EASY—an Expert Analysis SYstem for interpreting database search outputs

J. N. Selley1,∗, J. Swift2 and T. K. Attwood1

1School of Biological Sciences, The University of Manchester, Oxford Road, Manchester, M13 9PT, UK and 2Zeneca Pharmaceuticals, Mereside, Alderley Edge, Macclesfield, Cheshire, SK10 4TG, UK

Received on March 13, 2000; revised on September 14, 2000; accepted on September 28, 2000

ABSTRACT
Summary: With the ever-increasing need to handle large volumes of sequence data efficiently and reliably, we have developed the EASY system for performing combined protein sequence and pattern database searches. EASY runs searches simultaneously and distils results into a concise 1-line diagnosis. By bringing together results of several different analyses, EASY provides a rapid means of evaluating biological significance, minimising the risk of inferring false relationships, for example from relying exclusively on top BLAST hits. The program has been tested using a variety of protein families and was instrumental in resolving family assignments in a major update of the PRINTS database.

Availability: http://www.bioinf.man.ac.uk/dbbrowser/easy/
Contact: selley@bioinf.man.ac.uk

In sequence analyses, it is now common practice to search both sequence and pattern databases. While some ‘expert systems’ (e.g. ‘GeneQuiz’, Scharf et al., 1994; MAGPIE, Gaasterland and Sensen, 1996; PEDANT, Frishman and Mewes, 1997) perform both types of query, few fully exploit the differences between them. Our aim was to make a range of diverse databases accessible under a single umbrella, to design a system for interpreting results of multiple searches and to facilitate the inference of protein function by exploiting the semantic differences between outputs. Such a system would differ from current tools by virtue not simply of reporting multiple results, but of digesting them and providing a consensus diagnosis in real time. Ultimately, our goal was to make sequence analysis EASY (Expert Analysis SYstem).

EASY is written in Perl and was designed in a modular way to be both extensible and configurable. It currently searches SWISS-PROT (Bairoch and Apweiler, 2000), PROSITE (Hofmann et al., 1999), PRINTS (Attwood et al., 2000), Pfam (Bateman et al., 2000), Profiles (Bucher et al., 1996), BLOCKS and BLOCKS-format PRINTS (Henikoff et al., 2000). SWISS-PROT is searched using BLAST (Altschul et al., 1990) and FastA (Pearson, 1990); pattern databases are searched using the custom-built tools at their host sites using default parameters.

Query submission is via a web form, which accepts a variety of formats and allows users to select which database to search. Searches are run in parallel, so execution time (∼2 min on a dual Pentium II 450 with 1 GB memory) is dependent on the slowest method (usually FastA). Currently, 2 databases are searched locally and all others are queried across the Internet. However, the system can be configured to run all searches locally, or remotely, as desired.

When the outputs are returned, EASY extracts and digests the description lines from each match using rudimentary information retrieval techniques. The system has been designed to recognise the semantic difference between sequence database search outputs, which ‘know’ about proteins, and pattern database outputs, which ‘know’ about families. The program first calculates the most frequently occurring word (disregarding uninformative words like ‘protein’), and then seeks the longest, most commonly occurring phrase. Results are presented in the form of a concise distillation of all outputs, initially as a synthetic phrase that endeavours to diagnose the specific protein matched and the family to which it might belong. Next, the top five words and phrases are given, together with options to view the full results from each search.

Figure 1 illustrates a query with a rhodopsin sequence. The diagnostic phrase indicates that ‘the query may be a rhodopsin protein and may belong to the rhodopsin-like GPCR’ family, ‘rhodopsin’ being the top word and ‘rhodopsin-like GPCR’ the top phrase. Of itself, this is not surprising, but the precision of the answer recovered from eight different searches is powerful, given the disparate terms used by the different databases to describe the superfamily (i.e. G-protein coupled receptors, seven transmembrane receptor, rhodopsin-like GPCR, 7-Helix G-protein coupled receptor).

∗To whom correspondence should be addressed.
Here, OPSD view the raw data from each search, so no information is discarded. from which this phrase was constructed; below this are options to top of the page, followed by the top five phrases and top five words from which this phrase was constructed: below this are options to view the raw data from each search, so no information is discarded. Here, OPSD_SHEEP is diagnosed as a rhodopsin protein belonging to the rhodopsin-like GPCR family.

EASY’s phrasing has been refined using many different families and will mature with increasing use. An important test of the program came during a recent update of PRINTS (Attwood et al., 2000), which revealed large numbers of TrEMBL (Bairoch and Apweiler, 2000) matches that could not be unambiguously assigned to a given family. As the volume of data was too large to tackle efficiently by hand, we submitted the sequences to EASY. The resulting 1-line diagnoses made light work of what would otherwise have been an impractically slow and difficult task.

Future enhancements of EASY will include: links to thesauri and ontologies, to help address the lack of consistency in protein and family names; incorporation of additional resources such as InterPro (Apweiler et al., 2000), to exploit its deeper knowledge of family relationships and broaden the annotation base available to EASY; re-coding in Java, to facilitate web navigation and provide a more powerful user interface (e.g. using graphs to visualise word frequencies).

Today, comprehensive sequence analysis demands the exploitation of different data sources using disparate search tools, which require an assortment of inputs and provide diverse outputs. Consequently, interpretation can be arduous, especially for inexperienced users. To address these problems, we have developed EASY, which delivers a 1-line diagnosis from the results of multiple searches.

This is particularly important in the context of genome analysis, where data are generated in such vast quantities that efficiency and reliability of automatic annotation is paramount and easy-to-use tools for human curators are essential.

ACKNOWLEDGEMENTS
We are grateful for individual support from Zeneca Pharmaceuticals (JNS). TKA is a Royal Society University Research Fellow.

REFERENCES