Bezbedno traženje biomarkera korišćenjem hibridne homomorfne enkripcione šeme

Seminarski rad u okviru kursa
Kriptografija
Matematički fakultet

Stanković Una
una_stankovic@yahoo.com

16. maj 2017.

Sažetak

Sa sve bržim razvojem tehnologija za sekvenciranje genoma, raste i potreba da se očuva bezbednost podataka, posebno jer se oni, sve češće, čuvaju u oblaku i odatle koriste za istraživanje. U radu će biti predstavljen efikasan način da se bezbedno pretraži pozicija koja odgovara podacima iz upita i izvuku neke informacije sa te pozicije. Nakon dekriptovanja, postoji samo mala kolicina poređenja sa informacijama iz upita koja se vrši nad običnim tekstom. Ovaj metod će biti primenjen da se nađe skup biomarkera u enkriptovanim genomima.

Sadržaj

1 Uvod

Brz razvoj tehnologije sekvenciranja genoma dozvoljava nam da pristupimo velikim skupovima genoma i što bi moglo da nam omogući bitne pomake u medicinskim istraživanjima. 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.

Predloženo je da se privatnost podataka očuva korišćenjem homomorfne enkripcije (engl. Homomorphic Encryption - HE), koja 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.

Jasuda i drugi autori (engl. Yasuda et al.) su dali praktično rešenje za pronalazak lokacije uzorka u tekstu izračunavanjem više Hamingovih rastojanja nad enkriptovanim podacima. Loter i drugi (engl. Lauter et al.) su dali rešenje kako da se bezbedno izvrše osnovni genomski algoritmi korišćeni u mnogim sprovedenim studijama.

Homomorfna enkripcija se može primeniti za očuvanje privatnosti sekvence koju analiziramo, ali se pokazuje kao veoma nepraktična za analizu informacija kod kompletnog ljudskog genoma, već nam je pogodna za neke delove genoma. Glavna ideja, koja će biti korišćena za bezbedno pretraživanje skupa biomarkera korišćenjem Ring-GSW homomorfne enkripcione šeme, je da se enkodira baza genoma kao jedan element polinomijalnog prstena. Operacija pretrage u bazi se radi vršenjem jednog množenja sa genomom upita. Potom vršimo proceduru izvlačenja kako bismo dobili DNK sekvencu i nakon dekripcije je poredimo sa DNK upita u običnom tekstu.

2 Postavka problema

Zadatak je da se na bezbedan način izračuna verovatnoća genetskih bolesti kroz uparivanje skupa biomarkera sa enkriptovanim genomima koji se čuvaju u javnom "oblaku"(engl. cloud). Zahtev je da ceo proces uparivanja bude izvršen korišćenjem homomorfne enkripcije tako da se nikakve informacije o bazi i upitu ne otkriju serveru prilikom izračunavanja.
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 uz očuvanje privatnosti

Posmatrajmo bazu koja je skup od n torki (engl. tuples). Svaka torka se 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 $Z_t \backslash \{0\}$. Primetiti da sve oznake podataka treba da se međusobno razlikuju. Na primer, 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,α) = (d_i,α_i). Posmatrajmo sada pojednostavljen upit za pretragu: odaberi α_i ako postoji indeks i takav da je $d_i = d$, inače nula.

Glavni cilj nam je da server ne nauči nista iz enkriptovanog upita, kao ni da korisnik ne dobije bilo kakve informacije osim onih koje predstavljaju krajnji rezultat.

3.0.1 Metod za bezbednu pretragu i izvlačenje podataka

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 enkriptuje monom X^{-d} sa simetričnom enkripcijom, a šifrovani tekst šalje ka serveru.

3.1 Bezbedna pretraga biomarkera

U nastavku će biti opisano kako da se enkodira i enkriptuje informacija o genotipu iz VCF fajla kako bi se primenila bezbedna pretraga i izvlačenje iz baze.

3.1.1 Enkodiranje informacija o genomu

VCF fajl sadrži više linija sa informacijama 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.

Predstavićemo hromozome pola, X i Y, sa 0 i 23, a potom definisati funkciju enkodiranja sa ϵ: $\mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ sa

$$(ch, pos) \to d = ch + 24pos$$

Sada treba predstaviti enkodiranje za SNP. 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.

3.1.2 Enkriptovanje informacija o genomu

Fajl baze podataka je enkodiran kao skup parova (d_i, α_i) za i = 1,...,n, tako da je $d_i = \epsilon(\text{ch}_i, pos_i)$, a α_i je enkodirani ceo broj koji predstavlja i-ti SNP niz alela. Tada posmatramo enkodiranje d_i i α_i kao oznaku podataka i vrednost atributa. Korisnik podataka konstruiše polinom $DB(X) = \sum_k c_k X^k$ takav da

$$c_k = \begin{cases} \alpha_i, \text{ako je } k = d_i \text{ za neko} i \\ \alpha \leftarrow \mathbb{Z}_t, \text{inače} \end{cases}$$

Korisnik enkriptuje polinom sa RLWE šemom za enkripciju sa javnim ključem. RLWE (engl. Ring Learning With Errors) pretpostavka je da je je nemoguće razlikovati naredne dve raspodele. Prva se sastoji iz parova (a_i, u_i) gde se a_i i u_i izvlače na slučajan način iz prostora šifrovanog teksta R_Q. Druga raspodela se sastoji iz parova (a_i, b_i) = (a_i, $a_i\mathbf{s} + e_i$), gde je a_i slučajno odabrano iz R_Q, a $\mathbf{s}$, e_i su izvučeni iz raspodele greške χ, gde ono predstavlja raspodelu buke u prostoru R.

4 Zaključak