W170–W172 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh369

SNPbox: web-based high-throughput primer design from gene to genome

Stefan Weckx, Peter De Rijk, Christine Van Broeckhoven and Jurgen Del-Favero*

Department of Molecular Genetics, Bioinformatics Unit, Flanders Interuniversity Institute for Biotechnology, University of Antwerp, Belgium

Received February 12, 2004; Revised and Accepted March 9, 2004

ABSTRACT

SNPbox is a modular software package that automates the design of PCR primers for large-scale amplification and sequencing projects in a standardized manner resulting in high-quality PCR amplicons with a low failure rate. Here, we present the SNPbox web server at http://www.SNPbox.org, which hosts the SNPbox web service as well as the data from SNPbox analysis of all Ensembl exons. The data of this genome-wide SNPbox application can be visualized in Ensembl's *ContigView* through a DAS (distributed annotation system) annotation server.

INTRODUCTION

Designing primers has become an everyday routine in life science laboratories. Hereto, numerous primer design tools are available as web applications or as stand-alone programs (1–6). Although the efficiency of these programs is beyond dispute, most of these programs can design only one primer set at a time and are therefore less suited for large-scale primer design projects. Given that current laboratory equipment allows high-throughput PCR sample preparation and analysis, the primer design step has become a major bottleneck in the workflow, especially in large-scale projects that involve templates in the kb to Mb range. Therefore, a high-throughput approach towards automated primer design became essential. Here the designed primers should amplify under uniform conditions resulting in high-quality PCR amplicons with a low failure rate.

SNPbox was designed with these specific criteria in mind and allows automated primer design for large template sequences in a highly standardized way taking repeat regions into account (Weckx *et al*., submitted). SNPbox consists of three modules: an SNP module, an exon module and a saturation module. In short, primers are designed for a well-defined target sequence which can contain either known SNPs from public databases for validation (SNP module), coding regions indicated by aligning cDNA/EST sequences to the genomic DNA (exon module), or promoter or whole genomic regions (saturation module). Coding regions are extended with 50 bp of intronic sequence 5′ and 3′ to optimize for inclusion of the branch point and splice sites in the target. Non-polymorphic repeats less then 300 bp can be included in the target. Polymorphic repeats cannot be included and primers can never be selected in any kind of repeat sequences (Weckx *et al*., submitted). Of the 2500 primer sets we designed using SNPbox, 95% successfully amplified genomic DNA resulting in one specific amplicon of expected size using the built-in PCR conditions without further need for PCR optimization. The overall success rate also depended on the quality of the oligonucleotide synthesis.

Here, we introduce the SNPbox web server at http://www.SNPbox.org containing three components: a web service for primer design, a distributed annotation system (DAS) annotation server providing information of the genome-wide application of SNPbox and an interface to a local database where primer information of the genome-wide application can be retrieved.

SNPbox WEB SERVICE

A web interface, based on the CGI-Tcl library (http://expect.nist.gov/cgi.tcl/), guides the visitor in three steps through the process of job submission: in the first step, the user selects a module: SNP, exon or saturation module. In the second step, required and optional input is requested. For all modules, a genomic sequence is required, either as a sequence file in FASTA format or as a GenBank accession number or gi number, in which case the server will retrieve the sequence directly from GenBank. Optionally, the user can provide an email address so that a notification message can be sent when the SNPbox job is finished. The default primer conditions, such as melting temperature and primer length ranges, as well as the optimal target length, can be defined by the user.

For the SNP module, no further input is required. The genomic sequence is aligned to a local copy of HGVbase (7,8) using the BLAST algorithm (9) to determine the positions of public SNPs. For the exon module, cDNA/EST

*To whom correspondence should be addressed. Tel: +32 3 820 2321; Fax: +32 3 820 2541; Email: jurgen.delfavero@ua.ac.be

sequences are required that will be aligned to the genomic sequence using Spidey (10). The cDNA sequences can be uploaded as a FASTA file, or can be provided as GenBank accession or gi numbers. For the saturation module, a start and end position in the genomic sequence is required in order to define the genomic region for amplicon coverage. In the third step, the input is checked for inconsistencies before submission. In case of nonconformities, the user is asked to change the input according to the guidelines.

The output of SNPbox is visualized as a web page; the URL is displayed directly after job submission allowing immediate opening or bookmarking of the result web page. The URL incorporates an eight-digit random-generated code ensuring data confidentiality to the users. Once SNPbox has finished, the output can be viewed on the web page. Furthermore, a tab-delimited text file containing all information about the primers and amplicons can be downloaded, as well as a compressed file containing the experiment folder. The directory containing the web page, as well as the database entries of an experiment, will be removed automatically after one week or can be deleted sooner by providing the experiment ID number and eight-digit code.

For server performance, there is a limit on the length of genomic sequence to be analyzed, which is currently set to 200 kb, and is thus within the size range of most genes.

SNPbox AND THE GENOME: GENOME-WIDE, EXON-BASED PCR AMPLICONS

Ensembl, a joint project of the Wellcome Trust Sanger Institute and the European Molecular Biology Laboratory (EMBL) European Bioinformatics Institute, has become a reliable source of annotation information for the human genome. Gene prediction is an important part of the automated annotation pipeline behind Ensembl and the combination of all generated evidence based on different gene prediction strategies and sequence database homology searches results in the Ensembl genes (11–13). This gene annotation can be used for identifying potential disease genes in a candidate region under investigation. Since many scientists are searching for polymorphisms and mutations in candidate genes using resequencing, all are frequently designing primers for PCR and sequencing. To facilitate this process, we used the exon mode of SNPbox to design primer pairs for all exons of the Ensembl genes. Hereto, the tables *contig*, *assembly*, *exon*, *exon_stable_id*, *exon_transcript*, *transcript_stable_id*, *transcript* and *gene_stable_id* of the Ensembl's homo_sapiens_core MySQL database version 18.34.1 were installed locally; the annotation in this release is based on the NCBI assembly 34. SNPbox was run in parallel on a cluster with 15 nodes and one server, communicating over AFS, and the data were stored in a SQLite database (http://www.sqlite.org). SQLite was chosen because it requires a low overhead. The use of a client–server database management system would have caused a large overhead on the cluster head and would have slowed the whole process. The contig sequences were retrieved from the Ensembl FTP server (ftp://ftp.ensembl.org) and saved on the cluster head. The dna_id of each contig was retrieved from the Ensembl database and incorporated into the file names for easy access of the sequences.

We used the RepeatMasker (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker) annotation provided in the *repeat_consensus* and the *repeat_feature* tables. Instead of using the 'dust' repeat annotation, the output of an adopted version of Sputnik (http://rast.abajian.com/sputnik/) was used to annotate micro-satellite repeats and single base stretches. Comparison of the 'dust' repeats with the output of Sputnik showed that both annotations were very similar. Although the integration of an adapted Sputnik version resulted in additional computational steps, we preferred to use Sputnik because the repeats are classified by SNPbox in three classes with different consequences for the decision-making in SNPbox (Weckx *et al*., submitted): small micro-satellite repeats with <8 repeat units, moderate micro-satellite repeats with ⩾8 repeat units, and the single base stretches with >8 equal bases. Overall, moderate micro-satellite repeats and single base stretches cause problems in subsequent sequencing reactions and can therefore not be included in the PCR amplicons.

To provide the results of the genome-wide application of SNPbox on the Ensembl genes, we have set up a Lightweight DAS annotation server (LDAS), based on the DAS (14), following the guidelines on the DAS website (http://www.biodas.org). The results can easily be added to Ensembl's *ContigView*. Detailed instructions are available on the SNPbox web server. The SNPbox LDAS annotation server is at http://www.snpbox.org/cgi-bin/das/; the corresponding reference server is at http://servlet.sanger.ac.uk:8080/das/. The results of SNPbox can best be evaluated when displaying only the Ensembl Transcripts and repeats. By clicking on the SNPbox annotation feature, the browser will redirect to the SNPbox server and will show additional information such as primer sequences, annealing temperatures and amplicon length. Additionally, by providing the Ensembl gene ID to a search tool on the SNPbox web server, primer information for all exons of the specified gene can be obtained.

The overall success rate of the *in silico* primer design by SNPbox on the Ensembl exons was 97.17%. For the 208 202 exons, 227 187 objects were initially indicated and a target sequence could be defined for 98.62% of them. The remaining 3176 objects had a significant overlap with repeats not suitable for inclusion in a target. SNPbox was able to design primer sets for 223 844 targets. For 3343 targets, the primer design failed mainly because of an unfavorable base composition in the regions where the primers should be picked. The success rate for primer design was 98.53%.

CONCLUSION

Primer design has become a time-determining step in the workflow of high-throughput projects in life science laboratories. SNPbox allows highly standardized primer design in an automated way. With the SNPbox web service and the application of SNPbox to the Ensembl genes, we aim to make SNPbox known and available to a broad group of interested scientists. They can use the SNPbox software on-line to facilitate their primer design for projects up to 200 kb of genomic sequence, or can use the preprocessed primer sets designed for the Ensembl exons. To use SNPbox in projects beyond the scope of the server, the authors can be contacted for information about a standalone version of the software.

ACKNOWLEDGEMENTS

This work was in part funded by the Special Research Fund of the University of Antwerp, the Fund for Scientific Research Flanders and the Inter University Attraction Poles program P5/19 of the Federal Science Policy Office, Belgium.

REFERENCES

1. Chen,S.H., Lin,C.Y., Cho,C.S., Lo,C.Z. and Hsiung,C.A. (2003) Primer Design Assistant (PDA): a web-based primer design tool. *Nucleic Acids Res.*, **31**, 3751–3754.
2. Haas,S., Vingron,M., Poustka,A. and Wiemann,S. (1998) Primer design for large scale sequencing. *Nucleic Acids Res.*, **26**, 3006–3012.
3. Haas,S.A., Hild,M., Wright,A.P., Hain,T., Talibi,D. and Vingron,M. (2003) Genome-scale design of PCR primers and long oligomers for DNA microarrays. *Nucleic Acids Res.*, **31**, 5576–5581.
4. Li,P., Kupfer,K.C., Davies,C.J., Burbee,D., Evans,G.A. and Garner,H.R. (1997) PRIMO: a primer design program that applies base quality statistics for automated large-scale DNA sequencing. *Genomics*, **40**, 476–485.
5. Proutski,V. and Holmes,E.C. (1996) Primer Master: a new program for the design and analysis of PCR primers. *Comput. Appl. Biosci.*, **12**, 253–255.
6. Rozen,S. and Skaletsky,H.J. (2003) Primer3 on the WWW for general users and for biologist programmers. *Meth. Mol. Biol.*, **132**, 365–286.
7. Fredman,D., Siegfried,M., Yuan,Y.P., Bork,P., Lehvaslaiho,H. and Brookes,A.J. (2002) HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res.*, **30**, 387–391.
8. Fredman,D., Munns,G., Rios,D., Sjoholm,F., Siegfried,M., Lenhard,B., Lehvaslaiho,H. and Brookes,A.J. (2004) HGVbase: a curated resource describing human DNA variation and phenotype relationships. *Nucleic Acids Res.*, **32**, D516–D519.
9. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids. Res.*, **25**, 3389–3402.
10. Wheelan,S.J., Church,D.M. and Ostell,J.M. (2001) Spidey: a tool for mRNA-to-genomic alignments. *Genome Res.*, **11**, 1952–1957.
11. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J. *et al.* (2004) Ensembl 2004. *Nucleic Acids. Res.*, **32**, D468–D470.
12. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.
13. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
14. Dowel,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L.D. (2002) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.