Approximate Runs - Revisited

Gad M. Landau[1,2]

[1] Department of Computer Science,
University of Haifa, Haifa, Israel
landau@cs.haifa.ac.il
[2] Department of Computer and Information Science,
Polytechnic University, New York, USA
landau@poly.edu

Abstract. The problem of finding repeats within a string is an important computational problem with applications in data compression and in the field of molecular biology. Both exact and inexact repeats occur frequently in the genome, and certain repeats are known to be related to human diseases.

A multiple tandem repeat in a sequence S is a (periodic) substring r of S of the form $r = u^a u'$, where u (the period) is a prefix of r, u' is a prefix of u and $a \geq 2$. A run is a maximal (non-extendable) multiple tandem repeat. An *approximate* run is a run with errors (i.e. the repeated subsequences are similar but not identical).

Many measures have been proposed that capture the similarity among all periods. We may measure the number of errors between consecutive periods, between all periods, or between each period and a consensus string. Another possible measure is the number of positions in the periods that may differ.

In this talk I will survey a range of our results in this area. Various parts of this work are joint work with Maxime Crochemore, Gene Myers, Jeanette Schmidt and Dina Sokol.

A. Amir, A. Turpin, and A. Moffat (Eds.): SPIRE 2008, LNCS 5280, p. 2, 2008.