Liet. matem. rink., T. 43, spec. nr., 660–665



# Antrinių baltymo struktūrų nuspėjimo metodų modifikavimo galimybės

Alvydas ŠPOKAS, Albertas TIMINSKAS (VGTU)
*el. paštas:* alvydas.spokas@fm.vtu.lt, timis@ibt.lt

## 1. Įvadas

Baltymai – tai biologiškai aktyvios molekulės, sudarytos iš įvairių amino rūgščių, susijungusių į ilgą peptidinę grandinę. Biologinei funkcijai atlikti tokia amino rūgščių seka turi įgauti savitą erdvinę struktūrą (1 pav.).

Baltymų erdvinės struktūros sudarytos iš tam tikrų tarpinės (antrinės) struktūros elementų ($H$ – alfa spiralė, $S$ – beta klostė, $C$ – jungiamieji fragmentai).

Dauguma baltymų savarankiškai sugeba susivynioti į reikiamą erdvinę globulę. Tai leidžia teigti, jog amino rūgščių sekoje yra užkoduota visa reikiama informacija apie baltymo molekulės erdvinę struktūrą. Nuspėti iš karto tretinę baltymo struktūrą yra labai sudėtinga dėl per didelio spėjimų laisvės laipsnių skaičiaus, todėl stengiamasi tiksliai nuspėti tarpinę – antrinę – struktūrą siekiant palengvinti tretinės struktūros spėjimą. Teisingai iškodavus amino rūgščių sekoje slypinčią struktūros susidarymo informaciją, atsivertų didelės jos panaudojimo galimybės:

1. Tiksliai o svarbiausia greitai ir su minimaliom išlaidom nuspėti praktiškai bet kokios amino rūgščių sekos antrinę struktūrą. Nuspėjimo tikslumas bus apspręstas iškodavimo tikslumo.
2. Turima pakankamai tiksli amino rūgščių sekos antrinė struktūra leistų nesunkiai prognozuoti (sumodeliuoti) šios sekos erdvinę struktūrą.
3. Žinoma erdvinė struktūra (molekulės vaizdas) suteikia daug reikšminės informacijos palengvinančios tolimesnius baltymų tyrimus.

1 pav. Erdvinė baltymo struktūra. Atitinkamai išskirtos skirtingos baltymo antrinės struktūrinės būsenos.

4. Suprantamas baltymo molekulės erdvinės struktūros susidarymo mechanizmas leistų konstruoti dirbtinius baltymus su norima (iš anksto žinoma) funkcija.

Šiuo metu ypač aktualūs baltymo antrinių struktūrų nuspėjimo metodai, kaip tarpinė grandis leidžianti ženkliai sumažinti laisvės laipsnių skaičių sprendžiant baltymų erdvinės struktūros nuspėjimo uždavinius.

## 2. Nuspėjimo metodai

Gamtoje aptinkamų baltymų sudėtyje yra 20 skirtingų amino rūgščių, tačiau kiekvieno baltymo amino rūgščių sudėtis bei antrinių struktūrinių būsenų kiekis yra saviti. Vienos amino rūgštys baltymuose yra populiaresnės už kitas, be to amino rūgštys savitai pasiskirsto antrinių struktūrų būsenose. Pasinaudojant baltymų duomenų bazės stebėjimais yra įvertinamas kiekvienos amino rūgšties struktūriškumas (gebėjimas įtakoti tam tikros antrinės struktūrinės būsenos susidarymą).

$$I(S;P)=\ln\left(\frac{P(S|R)}{P(S)}\right)=\ln\left(\frac{f(S,R)}{f(R)}/\frac{f(S)}{f(N)}\right)=\ln\left(\frac{f(S,R)}{f(R)}\right)-\ln\left(\frac{f(S)}{f(N)}\right), \quad (1)$$

čia:

$S$ – viena iš trijų struktūrinių būsenų ($H$, $S$ arba $C$),
$R$ – viena iš 20 amino rūgščių,
$P(S;R)$ – tikimybė, kad amino rūgštis $R$ yra struktūrinėje būsenoje $S$,
$P(S)$ – tikimybė aptikti struktūrinę būseną,
$f(S,R)$ – amino rūgšties dažnumas antrinės struktūros būsenoje,
$f(S)$ – struktūrinės būsenos dažnumas,
$f(R)$ – amino rūgšties dažnumas.

### 2.1. *Fasman'o metodas*

Struktūrinė informacija pati savaime nėra patikima dėl normavimo, todėl plačiau taikomas struktūrinės informacijos pokytis (struktūriškumo matas), kuris parodo amino rūgšties vienos struktūrinės informacijos pranašumą prieš kitas.

$$I(S) = I(S;R) - I(\bar{S};R) = \ln\left(\frac{f(S,R)}{f(\bar{S},R)}\right) - \ln\left(\frac{f(S)}{f(\bar{S})}\right), \quad (2)$$

čia:

$I(S;R)$ – struktūrinė informacija susidaryti tam tikrai struktūrinei būsenai $S$,
$I(\bar{S};R)$ – struktūrinė informacija susidaryti kitoms struktūrinėms būsenoms (visoms išskyrus $S$).

Paremtas amino rūgščių struktūriškumo matais sukurtas vienas pirmųjų antrinių struktūrų nuspėjimo metodų – Fasman'o metodas. Šio metodo efektyvumas (tikslumas) siekia 50%, nors naudojamos struktūriškumo matų matricos dydis yra tik $3 \times 20$ (kiekviena iš 20 amino rūgščių gali būti vienoje iš 3 struktūrinių būsenų). Baltymo antrinių

struktūrinių būsenų nuspėjimo metodų tikslumas apskaičiuojamas įvertinant kiekvienos struktūrinės būsenos spėjimo tikslumą.

$$Q = \frac{Q_H + Q_S + Q_C}{3}, \tag{3}$$

čia $Q_H$, $Q_S$, $Q_C$ – atitinkamų struktūrinių būsenų spėjimo tikslumai.

### 2.2. *GOR metodai*

Amino rūgščių struktūriškumo matai įtakoja vienas kitą, t.y. vienos amino rūgšties struktūriškumo matas dalinai apsprendžia kaimyninės amino rūgšties (kaimynės) struktūrinę būseną. Todėl GOR algoritmu paremti antrinių baltymo struktūrų nuspėjimo metodai įvertina ne tik centrinės (analizuojamoje pozicijoje esančios), bet ir kaimynių (po 8 į abi puses) struktūriškumo matus. Taigi (2) išraiška taikoma ir visiems kaimynams sumuojant jų struktūriškumo matų įtakas.

$$I(\Delta S) = \sum_{|n|=1}^{8} \big(I(S; R_n) - I(\bar{S}; R_n)\big) = \sum_{|n|=1}^{8} \left(\ln\Big(\frac{f(S, R_n)}{f(\bar{S}, R_n)}\Big) - \ln\Big(\frac{f(S)}{f(\bar{S})}\Big)\right), \tag{4}$$

čia:

$R_n$ – kaimyninė, per n pozicijų nuo centrinės nutolusi amino rūgštis,

$I(S; R_n)$ – kaimyninės amino rūgšties struktūriškumo matas.

Tokiu atveju metode naudojamų struktūriškumo matų matricos dydis yra $3 \times 20 \times 17$, o metodo tikslumas siekia 63%.

Kaimyninės amino rūgšties įtaka centrinei struktūrinei būsenai priklauso nuo centrinės bei kaimyninės amino rūgščių prigimties, todėl (4) išraiška GOR metodo trečiojoje versijoje (GOR III) naudojama žinomos prigimties amino rūgšties struktūrinei būsenai nusakyti.

$$\begin{aligned} I(\Delta S) &= \sum_{|n|=1}^{8} \big(I(S; R_n|R) - I(\bar{S}; R_n|R)\big) \\ &= \sum_{|n|=1}^{8} \left(\ln\Big(\frac{f(S, R_n, R)}{f(\bar{S}, R_n, R)}\Big) - \ln\Big(\frac{f(S, R)}{f(\bar{S}, R)}\Big)\right). \end{aligned} \tag{5}$$

Tokiu būdu metode naudojamų struktūriškumo matų matricos dydis padidėja iki $20 \times 3 \times 20 \times 16$, o metodo tikslumas siekia 64%.

### 2.3. *Metodo išplėtimas*

GOR metodo autoriai daug dėmesio skiria kaimynių struktūriškumo matams įvertinti, tačiau nekreipiamas dėmesys į susidarančių antrinių struktūrų tarpusavio sąveikas. Baltymo

sintezės metu susidarantys antrinių struktūrinių būsenų komponentai sąveikauja tarpusavyje ir neretai tolimesni struktūrinių būsenų pokyčiai apsprendžiami ne kaimynų prigimtis o jų struktūrinių būsenų susidarymas (struktūrinė būsena įtakoja struktūrinei būsenai). Į (5) išraišką įtraukus kaimynės struktūrinę būseną, gauname:

$$I(\Delta S, \Delta S_n) = \sum_{|n|=1}^{8} \big(I(S; R_n|S_n) - I(\bar{S}; R_n|S_n)\big)$$
$$= \sum_{|n|=1}^{8} \left(\ln\Big(\frac{f(S, R_n, S_n)}{f(\bar{S}, R_n, S_n)}\Big) - \ln\Big(\frac{f(S, S_n)}{f(\bar{S}, S_n)}\Big)\right). \quad (6)$$

Taip peržiūrint visas kaimynių struktūras, įvertinama centrinės amino rūgšties struktūrinės būsenos susidarymo informacija.

$$I(\Delta S) = \sum_{i=1}^{3} I(\Delta S, \Delta S_n^i) = \sum_{|n|=1}^{8}\sum_{i=1}^{3} \left(\ln\Big(\frac{f(S, R_n, S_n^i)}{f(\bar{S}, R_n, S_n^i)}\Big) - \ln\Big(\frac{f(S, S_n^i)}{f(\bar{S}, S_n^i)}\Big)\right). \quad (7)$$

Šios lygties sprendimo variantai yra keli. Dabar pateikiamas supaprastintas sprendimo variantas, kuris buvo išbandytas praktiškai nuspėjant baltymo antrinę struktūrą. Struktūrinių būsenų nuspėjimo metu analizuojamos iš karto dviejų (centrinės ir kaimyninės) struktūrinių būsenų susidarymas, todėl uždavinio sprendimo metu daroma prielaida:

$$I(\Delta S) = \sum_{i=1}^{3} I(\Delta S; \Delta S_n^i), \quad (8)$$

kurios dėka (7) išraiška tampa ekvivalentiška (5). Tokiu būdu įvertinus centrinės amino rūgšties struktūrinės būsenos susidarymą, yra įvertinama kaimynių struktūrinių būsenų susidarymas remiantis centrinės ir kaimyninės struktūrinių būsenų populiarumu.

$$I(\Delta S_n) = I(\Delta S_0) \times P(S_0, S_n|R, R_n), \quad (9)$$

čia $P(S_0, S_n|R, R_n)$ – tikimybė aptikti struktūrinių būsenų kombinaciją (1 iš 9), prie tam tikros amino rūgščių kombinacijos.

Bendru atveju kiekvieno baltymo sekos taško galimas struktūrinis būvis nusakomas remiantis dviejų komponentų rezultatais:

1) centrinės ir kaimynių struktūriškumo matų įtakos,
2) centrinės ir kaimynių struktūrinių būsenų kombinacijų populiarumu.

Taigi ir bendras kiekvieno sekos taško struktūrinis būvis nusakomas dviejų komponenčių pagalba.

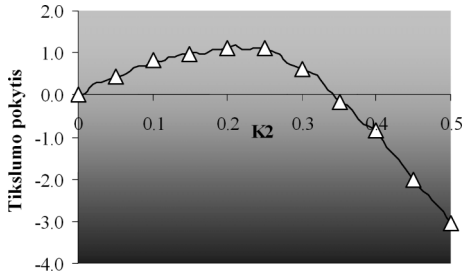
$$I(\Delta S) = I(\Delta S_0) \times K_1 + I(\Delta S_n) \times K_2. \quad (10)$$

## 2.4. *Sprendimas*

Varijuojant komponenčių svorio koeficientais kaimynių struktūrinių būsenų įtakos įvedimas leido klasikinio GOR metodo tikslumą padidinti apie 1% (2 pav.).

Gautas metodo tikslumo padidėjimas yra reikšmingas, kadangi metode naudojamų struktūriškumo matų matricos dydis padidėjo nežymiai, lyginant su GOR metodo evoliucijos metu gautais tikslumo bei matricos dydžio pokyčiais (1 lentelė).

Kitas (6) lygties sprendimo variantas yra subtilesnis ir jis nėra išbandytas praktiškai. Siūloma, pagal (6) lygtį analizuoti visas įmanomas struktūrinių būsenų kombinacijas, kiekvienai kaimynei ir iš gautų duomenų (3 pav.) parinkti optimalią centrinės struktūrinę būseną.




2 pav. Kaimyninių struktūrinių būsenų įvertinimo rezultatai, priklausomybėje nuo suteikto jiems svorio $K2$ ($K1 + K2 = 1$).

1 lentelė. Antrinių baltymo struktūrų nuspėjimo metodų tikslumo pokytis

| **Eil. Nr.** | Metodai | Parametrų skaičius | Parametrų pokytis kartais | Tikslumo pokytis |
|---|---|---|---|---|
| **1** | GOR I | $20 \times 17 \times 3$ | 1 | 0 |
| **2** | GOR III | $20 \times 17 \times 3 \times 20$ | 20 | 0,5 |
| **3** | GOR Is | $20 \times 3 \times 17 \times 3$ | 3 | 1 |

a)
87654321012...
MADFRTIJKLM...
$H * * * * * * * H * * * * \Rightarrow N81$
$H * * * * * * * S * * * * \Rightarrow N82$
$H * * * * * * * C * * * * \Rightarrow N83$
$S * * * * * * * H * * * * \Rightarrow N84$
$S * * * * * * * S * * * * \Rightarrow N85$
$* * * * * * * * * * * * *$
$*H * * * * * *H * * * * \Rightarrow N71$
$* * * * * * * * * * * * *$
$* * * * * * *HH * * * * \Rightarrow N11$
$* * * * * * * * * * * * *$

b)
87654321012...
MADFRTIJKLM...
$SCHHSSC'N16$
$*SHHCCSC'N26$
$**SHHCCSC'N33$
$***CSHHHSC'N49$
$****CCHSSSC'N59$

3 pav. Optimalių struktūrinių būsenų kombinacijų parinkimas. a) Įvertinamas kiekvienos struktūrinių būsenų poros populiarumas, b) Išrenkamos populiariausios poros ir atliekama optimizacija.

Pagal pasirinktą centrinę struktūrinę būseną atitinkamai parenkamos optimalios kaimynių struktūrinės būsenos ir gaunami lokalių 17-tukų struktūrinių būsenų spėjimai. Kiekvienas sekos taškas nusakomas 17 kartų, iš kurių turi būti parenkamas optimalus. Kadangi struktūrinės būsenos parinkimas turi grįžtamąjį ryšį (įtakoja prieš tai pasirinktų struktūrinių būsenų pokyčius), tikslinga naudoti iteracinį optimalios struktūrinės būsenos nusakymą.

Be abejo, galima šį optimizavimo uždavinį spręsti be tarpinių (lokaliai optimalių fragmentų) nusakymų, tačiau tokiu atveju pernelyg didelis laisvės laipsnių skaičius labai apsunkintų uždavinio sprendimą.

## 3. Išvados

Sėkmingai pritaikytas vienas iš galimų baltymo antrinių struktūrų nuspėjimo metodo GOR išplėtimas panaudojant struktūrinių būsenų tarpusavio sąveikas. Gauti rezultatai rodo, kad struktūrinių būsenų sąveikos įvertinimas yra reikšmingas ir suteikia papildomos informacijos apie baltymo antrinių struktūrų susidarymo ypatybes.

## Literatūra

[1] *Bioinformatics: Sequence, Structure, and Databanks*, Oxford Press (2000).
[2] J. Garnier, D.J. Osguthorpe, B. Robson, *J. Mol. Biol.*, **120**, 97–20 (1978).
[3] M. Ouali, King RD, *Protein Sie 2000*, **9**(6), 9998–99.
[4] J. Garnier, D.J. Osguthorpe, B. Robson, *Methods Enzymol*, **266**, 541–53 (1996).
[5] J. Westbrook, Z. Feng, S. Jain, T.N. Bhat, N. Thanki, V. Ravichandran, G.L. Gilliland, W. Bluhm, H. Weissig, D.S. Greer, P.E. Bourne, H.M. Berman, *Nucleic Acids Res.*, **30**(1), 245–48 (2002).
[6] R.D. King, M.J.E. Sternberg, *Protein Sci.*, **5**(11), 2298–10 (1996).
[7] `http://www.rcsb.org/pdb/`

## Modify possibilities of the secondary structures prediction method

A. Špokas, A. Timinskas

It was analyzed dependence of the average accuracy of secondary protein structure prediction on various GOR algorithm modifications. In essence new modification has expanded informational parameter set by taking into account secondary structure of neighboring amino acid. The accuracy of the method with this modification increases almost for 1 percent. This increase in accuracy is quite significant, as the amount of informational parameters increases not much.