GENEQUIZ II: AUTOMATIC FUNCTION ASSIGNMENT FOR GENOME SEQUENCE ANALYSIS

Georg Casari, Christos Ouzounis,[a] Alfonso Valencia,[b] & Chris Sander [c]
Protein Design Group
EMBL Heidelberg, Meyerhofstrasse 1
D-69012 Heidelberg, Germany
e-mail: {surname}@embl-heidelberg.de

GeneQuiz II, a system for large-scale biological sequence analysis, has been improved significantly. The most important new features are a module for fully automatic function assignment and a web interface. The automatic function assignment module has been developed on the basis of past experience and a detailed benchmarking for a large number of manually assigned sequences. It is estimated that its performance reaches over 90% accuracy with no false positives. Therefore, the interactive data evaluation step, which in the past has been a serious bottleneck, is becoming increasingly non-essential. We have also modified the user interface, replacing the previous viewer with a html interface to the relational database, so the results can be accessible through the world wide web. A recent analysis of the complete genome of *Haemophilus influenzae* shows the spectacular improvements in speed and accuracy.

Keywords: Automated sequence analysis, molecular databases, rule-based expert systems, knowledge representation, data visualization, web interface

[a]AI Center, SRI International, 333 Ravenswood Ave., Menlo Park CA 94025-3493, USA
[b]CNB-CSIC, Campus Univesidad Autonoma, Cantoblanco, Madrid E-28049, Spain
[c]European Bioinformatics Institute (EMBL-EBI), Hinxton Hall, Cambridge CB10 1RQ, UK

We present GeneQuiz II, the updated version of an integrated system for large-scale biological sequence analysis. The main improvement of the system consists of a module that allows a fully automatic evaluation of results and function assignment. This component of GeneQuiz II has been modeled on our experience with the analysis of yeast chromosomes III, XI and VIII, a collection of protein sequences from Archaea and a contig dataset from *Mycoplasma capricolum*. GeneQuiz II has been used for the automated analysis of yeast chromosomes I, II, V and IX, with a remarkably improved performance. The original prototype consisted of three parts: (i) the database update and (ii) search systems - driven by perl programs and rdb, a simple relational database engine also written in perl - and (iii) the visualization and browsing system (developed under C++/ET++). The principal design requirement for the first part was the complete automation of all repetitive actions: database updates, efficient sequence similarity searches and sampling of results in a uniform fashion. Now, there is a principal intermediate component, which has been designed to automate the last part of data evaluation. The user is presented with a compact summary consisting of 'features' that summarize the results from heterogeneous database searches and diagnostics on protein sequences. The expert's primary task of result interpretation has now been replaced by procedures which characterize the query protein sequences in subtle ways based on a number of expert rules, that perform remarkably well. Thus, the ultimate task of interactive data evaluation is becoming increasingly non-essential, and browsing through a database of results forms an optional requirement in the whole analysis. The present paper discusses the recent significant advances towards the complete automation of the sequence analysis process, based on the 'rule-based' component for automated evaluation. The hierarchical structure of the system allows future extensions to incorporate other, more general, modules that can reason about function distribution and phylogenetic comparisons of genome data. Benchmarking with yeast chromosomes VIII and II shows that the performance of the system reaches over 90% accuracy, relative to manual annotation. A novel method of eliminating composition bias in protein sequences allows lowering of thresholds for automatic homology detection, boosting accuracy close to 100%. Also, an automatic classification scheme into major functional classes has been implemented. A recent successful application was the fully automatic analysis of the 1743 open reading frames of *Haemophilus influenzae*, the first completely sequenced genome of a free-living organism (see http://www.sander.embl-heidelberg.de/genequiz/haemophilus.html). The original browser has been replaced by a www interface for wide access through the web.

References

1. Scharf M, Schneider R, Casari G, Bork P, Valencia A, Ouzounis C, Sander C (1994) GeneQuiz: a workbench for sequence analysis. In: *Intelligent Systems for Molecular Biology 1994*. AAAI Press, pp 348-353. Menlo Park, CA.

2. Bork P, Ouzounis C, Sander C (1994) From genome sequences to protein function. Curr. Opinion Struct. Biol. 4, 393-403.

3. Bork P, Ouzounis C, Sander C, Scharf M, Schneider R, Sonnhammer E (1992) What's in a genome? Nature 358, 287-287.

4. Ouzounis C, Bork P, Casari G, Sander C (1995) New protein functions in yeast chromosome VIII. Protein Sci., in press.

5. Ouzounis CA, Kyrpides NC, Sander C (1995) Novel protein families in Archaean genomes. Nucl. Acids Res. 23, 565-570.

6. Bork P, Ouzounis C, Casari G, Schneider R, Sander C, Dolan M, Gilbert W, Gillevet PM (1995) Exploring the *Mycoplasma*

capricolum genome: a minimal cell reveals its physiology. Mol. Microbiol. 16, 955-967.

7. Casari G, Ouzounis C, Sander C (1995) A general method to detect compositionally biased regions in biological sequences. In preparation.

8. Casari G, Andrade M, Bork P, Boyle J, Daruvar A, Ouzounis C, Schneider R, Tamames J, Valencia A, Sander C (1995) Challenging times for bioinformatics. Nature 376, 247-248.

GENEQUIZ II: AUTOMATIC FUNCTION ASSIGNMENT FOR GENOME SEQUENCE ANALYSIS

Georg Casari, Christos Ouzounis,[a] Alfonso Valencia,[b] & Chris Sander [c]
Protein Design Group
EMBL Heidelberg, Meyerhofstrasse 1
D-69012 Heidelberg, Germany
e-mail: {surname}@embl-heidelberg.de

GeneQuiz II, a system for large-scale biological sequence analysis, has been improved significantly. The most important new features are a module for fully automatic function assignment and a web interface. The automatic function assignment module has been developed on the basis of past experience and a detailed benchmarking for a large number of manually assigned sequences. It is estimated that its performance reaches over 90% accuracy with no false positives. Therefore, the interactive data evaluation step, which in the past has been a serious bottleneck, is becoming increasingly non-essential. We have also modified the user interface, replacing the previous viewer with a html interface to the relational database, so the results can be accessible through the world wide web. A recent analysis of the complete genome of *Haemophilus influenzae* shows the spectacular improvements in speed and accuracy.

Keywords: Automated sequence analysis, molecular databases, rule-based expert systems, knowledge representation, data visualization, web interface

[a] AI Center, SRI International, 333 Ravenswood Ave., Menlo Park CA 94025-3493, USA
[b] CNB-CSIC, Campus Univesidad Autonoma, Cantoblanco, Madrid E-28049, Spain
[c] European Bioinformatics Institute (EMBL-EBI), Hinxton Hall, Cambridge CB10 1RQ, UK

We present GeneQuiz II, the updated version of an integrated system for large-scale biological sequence analysis. The main improvement of the system consists of a module that allows a fully automatic evaluation of results and function assignment. This component of GeneQuiz II has been modeled on our experience with the analysis of yeast chromosomes III, XI and VIII, a collection of protein sequences from Archaea and a contig dataset from *Mycoplasma capricolum*. GeneQuiz II has been used for the automated analysis of yeast chromosomes I, II, V and IX, with a remarkably improved performance. The original prototype consisted of three parts: (i) the database update and (ii) search systems - driven by perl programs and rdb, a simple relational database engine also written in perl - and (iii) the visualization and browsing system (developed under C++/ET++). The principal design requirement for the first part was the complete automation of all repetitive actions: database updates, efficient sequence similarity searches and sampling of results in a uniform fashion. Now, there is a principal intermediate component, which has been designed to automate the last part of data evaluation. The user is presented with a compact summary consisting of 'features' that summarize the results from heterogeneous database searches and diagnostics on protein sequences. The expert's primary task of result interpretation has now been replaced by procedures which characterize the query protein sequences in subtle ways based on a number of expert rules, that perform remarkably well. Thus, the ultimate task of interactive data evaluation is becoming increasingly non-essential, and browsing through a database of results forms an optional requirement in the whole analysis. The present paper discusses the recent significant advances towards the complete automation of the sequence analysis process, based on the 'rule-based' component for automated evaluation. The hierarchical structure of the system allows future extensions to incorporate other, more general, modules that can reason about function distribution and phylogenetic comparisons of genome data. Benchmarking with yeast chromosomes VIII and II shows that the performance of the system reaches over 90% accuracy, relative to manual annotation. A novel method of eliminating composition bias in protein sequences allows lowering of thresholds for automatic homology detection, boosting accuracy close to 100%. Also, an automatic classification scheme into major functional classes has been implemented. A recent successful application was the fully automatic analysis of the 1743 open reading frames of *Haemophilus influenzae*, the first completely sequenced genome of a free-living organism (see http://www.sander.embl-heidelberg.de/genequiz/haemophilus.html). The original browser has been replaced by a www interface for wide access through the web.

References

1. Scharf M, Schneider R, Casari G, Bork P, Valencia A, Ouzounis C, Sander C (1994) GeneQuiz: a workbench for sequence analysis. In: *Intelligent Systems for Molecular Biology 1994*, AAAI Press, pp 348-353. Menlo Park, CA.

2. Bork P, Ouzounis C, Sander C (1994) From genome sequences to protein function. Curr. Opinion Struct. Biol. 4, 393-403.

3. Bork P, Ouzounis C, Sander C, Scharf M, Schneider R, Sonnhammer E (1992) What's in a genome? Nature 358, 287-287.

4. Ouzounis C, Bork P, Casari G, Sander C (1995) New protein functions in yeast chromosome VIII. Protein Sci., in press.

5. Ouzounis CA, Kyrpides NC, Sander C (1995) Novel protein families in Archaean genomes. Nucl. Acids Res. 23, 565-570.

6. Bork P, Ouzounis C, Casari G, Schneider R, Sander C, Dolan M, Gilbert W, Gillevet PM (1995) Exploring the Mycoplasma

capricolum genome: a minimal cell reveals its physiology. Mol. Microbiol. 16, 955-967.

7. Casari G, Ouzounis C, Sander C (1995) A general method to detect compositionally biased regions in biological sequences. In preparation.

8. Casari G, Andrade M, Bork P, Boyle J, Daruvar A, Ouzounis C, Schneider R, Tamames J, Valencia A, Sander C (1995) Challenging times for bioinformatics. Nature 376, 247-248.