BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 12 2001
Pages 1234–1235

mmsearch: a motif arrangement language and search program

Thomas Junier, Marco Pagni and Philipp Bucher

Swiss Institute of Bioinformatics, Switzerland

Received on April 5, 2001; revised on June 8, 2001; accepted on July 2, 2001

ABSTRACT

Summary: This paper presents a language for describing arrangements of motifs in biological sequences, and a program that uses the language to find the arrangements in motif match databases. The program does not by itself search for the constituent motifs, and is thus independent of how they are detected, which allows it to use motif match data of various origins.

Availability: The program can be tested online at http://hits.isb-sib.ch and the distribution is available from ftp://ftp.isrec.isb-sib.ch/pub/software/unix/mmsearch-1.0.tar.gz

Contact: Thomas.Junier@isrec.unil.ch

Supplementary information: The full documentation about `mmsearch` is available from http://hits.isb-sib.ch/~tjunier/mmsearch/doc.

INTRODUCTION

In biological sequences, the joint occurrence of several conserved motifs is often more informative than the presence of a single one, and the arrangement of the motifs along the sequence can carry even more information as to the sequence's function. One way of modelling arrangements is to build a descriptor for the whole region. This is analogous to the concatenation of individual profiles for the constituent motifs, with the intervening regions modelled by states with higher entropy and insertion probability. This approach is used by MetaMEME (Grundy *et al.*, 1997), and a similar one by FingerPRINTScan (Scordis *et al.*, 1999).

The `mmsearch` program differs from these in two ways. First, it does not look for the motifs themselves, but only for arrangements, which the user specifies at runtime. The position of motifs along sequences (or *match data*) is supplied to `mmsearch` in a tab-separated format such as GFF (Durbin, 1997). The advantage of this approach is that it does not depend on a particular motif search algorithm, and that motif match data of diverse origin (including, for example, database annotations) can be freely intermixed. Secondly, and more importantly, the technique can handle motif overlaps and inclusions, which are hard or impossible to handle with HMMs, profiles or similar techniques (see Section **Examples**).

EXAMPLES

Simple arrangements

Most cytokine receptors share two conserved regions: a series of Cys residues near the *N*-terminus, and a W-S-x-W-S pattern near the membrane. Furthermore, all such receptors feature a signal peptide, and most are membrane-bound (Figure 1a). The two conserved regions are modeled by PROSITE (Hofmann *et al.*, 1999) patterns PS00241 and PS00340. These have 20 and 53 false positives in SWISS-PROT release 38, respectively. (see PROSITE documentation entry `PDOC00214`.) These figures can be lowered by running pattern search programs and predictors for signal peptides and transmembrane domains, and by conserving only the proteins which match the following arrangement:

$$\texttt{SIGNAL} = \texttt{PS00241} = \texttt{PS00340} = \texttt{TRANSMEM}. \quad (1)$$

The patterns also miss a number of true matches (false negatives). This can be solved by rewriting the patterns in a less restrictive manner, the arrangement condition keeping out most false positives.

Feature ends

The thioredoxin domain is shared by a variety of proteins, some of which have an active site consisting of a disulfide bridge (Figure 1b). If we wish to select only those proteins that contain a thioredoxin domain with an active site, we could use an expression of the form

$$\texttt{<THIORED} = \texttt{SS} = \texttt{THIORED>} \quad (2)$$

where '<' and '>' mean the *start* and *end* of a motif, respectively; and `SS` stands for a protein region flanked by cysteine residues forming a disulfide bridge. These are handy for dealing with overlaps and inclusions.

Equivalence

There may be more than one predictor for the same motif, for example a PROSITE profile and a Pfam HMM (Bateman *et al.*, 2000) for the same domain. The different predictors rarely agree exactly: they may start or stop a

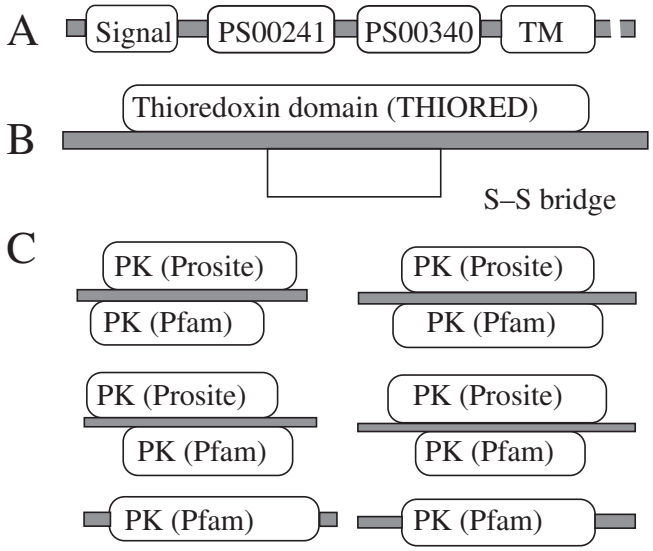

Fig. 1. Examples of domain arrangements that can be located with a metamotif search.

few residues apart, or either one may be absent altogether. Figure 1c presents six different arrangements which in fact represent the same biological fact, i.e. the presence of a protein kinase domain. This is handled by the 'equivalence' operator, in which the possible motifs are separated by pipes ('`|`') and included in square brackets ('`[]`'):

```
[PFAM_PK|PROSITE_PK]                (3)
```

which reads 'a PROSITE PK, or a Pfam PK, or both; and in that case, arrangement is irrelevant'.

Other features

`mmsearch`'s syntax include controlling the distance between motifs, repeating part of a metamotif, alternating between possible arrangements, and anchoring the metamotif at the beginning or end of the sequence. This is explained in detail, along with a complete grammar, in the documentation (see Section **Abstract**).

THE SEARCH PROGRAM

The program generates a state automaton by parsing the user-supplied expression. The match data are converted into a string containing motif names and positions. This string is fed to the automaton, but unlike in pure pattern matching, some state transitions are governed not only by the next input symbol, but by the values of numbers or names parsed from the input string.

The `mmsearch` program was written as a stand-alone, command-line application using the Python language. A Web interface to `mmsearch` is available on the Hits Web site (URL given in the Section **Abstract**) where it can be used with PROSITE and Pfam motifs in the SwissProt, TrEMBL, trGEN and trEST databases (Pagni *et al.*, 2001).

ACKNOWLEDGEMENT

This work was partly supported by grant 3100-49669.96 of the Swiss National Science Foundation.

REFERENCES

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

Durbin,R. (1997) http://www.sanger.ac.uk/Users/rd/gff.shtml.

Grundy,W.N., Bailey,T.L., Elkan,C.P. and Baker,M.E. (1997) Meta-MEME: motif-based hidden Markov models of protein families. *Comput. Appl. Biosci.*, **13**, 397–406.

Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database: its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.

Pagni,M., Iseli,C., Junier,Th., Falquet,L., Jongeneel,V. and Bucher,Ph. (2001) trEST, trGEN and Hits: access to databases of predicted protein sequences. *Nucleic Acids Res.*, **29**, 148–151.

Scordis,P., Flower,D.R. and Attwood,T.K. (1999) FingerPRINTSScan: intelligent searching of the PRINTS motif database. *Bioinformatics*, **15**, 799–806.