COMPUTATIONAL SCIENCE AND TECHNIQUES
Volume 1, Number 2, 2013, 153-166

Publisher: Klaipėda University
http://journals.ku.lt/index.php/CST
Online ISSN: 2029-9966

GENOTIPO IR FENOTIPO DUOMENŲ ANALIZĖ IR VIZUALIZAVIMAS

Alma Molytė, Vaidutis Kučinskas, Aušra Matulevičienė, Eglė Preikšaitienė
Žmogaus ir medicininės genetikos katedra, Medicinos fakultetas,
Vilniaus universitetas, Universiteto g. 3, LT-01513, Vilnius,
alma.molyte@mf.vu.lt, vaidutis.kucinskas@mf.vu.lt, ausra.matuleviciene@mf.vu.lt,
egle.preiksaitiene@mf.vu.lt

Anotacija. Darbe pateikiama genotipo ir fenotipo duomenų analizei taikomų hierarchinio klasterizavimo ir daugiamačių skalių metodų lyginamoji analizė. Priklausomybės tarp įgimtų anomalijų nustatymui taikytas tikslusis Fišerio kriterijus. Įgimtų raidos anomalijų, deformacijų bei mikroanomalijų susijusių su veido nesuaugimais tarpusavio ryšių nustatymui, taikytas Spirmeno ir Kendalo koreliacijos koeficientai. Išsiaiškinta, kuris iš metodų yra tinkamesnis genetinių duomenų atvaizdavimui.

Pagrindiniai žodžiai: genotipo ir fenotipo duomenys, Spirmeno ir Kendalo koreliacijos koeficientai, hierarchinis klasterizavimas, duomenų vizualizavimas.

Įvadas

Dabartinės technologijos leidžia kaupti didelius duomenų masyvus, tai ypatingai svarbu genominių duomenų analizei, kadangi genetinėje medžiagoje saugomi milžiniški informacijos kiekiai. Kiekvienam gyvam organizmui yra būdingas specifinis genomas, koduojantis biologinę informaciją, reikalingą suformuoti ir palaikyti gyvybines funkcijas. Absoliuti dauguma genomų yra sudaryta iš deoksiribonukleorūgšties (DNR) (Kučinskas, 2012).

Klinikinėje genetikoje be genomo, svarbi ir fenomo informacija. Fenomas – ląstelės, organo, audinio bei viso organizmo fenotipinių požymių visuma. Tiriamiesiems su įvairiais veido nesuaugimais, kuriems neidentifikuotos jų atsiradimo priežastys, naudojant įvairius statistinius fenotipo duomenų analizės metodus svarbu nustatyti vyraujančias kitas kartu esančias įgimtas raidos anomalijas bei jų tarpusavio ryšį su veido nesuaugimais, palyginti šiuos rezultatus tarpusavyje, suskirstant tiriamuosius į mažesnes grupes pagal veido nesuaugimo tipą, o nustačius glaudžią kitų įgimtų raidos anomalijų priklausomybę tarp jų ir veido nesuaugimų šiuos duomenis pritaikyti klinikinėje genetikoje (Matulevičienė ir kt., 2013). Tai svarbu tiriant tokius asmenis bei parenkant jiems tikslesnius instrumentinius tyrimo metodus, tokius kaip rentgenas, kompiuterinė tomografija, magnetinis rezonansas ir kiti. Kuo tiksliau įvertinti fenotipo duomenys, tuo kryptingiau galima būtų pritaikyti genetinius tyrimo metodus, tiriamojo genotipui įvertinti, kas ateityje leistų identifikuoti veido nesuaugimų priežastis didesnei tiriamųjų grupei.

Analizuojant tiriamųjų su intelektine negalia (INN) genominius (molekulinio kariotipavimo) duomenis ir fenotipinius požymius, tikimasi pagerinti pacientų su INN atranką molekulinio kariotipavimo tyrimui (Preikšaitienė ir kt., 2012). Rekomenduojama,

tiriamiesiems su kognityviniais sutrikimais, atlikti pirminį molekulinio kariotipavimo tyrimą, tačiau jis yra brangus ir ne visos laboratorijos gali jį atlikti visiems tiriamiesiems su kognityviniais sutrikimais ar įgimtomis anomalijomis. Tikėtina, kad diagnostinį efektyvumą padidintų tikslinga tokių tiriamųjų atranka tyrimui. Nors žinoma, kad submikroskopiniai pokyčiai chromosomose dažnai susiję su INN, tačiau nėra pakankamai duomenų, ar jie lemia specifinius fenotipinius (arba fenomo) ypatumus. Tiriamųjų su INN bei chromosominiais pokyčiais fenomo detalizavimas, lyginant juos su tiriamaisiais be chromosominių pokyčių, leistų apibrėžti klinikinius kriterijus, padedančius atpažinti tiriamuosius su galimais patogeniniais chromosominiais pakitimais.

Genetinių duomenų analizei yra sukurta daug įvairių metodų: statistinės analizės, klasifikavimo, klasterizavimo, vizualizavimo bei kt. Jais galima nustatyti duomenų ar jų grupių artimumus, panašumus, vertinti atskirų parametrų (požymių) įtaką daromam sprendimui.

Šiame darbe tiriamos genotipo ir fenotipo duomenų priklausomybės ir jų vizualizavimas. Tiriamųjų su patogeniniais chromosominiais pokyčiais klinikiniai duomenys palyginti su tiriamųjų be patogeninių chromosominių pokyčių. Nagrinėtos įgimtos raidos anomalijos, susijusios su veido srities nesuaugimais bei šių raidos anomalijų tarpusavio ryšys su lūpos ir (arba) gomurio nesuaugimu.

1. Kintamųjų ryšių matai ir χ^2 kriterijus

Dažnai reikia išsiaiškinti, ar stebimi kintamieji yra nepriklausomi, ar priklausomi. Kokybiniams kintamiesiems taikome χ^2 kriterijų. Esant teisingai hipotezei H_0 statistika χ^2 yra asimptotiškai pasiskirsčiusi pagal χ^2 dėsnį su $(r-1)(c-1)$ laisvės laipsnių. Hipotezė H_0 apie kintamųjų nepriklausomumą yra atmetama, kai apskaičiuotos statistikos χ^2 reikšmė yra didesnė už χ^2 skirstinio su $(r-1)(c-1)$ laisvės laipsnių α lygmens kritinę reikšmę, čia α – pasirinktas reikšmingumo lygmuo, r – eilučių skaičius, c – stulpelių skaičius (Čekanavičius, 2006).

Jeigu apskaičiuoti tikėtini dažniai yra maži (< 5), tai vietoje χ^2 kriterijaus naudojamas Fišerio kriterijus, kuris yra konservatyvus – atmeta hipotezę tik esant dideliems skirtumams (Čekanavičius, 2006).

Šiame darbe naudojamas Fišerio tikslusis kriterijus, o reikšmingumo lygmuo α yra 0,05 ir 0,01.

1.1. Spirmeno koreliacijos koeficientas

Dauguma genetinių duomenų yra kokybiniai, todėl norint nustatyti ryšio stiprumą tarp analizuojamų kintamųjų, skaičiuojamas Spirmeno ir Kendalo koreliacijos koeficientai.

Tarkime, jog kintamųjų poros (X, Y) stebėjimai yra poros $(x_i, y_i), \dots, (x_n, y_n)$. Spirmeno koreliacijos koeficientą skaičiuojame, kai duomenys netenkina normalumo prielaidos arba duomenų mažai (< 20 stebėjimų). Iš pradžių duomenys ranguojami. Po rangavimo duomenis sudaro poros $(R_{x1}, R_{y1}), \dots, (R_{xn}, R_{yn})$.

Spirmeno koreliacijos koeficientas apskaičiuojamas pagal formulę (Chok, 2010):

$$r_s = \frac{\sum_{i=1}^{n}\left(R_{xi}-\frac{n+1}{2}\right)\left(R_{yi}-\frac{n+1}{2}\right)}{\sqrt{\sum_{i=1}^{n}\left(R_{xi}-\frac{n+1}{2}\right)^2}\sqrt{\sum_{i=1}^{n}\left(R_{yi}-\frac{n+1}{2}\right)^2}},$$

čia R_{xi} yra x_i rangas, o R_{yi} – y_i rangas. Tarp kintamųjų priklausomybė yra stipresnė, kai koeficientas absoliučiuoju didumu yra didesnis.

1.2. Kendalo koreliacijos koeficientas

Kendalo koreliacijos koeficientas, kaip ir Spirmeno, naudojamas ranginių kintamųjų ryšio stiprumui įvertinti. Koeficientas apskaičiuojamas pagal formulę (Chok, 2010):

$$\tau = \frac{S}{\frac{n(n-1)}{2}},$$

čia S – suderintų, nesuderintų porų skaičiaus skirtumas, n yra imties didumas. Dvi duomenų poros (x_i, y_i) ir (x_j, y_j), $(i \neq j)$yra suderintos tada, kai $(x_i > x_j$ ir $y_i > y_j)$ arba $(x_i < x_j$ ir $y_i < y_j)$. Poros nesuderintos, kai $(x_i > x_j$ ir $y_i < y_j)$ arba $(x_i < x_j$ ir $y_i > y_j)$. Jeigu visos poros suderintos, tai $\tau = 1$, jei visos poros nesuderintos, tai $\tau = -1$.

2. Genotipo ir fenotipo duomenų klasterizavimas ir vizualizavimas

2.1. Hierarchinio klasterizavimo metodai

Klasterizavimas – tai analizuojamų objektų suskirstymas į skirtingas grupes (klasterius) taip, kad grupės viduje esantys objektai būtų panašūs tarpusavyje, o objektai iš skirtingų grupių būtų nepanašūs. Klasterizavimas plačiai taikomas augalų ir gyvūnų klasifikacijoje, ligų klasifikacijoje, vaizdų apdorojime, atpažinime ir kt. Klasterizavimo metodai gali būti suskirstyti į: hierarchinius, padalijimo metodus bei tankiu, tinkleliu ir modeliu pagrįstais metodais (Kurasova, 2005).

Hierarchiniuose klasterizavimo metoduose du maži klasteriai jungiami į vieną didesnį arba didelis klasteris skaidomas į kelis mažesnius. Metodai skiriasi vienas nuo kito pagal taisykles, kuriomis remiantis nusprendžiama, kurie du maži klasteriai yra sujungiami arba keli dideli klasteriai suskaidomi (Dash, 2003).

Hierarchinio klasterizavimo rezultatai gali būti pateikti dendrograma, kuri parodo kaip arti duomenų elementai yra vieni nuo kitų. Pradžioje apskaičiuojami atstumai tarp analizuojamų duomenų objektų pagal pasirinktą matą: Čebyševo, Minkovskio, Manhatino, Euklido ar kt. (Dzemyda ir kt., 2008). Išrenkami objektai, tarp kurių atstumai yra mažiausi. Jie sudarys vieną klasterį. Kitame žingsnyje perskaičiuojami atstumai nuo gauto klasterio iki kitų objektų ir išrenkami objektai arba klasteriai tarp kurių atstumai yra mažiausi. Jie sudarys naują klasterį arba objektas bus priskirtas jau egzistuojančiam klasteriui. Klasterių sujungimui naudojamas vienas iš metodų: vienetinės jungties („artimiausio kaimyno“), pilnos jungties („tolimiausio kaimyno“), vidutinės jungties, vordo (*angl.* ward‘s) bei centrų metodas.

Šiame darbe taikomas hierarchinio klasterizavimo metodas, kuriuo gauti rezultatai pateikti dendrograma. Klasterių sujungimui taikomi vienetinės, pilnosios ir vidutinės jungties metodai.

2.2. Daugiamačių skalių metodas

Daugiamačių duomenų vizualizavimo metodais didelės dimensijos duomenys transformuojami į mažesnės dimensijos erdvę taip, kad išliktų arba būtų atrastos „užslėptos" analizuojamų duomenų savybės.

Naudojantis daugiamačių skalių (*angl.* multidimensional scaling, MDS) metodu, ieškoma daugiamačių duomenų projekcijos mažesnio skaičiaus matmenų erdvėje (dažniausiai R^2, R^3), siekiant išlaikyti analizuojamos aibės objektų artimumus – panašumus arba skirtingumus (Borg, Groenen, 2005). Vienas daugiamačių skalių metodų tikslų yra rasti optimalų daugiamačius objektus atitinkančių taškų vaizdą mažo skaičiaus matmenų erdvėje (Dzemyda ir kt, 2008).

Tarkime kiekvieną n-matį vektorių $X_i \in R^n$, $i \in \{1, \dots, m\}$, atitinka mažesnio dimensijų skaičiaus vektorius $Y_i \in R^d$, $d < n$. Artumą (panašumą arba skirtingumą) tarp n-mačių vektorių X_i ir X_j pažymėkime $\delta(X_i, X_j)$, o atstumą tarp dvimačių vektorių Y_i ir Y_j – $d(Y_i, Y_j)$, $i, j = 1, \dots m$. Jeigu artumas yra Euklido atstumas, tai $\delta(X_i, X_j) = d(X_i, X_j)$.

Naudojantis MDS algoritmu, bandoma atstumą $d(Y_i, Y_j)$ priartinti prie atstumo $d(X_i, X_j)$. Jei naudojama kvadratinė paklaidos funkcija, tai minimizuojama tikslo funkcija E_{MDS} gali būti užrašyta taip: $E_{MDS} = \sum_{i<j} w_{ij} \left(\delta(X_i, X_j) - d(Y_i, Y_j)\right)^2$. Paklaidos funkcija E_{MDS} dar vadinama Strees funkcija, kurios reikšmė parodo, kaip tiksliai modelis atitinka pradinius duomenis. Taškų koordinatės dvimatėje erdvėje transformuojamos taip, kad Stress funkcijos reikšmė būtų minimali. Iteracinis procesas kartojamas tol, kol funkcijos Stress reikšmę gauname mažesnę už nurodytą. Rekomenduojama, kad Stress reikšmė būtų mažesnė už 0,15, o geriausia kai ji neviršija 0,1.

Šiame darbe naudojamas daugiamačių skalių PREFSCAL algoritmas (Borg, Groenen, 2005).

3. Tyrimų rezultatai

Tyrimuose naudotos duomenų aibės, turinčios tam tikrų specifinių savybių:

1. Asmenų su INN duomenys [211; 17]. Tiriamieji buvo su izoliuota arba sindromine INN. Esant sindrominei INN be kognityvinių funkcijų sutrikimų stebėti fiziniai pokyčiai (dismorfiniai požymiai ir / arba įgimtos anomalijos), neurologiniai ar elgesio sutrikimai. Nesindrominės INN vienintelis patologinis požymis – INN.
2. Veido nesuaugimo (VN) duomenys [142; 10]. Tiriamieji su įvairiais veido srities nesuaugimais (vienpusis lūpos ir gomurio nesuaugimas, abipusis lūpos ir gomurio nesuaugimas, vienpusis lūpos nesuaugimas be gomurio nesuaugimo, abipusis lūpos nesuaugimas be gomurio nesuaugimo ir gomurio nesuaugimas) ir kitomis įgimtomis raidos anomalijomis (tiek didžiosiomis, tiek mažosiomis), kurių atsiradimo priežastys yra neidentifikuotos.

Norint įvertinti struktūrinių pokyčių chromosomose reikšmę asmenų su INN klinikiniams požymiams, buvo palygintos dvi grupės tiriamųjų, kurių vieniems molekulinio kariotipavimo tyrimu nustatytas chromosominis pokytis, o kitiems – patogeninių pakitimų

nenustatyta. Tarp visų galimų įgimtų anomalijų porų priklausomybės nustatymui buvo skaičiuojamas Fišerio tikslusis kriterijus.

Analizuotos mikroanomalijos bei įgimtos formavimosi ydos ir deformacijos: įgimtos nervų sistemos formavimosi ydos (Q00-Q07CA), įgimtos akies, ausies, veido ir kaklo formavimosi ydos (Q10-Q18CA), įgimtos kraujo apytakos sistemos formavimosi ydos (Q20-Q28CA), įgimtos kvėpavimo sistemos formavimosi ydos (Q30-Q34CA), kitos įgimtos virškinimo sistemos formavimosi ydos (Q38-Q45CA), įgimtos lyties organų formavimosi ydos (Q50-Q56CA), įgimtos šlapimo organų sistemos formavimosi ydos (Q60-Q64CA) ir kitos įgimtos formavimosi ydos (Q80-Q89CA), kurios pateiktos TLK-10–AM klasifikacijoje; mikroanomalijos pateiktos EUROCAT registre (Viso_kituMA) bei fenotipiniai ypatumai (Fen_ypat), kurie neįtraukti nei į TLK-10–AM Q.00–Q.99 skyrių „Įgimtos formavimosi ydos, deformacijos ir chromosomų anomalijos“, nei į mikroanomalijų sąrašą, tačiau pažymėti konsultavusio gydytojo genetiko vertinant fenotipą, pvz. žema kakta. Anomalijos, kurios nustatytos tiriamiesiems kartu su veido nesuaugimais buvo koduotos pagal Britų pediatrų asociacijos (*angl.* British Pediatric Association –BPA) sudarytą klasifikaciją: virškinimo organų sistemos anomalijos (BPA code 750-751), šlapimo organų sistemos anomalijos (BPA code 753), akių anomalijos (BPA code 743), centrinės nervų sistemos anomalijos (BPA code 740-742), ausų, veido ir kaklo anomalijos (BPA code 744), kvėpavimo organų sistemos anomalijos (BPA code 748), lyties organų anomalijos (BPA code 752), širdies-kraujagyslių sistemos anomalijos (BPA code 745-747), galūnių ir skeleto anomalijos (BPA code 754-756) ir nespecifinės anomalijos.

Tiriamiesiems (INN) su patogeniniais chromosominiais pokyčiais nustatyta priklausomybė tarp įgimtų akies, ausies, veido ir kaklo formavimosi ydų ir įgimtų raumenų ir skeleto sistemos formavimosi ydų bei deformacijų ($p = 0{,}010$), įgimtų raumenų ir skeleto sistemos mikroanomalijų ir įgimtų kraujo apytakos sistemos formavimosi ydų ($p = 0{,}024$); įgimtų raumenų ir skeleto sistemos mikroanomalijų ir kitų įgimtų virškinimo sistemos mikroanomalijų ($p = 0{,}008$); fenotipinių ypatumų ir įgimtų raumenų ir skeleto sistemos formavimosi ydų bei deformacijų ($p = 0{,}049$).

Tiriamiesiems su intelektine negalia be patogeninių chromosominių pokyčių nustatyta priklausomybė tarp:

- įgimtų nervų sistemos formavimosi ydų ir įgimtų akies, ausies, veido, kaklo mikroanomalijų ($p = 0{,}033$);
- įgimtų akies, ausies, veido ir kaklo formavimosi ydų ir šios sistemos mikroanomalijų ($p = 0{,}047$), įgimtų kraujo apytakos sistemos formavimosi ydų ($p = 0{,}003$), įgimtų kvėpavimo sistemos formavimosi ydų ($p = 0{,}004$), kitų įgimtų virškinimo sistemos mikroanomalijų ($p = 0{,}027$), įgimtų lyties organų formavimosi ydų ($p = 0{,}005$), įgimtų šlapimo organų sistemos formavimosi ydų ($p = 0{,}033$) bei visų kitų mikroanomalijų ($p = 0{,}007$);
- įgimtų akies, ausies, veido ir kaklo mikroanomalijų ir įgimtų kvėpavimo sistemos formavimosi ydų ($p = 0{,}035$), įgimtų lyties organų formavimosi ydų ($p = 0{,}002$), įgimtų raumenų ir skeleto sistemos mikroanomalijų ($p = 0{,}010$) bei visų kitų mikroanomalijų ($p = 0{,}001$);

- įgimtų kraujo apytakos sistemos formavimosi ydų ir kitų įgimtų virškinimo sistemos mikroanomalijų ($p = 0{,}046$), įgimtų šlapimo organų sistemos formavimosi ydų ($p = 0{,}010$) bei visų kitų mikroanomalijų ($p = 0{,}048$);
- įgimtų kvėpavimo sistemos formavimosi ydų ir visų kitų mikroanomalijų ($p = 0{,}045$);
- kitų įgimtų virškinimo sistemos mikroanomalijų ir įgimtų raumenų ir skeleto sistemos formavimosi ydų bei deformacijų ($p = 0{,}004$), visų kitų mikroanomalijų ($p = 0{,}018$);
- įgimtų lyties organų formavimosi ydų ir šios sistemos mikroanomalijų ($p = 0{,}008$), kitų įgimtų formavimosi ydų mikroanomalijų ($p = 0{,}001$), visų kitų mikroanomalijų ($p = 0{,}012$) bei fenotipinių ypatumų ($p = 0{,}044$);
- įgimtų lyties organų mikroanomalijų ir kitų įgimtų formavimosi ydų ($p = 0{,}038$) bei visų kitų mikroanomalijų ($p = 0{,}008$);
- įgimtų šlapimo organų sistemos formavimosi ydų ir visų kitų mikroanomalijų ($p = 0{,}007$);
- įgimtos raumenų ir skeleto sistemos formavimosi ydų bei deformacijų ir kitų įgimtų formavimosi ydų ($p = 0{,}038$) bei fenotipinių ypatumų ($p = 0{,}041$);
- įgimtų raumenų ir skeleto sistemos mikroanomalijų ir visų kitų mikroanomalijų ($p < 0{,}001$) bei fenotipinių ypatumų ($p < 0{,}001$);
- kitų įgimtų formavimosi ydų ir visų kitų mikroanomalijų, kurios neįtrauktos TLK-10–AM ($p = 0{,}007$).

Įgimtų formavimosi ydų ir deformacijų bei mikroanomalijų tarpusavio ryšių nustatymui taikytas Spirmeno ir Kendalo koreliacijos koeficientai (1–2 lentelės). Tiriamiesiems su patogeniniais chromosominiais pokyčiais nustatytas vidutinis statistiškai reikšmingas ryšys tarp įgimtų formavimosi ydų, mikroanomalijų, kurios yra pateiktos 1 lentelėje, o pasikliovimo lygmuo $\alpha = 0{,}05$.

1 lentelė. INN duomenų, kai tiriamiesiems nustatyti patogeniniai chromosominiai pokyčiai, įgimtų formavimosi ydų ir deformacijų bei mikroanomalijų koreliacija.

Nr.	Įgimtų formavimosi ydų ir deformacijų bei mikroanomalijų pavadinimai		Koreliacijos koeficientai				α
			Spirmeno(r_s)	p-reikšmė	Kendalo (r_k)	p-reikšmė	
1	Q10–Q18CA	Q65–Q79CA	0,459	0,014	0,447	0,017	0,05
2	Q20–Q28CA	Q65–Q79MA	0,462	0,013	0,430	0,014	0,05
3	Q20–Q28CA	Viso_kituMA	0,403	0,033	0,363	0,038	0,05
4	Q38–Q45MA	Q65–Q79MA	0,453	0,016	0,418	0,019	0,05
5	Q65–Q79CA	Fen_ypat	0,407	0,031	0,337	0,042	0,05

Iš 2 lentelės matyti, kad egzistuoja vidutinis statistiškai reikšmingas ryšys tarp įgimtų akies, ausies, veido ir kaklo formavimosi ydų ir įgimtų kvėpavimo sistemos formavimosi ydų, kai reikšmingumo lygmuo $\alpha = 0{,}01$. Silpnai, tačiau statistiškai reikšmingai koreliuoja, įgimtos akies, ausies, veido ir kaklo formavimosi ydos su šios sistemos mikroanomalijomis, įgimtomis kraujo apytakos sistemos formavimosi ydomis ir įgimtomis lyties organų

formavimosi ydomis, kai reikšmingumo lygmuo $\alpha = 0{,}01$ bei įgimtomis virškinimo sistemos mikroanomalijomis, įgimtomis raumenų ir skeleto sistemos formavimosi ydomis bei deformacijomis, kai $\alpha = 0{,}05$. Egzistuoja silpnas statistiškai reikšmingas ryšys tarp įgimtų akies, ausies, veido ir kaklo mikroanomalijų ir įgimtų kvėpavimo sistemos formavimosi ydų, tarp įgimtų raumenų ir skeleto sistemos formavimosi ydų bei deformacijų ir šios sistemos mikroanomalijų, tarp įgimtų raumenų ir skeleto sistemos formavimosi ydų bei deformacijų ir kitų įgimtų formavimosi ydų bei fenotipinių ypatumų, kai $\alpha = 0{,}05$. Kitos 2 lentelėje pateiktos įgimtos anomalijos koreliuoja silpnai, kai $\alpha = 0{,}01$.

Analizuojant įgimtas raidos anomalijas susijusias su veido srities nesuaugimais, duomenys (VN) buvo suskirstyti į tris grupes, t.y. gomurio nesuaugimai (GN), lūpos nesuaugimai (LN) bei lūpos ir gomurio nesuaugimai (LGN). Gomurio nesuaugimų grupėje nustatyta priklausomybė tarp šlapimo organų sistemos anomalijų ir virškinimo organų sistemos anomalijų ($p = 0{,}036$), tarp ausų, veido ir kaklo anomalijų ir širdies-kraujagyslių sistemos anomalijų ($p < 0{,}0001$), galūnių ir skeleto anomalijų ($p = 0{,}035$) bei nespecifinių anomalijų ($p = 0{,}035$), širdies-kraujagyslių sistemos anomalijų ir virškinimo organų sistemos anomalijų ($p = 0{,}043$), tarp galūnių ir skeleto anomalijų ir virškinimo organų sistemos anomalijų ($p = 0{,}004$) bei tarp virškinimo organų sistemos anomalijų ir nespecifinių anomalijų ($p = 0{,}005$). Lūpos ir gomurio nesuaugimų grupėje nustatyta priklausomybė tarp centrinės nervų sistemos anomalijų ir akių anomalijų ($p < 0{,}0001$), nespecifinių anomalijų ($p = 0{,}045$), tarp ausų, veido ir kaklo anomalijų ir nespecifinių anomalijų ($p = 0{,}037$) bei tarp širdies-kraujagyslių sistemos anomalijų ir galūnių ir skeleto anomalijų ($p = 0{,}026$). Lūpos nesuaugimų grupėje priklausomybių tarp kitų įgimtų raidos anomalijų nenustatyta.

2 lentelė. INN duomenų, kai tiriamiesiems nenustatyti patogeniniai chromosominiai pokyčiai, įgimtų formavimosi ydų ir deformacijų bei mikroanomalijų koreliacija.

Nr.	Įgimtų formavimosi ydų ir deformacijų bei mikroanomalijų pavadinimai		Koreliacijos koeficientai				α
			Spir-meno(r_s)	p-reikšmė	Kendalo (r_k)	p-reikšmė	
1	**Q10–Q18CA**	**Q10–Q18MA**	0,192	0,009	0,182	0,010	0,01
2	Q10–Q18CA	Q20–Q28CA	0,219	0,003	0,215	0,003	0,01
3	**Q10–Q18CA**	**Q30–Q34CA**	0,424	0,001	0,422	0,001	0,01
4	**Q10–Q18CA**	**Q38–Q45MA**	0,178	0,016	0,177	0,018	0,05
5	**Q10–Q18CA**	**Q50–Q56MA**	0,308	0,001	0,306	0,001	0,01
6	**Q10–Q18CA**	**Q65–Q79CA**	0,153	0,039	0,150	0,039	0,05
7	Q10–18MA	Q30–Q34CA	0,178	0,016	0,170	0,016	0,05
8	Q10–18MA	Q50–Q56MA	0,252	0,001	0,241	0,001	0,01
9	**Q10–18MA**	**Q65–Q79MA**	0,227	0,002	0,199	0,003	0,01
10	**Q10–18MA**	**Viso kituMA**	0,277	0,001	0,251	0,001	0,01
11	**Q20–Q28CA**	**Q60–Q64CA**	0,295	0,001	0,291	0,001	0,01
12	**Q38–Q45MA**	**Q65–Q79CA**	0,224	0,002	0,221	0,002	0,01
13	Q38–Q45MA	Viso kituMA	0,217	0,003	0,205	0,003	0,01
14	**Q50–Q56CA**	**Q80–Q89MA**	0,336	0,001	0,335	0,001	0,01
15	**Q50–Q56MA**	**Q80–Q89MA**	0,197	0,008	0,196	0,008	0,01
16	Q65–Q79CA	Q65–Q79MA	0,154	0,037	0,142	0,038	0,05
17	**Q65–Q79CA**	**Q80–Q89CA**	0,204	0,006	0,201	0,006	0,05

18	Q65–Q79CA	Fen_ypat	0,190	0,010	0,165	0,011	0,05
19	Q65–Q79CA	Viso_kituMA	0,213	0,004	0,190	0,004	0,01
20	Q65–Q79CA	Fen_ypat	0,381	0,001	0,311	0,001	0,01

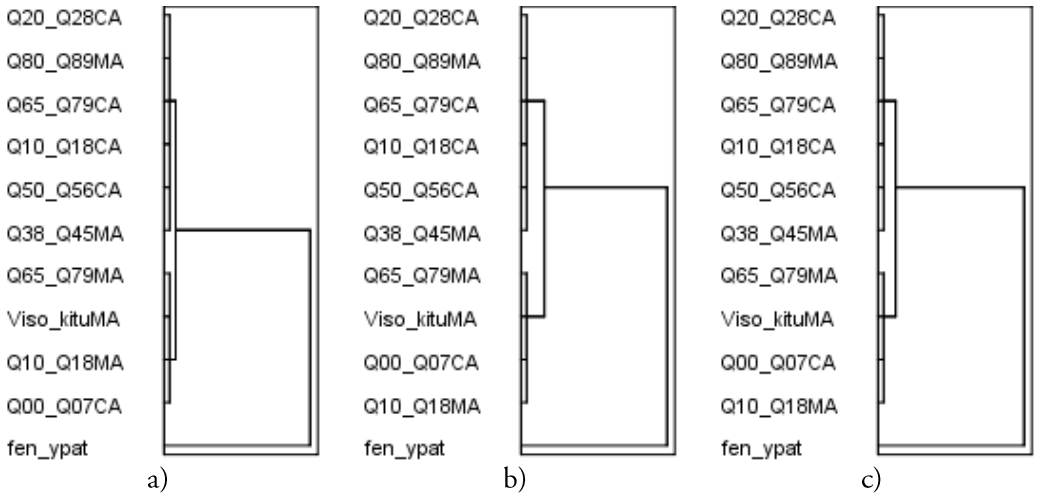
Atlikus koreliacinę analizę įgimtų raidos anomalijų GN grupėje, nustatyta vidutinė koreliacija tarp virškinimo organų sistemos anomalijų bei galūnių ir skeleto anomalijų ir nespecifinių anomalijų, kai $\alpha = 0{,}01$. Silpnas, tačiau statistiškai reikšmingas ryšys egzistuoja tarp virškinimo organų sistemos anomalijų ir akių anomalijų bei širdies-kraujagyslių sistemos anomalijų, kai $\alpha = 0{,}05$ (3 lentelė). LGN grupėje egzistuoja vidutinė koreliacija tarp centrinės nervų sistemos anomalijų ir akių anomalijų, kai reikšmingumo lygmuo $\alpha = 0{,}01$ bei tarp širdies-kraujagyslių sistemos anomalijų ir galūnių ir skeleto anomalijų, kai $\alpha = 0{,}05$.

3 lentelė. Įgimtų raidos anomalijų, susijusių su veido srities nesuaugimais, koreliacija.

Grupė	Įgimtų raidos anomalijų, susijusių su veido srities nesuaugimais, kodai pagal BPA		Koreliacijos koeficientai				α
			Spir-meno (r_s)	p-reikšmė	Kendalo (r_k)	p-reikšmė	
GN	**BPA 750-751**	**BPA 743**	0,343	0,028	0,337	0,030	0,05
	BPA 750-751	BPA 745-747	0,382	0,014	0,360	0,015	0,05
	BPA 750-751	BPA 754-756	0,411	0,008	0,371	0,010	0,01
	BPA 750-751	Nespecifinės	0,453	0,003	0,428	0,004	0,01
LGN	BPA 740-742	BPA 743	0,520	0,001	0,511	0,001	0,01
	BPA 745-747	BPA 754-756	0,235	0,035	0,213	0,036	0,05

3.1 Genotipo ir fenotipo duomenų vizualizavimas

Panaudoję hierarchinius klasterizavimo metodus suklasifikuosime įgimtas formavimosi ydas ir deformacijas, mikroanomalijas bei fenotipinius ypatumus. Klasterių sujungimui naudosime vienetinės jungties („artimiausio kaimyno“), pilnos jungties („tolimiausio kaimyno“) ir dažniausiai naudojamą vidutinės jungties metodą. Vienetinės jungties metodu atstumas tarp klasterių yra apibrėžiamas kaip atstumas tarp dviejų artimiausių objektų, priklausančių skirtingiems klasteriams, pilnosios jungties – kaip atstumas tarp dviejų tolimiausių objektų, priklausančių skirtingiems klasteriams, o vidutinės jungties – vidutinis atstumas tarp visų galimų dviejų skirtingų klasterių objektų porų.

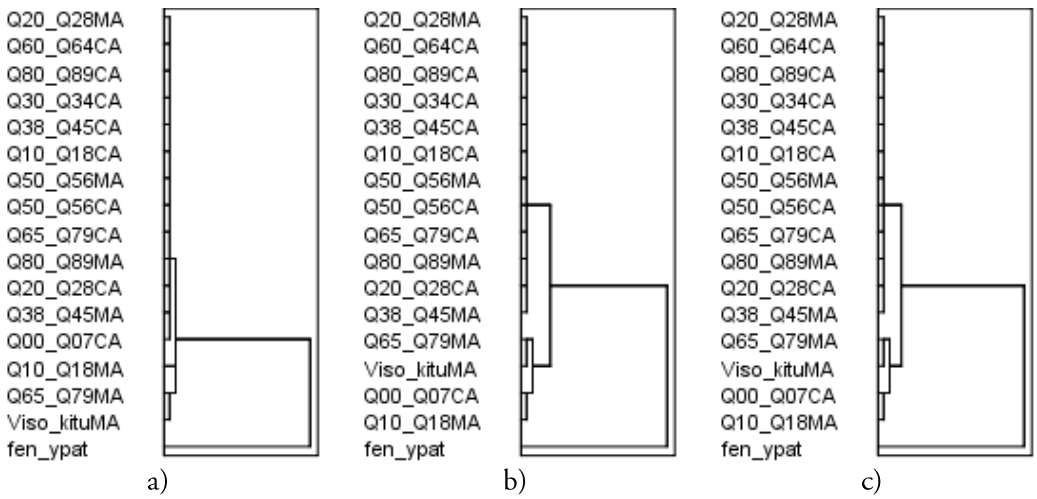


1 pav. Klasterizuoti INN duomenys, kai tiriamiesiems nustatyti patogeniniai chromosominiai pokyčiai: a) vienetinės jungties; b) pilnosios jungties; c) vidutinės jungties.

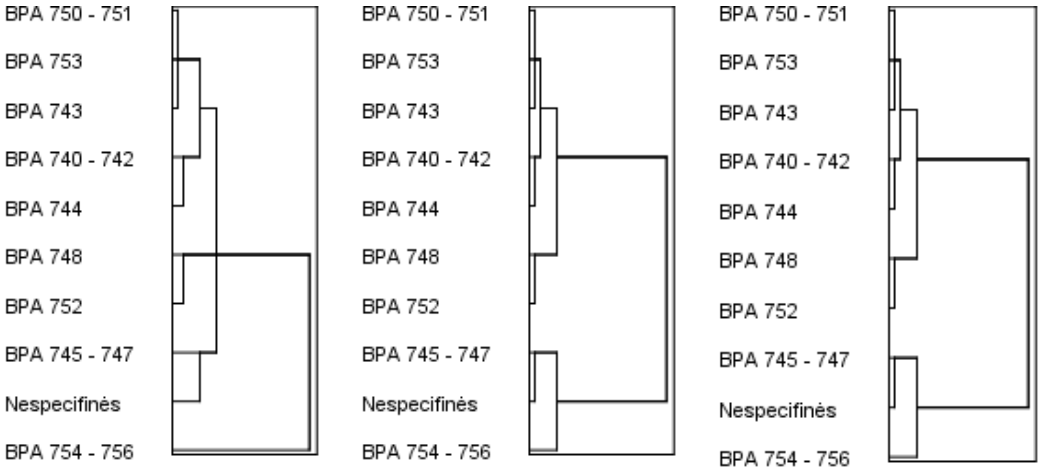
Atstumas tarp klasterių yra lygus atstumo tarp visų galimų stebėjimų porų vidurkiui, kai vienas stebėjimas yra imamas iš vieno, o antras – iš kito klasterio. Šis metodas yra pranašesnis už vienetinės ir pilnos jungties metodus, nes naudoja informaciją apie visas dviejų klasterių objektų poras, o ne tik apie artimiausius arba tolimiausius kaimynus.

Iš 1 ir 2 paveikslų matyti, kad, taikant klasterių sujungimui pilnosios ir vidutinės jungties metodus, gauname identiškus vaizdus. Klasterizuojant įgimtas formavimosi ydas, mikroanomalijas bei fenotipinius ypatumus, kai tiriamiesiems nustatyti patogeniniai chromosominiai pokyčiai, gauname vienodus susidariusius klasterius nepriklausomai nuo klasterių jungimo būdo parinkimo (1 pav.).

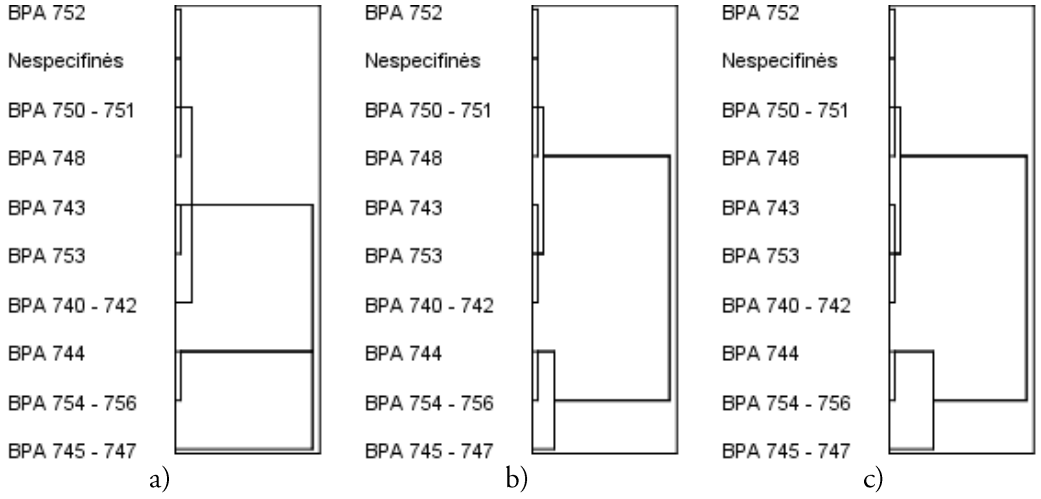
Tiriamiesiems su patogeniniais chromosominiais pokyčiais ir be jų, fenotipiniai ypatumai sudaro atskirą klasterį (1–2 paveikslai).



2 pav. Klasterizuoti INN duomenys, kai tiriamiesiems nenustatyti patogeniniai chromosominiai pokyčiai: a) vienetinės jungties; b) pilnosios jungties; c) vidutinės jungties.



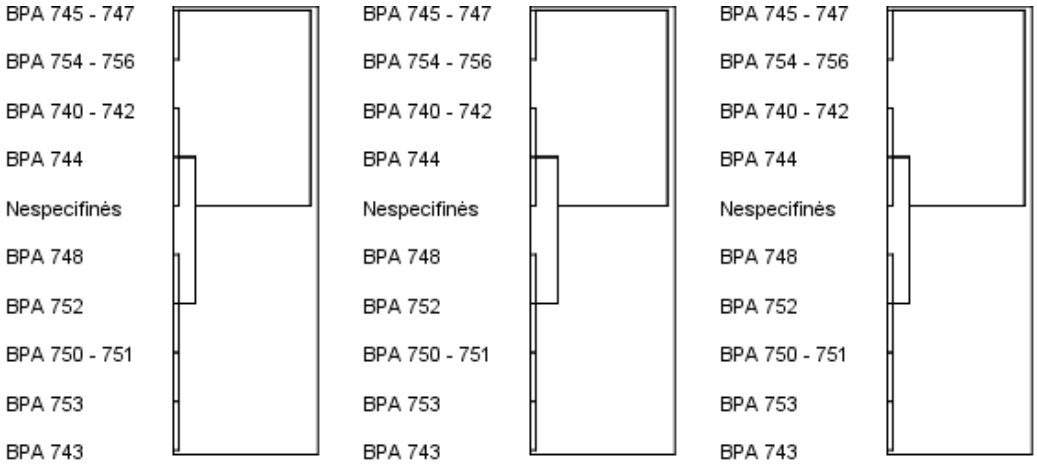
3 pav. Klasterizuoti VN duomenys, kai tiriamiesiems nustatyti gomurio nesuaugimo defektai: a) vienetinės jungties; b) pilnosios jungties; c) vidutinės jungties.



4 pav. Klasterizuoti VN duomenys, kai tiriamiesiems nustatytos lūpos nesuaugimo defektai: a) vienetinės jungties; b) pilnosios jungties; c) vidutinės jungties.

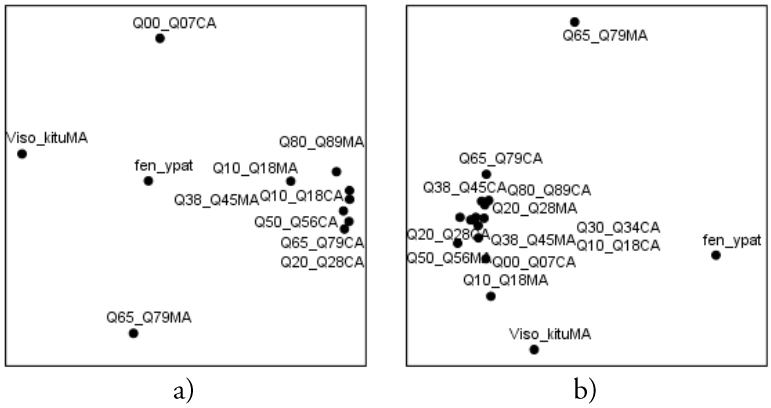
Klasterizuojant intelektinės negalios duomenis, kai tiriamiesiems nenustatyti patogeniniai chromosominiai pokyčiai, pastebėta, jog pirmame klasterių jungimo žingsnyje 60 proc. tarpusavyje koreliuojančių anomalijų patenka į tą patį klasterį ir tik 20 proc., kai analizuojame tiriamuosius su chromosominiais pokyčiais (1-2 lentelės ir 1-2 paveikslai).

Klasterizuojant įgimtas raidos anomalijas, susijusias su veido srities nesuaugimais, kai tiriamiesiems nustatyti lūpos (LN) ir gomurio (GN) nesuaugimai, gauname identiškus vaizdus, kai klasterių sujungimui taikome pilnosios ir vidutinės jungties metodus (3–4 pav.). Tiriamiesiems, kuriems yra nustatyti lūpos ir gomurio nesuaugimai (LGN) kartu, taikant vienetinės, pilnosios ir vidutinės jungties klasterių sujungimo metodus, gauname identiškus klasterius (5 pav.).



5 pav. Klasterizuoti VN duomenys, kai tiriamiesiems nustatytas gomurio ir lūpos nesuaugimas kartu: a) vienetinės jungties; b) pilnosios jungties; c) vidutinės jungties.

Klasterizuojant veido nesuaugimo duomenų, GN grupę, pastebėta, kad pirmu klasterių sujungimo žingsniu tarpusavyje koreliuojančios virškinimo organų sistemos ir akių anomalijos patenka į tą patį klasterį ir tik ketvirtajame žingsnyje likusios koreliuojančios anomalijos sudaro vieną bendrą klasterį. Išanalizavus LGN grupę, nustatyta, jog tarpusavyje koreliuojančios anomalijos ketvirtajame žingsnyje patenką taip pat į tą patį klasterį.



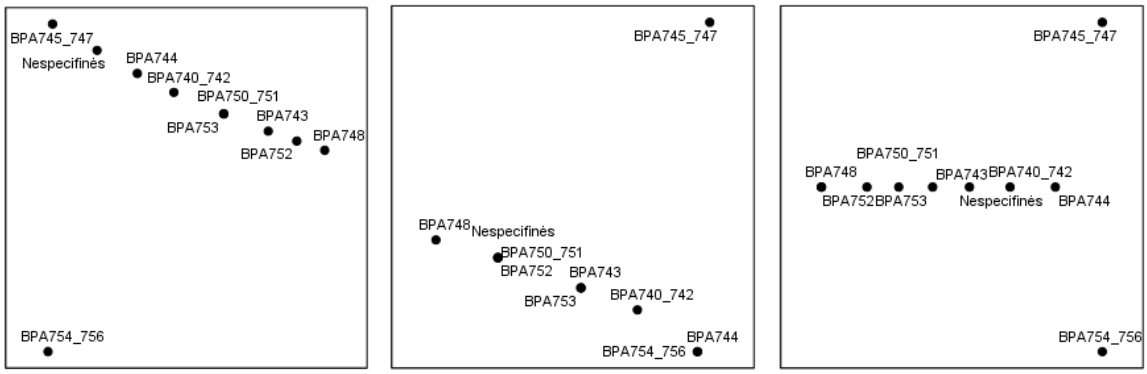
6 pav. Vizualizuoti INN duomenys, kai: a) tiriamiesiems nustatyti patogeniniai chromosominiai pokyčiai; b) tiriamiesiems nenustatyti patogeniniai chromosominiai pokyčiai.

Taikant daugiamačių skalių metodą vizualizuotos įgimtos raidos anomalijos, mikroanomalijos bei fenotipiniai ypatumai, susiję su veido srities nesuaugimais. Daugiamačių skalių analizės sprendiniui gauti, buvo atlikta po dvi iteracijas, analizuojant GN ($E_{MDS_g} < 0{,}0000013$), LN ($E_{MDS_l} < 0{,}0000012$) ir LGN ($E_{MDS_{gl}} < 0{,}0000015$) grupes.

Gautos *Stress* funkcijos reikšmės yra mažesnės už 0,1. Vizualizuojant įgimtas formavimosi ydas, mikroanomalijas bei fenotipinius ypatumus, buvo atlikta 152 iteracijos analizuojant tiriamuosius su patogeniniais chromosominiais pokyčiais ir 254 iteracijos – be patogeninių chromosominių pokyčių. Gautos *Stress* funkcijos reikšmės yra didesnės už 0,1 ($E_{MDS_1} = 0{,}3921635$, $E_{MDS_2} = 0{,}1413160$).

Iš 7a ir 7b paveikslų matyti, kad vizualizuojant įgimtas raidos anomalijas, kai tiriamiesiems nustatyti gomurio arba lūpos nesuaugimai, susidaro du klasteriai, o vizualizuojant įgimtas raidos anomalijas tiriamiesiems, kuriems nustatyti lūpos ir gomurio nesuaugimai – susidaro trys pagrindinai klasteriai.

Vizualizuojant įgimtas formavimosi ydas, mikroanomalijas bei fenotipinius ypatumus, atsižvelgiant į tiriamuosius su patogeniniais chromosominiais pokyčiais ir be jų susidaro keturi pagrindiniai klasteriai.



a) b) c)

7 pav. Klasterizuoti VN duomenys, kai tiriamiesiems nustatytas:
a) gomurio defektai; b) lūpos defektai; c) lūpos ir gomurio defektai kartu.

Išvados

Straipsnyje nagrinėtos asmenų turinčių INN su arba be chromosominiais pokyčiais grupėse įgimtų formavimosi ydų, mikroanomalijų bei fenotipinių ypatumų ir asmenų su VN įgimtų raidos anomalijų, susijusių su veido srities nesuaugimais, tarpusavio priklausomybės.

Nustatyta, kad hierarchinio klasterizavimo metodas yra tinkamesnis genetiniams duomenims vizualizuoti, kai nagrinėjamos asmenų su INN įgimtos formavimosi ydos, mikroanomalijos bei fenotipiniai ypatumai, nes susidarę maži klasteriai yra geriau matomi, negu vizualizuojant daugiamačių skalių metodu.

Daugiamačių skalių metodu, duomenų struktūra geriau matoma, kai vizualizuojame įgimtas raidos anomalijas, susijusias su veido nesuaugimais.

Tiek INN, tiek ir VN duomenų klasterizavimui galime taikyti hierarchinio klasterizavimo metodą, nes duomenų struktūra gerai matoma.

Vizualizavimo rezultatai svarbūs sudarant asmenų su INN genetinio ištyrimo gaires bei tikslingesnei tiriamųjų atrankai molekulinio kariotipavimo tyrimui. Asmenų su VN įgimtų raidos anomalijų susijusių su VN tarpusavio priklausomybės analizė naudinga tiriamųjų su VN kryptingam sveikatos priežiūros plano sudarymui bei ištyrimui, o taip pat yra svarbi ir ankstyvai galimų komplikacijų prevencijai.

Literatūra

AmpFlSTR® Yfiler® PGR rinkinio protokolas (2012). http://www3.appliedbiosystems.com/cms/groups/appliend_markets_support/documents/generaldocuments/cms_041477.pdf

Borg, I., Groenen, P. (2005). Modern Multidimensional Scaling. New York: Springer-Verlag. 616, 330–331.

Chok, N. S. (2010). Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data. http://d-scholarship .pitt.edu/8056/1/Chokns_etd2010.pdf

Čekanavičius, V., Murauskas, G. (2006). Statistika ir jos taikymai. Vilnius: TEV leidykla. I dalis, 199–207 p.

Dash, M., Liu, H. (2003). Efficient Hierarchical Clustering and Its Validation. Data and Knowledge Engineering Journal. Vol 44, 109–138.

Dzemyda, G., Kurasova, O., Žilinskas, J. (2008). Daugiamačių duomenų vizualizavimo metodai. Vilnius: Mokslo aidai. 13–14 p., 52 p.

Kučinskas, V. (2012). Genetikos ir genomikos pagrindai. Vilnius: Vilniaus universiteto leidykla. 11–12 p.

Kurasova, O. (2005). Daugiamačių duomenų vizuali analizė taikant saviorganizuojančius neuroninius tinklus (SOM). Vilnius: Technika. 46–52 p.

Matulevičienė, A., Preikšaitienė, E., Linkevičienė, L., Radavičius, M., Molytė, A., Utkus, A., Kučinskas, V. (2013). Heterogeneity of oral clefts in relation to associated congenital anomalies. Kaunas, Vol. 49 (2), 61–66.

Preikšaitienė, E., Kasnauskienė, J., Utkus, A., Kučinskas, V. (2012). Asmenų su intelektine negalia genetinio ištyrimo gairės. Neurologijos seminarai, Vilnius: Rotas. T. 16, Nr. 4(54), 283-288 p.

TLK-10-AM / ACHI / ACS elektroninis vadovas (2008). http://ebook.vlk.lt/e.vadovas /index.jsp

GENOTYPE AND PHENOTYPE DATA ANALYSIS AND VISUALIZATION

Alma Molytė, Vaidutis Kučinskas, Aušra Matulevičienė, Eglė Preikšaitienė

Summary

In this paper, we present a comparative analysis of hierarchical clustering and multidimensional scaling methods for genotype and phenotype data analysis. Fisher's exact test was applied to determinate dependencies between congenital anomalies. In order to determine the relationship between the dependences of congenital anomalies, deformations, these systems' micro anomalies and congenital anomalies associated with orofacial clefs, the Spearman and Kendall correlation coefficients were applied. It has been detected which methods are better for genetic data visualization.

Key words: genotype and phenotype data, Spearman and Kendall correlation coefficients, hierarchical clustering, data visualization.