B-16

ProSeS: Protein Sequence Search based on N-gram Indexing

Mi-Nyeong Hwang, Sul-Ah Ahn, and JinSuk Kim
Center for Computational Biology & Bioinformatics,
Korea Institute of Science and Technology Information (KISTI), Korea

Abstract

We present an N-gram indexing method to retrieve similar protein sequences fast, and comparably precisely. This method regards a protein sequence as a text written in language of 20 amino acid codes, adapts N-gram tokens of fixed-length as its indexing scheme for sequence strings. After such tokens are indexed for all the sequences in the database, sequences can be searched with information retrieval algorithms. We show experimentally that the N-gram indexing approach saves the retrieval time significantly, and that it is comparably as accurate as current popular search tool BLAST.

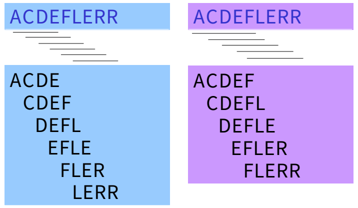
1. Introduction

Popular systems for searching databases match queries to answers by comparing a query to each of the sequences in the databases. Efficiency in such exhaustive systems is not satisfactory, since servers must process many queries simultaneously and solution of each query requires comparison to over the huge sequence databases, for example, BLAST (Basic Local Alignment Search Tool) [1]. BLAST uses a heuristic method to find the highest scoring and locally optimal alignments between a query sequence and sequences in the database. The program has been developed by NCBI (National Center for Biotechnology Information) and benefits from technical supports for strong and continuing refinement. Although BLAST also adopts simple indexing scheme to build sequence databases and to choose candidates from the database, it does not fully utilize indexing features of information retrieval. Furthermore BLAST requires powerful computational facilities of the CPU processors for the reason of its origin in dynamic programming. This leads BLAST to a problem, where many simultaneous users through the Internet do not satisfied with search speed.

2. Method

N-gram based indexing

Protein sequences can be regarded as texts written in language of 20 amino acid codes. ProSeS adopts N-gram tokens of fixed length as its indexing scheme for sequence strings. For example, if cutting interval is 4, which is known as tetra-gram, sequence 'ACDEFLERR' is segmented into 'ACDE', 'CDEF', 'DEFL', 'EFLE', 'FLER', and 'LERR'.



(a) (b)

Figure 1: N-gram indexing example (a) tetra-gram, (b) penta-gram

After the N-gram indexing, N-grams are stored into an inverted index. The inverted file indexing scheme is extended so that within each postings list is stored not only the ordinal sequence number that contains the interval, but also offset information. For example, consider the following posting list

"ACEP" 36(2), 127(3), 1074(1),……

in which the indexed sequences, the 36th sequence contains the interval 'ACEP'. The interval occurs twice in the 36th sequence.

Retrieval based on Vector Space Model

The vector space model has been investigated in depth for information retrieval [2]. ProSeS uses this retrieval method for searching similar sequences. The similarity measure (*Sim*(*q,d*)) between query sequence *q* and target sequence *d* is defined as follows:

$$Sim(q,d) = \frac{1}{W_d} \cdot \sum_{t \in q \wedge d} w_{d,t} w_{q,t}$$

with:

$$W_d = \log(1 + \sum_{t \in d} f_{d,t})$$

$$w_{d,t} = \log(f_{d,t} + 1) \bullet \log(\frac{N}{f_t} + 1)$$

$$w_{q,t} = \log(f_{q,t} + 1) \bullet \log(\frac{N}{f_t} + 1)$$

where $f_{s,t}$ is the frequency of N-gram token *t* in sequence *s*; *N* is the total number of sequences in the data collection; f_t is the number of sequence where token *t* occurs more than or equal to once; $w_{s,t}$ is the weight of term *t* in the query or target sequence *s*; W_d is the normalization factor for the length of target sequence *d*.

3. Experiments and Results

Test collection

To make a comparison of retrieval effectiveness between BLAST and ProSeS, we use the PIR-NREF database [3]. We select randomly a set of 100 sequences out of 1,292,569 sequences with sequence lengths from 50 to 1000. We ran BLASTP against these 100 sequences and generated local alignments for each test sequences up to 1000 entries. This data set was regarded as relevant answers that retrieval effectiveness was measured.

To quantify the relative performance or retrieval effectiveness of ProSeS, we use the relevance-based measures of recall and precision. Recall and precision are frequently used to demonstrate the retrieval effectiveness of systems, particularly those used for information retrieval. Sequences of resultant query set searched by BLAST with filtering option on/off for each test data assigned collection.

Precision is a measure of the fraction of relevant items retrieved in the result set at a particular point, that is

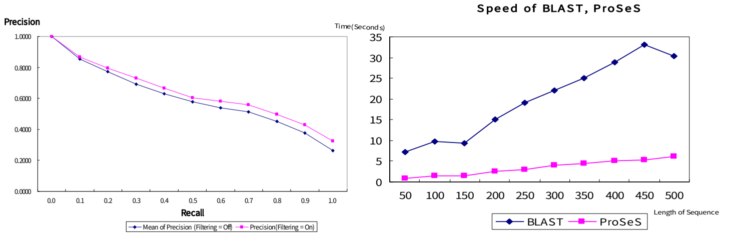
$$P = \frac{\text{\# of relevant items retrieved}}{\text{Total \# of items retrieved}}$$

Recall, in contrast, measures the fraction of the relevant answers retrieved from the relevant answers at a particular point, or

$$R = \frac{\text{\# of relevant items retrieved}}{\text{\# of relevant items in the collection}}$$

Results

Figure 2 (a) shows a mean recall-precision graph for the search results from ProSeS. Results are shown for searching the test collection with our query test set comprised of 100 sequences. It is noted that ProSeS uses an interval length of *n* = 5, penta-gram.



(a) (b)

Figure 2: (a)11pt – average recall-precision (b) Performance report

These results suggest that ProSeS is almost as effective as BLAST in finding homologies at lower recall levels. We show in Figure 2(b) the relative speeds of both BLAST and ProSeS in searching the PIR-NREF database with our 100-query test set, limiting the maximum number of results as 1000. The parameters are the same as previous experiment. This experiment was carried out on a machine with Intel XEON dual CPU 2.4GHz processors and 3GB RAM which is operated with Linux.

4. Application

ProSeS is a protein sequence analysis system which provides overall analysis results such as similar sequences with significant homologies, predicted subcellular locations of the query sequence, and major keywords extracted from annotations of similar sequences [4].

(a) (b)

Figure 3: ProSeS (a) Main Page (b) Result page

5. Acknowledgement & Availability

This work was supported in part by the IBM Korea. The Protein Sequence Search (ProSeS) service is available at http://proses.kisti.re.kr.

6. Reference

[1] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. (1990) Basic local alignment search tool. Journal of Molecular biology, 215: 403-410

[2] V. Raghavan and S. Wong. (1986) A critical analysis of vector space model for information retrieval. Journal of the American Society for Information Science, 37(5), p. 279-87

[3] http://pir.georgetown.edu/pirwww/search/pirnref.shtml

[4] H. E. Park and J. Kim (2003) ProSLP: a novel predictor for subcellular localization based on N-gram Features. Submiited to Journal of bioinformatics, available at http://proslp.kisti.re.kr.