Sequence Database Search Using Jumping Alignments

| Constantin Bannert\(^1\), Marc Rehmsmeier\(^2\), Rainer Spang\(^1\) and Jens Stoye\(^1\) |
| Please visit [http://jali.molgen.mpg.de](http://jali.molgen.mpg.de) for further information. |

\(^1\) Computational Molecular Biology, Max-Planck-Institut für Molekule Genetik
Bösenstraße 73, D-14195 Berlin, Germany ((bannert|spang|stoye)@molgen.mpg.de)

\(^2\) Theoretical Bioinformatics, Deutsches Krebsforschungszentrum
Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany (M.Rehmsmeier@dkfz.de)

---

**Overview**

We present an algorithm for amino acid sequence classification and the detection of remote homologies. The rationale is to exploit vertical and horizontal information of a multiple alignment in a well-balanced manner. Established methods like profile and hidden Markov models (HMMs) focus on vertical information only, saturating the individual columns of the multiple alignment. We also take into account row dependencies. Our setting is the following. For a given, uncharacterized protein “candidate sequence”, we want to find from a database of protein families that family where this sequence is likely to belong to. For each protein family from the database, a multiple alignment is constructed. The candidate sequence is then tested against this alignment by means of a new Jumping Alignment algorithm (Jali). It computes a local alignment of the candidate sequence and the protein family alignment. The method is published in [5].

**Motivation**

The characterization of new sequence data produced by the genome projects requires fast and easy methods of functional annotation. Most proteins with similar function descend from common ancestors. The common function is reflected by conserved regions at the sequence level. Given an uncharacterized candidate protein sequence, the question is whether this sequence fits into one of the known families. The following figure shows a possible local alignment of a candidate sequence (below the dashed line) to a multiple alignment.

---

**Evaluation**

We evaluate our method using the SCOP database [2]. It contains classified protein families. Superfamilies are defined as sets of homologous sequences. Each superfamily consists of one or more less divergent families.

---

**Algorithm and Implementation**

**Algorithm**

The algorithm is based on the dynamic programming paradigm. It can be viewed as an extension of the Smith-Waterman algorithm for aligning pairs of sequences. For every candidate sequence we compute the optimal local jumping alignment score, where hopping between the sequence of the seed alignment is penalized.

**Setting**

Our evaluation procedure was first described by [3]. From a SCOP superfamily, one family is split off, called the excluded subfamily. The remaining sequences form the seed, which is used to construct a multiple alignment. The database is then searched for members of the excluded subfamily.

---

**References**