auftreten. Das Double Digest Problem ist NP-schwer. Wir zeigen, dass es sogar stark NP-schwer ist, selbst wenn die beiden Enzyme die DNS stets an verschiedenen Positionen aufspalten. Ausserdem zeigen wir, dass Double Digest schwer zu approximieren ist, wenn ein Enzym die DNS an einer Position möglicherweise nicht spaltet, obwohl dort das Restriktionsmuster des Enzyms vorliegt (*partial cleavage error*).

Beim Partial Digest Problem sind Fragment-Längen aus Verdau-Experimenten mit nur einem Enzym gegeben, und wie bei Double Digest soll eine physikalische Karte der DNS berechnet werden. Es ist nicht bekannt, ob Partial Digest polynomiell lösbar oder NP-schwer ist. Wir untersuchen das Problem für die Fälle, dass einige Fragment-Längen in der Eingabe fehlen oder dass zusätzliche Längen vorkommen oder dass die Längen nicht exakt gemessen wurden. Wir zeigen, dass die entsprechenden Varianten von Partial Digest NP-schwer bzw. schwer zu approximieren bzw. stark NP-schwer sind.

Das Equal Sum Subsets Problem, bei dem n natürliche Zahlen ge-

geben sind und wir nach zwei Teilmengen suchen, deren Elemente sich zur selben Summe aufaddieren, tritt im Zusammenhang mit PARTIAL DIGEST auf. EQUAL SUM SUBSETS ist NP-schwer. Wir untersuchen verschiedene Varianten von EQUAL SUM SUBSETS, z.B. wenn ein beliebiger positiver rationaler Faktor zwischen den Summen der beiden Teilmengen erlaubt ist, und geben (pseudo–)polynomielle Algorithmen an oder beweisen, dass sie (stark) NP-schwer sind.

Im zweiten Teil dieser Arbeit beschäftigen wir uns mit der Identifikation von Proteinen. Der *Fingerabdruck* eines Proteins enthält die Massen von Protein-Fragmenten, die beim Verdauen des Proteins entstehen. Fingerabdrücke werden z.B. verwendet, um ein Protein in einer Proteindatenbank zu suchen. In diesem Zusammenhang tritt das MASS FINDING Problem auf, bei dem eine Masse M und eine Proteinsequenz gegeben sind und entschieden werden soll, ob das Protein ein Fragment der Masse M enthält. Das MASS FINDING Problem kann in Zeit linear in n, der Länge der Proteinsequenz, gelöst werden. Wir präsentieren einen Algorithmus mit sublinearer Laufzeit $O(\frac{n}{\log n})$. Dieser Algorithmus verwendet eine Datenstruktur, die vorab berechnet wird und die nur linearen Speicherplatz benötigt.

Proteine können auch identifiziert werden, indem man ihre Aminosäuren-Sequenz bestimmt (*de novo sequencing*). Eine Methode hierfür spaltet zunächst das Protein in Fragmente (*Peptide*) auf. Die Peptide werden dann einzeln weiter zerkleinert, und die Massen der entstehenden Teilstücke werden mittels Massenspektrometrie bestimmt. Dies liefert ein *Tandem–Massenspektrum (MS/MS Spektrum)* für jedes einzelne Peptid. Es existieren effiziente Algorithmen, die aus einem MS/MS Spektrum die Aminosäuren-Sequenz des Peptids berechnen, falls die Daten fehlerfrei sind. Da diese Annahme jedoch i.d.R. auf reale Spektren nicht zutrifft, haben wir ein Sequenzierungs-Programm (Audens) implementiert, das auch Fehler in den Daten zulässt, und seine Qualität anhand von realen Spektren evaluiert.

Ein Problem, das im Zusammenhang mit Audens auftaucht, ist das DECOMPOSITION Problem, bei dem entschieden werden soll, ob eine gegebene Zahl sich als Summe von Aminosäuren-Massen darstellen lässt. Dieses Problem ist NP-schwer. Wir zeigen, dass das Problem in polynomieller Zeit lösbar ist, wenn die Anzahl der verschiedenen Aminosäuren-Massen konstant ist oder wenn es nur konstant viele Aminosäuren gibt, deren Masse nicht polynomiell beschränkt ist. Ausserdem betrachten wir die beiden Optimierungsvarianten, bei denen wir nach einer maximalen bzw. minimalen Anzahl von Aminosäuren fragen, deren Massen sich zu einer bestimmten Zahl aufsummieren. Wir zeigen, dass kein polymieller Algorithmus für diese beiden Optimierungsprobleme existiert, der einen konstanten Approximationsfaktor garaniert (falls $\mathsf{P} \neq \mathsf{NP}$).