Today’s molecular biologists often find themselves confronted with a paradox: although accurate sequence information on genes of interest is obtained routinely, the subsequent question: ‘What is the function of this putative protein?’ is much harder to answer. This is true even for experienced experimental researchers, who find themselves struggling suddenly with an overwhelming variety of bioinformatics methods, as the lack of experimental evidence makes database searches and sequence analyses necessary.

The GeneQuiz system for automatic functional annotation of protein sequences, available to the scientific community via a WWW server at the European Bioinformatics Institute, provides easy access to a first answer to this crucial question. It provides a single, uniform, Web-based user interface and, being fully automatic, hides the complexity of the different analysis methods used. It attempts to derive a single functional annotation for each query sequence, which is presented in the context of an extensive report that allows the user to track the various aspects of the analyses in detail, if desired (Figs 1, 2).

How this system works is explained in Box 1. Briefly, the user inputs one or more amino acid sequences, and the server runs a variety of automated analyses, returning a Web-based report that summarizes the results that have now become available. A GeneQuiz analysis takes usually between 5 and 15 minutes per sequence, although the actual time until results are accessible might be considerably larger if the server is heavily used.

The system is used routinely for analyses of large sets of open reading frames (ORFs) typically derived from complete genome sequences. As a result, the majority of manually assigned functions are confirmed and a substantial number of newly assigned ORFs are added, as first shown for Haemophilus influenzae15, followed by Mycoplasma genitalium16, Methanococcus jannaschii17, and others. The advantage of GeneQuiz using up-to-date databases has been demonstrated, for example, in the case of the Synechocystis sp. ORF slr0665, the GeneQuiz annotation ‘cis-aconitase’ correctly anticipated the eventually updated annotation in SwissProt by several months.

Caveats
When using the GeneQuiz system, one should bear in mind several caveats. First, GeneQuiz does not perform analysis of DNA sequences – nucleotide sequences have to be translated into amino acid sequences prior to a GeneQuiz analysis. Second, GeneQuiz performs a set of analyses that result in a functional annotation, but does not allow additional methods to be run, nor does it provide means to do comparative analyses between several query sequences. Third, automatic functional annotation is not perfect. To decide about the validity of the automatically assigned function, it is recommended to make use of the comprehensive and detailed report features.

Summary
We believe that the GeneQuiz service is ideal to help lab-based researchers to obtain rapidly functional information for a limited number of novel sequences, for which none or very little is available. Sequences can be resubmitted for a GeneQuiz analysis to ensure that the results remain up-to-date.

The results of a GeneQuiz analysis would typically guide further analysis or experiments, or lead to some kind of research prioritization among a number of sequences of interest.

Access to the GeneQuiz server
GeneQuiz is accessible through a public Web server at http://www.sander.ebi.ac.uk/gqsrv/submit. When submitting protein sequences for analysis, the user must supply a valid e-mail address, to which the system will send automatically a message with the identifier of the analysis and a URL address from where
Figure 2
Alignment by MView (Ref. 14) for the query protein the report of which is displayed in Fig. 1. (a) List of homologs including database identifier (blue), brief description and percentage of identity to the query (last column), as found by BLASTP. The top 75 hits were taken and then filtered so that no two sequences displayed have more than 70% sequence identity (see top-left corner). (b) List of structural properties of the query sequence and motifs (Box 1). (c) Alignment to the query sequence (only 60 amino acids are shown), with structural properties and motifs (top), and similar fragments detected in homologs by BLASTP (bottom). The composite alignment of similar fragments to the query sequence indicates which regions of the query are conserved in similar sequences. The coloring scheme (by identity to the query, color-coded by physicochemical property: green, hydrophobic; blue, negatively charged; red, positively charged; purple, polar; orange, cysteine) reveals patterns that are consistently conserved across sets of sequences. A BLOCKS motif and low-complexity regions are also found in this region of the query protein. They are shown at the top together with the predicted residue exposure (green, buried; red, exposed), secondary structure (h, helix; b, structure), transmembrane helix (H) and transmembrane helix and loop topology (red i, in; purple o, out; green T; transmembrane helix) as determined using PHD (Ref. 10). Note that MView is not an alignment tool: homologous fragments are pasted below the sequence and the resulting sequences are constructions.

Box 1. How GeneQuiz works
Input: An amino acid sequence
Analyses performed:
- Primary structure analysis:
  - Search for repeats and coiled-coil regions [REPEATS (M. Vingron, unpublished), COILS (Ref. 4)]
  - Search for composition-biased regions with subsequent masking [SEG (Ref. 5), BIASDB (G. Casari and C. Ouzounis, unpublished)]
- Similarity searches [BLAST (Ref. 6), FASTA (Ref. 7)], performed on a non-redundant protein database comprising SwissProt, SwissNew, SPTREMBL, PIR, GenPept and GenPeptNew (Ref. 8), and on a non-redundant nucleotide database comprising EMBL and EMBLNEW (Ref. 8). The databases are updated bi-weekly
- Protein-motif searches [PROSITE (Ref. 8), BLOCKS (Ref. 8)]
- Secondary-structure prediction including accessibility and transmembrane regions [PHD (Ref. 10)]
- 3D-structure modeling, if a sufficiently similar protein of known structure is available in the PDB databank [WHATIF (Ref. 12)]
- Functional annotation with reliability measure
- Assessment of functional class
- Assessment of species distribution for homologs
Output: A Web-based report summarizing the results of the above analyses (see example in Fig. 1). Hyperlinks allow easy navigation and access to underlying data and database entries via SRS (Ref. 13). Also accessible from the report are visualizations of the BLAST and FASTA search results, with hit sequences being stacked under the matching regions of the query sequence via MView (Ref. 14; see example in Fig. 2)
the results can be accessed. This URL cannot be guessed, which provides a minimum degree of privacy for the results.

Usage of the GeneQuiz server is subject to the following restrictions (for a given e-mail address): 12 sequences per month; 84 sequences per month maximum; 100 sequences maximum total. Researchers whose needs exceed these limits should contact the maintainers of GeneQuiz per e-mail (genequiz@ebi.ac.uk) to discuss a special agreement.

References


P2Y-receptor-ligand database

Since the mediatory role of extracellular nucleotides was proposed in the early 1970s, nine nucleotide receptors belonging to the P2Y-receptor family have been cloned and expressed. This family of receptors are G-protein-coupled receptors, which are able to evoke intracellular responses through activation of the classical secondary messenger pathway. These downstream signalling effects have been used to analyse the ligand-binding properties of these receptors. We have collected most of this information into a computer-operated database and have made this available for the purinoceptor research community at http://bioorg.chem.ut.ee/p2y/.

This database can be searched online and it includes information about the biological system used in experiments, the putative receptor subtype(s) involved, the activity of the ligands, the assay methods and the references to the original papers. The present version involves 1976 records for 267 distinct compounds. Although the database is focused on G-protein-coupled P2Y receptors, also some data for P2X receptors (ligand-gated ion channels activated by extracellular nucleotides) were included. No critical analysis of the data was made before their listing. Thus, the data collected provide also information about the development of the assay methods as well as our understandings of the purity requirements for the ligands tested.

Any comments and references, but also any new data to add to the database should be addressed to: p2y-data@bioorg.chem.ut.ee.

KATRIN SAK, ANDRES KREEGIPUU, JAAN JARV
Institute of Chemical Physics, Tartu University, 2 Jakobi street, 51014 Tartu, Estonia.

References