



(12) Translation of
European patent specification

(11) NO/EP 3583249 B1

NORWAY

(19) NO

(51) Int Cl.

G16B 50/50 (2019.01)

G16B 30/20 (2019.01)

G06F 21/62 (2013.01)

H03M 7/30 (2006.01)

Norwegian Industrial Property Office

(45)	Translation Published	2023.11.27
(80)	Date of The European Patent Office Publication of the Granted Patent	2023.07.19
(86)	European Application Nr.	17896453.2
(86)	European Filing Date	2017.12.14
(87)	The European Application's Publication Date	2019.12.25
(30)	Priority	2017.02.14, WO, PCT/US17/017842 2017.07.11, WO, PCT/US17/041579
(84)	Designated Contracting States:	AL; AT; BE; BG; CH; CY; CZ; DE; DK; EE; ES; FI; FR; GB; GR; HR; HU; IE; IS; IT; LI; LT; LU; LV; MC; MK; MT; NL; NO; PL; PT; RO; RS; SE; SI; SK; SM; TR
(73)	Proprietor	Genomsys SA, Fondation EPFL Innovation Park Batiment C, 1015 Lausanne, Sveits
(72)	Inventor	BALUCH, Mohamed Khoso, 4439 Woodsedge Court, Chantilly, Virginia 20151, USA ALBERTI, Claudio, 1 chemin des Esserts, 1213 Petit-Lancy, Sveits
(74)	Agent or Attorney	Nordic Patent Service A/S, Bredgade 30, 1260 KØBENHAVN K, Danmark

(54)	Title	METHOD AND SYSTEMS FOR THE RECONSTRUCTION OF GENOMIC REFERENCE SEQUENCES FROM COMPRESSED GENOMIC SEQUENCE READS
------	-------	--

(56)	References Cited:	Mihai Pop ET AL: "Comparative genome assembly", Briefings in bioinformatics, 1 September 2004 (2004-09-01), pages 237-248, XP55742204, England DOI: 10.1093/bib/5.3.237 Retrieved from the Internet: URL: http://ccb.jhu.edu/people/salzberg/doc s/AMOScmp-reprint.pdf K. SCHNEEBERGER ET AL: "Reference-guided assembly of four diverse Arabidopsis thaliana genomes", PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES, vol. 108, no. 25, 21 June 2011 (2011-06-21), pages 10249-10254, XP055222736, ISSN: 0027-8424, DOI: 10.1073/pnas.1107739108 Anonymous: "CRAM format specification (version 3.0)", , 8 September 2016 (2016-09-08), XP002771305, Retrieved from the Internet: URL: https://samtools.github.io/hts-specs/CRAMv3.pdf [retrieved on 2017-06-22] PANNEEL: "Implementation and optimization of CABAC in a genomic data compression framework with support for random access", Department of Electronics and Information Systems, August 2015 (2015-08), XP055535469, Retrieved from the Internet: URL: https://lib.ugent.be/tulltxt/ROG01/002/224/797/RUG01-002224797_2015_000_1_AC.pdf [retrieved on 2018-03-27] DAILY et al.: "Data structures and compression algorithms for high-throughput sequencing technologies", BMC Bioinformatics, vol. 11, 14 October 2010 (2010-10-14), pages 1-12, XP021071841, BAO et al.: "AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references. AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references", Bioinformatics, vol. 30, no. 12 15 June 2014 (2014-06-15), pages i319-i328, XP055535457, Retrieved from the Internet: URL: https://doi.org/10.1093/bioinformatics/btu291
------	----------------------	---

TATWAWADI et al.: "GTRAC: fast retrieval from compressed collections of genomic variants",
Bioinformatics, vol. 32, no. 17 1 September 2016 (2016-09-01), pages i479-i486, XP055535444,
Retrieved from the Internet: URL:doi: 10.1093/bioinformatics/btw437

Enclosed is a translation of the patent claims in Norwegian. Please note that as per the Norwegian Patents Acts, section 66i the patent will receive protection in Norway only as far as there is agreement between the translation and the language of the application/patent granted at the EPO. In matters concerning the validity of the patent, language of the application/patent granted at the EPO will be used as the basis for the decision. The patent documents published by the EPO are available through Espacenet (<http://worldwide.espacenet.com>) or via the search engine on our website here: <https://search.patentstyret.no/>

Krav

1. En datamaskin implementert metode for koding av justerte genomsekvensdata som omfatter avlesninger av nukleotidsekvenser, der metoden omfatter trinnene:

- justere avlesningene for én eller flere referansesekvenser, og dermed opprette justerte avlesninger,
- 5 • kartlegge, på én eller flere referansesekvenser, og de justerte avlesningene som skal kodes,
- sammenfatte de justerte avlesningene for derved å skape en overlappende sekvens,
- sammenligne én eller flere referansesekvenser og den overlappende sekvensen, og dermed innhente informasjon relatert til posisjonene til uoverensstemmelser og typen av uoverensstemmelser mellom én eller flere referansesekvenser og overlappende sekvensen,
- 10 • entropikoding av den informasjon som relaterer til posisjonen til uoverensstemmelsene og typen av uoverensstemmelser, og
- der hvor informasjonen som relaterer til posisjonen til uoverensstemmelsene og typen av uoverensstemmelser indikeres ved hjelp av henholdsvis den første deskriptoren (203) og den andre deskriptoren (204), og den første deskriptoren og den andre deskriptoren er innkapslet i samme
- 15 Tilgangsenhet for å muliggjøre, ved bruke av dekodingsenheten, den selektive rekonstruksjonen av en eller flere referansesekvenser som brukes til justeringen.

2. Metoden i krav 1, der sammenfatningen av de justerte avlesningene omfatter trinnet med å velge ut, for hver posisjon på en eller flere av referansesekvensene, det nukleotidet som er tilstede med den høyeste frekvensen i de justerte avlesningene ved den posisjonen.

20

3. Metoden i krav 1 eller 2, der lengden på den overlappende sekvensen defineres som inndataparameter for koderen, eller dynamisk tilpasset av koderen, fortrinnsvis så blir den første deskriptoren binærisert ved hjelp av en oppdelt enhetsmessig avkortet unary-binærisering, der den oppdelte enhetsmessige avkortede unary-binæriseringen er en sammensetting av gjentatte avkortede

25 unary-binæriseringer. Og hver avkortede unary-binærisering blir anvendt på deler av verdien som skal binæriseres er N-bits lange, der N er et forhåndsvalgt parameter, der lengden på den overlappende sekvensen innbefattes i en syntaksoverskrift.

4. Metoden i krav 3, der lengden på den overlappende sekvensen innbefattes i en syntaksoverskrift, og den første deskriptoren blir binærisert ved hjelp av en delt enhetsmessig avkortet unary-binæriseringer. Og den delte enhetsmessig avkortede unary-binæriseringen er en sammensetting av gjentatte avkortede unary-binæriseringer, der hver avkortede unary-binærisering som blir brukt på
- 5 deler av verdien som skal binæriseres er N bit lange, der N er et forhåndsvalgt parameter, og der den andre deskriptoren blir binærisert ved hjelp av en avkortet unary-binærisering, der verdien til den andre deskriptoren blir fulgt av en null, og hvis verdien er lik den største mulige verdien som skal binæriseres blir det etterfølgende 0-bittet forkastet.
- 10 5. Metoden i krav 3, der metoden ikke koder informasjon som tilsier bruken av en spesifikk referansesekvens.
6. Et apparat for koding av genomsekvensdata, der genomsekvensdataene ombefatter avlesninger av sekvenser for nukleotider, og dette apparatet gjør det mulig å:
- 15 • justere avlesningene for én eller flere referansesekvenser, og dermed opprette justerte avlesninger,
- kartlegge, på én eller flere referansesekvenser, og de justerte avlesningene som skal kodes,
- sammenfatte de justerte avlesningene for derved å skape en overlappende sekvens,
- sammenligne én eller flere referansesekvenser og den overlappende sekvensen, og dermed innhente informasjon relatert til posisjonene til uoverensstemmelser og typen av
- 20 uoverensstemmelser mellom én eller flere referansesekvenser og overlappende sekvensen,
- entropikode den informasjon som relaterer til posisjonen til uoverensstemmelsene og typen av uoverensstemmelser, og
- der hvor informasjonen som relaterer til posisjonen til uoverensstemmelsene og typen av uoverensstemmelser indikeres ved hjelp av henholdsvis den første deskriptoren (203) og den andre
- 25 deskriptoren (204), og den første deskriptoren og den andre deskriptoren er innkapslet i samme Tilgangsenhet for å muliggjøre, ved bruke av dekodingsenheten, den selektive rekonstruksjonen av en eller flere referansesekvenser som brukes til justeringen.
7. Apparatet i krav 6, som i tillegg omfatter mulighetene for å gjennomføre kodingsmetodene i hvilket som helst av kravene 2 til 5.

8. En datamaskin implementert metode for dekoding av kodede genomsekvensdata, som omfatter trinnene:

analysere den kodede inndatafilen, slik at du får tilgangsenheter for de genomiske dataene,

entropidekodings informasjon som relaterer til posisjonen til en uoverensstemmelse og typen

5 uoverensstemmelser som finnes i en overlappende sekvens, og

Modifisere den overlappende sekvensen ved hjelp av informasjonen som relaterer til posisjoner og typer av uoverensstemmelser, for derved å hente ut en eller flere referansesekvenser som brukes til justeringen før komprimering. Der modifisering av den overlappende sekvensen ved hjelp av informasjonen som relaterer til posisjoner og typer uoverensstemmelser, for derved hente ut en eller
10 flere referansesekvenser omfatter videre entropidekoding av den første deskriptoren (203) og den andre deskriptoren (204), og videre omfatter det dekapsulering fra samme tilgangsenhet av den første deskriptoren og den andre deskriptoren, for å derved oppnå en selektiv rekonstruksjon av en eller flere referansesekvenser.

9. Metoden i krav 8, som videre omfatter dekoding av lengden på den overlappende sekvensen
15 fra en syntaksoverskrift som er tilstede i inndatafilen.

10. Metoden i krav 8, som videre omfatter en omvendt binærisering av den første deskriptoren, der den første deskriptoren blir binærisert ved hjelp av en oppdelt enhetsmessig avkortet unary-binærisering, der den delte enhetsmessige avkortede unary-binæriseringen er en sammensetting av
20 gjentatte avkortede unary-binæriseringer. Der hver avkortet unary-binærisering anvendes på deler av verdien som skal binæriseres er N-bit lange, der N er et forhåndsvalgt parameter, som fortrinnsvis også omfatter en omvendt binærisering av den andre deskriptoren der den andre deskriptoren blir binærisert ved hjelp av en avkortet unary-binærisering, der verdien av den andre deskriptoren etterfølges av null, og hvis verdien er lik den største mulige verdien som skal
25 binæriseres, blir den etterfølgende 0-bittet forkastet.

11. Metoden i krav 8, der inndatafilen ikke inneholder informasjon som tilsier bruken av en bestemt referansesekvens.

30 12. Et apparat for dekoding av kodede genomsekvensdata, som gjør det mulig å:

analysere den kodede inndatafilen, slik at du får tilgangsenheter for de genomiske dataene,

entropidekode informasjon som relaterer til posisjonen til en uoverensstemmelse og typen uoverensstemmelser som finnes i en overlappende sekvens, og

Modifisere den overlappende sekvensen ved hjelp av informasjonen som relaterer til posisjoner og typer av uoverensstemmelser, for derved å hente ut en eller flere referansesekvenser som brukes til
5 justeringen før komprimering. Der modifisering av den overlappende sekvensen ved hjelp av informasjonen som relaterer til posisjoner og typer uoverensstemmelser, for derved hente ut en eller flere referansesekvenser ombefatter videre entropidekoding av den første deskriptoren (203) og den andre deskriptoren (204), og videre ombefatter det dekapsulering fra samme tilgangsenhet av den første deskriptoren og den andre deskriptoren, for å derved oppnå en selektiv rekonstruksjon av en
10 eller flere referansesekvenser.

13. Apparatet i krav 12, som videre gjør det mulig å utføre dekodingsmetoden i et hvilket som helst av kravene 9 til 11.

14. Et datamaskin avlesbart medium som kan lagre instruksjoner som, når de utføres av én eller
15 flere prosessorer, får én eller flere prosessorer til å gjennomføre fremgangsmåten i hvilket som helst av kravene 1-5.

15. Et datamaskin avlesbart medium som kan lagre instruksjoner som, når de utføres av én eller flere prosessorer, får én eller flere prosessorer til å gjennomføre fremgangsmåten i hvilket som helst
20 av kravene 8-11.