Matematički fakultet

Seminarski rad
iz tehničkog i naučnog pisanja

Bioinformatika u genomici

Student
Andrija Radović 133/23

Profesor
dr Jelena Graovac

Beograd, 19. novembar 2023.

Sadržaj

1 Uvod

Bionformatika je interdisciplinirana oblast koja se bavi razvojem metoda i alata za razumevanje bioloških podataka. Bioinformatika ima široku i veoma značaju primenu, a njena primena je u: genskoj terapiji, evolucionoj biologiji, razvijanju novih lekova, agronomiji i genomici. U ovom tekstu ćemo se susreti sa njenom primenom samo u genomici. Takođe ćemo preći preko osnovnih softverskih alata, programa i baza podataka koje se danas koriste.

Definicija 1.1 *Genomika je oblast molekularne biologije koja proučava strukturu, funkciju, organizaciju, evoluciju i mapiranje genoma.*

2 Alati u bionformatici

Radi uspešnog i olakšanog rada, u bionformatici se razvijaju različita pomagala koja pomažu programeru da obave svoj posao. Glavna podela ovih pomagala je:

1. softveri i alati
2. algoritmi
3. baze podataka

Softverski alati koji se koriste mogu biti jednostavne komande, napredni grafički programi ili veb-usluge.

Koriste se otvoreni programi (eng. *open source software*) koji obično predstavljaju mesto gde sve ideje nastaju, takođe su zaslužni za određivanje standarada. Neki od bioinformatičkih softvera su: Biopython, BioJava i mnogi drugi.

Veb usluge omogućavaju naučnicima da koriste algoritme i podatke iz drugih servera sa drugog kraja sveta. Glavna prednost ovih usluga u odnosu na obične programe je ta što njihovi korisnici ne moraju da razmišljaju o samom softveru i njegovim podacima.

Jedan od najkorišćenijih bioinformatičkih algoritama jeste BLAST. BLAST je algoritam za poređenje primarnih bioloških sekvenci, poput proteinskih i nukleinskih sekvenci. Omogućava naučnicima da porede sekvence sa bibliotekom ili sekvencionom bazom podataka kako bi identifikovali sekvencu. Kada se u nekoj vrsti otkrije neki novi gen koji želimo da potražimo u genomu čoveka, koristićemo BLAST kako bi oktrili sekvence u genomu koje su slične novom genu.

Baze podataka su neophodne za bioinformatička istraživanja i njihovu primenu. Postoje u različitim oblicima i sadrže razne vrste informacija, poput: DNK sekvenci, biodiverziteta... (više će ih biti navedeno u tabeli 1). U bazama se mogu naći i empirijski podaci kao i pretpostavljeni podaci (eng. *predicted data*). Neke od upotreba baza podataka:

- strukturna analiza
- pronalaženje proteinskih porodica
- dizajniranje sintetičkih genetskih kola

Tabela 1: Neke od baze podataka u bioformatici.

Tip baze podataka	Opis	Primer
Bibliografska	Sadrži istraživačke i naučne radove	MEDLINE
Genomska	Sadrži genomske sekvence različitih organizama	GIB
Sekvenciona	Sadrži sekvence proteina i nukleotida	DDBJ
Strukturna	Sadrži 3D strukture proteina i nukleinskih kiselina	NDB
Metabolička	Razni biološki putevi	KEGG
Enzimska	Sadrži podatke o strukturi, funkciji i putevima različitih enzima	REBASE
Hemijska	Podaci o biološkoj aktivnosti nekoliko malih molekula	PubChem
Mikromatrična	Podaci dobijeni iz mikromatričnih eksperimentima	GEO

3 Primena bioinfromatike u genomici

Jedna od najširih primena bionformatike jeste u genomici. Razlog toga je to što se u genomici susrećemo sa genomima koji su često preveliki za ručnu obradu, pa se sve više koriste algoritmi i bionformatički alati.

Definicija 3.1 *Genom je skup gena koje sadrži jedna haploidna ćelija.*

3.1 Genetika bolesti

Što se tiče genetike bolesti (eng. *genetics of disease*) primena bioinformatike se svodi na razvijanje i poboljšavanje već postojećih tretmana za različite bolesti. To možemo uraditi na više načina.

Jedan od načina na koji bioinformatika pozitivno utiče na razvoj genetike bolesti jeste oktirvanje gena koji su vezani za neku specifičnu bolest koju istražujemo. Informacije o ovim genima nam mogu biti korisne pri izradi novih lekova i terapija koje se fokusiraju na te gene. Takođe može pomoći pri dijagnozi bolesti pre nego što one pokažu prve simptome analizom pacijentovog DNK.

Danas se bionformatika u ovoj oblasti sve više koristi za razvijanje personalizovane medicine, medicine koja je namenjena isključivo jednoj osobi na osnovu njihove genetike. Ovakva primena bioinformatike već ima uticaj na način lečenja određenih oblika raka. [1]

3.2 Pangenom

Pangenom (eng. *pangenome*) je jedan od novijih koncepata koji je definisao Tettelin (Tetelin) kao kompletan repertoar gena neke konkretne taksonomske grupe, mada ju je Tettlin specifično namenio bakterijama čiji pangenom sadrži izvorni genom koga čine geni prisutni u svim sojevima kao i promenjivi genom koga čini set gena koji su jedinstveni za svaki soj.

Pangenomska analiza nam omogućava razvijanje univerzalnih vakcina koje bi mogle biti korišćene na sve sojeve jedne vrste, ili čak na više srodnih vrsta bakterija. Zbog toga se javlja potreba za sve većom efikasnosti algoritama i strukture podataka, a time se u ovom slučaju bavi bioinformatika. [2]

Neki od bioinformatičkih alata i algoritama koji se koriste su:

- OrthoMCL: "genome scale"algoritam koji grupiše proteine u tzv. "ortologne grupe". [3]
- COG (Clusters of Orthologous Groups), InterPro i KEGG (Kyoto Encyclopedia of Genes and Genomes) su bioinformatičke baze podataka koje

sadrže informacije o načinu raspodele gena unutar izvornog i promenjivog genoma.

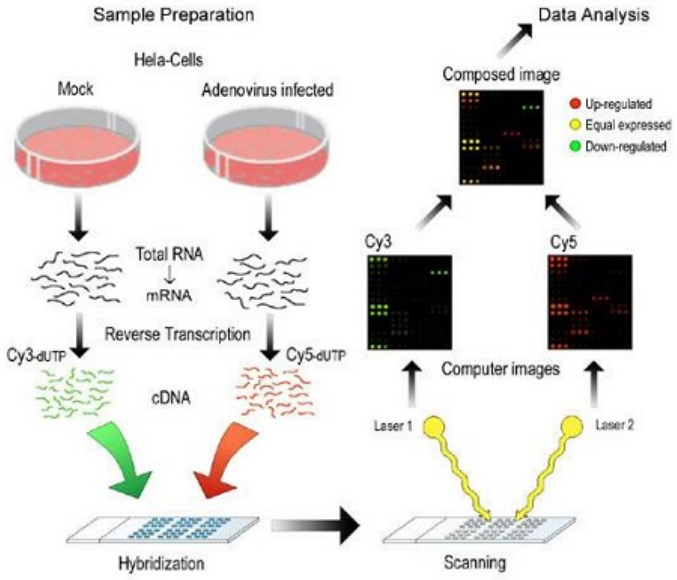
- P2RP (Predicted Prokaryotic Regulatory Proteins): Onlajn alat koji služi za istraživanje i identifikaciju regulatora proteinske ekspresije i povezanih faktora transkripcije.

3.3 Projekat ljudskog genoma

Projekat ljudskog genoma (eng. *Human Genome Project*) je bio naučno istraživački projekat čiji je cilj bio određivanje sekvenci parova baza koje čine DNK čoveka, kao i mapiranje ljudskog genoma.

Bionformatika je imala ključnu ulogu u Projektu ljudskog genoma, a njena uloga je bila upravljanje i analiza ogromne količine genomskih podata koji su nastali u toku ovog projekta. Bionformatički alati su obrađivali i analizirali milijarde DNK sekvenci, a potom su ih organizovali u baze podataka. DNK sekvence su zatim bile slagane u potpune genomne sekvence koje su bile obeležene raznim infromacijama o genima.

Posledica Projekta ljudskog genoma je razvoj novih i moćnih bioinformatičkih alata poput mikromatrica (eng. *microarrays*). Mikromatrice su omogućile detekciju hiljade gena u isto vreme, tj. analizu genoma bez sekvenciranja, što je smanjilu cenu velikih istraživanja u različitim oblastima biologije.



Slika 1: Mikromatrica

Na slici 1 je prikazana mimromatrica, njen princip rada kao i rezultat njenog rada.

3.4 Komparativna genomika

Komparativna genomika (eng. *comparative genomics*) je oblast bioinformatike koja uključuje poređenje genoma dve različite vrste, ili dva različita soja/rase iste vrste.

Komparativna genomika nam može pomoći u boljem razumevanju evolucije, razumevanju koji geni utiču na različite biološke sisteme, što nam možda u budućnosti bude pomoglo pri razvijanju novih načina za lečenje različitih oboljena i poboljšanju celokupnog ljudskog zdravlja.

Neki od zanimljivh rezultata istraživanja na ovu temu su:

- mapiranje gena koji povećavaju prinos mleka kod mlečnih krava.
- poređenjem genoma pedeset ptica pevačica pronađen je mreža gena koja je možda imala veliki uticaj na razvoj govora kod ljudi.
- mapiranje genomike različith bolesti i raka kod pasa, što nam može omogućiti novi pogled i pristup raku kod ljudi.

Zbog svega ovoga je bitan razvoj novih bioinformatičkih algoritama i programa. Bionformatički alati korišćeni u komparativnoj genetici:

- Opscan je program koji detektuje ortologne gene u dve sekvence gena, koji su potrebni za otrkivanje srodnosti između vrsta. U odnosu na druge programe poput BLAST Opscan je znatno brži.

  Definicija 3.2 *Ortologni geni su geni različitih vrsta koji su nastali vertikalnim nasleđivanjem iz jednog gena najbližeg zajedničkog pretka.*

- GeneOverlap je R paket koji na brz i lak način može porediti dva niza gena koji imaju slične atribute ili karakteristike.

  Definicija 3.3 *R paketi su su skupovi funkcija, podataka i kompiliranih kodova skladištenih u specifičnom formatu. Oni se koriste u R programskom jeziku koji se koristi za statističke izračune i grafikone.*

- Phylostratr je takođe R paket koji služi za pravljenje filostratigrafije (eng. *phylostratigraphy*) za sve gene u genomu.

  Definicija 3.4 *Filostratigrafija je metoda razvijena za određivanje porekla specifičnih gena postmatrajući homogoliju više vrsta.*

3.5 Označavanje genoma

Označavanje genoma je postupak identifikovanja lokacije gena i svih kodirajućih regija u genomu, kao i utvrđivanje njihovog delovanja. Pošto genomi predstavljaju ogromni skup podataka, teško je označiti ceo genom ručno.

Označavanje gena trenutno predstavlja „usko grlo“ (eng. *bottleneck*) u bioinformatici. Tehnologija za označavanje genoma ostaje usko grlo zato što sam proces označavanja nije u potpunosti jasan. Veliki centri poput NCBI-a, nemaju dovoljno ljudi, a ni računara kako bi sve svoje kanale za označavanje primenili na sve nove genome, dok mali i neiskusni timovi ne znaju odakle da počnu jer kanali za označavanje ne rade podjednako dobro na svim genomima. [4]

Sekvenciranjem gena dobijamo sekvencu informacija bez njihove funkcionalne uloge, i zato je bitno označiti ga kako bi mu pridodali značenje o njegovoj strukturi i ulozi. Zbog toga je označavanje gena bitno u dijagnozi bolesti jer može otkriti gene koji izazivaju različita oboljenja. Zbog dijagnoza bolesti i otkrivanja njihovog porekla, važno je razvijati genomske baze podataka i softverske alate za sekvenciranje genoma.

Softveri koji se koriste:

- za označavanje genoma eukariota koristi se FINDER
- za označavanje genoma prokariota koriste se Bakta, Prokka i PGAP

4 Zaključak

U ovom seminarskom radu smo prešli preko osnovnih softverskih alata i baza podataka koje se koriste u bioinformatici kao i njenoj primeni u genomici. Bioinformatika ima široku primenu u genomici, a neke od oblasti gde se primenjuje su: genetika bolesti, pangenomika, komparativna genomika... Videli smo da je razvoj ove interdisciplinirane oblasti bitan, jer nam može pružiti uvod u razvoj bolesti, njihovo lečenje, razvoj lekova, razumevanje evolucije i odnosa između organizama.

Literatura

[1] Khalid, R. and Nilanjan, D. *Translational Bioinformatics in Healthcare and Medicine*. London, 2021.

[2] Debmalya B., Siomar C. Soares, Sandeep T. and Vasco Ariston De Car Azevedo *Pan-genomics: Applications, Challenges, and Future Prospects*. London, 2020.

[3] *OrthloMCL, on-line at: https://orthomcl.org/orthomcl/app*

[4] Tomáš B., Heng Li, Joseph G., Daniel H., Steffen H., Mario S., Natalia N., Matthis E., Lars G. and Katharina J. H. *GALBA: Genome Annotation with Miniprot and AUGUSTUS*.