

**VILNIAUS UNIVERSITETO
FIZIKOS FAKULTETO
CHEMINĖS FIZIKOS INSTITUTAS**

Tautvydas Taraškevičius

**SPEKTRINIŲ ŽYMENŲ IDENTIFIKAVIMAS BIOLOGINIŲ BANDINIŲ
INFRARAUDONOSIOS SUGERTIES SPEKTRUOSE STATISTINĖS ANALIZĖS
METODAIS**

MAGISTRANTŪROS STUDIJŲ BAIGIAMASIS DARBAS

Cheminės fizikos studijų programa

Studentas	Tautvydas Taraškevičius
Leista ginti	2026-05-20
Darbo vadovas	prof. Justinas Čeponkus
Instituto/Centro direktorius/atstovas	prof. Darius Abramavičius

Vilnius 2026

Turinys

Įvadas	3
1. Teorinė dalis.....	4
1.1. Virpesinės spektroskopijos pritaikymas biologiniuose tyrimuose.....	4
1.2. Spektrų pirminio apdorojimo procedūros	5
1.2.1. Bazinės linijos korekcija	6
1.2.2. Spektrų normavimas	7
1.2.3. Pašalinių medžiagų kompensavimas.....	8
1.3. Spektrinių žymenų išgavimo metodai.....	8
1.3.1. Principinė komponentų analizė (PCA)	9
1.3.2. Tiesinio diskriminanto klasifikacija (LDC)	11
1.3.3. Išankstinis požymių atrinkimas (FFS)	12
1.4. Darbe nagrinėjami biologiniai dariniai ir skirtumai tarp jų	13
2. Eksperimentinė dalis	14
2.1. Darbe naudojami spektrinių duomenų rinkiniai	17
2.2. Rezultatai	18
2.2.1. Kriterijaus funkcijos pasirinkimas FFS algoritmui.....	18
2.2.2. Gautos spektrinių žymenų histogramos ir jų interpretacijos.....	19
2.2.3. Spektrinių žymenų panašumo rezultatai	26
2.2.4. Spektrinių žymenų stabilumo rezultatai.....	30
2.3. Rezultatų aptarimas.....	32
Išvados	36
Padėka	36
Priedas.....	37
Literatūros sąrašas.....	38
Summary	40

Išvadas

Virpesinės spektroskopijos metodų pritaikymas nedestruktyviai biologinių bandinių analizei yra palyginti nauja ir besivystanti sritis. Pagrindiniai virpesinės spektroskopijos metodai yra Furjė transformacijos infraraudonosios sugerties (FT-IR) ir Ramano sklaidos spektroskopijos. Jų veikimo principas pagrįstas elektromagnetinės spinduliuotės sąveika su molekulių virpesiais. Infraraudonosios sugerties spektroskopijos atveju molekulės gali sugerti vidurinės infraraudonosios srities šviesą ir pereiti į aukštesnės energijos virpesinį lygmenį. Bangos ilgiai, ties kuriais molekulės sugeria infraraudonąją šviesą, priklauso nuo jų normaliųjų virpesių dažnio, o sugerties intensyvumas priklauso nuo molekulės elektrinio dipolio momento pokyčio vykstant atitinkamam virpesiui. Pagal tai iš infraraudonosios sugerties spektro galima išgauti informaciją apie tiriamų molekulių struktūrą ir cheminę sudėtį. FT-IR spektroskopija yra sėkmingai pritaikoma įvairiuose biochemijos tyrimuose, pavyzdžiui, farmacijoje [1], biologinių skysčių analizėje [2], augalų tyrimuose [3], ląstelių ir audinių cheminiame vaizdinime naudojant mikroskopus-spektrometrus [4], vėžio tyrimuose ir diagnostikoje [5], maisto kokybės kontrolėje [6] ir taip toliau.

Infraraudonosios sugerties spektroskopijos pagrindiniai privalumai, padarantys ją patrauklią biologijos ir medicinos tyrimuose, yra lengvas naudojimas, greitas spektrų registravimas, nedestruktyvumas ir nespecifiškumas. Vienas iš virpesinės spektroskopijos išskirtinių pranašumų yra tai, kad dauguma spektrų analizei naudojamų statistinių ir mašininio mokymosi metodų leidžia nustatyti, kurios spektrinės sritys konkrečiam tyrimui yra statistškai svarbiausios. Šios sritys yra susijusios su tiriamais biocheminiais vyksmais ar skirtumais tarp bandinių. Tai reiškia, kad naudojantis virpesine spektroskopija galima išgauti bandinių biocheminį „piršto antspaudą“, arba kitaip tariant, spektrinius žymenis (taip pat vadinami biožymenimis). Spektrinių žymenų nustatymas yra dažna biologinių tyrimų su virpesine spektroskopija dalis, įprastai atliekama kartu su bandinių spektrų klasifikacija. Biologinių bandinių FT-IR spektrų klasifikacija yra plačiai aptarta ir išnagrinėta tema, tačiau spektrinių žymenų identifikavimo metodų patikrinimui ir vystymui skiriama santykinai mažiau dėmesio [7]. Esminė problema su spektrinių žymenų identifikavimu virpesiniuose spektruose yra ta, kad, skirtingai nei su klasifikavimu, dažniausiai nėra iš anksto žinomų biožymenų, su kuriais būtų galima palyginti gautus rezultatus ir įvertinti pasirinkto metodo tikslumą. Spektrinių žymenų išgavimą taip pat apsunkina nepageidaujami fizikiniai reiškiniai - šviesos sklaida, bandinio nevienalytiškumas, koncentracijos variacijos ir kiti - kurie įveda nuokrypius registruojant spektrus. Šie nuokrypiai turi būti iškompensuojami pritaikant pirminio apdorojimo procedūras, tačiau pirminio apdorojimo įtaka spektrinių žymenų nustatymui dar nėra ištirta.

Šio darbo tikslas – nustatyti, kaip skirtingais statistiniais metodais gauti biologinių bandinių FT-IR spektriniai žymenys priklauso nuo spektrų pirminio apdorojimo. Šiame tyrime kaip konkretūs pavyzdžiai buvo naudojami du FT-IR spektrinių duomenų rinkiniai:

- patologinių bakterijų ir mielių;
- ekstraląstelių vezikulių, išgautų iš sveikų žmonių ir miokardo infarkto pacientų kraujo.

Tyrimo tikslui pasiekti buvo iškelti šie uždaviniai:

- turimus spektrų rinkinius apdoroti įvairiais skirtingais FT-IR spektrų pirminio apdorojimo metodais;
- iš spektrinių duomenų rinkinių su skirtingais apdorojimais išgauti spektrinius žymenis naudojantis išankstinio požymių atrinkimo (angl. *forward feature selection*) ir principinės komponentų analizės – tiesinio diskriminanto klasifikacijos (angl. *principal component analysis – linear discriminant classifier*) algoritmais;
- įvertinti gautų spektrinių žymenų pagrįstumą ir priklausomybę nuo spektrų pirminio apdorojimo remiantis žymenų stabilumo ir panašumo kriterijais, taip pat pagal spektrinę/biocheminę interpretaciją.

1. Teorinė dalis

1.1. Virpesinės spektroskopijos pritaikymas biologiniuose tyrimuose

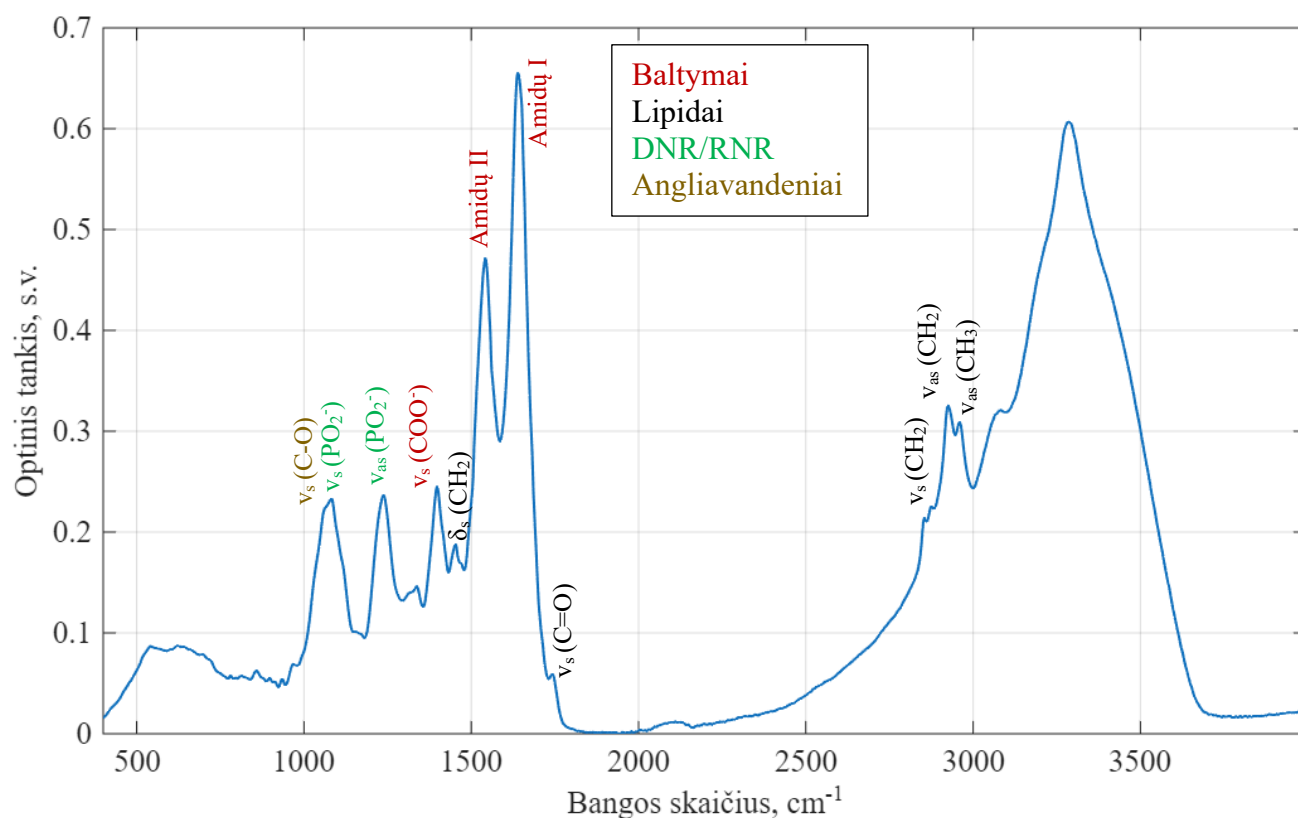
Pagrindinės biologinės sistemos sudarančios biomolekulės – baltymai, aminorūgštys, lipidai, nukleorūgštys, angliavandeniai – turi savo charakteringus virpesius, kuriuos galima stebėti virpesiniuose spektruose. Tipinis biologinio bandinio infraraudonosios sugerties spektras pavaizduotas 1 pav.

Biologinių darinių spektruose tipiškai aktualiausia yra $1700 - 900 \text{ cm}^{-1}$ sritis, vadinama „pirštų antspaudo“ sritimi. Šioje srityje yra dauguma biomolekulių virpesinių juostų, iš kurių svarbiausios ir intensyviausios yra:

- Amidų I ir II virpesinės juostos, atitinkančios baltymų peptidinės jungties virpesius, kurių dažnis priklauso nuo baltymų antrinės struktūros;
- $-\text{CH}_2$ ir $-\text{CH}_3$ grupių baltymuose ir lipiduose deformacinių virpesių juostos;
- Fosfato $-\text{PO}_2$ grupės valentinių simetrinių ir antisimetrinių virpesių juostos, priskiriamos DNR ir RNR fosfodiesteriniai jungčiai.

FT-IR spektroskopija turi didelį potencialą mikrobiologijoje [8], maisto apsaugoje [6], diagnostinėje medicinoje, taip pat biomolekulių struktūros, sąveikų ir dinamikos tyrimuose [9]. Ji gali būti pritaikyta

plačiam biologinių bandinių pasirinkimui, įskaitant diagnostinės medicinos mėginius – kraują, šlapimą, audinius. Pavyzdžiui, FT-IR spektroskopija buvo panaudota tirti žmogaus serumo albumino antrinės struktūros pokyčius įvairiomis denatūravimo sąlygomis [10]. Šio baltymo struktūros pokyčiai buvo nustatyti pagal amidų I spektrinę juostą jai atlikus spektrinę dekonvoliuciją, kuri išskiria kelias persiklojusias juostas. Tuomet pagal amidų I juostos vidinę sandarą galima nustatyti santykį tarp baltymo α -spiralių, β -klosčių ir atsitiktinių struktūrų kiekio.



1 pav. Bakterijos *Staphylococcus aureus* IR sugerties spektras. Spektre pažymėtos dažniausiai biologiniuose bandiniuose pasitaikančios spektrinės juostos ir kokioms junginių grupėms jos priklauso. v_s žymi simetrinius valentinius virpesius, v_{as} – antisimetrinius valentinius virpesius, δ – deformacinius virpesius.

Vienas iš perspektyviausių virpesinės spektroskopijos pritaikymų biologijoje yra cheminis vaizdinimas naudojant mikroskopus-spektrometrus. Ši sritis yra aktuali ląstelių ir audinių tyrimuose, histopatologijoje ir vėžio diagnostikoje, kuriuose virpesinės spektroskopijos vaizdinimas leistų aptikti biocheminės sandaros skirtumus, neaptinkamus įprastais audinių dažymo metodais [5][11].

1.2. Spektrų pirminio apdorojimo procedūros

Spektrų pirminis apdorojimas yra svarbiausias žingsnis biologinių virpesinių spektrų analizėje, kadangi jis daro esminę įtaką visų sekančių operacijų tikslumui. Pirminio apdorojimo tikslas yra kompensuoti įvairius nepageidaujamus veiksnus, sukeliančius spektrų iškraipymus ir geriau išskirti biocheminius skirtumus tarp bandinių. Pirminio apdorojimo procedūrų ir algoritmų pasirinkimą

diktuoja bandymo specifikacijos ir kiekvienam tyrimui optimaliausios procedūros dažniausiai būna skirtingos. Pirminis apdorojimas parenkamas vadovaujantis fizikiniais principais ir optimizuojant duomenų analizės tikslumą.

1.2.1. Bazinės linijos korekcija

Be elektromagnetinės spinduliuotės sąveikos su molekulių virpesiais, bandiniuose vyksta ir kiti šviesos reiškiniai, kurie infraraudonosios sugerties spektruose pasireiškia kaip kylanti arba osciliuojanti bazinė linija. Biologinių bandinių FT-IR spektruose dažniausia kylančios bazinės linijos priežastis yra žadinančios spinduliuotės rezonansinė Mie sklaida [11] (sklaida, kai krentančios šviesos bangos ilgis yra palyginamas su sklaidančios dalelės dydžiu). Sklaida dažnai būna mažesnė FT-IR spektrus registruojant naudojantis pažeistojo vidaus atspindžio (angl. *attenuated total reflection*, ATR) spektrometrais. Bazinės linijos kilimas dažniausiai taisomas aproksimuojant ir atimant bazinę liniją, arba modeliuojant kilimą sukeliančius reiškinius [12].

Bazinės linijos korekcijai (angl. *baseline correction*) naudojami šie metodai:

1. „*Rubberband*“, arba „guminės juostos“, korekcija. Šis algoritmas randa liniją, jungiančią taškus, atitinkančius spektro minimumus. Taip pat naudojamas rankinis šio metodo variantas, kuomet linijos taškai parenkami vartotojo nuožiūra.
2. Bazinės linijos aproksimavimas tiesine arba kvadratine priklausomybe. Šis metodas skiriasi nuo „*rubberband*“ tuo, kad vietoj kreivos ir kampuotos linijos, jungiančios spektro minimumų taškus, pagal tuos taškus atliekama funkcijos aproksimacija.
3. Diferencijavimas. Atliekant pirmos eilės diferencijavimą teoriškai panaikinama tiesinė spektro komponentė. Šis metodas pakeičia spektro kontūrą, bet geriau išskiria spektrines juostas. Taip pat naudojamas ir antros eilės diferencijavimas, kuris, kitaip nei pirmos eilės, duoda simetrines spektrines juostas. Pagrindinis diferencijavimo kaip spektrinio apdorojimo metodo trūkumas yra tai, kad kartu su juostų išskyrimu jis taip pat sustiprina spektro triukšmą. Dėl šios priežasties praktikoje spektrų diferencijavimui dažnai naudojamas *Savitzky-Golay* algoritmas su triukšmo išlyginimu.
4. „*Extended multiplicative signal correction*“, arba sutrumpintai EMSC, metodas. Tai yra sudėtingesnis infraraudonosios sugerties spektrų apdorojimo metodas, kuris atlieka ir bazinės linijos korekcijos, ir normalizavimo funkcijas. Jis yra paremtas Lamberto – Bugerio – Bero dėsnio modeliavimu. EMSC remiasi prielaida, kad užregistruotą infraraudonosios sugerties spektrą $A_{užr}(ν)$ galima išskleisti kokio nors atraminio spektro $A_{atr}(ν)$ atžvilgiu [13]:

$$A_{užr}(ν) = a + b \cdot A_{atr}(ν) + c \cdot ν + d \cdot ν^2 + ε(ν), \quad (1)$$

čia a, b, c, d – skleidimo parametrai, o $\varepsilon(v)$ – liekana, atitinkanti nemodeliuotus skirtumus tarp $A_{užr}(v)$ ir $A_{atr}(v)$. Laikoma, kad $A_{atr}(v)$ yra „idealus“ spektras, neturintis artefaktų, o pirmas, trečias ir ketvirtas lygties nariai atitinka spektro bazinę liniją dėl sklaidos. Apdorojimas EMSC metodu veikia taip: pirma parenkamas atraminis spektras, dažniausiai kaip spektrinių duomenų rinkinio vidutinis spektras. Tuomet visiems rinkinio spektrams pritaikoma (1) aproksimacija. Galiausiai naudojantis gautais koeficientais iš $A_{užr}(v)$ gaunamas apdorotas, normalizuotas spektras su kompensuota bazine linija $A_{apd}(v)$:

$$A_{apd}(v) = \frac{A_{užr}(v) - a - c \cdot v - d \cdot v^2}{b}. \quad (2)$$

1.2.2. Spektrų normavimas

Pagal Lamberto – Bero – Bugo dėsnį, bandinio pralaidumas šviesai priklauso nuo jo koncentracijos ir storio. Praktikoje atliekant spektroskopinius matavimus išlaikyti bandinių storį, koncentraciją bei šviesos kelią juose tokius pačius yra sudėtinga, ypač dirbant su biologiniais bandiniais. Dėl to registruojami virpesiniai spektrai dažnai varijuoja savo intensyvumais, ir šios variacijos gali paslėpti realius biocheminius skirtumus tarp bandinių. Siekiant suvienodinti spektrų intensyvumus, jie yra normalizuojami pagal įvairius parametrus.

FT-IR spektrų normavimui naudojami šie metodai:

1. Vektorinis normavimas. Šis metodas padalina spektrą iš jo euklidinės normos:

$$\text{Euklidinė norma} = \sqrt{x_1^2 + x_2^2 + \dots + x_N^2}, \quad (3)$$

čia $x_1, x_2, \dots, x_N$ – spektro optinio tankio vertės duotam bangos skaičiui, kai spektre iš viso yra N taškų. Vektoriniam normavimui galima suteikti fizikinę interpretaciją, kad spektrai transformuojami taip, jog jie tartum būtų gauti iš vienodo storio bandinių.

2. *Min – max* normavimas. Šis metodas pirma pastumia spektrą taip, kad jo optinio tankio minimumas būtų ties 0 (kitais tarant, atima spektro intensyvumą iš jo minimumo vertės) ir tuomet padalina skirtumą iš skirtumo tarp spektro maksimumo ir minimumo verčių x_{max} ir x_{min} :

$$x_{i \text{ norm.}} = \frac{x_i - x_{min}}{x_{max} - x_{min}}, \quad (4)$$

čia $x_{i \text{ norm.}}$ – normuoto spektro i -tasis elementas, o x_i – nenormalizuoto spektro i -tasis elementas, $i=1, 2, \dots, N$.

3. Normavimas pagal vieną juostą. Šis metodas padalina spektrus pagal jų tam tikros spektrinės juostos integrinį intensyvumą. Dažniausiai tam parenkama amidų I arba II juosta, kadangi šios juostos yra intensyvios ir stebimos didžiojoje daugumoje biologinių bandinių. Tokiam normavimui galima suteikti fizikinę interpretaciją, kad jie būtų gauti iš bandinių su vienodu kiekiu

baltymų/aminorūgščių (arba kitos medžiagos, kurios spektrinės juostos buvo pasirinktos). Kita vertus, šis metodas netinka, jei šios (ar kitos pasirinktos) juostos reikšmingai skiriasi tarp bandinių.

4. Normavimas pagal spektro plotą. Šis metodas padalina spektrą iš jo „ploto“, apskaičiuojamo kaip:

$$Plotas = x_1 + x_2 + \dots + x_N, \quad (5)$$

čia $x_1, x_2, \dots, x_N$ – spektro optinio tankio vertės duotam bangos skaičiui, kai spektre iš viso yra N taškų. Ploto normavimas veikia labai panašiai, kaip ir vektorinis normavimas.

1.2.3. Pašalinių medžiagų kompensavimas

Vanduo turi intensyvią ir plačią infraraudonosios sugerties spektrinę juostą ties $\sim 3400 \text{ cm}^{-1}$, susijusią su O-H valentiniais simetriniais ir antisimetriniais virpesiais, taip pat plačią $\sim 1640 \text{ cm}^{-1}$ deformacinių virpesių juostą. Šios juostos veikia kaip papildomas fonas šalia esančioms aktualioms virpesinėms juostoms, pvz. $-\text{CH}_2$ ir $-\text{CH}_3$ valentinių bei amidų I ir II virpesių juostoms. Beveik visi biologiniai bandiniai arba turi savyje reikšmingus kiekius vandens, arba yra laikomi vandens terpėje. Atmosferinės dujos – vandens garai, anglies dioksidas – taip pat įneša papildomas juostas. Atmosferinių dujų virpesinės juostos dažniausiai kompensuojamos pagal bandymo pradžioje užregistruotą atraminį spektrą. Spektrinius priedus taip pat gali įnešti kai kurių biologinių bandinių paruošimui ir išlaikymui naudojamos medžiagos, pvz. parafinas. Pašalinių medžiagų spektrinės juostos gali būti kompensuojamos kai kuriais sudėtingesniais apdorojimo metodais (pvz. ESMC), tačiau dažniausiai jos tiesiog iškerpamos iš spektrų.

1.3. Spektrinių žymenų išgavimo metodai

Kartu su panaudojimais spektrų analizėje bei klasifikacijos ir diagnostikos uždaviniuose, statistiniai metodai biologinių bandinių virpesinėje spektroskopijoje taip pat leidžia nustatyti, kurios spektrinės sritys yra svarbiausios statistiškai. Tos sritys yra susijusios su biomolekulėmis ir cheminiais ryšiais, kurie indikuoja biocheminius pokyčius tarp bandinių grupių (pavyzdžiui, cheminis paveikimas, ligos tipas ir taip toliau). Statistinių metodų pritaikymas eksperimentinių rezultatų susiejimui su skirtumais tarp bandinių yra plati bioinformatikos tema, neapribota vien tik virpesine spektroskopija. Biožymenų nustatymas taip pat atliekamas genų/baltymų sekų analizėje, DNR mikrogardelių analizėje, masių spektrometrijoje ir kitur.

Biožymenų išgavimas remiasi prielaida, kad konkrečiame bandyme yra stebimi nuo tiriamų biologinių mechanizmų priklausantys pokyčiai, kurie palieka savo žymenys eksperimentiniuose duomenų rinkiniuose, šiuo atveju – FT-IR spektruose. Pagrindiniai metodai spektrinių žymenų išgavimui iš virpesinių spektrų yra šie [7][12]:

1. Statistiniai hipotezės testai, pvz. Krusalo ir Valiso (*Kruskal-Wallis*) testas, Mano ir Vitnio (*Mann-Whitney's*) testas ir kiti. Šiais metodais kiekvienam spektrų rinkinio bangos skaičiui gaunama p vertė, pagal kurią įvertinamos svarbiausios spektrinės sritys ir viršūnės.
2. Tiesiniai statistiniai ir klasifikacijos metodai, kurie kiekvienam bangos skaičiui priskiria tam tikrus svorius, pavyzdžiui, principinių komponentų analizė (PCA), tiesinio diskriminanto klasifikacija (LDC), mažiausiųjų kvadratų metodai ir kiti;
3. Požymių atrinkimo metodai, pavyzdžiui, išankstinis požymių atrinkimas (angl. *forward feature selection*, sutrumpintai FFS), genetinio algoritmo (angl. *genetic algorithm*) metodai ir kiti. Jie iš visų spektrų rinkinio parenka tik kelis bangos skaičių kintamuosius, geriausiai atskleidžiančius skirtumus rinkinyje pagal tam tikrą kriterijų.

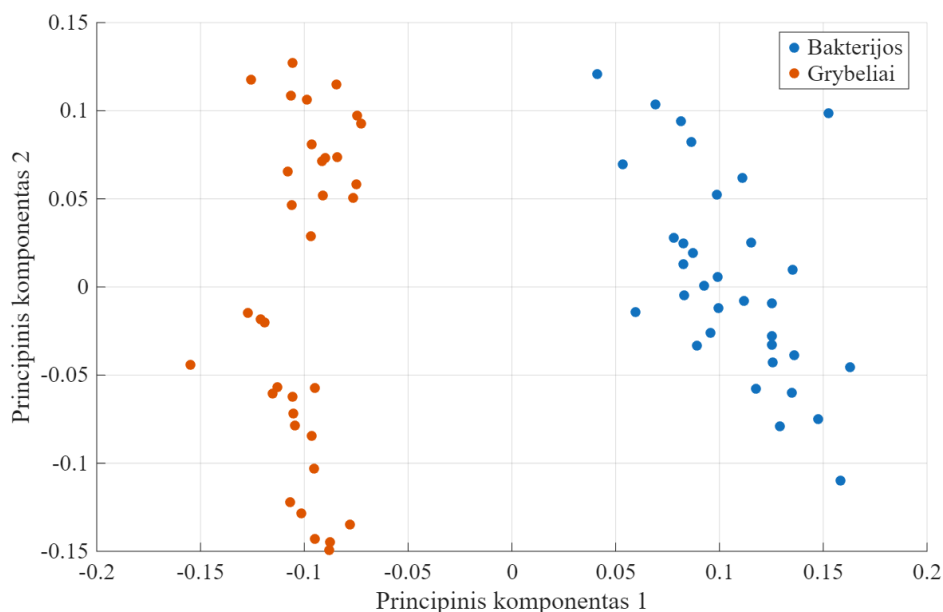
Minėti metodai turi savo privalumų ir trūkumų. Problema su hipotezės testais yra ta, kad juos atliekant kiekvienam bangos skaičiui individualiai neįvertinama, kaip spektrinės sritys gali būti svarbios kolektyviai. Be to, p vertės mažėja imčiai didėjant, taigi didelėse imtyse mažos p vertės nebūtinai koreliuoja su bangos skaičiaus aktualumu. Niuansas su tiesiniais metodais yra tas, kad gautiems svorių vektoriams reikšmingą įtaką turi ne tik skirtumai tarp spektrų grupių, bet ir variacijos grupių viduje [7]. Vienas iš tiesinių metodų privalumų yra tas, kad jų svorių kryptis (tai yra, ar svorio vertė teigiama, ar neigiama) indikuoja, kaip spektro intensyvumas tam tikrame bangos skaičiuje koreliuoja su spektro priklausymui vienai ar kitai grupei. Tuo tarpu kiti metodai gali tik pasakyti, ar bangos skaičius yra svarbus bendrai. Pagrindinis požymių atrinkimo metodų trūkumas yra ilgas skaičiavimo laikas, kadangi jie reikalauja apskaičiuoti tikslumo kriterijus dideliame skaičiui kintamųjų derinių. Iš čia minėtų metodų kaip patikimiausi biožymenų nustatymui literatūroje yra laikomi požymių atrinkimo metodai [7][12][14][15].

1.3.1. Principinė komponentų analizė (PCA)

Principinė komponentų analizė (angl. *principal component analysis*, PCA) yra vienas populiariausių virpesinių spektrų apdorojimo ir analizės metodų. Tai yra tiesinis dimensijos sumažinimo metodas, kuris N – kiekį kintamųjų (spektroskopijos atveju – sugerties intensyvumų ties tam tikrais bangos skaičiais) sutraukia į kelis naujus kintamuosius, vadinamus principiniais komponentais. Matematiškai principinė komponentų analizė atitinka spektrinio duomenų rinkinio matricos $\mathbf{X}$, kurios dimensija yra $n \times p$, kur n – rinkinyje esančių spektrų skaičius, o p – viename spektre esančių kintamųjų (sugerties intensyvumų) skaičius, bazės pakeitimą:

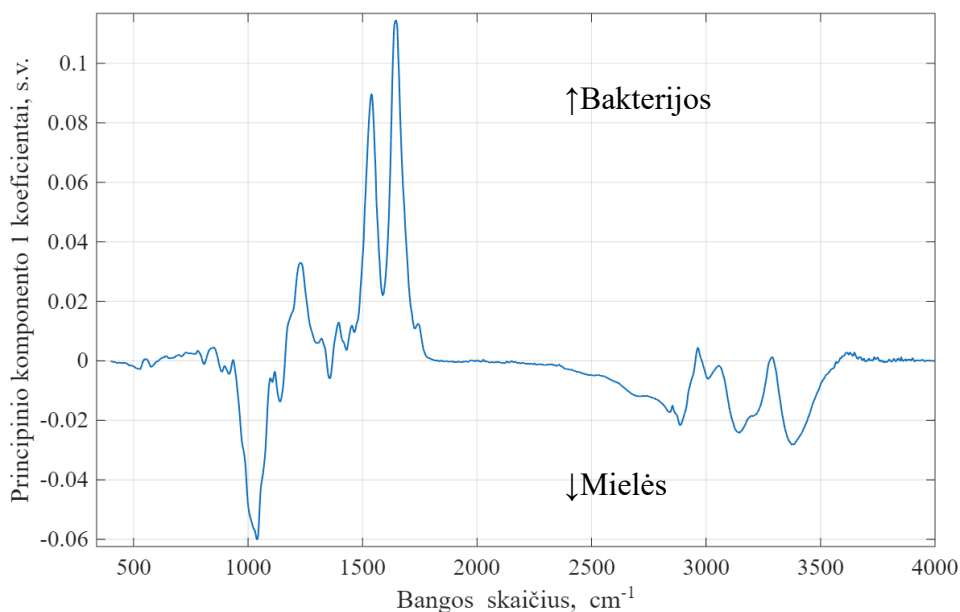
$$\mathbf{XP} = \mathbf{Y}, \quad (6)$$

čia $\mathbf{P} - p \times k$ transformacijos matrica, kur k – pasirinktas principinių komponentų skaičius, o $\mathbf{Y} - n \times k$ gauta principinių komponentų matrica. Transformacijos matrica $\mathbf{P}$ parenkama tokia, kad gauti komponentai būtų ortogonalūs ir kuo geriau aprašytų duomenų variaciją. FT-IR spektrų analizėje PCA dažnai naudojama aptikti esminius dėsningumus, kiek ir kaip skiriasi įvairios bandinių grupės. Šie skirtumai dažnai lengvai pastebimi grafike atvaizduojant duomenų rinkinį pagal pirmus du (arba tris) principinius komponentus, kadangi jie apima didžiąją dalį duomenų variacijos (2 pav. rodomas pavyzdys).



2 pav. Patologinių mikroorganizmų FT-IR spektrų rinkinio pirmų dviejų PCA principinių komponentų pasiskirstymas pagal jų tipą. Prieš PCA spektrai buvo apdoroti bazinės linijos korekcija ir vektoriniu normavimu.

Jeigu tarp gautų principinių komponentų spektrinių duomenų grupės gerai atsiskiria, principinių komponentų koeficientai, gauti iš transformacijos matricos, gali būti panaudojami spektrinių žymenų nustatymui. Žymenų nustatymo pavyzdys pateiktas 2 ir 3 pav. 2 pav. matomas šiame darbe naudojamo patologinių mikroorganizmų FT-IR spektrų rinkinio pirmųjų dviejų principinių komponentų pasiskirstymas. Kadangi mikroorganizmų tipai (mielės/bakterijos) puikiai atsiskiria pagal pirmąjį principinį komponentą, galima daryti prielaidą, jog jo koeficientai gerai atskleidžia spektrinių duomenų variaciją tarp rinkinio grupių. Svarbu paminėti, kad PCA yra „neprižiūrimas“ metodas ir tiesiogiai į spektrų priklausymą kokiai nors grupei neatsižvelgia. Tai reiškia, jog PCA netinka žymenų nustatymui, jei nėra gero atsiskyrimo tarp duomenų grupių. PCA taip pat naudojamas kaip apdorojimo žingsnis prieš kitus duomenų analizės metodus siekiant sumažinti skaičiavimams reikalingą kintamųjų kiekį ir išfiltruoti triukšmus bei neaktualią informaciją spektruose.



3 pav. Patologinių mikroorganizmų FT-IR spektrų rinkinio pirmo PCA principinio komponento koeficientai.

Grafike pažymėta, kurios koeficientų kryptys koreliuoja su kokiais mikroorganizmų tipais, atsižvelgiant į 2 pav. Prieš PCA spektrai buvo apdoroti bazinės linijos korekcija ir vektoriniu normavimu.

Principinei komponentų analizei yra aktualus mašininio mokymosi pirminio apdorojimo žingsnis, vadinamas standartizavimu (angl. *standardization*, taip pat vadinamas *feature scaling* arba *Z-score normalization*). Jis normalizuoja kiekvieno duomenų rinkinio kintamojo (spektroskopijos atveju – spektrų intensyvumus ties tam tikrais bangos skaičiais) vidurkį į 0 ir dispersiją į 1. Standartizavimas naudojamas dėl to, kad kai kurie mašininio mokymosi modeliai pervertina kintamuosius, kurie varijuoja didesniais intervalais. Pavyzdžiui, kintamasis, kurio vertės varijuoja intervale nuo 1 iki 100, būtų labiau vertinamas nei kintamasis, varijuojantis intervale 1 – 10, nepaisant to, kad pastarajame galbūt geriau atsiskiria duomenų klasės. Principinei komponentų analizei standartizavimas dažnai pagerina duomenų grupių atsiskyrimą ir padidina principinių komponentų skaičių, reikalingą pasiekti aukštą paaiškindantą variaciją (kitais tariant, padaro principinius komponentus informatyvesnius). Kitiems statistiniams/mašininio mokymosi metodams standartizavimo poveikis nevienodas, kai kuriems standartizavimas yra būtinas geram tikslumui, kitiems jis nereikalingas.

1.3.2. Tiesinio diskriminanto klasifikacija (LDC)

Tiesinio diskriminanto klasifikacija (angl. *linear discriminant classifier*, sutrumpintai LDC, taip pat vadinama Fišerio (*Fischer*) diskriminanto klasifikacija) yra statistikos metodas, kuris randa kintamųjų tiesinę kombinaciją, leidžiančią geriausiai atskirti dvi arba daugiau duomenų grupių. Kitaip tariant, LDC suranda naują ašį kintamųjų N -matėje erdvėje, kurioje skirtingų grupių duomenų projekcijos geriausiai atsiskiria. Tuomet objektas gali būti klasifikuotas pagal tai, kurioje ašies pusėje jis yra. Dvejų klasių atveju LDC matematiškai aprašoma taip [16]: duotas vienas išmatavimas - vektorius $x =$

$(x_1, x_2, \dots, x_n)$, čia $x_1, x_2, \dots, x_n$ – bandymo metu išmatuoti požymiai/savybės (spektroskopijos atveju – sugertis ties tam tikrais bangos skaičiais). Tiesinio diskriminanto klasifikacija remiasi prielaidomis, kad kiekvienos klasės išmatavimai yra pasiskirstę normaliuoju skirstiniu ir kad visos klasės turi vienodą kovariacijos matricą. Taip pat žinomi abiejų klasių išmatavimų vektorių vidurkiai $\bar{X}^{(1)}$ ir $\bar{X}^{(2)}$ bei abiejų klasių kovariacijos matrica Σ . Tuomet išmatavimas x priskiriamas klasei pagal funkciją $g(x)$:

$$g(x) = x^T \beta + \omega_0, \quad \beta = \Sigma^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}), \quad (7)$$

čia ω_0 – funkcijos slenksčio konstanta. Jei $g(x) > 0$, išmatavimas priskiriamas pirmai klasei, jei ne – antrai klasei.

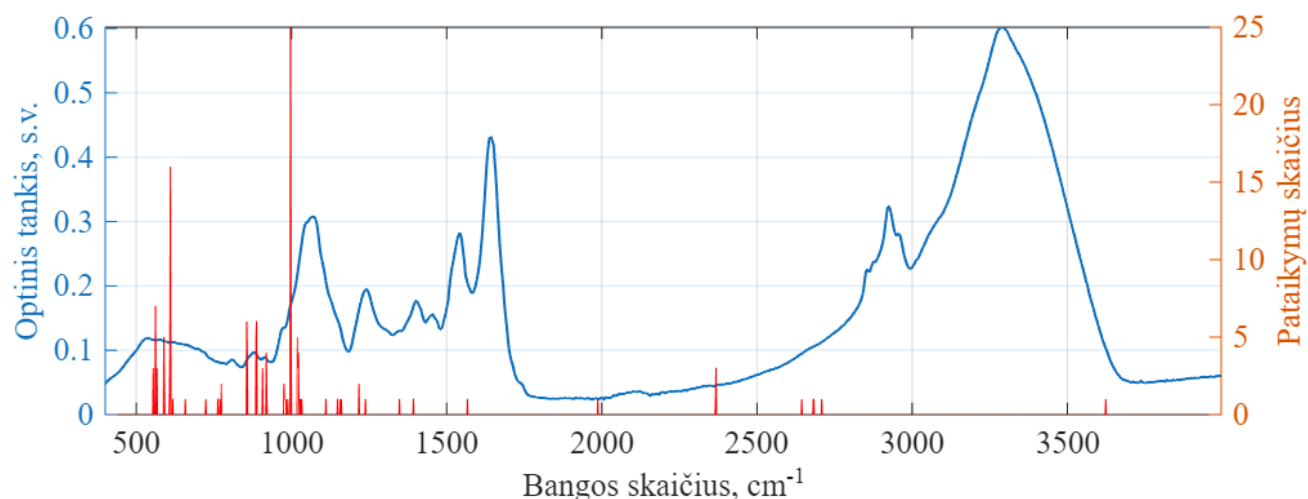
LDC gali būti naudojamas spektrinių žymenų išgavimui iš FT-IR spektrų dėka jo parametrų vektoriaus β , kuris, panašiai kaip PCA, parodo, kaip spektrų intensyvumas ties tam tikrais bangos skaičiais koreliuoja su spektro priklausymu tam tikrai klasei. Be klasifikavimo ir žymenų išgavimo, šis metodas taip pat kartais naudojamas kaip dimensijos sumažinimo metodas (tuomet jis vadinamas tiesinio diskriminanto analize), panašiai, kaip principinė komponentų analizė. LDC nuo PCA skiriasi tuo, kad LDC yra prižiūrimas metodas, kuris bando atskirti duomenis tiesiogiai pagal jų priklausymą kokiai nors grupei, taigi spektrinių žymenų išgavimui yra patikimesnis ir universalesnis. Praktikoje LDC dažnai naudojamas kartu su PCA, kadangi LDC tipiškai geriau veikia su mažesniu skaičiumi kintamųjų [12]. Spektrinių duomenų analizėje LDC yra vienas paprasčiausių ir dažniausiai naudojamų klasifikavimo modelių. Jis yra greitai apmokomas, lengvai interpretuojamas ir įprastai nepersimoko, tačiau taip pat gali būti nepakankamai sudėtingas tiksliai nustatyti ribas tarp grupių. Taip pat yra ir kvadratinio arba kubinio diskriminanto klasifikacijos metodai, kurie gali nustatyti sudėtingesnes ribas.

1.3.3. Išankstinis požymių atrinkimas (FFS)

Išankstinio požymių atrinkimo (toliau – FFS) algoritmas suranda mažą kiekį kintamųjų (bangos skaičių) spektrų rinkinyje, su kuriais pasirinktas klasifikacijos modelis (šio darbo atveju – tiesinio diskriminanto klasifikacija) pasiekia geriausią tam tikro kriterijaus, susijusio su klasifikacijos tikslumu, vertę. Algoritmas tai atlieka pradėdamas nuo tuščios kintamųjų imties. Tuomet jis apmoko klasifikacijos modelį naudojant vieną rinkinio kintamąjį ir apskaičiuoja apmokyto modelio kriterijaus vertę. Šis procesas pakartojamas kiekvienam kintamajam. Kintamasis, su kuriuo gauta kriterijaus vertė yra optimaliausia, yra pridedamas į imtį. Apmokymo procesas pakartojamas šįkart ieškant optimaliausio kintamojo imtyje kartu su prieš tai pasirinktu kintamuoju. Taip kintamieji yra vienas po kito pridedami į imtį iki kol yra pasiekiamas iš anksto nustatytas kintamųjų kiekis arba kai kriterijaus vertė konverguoja. FFS algoritmo pasirinkti spektriniai kintamieji ir bus ieškomi spektriniai žymenys,

kadangi jie nurodo, kuriose spektrinėse srityse yra geriausias atsiskyrimas tarp duomenų rinkinio grupių.

Mokslinėje literatūroje yra aptarta, kad naudojant požymių atrinkimo metodus žymenų identifikavimui, išrinkimo procesas turėtų būti kartojamas kelis kartus, kas kartą nežymiai pakeičiant duomenų imtį [14][17][18]. Taip daroma dėl to, jog požymių atrinkimo metodai dažnai būna nestabilūs, tai yra, nedideli pokyčiai duomenų imtyje gali sukelti didelius pokyčius išrinkimo rezultate. Kartojant požymių atrinkimo procesą keliolika kartų, iš tikrųjų svarbiausi kintamieji bus parinkti daugiau nei kiti, ir taip sumažinama rizika išgauti netinkamus žymenis. Kartojamo požymių atrinkimo rezultatai dažnai vizualizuojami kintamųjų histograma, kur x ašyje yra duomenų rinkinio kintamieji (virpesinės spektroskopijos atveju – spektrų bangos skaičiai), o y ašyje – kartų, kiek buvo pasirinktas atitinkamas kintamasis, skaičius (pavyzdys parodytas 4 pav.). Kuo daugiau kartų FFS algoritmas pataikė į tą patį bangos skaičių, tuo jis yra santykinai reikšmingesnis.



4 pav. Kartojamo FFS kintamųjų histograma (raudona spalva) mikroorganizmų FT-IR spektrų rinkiniui. Dėl aiškumo kartu įdėtas atsitiktinis rinkinio spektras (mėlyna spalva).

1.4. Darbe nagrinėjami biologiniai dariniai ir skirtumai tarp jų

Šiame darbe kaip biologinių darinių pavyzdžiai naudojami ekstraląstelinės vezikulės ir patologiniai mikroorganizmai – bakterijos ir mielės.

Bakterijos yra prokariotiniai (t.y. neturintys ląstelių branduolio) vienaląsčiai organizmai. Bakterijų ląstelės yra paprastesnės vidinės struktūros palyginus su eukarotinėmis ląstelėmis, neturi branduolio ar kitų organelių (su keliomis išimtimis). Bakterijos išsiskiria savo išorinėmis struktūromis – ląstelės membrana ir sienele. Bakterijų sienelės yra sudarytos iš peptidoglikano, būdingo tik bakterijoms. Pagal ląstelės sienelės sandarą bakterijos skirstomos į dvi grupes: gram - teigiamas ir gram – neigiamas

(pagal Gramo dažymo testą). Gram – teigiamos bakterijos pasižymi storesne sienele, o gram – neigiamos turi plonesnę sienelę, bet papildomą membraną.

Mielės yra grybeliams priskiriami vienaląsčiai mikroorganizmai. Grybeliai yra eukariotinių organizmų grupė, kuriai priskiriami grybai, mielės ir pelėsiai. Jie, kaip ir kiti eukariotai, turi branduolį ir organeles, tokias kaip mitochondrijos, ribosomos ir vakuolės. Grybelių sienelės yra sudarytos iš chitino ir glukanų komplekso. Grybeliai kaupia energiją glikogeno forma (panašai, kaip gyvūnai). Kai kurios mielės yra patogeniškos ir sukelia užkrečiamas ligas. Palyginus su bakterijomis, mielių sausą masę grubiai vidutiniškai sudaro didesnis santykinis kiekis angliavandenių, bet mažesnis kiekis DNR/RNR ir baltymų [19].

1 lentelė. Apytikslė bakterijų ir mielių biocheminė sudėtis pagal sausos masės procentinį svorį [19].

	Bakterijos	Mielės
DNR, %	2-4	1-3
RNR, %	5-15	3-10
Baltymai, %	40-60	40-50
Lipidai, %	10-15	5-20
Angliavandeniai, %	10-20	10-25

Vezikulės yra gyvose ląstelėse susiformuojančios nanostruktūros, sudarytos iš skysčio ar citoplazmos, uždarytos lipidų sluoksnyje tartum lašeliai ir savyje nešančios įvairias biomedžiagas. Ekstraląstelinės vezikulės - vezikulės, išsilaisvinusios iš jas pagaminusių ląstelių. Ląstelėse ir organizmuose ekstraląstelinės vezikulės atlieka homeostazės palaikymo funkciją pašalindamos toksiškas ar perteklines medžiagas iš ląstelių, ir bioaktyvių medžiagų perkėlimo tarp ląstelių funkciją.

2. Eksperimentinė dalis

Siekiant nustatyti, kokią įtaką spektrinis pirminis apdorojimas turi išgautiems spektriniams žymenims ir jų tikslumui, šiame tyrime buvo išbandytos įvairios pirminio apdorojimo operacijų kombinacijos. Šiame darbe naudojamų FT-IR spektrinių rinkinių spektrinis diapazonas yra 4000 - 400 cm^{-1} . Biologiniuose bandiniuose aktualiausia spektrinė sritis yra „pirštų antspaudo“ sritis ties maždaug 1700 – 600 cm^{-1} . Šiame darbe buvo bandomi apdorojimai tiek su pilnais spektrais ir su spektrais, apkarpytais iki pirštų antspaudo srities. Taip buvo bandoma įvertinti gaunamų spektrinių žymenų jautrumą spektrinės informacijos kiekiui. Iš viso buvo išbandytos šios apdorojimo procedūros ir jų deriniai (skliausteliuose nurodyti toliau darbe naudojami procedūrų trumpiniai):

1. jokio apdorojimo;
2. tiesinė bazinės linijos korekcija (t. b. l. k.);

3. antros eilės polinominė bazinės linijos korekcija (p. b. l. k.);
4. „*rubberband*“ bazinės linijos korekcija (r. b. l. k.);
5. tiesinė bazinės linijos korekcija + vektorinis normavimas (t. b. l. k. + vekt. n.);
6. „*rubberband*“ bazinės linijos korekcija + vektorinis normavimas (r. b. l. k. + vekt. n.);
7. pirma išvestinė (1 išv.);
8. antra išvestinė (2 išv.);
9. tiesinė bazinės linijos korekcija + vektorinis normavimas + standartizavimas (t. b. l. k. + vekt. n. + st.);
10. antra išvestinė + standartizavimas (2 išv. + st.);
11. antra išvestinė + vektorinis normavimas (2 išv. + vekt. n.);
12. „*extended multiplicative signal correction*“ (EMSC);

Toliau išvardinti apdorojimai su spektrų apkarpymu iki $1700 - 600 \text{ cm}^{-1}$ sities:

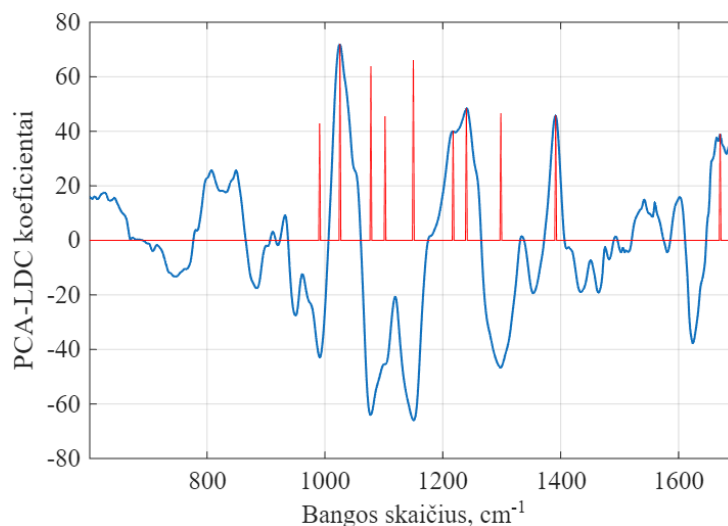
13. apkarpymas iki $1700 - 600 \text{ cm}^{-1}$ sities (apk.);
14. apkarpymas + tiesinė bazinės linijos korekcija (apk. + t. b. l. k.);
15. apkarpymas + tiesinė bazinės linijos korekcija + vektorinis normavimas (apk. + t. b. l. k. + vekt. n.);
16. apkarpymas + tiesinė bazinės linijos korekcija + vektorinis normavimas + standartizavimas (apk. + t. b. l. k. + vekt. n. + st.);
17. apkarpymas + EMSC (apk. + EMSC)
18. apkarpymas + EMSC + standartizavimas (apk. + EMSC + st.)
19. apkarpymas + pirma išvestinė (apk. + 1 išv.)
20. apkarpymas + antra išvestinė (apk. + 2 išv.)
21. apkarpymas + antra išvestinė + vektorinis normavimas (apk. + 2 išv. + vekt. n.)

Su kiekvienu iš čia išvardintu apdorojimu buvo išgauti spektriniai žymenys naudojantis dvejais metodais: išankstiniu požymių atrinkimu (angl. *forward feature selection*, FFS) su tiesinio diskriminanto klasifikacija ir principine komponentų analize – tiesinio diskriminanto klasifikacija (angl. *principal component analysis – linear discriminant classifier*, PCA-LDC). Abu metodai yra paremti spektrų klasifikacijos modeliais. Spektrų apdorojimas ir žymenų išgavimo skaičiavimai buvo atlikti naudojantis MATLAB programine aplinka (versija R2025b), MATLAB statistikos ir mašininio mokymosi bei lygiagretaus skaičiavimo paketais „*Statistics and Machine Learning Toolbox*“ bei „*Parallel Computing Toolbox*“, taip pat viešai prieinamu virpesinės spektroskopijos duomenų analizės MATLAB paketu „*IRootLab*“ [20].

Šiame darbe FFS algoritmui naudotas klasifikacijos modelis yra tiesinio diskriminanto klasifikacija be reguliarizacijos. Kriterijaus funkcijos įvertinimas FFS algoritmo metu buvo atliekamas su penkeriopa kryžmine validacija (angl. *5-fold cross validation*). Tai reiškia, kad prieš klasifikacijos modelio apmokymą spektrinių duomenų rinkinys buvo padalinamas į penkias dalis. Tuomet modelio apmokymas buvo kartojamas penkis kartus, kaskart vieną dalį paliekant tik modelio testavimui ir kriterijaus funkcijos skaičiavimui. Tuomet galutinė FFS kriterijaus funkcijos vertė yra visais penkiais atvejais suskaičiuotų kriterijaus funkcijos verčių vidurkis.

Kiekvienai šiame darbe nagrinėtai apdorojimų sekai spektrinių žymenų išgavimas FFS ir PCA-LDC algoritmais buvo kartojamas 30 kartų, kas kartą naudojant tik 90 % atsitiktinai parinktą spektrinių duomenų rinkinio dalį. Taip buvo daroma siekiant išmatuoti žymenų išgavimo metodų stabilumą ir jo priklausomybę nuo spektrų pirminio apdorojimo. Iš visų 30 FFS pakartojimų rezultatų buvo sudarytos kintamųjų histogramos kiekvienai pirminio apdorojimo sekai, atitinkančios duomenų rinkinio spektrinius žymenis. Pastebėtina, kad moksliniuose tyrimuose FFS skaičiavimai dėl patikimumo įprastai kartojami 100-1000 kartų, tačiau dėl itin ilgo FFS algoritmo skaičiavimo laiko šiame tyrime buvo apsiribota tik su 30 pakartojimų. Apskaičiuoti FFS žymenis vienam spektrinių duomenų rinkiniui, visiems 21 apdorojimo atvejų ir su 30 pakartojimų naudojantis MATLAB lygiagretaus skaičiavimo paketu „*Parallel Computing Toolbox*“ su 6 procesoriaus branduoliais užtruko apytiksliai 10 val.

PCA-LDC rezultatai buvo paversti į kintamųjų histogramas tam, kad jie galėtų būti palyginami su FFS rezultatais. Apmokytų PCA-LDC modelių koeficientai buvo paversti į histogramas laikant, kad koeficientų spektrų viršūnės atitinka FFS kintamojo pasirinkimą, kur viršūnės padėtis atitinka bangos skaičiaus kintamąjį, o viršūnės intensyvumas – požymio atrinkimo „dažnį“ (kuo PCA-LDC koeficiento vertės modulis didesnis, tuo atitinkamas bangos skaičius yra svarbesnis analogiškai, kaip kartojamo FFS dažniau pasirinktas požymis yra svarbesnis). Tokio pakeitimo pavyzdys pateiktas 5 pav. Siekiant išvengti nereikalingo triukšmo, į histogramas buvo paverčiamos tik pirmos 10 absoliutine verte intensyviausių viršūnių. Reikia paminėti, kad toks PCA-LDC koeficientų pavertimas į kintamųjų histogramą panaikina informaciją apie koeficientų kryptį (tai yra, ar koeficientas teigiamas ar neigiamas), iš kurios galima nusakyti, kaip atitinkamos spektrinės juostos intensyvumas koreliuoja su bandinio priklausymu konkrečiai klasei.



5 pav. PCA-LDC koeficientų pavertimo į kintamųjų histogramą pavyzdys. Mėlyna spalva - PCA-LDC koeficientai, raudona – iš jų gauta histograma.

Šiame darbe spektrinių žymenų išgavimas ir analizė buvo atliekami remiantis J. Trevisan *et al.* moksliniu straipsniu „*Measuring similarity and improving stability in biomarker identification methods applied to Fourier-transform infrared (FTIR) spectroscopy*“ [7]. Šiame straipsnyje tyrėjai aptarė žymenų iš FT-IR spektrų išgavimo tikslumo patikrinimo metodus. Kaip kriterijus patikrinimui jie pasiūlė žymenų, išgautų skirtingais metodais, panašumą ir stabilumą lyginant žymenis, išgautus tuo pačiu metodu, bet su nedideliais pakeitimais spektrų rinkiniuose. Kartu su minėtais kriterijais šiame tyrime žymenų tikslumas taip pat buvo tikrinamas atliekant gautų žymenų fizikinę interpretaciją.

2.1. Darbe naudojami spektrinių duomenų rinkiniai

Šiame darbe spektrinių žymenų išgavimo tyrimui kaip pavyzdžiai naudojami du biologinių bandinių FT-IR spektrų rinkiniai:

- pataloginių bakterijų ir mielių;
- ekstraląstelių vezikulių, išgautų iš sveikų žmonių ir miokardo infarkto pacientų kraujo;

Patologinių bakterijų ir grybelių duomenų rinkinį sudaro FT-IR spektrai iš 4 skirtingų mikroorganizmų rūšių: dvi mielių rūšys *Candida guilliermondii* (*C. guilliermondii*), *Candida parapsilosis* (*C. parapsilosis*) ir dvi bakterijų rūšys *Streptococcus pyogenes* (*S. pyogenes*) bei *Staphylococcus aureus* (*S. aureus*). Šiame darbe šis spektrų rinkinys buvo panaudotas trejuose žymenų išgavimo uždaviniuose:

- 1) bakterijų rūšių atskyrimui nuo mielių;
- 2) bakterijų rūšių *S. pyogenes* ir *S. aureus* atskyrimui;
- 3) mielių rūšių *C. guilliermondii* ir *C. parapsilosis* atskyrimui.

Dalis mikroorganizmų FT-IR spektrų buvo užregistruoti šio darbo autoriaus, o kita dalis spektrų buvo paimta iš Vilniaus universiteto Fizikos fakulteto Cheminės fizikos instituto doktorantės Gerdos Anužienės parengto mikroorganizmų spektrų rinkinio. Duomenų rinkinį iš viso sudaro 100 spektrų.

Ekstraląstelinių vezikulių duomenų rinkinį sudaro vezikulių, išgautų iš sveikų žmonių ir miokardo infarkto pacientų kraujo, FT-IR spektrai. Ekstraląstelinių vezikulių bandiniai buvo paruošti Vilniaus universiteto Medicinos fakulteto laboratorijose. Jų spektrai buvo užregistruoti Vilniaus universiteto Fizikos fakulteto Cheminės fizikos instituto mokslo darbuotojos dr. Rimantės Bandzevičiūtės. Duomenų rinkinyje iš viso yra 84 spektrai, iš kurių 26 buvo miokardo infarkto pacientų vezikulių bandinių.

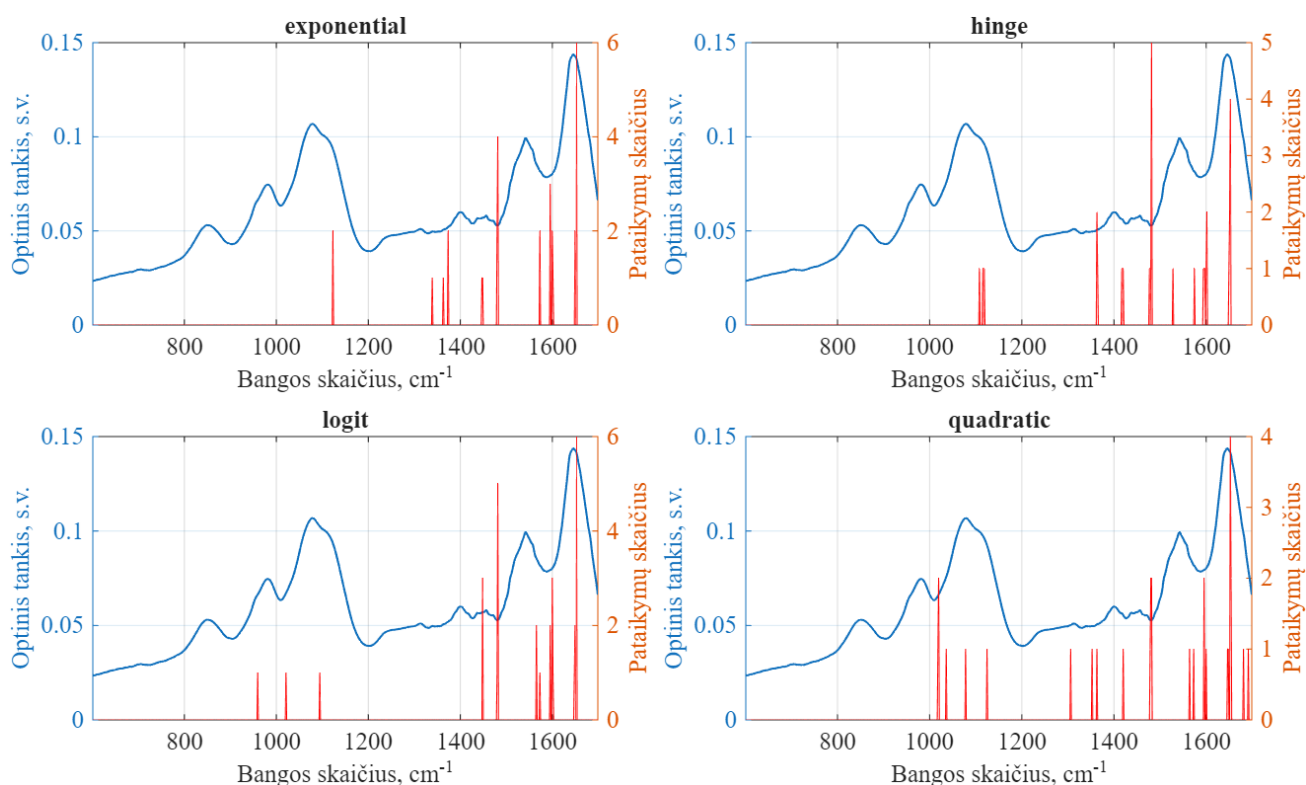
2.2. Rezultatai

2.2.1. Kriterijaus funkcijos pasirinkimas FFS algoritmui

Siekiant pasirinkti FFS kriterijaus funkciją buvo atlikti trumpi FFS skaičiavimai su skirtingomis kriterijaus funkcijomis apkirptiems, bet neapdorotiems spektrams (6 kartojimai su 5 kintamaisiais vezikulėms, 12 kartojimų su 3 kintamaisiais mikrobams, kadangi šie per greitai konverguoja). Kriterijaus funkcijos buvo pasirinktos iš MATLAB esančių paruoštų funkcijų, jų apibrėžimas pateiktas šio darbo priede 8 lentelėje ir taip pat randamas MATLAB dokumentacijoje [21]. Kriterijaus funkcija buvo pasirinkta remiantis gautų histogramų stabilumo indekso (jis apibrėžtas skyriuje 2.2.4.) priklausomybe nuo pasirinktos funkcijos, taip pat pagal histogramų logiškumą ir polinkį į ankstyvą konvergavimą. 2 lentelėje matoma, kad mikroorganizmams stabiliausia kriterijaus funkcija buvo „hinge“, o vezikulėms – „quadratic“. 6 pav. matoma, kad praktiniai skirtumai tarp histogramų yra nežymūs, o mikroorganizmų atveju kvadratinė kriterijaus funkcija pagerina ankstyvą konvergavimą nedaug palyginus su „hinge“ funkcija. Taigi tolesniems skaičiavimams buvo pasirinkta „hinge“ funkcija.

2 lentelė. Mikroorganizmų ir vezikulių spektrų rinkinių FFS histogramų, gautų su skirtingomis FFS kriterijaus funkcijoms, stabilumo indeksai. Spektrai prieš FFS buvo apdoroti tik apkarpančiais 1700-600 cm⁻¹ srities.

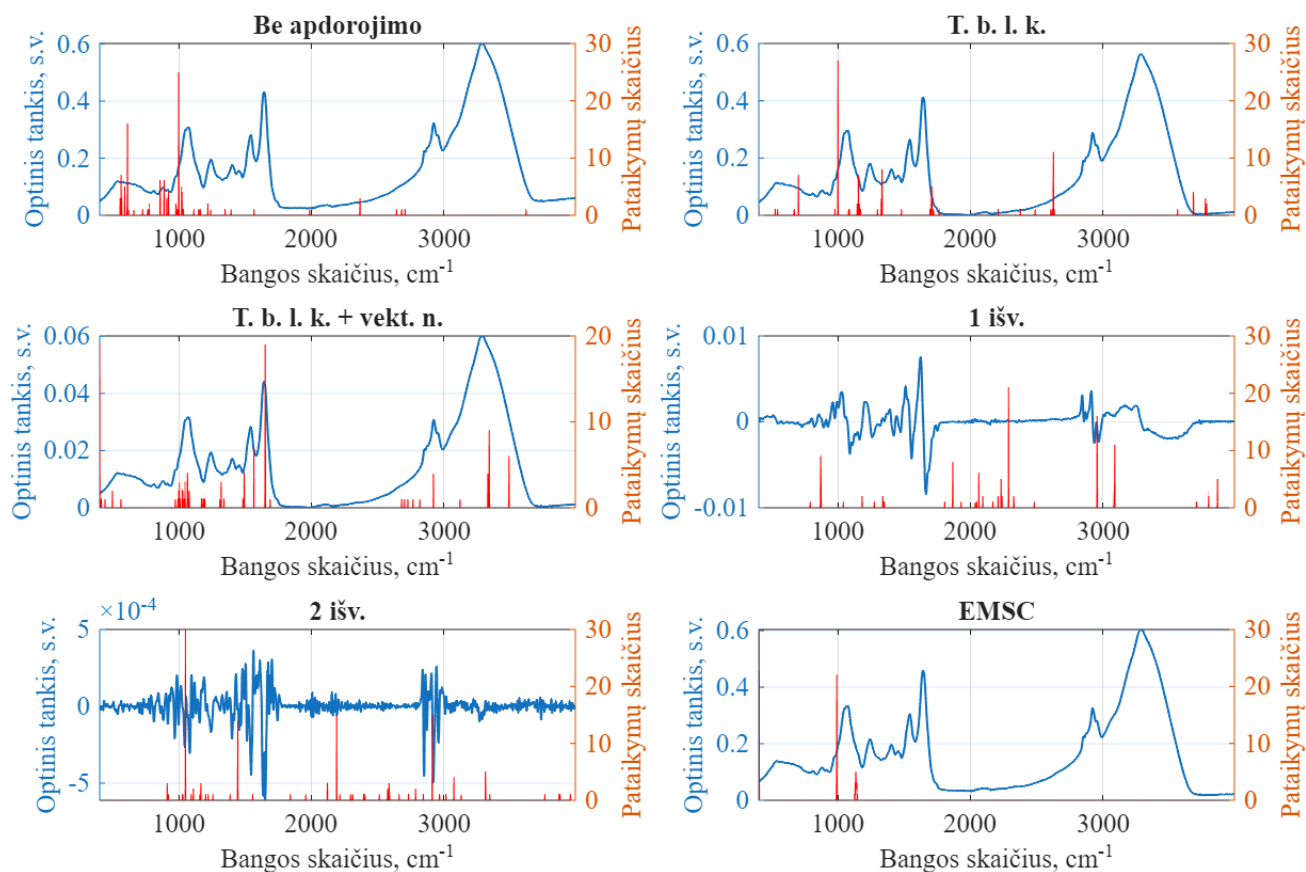
Kriterijaus funkcija	Mikroorganizmų stabilumo ind.	Vezikulių stabilumo ind.
<i>binodeviance</i>	0,341	0,356
<i>classifcost</i>	0,402	0,275
<i>classiferror</i>	0,402	0,275
<i>exponential</i>	0,346	0,383
<i>hinge</i>	0,412	0,463
<i>logit</i>	0,361	0,476
<i>quadratic</i>	0,356	0,584



6 pav. Vezikulių FT-IR spektrų FFS histogramos, gautos naudojantis skirtingomis kriterijaus funkcijomis.

2.2.2. Gautos spektrinių žymenų histogramos ir jų interpretacijos

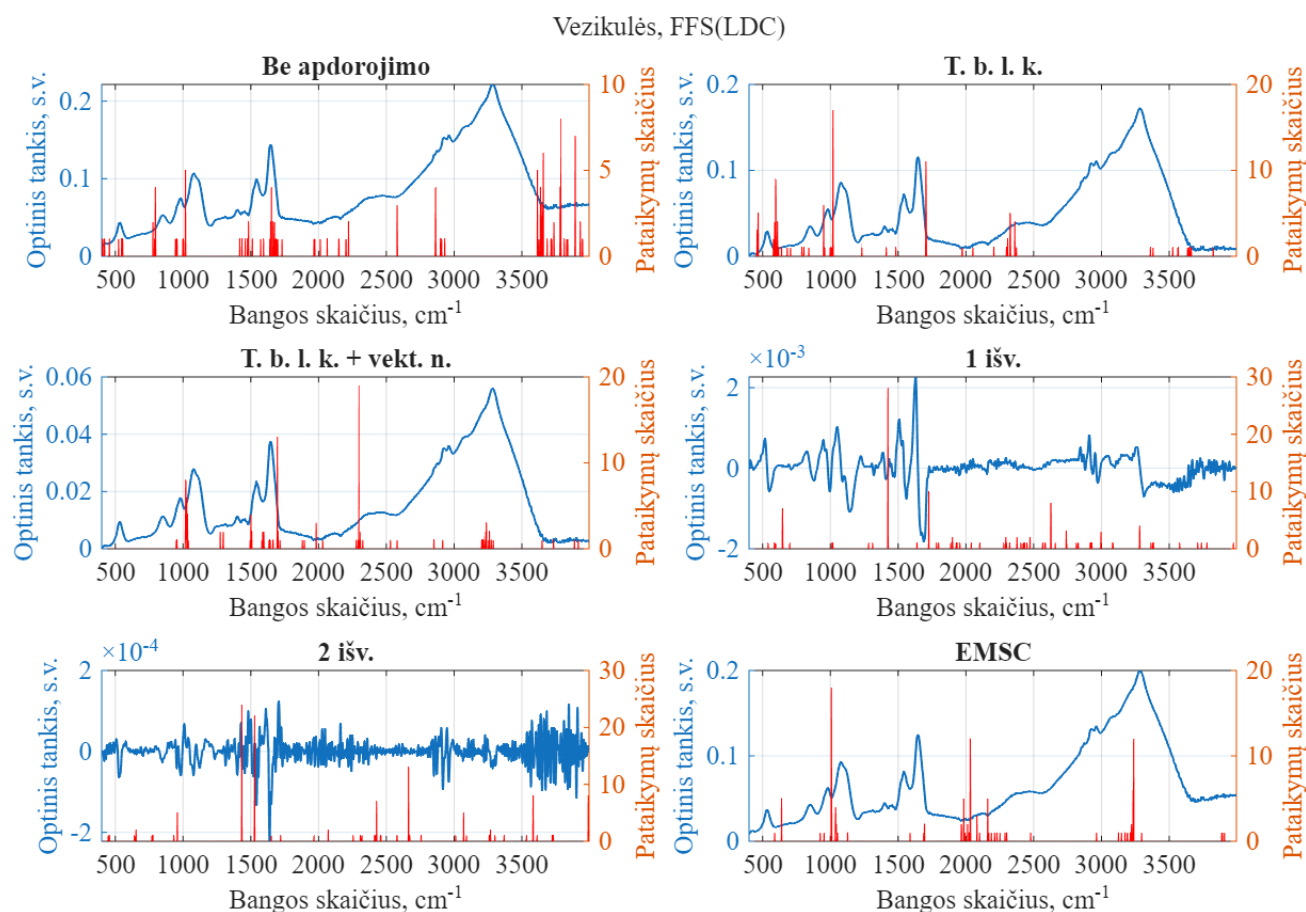
Toliau pateiktos kelios šiame tyrime su FFS(LDC) ir PCA-LDC metodais gautos spektrinių žymenų histogramos ekstralastelinių vezikulių ir mikroorganizmų (bakterijų/mielų atskyrimas) rinkiniams su skirtingais apdorojimais. Būtent šie rinkiniai pasirinkti dėl to, kad biocheminiai skirtumai tarp juose esančių bandinių grupių yra gerai žinomi, kas leidžia patikrinti gautų žymenų spektrinį pagrįstumą. Dėl glaustumo histogramose pavaizduota tik dalis tyrime nagrinėtų spektrų pirminio apdorojimo procedūrų. Apdorojimai su ir be spektrų apkarpymo buvo atvaizduoti atskiruose grafikuose.



7 pav. Mikroorganizmų pilni FT-IR spektrai, apdoroti skirtingais metodais (mėlyna spalva), kartu su spektrinių žymenų histogramomis, išgautomis FFS algoritmu (raudona spalva). Grafikuose vaizduojamas tas pats spektras, atsitiktinai parinktas iš duomenų rinkinio.

7 pav. matomi spektrinių žymenų išgavimo FFS algoritmu rezultatai mikroorganizmų FT-IR spektrų rinkiniui su apdorojimais be apkarpyto. Pradedant nuo atvejo be apdorojimo, didžioji dauguma žymenų pataikymų yra $<1000\text{ cm}^{-1}$ srityje, o intensyviausia viršūnė yra ties 997 cm^{-1} . Pritaikius bazinės linijos korekciją žymenų stabilumas pagerėjo (sumažėjo pataikymų išsibarstymas), intensyviausia viršūnė išliko 997 cm^{-1} , tačiau atsirado su spektrinėmis juostomis nesusijęs pataikymas ties 2628 cm^{-1} . Vektorinio normavimo pridėjimas sukėlė persimokymą, matomą kaip žymenų pataikymą pačiame spektro gale. Intensyviausios žymenų viršūnės šiuo atveju yra ties amidų I ir II juostomis, taip pat yra platus žymenų pasiskirstymas ties $\sim 1060\text{ cm}^{-1}$ su angliavandeniais susijusioje juostoje. Taip pat yra pataikymų plačioje vandens juostoje ties ~ 3340 ir $\sim 3490\text{ cm}^{-1}$. Apdorojus pirmos eilės diferencijavimu dauguma FFS pataikymų atsidūrė tuščioje srityje nuo 1800 iki 2600 cm^{-1} , vieninteliai aktualūs pataikymai yra ties $-\text{CH}_3$ asimetriniais valentiniais virpesiais ties 2958 cm^{-1} . Antros eilės diferencijavimo rezultatai buvo geresni, daugiau pataikymų buvo pirštų antspaudų srityje, o intensyviausia viršūnė buvo ties 1048 cm^{-1} angliavandenių juosta, tačiau intensyvus pataikymas CO_2 virpesių srityje išliko. EMSC apdorojimo metodas taip pat sukėlė persimokymą panašiai, kaip bazinės

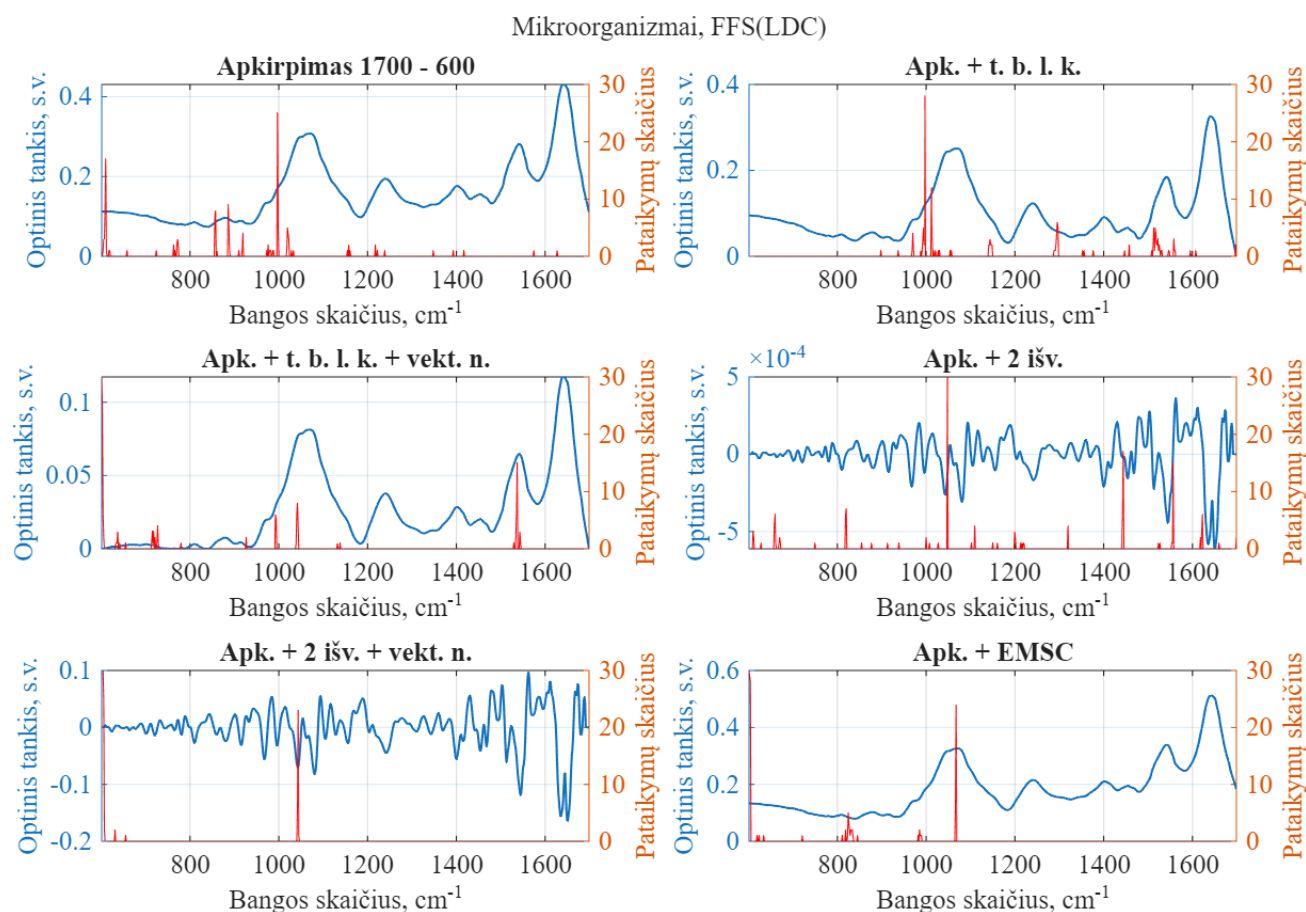
linijos k. su vektoriniu normavimu atveju, tačiau EMSC buvo daug stabilesnis, vieninteliai pataikymai yra ties 990 ir 1135 cm^{-1} . Iš minėtų apdorojimų laikoma, kad tiksliausias spektrinių žymenų rezultatas gautas su apdorojimu bazinės linijos korekcija + vektoriniu normavimu. Taip manoma todėl, kad nepaisant persimokymo, pagrindiniai žymenys yra ties baltymų amidų I ir II juostomis, taip pat 1060 cm^{-1} angliavandenių juostoje, o vienas iš esminių biocheminių skirtumų tarp mielių ir bakterijų yra koncentracijų skirtumai tarp minėtų biomolekulių.



8 pav. Ekstraląstelių vezikulių pilni FT-IR spektrai, apdoroti skirtingais metodais (mėlyna spalva), kartu su spektrinių žymenų histogramomis, išgautomis FFS algoritmu (raudona spalva). Grafikuose vaizduojamas tas pats spektras, atsitiktinai parinktas iš duomenų rinkinio.

8 pav. matomi ekstraląstelių vezikulių neapkarpytų FT-IR spektrų FFS spektrinių žymenų išgavimo rezultatai. Atveju be jokio apdorojimo gauti žymenys yra labai nestabilūs ir išsibarstę, dauguma jų sukoncentruoti tuščioje 3600 – 4000 cm^{-1} srityje. Bazinės linijos korekcija pagerina situaciją: žymenys daug stabilesni, taip pat daug mažiau jų yra neaktualiose srityse. Pridėjus vektorinį normavimą atsiranda intensyvi, bet neaktuali pataikymų viršūnė ties 2300 cm^{-1} , o kiti svarbesni pataikymai ties 1696 ir 1020 cm^{-1} išlieka. Pirmos išvestinės spektruose pasirodė intensyvus pataikymas ties 1424 cm^{-1} , kuris yra šalia COO^- grupės simetrinių valentinių virpesių aminorūgštyse juostos ir CH_2 grupės

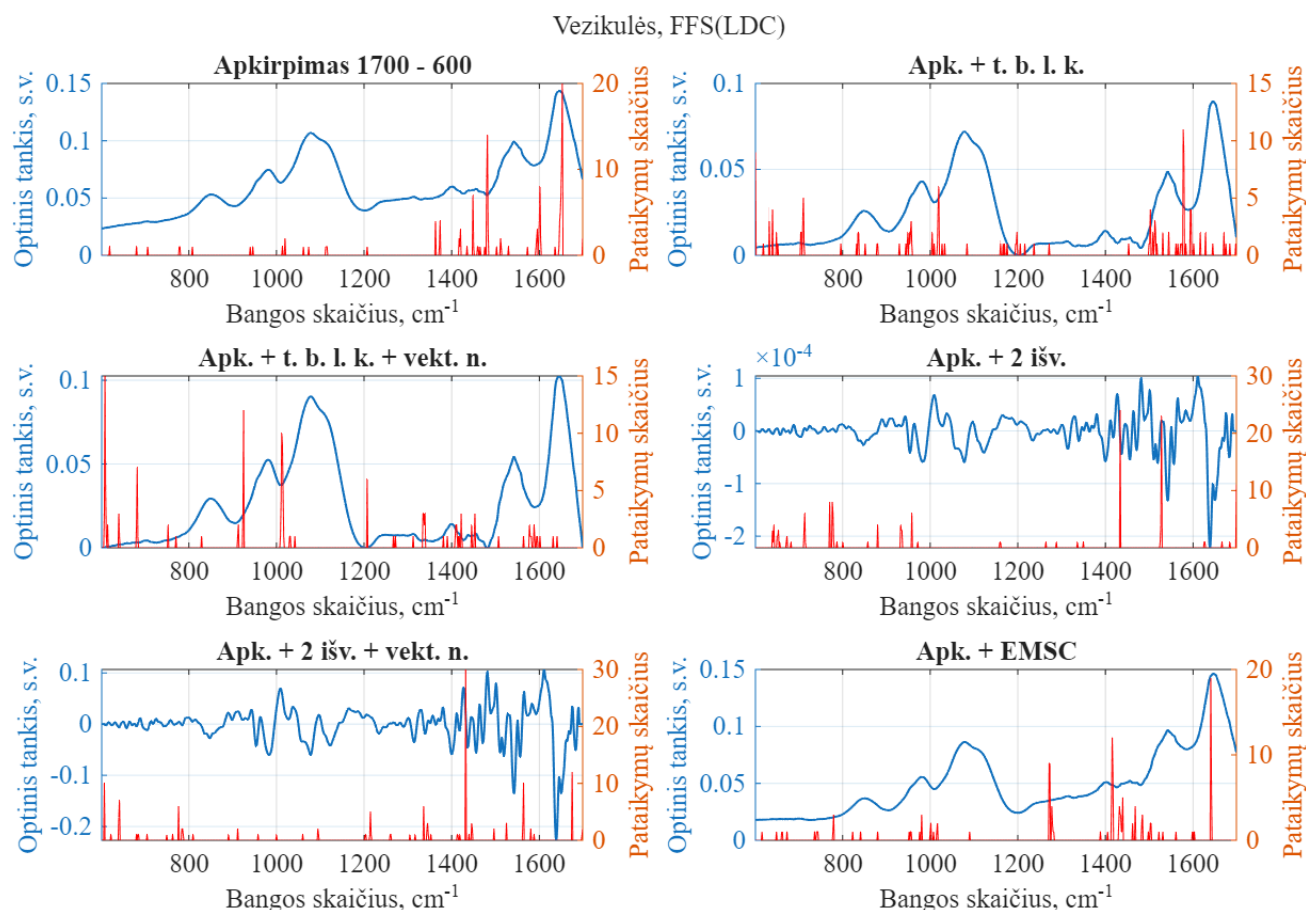
deformacinių virpesių lipiduose ir baltymuose juostos. Antros išvestinės spektruose yra tas pats pataikymas kartu su nauju pataikymu ties 1530 cm^{-1} , atitinkančiu amidų II juostą. Spektruose, apdorotuose EMSC metodu, yra intensyvus pataikymas ties 1000 cm^{-1} , tačiau kiti pataikymai yra neaktualiose srityse, prie CO_2 ir H_2O virpesių. Kaip tiksliausi žymenys laikomi atveju su antros išvestinės apdorojimu, kadangi jie arčiausiai koreliuoja su vezikulėse esančių baltymų galimais pokyčiais.



9 pav. Mikroorganizmų apkarpyti FT-IR spektrai, apdoroti skirtingais metodais (mėlyna spalva), kartu su spektrinių žymenų histogramomis, išgautomis FFS algoritmu (raudona spalva). Grafikuose vaizduojamas tas pats spektras, atsitiktinai parinktas iš duomenų rinkinio.

9 pav. matomi FFS žymenų išgavimo rezultatai mikroorganizmų FT-IR spektrų rinkiniui, kai kartu su pirminiu apdorojimu spektrai buvo apkarpyti iki $1700\text{--}600\text{ cm}^{-1}$ „pirštų antspaudų“ srities. Pirma pastebima, kad spektruose tik su apkarpymu stebima viena intensyvi pataikymų viršūnė ties 997 cm^{-1} , taip pat stebimas pataikymas dėl FFS algoritmo persimokymo spektro gale prie 600 cm^{-1} . Tiesinė bazinės linijos korekcija panaikino persimokymą ir išlaikė tą pačią pataikymų viršūnę. Tuo tarpu vektorinis normavimas vėl sukėlė persimokymą - spektro gale matomas intensyvi viršūnė dėl ankstyvo konvergavimo. Be persimokymo viršūnės, kiti intensyviausi pataikymai yra ties 1537 cm^{-1} amidų II juosta, taip pat ties 994 ir 1042 cm^{-1} juostomis. Antros išvestinės spektruose intensyviausia viršūnė yra

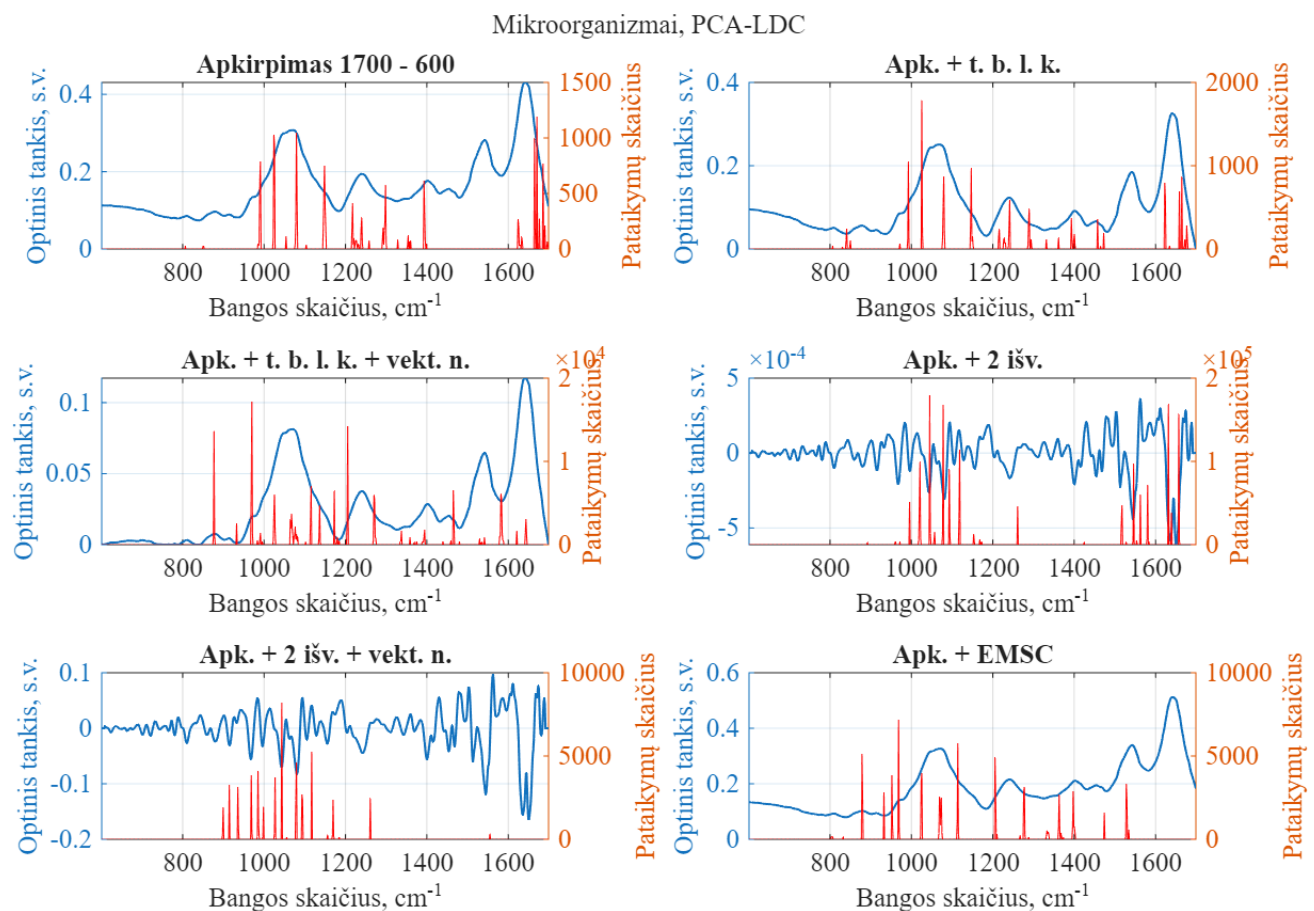
ties 1048 cm^{-1} , o sekančios dvi yra ties 1444 ir 1557 cm^{-1} $-\text{CH}_2$ deformacinių ir amidų II virpesių spektrinėmis juostomis atitinkamai. Vektorinio normavimo pridėjimas prie spektrų antrų išvestinių irgi sukėlė persimokymą, bet taip pat gerai stabilizavo žymenis ir paliko tik vieną tą pačią viršūnę ties 1042 cm^{-1} . Rezultatas su EMSC apdorojimu labai panašus, kaip ir su antros išvestinės + vektorinio normavimo apdorojimu, tik viršūnės padėtis pasislinko į 1068 cm^{-1} . Tiksliausi FFS metodų gauti žymenys laikomi su spektrų antromis išvestinėmis dėl tų pačių argumentų, kaip ir prieš tai aptartu atveju su neapkarpytais spektrais.



10 pav. Ekstraląstelių vezikulių apkarpyti FT-IR spekrai, apdoroti skirtingais metodais (mėlyna spalva), kartu su spektrinių žymenų histogramomis, išgautomis FFS algoritmu (raudona spalva). Grafikuose vaizduojamas tas pats spektras, atsitiktinai parinktas iš duomenų rinkinio.

10 pav. matomi FFS žymenų išgavimo rezultatai ekstraląstelių vezikulių apkarpytų FT-IR spektrų rinkiniui su skirtingais apdorojimais. Atveju be apdorojimo FFS žymenys buvo sukoncentruoti ties baltymų virpesinėmis juostomis - ties amidų I juosta ir prie $-\text{CH}_2$ deformacinių virpesių baltymuose bei lipiduose (pataikymai ties 1480 ir 1448 cm^{-1}). Pridėjus tiesinę bazinės linijos korekciją žymenų stabilumas pablogėjo, intensyviausias pataikymas stebimas ties 1578 cm^{-1} tarp amidų I ir II juostų, bet kiti pataikymai labiau išsibarstę, taip pat stebimas pataikymas spektro gale dėl persimokymo. Pridėjus

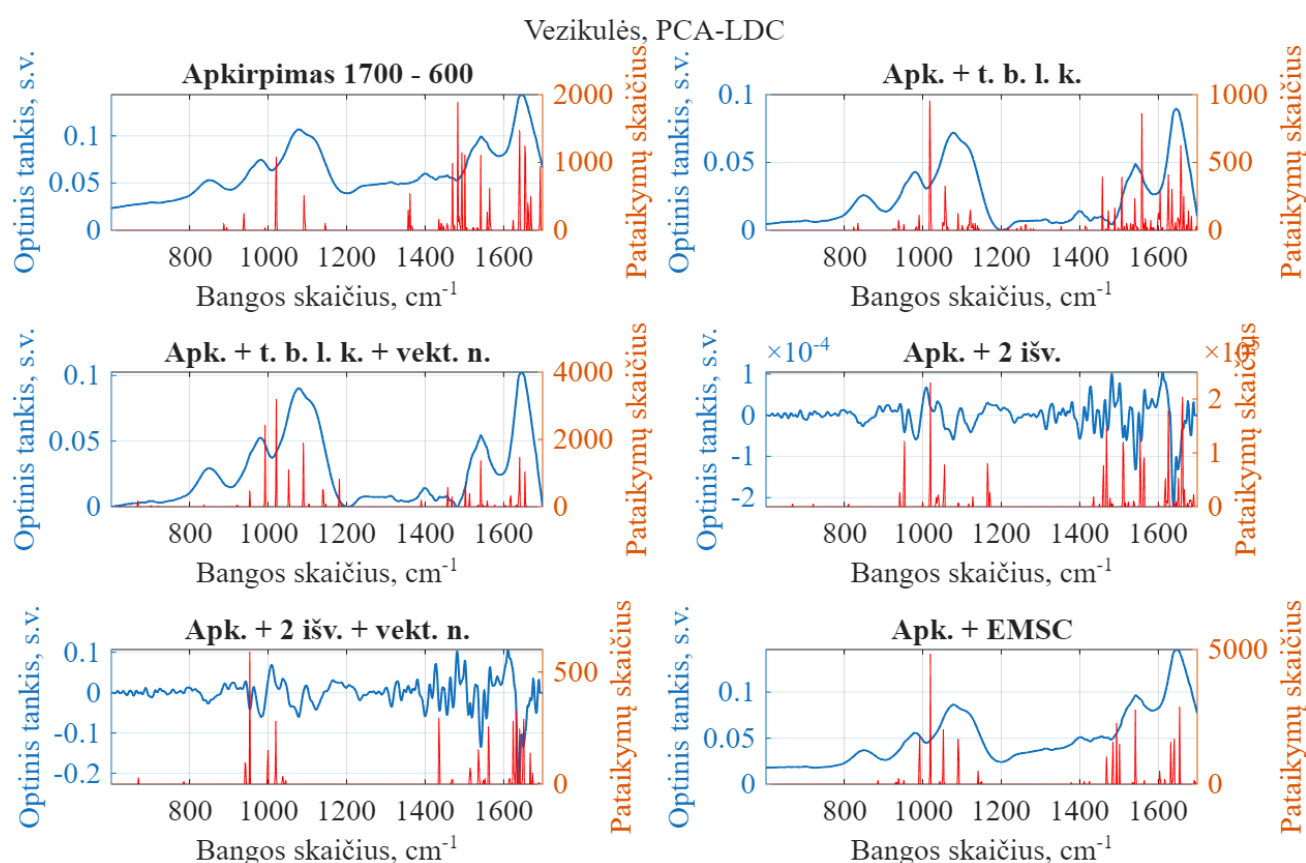
vektorinį normavimą taip pat buvo stebimas persimokymas ir žymenų išsibarstymas, intensyviausi žymenys yra ties 1012 ir 925 cm^{-1} . Spektrus apdorojus antros eilės diferencijavimu gauti žymenys yra stabilesni, du pagrindiniai pataikymai yra ties 1528 cm^{-1} amidų II juosta ir 1433 cm^{-1} , šalia COO^- grupės simetrinių valentinių virpesių aminorūgštyse juostos. Prie antros išvestinės pridėjus vektorinį normavimą gaunamas vienas intensyviausias pataikymas vėl ties 1433 cm^{-1} . Galiausiai su EMSC buvo stebimi trys intensyvūs pataikymai ties 1641 cm^{-1} amidų I juosta, 1416 cm^{-1} COO^- juosta ir 1272 cm^{-1} juosta. Tiksliausi žymenys buvo gauti su spektrų antromis išvestinėmis, kadangi jie buvo stabiliausi ir labiausiai susiję su pokyčiais vezikulių baltymuose.



11 pav. Mikroorganizmų apkarpyti FT-IR spektrai, apdoroti skirtingais metodais (mėlyna spalva), kartu su spektrinių žymenų histogramomis, išgautomis PCA-LDC algoritmu (raudona spalva). Grafikuose vaizduojamas tas pats spektras, atsitiktinai parinktas iš duomenų rinkinio.

11 pav. matomi mikroorganizmų FT-IR apkarpytų spektrų rinkinio spektriniai žymenys, gauti su PCA-LDC algoritmu ir su skirtingais apdorojimais. Visų pirma, atveju be papildomo apdorojimo žymenys yra daugiausia pasiskirstę 1150–990 cm^{-1} srityje, daugiausiai susijusioje su angliavandenių virpesiais, ir aplink 1670 cm^{-1} prie amidų I juostos. Žymenų rezultatas išlieka panašus pridėjus bazinės linijos korekciją. Pritaikius vektorinį normavimą spektriniai žymenys pasikeitė, šiuo atveju intensyviausi

žymenys buvo ties 1205, 970 ir 877 cm^{-1} sritimis, o žymenys ties angliavandenių ir amidų spektrinėmis juostomis buvo santykinai silpnėsni. Žymenys, gauti su spektrų antromis išvestinėmis, buvo gerai sukoncentruoti prieš tai minėtoje 1150 – 990 cm^{-1} srityje, taip pat ties amidų I ir II juostomis. Prie išvestinių pridėjus vektorinį normavimą visi žymenys susikoncentravo aplink angliavandenių sritį. Galiausiai su EMSC apdorojimu gauti žymenys buvo tolygiau pasiskirstę per visą spektrą ir pagal intensyvumą, svarbiausi žymenys buvo ties 970 ir 1114 cm^{-1} . Iš visų apdorojimų tiksliausi spektriniai žymenys pasirodė spektrų antrose išvestinėse – jie buvo gerai sukoncentruoti ties amidų ir angliavandenių spektrinėmis juostomis.



12 pav. Ekstraląstelinių vezikulių apkarpyti FT-IR spektrai, apdoroti skirtingais metodais (mėlyna spalva), kartu su spektrinių žymenų histogramomis, išgautomis PCA-LDC algoritmu (raudona spalva). Grafikuose vaizduojamas tas pats spektras, atsitiktinai parinktas iš duomenų rinkinio.

12 pav. matomi PCA-LDC žymenų išgavimo rezultatai apkarpytiems ekstraląstelinių vezikulių FT-IR spektrams su skirtingais apdorojimais. Atveju be apdorojimo matoma, kad dauguma žymenų yra su baltymais susijusiose srityse, ties amidų I ir II juostomis, taip pat srityje aplink 1480 cm^{-1} . Spektrus apdorojus bazinės linijos korekcija žymenys prie amidų I/II juostų šiek tiek persiskirstė, taip pat žymuo ties 1020 cm^{-1} tapo santykinai svarbesnis. Pridėjus vektorinį normavimą žymenys tapo stabilesni, daugiau santykinai intensyvesnių žymenų pasirodė 1180 – 990 cm^{-1} srityje. Spektrų antrose išvestinėse

su ir be vektorinio normavimo žymenys buvo išsidėstę panašiai kaip atveju su bazinės linijos korekcija, tik šiek tiek stabiliau. Apdorojimo su EMSC rezultatai taip pat yra panašūs į kitus atvejus, arčiausiai į apdorojimą bazinės linijos korekcija + vektoriniu normavimu. PCA-LDC žymenų padėtys vezikulių spektruose buvo labai panašūs visais čia aptartais atvejais. Todėl kaip tiksliausi žymenys yra laikomi gauti su spektrais, apdorotais bazinės linijos korekcija + vektoriniu normavimu, nes jie buvo stabiliausi. Žymenys su EMSC taip pat buvo panašiai stabilūs.

Šiame skyriuje aptartų FT-IR spektrų apdorojimo metodų įtaka spektrinių žymenų išgavimui, remiantis gautomis žymenų histogramomis, buvo apibendrinta 3 lentelėje.

3 lentelė. Šiame skyriuje aptartų FT-IR spektrų apdorojimo procedūrų poveikio FFS ir PCA-LDC metodais gautiems spektriniais žymenims apibendrinimas.

Apdorojimai	Poveikis gautiems žymenims
Be apdorojimo	Nestabilu, daug nereikšmingų žymenų
Tiesinė bazinės linijos korekcija	Geresnis stabilumas, mažiau nereikšmingų žymenų
Tiesinė bazinės linijos korekcija + vektorinis normavimas	Neblogas stabilumas, yra nereikšmingi žymenys dėl persimokymo
Pirma išvestinė	Stabilu, bet yra nereikšmingų žymenų
Antra išvestinė	Stabilu, gera interpretacija
EMSC	Stabilu, bet yra nereikšmingų žymenų
Apkarpymas iki 1700-600 cm^{-1}	Mažiau išsibarstymo ir nestabilumo palyginus su neapkarpytais spektrais
Apkarpymas + tiesinė bazinės linijos korekcija	Kai kuriais atvejais sumažino nereikšmingų žymenų kiekį, bet kitais atvejais pablogino stabilumą
Apkarpymas + tiesinė bazinės linijos korekcija + vektorinis normavimas	Didelis polinkis persimokyti, prastesnė interpretacija
Apkarpymas + antra išvestinė	Stabilu, mažai išsibarstymo, gera interpretacija
Apkarpymas + antra išvestinė + vektorinis normavimas	Didelis polinkis persimokyti, labai stabilu
Apkarpymas + EMSC	Didelis polinkis persimokyti, labai stabilu, nebloga interpretacija

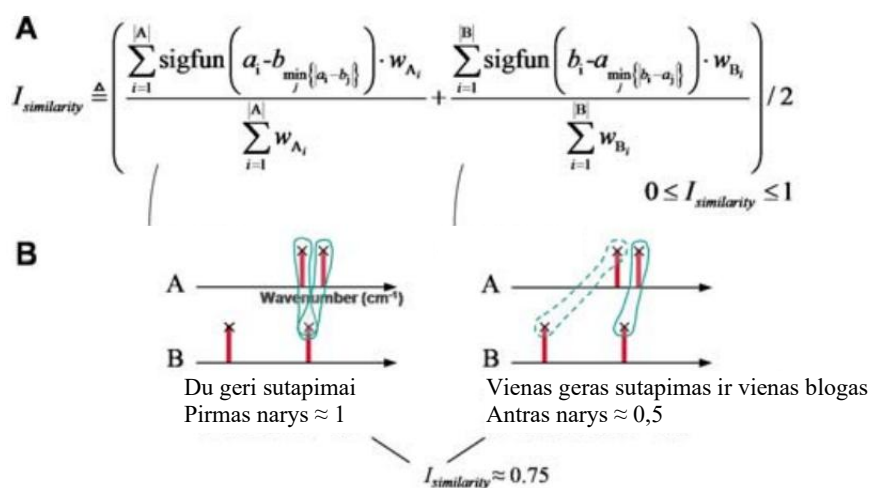
2.2.3. Spektrinių žymenų panašumo rezultatai

Šiame darbe skirtingais metodais gauti spektriniai žymenys buvo palyginami naudojantis žymenų panašumo indeksu $I_{similarity}$, kurį įvedė J. Trevisan *et al.* [7] Jo apibrėžimas pateiktas 13 pav. Indekso principas yra toks: duotos dvi žymenų histogramos - bangos skaičių aibės $A = \{a_1, a_2, \dots\}$ ir $B = \{b_1, b_2, \dots\}$ bei su jomis susiję svoriai (žymenų „intensyvumai“) $W_A = \{w_{A1}, w_{A2}, \dots\}$ ir $W_B = \{w_{B1}, w_{B2}, \dots\}$.

...}. Kiekvienam vienos histogramos žymeniui (nenuliniam svoriui) surandamas arčiausias atitinkamas žymuo kitoje histogramoje, ir jų panašumas įvertinamas naudojantis sigmoidine funkcija $\text{sigfun}(x)$:

$$\text{sigfun}(x) = \frac{1 + e^{-s}}{1 + e^{\frac{s}{hh}(|x| - hh)}}; \quad (8)$$

čia s – funkcijos statumo parametras, hh – pločio parametras. J. Trevisan *et al.* šios funkcijos skaičiavimui naudojo šias parametrų vertes: $s = 6$, $hh = 45 \text{ cm}^{-1}$ [7], tokios pačios vertės buvo naudojamos ir šiame tyrime. Žymenų panašumo indekso skaičiavimo metu $\text{sigfun}(x)$ argumentas yra bangos skaičių atstumas tarp anksčiau minėtų dvejų žymenų. Indeksas suskaičiuojamas susumuojant visas sigmoidinės funkcijos vertes atstumams tarp vienos histogramos žymenų ir kitos. Žymenų panašumo indeksas įgyja reikšmes nuo 0 iki 1, kur 1 reiškia, kad žymenų histogramos yra identiškos, o 0 – jog histogramos yra visiškai skirtingos.



13 pav. A formulė apskaičiuoti žymenų panašumo indeksą. B Indekso veikimo principo iliustracija.

Adaptuota iš [7].

4 lentelėje matomi FFS algoritmu gautų žymenų panašumo indeksai tarp įvairių apdorojimo procedūrų porų, suvidurkinti per visus keturis šiame darbe naudotus duomenų rinkinius. Visų pirma galima pastebėti, kad atvejis be apdorojimo (tik su iškirpimu iki pirštų antspaudų srities) gerai sutampa su apdorojimu su bazinės linijos korekcija bei vektoriniu normavimu, taip pat su išvestinėmis (panašumo indeksas $\approx 0,90$ su standartiniu nuokrypiu $\leq 0,05$). Išvestinės taip pat neblogai sutapo su kitais metodais, antros eilės išvestinė šiek tiek geriau nei pirmos. Antra, apdorojimas tik bazinės linijos korekcija išsiskyrė blogu panašumu į su kitais apdorojimais gautais žymenimis. EMSC metodas bei apdorojimas antros eilės išvestine ir vektoriniu normavimu taip pat prasčiau sutapo su kitais metodais. Taip pat pastebima, kad kadangi FFS žymenų rezultatai atvejais su ir be standartizacijos yra identiški, panašumo indeksai tarp šių atvejų lygūs 1.

4 lentelė. Tyrime naudotų spektrinių rinkinių FFS metodu gautų žymenų vidutiniai panašumo indeksai ir jų standartiniai nuokrypiai įvairioms apdorojimų poroms. Lentelės langeliai nuspalvinti gradientu nuo žalios iki raudonos spalvos, kur žalia spalva – didžiausias panašumas, o raudona – mažiausias panašumas.

	Apkirpimas 1700-600	Apk. + t. b. l. k.	Apk. + t. b. l. k. + vekt. n.	Apk. + t. b. l. k. + vekt. n. + st.	Apk. + EMSC	Apk. + EMSC + st.	Apk. + 1 išv.	Apk. + 2 išv.	Apk. + 2 išv. + vekt. n.
Apkirpimas 1700 - 600	1,00±0,00	0,85±0,08	0,91±0,03	0,91±0,03	0,82±0,12	0,82±0,12	0,89±0,03	0,90±0,05	0,75±0,24
Apk. + t. b. l. k.	0,85±0,08	1,00±0,00	0,71±0,14	0,71±0,14	0,53±0,37	0,53±0,37	0,73±0,19	0,72±0,21	0,49±0,40
Apk. + t. b. l. k. + vekt. n.	0,91±0,03	0,71±0,14	1,00±0,00	1,00±0,00	0,80±0,11	0,80±0,11	0,85±0,21	0,89±0,05	0,86±0,15
Apk. + t. b. l. k. + vekt. n. + st.	0,91±0,03	0,71±0,14	1,00±0,00	1,00±0,00	0,79±0,12	0,79±0,12	0,85±0,21	0,89±0,05	0,85±0,15
Apk. + EMSC	0,82±0,12	0,53±0,37	0,80±0,11	0,79±0,12	1,00±0,00	1,00±0,00	0,83±0,11	0,85±0,16	0,79±0,17
Apk. + EMSC + st.	0,82±0,12	0,53±0,37	0,80±0,11	0,79±0,12	1,00±0,00	1,00±0,00	0,83±0,11	0,85±0,16	0,79±0,17
Apk. + 1 išv.	0,89±0,03	0,73±0,19	0,85±0,21	0,85±0,21	0,83±0,11	0,83±0,11	1,00±0,00	0,91±0,11	0,79±0,18
Apk. + 2 išv.	0,90±0,05	0,72±0,21	0,89±0,05	0,89±0,05	0,85±0,16	0,85±0,16	0,91±0,11	1,00±0,00	0,84±0,15
Apk. + 2 išv. + vekt. n.	0,75±0,24	0,49±0,40	0,86±0,15	0,85±0,15	0,79±0,17	0,79±0,17	0,79±0,18	0,84±0,15	1,00±0,00

5 lentelė. Tyrime naudotų spektrinių rinkinių PCA-LDC metodu gautų žymenų vidutiniai panašumo indeksai ir jų standartiniai nuokrypiai įvairioms apdorojimų poroms. Lentelės langeliai nuspalvinti gradientu nuo žalios iki raudonos spalvos, kur žalia spalva – didžiausias panašumas, o raudona – mažiausias panašumas.

	Apkirpimas 1700-600	Apk. + t. b. l. k.	Apk. + t. b. l. k. + vekt. n.	Apk. + t. b. l. k. + vekt. n. + st.	Apk. + EMSC	Apk. + EMSC + st.	Apk. + 1 išv.	Apk. + 2 išv.	Apk. + 2 išv. + vekt. n.
Apkirpimas 1700 - 600	1,00±0,00	0,99±0,02	0,96±0,03	0,67±0,17	0,93±0,09	0,81±0,10	0,95±0,03	0,91±0,07	0,87±0,11
Apk. + t. b. l. k.	0,99±0,02	1,00±0,00	0,97±0,02	0,77±0,07	0,94±0,06	0,89±0,07	0,96±0,02	0,91±0,06	0,87±0,10
Apk. + t. b. l. k. + vekt. n.	0,96±0,03	0,97±0,02	1,00±0,00	0,77±0,07	0,97±0,02	0,94±0,03	0,97±0,02	0,93±0,09	0,88±0,11
Apk. + t. b. l. k. + vekt. n. + st.	0,67±0,17	0,77±0,07	0,77±0,07	1,00±0,00	0,76±0,11	0,82±0,11	0,66±0,20	0,67±0,31	0,65±0,29
Apk. + EMSC	0,93±0,09	0,94±0,06	0,97±0,02	0,76±0,11	1,00±0,00	0,95±0,02	0,93±0,04	0,87±0,13	0,88±0,10
Apk. + EMSC + st.	0,81±0,10	0,89±0,07	0,94±0,03	0,82±0,11	0,95±0,02	1,00±0,00	0,86±0,05	0,81±0,12	0,79±0,13
Apk. + 1 išv.	0,95±0,03	0,96±0,02	0,97±0,02	0,66±0,20	0,93±0,04	0,86±0,05	1,00±0,00	0,96±0,03	0,90±0,07
Apk. + 2 išv.	0,91±0,07	0,91±0,06	0,93±0,09	0,67±0,31	0,87±0,13	0,81±0,12	0,96±0,03	1,00±0,00	0,94±0,05
Apk. + 2 išv. + vekt. n.	0,87±0,11	0,87±0,10	0,88±0,11	0,65±0,29	0,88±0,10	0,79±0,13	0,90±0,07	0,94±0,05	1,00±0,00

5 lentelėje matomi PCA-LDC algoritmu gautų žymenų panašumo indeksai tarp įvairių apdorojimo procedūrų porų, suvidurkinti per visus keturis šiame darbe naudotus duomenų rinkinius. Pastebima, kad atvejai su tik apkarpyimu, bazinės linijos korekcija ir korekcija su normavimu yra labai panašūs. Tai yra pokytis palyginus su FFS algoritmu gautais žymenimis, kur apdorojimas su tik bazinės linijos korekcija reikšmingai skyrėsi nuo kitų. Rezultatai su EMSC metodu taip pat pasirodė panašesni į kitus palyginus su FFS panašumo rezultatais. Prasčiausiai pasirodė apdorojimai su standartizavimu, kurie buvo mažiausiai panašūs ir tarp kitų metodų, ir tarp vienas kito. Apdorojimas su antros eilės išvestine ir vektoriniu normavimu taip pat buvo mažiau panašus nei kiti, nors jo vidutinis panašumas yra šiek tiek geresnis nei FFS atveju.

6 lentelė. FHG žymenų panašumo indeksai tarp FFS ir PCA-LDC histogramų gavimo metodų tiems patiems duomenų rinkiniams su skirtingais apdorojimais. Lentelėje taip pat pateikti panašumo indeksų vidurkiai kiekvienam apdorojimui kartu su standartiniu nuokrypiu. Lentelės langeliai nuspalvinti gradientu nuo žalios iki raudonos spalvos, kur žalia spalva – didžiausias panašumas, o raudona – mažiausias panašumas.

Apdorojimai	Mikrobai	Vezikulės	Bakterijos	Mieles	Vidurkis
Be apdorojimo	0,42	0,63	0,32	0,88	0,56±0,22
T. b. l. k.	0,72	0,66	0,57	0,58	0,63±0,06
P. b. l. k.	0,45	0,45	0,19	0,88	0,49±0,25
R. b. l. k.	0,49	0,51	0,19	0,43	0,40±0,13
T. b. l. k. + vekt. n.	0,80	0,88	0,79	0,53	0,75±0,13
R. b. l. k. + vekt. n.	0,61	0,87	0,58	0,57	0,66±0,12
1 išv.	0,34	0,68	0,51	0,47	0,50±0,13
2 išv.	0,59	0,55	0,66	0,51	0,58±0,06
T. b. l. k. + vekt. n. + st.	0,49	0,91	0,76	0,67	0,71±0,15
2 išv. + st.	0,63	0,50	0,87	0,83	0,71±0,15
2 išv. + vekt. n.	0,49	0,56	0,34	0,62	0,50±0,11
EMSC	0,59	0,86	0,78	0,78	0,75±0,10
Apkirpimas 1700 – 600 cm ⁻¹	0,84	0,98	0,86	0,80	0,87±0,07
Apk. + t. b. l. k.	0,83	0,96	0,98	0,63	0,85±0,14
Apk. + t. b. l. k. + vekt. n.	0,53	0,88	0,91	0,90	0,81±0,16
Apk. + t. b. l. k. + vekt. n. + st.	0,83	0,92	0,89	0,96	0,90±0,05
Apk. + EMSC	0,62	0,94	0,98	0,72	0,82±0,15
Apk. + EMSC + st.	0,47	0,89	0,98	0,73	0,77±0,19
Apk. + 1 išv.	0,72	0,96	0,80	0,93	0,85±0,10
Apk. + 2 išv.	0,92	0,86	0,68	0,79	0,81±0,09
Apk. + 2 išv. + vekt. n.	0,65	0,73	0,82	0,89	0,77±0,09

6 lentelėje matomi panašumo indeksai, kurie palygina spektrinius žymenis, išgautus naudojantis FFS ir PCA-LDC algoritmais. Visų pirma aiškiai pastebima, kad vidutinis panašumas tarp minėtų metodų geresnis, kai spektrai yra apkarpyti. Mažiausias panašumas apkarpytų spektrų atveju (0,78 su EMSC

+ standartizavimu) buvo geresnis už didžiausią panašumą su neapkarpytais spektrais (0,75 su EMSC arba tiesine bazinės linijos korekcija + vektoriniu normavimu). Didžiausias panašumas (0,90) gautas su apkarpymu + tiesine bazinės linijos korekcija + vektoriniu normavimu + standartizavimu, be to, standartinis nuokrypis šiuo atveju (0,05) irgi buvo geriausias. Taip pat matoma, kad tik su dvejais apdorojimais panašumo indeksai buvo didesni nei 0,80 visiems spektrų rinkiniams. Pastebėtina, kad J. Trevisan *et al.* biologinių bandinių FT-IR spektrinių žymenų panašumo tyrime vidutinė panašumo tarp FFS(LDC) ir PCA-LDC indekso vertė yra $0,68 \pm 0,12$ [7]. Tame tyrime FT-IR spektrų rinkiniai buvo apdoroti apkirpimu iki $1800 - 900 \text{ cm}^{-1}$ srities, „rubberband“ bazinės linijos korekcija ir normavimu pagal amidų I juostą, o šiame tyrime šis apdorojimas nebuvo išbandytas. Be to, prieš tai minėta panašumo indekso vertė gauta PCA-LDC skaičiuojant tik su pirmais 10 PCA komponentų, tuo tarpu šiame tyrime buvo naudojami pirmi 15 komponentų. Turint omeny šiuos skirtumus, šiame tyrime visiems apdorojimams su spektrų apkarpymu panašumo indeksas patenka į intervalą nuo 0,77 iki 0,90, kas yra reikšmingai daugiau nei J. Trevisan *et al.* gauta 0,68 vertė.

2.2.4. Spektrinių žymenų stabilumo rezultatai

Šiame darbe spektrinių žymenų išgavimo stabilumo priklausomybės nuo spektrų rinkinių pirminio apdorojimo tyrimui buvo naudojamas požymių atrinkimo stabilumo indeksas, įvestas L. Kuncheva [22]. Dvejomis požymių atrinkimo sekoms, kai iš kintamųjų aibės X išrenkamos dvi poaibės A ir B , stabilumo indeksas $I_C(A, B)$ apibrėžiamas kaip:

$$I_C(A, B) = \frac{rn - k^2}{k(n - k)}; \quad (9)$$

čia $n = |X|$ - visos kintamųjų aibės dydis (visų kintamųjų skaičius), $k = |A| = |B|$ - poaibių dydis (išrinktų kintamųjų skaičius), o $r = |A \cap B|$ - poaibių sankirtos dydis (skaičius kintamųjų, bendrų abejoms poaibėms). Stabilumo indeksas įgyja vertes nuo 0 iki 1 imtinai, kur 1 reikštų, kad poaibės yra identiškos ($r = k$), o 0 – kad poaibės yra visiškai skirtingos ($r = 0$). Atveju, kai iš viso yra K požymių atrinkimo sekų, stabilumo indeksas yra apibrėžiamas kaip visų sekų porų stabilumo indeksų vidurkis:

$$I_S(S_1, S_2, \dots, S_K) = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^K I_C(S_i, S_j); \quad (10)$$

čia $S_1, S_2, \dots, S_K$ – kintamųjų poaibės, gautos iš K požymių atrinkimo sekų. Nors stabilumo indeksas originaliai buvo apibrėžtas požymių atrinkimo algoritmui, šiame tyrime jis taip pat buvo naudojamas tikrinti PCA-LDC metodo stabilumą pavertus iš jo gautus koeficientus į kintamųjų histogramas.

7 lentelė. Visų duomenų rinkinių žymenų stabilumo indeksų vidurkiai, suskaičiuoti skirtingiems žymenų gavimo metodams (FFS ir PCA-LDC) ir apdorojimo metodams. Taip pat pateiktas indekso skirtumas tarp minėtų dviejų metodų. Lentelės langeliai nuspalvinti gradientu nuo žalios iki raudonos spalvos, kur žalia spalva – didžiausias stabilumas, o raudona – mažiausias stabilumas.

Apdorojimas\Stabilumas	FFS(LDC)	PCA-LDC	Skirtumas
Be apdorojimo	0,203	0,421	0,218
T. b. l. k.	0,202	0,331	0,129
P. b. l. k.	0,282	0,359	0,077
R. b. l. k.	0,303	0,355	0,052
T. b. l. k. + vekt. n.	0,251	0,446	0,196
R. b. l. k. + vekt. n.	0,230	0,496	0,266
1 išv.	0,265	0,449	0,184
2 išv.	0,349	0,613	0,264
T. b. l. k. + vekt. n. + st.	0,251	0,359	0,108
2 išv. + st.	0,349	0,489	0,140
2 išv. + vekt. n.	0,466	0,665	0,199
EMSC	0,380	0,501	0,121
Apkirpimas 1700 – 600 cm ⁻¹	0,228	0,422	0,194
Apk. + t. b. l. k.	0,236	0,412	0,176
Apk. + t. b. l. k. + vekt. n.	0,297	0,468	0,171
Apk. + t. b. l. k. + vekt. n. + st.	0,308	0,477	0,169
Apk. + EMSC	0,341	0,496	0,156
Apk. + EMSC + st.	0,341	0,413	0,072
Apk. + 1 išv.	0,307	0,373	0,066
Apk. + 2 išv.	0,412	0,540	0,128
Apk. + 2 išv. + vekt. n.	0,588	0,636	0,048

7 lentelėje matomi spektrinių žymenų stabilumo indekso vidurkiai visiems šiame darbe naudotiems FT-IR spektrų rinkiniams ir pirminio apdorojimo procedūroms. Galima pastebėti, kad visais atvejais PCA-LDC metodu gauti žymenys buvo stabilesni už FFS žymenis. Taip pat pastebima, kad apdorojimai su spektro apkarpytųjų yra stabilesni už atitinkamus apdorojimus su pilnais spektrais (viena išimtis – EMSC apdorojimas). Tai yra logiška, kadangi spektrų apkarpymas pašalina nereikalingą spektrinę informaciją, kas turėtų sumažinti tikimybę statistiniam modeliui padaryti klaidą. Be to, siauresnė spektrinė sritis (kitais tariant, mažesnis kintamųjų kiekis) savaime apriboja galimų žymenų variacijų kiekį. Kaip apdorojimo metodas, duodantis stabiliausius spektrinius žymenis, pasirodė antros eilės išvestinė + vektorinis normavimas. Šis metodas buvo stabiliausias tiek neapkarpytiems ir apkarpytiems spektrams, tiek FFS ir PCA-LDC metodams. Antros eilės išvestinė be normavimo taip pat pasiekė palyginti aukštą stabilumą.

2.3. Rezultatų aptarimas

Kalbant apie spektrinių žymenų išgavimo su išankstiniu požymių atrinkimu (FFS) rezultatus, vienas iš šio metodo trūkumų yra vidutiniškai didelis žymenų jautrumas mažiems duomenų imties pokyčiams. Tai pastebima tiek gautose žymenų histogramose (žr. 7-12 pav.) kaip didelis kiekis mažo dažnio pataikymų, taip pat stabilumo indeksuose su skirtingais apdorojimais (žr. 7 lentelę), ypač palyginus su PCA-LDC metodo stabilumu. Blogas stabilumas indikuotų, kad metodas yra nepatikimas ir juo gauti spektriniai žymenys yra labiau atsitiktiniai nei susiję su tikrais pokyčiais bandiniuose. Ši problema taip pat buvo pastebėta ir aptarta J. Trevisan *et al.* straipsnyje [7]. Jame tyrėjai pasiūlė FFS stabilizavimo metodą: požymių atrinkimo proceso metu, vykstant kryžminei validacijai kiekvienas apmokymo rinkinys yra papildomai padalijamas į dar kelias apmokymo – testavimo poras, ir kriterijaus funkcija tai kryžminės validacijos iteracijai yra apskaičiuojama kaip tų porų kriterijaus vidurkis. Tyrėjai pademonstravo, kad tokio stabilizavimo pritaikymas sėkmingai sumažino mažo dažnio pataikymų kiekį ir sukonzentravo pataikymus ties pagrindiniais spektriniais žymenimis, tačiau taip pat ir pareikalavo ilgesnio skaičiavimo laiko. Šiame darbe šis stabilizavimo metodas nebuvo taikomas, tačiau buvo pastebėta, kad kai kurie spektrų apdorojimai – spektrų apkarpymas, diferencijavimas, EMSC – reikšmingai pagerino FFS žymenų stabilumą.

Kita pastebėta problema su žymenų išgavimu FFS metodu yra dažnas persimokymas/ankstyvas konvergavimas. Gautuose žymenyse tai dažniausiai pasireiškė kaip pertekliniai pataikymai spektrų galuose (žr. 9 pav.). Verta paminėti, kad FFS persimokymas dažniausiai buvo stebimas su spektriniais apdorojimais, turinčiais normavimą, būtent su vektoriniu normavimu arba EMSC. Tai nebūtinai yra šių apdorojimų trūkumas, galimai problema yra ta, kad FFS metodas prastai veikia, kai atsiskyrimas tarp spektrų grupių yra per lengvas. Šią išvadą paremia tai, kad ankstyvas konvergavimas daugiau pasireiškė mikroorganizmų spektrų rinkinyje nei vezikulių rinkinyje, kadangi bakterijų ir mielių spektrai labai gerai atsiskiria vien dėl baltymų/angliavandenių koncentracijų skirtumo. FFS persimokymo problema gali būti sprendžiama bandant požymių atrinkimui panaudoti kitokius klasifikacijos modelius, pavyzdžiui, kvadratinio diskriminanto klasifikaciją (angl. *quadratic discriminant classifier*) arba k – arčiausių kaimynų (angl. *k - nearest neighbours*) klasifikaciją. Taip pat galima išbandyti tą pačią tiesinio diskriminanto klasifikaciją, bet su kitokiais hiperparametrais, pavyzdžiui, reguliarizavimu (šiam tyrime LDC buvo naudojamas be reguliarizavimo). Galiausiai galima išbandyti sudėtingesnius požymių išgavimo algoritmus, pavyzdžiui, genetinio algoritmo (angl. *genetic algorithm*) optimizacija pagrįstus metodus.

Ir FFS, ir PCA-LDC žymenims kai kuriems spektrų apdorojimams be apkarpymo buvo būdingi dažni pataikymai $1800 - 2800 \text{ cm}^{-1}$ srityje, kurioje nėra biologiniuose objektuose aktualių virpesinių juostų

(žr. 7 ir 8 pav.). Nepaisant to ir kitų prieš tai minėtų trūkumų, FFS metodas kartu su tam tikrais spektriniais apdorojimais sėkmingai identifiko spektrinius žymenis su gera spektrine interpretacija bei žymenų stabilumu. Geri FFS algoritmo veikimo pavyzdžiai matomi mikroorganizmų FT-IR spektrų rinkinyje su tiesinės bazinės linijos korekcijos + vektorinio normavimo (žr. 7 pav.), antros išvestinės su ir be spektrų apkarpyimo iki $1700 - 600\text{ cm}^{-1}$ (žr. 7 ir 9 pav.) apdorojimo sekomis, taip pat vezikulių spektrų rinkinyje su antros išvestinės su ir be spektrų apkarpyimo (žr. 8 ir 10 pav.), apkarpyimo + EMSC (žr. 10 pav.) apdorojimais.

Pereinant prie žymenų išgavimo rezultatų su principinių komponentų analize – tiesinio diskriminanto klasifikacija (PCA - LDC), šiuo metodu išgauti žymenis buvo stabilesni nei gauti FFS metodu (žr. 7 lentelę), be to, su skirtingais apdorojimais gauti žymenis buvo panašesni į vienas kitą palyginus su FFS žymenimis (žr. 5 lentelę). PCA-LDC taip pat neturėjo FFS persimokymo problemos – didžioji dauguma žymenų buvo jei ne ties svarbiausiomis, tai bent ties bendrai spektriškai aktualiomis juostomis, ypač su apkarpytais spektrais (pavyzdžiui, mikroorganizmų rinkiniui su apkirpimo + EMSC apdorojimu, žr. 11 pav.). Kita vertus, gauti PCA-LDC žymenis buvo labiau išsibarstę nei FFS žymenis. Pavyzdžiui, mikroorganizmų spektrų atveju su apkirpimu + 2 išvestine + vektoriniu normavimu ir atveju su apkirpimu + EMSC PCA-LDC svarbiausi žymenis pagal interpretaciją yra ties maždaug 1050 cm^{-1} angliavandenių juosta, tačiau daug kitų panašaus intensyvumo žymenų yra stebimi šalia minėtos juostos, plačiame intervale nuo 900 iki 1260 cm^{-1} . Be išankstinių žinių būtų sunku nustatyti, kurie žymenis yra svarbesni (ar iš viso svarbūs). Šitoje situacijoje pasireiškia tiesinių klasifikacijos metodų interpretacijos problema, aptarta teoriniame įvade (žr. 1.3. skyrių). Dar vienas niuansas su principinės komponentų analizės pritaikymu žymenų išgavimui yra tas, kad klasifikacijos modelių tikslumas, kartu ir su tais modeliais išgautų žymenų tikslumu, priklauso nuo pasirinkto principinių komponentų skaičiaus. Šiame tyrime žymenų tikslumo priklausomybė nuo principinių komponentų skaičiaus nebuvo tiriama, PCA buvo naudojama su pastoviu komponentų skaičiumi (15).

Toliau bus aptariami šiame darbe nagrinėti spektrų pirminio apdorojimo metodai ir jų įtaka spektrinių žymenų išgavimui. Pirma galima pastebėti, kad spektriniai žymenis, gauti iš spektrų be jokio apdorojimo, turėjo blogą stabilumą ir buvo linkę turėti daug atsitiktinių ir nereikšmingų pataikymų (pavyzdžiui, ties CO_2 spektrinėmis juostomis). Spektrų apkarpymas iki pirštų antspaudo srities reikšmingai pagerino žymenų stabilumą ir padėjo išvengti pataikymų nereikšmingose srityse. Be to, išskyrus kelias išimtis, visi spektrai su apkarpymu tarpusavyje turėjo bent patenkinamą panašumą (indekso vertė $>0,80$), kas indikuoja bendrą patikimumą. Vienas niuansas būtent su FFS metodu yra tas, kad apkarpymas padidino jo riziką persimokyti. Bazinės linijos korekcijos pritaikymas padėdavo sumažinti blogų pataikymų kiekį, tačiau jo poveikis žymenų stabilumui ir panašumui buvo

nevienareikšmis. FFS žymenims bazinės linijos korekcija padidino stabilumą (turint omeny atvejus be tolesnio apdorojimo), o PCA-LDC žymenims – sumažino (žr. 7 lentelę). Tačiau FFS žymenys su apkarpymu ir bazinės linijos korekcija taip pat turėjo stebėtinai mažą panašumą su kitaip apdorotų spektrų žymenimis (žr. 4 lentelę), o su PCA-LDC panašumas buvo puikus (žr. 5 lentelę). Prie bazinės linijos korekcijos pridėjus vektorinį normavimą gaunami žymenys tapo stabilesni, tačiau jų fizikinė interpretacija buvo nenuosekli. Be to, FFS algoritmas daug dažniau persimokė su spektrais, apdorotais vektoriniu normavimu, kas kartu reiškia, kad jo teigiama įtaka stabilumui gali būti pervertinta. Kaip minėta aptariant FFS metodo rezultatus, spektrų normavimas gali būti nesusijęs tiesiogiai su FFS persimokymu. Spektrų diferencijavimas pasirodė gerai – iš spektrų antrų išvestinių išgauti žymenys turėjo ir gerą interpretaciją, ir buvo stabiliausi iš visų tirtų apdorojimo metodų. Be to, naudojant FFS metodą antrų išvestinių žymenys išvengė persimokymo. Tiesa, kai kuriais atvejais su antros eilės diferencijavimu ir vektoriniu normavimu pasireiškė FFS persimokymas, tačiau šiems atvejams taip pat buvo būdingas labai aukštas stabilumas (beveik jokių žemo dažnio pataikymų, pavyzdžiui, kaip 9 pav. su apk. + 2 išv. + vekt. n.). EMSC apdorojimas taip pat pasirodė neblogai: žymenys iš EMSC spektrų buvo vieni stabiliausių ir turėjo gerą interpretaciją, tačiau turėjo tą pačią persimokymo problemą kaip ir vektorinis normavimas. Šiame tyrime dėl geresnės spektrinių žymenų interpretacijos EMSC apdorojimo metodas laikomas pranašesniu už bazinės linijos korekciją + vektorinį normavimą. Standartizavimas neturėjo įtakos FFS žymenims (gaunami vienodi žymenys, kaip ir be standartizavimo). PCA-LDC žymenims standartizavimas daugeliu atveju tik sumažino stabilumą, be to, PCA-LDC žymenys iš standartizuotų spektrų taip pat buvo mažiausiai panašūs į kitus žymenis (žr. 5 lentelę). Tai indikuoja, kad spektrinių žymenų identifikavimui spektrų standartizavimas yra nereikalingas bent jau naudojant žymenų išgavimo metodus, paremtus tiesinio diskriminanto klasifikacija.

Galiausiai reikia paminėti šio tyrimo apribojimus. Pirma, šiame tyrime buvo nagrinėjami tik du spektrinių žymenų išgavimo metodai: vienas, paremtas tiesiniais klasifikacijos modeliais (principinė komponentų analizė + tiesinio diskriminanto klasifikacija), ir vienas, paremtas požymių atrinkimu (išankstinis požymių atrinkimas naudojant tiesinio diskriminanto klasifikaciją). Antra, tiriant FT-IR spektrų pirminio apdorojimo įtaką gautiems žymenims buvo išbandytas ribotas kiekis apdorojimo operacijų ir jų kombinacijų. Pavyzdžiui, šiame tyrime buvo nagrinėjami tik du spektrų normavimo metodai: vektorinis normavimas ir EMSC, o kiti populiarūs normavimo metodai, kaip normavimas pagal amidų I juostą, nebuvo išbandyti. Skirtingų bazinės linijos korekcijos metodų (tiesinės, polinominės, „rubberband“) tyrimas taip pat buvo ribotas. Trečia, šiame tyrime spektrinių žymenų išgavimas buvo patikrintas tik su keturiais spektrinių duomenų atvejais, iš kurių trys buvo iš to paties

spektrų rinkinio. Tolesni spektrinių žymenų tyrimai galėtų išplėsti šio darbo rezultatus su platesniu žymenų išgavimo ir spektrų apdorojimo metodų pasirinkimu, taip pat patikrinti, ar šio darbo išvados galioja ir su kitokių biologinių objektų FT-IR spektrų rinkiniais.

Išvados

Apibendrinant, šiame darbe buvo tiriami biologinių bandinių Furjė transformacijos infraraudonosios sugerties (FT-IR) spektrų spektrinių žymenų išgavimo metodų patikimumas ir jo priklausomybė nuo spektrų pirminio apdorojimo. Tyrimo metu dviejų biologinių objektų – ekstraląstelių vezikulių ir patologinių mikroorganizmų - spektrų rinkiniai buvo apdorojami skirtingomis pirminio apdorojimo operacijomis, ir tuomet iš gautų apdorotų spektrų buvo išgauti spektriniai žymenys. Tyrime buvo išbandyti du spektrinių žymenų išgavimo algoritmai: išankstinis požymių atrinkimas (angl. *forward feature selection*, FFS) ir principinių komponentų analizė – tiesinio diskriminanto klasifikacija (angl. *principal component analysis – linear discriminant classifier*, PCA-LDC). Atsižvelgiant į gautus rezultatus, buvo padarytos šios išvados:

1. Išgavus spektrinius žymenis iš FT-IR spektrų rinkinių, apdorotų skirtingais metodais ir jų kombinacijomis pastebėta, kad spektrų pirminis apdorojimas turi didelę įtaką spektrinių žymenų stabilumui ir fizikinei interpretacijai.
2. Palyginus spektrinius žymenis, gautus su pilnais ($400 - 4000\text{ cm}^{-1}$ diapazonas) ir apkarpytais iki „pirštų antspaudų srities“ ($600 - 1700\text{ cm}^{-1}$ diapazonas) FT-IR spektrais nustatyta, kad spektrų apkarpymas iki pirštų antspaudų srities reikšmingai pagerina žymenų stabilumą ir panašumą, taip pat padeda išvengti biochemiškai nereikšmingų žymenų.
3. Peržiūrint žymenų išgavimo rezultatus FFS algoritmui pastebėta, kad spektrų normavimo metodai skatino jo persimokymą ir ankstyvą konvergavimą. Spėjama, kad ši problema yra labiau susijusi su FFS metodo ypatybėmis, ir ją išsprendus spektrų normavimas pagerintų gautų žymenų stabilumą.
4. Palyginus spektrinius žymenis, gautus PCA-LDC ir FFS algoritmais, buvo nustatyta, kad PCA-LDC žymenys buvo stabilesni už FFS žymenis ir neturėjo persimokymo problemos, tačiau FFS žymenys buvo vidutiniškai labiau koncentruoti, taigi ir aiškiau interpretuojami. Dėl šios priežasties FFS žymenų išgavimo metodas laikomas pranašesniu už PCA-LDC.
5. Išanalizavus visų tyrime išbandytų FT-IR spektrų apdorojimo procedūrų poveikį gautų spektrinių žymenų stabilumui ir fizikinei interpretacijai laikoma, kad geriausia spektrų apdorojimo procedūra spektrinių žymenų išgavimui yra spektrų apkarpymas iki $1700 - 600\text{ cm}^{-1}$ srities kartu su antros eilės diferencijavimu.

Padėka

Norėčiau padėkoti Vilniaus universiteto Fizikos fakulteto Cheminės fizikos instituto doktorantei Gerdai Anužienei ir Vilniaus universiteto Fizikos fakulteto Cheminės fizikos instituto mokslo darbuotojai dr. Rimantei Bandzevičiūtei už leidimus naudotis jų parengtais FT-IR spektrų rinkiniais.

Priedas

8 lentelė. Darbe tirtų FFS algoritmo kriterijaus funkcijų L apibrėžimai. n yra imties dydis; y_j – j -ojo išmatavimo klasė, dvejų klasių atveju programiškai koduojama kaip -1 arba 1; m_j – klasifikacijos įvertis, LDC atveju $m_j = y_j g(\mathbf{x}_j)$, $g(\mathbf{x}_j)$ gaunamas pagal (7), jei $g(\mathbf{x}_j) > 0$, tai $y_j = 1$; w_j – svoriai, priklausantys nuo klasių pradinio pasiskirstymo imtyje; c_j – nustatyta klasifikacijos kaina (pagal nutylėjimą 0, kai klasifikavimas teisingas, ir 1, kai klasifikavimas neteisingas); $I_j = 0$, kai klasifikavimas teisingas, arba $I_j = 1$, kai klasifikavimas neteisingas [21].

Kriterijaus funkcijos pavadinimas	Apibrėžimas
<i>binodeviance</i>	$L = \sum_{j=1}^n w_j \ln\{1 + \exp[-2m_j]\}$
<i>classifcost</i>	$L = \sum_{j=1}^n w_j c_j$
<i>classiferror</i>	$L = \sum_{j=1}^n w_j I_j$
<i>exponential</i>	$L = \sum_{j=1}^n w_j \exp[-m_j]$
<i>hinge</i>	$L = \sum_{j=1}^n w_j \max\{0, 1 - m_j\}$
<i>logit</i>	$L = \sum_{j=1}^n w_j \ln\{1 + \exp[-m_j]\}$
<i>quadratic</i>	$L = \sum_{j=1}^n w_j (1 - m_j)^2$

Literatūros sąrašas

- [1] P. C. Groß and M. Zeppezauer, “Infrared spectroscopy for biopharmaceutical protein analysis,” *J. Pharm. Biomed. Anal.*, vol. 53, no. 1, pp. 29–36, Sep. 2010, doi: 10.1016/J.JPBA.2010.03.009.
- [2] A. G. Theakstone *et al.*, “Fourier-transform infrared spectroscopy of biofluids: A practical approach,” *Transl. Biophotonics*, vol. 3, no. 2, p. e202000025, May 2021, doi: 10.1002/TBIO.202000025.
- [3] H. J. Butler, M. R. McAinsh, S. Adams, and F. L. Martin, “Application of vibrational spectroscopy techniques to non-destructively monitor plant health and development,” *Analytical Methods*, vol. 7, no. 10, pp. 4059–4070, May 2015, doi: 10.1039/C5AY00377F.
- [4] M. Pilling and P. Gardner, “Fundamental developments in infrared spectroscopic imaging for biomedical applications,” *Chem. Soc. Rev.*, vol. 45, no. 7, pp. 1935–1957, Mar. 2016, doi: 10.1039/C5CS00846H.
- [5] G. Bellisola and C. Sorio, “Infrared spectroscopy and microscopy in cancer research and diagnosis,” *Am. J. Cancer Res.*, vol. 2, no. 1, p. 1, 2011, Accessed: Apr. 17, 2025. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC3236568/>
- [6] F. Gómez-De-Anda *et al.*, “Determination of *Trichinella spiralis* in pig muscles using Mid-Fourier Transform Infrared Spectroscopy (MID-FTIR) with Attenuated Total Reflectance (ATR) and Soft Independent Modeling of Class Analogy (SIMCA),” *Meat Sci.*, vol. 91, no. 3, pp. 240–246, Jul. 2012, doi: 10.1016/J.MEATSCI.2012.01.019.
- [7] J. Trevisan *et al.*, “Measuring similarity and improving stability in biomarker identification methods applied to Fourier-transform infrared (FTIR) spectroscopy,” *J. Biophotonics*, vol. 7, no. 3–4, pp. 254–265, Apr. 2014, doi: 10.1002/jbio.201300190.
- [8] M. M. Jansson, M. Kögler, S. Hörkkö, T. Ala-Kokko, and L. Rieppo, “Vibrational spectroscopy and its future applications in microbiology,” *Appl. Spectrosc. Rev.*, vol. 58, no. 2, pp. 132–158, 2023, doi: 10.1080/05704928.2021.1942894.
- [9] K. B. Beć, J. Grabska, and C. W. Huck, “Biomolecular and bioanalytical applications of infrared spectroscopy – A review,” *Anal. Chim. Acta*, vol. 1133, pp. 150–177, Oct. 2020, doi: 10.1016/J.ACA.2020.04.015.
- [10] D. Usoltsev, V. Sitnikova, A. Kajava, and M. Uspenskaya, “Systematic FTIR Spectroscopy Study of the Secondary Structure Changes in Human Serum Albumin under Various Denaturation Conditions,” *Biomolecules 2019, Vol. 9, Page 359*, vol. 9, no. 8, p. 359, Aug. 2019, doi: 10.3390/BIOM9080359.
- [11] M. J. Baker *et al.*, “Using Fourier transform IR spectroscopy to analyze biological materials,” 2014, doi: 10.1038/nprot.2014.110.
- [12] J. Trevisan, P. P. Angelov, P. L. Carmichael, A. D. Scott, and F. L. Martin, “Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives,” *Analyst*, vol. 137, no. 14, pp. 3202–3215, Jun. 2012, doi: 10.1039/c2an16300d.
- [13] J. H. Solheim, B. Zimmermann, V. Tafintseva, S. Dzurendová, V. Shapaval, and A. Kohler, “The Use of Constituent Spectra and Weighting in Extended Multiplicative Signal Correction in Infrared Spectroscopy,” *Molecules*, vol. 27, no. 6, p. 1900, Mar. 2022, doi: 10.3390/MOLECULES27061900/S1.

- [14] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, “Robust biomarker identification for cancer diagnosis with ensemble feature selection methods,” *Bioinformatics*, vol. 26, no. 3, pp. 392–398, Feb. 2010, doi: 10.1093/BIOINFORMATICS/BTP630.
- [15] R. L. Somorjai, “Creating robust, reliable, clinically relevant classifiers from spectroscopic data,” *Biophysical Reviews* 2009 1:4, vol. 1, no. 4, pp. 201–211, Nov. 2009, doi: 10.1007/S12551-009-0023-6.
- [16] V. Bagdonavičius, J. J. Kruopis, MATEMATINĖ STATISTIKA IV, 2015, Vilniaus universiteto leidykla, 230 p., ISBN 978-609-459-518-9.
- [17] C. Beleites and R. Salzer, “Assessing and improving the stability of chemometric models in small sample size situations,” *Analytical and Bioanalytical Chemistry* 2007 390:5, vol. 390, no. 5, pp. 1261–1271, Jan. 2008, doi: 10.1007/S00216-007-1818-6.
- [18] J. Ullo Trevisan *et al.*, “Syrian hamster embryo (SHE) assay (pH 6.7) coupled with infrared spectroscopy and chemometrics towards toxicological assessment †”, doi: 10.1039/c0an00586j.
- [19] H.-Ulrich. Gremlich and Bing. Yan, “Infrared and raman spectroscopy of biological materials,” p. 581, 2001, Accessed: Jan. 09, 2026. [Online]. Available: https://books.google.com/books/about/Infrared_and_Raman_Spectroscopy_of_Biolo.html?hl=lt&id=vYyxH-hgMzwC
- [20] J. Trevisan, P. P. Angelov, A. D. Scott, P. L. Carmichael, and F. L. Martin, “IRootLab: a free and open-source MATLAB toolbox for vibrational biospectroscopy data analysis,” *Bioinformatics*, vol. 29, no. 8, pp. 1095–1097, Apr. 2013, doi: 10.1093/BIOINFORMATICS/BTT084.
- [21] “Documentation - MATLAB & Simulink”
<https://se.mathworks.com/help/releases/R2025b/stats/classificationdiscriminant.loss.html#bs2es1z-2>.
- [22] L. I. Kuncheva, Proceedings of the 25th IASTED on Artificial Intelligence and Applications, Innsbruck, Austria, February 2007, pp. 390–395.

Identification of spectral markers in Fourier transform infrared spectra of biological samples by means of statistical analysis

Tautvydas Taraškevičius

Summary

Fourier transform infrared (FT-IR) spectroscopy is a versatile, non-destructive and label-free vibrational spectroscopy method with great potential in biomedical research and diagnosis. One of the advantages of this method is that many data analysis methods used in spectral analysis provide the means to detect the most important spectral regions that relate to biological mechanisms under investigation. However, the validity of these spectral markers is not well understood because of the complexity of vibrational spectra of biological samples, as well as the fact that usually there are no reference biomarkers to compare to. Furthermore, the effect spectral preprocessing has on spectral marker extraction has not yet been investigated. The aim of this study is to evaluate the validity of FTIR spectral biomarkers obtained by various statistical and machine learning methods and how they depend on spectral preprocessing. For this, two methods of spectral marker identification were tested: forward feature selection (FFS) wrapped with a linear discriminant classification model and extraction of loadings from a principal component analysis - linear discriminant classification (PCA-LDC) model. Two experimental FT-IR spectral datasets were used as examples, one containing spectra from pathological microorganisms and the another containing spectra from extracellular vesicles extracted from human blood of healthy individuals and myocardial infarction patients. Spectral markers extracted from spectra preprocessed by various means were compared and evaluated by their similarity, stability and spectral interpretation. The results show that spectral preprocessing has a major impact on spectral marker extraction results. Despite poorer stability and similarity, the FFS method is considered overall more reliable than PCA-LDC on the grounds of clearer spectral interpretation. The best FT-IR spectral preprocessing procedure for spectral marker extraction is believed to be spectral truncation to the fingerprint region + second order differentiation.