This information has not been peer-reviewed. Responsibility for the findings rests solely with the author(s).

Deposited research article

AutoPrime: selecting primers for expressed sequences

Gunnar Wrobel*‡, Felix Kokocinski†‡ and Peter Lichter†

Addresses: *Divisions of Bioinformatics and Biochemistry, Swiss Institute of Bioinformatics, Klingelbergstrasse 50/70, 4056 Basel, Switzerland. †Molecular Genetics, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany. ‡These authors contributed equally to this work.

Correspondence: Peter Lichter. E-mail: m.macleod@dkfz.de

Posted: 15 April 2004

Genome ***Biology*** 2004, **5:**P11

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/5/P11

© 2004 BioMed Central Ltd

Received: 13 April 2004

This is the first version of this article to be made available publicly.

.deposited research

AS A SERVICE TO THE RESEARCH COMMUNITY, GENOME **BIOLOGY** PROVIDES A 'PREPRINT' DEPOSITORY TO WHICH ANY ORIGINAL RESEARCH CAN BE SUBMITTED AND WHICH ALL INDIVIDUALS CAN ACCESS FREE OF CHARGE. ANY ARTICLE CAN BE SUBMITTED BY AUTHORS, WHO HAVE SOLE RESPONSIBILITY FOR THE ARTICLE'S CONTENT. THE ONLY SCREENING IS TO ENSURE RELEVANCE OF THE PREPRINT TO GENOME **BIOLOGY**'S SCOPE AND TO AVOID ABUSIVE, LIBELLOUS OR INDECENT ARTICLES. ARTICLES IN THIS SECTION OF THE JOURNAL HAVE **NOT** BEEN PEER-REVIEWED. EACH PREPRINT HAS A PERMANENT URL, BY WHICH IT CAN BE CITED. RESEARCH SUBMITTED TO THE PREPRINT DEPOSITORY MAY BE SIMULTANEOUSLY OR SUBSEQUENTLY SUBMITTED TO GENOME **BIOLOGY** OR ANY OTHER PUBLICATION FOR PEER REVIEW; THE ONLY REQUIREMENT IS AN EXPLICIT CITATION OF, AND LINK TO, THE PREPRINT IN ANY VERSION OF THE ARTICLE THAT IS EVENTUALLY PUBLISHED. IF POSSIBLE, GENOME **BIOLOGY** WILL PROVIDE A RECIPROCAL LINK FROM THE PREPRINT TO THE PUBLISHED ARTICLE.

AutoPrime: Selecting Primers for Expressed Sequences

Gunnar Wrobel[1,4], Felix Kokocinski[2,4] and Peter Lichter*[2]

[1]Divisions of Bioinformatics and Biochemistry, Swiss Institute of Bioinformatics, Klingelbergstrasse 50/70, 4056 Basel, Switzerland

[2]Molecular Genetics, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

Email: GunnarWrobel - gunnar.wrobel@unibas.ch; Felix Kokocinski - f.kokocinski@dkfz.de; Peter Lichter - m.mcleod@dkfz.de;

*Corresponding author

[4]Both authors contributed equally to this work.

Abstract

Background: Primers designed for expressed sequences should be specific for the mRNA sequence of a gene but not yield a product from its genomic sequence. Already slight contaminations of genomic DNA can lead to incorrect measurements for sensitive methods like real-time PCR (RT-PCR).

Results and Conclusions: The software AutoPrime automates the task of generating such primers for RT-PCR experiments by combining the information from sequence databases with primer design software. Thus the software is able to cope with the high demand of RT-PCR validation experiments following the recent increase in microarray expression profiling studies. The software generates high quality primers more efficiently than possible by manual design methods and has been successfully employed for RT-PCR primer design in our as well as other laboratories.

Availability http://www.AutoPrime.de

Background

Primers designed for the RT-PCR measurement of eukaryotic gene expression are usually selected so that they will amplify cDNA generated from the mRNA but not from the genomic sequence of a gene. Small contaminations of only 0.1% of genomic DNA can already lead to a multi-fold excess of genomic over mRNA sequences. Thus the primer design process needs to take advantage of the eukaryotic splicing system for mRNA. Primer pairs can be selected so that one of the primers matches an exon-exon border sequence that is not present on the genomic level. Alternatively the primer pair can be designed by placing each primer in a different exon so that a product based on the genomic sequence would include a long intronic sequence resulting in a lower amplification efficiency.

One of the tools often used for primer design in RT-PCR measurements is the Primer Express Software (Applied Biosystems, Foster City, USA) provided in conjunction with RT-PCR-cyclers. Others are Oligo (Molecular Biology Insights, Inc., Cascade, USA), PRIDE [1] or primer3 [2], the latter being freely available. While these programs provide high quality primer design and some also allow the generation of hybridization probes, they do not provide an environment that facilitates the design of primers discriminating cDNA from genomic DNA.

Because of the recent increase in expression profiling studies and the subsequent need for result verification by RT-PCR two databases for tested RT-PCR primer pairs have been created recently [3, 4]. In addition a third database containing automatically designed primers has been established and provides primers for human and mouse genes [5]. The creation of these databases underlines the necessity for high quality primers in the area of RT-PCR.

Currently there exists no software that is capable of automatically performing the primer design following the quality criteria described above and directly takes advantage of the sequence information available online. Instead most rely on manual copy and paste interactions for the sequence input.

Methods

The software AutoPrime has been designed to automate the primer design task for RT-PCR. It is written in the language perl and combines the benefits of the primer3 program with the information of the Ensembl database [6].

To design primers that take the exon boundaries of a transcript into consideration, the genomic sequence and the exon structure of the corresponding gene have to be known. AutoPrime automatically retrieves

this information from the Ensembl databases and uses it to create a first primer either on an exon border or inside an exon. The selected sequence regions are then explicitely specified to the primer design software. For a primer atop an exon-exon border, the pairing primer is subsequently searched without positional restriction, but for a primer pair inside the exons, AutoPrime will search a second primer only in the confines of an adjacent exon. Each primer pair generated is subsequently checked for problems like the potential to form primer dimers, self-homologies, problematic melting temperatures, etc.

Automating the primer design process for RT-PCR primers provides several advantages. The direct retrieval of sequence and exon information for a gene is faster and can be considered less error-prone than manual processing. Once the sequences are available AutoPrime can generate the possible priming ranges from the exon structure. These in conjunction with the transcript sequence are then passed as input to the primer generation program primer3 for the automatic generation of primer pairs (Figure 1). Additionally, when choosing a primer on an exon-exon border, AutoPrime can enter the genomic sequence of the gene into the mispriming library. This further helps in avoiding products from genomic contaminations which is the primary reason for choosing primers on such a border. Additionally, hybridization probes can be generated to be used as a reporter for the RT-PCR.

AutoPrime is designed to be accessed via the command-line or called from within other programs. We have also set up a web front-end, allowing comfortable retrieval of primer sequences using a graphical user mask via any web browser [7]. The user needs to supply no more than the gene short name or the gene Ensembl id for which the primers are to be designed. The default parameters can be adjusted if the result is not satisfying. All parameters are further explained in the supplied documentation. The default output of the program is an XML file which includes information about all parameter settings, the transcript sequences of the gene and statistics about the primers found along with their sequences. Using the web interface, the output can also be formatted as an HTML or text file.

Currently primers can be designed for M. musculus, R. norvegicus, D. rerio, F. rubripes, A. gambiae, D. melanogaster, C. elegans, C. brigsae and H. sapiens. For all of these organisms sequence data is retrieved online from Ensembl. Mispriming libraries were obtained from RepBase [8] for most of these species. Only in case of R. norvegicus the user has to use the same library as provided for M. musculus. The AutoPrime programm and the mispriming libraries are frequently updated to work on the newest datasets.

Discussion

The best way to find suitable primer pairs for expressed sequences are databases of experimentally validated primers. But this validation also precludes coverage of a large amount of genes and consequently the two primer databases mentioned above currently contain only about several hundred entries each. While automatic generation of primers for expressed sequences is possible, current primer design software is inable to cope with the positional restrictions imposed by the quality requirements for RT-PCR primers. Thus the identification of possible primer sites is often performed by manual selection of a potential sequence, which is audited by a suitable primer design program for any difficulties. This requires manual processing of the sequence data and has the drawback of providing a high number of primers that will be rejected. Even in case the full sequence is provided to the primer design software, the exon border positions still need to be identified and sequentially specified as potential target regions until a suitable pair of primers is found. These difficulties often lead to the violation of sensible primer design constraints and thus to inefficient primer pairs.

AutoPrime provides a solution that tries to optimize the trade-off between quality of the primers and coverage of genes. The software allows the design of RT-PCR primers for all known genes of a large variety of organisms while at the same time following the stringent quality criteria established in the field of expression measurements.

In addition AutoPrime implements an online database access to the Ensembl database. While a variety of primer design tasks exist that need sequence information from biological databases like Genbank or Ensembl as input, nearly all of the available software tools rely on cut and paste operations for entering the sequences of interest. Some tools work with locally stored sequence files for the generation of large numbers of primers (e.g. PrimeArray [9]). Tools like BioPerl [10] and the Ensembl perl modules [6] provide a software interface to biological databases and allow the integration of sequence information into any software tool. AutoPrime employs these facilities to provide an automated process by interfacing Ensembl with a primer design software. Thus it accelerates the primer design process and increases its realiability at the same time. The access to online sequence information also significantly extends the number of organisms covered as compared to PrimerBank.

AutoPrime is being successfully used by scientist and medical personnel in our laboratory. For testing purposes 30 primer pairs were designed and used in RT-PCR expression measurements. 22 of these did yield a positive result and amplified no genomic background which is higher than comparable success rates for “manually” designed primers.

This suggests that AutoPrime is indeed able to provide high quality primers while shortening the design process dramatically.

Acknowledgements

The authors would like to thank Kai Neben and Christian Korz for critical testing of the program.

References

1. Haas S, Vingron M, Poustka A, Wiemann S: **Primer design for large scale sequencing**. *Nucleic Acids Res* 1998, **26**(12):3006–3012.

2. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers**. *Methods Mol Biol* 2000, **132**:365–386.

3. Pattyn F, Speleman F, De Paepe A, Vandesompele J: **RTPrimerDB: the real-time PCR primer and probe database**. *Nucleic Acids Res* 2003, **31**:122–123.

4. Wang X, Seed B: **A PCR primer bank for quantitative gene expression analysis**. *Nucleic Acids Res* 2003, **31**(24).

5. **RealTime-PCR Primer Database**[http://www.realtimeprimers.org/].

6. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M: **The Ensembl genome database project**. *Nucleic Acids Res* 2002, **30**:38–41.

7. **AutoPrime**[http://www.autoprime.de/].

8. Jurka J: **Repbase update: a database and an electronic journal of repetitive elements**. *Trends Genet* 2000, **16**(9):418–420.

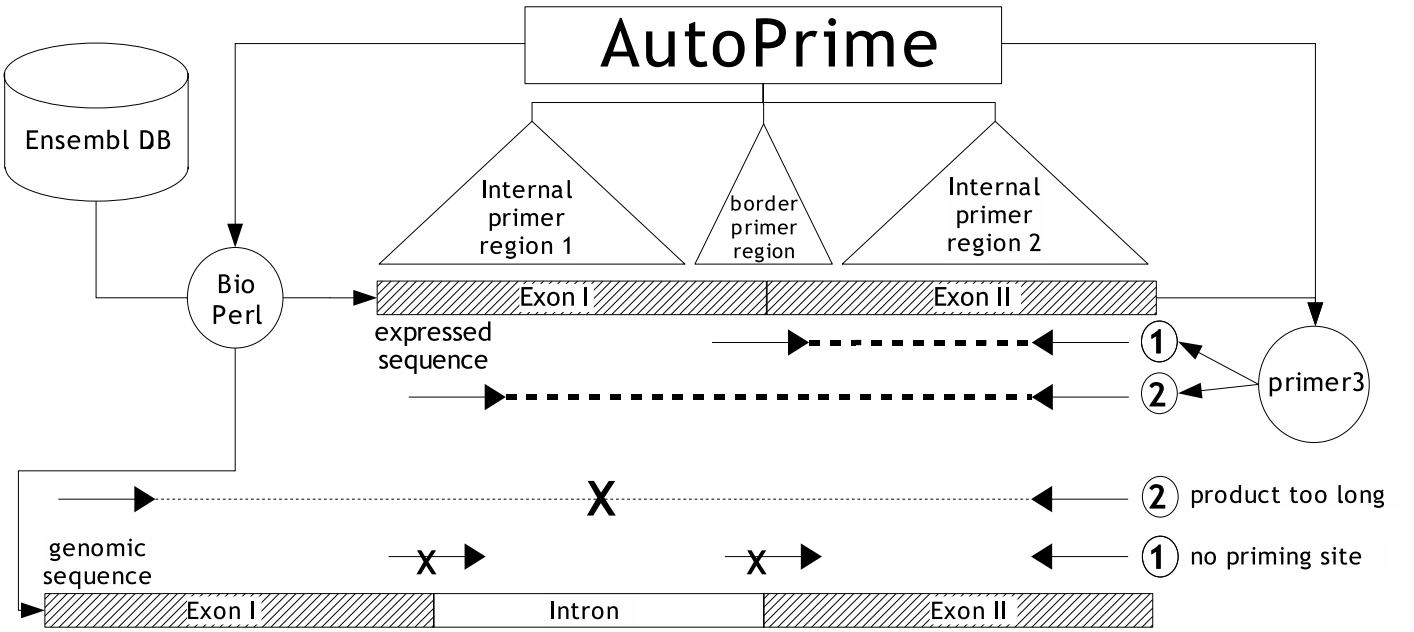
9. Raddatz G, Dehio M, Meyer TF, Dehio C: **PrimeArray: genome-scale primer design for DNA-microarray construction**. *Bioinformatics* 2001, **17**:98–99.

10. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The Bioperl toolkit: Perl modules for the life sciences**. *Genome Res* 2002, **12**(10):1611–1618.

Figures

Figure 1 - Organization of AutoPrime

The tool accesses the Ensembl database using perl modules and instructs the primer design software primer3 to generate primers for specific regions of the expressed sequence. The resulting primers are specific for the mRNA since one of the primers is designed specifically for the exon border sequence (1) or the primers lie in two adjacent exons so that a product from genomic DNA would include the intronic sequence (2)


AutoPrime
Ensembl DB
Internal primer region 1
border primer region
Internal primer region 2
Bio Perl
Exon I
Exon II
expressed sequence
1
2
primer3
2 product too long
genomic sequence
1 no priming site
Exon I
Intron
Exon II