ENTREZ DIRECT - INTRODUCTION

Entrez Direct (EDirect) constructs an Entrez service request from commands typed into a terminal window, taking search terms from command-line arguments and combining operations into multi-step queries with UNIX pipes.

Using EDirect, a biologist can perform sophisticated queries against the NCBI's interconnected Entrez data domains (publication, nucleotide, protein, structure, gene, variation, expression, etc.) without having to write a program.

For example, given a published article, or an interesting topic, EDirect can look up the relevant paper(s) by searching on author names or text words:

```
  esearch -db pubmed -query "Casadaban AND transposition immunity"
```

Linking can find related records within a database:

```
  elink -related
```

and look up associated records between databases:

```
  elink -target protein
```

Results can be refined by further searching in Entrez:

```
  efilter -query "NOT (bacteria [ORGN] OR fungi [ORGN])"
```

Records can be retrieved in a variety of report formats, such as GenPept flatfile:

```
  efetch -format gp
```

or as document summaries in XML format:

```
  esummary
```

The xtract function uses command-line arguments to direct the selective conversion of XML structured data into a tab-delimited table. Piping the summary XML to:

```
  xtract -pattern DocumentSummary -element AccessionVersion Slen
```

extracts sequence accession numbers and sequence lengths into a 2-column table. (Additional arguments can limit data extraction to specified regions of the XML, filter by data content, and customize the table presentation.)

Queries can move seamlessly between EDirect commands and UNIX utilities to perform actions that cannot be accomplished entirely within Entrez. Piping the table to:

```
  sort -k 2,2nr | head -n 10 | cut -f 1 -
```

sorts by sequence length and prints the accessions of the ten longest proteins. These are then posted to Entrez and downloaded as FASTA sequence reports by:

```
  epost -db protein -format acc | efetch -format fasta
```

Tabular data can also be piped to a script or saved to a file for importing into a database or spreadsheet program.

EDirect will run on UNIX and Macintosh computers that have the Perl language installed, and under the Cygwin UNIX-emulation environment on Windows PCs.