Liet. matem. rink, **46***, spec. nr.*, 2006, 280–285

Genetinių sekų markoviškumo tyrimas

Jurgita ŽIDANAVIČIŪTĖ, Tomas REKAŠIUS (VGTU)
el. paštas: jurz@fm.vtu.lt, tomas.rekasius@fm.vtu.lt

Reziumė. Šiame darbe nagrinėjamos DNR genetinės sekos kaip diskrečių būsenų Markovo grandinė. Statistinėje analizėje naudojami duomenys pateikiami dažnių lentelių pavidalu, ir taikomas apibendrintas logit modelis pirmos eilės Markovo grandinės savybei patikrinti visose koduojančiose ir nekoduojančiose DNR pirminės ir antrinės grandinės sekose.

1. Įvadas

Visa paveldima organizmo genetinė informacija, esanti DNR sekose, sudarytose iš 4 tipų nukleotidų, nėra pastovi. Vyksta atsitiktinės tų sekų mutacijos, vienus nukleotidus keičia kiti, atskiri sekų fragmentai prapuola, įsiterpia kiti. Paprastai laikoma, kad nukleotidų skirstinį sekoje aprašo homogeninė Markovo grandinė. Šiame darbe statistiniais metodais parodoma, kad net paprasčiausių organizmų (*bakterijos E. Coli*) DNR sekoms ši prielaida negalioja.

Natūralu laikyti, kad DNR sekos evoliucija laike yra Markovo procesas. Ilgą laiką buvo daroma prielaida, kad nukleotidai mutuoja nepriklausomai vienas nuo kito, ir tik pastaruoju metu pradėti nagrinėti modeliai, kuriuose kiekvieno nukleotido mutavimas priklauso nuo gretimų (kaimyninių) nukleotidų. Platesnį minėtų modelių aprašymą galima rasti, pavyzdžiui, straipsniuose [4] ir [5].

J.L. Jensen [4] parodė, kad jeigu DNR sekų atsitiktinės mutacijos priklauso tik nuo artimiausiųjų kaimyninių nukleotidų, tai stacionarusis nukleotidų skirstinys jose sudaro pirmos eilės Markovo grandinę. Ar tai galioja realioms DNR sekoms? Šį klausimą nagrinėjo P.J. Avery ir D.A. Henderson [2, 3]. Jie naudojo konkretaus geno (*preproglucagon*) nekoduojančią dalį (*introną*) iš žmogaus DNR. Laikydami, kad DNR seka sudaro baigtinės eilės Markovo grandinę, jie taikė logtiesinius modelius nukleotidų porų ir jų tripletų dažnių bendram tikimybiniam skirstiniui įvertinti bei Markovo grandinės eilei nustatyti. Pirmos eilės Markovo modelis nepakankamai gerai aprašė turimus duomenis, o antros eilės – nebuvo atmestas. Jų sudaryta imtis dėl stebėjimų tarpusavio priklausomybės netenkina standartinių logtiesinio modelio prielaidų, ir todėl modeliavimo būdu buvo ieškoma sprendimų tinkamai χ^2 statistikos aproksimacijai.

Šio darbo tikslas – patikrinti DNR sekos pirmos eilės markoviškumo savybę, naudojant apibendrintą logit modelį [1,6]. Specialiu būdu sudaroma imtis, kuri tenkina šio modelio standartines sąlygas. Apibendrintam logit modeliui įvertinti naudojama SAS programinė įranga ir jos procedūra CATMOD [6].

Antrajame skyrelyje trumpai aprašomos nukleotidų sekos, jų markoviškumo savybė ir šiai savybei patikrinti taikomas apibendrintas logit modelis. Trečiajame – aptariama atlikta statistinė analizė ir pateikta gautų rezultatų suvestinė. Gale pateikiamos išvados.

2. Markovo savybė nukleotidų sekoms. Apibendrintas logit modelis

2.1. Nukleotidų sekos

DNR molekulė turi keturių tipų nukleotidus: du pirimidinus – timiną (T) ir citoziną (C), ir du purinus – guaniną (G) ir adeniną (A). Jų aibę žymėsime $\mathcal{A} = \{A, C, G, T\}$.

Taisyklingai DNR grandinei susisukti į dvigubą spiralę reikalinga, jog purinas vienoje grandinės pusėje būtų pakeičiamas pirimidinu kitoje. Todėl nukleotidas A stovi priešais T, o nukleotidas C – priešais G. Be to, A ir T jungiasi per dvi, o C ir G – per tris vandenilines jungtis. Tokios bazių poros ir iš jų sudarytos DNR grandinės vadinamos komplementariomis.

Pagal susitarimą, dvigubos DNR spiralės viršutinė (pirminė) grandinė skaitoma iš kairės į dešinę, o apatinė (antrinė) – iš dešinės į kairę.

Pirminės ir antrinės grandinės yra suskirstytos į koduojančias sritis (genus) ir nekoduojančias (tarpus). Pirminės grandinės genai ir tarpai nebūtinai atitinka antrinės grandinės genų ir tarpų sritis.

2.2. Markovo savybė

Tarkime, kad nukleotidų sekos evoliuciją aprašanti homogeninė Markovo grandinė yra apverčiama ir joje nevyksta nukleotidų įterpimai ar išmetimai, o tik nukleotidų mutacijos. Be to, per laiko vienetą gali įvykti tik vieno atsitiktinai parinkto nukleotido mutacija (*Glauberio dinamika*) ir jos tikimybė priklauso tik nuo artimiausiųjų to nukleotido kaimynų. Tada, iš [4] išplaukia, kad stacionarusis šios evoliucijos skirstinys $\mathbf{P}$ nukleotidų sekų $X = (x_0, x_1, \ldots, x_n)$ aibėje $\mathcal{A}^{n+1}$ taip pat turi Markovo savybę:

$$\mathbf{P}\{x_l = a | x_s, s < l\} = \mathbf{P}\{x_l = a | x_{l-1}\}, \quad a \in \mathcal{A},\ l = 1, \ldots, n. \tag{1}$$

Simetrinė šios sąlygos forma yra tokia

$$\mathbf{P}\{x_l = a | x_s, s \neq l\} = \mathbf{P}\{x_l = a | x_{l-1}, x_{l+1}\}, \quad l = 1, \ldots, n-1. \tag{2}$$

Paprasčiausiame nukleotidų sekų evoliucijos modelyje natūralu laikyti, kad nukleotidų mutacijos nepriklauso nuo jų vietos (*cite, location*) l. Šią prielaidą perkėlę į stacionarųjį skirstinį $\mathbf{P}$ gauname, kad Markovo grandinė $X = (x_0, x_1, \ldots, x_n)$ yra *homogeninė*: jos perėjimo tikimybės dešinėje lygybės (1) pusėje taip pat nepriklauso nuo l. Trumpumo dėlei tas perėjimo tikimybes pažymėkime $p_{ab} := \mathbf{P}\{x_1 = b | x_0 = a\}$ ir tarkime, kad $p_{ab} > 0\ \forall\, a, b \in \mathcal{A}$. Gerai žinoma, kad tada egzistuoja vienintelis stacionarusis homogeninės Markovo grandinės X skirstinys $\pi = \{\pi_a, a \in \mathcal{A}\}$.

Galimybė arba *šansas (odds)* apibrėžiamas formule

$$O_{b/a|k,d} := \frac{\mathbf{P}\{x_l = b | x_{l-1} = k, x_{l+1} = d\}}{\mathbf{P}\{x_l = a | x_{l-1} = k, x_{l+1} = d\}}. \tag{3}$$

Jis nusako, kiek kartų skiriasi centrinio nukleotido C tikimybė įgyti reikšmę b nuo jo bazinės reikšmės (*reference value*) a tikimybės, kai kairiojo K ir dešiniojo D nukleotidų reikšmės yra atitinkamai k ir d. Nesunku matyti, kad

$$O_{b/a|k,d} := \frac{p_{kb}\,p_{bd}}{p_{ka}\,p_{ad}}. \tag{4}$$

Vadinasi, Markovo modelio atveju galimybių santykio logaritmas (*generalized logit: apibendrinta logit funkcija*) yra

$$\ln\Big(O_{b/a|k,d}\Big) := \lambda_b^C + \lambda_{kb}^{KC} + \lambda_{bd}^{CD}. \tag{5}$$

Čia

$$\lambda_b^C = \ln p_{ab} + \ln p_{ba} - 2\ln p_{aa}, \tag{6}$$

$$\lambda_{kb}^{KC} = \ln\big(p_{kb}/p_{ab}\big) - \ln\big(p_{ka}/p_{aa}\big), \tag{7}$$

$$\lambda_{bd}^{CD} = \ln\big(p_{bd}/p_{ba}\big) - \ln\big(p_{ad}/p_{aa}\big), \tag{8}$$

Pastebėkime, kad $\lambda_a^C = \lambda_{ab}^{KC} = \lambda_{ba}^{CD} = 0$ (parametrų λ *identifikuojamumo sąlyga*). Modelis (5), išplaukiantis iš 1-os eilės Markovo savybės pozicijoje l, skiriasi nuo pilnojo logtiesinio modelio (*saturated loglinear model*) tuo, kad neturi kairiojo ir dešiniojo kaimynų sąveikos (*interaction*) nario λ_{kbd}^{KCD}. Taigi, uždavinys būtų kiekvienoje pozicijoje $l = 1, ..., n-1$ patikrinti nulinę hipotezę H_0: $\lambda_{kbd}^{KCD} = 0$ su alternatyva H_1: $\lambda_{kbd}^{KCD} \neq 0$.

3. Statistinė analizė

Statistinei analizei pasirinkta (*bakterijos E.Coli*) DNR seka (genomas). Ši bakterija yra populiari bioinžineriniuose ir bioinformatikos srities tyrimuose.

Nagrinėjami nukleotidų tripletai (*trys iš eilės einatys nukleotidai*) pirminėje ir antrinėje DNR grandinėse, išskiriant atskirai koduojančias sritis ir tarpus. Imčiai $\mathcal{D}$ sudaryti išrenkamas kas antras nukleotidas iš DNR sekos, kurį vadinsime tripleto viduriniąja reikšme, o nukleotidai jam iš kairės ir dešinės bus kaimyniniais. Greta esančių dviejų tripletų šoninis nukleotidas yra bendras: vienam – jis kaimyninis iš kairės, kitam – iš dešinės. Nukleotidų tripletų aibę pažymėkime

$$\mathcal{D} = \big\{(y_l, z_l), l = 1, \ldots, N\big\},$$

čia

$$y_l = x_{2l}, \quad z_l = (x_{2l-1}, x_{2l+1}), \quad l = 1, \ldots, N.$$

Turimiems duomenims daroma prielaida:
(P) $\{y_l,\ l = 1, \ldots, N\}$ yra sąlyginai nepriklausomi, kai žinomas $\{z_l,\ l = 1, \ldots, N\}$, ir z poveikis y'ui nepriklauso nuo l.

Ši prielaida leidžia tiesiogiai taikyti apibendrintą logit modelį [1,6]. Pastebėkime, kad šios prielaidos yra išpildytos, jeigu X yra 1-os eilės homogeninė Markovo

1 lentelė. Apibendrinto logit modelio rezultatai; Skliaustuose nurodytas stebėjimų ir sekų skaičius

Kaimyninių nukleotidų savybės	**DF**	**Tarpai-I (1252795), (622)**	**Tarpai-II (1194109), (618)**	**Genai-I (843894), (1351)**	**Genai-II (892624), (1454)**	***p***
			Chisq			
Konstanta	3	1330,8	1321,1	2575,9	2812,9	<0,0001
Kairėje savybė (p)	3	4436,6	4056,2	14830,1	16219,9	<0,0001
Kairėje savybė (j)	3	15193,4	14275,4	4801,9	5033,2	<0,0001
Dešinėje savybė (p)	3	16384,9	15913,6	8317,5	9222,4	<0,0001
Dešinėje savybė (j)	3	17428,9	15945,3	3707,7	4019,8	<0,0001
Kairėje sav. (p*j)	3	7782,6	7583,88	11045,0	11494,5	<0,0001
Dešinėje sav. (p*j)	3	11722,7	10778,3	13467,0	14074,2	<0,0001
LR	**27**	**18735,2**	**18008,78**	**18029,6**	**19127,3**	**<0,0001**

grandinė. Prielaida (P) kartu su H_1 reiškia, kad sekos X kas antras elementas netenkina 1-os eilės homogeninės Markovo grandinės sąlygos.

Kiekvieną nukleotidą apibūdina dvi savybės:

(p) ar jis pirimidinas ar purinas,

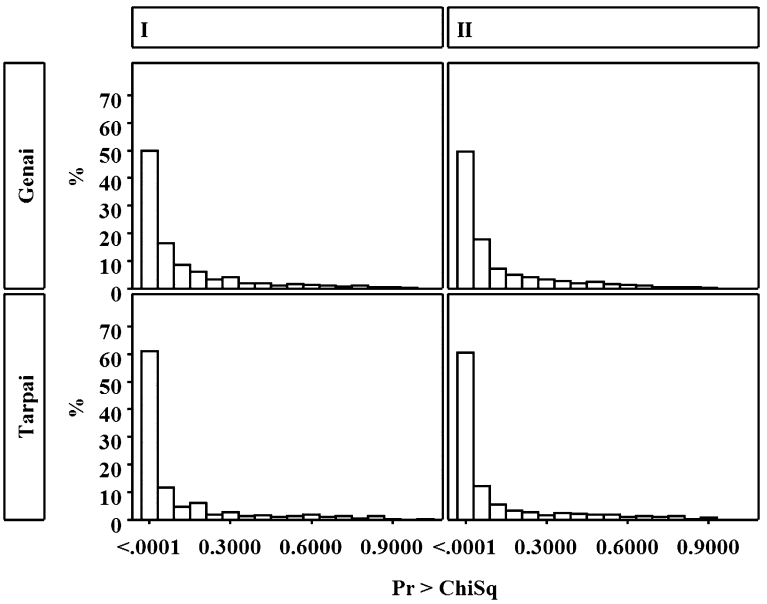
(j) ar jis turi dvi ar tris jungtis.

Todėl kaimyniniams nukleotidams naudojamas šių savybių binarinis kodavimas. Tai leidžia atskirai nagrinėti kiekvienos kaimyninio nukleotido savybės (p), (j) ir jų sąveikos (p*j) įtaką viduriniajam nukleotidui.

Viso yra $4^3 = 64$ skirtingos nukleotidų tripletų kombinacijos, kurios sudaro penkių požymių dažnių lentelę. Kad tikėtinumo santykio (*LR, Likelihood Ratio*) statistikos skirstinio χ^2 aproksimacija būtų pakankamai tiksli rekomenduojama, kad minimalus prognozuojamas dažnis lentelėje būtų nemažesnis už 5. Jeigu reikalauti, kad vidutiniškai kiekvienoje dažnių lentelės ląstelėje būtų $\geqslant 5$ stebėjimai, tai imtyje kiekviena kombinacija turėtų pasikartoti nemažiau kaip penkis kartus. Kadangi DNR sekoje kas antras tripletas praleidžiamas, tai iš vienos sekos galima išrinkti tik perpus mažesnio ilgio imtį. Dėl to apibendrinto logit modelio vertinimui tenka naudoti ne trumpesnes kaip $64 \cdot 5 \cdot 2 = 640$ ilgio DNR sekas.

Pirmiausia įvertintas apibendrintas logit modelis pirmos eilės Markovo savybei (6) patikrinti visose, tiek koduojančiose, tiek nekoduojančiose, DNR pirminės ir antrinės grandinės sekose (1 lentelė), o vėliau tas pats modelis – kiekvienai sekai atskirai (1 pav.).

Visoms koduojančioms ir nekoduojančioms sekoms pirminėje ir antrinėje grandinėse visi parametrai modelyje yra statistiškai reikšmingi ($p < 0.0001$). Vadinasi, kaimyninių nukleotidų iš kairės ir iš dešinės savybės įtakoja viduriniojo nukleotido tipą. Ar grandinė yra pirmos eilės Markovo, sprendžiama pagal kairiojo ir dešiniojo nukleotidų sąveikos koeficientus. Kadangi jie neįtraukti į modelį (6), tai jų reikšmingumą nusako LR statistika, kuri lygina vertinamą modelį su pilnuoju (*saturated*) modeliu. Visais atvejais hipotezė apie skirtumų nebuvimą tarp vertinamo ir pilnojo modelių buvo atmesta. Taigi, ir koduojančioms ir nekoduojančioms DNR sekoms pirmos eilės Markovo savybė negalioja.



1 pav. LR statistikos p reikšmių skirstiniai; I – pirminė grandinė, II – antrinė grandinė.

Vertinat modelį (6) kiekvienai sekai atskirai, gauta, kad pirmos eilės Markovo savybė galioja tik nedidelei daliai sekų (1 lentelė).

4. Išvados

Priklausomybei tarp nukleotidų DNR grandinėje aprašyti pirmos eilės Markovo modelio neužtenka. Tik labai nedidelei daliai sekų pirmos eilės Markovo savybė nebuvo atmesta.

Aukštesnės eilės Markovo modeliui nesunku užrašyti teorinę apibendrintą logit funkciją, bet pakankamai sudėtinga modelį įvertinti praktiškai. Didėjant Markovo modelio eilei, didėja parametrų skaičius, o tuo pačiu reikalingos ir ilgesnės sekos pakankamai tiksliam modelio parametrų statistiniam įvertinimui. Dauguma sekų DNR grandinėje nėra ilgos, todėl, jeigu reikalauti, kad kiekviena tripleto nukleotidų kombinacija pasikartotų sekoje ne mažiau kaip 5-is kartus, tai didelė dalis sekų iškrenta iš tyrimo. Minėtas reikalavimas atsiranda, nagrinėjant dažnių lenteles, kur chikvadrat aproksimacijos tikslumui pageidautina prielaida, kad vienoje ląstelėje vidutiniškai būtų $\geqslant 5$ stebėjimai. Todėl šiuo atveju iškyla aktualus uždavinys išvystyti statistinius metodus tokioms lentelėms, kurių daugumoje ląstelių stebėjimų skaičius yra mažesnis už penkis arba atskiru atveju dauguma ląstelių yra tuščios, t.y. dažnių lentelės yra stipriai išretintos.

Literatūra

1. A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, New York (1990).

2. P.J. Avery, Fitting interconnected Markov chain models-DNA sequences and test cricket matches, *The Statistician*, **51**, 267–278 (2002).
3. P.J. Avery, Henderson, D.A. Fitting, Markov chain models to discrete state series such as DNA sequences, *Appl. Statist.*, **48**, 53–61 (1999).
4. J.L. Jensen, Context dependent DNA evoliutionary models, *Research Reports*, **458** (2005).
5. T. Rekašius, Priklausomybių modeliuotose DNR sekose tyrimas, *Liet. matem. rink.*, **45**(spec. nr.), 363–368 (2005).
6. M.E. Stokes, C.S. Davis, G.S. Koch, *Categorical Data Analysis Using the SAS(R) System*, SAS Institute, Cary, NC (2001).

SUMMARY

J. Židanavičiūtė (VGTU), T. Rekašius. Fitting Markov property to genetic sequences

In this paper DNA sequents are modelled as discrete-state Markov chains. Statistical data is presented in contingency tables form. The generalized logit model is used to test the first-order Markov property for all coding and non-coding subsequences of DNA.

Keywords: contingency tables, generalized logit, Markov chains.