(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 September 2013 (12.09.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/134341 A1

(51) International Patent Classification:
G06F 19/22 (2011.01)

(21) International Application Number:
PCT/US2013/029268

(22) International Filing Date:
6 March 2013 (06.03.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/607,630 7 March 2012 (07.03.2012) US

(71) Applicant: **DOW AGROSCIENCES LLC** [US/US]; 9330 Zionsville Road, Indianapolis, IN 46268 (US).

(72) Inventors; and

(71) Applicants : **THOMAS, Adam J.** [US/US]; 509 Lawndale Dr, Plainfield, IN 46168 (US). **BUYYARAPU, Ramesh** [IN/US]; 6103 Eagles Nest Blvd, Zionsville, IN 46077 (US). **GANDRA, Premchand** [IN/US]; 2061 Norcross Cr, Apt C, Indianapolis, IN 46260 (US). **ARORA, Kanika** [IN/US]; 2903 Spring Meadow Ct, Indianapolis, IN 46268 (US). **ELANGO, Navin** [IN/US]; 8385 Chadwood Lane West Drive; Apt. 1B, Indianapolis, IN 46268 (US). **PERIANAYAGAM, Rajesh** [IN/US]; 2849 Spring Meadow Ct, Indianapolis, IN 46268 (US). **LU, Fang** [CN/US]; 3442 Burlingame Blvd, Westfield, IN 46074 (US).

(74) Agent: **LEE, Yung-Hui**; DOW AGROSCIENCES LLC, 9330 Zionsville Rd, Indianapolis, Indiana 46268 (US).

[Continued on next page]

(54) Title: PRIMER DESIGNING PIPELINE FOR TARGETED SEQUENCING

(57) Abstract: Provided are systems and methods for customized primer/amplicon designing programs which enable users to design overlapping primers/amplicons in a target region or multiple targeted regions. The target region can be a small (<1 kb) to large contiguous or repeat-masked regions (even entire genomes provided sufficient hardware and memory to handle the processes). The systems and methods provided herein can be used to design multiple sets of overlapping or non-overlapping primers/amplicons for a target region.

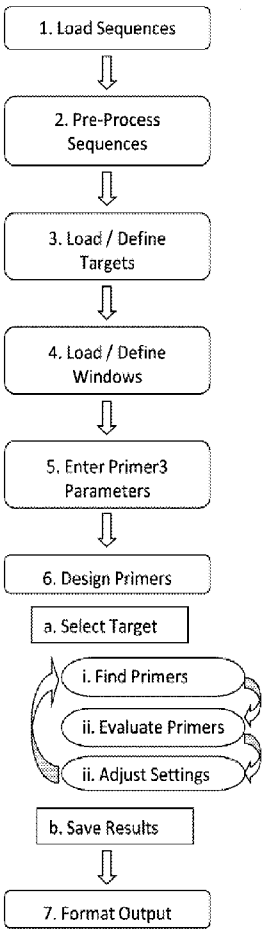


Figure 1

WO 2013/134341 A1

(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

PRIMER DESIGNING PIPELINE FOR TARGETED SEQUENCING

FIELD OF THE INVENTION

[0001] This invention is generally related to the field of molecular biology, and more specifically the field of primer design for targeted/high-throughput sequencing.

BACKGROUND OF THE INVENTION

[0002] Primer designing for DNA sequencing is a vital part for modern biological research. Traditional primer designing programs, for example publically available "Primer3," use a single input DNA sequence to process and design optimal primers under specified parameters including primer length, GC content, melting temperature (Tm) and others. However, there is no dynamic primer designing program which can design overlapping primers simultaneously to facilitate coverage across an entire region or multiple targeted regions. In addition, these traditional primer designing programs are not suitable for handling large files with long or multiple sequences. Thus, there remains a need for an efficient primer designing pipeline for targeted sequence spanning a broad region or multiple targeted regions.

SUMMARY OF THE INVENTION

[0003] Provided are systems and methods for customized primer/amplicon designing programs which enable users to design overlapping primers/amplicons in a target region or multiple targeted regions. The target region can be a small (<1 kilo bases or kb) to large contiguous or repeat-masked regions (even entire genomes provided sufficient hardware and memory to handle the processes). The systems and methods provided herein can be used to design multiple sets of overlapping or non-overlapping primers/amplicons for a target region. In some embodiments, the primers designed using the systems and methods provided herein can also be used for multiplex PCR analysis.

[0004] In one aspect, provided is a computerized system for primer/amplicon design for sequencing. The system comprises:

(a) an input device and an output device/interface;

(b) an analysis system interface coupled to memory of a computer;

(c) an operating system optionally comprising a database;

(d) a load sequence module for loading nucleic acid sequences; and

(e) a primer design module for design primer pairs.

[0005] In one embodiment, the system further comprises at least one of format output module, and a BLAST verification module. In a further embodiment, the system further

comprises at least one of format output module, BLAST verification module, and adaptor verification module. In another embodiment, the input device is selected from the group consisting of automated sequencer, sequencing data input device, and sequencing data storage device. In another embodiment, the output interface comprises interface for WebGBrowse or GenomeBrowser.

[0006] In one embodiment, the database described herein contains information selected from the group consisting of genomic sequences, previously generated primers, and sequences for BLAST analysis. In another embodiment, the load sequence module processes sequences in FASTA format. In another embodiment, the load sequence module uses random file access. In another embodiment, the load sequence module does not use sequential file access.

[0007] In one embodiment, the primer design module performs at least one of (1) automatically adding standard 5' tag or tail to each primer; (2) selecting nested primer pairs; (3) selecting primers for multiplexed amplifications; (4) designing a tiling of amplicons across a sequence; (5) picking primers from a reverse-translated amino acid sequence; and (6) selection from multiple primer sets. In another embodiment, the primer design module processes primer design in parallel. In another embodiment, the primer design module does not design primers in a non-parallel or sequential manner. In another embodiment, the primer design module generates primers or processes primer design at a speed greater than 10 primers per minute. In another embodiment, the primer design module generates primers or processes primer design at a speed greater than 100 primers per minute. In a further or alternative embodiment, the primer design module generates primers or processes primer design at a speed between 200 and 500 primers per minute. In another embodiment, the primers constitute overlapping amplicons for sequence assembly. In another embodiment, the primers constitute overlapping amplicons for sequencing and assembly. In a further embodiment, the overlapping region of amplicons comprises at least 50 bp or minimal overlap. In a further embodiment, the overlapping region of amplicons comprises at least 100 bp. In a further embodiment, the overlapping region of amplicons comprises between 100 bp and 1000 bp.

[0008] In another aspect, provided is a method for use in a computerized system for primer/amplicon design for sequencing. The method comprises:

(a) upload sequence data using a load sequence module;

(b) designing multiple primers in parallel using a primer design module; and

(c) outputting primer design through an output interface.

[0009] In one embodiment, the method further comprises pre-processing sequences by modifying sequences before primer design. In a further or alternative embodiment, the method further comprises defining target regions/sequences. In a further or alternative embodiment, the method further comprises defining windows for primer design.

[0010] In one embodiment, the computerized system of the method comprises a system described herein. In another embodiment, the sequence data is larger than 100 kilo bases (kb). In a further or alternative embodiment, the sequence data is larger than 10 Mega bases (mb). In a further or alternative embodiment, the sequence data is between 10 mb and 1 giga bases (gb).

[0011] In one embodiment, the load sequence module processes sequences in FASTA format. In another embodiment, the load sequence module uses random file access. In another embodiment, the method provided further comprises at least one of (1) automatically adding standard 5' tag or tail to each primer; (2) selecting nested primer pairs; (3) selecting primers for multiplexed amplifications; (4) designing a tiling of amplicons across a sequence; (5) picking primers from a reverse-translated amino acid sequence; and (6) selection from multiple primer sets.

[0012] In one embodiment, the method provides primers at a speed greater than 10 primers per minute. In a further or alternative embodiment, the method provides primers at a speed greater than 100 primers per minute. In a further or alternative embodiment, the method provides primers at a speed between 200 and 500 primers per minute. In another embodiment, the primers constitute overlapping amplicons for sequence assembly. In a further embodiment, the overlapping region of amplicons comprises at least 50 bp or minimal overlap. In a further embodiment, the overlapping region of amplicons comprises at least 100 bp. In a further embodiment, the overlapping region of amplicons comprises between 100 bp and 1000 bp.

[0013] In one embodiment, the method further comprises verifying the primers using a BLAST verification module. In another embodiment, the method further comprises verifying secondary structure of primers using an adaptor verification module. In another embodiment, the method further comprises simulating the sequencing using a sequencing simulation module. In another embodiment, the method further comprises an output format module for outputting for visualization using WebGBrowse.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] Figure 1 shows an exemplary flowchart of the Primer Designing Pipeline provided herein.

[0015] Figure 2 shows an exemplary embodiment for overlapping amplicon design for high-throughput sequencing.

[0016] Figure 3 shows an exemplary automated process for the systems and methods provide herein.

[0017] Figure 4 shows an exemplary process for the BLAST verification modules and methods provided herein.

[0018] Figure 5 shows exemplary FASTA sequences to be loaded into the Primer Designing Pipeline provided herein.

[0019] Figure 6 shows an exemplary screen shot when the Primer Designing Pipeline provided loads FASTA files for downstream analysis.

DETAILED DESCRIPTION OF THE INVENTION

[0020] Various algorithms have been descried previous and can be incorporated in the systems and methods provided to design multiple pairs of primers simultaneously. For example, primer design methods have been disclosed in U.S. Patent Nos. 5,512,458, 5,556,749, 6,928,368, 7,565,248, 7,698,069, and 8,014,955; patent applications US2003/0108919, US2003/0215834, US2003/0215834, US2004/0012633, US2005/0032074, US2006/0281105, US2007/0032963, US2010/0070452, US2010/0184067, JP2003079366, JP2005301532, JP2009268360, JP2011004621, JP2011062085, and EP1136932; international patent applications WO2009/063270, WO2009/152336, WO2010/113789, and WO 2011/053241, the content of which are incorporated by reference in their entireties. In some embodiments, a publically available "Primer3" program is incorporated by the systems and methods provided to process the overlapping primer designing task in targeted regions while also combining the utility of Batch-Primer3 using a customized and compiled program. The "Primer3" program has been previously described in Steve Rozen and Helen J. Skaletsky. (2000) "Primer3 on the WWW for general users and for biologist programmers." In: Krawetz S, Misener S (eds) *Bioinformatics Methods and Protocols: Methods in Molecular Biology.* Humana Press, Totowa, NJ, pp 365-386, the content of which is hereby incorporated by reference in its entirety. In some embodiments, the Primer Designing Pipeline provided is programmed using .NET framework and allows multiple "Primer3" processes to be performed in parallel while validating for overlap. In some embodiments, the

Primer Designing Pipeline described herein provides at least one of the advantages below: (1) automatically adding standard 5' tag or tail to each primer; (2) selecting nested primer pairs; (3) selecting primers for multiplex amplifications; (4) designing a tiling of amplicons across a sequence; and (5) picking primers from a reverse-translated amino acid sequence.

[0021] In some embodiments, the systems and methods provided herein enable primer designing especially for "targeted re-sequencing" applications, for example, using high-throughput (HTP) next-generation sequencing (NGS) instruments. The Primer Designing Pipeline provided can be modular to take small to large sequences as input and also allows changes of the amplicon lengths to suit various NGS platforms as requested by users.

[0022] In some embodiments, primers designed using the systems and/or methods provided herein can be used with Fluidigm AccessArray system, a HTP multiplexed amplicon library generation system for efficient and cost-effective generation of sequencing data for further analysis. For targeted re-sequencing projects, the systems and methods described herein can provide HTP overlapping primer design for complementing the utility of Fluidigm AccessArray system for marker development, gene confirmation, transgene region validation for regulatory affairs, QTL mining and genotyping-by-sequencing.

[0023] An exemplary primer design workflow is illustrated in Figure 1. In the Load Sequence step, the users can select sequence files for which they want to design primers. Files are typically not loaded into memory but instead analyzed for random access. First introduced by Bill Pearson and David Lipman in 1988 for representing either nucleotide or amino acid sequences (see Pearson and Lipman, "Improved tolls for biological sequence comparison" (1988) *Proc. Natl. Acad. Sci. USA* 85:2444-2448; the content of which is hereby incorporated by reference in its entirety), the FASTA file format is a common platform for displaying biological sequences. When designing primers, the target sequence(s) can typically be provided in a FASTA format. The FASTA file can contain one or multiple sequences. Each sequence is always preceded by a header line which is prefixed with ">" followed by the ID, description, and/or other pertinent information about the sequence. The sequence information is then listed on subsequent lines and usually wraps (carriage return/line feed) every 70 or 80 characters depending on the program that generated the file. Typically DNA sequences in FASTA format are shown in Figure 5.

[0024] Especially for re-sequence projects and/or high-throughput sequencing, FASTA files can become quite large when dealing with long sequences, a large number sequences, or a combination of both. For example most FASTA files which represent the entire genome of

complex eukaryotes may easily exceed 2 gigabytes. This size issue poses a problem for the primer design process because normal file handling methods involve starting at the beginning and reading the file into memory until the data of interest is found. For example, a sequential file access for a file larger than 2 gigabytes may take longer than 30 seconds to sequentially read to the spot in the file where the data of interest is. Consequently, this process has to take place for each set of primers which need to be designed for traditional primer designing programs.

[0025] Accordingly, provided is a Load Sequence Module which runs the sequence loading processes in parallel. Random file access is combined with sequential file access to speed up the process. Each character in the file resides at a specific addressable location on a disk. In addition to starting at the beginning of the file and reading each character in order (sequential file access), the Load Sequence Module provides a means to access any location in the file at random as long as the address is known. Random file access can speed processes up considerably because the process does not have to read all the characters/data that came before the data in the file that it's interesting in extracting. Instead of being obligated to perform sequential file access faster, the Load Sequence Module provided is able to determine the starting address in the file at which the data of interest is located.

[0026] In some embodiments, the Primer Designing Pipeline provided initially reads the entire file sequentially through once, analyzes it to determine how it's formatted, and stores that analysis in a SQL Server Compact database. Then when the Primer Designing Pipeline provided is designing the primers and needs to extract a sequence from one of the files, the Primer Designing Pipeline provided uses the analysis results stored in the SQL Server Compact database to calculate the location within the file (or address) of the data of interest. Thus, that address is used to extract/read the data using random file access.

[0027] In a further embodiment, when the program initially reads through a file sequentially it breaks the file into blocks where each block has the same line length, and it stores information about how many lines, characters per line, and file start and stop position for each block. For example as shown in Figure 5, the first block would start at file position 0, contain 1 line with 13 characters per line (a newline and carriage return exist at the end of the line), and the block would end at position 12. The block after that would start at file position 13, contain 2 lines with 72 characters long, and end at position 156. Additionally, since Load Sequence Module assumes that each header block is unique, each header block can be loaded into a hash table (dictionary) in memory for quick access when working with

the file. The blocks following the header which contain the sequence are then linked to the header block in the hash table.

[0028] Later when the Primer Designing Pipeline provided needs to access a sequence or sub-sequence within the file, the Primer Designing Pipeline provided would look up the header up in the hash table first. Then it iterates through each block sequentially to see if the sequence starts in that block. If the file format follows the normal FASTA format, then the Primer Designing Pipeline provided should at most only have to check two blocks because there should be only be two blocks for each sequence in the file. Once the block containing the starting position is determined, the position within that block can be calculated because each character takes up one position/byte in the file.

[0029] The challenging part of calculating the starting position is taking into account the newline characters that occur at the end of each line. In some embodiments, the newline character can also be preceded by a carriage return character. In some embodiments, each block is also analyzed for what type of newline characters as well as other whitespace characters occur at the end each line in the block. For example, if the following target "SEQUENCE_2 | Corn Sequence Gene A45: 136,8" (header : start, length) is needed to be extracted, then the Primer Designing Pipeline provided would determine that it fell in the first block following the header. For example, the block starts at file position 222, contains 3 lines with 72 characters per line, and there are 2 ending whitespace characters in each line. Using these statistics the Primer Designing Pipeline provided can divide the sequence position by the number of sequence characters per line: 136/70 = 1 line plus remainder 66. It then takes the dividend and multiplies it by the whitespace character count: 1 x 2 = 2. And then it adds that to the remainder: 66 + 2 = 68. That result is then added to the starting file position for the block to determine the actually starting file position for the sub-sequence. 222 + 68 = 290. The consequence is that the file position is determined random file access which can be used to read the sub-sequence more quickly than with sequentially file access.

[0030] Figure 6 shows an exemplary screenshot illustrating the part of the Load Sequence Module used to load the example FASTA file as shown in Figure 5, where the first block of each section is the header block.

[0031] Back to Figure 1, in the Pre-Process Sequences step, modifications may be added to the original sequences including masking and/or converting bases for methylation. In the Load / Define Targets step, segments for which primers to be designed are define by the user(s) (for example all sequences or only masked regions greater than a specified length) or

loaded from a file such as a GFF file. The GFF format is useful as input files for programs like WebGBrowse, which is previously disclosed in Ram Podicheti, Rajesh Gollapudi, and Qunfeng Dong. (2009) "WebGBrowse - a web server for GBrowse." Bioinformatics, 25(12):1550–1551, the content of which is incorporated by reference in its entirety.

[0032] Next in Figure 1 is the Load / Define Window step, where the area in which primers to be placed is defined (for example 100 bps up and downstream from the target) or loaded from a file. The next step is Enter Primer Parameters, where parameters including primer length, melting temperature, GC content, 3' stability, and/or estimated secondary structure. In some embodiment, Primer3 is used as the primary design engine and use(s) can enter parameters as required by the Primer3 program.

[0033] Provided is a Primer Design Module having two major functions: (1) selecting and processing the target and (2) saving the results. In some embodiments, when processing a target, the algorithm of the Primer Design Module may adjust settings internally. For example with HTP primer design, the Primer Design Module can start at the beginning of the target area and design overlapping primer sets until it reaches the end. In one embodiment, additional evaluation may be needed such as BLASTing or secondary structure prediction before moving to the next set of primers. Basic Local Alignment Search Tool (BLAST) is a commonly used sequence alignment tool. See Altschul *et al.* (1990) *J. Mol. Biol.* 215: 403-410, the content of which is hereby incorporated by reference in its entirety. In some embodiments, the Primer Designing Pipeline provided automatically generates and adds specified adaptor sequences to the designed primers.

[0034] When performing targeted genome sequencing where only specific sub-sequences within a genome are desired, the user(s) must manually design primers to create overlapping amplicons where the targeted region is larger than the maximum read length of the sequencing equipment. To date, most high throughput sequencing machines can only sequentially read a limited number of base pairs in one run. Thus, the source genetic material needs to be chopped up into segments that are less than the maximum read length. In order to assembly these segments back into one sequence, the segments need to have some overlap sequence (usually at least 20 base pairs).

[0035] Figure 2 shows an exemplary embodiment of high-throughput sequencing where a sequencer is used to sequence the target region. The target region is 5,000 base pairs long and the sequencer can only read segments of DNA up to 700 base pairs – *i.e.*, the 5,000 bp sequence needs to be "chopped" in 700 bp (base pairs) or less segments. In reality, the

source sequence isn't chopped but the 5,000 bp sequence is amplified into 700 bp segments for sequencing. In the amplification step of the sequencing process where the multiple copies of the sequence are made to be sequenced, shorter overlapping copies are made instead. After the sequencing machine finishes reading/sequencing all the short copies, the reads are stored in a data file and an assembly program assembles them back into one continuous sequence.

[0036] In order to make the shorter copies or amplicons during the amplification stage of the sequencing, primers, short sequences that mark the beginning and end of an amplicon, have to be designed and created. Traditional tools for designing primers can only design one set of primers (or one amplicon) at a time. These overlapping amplicons have to be designed serially (one after the other) and usually in a fairly manual process. First the user designs primers for the first amplicon using some software, then sees where that amplicon ends and then designs primers for the next amplicon making sure the beginning of that amplicon overlaps the end of the previous. This is very tedious and requires considerable amounts of copying and pasting and calculating overlaps by hand. Additionally for targeted sequencing, there can be more than one target so this design process has to be performed for each target.

[0037] Provided is an automated systems and methods for designing primers for overlapping amplicons. In some embodiments, the automated systems and methods provided start from a traditional primer design program/software, for example Primer3, which can be downloaded locally and run from a command line interface from either a Linux or Windows machine.

[0038] In some embodiments, the automated systems and methods provided are generated using Perl script (parallelized or non-parallelized). In other embodiment, the automated systems and methods provided are generated using Microsoft .NET 4.0., which contains functionality for parallelizing processes. In some embodiments, the systems and methods provided using Microsoft .NET 4.0 can design large batches of primers in a few minutes as compared to hours using non-parallelized Perl script or days with the traditional approaches. In some embodiments, the automated systems and methods provided use a parallelized approach. In some embodiments, the automated systems and methods provided does not use a non-parallelized approach.

[0039] Figure 3 shows an exemplary system provided using Microsoft .NET 4.0. The output from the .NET program is a tab delimited test file containing the primer sequences and information about the quality of the primers. In some embodiments, in addition to the final

file containing the aggregated results, the Primer Designing Pipeline provides also saves copies of the input sent to Primer3 and the output Primer3 generates. This can be useful in troubleshooting any issues that may arise or manually re-running one portion of the process if necessary.

[0040] Back to Figure 1, provided is a Format Output Module which reads the output from the automation program and generate a general feature file (GFF) formatted file that can be used by other programs including GBrowse to visually overlay the primers and amplicons on the source sequence. In some embodiments, the raw results from the Primer Designing Pipeline are compiled and formatted into a tab delimited format as well as optionally a GFF format for feeding into GBrowse for visualization. For one example, 14 pairs of primers/amplicons (*i.e.*, 28 primers) are created within four minutes using the systems and methods provided. These 14 pairs of primers form overlapping amplicons over one broad targeted sequence. For another example, 9 pairs of primers/amplicons (*i.e.*, 18 primers) are created within two minutes using the systems and methods provided. These 9 pairs of primers from overlapping amplicons over two separated targeted sequences, where the two targeted sequences are still within the same genome (*i.e.*, skipping one region in between for sequencing).

[0041] In further embodiments, the systems and/or methods provided also comprise a BLAST Verification Module. An exemplary BLAST Verification Module is illustrated in Figure 4. In one embodiment, the BLAST Verification Module verifies target redundancy for amplicon primer design. In another embodiment, after the previous steps of the primer design process, the BLAST Verification Module will take the outputs and BLAST them against the targeted genome or sequence library available in database. Typically the BLAST Verification Module allows a primer set (pair) or amplicon to be unique based on BLAST analysis. In some embodiments, the BLAST Verification Module provides BLAST analysis in parallel, thus saving time as compared to sequential analysis. Typically one primer from the first set has to be BLASTed then the next one, and then the results from both BLAST queries must be compared to see if both primers land within pre-determined number of base pairs from each other (usually 1000) and are on opposite strands of the DNA pointing the correct direction for amplification to occur. If non-unique primers are found, then those specific sequences need to be re-run through the primer design process with different parameters.

[0042] In some embodiment, the BLAST Verification Module specifies a BLAST

database to use before running the primer design process. As the primer design got back individual results, the Primer Designing Pipeline disclosed can automatically check them for uniqueness and try to re-run that sequence if necessary. In some embodiments, a copy of the BLAST database is created locally on the user's workstation. In other embodiments, the BLAST databases is located on a serve and accessed remotely by the Primer Designing Pipeline.

[0043] In further embodiments, the systems and/or methods provided also comprise an Adaptor Verification Module. In some embodiments, the Primer Designing Pipeline adds adapter/tag sequences to the primers in order for the sequencing machine to be able to sequence the amplicon. This adapter/tag sequence is often the same for all primers. The secondary structure of a designed primer may change significantly after adding such adaptor/tag sequence. Currently there are a few programs capable of analyzing secondary structures for nucleic acids. For example, RNAstructure (for both DNA and RNA) has been developed by the University of Rochester that can be used to predict the most likely binding structure a sequence or pair sequences can make.

[0044] In some embodiments, the Adaptor Verification Module of the Primer Designing Pipeline can automate the RNAstructure program through the command line and enable prediction whether after adding an adapter sequence, a primer set is still a scientifically good choice. In some embodiment, the Adaptor Verification Module comprises an internal scoring system for classifying primers based on the predicted secondary structure. Other programs can be used for the Adaptor Verification Module provided including Mfold (as disclosed in M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31 (13), 3406-3415, 2003) and UNAFold (as disclosed in N. R. Markham & M. Zuker. UNAFold: Software for Nucleic Acid Folding and Hybridization. In *Data, Sequence Analysis, and Evolution*, J. Keith, ed., *Bioinformatics: Volume 2*, Chapter 1, pp 3-31, Humana Press Inc., 2008), the content of both are hereby incorporated by reference in their entireties.

[0045] In addition, simulation of amplification and/or sequencing can be performed. Several programs have been disclosed to simulate the entire sequencing process, for example *in silico* PCR amplification for amplification, MetaSim for simulating sequencing, and CAP3 for assembly. The Primer Designing Pipeline can integrate the ability to simulate the entire sequencing process using a series of simulators.

We claim:

1. A computerized system for primer/amplicon design for sequencing, comprising,
   (a) an input device and an output device/interface;
   (b) an analysis system interface coupled to memory of a computer;
   (c) an operating system comprising a database;
   (d) a load sequence module for loading nucleic acid sequences; and
   (e) a primer design module for design primer pairs.

2. The computerized system of claim 1, further comprising at least one of format output module, BLAST verification module, and adaptor verification module.

3. The computerized system of claim 1, wherein the input device is selected from the group consisting of automated sequencer, sequencing data input device, and sequencing data storage device.

4. The computerized system of claim 1, wherein the output interface comprises interface for WebGBrowse or GenomeBrowser.

5. The computerized system of claim 1, wherein the database contains information selected from the group consisting of genomic sequences, previously generated primers, and sequences for BLAST analysis.

6. The computerized system of claim 1, wherein the load sequence module processes sequences in FASTA format.

7. The computerized system of claim 1, wherein the load sequence module uses random file access.

8. The computerized system of claim 1, wherein the primer design module performs at least one of (1) automatically adding standard 5' tag or tail to each primer; (2) selecting nested primer pairs; (3) selecting primers for multiplexed amplifications; (4) designing a tiling of amplicons across a sequence; (5) picking primers from a reverse-translated amino acid sequence; and (6) selection from multiple primer sets.

9. The computerized system of claim 1, wherein the primer design module processes primer design in parallel.

10. The computerized system of claim 1, wherein the primer design module generates primers at a speed greater than 10 primers per minute.

11. The computerized system of claim 10, wherein the primers constitute overlapping amplicons for sequencing and assembly.

12. A method for use in a computerized system for primer/amplicon design for sequencing, comprising,
(a) upload sequence data using a load sequence module;
(b) designing multiple primers in parallel using a primer design module; and
(c) outputting primer design through an output interface.

13. The method of claim 12, further comprising pre-processing sequences by modifying sequences before primer design.

14. The method of claim 12, further comprising defining target sequences.

15. The method of claim 12, further comprising defining windows for primer design.

16. The method of claim 12, wherein the computerized system comprises a system of claim 1.

17. The method of claim 12, wherein the sequence data is larger than 100 kb.

18. The method of claim 12, wherein the load sequence module processes sequences in FASTA format.

19. The method of claim 12, wherein the load sequence module uses random file access.

20. The method of claim 12, further comprising at least one of (1) automatically adding standard 5' tag or tail to each primer; (2) selecting nested primer pairs; (3) selecting primers for multiplexed amplifications; (4) designing a tiling of amplicons across a sequence; (5) picking primers from a reverse-translated amino acid sequence; and (6) selection from multiple primer sets.

21. The method of claim 12, wherein the method provides primers at a speed greater than 10 primers per minute.

22. The method of claim 21, wherein the primers constitute overlapping amplicons for sequence assembly.

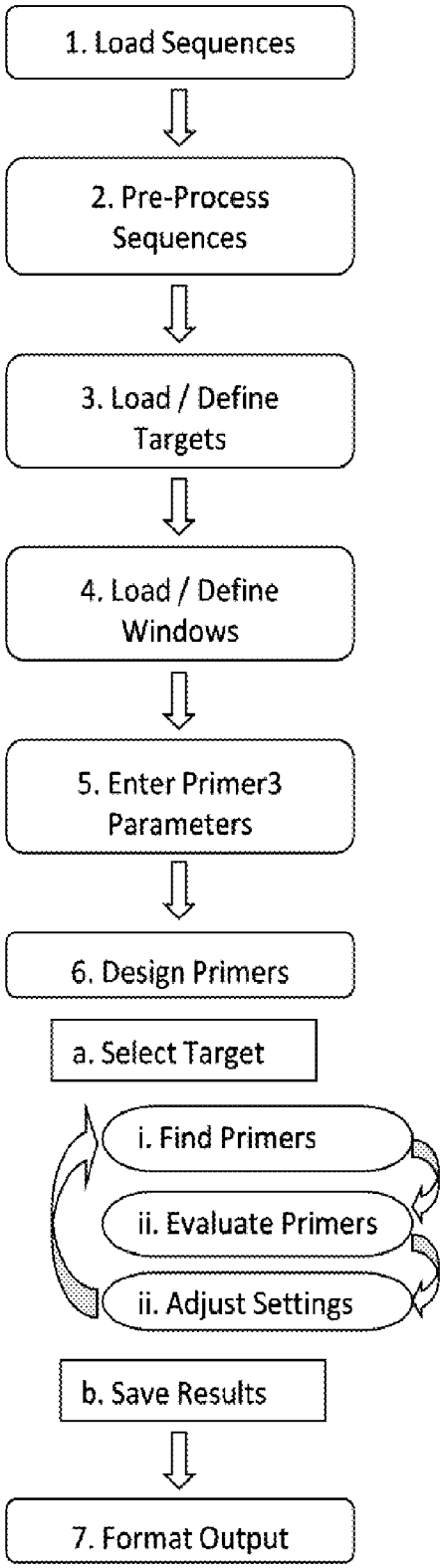

1. Load Sequences
2. Pre-Process Sequences
3. Load / Define Targets
4. Load / Define Windows
5. Enter Primer3 Parameters
6. Design Primers
a. Select Target
i. Find Primers
ii. Evaluate Primers
ii. Adjust Settings
b. Save Results
7. Format Output

Figure 1

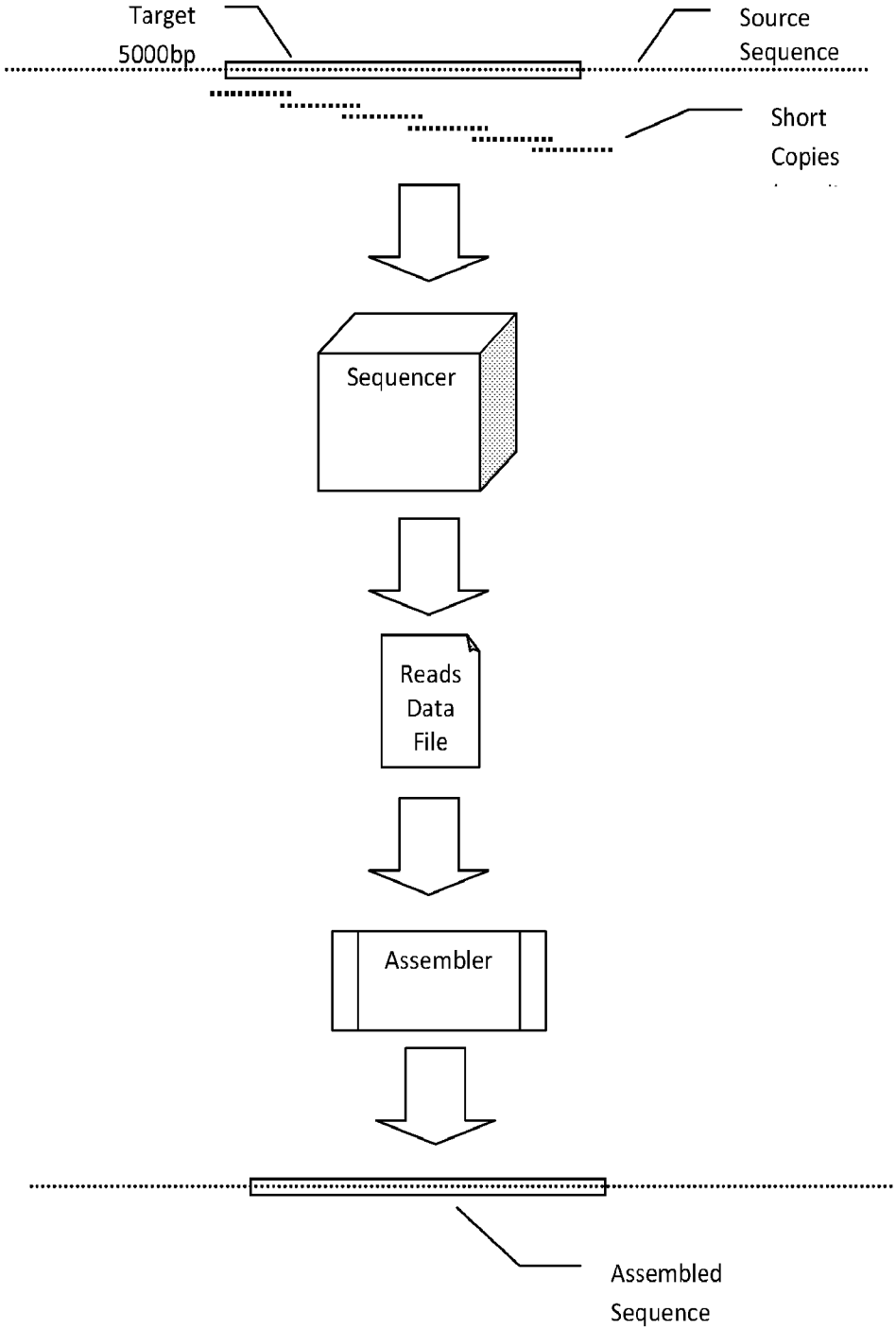

Target
5000bp
Source
Sequence
Short
Copies
Sequencer
Reads
Data
File
Assembler
Assembled
Sequence

Figure 2

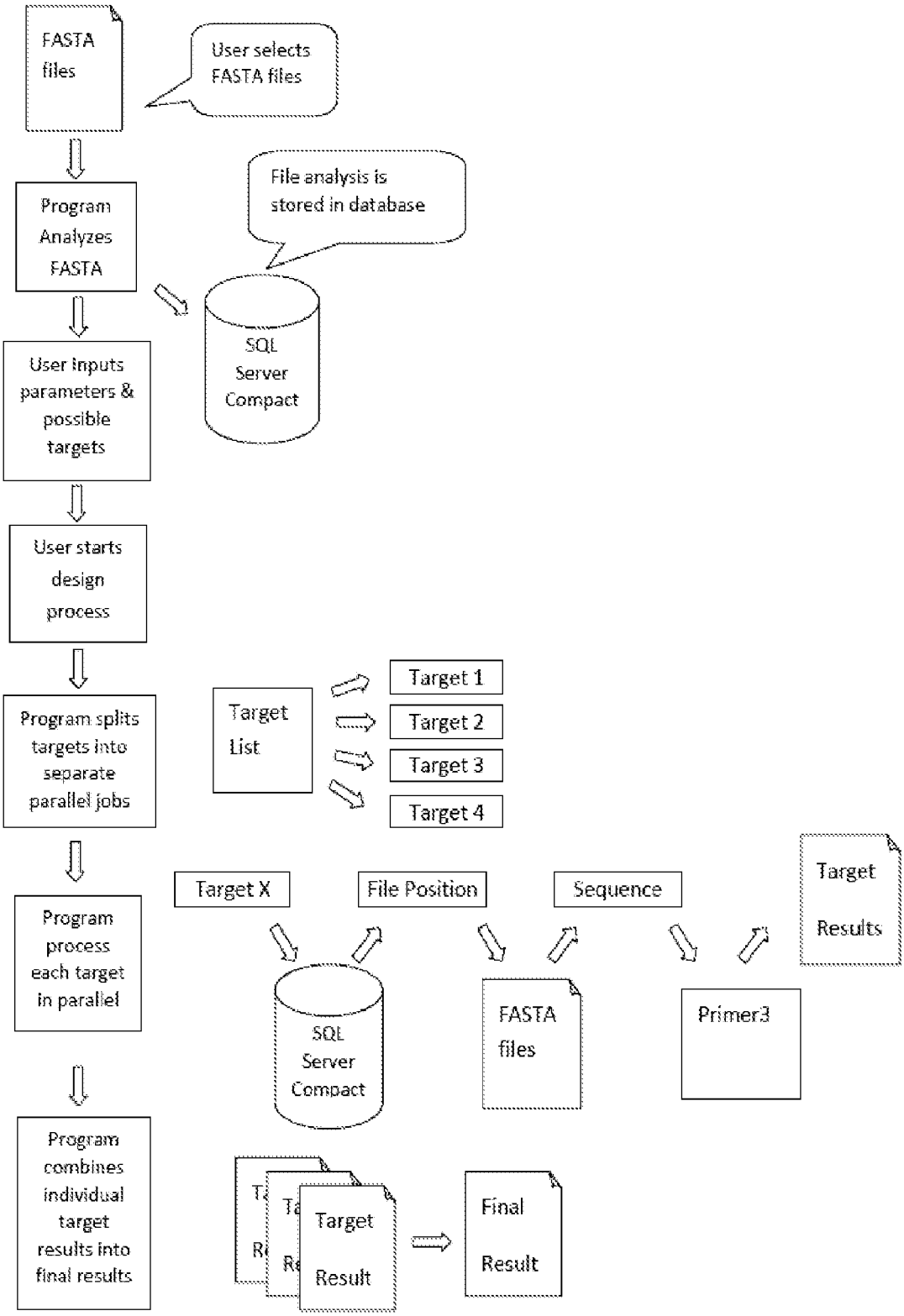

FASTA files
User selects FASTA files
Program Analyzes FASTA
File analysis is stored in database
SQL Server Compact
User inputs parameters & possible targets
User starts design process
Program splits targets into separate parallel jobs
Target List
Target 1
Target 2
Target 3
Target 4
Program process each target in parallel
Target X
File Position
Sequence
Target Results
SQL Server Compact
FASTA files
Primer3
Program combines individual target results into final results
Target Result
Final Result

Figure 3

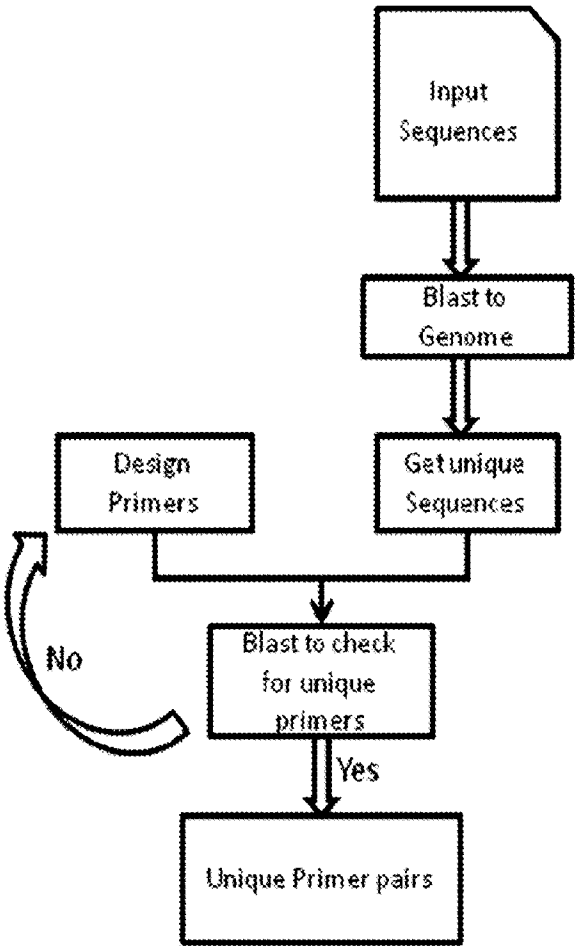
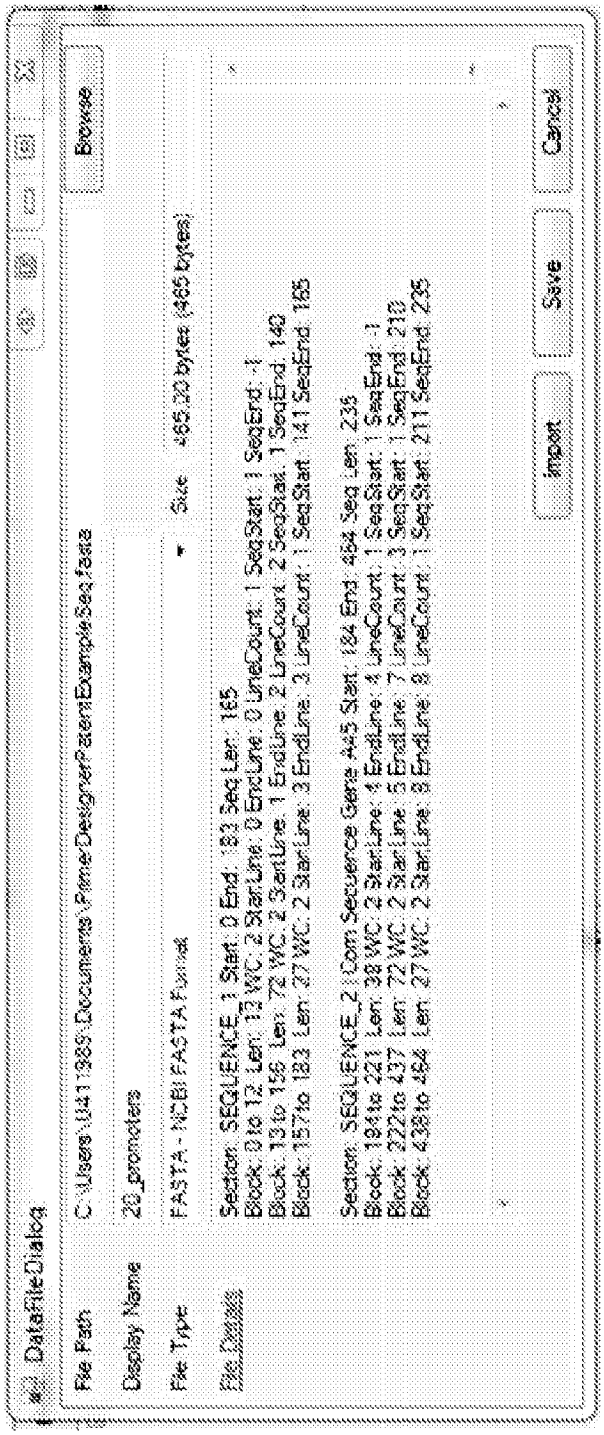

Input Sequences
Blast to Genome
Design Primers
Get unique Sequences
No
Blast to check for unique primers
Yes
Unique Primer pairs

Figure 4

>SEQUENCE_1
ATCGGGTCAATACGGTCAATACGCGGGTCAATACGAATACGCGGGTCAATACGGTCAATAC
AATACGGTCAATACGAATACGCGGGTCAATACGGTCAATACTCAATACGGTCAATACGAACAATACGAAG
ATACGCGGGTCAATACGGTCAATAC
>SEQUENCE_2 | Corn Sequence Gene A45
ATCGGGTCAATACGGTCAATACGCGGGTCAATACGGTCAA TACGAATACGCGGGTCAATACGGTCAATAC
AATACGGTCAATACGAATACGCGGGTCAATACGGTCAATACTCAATACGGTCAATACGAACAATA CGAAG
TACTCAATACGGTCAATACGAACAATACGAAGTACTCAATACGGTCAATACGAACAATACGAAGGGATTC
ATACGCGGGTCAATACGGTCAATAC

Figure 5


DataFileDialog
File Path
C:\Users\U411989\Documents\PrimeDesigner\PrimeExampleSeq.fasta
Browse
Display Name
20_promoters
File Type
FASTA - NCBI FASTA Format
Size
465.20 bytes (465 bytes)
File Details
Section: SEQUENCE_1 Start: 0 End: 83 Seq Len: 165
Block: 0 to 12 Len: 13 WC: 2 StartLine: 0 EndLine: 0 LineCount: 1 SeqStart: 1 SeqEnd: -1
Block: 13 to 156 Len: 72 WC: 2 StartLine: 1 EndLine: 2 LineCount: 2 SeqStart: 1 SeqEnd: 140
Block: 157 to 183 Len: 27 WC: 2 StartLine: 3 EndLine: 3 LineCount: 1 SeqStart: 141 SeqEnd: 165
Section: SEQUENCE_2 (Com Sequence Gene 445 Start: 184 End: 464 Seq Len: 235
Block: 184 to 221 Len: 38 WC: 2 StartLine: 4 EndLine: 4 LineCount: 1 SeqStart: 1 SeqEnd: -1
Block: 222 to 437 Len: 72 WC: 2 StartLine: 5 EndLine: 7 LineCount: 3 SeqStart: 1 SeqEnd: 210
Block: 438 to 464 Len: 27 WC: 2 StartLine: 8 EndLine: 8 LineCount: 1 SeqStart: 211 SeqEnd: 235
Import
Save
Cancel

Figrue 6

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2013/029268

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F19/22
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	LI KELVIN ET AL: "Novel computational methods for increasing PCR primer design effectiveness in directed sequencing", BMC BIOINFORMATICS, BIOMED CENTRAL, LONDON, GB, vol. 9, no. 1, 11 April 2008 (2008-04-11), page 191, XP021031763, ISSN: 1471-2105 the whole document ----- -/--	1-22

[X] Further documents are listed in the continuation of Box C.

[ ] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
30 July 2013	06/08/2013

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Kürten, Ivayla

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2013/029268

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	BROWN ANDREW MK ET AL: "Optimus Primer: A PCR enrichment primer design program for next-generation sequencing of human exonic regions", BMC RESEARCH NOTES, BIOMED CENTRAL LTD, GB, vol. 3, no. 1, 7 July 2010 (2010-07-07), page 185, XP021083073, ISSN: 1756-0500, DOI: 10.1186/1756-0500-3-185 the whole document	1-22
A	GARIMA KUSHWAHA ET AL: "PRIMEGENSw3: A Web-Based Tool for High-Throughput Primer and Probe Design", BIOINFORMATICS AND BIOMEDICINE (BIBM), 2011 IEEE INTERNATIONAL CONFERENCE ON, IEEE, 12 November 2011 (2011-11-12), pages 345-351, XP032087106, DOI: 10.1109/BIBM.2011.43 ISBN: 978-1-4577-1799-4 the whole document	1-22
A	SIMMLER H ET AL: "Real-time primer design for DNA chips", PARALLEL AND DISTRIBUTED PROCESSING SYMPOSIUM, 2003. PROCEEDINGS. INTE RNATIONAL APRIL 22-26, 2003, PISCATAWAY, NJ, USA,IEEE, 22 April 2003 (2003-04-22), pages 153-160, XP010645717, ISBN: 978-0-7695-1926-5 the whole document	1-22
A	KADERALI L ET AL: "Primer-design for multiplexed genotyping", NUCLEIC ACIDS RESEARCH, OXFORD UNIVERSITY PRESS, SURREY, GB, vol. 31, no. 6, 15 March 2003 (2003-03-15) , pages 1796-1802, XP002996256, ISSN: 0305-1048, DOI: 10.1093/NAR/GKG267 the whole document	1-22