Liet. matem. rink., T. 43, spec. nr., 532–537

Psi blast algoritmo efektyvumas priklausomybėje nuo skenavimo langa dydžio

Mindaugas MARGELEVIČIUS, Albertas TIMINSKAS (Biotechnologijos inst.)
el. paštas: minmar@ibt.lt

1. Įvadas

Baltymai yra ilgi amino rūgščių polimerai. Iš viso yra 20 skirtingų amino rūgščių (a. r.). Amino rūgščių išsidėstymas baltymo polimere nėra atsitiktinis, ir jų konkretus derinys lemia trimačią baltymo struktūrą. Kuriami įvairūs homologinės (genealogiškai susijusių sekų) paieškos algoritmai, kurie naudojami baltymų tretinei struktūrai prognozuoti, baltymų ir jų segmentų klasifikavimui, naujų baltymų savybių tyrimui. Kadangi baltymų duomenų bazių apimtys yra didelės, išvedami euristiniai algoritmai, kuriuose optimizuojamas spartumo ir tikslumo santykis.

Šiame straipsnyje aptarsime PSI-BLAST (*position-specific iterated Basic Local Alignment Search Tool*) euristinio algoritmo ypatumus bei jo modifikacijos aspektus. Taikysime modifikuotą algoritmą baltymų paieškai, aptarsime rezultatus.

2. PSI-BLAST algoritmas

Išlyginimo (*alignment*) algoritme PSI-BLAST įvesta aukšto įverčio segmentų poros (HSP) sąvoka, reiškianti lokaliai optimalius langa dydžio w porų segmentus, kurių įvertis negali būti padidintas. Tokių HSP sekoje gali būti keletas, ir, jei įvertis didesnis už nurodytą slenkstinį parametrą T (vadinsime kirčiu), pradedamas plėtimo procesas. Plėtimas stabdomas, kuomet porų segmentų įvertis sumažėja per leistino pokyčio dydį X. Galutiniam įverčiui viršijant užduotą reikšmę S, seka laikoma panaši į užklausą (baltyminė seka, pagal kurią atliekamas išlyginimas), ir rezultatas išsaugojamas.

Sekos plėtimo žingsnis, skaičiuojant įverčius, yra brangus, nes sudaro >90% BLAST skaičiavimo laiko, o jo atlikimų skaičius priklauso nuo parametrų T ir X reikšmių. PSI-BLAST patobulino BLAST algoritmą: (1) „dviejų kirčių“ (*two-hit*) metodo pritaikymu, kuris įverčio skaičiavimo plėtimą pradeda, jei toje pačioje matricos (kur eilutės ir stulpeliai atspindi dvi lygiuojamas sekas) *įstrižainėje* du nepersidengiančių langa dydžio segmentų įverčiai yra ne mažesni nei T, o segmentai vienas nuo kito nutolę per nuotolį A; (2) naudojamas su tarpeliais (*gapped*) algoritmas, kur užtenka rasti vieną išlyginimą (HSP) vietoje jų eilės, tai žymiai padidina algoritmo spartumą, o tikimybė, kad bus prarandamas reikšmingas išlyginimas, sumažėja; (3) paieška gali būti iteruojama naudojant ankstesnėje iteracijoje sugeneruotą nuo pozicijų priklausomą įverčių matricą; tokie paieškos metodai yra jautresni nei lyginimo poromis metodai [2].

Segmentų įvertis skaičiuojamas $S' = \sum_{i,j} P_i P_j s_{ij}$, čia P_i a. r. pasirodymo tikimybė i-oje pozicijoje, s_{ij} – dviejų skirtingų sekų a. r. panašumo įvertinimas pagal a. r. panašumo lenteles (PAM, BLOSUM) [3]. Du empiriškai (lyginant nesusijusias sekas) nustatomi parametrai λ ir K naudojami nominalaus HSP įverčio *S'* normalizavimui (bitais):

$$S = \frac{\lambda S' - \ln K}{\ln 2}. \tag{1}$$

Sekų segmentų, kurių įvertis didesnis už S, skaičiaus z pasiskirstymas yra puasoninis su vidurkiu $\mu(S) = Kmn\mathrm{e}^{-\lambda S}$, čia *mn* duomenų bazės paieškos apimtis (žr. žemiau). Tuomet taikant ekstremalių reikšmių pasiskirstymą (EVD) [4], tikimybė rasti c ar daugiau skirtingų HSP, kurių įvertis didesnis ar lygus S:

$$P(z \geqslant c) = 1 - \mathrm{e}^{-\mu(S)} \sum_{k=0}^{c-1} \frac{\mu(S)^k}{k!}. \tag{2}$$

Pageidautina, kad vidurkis μ būtų žymiai mažesnis už 1, ir parametrai λ ir K parenkami tenkinantys šią sąlygą. PSI-BLAST algoritme įverčių statistinio reikšmingumo įvertinimui parinkta $c = 1$. Parametrai, išlyginimo algoritmui su tarpeliais, nustatomi generuojant atsitiktines sekas. Nustatyta, kad šių parametrų svyravimas neviršija 2% ribos [2].

Esminį vaidmenį algoritmo skaičiavimuose vaidina E reikšmė, kuri nusako tikėtiną skirtingų HSP, kurių normalizuoti įverčiai ne mažesni nei S, skaičių m ir n ilgio sekų palyginimui:

$$E = N/2^S, \tag{3}$$

čia $N = mn$ yra paieškos erdvės apimtis; bendru atveju n galima laikyti a. r. kiekį duomenų bazėje. Algoritmas naudoja pasirinktą E reikšmę normalizuotam slenkstiniam įverčiui apskaičiuoti. Plėtimas pradedamas, jei sekų poros segmentų įvertis S yra didesnis už slenkstinį, ir pradedamas nuo rasto maksimalaus įverčio pozicijos.

PSI-BLAST jautrumas pagerintas, naudojant įverčių pagal pozicijas matricas (motyvus). Tikimybės motyvuose priklauso nuo konkrečios pozicijos, todėl įverčių sistema tampa jautresnė. Sudėtinis (*multiple*) išlyginimas M, reikalingas sudaryti jautresnį motyvą, atliekamas surenkant visus paieškos sekų segmentus, kurie buvo išlyginti pagal užklausą su E reikšme, mažesne už slenkstinę. Taip pat įvertinamas nepriklausomų išlyginimo M sekų segmentų santykinis skaičius N_C.

Motyvo įverčiai išreiškiami logaritmo pavidalu $\log(Q_i/P_i)$, čia Q_i – tikimybė, kad i-oji a. r. bus randama kažkuriame stulpelyje:

$$Q_i = \frac{\alpha f_i + \beta g_i}{\alpha + \beta}. \tag{4}$$

PSI-BLAST svorius parenka empiriškai: $\alpha = N_C - 1$, $\beta = 10$. Formulė (4) remiasi pseudo-dažnumų g_i įvertinimais ir išlyginime M apskaičiuotais a. r. dažnumais f_i.

Dažnumų įvertinimui turi būti panaudojamas *a priori* skirstinys (pvz., Dirichlet). PSI-BLAST naudoja Tatusov [5] priklausomiems duomenims pasiskirstymą. Tuomet pseudo-dažnumai stulpeliui išreiškiami

$$g_i = \sum_j \frac{f_j}{P_j} q_{ij}. \tag{5}$$

Čia i ir j a. r. pasirodymo (išlyginimo) tikimybė q_{ij}, išreiškiama per panašumo lentelės reikšmes s_{ij}:

$$q_{ij} = P_i P_j \exp(\lambda s_{ij}). \tag{6}$$

3. Algoritmo realizacijos modifikacija

Mes modifikavome NCBI (*National Center for Biotechnology Information*) PSI-BLAST kompiuterinę programą *blastpgp* v 6.104 2001 (`http://www.ncbi.nlm.nih.gov/BLAST`), praplėsdami lango dydį nuo 3 iki 5. Modifikavimo tikslas, praplečiant skenavimo langą,– padidinti algoritmo selektyvumą spartumo kaina. Statistinei analizei buvo pasirinkta citozino-5 metiltransferazės fermentų šeima. SWISS-PROT duomenų bazė (DB, `http://us.expasy.org/sprot`), kurioje buvo atliekama šių fermentų panašumo paieška, buvo papildyta šiomis fermentų šeimomis: (1) citozino-5 metiltransferazės šeima (336 nariai); (2–4) amino-metiltransferazės alfa, beta ir gama šeimomis (175, 251, 343 nariai atitinkamai); (5) visi kiti žinomi metiltransferazių fermentai (233 nariai). Bendras metiltransferazių, papildžiusių DB, skaičius – 1338.

4. Statistinė analizė

Skaičiavimo priemonių aplinka, kurioje atlikome tyrimus, apibrėžiama tokiomis parametrų reikšmėmis: a. r. panašumų lentelė – BLOSUM62; bendras sekų skaičius DB (įskaitant papildymą) – 125 802; parametras $T = 11$; parametras $A = 40$.

Statistinę analizę atlikome taikydami paieškos algoritmą kiekvienam citozino-5 metiltransferazės šeimos fermentui, keičiant lango dydį w nuo 3 iki 5. Bendras rastų baltyminių sekų skaičius kiekvienam iš atvejų matomas 1a pav. Matyti, kad didėjant lango dydžiui, rezultate atsiranda naujų homologijų (genealogiškai susijusių sekų). Kai $w = 5$, beveik kiekvienam fermentui atrandama po kelis ar daugiau panašių sekų nei ieškant, kai $w = 3$.

Fermentui 245 PgiAORF6P (1b pav.) paieška su $w = 5$ surado 173 homologijomis mažiau nei paieška su $w = 3$ (911 prieš 1084). Šiam fermentui skaičiavimo laikai (kai $w = 3$ ir $w = 5$) skiriasi per pusę, tačiau neviršija 7 min. Padidėjęs kirčių skaičius (387 805 136, $w = 5$; 203 715 261, $w = 3$) paaiškinamas tuo, kad skaičiavimams nebuvo pakeista slenkstinio parametro T reikšmė. Tokiu būdu orientavomės daugiau į algoritmo jautrumą nei į spartumą. Suprastėjusio selektyvumo priežastys šiam fermentui

gali būti aiškinamos tik analizuojant atrinktų baltyminių sekų biologinį artumą. Vidutinis paieškų skaičiaus skirtumas yra aiškiai teigiamas. Didžiausias teigiamas skirtumas (483 prieš 368) gautas fermentui 248 PhiHIBP (kirčių skaičius – 705 211 395, plėtimų skaičius 28 046 776, kai $w = 5$).

Bendras vidutinis rastų sekų skaičius vienam MF nežymiai didėja, didėjant lango dydžiui: 504,76 ($w = 3$), 513,62 ($w = 4$) ir 516,25 ($w = 5$). Tačiau tarp šių paieškos rezultatų vidutinis rastas skirtingų MF iš jų šeimų aibės, prijungtos prie duomenų bazės, skaičius yra bemaž vienodas – 345,03 ($w = 3$), 345,49 ($w = 4$) ir 345,22 ($w = 5$).

Paieškos radimų skirtumas metiltransferazės fermentams (MF), kai $w = 5$ ir $w = 3$, matomas 2a pav. Galime teigti, kad didžiausias neigiamas skirtumas matomas 1b pav., yra sąlygojamas MF radimų skaičiaus skirtumo. PSI-BLAST fermentui 245 PgiAORF6P, kai $w = 3$, surado 141 daugiau MF nei esant $w = 5$. Šis skirtumas labiausiai įtakoja vidutinį radimų skaičių. Tas pats rezultatas stebimas ir fermentui 248 PhiHIBP, kuriam algoritmas surado 54 MF sekų daugiau lango dydžiui 5 nei lango dydžiui 3. Taip pat galima pastebėti (2a pav.), kad fermentui 196 NgoAXAP rezultatai išsiskiria savo selektyvumu (kirčių skaičius – 858 424 454, plėtimų skaičius – 31 690 567, kai $w = 5$).

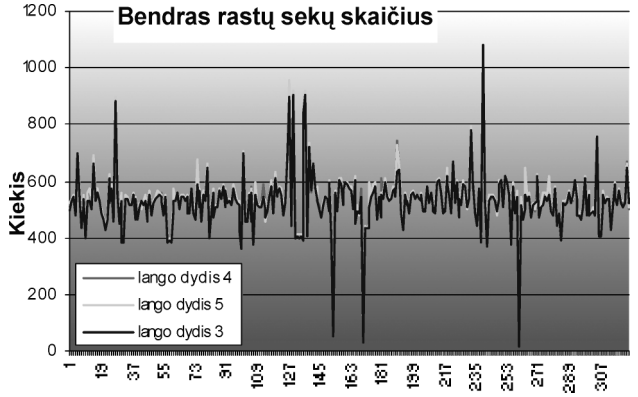
Didelius skirtumo nuokrypius stebime tik keliems MF, o neigiamą skirtumą nedaugeliui; kitiems bendra tendencija yra atrasti daugiau homologijų didėjant lango dydžiui. Tai galėtų būti aiškinama parametrų λ ir K euristiniu parinkimu, kuris kai kurioms sekoms gali iššaukti minėtus nelauktus pokyčius. Taip pat atlikti eksperimentai parodė, kad sėkmingų plėtimų ir HSP skaičių pasiskirstymai yra proporcingi (tie skaičiai auga didėjant w). Tai rodo, kad sėkmingi plėtimai priklauso nuo HSP skaičiaus. Be to didelis HSP skaičius tenkantis vienai duomenų bazės sekai nereiškia, kad bus surasta daugiausia skirtingų baltyminių sekų.

PSI-BLAST algoritmo vykdymo laikai pavaizduoti 2b pav. Pagal grafikus matyti, kad skirtingiems langams paieškos pagal kai kuriuos MF laikai skiriasi net per eilę. Tokių MF pavyzdžiai – 252 PliMCI, 329 XmaXhI; jų ilgiai dideli: ~ 1500 a. r. – ir jų homologinės paieškos laikas, didėjant lango dydžiui, smarkiai išauga. Kaip jau buvo minėta, tyrimams slenkstinio parametro T reikšmė nebuvo keičiama. Padidinus šią reikšmę, skaičiavimo laikas turėtų sumažėti. Vidutiniai laikai, tenkantys vienam MF – 1301 s ($w = 5$), 458 s ($w = 4$), 310 s ($w = 3$).

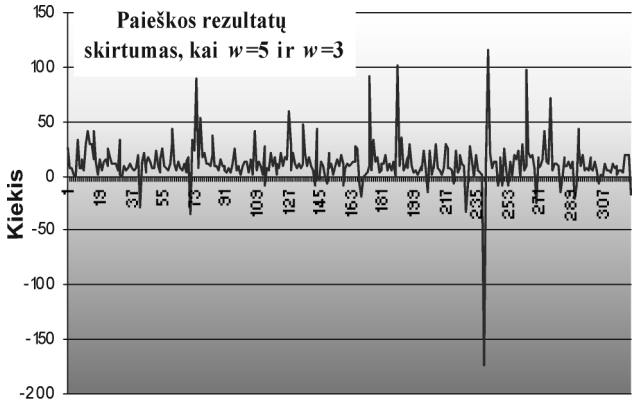
5. Išvados

Tyrimuose praplėstas lango dydis nuo $w = 3$ iki $w = 5$ sąlygojo vidutiniškai 12 atrastomis biologiškai panašiomis sekomis daugiau vienam metiltransferazės fermentui. Visiems w vidutinis rastas MF kiekis yra artimas ~ 345. Tačiau, kai $w = 3$, vidutiniam reikšmės skaičiavimui didžiausią įtaką turėjo vienintelis užklausos MF. Bendra tendencija atrasti daugiau panašių sekų metiltransferazėms didinant lango dydį, nors ir nežymiai, išlieka.

Vidutinis laikas vienam MF, skaičiuojant su $w = 5$, gautas 1301 s; su $w = 3$ – 310 s. Skaičiavimo laikai gali būti mažinami, didinant slenkstinių parametrų T ir A reikšmes.

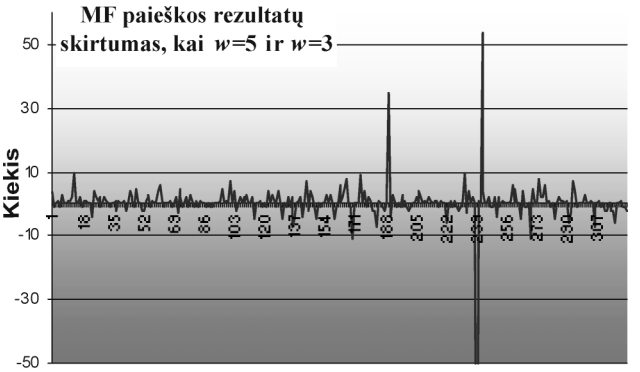


a)

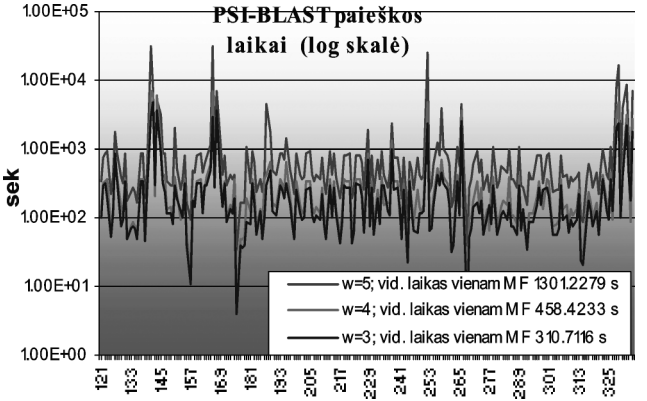


b)

1 pav. Paieškos rezultatai kintant lango dydžiui (abscisėje sužymėti numeriai identifikuoja konkretų fermentą iš jų šeimos).



a)



b)

2 pav. a) Rastų MF sekų skirtumas lango dydžiams 5 ir 3; b) algoritmo vykdymo laikų pasiskirstymas, kai w kinta nuo 3 iki 5.

Literatūra

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403–410 (1990).

[2] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919 (1992).

[3] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, **25**(17), 3389–3402 (1997).

[4] R. Durbin, S. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*, University Press, Cambridge (2001).

[5] R.L. Tatusov, S.F. Altschul, E.V. Koonin, Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks, *Proc. Natl. Acad. Sci. USA*, **91**(25), 12091–12095 (1994).

Effectiveness of the PSI BLAST algorithm in dependence on scanning word size

M. Margelevičius, A. Timinskas

The alignment searching algorithms for biomolecular sequences PSI-BLAST are reviewed in the article. We asses the feasibility to modify NCBI (National Center for Biotechnology Information) PSI-BLAST implementation. We modify the implementation of the iterative algorithm by increasing scanning word size. We also accomplish statistical analysis for each enzyme of the cytosine-5 methyltransferase family after we apply alignment searching for the biologically related protein sequences.