INTRODUCTION

Over the past year or so Amos Bairoch (bairoch @cgecmu51.BITNET) has released an number of versions of his Prosite database. This is a database of patterns which have been associated with particular enzymatic activities or structures. For example, the well known pattern for N-link glycosylation Asn-Xxx-Ser/Thr.

Amos has compiled a database that consists of references about each pattern, validity of the patterns, occurrences, and a host of other details. This database is of general use, and has been used by Amos in his PC/Gene Suite of programs for analysis of DNA and Protein sequences.

I wanted to use this database on a Unix machine and be able to ask the question, "Which of these patterns occur in sequence X?"

This is the second release of Prosearch. It completely supersedes the first version with one important bug fix, and support for VMS, MS-DOS, and UNIX. Also, by using ReadSeq, a fine program from Don Gilbert <gilbertd@silver.ucs.indiana.edu>, more protein data formats are accessible.

IMPLEMENTATION

Most patterns can be expressed as regular expressions. For example the pattern '^P' when used with the unix utility grep matches any line in the input that begins with a 'P'.

I translated all but 1 of the 337 patterns in Prosite to Unix style regular expressions and wrote a simple searching program to search a protein sequence for their occurrence. The pattern I did not translate was the pattern PS0003 which is Tyrosine Sulfation. There is no clean pattern for this modification.

The program is written in the Awk language, and runs on machines which have either Nawk from AT&T, Gawk from the Free Software Foundation, or one of several versions of Awk which run on MSDOS compatibles. Read the approriate INSTALL file for details.

INPUT FILES

In put file are any protein sequence files in an unstructured format. AWK will accept the input on any number of lines of any length (I've tried proteins sequences up to 2500 amino acids on one line with no problem). Each ASCII character will be interpreted as an amino acid, and all letters must be capitalized. With 'readseq' any of a number of formats can be used.

OUTPUT

There are two possible forms of output. The "short" form is a table of accession numbers, positions in the sequence and short names for patterns. The "long" form is the same except that the relevant sections from the Prosite Database is also printed.

Here is an example of the short output for Bovine Rhodopsin.

```
Prosite Database -- Release 5.0 of April 1990 Copyright: Amos Bairoch
ProSearch Software -- Release 0.1beta -- Copyright: Lee Kolakowski
The following patterns are in < test.ops >:

Access#     From->To    Name
_______     ________    ____
PS00001         2->6    ASN_GLYCOSYLATION
PS00001       15->19    ASN_GLYCOSYLATION
PS00001     200->204    ASN_GLYCOSYLATION
PS00005       14->17    PKC_PHOSPHO_SITE
PS00005     229->232    PKC_PHOSPHO_SITE
PS00005     243->246    PKC_PHOSPHO_SITE
PS00006       22->26    CK2_PHOSPHO_SITE
PS00006     193->197    CK2_PHOSPHO_SITE
PS00006     198->202    CK2_PHOSPHO_SITE
PS00006     229->233    CK2_PHOSPHO_SITE
PS00006     338->342    CK2_PHOSPHO_SITE
PS00007       21->30    TYR_PHOSPHO_SITE
PS00008       89->95    MYRISTYL
PS00008     120->126    MYRISTYL
PS00008     156->162    MYRISTYL
PS00008     182->188    MYRISTYL
PS00013     157->168    PROKAR_LIPOPROTEIN
PS00237       68->85    G_PROTEIN_RECEPTOR
PS00238     296->314    OPSIN
```

USAGE

```
pros file ...
prosearch file ...
```

BUGS

Please send bug reports or improvements to me.

NOTICES

Frank Kolakowski

lfk@athena.mit.edu	Lee F. Kolakowski
lfk@eastman2.mit.edu	M.I.T.
kolakowski@wccf.mit.edu	Dept of Chemistry
lfk@mbio.med.upenn.edu	Room 18-506
lfk@hx.lcs.mit.edu	77 Massachusetts Ave.
AT&T: 1-617-253-1866	Cambridge, MA 02139