

Improving the GPM's peptide subsequence searching algorithm

Dan Evans (evans.dan@gmail.com)
Biomedical Research Centre
University of British Columbia

Overview

In order to increase the speed of subsequence (or "tag") searches, some changes have been made to the GPM. An index table of residue "words" has been created to enhance tag searches, which has resulted in a large speed increase. Protein subsequence searches that used to take several minutes may now take roughly ten seconds.

The speed increase is due to the creation of a database table, used as an index, called `peptide_words`. This table has one row in it for each four-residue "word" (e.g., "AAAA", "RUQN", "SPSS", etc.). The second column in this table is the corresponding list of peptide identifiers belonging to sequences which contain that four-residue word. By iteratively comparing the contents of those lists, a candidate set of sequences which may contain the requested subsequence may be found. This is much faster than doing a scan of every stored protein sequence.

The table will be updated automatically once per day when newly added sequence information is processed on the GPM.

Technical Details

The `peptide_words` table contains 331,776 (24^4) rows, corresponding to all possible combinations of 24 letters used as amino acid abbreviations. The only letters not used are J and O; U is used as selenocystine in many sequences and the three placeholder letters are also included. With the new index, searching for a subsequence of N residues will compare the peptide identifier sets of (N - 3) peptide words as opposed to doing a substring match on every one of more than 30 million rows of sequence data.

Given that an average list of peptide identifiers contains D unique identifiers, the number of comparisons required for an N-residue subsequence will be:

$$\# \text{ comparisons} \approx D(N - 3)$$

in the average case. Given that the average peptide sequence is A residues long, the number of comparisons in a straight search of P sequences for an N-residue

subsequence will be:

$$\# \text{ comparisons} \approx P(A - N)$$

in the average case. Since the average number of peptide identifiers per word will always be much smaller than the total number of stored sequences ($D < P$), and because the average number of peptide identifiers will grow much more slowly than the number of peptide sequences, the new approach will always be faster.

In a timing test of 15 peptides, with an average length of 15 residues, a previous search method required approximately six times as long as the the new word-based algorithm.

A Worked Example

1. A user requests a subsequence search for the residues "DAALGS". (Note: the field is case insensitive and will return more results than the example below.)
2. The GPM breaks the subsequence into a set of four sequential words: "DAAL", "AALG" and "ALGS".
3. For each word, the list of all peptide identifiers for sequences which contain that word is retrieved. E.g.,
 1. "DAAL" - (1, 2, 3, 4, 15)
 2. "AALG" - (3, 4, 5, 6, 7, 15, 21)
 3. "ALGS" - (2, 3, 7, 9, 15)
4. The intersection of the sets of identifiers is obtained subtractively. I.e., if the identifier is in the first set but not the second, it is removed from the master list:
 1. Starting with the "DAAL" set, the "AALG" set also contains (3, 4, 15), but not (1, 2), so identifiers 1 and 2 are removed from the set of candidates.
 2. Comparing the remaining candidate set (3, 4, 15) against the "ALGS" set, the candidate set also contains (3, 15), but not (4), so 4 is removed, leaving (3, 15) as the final list of candidates.
5. The sequence corresponding to each peptide identifier in the candidate set is then compared against the requested subsequence. All peptides whose sequence contains the full subsequence is counted as a hit and returned.

Due to the subtractive nature of the peptide identifier list generation, the longer the requested subsequence is, the less time the search may take to complete: each word searched for is an additional discriminatory step the algorithm can take. A four-residue search may take longer to complete as the search for a six residue subsequence.