



(51) International Patent Classification:

G01N 33/50 (2006.01) C40B 50/02 (2006.01)
C40B 30/02 (2006.01)

(21) International Application Number:

PCT/US2014/048265

(22) International Filing Date:

25 July 2014 (25.07.2014)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

61/858,600 25 July 2013 (25.07.2013) US

(71) Applicant: KBIOBOX INC. [US/US]; 4 Tyson Terrace,
Worcester, MA 01606 (US).(72) Inventors: DALY, Kristopher, Edward; 4 Tyson Terrace,
Worcester, MA 01606 (US). STEPHENS, Kurtis, Phillip;
4 Tyson Terrace, Worcester, MA 01606 (US).(74) Agent: DETWEILER, Sean, D.; Morse, Barnes-brown &
Pendleton, P.c., Citypoint, 230 Third Avenue, 4th Floor,
Waltham, MA 02451 (US).

(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, IR, IS, JP, KE, KG, 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, SA, 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, KM, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: METHOD AND SYSTEM FOR RAPID SEARCHING OF GENOMIC DATA AND USES THEREOF

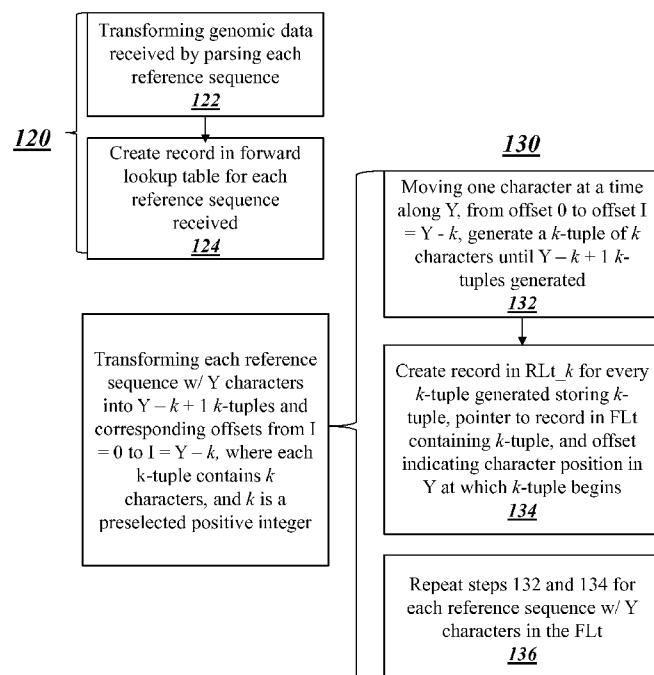


FIG. 2

(57) Abstract: A method, apparatus and system for transforming genomic data into a computer database environment comprising a forward lookup table and a plurality of reverse lookup tables which relate consecutive overlapping reference sequence segments to reference sequences stored in the forward lookup table enables rapid and precise matching of undefined biological sequences with reference sequences.

Published:

(88) Date of publication of the international search report:

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

9 April 2015

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 14/48265

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - G01N 33/50; C40B 30/02; C40B 50/02 (2014.01)

CPC - G06F 19/22; C40B 30/02; C40B 50/02

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)
CPC: G06F 19/22; C40B 30/02; C40B 50/02

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
CPC: G06F 19/22; C40B 30/02; C40B 50/02 (text search)
USPC: 506/8; 506/24; 702/19 (text search)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Electronic data bases: PatBase; Google Scholar; Google Patents
Search terms: unknown DNA, identify or classify, K-mer, reference database or library, catalog, hash table, forward or reverse look-up table, non-transitory memory (e.g. hard disk drive, SSD, computer readable storage medium), nucleic acid, thread manager

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	ENSEMBL. Ensembl FTP Download [online] 23 March 2009 [retrieved 9 January 2015]. Available on the internet: <URL: https://web.archive.org/web/20090323074310/http://www.ensembl.org/info/data/ftp/index.html >. Especially pg 2	1-15, 43
Y ----- A	US 2009/0270277 A1 (GLICK et al.) 29 October 2009 (29.10.2009). Especially para [0009], [0019], [0020], [0021], [0025]	1-15, 43 ----- 16-42, 44
Y	DeSantis et al. Simrank: Rapid and sensitive general-purpose k-mer search tool. BMC Ecol 27 April 2011 Vol 11 No 11 Pages 1-8. Especially pg 3 col 1 para 1.	4, 43/4
Y ----- A	US 2012/0004111 A1 (COLWELL et al.) 05 January 2012 (05.01.2012). Especially para [0190].	5-11, 43/(5-11). ----- 42
A	US 6,728,960 B1 (LOOMANS) 27 April 2004 (27.04.2004). Especially col 4 ln 9-21.	42

☐ Further documents are listed in the continuation of Box C.

* 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

05 January 2015 (05.01.2015)

Date of mailing of the international search report

22 JAN 2015

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-3201

Authorized officer:

Lee W. Young

PCT Helpdesk: 571-272-4300

PCT.OSP: 571-272-7774

Continuation of Box III (Lack of Unity of Invention)

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fees must be paid.

Group I: Claims 1-15 and 43 drawn to a computer-implemented method comprising:

- a) receiving, via a data processing system, genomic data comprising a plurality of reference sequences;
- b) constructing, via the data processing system, a forward lookup table and at least one reverse lookup table which relates reference sequence segments stored therein to reference sequences stored in the forward lookup table, wherein the reference sequence segments each comprise a first string of contiguous characters having a string length equal to a preselected positive integer k and each reference sequence comprises a second string of contiguous characters having a string length equal to a positive integer Y , wherein constructing the forward lookup table comprises transforming the genomic data into a plurality of records in the forward lookup table, wherein transforming comprises parsing the genomic data to create a reference sequence record in the forward lookup table for each reference sequence in the genomic data, wherein constructing the at least one reverse lookup table comprises: i) transforming each reference sequence record in the forward lookup table into a set of $Y - k + 1$ consecutive overlapping reference sequence segments; and ii) storing each reference sequence segment thus transformed in a record in the at least one reverse lookup table; and
- c) storing, via the data processing system, the forward lookup table and the at least one reverse lookup table in a computer database environment maintained in a non-transitory computer readable storage medium and accessible by the data processing system.

Group II: Claims 16-42 and 44, drawn to a computer-implemented method comprising:

- a) receiving, via a data processing system, a query comprising at least one query sequence comprising a third string of contiguous characters having a string length equal to L ;
 - b) selecting, via the data processing system, at least a first reverse lookup table of at least one reverse lookup table in a computer database environment that has the largest preselected positive integer k that is less than or equal to L , wherein the computer database environment is maintained in a non-transitory computer readable storage medium and associated with the data processing system, wherein the computer database environment comprises a forward lookup table and the at least one reverse lookup table which relates reference sequence segments stored therein to reference sequences stored in the forward lookup table, wherein the reference sequence segments each comprise a first string of contiguous characters having a string length equal to a preselected positive integer k which differs for each of the at least one reverse lookup tables and each reference sequence comprises a second string of contiguous characters having a string length equal to a positive integer Y , and wherein each of the at least one reverse lookup tables comprises a set of $Y - k + 1$ consecutive overlapping reference sequence segments for each reference sequence record stored in the forward lookup table;
 - c) transforming, via the data processing system, the at least one query sequence into at least a first set of L/k consecutive non-overlapping query sequence segments by partitioning the third string of contiguous characters into L/k consecutive non-overlapping segments where k is equal to the preselected positive integer k for the at least the first reverse lookup table selected in b);
 - d) querying, via the data processing system, the at least the first set of L/k consecutive non-overlapping query sequence segments transformed in step c) against each set of $Y - k + 1$ consecutive overlapping reference sequence segments in the at least the first reverse lookup table selected in b) for matching reference and query sequence segments; and
 - e) outputting, via the data processing system, at least one reference sequence that matches the at least one query sequence.
- Group II may also comprise (claim 42) iii) gap filling, for each set of equally sized matching reference sequence segments via the third at least one thread, to match the at least one query sequence with at least one reference sequence in the forward lookup table.

The inventions listed as Groups I-II do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

Special Technical Features:

Group I has the special technical features of receiving genomic data comprising a plurality of reference sequences and CONSTRUCTING a forward lookup table and at least one reverse lookup table which relates reference sequence segments stored therein to reference sequences stored in the forward lookup table from the genomic data, not required by Group II.

Group II has the special technical feature of receiving a query comprising at least one query sequence comprising a third string of contiguous characters having a string length equal to L and USING, via the data processing system, at least a first reverse lookup table of at least one reverse lookup table in a computer database environment that has the largest preselected positive integer k that is less than or equal to L , and elements c), d) and e) of claim 16, as listed above, not required by Group I.

-----continued on next sheet-----

continuation of previous sheet

Common Technical Feature:

Groups I and II share the common technical features of a forward lookup table and at least one reverse lookup table which relates reference sequence segments stored therein to reference sequences stored in the forward lookup table, wherein the reference sequence segments each comprise a first string of contiguous characters having a string length equal to a preselected positive integer k and each reference sequence comprises a second string of contiguous characters having a string length equal to a positive integer Y wherein the forward lookup table comprises transforming the genomic data into a plurality of records in the forward lookup table, wherein transforming comprises parsing the genomic data to create a reference sequence record in the forward lookup table for each reference sequence in the genomic data and wherein the at least one reverse lookup table comprises: i) transforming each reference sequence record in the forward lookup table into a set of $Y - k + 1$ consecutive overlapping reference sequence segments; and ii) storing each reference sequence segment thus transformed in a record in the at least one reverse lookup table; and c) storing, via the data processing system, the forward lookup table and the at least one reverse lookup table in a computer database environment maintained in a non-transitory computer readable storage medium and accessible by the data processing system. Group II is related to Group I as a method of using the constructed and stored forward look-up table and reverse look-up table to query a third string of contiguous characters.

However, said common technical feature do not represent a contribution over the prior art and is obvious over US 2009/0270277 AI to GLICK et al. (hereinafter "Glick").

Glick teaches a file equivalent to a "first lookup table" wherein the reference sequence segments each comprise a first string of contiguous characters having a string length equal to a preselected positive integer k and each reference sequence comprises a second string of contiguous characters having a string length equal to a positive integer Y (para [0025]; "In the present invention, a nucleic acid sequence is indexed by scanning the sequence with a window of a selected width K (width being the number of bases). Each sequence of width K that has a specific combination of bases is a unique K -mer. Each unique K -mer is repeated multiple times in a sequence. For example, when $K=4$, examples of some unique K -mers could be ACCT, ACTG, ACTT, TCGA, TTCG, etc., with each of these unique K -mers being repeated multiple times in a DNA sequence. For a sequence of length L "). Glick further teaches generating additional data structures comprising the equivalent of "reverse lookup tables" by transforming each reference sequence record in the forward lookup table into a set of $Y - k + 1$ consecutive overlapping reference sequence segments (para [0025]; "For a sequence of length L , sliding the window one base at a time yields $L-K+1$ overlapping K -mers" and ii) storing each reference sequence segment thus transformed in a record in the at least one reverse lookup table (para [0025]; "A MICA index uses arrays to store all of the positions for each K -mer in the subject nucleic acid sequence. Because chromosomal DNA molecules can be up to several hundred million bases long, the K -mer position values would normally be represented as 4-byte integers. The present invention reduces this data storage requirement by dividing the sequence into "chunks." Suitably a sequence is divided into C separate "chunks" of 65,535 (2.sup.16 -1) bases"; and c) storing, via the data processing system, the forward lookup table and the at least one reverse lookup table in a computer database environment maintained in a non-transitory computer readable storage medium and accessible by the data processing system. (para [0020]; "The term "computer file", as used herein, refers to a stream (sequence) of bits stored as a single unit, typically in a file system on electronic storage medium such as a computer disk, a computer hard drive, a CD-ROM, a DVD-ROM or a magnetic tape. While a file is usually presented as a single stream, it can be, and is most often, stored as multiple fragments of data at different places on a disk or hard drive"). Although Glick does not specifically teach a forward look-up table and a reverse look-up table which relates to reference sequences stored in the forward look-up table, the method taught by Glick essentially does teach the equivalent of this because a first k -mer string of contiguous characters is created and stored from a reference source, and then from that, overlapping sets of reference sequences $Y-K+1$, are created and stored.

As the common technical feature was known in the art at the time of the invention, this cannot be considered a common special technical feature that would otherwise unify the groups. The inventions lack unity with one another.

Therefore, Groups I and II lack unity of invention under PCT Rule 13 because they do not share a same or corresponding special technical feature

Note concerning Item 4:

Claims 45 and 46 are unsearchable omnibus claims under PCT Rule 6.2(a). These claims include omnibus reference to (for example) "A method as described in any of FIGS. 1-15B" and thus do not clearly set forth a specific invention for which protection is being sought.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 14/48265

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☒ Claims Nos.: 45, 46
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
Claims 45 and 46 are unsearchable omnibus claims under PCT Rule 6.2(a). These claims include omnibus reference to (for example) "A method as described in any of FIGS. 1-15B" and thus do not clearly set forth a specific invention for which protection is being sought.

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:
----go to Extra Sheet for continuation----

1. ☒ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☒ No protest accompanied the payment of additional search fees.