UPS 2.0: Unique Probe Selector for Probe Design and Oligonucleotide Microarrays in Pangenomic/ Genomic Level

Shu-Hwa Chen[1], Chen- Zen Lo[1], Sheng-Yao Su[1], Bao-Han Kuo[1], Chao A. Hsiung[2] and Chung-Yen Lin*[1, 2, 3, 4§]

[1] Institute of Information Science, Academia Sinica, No. 128 Academia Rd., Sec. 2, Taipei 115, Taiwan

[2] Division of Biostatistics and Bioinformatics, National Health Research Institutes. No. 35 Keyan Rd. Zhunan, Miaoli County 350, Taiwan

[3] Institute of Fisheries Science, College of Life Science, National Taiwan University, No. 1, Roosevelt Rd. Sec 4, Taipei, Taiwan

[4] Research Center of Information Technology Innovation, Academia Sinica, No. 128 Yan-Chiu-Yuan Rd., Sec. 2, Taipei 115, Taiwan

[§]Corresponding author

Email addresses:

SHC: Sophia@iis.sinica.edu.tw

CZL: vandomlo@gmail.com

SYS: daniel0523@gmail.com

BHK: phkuo@iis.sinica.edu.tw

CAH: hsiung@nhri.org.tw

CYL: cylin@iis.sinica.edu.tw

Abstract

Background

One of the most popular principles in biomedical research field is nucleic acid hybridization. The performance of any hybridization-based methods depends on the specificity between probes to their targets. To find best, unique probe(s) to detect the target(s) from a sample cocktail, we had developed an algorithm and implemented it into a workflow of probe design and upgrade this method to match the experiment requirement of probe design upon the increasing volume of sequence dataset.

Results

The algorithms and probe parameters applied in UPS 2.0 include the GC content, the secondary structure, Tm, the stability of probe-target duplex estimated by thermodynamic model, the sequence complexity, similarity of probes to non-target sequences, and some other empirical parameters used in lab. Several probe background options, ***Unique probe within group, Unique probe in the specific Unigene set, Unique probe based on pangenomic level, and Unique Probe in the user-defined genome/ transcriptome,*** are available to meet the scenarios that the experiments will be conducted. Parameters such as salt concentration, the lower-bound Tm of probes are opened for users to optimize their probe design query. The output files are ready for download in the result page. Probes designed by UPS algorithm are suitable for generating microarrays and probes had been designed by UPS 2.0. The performance validated by experiments.

Conclusions

UPS 2.0 evaluates probe-to-target hybridization under a user-defined condition to ensure high-performance hybridization with minimal chance of non-specific binding

in pangenomic and genomic level. It mimic the target/ non-target mixture in an experiment and are very useful in developing diagnostic kits and microarrays. There are more than 1,300 visits with 360,000 sequences performed the analysis through UPS 2.0 in the past 30 months. It is freely accessible by http://array.iis.sinica.edu.tw/ups/.
Screen cast: http://array.iis.sinica.edu.tw/ups/ups_demo_en/ups_view.htm

Keywords: Probe design, thermodynamics, hybridization, microarray, Oligonucleotide

Background

One of the most popular principles in biomedical research field is nucleic acid hybridization. Several methods are derived from this essential principle to detect targets by sequence-specific probes, *e.g.*, northern blot, southern blot, and in situ hybridization. The experimental condition of hybridization is optimized and miniaturized into microarray to detect the transcriptional activity of thousands of genes simultaneously [1]. The performance of these widely applied methods depends on the specificity between probes to their targets. Actually, the selection of suitable oligonucleotide probes remains a bottleneck of the microarray workflow [2]. To find best, unique probe(s) to detect the target(s) from a sample cocktail, we developed an algorithm and implemented it into a workflow of probe design, the UPS [3]. The parameters used in UPS include GC content, GC clamps, thermodynamic theory model of duplex stability, and the secondary structure of the probes, low-complexity mask, and some other empirical preferences of wet-lab researchers. This probe design tool ponders the problem of background noise of hybridization.

Several tools are available for probe design. OligoArray 2.1 [4, 5], GoArrays [6], OligoPicker [7], ArrayOligoSelector [8], YODA [9], ProbeSelect [10], and Picky [11] are stand-alone tools running in command mode and provide unique probes for input sequences under the constraints of GC%, Tm, the absence of low complexity and position near the 3'end. OligoWiz2 [12] is a java-based server-client solution for probe design in a graphical user interface. Oligodb[13] and OligoArrayDb are databases providing pre-calculated probes for human/ mouse transcriptome and sequenced Archaea/ Bacteria/ Eukaryote species, respectively. Golfier et al. [14] developed an informatics tool named SOL which updates the mouse sequences in NCBI/ RefSeq semi-automatically and provides a collection of mouse oligonucleotide probe candidates. Researchers may set up a SOL themselves and refine the parameters to work on their probe design tasks on other species. Although most of these tools are useful for designing nucleotide probes, few of them take the background noise of cross-hybridization into consideration, or require users' skills to set up and optimize the tool for their experiment. Thus, we implemented a user-friendly, web-based tool UPS [3] for designing probes with small likelihoods of forming non-specific duplexes to the other submitted targets in a same query batch or to the sequence set of a NCBI/ Unigene listed species, and upgrade its capability of customized reference sequence set in the present version.

Recent advances in sequencing method, the next generation sequencing (NGS) technology, allow researchers to decrypt of gene contexts of unknown genome faster and cheaper. The basic principles of these methods are running millions of sequencing reaction in parallel, generating short sequence reads in the gigabase-scale, and harvesting assembled contigs by algorithms. These methods are successfully applied in genome re-sequencing and have been expended to resolve novel genomes and

multiple genome mixture (metagenome). The application of sequencing a transcriptome for resolving individual transcriptome is also available. Since both time and cost of sequencing a transcriptome or a genome are greatly reduced, researchers are attracted and are willing to take a sequencing approach to leap over blockages of conducting a classical genome/ EST library sequencing project. Thus, the background of a hybridization experiment is in higher possibility to know in prior even if the target organism is not a well-studied species. The NGS assembles also can be raw material to design probes for array fabrication like 16S rRNA gene-based diagnostic arrays [15]. There are several widely used tools like ORMA[16], ARB[17], and PRIMROSE[18] for the classification and the phylogenetic analysis of bacterial species. These tools operate the probe design process upon the data renewing from the databases for the specific molecular targets (i.e. 16S rRNA) and on the basis of the phylogenesis of the species under analysis. However, each program adopts its own strategy, which makes the choice of the best solution difficult.

As we mentioned previously, the algorithm of UPS is for finding best probes of input target sequences using parameters of sequence uniqueness and the duplex stability to achieve good hybridization uniformity (Figure 1). In this updated version UPS 2.0, we add two new options for probe designing, and upgrade the reference sequence database and server hardware framework. The newly added option "Unique Probe based on pangenomic level" allows users using environmental nucleotide sequences (Env_NT) or non-redundant NCBI nucleotides database (NCBI_NT) as the potential background noise source to design probes for submitted target sequences. Probes designed in this option will be adequate to detect sequences from samples with mixed genomes or transcriptomes such as specimens in metagenomic study. The other new function, "Unique Probe in the user-defined genome/ transcriptome", will select

probes for target sequences using the genome/ transcriptome sequence set (in FASTA format) as the background reference uploaded together with the probe designing query. The background reference in option “Unique probe in the specific Unigene set” is updated; currently, 128 species derived from Unigene (on Feb, 2010) are listed in this option. Some probe parameters, like the maximum number of best probe for each submitted sequence, GC%, salt concentration, the lower-bound melting temperature, are opened for adjusting with more flexibility. Regarding to the increasing of sequence entries that a user may submit in a query, the redundancy of sequence ID is checked. The outputs of UPS 2.0 are presented in several layout formats which may fit the special need of usage (Figure 2).

Results

Criteria for Probe Selection in UPS 2.0

To reduce the possibility of non-target hybridization, we design the workflow for finding best probes of input sequences regarding to parameters of sequence uniqueness and duplex stability. The Nearest-Neighbor model[19] and blastn were applied to identify probe candidates for each input sequence with a small likelihood of cross-hybridization. First, probes with single, di- or tri- nucleotide repeats are excluded as described previously. Next, the similarity of probes to non-target sequence is evaluated. We uses criteria established by Li et al [20] to further exclude unsuitable oligonucleotides: 1) the overall similarity of a probe to non-target sequences should below 85%; 2) a continuous identical stretch of a probe to non-target sequences should not exceed 17 bases. Both rules integrated with blastn were called as *in silico* hybridization in this study. Furthermore, the uniqueness of probes is evaluated on the frequency table built from the reference sequence set given in one of the three customized options. The simplest condition is finding unique probes for each

sequence among the input sequence set (option “***Unique probe within group***”). The second option “***Unique probe in the specific Unigene***” is for devising unique probes for each submitted sequence with the least chance to cross-hybridize to the selected Unigene set. For the request from cooperatiors conducting metagenomic studies, we developed the new option, "***Unique Probe based on pangenomic level***", which allows users to design probe based on environmental nucleotide sequences (Env_NT) or non-redundant NCBI nucleotides database (NCBI_NT). Another novel option, “***Unique Probe in the user-defined organism***”, allows users to design probes by taking uploaded sequences as the background reference. The uploaded reference sequence set is up to 30 Mb in file size and up to 5,000,000 bases for a single sequence entry. Probe candidates passing the criteria, eg., Tm cutoff, GC%, the proximity of 3’ end, in silico hybridization (sequence complexity, homology to non-target sequence), are ranked by the E-value of probe-target duplex (blastn) using the scoring matrix built from the customized reference sequence. More details for *in silico* hybridization can be found at UPS 2.0 in the online HELP page (http://array.iis.sinica.edu.tw/ups/help1.html).

Applications and Experimental Validation

Scenario 1. Best probes for detecting specific targets from PCR fragments

Using common primer sets, such as a consensus primer set targeting to amplify small subunit rRNA gene region, a bundle of PCR products could be amplified from a sample with mix microbial genomes. To detect the presence of some particular species, a dot blotting assay can be developed based on a small set of probes with discrimination power. Using the UPS option “Unique Probe within group”, user can find unique probes for each sequence with less possibility to match non-target sequences in the uploaded sequence set. Such probe set is suitable for detecting

malicious pathogen on environmental samples or food contaminants. If the target sequences are mixed with various genomes, such as sludges from sanitary sewers, we can use the option “Unique Probe based on pangenomic level” to assign all known environmental sequences as the reference background. The assay could be started with degenerated primer sets to amplify the target sequences, following with a hybridization on a small array to provide sequence-specific signals.

Scenario 2. Finding probes to detect exogenous/ pathogenous genes on a Unigene- listed species

The spatiotemporal activity of a pathogenic genome is an essential issue. If we want to design probes for detecting pathogen sequences on a specimen from a particular organism that is in the Unigene species list (eg., human) or closely related to a species in Unigene, we can use pathogen sequences as input set in the option “Unique Probe in specific Unigene set” and assign the sequence set of the species (Homo_sapiens, in this case) as the reference background. UPS 2.0 will help us to find suitable probes for identifying pathogen sequences with low the false-positive hits on host sequences.

Scenario 3. Unique probes working on the user-uploaded sequence(s) as reference background

Unigene collects EST clustering from more than a hundred species; however, partially or unpublished sequence assemblies from species other than those in Unigene list are possible available in individual labs. If we want to design probes working in a biopsy from a particular organism, such as detecting pathogen’s transcripts, we can assign the background by uploading a set of reference sequences (a file in FASTA format) of this species in “Unique Probe in the user-defined organism”. Thus, pathogen-specific probes will be selected using UPS criteria which ensure the less chance of cross-hybridization of probe to non-target sequence.

Verification by Experiment

We are running a cooperating project with researchers who are devoted to shrimp virology. Our cooperators had carried out intensive sequencing tasks for deciphering the shrimp virus (WSSV) genome [21] and shrimp transcriptome [22]. The major purpose of this project is to monitor the expression of viral genes through the whole infection process. Here we utilized the new strategy on UPS2.0 to design probes for a shrimp/pathogen microarray. Briefly, EST reads from Penaeus monodon (~25,000 sequences, public database + unpublished data) were assembled into ~7,000 clusters (contigs). Together with 701 putative viral ORFs, these sequences were submitted to UPS 2.0 on the option "*Unique Probe within group*" to produce customized microarray in 44k format (six probe replicates for each submitted sequence) manufactured by in situ synthesis technology (Agilent). The mRNA was extracted from the time-coursed samples (the infected shrimp tissue, gills in this case, from 0 to 48 hours) of viral challenging experiments and the gene profiling was sketched using the microarray platform. Our cooperators randomly picked several viral ORFs to check the correlation between array data and QPCR data (Figure 3). In most of the cases, the correlation of gene expression level quantified by array and by QPCR is splendid, evaluated by the correlation of two expression profiles (Pearson's correlation coefficient) and the linear correlation R^2 of each probe/gene pairs. The coefficients of linear regression of gene-to-probe pairs are above 0.7 and are highly correlated (Pearson's correlation coefficient, ranging from 0.86 to 0.99). Text for this sub-section.

…

Discussion and Conclusions

UPS 2.0 is updated from the previous version which is the probe design tool to identify the unique segment of each submitted sequence among the query dataset, or

the unique segment of input sequences in the user-defined background. With the flexible and intuitive interface, users can easily customize their own probe design parameters by following the friendly online instruction and fetch the probes with information of quality. Probes designed in UPS 2.0 can be used to detect the target in a classical hybridization-based assay; probes from a batch-wise query can fulfill the requirement for microarray probe design. In this new version UPS 2.0, the Unigene data is updated, and it can take the user-uploaded sequence set for the background reference. To date, more than a hundred of species are available in the current Unigene dataset and number is kept growing. The lately new option allows user to upload the reference set other than the Unigene list. These improvements expand the choice of user-defined reference in probe design.

UPS workflow mimics the condition of target/ non-target mixture in a true experiment to be carried out in lab bench. Thus, probes designed by UPS 2.0 are especially suitable for detecting infectious agents such as bacteria and virus with fewer chances of cross hybridization on the host transcripts, or for detecting a specially selected gene list (gene signatures) by small arrays with low background noise. It is a useful merit in developing a diagnostic kit. A single entry in user-uploaded reference set is allowed to 5x106 bps; that means most prokaryote genomes can serve as the probe working background in UPS 2.0.

UPS has been run stably for more than two years. There are more than 1,300 visits and about 360,000 sequences have been submitted to UPS 2.0. We will continue to polish the performance of UPS by applying more sophisticated algorithm and reducing the calculation time. The spectrum of reference backgrounds have been expanded by incorporating published metagenomic datasets, environmental nr, and

the whole genome sequence set in public domain into customized options to optimize oligonucleotide probe designed for experiment in pangenomic/ genome scale.

Methods

System Implementation

For improving the performance of UPS 2.0, we upgraded the hardware to symmetrical multi-processor (SMP) PCs equipped with quad CPUs (Intel Xeon E5420 2.5GHz) and 16 GB of RAM. The LAMP structure (SUSE Linux Enterprise Server 10, Apache v.2.04, PostgreSQL v.8.2.4, PHP v.5.1.0), to provide the web accessing, file uploading, format checking, result storing/ downloading, job controlling, and the e-mail notifying functions.

All calculations in our core algorithm concerning suffix arrays, constraints for probe selection (GC%, low complexity masking, distance from the 3' end), free energy for probe secondary structure, melting temperature estimated in thermodynamic model, and in silico hybridization based on NCBI blastn [19] with several checking rules, i.e. percentage of similarity and maximum overlaped bases of continued sequences between probe and possibile targets, are performed in Boland Delphi 2006 on the MS-Window 2003 server. To provide the most updated background reference, an automatic procedure to rebuild the suffix arrays for each species from the updated Unigene databases is scheduled.

Parameters in Hybridization

The parameter details of UPS were described in Chen *et al* [3]. Some probe parameters are opened adjustable with more flexibility in this new UPS 2.0 upon the request from users for their applications. Briefly, the probe length was adjustable ranging from 30 to 120 mers in length. The salt concentration is set at 0.58M and is adjustable, ranging from 0 to 1M, to fit the ionic strength in the user-defined condition.

Probe candidates with a low complexity segment(s) such as five or more continual nucleotides and continuous di- /tri-nucleotide repeats are excluded. The acceptable GC% of probes is 30–70% according to empiricism and can be adjusted by user. Further constraints on probe features, the GC% and the lower-bound value of Tm, are available in the output browsing interface. The melting temperature of probes is calculated based on Nearest-Neighbor model [20], and the secondary structure of designed probe is calculated by the Perl program UNAFold.pl [23]. Thermodynamic theory was integrated to estimate the ΔG of probe-target duplexes, probe-non target duplexes and probe secondary structure.

Figure Legends

Figure 1. The concept of UPS 2.0 for pangenomic/ genomic studies.

Figure 2. The output of UPS 2.0. (A) The advanced filter by Tm and GC%, and the probe list with download files. (B) The output for best probe without cross-hybridization in fasta. (C) All selected probes (Top 3) for each sequence. (D) *In silico* hybridization result based on blastn with several checking rules. (E) Detail information in tabular format.

Figure 3. Experimental validation on probes designed by UPS 2.0 by QPCR. The expression level of nine WSSV genes, determined either by QPCR or by microarray, were normalized to the measurement at 48hr post virus infection. The coefficients of linear regression of gene-to-probe pairs are above 0.7. The gene profile of an example, gene E, is plotted in the insert panel. The normalized measurement by QPCR and microarray are highly correlated in gene E (Pearson's correlation coefficient: 0.99) and all the others in this test (data not shown).

Competing interests

The authors declare that they have no competing interests.

Authors' Contributions

SHC and CYL conceptualized the algorithm, design the method, drafted the manuscript together. CZL, SYS and BHK were responsible for the implementation and system integration. SHC and CAH participated in discussion and conceptualization as well as revising the draft. All the authors read and approved the manuscript.

Acknowledgements

The authors would like to thank National Science Council (NSC), Taiwan, for financially supporting this research through NSC 98-2221-E-001-018- and NSC 98-3112-B-400-010- . The authors wish to express gratefulness to Yu-Bin Wang, Jan-Hrong Lu and Chu-Fang Lo from National Taiwan University for their kindness help on experimental validation.

Figures

Figure 1. The concept of UPS 2.0 for pangenomic/ genomic studies.

Figure 2. The output of UPS 2.0.

(A) The advanced filter by Tm and GC%, and the probe list with download files. (B) The output for best probe without cross-hybridization in fasta. (C) All selected probes (Top 3) for each sequence. (D) *In silico* hybridization result based on blastn with several checking rules. (E) Detail information in tabular format.

Figure 3. Experimental validation on probes designed by UPS 2.0 by QPCR. The expression level of nine WSSV genes, determined either by QPCR or by microarray, were normalized to the value of 48hr post virus infection. The coefficients of linear regression of gene-to-probe pairs are above 0.7. The gene profile of an example, gene E, is plotted in the insert panel. The normalized measurement by QPCR and microarray are highly correlated in gene E (Pearson's correlation coefficient: 0.99) and all the others in this test (data not shown).

References

1. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science (New York, NY)* 1995, **270**(5235):467-470.
2. Lemoine S, Combes F, Le Crom S: **An evaluation of custom microarray applications: the oligonucleotide design challenge**. *Nucleic acids research* 2009, **37**(6):1726-1739.
3. Chen SH, Lo CZ, Tsai MC, Hsiung CA, Lin CY: **The unique probe selector: a comprehensive web service for probe design and oligonucleotide arrays**. *BMC bioinformatics* 2008, **9 Suppl 1**:S8.

4. Rouillard JM, Herbert CJ, Zuker M: **OligoArray: genome-scale oligonucleotide design for microarrays**. *Bioinformatics (Oxford, England)* 2002, **18**(3):486-487.
5. Rouillard JM, Zuker M, Gulari E: **OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach**. *Nucleic acids research* 2003, **31**(12):3057-3062.
6. Rimour S, Hill D, Militon C, Peyret P: **GoArrays: highly dynamic and efficient microarray probe design**. *Bioinformatics (Oxford, England)* 2005, **21**(7):1094-1103.
7. Wang X, Seed B: **Selection of oligonucleotide probes for protein coding sequences**. *Bioinformatics (Oxford, England)* 2003, **19**(7):796-802.
8. Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, DeRisi JL: **Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray**. *Genome biology* 2003, **4**(2):R9.
9. Nordberg EK: **YODA: selecting signature oligonucleotides**. *Bioinformatics (Oxford, England)* 2005, **21**(8):1365-1370.
10. Li F, Stormo GD: **Selection of optimal DNA oligos for gene expression arrays**. *Bioinformatics (Oxford, England)* 2001, **17**(11):1067-1076.
11. Chou HH, Hsia AP, Mooney DL, Schnable PS: **Picky: oligo microarray design for large genomes**. *Bioinformatics (Oxford, England)* 2004, **20**(17):2893-2902.
12. Wernersson R, Nielsen HB: **OligoWiz 2.0--integrating sequence feature annotation into the design of microarray probes**. *Nucleic acids research* 2005, **33**(Web Server issue):W611-615.
13. Mrowka R, Schuchhardt J, Gille C: **Oligodb--interactive design of oligo DNA for transcription profiling of human genes**. *Bioinformatics (Oxford, England)* 2002, **18**(12):1686-1687.
14. Golfier G, Lemoine S, van Miltenberg A, Bendjoudi A, Rossier J, Le Crom S, Potier MC: **Selection of oligonucleotides for whole-genome microarrays with semi-automatic update**. *Bioinformatics (Oxford, England)* 2009, **25**(1):128-129.
15. Stedtfeld RD, Wick LM, Baushke SW, Tourlousse DM, Herzog AB, Xia Y, Rouillard JM, Klappenbach JA, Cole JR, Gulari E *et al*: **Influence of dangling ends and surface-proximal tails of targets on probe-target duplex formation in 16S rRNA gene-based diagnostic arrays**. *Applied and environmental microbiology* 2007, **73**(2):380-389.
16. Severgnini M, Cremonesi P, Consolandi C, Caredda G, De Bellis G, Castiglioni B: **ORMA: a tool for identification of species-specific variations in 16S rRNA gene and oligonucleotides design**. *Nucleic acids research* 2009, **37**(16):e109.
17. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G *et al*: **ARB: a software environment for sequence data**. *Nucleic acids research* 2004, **32**(4):1363-1371.
18. Ashelford KE, Weightman AJ, Fry JC: **PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database**. *Nucleic acids research* 2002, **30**(15):3481-3489.
19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *Journal of molecular biology* 1990, **215**(3):403-410.

20. Li X, He Z, Zhou J: **Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation**. *Nucleic acids research* 2005, **33**(19):6114-6123.
21. Leu JH, Yang F, Zhang X, Xu X, Kou GH, Lo CF: **Whispovirus**. *Current topics in microbiology and immunology* 2009, **328**:197-227.
22. Leu JH, Chen SH, Wang YB, Chen YC, Su SY, Lin CY, Ho JM, Lo CF: **A Review of the Major Penaeid Shrimp EST Studies and the Construction of a Shrimp Transcriptome Database Based on the ESTs from Four Penaeid Shrimp**. *Mar Biotechnol (NY)* 2010.
23. Markham NR, Zuker M: **DINAMelt web server for nucleic acid melting prediction**. *Nucleic acids research* 2005, **33**(Web Server issue):W577-581.