ENTREZ DIRECT - EXECUTIVE SUMMARY

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 publication describing a biosynthetic enzyme with unexpected specificity, a scientist might want to find the heaviest functionally equivalent proteins that could be isolated from microorganisms for in vitro studies.

EDirect can look up the paper by searching on author names or text words, link to the protein sequence(s) published in the article, find similar protein sequences (BLAST neighbors), and restrict the results to suitable microbial eukaryotes:

```
  esearch -db pubmed -query "carotene catechol" | \
  elink -target protein | \
  elink -related | \
  efilter -query "yeast [ORGN]" | \
```

Records can be retrieved as standard document summaries or in a variety of report formats, such as an XML structured version of the GenPept flatfile:

```
  efetch -format gpc -mode xml | \
```

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

```
  xtract -pattern INSDSeq -ACCN INSDSeq_accession-version \
    -group INSDFeature -avoid ">proprotein<" -and ">sig_peptide<" \
      -block INSDQualifier -match ">calculated_mol_wt<" \
      -sfx "\n" -tab "" -ret "" -element "&ACCN",INSDQualifier_value | \
```

extracts the sequence accessions and protein molecular weights (including those of mature peptides from multi-product precursor proteins) into a 2-column table.

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 | grep "NP_" | head -n 5 | cut -f 1 - | \
```

sorts by molecular weight and prints the accessions of the five heaviest curated RefSeq proteins. These are then posted to Entrez and displayed in GenPept format:

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

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.