

CLAIMS

What is claimed is:

Claim 1. A method for generating a hash table for software-accelerated genomic data read mapping, the method comprising:

receiving, by one or more computers, genomic data, wherein the genomic data is derived from parent genomic data;

generating, by one or more computers, a first set of values based on the genomic data;

generating, by one or more computers, a subset of the genomic data based on the first set of values;

computing, by one or more computers, a signature for each k-mer of the subset of the genomic data, wherein the signature is computed based on a first hash function;

computing, by one or more computers, a first attribute for each k-mer of the subset of the genomic data, wherein the first attribute comprises a position of a given k-mer of the genomic data within a sequence of the genomic data;

computing, by one or more computers, an index for each k-mer of the subset of the genomic data, wherein the index is computed based on a second hash function; and

storing, by one or more computers, the signature and the first attribute for each k-mer of the subset of the genomic data within a hash data structure based on the index for each k-mer of the subset of the genomic data.

Claim 2. The method of claim 1, wherein each k-mer of the subset of the genomic data is a k-mer comprising k letters representing a string of one or more nucleotides.

Claim 3. The method of any one of claims 1 or 2, wherein the first set of values comprises a representation of a number of times that a given k-mer of the genomic data occurs within the parent genomic data.

Claim 4. The method of any one of the preceding claims 1 to 3, wherein the first set of values comprises a representation of a hash value computed based on a corresponding k-mer of the genomic data.

Claim 5. The method of any one of the preceding claims 1 to 4, wherein a memory allocation size used to store the signature for a given k-mer of the subset is smaller than a memory allocation size used to store the given k-mer.

Claim 6. The method of any one of the preceding claims 1 to 4, the method further comprising:

sending, by one or more computers, data corresponding to the hash data structure as a data package to a first device.

Claim 7. The method of claim 6, wherein the first device is a memory storage device.

Claim 8. The method of claim 6 or claim 7, wherein a second device reads the data corresponding to the hash data structure from the first device, and wherein the second device performs a series of operations to generate a second hash data structure based on the data corresponding to the hash data structure.

Claim 9. A system for generating a hash table for software-accelerated genomic data read mapping, the system comprising:

one or more computers; and

one or more memories storing instructions that, when executed by the one or more computers, cause the one or more computers to perform operations, the operations comprising:

receiving, by one or more computers, genomic data, wherein the genomic data is derived from parent genomic data;

generating, by the one or more computers, a first set of values based on the genomic data;

generating, by the one or more computers, a subset of the genomic data based on the first set of values;

computing, by the one or more computers, a signature for each k-mer of the subset of the genomic data, wherein the signature is computed based on a first hash function;

computing, by the one or more computers, a first attribute for each k-mer of the subset of the genomic data, wherein the first attribute comprises a position of a given k-mer of the genomic data within a sequence of the genomic data;

computing, by the one or more computers, an index for each k-mer of the subset of the genomic data, wherein the index is computed based on a second hash function; and

storing, by the one or more computers, the signature and the first attribute for each k-mer of the subset of the genomic data within a hash data structure based on the index for each k-mer of the subset of the genomic data.

Claim 10. The system of claim 9, wherein each k-mer of the subset of the genomic data is a k-mer comprising k letters representing a string of one or more nucleotides.

Claim 11. The system of any one of the preceding claims 9 or 10, wherein the first set of values comprises a representation of a number of times that a given k-mer of the genomic data occurs within the parent genomic data.

Claim 12. The system of any one of the preceding claims 9 to 11, wherein the first set of values comprises a representation of a hash value computed based on a corresponding k-mer of the genomic data.

Claim 13. The system of any one of the preceding claims 9 to 12, wherein a memory allocation size used to store the signature for a given k-mer of the subset is smaller than a memory allocation size used to store the given k-mer.

Claim 14. The system of any one of the preceding claims 9 to 13, wherein the operations further comprise:

sending, by one or more computers, data corresponding to the hash data structure as a data package to a first device.

Claim 15. The system of claim 14, wherein the first device is a memory storage device.

Claim 16. The system of claim 14 or claim 15, wherein a second device reads the data corresponding to the hash data structure from the first device, and wherein the second device performs a series of operations to generate a second hash data structure based on the data corresponding to the hash data structure.

Claim 17. A computer-readable medium storing instructions that, when executed by one or more computers, cause the one or more computers to perform operations for generating a hash table for software-accelerated genomic data read mapping, the operations comprising:

receiving genomic data, wherein the genomic data is derived from parent genomic data;

generating a first set of values based on the genomic data;

generating a subset of the genomic data based on the first set of values;
computing a signature for each k-mer of the subset of the genomic data, wherein the signature is computed based on a first hash function;

computing a first attribute for each k-mer of the subset of the genomic data, wherein the first attribute comprises a position of a given k-mer of the genomic data within a sequence of the genomic data;

computing an index for each k-mer of the subset of the genomic data, wherein the index is computed based on a second hash function; and

storing the signature and the first attribute for each k-mer of the subset of the genomic data within a hash data structure based on the index for each k-mer of the subset of the genomic data.

Claim 18. The computer-readable medium of claim 17, wherein each k-mer of the subset of the genomic data is a k-mer comprising k letters representing a string of one or more nucleotides.

Claim 19. The computer-readable medium of claim 17 or claim 18, wherein the first set of values comprises a representation of a number of times that a given k-mer of the genomic data occurs within the parent genomic data.

Claim 20. The computer-readable medium of any one of the preceding claims 17 to 19, wherein the first set of values comprises a representation of a hash value computed based on a corresponding k-mer of the genomic data.

Claim 21. The computer-readable medium of any one of the preceding claims 17 to 20, wherein a memory allocation size used to store the signature for a given k-mer of the subset is smaller than a memory allocation size used to store the given k-mer.

Claim 22. The computer-readable medium of any one of the preceding claims 17 to 20, the operations further comprising:

sending data corresponding to the hash data structure as a data package to a first device.

Claim 23. The computer-readable medium of claim 22, wherein the first device is a memory storage device.

Claim 24. The computer-readable medium of claim 22 or claim 23, wherein a second device reads the data corresponding to the hash data structure from the first device, and wherein the second device performs a series of operations to generate a second hash data structure based on the data corresponding to the hash data structure.