Κεφάλαιο 8

Λειτουργική Ανάλυση της Γονιδιακής Έκφρασης

Σύνοψη

Το κεφάλαιο αυτό αποτελεί φυσική συνέχεια του αμέσως προηγούμενου, παρουσιάζοντας την ανάλυση πειραμάτων γονιδιακής έκφρασης σε λειτουργικό επίπεδο. Οι βασικές βιολογικές έννοιες του συγκεκριμένου κεφαλαίου περιλαμβάνουν τις διαφόρων τύπων οργανώσεις σημασιολογικής πληροφορίας όπως οι βιολογικές οντολογίες και τα βιολογικά μονοπάτια. Επίκεντρο του κεφαλαίου θα είναι το πραγματικό βιολογικό πρόβλημα να προσδιοριστούν οι διαφορικά ρυθμιζόμενες βιολογικές λειτουργίες από ένα πείραμα γονιδιακής έκφρασης. Αφού εισαχθούν οι βασικές μαθηματικές έννοιες του εμπλουτισμού και των λόγων πιθανοτήτων (odds-ratios) θα περιγραφεί μεθοδολογικά η διαδικασία της λειτουργικής ανάλυσης μέσω της χρήσης της υπεργεωμετρικής κατανομής. Στη συνέχεια θα αναλυθούν οι σημαντικές έννοιες, τόσο για τη λειτουργική ανάλυση όσο και για τη μελέτη της γονιδιακής έκφρασης γενικότερα, του ελέγχου πολλαπλών υποθέσεων και τις σχετικές διορθώσεις των αποτελεσμάτων. Στο τελευταίο μέρος του κεφαλαίου θα παρουσιαστούν οι σημαντικότερες κατηγορίες υπολογιστικών μεθοδολογίων για τη λειτουργική ανάλυση της γονιδιακής έκφρασης.

Στο τέλος του Κεφαλαίου θα πρέπει να μπορείτε:

- *Να αποκομίσετε από το διαδίκτυο πληροφορίες σχετικά με τις λειτουργίες ενός ή περισσοτέρων γονιδίων.*
- *Να προσδιορίσετε τις λειτουργίες εκείνες που είναι σημαντικότερες για ένα δεδομένο πείραμα μέσω της ανάλυσης των σχετικών επιπέδων έκφρασης των γονιδίων.*
- *Να αξιολογήσετε στατιστικά τα αποτελέσματα μιας λειτουργικής ανάλυσης διενεργώντας έλεγχο πολλαπλών υποθέσεων.*

Εισαγωγή

Στο προηγούμενο κεφάλαιο είδαμε πώς μπορούμε να αναλύσουμε τις σχετικές συγκεντρώσεις του συνόλου (σχεδόν) των γονιδίων ενός οργανισμού σε μια δεδομένη συνθήκη, με σκοπό να διερευνήσουμε το "πρόγραμμα έκφρασης", τη διαμόρφωση δηλαδή των επιπέδων των mRNA μορίων που προέρχονται από κάθε γονίδιο, ως μέρος μιας σφαιρικότερης αποτύπωσης των κυτταρικών λειτουργιών. Τα αποτελέσματα της μελέτης της γονιδιακής έκφρασης οργανώνονται σε σύνολα γονιδίων, των οποίων τα επίπεδα έκφρασης μεταβάλλονται σημαντικά μεταξύ δύο (ή περισσότερων) συνθηκών και τα οποία ονομάζουμε "διαφορικά εκφραζόμενα" (differentially expressed genes, DEG). Εναλλακτικά, μέσω μεθόδων ομαδοποίησης, τις οποίες συζητήσαμε στο αμέσως προηγούμενο κεφάλαιο, τα γονίδια μπορούν να καταταχθούν σε υποσύνολα, τα οποία εκφράζονται με παρόμοιο τρόπο σε διάφορες καταστάσεις. Σε κάθε περίπτωση, η πληροφορία που προκύπτει από την ανάλυση πειραμάτων έκφρασης συνοψίζεται σε καταλόγους γονιδίων που είναι συχνά αρκετά εκτεταμένοι. Η εξαγωγή συμπερασμάτων για τις κυτταρικές λειτουργίες, τις μοριακές διεργασίες που επισυμβαίνουν στη συνθήκη που μελετούμε είναι δύσκολη, αν όχι αδύνατη, μέσω της απλής επισκόπησης τέτοιων καταλόγων. Είναι έτσι απαραίτητο ένα επόμενο στάδιο ανάλυσης που θα συνδέσει τα υποσύνολα διαφορικά εκφραζόμενων, ή συνεκφραζόμενων γονιδίων με συγκεκριμένα λειτουργικά χαρακτηριστικά και βιολογικές ιδιότητες.

Στο κεφάλαιο αυτό, που αποτελεί φυσική συνέχεια του προηγούμενου, θα δούμε πώς μπορούμε να αναλύσουμε αποτελέσματα αυτής της μορφής με τρόπο που να γίνεται αυτή η σύνδεση. Το αντικείμενο αυτού του κεφαλαίου, είναι η λειτουργική ανάλυση της γονιδιακής έκφρασης που συνίσταται στη μελέτη αποτελεσμάτων πειραμάτων έκφρασης σε συνδυασμό με προϋπάρχουσα γνώση για τις κυτταρικές διεργασίες, τη μοριακή λειτουργία και τα βιολογικά μονοπάτια που ανά πάσα στιγμή μπορούν να είναι ενεργοποιημένα (ή αντίθετα κατεσταλμένα) σ' ένα κύτταρο. Η προϋπάρχουσα γνώση και ο τρόπος με τον οποίον αυτή έχει αποδελτιωθεί και οργανωθεί σε συγκεκριμένες δομές δεδομένων, αποτελεί ένα βασικό κομμάτι αυτού του κεφαλαίου, το οποίο αφιερώνεται στις έννοιες των γονιδιακών οντολογιών και σε μια περίληψη των διαθέσιμων "αποθετηρίων" (repositories) της γνώσης μας για τη γονιδιακή λειτουργία σε διάφορα επίπεδα. Η διαρκώς εμπλουτιζόμενη αυτή γνώση, όπως προκύπτει από τη συνεχή μελέτη ενός τεράστιου αριθμού βιολογικών συστημάτων σε μικρή, μεσαία και μεγάλη κλίμακα, αλλά κυρίως ο τρόπος με τον οποίον αυτή είναι οργανωμένη, αποτελεί βασικό στοιχείο της λειτουργικής ανάλυσης.

Στο δεύτερο μέρος του Κεφαλαίου θα δούμε με ποιες τεχνικές μπορούμε να συνδυάσουμε την προϋπάρχουσα γνώση με τα αποτελέσματα ενός συγκεκριμένου πειράματος προκειμένου να εξάγουμε χρήσιμες πληροφορίες για αυτό. Θα συζητήσουμε κάποιες απλές (και κάποιες λιγότερο απλές) αρχές στατιστικής ανάλυσης για το συγκερασμό αυτών των δεδομένων και θα εξετάσουμε μεθόδους για την καλύτερη δυνατή ερμηνεία των αποτελεσμάτων ενός πειράματος γονιδιακής έκφρασης.

Θα ξεκινήσουμε -όπως συνήθως- με τη διατύπωση ενός βιολογικού προβλήματος.

Το βιολογικό πρόβλημα

Φανταστείτε ότι έχετε μόλις ολοκληρώσει ένα πείραμα γονιδιακής έκφρασης. Το πείραμα συνίσταται στη μέτρηση της σχετικής έκφρασης 20000 γονιδίων μιας ανθρώπινης κυτταρικής σειράς πριν και μετά την επίδραση μιας ουσίας με κυτταροστατική δράση, που είναι γνωστό δηλαδή ότι αναστέλλει την κυτταρική διαίρεση. Ο σκοπός του πειράματος είναι αρχικά να επιβεβαιωθεί ότι η κυτταροστατική δράση της ουσίας μεσολαβείται από συγκεκριμένα γονίδια που εμπλέκονται στη διεργασία της κυτταρικής διαίρεσης και στη συνέχεια να μελετηθούν οι πιθανές διαταραχές που η ουσία προκαλεί σε άλλες λειτουργίες τους κυττάρου. Από την ανάλυση της διαφορικής έκφρασης (με τυπικά κριτήρια $\log_2$FC και p-value, βλ. Κεφάλαιο 7) προκύπτει μια λίστα με 1000 διαφορικά εκφραζόμενα γονίδια και μετά από μια πρώτη επισκόπηση της λίστας αυτής, εντοπίζετε αρκετά από τα γονίδια, τα οποία, από την ερευνητική σας δουλειά, γνωρίζετε ότι σχετίζονται με τις λειτουργίες του κυτταρικού κύκλου και της κυτταρικής διαίρεσης. Αναλογιζόμενοι τις συνθήκες του πειράματος που διενεργήσατε, θεωρείτε ότι η δράση της ουσίας αντανακλάται στο πρότυπο γονιδιακής έκφρασης και πως κάτι τέτοιο είναι απολύτως λογικό. Σε ποιο βαθμό, ωστόσο θα μπορούσατε να υποστηρίξετε αυστηρά αυτή σας την παρατήρηση; Υπάρχει η δυνατότητα να ποσοτικοποιήσετε το επίπεδο επίδρασης της κυτταροστατικής ουσίας στη λειτουργία της κυτταρικής διαίρεσης και αν ναι, θα μπορούσατε να ελέγξετε και ποιες άλλες λειτουργίες ενδεχομένως έχει επηρρεάσει;

Έχοντας στα χέρια σας έναν μεγάλο αριθμό γονιδίων και έχοντας διενεργήσει μια ανάλυση έκφρασης σε γονιδιωματικό επίπεδο θα είναι λάθος να μείνετε στον εντοπισμό μιας δεκάδας γονιδίων από τη λίστα σας και να μην προχωρήσετε σε πιο αναλυτικά ερωτήματα όπως:

- *Είναι αυτοί οι αριθμοί γονιδίων μεγάλοι ή μικροί και πώς θα μπορούσαμε να αξιολογήσουμε κάτι τέτοιο;*
- *Πώς θα μπορούσαμε να επεκτείνουμε αυτήν την ανάλυση και σε άλλες λειτουργίες; Πώς θα μπορούσαμε να προσδιορίσουμε τις λειτουργίες εκείνες, για τις οποίες οι αριθμοί των διαφορικά εκφραζόμενων γονιδίων είναι σημαντικά μεγάλοι (ή αντίστοιχα σημαντικά μικροί);*
- *Πόσα από τα διαφορικά εκφραζόμενα γονίδια συνδέονται με τις λειτουργίες του κυτταρικού κύκλου και της κυτταρικής διαίρεσης;*

Τα παραπάνω ερωτήματα συνθέτουν τον πυρήνα αυτού που ονομάζουμε λειτουργική ανάλυση της γονιδιακής έκφρασης. Στη συνέχεια αυτού του κεφαλαίου θα δούμε με ποιον τρόπο μπορούμε να προσπαθήσουμε να τα προσεγγίσουμε ξεκινώντας από κάποια απλούστερα ερώτηματα που προηγούνται και έχουν να κάνουν με την οργάνωση των απαραίτητων δεδομένων. Έπειτα, θα περάσουμε στη συζήτηση γύρω από μεθοδολογίες που σχετίζονται με τη συνδυαστική μελέτη των αποτελεσμάτων ενός πειράματος στο πλαίσιο της προϋπάρχουσας γνώσης.

Αντικείμενα Λειτουργικής Ανάλυσης

Ένα από τα ερωτήματα που θέσαμε παραπάνω ήταν το εξής:

Πόσα από τα διαφορικά εκφραζόμενα γονίδια συνδέονται με τις λειτουργίες του κυτταρικού κύκλου και
της κυτταρικής διαίρεσης;

Το ερώτημα αυτό αναδεικνύει την πρωταρχική ανάγκη για σαφώς ορισμένα και αξιόπιστα δεδομένα σχετικά με τη λειτουργία, ή καλύτερα τις λειτουργίες των γονιδίων του οργανισμού που μελετούμε. Οι σχέσεις γονιδίων-λειτουργιών βρίσκονται προφανώς στη βάση της λειτουργικής ανάλυσης κι έτσι, προκειμένου να απαντήσουμε στην πιο πάνω ερώτηση, θα πρέπει να έχουμε στη διάθεσή μας τα ονόματα των γονιδίων, τα οποία γνωρίζουμε ότι σχετίζονται με τη λειτουργία του κυτταρικού κύκλου. Χρειαζόμαστε δηλαδή με κάποιον τρόπο ένα "ευρετήριο" γονιδίων που θα τα κατηγοριοποιεί ανάλογα με τις διαπιστωμένες (ή προβλεπόμενες) λειτουργίες τους. Αν γνωρίζαμε τα γονίδια που σχετίζονται (σε κάποιο επίπεδο) με τις λειτουργίες του κυτταρικού κύκλου και της κυτταρικής διαίρεσης δεν θα είχαμε παρά να απαριθμήσουμε πόσα από αυτά βρίσκονται στη λίστα με τα 1000 διαφορικά εκφραζόμενα γονίδια του πειράματός μας για να απαντήσουμε στο ερώτημά μας και να κάνουμε έτσι το πρώτο βήμα για τη λειτουργική ανάλυσή του.

Τέτοια "ευρετήρια" καταλόγων γονιδίων ευτυχώς υπάρχουν. Είναι το αποτέλεσμα του συνδυασμού της ερευνητικής δουλειάς χιλιάδων ανθρώπων επί δεκαετίες και της οργανωτικής προσπάθειας κάποιων οξυδερκών βιοεπιστημόνων που πριν από μερικά χρόνια διέβλεψαν πως η αποδελτίωση της υπάρχουσας βιολογικής γνώσης, σχετικα με τη γονιδιακή λειτουργία, σε καλά οργανωμένες δομές δεδομένων θα επέτρεπε μια σειρά από περαιτέρω συνδυαστικές αναλύσεις. Δομές δεδομένων αυτού του τύπου είναι οι διάφορες κατηγοριοποιήσεις γονιδίων σε λειτουργικές ομάδες μέσω σχημάτων όπως είναι οι οντολογίες (Lambrix, Tan, Jakonienė & Strömbäck, 2007) που θα συζητήσουμε αμέσως μετά.

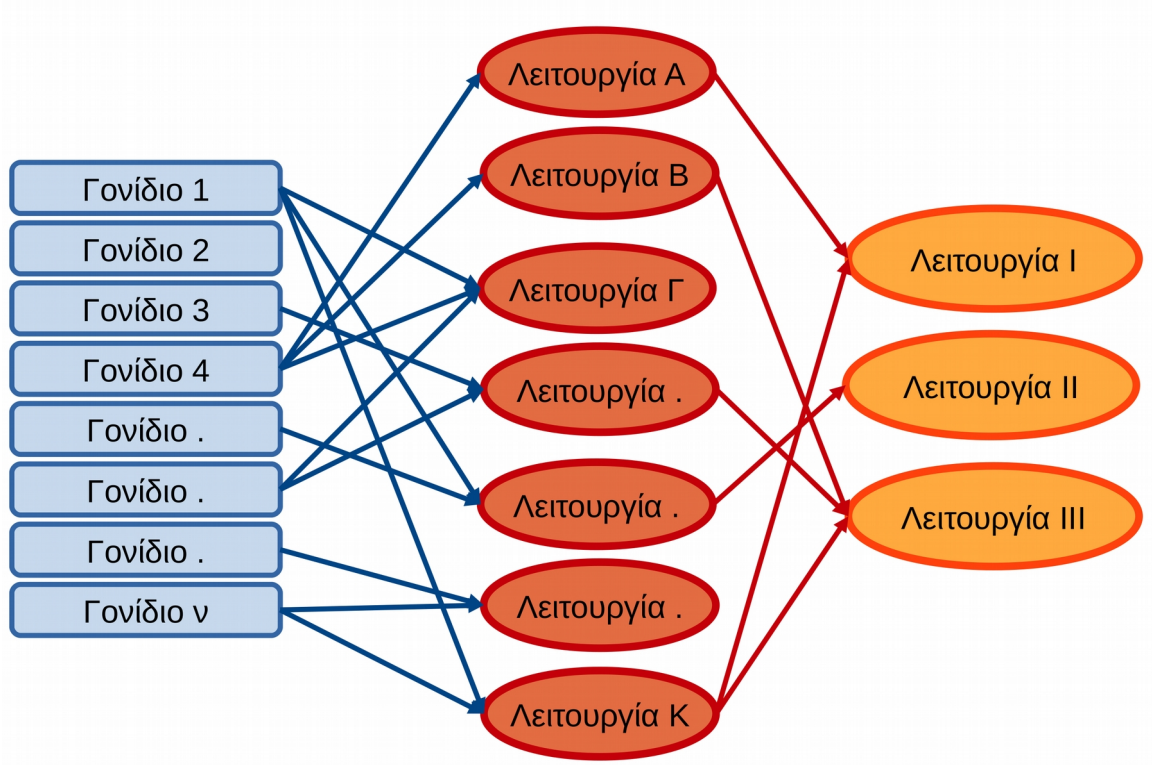
Κατηγοριοποιήσεις γονιδίων. Γονιδιακές οντολογίες.

Η βάση για τη διενέργεια μιας οποιασδήποτε ανάλυσης γονιδίων σε λειτουργικό επίπεδο είναι μια κατηγοριοποίηση αυτών των γονιδίων με τρόπο συνεκτικό, σαφή και όσο το δυνατόν αναλυτικότερο. Με τον όρο "κατηγοριοποίηση" αναφερόμαστε σε οποιοδήποτε σχήμα καταλόγων αντιστοιχεί τα αντικείμενα της ανάλυσής μας, δηλαδή τα γονίδια, με τις ιδιότητες τις οποίες θέλουμε να εξετάσουμε, δηλαδή τις λειτουργίες τους. Μπορούμε έτσι να φανταστούμε ένα σχήμα όπως αυτό που φαίνεται στην Εικόνα 8.1 όπου ένας κατάλογος από γονίδια αντιστοιχίζονται σε λειτουργίες. Οι σχέσεις μεταξύ γονιδίων και λειτουργιών δεν είναι αλληλο-αποκλειόμενες και οι αντιστοιχίσεις είναι ελεύθερες, έτσι ώστε ένα γονίδιο να μπορεί να αντιστοιχεί σε περισσότερες από μία λειτουργίες. Η πληρότητα της κατηγοριοποίησης συνίσταται στην "κάλυψη" του καταλόγου των γονιδίων, καθώς το κατά πόσο μία τουλάχιστο λειτουργία θα αποδίδεται σε κάθε γονίδιο εξαρτάται από την έκταση της προϋπάρχουσας γνώσης, την οποία αποδελτιώνει η συγκεκριμένη κατηγοριοποίηση. Είναι προφανές ότι συγκεκριμένες κατηγοριοποιήσεις δεν είναι πλήρης καθώς εξ ορισμού δεν μπορούν να αφορούν το σύνολο των γονιδίων. Περισσότερα για επιμέρους θέματα όπως αυτό θα δούμε παρακάτω.

Κατηγοριοποιήσεις που δε βασίζονται σε μια απλή αντιστοίχηση, αλλά εμπεριέχουν και ιεραρχικές σχέσεις μεταξύ των ιδιοτήτων με τρόπο τέτοιο που να οδηγούν σε σύνθετες ταξινομήσεις, ονομάζονται οντολογίες. Η έννοια της **οντολογίας** είναι πρωταρχικά φιλοσοφική και αφορά τη μελέτη ζητημάτων του "είναι" της ίδιας δηλαδή της ύπαρξης, αποσκοπώντας ουσιαστικά στην κατατάξη διαφόρων οντοτήτων με βάση τα χαρακτηριστικά που τα προσδιορίζουν. Οι κατατάξεις αυτές είναι συχνά ιεραρχικές και προσομοιάζουν με **ταξινομήσεις**, διαδικασίες που δεν

είναι άγνωστες στο χώρο των βιολογικών επιστημών. Στην επιστήμη των υπολογιστών, η οντολογία συνδυάζει την έννοια μας ιεραρχικής ταξινόμησης με την οργανωμένη απόδοση χαρακτηριστικών σε μια σειρά από οντότητες που ορίζονται με ακριβή τρόπο. Στην περίπτωσή μας, οι οντότητες αυτές είναι τα γονίδια και τα χαρακτηριστικά με τα οποία οργανώνονται είναι οι λειτουργικές τους ιδιότητες, όπως οι διεργασίες στις οποίες συμμετέχουν μέσα στο κύτταρο, τα βιολογικά μονοπάτια των οποίων αποτελούν μέρος ή τα διάφορα οργανίδια του κυττάρου (πυρήνας, μιτοχόνδριο, ενδοπλασματικό δίκτυο κλπ.) στα οποία εντοπίζονται. Οι οντολογίες που οργανώνονται σε αυτή τη βάση ονομάζονται, εύλογα, **γονιδιακές οντολογίες**.

Στη συνέχεια αυτής της ενότητας θα δούμε χαρακτηριστικές κατηγοριοποιήσεις γονιδίων που χρησιμοποιούνται στη λειτουργική ανάλυση και που με αυτόν τον τρόπο ορίζουν τα επίπεδα στα οποία διενεργείται η ανάλυση της γονιδιακής έκφρασης.



Εικόνα 8.1: *Σχηματική αναπαράσταση της δημιουργίας μιας κατηγοριοποίησης γονιδίων. Οι σχέσεις μεταξύ των αντικειμένων και των ιδιοτήτων τους δεν είναι αλληλο-αποκλειόμενες και ένα γονίδιο μπορεί να αντιστοιχηθεί σε περισσότερες από μία ιδιότητες. Στο σχήμα της Εικόνας το Γονίδιο 2 δεν αποδίδεται σε καμία λειτουργία, ενώ τα 1 και 4 αποδίδονται σε τρεις διαφορετικές λειτουργίες. Οι ίδιες οι λειτουργίες οργανώνονται ιεραρχικά από ειδικότερες (Α, Β, Γ κλπ) σε γενικότερες (Ι, ΙΙ, ΙΙΙ).*

Επίπεδα λειτουργικής ανάλυσης

Γονιδιακή Οντολογία (GO)

Η Γονιδιακή Οντολογία αποτέλεσε ένα από τα πιο σημαντικά πρώτα βήματα προς την

επίλυση ενός σημαντικού βιολογικού προβλήματος: την αποτύπωση του συνόλου της βιολογικής γνώσης σχετικά με τη γονιδιακή λειτουργία με ένα σαφή και οργανωμένο τρόπο. Αποτελεί το αποτέλεσμα της δουλειάς του αντίστοιχου Consortium που συστήθηκε το 1998 με σκοπό τη δημιουργία μιας συνεκτικής, παγκόσμιας ονοματολογίας γονιδίων για όλους τους οργανισμούς (Ashburner et al., 2000). στο πλαίσιο αυτής της προσπάθειας, η βάση δεδομένων της Γονιδιακής Οντολογίας (www.geneontology.org) έχει δημιουργήσει τρεις βασικές, ανεξάρτητες μεταξύ τους, κατηγορίες (οντολογίες) που περιγράφουν τα γονίδια όλων των μελετημένων οργανισμών στη βάση:

1. Της κυτταρικής διεργασίας στην οποία συμμετέχουν οι πρωτεΐνες τους. Ποια είναι δηλαδή η λειτουργία του γονιδίου.
2. Της μοριακής λειτουργίας των πρωτεϊνικών προϊόντων τους. Με ποιον τρόπο επιτελεί μοριακά τη λειτουργία της η πρωτεΐνη που αντιστοιχεί στο γονίδιο.
3. Του κυτταρικού εντοπισμού των πρωτεϊνών τους. Σε ποιο τμήμα/οργανίδιο του κυττάρου επιτελείται αυτή η λειτουργία.

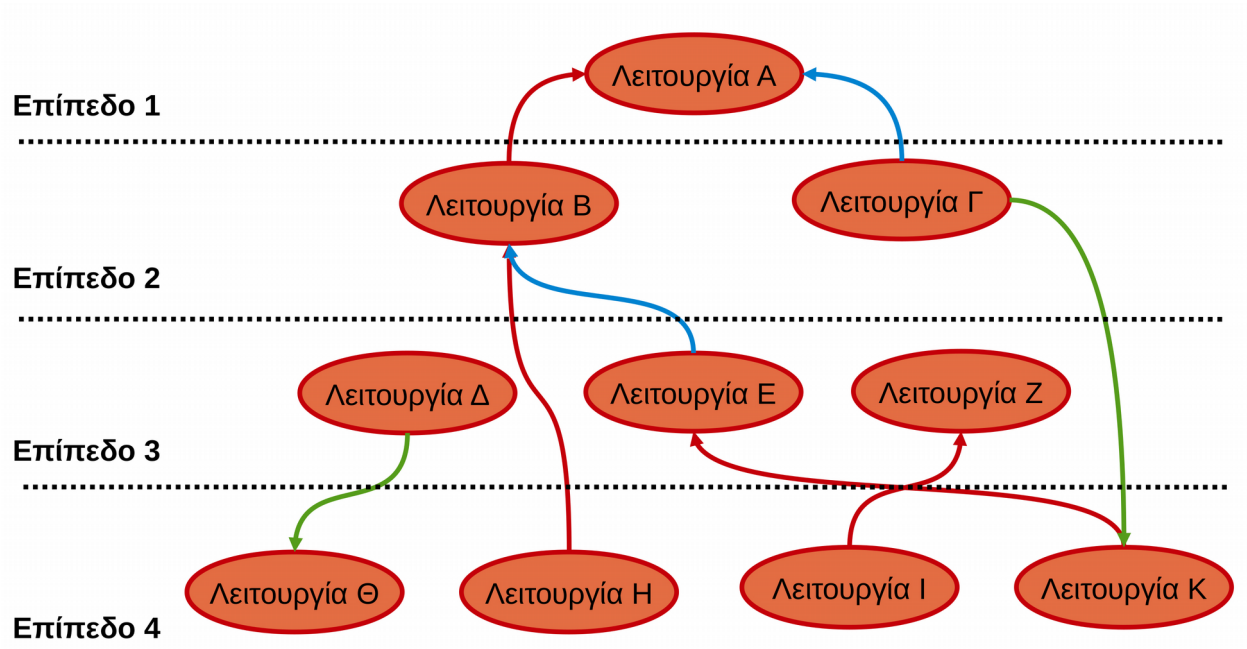
στο πλαίσιο της γονιδιακής οντολογίας, τα γονίδια αναπαρίστανται ως τμήματα-μονοπάτια στους κατευθυνόμενους ακυκλικούς γράφους που ορίζονται από την οντολογία. Μπορούμε να καταλάβουμε καλύτερα τι σημαίνει αυτό μέσα από το σχήμα που φαίνεται στην Εικόνα 8.2. Η οντολογία ορίζεται από τις επιμέρους σχέσεις μεταξύ των καταλογογραφημένων ιδιοτήτων που είναι οι διάφορες γονιδιακές λειτουργίες και οι οποίες εκφράζονται ως όροι (terms) της οντολογίας. Παραδείγματα τέτοιων όρων είναι π.χ. η “κυτταρική διαίρεση”, η “μίτωση” ή ο “μεταβολισμός υδατανθράκων”. Οι σχέσεις μεταξύ των όρων αυτών ορίζονται ιεραρχικά και οργανώνονται σε κατευθυνόμενους ακυκλικούς γράφους (directed acyclic graph) όπως αυτός της Εικόνας 8.2. Έτσι π.χ. ένας όρος που περιγράφεται ως “βιοσύνθεση υδατανθράκων” αποτελεί μέρος της γενικότερης κατηγορίας “μεταβολισμός υδατανθράκων”, ένας όρος που ονομάζεται “κυτταρική μετανάστευση κατά τη διαφοροποίηση” είναι μια από τις υποκατηγορίες του όρου “κυτταρική μετανάστευση” ενώ, τέλος, ο όρος “ρύθμιση της κυτταρικής διαίρεσης” αντιστοιχεί σε μια ρυθμιστική λειτουργία για τον όρο “κυτταρική διαίρεση”. Τα παραπάνω παραδείγματα συνοψίζουν τα τρία βασικά είδη σχέσεων που μπορούν να ενώσουν στοιχεία/όρους σ' έναν ακυκλικό γράφο γονιδιακής οντολογίας. Αυτά είναι:

1. “**είναι**” (is). Αντιστοιχεί σε μια οντολογική σχέση ιεραρχικών ιδιοτήτων. Έτσι ο όρος “ενδο-ριβο-νουκλεάση” είναι θυγατρικός όρος του “ριβονουκλεάση”. Ταυτόχρονα είναι και θυγατρικός όρος του “ενδονουκλεάση”.
2. “**αποτελεί μέρος του**” (is part). Αντιστοιχεί σε μια πιο αυστηρή σχέση που δεν μπορεί να έχει περισσότερο από έναν μητρικό όρο. Για παράδειγμα ο όρος που αντιστοιχεί στον κυτταρικό εντοπισμό “P-σωμάτια” (p-bodies) είναι αποκλειστικά και μόνος μέρος του όρου “κυτταρόπλασμα”.
3. “**ρυθμίζει**” (regulates). Είναι μια προφανής σύνδεση που σχετίζεται με τη λειτουργία της ρύθμισης της έντασης μια διεργασίας.

Στην Εικόνα 8.2 οι διαφορετικές αυτές σχέσεις συμβολίζονται με ακμές διαφορετικών χρωμάτων. Τα γονίδια αποτελούν τμήματα αυτού του γράφου και λόγω ακριβώς του ιεραρχικού χαρακτήρα των ιδιοτήτων διατρέχουν περισσότερους από έναν κόμβους, κάποιες φορές σε διαφορετικούς κλάδους ταυτόχρονα. Ένα γονίδιο μπορεί να είναι έτσι ένας μεταγραφικός παράγοντας που εμπλέκεται στις διαδικασίες της διαφοροποίησης αλλά να επιτελεί τη λειτουργία

του μέσω αλλαγών στο μεταβολισμό ή την κυτταρική διαίρεση. Μ' αυτόν τον τρόπο κάθε γονίδιο αντιστοιχεί μ' ένα γράφο γονιδιακής οντολογίας. Ο γράφος αυτός περιλαμβάνει όλες τις λειτουργίες που είναι γνωστές για το συγκεκριμένο γονίδιο και δεν είναι παρά ένα μικρό υποσύνολο του γενικότερου υπερ-γράφου που ορίζεται από το σύνολο της οντολογίας (Bourne, 2009). Κάθε γονίδιο έτσι, αντιστοιχίζεται σε μια σειρά από λειτουργίες με τον τρόπο που φαίνεται στην Εικόνα 8.1, αλλά οι λειτουργίες αυτές είναι οργανωμένες ιεραρχικά όπως φαίνεται στην Εικόνα 8.2.

Η δημιουργία, διατήρηση και ενημέρωση της γονιδιακής οντολογίας είναι διαρκής και γίνεται μέσω της συνδυασμένης προσπάθειας ενός μεγάλου αριθμού εργαστηρίων, των οποίων ο ρόλος είναι η προσθήκη νέων όρων και η σύνδεση όρων με γονίδια (ή αντίστοιχα η αποσύνδεσή τους) με βάση το σύνολο της αποδελτιωμένης βιολογικής γνώσης (Osborne et al., 2009). Η σύνδεση των γονιδίων με τους όρους που τους αποδίδονται γίνεται στη βάση α) πειραματικών δεδομένων (experimental evidence) ή β) υπολογιστικών προβλέψεων (computational evidence). Χαρακτηριστικά παραδείγματα πειραματικής επιβεβαίωσης είναι η γενετική αλληλεπίδραση (π.χ. πειράματα απαλοιφής γονιδίων) και οι μετρήσεις γονιδιακής έκφρασης ενώ οι υπολογιστικές προβλέψεις περιλαμβάνουν την ομοιότητα σε επίπεδο αλληλουχίας, τις φυλογενετικές σχέσεις κλπ. Είναι προφανές ότι είναι σημαντικό να γνωρίζουμε (και να ελέγχουμε) τη βάση πάνω στην οποία μια λειτουργία έχει αποδοθεί σε ένα γονίδιο.



Εικόνα 8.2: *Σχηματική αναπαράσταση ενός κατευθυνόμενου ακυκλικού γράφου που περιγράφει σχέσεις γονιδιακής οντολογίας. Οι διαφορετικές σχέσεις μεταξύ των όρων ("είναι", "μέρος του" και "ρυθμίζει") συμβολίζονται με κόκκινο, γαλάζιο και πράσινο αντίστοιχα. Οι σχέσεις μεταξύ των διαφόρων όρων διαρθρώνονται ιεραρχικά σε διακριτά επίπεδα (1ο, 2ο, 3ο κ.ο.κ)*

Βιολογικά Μονοπάτια

Τα βιολογικά μονοπάτια δεν είναι γονιδιακές οντολογίες καθώς δεν αναπαριστούν ιεραρχικές

σχέσεις. Ωστόσο από πλευράς οργάνωσης της πληροφορίας θεωρούνται σύνθετες κατηγοριοποιήσεις καθώς αναπαριστούν τις λειτουργικές σχέσεις μεταξύ των γονιδίων (για την ακρίβεια των πρωτεϊνικών τους προϊόντων) που συμμετέχουν σε κοινά βιολογικά μονοπάτια. Τέτοια μονοπάτια μπορεί να σχετίζονται με:

- Μεταβολικές διεργασίες (π.χ. "Γλυκόλυση")
- Μεταφορά και επεξεργασία γονιδιωματικής πληροφορίας (π.χ. "Μεταγραφή mRNA")
- Μεταφορά και επεξεργασία περιβαλλοντικής πληροφορίας (π.χ. "Σηματοδότηση μέσω NF-κΒ")
- Κυτταρικές διεργασίες (π.χ. "Ενδοκύττωση")
- Βιολογικά συστήματα οργάνων (π.χ. "Έκκριση ινσουλίνης")
- Ασθένειες (π.χ. "Μόλυνση από Salmonella")

Η οργάνωση της πληροφορίας στα βιολογικά μονοπάτια είναι της μορφής του πιο απλού σχήματος της Εικόνας 8.1 όπου ένα γονίδιο αντιστοιχίζεται σε ένα ή περισσότερα μονοπάτια, αυτό που διαφέρει ωστόσο είναι ότι μέσω της λεπτομερούς αναπαράστασης της προϋπάρχουσας γνώσης για κάθε μονοπάτι, μπορούμε να έχουμε άμεση πρόσβαση στην πληροφορία σχετικά με την εγγύτητα των αλληλεπιδράσεων μεταξύ γονιδίων.

Σε αντίθεση με τη γονιδιακή οντολογία, όπου ένα διεθνές consortium είναι υπεύθυνο για την ενημέρωση και διατήρηση της βάσης δεδομένων, υπάρχουν διάφορα αποθετήρια πληροφορίας για τα βιολογικά μονοπάτια:

Η KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/pathway.html) (Kanehisa & Goto, 2000) αποτελεί την παλαιότερη βάση δεδομένων βιολογικών μονοπατιών. Περιέχει ίσως το μεγαλύτερο όγκο πληροφορίας κι είναι, από άποψη κατηγοριών, η πληρέστερη βάση, ωστόσο μειονεκτεί σε ό,τι αφορά την αναπαράσταση των μονοπατιών. Από την άλλη πλευρά, λόγω ακριβώς της παλαιότητάς της είναι ενσωματωμένη στα περισσότερα διαδικτυακά εργαλεία λειτουργικής ανάλυσης (όπως θα δούμε παρακάτω) και γι' αυτό το λόγο είναι η πιο ευρέως χρησιμοποιούμενη.

Η Biocarta (http://www.biocarta.com/) (Nishimura, 2001) παρέχει οργανωμένη πληροφορία για μεγάλο αριθμό μονοπατιών που διακρίνεται μεταξύ του ανθρώπινου γονιδιώματος κι αυτού του ποντικού. Παρέχει εικόνες καλύτερης ποιότητας που είναι όμως στατικές όπως και αυτές της KEGG. Η Reactome (http://www.reactome.org/) (Joshi-Tope et al., 2005) αντίθετα, οργανώνει τα μονοπάτια με μια ιεραρχική αντίληψη που προσομοιάζει λίγο τη λογική μιας οντολογίας μονοπατιών και που επιτρέπει τη διαδραστική αναζήτηση μονοπατιών-γονιδίων ενώ συνδυάζει και εξωτερικά δεδομένα όπως δεδομένα έκφρασης, δομές πρωτεϊνών κλπ. Λόγω της πιο πρόσφατης δημιουργίας της και της πολύπλοκης δομής της, έχει το μειονέκτημα να μην εμπεριέχεται στις τυπικές συλλογές διαδικτυακών εργαλείων λειτουργικής ανάλυσης.

Άλλες Κατηγοριοποιήσεις

Οι γονιδιακές οντολογίες και τα βιολογικά μονοπάτια αποτελούν τις πιο χαρακτηριστικές κατηγοριοποιήσεις γονιδίων. Ιστορικά, ήταν οι πρώτες που οργανώθηκαν σε αναζητήσιμες βάσεις

δεδομένων και που ενσωματώθηκαν σε διαδικτυακά εργαλεία λειτουργικής ανάλυσης. Δεν είναι ωστόσο οι μόνες και ο ολοένα αυξανόμενος όγκος γονιδιωματικών δεδομένων έχει οδηγήσει στην ανάπτυξη διαφόρων επιπλέον ομαδοποιήσεων γονιδίων (gene groupings) που περιέχουν πληροφορία πέρα από την τυπική που σχετίζεται με τη φυσική λειτουργία των πρωτεϊνών που κωδικοποιείται από τα γονίδια. Πρακτικά, κανείς μπορεί πλέον να συγκεντρώσει υποσύνολα γονιδίων και να τα κατηγοριοποιήσει ανάλογα με οποιαδήποτε ιδιότητά τους μπορεί να μετρηθεί σε μεγάλη κλίμακα. Έτσι, υπάρχουν κατηγοριοποιήσεις γονιδίων ανάλογα με το μεταγραφικό παράγοντα που τα ενεργοποιεί ή το μη-κωδικό RNA που τα καταστέλλει, κατηγοριοποιήσεις με βάση τη συσχέτιση γονιδίων με συγκεκριμένες ασθένειες κλπ. Στη συνέχεια παραθέτουμε συνοπτικά κάποιες από τις πιο χαρακτηριστικές ομαδοποιήσεις αυτού του τύπου, οργανωμένες σε τρεις μεγάλες κατηγορίες: ομαδοποιήσεις γονιδίων που βασίζονται α) σε πρωτεϊνικές αλληλεπιδράσεις (βλ. Κεφάλαιο 9), β) σε μελέτες σύνδεσης γονιδίων με ασθένειες (βλ Κεφάλαιο 10) και γ) σε πειράματα γονιδιωματικής σε μεγάλη κλίμακα (βλ. Κεφάλαιο 11).

***Ερώτηση**: Μπορείτε να σκεφτείτε άλλες ομαδοποιήσεις γονιδίων που θα βοηθούσαν στην πραγματοποίηση συγκεκριμένων λειτουργικών αναλύσεων;*

Δίκτυα πρωτεϊνικών αλληλεπιδράσεων

Μέσα στο κύτταρο, οι πρωτεΐνες πολύ συχνά δεν επιτελούν αυτόνομες λειτουργίες αλλά αντίθετα οργανώνονται σε πολυ-πρωτεϊνικά σύμπλοκα προκειμένου να ολοκληρώσουν πολύπλοκες διεργασίες. Έτσι, π.χ. η αποσύνθεση των πρωτεϊνών από το πρωτεάσωμα ή η σύνθεσή τους από το ριβόσωμα επιτελούνται από μεγάλα πρωτεϊνικά σύμπλοκα στα οποία συμμετέχει μεγάλος αριθμός πρωτεϊνών. Πιο σύνθετες διεργασίες, όπως π.χ. η ανοσοαπόκριση, απαιτούν τη συνδυασμένη λειτουργία περισσότερων του ενός συμπλόκων και συνεπώς ακόμα μεγαλύτερου αριθμού πρωτεϊνών. Η οργάνωση των πρωτεϊνικών μορίων σε σύμπλοκα και η συνδυασμένη δράση συμπλόκων αποτελεί ανοιχτό πεδίο έρευνας, η βάση του οποίου είναι η ταυτοποίηση και μελέτη του τρόπου με τον οποίο οι πρωτεΐνες αλληλεπιδρούν μεταξύ τους. Σταδιακά, η ταυτοποίηση ενός μεγάλου αριθμού αλληλεπιδράσεων του τύπου “το Α συνδέεται με το Β” οδηγούν στη δημιουργία δικτύων αλληλεπιδράσεων πρωτεϊνών (Vazquez, Flammini, Maritan & Vespignani, 2003).

Κατηγοριοποιήσεις γονιδίων που βασίζονται σε τέτοια δίκτυα είναι αρκετά χρήσιμες για τη μελέτη σύνθετων κυτταρικών διεργασιών που εμπλέκουν περισσότερες από μια λειτουργίες. Παρόλο που η ιεραρχική δομή των γονιδιακών οντολογιών περιλαμβάνει (σε ανώτερα επίπεδα οργάνωσης) τέτοιες σύνθετες διεργασίες, ο αριθμός των γονιδίων που αναφέρονται σε αυτές είναι συχνά πολύ μεγάλος, τέτοιος που δεν επιτρέπει την εξαγωγή χρήσιμων συμπερασμάτων. Τα δίκτυα πρωτεϊνικών αλληλεπιδράσεων όμως, ιδίως αυτά που βασίζονται σε αυστηρά κριτήρια περιέχουν συχνά έναν “σκληρό πυρήνα” γονιδίων και μπορούν έτσι να δώσουν χρήσιμες πληροφορίες για ένα σύστημα.

Λόγω του γεγονότος ότι η διαμόρφωση των δικτύων βασίζεται σε μεγάλο αριθμό πολύπλοκων πειραμάτων πρωτεϊνικών αλληλεπιδράσεων, η συντριπτική πλειοψηφία των γνωστών δικτύων αφορούν ανθρώπινους κυτταρικούς τύπους. Πιο χαρακτηριστικές βάσεις δεδομένων

τέτοιων δικτύων είναι η Human Protein Reference Database HPRD (http://www.hprd.org/) (Keshava Prasad et al., 2009) και η STRINGdb (http://string-db.org/) (Szklarczyk et al., 2011) που περιέχει σημαντικό αριθμό δικτύων και για άλλους οργανισμούς. Τα δίκτυα πρωτεϊνικών αλληλεπιδράσεων, αποτελούν ένα μόνο μέρος του πολύ ενεργού πεδίου της μελέτης των βιολογικών δικτύων, το οποίο αποτελεί και το αντικείμενο του επόμενου κεφαλαίου (βλ. Κεφάλαιο 9).

Κατηγοριοποιήσεις από πειράματα γονιδιωματικής σε μεγάλη κλίμακα

Η τεχνολογική προόδος στο πεδίο της γονιδιωματικής μέσω των μεθόδων αλληλούχισης νέας γενιάς έχει οδηγήσει σε μια μικρή επανάσταση στο χώρο της μελέτης των γονιδιωμάτων. Ένας πολύ μεγάλος αριθμός από γονιδιωματικές ιδιότητες όπως η ρύθμιση της μεταγραφής, οι επιγενετικές τροποποιήσεις και η δομή της χρωματίνης μπορεί πλέον να ποσοτικοποιηθεί σε γονιδιωματική κλίμακα μέσω της σύνδεσης συγκεκριμένων πειραματικών πρωτοκόλλων με τεχνικές αλληλούχισης (βλ. Κεφάλαιο 11). Πειράματα αυτού του τύπου πάραγουν έναν εξαιρετικά μεγάλο όγκο δεδομένων ο οποίος σε κάποιες περιπτώσεις μπορεί να τυποποιηθεί και να οργανωθεί με τρόπο που να οδηγεί σε χρήσιμες κατηγοριοποιήσεις. Έτσι, με βάση πειράματα μεγάλης κλίμακας μπορούμε να έχουμε γονίδια που ομαδοποιούνται με βάση:

1. Την κοινή μεταγραφική τους ρύθμιση μέσω της πρόσδεσης στους υποκινητές τους του ίδιου μεταγραφικού παράγοντα. Οι ομαδοποιήσεις αυτού του τύπου αντιστοιχούν μεταγραφικούς παράγοντες με καταλόγους γονιδίων στόχων σε συγκεκριμένους κυτταρικούς τύπους (Essaghir et al., 2010).
2. Την κοινή ρύθμιση από μη-κωδικά RNA. Παρομοίως με τους μεταγραφικούς παράγοντες γίνεται αντιστοίχηση μη-κωδικών RNA και γονιδίων στόχων με τρόπο ειδικό για κάθε κυτταρικό τύπο (Vlachos et al., 2014).
3. Την κοινή ή παρόμοια επιγενετική σήμανσή τους από μεθυλιωμένο DNA, μετα-μεταφραστικά τροποποιημένα μόρια ιστονών. Γονίδια που μοιράζονται συγκεκριμένα πρότυπα κατάληψης από επιγενετικά στοιχεία (π.χ. μεθυλίωση στο DNA του υποκινητή ή σήμανση από συγκεκριμένες μετα-μεταφραστικές τροποποιήσεις ιστονών) οργανώνονται σε κοινούς καταλόγους.

Ανάλογα με την προέλευση των δεδομένων, υπάρχουν διάφορες βάσεις που οργανώνουν αυτήν την πληροφορία. Σημείο αναφοράς γι' αυτού του είδους τις μελέτες είναι το ENCODE Consortium το οποίο αποτέλεσε τη μεγαλύτερη και πιο φιλόδοξη προσπάθεια καταγραφής και οργάνωσης γονιδιωματικής πληροφορίας για το ανθρώπινο γονιδίωμα (https://www.encodeproject.org/) (Dunham et al., 2012) ενώ πρόσφατα επέκτεινε τις μελέτες αυτές και στο γονιδίωμα του ποντικού (http://www.mouseencode.org/) (Consortium et al., 2014). Η ενσωμάτωση πολλών ειδών δεδομένων που προέκυψαν (και προκύπτουν) από μεγάλης κλίμακας πειράματα γίνεται με αργούς ρυθμούς λόγω της εγγενούς ανομοιομορφίας τους (κυτταρικές σειρές, πειραματικά πρωτόκολλα) αλλά και της δυσκολίας που εμπεριέχεται στην πρωτογενή τους ανάλυση. Στο Κεφάλαιο 11 θα συζητήσουμε πιο αναλυτικά τόσο τις πειραματικές αλλά κυρίως τις υπολογιστικές πτυχές της γονιδιωματικής μεγάλης κλίμακας.

Ομαδοποιήσεις από ανώτερου τύπου λειτουργικές σχέσεις

Ομαδοποιήσεις που σχετίζονται με ανώτερου τύπου λειτουργικές σχέσεις περιλαμβάνουν τις συσχετίσεις που προκύπτουν μεταξύ γονιδίων και παθολογικών καταστάσεων ή την επιλεκτική έκφρασή τους σε συγκεκριμένους κυτταρικούς τύπους. Τέτοιου τύπου ομαδοποιήσεις είναι αποτέλεσμα της ανάλυσης ενός μεγάλου αριθμού σύνθετων πειραμάτων που δεν αφορούν μόνο τη λειτουργία των γονιδίων (αν π.χ. το Α κωδικοποιεί μια φωσφατάση ή έναν μεταγραφικό παράγοντα) αλλά πιο σύνθετες ιδιότητες όπως π.χ. αν το Α έχει βρεθεί να συσχετίζεται με την εμφάνιση ή όχι της ασθένειας Χ στον άνθρωπο ή αν το Α εκφράζεται κυρίως στο πάγκρεας αλλά όχι στον εγκέφαλο.

Οι κατηγοριοποιήσεις με βάση τη συσχέτιση με παθολογικές καταστάσεις βασίζονται συχνά σε μεγάλης κλίμακας μελέτες που αποσκοπούν στον προσδιορισμό της γενετικής βάσης ασθενειών, την ύπαρξη δηλαδή ή όχι μεταλλαγών σε συγκεκριμένα γονίδια μεταξύ ασθενών και υγιών ατόμων ή ακόμα και τη διαφορετική έκφραση γονιδίων μεταξύ συγκεκριμένων ομάδων ατόμων. Τέτοιες μελέτες είναι οι Μελέτες Συσχέτισης σε Γονιδιωματική Κλίμακα (Genome Wide Association Studies, GWAS) για τις οποίες θα συζητήσουμε αναλυτικά στο Κεφάλαιο 10. Οι πιο σημαντικές βάσεις δεδομένων που περιέχουν ομαδοποιήσεις γονιδίων σε σχέση με ασθένειες είναι η OMIM (http://www.omim.org/) (Hamosh, Scott, Amberger, Bocchini & McKusick, 2005) και η GWAS-central (http://www.gwascentral.org/) (Beck, Hastings, Gollapudi, Free & Brookes, 2014).

Ομαδοποιήσεις με βάση την επιλεκτική έκφραση σε διαφορετικούς κυτταρικούς τύπους υπάρχουν για καλά μελετημένους οργανισμούς όπως ο άνθρωπος και το ποντίκι. Εκτός από την πληροφορία αυτού του τύπου, που μπορεί κανείς να βρει σε βάσεις δεδομένων μεγάλων προγραμμάτων όπως του ENCODE (βλ. παραπάνω) υπάρχουν στοιχεία στις βάσεις δεδομένων των γονιδιωματικών ατλάντων (Gene Atlas) (Petryszak et al., 2014) για τον άνθρωπο (http://www.proteinatlas.org/) και για το ποντίκι (http://www.emouseatlas.org/emap/home.html).

Ποσοτικοποίηση της λειτουργικής Ανάλυσης. Η έννοια του εμπλουτισμού

Προσπαθώντας να απαντήσουμε στο πρώτο από τα βασικά ερωτήματα που είχαμε θέσει εξαρχής (“Πόσα γονίδια αντιστοιχούν σε μια συγκεκριμένη βιολογική λειτουργία;”) ολοκληρώσαμε μια επισκόπηση των πηγών της προϋπάρχουσας γνώσης. Θα περάσουμε τώρα στο επόμενο ερώτημα που μπορεί να διατυπωθεί καλύτερα ως εξής:

Δεδομένου ότι μεταξύ των Ν γονιδίων που βρίσκουμε να είναι διαφορικά εκφραζόμενα σε ένα πείραμα, Χ σχετίζονται με την γονιδιακή λειτουργία Υ, είναι αυτός ο αριθμός Χ μεγάλος ή μικρός και πώς θα μπορούσαμε να αξιολογήσουμε κάτι τέτοιο;

Μια άλλη διατύπωση της παραπάνω ερώτησης θα ήταν:

Πότε μπορούμε να πούμε ότι μέσα σε ένα σύνολο γονιδίων, τα γονίδια που σχετίζονται με μια λειτουργία είναι περισσότερα/λιγότερα ή κοντά σε αυτό που θα περιμέναμε αν τα διαλέγαμε στην τύχη;

Διατυπωμένη μ' αυτόν τον τρόπο, η ερώτηση μας καθοδηγεί στο να προσδιορίσουμε τη βασική έννοια για την ποσοτικοποίηση του προβλήματος. Αυτή είναι η έννοια του εμπλουτισμού που

με διαφορετική μορφή έχουμε ήδη συναντήσει στα Κεφάλαια 1 και 2 (όταν συζητούσαμε για τις σχετικές συχνότητες εμφάνισης ολιγονουκλεοτιδίων) και που αποτελεί έναν τρόπο αξιολόγησης αποκλίσεων από την τυχαιότητα. Πράγματι, το βασικό στοιχείο στη λειτουργική ανάλυση είναι να εκτιμήσουμε το βαθμό, στον οποίο ένα σύνολο γονιδίων μπορεί να έχει προκύψει από μια τυχαία δειγματοληψία ή όχι, επειδή όμως η δειγματοληψία μπορεί να γίνει από πολλές διαφορετικές λειτουργίες η ανάλυση εμπλουτισμού έχει συγκεκριμένα χαρακτηριστικά. Στη βιβλιογραφία αλλά και στο διαδίκτυο υπάρχει ένας μεγάλος αριθμός υπολογιστικών εργαλείων για τη λειτουργική ανάλυση της γονιδιακής έκφρασης και οι προσεγγίσεις τους διαφέρουν λιγότερο ή περισσότερο στον τρόπο με τον οποίο αξιολογούν το βαθμό εμπλουτισμού του δείγματος γονιδίων σε συγκεκριμένες λειτουργίες. Στη συνέχεια θα περιγράψουμε αναλυτικά την τόσο βασική έννοια του εμπλουτισμού σ' ένα μη-βιολογικό παράδειγμα πριν περάσουμε σε εφαρμογές από το χώρο της ανάλυσης της γονιδιακής έκφρασης.

Ένα μη-βιολογικό πρόβλημα

Το 1943 ο Herman Hesse δημοσίευσε το τελευταίο του μυθιστόρημα με τίτλο “Το παιχνίδι με τις χάντρες” (Das Glasperlenspiel), στο οποίο περιγράφει μια φανταστική, ουτοπική κοινωνία διανοουμένων που σκοπό της ζωής τους έχουν να τελειοποιήσουν την τεχνική τους στο ομώνυμο, πολύπλοκο παιχνίδι, κάτι που απαιτεί χρόνια παράλληλης μελέτης μουσικής, μαθηματικών και ιστορίας. Χωρίς να είναι τόσο πολύπλοκη, η λειτουργική ανάλυση μπορεί να τελειοποιηθεί αν προσπαθήσουμε να φανταστούμε ένα αρκετά απλούστερο παιχνίδι με χάντρες σαν κι αυτό που θα περιγράψουμε στη συνέχεια.

Φανταστείτε ένα κουτί σαν αυτό που φαίνεται στην Εικόνα 8.3 και που περιέχει χάντρες ίδιου μεγέθους αλλά διαφορετικών χρωμάτων. Συνολικά περιέχονται 100 χάντρες από τις οποίες επιλέγεται κάθε φορά ένας συγκεκριμένος αριθμός από το διαιτητή του παιχνιδιού. Ο παίχτης καλείται να απαντήσει στο εξής ερώτημα:

Έχει διαλέξει ο διαιτητής στην τύχη τις χάντρες ή όχι;

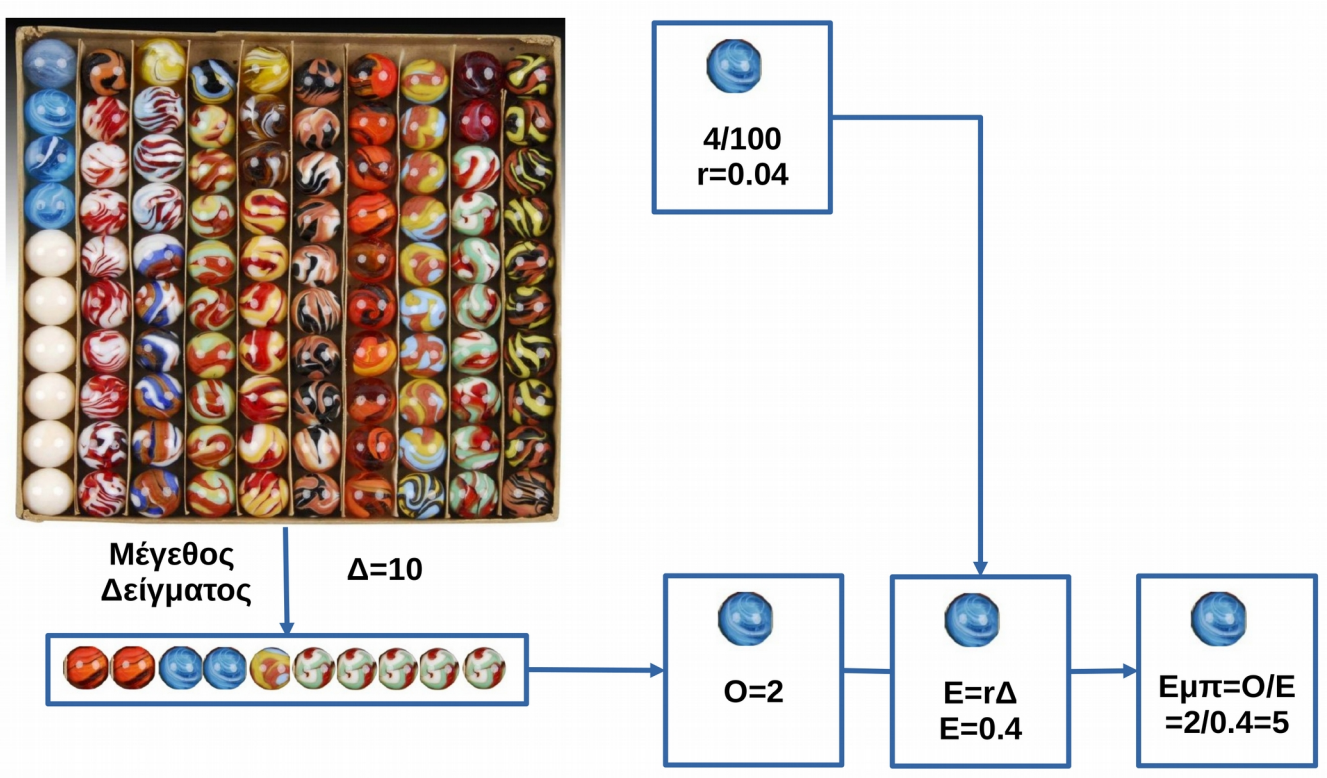
Η στρατηγική για το απλό αυτό “Παιχνίδι με τις χάντρες” θα βασιστεί στην ανάλυση της έννοιας του εμπλουτισμού. Φανταστείτε ότι ο διαιτητής σας παρουσιάζει την επιλογή του από δέκα χάντρες που φαίνονται πάνω από το κουτί στην Εικόνα 8.3 μεταξύ των οποίων, διάφοροι χρωματικοί συνδυασμοί εκπροσωπούνται σε διαφορετικό αριθμό. Ο παίχτης πρακτικά καλείται ν' αποφασίσει αν αυτή η επιλογή θα μπορούσε να έχει γίνει στην τύχη και ο τρόπος για να κάνει κάτι τέτοιο είναι να εκτιμήσει ποιοι θα ήταν οι σχετικοί αριθμοί για κάθε είδος χάντρας του δείγματος αν ο διαιτητής πραγματικά διάλεγε στην τύχη. Να υπολογίσει δηλαδή τις αναμενόμενες (βάση τυχαιότητας) τιμές εκπροσώπησης για κάθε χρώμα χάντρας πριν περάσει στην εκτίμηση αν αυτό υπέρ- ή ύπο-εκπροσωπείται. Η στρατηγική του μπορεί να συνοψιστεί στα εξής βήματα:

- Εντόπισε τα είδη που περιέχονται στο δείγμα.
- Υπολόγισε ποια θα ήταν η αναμενόμενη τιμή για κάθε είδος βάση τυχαιότητας.
- Εκτίμησε το βαθμό που η αναμενόμενη αυτή τιμή διαφέρει από την πραγματική.

Για το παράδειγμα της Εικόνας 8.3 τα τρία αυτά βήματα οδηγούν στους εξής υπολογισμούς:

Στο δείγμα υπάρχουν τέσσερα διαφορετικά είδη χαντρών: δύο πορτοκαλί, δύο γαλάζιες, μία

κιτρινωπή και πέντε κοκκινόλευκες.



Εικόνα 8.3: *Σχηματική αναπαράσταση του υπολογισμού του εμπλουτισμού. Το κουτί στα αριστερά περιέχει 100 χάντρες από τις οποίες οι 4 είναι γαλάζιες. Η πιθανότητα έτσι να διαλέξει κανείς μία γαλάζια είναι 0.04 και σε ένα δείγμα από 10 χάντρες η αναμενόμενη τιμή είναι 0.04*10=0.4 γαλάζιες χάντρες. Η σχέση της παρατηρούμενης (O=2) με την αναμενόμενη αυτή τιμή μας δίνει την τιμή εμπλουτισμού.*

Για να υπολογίσουμε την αναμενόμενη τιμή για καθένα απ' αυτά τα είδη θα πρέπει να σκεφτούμε με τον ακόλουθο τρόπο. Αν το κουτί περιέχει συνολικά Χ χάντρες από το είδος Α τότε η πιθανότητα *p(A)* να τραβήξουμε μία χάντρα Α, εφόσον το κουτί περιέχει συνολικά 100 χάντρες, είναι ίση με *p(A)=X/100*. Η αναμενόμενη τιμή σε ένα δείγμα Ν χαντρών προκύπτει από τον πολλαπλασιασμό του πλήθους των χαντρών του δείγματος επί τη σχετική πιθανότητα του κάθε είδους και είναι τότε ίση με *E(A)=Np(A)*. Για παράδειγμα, στο κουτί υπάρχουν συνολικά 10 πορτοκαλί χάντρες και το μέγεθος τους δείγματος είναι (N=10) οπότε η αναμενόμενη τιμή είναι Ε(πορτοκαλί)=10*(10/100)=1. Με παρόμοιο τρόπο υπολογίζουμε για τις άλλες χάντρες ότι Ε(γαλάζια)=0.4, Ε(κιτρινωπή)=0.9, Ε(κοκκινόλευκη)=0.8.

Στο δεύτερο στάδιο, έχοντας υπολογίσει τις αναμενόμενες τιμές θα πρέπει να συγκρίνουμε τις τιμές αυτές με τις παρατηρούμενες. Ο καλύτερος τρόπος είναι μέσω του λόγου παρατηρούμενης προς αναμενόμενη τιμή, *r(A)=P(A)/E(A)*. Ο λόγος αυτός αντιστοιχεί στο βαθμό που το δείγμα περιέχει περισσότερες ή λιγότερες χάντρες από τις αναμενόμενες και για το λόγο αυτό ονομάζεται λόγος εμπλουτισμού (enrichment rate) ή απλά εμπλουτισμός (enrichment). Τιμές r>1 αντιστοιχούν σε θετικό εμπλουτισμό δηλαδή παρατηρούμενες τιμές μεγαλύτερες από τις αναμενόμενες. Το αντίθετο ισχύει για r<1 ενώ για r=1 λέμε ότι δεν υπάρχει εμπλουτισμός καθώς η παρατηρούμενη τιμή είναι

ίση με την αναμενόμενη. Τιμές που αποκλίνουν πολύ από τη μονάδα είναι έτσι ενδεικτικές ότι ο διαιτητής δεν έχει επιλέξει στην τύχη αλλά με συγκεκριμένες προτιμήσεις για τη μία ή την άλλη κατηγορία[1]. Οι τιμές που προκύπτουν είναι r(πορτοκαλί)=2/1=2, r(γαλάζια)=2/0.4=5, r(κιτρινωπή)=1/0.9=1.1 και r(κοκκινόλευκη)=5/0.8=6.25. Παρατηρούμε ότι κάποιοι από τους λόγους αυτούς (για τις γαλάζιες και τις κοκκινόλευκες) είναι πολύ μεγαλύτεροι από την μονάδα, για τις πορτοκαλι η τιμή είναι διπλάσια από την μονάδα ενώ για τις κιτρινωπές η τιμή υποδεικνύει μάλλον απουσία εμπλουτισμού (r~1).

Τι σημαίνει αυτό για τον επίδοξο νικητή του παιχνιδιού μας; Πόσο βέβαιος μπορεί να είναι για την απάντηση που θα δώσει; Το γεγονός ότι, τουλάχιστο για κάποιες από τις κατηγορίες χαντρών, η παρατηρούμενη τιμή διαφέρει πολύ από την αναμενόμενη είναι αρκετό για να τον πείσει να πει πως η επιλογή δεν ήταν τυχαία; Πριν το κάνει θα πρέπει να σκεφτεί πώς θα απαντήσει σε δύο βασικά ερωτήματα:

- Πώς θα αξιολογήσει στατιστικά την τιμή ενός εμπλουτισμού; Πόσο μεγάλη τιμή είναι η r=4 και γιατί να μην λάβει υπόψη του μια τιμή r=2; κι έπειτα
- Πόσο πιθανό είναι να εντοπίσει μια κατηγορία με σημαντική απόκλιση από την τιμή r=1 κι αυτό να είναι τυχαίο;

Θα πρέπει να έχετε ήδη μια καλή ιδέα για το πώς θα απαντήσουμε τόσο στην πρώτη αλλά κυρίως στη δεύτερη ερώτηση. Όμως πριν περάσουμε στις απαντήσεις τους ας δούμε πώς το παιχνίδι με τις χάντρες μεταφράζεται σε ένα αμιγώς βιολογικό πρόβλημα.

***Ερώτηση**: Στο απλουστευμένο παράδειγμα που μόλις είδαμε εξετάσαμε μόνο τους συνδυασμούς των χαντρών που βρίσκονται μέσα στο δείγμα του διαιτητή. Τι συμβαίνει με αυτούς που δεν υπήρχαν στο δείγμα; Τι μπορούμε να πούμε γι' αυτούς και πώς αυτό θα μας βοηθούσε στο ν' απαντήσουμε στο ερώτημα αν ο διαιτητής διαλέγει στην τύχη;*

Επιστροφή στο βιολογικό πρόβλημα

Με ποιον τρόπο προσομοιάζει το "Παιχνίδι με τις Χάντρες" το πρόβλημα της λειτουργικής ανάλυσης της γονιδιακής έκφρασης; Οι αναλογίες είναι μάλλον προφανείς. Το κουτί με τις χάντρες αντιστοιχεί στο σύνολο των γονιδίων, των οποίων έχουμε μετρήσει την έκφραση. Τα γονίδια δεν έχουν όλα τις ίδιες ιδιότητες, οι οποίες αντιστοιχούν στα διάφορα χρώματα που έχουν οι χάντρες. Το σύνολο των χαντρών που μας παρουσιάζει ο διαιτητής αντιστοιχεί στα γονίδια εκείνα που προκύπτει πως είναι διαφορικά εκφραζόμενα και για τα οποία θέλουμε να διενεργήσουμε την ανάλυση εμπλουτισμού. Η ερώτηση που τίθεται στον "παίχτη" αντιστοιχεί στην ερώτηση που θέτει ο ερευνητής στον εαυτό του:

Πότε μπορούμε να πούμε ότι μέσα σε ένα σύνολο γονιδίων, τα γονίδια που σχετίζονται με μια λειτουργία είναι περισσότερα/λιγότερα ή κοντά σε αυτό που θα περιμέναμε αν τα διαλέγαμε στην τύχη;

Η απάντηση είναι εύκολη και απλώς περιλαμβάνει έναν μεγάλο αριθμό υπολογισμών τιμών εμπλουτισμού για όλες τις λειτουργικές κατηγορίες που θέλουμε να εξετάσουμε. Έτσι για ένα

[1] *Εναλλακτικά ο "διαιτητής" μπορεί σκόπιμα να αποφεύγει να διαλέξει από μία κατηγορία (κάτι που συμβαίνει στο δείγμα μας). Αυτό θα αντανακλάται σε τιμές r που είναι σημαντικά μικρότερες από 1.*

πείραμα στο οποίο μετρήθηκε η έκφραση *N* γονιδίων από τα οποία *n* βρέθηκαν να είναι διαφορικά εκφραζόμενα, ο εμπλουτισμός της λειτουργικής κατηγορίας Α δίνεται από την τιμή:

$$r(A)=\frac{\frac{k}{K}}{\frac{n}{N}}=\frac{kN}{nK} \quad 8.1$$

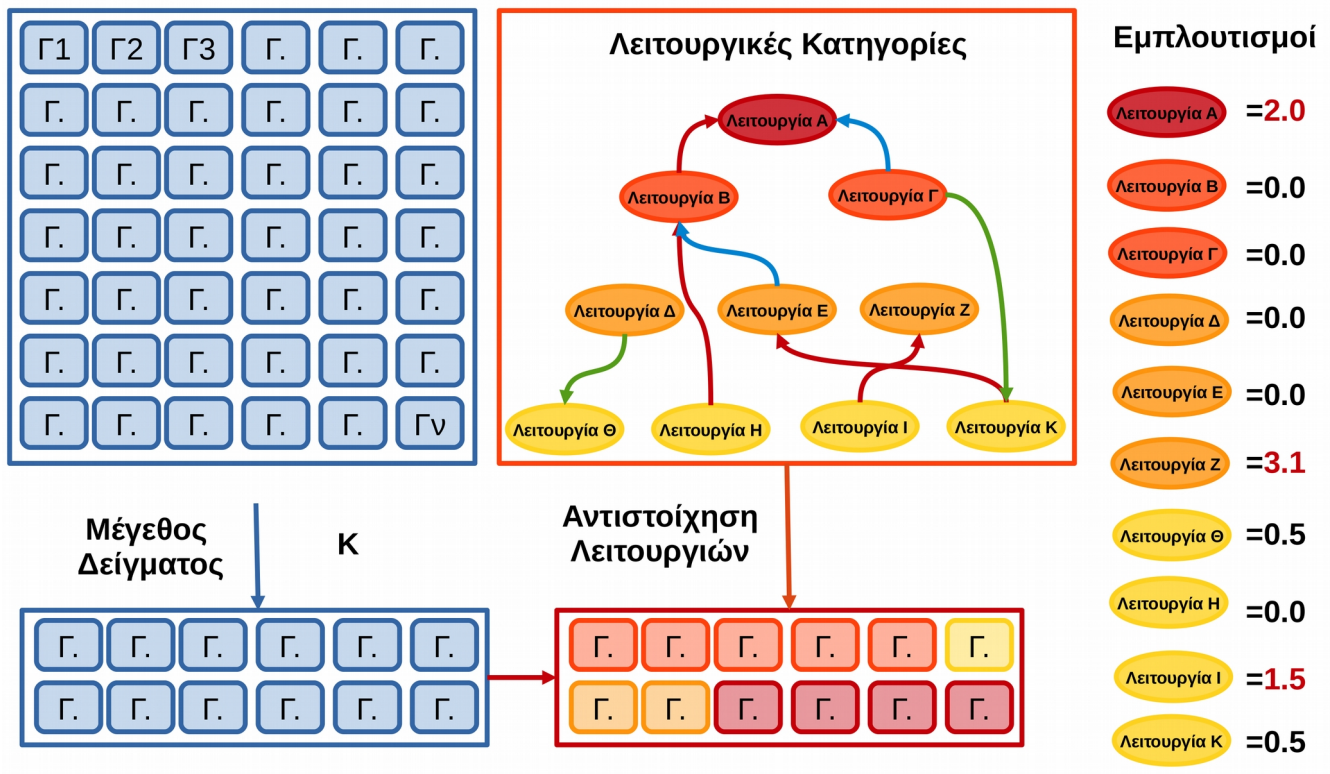
όπου *K* είναι το σύνολο των *N* γονιδίων που ανήκουν στην κατηγορία Α και *k* είναι ο αριθμός των γονιδίων της κατηγορίας αυτής που βρίσκονται εντός του δείγματος *n* *(k<K, k<n)*. Αναλογιζόμενοι ότι ο αριθμός των γονιδίων της Α που θα μπορούσαμε να περιμένουμε σε δείγμα μεγέθους *n* είναι ίσος με:

$$E(A)=n\frac{K}{N} \quad 8.2$$

προκύπτει ότι:

$$r(A)=\frac{k}{E(A)} \quad 8.3$$

δηλαδή ο εμπλουτισμός είναι ίσος με το λόγο της παρατηρούμενης προς την αναμενόμενη τιμή όπως τον ορίσαμε και παραπάνω.



Εικόνα 8.4: *Η ανάλυση εμπλουτισμού σ' ένα δείγμα γονιδίων με βάση την κατηγοριοποίησή τους σε λειτουργικές κατηγορίες. Από το αρχικό σύνολο Ν γονιδίων επιλέγουμε ένα δείγμα Κ γονιδίων, τα οποία και αντιστοιχούμε στις λειτουργίες που προκύπτουν από το σχήμα της Εικόνας. Με βάση τις σχετικές τιμές αναμενόμενων και παρατηρούμενων τιμών για κάθε λειτουργία προκύπτει μια αντίστοιχη τιμή εμπλουτισμού.*

Στην Εικόνα 8.4 παρουσιάζεται σχηματικά η διαδικασία για ένα μικρό αριθμό κατηγοριών γονιδίων. Θα πρέπει όμως να σημειωθεί πως το πλήθος των κατηγοριών είναι συχνά πολύ μεγάλο (ενδεικτικά το σύνολο των όρων GO για το ανθρώπινο γονιδίωμα είναι ~8000) κάτι που σημαίνει ότι το αποτέλεσμα τέτοιων αναλύσεων είναι συχνά μακρείς κατάλογοι τιμών εμπλουτισμού. Παραμένει τέλος να απαντηθούν τα δύο ερωτήματα που θέσαμε στον παίχτη του παιχνιδιού με τις χάντρες. Πρώτον, πώς θα αξιολογήσει αν μια τιμή εμπλουτισμού είναι σημαντικά μεγάλη/μικρή ή όχι και δεύτερον πώς μπορεί να ξέρει για πόσες κατηγορίες μπορεί να περιμένει σημαντικά μεγάλους/μικρούς εμπλουτισμούς. Για την απάντηση και στα δύο ερωτήματα θα πρέπει να καταφύγουμε στη στατιστική.

Μαθηματικό Ιντερμέδιο Ι. Η υπεργεωμετρική κατανομή

Είδαμε παραπάνω ότι ο εμπλουτισμός που υπολόγισε ο παίχτης στο παράδειγμα της Εικόνας 8.3 για τις πορτοκαλί χάντρες είναι r(πορτοκαλί)=2, μια τιμή που υποδηλώνει διπλάσιες από το αναμενόμενο χάντρες αυτού του χρώματος. Ο παίχτης όμως είναι διστακτικός να θεωρήσει ότι δύο πορτοκαλί χάντρες αντί για μία είναι ασφαλής ένδειξη ότι ο διαιτητής δε διαλέγει στην τύχη. Ανησυχεί ότι η σχέση μεταξύ του αριθμού των πορτοκαλί χαντρών στο κουτί (10) και του μεγέθους του δείγματος (10 από τις 100) είναι τέτοια που δε θα απέκλειε οι 2 πορτοκαλί χάντρες να είναι αποτέλεσμα τυχαίας δειγματοληψίας. Πώς όμως μπορεί να ποσοτικοποιήσει αυτή του τη διαίσθηση, ώστε να έχει ένα αντικειμενικό κριτήριο για να στηρίξει την απόφασή του;

Θα προσπαθήσουμε να αναλύσουμε το πρόβλημά του ποσοτικά ως εξής:

Εξετάζοντας μόνο τις πορτοκαλί χάντρες στο παιχνίδι, γνωρίζουμε ότι 10 από τις χάντρες στο κουτί είναι πορτοκαλί και 90 δεν είναι. Αν ο διαιτητής επιλέξει μία χάντρα, η πιθανότητα αυτή να είναι πορτοκαλί είναι 10/100=0.1. Αν η χάντρα είναι πορτοκαλί, τότε το κουτί περιέχει πλέον 99 χάντρες εκ των οποίων οι 9 είναι πορτοκαλί και οι 90 δεν είναι. Η πιθανότητα επιλογής μίας επόμενης πορτοκαλί χάντρας είναι τώρα 9/99=0.091, έχει δηλαδή αλλάξει από πριν καθώς κάθε χάντρα που επιλέγεται δεν ξανατοποθετείται στο κουτί πριν επιλεχθεί η επόμενη. Στη θεωρία των πιθανοτήτων μια τέτοια διαδικασία ονομάζεται “δειγματοληψία χωρίς επανάθεση”. Ποια είναι λοιπόν η πιθανότητα να επιλέξουμε ακριβώς δύο πορτοκαλί χάντρες μέσα σε ένα δείγμα από δέκα χάντρες; Αν οι δύο πορτοκαλί χάντρες ήταν η πρώτη και η δεύτερη τότε η πιθανότητα θα ήταν:

P(1η και 2η πορτοκαλί στις 10)=P(πορτοκαλι 10:100)P(πορτοκαλί 9:99)P(άλλη 90:98)P(άλλη 89:97)P(άλλη 88:96)....

Η παραπάνω σχέση προκύπτει στη λογική ότι στην αρχή έχουμε 10/100 πορτοκαλί, έπειτα 9/99 πορτοκαλί αλλά στη συνέχεια θέλουμε να επιλέξουμε χάντρες που δεν είναι πορτοκαλί. Αυτές είναι (μετά από 2 πορτοκαλί επιλογές) αρχικά 90 από τις 98, έπειτα 89 από τις 97, 88 από τις 96 κ.ο.κ. Η πιθανότητα που υπολογίζεται έτσι είναι:

P(1η και 2η πορτοκαλί στις 10)=0.0039

Πριν βιαστούμε να απαντήσουμε στον προβληματισμό του παίχτη αναφέροντας μια τόσο μικρή πιθανότητα θα πρέπει να σκεφτούμε κάτι πολύ σημαντικό. Αυτό είναι ότι η διάταξη των χαντρών στο δείγμα δε μας ενδιαφέρει καθόλου. Η παραπάνω πιθανότητα αντιστοιχεί στην πιθανότητα να έχουμε ένα οποιοδήποτε δείγμα στο οποίο ακριβώς δύο πορτοκαλί χάντρες

τραβήχτηκαν πρώτη και δεύτερη. Ωστόσο ο παίχτης δεν ενδιαφέρεται για τη σειρά επιλογής, ούτε μπορεί να τη γνωρίζει, όπως και για το δικό μας πρόβλημα δεν έχει κανένα νόημα να σκεφτόμαστε ποιο γονίδιο είναι πρώτο και δεύτερο στη λίστα με τα διαφορικά εκφραζόμενα γονίδια[2]. Αυτό που χρειαζόμαστε είναι η πιθανότητα δύο πορτοκαλί χάντρες να υπάρχουν στο δείγμα ανεξάρτητα από τη σειρά με την οποία επιλέχθηκαν. Οι πιθανοί συνδυασμοί που δίνουν δύο πορτοκαλί χάντρες σε οποιαδήποτε από την 1η ως τη 10η θέση επιλογής είναι πάρα πολλοί και ως εκ τούτου η πιθανότητα *P(2 πορτοκαλί στις 10)* είναι πολύ μεγαλύτερη από 0.0039. Η ακριβής τιμή της δίνεται από τον τύπο:

$$P(X=k)=\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad 8.4$$

Ο συμβολισμός $\binom{x}{y}$ είναι ο διωνυμικός τελεστής για τα *x, y* και αντιστοιχεί στο πλήθος των δυνατών συνδυασμών *y* αντικειμένων από ένα σύνολο *x* δυνατών επιλογών και που ισούται με:

$$\binom{x}{y}=\frac{x!}{(x-y)!y!} \quad 8.5$$

***Ερώτηση**: Πόσοι είναι οι πιθανοί τρόποι να πάρουμε 2 πορτοκαλί χάντρες σ' ένα δείγμα 10 χαντρών από το κουτί της Εικόνας 8.3; Πως αλλάζει ο αριθμός για δείγμα 15 και 20 χαντρών;*

Στην εξίσωση 8.4, *N* είναι το σύνολο των χαντρών στο κουτί (100), *K* είναι το σύνολο των πορτοκαλί χαντρών στο κουτί (10), *n* είναι το πλήθος των χαντρών που συνολικά επιλέγονται (10) και *k* το πλήθος των πορτοκαλί χαντρών στο δείγμα (2). Οι τιμές αυτές μπορούν να παρασταθούν σε έναν πίνακα όπως ο παρακάτω:

	Στο δείγμα	**Στο κουτί**	**Σύνολο**
Πορτοκαλί	k=2	K-k=8	K=10
Άλλες	n-k=8	N+k-n-K=82	N-K=90
Σύνολο	n=10	N-n=90	N=100

Η εξίσωση 8.4 υπολογίζει την πιθανότητα το k να είναι ακριβώς ίσο με 2 με βάση το γεγονός ότι πρόκειται για μια υπεργεωμετρική τυχαία μεταβλητή. Η υπεργεωμετρική κατανομή είναι η κατανομή που χρησιμοποιούμε για να προσομοιάσουμε τυχαία γεγονότα επιλογής δείγματος αντικειμένων από ένα σύνολο χωρίς επανάθεση, όπως συμβαίνει στο πρόβλημα του "Παιχνιδιού με τις Χάντρες" (και της λειτουργικής ανάλυσης γονιδίων). Αντικαθιστώντας στην εξίσωση 8.4 τις τιμές για τις πορτοκαλί/μη-πορτοκαλί χάντρες έχουμε:

2 *Η διάταξη των διαφορικά εκφραζόμενων γονιδίων χρησιμοποιείται μόνο σε κάποιες περιπτώσεις σε κατάταξη από το πιο ενεργοποιημένο στο πιο κατεσταλμένο, όπως στην Ανάλυση Εμπλουτισμού σε Σύνολα Γονιδίων (GSEA), που θα συζητήσουμε παρακάτω.*

$$P(X=2)=\frac{\binom{10}{2}\binom{90}{8}}{\binom{100}{10}}=0.2$$

Η πιθανότητα δηλαδή *P(2 πορτοκαλί στις 10)* είναι περίπου 1/5. Όχι ιδιαίτερα μικρή σε κάθε περίπτωση. Ο παίχτης θα πρέπει να είναι ιδιαίτερα επιφυλακτικός πριν αποφασίσει ότι ο εμπλουτισμός r(πορτοκαλί)=2 είναι σημαντικός. Για την ακρίβεια μπορεί να εφαρμόσει κατάλληλα την υπεργεωμετρική κατανομή ώστε να υπολογίσει μια ακόμα ενδιαφέρουσα πιθανότητα που μπορεί να χρησιμοποιήσει ως αντικειμενικό, ποσοτικό κριτήριο για την απόφασή του. Μπορεί να αναρωτηθεί ποια είναι η πιθανότητα το δείγμα του διαιτητή να περιέχει 2 ή περισσότερες πορτοκαλί χάντρες δηλαδή *P(τουλάχιστον 2 πορτοκαλί στις 10)*. Η πιθανότητα αυτή αντιστοιχεί στο άθροισμα των πιθανοτήτων:

$$P(X\geq 2)=\sum_{i=2}^{i=K} P(X=i) \qquad \text{με } (K=10) \quad 8.6$$

και γι' αυτό ονομάζεται αθροιστική (cumulative) πιθανότητα. Η λογική πίσω απ' αυτόν τον υπολογισμό είναι ότι ο παίχτης θέλει να αποφασίσει αν οι δύο πορτοκαλί χάντρες ξεπερνούν ένα όριο πέρα από το οποίο μπορεί να πει με ασφάλεια ότι η επιλογή του διαιτητή είναι συνειδητή υπέρ των πορτοκαλί χαντρών. Η εξίσωση 8.6 δίνει ότι *P(τουλάχιστον 2 πορτοκαλί στις 10)=0.26*, υπάρχει δηλαδή περισσότερο από 25% πιθανότητα μια τέτοια τιμή να προκύψει τυχαία. Μια τέτοια τιμή δεν επιτρέπει στον παίχτη μας να κρίνει το διαιτητή προκατειλημένο κι έτσι θα πρέπει μάλλον να στραφεί σε μια άλλη κατηγορία χαντρών για να εξετάσει την προτίμηση του διαιτητή. Με βάση τα παραπάνω, για τις τέσσερις κατηγορίες χαντρών ο παίχτης του "Παιχνιδιού με τις Χάντρες" μπορεί να υπολογίσει ότι P(πορτοκαλί>=2)=0.26, P(γαλάζια>=2)=0.048, P(κιτρινωπή>=1)=0.62 και P(κοκκινόλευκη>=5)=0.0007 και να καταλήξει έτσι ότι για δύο από τις τέσσερις κατηγορίες, για τις οποίες ισχύει p<0.05 ο εμπλουτισμός είναι στατιστικά σημαντικός. Η αθροιστική πιθανότητα της υπεργεωμετρικής κατανομής μπορεί να λειτουργήσει ως εκτιμητής της τυχαιότητας μιας διαδικασίας δειγματοληψίας χωρίς επανάθεση. Η συγκεκριμένη τιμή χρησιμοποιείται ως τιμή p-value, που όπως έχουμε δει και σε προηγούμενα κεφάλαια μπορεί να χρησιμεύσει στον έλεγχο υποθέσεων.

Αν αφήσουμε για λίγο τον παίχτη και στραφούμε στο δικό μας βιολογικό πρόβλημα μπορούμε να θυμηθούμε το τρίτο και τελευταίο από τα αρχικά ερωτήματα:

Πώς θα μπορούσαμε να προσδιορίσουμε τις λειτουργίες για τις οποίες οι αριθμοί των διαφορικά εκφραζόμενων γονιδίων είναι σημαντικά μεγάλοι (ή αντίστοιχα σημαντικά μικροί);

Η ερώτηση αυτή είναι πρακτικά ισοδύναμη με το πρόβλημα που μόλις επιλύσαμε. Προκειμένου να εξετάσουμε τον εμπλουτισμό μιας οποιασδήποτε λειτουργικής κατηγορίας, αρκεί να δημιουργήσουμε τον αντίστοιχο πίνακα:

	Διαφορικά Εκφραζόμενα	**Μη-διαφορικά Εκφραζόμενα**	**Σύνολο Γονιδίων**
Σχετικά με τη Λειτουργία Α	*K*	*K-k*	*K*
Μη-σχετικά με	*n-k*	*N+k-n-K*	*N-K*

τη λειτουργία			
Σύνολο	*n*	*N-n*	*N*

και να υπολογίσουμε ένα p-value μέσω τις πιθανότητας:

$$P(X \geq k) = \sum_{i=k}^{i=K} P(X=i) \quad 8.7$$

Με αυτό το κριτήριο έχουμε στα χέρια μας το βασικό έλεγχο που απαιτείται για την ανάλυση της στατιστικής σημασίας θετικών εμπλουτισμών. Τι συμβαίνει στην περίπτωση που εξετάζουμε έναν αρνητικό εμπλουτισμό, όταν δηλαδή η παρατηρούμενη τιμή είναι αρκετά μικρότερη από την αναμενόμενη; Η διαδικασία είναι συμμετρικά αντίθετη, αρκεί δηλαδή τώρα να υπολογίσουμε την αθροιστική πιθανότητα για:

$$P(X \leq k) = \sum_{i=0}^{i=k} P(X=i) \quad 8.8$$

Με τη χρήση της υπεργεωμετρικής κατανομής έχουμε απαντήσει και στο τελευταίο ερώτημα που θέσαμε εξαρχής. Για κάθε λειτουργία που εξετάζουμε μπορούμε να υπολογίσουμε τόσο το βαθμό στον οποίο αυτή είναι εμπλουτισμένη σε γονίδια που είναι διαφορικά εκφραζόμενα, όσο και το κατά πόσο αυτός ο εμπλουτισμός είναι στατιστικά σημαντικός. Για συγκεκριμένες λειτουργικές αναλύσεις αυτό θα σημαίνει χιλιάδες από p-values που θα πρέπει να αναλυθούν περαιτέρω.

***Ερώτηση**: Πόσο ασφαλείς είναι οι τιμές p-value όταν εξετάζονται για χιλιάδες διαφορετικές λειτουργίες;*

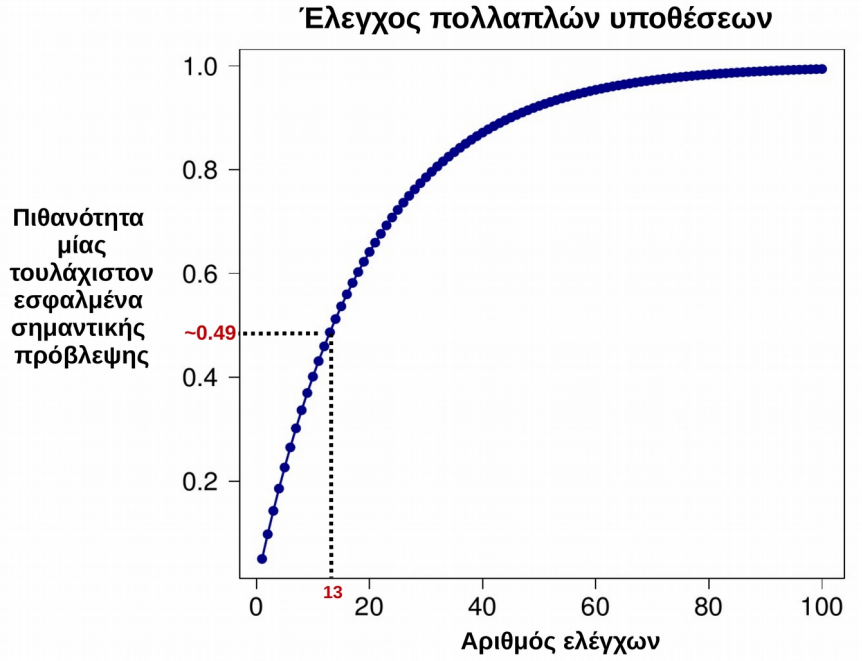
Στη συνέχεια θα συζητήσουμε το λεπτό πρόβλημα που προκύπτει από την παραπάνω ερώτηση και το οποίο θίξαμε κάπως επιφανειακά στο προηγούμενο κεφάλαιο. Πρόκειται για ένα πρόβλημα που δεν μπορούμε να συνεχίσουμε να "αποφεύγουμε". Το πρόβλημα των πολλαπλών υποθέσεων.

Μαθηματικό Ιντερμέδιο ΙΙ. Έλεγχος πολλαπλών υποθέσεων

Στο σημείο αυτό θα επανέλθουμε σ' ένα ερώτημα που έχουμε αφήσει αναπάντητο. Θυμηθείτε τον παίχτη του "Παιχνιδιού με τις Χάντρες" ο οποίος έχει υπολογίσει τις τιμές εμπλουτισμού για τις τέσσερις κατηγορίες χαντρών που περιέχονται στο δείγμα που του παρουσιάζει ο διαιτητής. Όπως είδαμε παραπάνω για δύο από τις τέσσερις κατηγορίες προέκυψε ότι ο εμπλουτισμός είναι στατιστικά σημαντικά μεγάλος, ωστόσο ένας έμπειρος παίχτης, κάποιος που έχει παίξει αρκετές φορές κι έχει αναπτύξει και μια στατιστική διαίσθηση αναλογίζεται σε ποιο βαθμό μία ή δύο κατηγορίες θα μπορούσαν να φαίνονται σημαντικά εμπλουτισμένες από καθαρή τύχη.

Η ανησυχία του παίχτη έχει στατιστική βάση και για το λόγο αυτό αξίζει να κάνουμε μια ακόμα μαθηματική παρέκβαση για να συζητήσουμε ένα πρόβλημα που προκύπτει κατά τον υπολογισμό πολλών τιμών p-value. Ξεχάστε το "Παιχνίδι με τις Χάντρες" και φανταστείτε το εξής

υποθετικό πρόβλημα. Κάποιος σας ζητά να ελέγξετε αν ένα νόμισμα είναι κάλπικο, επιτρέποντάς σας να κάνετε μόνο ρίψεις "κορώνα ή γράμματα". Εσείς αποφασίζετε να κάνετε έναν μεγάλο αριθμό ρίψεων και να βασίσετε την απάντησή σας στο κατά πόσο ο λόγος των αποτελεσμάτων διαφέρει από το 1:1. Αφού πραγματοποίησετε 100 ρίψεις προκύπτει ότι τα αποτελέσματα ήταν 55 κορώνες και 45 γράμματα. Με έναν λόγο ~1.2 κλίνετε μάλλον προς το να απαντήσετε ότι το νόμισμα είναι γνήσιο αλλά προκειμένου να σιγουρευτείτε αποφασίζετε να επαναλάβετε το πείραμα των 100 ρίψεων όσες περισσότερες φορές μπορείτε. Καθώς έχετε την τύχη να διδάσκετε σ' ένα τμήμα με 100 φοιτητές, αποφασίζετε να καταχραστείτε (για λίγο μόνο) την εξουσία που σας δίνει η θέση σας και ζητάτε από τον καθένα να ρίξει το νόμισμα 100 φορές κι έπειτα να σας αναφέρει το λόγο των αποτελεσμάτων. Όταν η διαδικασία ολοκληρώνεται και παίρνετε στα χέρια σας τα αποτελέσματα των ρίψεων των φοιτητών παρατηρείτε με έκπληξη ότι πέντε (5) απ' αυτούς αναφέρουν λόγους που είναι μεγαλύτεροι από 4:1. Για πέντε δηλαδή από τους φοιτητές σας το νόμισμα θα είναι κατά πάσα πιθανότητα κάλπικο. Τι συμβαίνει στην πραγματικότητα;



Εικόνα 8.5: *Γραφική παράσταση του φαινομένου του Ελέγχου Πολλαπλών υποθέσεων. Η πιθανότητας μίας τουλάχιστον εσφαλμένης απόδοσης σημαντικών προβλέψεων αυξάνεται εκθετικά με τον αριθμό των ελέγχων που πραγματοποιεί κανείς. Για 100 ελέγχους η πιθανότητα είναι >0.99.*

Αυτό που έχει συμβεί είναι ότι έχετε πέσει στην "παγίδα" του ελέγχου πολλαπλών υποθέσεων. Προσπαθώντας να ελέγξετε μια υπόθεση πολλές φορές, καταλήγετε μοιραία να την επιβεβαιώσετε σ' ένα μικρό ποσοστό των δοκιμών που κάνετε. Το νόμισμα είναι κατά πάσα πιθανότητα γνήσιο (όπως προκύπτει από τη μεγάλη πλειοψηφία των αποτελεσμάτων των φοιτητών σας) αλλά επειδή ακριβώς ο ίδιος έλεγχος της γνησιότητας έγινε 101 φορές είναι στατιστικά αναπόφευκτο ότι σε κάποιες από αυτές θα προέκυπτε αρνητικός. Για να καταλάβετε καλύτερα αυτήν την "παγίδα" αναλογιστείτε ότι διενεργείτε λειτουργική ανάλυση σε επίπεδο γονιδιακής οντολογίας για 8000 διαφορετικούς όρους (λειτουργίες) σ' ένα ιδιότυπο πείραμα γονιδιακής έκφρασης. Η ιδιοτυπία του έγκειται στο γεγονός ότι έχετε επίτηδες "ανακατέψει" τις τιμές έκφρασης

με τρόπο που τα διαφορικά εκφραζόμενα γονίδια να είναι πρακτικά επιλεγμένα στην τύχη. Κάτω από αυτές τις συνθήκες δε θα περιμένετε να υπάρχει εμπλουτισμός σε καμία λειτουργική κατηγορία. Είναι όμως λογικό να περιμένετε ότι για κανέναν από τους 8000 όρους για τους οποίους θα υπολογίσετε την πιθανότητα εμπλουτισμού, δε θα παρατηρήσετε μια τιμή p-value μικρότερη από 0.05; Η απάντηση είναι όχι. Για την ακρίβεια η πιθανότητα να παρατηρήσετε τουλάχιστο μία τιμή p-value<0.05 είναι ίση με $1-0.95^{8000}$~1. Το κλειδί εδώ είναι η τιμή του εκθέτη που αντιστοιχεί στον αριθμό των φορών που ελέγχθηκε η υπόθεσή σας. Στην Εικόνα 8.5 φαίνεται πώς ακόμα και για ένα μέτριου μεγέθους αριθμό ελέγχων η πιθανότητα για ένα τουλάχιστον εσφαλμένα σημαντικό αποτέλεσμα είναι 100%. Πολύ μεγάλοι αριθμοί πολλαπλότητας υποθέσεων οδηγούν αναπόφευκτα σε ένα ποσοστό εσφαλμένων "ανακαλύψεων" (false discoveries)[3].

Πώς όμως ξεπερνούμε το πρόβλημα των εσφαλμένων ανακαλύψεων που προκύπτουν από τον έλεγχο πολλαπλών υποθέσεων; Στη βιβλιογραφία έχουν προταθεί διάφοροι τρόποι για τη διόρθωση των αποτελεσμάτων. Όλες οι μέθοδοι συγκλίνουν στην κοινή διαδικασία απόρριψης ενός μέρους των αρχικά στατιστικά σημαντικών p-values. Η πιο απλή διόρθωση είναι αυτή που προτάθηκε αρχικά από τον Bonferroni, αλλά έγινε ευρύτερα γνωστή από τον Dunn (Dunn, 1959), σύμφωνα με την οποία η αρχική τιμή κατωφλίου p-value διαιρείται με τον αριθμό των ελέγχων και το πηλίκο που προκύπτει χρησιμοποιείται ως νέα τιμή κατωφλίου. Για το πείραμα έκφρασης που συζητήσαμε παραπάνω η νέα αυτή τιμή (που ονομάζεται και q-value) θα είναι ίση με q-value=0.05/20000=0.0000025! Το πρόβλημα με τη διόρθωση Bonferroni είναι ότι δεν αποδίδει ρεαλιστικές τιμές κατωφλίων σε πειράματα με μεγάλο αριθμό ελέγχων και τείνει να περιορίζει πολύ τα τελικά αποτελέσματα στα πειράματα έκφρασης. Μια πιο ρεαλιστική μέθοδος είναι αυτή των Benjamini και Hochberg (Benjamini & Hochberg, 1995) που είναι γνωστή και ως Ποσοστό Εσφαλμένων Ανακαλύψεων (False Discovery Rate, FDR). Η ανάλυση μέσω του FDR ουσιαστικά γίνεται με την οριοθέτηση ενός δεύτερου κατωφλίου *q* μετά την τιμή όριο για το p-value. Η τιμή *q* αντιστοιχεί τώρα στο ποσοστό του συνόλου των "ανακαλύψεων" (των σημαντικών δηλαδή γονιδίων με βάση μόνο το p-value) που αναμένεται να είναι εσφαλμένες. Με βάση έτσι ένα q=0.05 (ή 5% FDR) η μέθοδος των Benjamini και Hochberg καθορίζει ένα νέο όριο για το p-value με βάση την κατανομή των τιμών του. Είναι λογικό οι πολύ χαμηλές τιμές p-value να έχουν μικρότερη πιθανότητα να αντιστοιχούν σε εσφαλμένες ανακαλύψεις. Έτσι, μετατοπίζοντας το όριο του κατωφλίου του p-value σε συνάρτηση με το όριο του q-value ένα μέρος των αρχικών τιμών p-value απορρίπτονται κι αυτές που απομένουν είναι εμπλουτισμένες σε πιθανότητα να είναι πραγματικές ανακαλύψεις.

Είναι προφανές ότι τα όρια για τις τιμές p- και q-value είναι εξίσου αυθαίρετα, ωστόσο είναι χρήσιμο να έχουμε στο μυαλό μας τη λογική που βρίσκεται πίσω από αυτές. Όλα τα γνωστά προγράμματα στατιστικής που αναφέραμε και παραπάνω προσφέρουν την επιλογή διόρθωσης για πολλαπλές υποθέσεις κι έτσι οι λεπτομέρειες για τις διορθώσεις παραλείπονται.

***Ερώτηση**: Μπορείτε να σκεφτείτε έναν τρόπο γραφικής παράστασης μιας λειτουργικής ανάλυσης πολλών ταυτόχρονα λειτουργικών κατηγοριών που θα αναπαριστά ταυτόχρονα τις τιμές εμπλουτισμού και τις τιμές q-value;*

[3] *Οι δεισιδαίμονες μπορούν να θυμούνται ότι με όριο σημαντικότητας το 0.05 αρκούν 13 έλεγχοι για να εξασφαλίσουν πιθανότητα 50% να προκύψει τουλάχιστο μία εσφαλμένη ανακάλυψη.*

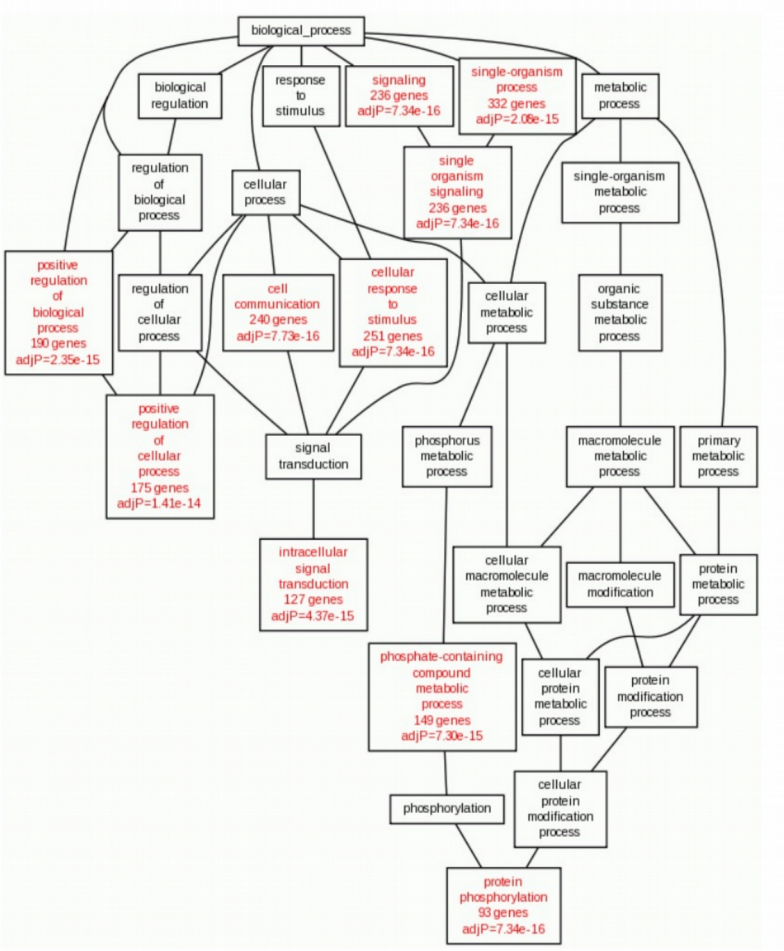
Μέθοδοι Λειτουργικής Ανάλυσης της Γονιδιακής Έκφρασης

Η λειτουργική ανάλυση της γονιδιακής έκφρασης αποτελεί βασικό στάδιο για την αξιοποίηση ενός ολοένα αυξανόμενου αριθμού πειραμάτων. Ταυτόχρονα, η εύκολη εφαρμογή τους και η ερμηνεία των αποτελεσμάτων τους ακόμα και από ερευνητές που δεν είναι εξοικειωμένοι με υπολογιστικές μεθοδολογίες, καθιστούν τις αναλύσεις αυτού του τύπου ιδιαίτερα δημοφιλείς. Ένας μεγάλος αριθμός διαδικτυακών εργαλείων για την ανάλυση της γονιδιακής έκφρασης είναι έτσι διαθέσιμος (Huang, Sherman & Lempicki, 2009a) όμως η επιμέρους παρουσίαση ακόμα και των πιο ευρέως χρησιμοποιούμενων από αυτά ξεφεύγει από τους σκοπούς αυτού του βιβλίου. Είναι σημαντικό ωστόσο να αναδειχτούν οι μεθοδολογικές διαφορές που υπάρχουν μεταξύ τους και που ορίζουν ευρύτερες κατηγορίες υπολογιστικών προσεγγίσεων στο πρόβλημα της ανάλυσης του εμπλουτισμού λειτουργικών κατηγοριών.

Στην επόμενη, τελευταία ενότητα αυτού του κεφαλαίου θα παρουσιάσουμε τις τρεις βασικότερες κατηγορίες μεθόδων παραθέτοντας παράλληλα κάποια χαρακτηριστικά προγράμματα ανάλυσης για την καθεμία από αυτές.

Απλή Ανάλυση Εμπλουτισμού (SEA)

Η Απλή Ανάλυση Εμπλουτισμού (Single Enrichment Analysis, SEA), όπως περιγράφηκε παραπάνω περιλαμβάνει, σε δύο στάδια, τον υπολογισμό των σχετικών εμπλουτισμών και το στατιστικό έλεγχο των τιμών τους μέσω της υπεργεωμετρικής κατανομής. Ο συγκεκριμένος τύπος ανάλυσης είναι ο απλούστερος αλλά αποτελεί τη βάση για όλες τις πιο σύνθετες μεθοδολογίες. Μια σειρά από διαδικτυακά εργαλεία έχουν αναπτυχθεί στη βάση της SEA, μεταξύ των οποίων βρίσκονται κάποια από τα πιο ευρέως χρησιμοποιούμενα. Ενδεικτικά αναφέρουμε το WebGestalt (http://bioinfo.vanderbilt.edu/webgestalt/) (B. Zhang, Kirov & Snoddy, 2005), το οποίο διενεργεί λειτουργική ανάλυση σε διάφορα επίπεδα και παρουσιάζει τα αποτελέσματα της γονιδιακής οντολογίας με μορφή ακυκλικών γράφων και το DAVID (https://david.ncifcrf.gov/) (Huang, Sherman & Lempicki, 2009b) το οποίο προσφέρει ανάλυση για έναν πολύ μεγάλο αριθμό διαφορετικών λειτουργικών κατηγοριών, που περιλαμβάνει μεταξύ άλλων και δίκτυα πρωτεϊνικών αλληλεπιδράσεων. Στην Εικόνα 8.6 μπορεί κανείς να δει το αποτέλεσμα μιας ανάλυσης μέσω του WebGestalt για ένα απλό παράδειγμα ενός πειράματος γονιδιακής έκφρασης με ~300 υπερεκφραζόμενα γονίδια.



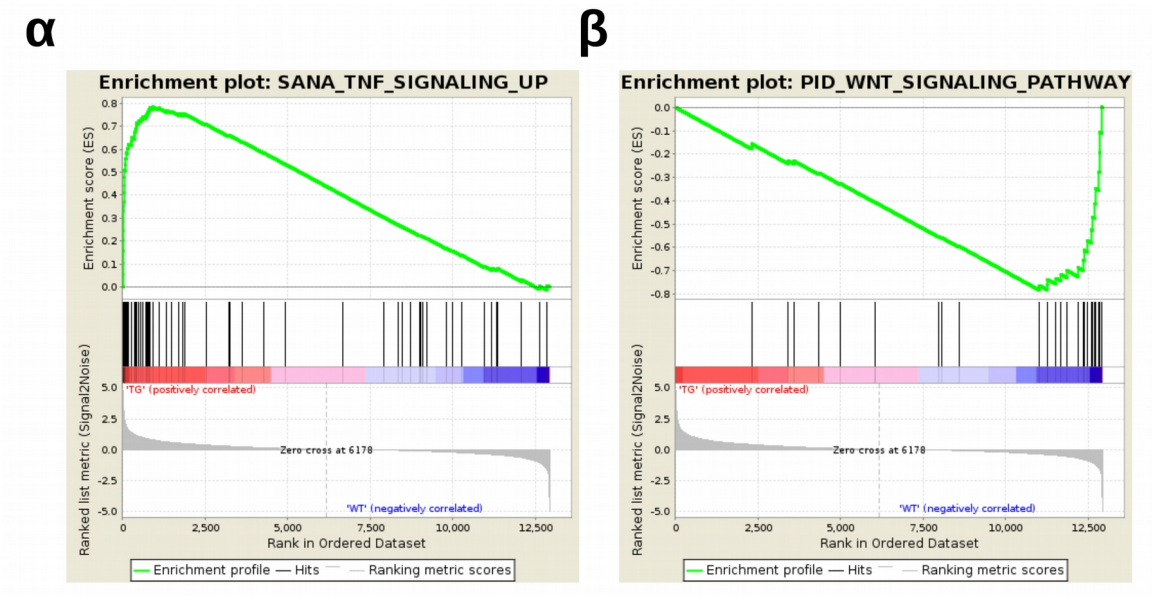
Εικόνα 8.6: *Παράδειγμα αποτελεσμάτων μετά την εφαρμογή μιας μεθόδου SEA (WebGestalt). Οι στατιστικά σημαντικά εμπλουτισμένες λειτουργίες σε επίπεδο γονιδιακής οντολογίας αναπαρίστανται σε έναν ακυκλικό γράφο που περιγράφει τις σχέσεις μεταξύ των εμπλουτισμένων λειτουργιών (με κόκκινο) και αυτών που σχετίζονται με αυτές (μαύρο) για την κατηγορία "Κυτταρική Διεργασία".*

Ανάλυση Εμπλουτισμού σε Σύνολα Γονιδίων (GSEA)

Η Ανάλυση Εμπλουτισμού σε Σύνολα Γονιδίων (Gene Set Enrichment Analysis, GSEA) είναι μια μέθοδος με δύο βασικά πλεονεκτήματα σε σχέση με την απλούστερη SEA. Πρώτον, δεν απαιτεί την εφαρμογή αυθαίρετων κριτηρίων για τον ορισμό ενός υποσυνόλου διαφορικά εκφραζόμενων γονιδίων. Δεύτερον, ακριβώς εξαιτίας αυτής της απουσίας τιμών-κατωφλίων, χρησιμοποιεί το σύνολο των δεδομένων αντί για ένα περιορισμένο μέρος τους. Τα δεδομένα εισόδου στην GSEA είναι οι τιμές έκφρασης από ολόκληρο το πείραμα. Το τελικό αποτέλεσμα είναι κι εδώ μια σειρά από τιμές p-value που αξιολογούν το βαθμό εμπλουτισμού μιας δεδομένης λειτουργίας, όμως ο τρόπος που υπολογίζεται τόσο η κάθε τιμή p-value αλλά και ο εμπλουτισμός διαφέρουν από την SEA. Η διαδικασία περιγράφεται σχηματικά στην Εικόνα 8.7 και έχει ως εξής:

Αλγόριθμος :: GSEA
Δεδομένης μιας σειράς κατάταξης R(N) από τιμές $\log_2$FC για N γονίδια
Δεδομένης μιας λειτουργίας A στην οποία ανήκουν K γονίδια
1. Εντόπισε στην κατάταξη R τα K γονίδια της A
2. Υπολόγισε την αθροιστική καμπύλη των τιμών R για τα K γονίδια ως εξής:
S(A)=S(A)+R[i] για i=1 έως K
3. Απόδωσε τιμή Εμπλουτισμού r(A)=max(S(A)) #ή αντίστοιχα r(A)=min(S(A)
4. Απόδωσε p-value p(A) την τιμή που προκύπτει από τη σύγκριση S(A) με μια σειρά από τυχαιοποιημένες κατανομές
Τερματισμός

Ο παραπάνω αλγόριθμος περιγράφει ουσιαστικά μια διαδικασία που συγκρίνει μια κατάταξη γονιδίων με βάση τις τιμές διαφορικής έκφρασης με το υποσύνολο των γονιδίων που ανήκουν σε μια συγκεκριμένη λειτουργία. Αν τα γονίδια που σχετίζονται με τη λειτουργία είναι σημαντικά υπερ-εκφραζόμενα τότε η αθροιστική καμπύλη των τιμών έκφρασης θα έχει ένα χαρακτηριστικό ασύμμετρο σχήμα προς τα πάνω και αριστερά (υψηλές τιμές) όπως φαίνεται στην Εικόνα 8.7.



Εικόνα 8.7: *Γραφική παράσταση αποτελεσμάτων της μεθόδου GSEA για ένα πείραμα επαγωγής μιας κυτταρικής σειράς με TNF. α) Αριστερά το διάγραμμα εμπλουτισμού για τη λειτουργική κατηγορία "Σηματοδότηση μέσω TNF", τα γονίδια της οποίας (λογικά) ενεργοποιούνται και βρίσκονται μεταξύ των πιο υψηλά εκφραζόμενων (μαύρες γραμμές στα αριστερά). β) Δεξία το ίδιο διάγραμμα για τα γονίδια της κατηγορίας "Σηματοδότηση μέσω WNT", μιας λειτουργίας που στο συγκεκριμένο σύστημα καταστέλλεται σημαντικά, με σημαντικό αριθμό γονιδίων να κατατάσσονται στις θέσεις με τη χαμηλότερη έκφραση (μαύρες γραμμές στα δεξιά).*

Το μέγιστο ύψος της καμπύλης δίνει, σ' αυτήν την περίπτωση, την τιμή εμπλουτισμού. Αντίστοιχα το σχήμα θα είναι ασύμμετρο προς τα κάτω και δεξιά εφόσον ο εμπλουτισμός είναι σε υπο-εκφραζόμενα γονίδια. Σε κάθε περίπτωση, η στατιστική σημαντικότητα της τιμής εμπλουτισμού δίνεται από τη σύγκριση της καμπύλης S(A) με μια σειρά από τυχαιοποιημένες κατανομές. Ο

στατιστικός έλεγχος που εφαρμόζεται είναι ο έλεγχος Kolmogorov-Smirnov (K-S test)[4]. Η GSEA παρουσιάστηκε αρχικά με τη μορφή που έχει στο (http://www.broadinstitute.org/gsea/index.jsp) (Subramanian, Tamayo, Mootha, Mukherjee & Ebert, 2005), ενώ πλέον και άλλες μέθοδοι όπως η ErmineJ (http://erminej.chibi.ubc.ca/) (Lee, Braynen, Keshav & Pavlidis, 2005) χρησιμοποιούν την ίδια μεθοδολογική βάση.

Όπως αναφέραμε και παραπάνω, το βασικό πλεονέκτημα της GSEA είναι ότι χρησιμοποιεί το σύνολο της πληροφορίας όπως αυτό προκύπτει από ένα πείραμα μέτρησης γονιδιακής έκφρασης χωρίς να απαιτεί από τον πειραματιστή να παράσχει έναν κατάλογο γονιδίων θέτοντας αυθαίρετα όρια διαφορικής έκφρασης. Από την άλλη πλευρά υπάρχουν κάποιοι περιορισμοί στη χρήση της που προκύπτουν ακριβώς από την ανάγκη ύπαρξης αριθμητικών δεδομένων για όλα τα γονίδια. Συχνά μας ζητείται να διενεργήσουμε μια ανάλυση σ' ένα ήδη δεδομένο υποσύνολο γονιδίων (που μπορεί να έχει προκύψει με διάφορους τρόπους) χωρίς να έχουμε τις αντίστοιχες τιμές έκφρασής τους. Από την άλλη πλευρά η GSEA εμφανίζει μια ελαφρώς μεγαλύτερη εξάρτηση στα γονίδια με ακραίες τιμές έκφρασης σε σχέση με την SEA. Σε κάθε περίπτωση τα αποτελέσματα που προκύπτουν από τις δύο μεθόδους είναι απόλυτα συγκρίσιμα εφόσον προηγηθεί μια προσεκτική επιλογή τιμών κατωφλίων για την SEA.

Σπονδυλωτή (modular) Ανάλυση Εμπλουτισμού (MEA)

Κληρονομώντας το βασικό υπολογισμό εμπλουτισμών από την SEA, η Σπονδυλωτή Ανάλυση Εμπλουτισμού (Modular Enrichment Analysis, MEA) ενσωματώνει επιπλέον αλγορίθμους που αποσκοπούν στην ανάδειξη ιδιοτήτων δικτύων που λαμβάνουν υπόψη τις σχέσεις μεταξύ λειτουργικών όρων. Έτσι αν δύο όροι π.χ. γονιδιακής οντολογίας βρεθούν να είναι σημαντικά εμπλουτισμένοι αλλά ταυτόχρονα βρίσκονται και σε γειτονικές θέσεις στο γράφο της ιεραρχίας των όρων, θα θεωρηθούν ακόμα μεγαλύτερης σημασίας. Το βασικό πλεονέκτημα αυτών των μεθόδων είναι ότι επιτρέπουν την εξόρυξη πληροφορίας που σχετίζεται με βαθύτερες βιολογικές σχέσεις όπως η ιεραρχική οργάνωση κυτταρικών διεργασιών ή οι αλληλεπιδράσεις μεταξύ μονοπατιών. Στον αντίποδα, ένας βασικός περιορισμός είναι ότι η MEA απαιτεί την ύπαρξη ιεραρχίας στην οργάνωση των λειτουργικών κατηγοριών (όρων) και έτσι μπορεί να χρησιμοποιηθεί κυρίως στην περίπτωση των γονιδιακών οντολογιών. Οι προσπάθειες για την ιεραρχική οργάνωση κι άλλων λειτουργικών κατηγοριοποιήσεων έχουν ενταθεί καθώς είναι προφανές πως θα συμβάλλουν σημαντικά στην καλύτερη και πληρέστερη βιολογική ερμηνεία των πειραμάτων έκφρασης με μικρότερη εξάρτηση από ακραίες τιμές έκφρασης που δημιουργούν μια χαρακτηριστική τάση υπερ-εκπροσώπησης συγκεκριμένων λειτουργικών κατηγοριών (π.χ. ρύθμισης της μεταγραφής). Σημαντική είναι η συνεισφορά, σε αυτό το πεδίο, της ανάλυσης βιολογικών δικτύων η οποία αποτελεί αντικείμενο του αμέσως επόμενου κεφαλαίου.

Οι μέθοδοι που ενσωματώνουν χαρακτηριστικά MEA ανήκουν στην τελευταία γενιά εργαλείων λειτουργικής ανάλυσης με πιο χαρακτηριστικά από αυτά να είναι τα TopGO (http://topgo.bioinf.mpi-inf.mpg.de/) (Alexa, Rahnenfuhrer & Lengauer, 2006) και Ontologizer (http://compbio.charite.de/contao/index.php/ontologizer2.html) (Bauer, Grossmann, Vingron & Robinson, 2008) ενώ στοιχεία MEA υπάρχουν και σε αναθεωρημένες εκδόσεις συγκεκριμένων

[4] *Η συζήτηση για το συγκεκριμένο στατιστικό έλεγχο ξεφεύγει από τους σκοπούς του βιβλίου.*

προγραμμάτων όπως το DAVID.

Συμπεράσματα

Η λειτουργική ανάλυση παρουσιάστηκε σ' αυτό το κεφάλαιο από τη σκοπιά της περαιτέρω διερεύνησης πειραμάτων γονιδιακής έκφρασης. Κάτω απ' αυτό το πρίσμα εξετάσαμε αποκλειστικά λειτουργικές ιδιότητες γονιδίων και των πρωτεϊνικών προϊόντων τους όπως αυτές εκφράζονται από τη γονιδιακή οντολογία, τη συμμετοχή τους σε βιολογικά μονοπάτια κλπ. Όμως ο διαρκώς αυξανόμενος όγκος γονιδιωματικών δεδομένων επιτρέπει, αν δεν επιβάλλει, την επέκταση της λειτουργικής ανάλυσης τόσο σε μη-γονιδιακές οντότητες όσο και σε λειτουργικές κατηγοριοποιήσεις διαφορετικών τύπων.

Τα τελευταία χρόνια, σύγχρονες προσεγγίσεις γονιδιωματικής που διενεργούνται σε συνδυασμό με αλληλούχιση DNA νέας γενιάς οδηγούν στην αποκομιδή δεδομένων όπως π.χ. οι θέσεις πρόσδεσεις μεταγραφικών παραγόντων στο DNA, θέσεις διαφορικής μεθυλίωσης ή περιοχές ανοιχτής χρωματίνης σε γονιδιωματική κλίμακα. Η ανάλυση τέτοιου τύπου δεδομένων (τα οποία θα συζητήσουμε πιο αναλυτικά στο Κεφάλαιο 11) έχει οδηγήσει στην ανάγκη εφαρμογής διευρυμένων προσεγγίσεων λειτουργικής ανάλυσης. Από τη στιγμή που πρόκειται για δεδομένα που εκφράζονται ως γονιδιωματικές συντεταγμένες, μια πρώτη ανάλυση σχετίζεται με τη διερεύνηση των τοπολογικών τους προτιμήσεων (McLean et al., 2010). Προσεγγίσεις γενικευμένης λειτουργικής ανάλυσης εξετάζουν έτσι όχι μόνο αν οι θέσεις πρόσδεσης ενός συγκεκριμένου μεταγραφικού παράγοντα τείνουν να βρίσκονται κοντά σε συγκεκριμένες λειτουργικές κατηγορίες γονιδίων, αλλά επεκτείνονται και σε τοπολογικά χαρακτηριστικά, υπολογίζοντας π.χ. κατά πόσο οι θέσεις αυτές τείνουν να βρίσκονται πλησιέστερα ή μακρύτερα από σημεία έναρξης της μεταγραφής, αν έχουν την τάση να συνεντοπίζονται με ενισχυτές ή υποκινητές κλπ.

Γενικευμένες προσεγγίσεις αυτού του τύπου είναι ιδιαίτερα σημαντικές στην προσπάθειά μας να αναλύσουμε ταυτόχρονα δεδομένα διαφορετικών τύπων από το ίδιο σύστημα. Ένα παράδειγμα μιας τέτοιας προσέγγισης σε επίπεδο “βιολογίας συστημάτων” θα δούμε στο τελευταίο κεφάλαιο αυτού του βιβλίου (Κεφάλαιο 12). Πριν φτάσουμε όμως εκεί θα εξετάσουμε ένα χαρακτηριστικό των βιολογικών συστημάτων που αποκτά όλο και μεγαλύτερη σημασία σε συστημικές προσεγγίσεις. Αυτό δεν είναι άλλο από τα βιολογικά δίκτυα.

Ερωτήσεις/Ασκήσεις

1. Από ένα πείραμα ολικής ανάλυσης της γονιδιακής έκφρασης που συμπεριέλαβε τα 6000 γονίδια του Σακχαρομύκητα κάτω από κανονικές συνθήκες και σε συνθήκες έλλειψης γλυκόζης, προέκυψαν 250 γονίδια με ανεβασμένη έκφραση στις συνθήκες έλλειψης γλυκόζης. Με βάση τη γονιδιακή οντολογία 50 από αυτά τα γονίδια σχετίζονταν με τον κυτταρικό κύκλο και 200 με το μεταβολισμό των υδατανθράκων. Πιστεύετε ότι κάποια από τις δύο λειτουργίες είναι υπερεκπροσωπούμενη στο δείγμα σας των 250 γονιδίων; Να δικαιολογήσετε την απάντησή σας.

2. Οι κατηγορίες Α-Κ είναι οι 10 κατηγορίες γονιδιακής οντολογίας του οργανισμού Χ. Στον πίνακα φαίνονται α) ο αριθμός των συνολικών γονιδίων που ανήκουν στην καθεμία και β) ο αριθμός των γονιδίων της καθεμίας που βρέθηκαν να ενεργοποιούνται σε ένα πείραμα έκφρασης σε γονιδιωματικό επίπεδο (μετρήθηκαν όλα τα γονίδια του Χ). Να αναφέρετε α) ποιες λειτουργίες είναι εμπλουτισμένες σε βαθμό τουλάχιστο διπλάσιο από τον αναμενόμενο και β) ποιες είναι απεμπλουτισμένες σε βαθμό τουλάχιστο μισό του αναμενόμενου.

GO	Α	Β	Γ	Δ	Ε	Ζ	Η	Θ	Ι	Κ
Συνολικά Γονίδια	100	80	200	1000	10	120	40	150	200	100
Ενεργοπ. Γονίδια	10	16	5	80	2	60	0	15	10	2

3. Στην παραπάνω ερώτηση, να υπολογίσετε τις τιμές p-value μέσω της υπεργεωμετρικής κατανομής λαμβάνοντας υπόψη την κατεύθυνση του εμπλουτισμού (θετικός/αρνητικός). Μπορείτε να χρησιμοποιήσετε οποιαδήποτε εφαρμογή υπολογισμού της κατανομής επιθυμείτε.

Διαβάστε Περισσότερα

Για τη γονιδιακή οντολογία και τα βιολογικά μονοπάτια:
Στο Κεφάλαιο 10 του *Bioinformatics and Functional Genomics* (Pevsner, 2015) μπορείτε να βρείτε μια αναλυτική περιγραφή των περιεχομένων της βάσης GO χωρίς όμως περισσότερες λεπτομέρειες για τη δομή της. Οι βασικές αναφορές είναι αυτές που παρατίθενται στο κείμενο. Περισσότερα μπορείτε να βρείτε στις αντίστοιχες ιστοσελίδες. Για τη Γονιδιακή Οντολογία: (http://geneontology.org/). Για την KEGG (http://www.genome.jp/kegg/).

Για τα δίκτυα πρωτεϊνικών αλληλεπιδράσεων:
Μια καλή εισαγωγή είναι το άρθρο της βάσης δεδομένων STRING-db, που είναι μάλλον η βάση αναφοράς για πρωτεϊνικές αλληλεπιδράσεις (Franceschini et al., 2013), ενώ στο (Vazquez et al., 2003) παρουσιάζεται με μορφή επισκόπησης η σύνδεση των πρωτεϊνικών δικτύων αλληλεπίδρασης με το λειτουργικό χαρακτηρισμό τους. Για τον αναγνώστη που επιθυμεί να εμβαθύνει, ένα εξαιρετικό βιβλίο είναι το *Protein Interaction Networks: Computational Analysis* (A. Zhang, 2009).

Για την ανάλυση εμπλουτισμού:
Μια πολύ καλή συνοπτική παρουσίαση των διαφορετικών ειδών ανάλυσης γίνεται στο άρθρο επισκόπησης (Huang et al., 2009a).

Για το μαθηματικό υπόβαθρο της υπεργεωμετρικής κατανομής:
Βασικά συγγράμματα πιθανοτήτων και στατιστικής περιέχουν αναπόφευκτα αναφορές σε όλα τα είδη κατανομών. Στο Κεφάλαιο 4 του *A first Course in Probability* (Ross, 2014) υπάρχει η σχετική συζήτηση για την υπεργεωμετρική κατανομή.

Για το μαθηματικό υπόβαθρο του Ελέγχου Πολλαπλών Υποθέσεων:
Μια πολύ καλή εισαγωγή ακόμα και για τους λιγότερο εξοικειωμένους με τα μαθηματικά δίνεται στα Κεφάλαια 22-23 του *Intuitive Biostatistics* (Motulsky, 2010).

Για τις μεθόδους λειτουργικής ανάλυσης γονιδιακής έκφρασης:
Για τα επιμέρους χαρακτηριστικά της κάθε μεθόδου, οι εργασίες που αναφέρονται στο κείμενο είναι το καλύτερο σημείο εκκίνησης.

Βιβλιογραφία

Alexa, A., Rahnenfuhrer, J. & Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, *22*(13), 1600–1607. http://doi.org/10.1093/bioinformatics/btl140

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29. http://doi.org/10.1038/75556

Bauer, S., Grossmann, S., Vingron, M. & Robinson, P. N. (2008). Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics (Oxford, England)*, *24*(14), 1650–1. http://doi.org/10.1093/bioinformatics/btn250

Beck, T., Hastings, R. K., Gollapudi, S., Free, R. C. & Brookes, A. J. (2014). GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *European Journal of Human Genetics : EJHG*, *22*(7), 949–52. http://doi.org/10.1038/ejhg.2013.274

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, *57*(1), 289–300. http://doi.org/10.2307/2346101

Bourne, P. E. (Ed.). (2009). The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Computational Biology*, *5*(7), e1000431. http://doi.org/10.1371/journal.pcbi.1000431

Consortium, M. E., Tables, S., Yue, F., Cheng, Y., Breschi, A., Vierstra, J., … Ren, B. (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, *515*(7527), 355–364. http://doi.org/10.1038/nature13992

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C., Doyle, F., … Gillespie, S. (2012, September 6). An integrated encyclopedia of DNA elements in the human genome. *Nature*.

Dunn, O. J. (1959). Estimation of the Medians for Dependent Variables. *The Annals of Mathematical Statistics*, *30*(1), 192–197.

Essaghir, A., Toffalini, F., Knoops, L., Kallin, A., van Helden, J. & Demoulin, J.-B. (2010). Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. *Nucleic Acids Research*, *38*(11), e120. http://doi.org/10.1093/nar/gkq149

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., … Jensen, L. J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, *41*(Database issue), D808–15. http://doi.org/10.1093/nar/gks1094

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic*

Acids Research, *33*(Database issue), D514–7. http://doi.org/10.1093/nar/gki033

Huang, D. W., Sherman, B. T. & Lempicki, R. a. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13. http://doi.org/10.1093/nar/gkn923

Huang, D. W., Sherman, B. T. & Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. http://doi.org/10.1038/nprot.2008.211

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., … Stein, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, *33*(Database issue), D428–32. http://doi.org/10.1093/nar/gki072

Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., … Pandey, A. (2009). Human Protein Reference Database--2009 update. *Nucleic Acids Research*, *37*(Database issue), D767–72. http://doi.org/10.1093/nar/gkn892

Lambrix, P., Tan, H., Jakonienė, V. & Strömbäck, L. (2007). Biological Ontologies. Springer.

Lee, H. K., Braynen, W., Keshav, K. & Pavlidis, P. (2005). ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, *6*, 269. http://doi.org/10.1186/1471-2105-6-269

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., … Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, *28*(5), 495–501. http://doi.org/10.1038/nbt.1630

Motulsky, H. (2010). *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking*. Oxford University Press.

Nishimura, D. (2001). BioCarta. *Biotech Software & Internet Report*, *2*(3), 117–120. http://doi.org/10.1089/152791601750294344

Osborne, J. D., Flatow, J., Holko, M., Lin, S. M., Kibbe, W. A., Zhu, L. J., … Chisholm, R. L. (2009). Annotating the human genome with Disease Ontology. *BMC Genomics*, *10 Suppl 1*, S6. http://doi.org/10.1186/1471-2164-10-S1-S6

Petryszak, R., Burdett, T., Fiorelli, B., Fonseca, N. a, Gonzalez-Porta, M., Hastings, E., … Brazma, A. (2014). Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Research*, *42*(Database issue), D926–32. http://doi.org/10.1093/nar/gkt1270

Pevsner, J. (2015). *Bioinformatics and Functional Genomics*. Wiley.

Ross, S. M. (2014). *A First Course in Probability*. Pearson Education.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S. & Ebert, B. L. (2005). Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. *Proc Natl Acad Sci U S A*,

102(43), 15545–15550.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., … von Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, *39*(Database issue), D561–8. http://doi.org/10.1093/nar/gkq973

Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, *21*(6), 697–700. http://doi.org/10.1038/nbt825

Vlachos, I. S., Paraskevopoulou, M. D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., … Hatzigeorgiou, A. G. (2014). DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Research*, *43*(D1), D153–159. http://doi.org/10.1093/nar/gku1215

Zhang, A. (2009). *Protein Interaction Networks: Computational Analysis*. Cambridge University Press.

Zhang, B., Kirov, S. & Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*, *33*(Web Server issue), W741–8. http://doi.org/10.1093/nar/gki475