# Uvod

Stablo valića (eng. Wavelet tree) struktura je koja se koristi za sažimanje niza znakova. Stablo se gradi rekurzivnim dijeljenjem abecede ulaznog niza u dva skupa. Listovi stabla odgovaraju točno određenom znaku abecede, a svaki čvor u *bit-vektor* pohranjuje u koji skup abecede znak spada (Slika 1).



Slika 1. – primjer stabla valića

Opisano i prikazano stablo valića je tzv. binarno stablo valića (eng. binary wavelet tree).

Ovakva struktura pogodna je za provođenje operacije *rank(i)*. Rank (i) operacija nad nekim *bit-vektorom* će vratiti broj bitova koji su postavljeni u 1, od nulte do *i*-te pozicije ( [0,i] ). Npr za bit-vektor „0010110“ operacija Rank (3) vratiti će 1 jer se od nulte do treće pozicije nalazi samo jedna jedinica. Operacija rank zapravo je generalizacija operacije popcount tj. Hammingove težine.

Operacija rank može se proširiti tako da se uvede dodatni argument koji predstavlja znak u nekom nizu. Npr. operacija rank (a, 4) nad nizom „banana“ vratit će 2 jer se znak 'a' dvaput pojavljuje u do pozicije 4 ulaznog niza.

Složenost ovakve operacije očigledno je O(N) gdje je N duljina ulaznog niza. Konstrukcijom stabla valića za ulazni niz duljine N, složenost operacije postaje O(log2(A)) gdje je A veličina abecede. Ako spremamo ulazni niz kao bit-vektor (bez RRR) ukupna složenost je O(N\*log2(A)) - log2(A) operacija rank (c, i) koje pozivaju binarne rank operacije rank(i) koje imaju složenost O(n).

Operaciju rank (c, i) nad stablom valića provodimo na sljedeći način (krećemo od korijenskog čvora):

1. Nad bit-vektorom pohranjenim u čvoru provedemo operaciju i = rank (value(c), i) gdje je value (c) vrijednost znaka c u tom čvoru (0 ili 1).
2. Ako je value(c) 0 krećemo u lijevo dijete, ako je 1 u desno dijete.
3. Ponavljamo korake 1. i 2. sve dok ne dođemo do lista, ali tako da je parametar ***i*** u operaciji rank uvijek rezultat prethodno izračunate operacije (u roditeljskom čvoru).

Sljedeća operacija koja je pogodna za izvođenje nad stablom valića je select(i). Operacija select, svrojevrsni je inverz operacije rank – rezultat je pozivija i-te jedinice u bit-vektoru. Stoga primjećujemo da vrijedi:

Rank (Select(i)) = i (napomena: obrat ne vrijedi nužno)

Operacija select (c, i) provodi se nad stablom valića slično kao i rank, samo ovdje ne počinjemo u korijenu, nego u listu u kojem je spremljena „konačna vrijednost“ znaka c:

1. Nad bit-vektorom pohranjenim u čvoru provedemo operaciju i =select (value(c), i) gdje je value (c) vrijednost znaka c u tom čvoru (0 ili 1).
2. Prelazimo na roditeljski čvor.
3. Ponavljamo korake 1. i 2. sve dok ne dođemo do korijena, ali tako da je parametar ***i*** u operaciji rank uvijek rezultat prethodno izračunate operacije (u čvoru djetetu).

RRR struktura koristi se za sažimanje bit-vektora te svodi složenost binarne rank operacije nad pojedinim čvorom na O(1). Operaciju je moguće obaviti bez dekompresije (eng. Succinct data structure). Za izgradnju RRR strukture, bit-vektor rastavljamo na blokove i superblokove. Npr. za b=5 i f=3 gdje je b veličina bloka, a f veličina superbloka nad bit-vektorom dobivamo slučaj sa slike 2.



Slika 2. – podijela bit-vektora na blokove i superblokove

U RRR strukturi blokovi se zamjenjuju uređenim parovima (c, o) gdje je **c** klasa (eng. class), a **o** odmak (eng. offset). Klasa označava popcount bloka, a odmak označava indeks permutacije bloka te klase. Te permutacije spremljene su u tablici – ključ je klasa a vrijednosti su svi blokovi (permutacije) te klase. Primjer je na slikama 3. i 4.



Slika 3. – pretvorba bit-vektora u RRR strukturu



Slika 4. – izgled tablice u kojoj se pohranjuju odmaci (offsets)

Broj klasa će uvijek odgovarati veličini bloka + 1 (jer je najveći popcount bloka njegova duljina – slučaj gdje su sve jedinice). Sažimanje se dobiva uporabom varijabilnih duljina odmaka u RRR strukturi. Broj bitova za klase je uvijek konstantan i iznosi log2(b) + 1 gdje je b duljina bloka. Broj bitova za odmake je kao što je već rečeno varijabilan, odnosno varira od klase do klase i iznosi log2() gdje je b veličina bloka, a C je vrijednost klase (zapravo se radi o broju permutacija).

Kako bi postigli složenost O(1) koristimo superblokove. Superblokovi grupiraju blokove u cjeline te spremaju sumu cijelog bloka te pokazivač na prvi blok sljedećeg superbloka (slika 5.). Time postižemo da će se za izračun operacije rank iterirati po najviše f blokova (f je „duljina“ superbloka) jer će svi prethodni superblokovi biti unaprijed izračunati.



Slika 5. – superblokovi

Računanje operacija rank (i) sada izgleda ovako:

1. Izračunamo indeks bloka: (ib je globalni indeks bloka).
2. Izračunamo indeks superbloka u kojem se nalazi dobiveni blok: (is je indeks superbloka).
3. Kao početnu sumu postavljamo sumu svih prethodnih superblokova (koja je izračunata prilikom izgradnje RRR strukutre).
4. Iteriramo po blokovima dobivenog superbloka pa sve do dobivenog bloka, te iz (c, o) parova dodajemo c u sumu (klasa je ujedno i popcount bloka).
5. Kada dođemo do dobivenog bloka, pomoću vrijednosti c i o dohvaćamo ga iz tablice odmaka, te računamo rank tog bloka za j = i mod b bitova.

Na sličan način se iz RRR strukture računa i operacija select (i):

1. Pronađi prethodnik superbloka za koji vrijedi da je pohranjena suma > i
2. Postavi trenutni rezultat na j \* veličina superbloka gdje je j indeks dobivenog superbloka
3. Iteriraj po blokovima dobivenog superbloka, te dodaj njihove klase ukupnoj sumi i povećavaj rezultat dodajući veličinu bloka , dok ne dođeš do bloka za kojeg vrijedi: suma bloka + ukupna suma > i
4. Dobiveni blok dohvati iz RRR tablice te na ukupnu sumu pridodaj jedinice (ili nule, ovisi o vrijednosti ulaznog znaka u tom čvoru) iz tog bloka, sve do ukupna suma ne bude jednaka i.

Dakle, ako spremimo ulazni niz kao RRR ukupna složenost rank (i select) operacija bit će O(log2(A)) - log2(A) poziva rank (c, i) od kojih svaki poziva binarnu rank operaciju nad RRR strukturom čija je složenost O(1).

# Općeniti primjer algoritma

## Izgradnja stabla

Uzmimo za primjer neki kratki genom: **ATGATCGTAAGGCTCA** -> Abeceda: {A, T, G, C}

Prvo gradimo binarno stablo valića:

ATGATCGTAAGGCTCA

0010011000111010

ATATTA ATA GCGGGCC

010110010 0100011

U korijenskom čvoru abeceda je {A, T, G, C}. Nju dijelimo na 2 dijela te pridružujemo svakom znaku abecede vrijednost 0 ili 1. Time dobivamo {A=0, T=0, G=1, C=1} te zapisujemo ulazni niz u bit-vektor preslikavajući znak u odgovarajući bit. U lijevo dijete ulazi niz koji sadrži samo znakove koji su u abecedi korijena imali vrijednost 0 (isto tako u desno dijete idu oni s vrijednosti 1). Abeceda se ponovno dijeli {A=0, T=1} u lijevom dijetetu i {G=0, C=1} u desnom dijetetu. Ponovno preslikavamo niz u bit-vektor. U ovom trenutku izgradnja stabla završava jer u krajnjim čvorovima imamo točno definirano koji znak ima koju vrijednost. Napomena: ulazni nizovi se ne spremaju u čvorovima, njega se može dobiti rekonstrukcijom iz bit-vektora.

Nakon toga potrebno je iz bit-vektora dobiti RRR strukturu. Prvo određujemo veličinu blokova i superblokova po formulama:

Veličina bloka: , gdje je N duljina ulaznog niza.

Veličina superbloka: , gdje je b veličina bloka, a N duljina ulaznog niza.

U našem slučaju N = 16, uvrštavanjem u gore navedene formule, dobivamo b = 2 i f = 8. To znači da ćemo bit-vektor predstaviti sa 8 blokova i 2 superbloka (4 bloka su 1 superblok). No prvo treba napraviti „lookup“ tablicu odmaka. Pošto je duljina bloka 2 postoje 2 +1 = 3 različite klase, svaka sa različitih odmaka (eng. offset).

Tablica:

[0] -> {00};

[1] -> {01, 10};

[2] -> {11};

Pomoću tablice sada gradimo RRR strukturu za bit-vektor u korijenu čvora:

00 | 10 | 01 | 10 | 00 | 11 | 10 | 10

(0,0)(1,1)(1,0)(1,1)(0,0)(2,0)(1,1)(1,1)

Za zapis klase nam je potrebno 2 bita, a za prikaz odmaka 1 bit (uobičajno će se odmaci u pojedinim klasama prikazivati različitim brojem bitova, no ovdje se zbog jednostavnosti primjera svi prikazuju s jednim bitom).

Kada kodiramo uređene parove dobivamo:

000 | 011 | 010 | 011 | 000 | 100 | 011 | 011

Prilikom konstrukcije pamtimo i sume unutar superblokova te pokazivače na početak sljedećeg superbloka tj. ukupne duljine prethodnih superblokova (kako bi znali gdje idući počinje).

Sume: 3 | 7

Duljine: 12 | 24

Ekvivalentne pretvorbe obavljaju se za lijevi i desni čvor. Sada imamo potpuno inicijalizirano binarno stablo valića s RRR strukturom.

## Rank operacija

Rank operaciju pozivamo s dva parametra: prvi je znak čija pojavljivanja u nizu želimo izbrojati, a drugi je granica do koje brojimo. Npr. Rank (A, 6) će nam reći koliko se puta A pojavljuje u prvih 6 znakova.

Pozovimo npr. Rank ('G', 13).

Algoritam izvodimo od korijenskog čvora. Znak G u korijenskom je čvoru predstavljen s 1 što znači da je potrebno izbrojati jedinice do 13. znaka. Pošto smo pohranili bit-vektor kao RRR računamo rank na sljedeći način:

1. Računamo indeks bloka: (cjelobrojno dijeljenje).
2. Računamo indeks superbloka: (faktor superbloka je 4 jer superblok obuhvaća 4 bloka).
3. Trenutna suma postaje suma prethodnih superblokova (suma[0] = 3), a superblok s traženim blokom počinje na odmak[0] = 12.
4. Iteriramo po blokovima do traženog bloka (u ovom slučaju to su 2 bloka: 000 | 100) te zbrajanjem njihovih klasa i dodavanjem sumi dobivamo:

suma = suma[0] + 0 + 2 = 3 + 0 + 2 = 5

1. Sada iz idućeg bloka (011) treba dobiti originalni blok te izračunati njegov popcount. Iz tablice dobivamo da je [1][1] -> 10, te nam je konačna suma = 6.

Ovdje, postupak nad korijenskim čvorom završava, a operacija rank se poziva nad desnim dijetetom korijenskog čvora (jer je vrijednost od G u riječniku 1). Kao što je i rečeno, rezultat roditeljske rank operacije postaje granica rank operacije u dijetetu, tako da se nad tim čvorem poziva rank (G, 6) te će rezultat te operacije biti konačno rješenje (nema više djece). Konačni rezultat: rank (G, 13) = 3.

## Select operacija

Kao ulaznu strukturu uzet ćemo stablo valića iz prethodnog primjera. Pokušajmo sada izračunati select (G, 3). Dakle tražimo poziciju trećeg slova G u ulaznom nizu. Kao što algoritam kaže počinjemo iz lista u kojem se nalazi G. U našem slučaju to je desno dijete korijena u kojem je G = 0.

Dakle brojimo nule dok ne izbrojimo 3 i vratimo zadnju poziciju. Lako možemo vidjeti da je rezultat select (G, 3) = 4

U ovom trenutku se pomičemo u roditeljski čvor (korijen). U korijenu je G = 1, dakle brojimo jedinice.

Na sličan način kao za rank operaciju nad RRR-om (detalji su objašnjeni u Uvodu), dobivamo da je rezultat operacije select (G, 4) = 11.

# Implementacije

## Implementacija u programskom jeziku C

Implementacija u programskom jeziku C strukturirana je u nekoliko datoteka.

U datotekama WaveletTree.c i WaveletTree.h napisane su funkcije za izgradnju stabla valića i dane su implementacije operacija rank i select (visoke razine) te rank i select koje rade s bit-vektorima.

Funkcija *buildWaveletTree* gradi stablo iz ulaznog niza koji joj se šalje kao parametar. Ona će iz niza „izvući“ abecedu (funkcija *extractAlphabet*) te će predati abecedu i niz funkciji *buildWaveletNode.* Funkcija *buildWaveletNode* rekurzivno gradi stablo stvaranjem struktura WaveletNode. Ona će prvo izgraditi bit-mapu pa iz nje i RRR strukturu, a zatim će podijeliti abecedu i ulazni niz na lijevi i desni te će ih rekurzivno proslijediti za lijevo i desno dijete.

Funkcije *rankOperation* i *selectOperation* su implementacije algoritama rank(c, i) i select (c, i) koje su opisane u uvodu.

Funkcije popcount i selectOnBitmap su implementacije binarnih rank i select operacija. U oba slučaja iterira se po bit-vektoru te se broje jedinice ili nule (uvjet zaustavljanja je drukčiji).

Deklarirane strukture su WaveletTree i WaveletNode. WaveletTree sadrži samo pokazivač na korijenski WaveletNode. WaveletNode sadrži pokazivače na bit-vektor, RRR-strukturu, abecedu te čvorove-djecu i roditelja.

U datotekama RRR.c i RRR.h nalaze se implementacije funkcija za izgradnju RRR strukture iz bit-vektora, računanje rank i select operacija nad RRR strukturom te funkcija koja računa veličine bloka i superbloka. Također sadrži funkciju za izgradnju RRR tablica.

Funkcija bitmapToRRR prima ulazni bit-vektor te na temelju izračunatih veličina blokova i superblokova gradi RRR strukturu na način koji je opisan u uvodu.

Funkcije popcountRRR i selectRRR implementacije su binarnih rank(i) i select(i) operacija koje su također opisane u uvodu.

Funkcija buildRRRTable na temelju izračunatih veličina blokova i superblokova gradi RRR „lookup“ tablicu.

Deklarirane strukture su RRRStruct, RRRTable i RRRTableEntry. RRRStruct predstavlja RRR strukturu te sadrži pokazivač na bit-vektor koji je predstavlja. Također sadrži spremljene vrijednosti suma i odmaka superblokova u tom bit-vektoru. RRRTable predstavlja „lookup“ tablicu te sadrži niz RRRTableEntry struktura. RRRTableEntry predstavlja zapis jedne klase u tablici. Sastoji se od klase (suma bloka) te pokazivača na blokove koje pripadaju toj klasi.

U datotekama Utils.c i Utils.h nalaze se funkcije za rad sa strukturom Dictionary. Struktura Dictionary predstavljat će abecedu pohranjenu u čvoru– sastoji se od varijable char i njezine vrijednosti boolean.

Funkcija *extractAlphabet* izvlači abecedu iz danog niza te im pridjeljuje vrijednosti. Funkcija *splitAlphabet* dijeli zadanu abecedu na lijevu i desnu te ih vraća preko pokazivača. Funkcije *charInDictionary* i *getDictionarValue* provjeravaju postojanje odnosno vraćaju vrijednost danog znaka.

Uz Dictionary, deklarirana je i struktura BitMap koja predstavlja bit-vektor. BitMap sadrži pokazivač na polje char znakova, ali tako da se unutar jednog char znaka nalazi 8 bitova.

U datoteci main.c nalazi se funkcija main() koja prima ime ulazne datoteke, rank ili select, znak i granicu. Funkcija parsira FASTA datoteku te učitava ulazni niz. Zatim poziva buildWaveletTree za izgradnju stabla te na kraju poziva (ovisno o ulaznom parametru) rankOperation ili selectOperation.

# Analiza rezultata

Rezultati performansi implementacije zadanih struktura dani su po tablicama.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Ulazna datoteka,**  **Broj znakova,**  **Veličina abecede** | **C** | | **C++** | | **C# - v1** | | **C# - v2** | | **Java** | | **Python** | |
| Input.fa,  n = 502,  Σ = 4 | 46 us | 143 B |  |  |  |  |  |  | 32.7 ms | 1 kB |  |  |
| Input1.fa,  n = 4 570 937,  Σ = 4 | 350 ms | 1.09 MB |  |  |  |  |  |  | 2.2 s | 8.71 MB |  |  |
| Input2.fas,  n = 5 524 970,  Σ = 4 | 673 ms | 2.47 MB |  |  |  |  |  |  | 4.3 s | 19.77 MB |  |  |

Tablica 1. Usporedba vremena izgradnje stabla valića s binarnim vektorima kao čvorovima

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Ulazna datoteka,**  **Broj znakova,**  **Veličina abecede** | **C** | | **C++** | | **C# - v1** | | **C# - v2** | | **Java** | | **Python** | |
| Input.fa,  n = 502,  Σ = 4 | 125 us | 368 B |  |  |  |  |  |  | 63.2 ms | 3.52 kB |  |  |
| Input1.fa,  n = 4 570 937,  Σ = 4 | 1.5 s | 1.33 MB |  |  |  |  |  |  | 7.5 s | 15.55 MB |  |  |
| Input2.fas,  n = 5 524 970,  Σ = 4 | 2.4 s | 2.19 MB |  |  |  |  |  |  | 11 s | 35.21 MB |  |  |

Tablica 2. Usporedba vremena izgradnje stabla valića s RRR strukturama kao čvorovima

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Ulazna datoteka,**  **Broj znakova,**  **Veličina abecede** | **Testovi** | **Referentni** | **C** | **C++** | **C# - v1** | **C# - v2** | **Java** | **Python** |
| Input.fa,  n = 502,  Σ = 4 | Rank('A', 234) = 58 | 2 us | 6 us |  |  |  | 85 us |  |
| Select('C', 23) = 107 | 1 us | 4 us |  |  |  | 67 us |  |
| Rank('T', 432) = 114 | 4 us | 9 us |  |  |  | 65 us |  |
| Select('G', 100) = 389 | 4 us | 11 us |  |  |  | 84 us |  |
| Input1.fa,  n = 4 570 937,  Σ = 4 | Rank('G', 12345) = 3383 | 106 us | 248 us |  |  |  | 859 us |  |
| Select('A', 12345) = 52553 | 411 us | 1262 us |  |  |  | 16.5 ms |  |
| Rank('C', 1234567) = 305484 | 12168 us | 33365 us |  |  |  | 47.9 ms |  |
| Select('T', 1000000) = 4053864 | 33688 us | 110117 us |  |  |  | 115.5 ms |  |
| Input2.fas,  n = 5 524 970,  Σ = 4 | Rank ('t', 4000000) = 990024 | 32252 us | 83350 us |  |  |  | 75.3 ms |  |
| Select('a', 40000) = 167264 | 1308 us | 6410 us |  |  |  | 29 ms |  |
| Rank ('a', 543210) = 132287 | 4278 us | 15962 us |  |  |  | 56.4 ms |  |
| Select('c', 543210) = 2226555 | 18492 us | 111657 us |  |  |  | 119.7 ms |  |

Tablica 3. Usporedba vremena operacija „Rank“ i „Select“ gdje je čvor stabla binarni vektor

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Ulazna datoteka,**  **Broj znakova,**  **Veličina abecede** | **Testovi** | **Referentni** | **C** | **C++** | **C# - v1** | **C# - v2** | **Java** | **Python** |
| Input.fa,  n = 502,  Σ = 4 | Rank('A', 234) = 58 | 2 us | 2 us |  |  |  | 102 us |  |
| Select('C', 23) = 107 | 1 us | 2 us |  |  |  | 149 us |  |
| Rank('T', 432) = 114 | 4 us | 1 us |  |  |  | 102 us |  |
| Select('G', 100) = 389 | 4 us | 4 us |  |  |  | 134 us |  |
| Input1.fa,  n = 4 570 937,  Σ = 4 | Rank('G', 12345) = 3383 | 106 us | 3 us |  |  |  | 180 us |  |
| Select('A', 12345) = 52553 | 411 us | 161 us |  |  |  | 482 us |  |
| Rank('C', 1234567) = 305484 | 12168 us | 5 us |  |  |  | 21.3 ms |  |
| Select('T', 1000000) = 4053864 | 33688 us | 16950 us |  |  |  | 22.6 ms |  |
| Input2.fas,  n = 5 524 970,  Σ = 4 | Rank ('t', 4000000) = 990024 | 32252 us | 6 us |  |  |  | 21.4 ms |  |
| Select('a', 40000) = 167264 | 1308 us | 931 us |  |  |  | 1.5 ms |  |
| Rank ('a', 543210) = 132287 | 4278 us | 4 us |  |  |  | 3.3 ms |  |
| Select('c', 543210) = 2226555 | 18492 us | 14679 us |  |  |  | 31.6 ms |  |

Tablica 4. Usporedba vremena operacija „Rank“ i „Select“ gdje je čvor stabla RRR struktura