Formati podataka

FASTA

FASTA format je format tekstualne datoteke koja sadrži u sebi nizove nukleinskih baza ili aminokiselina s njihovim opisima, odnosno imenima. Svaki niz najčešće započinje s linijom opisa koja započinje sa znakom > koju nazivamo zaglavljem sekvence. U zaglavlju nalazi se ime ili identifikator sekvence uz mogući dodatak njenog opisa. Iza ili umjesto zaglavlja sekvence mogu se nalaziti linije komentara koje sadrže neke dodatne opise sekvenci. Linije komentara započinju sa znakom ;. Iza zaglavlja slijedi niz slova znakova kod kojeg svaki znak predstavlja određenu nukleinsku bazu ili aminokiselinu. Ne postoji dogovorena ekstenzija datoteke u FASTA formatu, stoga neki alati ne mogu prepoznati neke datoteke kao datoteke u FASTA formatu. Najčešće ekstenzije su: .fasta, .fna, .ffn, .faa, .frn, .fa.

|  |  |
| --- | --- |
| Kodni izraz nukleinske baze | Značenje |
| A | adenin |
| C | citozin |
| G | guanin |
| T | timin |
| U | uracil |
| (i) | inozin |
| R | A ili G (purin) |
| Y | C, T ili U (pirimidin) |
| K | G, T ili U (baze koje su ketoni) |
| M | A ili C (baze s amino skupinama) |
| S | C ili G (jaka veza) |
| W | A, T ili U (slaba veza) |
| B | C, G, T ili U |
| D | A, G, T ili U |
| H | A, C, T ili U |
| V | A, C ili G |
| N | A, C, G, T ili U |
| - | praznina određene duljine |

Tablica 1.1: opis i značenje kodnih izraza za nukleinske baze

unutar sekvenci FASTA datoteke



FASTQ

FASTQ format je format tekstualne datoteke koja, uz sve podatke koje sadrži i datoteka u FASTA formatu, sadrži informacije o kvaliteti očitanja. Svaka sekvenca u datoteci FASTQ formata sadrži 4 linije. Prva linija započinje znakom @ i sadrži ime i/ili identifikator sekvence, po mogućnosti uz još neke dodatne informacije. Druga linija predstavlja niz nukleinskih baza sekvence. Treća linija sastoji se od znaka + koji predstavlja separator uz koji se ponekad nalazi ponovljeno ime i opis iz prve linije. Četvrta linija sadrži niz znakova koji predstavljaju kvalitetu očitanja sekvence iz druge linije. Četvrta linija mora imati jednak broj znakova kao druga linija. Kao i kod FASTA formatiranih datoteka, ne postoji dogovorena ekstenzija za datoteke u FASTQ formatu zbog čega ih neki alati ne mogu prepoznati. Najčešće korištene ekstenzije su: .fq i .fastq.



SAM

|  |  |  |  |
| --- | --- | --- | --- |
| Indeks polja | Naziv polja | Vrsta podatka | Opis |
| 1 | QNAME | string | naziv fragmenta |
| 2 | FLAG | integer | bitovi koji predstavljaju zastavice |
| 3 | RNAME | string | naziv reference |
| 4 | POS | integer | početak mapiranja  (počinje od 1) |
| 5 | MAPQ | integer | kvaliteta mapiranja |
| 6 | CIGAR | string | CIGAR string |
| 7 | RNEXT | string | naziv referentne sekvence idućeg očitanja fragmenta |
| 8 | PNEXT | integer | početak referentne sekvence idućeg očitanja fragmenta (počinje od 1) |
| 9 | TLEN | integer | duljina fragmenta |
| 10 | SEQ | string | niz baza u sekvenci  (A, C, G, T, U ili N) |
| 11 | QUAL | string | niz koji opisuje kvalitetu očitanja sekvence |

SAM je tekstualni format koji opisuje poravnanja između fragmenata i referentnog genoma. Sastoji se od dva dijela: zaglavlja i dijela poravnanja. Skraćenica je za „*Sequence Alignment/Map format*“. Sve vrijednosti unutar SAM datoteke odvojene su znakom TAB. Svaka linija zaglavlja počinje znakom @ nakon čega slijedi oznaka linije zaglavlja koja se sastoji od dva slova. Nakon toga u svakoj liniji zaglavlja slijedi niz parova OZNAKA:VRIJEDNOST odvojenih znakom TAB. U drugom dijelu datoteke opisuje se poravnanje svakog fragmenta na referentni genom. Svaka linija sastoji se od najmanje 11 vrijednosti međusobno odvojenih znakom TAB koje nam daju opis i informacije o poravnanju.

Tablica 1.2: prikaz opisa i objašnjenja svakog polja u liniji koja predstavlja poravnanje u SAM datoteci

PAF

PAF je tekstualni format datoteka koji služi za opis regija mapiranja između sekvenci. Svaka linija unutar PAF datoteke sadrži vrijednosti odvojene znakom TAB koje opisuju mapiranje sekvence na referentni genom.

|  |  |  |
| --- | --- | --- |
| Indeks polja | Vrsta podataka | Opis |
| 1 | string | ime fragmenta |
| 2 | integer | duljina fragmenta |
| 3 | integer | početak fragmenta (počinje od nule, zatvoren kraj intervala) |
| 4 | integer | završetak fragmenta (počinje od nule, otvoren kraj intervala) |
| 5 | char | odnos između sekvenci  (originalan položaj ili reverzno komplementaran; + ili - ) |
| 6 | string | ime sekvence referentnog genoma |
| 7 | integer | duljina sekvence referentnog genoma |
| 8 | integer | početak sekvence referentnog genoma na originalnom lancu (počinje od nule) |
| 9 | integer | završetak sekvence referentnog gena na originalnom lancu (počinje od nule) |
| 10 | integer | broj usklađenih parova baza |
| 11 | integer | duljina bloka poravnanja |
| 12 | integer | kvaliteta mapiranja  (0-255; 0 je apsolutno podudaranje) |

Tablica 1.3: Opis vrijednosti svakog polja unutar svake linije PAF datoteke