

Retsensioon Anton Stalnuhhini bakalaureusetööle

“Genetic Algorithm for the Improved Discovery of DNA Regulatory Elements”

DNA reguleerivate elementide esitamiseks on mitmeid võimalusi, üheks enimkasutatud on positsioonispetsiifilised kaalumaatriksid (ingl.k. lühend PWM). Anton Stalnuhhini bakalaureusetöö uurib üht võimalust kaalumaatriksite parendamiseks geneetilise algoritmi abil. Seejuures on eesmärk jõuda kaalumaatriksini, millega sobivaid elemente on võimalikult palju ühes DNA lõikude kogumis ning võimalikult vähe teises.

Töö on suurepärast vormistatud ja sobib eeskujuks teistele. Viitamine on korrektne, trükivigu ma peaaegu ei leidnud, töös on sobival hulgal illustratsioone, vajadusel on kasutatud värve. Kõik nõutud komponendid on olemas ja lisaks veel ingliskeelne abstrakt, mida vist ingliskeelse töö korral tingimata olema ei pea. Eestikeelne abstrakt tundub olemas ingliskeelse otsetõlge, sisaldades paari eesti keeles mittekasutatavat konstruktsiooni.

Töö ülesehitus on hea. Esimeses peatükis on antud sobiva detailsusastmega bioloogiline taust, teises lahti seletatud kaalumaatriksi mõiste. Kolmas peatükk kirjeldab geneetiliste algoritmide põhitõdesid ning selgitab, kuidas neid käesolevas töös on kasutatud. Viimane peatükk toob ära saadud tulemused. Lisadena on esitatud tulemuste tabelid ning CD-plaat kirjutatud Java-programmidega ning eksperimentide lähteandmete ja tulemustega.

Uurimistöö aktuaalsus on märkimisväärne, arvestades et töö kirjutamise ajal ilmus sisuliselt samal ideel põhinev artikkel.

Järgnevalt loetlen tööd lugedes tekkinud küsimused ja märkused:

- 1) Sissejuhatuses on öeldud, et kuivõrd mainitud artikkel avaldati enne bakalaureusetöö esitamist, siis kasutati võimalust tulemuste võrdlemiseks. Oleks eeldanud võrdluse esitamist (või vähemalt sarnasuste/erinevuste esitamist) ka bakalaureusetöös. Kahjuks pole selle artikli kohta rohkem midagi mainitud.
- 2) Lk. 17 on ära toodud teisendused, mida on kasutatud geneetilises algoritmis uue generatsiooni kaalumaatriksite saamiseks. See on bakalaureusetöö võtmeidee. Kahjuks jääb selgusetuks, kuidas neid rakendatakse. Kas valitakse juhuslikult üks loetletud moodustest? Kuidas toimitakse siis, kui kasutatakse viimast, nn. kompleksoperatsiooni, mis koosneb kümnest juhuslikult valitud operatsioonist? Mitmest algsest kaalumaatriksist ja kuidas siis tulemus saadakse?
- 3) Miks on kasutusele võetud motiivi praktiliselt vastupidiseks muutev inversiooni-teisendus? Kas seda ei võiks lihtsalt ära jätta?
- 4) Kokkuvõttes on öeldud, et tehisandmehulga korral näitab optimeeritud PWMi sekventsilogo selgelt, kuidas algse teksti sisse pandud motiiv taastatakse. See on just see, mida oleks soovinud, kuid kahjuks ju seda taastamist ei toimu. Rohkem oleks oodanud selgitust, miks see nii on.

- 5) Valk ei pruugi olla alati ahel aminohapetest nagu on väidetud lk.5, vaid võib koosneda ka mitmest ahelast.
- 6) Tiitellehel on "autor" ja "juhendaja" kirjutatud eesti keeles, samal real olev kuupäev on inglise keeles ("june").
- 7) Lk 15 ja 16 toodud joonistel krossingoveri ja mutatsioonide kohta oleks võinud värvilised pildid ja nende all olevad DNA järjestused omavahel vastavusse panna ja joondada. Praegusel kujul on väga raske jälgida, millised tähed on paksus kirjas ja kuidas DNA muutub.
- 8) Lk 23 on viidatud joonisele 4.3, kuigi ilmselt on mõeldud joonist 4.5.
- 9) Lk 24 on öeldud, et ROC AUC paranes 0.7866-lt 0.7842-le, kuigi selliste arvude korral oleks tegemist halvenemisega.

Kokkuvõttes on töö autor mõelnud välja sobivad geneetilises algoritmis kasutatavad teisendused, teinud ära vajaliku programmeerimistöö, teostanud eksperimendid ja kirjutanud kokku väga hea vormistusega töö. Leidsin kaks sisulist probleemi, mis nõuaksid rohkem vaeva (küsimused 1 ja 4). Anton Stalnuhhin väärrib minu hinnangul bakalaureusekraadi ja hea kaitsmise korral pakun hindeks "A".

Meelis Kull
25. juunil 2007