

Kahendklasterdamine

Ants Aader
Eesti Biokeskus
ants.ät.aader@punkt.com

Andmekaevandamise uurimisseminar MTAT.03.169.
Arvutiteaduse instituut, Tartu Ülikool
Detsember 2003, lk. 188–198

Kokkuvõte

Seoses geeniekspressiooni andmete mahu järsu suurenemisega tekkis vajadus ka tohutute andmete suures mahus töötlemiseks. Heaks meetodiks osutus kahendklasterdus, mille ülesehitus võimaldab välistada lihtklasterdamise puudusi. Kahendklasterdus sobib suuremahuliste kiipide andmete analüüsiks, kus on tegu erinevatele katsetingimustele vastavate geeniekspressiooni tasememuutustega, samuti võimaldab see uurida geene, mis korruga osalevad mitmes, näiteks sünteesi- ja metaboolses rajas ning stressivastuses, rakutsüklis ja mujal, kas ennustatud fenomen eksisteerib või mitte.

1 Kahendklasterdamise sissejuhatus

Bioloogias populaarne klasterdamine ei võimalda kiibiandmete puhul leida andmete hulgast kuigi palju informatsiooni.

Lihtsal klasterdamisel on kolm suurt puudust, mis on kiibiandmete analüüsil eriti olulised:

1. rühmadesse klasterdamisel saab iga liige olla vaid ühes rühmas;
2. iga liiga peab kuhugi rühma kuuluma;
3. rühmitatakse kõiki geene vastavalt tema käitumisele kõigis eksperimendi tingimustes.

Selline lihtne klasterdus sobib ainult **ühe** selge piiritlusega funktsionaalse või mõnel muul alusel liigendatud rühma uurimiseks. Näiteks lihtklasterdamise võimalustele ei vasta nii stressivastuses kui rakutsüklis osalevate geenide klasterdamine, kuna need funktsioonid on tihti lähedalt seotud. Teine piirang ilmneb näiteks vähi kliinilistes uuringutes, kus on mitmed eri vähikoed, aga klasterdamisel saame me leida vaid ühe efekti, mis kindlasti pole kõikidele vähikudedele omane.

Neid puudusi õnnestub suurel määral vähendada kahendklasterdusel, mis kirjeldati seitsmekümnendatel ja kasutati mitmetes valdkondades, enne kui Cheng ja Church (Cheng & Church 2000) seda geeniandmete analüüsil rakendasid. Kahendklaster oli nende määratluse järgi ühetaoline alammaatriks (millel on madal keskmise jäägi ruut hinnang). Lazzeroni ja Owen (Lazzeroni & Owen 2000) töid sisse ruutmustri (*plaid*) mudeli, mis vaatlleb sisendmaatriksit kui muutujate lineaarset funktsiooni, mis vastab kahendklastritele. Tanay, Sharan ja Shamir (Tanay, Sharan, & Shamir 2002) arendasid välja graafi teoorial ja statistilisel modelleerimisel põhineva meetodi SAMBA (Statistical-Algorithmic Method for Bicluster Analysis).

Järgnevalt vaatleme neid kolme meetodi lähemalt.

2 Cheng ja Church'i algoritm

Cheng ja Church'i algoritm on kõige esimene kahendklasterduse kasutus bioloogias, see on lihtne, toores ja töötav. Kõik hilisemad tööd on juba idee (bioloogiliste andmete kahendklasterdamine) erinevad teostused.

Olgu X geenide hulk ja Y katse tingimuste hulk. Olgu a_{ij} ekspressioonimaatriksi A element, mis väljendab i -nda geeni j -ndal tingimusel suhtelist mRNA hulga logaritmi. Olgu $I \subset X$ ja $J \subset Y$ geeni ja katsetingimuste alamhulgad. (I, J) paar kirjeldab A_{IJ} submaatriksi, mille keskmine ruutjäägi hinnang (mean squared residue) on:

$$H(I, J) = \frac{1}{|I||J|} \sum_{I \subseteq I, j \subseteq J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \quad (1)$$

kus

$$a_{iJ} = \frac{\sum_{j \in J} a_{ij}}{|J|}, \quad a_{Ij} = \frac{\sum_{i \in I} a_{ij}}{|I|} \quad (2)$$

ja

$$a_{IJ} = \frac{\sum_{i \in I, j \in J} a_{ij}}{|I||J|} = \frac{\sum_{i \in I} a_{iJ}}{|I|} = \frac{\sum_{j \in J} a_{Ij}}{|J|} \quad (3)$$

on rea ja veeru keskmised ning alammaatriksi (I, J) keskmine. A_{IJ} alammaatriksit nimetatakse δ -kahendklastriks, kui $H(I, J) \leq \delta$. Kahendklasterdamise algoritm otsib δ -kahendklastrit eeldades, et parameeter δ on valitud sobivalt, et ära hoida juhuslikku signaali äratundmist. Näiteks, me valime δ klasterdamisalgoritmi väljundi vähimaks (parimaks) hinnanguks.

Suurima δ -kahendklastrit identifitseerimise optimeerimisprobleemid (kus $|I| = |J|$ on suurim) on NP (*non-deterministic polynomial time*) keeruline ülesanne ning seda võib taandada täieliku tasakaalus kahealuselisest alamgraafi (balanced complete bipartite subgraph).

Lihtne ja toores δ -kahendklastrit otsimise algoritm võib alustada täismaatriksist ja igal sammul proovib lisada/kustutada rida/veergu, kui see parandab hinnangut ja lõpetab kui edasisi tehteid pole või kahendklastrit hinnang on alla δ väärtust. Selline lihtne keskmiste ja jääkväärtuste ümberarvutamine võib suure andmemaatriksi jaoks olla liiga töömahukas. Cheng ja Church'i algoritm kasutab keskmist jääkväärtust, mis võimaldab kiiremat sammu.

Järgnevalt toon ära Cheng ja Church'i toore biklasterdamise algoritmi.

Algoritm 1 (*Ühe sõlme kustutamine*)

Sisend: reaalarvudega maatriksi A ; suurim lubatud keskmine jääkide hinnangu ruut $\delta \geq 0$.

Väljund: δ -kahendklaster A_{IJ} , mis on I rea ja J veeruga A maatriks, mille hinnang pole suurem kui δ .

Algus: I ja J on vastavalt geeni ja tingimuste andmed ning $A_{IJ} = A$.

Iteratsioon:

1. Arvuta a_{iJ} üle $i \in I$, a_{Ij} üle $j \in J$, A_{IJ} ja $H(I, J)$. Kui $H(I, J) \leq \delta$, tagasta A_{IJ} .
2. Leia rida $i \in I$ suurima väärtusega

$$d(i) = \frac{1}{|J|} \sum_{j \in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2$$

ja veerg $j \in J$ suurima väärtusega

$$d(j) = \frac{1}{|I|} \sum_{i \in I} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2$$

kustuta suurima d väärtusega rida või veerg, uuendades kas I või J .

Algoritm 2 (Mitme sõlme kustutamine)

Sisend: reaalarvudega maatriks A ; suurim lubatud keskmine jääkide hinnangu ruut $\delta \geq 0$; mitme sõlme kustutamise läviväärtus $\alpha > 1$.

Väljund: δ -kahendklaster A_{IJ} , mis on A alamaatriks ridadega I ja veergudega J , mille hinnang pole suurem kui δ .

Algus: I ja J on vastavalt geeni ja tingimuste andmed ning $A_{IJ} = A$.

Iteratsioon:

1. Arvuta a_{iJ} üle $i \in I$, a_{Ij} üle $j \in J$, A_{IJ} ja $H(I, J) \leq \delta$, tagasta A_{IJ}
2. Kustuta read $i \in I$ kus

$$\frac{1}{|J|} \sum_{j \in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 > \alpha H(I, J)$$

3. Arvuta uuesti a_{Ij} , a_{IJ} ja $H(I, J)$.
4. Kustuta veerud $j \in J$ kus

$$\frac{1}{|I|} \sum_{i \in I} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 > \alpha H(I, J)$$

5. Kui iteratsiooni käigus pole midagi kustutada, lülitu algoritm 1 peale ümber.

Algoritm 3 (Sõlme lisamine)

Sisend: reaalarvude maatriks A , J ja I märgib δ -kahendklastrit.

Väljund: I' ja J' nii, et $I \subset I'$ ja $J \subset J'$ koos $H(I', J') \leq H(I, J)$ omadustega.

Iteratsioon:

1. Arvuta a_{iJ} üle i , a_{Ij} üle j , a_{IJ} ja $H(I, J)$.

2. Lisa veerg $j \notin J$, kus

$$\frac{1}{|I|} \sum_{i \in I} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \leq H(I, J)$$

3. arvuta uuesti a_{iJ} , a_{IJ} ja $H(I, J)$.

4. lisa rida $i \notin I$, kus

$$\frac{1}{|J|} \sum_{j \in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \leq H(I, J)$$

5. Iga rida i , mis pole veel I , lisa tema pööratud rida kui

$$\frac{1}{|J|} \sum_{j \in J} (-a_{ij} + a_{iJ} - a_{Ij} + a_{IJ})^2 \leq H(I, J)$$

6. kui iteratsiooni käigus pole midagi lisatud, tagasta lõplik I ja J kui I' ja J' .

Algoritm 4 (Etteantud arvu kahendklastrite leidmine)

Sisend: A on reaalarvude maatriks, kus võivad olla puuduvad elemendid; $\alpha \geq 1$ on mitme sõlme kustutamise parameeter; $\delta \geq 0$ on suurim lubatud keskmine jääkide hinnangu ruut ja n on etteantud leitav δ -kahendklastrite arv.

Väljund: n δ -kahendklastrit.

Alustamine: A maatriksi puuduvad väärtused asendatakse olemasolevate arvude ulatuses juhuslike arvudega. A' on A koopia.

n -kordne iteratsioon:

1. Kasuta algoritmi 2 A' maatriksil. Kui rida (veerg) on väike (alla 100), siis ei sooritata real (veerul) mitmest sõlme kustutamist. Tulemuseks on maatriks B .
2. (algoritmi 2 viies samm) kasuta algoritmi 1 B maatriksil, tulemuseks on C maatriks.
3. kasuta algoritmi 3 A ja C maatriksil, tulemuseks on kahendklaster D .
4. Esitle D ja asenda kõik elemendid maatriksil A' , mis on olemas ka maatriksil D , juhuslike arvudega.

3 Ruutmuster (*plaid*)

Geeniekspressiooni andmete analüüsimiseks töid Lazzeroni ja Owen sisse ruutmustri (*plaid*) mudeli (Tanay, Sharan, & Shamir 2002). Ka selle analüüsi idee pärineb statistikast. Ruutmustri mudel lubab geenil olla enam kui ühes klastris või siis mitte üheski, lubab geeniklastreid lahterdada vaid vastavalt osadele (mitte kõikidele) tingimustele, näiteks pärmi pungumisel (ja ainult selles protsessis) osalevad geenid ühte klastrisse ja teistes protsessides (ja mitte pungumisel) osalevad geenid teistesse klastritesse.

Ekspressioonimaatriksi sisend on $A = (A_{ij})$, kus $i = 1, \dots, n$ on geenid ja $j = 1, \dots, p$ on katsetingimused. A_{ij} kirjeldab j katse i geeni ekspressioonitaset. Kujutame $n \times p$ maatriksit kui tabelit, kus iga lahterkirjeldab vastavalt tema väärtusele teatud värvuse. Seejärel võime read ja veerud niimoodi ümber reastada, et sarnase väärtusega lahtrid satuksid võimalikult lähedale - millest tulenevalt tekib ruutmuster. Sarnaselt reageerib geenirühm tekitab „kihi“ (*layer*), mis on ruutmustri taustal „kahendklastrite“ sünonüüm.

Ideaalne massiivi ümberreastamine peaks looma pildi, kus diagonaalil on K arv ruudukujulisi blokke. Iga blokk võiks olla unikaalselt värvunud ja sellest blokkist väljapoole jääv osa peaks olema neutraalse taustaväärtusega. See ideaal vastab K eristunud geeniklastrile ja proovi (katse) K -osaks jagunemisele. Iga k geeniblokis olev geen ekspresseerub katse k blokis. Matemaatiliselt väljendude

$$A_{ij} = \mu_0 + \sum_{k=1}^K \mu_k p_{ik} k_{jk} \quad (4)$$

kus μ_0 on taustavärv, μ_k kirjeldab k bloki värvi, p_{ik} on geenibloki kuuluvust näitav indikaator — p_{ik} on 1 kui geen i kuulub k -ndasse geeniblokki ja k_{jk} on katsetingi-

mustebloki kuuluvust näitav indikaatormuutuja. Kui kattuvad kihid on keelatud, peame lisama kitsenduse $\sum_k k_{jk} = 1$ üle kõikide j ja $\sum_k p_{ip} = 1$ üle kõikide i . Üldine mudel kirjeldab andmeid kui tõenäoliselt ülekattuvate kihtide summat, mis ei pea katma tervet massiivi ja kus mitmetele kihtidele omistatakse sama geeni katmine.

Mudel 4 kirjeldab μ_k vastust, mis on jagatud üle kõikide geenide. Bioloogiliselt on huvitav leida geene, millel on identsed kuigi mitte konstantsed vastused teatud tingimustele. Ka vastupidi on huvitav, leida katsete rühm koos üldise lihtsa geeni expressioonimustriga. Järgnevad mudelid toetavad seda:

$$A_{i,j} = \mu_0 + \sum_{k=1}^K (\mu_k + \alpha_{ik}) p_{ik} \kappa_{jk} \quad (5)$$

$$A_{i,j} = \mu_0 + \sum_{k=1}^K (\mu_k + \beta_{ik}) p_{ik} \kappa_{jk} \quad (6)$$

$$A_{i,j} = \mu_0 + \sum_{k=1}^K (\mu_k + \alpha_{ik} + \beta_{jk}) p_{ik} \kappa_{jk} \quad (7)$$

kus iga $p_{ik} \in \{0, 1\}$ ja iga $\kappa_{ik} \in \{0, 1\}$. Kui α_{ik} või β_{ik} on kasutusel, siis $\sum_i p_{ik} \alpha_{ik} = 0$, $\sum_j \kappa_{ik} \beta_{ik} = 0$, et üleparameetriseerimist ära hoida. Siinkohal saame sügavama arusaama ruutmustrit pildimustrist, mis on $\mu_k + \alpha_{ik} \beta_{jk}$ joonis. Mudelid 4 kuni 7 lähendavad pilti kihtide summa abil. Võime kasutada märgendit θ_{ijk} , et kirjeldada kas μ_k , $\mu_k + \alpha_{ik}$ või $\mu_k + \alpha_{ik} + \beta_{jk}$, kui vaja. Segades kihtide tüüpe, saavutame suurema üldistatuse nii, et α_{ik} või β_{ik} võiks olla esindatud mõnes, aga mitte kõigis θ_{ijk} tüüpides. Mudeli võib seega ümber kirjutada kihtide summana:

$$A_{i,j} = \mu_0 + \sum_{k=1}^K \theta_{ijk} p_{ik} \kappa_{jk} \quad (8)$$

Saadud mudel on kahemõõtmelise dispersioonanalüüsi (two-way analysis of variance) superpositsioon üle geenide ja katsete alamrühmade.

Ruutmustrit regressiooni mudelil otsime ruutmustrit, mille puhul muutuja

$$\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^p (A_{ij} - \theta_{ij0} - \sum_{k=1}^K \theta_{ijk} p_{ik} \kappa_{jk})^2 \quad (9)$$

väärtus oleks väike. Igal k kihil on $(2^n - 1)(2^p - 1)$ lähenemist, kuidas valida osalevaid geene ja tingimusi ja kuna see probleem on NP-keeruline, seega kasutatakse teist lähenemist: uued kihid lisatakse mudelile ühekaupa. Eeldame, et meil

on $K - 1$ kihti ja otsime K -ndat kihti, mis vähendaks vigade ruutude summat. Olgu

$$Q = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^p (Z_{ij}^{(K-1)} - \theta_{ijK} p_{iK} \kappa_{jK})^2 \quad (10)$$

kus

$$Z_{ij}^{(K-1)} = A_{ij} - \theta_{ij0} - \sum_{k=1}^{K-1} \theta_{ijk} p_{ik} \kappa_{jk} \quad (11)$$

on esimeste $K - 1$ kihtide jääk.

Regressiooni algoritm kasutab iteratiivset lähenemist, kus iga samm optimeerib θ väärtust, p väärtust või κ väärtust samal ajal kahe teise parameetri perekonna väärtusi fikseerides. Samuti on hea valida κ ja p väärtused pidevas ulatuses, ainult viimas(t)es iteratsioonides omistada neile 0 ja 1 väärtused. Vahepealsetes staadiumides kirjeldab θ_{ijk} „hägust dispersioonanalüüsi“ (fuzzy analysis of variance), kus p_{iK} ja κ_{jK} alati pole 0 ja 1. Tähistagu $\theta^{(s)}$ iteratsioonis s kõiki θ_{ijk} väärtusi. Samaselt, tähistagu $p^{(s)}$ ja $\kappa^{(s)}$ iteratsioonis s kõiki p_{iK} ja κ_{jK} väärtusi. $s = 0$ korral on algväärtus juhuslikult ümber 0,5 leitud.

θ_{ijK} , κ_{jK} ja p_{iK} uuendamine: Üle iga K θ_{ijK} uuendamiseks peame minimeerima:

$$Q^{(K)} = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^p (Z_{ij}^{(K-1)} - (\mu_K + \alpha_{iK} + \beta_{jK}) p_{iK} \kappa_{jK})^2 \quad (12)$$

tingimuste $\sum_{i=1}^n p_{iK}^2 \alpha_{iK} = \sum_{j=1}^p \kappa_{jK}^2 \beta_{jK} = 0$ jaoks. Langrange kordaja argumentid näitavad, et

$$\mu_K = \frac{\sum_i \sum_j p_{iK} \kappa_{jK} Z_{ij}^{(K-1)}}{(\sum_i p_{iK}^2)(\sum_j \kappa_{jK}^2)} \quad (13)$$

$$\alpha_{iK} = \frac{\sum_j (Z_{ij}^{(K-1)} - \mu_K p_{iK} \kappa_{jK}) \kappa_{jK}}{(p_{iK})(\sum_j \kappa_{jK}^2)} \quad (14)$$

$$\beta_{jK} = \frac{\sum_i (Z_{ij}^{(K-1)} - \mu_K p_{iK} \kappa_{jK}) p_{iK}}{(\kappa_{jK})(\sum_i p_{iK}^2)} \quad (15)$$

Seega, s iteratsioonis kasutame neid võrrandeid $p^{(s-1)}$ ja $\kappa^{(s-1)}$ liikmetest $\theta^{(s)}$ uuendamiseks. Iga μ_K uuendus on sama, sõltumata sellest kas K -ndas kiht sisal-

dab α_{iK} või eta_{iK} . Uuendatud α_{iK} ja eta_{iK} on samad vaatamata sellele, kas teine on või mitte kihis kaasatud. p_{iK} ja κ_{jK} väärtused, mis Q väärtust miniseerivad on:

$$p_{iK} = \frac{\sum_j \theta_{ijK} \kappa_{jK} Z_{ij}^{K-1}}{\sum_j \theta_{ijK}^2 \kappa_{jK}^2} \quad (16)$$

$$\kappa_{jK} = \frac{\sum_i \theta_{ijK} p_{iK} Z_{ij}^{K-1}}{\sum_i \theta_{ijK}^2 p_{iK}^2} \quad (17)$$

Iteratsioonis s kasutame neid võrrandeid, et uuendada $p^{(s)}$ muutujatest $\theta^{(s)}$ ja $\kappa^{(s-1)}$ ning uuendada $\kappa^{(s)}$ muutujatest $\theta^{(s)}$ ja $p^{(s-1)}$. Selles iteratsioonis on geenid ja tingimused võrdselt koheldud ja uuendamine on samaväärne kui seda vastupidises järjekorras teha.

Kui K -nda kihi parameetrid $s = 1, \dots, S$ (S on iteratsioonide arv) iteratsioonis oleme uuendanud, kontrollime $(K + 1)$ -nda kihi parameetreid. Iga K kohta, vaadatakse, kas lõpetamiskriteerium kehtib või mitte. Toores algoritm, mis lisab ühe kihi korraga, vajab lõpetamis reeglit. Eeldame, et kõrgematel K väärtustel jääk muutub üha rohkem ja rohkem müraga sarnaseks. Selleks, et ära hoida struktureerimata mürast koosnevate kihtide lisamist, kasutatakse järgevat kriteeriumi: määratleme k kihi tähtsuse $\sigma_k^2 = \sum_{i=1}^n \sum_{j=i}^p p_{ik} \kappa_{jk} \theta_{ijk}^2$. Algoritm tunnistab kihti, kui kihi tähtsus on oluliselt suurem kui müral. Müra σ_k^2 jaotus on tundmatu. Selleks, et määrata müra σ_k^2 väärtus, Lazzeroni ja Owen permuteerisid andmeid:

1. olgu Z_{ij} jääkmaatriks, kus me k kihi jaoks otsime;
2. iga $r = 1, \dots, R$ jaoks, olgu $Z_{ij}^{(r)}$ maatriks, mis on saadud juhuslikult iga Z_{ij} rida ja veergu permuteerides. Kõik permutatsioonid on sõltumatud ja ühtlase jaotusega;
3. tähistagu $\sigma_k^{2,r}$ k kihi tähtsust, mis on leitud juhuslikest andmetest $Z_{ij}^{(r)}$;
4. kui $\sigma_k^2 < \max_{1 \leq r \leq R} \sigma_k^{2,r}$ ja $k < K_{\max}$. siis lisa mudelile uus k kiht.

4 SAMBA

SAMBA (Statistical Algorithmic Method of Bicluster Analysis) arendasid välja Tanay, Sharan ja Shamir (Tanay, Sharan, & Shamir 2002). Nende järgi on kahendklaster teatud tingimustele vastav sarnase ekspressiooni muustriga geenide rühm.

Ekspressiooniandmed on modelleeritud kui kahealuseline graafi d , mille kaks osa vastavad tingimustele ja geenidele.

Formaalselt võime iga geeniekspressiooniandmete kohta luua kaheosalise graafi $G = (U, V, E)$, kus U on tingimused, V on geenid ja $(u, v) \in E$ kus v vastab u tingimustele, mis on kui v ekspressioonitase oluliselt muutub u tingimustes. Kahendklaster vastab G graafi H alamgraafi $H = (U', V', E')$ ja kirjeldab alamrühma V' geene, mis on U' tingimustes koosreguleeritud. Kahendklastrite alamgraafi kaal on geenitingimuste paaride kaalude summa. Selleks, et alamgraafi kaalule statistilist tähendust omistada, arendasid autorid ekspressiooniandmete kahealuselise graafi esitamiseks statistilise mudeli. Mudeleid kasutades võib tuletada vaadeldud alamgraafi H olulisuse hindamise skeeme. Hinnang esitatakse alamgraafi paaride sõltumatu osaluse summamana, tänu millele võime kahendklasterdamise probleemi taandada raskete alamgraafi H leidmise probleemile.

Lihtne mudel on järgmine: olgu $H = (U', V', E')$ G graafi alamgraaf. Määratleme $|U'| = m'$, $|V'| = n'$. Olgu $p = \frac{|E|}{|U||V|}$, ja olgu $k' = |E'|$. Esimene mudel eeldab, et serv on olemas sõltumatult ja samatõenäoselt vastavalt p tihedusele. Määratledes, et $BT(k, p, n)$ on binomiaalne saba, s.o. k vaatluse tõenäosus n katsel, kus iga edukas vaatlus on vastav tõenäosusele p . Graafi nägemise tõenäosus on sama tihe kui H vastavalt mudelile $p(H) = BT(k', p', n'm')$.

Eesmärk on leida alamgraaf H koos madalaima $p(H)$. Eeldades, et $p < \frac{1}{2}$, saame järgmise ülemise raja $p(H) : p^*(H) = 2^{n'm'} p^k (1-p)^{n'm'-k}$. Otsides alamgraafi H , minimiseerides $\log p^*(H)$ on samaväärne kui leida suurima kaaluga G alamgraafi, kus igal serval on positiivne kaal $(-1 - \log p)$ ja igal mitteserval on negatiivne kaal $(-1 - \log(1-p))$.

Keerulisem nullmudel võtab arvesse G tasemete varieeruvuse, s.o. iga geeni ja katsetingimuse iseloomuliku käitumise. Olgu $H = (U', V', E')$ G graafi alamgraaf ja $\overline{E'} = (U' \times V') \setminus E'$. Igal tipul $w \in U' \cup V'$ tähistagu d_w tema astet G graafi s. Nullmudel eeldab, et iga serva (u, v) esinemine on Bernoulli muutujast parameetriga $p_{u,v}$ sõltumatu.

$p_{u,v}$ tõenäosus on kahealuselise graafi fraktsioon olles astme järgnevuses identne graafi G , mis sisaldab (u, v) serva. Tegelikult määratakse $p_{u,v}$ kasutades Monte-Carlo protsessi. Vaatluse H tõenäosus on $p(H) = \left(\prod_{(u,v) \in E'} p_{u,v} \cdot \left(\prod_{(u,v) \in \overline{E'}} (1 - p_{u,v}) \right) \right)$ Siiski ei saa alamgraafi vastavalt tema tõenäosusele võrrelda, sest see väheneb vastavalt H suuruse suurenemisele.

Selleks, et sellest probleemist üle saada, kasutatakse kahendklastrite olulisuse hindamiseks tõenäosuse suhet. Seega alternatiivse mudeli puhul eeldatakse, et iga kahendklastrite serv esineb muutumatu tõenäosusega $p_c > \max_{(u,v) \in U \times V} p_{u,v}$. See

model peegeldab arusaama, et lahendklastrid kirjeldavad ligikaudu ühtlast elementide vahelist suhet. H logaritmiline tõenäosuste suhe on:

$$\log L(H) = \sum_{(u,v) \in E'} \log \frac{p_c}{p_{u,v}} + \sum_{(u,v) \in \overline{E'}} \log \frac{1 - p_c}{1 - p_{u,v}}$$

Määrates igale (u, v) servale kaalu $\log \frac{p_c}{p_{u,v}} > 0$ ja igale (u, v) miteservale kaalu $\log \frac{1 - p_c}{1 - p_{u,v}} < 0$, teeme järelduse, et H hinnang on lihtsalt tema kaal. Kui me arvestame ka iga serva jaoks expressiooni muutuse suunda, on statistiline mudel veelgi keerulisem. Sellest hoolimata arvutatakse tõenäosuse hinnangut samal viisil nagu suunamata kujul.

5 Kokkuvõte

Kahendklasterdamine võimaldab viimaste aastate jooksul bioloogide poolt geeni-kiipide automatiseeritud tehnoloogia elluviimise tulemusena produtseeritud tohutute andmete analüüsi. Erinevalt lihtklasterdamisest võimaldab kahendklasterdamine väga paindlikult arvestada andmete iseloomu ja eripära. Ning kahendklasterdamise tulemused on kergesti visualiseeritavad.

Viited

- Cheng, Y., and Church, G. M. 2000. Biclustering of expression data. In *Proc. ISMB'00*, 93–103. AAAI Press.
- Lazzeroni, L., and Owen, A. 2000. Plaid models for gene expression data.
- Tanay, A.; Sharan, R.; and Shamir, R. 2002. Discovering statistically significant biclusters in gene expression data. *Bioinformatics* 1(1):1–9.