Journal of Automata, Languages and Combinatorics **10** (2005) 5/6, 609–626

APPROXIMATE SEEDS OF STRINGS [1]

MANOLIS CHRISTODOULAKIS, COSTAS S. ILIOPOULOS
Department of Computer Science, King's College London
Strand, London WC2R 2LS, UK
e-mail: {manolis,csi}@dcs.kcl.ac.uk

KUNSOO PARK[2]
School of Computer Science and Engineering, Seoul National University
Seoul, Korea
e-mail: kpark@theory.snu.ac.kr

and

JEONG SEOP SIM
School of Computer Science and Engineering, Inha University,
Inha, Korea
e-mail: jssim@inha.ac.kr

ABSTRACT

In this paper we study approximate seeds of strings, that is, substrings of a given string x that cover (by concatenations or overlaps) a superstring of x, under a variety of *distance* rules (the Hamming distance, the edit distance, and the weighted edit distance). We solve the *smallest distance approximate seed* problem and the *restricted smallest approximate seed* problem in polynomial time and we prove that the general *smallest approximate seed* problem is NP-complete.

Keywords: Regularities, approximate seeds, Hamming distance, edit distance, weighted edit distance

1. Introduction

Finding *regularities* in strings is useful in a wide area of applications which involve string manipulations. Molecular biology, data compression and computer-assisted music analysis are classic examples. By regularities we mean repeated strings of an approximate nature. Examples of regularities include repetitions, periods, covers and seeds. Regularities in strings have been studied widely the last 20 years.

[1] Full version of a submission presented at the *Prague Stringology Conference* (Czech Technical University in Prague, Czech Republic, September 22–24, 2003).

[2] Work supported by IMT 2000 Project AB02, MOST grant M1-0309-06-0003, and Royal Society grant.