Ανάλυση Γονιδιακής Έκφρασης

Εργασία για το μάθημα
"Εισαγωγή στην Βιοπληροφορική"

Ομάδα:
Ανθοπούλου Ολυμπία 2578
Μαυράκης Αλέξανδρος 2477
Τσιράκης Νικόλαος 2392

ΠΕΡΙΕΧΟΜΕΝΑ

Εισαγωγή

Τα τελευταία χρόνια έχουν γίνει τεράστια βήματα στους τομείς της μοριακής βιολογίας και της γενετικής μηχανικής και σημαντικός παράγοντας που βοήθησε σε αυτό είναι η επιστήμη των υπολογιστών. Και η βοήθεια δεν ήρθε μόνο με την παροχή ικανοποιητικής υπολογιστικής ισχύος και μεθόδων οργάνωσης των δεδομένων, αλλά και με την ενεργή συμμετοχή επιστημόνων των υπολογιστών στην βιολογική έρευνα συνδυάζοντας τις γνώσεις τους με τον τομέα της βιολογίας και αναπτύσσοντας λύσεις σε προβλήματα που είχαν οι βιολόγοι. Στην παρούσα αναφορά αναλύουμε τρόπους με τους οποίους οι ερευνητές αντλούν τις πληροφορίες από τα μακρομόρια, τις αναλύουν και τις συστηματοποιούν με τέτοιο τρόπο ώστε όταν παραδοθούν στον χρήστη να έχουν μορφή και δομή που να τον βοηθάνε να βγάλει εύκολα και γρήγορα κάποια πρώτα συμπεράσματα, ενώ παράλληλα να έχουν διαχωριστεί σε αρκετές περιπτώσεις τα χρήσιμα δεδομένα από τον θόρυβο.

Στην αναφορά αυτή μετά από μία γρήγορη αναφορά στην τεχνολογία των μικροπινάκων, μία τεχνολογία που επιτρέπει την μαζική σάρωση μακρομορίων DNA. Ακολουθεί η αναλυτική παρουσίαση των προσεγγίσεων που υπάρχουν για την οργάνωση των δεδομένων έκφρασης των γονιδίων. Η διαδικασία της οργάνωσης στην ουσία διαχωρίζει τα δεδομένα σε υποσύνολα(συστάδες), βάση κάποιων κριτηρίων. Αυτή η διαδικασία λέγεται κατηγοριοποίηση(clustering).

Εισαγωγή στην τεχνολογία των Μικροπινάκων (microarrays)

Γενικά:

Η τεχνολογία των μικροπινάκων DNA έκανε δυνατή την ταυτόχρονη παρακολούθηση εκατοντάδων γονιδίων κατά την διάρκεια σημαντικών βιολογικών διαδικασιών. Ο εντοπισμός κρυμμένων προτύπων στα δεδομένα έκφρασης των γονιδίων βοήθησε πολύ στην κατανόηση των γενετικών λειτουργιών. Παρόλ' αυτά ο τεράστιος αριθμός γονιδίων και η πολυπλοκότητα των βιολογικών δικτύων αύξησε τις προκλήσεις για κατανόηση και μετάφραση των μαζικών αποτελεσμάτων, τα οποία συχνά προκύπτουν από χιλιάδες μετρήσεις. Ένα σημαντικό βήμα για να αντιμετωπιστεί το πρόβλημα αυτό είναι η χρήση τεχνικών κατηγοριοποίησης (συσταδοποίησης), οι οποίες είναι απαραίτητες στην διαδικασία αλιείας δεδομένων για να ανακαλυφθούν φυσικές δομές και να αναγνωριστούν ενδιαφέροντα πρότυπα.

Η ανάλυση συστάδων (cluster analysis) διαχωρίζει τα δεδομένα σε ομάδες βασιζόμενη σε συγκεκριμένα χαρακτηριστικά, ώστε τα μέλη κάθε ομάδας να μοιάζουν περισσότερο μεταξύ τους από ότι με μέλη άλλων ομάδων. Σημαντικό έργο έχει γίνει τις 3 τελευταίες δεκαετίες στον τομέα αυτό, με συμβατικούς αλγόριθμους ανάλυσης συστάδων να εφαρμόζονται απευθείας στα δεδομένα, άλλους να τροποποιούνται και ακόμα να δημιουργούνται καινούριοι. Αυτοί οι αλγόριθμοι κατηγοριοποίησης έχουν φανεί χρήσιμοι στην αναγνώριση βιολογικά ομοίων ομάδων γονιδίων και δειγμάτων.

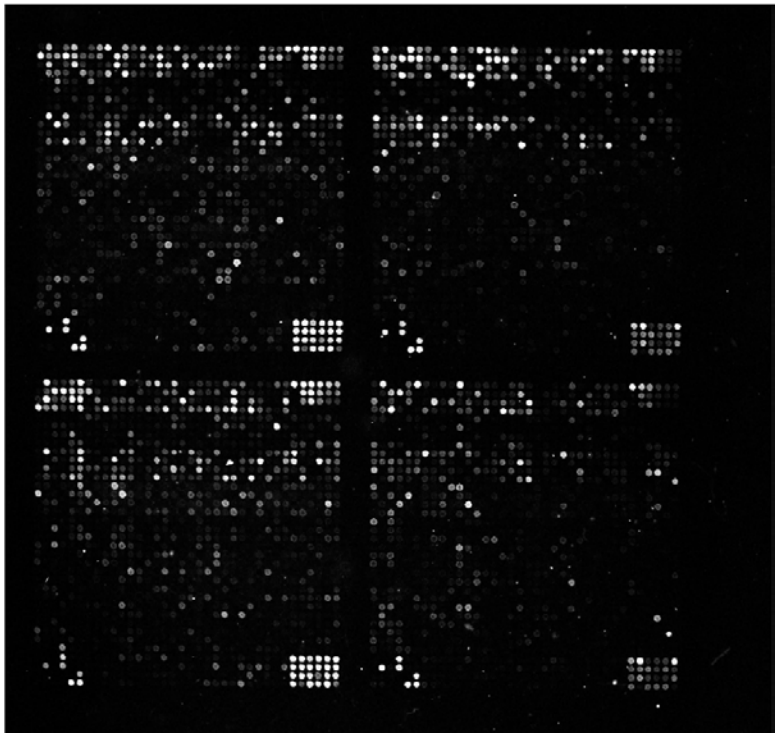
Τεχνολογία μικροπινάκων

Σε αντίθεση με τις συμβατικές προσεγγίσεις στην γονιδιακή έρευνα που εστίαζε στην τοπική μελέτη και συλλογή δεδομένων για μεμονωμένα γονίδια, τεχνολογία των μικροπινάκων έκανε εφικτή την ταυτόχρονη παρακολούθηση των επιπέδων έκφρασης χιλιάδων γονιδίων παράλληλα . Ξεφεύγοντας από την κλασική δομή των πινάκων όπου περιέχουν σειρές με απλά δεδομένα, τα Microarrays μετράνε την «*ποσότητα του στόχου*» έμμεσα μετρώντας μια άλλη φυσική ποσότητα – την ένταση του φθορισμού κάθε τοποθεσίας (*spot*) στον πίνακα αναπαράστασης φθορισμού. Έτσι στην πραγματικότητα οι σειρές δεδομένων που παράγονται από τα Microarrays είναι μονόχρωμες εικόνες. Μετασχηματίζοντας αυτές τις εικόνες μέσα στον "*gene expression matrix*", το οποίο αποτελεί μια δύσκολη διαδικασία, μπορούμε να αναλύσουμε αυτόν τον πίνακα και να προσπαθήσουμε να εξάγουμε από αυτόν κάποια συμπεράσματα και γνώσεις σχετικά με θεμελιώδης βιολογικές διαδικασίες

Δύο είναι οι κύριοι τρόποι πειραματισμού, ο ένας με cDNA μικροπίνακες και ο άλλος με ολιγονουκλεοτιδικούς πίνακες (oligo arrays). Γενικά παρά τις διαφορές τους σε λεπτομέρειες και τα δύο πειραματικά πρωτόκολλα πραγματοποιούν τα παρακάτω βήματα:

- Κατασκευή chip: ένα microarray είναι ένα μικρό chip (από nylon, γυαλί ή σιλικόνη) πάνω στο οποίο προσαρμόζονται μόρια DNA σε καθορισμένο πλέγμα. Κάθε κελί του πλέγματος σχετίζεται με μία ακολουθία DNA.
- Προετοιμασία, χαρακτηρισμός και υβριδοποίηση στόχου: δύο δείγματα mRNA (ένα δείγμα ελέγχου και ένα πειραματικό δείγμα) μεταγράφονται σε cDNA (στόχοι), μαρκάρονται με ραδιοϊσότοπα ή φθοριούχες ενώσεις, και υβριδοποιούνται με τα μόρια στην επιφάνεια του πλέγματος
- Τα chips σαρώνονται για να διαβαστεί η ένταση του σήματος που εκπέμπουν από τους μαρκαρισμένους υβριδοποιημένους στόχους.

Και οι δύο μέθοδοι μετράνε την έκφραση των επιπέδων για κάθε ακολουθία DNA από τον λόγο της έντασης των σημάτων μεταξύ του δείγματος ελέγχου και του πειραματικού δείγματος.

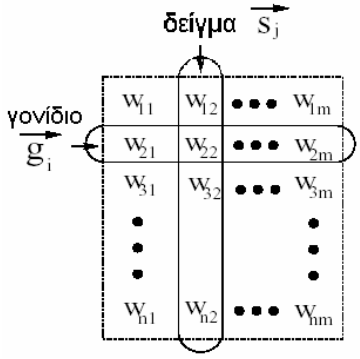


Εικόνα: σάρωση και υβριδοποίηση μικροπίνακα που περιέχει περισσότερα από 5000 γονίδια κάθε σημείο χαρακτηρίζει ένα σύνολο από πανομοιότυπα μονοκλωνικά μόρια DNA που αναπαριστούν ένα μόνο γονίδιο. Η ένταση της φωτεινότητας δείχνει την ποσότητα φθορίζουσας ουσίας που περιέχει το υβριδοποιημένο mRNA.

Προεργασία των δεδομένων έκφρασης των γονιδίων

Η τεχνολογία των μικροπινάκων έχει πρόσβαση σε μεγάλο μέγεθος ακολουθιών DNA κάτω από πολλές συνθήκες. Εδώ ανεπίσημα θα χαρακτηρίζουμε γονίδια κάθε ακολουθία DNA και δείγματα τις πειραματικές συνθήκες. Ένα σύνολο δεδομένων έκφρασης γονιδίων μπορεί να παρουσιαστεί σε έναν πίνακα έκφρασης πραγματικών τιμών

$M=\{w_{ij} \mid 1 \le i \le n, 1 \le j \le m\}$



n	αριθμός γονιδίων
m	αριθμός δειγμάτων
M	ένας πίνακας έκφρασης γονιδίων
w_{ij}	κελί στον πίνακα έκφρασης γονιδίων
$\vec{g_i}$	ένα γονίδιο
$\vec{s_j}$	ένα δείγμα
$G, G', G_0, \ldots$	ένα σετ γονιδίων
$S, S', S_0, \ldots$	ένα σετ δειγμάτων

όπου οι γραμμές είναι εκφράσεις προτύπων στα γονίδια και οι στήλες τα προφίλ των δειγμάτων.

Ο αρχικός πίνακας δεδομένων περιέχει θόρυβο, χαμένες τιμές και αποκλίσεις που προκύπτουν από τις πειραματικές διαδικασίες. Άρα απαιτείται ένα στάδιο προεπεξεργασίας , ώστε να λυθούν αυτά τα προβλήματα, πριν εφαρμόσουμε τις τεχνικές κατηγοριοποίησης. Με το στάδιο την προεπεξεργασίας δεν ασχολούμαστε σε αυτό το σημείο, και θα θεωρούμε πως τα δεδομένα μας έχουν περάσει από αυτό το στάδιο.

Εφαρμογές της κατηγοριοποίησης των δεδομένων έκφρασης γονιδίων

Η χρήση τεχνικών κατηγοριοποίησης έχουν αποδειχτεί ιδιαίτερα χρήσιμες στην κατανόηση της λειτουργικότητας των γονιδίων, των κανόνων που διέπουν τα γονίδια, τις κυτταρικές λειτουργίες και τους υποκατηγορίες των κυττάρων. Γονίδια με παρόμοια πρότυπα ομαδοποιούνται μαζί παρουσιάζοντας παρόμοιες κυτταρικές λειτουργίες και συχνά αποδεικνύεται ότι συνεργάζονται σε ίδιες διαδικασίες των κυττάρων, ενώ επιπλέον μπορεί να σχετίζονται με υποκατηγορίες μορφολογικά όμοιων κυττάρων.

Τεχνικές κατηγοριοποίησης

Κατηγοριοποίηση και τύποι κατηγοριοποίησης δεδομένων έκφρασης γονιδίων

Κατηγοριοποίηση είναι η διαδικασία ομαδοποίησης αντικειμένων σε κλάσεις διαφορετικές ξένες μεταξύ τους, ώστε αντικείμενα σε μία κλάση να έχουν μεγαλύτερη ομοιότητα μεταξύ τους από ότι αντικείμενα άλλων κλάσεων.

Σε δεδομένα έκφρασης γονιδίων, έχουμε την δυνατότητα να καταφύγουμε σε κατηγοριοποίηση βασιζόμενοι σε διάφορα κριτήρια. Αποτέλεσμα είναι να έχουμε:

- κατηγοριοποίηση βάση γονιδίων: στην περίπτωση αυτή τα γονίδια θεωρούνται σαν αντικείμενα και τα δείγματα(συνθήκες) σαν χαρακτηριστικά
- κατηγοριοποίηση βάση δειγμάτων: στην προκειμένη περίπτωση αντικείμενα είναι τα δείγματα, και χαρακτηριστικά πάνω στα οποία γίνεται η κατηγοριοποίηση είναι τα γονίδια.

Η διαφορά μεταξύ των δύο τύπων κατηγοριοποίησης είναι ότι βασίζονται σε διαφορετικά χαρακτηριστικά για να πετύχουν την κατηγοριοποίηση των δεδομένων. Παρόλ'αυτά αλγόριθμοι όπως ο K-means και οι ιεραρχικές προσεγγίσεις, μπορούν να εφαρμοστούν και για τις δύο περιπτώσεις.

Οι δύο μέθοδοι κατηγοριοποίησης βασίζονται στην εξαντλητική και αποκλειστική εύρεση του ίδιου χαρακτηριστικού στα αντικείμενα ώστε να τα χωρίσουν σε ομάδες. Νεότερες όμως προσεγγίσεις στην μοριακή βιολογία θεωρούν ότι μόνο ένα μικρό μέρος

των γονιδίων παίρνει μέρος στις κυτταρικές λειτουργίες. Έτσι έγινε απαραίτητη η εισαγωγή μίας νέας μεθόδου κατηγοριοποίησης:

- κατηγοριοποίηση υποδιαστημάτων (subspace clustering): Η κατηγοριοποίηση εφαρμόζεται σε ένα υποσύνολο των γονιδίων βάση ενός υποσυνόλου των δειγμάτων. Σε αυτή την προσέγγιση τα δείγματα και τα γονίδια αντιμετωπίζονται συμμετρικά, οπότε άλλοτε τα γονίδια και άλλοτε τα δείγματα θεωρούνται αντικείμενα ή χαρακτηριστικά.

Καθεμία από τις τρεις μεθόδους αναλύει τον πίνακα έκφρασης γονιδίων με διαφορετικούς τρόπους και αντιμετωπίζει διαφορετικές προκλήσεις.

Μέτρηση ομοιότητας για δεδομένα έκφρασης γονιδίων

Η *μέτρηση ομοιότητας* υπολογίζει την εγγύτητα (ή διαφορετικότητα) μεταξύ δύο αντικειμένων με δεδομένα.. Αντικείμενα με δεδομένα έκφρασης γονιδίων, μπορούν να τυποποιηθούν να αριθμητικοί πίνακες $\vec{O}_i = \{o_{ij} \mid 1 \leq j \leq p\}$, όπου o_{ij} είναι το j-στο χαρακτηριστικό για το i-στο αντικείμενο δεδομένων και p είναι το πλήθος των χαρακτηριστικών. Η εγγύτητα μεταξύ δύο αντικειμένων O_i και O_j μετριέται από μία συνάρτηση προσέγγισης των αντίστοιχων διανυσμάτων $\vec{O}_i$ και $\vec{O}_j$.
Συγκεκριμένα μία από τις ευρέως χρησιμοποιούμενες μεθόδους μέτρησης είναι η *Ευκλείδειος νόρμα* η οποία ορίζεται:

$$Euclidean(O_i, O_j) = \sqrt{\sum_{d=1}^{p} (o_{id} - o_{jd})^2}$$

Στα δεδομένα έκφρασης γονιδίων, η συνολική μορφή των προτύπων των γονιδίων παρουσιάζει μεγαλύτερο ενδιαφέρον από το ξεχωριστό μέγεθος κάθε χαρακτηριστικού χωριστά. Η Ευκλείδειος νόρμα δεν τα καταφέρνει καλά σε μετατοπισμένα ή κλιμακωμένα πρότυπα. Γι' αυτό πρέπει να καταφύγουμε σε μία κανονικοποίηση ως προς μία μονάδα μέτρησης ή μετατροπή των δεδομένων στην ίδια κλίμακα μεγέθους.

Μία εναλλακτική μέθοδος μέτρησης είναι *το γινόμενο συσχέτισης Pearson (Pearson's correlation coefficient)*, ο οποίος μετράει την σχηματική ομοιότητα δύο προτύπων. Δεδομένων O_i και O_j το Pearson's correlation coefficient ορίζεται:

$$Pearson(O_i, O_j) = \frac{\sum_{d=1}^{p} (o_{id} - \mu_{oi})(o_{jd} - \mu_{oj})}{\sqrt{\sum_{d=1}^{p} (o_{id} - \mu_{oi})^2} \sqrt{\sum_{d=1}^{p} (o_{jd} - \mu_{oj})^2}}$$

όπου o_{id} η τιμή της i μεταβλητής για τη d-στη οντότητα, μ_{oj} η μέση τιμή όλων των μεταβλητών για τη j οντότητα.
Το Pearson's correlation coefficient χρησιμοποιείται ευρέως και έχει αποδειχτεί ιδιαίτερα αποτελεσματικό στην μέτρηση της ομοιότητας για δεδομένα έκφρασης γονιδίων. Σε δυαδικά δεδομένα η εμπειρική μελέτη έχει δείξει πως η μέθοδος δεν είναι ιδιαίτερα "στιβαρή" και μπορεί να επιστρέψει ομοιότητα για τα δύο διανύσματα O_i και O_j ενώ στην πραγματικότητα η τοποθέτηση των διανυσμάτων στον χώρο διαφέρει. Μία διόρθωση έγινε με τη *συσχέτιση Jackknife* στην οποία ισχύει:

$$Jackknife(O_i, O_j) = \min\{\rho_{ij}^{(1)}, \ldots, \rho_{ij}^{(l)}, \ldots, \rho_{ij}^{(p)}\}$$

όπου $\rho_{ij}^{(l)}$ είναι το Pearson's correlation coefficient των O_i και O_j με το l-στο χαρακτηριστικό διαγραμμένο.

Υπάρχουν και άλλες προσεγγίσεις βασισμένες στο Jackknife, όμως η χρήση τους είναι σπάνια λόγω των μεγάλων απαιτήσεών τους σε υπολογιστική ισχύ.

Ένα άλλο πρόβλημα του Pearson's correlation coefficient είναι ότι υποθέτει μία Gauss-ιανή προσεγγιστική κατανομή σημείων η οποία μπορεί να μην είναι ανεκτική για μη Gauss-ιανές κατανομές.

Αλγόριθμοι κατηγοριοποίησης

Ακολουθεί η ανάλυση καθεμίας από τις τρεις μεθόδους κατηγοριοποίησης χωριστά, εστιάζοντας σε προβλήματα και αλγόριθμους επίλυσης τους

Κατηγοριοποίηση βάση γονιδίων

Παρακάτω αναλύονται τα προβλήματα της κατηγοριοποίησης γονιδίων βάση των προτύπων έκφρασης. Επίσης παρουσιάζονται οι προκλήσεις της κατηγοριοποίησης βάση γονιδίων και διάφοροι αλγόριθμοι κατηγοριοποίησης.

Προκλήσεις της κατηγοριοποίησης γονιδίων

Η ιδιαιτερότητα των δεδομένων έκφρασης γονιδίων και οι ιδιόμορφες απαιτήσεις του βιολογικού τομέα, δημιουργούν στην κατηγοριοποίηση γονιδίων προκλήσεις και προβλήματα, αρκετά από τα οποία εκκρεμούν.

- Η ανάλυση συστάδων είναι το πρώτο βήμα για data-mining και ανακάλυψη γνώσης (data mining). Ο σκοπός της κατηγοριοποίησης δεδομένων έκφρασης γονιδίων είναι να αποκαλύψουμε δομές δεδομένων και να επιτύχουμε μία πρώτη εικόνα σχετικά με τις κατανομές των δεδομένων. Για τον λόγο αυτό ένας καλός αλγόριθμος κατηγοριοποίησης πρέπει να έχει την ελάχιστη δυνατή εξάρτηση από προηγούμενη γνώση, η οποία συνήθως δεν είναι καν διαθέσιμη πριν την ανάλυση συστάδων.
- Τα δεδομένα που προκύπτουν από τα πειράματα με microarrays έχει συνήθως πολύ θόρυβο. Γι' αυτό τον λόγο ο αλγόριθμος κατηγοριοποίησης πρέπει να είναι ικανός να εξάγει χρήσιμη πληροφορία κάτω από συνθήκες έντονου θορύβου.
- Η εμπειρία έχει δείξει ότι τα δεδομένα έκφρασης γονιδίων είναι συχνά "ισχυρά συνδεδεμένα" μεταξύ τους, με τις διάφορες κατηγορίες να είναι επικαλυπτόμενες ή ακόμα να εξαπλώνεται η μία πάνω στην άλλη. Οπότε οι αλγόριθμοι gene-based κατηγοριοποίησης πρέπει να μπορούν να το αντιμετωπίσουν.
- Τέλος οι χρήστες μπορεί να μην ενδιαφέρονται μόνο για τις συστάδες γονιδίων, αλλά ακόμα και για τις σχέσεις μεταξύ των συστάδων, σχέσεις γονιδίων μέσα στην ίδια συστάδα κτλ. Γι' αυτό ένας αλγόριθμος κατηγοριοποίησης θα ήταν καλό να μην σταματά στον διαμερισμό των δεδομένων, αλλά να παρέχει και γραφική αναπαράσταση της δομής των συστάδων.

Αλγόριθμοι κατηγοριοποίησης γονιδίων

K-means

Ο αλγόριθμος K-means είναι μία τυπική μέθοδος κατηγοριοποίησης. Δεδομένου ενός αριθμού K, ο αλγόριθμος χωρίζει τα δεδομένα σε K ομάδες (όχι απαραίτητα ξένες μεταξύ τους) βελτιστοποιώντας την ακόλουθη συνάρτηση:

$$E = \sum_{i=1}^{K} \sum_{O \in C_i} | O - \mu_i |^2$$

όπου:

O : ένα αντικείμενο δεδομένων στην συστάδα C_i

μ_i : ο μέσος όρος των αντικειμένων του C_i

Ο K-means είναι ένας απλός και γρήγορος αλγόριθμος. Η πολυπλοκότητά του είναι O(l*k*n) όπου l ο αριθμός των επαναλήψεων και k το πλήθος των συστάδων. Μάλιστα η πράξη έχει δείξει ότι με λίγες επαναλήψεις έχουμε πετύχει μία ικανοποιητική σύγκλιση.

μειονεκτήματα:

1. Είναι δύσκολο να γνωρίζουμε εξαρχής το πλήθος των συστάδων που θα χρειαστούμε. Σαν λύση αρκετοί επαναλαμβάνουν τον αλγόριθμο με διαφορετικό k κάθε φορά, και συγκρίνουν τα αποτελέσματα. Για μεγάλο πλήθος δεδομένων όμως δεν είναι ιδιαίτερα βολική αυτή η προσέγγιση.
2. Αν τα δεδομένα έχουν πολύ θόρυβο(όπως συμβαίνει συνήθως), ο αλγόριθμος εξακολουθεί να τοποθετεί τα δεδομένα σε συστάδες · πράγμα που τον κάνει *ευαίσθητο στον θόρυβο*.

Μία βελτίωση που προτάθηκε ήταν η χρήση global παραμέτρων για τον έλεγχο της ποιότητας των εξαγόμενων συστάδων. Τότε όμως θέτονται τα ερωτήματα πως επιλέγονται αυτές οι παράμετροι και πως τελικά η ποιότητα των δεδομένων στις συστάδες μπορεί να διαφέρουν από τις παραμέτρους αυτές.

Αυτό-οργανωμένοι χάρτες S.O.M. (Self-organized maps)

Η μέθοδος προτάθηκε από την Kohonen και βασίζεται σε απλό επίπεδο νευρωνικό δίκτυο. Τα δεδομένα εμφανίζονται στους νευρώνες εισόδου, και οι νευρώνες εξόδου οργανώνονται σε μία δομή p*g πλέγματος. Κάθε νευρώνας συσχετίζεται με ένα διάνυσμα αναφοράς και κάθε δεδομένο εισόδου αντιστοιχίζεται στον νευρώνα με το "πλησιέστερο" διάνυσμα αναφοράς. Κατά την διάρκεια εκτέλεσης του αλγορίθμου κάθε δεδομένο εισόδου συμπεριφέρεται σαν ένα δείγμα εκπαίδευσης, το οποίο προσαρμόζει τα διανύσματα αναφοράς ώστε να αναπαριστούν την κατανομή του συνόλου των δεδομένων εισόδου. Όταν ολοκληρωθεί η εκπαίδευση, οι συστάδες αναγνωρίζονται αντιστοιχίζοντας τα δεδομένα εισόδου στους νευρώνες εξόδου.

Ο αλγόριθμος με χρήση S.O.M.s μοιάζει πολύ με κλασικούς αλγόριθμους ανάλυσης συστάδων (π.χ. K-means) όμως υποφέρει και από τα ίδια προβλήματα π.χ. δεν μπορούμε να καθορίσουμε το βέλτιστο πλήθος συστάδων. Επιπλέον η σύγκλιση ελέγχεται από πολλές παραμέτρους όπως ποσοστό εκμάθησης, ή το είδος των δεδομένων. Παρόλ' αυτά ο S.O.M. είναι αρκετά καλός για κατηγοριοποίηση πολυδιάστατων δεδομένων.

Ιεραρχική κατηγοριοποίηση

Σε αντίθεση με την partition-based κατηγοριοποίηση που τοποθετούν τα δεδομένα άμεσα σε συστάδες, οι ιεραρχικές κατηγοριοποιήσεις δημιουργούν μία ιεραρχική ακολουθία ένθετων συστάδων που γραφικά μπορούν να αναπαρασταθούν από ένα δέντρο -*δενδρόγραμμα (dendrogram)*-. Τα κλαδιά του δενδρογράμματος όχι μόνο καταγράφουν την ανάπτυξη των συστάδων, αλλά ακόμα υποδηλώνουν και τις διαφορές μεταξύ των συστάδων. "Κόβοντας" το δενδρόγραμμα σε κάποιο επίπεδο έχουμε τα κατηγοριοποίηση σε αντίστοιχο αριθμό συστάδων.

Οι αλγόριθμοι ιεραρχικής κατηγοριοποίησης μπορούν να χωριστούν παραπέρα σε συσσωρευτικές και διαιρετικές προσεγγίσεις, βασισμένοι στον τρόπο που αναπτύσσεται το δενδρόγραμμα. *Οι συγκεντρωτικοί αλγόριθμοι* αντιμετωπίζουν αρχικά κάθε αντικείμενο σαν χωριστή συστάδα και σε κάθε βήμα οι συστάδες ενώνονται με κριτήρια ομοιότητας μέχρι να φτάσουμε να έχουμε μόνο μία συστάδα. Τα διάφορα μέσα που χρησιμοποιούνται για την μέτρηση εγγύτητα των συστάδων όπως απλή ομοιότητα, πλήρης ομοιότητα και ελάχιστες διαφορές,, οδηγούν σε διάφορες στρατηγικές συγχώνευσης συστάδων. *Οι διαιρετικοί αλγόριθμοι* κάνουν το αντίθετο, ξεκινώντας από μία συστάδα που έχει όλα τα αντικείμενα και στην συνέχεια την χωρίζουν σε επιμέρους συστάδες. Το πρόβλημα εδώ είναι με ποιόν τρόπο γίνεται η διαίρεση συστάδων σε κάθε βήμα. Μερικοί βασίζονται σε μεθόδους ντετερμινιστικών αλγορίθμων ενώ άλλοι σε θεωρία γράφων.

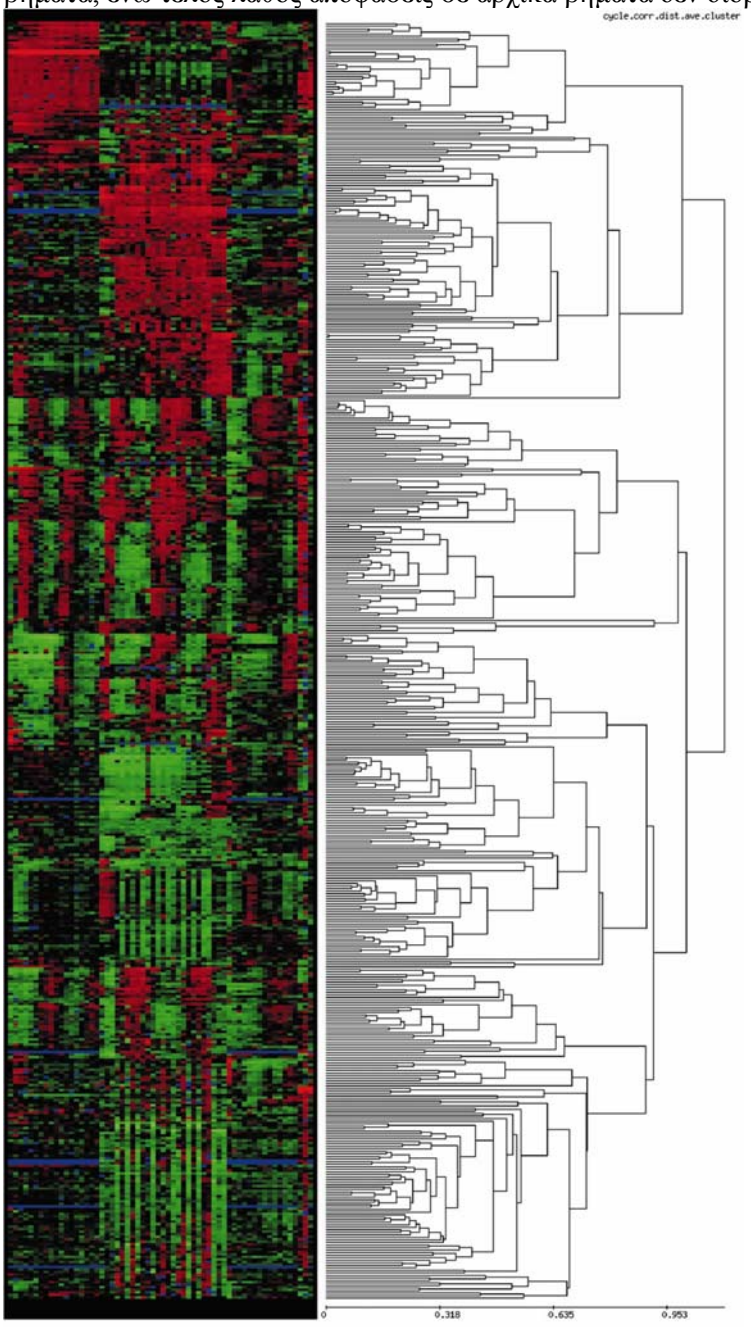
Ένα γνωστός συγκεντρωτικός αλγόριθμος είναι ο UPGMA (Unweighted Pair Group Method with Arithmetic Mean, προτάθηκε από τον Eisen) και υιοθετεί την γραφική αναπαράσταση των κατηγοριοποιημένων δεδομένων. Σε αυτή την μέθοδο κάθε κελί χρωματίζεται βάση του λόγου του μετρημένου φθορισμού, και οι γραμμές του πίνακα αναδιατάσσονται βασισμένες σε μία δομή ιεραρχικού δενδρογράμματος και ενός συνεπούς κανόνα κόμβο-διάταξης. Μετά την κατηγοριοποίηση ο αρχικός πίνακας έκφρασης γονιδίων αναπαρίσταται από ένα χρωματισμένο πίνακα (εικόνα συστάδας) όπου μεγάλα συνεχή χρωματικά μέρη αναπαριστούν ομάδες γονιδίων που μοιράζονται παρόμοια πρότυπα έκφρασης κάτω από πολλαπλές συνθήκες.

Μία διαιρετική προσέγγιση αποτελεί ο ντετερμινιστικός – δυναμωτικός αλγόριθμος (deterministic-annealing algorithm). Αρχικά επιλέγονται τυχαία δύο κέντρα συστάδας C_j, j=1,2 . Το πρότυπο έκφρασης του γονιδίου k αναπαρίσταται από διάνυσμα $\vec{g}_k$, και η πιθανότητα το γονίδιο k να ανήκει στην συστάδα j δίνεται από το Gauss-ιανό μοντέλο δύο συνιστωσών:

$$P_j(\vec{g}_k) = \exp(-\beta \,|\, \vec{g}_k - C_j \,|^2) \Big/ \sum_j \exp(-\beta \,|\, \vec{g}_k - C_j \,|^2)$$

Τα κέντρα συστάδας επαναϋπολογίζονται από την $Cj=\sum_k \vec{g}_k \, P_j(\vec{g}_k) \Big/ \sum_k P_j(\vec{g}_k)$. Μία επαναληπτική διαδικασία υπολογίζει τα P_j και C_j. Όταν το β μικραίνει με μικρά βήματα μέχρι να προσεγγιστεί ένα όριο, τα δύο κέντρα ενώνονται. Έτσι όλα τα σετ δεδομένων έχουν επαναληπτικά χωριστεί μέχρι το σημείο που κάθε στοιχείο είναι και μία συστάδα.

Η ιεραρχική κατηγοριοποίηση δεν διαχωρίζει μόνο τα δεδομένα σε συστάδες με παρόμοια πρότυπα έκφρασης, αλλά παρέχει και μία φυσική γραφική αναπαράσταση των σετ δεδομένων. Η γραφική αναπαράσταση βοηθάει τους χρήστες με μία γρήγορη ματιά να έχουν μία πρώτη εικόνα για την κατανομή των δεδομένων. Για τον λόγο αυτό η μέθοδος UPGMA είναι από τις πλέον δημοφιλείς στους βιολόγους. Παρόλ' αυτά οι ιεραρχικές μέθοδοι πάσχουν στον τομέα της στιβαρότητας αφού μικρές αλλαγές στα αρχικά δεδομένα μπορεί να προκαλέσουν μεγάλες τροποποιήσεις του δενδρογράμματος. Μία πλήρης διαδικασία ιεραρχικής κατηγοριοποίησης (διαιρετικής ή συγκεντρωτικής) απαιτεί $n^2 - n/2$ βήματα, άρα μιλάμε για πολυπλοκότητα $O(n^2 log n)$. Επιπλέον λόγω απληστίας των αλγορίθμων σε κάθε βήμα δεν μπορούν να γίνουν βελτιώσεις σε ενέργειες που έγιναν στα προηγούμενα βήματα, ενώ τέλος λάθος αποφάσεις σε αρχικά βήματα δεν διορθώνονται.



Εικόνα: Ιεραρχική κατηγοριοποίηση του πίνακα έκφρασης γονιδίων. Εδώ φαίνεται η μέση διασύνδεση της κατηγοριοποίησης 505 γονιδίων κατά την μελέτη 3 διαφορετικών κυτταρικών κύκλων, με ένα σύνολο 60 διαφορετικών χρονικών στιγμών να αναλύονται. Το πολύχρωμο μέρος δείχνει τις αριθμητικές τιμές χρωματικά κωδικοποιημένες.
Το κόκκινο δείχνει θετικές τιμές, το πράσινο αρνητικές και το μπλε χαμένες τιμές.

Γραφοθεωρητικές μέθοδοι κατηγοριοποίησης

Δεδομένου ενός σετ δεδομένων X, μπορούμε να κατασκευάσουμε ένα πίνακα εγγύτητας (proximity) P, όπου P[i,j] = proximity(O_i, O_j), και ένα γράφο εγγύτητας G(V, E), όπου κάθε δεδομένο αντιστοιχεί σε μια κορυφή. Για κάποιες από τις μεθόδους κατηγοριοποίησης, κάθε ζευγάρι αντικειμένων συνδέεται με μία ακμή με το βάρος να ορίζεται βάση της εγγύτητας των τιμών των δύο αντικειμένων. Για άλλες μεθόδους η εγγύτητα χαρτογραφείται μόνο από 0 και 1 με βάση κάποιο κατώφλι και ακμές υπάρχουν μόνο μεταξύ των αντικειμένων i, j όπου P[i, j] = 1. Οι γραφοθεωρητικές τεχνικές κατηγοριοποίησης είναι ρητά συνυφασμένες με τους γράφους, αφού μετατρέπουν το πρόβλημα της κατηγοριοποίησης σε πρόβλημα εύρεσης ελάχιστων – μέγιστων διαδρομών στον γράφο εγγύτητας.

Αλγόριθμοι:

- *CLICK*: Ο αλγόριθμος ψάχνει να βρει ισχυρά συνεκτικά συνιστώσες ως συστάδες στον γράφο εγγύτητας. Ο CLICK αρχικά κανονικοποιεί τα δεδομένα εισόδου, με αποτέλεσμα οι αποστάσεις μεταξύ δύο στοιχείων (ανεξάρτητα αν ανήκουν στην ίδια συστάδα) να εμφανίζουν ομοιόμορφη κατανομή. Με αυτή την παραδοχή το βάρος w_{ij} κάθε ακμής (i,j) αναπαριστά την πιθανότητα οι κορυφές i και j να ανήκουν στην ίδια συστάδα. Αυτός ο αλγόριθμος χρειάζεται ακόμα δύο βήματα, ένα για να διαχειριστεί τα μεμονωμένα στοιχεία και να ενημερώσει τις συστάδες, και ένα για να ενώσει δύο συστάδες που έχουν περάσει ένα προκαθορισμένο βαθμό ομοιότητας.
- *CAST*: Ο αλγόριθμος CAST (Cluster Affinity Search Technique), χρησιμοποιεί ως είσοδο ένα πίνακα ομοιότητας S και ένα κατώφλι ομοιότητας t, το οποίο ελέγχει τον αριθμό και το μέγεθος των παραγόμενων συστάδων. Σε κάθε επανάληψη δημιουργείται και μία νέα συστάδα. Ορίζουμε ως α(x) τη συνάρτηση ομοιότητας για το στοιχείο x που είναι ίσο με το άθροισμα των τιμών ομοιότητας μεταξύ του x και των στοιχείων που ανήκουν στην τρέχουσα συστάδα. Ένα στοιχείο x λέμε ότι παρουσιάζει υψηλή ομοιότητα αν ισχύει

  $\alpha(x) \geq t|C_{open}|$, όπου C_{open} η τρέχουσα συστάδα,

  διαφορετικά παρουσιάζει χαμηλή ομοιότητα.. Στη συνέχεια ο αλγόριθμος προσθέτει στοιχεία υψηλής ομοιότητας στην C_{open} και αφαιρεί από αυτή στοιχεία με χαμηλή ομοιότητα. Όταν θα σταθεροποιηθεί η κατάσταση, δηλαδή δεν θα προστίθενται ούτε θα αφαιρούνται άλλα στοιχεία, τότε ολοκληρώνεται μία συστάδα και ξεκινάει η δημιουργία μίας νέας συστάδας από την αρχή (cleaning step). Αυτό το βήμα κάνει τον αλγόριθμο CAST να αποφεύγει την αδυναμία της επανάληψης που παρουσιάζουν οι πιο δημοφιλείς τεχνικές ανάλυσης συστάδας.

Κατηγοριοποίηση βάση μοντέλου

Οι κατηγοριοποιήσεις βάση μοντέλου (model-based clustering) παρέχει ένα στατιστικό πλαίσιο εργασία για να μοντελοποιήσει την δομή των συστάδων με δεδομένα έκφρασης γονιδίων. Τα σετ δεδομένων υποθέτουμε ότι προέρχονται από ένα πεπερασμένο μείγμα υποκειμένων πιθανοτικών κατανομών. , κα κάθε συνιστώσα να αντιστοιχεί σε διαφορετική συστάδα. Ο στόχος είναι να υπολογιστούν τα

$\Theta = \{\theta_i \mid 1 \le i \le k\}$ και

$\Gamma = \{\gamma_r^i \mid 1 \le i \le k,\ 1 \le r \le n\}$ ώστε να μεγιστοποιείται η πιθανότητα

$L_{min}(\Theta,\Gamma) = \sum_{i=1}^{k} \gamma_r^i f_i(x_r \mid \theta_i)$ όπου n το πλήθος των αντικειμένων, k το πλήθος των συνιστωσών, x_r είναι ένα αντικείμενο δεδομένων, $f_i(x_r \mid \theta_i)$ είναι η συνάρτηση πυκνότητας του x_r της συνιστώσας του C_i με κάποιο άγνωστο σετ παραμέτρων θ_i και γ_r^i αναπαριστά την πιθανότητα το x_r να ανήκει στο C_i. Συνήθως οι παράμετροι Θ και Γ υπολογίζονται από τον αλγόριθμο ΕΜ. Ο αλγόριθμος ΕΜ επαναλαμβάνει διαδοχικά βήματα πρόβλεψης(E steps) και μεγιστοποίησης(M steps). Στο βήμα Ε, η κρυφή παράμετρος Γ υπολογίζεται υπό συνθήκη από τα δεδομένα στο τρέχον Θ. Στο βήμα Μ η παράμετρος μοντέλου Θ υπολογίζεται ώστε να μεγιστοποιηθεί η πιθανότητα να ολοκληρωθούν τα δεδομένα δεδομένων των κρυφών παραμέτρων. Όταν ο ΕΜ αλγόριθμος συγκλίνει κάθε αντικείμενο δεδομένων έχει τοποθετηθεί στη συστάδα με την μεγαλύτερη υποθετική πιθανότητα.

Το κύριο πλεονέκτημα της μεθόδου αυτής είναι ότι υπολογίζει την πιθανότητα γ_k^i που ένα αντικείμενο δεδομένων ανήκει σε μία συστάδα k. Γνωρίζοντας ότι τα γονίδια έχουν μεγάλη συσχέτιση μεταξύ τους, είναι πολύ πιθανόν κάποιο γονίδιο να ανήκει ταυτόχρονα σε δύο συστάδες. Επιπλέον το πιθανοτικό μοντέλο φαίνεται να ταιριάζει περισσότερο σε δεδομένα έκφρασης γονιδίων. Το μειονέκτημα της μεθόδου είναι η παραδοχή ότι τα δεδομένα ταιριάζουν σε μία συγκεκριμένη κατανομή. Αυτό όμως μπορεί να μην συμβαίνει πάντα. Γίνονται πολλές προσπάθειες πάνω στην βελτίωση της μεθόδου μοντελοποίησης.

Κατηγοριοποίηση βάση πυκνότητας (ιεραρχική προσέγγιση DHC)

Ένας νεότερος αλγόριθμος κατηγοριοποίησης που προτάθηκε από τους Jiang, Pei και Zhang για να αναγνωρίζονται γονίδια με όμοια πρότυπα, είναι ο DHC. Ο DHC αναπτύχθηκε με την αντίληψη της πυκνότητας και έλξης των αντικειμένων με δεδομένα. Η βασική ιδέα είναι να αντιμετωπιστούν οι συστάδες σαν πολυδιάστατες πυκνές περιοχές, όπου τα αντικείμενα δεδομένων έλκονται το ένα από το άλλο. Στον πυρήνα κάθε περιοχής υψηλής πυκνότητας υπάρχουν αντικείμενα με μεγάλη ομοιότητα. Όσο απομακρυνόμαστε η ομοιότητα χάνεται.

Ο DHC οργανώνει τις συστάδες με ιεραρχικές δομές δύο επιπέδων. Αρχικά ένα δέντρο "έλξης" δημιουργείται για να αναπαρασταθούν οι σχέσεις μεταξύ αντικειμένων σε πυκνές περιοχές. Κάθε κόμβος αυτού του δέντρου αντιστοιχίζεται με ένα αντικείμενο δεδομένων, και ο πατέρας του κάθε κόμβου είναι ο ελκών του. Η μόνη εξαίρεση είναι ότι τα δεδομένα τα οποία έχουν μεγαλύτερη πυκνότητα στο σετ δεδομένων. Τα προβλήματα παρουσιάζονται όταν τα δεδομένα γίνουν πολλά και η δομή δεδομένων καταντάει ιδιαίτερα πολύπλοκη. Για τον λόγο αυτό υπάρχει η δομή δευτέρου επιπέδου, η οποία είναι ένα δέντρο πυκνότητας. Κάθε κόμβος του δέντρου πυκνότητας συμβολίζει μία περιοχή υψηλής πυκνότητας. Αρχικά όλα τα δεδομένα θεωρούνται πως ανήκουν σε μία μεγάλη περιοχή πυκνότητας. Σε κάθε βήμα η περιοχή πυκνότητας διαιρείται σε επιμέρους περιοχές.

Η μέθοδος αυτή έχει πολλά πλεονεκτήματα, όπως αντοχή σε θόρυβο, και υψηλή συνεκτικότητα μεταξύ των δεδομένων (πρώτα εντοπίζει τον πυρήνα και στην συνέχεια τα όρια κάθε ομάδας). Παρόλ' αυτά η πολυπλοκότητά του $O(n^2)$ τον κάνει μη αποδοτικό. Τέλος χρειάζεται και αυτός με την σειρά του τον προκαθορισμό αρχικών παραμέτρων (το οποίο δεν είναι πάντα εφικτό).

Κατηγοριοποίηση βάση δειγμάτων

Στον πίνακα έκφρασης γονιδίων υπάρχουν συνήθως διάφοροι ιδιαίτεροι μακροσκοπικοί φαινότυποι δειγμάτων που σχετίζονται με ασθένειες ή επίδραση φαρμάκων, όπως δείγματα ασθενειών, φυσιολογικά δείγματα και δείγματα θεραπείας. Ο σκοπός της κατηγοριοποίησης βάση δειγμάτων είναι να βρει τις φαινοτυπικές δομές ή υπό-δομές των δειγμάτων. Προηγούμενες μελέτες έχουν δείξει ότι ο φαινότυπος των δειγμάτων μπορεί να διακριθεί μόνο μέσα από μία μικρή υπομονάδα γονιδίων των οποίων τα επίπεδα έκφρασης σχετίζονται στενά με την κλάση διάκρισης. Τα γονίδια αυτά αποκαλούνται και *πληροφοριακά γονίδια*. Τα υπόλοιπα γονίδια που απομένουν στον πίνακα έκφρασης γονιδίων θεωρούνται απλά θόρυβος.

Οι κλασικές μέθοδοι κατηγοριοποίησης, όπως αυτές που παρουσιάσαμε πριν (π.χ. αυτοοργανωμένοι χάρτες, K-means κτλ.), μπορούν να εφαρμοστούν θεωρώντας όλα τα γονίδια σαν χαρακτηριστικά, όμως ο υψηλός λόγος σήματος-θορύβου (που είναι της τάξης του 1/10) μειώνει δραματικά την αξιοπιστία των αποτελεσμάτων. Για τον λόγο αυτό έχουν προταθεί καινούριοι αλγόριθμοι, οι οποίοι χωρίζονται σε δύο κύριες κατηγορίες:
Την *εποπτευμένη ανάλυση* και την *μη-εποπτευμένη ανάλυση*.

Κατηγοριοποίηση βασισμένη σε εποπτευμένη επιλογή πληροφοριακών γονιδίων

Η εποπτευμένη προσέγγιση υποθέτει ότι η πληροφορία των φαινοτύπων είναι επισυνημμένη στα δείγματα, για παράδειγμα τα δείγματα είναι μαρκαρισμένα ως ασθενή ή φυσιολογικά. Με αυτή την πληροφορία μπορεί να κατασκευαστεί ένας ταξινομητής που περιέχει μόνο πληροφοριακά γονίδια. Βασισμένοι σε αυτόν τον ταξινομητή, τα δείγματα μπορούν να κατηγοριοποιηθούν ώστε να ταιριάζουν οι φαινότυποί τους, και οι ταμπέλες (ασθενή ή φυσιολογικά) που έχουν μπορούν να χρησιμοποιηθούν για προβλεφθούν τα μελλοντικά δείγματα από τα προφίλ έκφρασης. Οι εποπτευμένες μέθοδοι χρησιμοποιούνται ευρέως από τους βιολόγους για να επιλέγουν τα πληροφοριακά γονίδια.
Τα βήματα για να κατασκευαστεί ένας ταξινομητής (classifier) περιλαμβάνουν:

1. *Επιλογή δείγματος εκπαίδευσης*: Στο βήμα αυτό ένα υποσύνολο δειγμάτων επιλέγεται για να αποτελέσει το σετ εκπαίδευσης. Αφού το πλήθος των δειγμάτων είναι περιορισμένο (λιγότερα από 100), το μέγεθος του σετ εκπαίδευσης είναι συνήθως της ίδιας σημαντικότητας με το αρχικό πλήθος των δειγμάτων.
2. *Επιλογή πληροφοριακών γονιδίων*: Ο στόχος είναι η επιλογή γονιδίων των οποίων τα πρότυπα έκφρασης διαχωρίζουν τους διάφορους φαινοτύπους. Μερικές προσεγγίσεις που χρησιμοποιούνται για να επιτευχθεί αυτό είναι: η προσέγγιση γειτονικής ανάλυσης, οι εποπτευμένες μαθητευόμενες μέθοδοι και μία ποικιλία από μεθόδους βασισμένες σε ταξινόμηση.
3. *Κατηγοριοποίηση των δειγμάτων και ταξινόμηση*: μετά από περίπου 50~200 πληροφοριακά γονίδια τα οποία φανερώνουν μία κατηγορία φαινοτύπων στο επιλεγμένο δείγμα εκπαίδευσης, το όλο σύνολο δειγμάτων κατηγοριοποιείται χρησιμοποιώντας μόνο τα πληροφοριακά γονίδια σαν χαρακτηριστικό. Και αφού το σύνολο των χαρακτηριστικών είναι σχετικά μικρό, συμβατικοί αλγόριθμοι κατηγοριοποίησης χρησιμοποιούνται, όπως ο K-means ή ο SOM, για να κατηγοριοποιήσουν τα δείγματα.

Μη εποπτευομένη κατηγοριοποίηση και επιλογή πληροφοριακών γονιδίων

Η μη εποπτεύομενη κατηγοριοποίηση βάση δειγμάτων υποθέτει ότι καμία φαινοτυπική πληροφορία δεν έχει ανατεθεί σε κανένα δείγμα. Αφού η αρχική βιολογική αναγνώριση των κλάσεων είναι αργή, η υποθετική έρευνα και η αυτόματη ανακάλυψη φαινοτύπων των δειγμάτων έχει συνεισφέρει πολύ στην ανάλυση δεδομένων έκφρασης γονιδίων. Σαν μέθοδος μη-εποπτευμένης μεθόδου μαθήσεως, η κατηγοριοποίηση εξυπηρετεί και σαν εξερευνητική εργασία που βοηθάει να βρεθούν άγνωστες υπό-δομές στον αρχικό χώρο των δειγμάτων.

Η μη εποπτεύομενη κατηγοριοποίηση είναι πολύ πιο πολύπλοκη από την εποπτεύομενη, αφού δεν εκμεταλλεύεται τα δείγματα εκπαίδευσης αναφοράς για την επιλογή των πληροφοριακών γονιδίων και δεν προαπαιτεί βαθιά γνώση των φαινοτύπων. Γι' αυτό είναι δύσκολο σε μη εποπτεύομενη κατηγοριοποίηση να βρεθούν φαινότυποι και να επιλεγούν πληροφοριακά γονίδια.

Προκλήσεις:

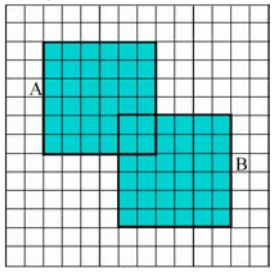
- Αφού έχουμε περιορισμένο πλήθος δειγμάτων και μεγάλο αριθμό γονιδίων, τα σετ δεδομένων είναι αραιά σε ένα πολυδιάστατο χώρο γονιδίων. Οπότε δεν είναι φανερές οι δομές των κλάσεων με τις συμβατικές τεχνικές
- Τα περισσότερα γονίδια δεν παρουσιάζουν ενδιαφέρον, άρα είναι θόρυβος. Αυτό κάνει αβέβαιο ποια από τα γονίδια είναι σχετικά με τον φαινότυπο, οπότε είναι δύσκολο να επιλεγούν πληροφοριακά γονίδια.

Δύο γενικές στρατηγικές που έχουν προταθεί σαν λύσεις είναι οι *μη-εποπτευόμενη επιλογή γονιδίων* και η *αλληλοσυσχετισμένη κατηγοριοποίηση*.

Μη-εποπτευόμενη επιλογή γονιδίων: Η επιλογή γονιδίων και η κατηγοριοποίηση δειγμάτων αντιμετωπίζονται σαν ανεξάρτητες διαδικασίες. Αρχικά η διάσταση των γονιδίων μειώνεται, και στην συνέχεια εφαρμόζονται κλασικοί αλγόριθμοι κατηγοριοποίησης. Αφού δεν υπάρχουν δείγματα εκπαίδευσης, η επιλογή γονιδίων βασίζεται μόνο σε στατιστικά μοντέλα που αναλύουν την διακύμανση στα δεδομένα έκφρασης γονιδίων.
Αλληλοσυσχετισμένη κατηγοριοποίηση: Παρατηρώντας προσεκτικά στο πρόβλημα επιλογής πληροφοριακών γονιδίων και στην κατηγοριοποίηση δειγμάτων, βλέπουμε ότι αυτά τα δύο έχουν μεγάλη σχέση μεταξύ του. Όταν βρούμε τα πληροφοριακά γονίδια είναι εύκολο σχετικά να κάνουμε κατηγοριοποίηση με συμβατικούς αλγορίθμους Από την άλλη μέσα σε μία συστάδα, τα γονίδια που έχουν μεγάλη συσχέτιση με την συστάδα θεωρούνται ως πληροφοριακά. Βασισμένη σε αυτή την παρατήρηση, η δεύτερη τεχνική χρησιμοποιεί δυναμική χρήση των σχέσεων μεταξύ γονιδίων και συστάδων, και επαναληπτικά συνδυάζει την διαδικασία κατηγοριοποίησης με την διαδικασία επιλογής γονιδίων. Διαισθητικά, αν και δεν ξέρουμε τις ακριβείς ομάδες (partitions) των δειγμάτων, σε κάθε επανάληψη μπορούμε να αναμένουμε μία προσέγγιση των ομάδων που είναι κοντά στη επιθυμητή ομάδα των δειγμάτων. Ο προσεγγιστικός διαμερισμός επιτρέπει την επιλογή μίας διαλλακτικά καλής υποομάδας γονιδίων, η οποία προσεγγίζει τη επιθυμητή ομάδα σε κάθε βήμα.

Κατηγοριοποίηση υποδιαστημάτων (subspace clustering)

Οι αλγόριθμοι κατηγοριοποίησης που παρουσιάστηκαν πριν είναι όλοι δείγματα "*ολικής κατηγοριοποίησης*" από την άποψη ότι για ένα δεδομένο σύνολο δεδομένων προς κατηγοριοποίηση, ο χώρος των χαρακτηριστικών είναι πλήρως καθορισμένος και μοιράζεται από όλες τις εξαγόμενες συστάδες, με τις συστάδες να είναι αποκλειστικές και εξαντλητικές. Παρόλ' αυτά στην μοντέρνα μοριακή βιολογία είναι γνωστό ότι μόνο ένα μικρό μέρος των γονιδίων παίρνει μέρος στις κυτταρικές λειτουργίες. Επιπλέον ένα γονίδιο μπορεί να συμμετέχει σε πολλαπλές λειτουργίες οι οποίες μπορεί ακόμα και να μην συνεργάζονται, οπότε ένα γονίδιο μπορεί να ανήκει σε πολλές συστάδες, ή και σε καμία. Πρόσφατα μία σειρά μεθόδων κατηγοριοποίησης υποδιαστημάτων προτάθηκε ώστε να κατανοήσει την σχέση που παρουσιάζεται στα μπλοκ μέσα στον πίνακα έκφρασης γονιδίων (μπλοκ είναι ένας υποπίνακας που ορίζεται από ένα υποσύνολο γονιδίων και ένα υποσύνολο δειγμάτων).



Η *κατηγοριοποίηση υποδιαστημάτων* αρχικά προτάθηκε από τον Agrawal για να βρει υποσύνολα αντικειμένων ώστε τα αντικείμενα να εμφανίζονται σαν μία συστάδα σε ένα υποχώρο που ορίζεται από ένα υποσύνολο των χαρακτηριστικών. Στο δίπλα σχήμα φαίνεται ένα παράδειγμα του υποσυνόλου των συστάδων (A, B) απλωμένου στο πίνακα έκφρασης γονιδίων. Στην κατηγοριοποίηση αυτή είναι δυνατόν το υποσύνολο των χαρακτηριστικών να διαφέρει από συστάδα σε συστάδα. Δύο υποσύνολα συστάδων μπορούν να μοιράζονται κοινά αντικείμενα και χαρακτηριστικά, ενώ μερικά αντικείμενα μπορεί να μην ανήκουν σε κανένα υποσύνολο συστάδας.

Για ένα πίνακα έκφρασης γονιδίων που έχει n γονίδια και m δείγματα, η πολυπλοκότητα για πλήρη συνδυασμό τους είναι 2^{n+m} οπότε το πρόβλημα της συνολικά βέλτιστης επιλογής μπλοκ είναι δυσκολίας NP. Οι μέθοδοι κατηγοριοποίησης υποδιαστημάτων συνήθως ορίζουν μοντέλα για να περιγράψουν το επιθυμητό μπλοκ και στην συνέχεια υιοθετούν κάποια μέθοδο να ψάξει στον χώρο των δειγμάτων-γονιδίων. Ακολουθούν μερικοί αλγόριθμοι κατηγοριοποίησης υποδιαστημάτων.

Συνδεδεμένη αμφίδρομη κατηγοριοποίηση (Coupled two-way clustering - CTWC)

Ο Getz μοντελοποιεί το μπλοκ σαν μία σταθερή συστάδα με χαρακτηριστικά F_i και αντικείμενα O_j , όπου τα F_i και O_j μπορεί να είναι είτε γονίδια είτε δείγματα. Η συστάδα είναι σταθερή από την άποψη ότι μόνο τα χαρακτηριστικά στο F_i χρησιμοποιούνται για να κατηγοριοποιηθούν τα αντίστοιχα O_j, τα O_j δεν διαχωρίζονται μετά από κάποιο κατώφλι. Ο CTWC παρέχει μία μέθοδο λύσης που αποφεύγει την "ωμή" απαρίθμηση όλων των δυνατών συνδυασμών. Μόνο υποσύνολα γονιδίων ή δειγμάτων που χαρακτηρίστηκαν σαν σταθερές συστάδες σε προηγούμενο βήμα είναι υποψήφια για την επόμενη επανάληψη.

Ο CTWC ξεκινά με ένα μόνο ζεύγος από σετ γονιδίων και σετ δειγμάτων (G_O,S_O), όπου G_O είναι ένα σετ που περιέχει όλα τα γονίδια και S_O περιέχει όλα τα δείγματα. Μία μέθοδος ιεραρχικής κατηγοριοποίησης, ονόματι υπέρ-παραμαγνητικός αλγόριθμος κατηγοριοποίησης, εφαρμόζεται σε κάθε σετ, και οι σταθερές συστάδες γονιδίων και δειγμάτων που αποδίδονται μετά από αυτή την πρώτη επανάληψη είναι οι G_1^i και S_1^j. Ο CTWC διατηρεί δυναμικά δύο λίστες, μία με τις σταθερές συστάδες (λίστα γονιδίων GL, λίστα δειγμάτων SL) και μία με τα ζεύγη των υποσυνόλων γονιδίων δειγμάτων (G_n^i, S_m^j). Σε κάθε επανάληψη ένα υποσύνολο γονιδίων από το GL και ένα υποσύνολο δειγμάτων από το SL, που δεν έχουν συνδιαστεί στο παρελθόν, συνδυάζονται και κατηγοριοποιούνται αμοιβαία σαν αντικείμενα και χαρακτηριστικά. Οι νέες συστάδες προστίθενται στις GL και SL, και ένας δείκτης γράφεται στην λίστα ζευγών γονιδίων για να δηλώνει την καταγωγή των συστάδων. Οι επαναλήψεις συνεχίζονται μέχρι να μην βρίσκονται νέες συστάδες που ικανοποιούν κάποιο κριτήριο, όπως σταθερότητα ή μέγεθος.

Μειονεκτήματα της μεθόδου είναι ότι υπάρχει περίπτωση κάποιοι συνδυασμοί μπορεί να μην γίνουν, λόγω του ντετερμινιστικού χαρακτήρα του αλγορίθμου, ενώ επιπλέον τα αποτελέσματα είναι δύσκολο να ερμηνευθούν.

Μοντέλο plaid

Το μοντέλο plaid (καρό μοντέλο) αντιμετωπίζει τα δεδομένα έκφρασης γονιδίων σαν ένα άθροισμα πολλαπλών "επιπέδων", όπου κάθε επίπεδο αναπαριστά την παρουσία μίας συγκεκριμένης βιολογικής διαδικασίας με μόνο ένα υποσύνολο των γονιδίων και ένα υποσύνολο των δειγμάτων που σχετίζονται με αυτήν. Το γενικευμένο plaid μοντέλο τυποποιείται ως:

$Y_{ij} = \sum_{k=0}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk}$, όπου

Y_{ij} το επίπεδο έκφρασης του γονιδίου i υπό το δείγμα j

θ_{ij0} είναι το υποφαινόμενο επίπεδο έκφρασης για ολόκληρο το σετ δεδομένων και το θ_{ijk} περιγράφει την συμβολή του επιπέδου k.

ρ_{ik} ισούται με 1 αν το γονίδιο(ή δείγμα) i ανήκει στο επίπεδο k, αλλιώς ισούται με 0

Η διαδικασία κατηγοριοποίησης αναζητεί τα επίπεδα στα σετ δεδομένων το ένα μετά το άλλο, χρησιμοποιώντας τον αλγόριθμο EM για να υπολογίσει τις παραμέτρους του μοντέλου. Υποθέτοντας ότι τα πρώτα K-1 επίπεδα έχουν εξαχθεί, το K-στο επίπεδο χαρακτηρίζεται ελαχιστοποιώντας το άθροισμα των τετραγώνων των λαθών:

$$Q = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}(Z_{ij} - \theta_{ijK}\rho_{iK}\kappa_{jK})^2$$

όπου $Z_{ij} = Y_{ij} - \theta_{ij0} - \sum_{k=0}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk}$ είναι το υπόλοιπο από τα K–1 επίπεδα. Η διαδικασία κατηγοριοποίησης σταματά όταν η μεταβολή των επιπέδων έκφρασης μέσα στο τρέχον επίπεδο είναι μικρότερη από ένα κατώφλι.

Το πρόβλημα του μοντέλου είναι ότι αν ένα γονίδιο συμμετέχει σε πολλές κυτταρικές διαδικασίες, τότε επίπεδο έκφρασής του είναι άθροισμα των όρων που μετέχουν σε κάθε διαδικασία χωριστά. Οπότε η αποτελεσματικότητα και η ικανότητα ερμηνείας των επιπέδων που βρέθηκαν, χρίζει περαιτέρω μελέτης.

Δικατηγοριοποίηση και δ-συστάδες (Biclustering and δ-Clusters)

Ο Cheng εισήγαγε την προσέγγιση bicluster για να μοντελοποιήσει ένα μπλοκ, μαζί με μία βαθμολογία που λέγεται μέσο-τετραγωνισμένο υπόλοιπο που μετράει την συνέπεια των γονιδίων και συνθηκών στο μπλοκ. Έστω G' και S' τα υποσύνολα των γονιδίων και των δειγμάτων. Το ζεύγος (G', S') καθορίζει έναν υποπίνακα με το μέσο-τετραγωνισμένο υπόλοιπο:

$$H(G', S') = \frac{1}{|G'||S'|} \sum_{i \in G', j \in S'} (w_{ij} - \eta_{iS'} - \eta_{G'j} + \eta_{G'S'})^2$$

όπου:

$$\eta_{iS'} = \frac{1}{|S'|} \sum_{j \in J} w_{ij}$$

$$\eta_{G'j} = \frac{1}{|G'|} \sum_{i \in I} w_{ij}$$

$$\eta_{S'J} = \frac{1}{|G'||S'|} \sum_{i \in G', j \in S'} w_{ij}$$

είναι οι μέσοι των γραμμών και των στηλών, και οι μέσοι του υποπίνακα. Ένας υποπίνακας καλείται δ-bicluster αν $H(G', S') \leq \delta$ για κάποιο $\delta > 0$. Μία μικρή μέσο-τετραγωνισμένη υπόλοιπη βαθμολογία μαζί με μία μεγάλη μεταβλητότητα από την υποδηλωμένη σταθερά είναι ένα καλό κριτήριο για να αναγνωριστεί ένα μπλοκ.

Παρόλ' αυτά το πρόβλημα εύρεσης ενός ελάχιστου σετ biclusters που καλύπτει όλα τα στοιχεία σε έναν πίνακα δεδομένων, έχει αποδειχτεί πως είναι δυσκολίας NP. Μία άπληστη μέθοδος που παρέχει μία προσέγγιση της βέλτιστης λύσης και μειώνει την πολυπλοκότητα σε πολυωνυμικό χρόνο προτάθηκε από τους Cheng και Church. Για να βρεθεί ένα bicluster, το σκορ H υπολογίζεται για κάθε πιθανή προσθήκη/αφαίρεση στήλης/γραμμής, και η ενέργεια που μειώνει περισσότερο το H εφαρμόζεται. Αν δεν γίνει καμία ενέργεια στον H ή αν $H \leq \delta$, η bicluster επιστρέφεται. Ο αλγόριθμος χρειάζεται χρόνο $O((n+m) \cdot mn)$, όπου n και m το πλήθος των γονιδίων και των δειγμάτων. Ο αλγόριθμος αυτός είναι ιδιαίτερα χρονοβόρος για μεγάλο αριθμό γονιδίων και δειγμάτων, όμως έχει γίνει βελτιστοποίηση ώστε να έχει πολυπλοκότητα O(nm).

Μειονεκτήματα του αλγορίθμου είναι ότι σταματάει όταν φτάσει σε ένα προκαθορισμένο αριθμό συστάδων. Ο biclustering αλγόριθμος δεν εγγυάται ότι οι νωρίτερα αναγνωρισμένες δισυστάδες είναι ανώτερης ποιότητας από αυτές που υπολογίζονται αργότερα.

Επαλήθευση κλάσεων

Μέχρι στιγμής είδαμε αλγόριθμους κατηγοριοποίησης οι οποίοι χώριζαν τα σετ δεδομένων βάση διαφόρων κριτηρίων. Στην κατηγοριοποίηση δεδομένων έκφρασης γονιδίων, τοποθετούσαμε σε ίδιες συστάδες γονίδια που εκδηλώνονται μαζί, δείγματα που εμφάνιζαν τον ίδιο φαινότυπο ή μπλοκ από γονίδια και δείγματα που είχαν σχέση με μία συγκεκριμένη βιολογική διαδικασία. Διαφορετικοί αλγόριθμοι κατηγοριοποίησης ή ακόμα και οι ίδιοι αλγόριθμοι αλλά με διαφορετικές παραμέτρους, γεννούν διαφορετικό σύνολο συστάδων. Η διαδικασία επαλήθευσης συστάδων στοχεύει στην αποτίμηση της ποιότητας και της αξιοπιστίας των διαφόρων συνόλων από συστάδων που προέκυψαν από τις διάφορες κατηγοριοποιήσεις.

Γενικά η διαδικασία επαλήθευσης έχει τρεις σκοπιές:

- Η ποιότητα των συστάδων μετριέται με κριτήρια *ομοιογένειας* και *διαφορετικότητας* σε σχέση με τον ορισμό την συστάδας: τα αντικείμενα στην ίδια συστάδα πρέπει να είναι όμοια μεταξύ τους και διαφορετικά με αυτά άλλων συστάδων
- Μία "*δεδομένη αλήθεια*" (ground truth), η δεδομένη αλήθεια προέρχεται από κεκτημένη γνώση, όπως η γνώση της λειτουργίας οικογένειας γονιδίων, ή από άλλες πηγές όπως κλινική διάγνωση. Η επαλήθευση συστάδας βασίζεται στη συμφωνία μεταξύ των αποτελεσμάτων κατηγοριοποίησης και της δεδομένης αλήθειας.
- Η επαλήθευση συστάδας βασίζεται στην αξιοπιστία της συστάδας, ή την πιθανότητα ότι η συστάδα δεν σχηματίστηκε από τύχη.

Ομοιογένεια και Διαφορετικότητα

Υπάρχουν διάφοροι ορισμοί για την ομοιογένεια των συστάδων οποίες μετράνε την ομοιότητα των στοιχείων στην συστάδα C. Για παράδειγμα:

$$H_1(C) = \frac{\sum_{O_i, O_j \in C, O_i \neq O_j} Similarity(O_i, O_j)}{\| C \| \cdot (\| C \| - 1)}$$

Ο ορισμός αυτός αναπαριστά την ομοιογένεια της C με την μέση ανά ζεύγος ομοιότητα των αντικειμένων στην C. Εναλλακτικά μπορεί να χρησιμοποιηθεί ένας μέσος και βάση αυτού να μετρηθεί η ομοιότητα π.χ.

$H_2(C) = \frac{1}{\| C \|} \sum_{O_i \in C} Similarity(O_i, \overline{O})$, όπου $\overline{O}$ ένα κέντρο βάρους

Ο υπολογισμός της διαφορετικότητας δύο συστάδων C_1, C_2 γίνεται κατ' αναλογία για παράδειγμα:

$$S_1(C_1,C_2) = \frac{\sum_{O_i \in C_1, O_j \in C_2} Similarity(O_i, O_j)}{\| C_1 \| \cdot \| C_2 \|} \quad \text{ή}$$

$S_2(C_1,C_2)$ = Similarity($\overline{O}_1, \overline{O}_2$)

Συμφωνία με σημείο αναφοράς

Αν η "δεδομένη αλήθεια" για την δομή της συστάδας είναι διαθέσιμη, μπορούμε να ελέγξουμε την απόδοση μίας διαδικασίας κατηγοριοποίησης συγκρίνοντας τα αποτελέσματα κατηγοριοποίησης με την "δεδομένη αλήθεια". Δεδομένων αποτελεσμάτων $C = \{C_1, \ldots, C_p\}$, κατασκευάζουμε n*n δυαδικό πίνακα C , όπου n το πλήθος των αντικειμένων με δεδομένα, $C_{ij} = 1$ αν οι O_i και O_j ανήκουν στην ίδια συστάδα, αλλιώς $C_{ij} = 0$. Παρόμοια κατασκευάζουμε δυαδικό πίνακα P για την "*δεδομένη αλήθεια*" $P = \{P_1, \ldots, P_s\}$. Η συμφωνία μεταξύ *C* και *P* μπορεί να αποκαλυφθεί από τις παρακάτω τιμές:

- n_{11} είναι το πλήθος των ζευγών των αντικειμένων (O_i,O_j), όπου $C_{ij}=1$ $P_{ij} = 1$,
- n_{10} είναι το πλήθος των ζευγών των αντικειμένων (O_i,O_j), όπου $C_{ij}=1$ $P_{ij} = 0$,
- n_{01} είναι το πλήθος των ζευγών των αντικειμένων (O_i,O_j), όπου $C_{ij}=0$ $P_{ij} = 1$,
- n_{00} είναι το πλήθος των ζευγών των αντικειμένων (O_i,O_j), όπου $C_{ij}=0$ $P_{ij} = 0$

Μερικοί συνηθισμένοι δείκτες που χρησιμοποιούνται για να μετρηθεί η ομοιότητα μεταξύ των *C* και *P* είναι:

Rand index: $Rand = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + {}_{00}}$,

Jaccard coefficient: $JC = \frac{n_{11}}{n_{11} + n_{01} + n_{10}}$,

Minkowski measure: $Minkowski = \sqrt{\frac{n_{10} + n_{01}}{n_{11} + n_{01}}}$

Αξιοπιστία των συστάδων

Ενώ ένα ευρετήριο αξιοπιστίας μπορεί να χρησιμοποιηθεί για να συγκρίνουμε διαφορετικά αποτελέσματα κατηγοριοποίησης, η σύγκριση αυτή δεν θα αποκαλύψει την αξιοπιστία των συστάδων. Για τον λόγο αυτό υπάρχει η πιθανότητα η συστάδες να μην κατασκευάστηκαν στην τύχη. Ακολουθούν δύο μέθοδοι μέτρησης της σπουδαιότητας των παραγομένων συστάδων:

<u>P-τιμή μίας συστάδας</u>

Ο Tavazoie χαρτογράφησε τα γονίδια σε κάθε παραγόμενη συστάδα σε 199 λειτουργικές κατηγορίες. Για κάθε συστάδα, η P-τιμή υπολογίζεται ώστε να μετρά την στατιστική σπουδαιότητα για εμπλουτισμό της λειτουργικής κατηγορίας. Για να είμαστε ακριβείς, χρησιμοποιήθηκε υπερ-γεωμετρική κατανομή για να υπολογιστεί η πιθανότητα της

παρατήρησης τουλάχιστον k γονιδίων από μία λειτουργική κατηγορία σε μία συστάδα μεγέθους n:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}},$$

όπου f είναι το συνολικό πλήθος γονιδίων στην λειτουργική κατηγορία και g είναι το συνολικό πλήθος γονιδίων στο γονιδίωμα. Αφού η πρόβλεψη του P μέσα στην συστάδα θα είναι μεγαλύτερη από 0.05%, ο συγγραφέας θεωρεί της συστάδες με P-τιμές μικρότερες από $3*10^{-4}$ σημαντικές.

Δύναμη της πρόβλεψης

Μία νεότερη προσέγγιση για να αξιολογηθεί η αξιοπιστία των συστάδων βασίζεται στην σκέψη ότι αν τα αποτελέσματα μίας κατηγοριοποίησης αντικατοπτρίζουν πραγματική δομή συστάδας, τότε ένας "μάντης"(predictor) βασισμένος στις παραγόμενες συστάδες θα μπορούσε να υπολογίσει με ακρίβεια τις ταμπέλες των συστάδων για νέα δείγματα ελέγχου. Για δεδομένα έκφρασης γονιδίων , επιπρόσθετα αντικείμενα δεδομένων σπάνια χρησιμοποιούνται σαν δείγματα ελέγχου, αφού το πλήθος των διαθέσιμων δειγμάτων είναι περιορισμένο. Αντίθετα, μία cross-validation μέθοδος χρησιμοποιείται. Οι παραγόμενες συστάδες αποτιμούνται από επαναληπτικές μετρήσεις της δύναμης πρόβλεψης τους, με ένα ή λίγα από τα αντικείμενα δεδομένων που ξέμειναν σαν δείγματα ελέγχου, καθώς τα εναπομείναντα αντικείμενα χρησιμοποιούνται στην κατηγοριοποίηση. Μία μέθοδος κατασκευής αυτού του "μάντη" προτάθηκε από τον Golub.

Τρέχουσες και μελλοντικές ερευνητικές κατευθύνσεις

Η τεχνολογία των μικροπινάκων έκανε εφικτή την παράλληλη παρατήρηση των επιπέδων έκφρασης χιλιάδων γονιδίων. Τα δεδομένα έκφρασης γονιδίων προσέφερε στην μοριακή βιολογία τεράστιες δυνατότητες. Είδαμε συνοπτικά μερικούς βασικούς αλγόριθμους κατηγοριοποίησης που αναπτύχθηκαν πρόσφατα, και εφαρμόζονται σε δεδομένα έκφρασης γονιδίων με ελπιδοφόρα αποτελέσματα.

Δυστυχώς δεν υπάρχει αλγόριθμος που θα θεωρείται κατάλληλος για όλα τα προβλήματα που έχουν οι βιολόγοι για ένα σύνολο δεδομένων. Δηλαδή δεν υπάρχει αλγόριθμος που να τα καταφέρνει καλύτερα από τους υπόλοιπους σε όλες τις εργασίες που επιτελεί. Γι' αυτό οι ερευνητές καταφεύγουν συνήθως σε μία ομάδα αλγορίθμων.

Τα δεδομένα έκφρασης περιέχουν συνήθως εκατοντάδες γονίδια. Οι βιολόγοι όμως έχουν συχνά διαφορετικές απαιτήσεις στην κατηγοριοποίηση υποσυνόλων των γονιδίων. Μπορεί για παράδειγμα κάποιος να ενδιαφέρεται μόνο για μία μικρή ομάδα γονιδίων. Παρόλ' αυτά όλοι περιμένουν από την ανάλυση να δουν κοινές δομές δεδομένων. Συνήθως οι γραφικές αναπαραστάσεις είναι πιο κατάλληλες και κατανοητές από τους χρήστες, αλλά δεν παρέχουν την δυνατότητα προσαρμογής ανάλογα με τις απαιτήσεις του χρήστη.

Τέλος οι τεχνικές κατηγοριοποίησης θεωρούνται "μη εποπτευόμενες" αφού δεν ξέρουμε εκ των προτέρων το πλήθος των συστάδων που θα δημιουργηθούν, ούτε την κατανομή των δεδομένων στον χώρο των αντικειμένων. Έχουμε συχνά όμως μερική γνώση σχετικά με τα δεδομένα έκφρασης γονιδίων. Επιπλέον ξέρουμε ότι μερικές πειραματικές συνθήκες είναι ισχυρά συνδεδεμένες και οι διαφορές στις δομές των συστάδων σε αυτές τις ομάδες μπορεί να έχουν ιδιαίτερο ενδιαφέρον.

Βιβλιογραφία:

- D. Jiang, A. Zhang, Cluster Analysis for Gene Expression Data: A Survey.
- M. B. Eisen, P. T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA, Vol. 95, pp. 14863–14868, December 1998, Genetics.
- Brazma, J. Vilo, " Minireview: Gene Expression Data Analysis", FEBS Letters 480 (2000) 17-24.