Poster H-71

A composite benchmark for local gapped alignments and scoring systems based on extended alphabets

Authors:
Lorenzo Cerutti (*Swiss Institute of Bioinformatics*)
Marco Pagni (*Swiss Institute of Bioinformatics*)

Short Abstract: We setup a benchmark to provide a complete evaluation of the scoring system for local alignment algorithms. We describe a general framework to extend standard amino acid alphabet with predicted information, which combined to our benchmark, permits to design new and more performing scoring systems rationally.

Long Abstract:
Motivation:
The Smith-Waterman algorithm was designed in early 1980's to align locally similar regions between two sequences.
Nowadays it remains a fundamental algorithm for more complex processes, for example building multiple sequences alignments.
The resulting gapped alignments depend on the chosen similarity matrix and gap penalties, i.e. the scoring system.
A number of tests have been published to benchmark alignment algorithms and their associated scoring systems.
These tests focus on the ability to detect homologous sequences in databases, or alternatively on the quality of the produced alignments.
However for local gapped alignments, none of these tests is sufficient for a complete evaluation of the scoring systems.

Result 1:
We establish a new composite benchmark for a complete evaluation of a scoring system based on four distinct tests: (1) the statistical properties of the scoring system, (2) the ability to detect homologous sequences in a database search, (3) the quality of the alignments, (4) the boundary correctness of the local alignments.
To our knowledge the last property has not been quantified before: local gapped alignments are often used by biologists to delineate the boundaries of the regions of homology, for example, in determining protein domain boundaries.

When our benchmark is applied to various scoring systems based on BLOSUM matrices, we observe that the commonly used BLOSUM62, with affine gap penalties of 11 and 1, results in the best compromise for the overall performance, although it does not produce the best results in database searching and alignment quality.
We want to emphasize that all the tests of the composite benchmark are required for the fair and complete evaluation of new scoring systems.
Introducing the new boundary test was crucial in the light of the use made of local pairwise alignment for determining the boundaries of biologically relevant domains.

Achieving such composite benchmark is a necessary step toward the rational optimization of new alignment algorithms.

Result 2:
The information content of the primary sequence of a protein can be increased by adding position dependent information to every amino acid of the sequence.
PSI-BLAST, for example, cycles over multiple sequence alignments of homologous regions to enrich the "model" of the initial query sequence by introducing position specific information.
Other methods add predicted features to the amino acid sequence, for example structure information.
We describe a general framework to combine predicted features with the primary sequence using extended alphabets. The new scoring system results from a weighted combination of the 2 original scoring sytems and can be used for a standard Smith-Waterman.
Together with our composite benchmark, this framework allows the optimization of the scoring system parameters rationally.

As a test we extended the representation of protein sequences (alphabet of 20 symbols) with predicted secondary structure (alphabet of 3 symbols), resulting in sequences encoded by an alphabet of 60 symbols.
Using our composite benchmark we show that extended alphabets can increase both database search performance and the quality of local alignments without reducing the performance in the correct boundary detection.

In conclusion we show that we can rationally design new scoring systems with significantly better performance than the ones based on the sole 20 aa symbols.