Identifying Periodic Occurrences of a Template with Applications to Protein Structure

Vincent A. Fischetti¹, Gad M. Landau²*, Jeanette P. Schmidt²**, and Peter H. Sellers¹

¹ Rockefeller University, 1230 York Avenue, New York, NY 10021.
² Polytechnic University, 333 Jay Street, Brooklyn, NY 11201.

Abstract. We consider a string matching problem where the pattern is a template that matches many different strings with various degrees of perfection. The quality of a match is given by a penalty matrix that assigns each pair of characters a score that characterizes how well the characters match. Superfluous characters in the text and superfluous characters in the pattern may also occur and the respective penalties for such gaps in the alignment are also given by the penalty matrix. For a text T of length n, and a template P of length m, we wish to find the best alignment of T with P', which is the concatenation of n copies of P, (m will typically be much smaller than n). Such an alignment can simply be obtained by solving a dynamic programming problem of size \(O(n^2m)\), and ignoring the periodic character of \(P'\). We show that the structure of \(P'\) can be exploited and the problem reduced to essentially solving a dynamic programming of size \(O(rnm)\). If the complexity of computing gap penalties is \(O(1)\), (which is frequently the case), our algorithm runs in \(O(mn)\) time. The problem was motivated by a protein structure problem.

1 Introduction

String matching and its many generalizations is a widely studied problem in computer science. One possible generalization that has been researched is approximate string matching - finding occurrences of a pattern in a text where differences (insertions and deletions) are allowed and matches may be defined by a function, with values in some range \((-r, r)\), which specifies how well a character from the pattern "matches" a given character in the text. (Positive values indicate "favorable matches", while negative values indicate "unfavorable matches").

Given a text T of length n, and a pattern P of length m, in the exact string matching problem one finds all the locations \(t_i\) in the text such that \(P = t_i t_{i+1} \ldots t_{i+m-1}\). When differences are allowed, however, every location in the text matches the pattern with some differences. A clarification of the definition of a "match" of the pattern is therefore needed. Most known algorithms for approximate string matching have two basic steps. In the first step each substring of the text receives a score, which reflects the quality of the match between the pattern and the given substring.

---

* Partially supported by NSF grant CCR-8908286.
** Partially supported by NSF grant CCR-9110255 and the New York State Science and Technology Foundation Center for Advanced Technology.
the second step some locations are chosen as sufficiently "good" matches, worthy
to be listed as "matches" in the output. The decision of what constitutes a "good"
match can be made in many ways: [U-83], [LV-89] and [GP-90] use the edit distance
between the matched strings as a measure (i.e the number of insertions, deletions
and mismatches, which have to be performed to obtain the matched string from
the pattern), and declare every match whose edit distance is below a predetermined
threshold a "good" match. Many other measures have been defined in the literature
[SK-83]. Note that the measure of the quality of a match between the pattern and a
substring of the text depends on the application. In many applications in molecular
biology a penalty table is given. This table gives the penalty value for the deletion
and insertion of each letter of the alphabet, as well as the value for matching any
pair of characters. In the simplest case the score of a match is simply the sum of the
corresponding values in the penalty matrix. In some cases however gaps (successive
insertions or deletions) get penalties that are different from the sum of the penalties
each insertion (deletion). A discussion on this subject is given in [GG-89].

In the approximate string matching problem discussed here, the pattern is not
uniquely defined. Given a pattern $P$, we would like to detect whether $P$ (or a cyclic
rotation of $P$) is periodically repeated in the text. We define $P = P^n$, such that
$P^n$ is built by concatenating $n$ copies of $P$. Our goal can be restated as finding
the best match between a substring of $P$ and a substring of the text. Notice that
the substrings of $P$ can be characterized by a suffix of $P$ followed by $i$ copies of $P$
followed by a prefix of $P$. Since $P^n$ is of length $mn$, it is very likely that the best
match with $T$ will be attained with a much shorter subsequence of $P$, presumably
of length no more than $n$, unless the penalty matrix gives very high rewards for
matches and very minor penalties to gaps. Our algorithm is particularly useful if $m$
is much smaller than $n$, and is designed to find many repetitions of $P$.

Section 2 contains a short description of the motivation of the problem, which
comes from protein structure determination. The exact definition of the problem is
given in Section 3. Section 4 describes a simple dynamic programming algorithm
that solves the problem by computing $O(mn^2)$ values. Section 5 describes the new
algorithm.

## 2 Protein structure motivation

The secondary structure of proteins is critical for their proper function. For the ma-
jority of proteins, the secondary structure can not simply be resolved by examining
the sequence of the amino acids that define the protein. While x-ray crystallography
is the only reliable means available today by which to solve the structure of a protein
molecule, it can not always be used. Some algorithms have been designed to help
estimate the conformation of proteins based on the amino acid sequence, employing
the chemical characteristics of the amino acids and their position within the sequence
([CF-74], [GOR-78]). Structural factors of a molecule are important in the overall
protein conformation. The $\alpha$-helical structure is the most common because of its
ability in stabilizing proteins through short regions of helix-helix packing, referred
to as a coiled-coil interaction ([CP-86], [FP-92]).

The ability to form an $\alpha$-helical coiled-coil structure is based on the presence of
a seven-residue repeat pattern denoted by (a-b-c-d-e-f-g). The first and the fourth