

Incremental Discovery of the Irredundant Motif Bases for all Suffixes of a String in $O(|\Sigma|n^2 \log n)$ Time [★]

Alberto Apostolico ^{a,b,1} Claudia Tagliacollo ^{a,2}

^a*Accademia Nazionale dei Lincei, Rome, Italy*

^b*College of Computing, Georgia Institute of Technology, 801 Atlantic Drive, Atlanta, GA 30318, USA. axa@cc.gatech.edu*

Abstract

Compact bases formed by motifs called “irredundant” and capable of generating all other motifs in a sequence have been proposed in recent years and successfully tested in tasks of biosequence analysis and classification. Given a sequence s of n characters drawn from an alphabet Σ , the problem of extracting such a base from s had been previously solved in time $O(n^2 \log n \log |\Sigma|)$ and $O(|\Sigma|n^2 \log^2 n \log \log n)$, respectively, through resort to the FFT-based string searching by Fischer and Paterson. More recently, a solution taking time $O(|\Sigma|n^2)$ without resort to the FFT was also proposed. In the present paper, the problem is considered of extracting the bases of all suffixes of a string incrementally. This problem was solved in previous work in time $O(n^3)$. A much faster incremental algorithm is described here, which takes time $O(|\Sigma|n^2 \log n)$. Whereas also this algorithm does not make use of the FFT, its performance is comparable to the one exhibited by the previous FFT-based algorithms computing only one base. The implicit representation of a single base requires $O(n)$ space, whence for finite alphabets the proposed solution is within a $\log n$ factor from optimality.

Key words: Design and Analysis of Algorithms, Pattern Matching, Motif Discovery, Irredundant Motif, Base.

[★] An extended abstract related to this work is due to appear in the proceedings of WABI 07.

¹ Work Supported in part by the Italian Ministry of University and Research under the Bi-National Project FIRB RBIN04BYZ7, and by the Research Program of Georgia Tech. Performed in part while visiting the Institute for Mathematical Sciences, National University of Singapore in 2006, and the Shanghai CAS-MPG Partner Institute for Computational Biology between the Chinese Academy of Sciences and the German Max Planck Society in 2007, with support provided by those Institutes.

² Work performed in part while visiting the College of Computing of the Georgia Institute of