

Konstruktura dendrogramov v taxonomii kvasiniek

ANNA KOCKOVÁ-KRATOCHVÍLOVÁ, Chemický ústav Slovenskej akadémie vied, Bratislava

582.282.232
683.13

V svojom poslednom článku v tomto časopise [1] som uviedla postup numerického vyhodnocovania vzťahov podobnosti medzi kmeňmi kvasiniek jednoduchým koeficientom podľa Jaccarda a spôsob kódovania znakov. Problematika, týkajúca sa výberu a štatistického spracovania znaku, rôzne spôsoby výpočtov koeficientov podobnosti alebo taxonomických dištancií, konštrukcie dendrogramov atp. je veľmi rozsiahla. Mnohé z týchto bodov som diskutovala v iných prácach [2, 3]. Toto pojednanie má slúžiť ako priame pokračovanie už uvedených metód a má uceliť najjednoduchší postup v numerickej taxonomii od výberu znaku až po vytvorenie prirodzených skupín taxonomických jednotiek.

Postupov, akými sa dendrogramy konštruujú, je niekoľko; tu však uvediem osvedčený a jednoduchý postup, ktorý tiež odporúča *Sneath* [4] na základe výpočtov skupinových priemerov. Príklad je uvedený na malej matici koeficientov podobnosti podľa Jaccarda (tabuľka 1),* ktoré boli vypočítané z matice 95 operatívnych znakových jednotiek pre prvý kvasný typ rodu *Candida*, *Torulopsis* a *Saccharomyces*. Pre názornosť metodického postupu tu uvádzam len skupinu druhov rodu *Candida* [*C. albicans*, 29-3-, *C. clausenii*, 29-31-, *C. stellatoidea*, 29-64- a dve prechodné formy [5]]. Matica je uvedená ako štvorcová, ačkoľvek by mohla byť udaná len ako trojuholníková, nakoľko ide o maticu symetrickú. Jednotky hlavnej diagonály tejto identitnej matice, sú nahradené komunalitami. Táto úprava matice je volená preto, aby lepšie ukázala výpočet prvého faktora a komunalit.

Pri konštrukcii dendrogramu na základe skupinových priemerov sa odporúča vychádzať z miesta najvyšších koeficientov v matici, napr. *

$$\bar{A}_3 = (0,925 + 0,950 + 0,950) : 3 = 0,942$$

pričom \bar{A}_n označujeme skupinový priemer trojuholníkovej parciálnej matice. Uvedený príklad ho-

vorí, že kmene 29-3-96, 29-3-10 a 29-3-17 sa ponášajú hladine 94,2 % podobnosti. Ak k nim priberieme ďalší kmeň 29-3-19, musíme pripočítat podobnosti ku každému z prvých troch:

$$\bar{A}_6 = (0,925 + 0,950 + 0,950 + 0,900 + 0,948 + 0,925) : 6 = 933, \text{ čo značí, že tieto štyri kmene majú } 93,3\% \text{ podobnosť.}$$

S pribúdaním kmeňov do skupiny, musí priemerná podobnosť stále klesať. Akonáhle by zrazu stúpila, ukázalo by to na zaradenie kmeňa na nesprávne miesto v skupine, tj. že kmeň môže patriť k susednej skupine, alebo dokonca že stojí osobitne medzi dvomi skupinami ako tranzitná forma. Výpočet týchto priemerných podobností môže byť preto dobrou kontrolou správnosti zostavenia matice koeficientov podobnosti.

Dobрым ukazateľom pre zostavu kmeňov v matici a pre tvorbu skupín dendrogramu je výpočet prvého faktora počtom centroidu. Tento počet je jednou z metód faktorovej analýzy [6], ktorá sa dá použiť pre malé matice aj bez samočinného počítača. Ačkoľvek faktorová analýza používa ako premenné korelačné koeficienty, je možné použiť tento počet aj tu pre koeficienty podobnosti, ktoré sú v rozmedzí 0 až 1. Prvým krokom ku každému spôsobu faktorizácie je výpočet komunality:

$$h^2_j = \frac{S_j^2}{T},$$

pričom h^2_j je komunalita, S_j je súčet koeficientov v stĺpci alebo v riadke a T (total) je súčet všetkých koeficientov v štvorcovej matici vrátane prvkov hlavnej diagonály. Pravda, toto je len jeden z mnohých spôsobov výpočtu komunality. Postupuje sa tak, že najprv sa na miesto jednotiek v hlavnej diagonále identickej matice dosadia najvyššie koeficienty príslušného stĺpca. Po vypočítaní h^2_j sa zmenia za ne vypočítané hodnoty. Štvorcová matica s vypočítanými komunalitami (tabuľka 1) slúži potom k výpočtu prvého faktora, a_1 (resp. a_{j1}):

* V tabuľkách sú uvádzané bodky pred číslom, čo značí 0,...

Tabuľka 1

Matica koeficientov podobnosti medzi kmeňmi

	-37	-52	-96	-17	-10	-15	-82	-19	-109	-75	-31-1	-31-2	-31-3	-64-1	-64-2	-64-3	-64-4	9	16	S _j	a _j
29-3-37																					
29-3-52	854																				
29-3-96	846	780																			
29-3-17	900	875	885																		
29-3-10	923	897	925	885																	
29-3-15	923	850	950	950	877																
29-3-82	880	829	860	857	837	831															
29-3-19	850	871	878	875	853	902	809														
29-3-109	948	846	900	948	925	904	875	859													
29-3-75	868	794	897	897	947	804	800	871	825												
29-31-1	850	850	925	875	900	837	853	875	846	831											
29-31-2	720	738	777	727	727	733	720	704	720	755	712										
29-31-3	813	785	777	795	795	800	790	772	750	733	906	749									
29-64-1	790	785	777	795	795	800	790	772	750	733	906	953	749								
29-64-2	743	641	769	743	769	731	692	769	743	659	644	644	738	916							
29-64-3	717	666	743	692	717	707	666	717	710	763	743	659	644	644	738	942					
29-64-4	717	680	743	717	743	707	641	743	736	743	659	622	622	916	916	699					
9	692	666	794	717	743	682	717	692	777	769	604	580	590	852	823	823	669				
16	733	680	723	760	760	744	733	739	750	702	702	708	708	681	636	653	622	677			
S _j	15,494	15,875	15,794	15,082	15,232	13,889	14,347	13,721	13,496	13,961	279,338										
a _j	926	885	949	940	944	910	902	932	911	915	830	858	858	845	820	822	807	812	835		

$$T = 279,338$$

$$\sqrt{T} = 16,713$$

$$1/\sqrt{T} = 0,0598$$

Tabuľka 2

Poradie prvých faktorov a príklad výpočtu skupinových priemerov

Kmeň	a _j	ΔS _n	n	ΔS _n	Skupiny kmeňov
29-3-96	0,949				
29-3-10	0,944				
29-3-17	0,940	2,825	3	0,942	-96, -10
29-3-19	0,932	5,598	6	0,933	-96, -10, -17, -19
29-3-37	0,926	9,292	10	0,929	-96, -10, -17, -19, -37
29-3-75	0,915	13,717	15	0,914	-96, -10, -17, -19, -37, -75
29-3-109	0,911	19,043	21	0,907	-96, -10, -17, -19, -37, -75, -109
29-3-15	0,910	25,022	28	0,894	-96, -10, -17, -19, -37, -75, -109, -15
29-3-82	0,902	31,908	36	0,886	-96, -10, -17, -19, -37, -75, -109, -15, -82
29-3-52	0,885	39,566	45	0,879	-96, -10, -17, -19, -37, -75, -109, -15, -82, -52
29-31-2	0,858	atď	atď	atď	atď
29-31-3	0,858				
29-64-1	0,845				
16	0,835				
29-31-3	0,830				
29-64-3	0,822				
29-64-2	0,820				
9	0,812				
29-64-4	0,807				

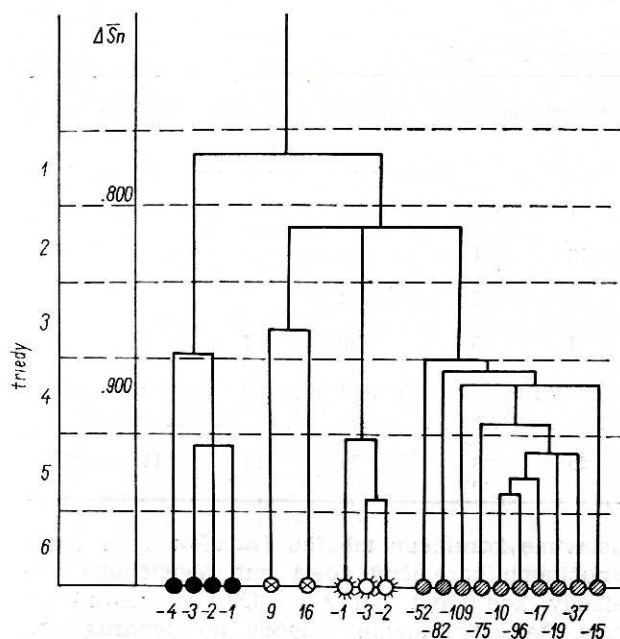
$$a_j = S_j \cdot 1/\sqrt{T}$$

Tak napr. v uvedenej matici bolo $\sqrt{T} = 16,713$ a $1/\sqrt{T} = 0,0598$. Súčty S_j sa preto násobia 0,0598.

Kmeň, pre ktorý bola vypočítaná najväčšia hodnota prvého faktora (v danom prípade prislúcha kmeňu 29-3-96 hodnotou $a_j = 0,949$) predstavuje typického reprezentanta danej skupiny. Pri výpočtoch priemerných podobností sa teda vychádza od tohoto kmeňa a k nemu sa hľadajú najtesnejšie vzťahy. Tabuľka 2 ukazuje, akým spôsobom výpočet prvého faktora usmerňuje tento výber do skupín. Veľkosti prvého faktora v tejto skupine ukazujú dôležitú skutočnosť, že forma 16 je blízka

skupinke kmeňov *C. clausenii* a forma 9 *C. stellatoidea*. To potvrdzuje správnosť nášho pôvodného zistenia (5), aj keď tentoraz uvedená matica koeficientov podobnosti bola vypočítaná na základe značne rozšíreného množstva nových informácií. No, kmene v skupinách sa neusporadúvajú jednoducho za sebou podľa veľkosti prvého faktora (tabuľka 2), ale posudzuje sa vzťah každého nového kmeňa ku kmeňom už zaradeným do skupiny z oboch strán. Kmeň sa potom umiestňuje tam, kde má tesnejší vzťah.

Testovanie príslušnosti kmeňa do skupín sa môže previesť aj *t*-testom (7, 8, 9) na základe výpočtu združenej smerodajnej odchylky:



Obr. 1. Dendrogram 19 kmeňov *C. albicans*, *C. claussenii* a *C. stellatoidea*

$$s^2 = \frac{(n_1 - 1) \cdot (s_1)^2 + (n_2 - 1) \cdot (s_2)^2}{n_1 + n_2 - 2}$$

$$t = \frac{d}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

príčom d je rozdiel medzi $\Delta \bar{S}_n$ a \bar{S}_i , teda medzi priemerným koeficientom skupiny a medzi koeficientom kmeňa (alebo priemerom druhej skupinky), s_1 je smerodajná odchylka priemernej hodnoty koeficientov skupiny, s_2 je smerodajná odchylka priemernej hodnoty koeficientov kmeňa (alebo skupinky), n_1 je počet koeficientov prvej skupiny, n_2 počet koeficientov hodnoteného kmeňa, odpovedajúcich vzťahom tohoto kmeňa ku všetkým

kmeňom porovnávané skupiny. Tabelárna hodnota t sa hľadá podľa združeného stupňa voľnosti $n_1 + n_2 - 2$ na zvolenej hladine pravdepodobnosti (0,05 alebo 0,01 P). Ak vypočítaná hodnota t prevyšuje hodnotu nájdenú v tabuľkách, bol rozdiel medzi priemernou hodnotou kmeňa a skupiny signifikantný a daný kmeň nemôže byť pojatý do skupiny. Pre výpočet smerodajnej odchylky odporúča Sokal a Sneath [4] aproximáciu:

$$\sigma = \sqrt{\frac{S \cdot (1 - S)}{n}}$$

kde S je koeficient podobnosti a n počet operatívnych znakových jednotiek. Napr. overenie príslušnosti kmeňa 29-3-52 do skupiny ostatných deviatich kmeňov *C. albicans*:

$$\Delta \bar{S}_{36} = 0,886, n = 36, s_1 = 0,03941 \text{ pre skupinu (tabuľka 2)}$$

$$\Delta \bar{S}_9 = 0,851, n_2 = 9, s_2 = 0,04417 \text{ pro kmeň 29-3-52}$$

Združená smerodajná odchylka:

$$\Delta s^2 = \frac{35,0,001553 + 8,0,001951}{43} = 0,001627$$

$$t = \frac{0,886 - 0,851}{0,040 \sqrt{\frac{1}{36} + \frac{1}{9}}} = \frac{0,035}{0,01503} = 2,329$$

Tabelárne hodnoty t : $t_{0,05,43} = 1,96$, $t_{0,01,43} = 2,57$

Rozdiel medzi priemerom skupiny a priemerom kmeňa je pri $P = 0,05$ preukazný, pri $P = 0,01$ je nepreukazný. Kmeň sa teda svojím umiestnením javí ako celkom krajný.

Podľa uvedenej matice (tabuľka 1) možno zostrojiť dendrogram (obr. 1), v ktorom jednotlivé skupiny sú zhrnuté na spoločnú priemernú hladinu podobnosti. Rozdelenie do skupín súhlasí s asimiláciou cukrov: sacharózy, trehalózy a sorbózy. Vytvorené prirodzené skupiny nazývajú Sokal a Sneath [4] fenóny. Fenóny sú charakterizované spoločnou hladinou priemernej podobnosti, ktorú možno označiť ako fenónová línia. Tak desať kme-

Tabuľka 3

Matice pre výpočet kofenetických korelácií

	-37	-52	-96	-17	-10	-15	-82	-19	-109	-75	-31-1	-31-2	-31-3	-64-1	-64-2	-64-3	-64-4	9	16
29-3-37	4	5	5	5	4	4	5	4	4	4	2	2	2	1	1	1	1	2	2
29-3-52	846	4	4	4	4	4	4	4	4	4	2	2	2	1	1	1	1	2	2
29-3-96	900	875	5	5	4	4	5	4	4	4	2	2	2	1	1	1	1	2	2
29-3-17	923	897	925	5	4	4	5	4	4	4	2	2	2	1	1	1	1	2	2
29-3-10	923	850	950	950	4	4	5	4	4	4	2	2	2	1	1	1	1	2	2
29-3-15	880	829	860	857	837	4	4	4	4	4	2	2	2	1	1	1	1	2	2
29-3-82	850	871	878	875	853	902	4	4	4	4	2	2	2	1	1	1	1	2	2
29-3-19	948	846	900	948	925	904	875	4	4	4	2	2	2	1	1	1	1	2	2
29-3-109	868	794	897	897	947	804	800	871	4	4	2	2	2	1	1	1	1	2	2
29-3-75	850	850	925	875	900	837	853	875	846	2	2	2	2	1	1	1	1	2	2
29-31-1	720	738	777	727	727	733	720	704	720	755	5	1	1	1	1	1	1	2	2
29-31-2	813	785	777	795	795	800	790	772	750	733	906	5	1	1	1	1	1	2	2
29-31-3	790	785	777	795	795	800	790	772	750	733	906	953	1	1	1	1	1	2	2
29-64-1	743	641	769	743	769	731	692	769	763	743	659	644	644	5	5	3	1	1	1
29-64-2	717	666	743	692	717	707	666	717	710	743	681	622	644	942	5	3	1	1	1
29-64-3	717	684	743	717	743	707	641	743	736	743	659	622	622	916	916	3	1	1	1
29-64-4	692	666	794	717	743	682	717	692	777	769	604	590	590	852	823	823	1	1	1
9	733	680	723	760	760	744	733	739	750	702	702	708	708	681	636	653	622	3	3
16	727	720	777	750	733	777	767	727	727	735	739	744	744	697	674	674	674	866	

Tabuľka 4

Korelačná tabuľka pôvodných koeficientov podobnosti a kofenetických tried

S	·560-·600	·600-·640	·640-·680	·680-·720	·720-·760	·760-·800	·800-·840	·840-·880	·880-·920	·920-·960	Σ
Kof. tr.											
1	2	6	14	19	12	7					60
2			1	7	29	18	1				56
3							2	2			4
4						1	5	20	7	2	35
5								1	6	9	16
Σ	2	6	15	26	41	26	8	23	13	11	171

ňov *C. albicans* má spoločnú fenónovú líniu 0,879, tri kmene *C. clausenii* 0,922 a štyri kmene *C. stellatoidea* 0,879.

Pretože podobnosť na určitej fenónovej línii je priemerná, znamená to, že niektoré koeficienty sú vyššie a iné nižšie ako hodnota tejto línie. V dôsledku toho musí nastať určité skreslenie pri konštrukcii dendrogramu oproti základnej matici koeficientov podobnosti. O veľkosti tohto skreslenia sa možno presvedčiť tzv. kofenetickou koreláciou, ktorú odporúča Sokal a Rohlf (10). Dendrogram rozdelíme priečnymi vodorovnými čiarami na pravidelné triedy tak, že najnižšia trieda je tam, kde je najnižšia podobnosť (obr. 1.). Pri počte kmeňov do 10 sa odporúča urobiť najviac štyri triedy, pri počte kmeňov 100 však najmenej 10 tried. Dendrogram na obr. 1 sme rozdelili na 6 tried, avšak užitočných bolo len 5. Potom sa zostrojí nová matica z kofenetických hodnôt. Kofenetické hodnoty, prvky novej matice, sú čísla tried, v ktorých sa nachádza spoločná priemerná hladina dvoch kmeňov. Tabuľka 3 predstavuje dve polovičné matice: dole vľavo pôvodnú maticu koeficientov podobnosti a hore vpravo maticu kofenetických hodnôt. Prvky hlavnej diagonály nie sú vyznačené. Z týchto dvoch matíc

zostavíme korelačnú tabuľku (tabuľka 4) a z nej vypočítame korelačný koeficient. Vypočítaný korelačný koeficient bol $r = 0,939$, čo nasvedčuje tomu, že usporiadanie kmeňov do dendrogramu bolo urobené pomerne presne (korelačná závislosť nie je síce 100 %, ale je veľmi tesná 93,9 %).

Dendrogramy nielen lepšie vystihujú vzťahy medzi kmeňmi, ako schematicky zakreslené matice koeficientov podobnosti, ale ukazujú aj priamo reprezentanta každej skupiny za daných podmienok. Tým sa môže označiť kmeň, ktorý má najvyššiu hodnotu centroidného prvého faktora (v tomto prípade kmeň 29-3-96). Ak sa tento výpočet použije pri typizácii dostatočne veľkej skupiny kmeňov jedného druhu, môže sa tak dospieť k určitému typickému predstaviteľu v štatistickom smysle. Okrajové kmene v takom prípade vymedzujú rozsah prirodzenej taxonomickej skupiny. Ak sa faktorizácia použije pre hodnotenie korelácií medzi znakmi (ktoré boli vopred štandardizované na nulový priemer a jednotkovú štandardnú deviáciu vzhľadom k ich meraniu v rôznych mierach), slúži pre stanovenie charakteristických vlastností taxonomickej skupiny a k extrakcii najvýznamnejších znakov. Ak vyberáme znaky pre numerické spracovanie

Gumovité látky ječmene a jejich chování při sladování a výrobě piva. I. Vymezení problému a analytické metody

Z rešeršního přehledu vyplývá, že otázkou vlivu gumovitých látek a hemicelulóz ječmene, popř. obilních surogátů, na technologii sladu a piva se zabývalo mnoho autorů. Ve výsledcích jsou však podstatné rozdíly, vysvětlitelné různou jakostí surovin i různým technologickým postupem. Porovnatelnost snižuje také nejednotnost používaných analytických metod.

Autoři přezkoušeli z širšího hlediska chování gumovitých látek v technologickém procesu pokusnými řadami téhož výchozího materiálu, definovaného odrůdou, výrobním místem a ročníkem a dobře reprodukovatelnými analytickými metodami. O výsledcích budou referovat v dalším sdělení.

Po podrobném přezkoušení metody stanovení gumovitých látek podle Preece a spol. a podle Meredith a spol., použili standardní metody, charakterizované tímto postupem:

povaření ječné nebo sladové moučky v etanolu (85 %) pod zpětným chladičem — filtrace — vysušení zbytku; nové mletí — extrakce roztokem papainu (0,025 %); filtrace — působení kyseliny trichloroctové na filtrát; filtrace — vysrážení gumovitých látek z filtrátu stejným objemem acetonu;

rozpuštění sraženiny ve vodě;

vysrážení roztokem síranu amonného 60 g/100 ml.

V takto získané směsi gumovitých látek lze chromatografií na papíru určit obsah glukózy, arabinózy a xylozy, z něhož se vypočítá množství β -glukanu, popř. pentozaňů.

V mladině a pivu se určují jako ve výluhu z ječmene po srážení kyselinou trichloroctovou.

Schuster, K. + Narziss, L. - Kumada, J.: Über die Gummistoffe der Gerste und ihr Verhalten während der Malz- und Bierbereitung. I. Problemstellung und Analysemethoden. = „Brauwissenschaft“, 20, 1967: 125.

Lhotský

covanie veľkých skupín kmeňov, vyvíjajúcich sa v podmienkach technológií (ako u *S. carlsbergensis*, *S. cerevisiae* atď.) slúži toto vyhodnocovanie súčasne aj k výberu najosvedčenejších produkčných kmeňov a k označeniu ich významných vlastností.

Literatura

- [1] Kocková-Kratochvílová, A.: Numerická metoda v taxonomii kvasinek. = „Kvasný průmysl“, 12, 1966: 169–175.
- [2] Kocková-Kratochvílová A.: Problem of the Choice and Number of Characters for Computing Coefficients of Similarity Between Strains of Yeasts. = „Folia microbiologica“, 13, 1968: 310–316.
- [3] Zur Problematik der numerischen Taxonomie bei der Hefen. = „Mitteilungen der Versuchstation für das Gärungsgewerbe in Wien“, 22, 1968: 65–73.
- [4] Sokal, R. R. - Sneath, P. H. A.: Principles of numerical taxonomy. W. H. Freeman San Francisco and London, 1963.
- [5] Kocková-Kratochvílová, A. - Šandula, J.: The genus *Candida* Berkhout. VI. Intermediate forms between fermentation types I and IV. = „Folia microbiologica“, 9, 1964: 369–373.
- [6] Harman, H. H.: Modern factor analysis, University of Chicago Press, 1962 2. Ed.
- [7] Bennett, C. A. - Franklin, N. L.: Statistical analysis in chemistry and chemical industry. New York, John Wiley & Sons, 1954.
- [8] Tsukamura, M.: Adansonian classification of Mycobacteria. = „J. gen. microbiol.“ 45, 1966: 253–273.
- [9] Sedransk, J.: An Application of sequential sampling to analytical surveys. = „Biometrika“, 53, 1966: 85–97.
- [10] Sokal, R. R. - Rohlf, F. J.: The comparison of dendrograms by objective methods. = „Taxon“, 11 [2], 1962: 33–40.

Došlo do redakcie 18. 1. 1968.

КОНСТРУКЦИЯ ДЕНДОГРАММ В ТАКСОНОМИИ ДРОЖЖЕЙ

Метод, разработанный Жакаром дает возможность аналитическо-математической оценки степени подобия разных штаммов дрожжей с помощью простого коэффициента и достаточно точного кодирования знаков. В статье рассматриваются разные методы конструирования дендрограмм дрожжей и рекомендуется метод, основанный на расчетах средних значений для групп, как дающий достоверные результаты и проверенный на практике.

KONSTRUKTION DER DENDROGRAMME IN DER TAXONOMIE DER HEFEN

Der Artikel knüpft an eine frühere Arbeit an, in der Methoden der numerischen Auswertung der Ähnlichkeitsverhältnisse zwischen Hefestämmen mittels des einfachen Koeffizienten nach Jaccard und des Code der Merkmale verfolgt wurden. Von den verschiedenen Verfahren zur Konstruktion der Dendrogramme erörtert die Autorin eine einfache und bewährte Methode, die auf der Berechnung der Gruppendurchschnitte basiert.

CONSTRUCTION OF DENDROGRAMS IN THE TAXONOMY OF YEAST

The Jaccard method permits to evaluate numerically the similarity of various strains of yeast by applying a simple coefficient and to code their signs. From several methods which are used at present for elaborating yeast dendrograms one, based on the calculations of mean values for whole groups, can be recommended as a well-proven, reliable and simple.