

INNOVATIVE TECHNOLOGIEN FÜR DIE ZUKUNFT

Informations- und Kommunikationstechnik

16985 Method for the coding of genomic variants

Einleitung / Abstract

Die Repräsentation und Codierung von genomischen Annotationen wird derzeit von ISO/IEC JTC 1/SC 29/WG 11 (MPEG) standardisiert.

Die vorliegende Erfindung wurde in den Standardisierungsprozess eingebracht und betrifft ein Verfahren zum Komprimieren und Dekomprimieren einer Information einer Erbinformation, insbesondere einer genetischen Variation.

Technology Readiness Level (TRL)

TRL 4

Patentsituation

Land: DE

Code: 10 2021 100 199 A1

Status: anhängig

Angebot

Lizenz zur gewerblichen Nutzung /
Kooperation möglich

Stichworte

Binärmatrix., Codierung,
Dekomprimierung, Direktzugriff,
Erbinformation, genomische Annotation,
Genotyp, Komprimierung, Sortierung,
Variant

Kontakt

Luise aus der Fünten, M. Sc.

Telefon: +49 (0) 511 . 850 308-0

ausderfuenten@ezn.de

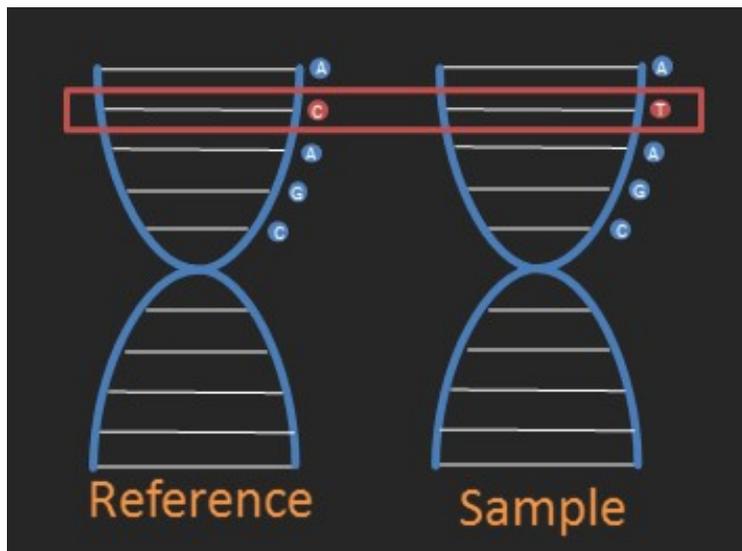


Abb. 1: Eine Veränderung von DNA Sequenz an bestimmter Genomeposition.

Hintergrund

Sequenzierungsexperimente können das Ziel haben, genetische Variationen zu identifizieren. Die Entdeckung von genetischen Variationen ist bei großen Populationen verwandter Proben eine der Hauptanwendungen der Sequenzierungstechnologien der nächsten und dritten Generation. Genetische Variationen können klassifiziert werden in:

- Einzelnukleotid-Polymorphismus (singlenucleotide polymorphisms (SNPs)),
- Einfügungen und Streichungen (insertions and deletions (indels)),

- Strukturvarianten (structural variants).
Genetische Variationen werden üblicherweise im textbasierten Variantenaufformat (VCF) gespeichert.

Lösung

Erfindungsgemäß wird ein verbessertes Verfahren für die Kodierung der Genotyp-Matrix vorgeschlagen, d.h. der Allel- und Phaseninformation. Die Kodierung einer Genotyp-Matrix G umfasst dabei die folgenden Schritte:

1. Reihenweise Aufspaltung der Genotyp-Matrix G in Blöcke, die anschließend separat prozessiert werden. Dabei können die Blöcke so konstruiert werden, dass sie nur eine bestimmte Klasse von genomischer Variation enthalten (z.B. SNPs, Indels oder Strukturvarianten). Zur Rekonstruktion der ursprünglichen Zeilenreihenfolge muss ein Index geführt werden.
2. Aufspaltung der Genotyp-Matrix G in eine Allel-Matrix A und eine Phasenmatrix P.
3. Optionale Binarisierung der Allel-Matrix A (dieser Prozess ergibt entweder die Bitebenen B_q oder eine binäre Allel-Matrix C).
4. Optionale zeilen- und spaltenweise Sortierung der Allelmatrix A oder der Bitebenen B_q oder der binären Allelmatrix C und der Phasenmatrix P.
5. Entropie-Kodierung der Allel-Matrix A oder der Bit-Ebenen B_q oder der binären Allelmatrix C und der Phasenmatrix P.

Vorteile

- Zusätzlich zur spaltenweisen Sortierung wird erfindungsgemäß eine zeilenweise Sortierung vorgeschlagen. Experimente zeigten, dass Entropie-Kodierungsschemata kleinere Bitströme ergeben, wenn die Zeilen auf bestimmte Weise sortiert werden.

Anwendungsbereiche

Repräsentation und Codierung von genomischen Annotationen.