Beleške za prezentovanje seminarskog rada

Kriptografija
Matematički fakultet

Una Stanković

09.06.2017.

1 Uvodna reč

Kao posledica sve bržeg razvoja tehnologija za sekvenciranje genoma javlja se poterba da se očuva bezbednost podataka posebno jer se oni sve češće čuvaju u oblaku odakle se uglavnom koriste za istraživanje. Čuvanje podataka u oblaku i brz razvoj tehnologija doprineli su tome da je moguće lako pristupiti velikim skupovima genoma što može doprineti bitnim pomacima u razvoju biomedicinskih istraživanja.

Informacije koje dobijamo iz skupova genoma se najčešće koriste u medicini, biomedicinskim istraživanjima, kao i uslugama koje se direktno pružaju korisnicima, zbog čega je veoma važno da se ovim podacima rukuje sa oprezom i da se oni obezbede od neovlašćenog pristupa i korišćenja.

Kako to otprilike izgleda? Podaci o genomima se prikupljaju i postavljaju na oblak iz različitih izvora kao što su:

- različiti vidovi biometrijskih autentikacija
- medicinska analiza
- laboratorijski rezultati
- razne studije i drugi.

1.1 Homomorfna enkripcija

Homomorfna enkripcija je pojam koji nam je veoma značajan, jer mi pomoću nje obezbeđujemo privatnost podataka, ona omogućava da izračunavanja budu prenesena u obliku šifrovanog teksta.

Homomorfna enkripcija je vid enkripcije koji nam dozvoljava da izvodimo operacije nad enkriptovanim tekstom i dobijemo enkriptovani rezultat, koji kada dekriptujemo, isti je kao rezultat koji bismo dobili da smo operaciju primenili nad neenkriptovanim tekstom.
HE je odlična za očuvanje privatnosti sekvence koju analiziramo, ali se pokazuje kao veoma nepraktična za analizu informacija kod kompletnog ljudskog genoma, zato želimo da je primenjujemo samo nad nekim delovima.

2 Postavka problema

Zadatak Mi želimo da na bezbedan način izračunamo verovatnoću genetskih bolesti kroz uparivanje skupa biomarkera sa enkriptovanim genomima koji se čuvaju u javnom oblaku.

2.1 Biomarkeri

Biomarker je supstanca u telu koja može da se izmeri i pruži informacije o stanju u telu pacijenta. Zamislite 2 osobe, svaka ima različiti skup molekula koji kruže u telu. Pomoću tih molekula formiramo nešto što se zove otisak, oni mogu da pokazuju razlike između dvoje ili više ljudi. Na primer, ako bismo uzeli zdravu osobu i osobu koja ima dijabetes možemo da vidimo da li je neka supstanca više zastupljena u telu bolesne osobe nego u telu zdrave i da tu informaciju kasnije koristimo kao biomarker za tu bolest. Biomarkeri nam pomažu pri dijagnostifikovanju bolesti, kao i prilikom uočavanja da li neka osoba dobro reaguje na terapiju i slično.

2.2 Proces

Osnovni zahtev koji imamo je da ceo proces uparivanja bude izvršen korišćenjem HE što nas dovodi do drugog uslova, a to je da serveru ne želimo da otkrivamo nikakve informacije o bazi niti o upitu.
Pretpostavimo da klijent ima VCF fajl (engl. Variation Call Format), koji sadrži informacije o genotipu, kao što su broj hromozoma i pozicija u genomu. Klijent enkriptuje informacije koristeći homomorfnu enkripciju i server računa tačno poklapanje nad enkriptovanim podacima. Ishod je prisustvo/odsustvo specifičnih biomarkera. Na kraju, klijent dekriptuje rezultat uz pomoć tajnog ključa homomorfne enkripcije.

3 Pretraga i izvlačenje podataka iz baze

Posmatrajmo bazu koja je skup od n torki (engl. tuples), gde se svaka torka sastoji iz para (d_i,α_i) za i = 1,...,n, gde je d_i oznaka podatka iz domena $\{0,1,...,\tau - 1\}$, a α_i odgovarajuća vrednost atributa u prostoru običnog teksta $Z_t\backslash\{0\}$. Želimo da se sve oznake teksta razlikuju tako da u slučaju baze podataka koja sadrži informacije o nekoj osobi, α_i može da bude broj godina korisnika čiji je identifikacioni broj d_i.

Kada imamo datu oznaku upita d iz domena oznaka i vrednost upita α iz prostora običnog teksta, problem spajanja se svodi na određivanje postojanja indeksa i, takvog da $(d,\alpha) = (d_i,\alpha_i)$. Posmatrajmo sada pojednostavljen upit za pretragu: odaberi α_i ako postoji indeks i takav da je $d_i = d$, inače nula.

3.1 Metod za bezbednu pretragu i izvlačenje podataka iz baze

Osnovna ideja je korišćenje narednog metoda za enkodiranje baze koji je pogodan za efikasno izračunavanje jednakosti i izvlačenje:

$$DB(X) = \sum_i \alpha_i X^{d_i} \in \mathbb{Z}$$

Korisnik enkriptuje polinom sa javnim ključem i šifrovani tekst čuva na serveru. U fazi ispitivanja upita, sa nazivom upita d, korisnik enkriptujeX^{-d} sa simetričnom enkripcijom, a šifrovani tekst šalje ka serveru.

3.2 Enkodiranje informacija o genomu

VCF fajl sadrži linije informacija o genotipu, gde se svaka od njih sastoji iz tripleta (ch_i,pos_i, SNPs_i) tj. broja hromozoma, pozicije i sekvence SNP alela (koji moraju biti A, T, G ili C).

Broj hromozoma je neki ceo broj od 1 do 22,X i Y. Nenegativni ceo broj pos predstavlja referencu na poziciju sa prvom bazom koja ima poziciju jedan, a SNPs je referenca na niz {A,T,G,C}.
Upit korisnika je, takođe, triplet ovakvog oblika, pa nam je cilj da odredimo postoji li ili ne prisustvo odgovarajućeg biomarkera u fajlu iz baze.

Postavićemo n_SNP za maksimalni broj alela za vršenje poređenja između genoma upita i korisničkog genoma. Svaki od SNP-ova je predstavljen sa dva bita tako da

$$A \to 00, T \to 01, G \to 10, C \to 11$$

a potom su ti bitovi konkatenirani. Potom stavljamo bit 1 na početak stringa sa leve strane da označimo početnu poziciju. I na kraju popunjavamo string dužine $l_{SNP} = 2n_{SNP} + 1$ i konvertujemo dobijeno u ceo broj, označen sa α_i. Ako se desi da je neki nukleotid nepoznat u datom delu, onda takav string enkodiramo nulama. Na primer: $'$GC$'$ je enkodiran binarno kao 11011, što je ceo broj 27.

4 Zaključak

Sa razvojem bioinformatike i sve većim potrebama za čuvanjem podataka u oblaku, možemo očekivati porast u broju, kvalitetu i pouzdanosti algoritama za enkripciju podataka, kao i razvoj sve boljih i bržih mehanizama za pretragu genoma.