**Globalaus sekų palyginimo programos (Needleman–Wunsch algoritmo) projektas ir analizė**

**Autorius:** Daniel Volcak
**Universitetas:** Vilniaus universitetas
**Fakultetas:** Informatikos fakultetas
**Studijų programa:** Bioinformatika
**Kursas:** Bioinformatika III
**Metai:** 2025

## 1. Įvadas ir užduoties formuluotė

Biologinių sekų palyginimas yra vienas iš fundamentalių bioinformatikos uždavinių, taikomų tiriant genetinę informaciją, baltymų struktūras bei evoliucinius ryšius tarp organizmų. Sekų palyginimo metodai leidžia nustatyti homologijas, identifikuoti konservuotus regionus ir daryti prielaidas apie biologinę funkciją.

Šio darbo tikslas – parengti **globalaus sekų palyginimo programos projektą (specifikaciją)**, pagrįstą **Needleman–Wunsch** algoritmu, bei pateikti šio metodo taikymo analizę. Programa projektuojama kaip komandinės eilutės (CLI) įrankis, galintis veikti Unix tipo sistemose ir būti naudojamas automatizuotuose bioinformatikos skaičiavimų srautuose.

Užduoties metu buvo siekiama: formaliai aprašyti programos funkcionalumą ir sąsają su naudotoju, apibrėžti įvesties ir išvesties duomenų formatus, pateikti algoritmo matematinį pagrindą, užtikrinti, kad pagal specifikaciją būtų įmanoma realizuoti suderinamą programos versiją.

## 2. Naudoti įrankiai ir metodai

### 2.1. Duomenų šaltiniai

Sekų palyginimui buvo naudojamos biologinės sekos **FASTA** formatu. FASTA formatas yra plačiai paplitęs bioinformatikoje ir naudojamas tiek DNR, tiek baltymų sekoms aprašyti. Pavyzdiniuose skaičiavimuose naudotos testinės sekos, leidžiančios aiškiai pademonstruoti algoritmo veikimą.

FASTA formato pasirinkimas užtikrina suderinamumą su daugeliu bioinformatinių įrankių ir leidžia lengvai integruoti programą į esamus analizės procesus.

### 2.2. Algoritmas ir metodas

Darbo pagrindą sudaro **Needleman–Wunsch** algoritmas, aprašytas 1970 m. (Needleman ir Wunsch, 1970). Tai klasikinis **dinaminio programavimo** metodas, skirtas **globalaus sekų palyginimo** uždaviniui spręsti.

Algoritmas remiasi balų matricos sudarymu, kur kiekvienas elementas apskaičiuojamas pagal rekursinę formulę:

- sutapimo arba nesutapimo balą;
- tarpo (gap) įvedimo baudą.

Užpildžius matricą, atliekamas **atsekimas (traceback)** nuo paskutinio matricos elemento iki pradžios, leidžiantis atkurti optimalų sekų išlygiavimą. Šis metodas garantuoja optimalaus globalaus sprendinio radimą pagal pasirinktą balų sistemą.

### 2.3. Programinė įranga ir versijavimas

Projektuojama programa aprašyta kaip **CLI įrankis**, veikiantis Debian, Ubuntu ir suderinamose sistemose. Specifikacijos kūrimo procesas buvo versijuojamas naudojant **Git**, užtikrinant visų tarpinių dokumento versijų išsaugojimą ir atsekamumą.

**3. Darbo eiga ir atkuriamumas**

Programos projektas numato, kad įrankis veikia kaip **Unix filtro tipo programa**: įvestis gali būti pateikiama per failus arba **STDIN**. Rezultatai rašomi į **STDOUT**, o klaidų pranešimai pateikiami per **STDERR**.

Tokiu būdu programa gali būti lengvai integruota į automatizuotas analizės grandines (pipeline). Skaičiavimų atkuriamumas užtikrinamas pateikiant:

- tikslią programos iškvietimo sintaksę;
- aprašytus įvesties ir išvesties formatus;
- Git repozitoriją su pilna projekto kūrimo istorija.

**4. Rezultatai**

Pritaikius Needleman–Wunsch metodą testinėms sekoms, gaunami globalūs sekų išlygiavimai, kuriuose aiškiai matomos sutampančios pozicijos, nesutapimai ir įvesti tarpai. Rezultatų analizė parodė, kad algoritmo veikimas yra stabilus ir jautrus balų sistemos parametrams.

Rezultatai gali būti apibendrinti **dažnių lentelėje**, kurioje pateikiamas:

- sutapimų skaičius;
- nesutapimų skaičius;
- įvestų tarpų skaičius.

Tokios lentelės leidžia kiekybiškai įvertinti sekų panašumą ir palyginti skirtingus parametrų rinkinius.

**5. Aptarimas**

Needleman–Wunsch algoritmas yra ypač tinkamas tada, kai lyginamos sekos yra panašaus ilgio ir tikimasi, kad jos yra homologinės per visą ilgį. Tačiau realiuose bioinformatiniuose duomenyse dažnai pasitaiko atvejų, kai homologija apsiriboja tik dalimis sekos. Tokiais atvejais efektyvesni gali būti lokalūs palyginimo metodai, tokie kaip Smith–Waterman algoritmas.

Nepaisant to, globalus palyginimas išlieka svarbus įrankis filogenetinėse analizėse, genų šeimų tyrimuose ir sekų anotacijoje.

**6. Išvados**

Šiame etaloniniame pavyzdyje pateikta globalaus sekų palyginimo programos specifikacija ir jos analizė. Darbas parodo, kaip: formaliai aprašyti CLI programą pagal pateiktas gaires, dokumentuoti algoritminį pagrindą, užtikrinti skaidrią kūrimo istoriją naudojant versijų valdymo sistemas.

Toks struktūruotas požiūris leidžia sukurti aiškiai apibrėžtą ir atkuriamą bioinformatinį įrankį.

**7. Didelių kalbos modelių (DKM) naudojimas**

Šio darbo rengimo metu buvo naudojami dideli kalbos modeliai (DKM), siekiant **pagerinti teksto kalbos taisyklingumą, stiliaus nuoseklumą ir akademinę formą**. Taip pat DKM buvo naudojami **pagalbai tvarkant Git komandų seką ir repozitorijos struktūrą**. DKM **nebuvo naudojami** generuoti darbo dalykinį turinį, skaičiavimus, rezultatus ar išvadas. Visas analizės turinys parengtas autoriaus savarankiškai.

Naudotas įrankis:
OpenAI (2025). *ChatGPT*, modelis GPT-5.2.
URL: https://chat.openai.com

**8. Literatūra**

Needleman, S. B. and Wunsch, C. D. (1970) *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. Journal of Molecular Biology, 48(3), pp. 443–453.

Chue Hong, N. P. et al. (2019) *Software citation checklist for authors*. Zenodo. https://doi.org/10.5281/ZENODO.3479198

Katz, D. S. et al. (2021) *Recognizing the value of software: a software citation guide*. F1000Research, 9, 1257.

University of Wolverhampton (2022) *Harvard referencing: the basics*. Available at: https://www.wlv.ac.uk/lib/media/departments/lis/skills/study-guides/LS134-Harvard-Quick-Guide.pdf