ProteoMix: an integrated and flexible system for interactively analyzing large numbers of protein sequences

Eisuke Chikayama\(^1\), Atsushi Kurotani\(^1\), Yutaka Kuroda\(^1\),† and Shigeyuki Yokoyama\(^1,2,3,4,\)∗

\(^1\)Protein Research Group, RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan, \(^2\)Cellular Signaling Laboratory and \(^3\)Structurome Research Group, RIKEN Harima Institute at SPring-8, 1-1-1 Kohto, Mikazuki-cho, Sayo, Hyogo 679-5148, Japan and \(^4\)Department of Biophysics and Biochemistry, Graduate School of Science, the University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

Received on March 11, 2004; revised and accepted on March 26, 2004
Advance Access publication April 22, 2004

ABSTRACT

Summary: ProteoMix is a suite of JAVA programs for identifying, annotating and predicting regions of interest in large sets of amino acid sequences, according to systematic and consistent criteria. It is based on two concepts (1) the integration of results from different sequence analysis tools increases the prediction reliability; and (2) the integration protocol is critical and needs to be easily adaptable in a case-by-case manner. ProteoMix was designed to analyze simultaneously multiple protein sequences using several bioinformatics tools, merge the results of the analyses using logical functions and display them on an integrated viewer. In addition, new sequences can be added seamlessly to an analysis performed on an initial set of sequences. ProteoMix has a modular design, and bioinformatics tools are run on remote servers accessed using the Internet Simple Object Access Protocol (SOAP), ensuring the swift implementation of additional tools. ProteoMix has a user-friendly interactive graphical user interface environment and runs on PCs with Microsoft OS.

Availability: ProteoMix is freely available for academic users at http://bio.gsc.riken.jp/ProteoMix/.
Contact: yokoyama@biochem.s.u-tokyo.ac.jp

It is being recognized that the integration of biological databases will play an essential role in yielding practical solutions to biological problems (Stein, 2003). Similarly, the integration of analysis tools can also yield effective approaches (Andrade et al., 1999; Eyrich et al., 2003; Frishman et al., 2003), as exemplified by meta-servers (Valencia, 2003), which appear to be among the most successful predictors of protein structures (Moult et al., 2003). In line with this reasoning, we have developed an integrated system, ProteoMix (Fig. 1), to analyze using multiple tools and select protein domain sequences for our structural proteomics project (Yokoyama et al., 2000).

As a result, ProteoMix was designed to interactively handle large numbers of sequences. However, the major characteristic distinguishing ProteoMix from meta-servers and most other integration systems is its ability to interactively and flexibly merge the results of several bioinformatics analyses. Here, we introduce the three major features of ProteoMix.

The first feature of ProteoMix is Tool-Mix [Fig. 1B (f)], which merges the results of several methods predicting the same property. Tool-Mix merges the results of annotations and predictions, using a logical statement specified by the user, usually in order to increase the prediction/annotation reliability. The logical statement consists of logical operators, such as ‘AND’, ‘OR’, ‘NOT’ and operands, which are the results of a prediction or annotation. For example, when secondary structure predictions from two methods, A and B, are available, the logical statement ‘[result A] AND [result B]’ means that a residue that is predicted as a helix (or a strand) by both methods will be predicted as a helix (or, respectively, a strand), and no prediction will be made otherwise. This example is similar to the joint method, which is known to enhance the effectiveness of secondary structure prediction (Nishikawa and Noguchi, 1991). Notably, Tool-Mix is not limited to merging secondary structure predictions, and the merging is not restricted to the ‘AND’ operator.

The second feature of ProteoMix is Complement-Mix. It is conceptually an extension of Tool-Mix, but it can
merge methods predicting different (although usually complementary) properties. As in merging prediction results for the same property, the predictions are expected to be more reliable when the complementary predictions are mutually consistent. For example, the structural domain prediction reliability can be enhanced by combining domain region predictions with those of their boundaries. Specifically, we used Complement-Mix to identify potential targets for our structural proteomics project. To this end, we analyzed 2000 sequences listed in the Kazusa HUGE cDNA sequence library (Kikuno et al., 2002), and selected putative domain regions with termini that overlapped with either the N- or C-terminus of a full-length protein or the predicted domain boundary [Fig. 1B (g)].

The third feature of ProteoMix is its versatility, especially with regard to extending the set of protein sequences
analyzed, and to adding new bioinformatics tools to the system. First, new sequences can be seamlessly added to an analysis performed on an initial set of sequences. For example, ProteoMix enables the prediction of structural domains from, e.g. 100 new sequences, using the protocol described above for the analysis of the HUGE library. It then merges the new results with those resulting from the initial analysis, and furthermore, it sorts and selects the structural domains from the updated list of putative domains. Second, the user can also extend the list of bioinformatics tools. The tools are incorporated in ProteoMix as external component files, and can thus be developed separately from the system. The tools may run on either the user’s PC or a remote server accessed on the Internet by the Simple Object Access Protocol (SOAP). In either case, the build-in design enables the swift implementation of newly released or updated tools, further adding to the versatility of ProteoMix.

ProteoMix runs on Microsoft Windows® with Java 1.4. An Intel® Pentium® IV/2.4 GHz or higher (or compatible) microprocessor and 256 MB RAM minimum (768 MB RAM recommended) are required. ProteoMix can be downloaded from our website. Detailed information is available at our website http://bio.gsc.riken.jp/ProteoMix/.

ACKNOWLEDGEMENTS

We thank Hiroaki Yamaguchi (Venture Safenet, Inc.) for technical support with the implementation ProteoMix and Tomohiro Takahashi (Borland Co., Ltd) for the design of the server system. This work was supported by the RIKEN Structural Genomics/Proteomics Initiative (RSGI), the National Project on Protein Structural and Functional Analyses, and the Ministry of Education, Culture, Sports, Science and Technology of Japan.

REFERENCES


