

GenericBioMatch: A novel generic pattern match algorithm for biological sequences

Youlian Pan and A. Fazel Famili
Integrated Reasoning Group, Institute for Information Technology,
National Research Council Canada,
1200 Montreal Road, M-50, Ottawa, ON K1A 0R6, Canada
youlian.pan@nrc-cnrc.gc.ca or fazel.famili@nrc-cnrc.gc.ca

Abstract

GenericBioMatch is a novel algorithm for exact match in biological sequences. It allows the sequence motif pattern to contain one or more wild card letters (eg. Y, R, W in DNA sequences) and one or more gaps of any number of bases. GenericBioMatch is a relatively fast algorithm as compared to probabilistic algorithms, and has very little computational overhead. It is able to perform exact match of protein motifs as well as DNA motifs. This algorithm can serve as a quick validation tool for implementation of other algorithms, and can also serve as a supporting tool for probabilistic algorithms in order to reduce computational overhead. This algorithm has been implemented in the BioMiner software (http://iit-iti.nrc-cnrc.gc.ca/biomine_e.trx), a suite of java tools for integrated data mining in genomics. It has been tested successfully with DNA sequences from human, yeast, and Arabidopsis.

1. Introduction

The grammar of biological sequence language varies markedly from any of the natural languages. The schema of DNA and protein sequences is less strict than that of natural languages, such that it allows deletion and insertion of letters in a sequence pattern. It also allows alternation of letters within a “word”. Attempts for readily use of text-searching algorithms, such as Boyer-Moore method[1] and Knuth-Morris-Pratt algorithm[2], in biological sequences have been tested unsuccessfully. In the past decades, various modifications of these early algorithms became available (e.g. [4] [6]); many probabilistic algorithms, such as Hidden Markov Model (e.g [3]), Gibbs Sampling (e.g.[3]), are being applied to biological sequence motif finding. However, there are very few algorithms that are suitable for quick, exact

pattern match in biological sequences. We report a new algorithm that is suitable for a quick pattern match in biological sequence search.

2. The algorithm

The algorithm consists of two components, one for match of individual bases and the other for match of the entire motif pattern. The base match component is sequence specific and differs among DNA, RNA and protein sequences. However, the motif match component is generic for DNA, RNA and protein.

2.1. Base match

The main cause of the unsuccessfulness in application of earlier text match algorithms in biological sequences is the alternation of bases in some motif pattern. To overcome such deficiency, baseMatch (Fig. 1) is designed to cope with inclusiveness of a wild card letter in a motif pattern. For example in DNA sequence, letter “B” include “G”, “C”, and “T”.

```
boolean baseMatch(motifChar, seqBase)
  if(motifChar includes seqBase)
    return true;
  else
    return false;
```

Figure 1. Base Match component of the algorithm.

2.2. Motif match

The other deficiency in application of the earlier text match algorithms in biological sequences is that some motif patterns allow insertion, represented by “-”, at

certain location. This is compensated in GenericBioMatch by looking for current or subsequent base that matches with the base after the insertion position.

```

Motif[] GenericBioMatch
Motif[] matches;
int matched = 0;
int sequenceLength;
int motifLength;
int startIndex = 1;
// j current sequence index
// k current motif index
int j = 1;
int k = 1;
while(j < SequenceLength)
    if(baseMatch(k, j))
        j++;
        k++;
        if(k > patLength)
            matches[matched++] =
                newMotif(motifFound, j);
            startIndex++;
            j = startIndex;
            k = 1;
// start search for next match;
elseif(isInsert(k))
    if(baseMatches(k+1, j+1))
        j++;
        k++;
    elseif(baseMatch(k+1, j))
        k++;
    else
        j++;
else
    startIndex++;
    j = startIndex;
    k = 1;
return matches;

```

Figure 2. The GenericBioMatch algorithm

3. Implementation and application

We have implemented this algorithm in Java in the BioMiner software, a suit of tools for data mining in genomics, and have been tested successfully with DNA sequences from human, vertebrate and plant consensuses (most of them contain wild card letters), yeast, and arabidopsis.

A comparison with Boyer-Moore algorithm (B-M)[1], the fastest text search algorithm to date, has been performed. In DNA sequences, B-M does not seem to be much advantageous over GenericBioMatch because the sequences consist of only four alphabets. However, B-M

and Knuth-Morris-Pratt algorithm[2] are not able to search a motif pattern that has a wild card letter or an insertion symbol from a sequence. A combination of either of the text search algorithms with current baseMatch component has been tested unsuccessfully. Further investigation is in progress.

4. Future direction of GenericBioMatch

GenericBioMatch is expected to be useful for DNA RNA and protein sequences, but tests are yet to be performed for RNA and protein.

GenericBioMatch is able to examine the existence of motif patterns that conformed to a consensus and extract such motifs. Therefore, it can be used as a preprocessing tool that acquires all "legal" motifs from the sequences in question with regard to the consensus, and then feed them to Hidden Markov Models and other probabilistic models. This will significantly reduce the computational overhead of those probabilistic algorithms. We have initiated such research. More tests are currently in progress.

5. Acknowledgement

This is publication NRC 45835 of National Research Council of Canada.

6. References

- [1] Boyer R. S. and J. S. Moore, "A fast string search algorithm", *Communications of the ACM* 20: 262-272, 1977.
- [2] Knuth D. E., H.H. Morris Jr. and V. R. Pratt, "Fast Pattern Matching in Strings", *SIAM Journal of Computing* 6(2): 323-350, 1977.
- [3] Krogh A, "An introduction to Hidden Markov Models for biological sequences", In *Computational Methods in Molecular Biology*, eds S. L. Salzberg, D.B. Searls and S. Kasif, Elsevier, Amsterdam. pp: 45-63, 1977.
- [4] Lefevre C. and I.E. Ikeda, "Pattern recognition in DNA sequences and its application to consensus foot-printing", *CABIOS* 9: 349-354, 1993.
- [5] Nussinov R., "Efficient algorithms for searching for exact repetition of nucleotide sequences", *J Mol Evol* 19:283-285, 1983.
- [6] Prunella N, S Liuni, M Attimonelli and G Pesole, "FASTPAT: a fast and efficient algorithm for string searching in DNA sequences", *CABIOS* 9: 541-545, 1993.
- [7] Thijs G., K. Marchal, M. Lescot, S. Rombauts, B. De Moor, P. Rouze and Y. Moreau, "A Gibbs Sampling Method to Detect overrepresented motifs in the upstream regions of coexpressed genes", *J. Comput. Biol.* 9: 447-464, 2002.