BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 4 2008, pages 591–593
doi:10.1093/bioinformatics/btm631

Gene expression

DECOMP—from interpreting Mass Spectrometry peaks to solving the Money Changing Problem

Sebastian Böcker[1,*], Zsuzsanna Lipták[2], Marcel Martin[2], Anton Pervukhin[1,†] and Henner Sudek[2]

[1]Lehrstuhl Bioinformatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, 07743 Jena and
[2]AG Genominformatik, Technische Fakultät, Universität Bielefeld, PF 100 131, 33501 Bielefeld, Germany

Received on August 13, 2007; revised on November 16, 2007; accepted on December 17, 2007

Advance Access publication January 2, 2008

Associate Editor: John Quackenbush

ABSTRACT

Summary: We introduce DECOMP, a tool that computes the sum formula of all molecules whose mass equals the input mass. This problem arises frequently in biochemistry and mass spectrometry (MS), when we know the molecular mass of a protein, DNA or metabolite fragment but have no other information. A closely related problem is known as the Money Changing Problem (MCP), where all masses are positive integers. Recently, efficient algorithms have been developed for the MCP, in which DECOMP applies to real-valued MS data. The excellent performance of this method on proteomic and metabolomic MS data has recently been demonstrated. DECOMP has an easy-to-use graphical interface, which caters for both types of users: those interested in solving MCP instances and those submitting MS data.

Availability: DECOMP is freely accessible at http://bibiserv.techfak.uni-bielefeld.de/decomp/

Contact: anton.pervukhin@minet.uni-jena.de

1 INTRODUCTION

Suppose you are given a DNA fragment of mass 2650 ± 1 Da and have no other information. What nucleotide combinations exist with this mass? In fact, there exist two such combinations: either 8 Guanines (totaling 2632.42 Da, plus 18.01 Da for a water molecule) or 7 Cytosines and 2 Thymines (2631.42 Da plus 18.01 Da). DECOMP solves this and similar problems efficiently.

Formally, the MASS DECOMPOSITION PROBLEM can be stated as follows: Given an ordered set of positive real numbers $(a_1, \ldots, a_k)$ (a weighted alphabet), an error bound ϵ, and a query M, we search for all non-negative integer vectors $(c_1, \ldots, c_k)$ such that $M - \epsilon \leq c_1 a_1 + \ldots + c_k a_k \leq M + \epsilon$. If the a_i's are positive integers and $\epsilon = 0$, this is known as the Money Changing Problem (MCP) or Coin Change Problem. In biochemical mass spectrometry (MS) applications, the alphabet corresponds to the molecular masses of the 20 amino acid residues (for protein fragments); of the four nucleotides (for DNA); or of the elements that are expected to occur (e.g. CHNOPS for most metabolites). The query M is the mass of the sample molecule.

A considerable amount of biochemical and mass spectrometry literature exists on the problem of determining the sum formula of a sample molecule from its mass, see for instance (Bertrand *et al*., 1987; Fürst *et al*., 1989; Pomerantz *et al*., 1993), and there are approaches that use known sum formulas for interpreting a mass spectrum (Grange *et al*., 2006; Kind and Fiehn, 2006; Spengler, 2004). There exist software packages to compute these sum formulas, for example Seth (http://www.zebra-crossing.de/software/), ElComp (http://medlib.med.utah.edu/masspec/), HiRes MS (http://homepage.sunrise.ch/mysunrise/joerg.hau/sci/), Elemental Composition Calculator (http://www.wsearch.com.au/) and MF finder (http://www.chemcalc.org/). To the best of our knowledge, all of these packages use exhaustive search to decompose the input mass. Since exhaustive search checks all potential solutions up to the input mass, it will slow down significantly when the input mass increases. To this end, note that there exist 1.9×10^{10} sum formulas with mass up to 2000 Da over the amino acid alphabet.

In addition, some of these packages, e.g. Seth or Elemental Composition Calculator, are available for one operating system only, while others are restricted to one type of alphabet, for example to the molecular masses of elements (HiRes MS, MF finder).

2 METHODS

The Money Changing Problem can be solved with a simple dynamic programming algorithm (Gilmore and Gomory, 1965). Recently, two of the authors presented new algorithms for solving the MCP and its variants (Böcker and Lipták, 2005a, b). We construct a data structure in which we backtrack in a smart order, ensuring that the runtime is proportional to the number of solutions. Our algorithm's main advantage over the classical DP algorithm is its vastly reduced space requirement, with equal or better runtimes, depending on the alphabet. Clearly, it is far superior to any type of exhaustive search.

For non-integer alphabets, the variables are scaled to integers using some precision $\delta \in \mathbb{R}$. The scaling and the error bound ϵ introduce rounding errors of two types. False positives, which do not lie in the interval $[M - \epsilon, M + \epsilon]$, can be dealt with by a simple consistency

*Authors are listed alphabetically.
†To whom correspondence should be addressed.

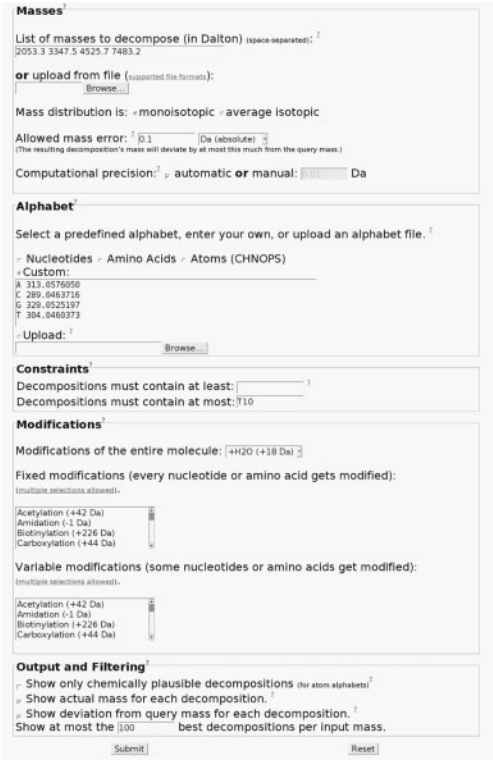

Fig. 1. Submission form for the real-valued mass decomposition problem.

check. False negatives, on the other hand, pose a non-trivial problem; we discuss in detail how to avoid false negatives in Böcker *et al.* (2006).

3 IMPLEMENTATION AND USE

DECOMP's core is written in C++ and runs on the Bielefeld University Bioinformatics Server (BiBiServ). DECOMP can be accessed interactively using a simple web interface. After submission, results are computed on the web server and can be retrieved as a text file.

DECOMP can also be used as a Web Service, which is useful for batch processing and other non-interactive uses. Java source code for an example Web Service client that can be used on the command line will be available for download.

The user can choose between two input forms, depending on whether he wants to solve a real-valued or an integer problem. The submission consists of (1) supplying the query mass or masses, and (2) defining the alphabet. For the real-valued case, we provide predefined alphabets of the common biomolecules (amino acid, nucleotides, CHNOPS) and give the choice between monoisotopic and average values. Alternatively, the user can define his own alphabet or upload it from a file. In addition, the error bound and computation precision can be set. Different formats of MS output, such as dta, mgf and others are supported to upload a query. The user can define modifications or choose from a list of predefined ones, where both modifications of the entire molecule (such as addition of a water molecule) and amino-acid-specific modifications (fixed or variable) are supported. The predefined modifications include those due to sample preparation or the ionization process, as well as common post-translational modifications. Moreover, the user can supply minimum and maximum constraints for each character (e.g. the solution should include at least one C). The output is ranked according to deviation from the query. For the atom alphabet, the sum formulas can be checked for chemical plausibility (Kind and Fiehn, 2007).

For the integer (MCP) case, the user can choose between different mass decomposition problems: compute all solutions (default), compute one solution, compute the number of solutions, or decide whether a solution exists. For a screenshot of the submission form, see Figure 1.

4 CONCLUSION

We have presented DECOMP, a new program to compute decompositions of an input query. Its two primary applications are (1) computing sum formulas of sample molecules from MS spectra, i.e. identifying all molecules with a certain molecular mass from different types of samples: protein, DNA, metabolites or others; and (2) solving instances of the Money Changing Problem. It employs recently developed, very efficient algorithms whose applicability to real-life MS data has been demonstrated (Böcker *et al.*, 2006). DECOMP is supplied with an easy-to-use web interface that allows users to modify all important parameters. Results can be used either independently or as a starting point for further evaluations in the identification pipeline of unknown sample fragments.

DECOMP fills a need for a simple and efficient tool that can quickly compute solutions for mass decomposition problems, both for helping to interpret MS data, and for solving MCP problems. A standalone program, which also allows users to integrate more data such as isotopic distributions, is currently under development.

ACKNOWLEDGEMENTS

The authors thank Henning Mersch and Jan Krüger at BiBiServ for their help in installing DECOMP and creating the DECOMP Web Service, and an anonymous referee for suggesting useful additional features. A.P. supported by Deutsche Forschungsgemeinschaft (BO 1910/1), Zs.L. by Alexander von Humboldt Foundation and the Bundesministerium für Bildung und Forschung, within the group 'Combinatorial Search Algorithms in Bioinformatics'.

Conflict of Interest: none declared.

REFERENCES

Bertrand,M.J. *et al.* (1987) Determination of the empirical formula of peptides by fast atom bombardment mass spectrometry. *Biomed. Environ. Mass Spectrom*, **14**, 249–256.

Böcker,S. and Lipták,Zs. (2005a) Efficient mass decomposition. In *Proceeding of ACM Symposium on Applied Computing (ACM SAC 2005)*. ACM Press, New York, pp. 151–157.

Böcker,S. and Lipták,Zs. (2005b) The Money Changing Problem revisited: computing the Frobenius number in time $O(k\ a_1)$. In *Proceeding of Conference*

on Computing and Combinatorics (COCOON 2005), Vol. 3595 of *Lectures Notes Computer Science*. Springer, Berlin Hedelberg, pp. 965–974.

Böcker,S. *et al*. (2006) Decomposing metabolomic isotope patterns. In *Proceeding of Workshop on Algorithms in Bioinformatics (WABI 2006)*, Vol. 4175 of *Lectures Notes Computer Science*. Springer, Berlin Heidelberg, pp. 12–23.

Fürst,A. *et al*. (1989) A computer program for the computation of the molecular formula. *Chemom. Intell. Lab. Syst.*, **5**, 329–334.

Gilmore,P.C. and Gomory,R.E. (1965) Multi-stage cutting stock problems of two and more dimensions. *Oper. Res.*, **13**, 94–120.

Grange,A. H. *et al*. (2006) Determination of ion and neutral loss compositions and deconvolution of product ion mass spectra using an orthogonal acceleration time-of-flight mass spectrometer and an ion correlation program. *Rapid Commun. Mass Spectrom.*, **20**, 89–102.

Kind,T. and Fiehn,O. (2006) Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, **7**, 234.

Kind,T. and Fiehn,O. (2007) Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, **8**, 105.

Pomerantz,S.C. *et al*. (1993) Determination of oligonucleotide composition from mass spectromectrically measured molecular weight. *J. Am. Soc. Mass Spectrom.*, **4**, 204–209.

Spengler,B. (2004) De novo sequencing, peptide composition analysis, and composition-based sequencing: a new strategy employing accurate mass determination by Fourier transform ion cyclotron resonance mass spectrometry. *J. Am. Soc. Mass Spectrom.*, **15**, 703–714.