



(12) Translation of
European patent specification

(11) NO/EP 3526711 B1

NORWAY

(19) NO
(51) Int Cl.
G16B 50/50 (2019.01)

Norwegian Industrial Property Office

(45)	Translation Published	2023.07.03
(80)	Date of The European Patent Office Publication of the Granted Patent	2023.03.29
(86)	European Application Nr.	16791320.1
(86)	European Filing Date	2016.10.11
(87)	The European Application's Publication Date	2019.08.21
(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	ZOIA, Giorgio, Chemin des Croix-Rouges 10, 1007 Lausanne, Sveits RENZI, Daniele, Chemin de Champrilly, 26, 1004 Lausanne, Sveits
(74)	Agent or Attorney	Nordic Patent Service A/S, Bredgade 30, 1260 KØBENHAVN K, Danmark

(54)	Title	METHOD AND APPARATUS FOR COMPACT REPRESENTATION OF BIOINFORMATICS DATA
(56)	References Cited:	US-A1- 2015 227 686 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/C RAMv3.pdf [retrieved on 2017-06-22] Anonymous: "SAM", , 11 March 2015 (2015-03-11), XP002771304, Retrieved from the Internet: URL: https://web.archive.org/web/2015031104_5750/http://davetang.org/wiki/tiki-index.p hp?page=SAM [retrieved on 2017-06-22]

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/>

Patentkrav

1. Datamaskinimplementert fremgangsmåte for komprimering av genomsekvensdata, idet nevnte genomsekvensdata omfatter reads av sekvenser av nukleotider,

idet nevnte fremgangsmåte omfatter trinnene av:

- 5 - å aligne nevnte reads med én eller flere referansesekvenser, idet det derved skapes alignede reads,
 - å klassifisere nevnte alignede reads i forskjellige klasser som i det minste omfatter:
 - en første klasse: når nevnte alignede reads stemmer overens med nevnte ene eller flere referansesekvenser uten noen som helst feil;
 - 10 - en andre klasse: når nevnte alignede reads stemmer overens med et område i nevnte ene eller flere referansesekvenser med et antall overensstemmelsesfeil som dannes av et antall posisjoner i hvilke sekvenseringsmaskinen ikke var i stand til å navngi en hvilken som helst base;
 - en tredje klasse: når nevnte alignede reads stemmer overens med et område i nevnte
 - 15 ene eller flere referansesekvenser med et antall overensstemmelsesfeil som dannes av et antall posisjoner i hvilke sekvenseringsmaskinen ikke var i stand til å navngi en hvilken som helst base eller den navnga en annerledes base enn den som var rapportert i referanseggenomet;
 - en fjerde klasse: når nevnte alignede reads stemmer overens med et område i nevnte
 - 20 ene eller flere referansesekvenser med et antall overensstemmelsesfeil som dannes av et antall posisjoner i hvilke sekvenseringsmaskinen ikke var i stand til å navngi en hvilken som helst base; eller den navnga en forskjellig base enn den som ble rapportert i referanseggenomet og ved tilstedeværelse av insersjoner eller delesjoner eller avskårne nukleotider;
 - 25 - en femte klasse: når nevnte alignede reads ikke finner noen gyldig mapping på nevnte ene eller flere referansesekvenser i henhold til spesifiserte alignmentbegrensninger,
- idet det derved skapes klasser av alignede reads;
- å kode nevnte klassifiserte og alignede reads som en multiplisitet av lag av syntakselementer som defineres av deskriptorer som entydig representerer
 - 30 genomsekvens-reads, idet nevnte deskriptorer omfatter for nevnte første klasse i det minste startposisjonen på referansesekvensen, et flagg som signalerer hvis readen skal betraktes som et revers-komplement i forhold til referansen, en avstand til det overensstemmende par i tilfelle parede reads, verdien av lengden i tilfelle

sekvenseringsteknologien produserer reads av variabel lengde, idet nevnte deskriptorer
omfatter for nevnte andre klasse i det minste deskriptorene av nevnte første klasse og en
overensstemmelsesfeil-posisjon for hver overensstemmelsesfeil, idet nevnte
deskriptorer omfatter for nevnte tredje klasse deskriptorene av nevnte andre klasse og
5 en overensstemmelsesfeil-posisjon og en overensstemmelsesfeil-type for hver
overensstemmelsesfeil, idet nevnte deskriptorer omfatter for nevnte fjerde klasse
deskriptorene av nevnte første klasse og, overensstemmelsesfeil-type for hver
overensstemmelsesfeil, og de avskårne baser,
hvor å kode nevnte klassifiserte alignede reads som en multiplisitet av lag av
10 syntakselementer, omfatter å velge ut nevnte syntakselementer omfattende nevnte
deskriptorer i henhold til nevnte klasser av alignede reads,
hvor kodingen av nevnte klassifiserte alignede reads som en multiplisitet av lag av
syntakselementer er tilpasset i henhold til de statistiske egenskaper av nevnte utvalgte
syntakselementer,
15 hvor kodingen av nevnte klassifiserte alignede reads som en multiplisitet av lag av
syntakselementer omfattende nevnte deskriptorer assosierer en spesifikk kildemodell og
en spesifikk entropikoder til hvert lag, idet kildemodellen er **karakterisert**
ved definisjonen av syntakselementene som emitteres av hver kilde, definisjonen av en
assosiert sannsynlighetsmodell og definisjonen av den assosierte entropikoder, hvor det
20 er dekomposisjon av sekvens-read-dataene og -metadataene til homogene lag av nevnte
deskriptorer for å oppnå distinkte informasjonskilder med redusert informasjonsentropi.

2. Fremgangsmåte ifølge krav 1, hvor nevnte lag av syntakselementer videre omfatter
25 posisjonen av en variant i forhold til referansesekvensen, typen av variant, posisjonen av en
delesjon i forhold til referansesekvensen, posisjonen av ett eller flere symboler som ikke er til
stede i referansesekvensen, men til stede i de alignede reads, typen av insersjon på en gitt
posisjon.

30 3. Fremgangsmåte ifølge krav 1, hvor nevnte entropi-koder er en kontekst-adaptiv aritmetisk
koder.

4. Fremgangsmåte for dekomprimering av en genomisk strøm som er komprimert ifølge
fremgangsmåten ifølge krav 1, idet nevnte fremgangsmåte omfatter trinnene av:

- å parse og dekode (212-214) den komprimerte genomiske strøm til genomiske lag av syntakselementer (215),
- å dekode nevnte genomisk lag til klasser av data (216-217),
- å ekspandere nevnte genomiske lag til klassifiserte reads av sekvenser av nukleotider, selektivt å dekode, under anvendelse av klassedekodere (219), nevnte klassifiserte reads av sekvenser av nukleotider og å slå sammen resultatet på én eller flere referansesekvenser for på den måte å produsere ukomprimerte reads av sekvenser av nukleotider.

5. Genomisk koderinnretning (2010) for komprimeringen av genomsekvensdata (209), idet nevnte genomsekvensdata (209) omfatter reads av sekvenser av nukleotider, idet nevnte genomiske koder (2010) omfatter:

- en aligner-enhet (201), som er konfigurert til å aligne nevnte reads med én eller flere referansesekvenser, idet det derved skapes alignede reads,
- en dataklassifiseringsenhet (204), som er konfigurert til å klassifisere nevnte alignede reads til forskjellige klasser, som i det minste omfatter:
 - en første klasse: når nevnte alignede reads stemmer overens med nevnte ene eller flere referansesekvenser uten noen som helst feil;
 - en andre klasse: når nevnte alignede reads stemmer overens med et område i nevnte ene eller flere referansesekvenser med et antall overensstemmelsesfeil som dannes av et antall posisjoner i hvilke sekvenseringsmaskinen ikke var i stand til å navngi en hvilken som helst base;
 - en tredje klasse: når nevnte alignede reads stemmer overens med et område i nevnte ene eller flere referansesekvenser med et antall overensstemmelsesfeil som dannes av et antall posisjoner i hvilke sekvenseringsmaskinen ikke var i stand til å navngi en hvilken som helst base eller den navnga en annerledes base enn den som var rapportert i referanseggenomet;
 - en fjerde klasse: når nevnte alignede reads stemmer overens med et område i nevnte ene eller flere referansesekvenser med et antall overensstemmelsesfeil som dannes av et antall posisjoner i hvilke sekvenseringsmaskinen ikke var i stand til å navngi en hvilken som helst base; eller den navnga en forskjellig base enn den som ble rapportert i referanseggenomet og ved tilstedeværelse av insersjoner eller delesjoner eller avskårne nukleotider;

- en femte klasse: når nevnte alignede reads ikke finner noen gyldig mapping på nevnte ene eller flere referansesekvenser i henhold til spesifiserte alignmentbegrensninger, idet det derved skapes klasser av alignede reads;

- én eller flere lag-kodeenheter (205-207), som er konfigurert til å kode nevnte klassifiserte alignede reads som lag av syntakselementer ved å velge nevnte syntakselementer i henhold til nevnte klasser av alignede reads, hvor nevnte deskriptorer omfatter for nevnte første klasse i det minste startposisjonen på referansesekvensen, et flagg som signalerer hvis readen skal betraktes som et revers-komplement i forhold til referansen, en avstand til det overensstemmende par i tilfelle parede reads, verdien av lengden i tilfelle sekvenseringsteknologien produserer reads av variabel lengde, idet nevnte deskriptorer omfatter for nevnte andre klasse i det minste deskriptorene av nevnte første klasse og en overensstemmelsesfeil-posisjon for hver overensstemmelsesfeil, idet nevnte deskriptorer omfatter for nevnte tredje klasse deskriptorene av nevnte andre klasse og en overensstemmelsesfeil-posisjon og en overensstemmelsesfeil-type for hver overensstemmelsesfeil, idet nevnte deskriptorer omfatter for nevnte fjerde klasse deskriptorene av nevnte første klasse og en overensstemmelsesfeil-type for hver overensstemmelsesfeil og de avskårne baser;

- entropi-koderenheter (2012-2014) for entropi-koding av nevnte lag av syntakselementer, hvor kodingen av nevnte klassifiserte alignede reads som en multiplisitet av lag av syntakselementer er tilpasset i henhold til de statistiske egenskaper av dataene som bæres av laget, hvor kodingen av nevnte klassifiserte alignede reads som en multiplisitet av lag av syntakselementer omfattende nevnte deskriptorer assosierer en spesifikk kildemodell og en spesifikk entropikoder til hvert lag, idet kildemodellen er **karakterisert ved** definisjonen av syntakselementene som emitteres av hver kilde, definisjonen av en assosiert sannsynlighetsmodell og definisjonen av den assosierte entropikoder, hvor det er dekomposisjon av sekvens-read-dataene og -metadataene til homogene lag av nevnte deskriptorer for å oppnå distinkte informasjonskilder med redusert informasjonsentropi.

6. Genomisk dekamerinnretning (218) for dekomprimeringen av en genomisk strøm (211) som er komprimert av den genomiske koder ifølge krav 5, idet nevnte genomiske dekamer (218) omfatter:

- parsings- og dekodingsmiddel (210, 212-214) som er konfigurert til å parse nevnte komprimerte genomiske strøm til genomiske lag av syntakselementer (215),
- én eller flere lag-dekamer (216-217), som er konfigurert til å dekode de genomiske lag til klasser av data, og videre konfigurert til å forarbeide nevnte genomiske lag til klassifiserte reads av sekvenser av nukleotider (211),
- genomiske dataklasser-dekamer (213), som er konfigurert til selektivt å dekode nevnte klassifiserte reads av sekvenser av nukleotider og konfigurert til å slå sammen resultatet på én eller flere referansesekvenser for på den måte å produsere ukomprimerte reads av sekvenser av nukleotider.

7. Genomisk dekamer-innretning ifølge krav 6, hvor den ene eller de flere referansesekvenser lagres i den komprimerte genomiske strøm (211).

8. Genomisk dekamer-innretning ifølge krav 6, hvor den ene eller de flere referansesekvenser tilveiebringes til dekameren via en out-of-band-mekanisme.

9. Genomisk dekamer-innretning ifølge krav 6, hvor den ene eller de flere referansesekvenser bygges i dekameren.

10. Datamaskinlesbart medium som omfatter anvisninger som når de utføres, bevirker at i det minste én prosessor utfører fremgangsmåten ifølge et hvilket som helst av kravene 1 til 4.