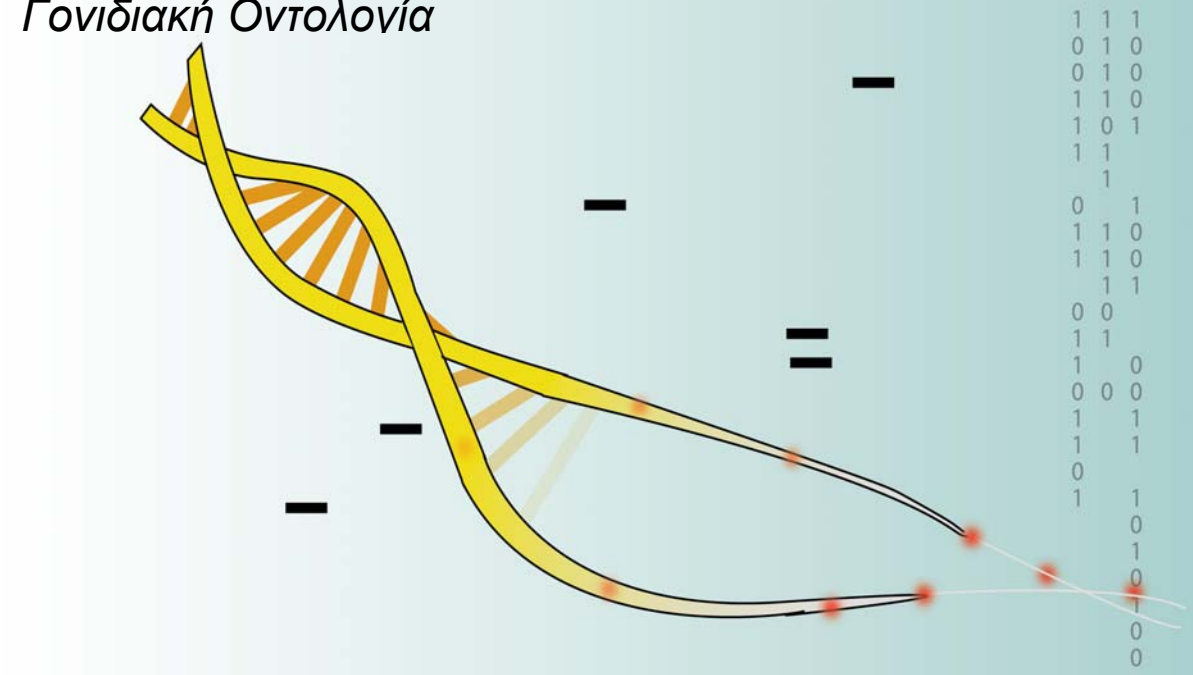
ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Διπλωματική Εργασία:

Βελτίωση της ακρίβειας ταξινόμησης συνόλων δεδομένων μικροσυστοιχιών με εκμετάλλευση γνώσης από τη Γονιδιακή Οντολονία



του Παπαχριστούδη Γεωργίου

Α.Ε.Μ.: 4691

Επιβλέπων Καθηγητής: Μήτκας Περικλής

Θεσσαλονίκη, Ιούνιος 2007

ΠΕΡΙΕΧΟΜΕΝΑ

Ευχαριστίες

Συνήθως η ενότητα των ευχαριστιών παρατίθεται στο τέλος ενός συγγράμματος ως μία υποτυπώδη εκδήλωση ευγνωμοσύνης στους ανθρώπους που στήριξαν είτε ηθικά είτε υλικά τον συγγραφέα. Ας μου επιτραπεί να κάνω την εξαίρεση και να τοποθετήσω αυτή την ενότητα στην αρχή αυτού του συγγράμματος ώστε να διασφαλιστεί το γεγονός της μνημόνευσης όλων αυτών που συνέβαλλαν σε αυτή την εργασία από κάποιον πιθανό αναγνώστη.

Αρχικά, θέλω να ευχαριστήσω τον Σωτήρη Διπλάρη, υποψήφιο διδάκτορα του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, ως επιβλέποντα της διπλωματικής μου για τη συνεργασία μας καθόλη τη διάρκεια εκπόνησης της. Πιο συγκεκριμένα, τον ευχαριστώ για τη βοήθεια που μου παρείχε κατά την πρώτη μου επαφή με θέματα Βιοπληροφορικής, την εξήγηση του βασικού βιολογικού υποβάθρου του εξεταζόμενου προβλήματος καθώς και για τη συνεισφορά του στην διπλωματική εργασία όποτε αυτή κρίθηκε απαραίτητη.

Επίσης, θα ήθελα να ευχαριστήσω τον κ. Michael Moorhouse, δόκτωρα του τμήματος Ιολογίας του Πανεπιστημίου Erasmus στην Ολλανδία, για τις πολύτιμες συμβουλές που μου έδωσε στην αρχή αυτής της διπλωματικής δίνοντάς μου αισιοδοξία όταν η πρόοδος της εργασίας αρχικά βρισκόταν σε στάσιμα επίπεδα.

Νιώθω επίσης την ανάγκη να ευχαριστήσω κ. Francisco Couto, επίκουρο καθηγητή του τμήματος Πληροφορικής του Πανεπιστημίου της Λισσαβόνας για την αποσαφήνιση ορισμένων σημείων της δημοσίευσής του [Couto Francisco, Silva Mário and Coutinho Pedro. "*Measuring Semantic Similarity between Gene Ontology Terms*". Data & Knowledge Engineering, Volume 61, Issue 1, April 2007, pp. 137–152.] και τη γόνιμη ανταλλαγή απόψεων κατά τις φορές που ήρθαμε σε επικοινωνία.

Εξάλλου, θα ήθελα να στείλω τις ευχαριστίες μου στην κ. Σοφία Κοσσίδα, επικεφαλή της ομάδας Βιοπληροφορικής στο Ινστιτούτο Ιατροβιολογικών Ερευνών της Ακαδημίας Αθηνών για την παροχή της δημοσίευσης [Pomeroy SL et al. "*Prediction of central nervous system embryonal tumour outcome based on gene*

expression". Nature, 415 (6870), 24 Jan 2002, pp. 436–442.] για τους σκοπούς της διπλωματικής.

Αισθάνομαι επίσης ευγνώμων για την πολύτιμη βοήθεια που μου προσέφερε ο φίλος μου Θανάσης Τερζής, πτυχιούχος του τμήματος Εφαρμογών Πληροφορικής στη Διοίκηση και Οικονομία του Ανωτάτου Τεχνολογικού Ιδρύματος Μεσολογγίου αναλαμβάνοντας την καλλιτεχνική επιμέλεια του εξωφύλλου.

Επιπλέον, θα ήθελα να ευχαριστήσω τον Φώτη Ψωμόπουλο, υποψήφιο διδάκτορα του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης για την αδιάλειπτη και πάντα με προθυμία παροχή της βοήθειας του όποτε του εζητήθη. Εξάλλου, στη διάρκεια της μονοετούς συνεργασίας μας αναπτύχθηκαν δεσμοί φιλίας καθώς και αισθήματα αμοιβαίας αλληλοεκτίμησης.

Ακόμη, θα ήθελα να ευχαριστήσω θερμά τον κ. Περικλή Μήτκα, καθηγητή του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης όχι μόνο για την εμπιστοσύνη που επέδειξε στο πρόσωπό μου αναθέτοντας μου την εκπόνηση αυτής της διπλωματικής εργασίας αλλά και γιατί με στήριξε σε προσπάθειες μου πέρα από τα πλαίσια αυτής της εργασίας και έδειξε κατανόηση σε περιόδους όπου η προτεραιότητα περάτωσης άλλων προσωπικών θεμάτων έφερναν σε δεύτερη μοίρα την ανάπτυξη της διπλωματικής.

Κλείνοντας τον κύκλο των ευχαριστιών, θα ήθελα να ευχαριστήσω τους γονείς μου και την αδερφή μου για την παροιμιώδη υπομονή που επέδειξαν σε ουκ ολίγες στιγμές εντάσεων, αποτέλεσμα κούρασης και κόπου, αλλά και για την άγρυπνη συμπαράστασή τους στην προσπάθειά μου να εκπληρώσω τους στόχους μου.

Τέλος, ξεφεύγοντας λίγο από τα στενά όρια απόδοσης ευχαριστιών σχετικά με την εκπόνηση αυτής της διπλωματικής εργασίας, δράττομαι της ευκαιρίας να χαρίσω σε όλους αυτούς που με στήριξαν ανιδιοτελώς και πίστεψαν σε μένα (συμπεριλαμβανομένων και κάποιων από τα παραπάνω πρόσωπα) βοηθώντας με να κυνηγήσω το όνειρό μου, το τρίτο στιχάκι από το παρακάτω απόφθεγμα του Νίκου Καζαντζάκη:

"*Το πρώτο χρέος σου εκτελώντας τη θητεία σου στη ράτσα είναι να νιώσεις μέσα σου όλους τους προγόνους.*

Το δεύτερο να φωτίσεις την ορμή τους και να συνεχίσεις το έργο τους.

Το τρίτο σου χρέος είναι να δώσεις στο γιό σου τη μεγάλη εντολή να σε ξεπεράσει"

Εικόνες

Πίνακες

1 Εισαγωγή

Τα τελευταία χρόνια η ολοένα αυξανόμενη υπολογιστική δύναμη έχει δώσει λύσεις σε προβλήματα πολλών διαφορετικών κλάδων επιστημών. Από την άλλη, η διάθεση απέραντων σχεδόν αποθηκευτικών χώρων δίνει τη δυνατότητα εκτέλεσης πειραμάτων που παράγουν τεράστιους όγκους δεδομένων και εσωκλείουν σημαντική πληροφορία. Η εφαρμογή μαθηματικών και στατιστικών μεθόδων βοηθούν στην απομόνωση των σημαντικών στοιχείων από πολύ μεγάλες πηγές δεδομένων και εν τέλει στην εξαγωγή αυτής της χρήσιμης πληροφορίας.

Τα παραπάνω έχουν δώσει ώθηση και σε βιολογικά προβλήματα. Οι όγκοι δεδομένων που παράγονται από τέτοιου είδους προβλήματα κάνουν την παρουσία της Πληροφορικής καθώς και άλλων εφαρμοσμένων επιστημών όχι μόνο επωφελή αλλά αναγκαία για την επεξεργασία τους. Η συνέργεια όλων των παραπάνω επιστημών για την επίλυση βιολογικών προβλημάτων οδήγησε στην ανάπτυξη ενός νέου κλάδου που είναι γνωστός με το όνομα Βιοπληροφορική. Με απλά λόγια, η Βιοπληροφορική είναι η ανάπτυξη και εφαρμογή εργαλείων για την οργάνωση και ανάλυση τέτοιων δεδομένων.

1.1 Ορισμός του προβλήματος

Ένα από τα πιο διαδεδομένα προβλήματα βιοπληροφορικής είναι αυτό της ανάλυσης μικροσυστοιχιών. Οι μικροσυστοιχίες είναι διατάξεις που επιτρέπουν την εποπτεία του τρόπου έκφρασης γονιδίων ή γονιδιακών προϊόντων υπό διαφορετικές συνθήκες. Παρόλα τα πλεονεκτήματα αυτών των διατάξεων, ο τεράστιος όγκος δεδομένων και κυρίως η μεγάλη διαφορά μεταξύ του πλήθους των γονιδίων και των διαφορετικών συνθηκών κάνει ιδιαίτερα δύσκολη την εξαγωγή χρήσιμης πληροφορίας. Πιο συγκεκριμένα, οι υπάρχοντες αλγόριθμοι εκμάθησης και φιλτραρίσματος γνωρισμάτων δεν έχουν κατασκευαστεί για να αντιμετωπίζουν

βέλτιστα τέτοια ογκώδη σύνολα δεδομένων. Έτσι, πέρα από την αύξηση του υπολογιστικού χρόνου είναι πιθανό να δημιουργηθούν και φαινόμενα υπερεκπαίδευσης, όπου καθίσταται αδύνατη η ασφαλής γενίκευση των συμπερασμάτων. Επιπλέον, η παρουσία πολλών άσχετων προς το πρόβλημα γονιδίων–γνωρισμάτων εμποδίζει την εμφάνιση της διακριτικός ισχύος των σχετικών (προς το πρόβλημα) γονιδίων–γνωρισμάτων. Τέλος, μεγάλη σημασία εμφανίζει η απόδοση βιολογικής ερμηνείας στα παραγόμενα αποτελέσματα κάτι που μέχρι στιγμής δεν καλύπτεται από τις υπάρχουσες μεθόδους. Τα παραπάνω έχουν ωθήσει στη δημιουργία ενός νέου εργαλείου, όπου γίνεται προσπάθεια εισαγωγής βιολογικής γνώσης στα δεδομένα με στόχο την παραγωγή αποτελεσμάτων με πληρέστερο βιολογικό νόημα.

1.2 Στόχοι διπλωματικής

Μέχρι τούδε ο συχνότερος τρόπος αντιμετώπισης αυτού του προβλήματος ήταν μέσω της εφαρμογής μεθόδων στατιστικής ομοιότητας όπου με αυτό τον τρόπο επιτυγχανόταν η απομόνωση των πιο πληροφοριακών γονιδίων βάσει των τιμών τους κατά μήκος του συγκεκριμένου συνόλου δεδομένων.

Αυτό που φιλοδοξούμε να επιτύχουμε σε αυτή τη διπλωματική θέση είναι:

i. Να βελτιώσουμε την ποιότητα των αποτελεσμάτων όχι μόνο με την αποκλειστική χρήση της εσωτερικής πληροφορίας του συνόλου δεδομένων μέσω της εφαρμογής μεθόδων επιλογής γνωρισμάτων αλλά να προχωρήσουμε ένα βήμα πιο μπροστά εκμεταλλευόμενοι τη γνώση που μας προσφέρει ένα δομημένο λεξιλόγιο γονιδιακών όρων, γνωστό με την ονομασία Γονιδιακή Οντολογία.
ii. Να αναπτύξουμε την κατάλληλη μεθοδολογία για την επίτευξη των προαναφερθέντων απαιτήσεων.
iii. Να κατασκευάσουμε το εργαλείο που θα βασίζεται στην παραπάνω μεθοδολογία και θα υλοποιείται σε μία μεταφέρσιμη γλώσσα προγραμματισμού (όπως είναι η Java).
iv. Να εκτελέσουμε ικανό αριθμό πειραμάτων που θα πιστοποιούν τη χρησιμότητα ή μη του παραπάνω εργαλείου.

Με λίγα λόγια, ο κύριος στόχος της διπλωματικής αυτής είναι η βελτίωση όχι μόνο της ακρίβειας ταξινόμησης καθώς και άλλων μέτρων ελέγχου της αξιοπιστίας

των αποτελεσμάτων αλλά και η απόδοση περισσότερης βιολογικής ερμηνείας σε αυτά αξιοποιώντας πληροφορία που περιέχει βιολογικό νόημα και είναι ήδη επιστημονικώς τεκμηριωμένη.

1.3 Μεθοδολογία

Το κύριο πρόβλημα αυτής της διπλωματικής ήταν η εύρεση ενός τρόπου ενοποίησης της υπάρχουσας βιολογικής γνώσης και της γνώσης που εμπεριέχεται στα εσωτερικά χαρακτηριστικά των συνόλων δεδομένων. Αναλυτικότερα, οι είσοδοι για το πρόβλημά μας ήταν σύνολα δεδομένων μικροσυστοιχιών καθώς και αρχεία Γονιδιακής Οντολογίας και Καταχώρησης Όρων Γονιδιακής Οντολογίας. Ο τελικός στόχος ήταν η άντληση της βιολογικής πληροφορίας που περιέχεται στη Γονιδιακή Οντολογία και σε συνδυασμό με την εφαρμογή μεθόδων φιλτραρίσματος γνωρισμάτων στο αρχικό σύνολο δεδομένων, η παραγωγή ενός μικρότερου συνόλου που θα περιείχε ως επί το πλείστον γονίδια–γνωρίσματα σχετικά προς το πρόβλημα. Με αυτό τον τρόπο επιχειρείται η βελτίωση της ποιότητας των αποτελεσμάτων.

Συνοπτικά, η διαδικασία που ακολουθήθηκε για την επίτευξη του παραπάνω στόχου ήταν η εξής:

Αρχικά εφαρμόζεται κάποια από τις γνωστές μεθόδους φιλτραρίσματος γνωρισμάτων. Αποτέλεσμα αυτής της διαδικασίας είναι η απομόνωση των πιο πληροφοριακών γονιδίων (γονιδίων–γνωρισμάτων) βάσει των εσωτερικών χαρακτηριστικών τους στο σύνολο δεδομένων. Από αυτό το στάδιο κατασκευάζεται και ένα αρχείο που περιέχει τα πιο πληροφοριακά γονίδια του αρχικού συνόλου δεδομένων (καθώς και τιμές τους κατά μήκος όλων των παραδειγμάτων του αρχικού συνόλου).

Στη συνέχεια, αφού εντοπιστούν όλοι οι όροι (από τις επιθυμητές οντολογίες–απόψεις) των γονιδίων του συνόλου δεδομένων μέσω των GOA αρχείων, υπολογίζεται η σημασιολογική ομοιότητα μεταξύ των πιο πληροφορικών γονιδίων (αυτών που περιέχονται στο μικρότερο αρχείο) και αυτών που απορρίφθηκαν κατά την εφαρμογή μεθόδων επιλογής γνωρισμάτων.

Κατόπιν, επαναεισάγονται ορισμένα από τα γονίδια που είχαν προηγουμένως απορριφθεί αλλά εμφανίζονται να έχουν υψηλή σημασιολογική ομοιότητα (άρα, κοινές λειτουργίες) με τα πιο πληροφοριακά γονίδια. Δημιουργείται, έτσι, ένα σύνολο δεδομένων που περιέχει τα πιο πληροφοριακά γονίδια και επιπροσθέτως γονίδια που έχουν μεγάλη (λειτουργική) ομοιότητα με αυτά.

Τέλος, για την αποτίμηση των αποτελεσμάτων γίνεται σύγκριση της ακρίβειας ταξινόμησης καθώς και άλλων μέτρων ελέγχου της αξιοπιστίας μεταξύ του αρχικού συνόλου, αυτού που περιέχει μόνο τα πιο πληροφοριακά γονίδια–γνωρίσματα και αυτού που παράγεται από το εργαλείο αυτής της διπλωματικής (που περιέχει δηλαδή τα πιο πληροφοριακά γονίδια–γνωρίσματα και επιπλέον τα γονίδια–γνωρίσματα που εμφανίζουν μεγάλη ομοιότητα με αυτά).

1.4 Οργάνωση της διπλωματικής

Η διπλωματική συνεχίζει με το Κεφάλαιο 2 όπου γίνεται μία εισαγωγή στις βασικές βιολογικές έννοιες που αφορούν το συγκεκριμένο πρόβλημα. Αρχικά, ορίζονται οι έννοιες της Βιοπληροφορικής και της Υπολογιστικής Βιολογίας (2.1), κλάδοι στους οποίους εντάσσεται το συγκεκριμένο πρόβλημα. Έπειτα, επεξηγώντας γενικές έννοιες όπως το DNA, τα γονίδια και τη γονιδιακή έκφραση (2.2) επιτυγχάνεται το ομαλό πέρασμα στις μικροσυστοιχίες (2.3), όπου γίνεται εκτενής αναφορά σε αυτές και τη χρησιμότητά τους ενώ στο τέλος παρατίθενται και τα μειονεκτήματα από τη χρησιμοποίησή τους.

Αφού έχουμε αναπτύξει το στέρεο βιολογικό υπόβαθρο του προβλήματος, προχωρούμε στο Κεφάλαιο 3 όπου αναλύεται αρχικά η έννοια της Οντολογίας (3.1) και κατόπιν η έννοια της Γονιδιακής Οντολογίας (3.3), που αποτελεί το βασικότερο εργαλείο σε αυτή τη διπλωματική.

Κατόπιν, το Κεφάλαιο 4 αποτελεί το τεχνικό κομμάτι της διπλωματικής αφού περιγράφονται αναλυτικά οι χρησιμοποιούμενες μέθοδοι. Αρχικά, γίνεται αναφορά σε γενικές έννοιες όπως η Εξόρυξη Δεδομένων, Απόκτηση Γνώσης και Εκμάθηση Μηχανών (4.1). Στη συνέχεια, παρουσιάζεται η έννοια του φιλτραρίσματος γνωρισμάτων (4.2) προτού παρατεθούν οι λόγοι χρησιμοποίησης αυτών των μεθόδων, ενώ επιχειρείται και η ταξινόμηση αυτών των μεθόδων βάσει των διαφορετικών κριτηρίων αξιολόγησης και διαδικασιών παραγωγής. Εν συνεχεία, περιγράφονται διεξοδικά οι μέθοδοι φιλτραρίσματος γνωρισμάτων που χρησιμοποιήθηκαν σε αυτήν εδώ τη διπλωματική (Information Gain, Chi Square, Relief). Συνεχίζει με την ανάπτυξη της θεωρίας των Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines) ως αλγορίθμου εκπαίδευσης των δεδομένων (4.3). Το Κεφάλαιο 4 κλείνει με την παράθεση της έννοιας της σημασιολογικής ομοιότητας (4.4) και την επεξήγηση ενός βασικότατου στοιχείου που χρησιμοποιείται κατά τον υπολογισμό της σημασιολογικής ομοιότητας, του

πληροφοριακού περιεχομένου. Τέλος, αναπτύσσονται οι μέθοδοι σημασιολογικής ομοιότητας που υλοποιήθηκαν σε αυτήν την εργασία.

Συνεχίζοντας, στο Κεφάλαιο 5 παρουσιάζεται αναλυτικά η πορεία ανάπτυξης του αλγορίθμου. Πιο συγκεκριμένα, αρχικά παρατίθενται οι λόγοι που οδήγησαν στην αναγκαιότητα ανάπτυξης ενός τέτοιου εργαλείου (5.1). Κατόπιν, γίνεται λόγος για το περιβάλλον ανάπτυξης του εργαλείου (5.2), ενώ στη συνέχεια παρουσιάζεται η δομή του (5.3). Αφού πια ο αναγνώστης έχει λάβει μία πρώτη γνώση του τρόπου λειτουργίας του εργαλείου, περιγράφονται οι διάφορες λειτουργίες του (5.4) καθώς και η κατάλληλη μορφή των δεδομένων εισόδου (5.5). Στο τέλος του Κεφαλαίου, παρατίθεται ο βασικός αλγόριθμος του SoFoCles (5.6).

Κατόπιν, στο Κεφάλαιο 6 επιχειρείται η αποτίμηση των αποτελεσμάτων. Αφού πρώτα περιγραφούν τα μέτρα ελέγχου της αξιοπιστίας των αποτελεσμάτων (6.1), γίνεται παρουσίαση και αξιολόγηση των αποτελεσμάτων σε δύο σύνολα δεδομένων· ενός που αφορά τις επιδράσεις του καπνίσματος στη μεταγραφή του γονιδιώματος των επιθηλιακών κυττάρων του ανθρώπινου αναπνευστικού συστήματος (D_S σύνολο δεδομένων (6.2)) και ενός που περιέχει πληροφορία σχετική με τα αποτελέσματα της θεραπείας σε ασθενείς που έχουν αναπτύξει εμβρυϊκούς όγκους στο Κεντρικό Νευρικό Σύστημα (D_{CNV} σύνολο δεδομένων (6.3)).

Στο Κεφάλαιο 7 γίνεται μία σύνοψη του αλγορίθμου, της λειτουργίας του καθώς και των στόχων που εξυπηρετεί ενώ προτείνονται και πιθανά σενάρια για ενδεχόμενη βελτίωσή του.

Εξάλλου, στο Κεφάλαιο 8, παρεμβάλλεται η Βιβλιογραφία και πιο συγκεκριμένα οι δημοσιεύσεις (8.1) και οι διαδικτυακές αναφορές (8.2) που αποτέλεσαν πηγή πληροφοριών για τη συγγραφή αυτής εδώ της αναφοράς καθώς και για την ανάπτυξη του εργαλείου.

Λίγο πριν το τέλος αυτής της αναφοράς, παρατίθενται τα Παραρτήματα, όπου στο Α γίνεται μία μικρή αναφορά στον τραγικό ποιητή Σοφοκλή και στην όποια ομοιότητα των καινοτομιών που εισήγαγε στα έργα του με τα στοιχεία αυτού του εργαλείου, SoFoCles. Από την άλλη, στο Παράρτημα Β αναφέρονται τα διάφορα πακέτα και τα εργαλεία που χρησιμοποιήθηκαν για την εκπόνηση αυτής της διπλωματικής.

Τέλος, η αναφορά αυτή κλείνει με την παράθεση ευρετηρίου για την ευκολότερη και λειτουργικότερη ανάγνωση.

2 Επιτομή βασικών βιολογικών εννοιών

Η εκθετική αύξηση του όγκου των βιολογικών και ιατρικών δεδομένων ιδίως μετά την αποκρυπτογράφηση του γονιδιώματος οργανισμών έθεσε ως απαραίτητο κριτήριο για την επεξεργασία τους την εκμετάλλευση της υπάρχουσας υπολογιστικής δύναμης και τη χρησιμοποίηση αλγορίθμων για την έγκυρη εξαγωγή συμπερασμάτων. Σε αυτό το σημείο, η Πληροφορική καθώς και άλλοι συγγενείς κλάδοι επιστρατεύτηκαν για την επίλυση τέτοιων απαιτητικών προβλημάτων. Οι δύο κλάδοι που επιφορτίζονται κυρίως με την επίλυση των παραπάνω προβλημάτων είναι η Βιοπληροφορική και η Υπολογιστική Βιολογία. Συνεπώς, απαραίτητη κρίνεται η αποσαφήνιση του αντικειμένου ενασχόλησης των δύο παραπάνω κατευθύνσεων καθώς και μία σύντομη επισκόπηση των σημαντικότερων περιοχών έρευνας.

2.1 Σχέση Βιοπληροφορικής και Υπολογιστικής Βιολογίας

Συνήθως οι όροι Βιοπληροφορική (Bioinformatics) και Υπολογιστική Βιολογία (Computational Biology) συγχέονται και όχι άδικα αφού υπάρχει σημαντική επικάλυψη όσον αφορά το αντικείμενο έρευνας τους. Η βάση τους προέρχεται από τις επιστήμες υγείας όπως και από επιστήμες υπολογιστών και πληροφορικής. Για την επίλυση των προβλημάτων των παραπάνω κλάδων επιστρατεύονται κλάδοι όπως εφαρμοσμένα μαθηματικά, φυσική, επιστήμη των υπολογιστών, μηχανική, στατιστική, τεχνητή νοημοσύνη, βιολογία, χημεία και επιστήμη της συμπεριφοράς (behavioral science). Παρόλα αυτά, οι δύο κλάδοι έχουν και διακριτούς ρόλους οι οποίοι σύμφωνα με τον λειτουργικό ορισμό (working definition) που δίνουν τα *Εθνικά Ινστιτούτα Υγείας (National Institutes of Health, NIH)* των Η.Π.Α., από τους επικεφαλής και πρωτοπόρους στην Ιατρική Έρευνα, συνοψίζονται ως εξής:

Βιοπληροφορική: *Έρευνα, ανάπτυξη, ή εφαρμογή υπολογιστικών εργαλείων και μεθόδων για την επέκταση της χρήσης βιολογικών, ιατρικών δεδομένων ή δεδομένων που σχετίζονται με συμπεριφορά και υγεία, συμπεριλαμβανομένων και αυτών (ενν. εργαλείων και μεθόδων) για την απόκτηση, αποθήκευση, οργάνωση, αρχειοθέτηση, ανάλυση ή οπτικοποίηση τέτοιων δεδομένων.*

Υπολογιστική Βιολογία: *Ανάπτυξη και εφαρμογή θεωρητικών μεθόδων και μεθόδων ανάλυσης δεδομένων, τεχνικών μαθηματικής μοντελοποίησης και υπολογιστικής προσομοίωσης για την μελέτη δεδομένων που εκπορεύονται από βιολογικά, κοινωνικά συστήματα καθώς και συστήματα συμπεριφοράς.* [15]

Με απλά λόγια, η Βιοπληροφορική έχοντας ένα εμφανώς πιο πρακτικό προσανατολισμό ασχολείται με την κατασκευή αλγορίθμων και μεθόδων για την επίλυση πρακτικών και φορμαλιστικών (δηλ. σαφώς ορισμένων) βιολογικών προβλημάτων, ενώ η Υπολογιστική Βιολογία ασχολείται με την κατευθυνόμενη μέσω υποθέσεων έρευνα (hypothesis-driven investigation) ενός συγκεκριμένου βιολογικού προβλήματος με την βοήθεια πειραμάτων και προσομοίωσης.

2.1.1 Περιοχές έρευνας

Οι περιοχές έρευνας των παραπάνω κλάδων περιλαμβάνουν την ανάλυση αλληλουχίας (sequence analysis), επισημείωση γονιδιώματος (genome annotation), υπολογιστική εξελικτική βιολογία (computational evolutionary biology), ανάλυση γονιδιακής έκφρασης (analysis of gene expression), ανάλυση γονιδιακής ρύθμισης (analysis of regulation), ανάλυση πρωτεϊνικής έκφρασης (analysis of protein expression), ανάλυση μεταλλάξεων σε περιπτώσεις καρκίνου (analysis of mutations in cancer), πρόβλεψη πρωτεϊνικής δομής (prediction of protein structure), συγκριτική γονιδιωματική (comparative genomics), μοντελοποίηση βιολογικών συστημάτων (modeling of biological systems) και ανάλυση εικόνας σε ευρεία κλίμακα (high-throughput image analysis).

Αναλυτικότερα, η ανάλυση αλληλουχίας αναφέρεται στην συστοίχιση της πρωτοταγούς δομής του DNA, RNA ή της πρωτεΐνης για τον εντοπισμό λειτουργικών, δομικών ή εξελικτικών σχέσεων μεταξύ διαφορετικών οργανισμών.

Η επισημείωση γονιδιώματος αναφέρεται στην απόδοση λειτουργικών χαρακτηριστικών σε γονίδια, ενώ η υπολογιστική εξελικτική βιολογία στην μελέτη της καταγωγής των ειδών και την εξέλιξή τους στην πάροδο του χρόνου.

Επιπλέον, η ανάλυση γονιδιακής ρύθμισης αναφέρεται στην προσπάθεια αποσαφήνισης των συμβάντων που διέπουν την γονιδιακή ρύθμιση, διαδικασία που περιλαμβάνει την εξωκυτταρική σηματοδότηση και επεκτείνεται μέχρι την μεταβολή της δραστηριότητας των πρωτεϊνικών μορίων. Ένα διάσημο πρόβλημα αυτής της περιοχής αποτελεί ο εντοπισμός ειδικών DNA ακολουθιών, των *προαγωγέων* (*promoters*), που περιβάλλουν το κωδικοποιημένο τμήμα ενός γονιδίου και είναι υπεύθυνες για τον βαθμό μεταγραφής αυτού σε mRNA.

Η ανάλυση μεταλλάξεων σε περιπτώσεις καρκίνου αναφέρεται στον εντοπισμό μεταλλάξεων μίας πληθώρας γονιδίων, ενώ η πρόβλεψη πρωτεϊνικής δομής περιλαμβάνει την εκτίμηση της δευτεροταγούς, τριτοταγούς ή τεταρτοταγούς δομής μίας πρωτεΐνης μέσω της αλληλουχίας των αμινοξέων της (πρωτοταγής δομή).

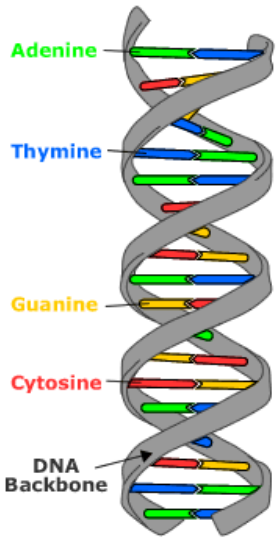
Εξάλλου, η συγκριτική γονιδιωματική ασχολείται με την εύρεση ορθολογίας (ομοιότητας δηλαδή) ανάμεσα σε γονίδια (ή άλλα γονιδιωματικά χαρακτηριστικά) μεταξύ διαφορετικών οργανισμών. Η μοντελοποίηση βιολογικών συστημάτων υλοποιεί την ανάλυση και οπτικοποίηση πολύπλοκων κυτταρικών διαδικασιών όπως είναι τα δίκτυα μεταβολιτών, μονοπάτια μεταγωγής σημάτων και δίκτυα γονιδιακής ρύθμισης, ενώ η ανάλυση εικόνας σε ευρεία κλίμακα χρησιμοποιείται για την επεξεργασία, ποσοτικοποίηση και ανάλυση μεγάλων ποσοτήτων δεδομένων.

Τελευταία και όχι τυχαία αφήσαμε την ανάλυση γονιδιακής και πρωτεϊνικής έκφρασης, που αποδίδουν τα επίπεδα έκφρασης των γονιδίων (ή αντίστοιχα την παρούσα κατάσταση των πρωτεϊνών) κάτω από ορισμένες συνθήκες. Μία διαδεδομένη τεχνική αναπαράστασης τέτοιων δεδομένων είναι μέσω των μικροσυστοιχιών (microarrays), στις οποίες θα αναφερθούμε εκτενέστερα στη συνέχεια. Σε αυτήν τη διπλωματική θέση θα ασχοληθούμε με τη βελτίωση της ακρίβειας (accuracy) ταξινόμησης (classification) ή ομαδοποίησης (clustering) microarray συνόλων.

Πριν όμως αναφερθούμε εκτενέστερα στις μικροσυστοιχίες παρουσιάζει ενδιαφέρον να αναπτύξουμε λίγο το βιολογικό υπόβαθρο πάνω στο οποίο κινούνται οι παραπάνω τεχνικές. Με άλλα λόγια, να αναφερθούμε σε κάποιες βασικές έννοιες της Μοριακής Βιολογίας.

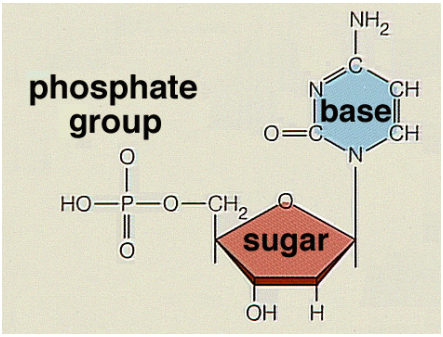
2.2 Δομή του DNA

Τα *κύτταρα* είναι οι θεμελιώδεις λειτουργικές μονάδες κάθε ζωντανού οργανισμού. Όλες οι απαραίτητες πληροφορίες για να δημιουργηθεί και να συντηρηθεί ζωή εσωκλείονται σε ένα μόριο ονόματι *δεοξυριβοζονουκλεϊκό οξύ* (*DNA, DeoxyriboNucleic Acid*). Πιο συγκεκριμένα, το DNA είναι ένα δίκλωνο μόριο πολυμεράσης αποτελούμενο από τέσσερις βασικές μοριακές μονάδες, τα *νουκλεοτίδια* (Εικ. 2.1).

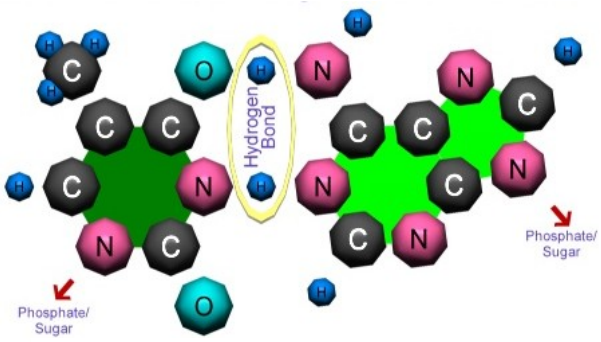


Εικόνα 2.1. Διπλή Έλικα του DNA
Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [xvii]

Κάθε νουκλεοτίδιο αποτελείται από μία ομάδα φωσφόρου, ένα δεοξυριβοϊκό σάκχαρο και μία από τις τέσσερις νιτρικές βάσεις (A: Adenine, T: Thymine, G: Guanine, C: Cytosine) (Εικ. 2.2). Το μόριο του DNA σχηματίζει μία διπλή έλικα όπου το νουκλεοτίδιο ενός κλώνου συνδέεται με δεσμό υδρογόνου με το συμπληρωματικό του νουκλεοτίδιο στον απέναντι κλώνο (Εικ. 2.3). Ως ζεύγη συμπληρωματικών νουκλεοτιδίων θεωρούμε αυτά που φέρουν τις βάσεις A (Αδενίνη) – T (Θυμίνη), G (Γουανίνη) – C (Κυτοσίνη). Οι βάσεις είναι κυκλικές ενώσεις που περιέχουν άζωτο στο δακτύλιό τους. Διακρίνονται σε δύο ομάδες: τις *πουρίνες* και τις *πυριμιδίνες*. Τις πουρίνες αποτελούν η Αδενίνη (A) και η Γουανίνη (G) και τις πυριμιδίνες η Ουρακίλη (U), η Κυτοσίνη (C) και η Θυμίνη (T). Ουσιαστικά, κάθε κλώνος στη διπλή έλικα του DNA μπορεί να ειδωθεί ως χημικό “κατοπτρικό είδωλο” του άλλου.



Εικόνα 2.2. Νουκλεοτίδιο
Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [xxii]



Εικόνα 2.3. Σύνδεση δύο συμπληρωματικών νουκλεοτιδίων με δεσμό υδρογόνου
Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [v]

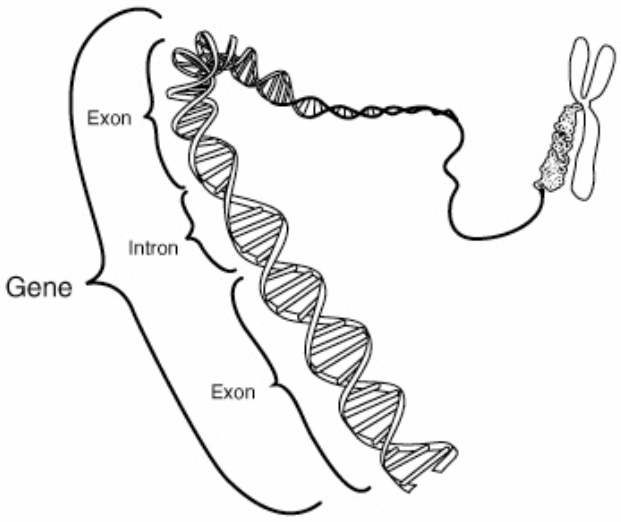
Προαναφέρθηκε, ότι το μόριο του DNA περιέχει όλες τις απαραίτητες πληροφορίες για τη δημιουργία ζωής. Πιο συγκεκριμένα, είναι η συγκεκριμένη αλληλουχία των νουκλεοτιδίων που επιτυγχάνει κάτι τέτοιο, που δίνει δηλαδή όλες τις απαραίτητες οδηγίες για την διαμόρφωση των ιδιαίτερων χαρακτηριστικών κάθε οργανισμού.

Εξάλλου, ολόκληρη η αλληλουχία ενός οργανισμού καλείται το *γονιδίωμά (genome)* του. Η λειτουργικότητα του γονιδιώματος δεν είναι συνεχής αλλά διαιρείται σε “λειτουργικές περιοχές”, τα *γονίδια*.

2.2.1 Γονίδια

Όπως προαναφέραμε τα γονίδια συνιστούν τις “λειτουργικές περιοχές” του γονιδιώματος. Πιο συγκεκριμένα, αποτελούνται από μεγάλους κλώνους DNA (ή RNA σε μερικούς ιούς). Αυτοί οι κλώνοι περιέχουν ακολουθίες ελέγχου της δραστηριότητας τους, τους προαγωγείς (promoters) και μία κωδική ακολουθία που καθορίζει τι παράγει το γονίδιο.

Τα περισσότερα γονίδια περιέχουν μη-κωδικές περιοχές που δεν κωδικοποιούν γονιδιακά προϊόντα αλλά ρυθμίζουν τη γονιδιακή έκφραση. Αυτές οι μη-κωδικές περιοχές ονομάζονται *ιντρόνια (introns)* και απομακρύνονται από το m-RNA κατά το δεύτερο στάδιο της γονιδιακής έκφρασης (μετάφρασης) σε μία διαδικασία που είναι γνωστή ως συρραφή (splicing). Οι περιοχές που κωδικοποιούν την παραγωγή γονιδιακών προϊόντων είναι στην πραγματικότητα πολύ μικρότερες από τα ιντρόνια και είναι γνωστές ως *εξόνια (exons)* (Εικ. 2.4). Όπως θα δούμε και παρακάτω, ένα μόνο γονίδιο μπορεί να οδηγήσει στη σύνθεση πολλών πρωτεϊνών μέσω των διαφορετικών διευθετήσεων των εξονίων, μίας διαδικασίας που είναι γνωστή ως *εναλλακτική συρραφή (alternative splicing)*.



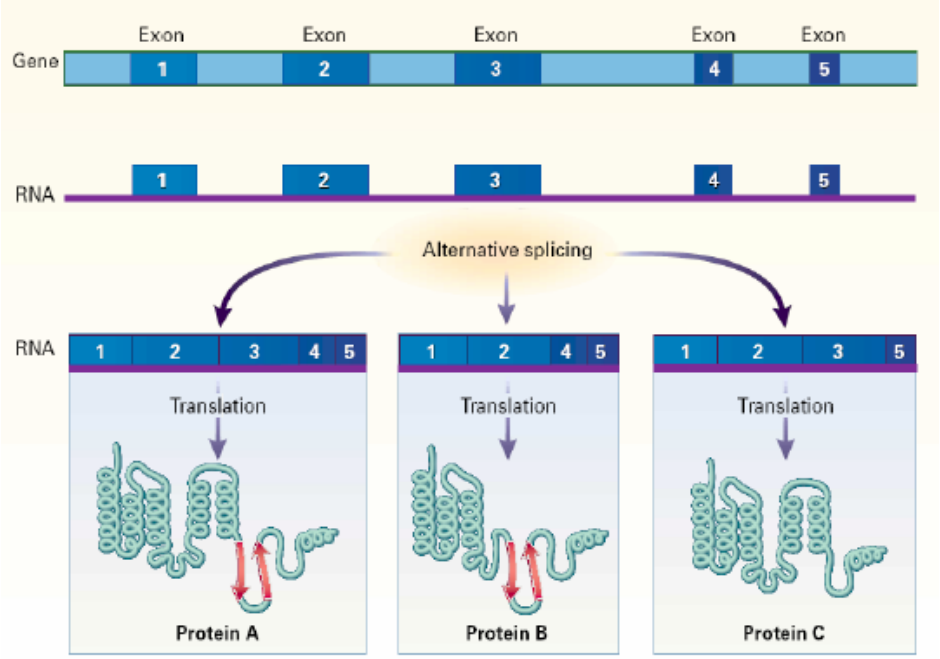
Εικόνα 2.4. Το γονίδιο αποτελείται από ιντρόνια και εξόνια. Τα ιντρόνια απομακρύνονται κατά τη διαδικασία συρραφής (splicing) ενώ τα εξόνια κωδικοποιούν την πρωτεΐνη.
Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [xxv]

Τα γονίδια χωρίζονται σε: *γονίδια καθορισμού RNA (RNA-specifying genes)*, *γονίδια κωδικοποίησης πρωτεϊνών (protein-coding genes)* και *γονίδια που δεν μεταγράφονται (untranscribed genes)*.

Η πρώτη ομάδα γονιδίων συντελεί στην πρώτη φάση της εκτέλεσης της γονιδιακής έκφρασης (gene expression), που είναι γνωστή ως *μεταγραφή (transcription)*. Μ' άλλα λόγια, αποτελεί κλώνο για την κατασκευή μορίων RNA, όπως είναι τα: r-RNA (δομικό και λειτουργικό στοιχείο του ριβοσώματος), t-RNA (προσαρμογέας μετάφρασης), mi-RNA (καταστολέας μετάφρασης), sn-RNA (συσκευή σύνδεσης RNA), sno-RNA (προεπεξεργασία r-RNA), large nc-RNA (γονιδιακή ρύθμιση) και m-RNA.

Η δεύτερη ομάδα γονιδίων κωδικοποιεί την κατασκευή πρωτεϊνών συμβάλλοντας έτσι στη δεύτερη και τελική φάση της γονιδιακής έκφρασης που είναι γνωστή ως *μετάφραση (translation)*, ενώ η τελευταία ουσιαστικά αποτελείται από περιοχές του γονιδιώματος που έχουν κάποια βιολογική λειτουργία, η οποία όμως δεν σχετίζεται με τη μεταγραφή ή τη μετάφραση. [14]

Στον άνθρωπο υπάρχουν περίπου 20000 γονίδια της δεύτερης ομάδας, γονίδια δηλαδή που κωδικοποιούν την παραγωγή πρωτεϊνών. Ο ανθρώπινος οργανισμός αποτελείται επίσης από 10 τρισεκατομμύρια (10^{13}) κύτταρα που κατανέμονται σε 210 περίπου είδη. Μερικά από αυτά τα είδη είναι τα επιθηλιακά κύτταρα, τα κύτταρα του ανοσοποιητικού συστήματος και τα νευρικά κύτταρα. Είναι όμως αξιοσημείωτο πως τα 20000 γονίδια μπορούν να καθορίσουν τη δομή ενός πολύπλοκου οργανισμού όπως αυτόν των θηλαστικών. Οι κύριοι λόγοι που το επιτρέπουν αυτό είναι η *καθορισμένη από το είδος του κυττάρου γονιδιακή έκφραση (cell-type specific gene expression)*, η *καθορισμένη από τον ιστό γονιδιακή μεταγραφή (tissue specific gene transcription)* και η *εναλλακτική συρραφή εξονίων (alternative splicing)* (Εικ. 2.5).

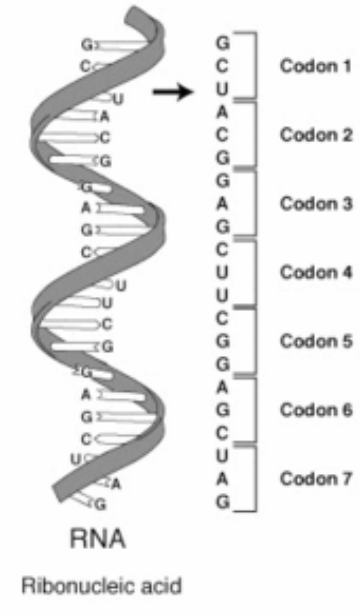


Εικόνα 2.5. Ένα γονίδιο μπορεί να παράγει πολλές διαφορετικές πρωτεΐνες μέσω της διαφορετικής συρραφής των εξονίων του.
Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [xviii]

2.2.2 Γονιδιακή έκφραση

Ως γνωστόν, οι πρωτεΐνες δεν είναι απλώς οι δομικοί λίθοι των κυττάρων αλλά επιτελούν και όλες τις κυτταρικές λειτουργίες. Η διαδικασία παραγωγής τους είναι η εξής:

Το κύτταρο μεταγράφει τη γενετική του ακολουθία σε *αγγελιαφόρο RNA (messenger RNA – mRNA)*. Στη συνέχεια, το mRNA μεταφράζεται σε τριπλέτες δομώντας μία ακολουθία αμινοξέων, των δομικών συστατικών της πρωτεΐνης. Πιο συγκεκριμένα, το *κωδικόνιο (codon)*, όπως ονομάζεται κάθε τριπλέτα νουκλεοτιδίων κωδικοποιεί τη σύνθεση ενός αμινοξέος σύμφωνα με τον τυποποιημένο γενετικό κώδικα (standard genetic code) που είναι σχεδόν κοινός σε όλους τους οργανισμούς. Μία οπτική επεξήγηση των παραπάνω παρέχεται στην Εικ. 2.6.



Εικόνα 2.6. Κάθε κωδικόνιο αποτελείται από μία τριάδα νουκλεοτιδίων που αντιστοιχεί σε ένα αμινοξύ.
Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [ii]

Ουσιαστικά, η τριπλέτα νουκλεοτιδίων προτάθηκε λόγω της ύπαρξης 20 διαφορετικών αμινοξέων στους ζώντες οργανισμούς. Έτσι, το 3 είναι ο μικρότερος δυνατός αριθμός ώστε $4^3 \geq 20$. Αναλυτικά, το $4^3 = 64$ είναι όλες οι διαφορετικές τριάδες των τεσσάρων νουκλεοτιδίων που προκύπτουν με αντικατάσταση. Παρόλα αυτά, διαπιστώνεται ότι υπάρχει ένα κενό μεταξύ του αριθμού των διαφορετικών τριάδων και του αριθμού των διαφορετικών αμινοξέων. Μ' άλλα λόγια, ο γενετικός κώδικας είναι υπεράριθμος. Τα δύο πρώτα νουκλεοτίδια κάθε κωδικονίου είναι σταθερά και χαρακτηρίζουν την κωδικοποίηση ενός μόνο αμινοξέος. Το τρίτο νουκλεοτίδιο ενός κωδικονίου μπορεί να μεταβάλλεται και παρόλα αυτά να συνεχίζει να κωδικοποιεί τη σύνθεση του ίδιου αμινοξέος. Ένας συνοπτικός πίνακας που αντιστοιχεί τα 20 αμινοξέα με τις 64 τριάδες νουκλεοτιδίων παρουσιάζεται παρακάτω.

Πίνακας Ι. Πίνακας αντιστοίχησης αμινοξέων–κωδικονίων.

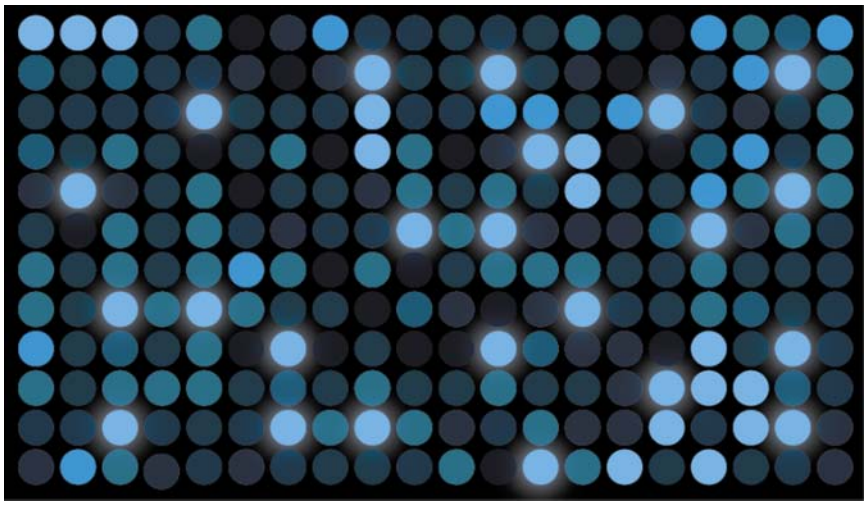
Αμινοξύ	**Κωδικόνια**	**Αμινοξύ**	**Κωδικόνια**
Ala	GCU, GCC, GCA, GCG	Leu	UUA, UUG, CUU, CUC, CUA, CUG
Arg	CGU, CGC, CGA, CGG, AGA, AGG	Lys	AAA, AAG
Asn	AAU, AAC	Met	AUG
Asp	GAU, GAC	Phe	UUU, UUC
Cys	UGU, UGC	Pro	CCU, CCC, CCA, CCG
Gln	CAA, CAG	Ser	UCU, UCC, UCA, UCG, AGU, AGC
Glu	GAA, GAG	Thr	ACU, ACC, ACA, ACG
Gly	GGU, GGC, GGA, GGG	Trp	UGG
His	CAU, CAC	Tyr	UAU, UAC
Ile	AUU, AUC, AUA	Val	GUU, GUC, GUA, GUG
START	AUG	STOP	UAG, UGA, UAA

Η διαδικασία όμως που καθορίζει το επίπεδο παραγωγής πρωτεϊνών από ένα γονίδιο ονομάζεται, όπως προαναφέρθηκε, *γονιδιακή έκφραση* [10]. Τα επίπεδα της γονιδιακής έκφρασης (ενός γονιδίου) καταδεικνύουν τον κατά προσέγγιση αριθμό

αντιγράφων RNA (αυτού του γονιδίου) και σχετίζονται με την ποσότητα παραγωγής των αντίστοιχων πρωτεϊνών (που κωδικοποιεί αυτό το γονίδιο). Έτσι, η έκφραση ενός γονιδίου παρέχει ένα μέτρο της δραστηριότητας αυτού του γονιδίου κάτω από συγκεκριμένες βιοχημικές συνθήκες [14]. Με αυτό τον τρόπο μπορούμε να παρακολουθήσουμε την επίδραση ενός γονιδίου σε μία συγκεκριμένη βιοχημική διαδικασία εξετάζοντας την κατανομή των επιπέδων της έκφρασής του.

2.3 Μικροσυστοιχίες

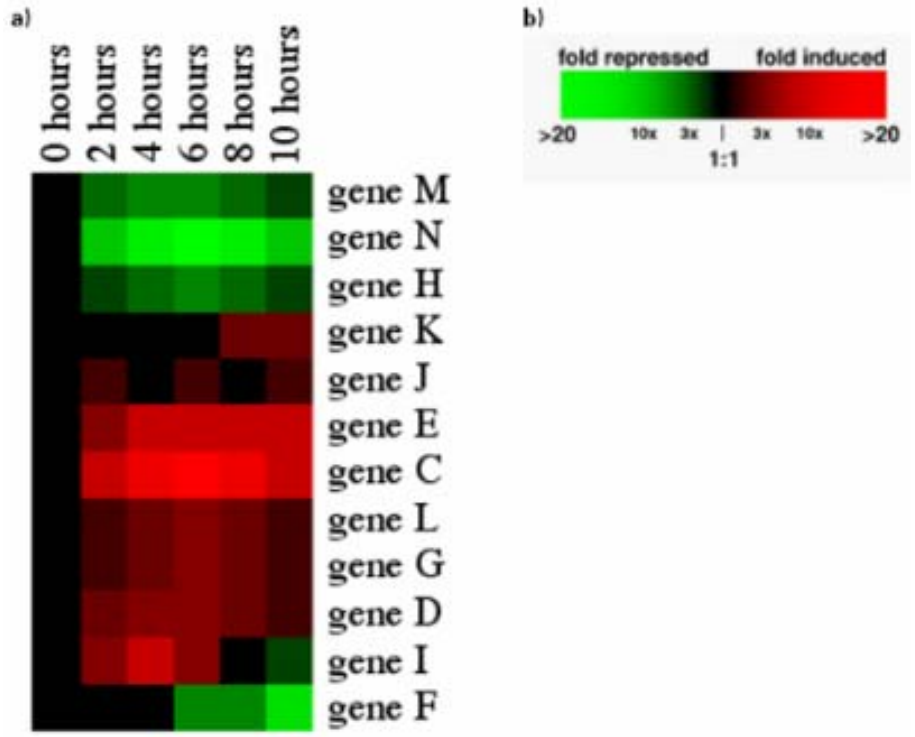
Αυτήν ακριβώς την πληροφορία μας δίνουν οι μικροσυστοιχίες. Αναλυτικότερα, οι μικροσυστοιχίες παρέχουν τη δυνατότητα απεικόνισης της έκφρασης χιλιάδων ή δεκάδων χιλιάδων γονιδίων σε δεκάδες ή εκατοντάδες δείγματα. Οι διατάξεις αυτές μπορούν να υποστούν ταξινόμηση, ομαδοποίηση, εκτίμηση πυκνότητας (Εικ. 2.7). Για παράδειγμα, μετρώντας επίπεδα έκφρασης σχετιζόμενα με δύο τύπους ιστών (κανονικός ή όγκος) δημιουργείται ένα σύνολο δεδομένων με ετικέτες που μπορεί να χρησιμοποιηθεί για διαγνωστική ταξινόμηση. Αναλυτικότερα, τα γονίδια συνήθως βρίσκονται σε δύο διακριτές βιολογικές καταστάσεις: καταστολή (repression, off) ή ενεργοποίηση (induction, on). Έτσι, κάθε γονίδιο παρουσιάζει διαφορετικά επίπεδα έκφρασης κατά μήκος παραδειγμάτων που δεν ανήκουν στο ίδιο είδος ιστών (κανονικός ή όγκος) [28].



Εικόνα 2.7. Μικροσυστοιχία

Ουσιαστικά, τα δεδομένα παρουσιάζουν μία χρωματική διαβάθμιση και καταγράφονται οι φωτεινότητες κάθε γονιδίου-γνωρίσματος κατά μήκος διαφόρων σταδίων (παραδειγμάτων) και εξάγονται (από αυτές) αριθμητικές τιμές. Ένα σύνηθες σχήμα κωδικοποίησης είναι η ένταση του πράσινου χρώματος να αντιπροσωπεύει ανάλογο επίπεδο καταστολής (repression) του συγκεκριμένου γονιδίου-γνωρίσματος και η ένταση του κόκκινου χρώματος ανάλογο επίπεδο ενεργοποίησης (induction). Μ' άλλα λόγια, το εντονότερο πράσινο αντιστοιχεί σε όλο και λιγότερη ποσότητα m-RNA που προέκυψε από τη μεταγραφή του γονιδίου για το συγκεκριμένο στάδιο (παράδειγμα), ενώ το εντονότερο κόκκινο αντιστοιχεί σε όλο και μεγαλύτερη

ποσότητα m-RNA αυτού του γονιδίου. Από την άλλη, το μαύρο χρώμα υποδηλώνει ανυπαρξία αλλαγής. Μία οπτικοποίηση των παραπάνω παρέχεται στην Εικ. 2.8.

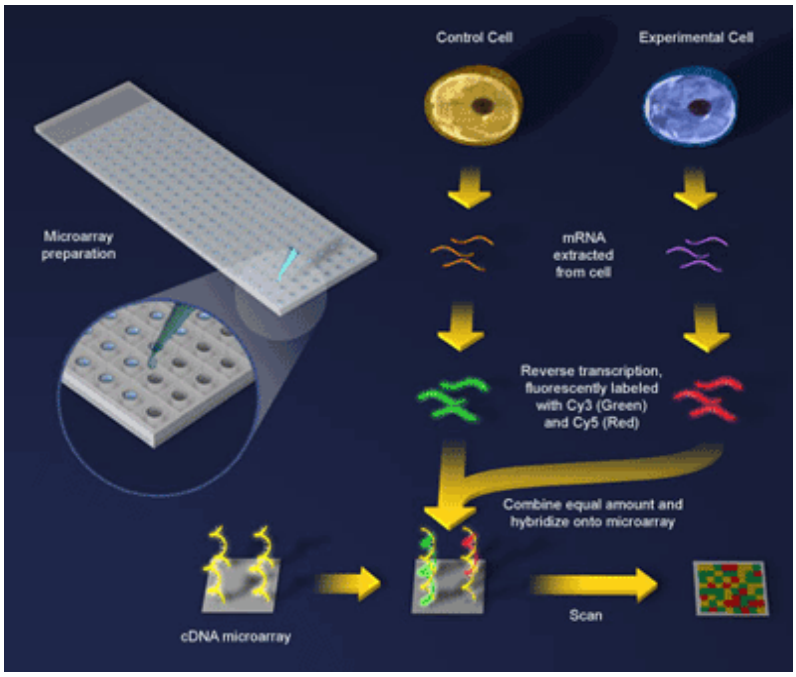


Εικόνα 2.8. Χρωματική διαβάθμιση μικροσυστοιχιών και επεξήγηση του χρωματικού κώδικα.
a) Μία μικροσυστοιχία με την παραπάνω χρωματική διαβάθμιση που παρουσιάζει την έκφραση γονιδίων κατά μήκος διαφορετικών χρονικών στιγμών
b) Χρωματικός κώδικας: Το πράσινο αντιπροσωπεύει την καταστολή του γονιδίου, ενώ το κόκκινο την ενεργοποίηση του.
Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [iv]

Αναφερθήκαμε προηγουμένως στη δυνατότητα διαγνωστικής ταξινόμησης που επιτρέπει η χρήση των μικροσυστοιχιών. Πριν όμως επεκταθούμε περισσότερο στη χρησιμότητα τους και τις μεθόδους εξόρυξης γνώσης από τέτοια σύνολα δεδομένων είναι καλό να αναφέρουμε συνοπτικά τη διαδικασία κατασκευής αυτών των διατάξεων.

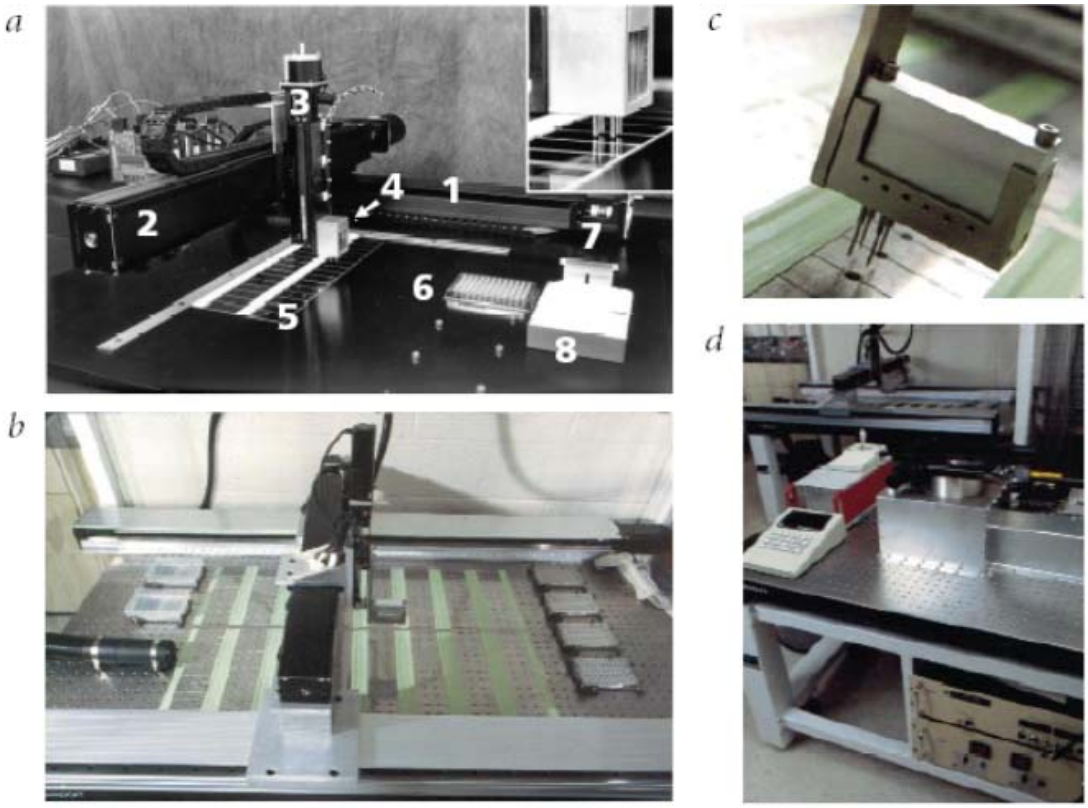
2.3.1 Διαδικασία κατασκευής μικροσυστοιχιών

Κατ' αρχήν, ένας DNA πίνακας αποτελείται από DNA ακολουθίες σε ακροδέκτες (probe DNA sequences) που ακινητοποιούνται πάνω σε μία επιφάνεια γυαλιού ή χρυσού. Στη συνέχεια, εξάγονται οι αντίστοιχες mRNA ακολουθίες από αυτές. Αυτές, στην πορεία, ενισχύονται (αντιγράφονται σε εκθετικό ρυθμό) και αναστροφο-μεταγράφονται σε DNA ακολουθίες. Οι αναστροφο-μεταγραφόμενες DNA ακολουθίες σηματοδοτούνται με φθόριο. Οι DNA ακολουθίες στους ακροδέκτες του πίνακα έχουν σχεδιαστεί έτσι ώστε να υβριδοποιούνται με αυτές τις ανάστροφες μεταγραφές (χάρη στη δυνατότητα της DNA – DNA υβριδοποίησης). Μετά τη DNA – DNA υβριδοποίηση, ο DNA πίνακας σαρώνεται για να ποσοτικοποιήσει το φθοριούχο φως που εκπέμπεται από κάθε ακροδέκτη (Εικ. 2.9).



Εικόνα 2.9. Διαδικασία παραγωγής ενός πειράματος μικροσυστοιχίας
Η παραπάνω εικόνα έχει από την τοποθεσία [iii]

Παρακάτω φαίνεται μία πραγματική διαδικασία κατασκευής μίας μικροσυστοιχίας καθώς και οι συσκευές που χρησιμοποιούνται σε αυτήν:



Εικόνα 2.10. Διαδικασία κατασκευής μίας μικροσυστοιχίας.

a) Penn microarray ρομπότ. Οι X-, Y-, Z- άξονες σηματοδοτούνται ως 1, 2 και 3 αντίστοιχα. Το
στοιχείο- κλειδί της διάταξης είναι οι κεφαλές εκτύπωσης (4). Αντικειμενοφόρες πλάκες γυαλιού μικροσκοπίων τοποθετούνται στην περιοχή αντικειμενοφόρας πλάκας (5). Τα δείγματα προετοιμάζονται και τοποθετούνται σε πινακωτή διάταξη σε ειδικές πλάκες (6). Οι ακίδες καθαρίζονται στα διαστήματα μεταξύ αποκτήσεων των δειγμάτων που λαμβάνουν χώρα στους σταθμούς πλυσίματος (7) και στεγνώματος (8).

b) AECOM microarray ρομπότ. Η πινακωτή διάταξη που παρουσιάζεται περιλαμβάνει 160 επιφάνειες με τέσσερις πλάκες, δύο σταθμούς πλυσίματος και ένα στεγνωτήρα.

c) Η κεφαλή εκτύπωσης δείχνει τέσσερις από τις πιθανές δώδεκα μύτες στυλό σε χρήση.

d) AECOM laser scanner. Στην εικόνα, είναι ορατά τα: οπτικό τραπέζι, τροφοδοσία ρεύματος για τα lasers και ψύξη PMT, το στάδιο Ludl, και lasers. Ο 20x αντικειμενικός φακός του μικροσκοπίου βρίσκεται εντός του ludl σταδίου, ενώ φακοί, καθρέφτες και άλλα οπτικά όργανα εσωκλείονται στο μεταλλικό περίβλημα. Οι PMTs είναι στα δεξιά και έξω από τη φωτογραφία.

Η παραπάνω εικόνα έχει από την τοποθεσία [iv]

Γενικά, υπάρχουν δύο κατηγορίες για την κατασκευή DNA πινάκων: 1) χρησιμοποιώντας ολόκληρη την ακολουθία κάθε γονιδίου (οι αντίστοιχοι πίνακες ονομάζονται cDNA πίνακες) και 2) χρησιμοποιώντας ένα μικρό μέρος βραχέων τμημάτων mRNA ακολουθίας που καλύπτει με μοναδικό τρόπο την mRNA ακολουθία (αυτοί οι πίνακες ονομάζονται DNA microarrays) [10].

2.3.2 Μειονεκτήματα από τη χρήση μικροσυστοιχιών

Η χρήση των μικροσυστοιχιών παρόλα τα πλεονεκτήματα που προσφέρει θέτει ένα σύνολο από εμπόδια που οφείλονται στην ίδια τη φύση του συνόλου δεδομένων:

i) Κατ' αρχήν, η ενίσχυση mRNA ενός κυττάρου είναι μία εξαιρετικά δύσκολη διαδικασία με την υπάρχουσα τεχνολογία. Έτσι, ιστοί που φαίνονται να έχουν κοινές λειτουργίες ουσιαστικά εντοπίζονται λόγω της επαρκούς ποσότητας mRNA που φέρουν. Αυτό έχει ως αποτέλεσμα, τα επίπεδα έκφρασης να υπολογίζονται ως ο μέσος όρος όλων των κυττάρων του πειράματος.

ii) Επίσης, η γενετική ποικιλομορφία επηρεάζει τη γονιδιακή έκφραση, για παράδειγμα τα επίπεδα έκφρασης δύο ατόμων μπορεί να είναι διαφορετικά [10].

iii) Τρίτον, επικρατεί αυτό που αποκαλείται *"κατάρα της διαστατικότητας" ("curse of dimensionality")*[1], τουτέστιν, υπάρχει μεγάλη διαφορά ανάμεσα στο πλήθος των γνωρισμάτων (attributes) και το πλήθος των δειγμάτων (instances). Πιο συγκεκριμένα, ο αριθμός των γνωρισμάτων είναι χαρακτηριστικά κατά πολύ μεγαλύτερος από τον αριθμό των δειγμάτων. Αυτό έχει ως αποτέλεσμα την κακή απόδοση των αλγορίθμων ταξινόμησης στατιστικής και μηχανικής εκμάθησης αφού δεν έχουν σχεδιαστεί για τέτοιο είδος δεδομένων. Επίσης, η πολύ μεγάλη διάσταση του χώρου γνωρισμάτων σε συνδυασμό με τα ελάχιστα δείγματα χειροτερεύει ακόμα περισσότερο την κατάσταση οδηγώντας στο φαινόμενο της *υπερεκπαίδευσης (overfitting)*. Μ' άλλα λόγια, το μοντέλο εκμάθησης που εξάγεται προσδιορίζει σχεδόν αποκλειστικά το συγκεκριμένο σύνολο δεδομένων πάνω στο οποίο εκπαιδεύτηκε χωρίς έτσι να καταστεί δυνατή η γενίκευση ασφαλών συμπερασμάτων. Επίσης, ένα άλλο ζήτημα που εγείρεται από το μέγεθος του συνόλου γνωρισμάτων είναι αυτό του χρόνου υπολογισμού (computation time)· ο χρόνος εκπαίδευσης και η διαδικασία εκμάθησης έχει άμεση εξάρτηση από τον αριθμό των γνωρισμάτων και η

[1] Για την ιστορία, αξίζει να αναφέρουμε ότι ο όρος αυτός επινοήθηκε το 1961 από το μαθηματικό Richard Ernest Bellman (1920 – 1984) στο βιβλίο του "Adaptive Control Processes" (Princeton University Press, Princeton, NJ) για να αποδώσει με αυτόν τον τρόπο τα προβλήματα που δημιουργούνται από την προσθήκη επιπλέον διαστάσεων στο (μαθηματικό) χώρο. [Cresswell Clio. "*Mathematics and Sex*". Allen and Unwin, 2003, Australia]

διαδικασία γίνεται πιο χρονοβόρα με ολοένα αυξανόμενο ρυθμό καθώς ο αριθμός αυτός αυξάνει.

iv) Τέταρτον, η ύπαρξη θορύβου στο σύνολο δεδομένων, ο οποίος μπορεί να είναι είτε βιολογικός είτε τεχνικός παραποιεί τα αποτελέσματα. Βιολογικός εξαιτίας της ύπαρξης γονιδίων – γνωρισμάτων που δεν είναι σχετικά με τον καθορισμό κλάσεων του υπάρχοντος συνόλου δεδομένων. Τεχνικός γιατί συμβαίνουν κάποιες ατέλειες ή παραβλέψεις κατά την προετοιμασία των δεδομένων, δηλαδή κατά το σχεδιασμό και την εκτέλεση του microarray πειράματος.

v) Πέμπτον, η αναλογία των άσχετων γονιδίων σε αυτό το είδος συνόλου δεδομένων είναι πολύ μεγαλύτερη από τη συνήθη. Ενώ στα περισσότερα σύνολα δεδομένων γονιδιακής έκφρασης ο αριθμός των σχετικών προς το πρόβλημα γνωρισμάτων είναι πολύ μικρός. Τα παραπάνω δυσχεραίνουν εξαιρετικά τη διακριτική ισχύ των σχετικών ως προς το σύνολο δεδομένων γονιδίων.

vi) Τέλος, έχει μεγάλη σημασία η εξαγωγή συμπερασμάτων να εμφανίζει βιολογική σημασία, αφού κάθε πληροφορία που εξορύσσεται κατά τη διάρκεια της διαδικασίας μπορεί να βοηθήσει σε περαιτέρω ανακάλυψη γονιδιακών λειτουργιών καθώς και σε άλλες βιολογικές μελέτες. Για παράδειγμα, ιδιαίτερα χρήσιμες πληροφορίες από τη διαδικασία ταξινόμησης θα μπορούσαν να είναι η αποκάλυψη γονιδίων που λειτουργούν ομαδικά ή ο εντοπισμός υπο-εκφρασμένων ή υπερ-εκφρασμένων γονιδίων σε συγκεκριμένους ιστούς ή κύτταρα. Έτσι, οι βιολόγοι θα μπορέσουν να αποκτήσουν βαθύτερη κατανόηση των γονιδίων, της λειτουργίας τους καθώς και του τρόπου διάδρασης μεταξύ τους [14].

Όπως προαναφέρθηκε, σημαντικό κριτήριο για την παραγωγή χρήσιμων συμπερασμάτων είναι η σχέση τους με το συγκεκριμένο είδος προβλήματος που εξετάζεται (εν προκειμένω, βιολογικό). Προηγουμένως, έγινε εμμέσως αναφορά και στη χαμηλή απόδοση που παρουσιάζουν οι υπάρχοντες αλγόριθμοι ταξινόμησης για τα σύνολα δεδομένων αυτού του είδους. Ένας συνήθης τρόπος βελτίωσης των αποτελεσμάτων είναι η προεπεξεργασία του συνόλου δεδομένων μέσω της εφαρμογής μεθόδων επιλογής γνωρισμάτων (feature selection methods). Για τις παραπάνω μεθόδους θα αναφερθούμε εκτενέστερα παρακάτω.

Σε πρώτη φάση, όμως, να αναφέρουμε ότι παρόλη την επιτυχή εφαρμογή τους σε προβλήματα πολλών πεδίων το ισχυρότερο μειονέκτημά τους είναι η αποκλειστική

στήριξή τους στα εσωτερικά χαρακτηριστικά του συγκεκριμένου συνόλου δεδομένων. Μ' άλλα λόγια, γίνεται προσπάθεια εντοπισμού των πιο χρήσιμων πληροφοριακά γνωρισμάτων μέσω στατιστικής ανάλυσης των τιμών τους για τα δείγματα του συγκεκριμένου προβλήματος. Αυτό όμως κρύβει δύο αδυναμίες: πρώτον δεν λαμβάνει υπόψη την πιθανώς εσφαλμένη απόδοση τιμών σε γνωρίσματα λόγω ατελειών διεξαγωγής του πειράματος (outliers) και αγνοεί την ήδη υπάρχουσα γνώση για το συγκεκριμένο είδος προβλήματος, εν προκειμένω, δηλαδή τη βιολογική γνώση.

3 Επεξήγηση των εννοιών της Οντολογίας και της Γονιδιακής Οντολογίας

3.1 Ορισμός της Οντολογίας

Την τελευταία ακριβώς αδυναμία, την έλλειψη δηλαδή εκμετάλλευσης της υπάρχουσας βιολογικής γνώσης προσπαθήσαμε να καλύψουμε μέσω αυτής της διπλωματικής εργασίας. Αυτό κατέστη δυνατόν με την ενσωμάτωση πληροφορίας από μία από τις πιο εγκεκριμένες Γονιδιακές Οντολογίες (The Gene Ontology), ενός δομημένου και ιεραρχημένου δηλαδή λεξιλογίου βιολογικών όρων. Κρίνεται όμως σκόπιμο να κάνουμε μία σύντομη αναφορά στις οντολογίες γενικότερα και στις ιδιότητες που εμφανίζουν.

Ο όρος *Οντολογία* εμφανίζεται από τα αρχαία χρόνια. Αρχικά, επινοήθηκε από τους φιλοσόφους, κυρίως της Πλατωνικής σχολής. Σαν κλάδος της μεταφυσικής, είχε ως προεξέχων αντικείμενο έρευνας το “τι πραγματικά υπάρχει”. Συνεπώς, μελετούσε τα όντα, τις κατηγορίες που αυτά ανήκαν και τις σχέσεις μεταξύ τους. Παρόλα αυτά, όσον αφορά την επιστήμη υπολογιστών έχει δοθεί ένας σαφώς πιο συγκεκριμένος ορισμός. Εδώ, η οντολογία αποτελεί μία συστηματική διάταξη όλων των σημαντικών κατηγοριών των αντικειμένων που ανήκουν σε ένα πεδίο έρευνας όπου παράλληλα καταδεικνύονται και οι σχέσεις μεταξύ τους (ενν. των αντικειμένων).

Η *Πρωτοβουλία Ψηφιακών Βιβλιοθηκών (Digital Libraries Initiative)* του Πανεπιστημίου του Illinois στο Urbana-Champaign ορίζει την οντολογία [xxiv] ως:

“*Ένας σαφής, επίσημος προσδιορισμός του τρόπου αναπαράστασης αντικειμένων, εννοιών και γενικά υποστάσεων που υποτίθεται ότι υπάρχουν σε μία περιοχή ενδιαφέροντος καθώς και των σχέσεων μεταξύ τους.*”

Η οντολογία μπορεί να είναι ένας Κατευθυνόμενος Άκυκλος Γράφος (DAG, Directed Acyclic Graph) με τους κόμβους να αναπαριστούν τα αντικείμενα (και τα εσωτερικά χαρακτηριστικά τους) και τις ακμές τις σχέσεις μεταξύ τους. Επιλέγεται αυτό το σχήμα γιατί αφήνει περιθώριο πολλαπλής συγγένειας των αντικειμένων

μεταξύ τους. Για παράδειγμα, στους γράφους επιτρέπεται η πολλαπλή γονεϊκότητα (ένα παιδί μπορεί να έχει πολλούς γονείς). Μία απλοποιημένη μορφή της οντολογίας μπορεί να περιλαμβάνει μία ιεραρχημένη ταξινόμηση που είναι ουσιαστικά μία δενδρική δομή και καλείται *ταξονομία*. Με άλλα λόγια, η οντολογία είναι ένα καλά δομημένο και ορισμένο λεξιλόγιο που αναπαριστά μία περιοχή έρευνας, τα αντικείμενά της και τις σχέσεις μεταξύ τους.

3.2 Βιοϊατρικές Οντολογίες

Δομές οντολογίας χρησιμοποιούνται ευρέως και στον χώρο της Βιοϊατρικής. Υπάρχουν πολλά δομημένα λεξιλόγια που είναι γνωστά ως *Ανοιχτές Βιοϊατρικές Οντολογίες (Open Biomedical Ontologies, OBO)* και το καθένα εξυπηρετεί συγκεκριμένο βιολογικό και ιατρικό τομέα. Υπάρχουν επίσης διαφορετικοί τύποι σχέσεων ανάμεσα στους όρους μίας οντολογίας. Μπορεί να έχουμε *θεμελιώδεις σχέσεις (foundational relations)*, σχέσεις δηλαδή που παίζουν κεντρικό ρόλο σχεδόν σε όλες τις οντολογίες, *χωρικές (spatial relations)* και *χρονικές σχέσεις (temporal relations)* καθώς και *σχέσεις συμμετοχής (participation relations)* (Πιν. ΙΙ). Επιπλέον, κάθε σχέση φέρει συγκεκριμένες ιδιότητες (Πιν. ΙΙΙ).

Πίνακας ΙΙ. Παρουσίαση των σχέσεων των βιοϊατρικών οντολογιών

Πρώτη Έκδοση της OBO Οντολογίας
Θεμελιώδεις σχέσεις
is-a *part-of*
Χωρικές σχέσεις (συνδέουν το ένα αντικείμενο με το άλλο στα πλαίσια των σχέσεων μεταξύ των χωρικών περιοχών που καταλαμβάνουν)
located_in *contained_in* *adjacent_to*
Χρονικές σχέσεις (συνδέουν αντικείμενα που συμβαίνουν σε διαφορετικές χρονικές στιγμές)
transformation_of *derives_from* *preceded_by*
Σχέσεις συμμετοχής (συνδέουν διαδικασίες με τους φέροντες (αυτών των διαδικασιών))
has_participant *has_agent*

Πίνακας III. Μερικές ιδιότητες των σχέσεων της OBO Οντολογίας

Σχέση	Μεταβατική	Συμμετρική	Ανακλαστική	Αντισυμμετρική
is-a	+	-	+	+
part-of	+	-	+	+
located_in	+	-	+	-
contained_in	-	-	-	-
adjacent_to	-	-	-	-
transformation_of	-	-	-	-
derives_ from	+	-	-	-
preceded_by	+	-	-	-
has_participant	-	-	-	-
has_agent	-	-	-	-

Η βιοϊατρική οντολογία που ασχολείται με την επισημείωση γονιδίων ή γονιδιακών προϊόντων είναι γνωστή ως Γονιδιακή Οντολογία (The Gene Ontology, GO). Η Γονιδιακή Οντολογία φέρει μόνο τις θεμελιώδεις σχέσεις “is-a” και ”part-of” οι οποίες είναι μεταβατικές και αντισυμμετρικές, ιδιότητες που επιτρέπουν τη συνέχεια της σύνδεσης και την ύπαρξη συνδέσεων μόνο προς μία κατεύθυνση, αντίστοιχα [21]. Αναλυτικότερα, για τη Γονιδιακή Οντολογία θα μιλήσουμε παρακάτω.

3.3 Γονιδιακή Οντολογία

Η διαρκώς αυξανόμενη αποκρυπτογράφηση μοριακών ακολουθιών, ιδιαίτερα ακολουθιών ολόκληρων γονιδιωμάτων, άλλαξε την οπτική υπό την οποία εξετάζεται η θεωρία και η πράξη στην πειραματική βιολογία. Παλαιότερα, οι βιοχημικοί συνήθιζαν να χαρακτηρίζουν τις πρωτεΐνες από τις διαφορές στις δραστηριότητές τους και οι γενετιστές τα γονίδια από το φαινότυπο των μεταλλάξεων τους. Τώρα πια έχει γίνει κοινά αποδεκτό ότι υπάρχουν χαρακτηριστικές ομάδες γονιδίων και πρωτεϊνών που έχουν διατηρηθεί (για τη χρησιμότητά τους) σε κύτταρα κάθε είδους οργανισμού. Αυτή η αποδοχή έχει ωθήσει στην ενοποίηση της βιολογίας. Μ’ άλλα λόγια, η γνώση του βιολογικού ρόλου μίας πρωτεΐνης σε ένα οργανισμό που εμφανίζεται να έχει κοινά στοιχεία με αντίστοιχες πρωτεΐνες σε άλλους οργανισμούς μπορεί να διαφωτίσει το ρόλο των αντίστοιχων γονιδιακών προϊόντων και στους άλλους οργανισμούς.

Η περιγραφή και ο ορισμός τέτοιων κοινών βιολογικών στοιχείων, τουτέστιν η απόδοση ονοματολογίας και λεπτομερούς χαρακτηρισμού των γονιδίων και των προϊόντων (καθώς και των ιδιοτήτων τους) δεν ακολουθεί κάποιο κοινό πρότυπο με αποτέλεσμα την έλλειψη ενός ενιαίου μοντέλου που θα συγκέντρωνε όλα τα

απαραίτητα στοιχεία για το χαρακτηρισμό γονιδίων (και των προϊόντων τους) υπό μία κοινώς αποδεκτή και τυποποιημένη φόρμα.

Η λύση, η ανάληψη δηλαδή πρωτοβουλίας για την ενοποίηση και σύνδεση όλων των γονιδιωματικών βάσεων δεδομένων αποτέλεσε την κυριότερη αιτία για τη δημιουργία της Γονιδιακής Οντολογίας [24].

Το έργο της Γονιδιακής Οντολογίας ξεκίνησε στα πλαίσια της συνεργασίας που αναπτύχθηκε μεταξύ των ερευνητικών ομάδων τριών οργανισμών της FlyBase (Drosophila), της Saccharomyces Genome Database (SGD) and της Mouse Genome Database (MGD) το 1998. Από τότε στην κοινοπραξία της Γονιδιακής Οντολογίας έχουν εισχωρήσει πολλές άλλες βάσεις δεδομένων, εκ των οποίων και μερικές από τις πιο σημαντικές πηγές πληροφορίας φυτικών, ζωικών και μικροβιακών γονιδιωμάτων.

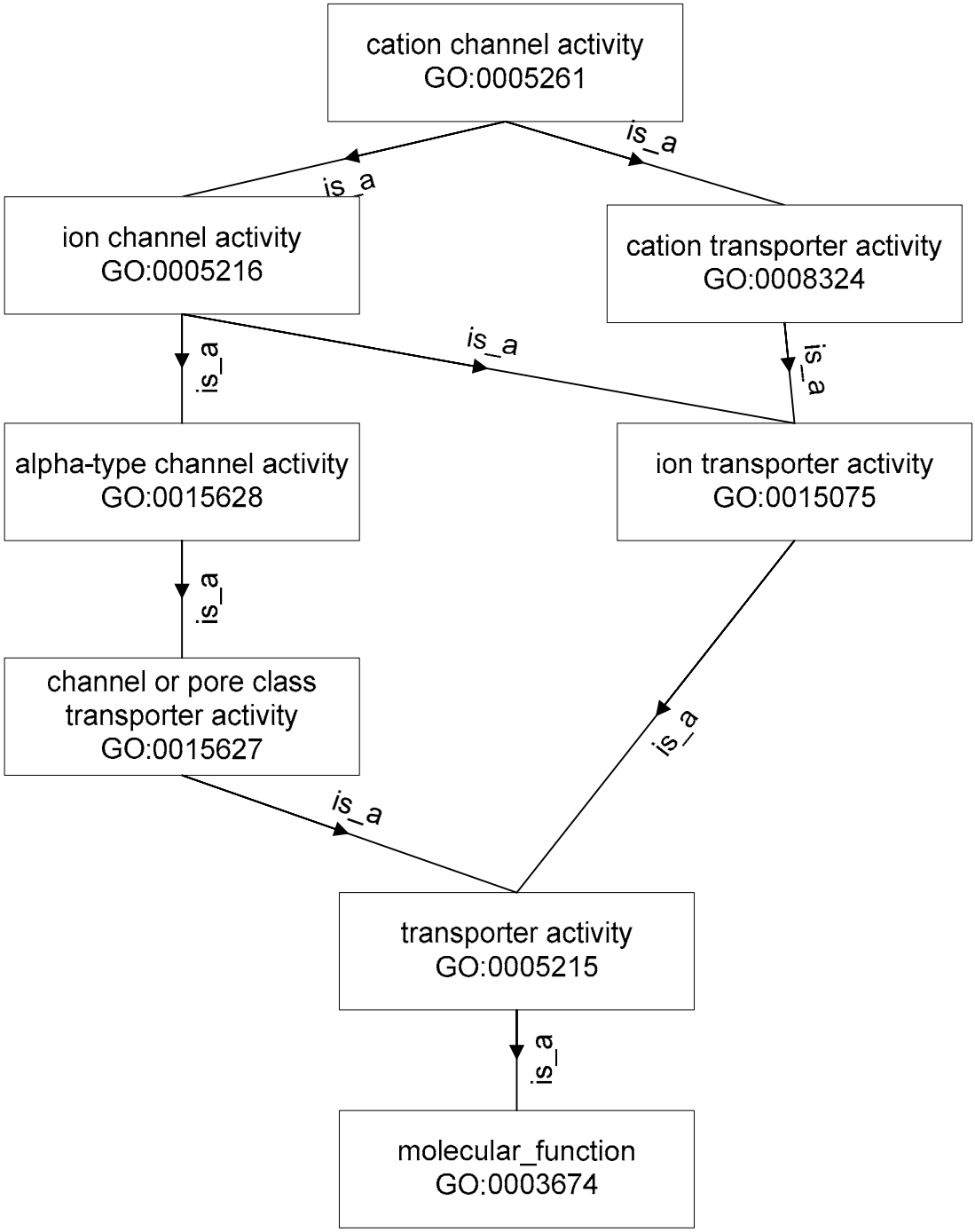
3.3.1 Περιγραφή της Γονιδιακής Οντολογίας και των απόψεών της

Αναλυτικότερα, η Γονιδιακή Οντολογία αποτελεί ένα ελεγχόμενο λεξιλόγιο που είναι δομημένο. Επιτρέπει δηλαδή την υποβολή ερωτημάτων σε διαφορετικά επίπεδα κάνοντάς την έτσι εξαιρετικά ευέλικτη. Χάρη στην προαναφερθείσα παρεχόμενη άνεση επιτρέπεται στους ειδικούς που επισημειώνουν όρους Γονιδιακής Οντολογίας, οι οποίοι είναι γνωστοί ως GO όροι, σε γονίδια (ή γονιδιακά προϊόντα) να επιλέγουν κάθε φορά τον όρο εκείνου του επιπέδου που αναλογεί στο μέτρο γνώσης που υπάρχει (για αυτό το γονίδιο ή γονιδιακό προϊόν) [i]. Η Γονιδιακή Οντολογία διαιρείται σε τρεις οντολογίες-απόψεις (aspects) οι οποίες παρέχουν πληροφορίες κοινές για όλα τα είδη οργανισμών (Εικ. 3.1). Αυτές οι τρεις οντολογίες-απόψεις είναι οι: *βιολογική διαδικασία (biological process)*, *μοριακή λειτουργία (molecular function)* και *κυτταρική σύσταση (cellular component)*. Το πιο σημαντικό στοιχείο αυτών των οντολογιών-απόψεων είναι η *ορθογωνιότητα*, η διασφάλιση δηλαδή ύπαρξης κάθε όρου μόνο σε μία από τις τρεις απόψεις. Η ιδιότητα αυτή εγγυάται τη μοναδικότητα ενός γνωρίσματος που αποδίδεται σε ένα γονίδιο. Με απλά λόγια, όταν ένας όρος Γονιδιακής Οντολογίας (GO όρος) αποδίδεται σε ένα γονίδιο αυτόματα αποκλείεται η

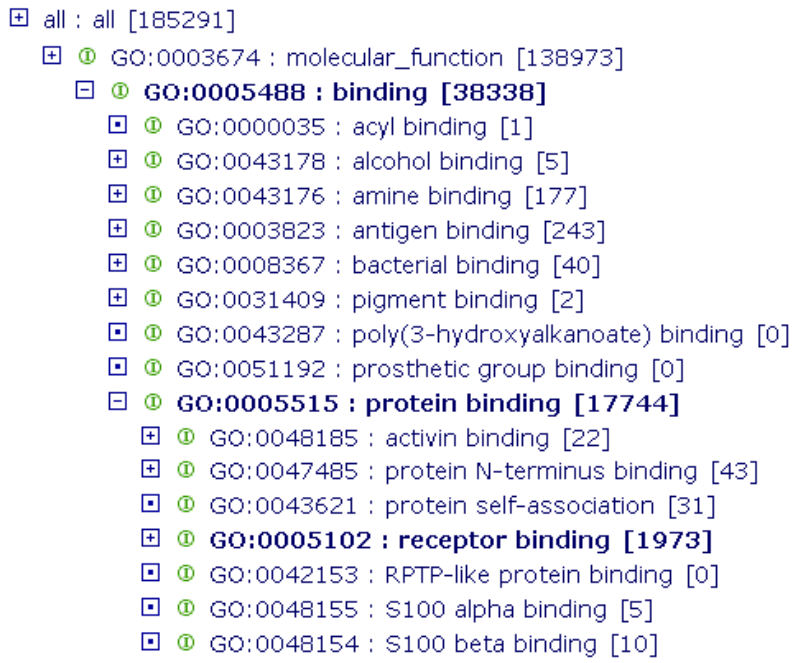
Εικόνα 3.1. Η Γονιδιακή Οντολογία Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [xxi]

ύπαρξη ενός αντίστοιχου όρου από άλλη οντολογία-άποψη που να φέρει πανομοιότυπες ιδιότητες.

Η οντολογία-άποψη της Μοριακής Λειτουργίας (MF) περιγράφει τη βιοχημική δραστηριότητα ενός γονιδιακού προϊόντος χωρίς να καθορίζει το χρόνο ή το χώρο όπου αυτή (η δραστηριότητα) συνέβη. Όπως μόλις προαναφέρθηκε περιγράφει τις δράσεις των αντικειμένων παρά τα ίδια τα αντικείμενα, ενώ συνήθως αναφέρεται σε δραστηριότητες που εκτελούνται από ένα μόνο γονιδιακό προϊόν. Είναι σύνηθες φαινόμενο για έναν όρο Μοριακής Λειτουργίας να χαρακτηρίζεται από το επισημειωμένο (annotated) του γονιδιακό προϊόν (Εικ. 3.2). Από εδώ και στο εξής, για λόγους συντομίας, θα αναφερόμαστε στην οντολογία-άποψη της Μοριακής Λειτουργίας (MF) με την ονομασία MF οντολογία.


cation channel activity
GO:0005261
is_a
is_a
ion channel activity
GO:0005216
cation transporter activity
GO:0008324
is_a
is_a
is_a
alpha-type channel activity
GO:0015628
ion transporter activity
GO:0015075
is_a
channel or pore class
transporter activity
GO:0015627
is_a
is_a
transporter activity
GO:0005215
is_a
molecular_function
GO:0003674

a)



b)

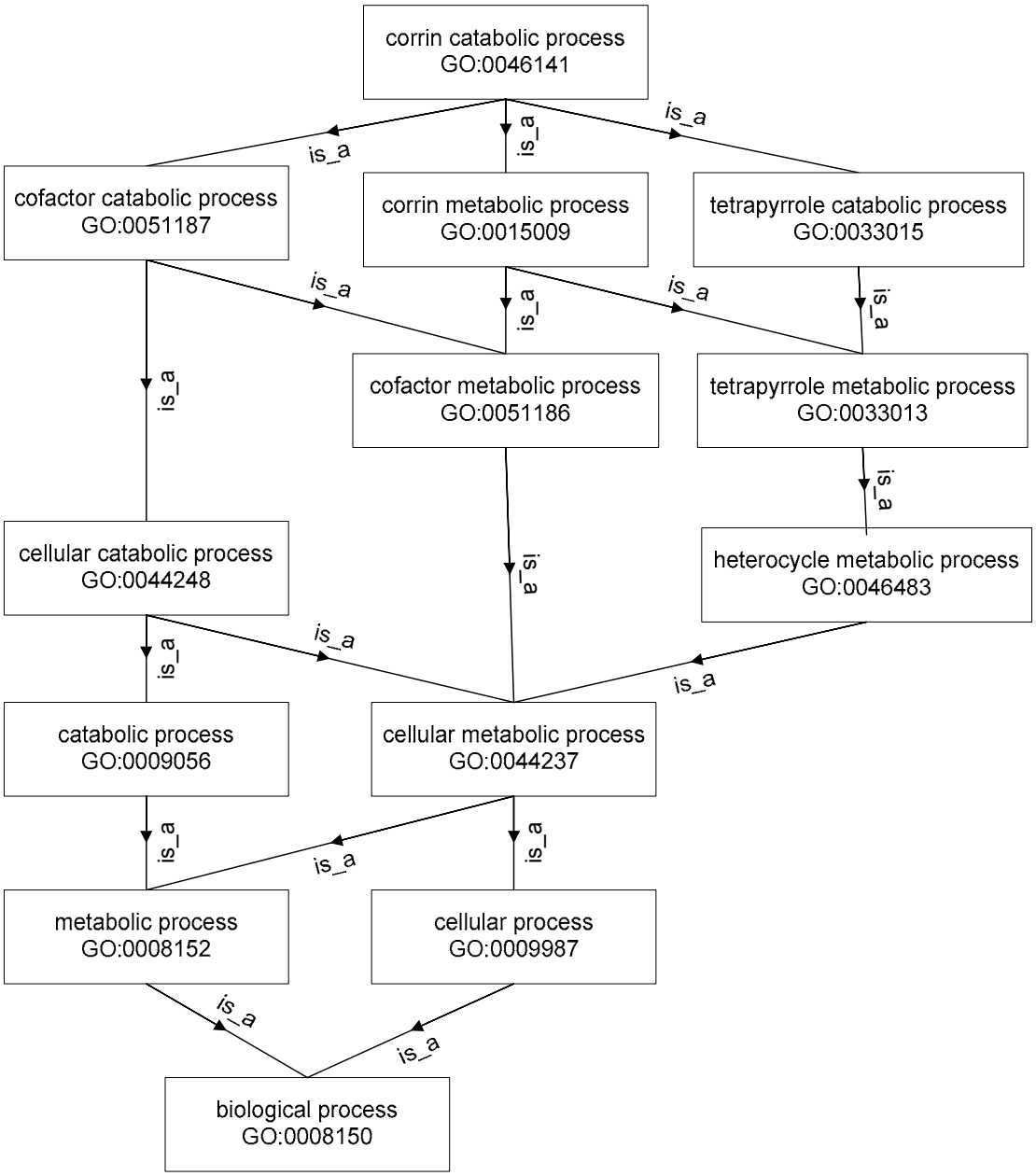
Εικόνα 3.2. Στιγμιότυπο της οντολογίας-άποψης της Μοριακής Λειτουργίας

a) Στιγμιότυπο της οντολογίας-άποψης της Μοριακής Λειτουργίας σε μορφή Κατευθυνόμενου Άκυκλου Γράφου (DAG). Όπως παρατηρείται, υπάρχουν κόμβοι που έχουν παραπάνω από έναν γονείς. Ρίζα της οντολογίας-άποψης είναι ο όρος: GO:0003674, molecular_function.

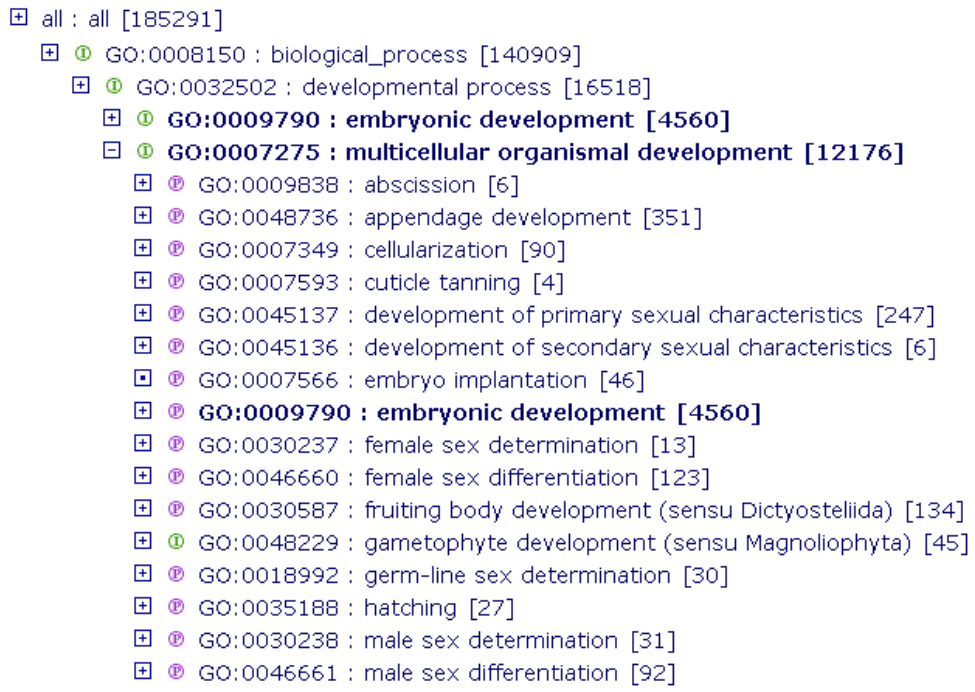
b) Στιγμιότυπο της οντολογίας-άποψης της Μοριακής Λειτουργίας σε δενδρική μορφή. Ρίζα της οντολογίας-άποψης είναι ο όρος: GO:0003674, molecular_function.

Η οντολογία-άποψη της Βιολογικής Διαδικασίας (BP) αποτελείται από ένα ή περισσότερα στάδια λειτουργιών. Υπάρχει μία σαφής σχέση μεταξύ των οντολογιών-απόψεων της Βιολογικής Διαδικασίας και της Μοριακής Λειτουργίας. Για παράδειγμα, η διαδικασία “apoptosis” περιλαμβάνει τη μοριακή λειτουργία που είναι γνωστή ως “caspase inhibitor activity” (Εικ. 3.3). Από εδώ και στο εξής, για λόγους συντομίας, θα αναφερόμαστε στην οντολογία-άποψη της Βιολογικής Διαδικασίας (BP) με την ονομασία BP οντολογία.

Εδώ να σημειώσουμε, ότι μία βιολογική διαδικασία δεν ισοδυναμεί με ένα βιολογικό μονοπάτι· για την ώρα δεν είναι μέσα στις προτεραιότητες της Γονιδιακής Οντολογίας να αναπαραστήσει τη δυναμική και τις εξαρτήσεις που περιγράφουν πλήρως ένα (βιολογικό) μονοπάτι [i].


corrin catabolic process
GO:0046141
is_a
is_a
is_a
cofactor catabolic process
GO:0051187
corrin metabolic process
GO:0015009
tetrapyrrole catabolic process
GO:0033015
is_a
is_a
is_a
is_a
cofactor metabolic process
GO:0051186
tetrapyrrole metabolic process
GO:0033013
is_a
is_a
is_a
cellular catabolic process
GO:0044248
heterocycle metabolic process
GO:0046483
is_a
is_a
is_a
catabolic process
GO:0009056
cellular metabolic process
GO:0044237
is_a
is_a
is_a
metabolic process
GO:0008152
cellular process
GO:0009987
is_a
is_a
biological process
GO:0008150

a)



b)

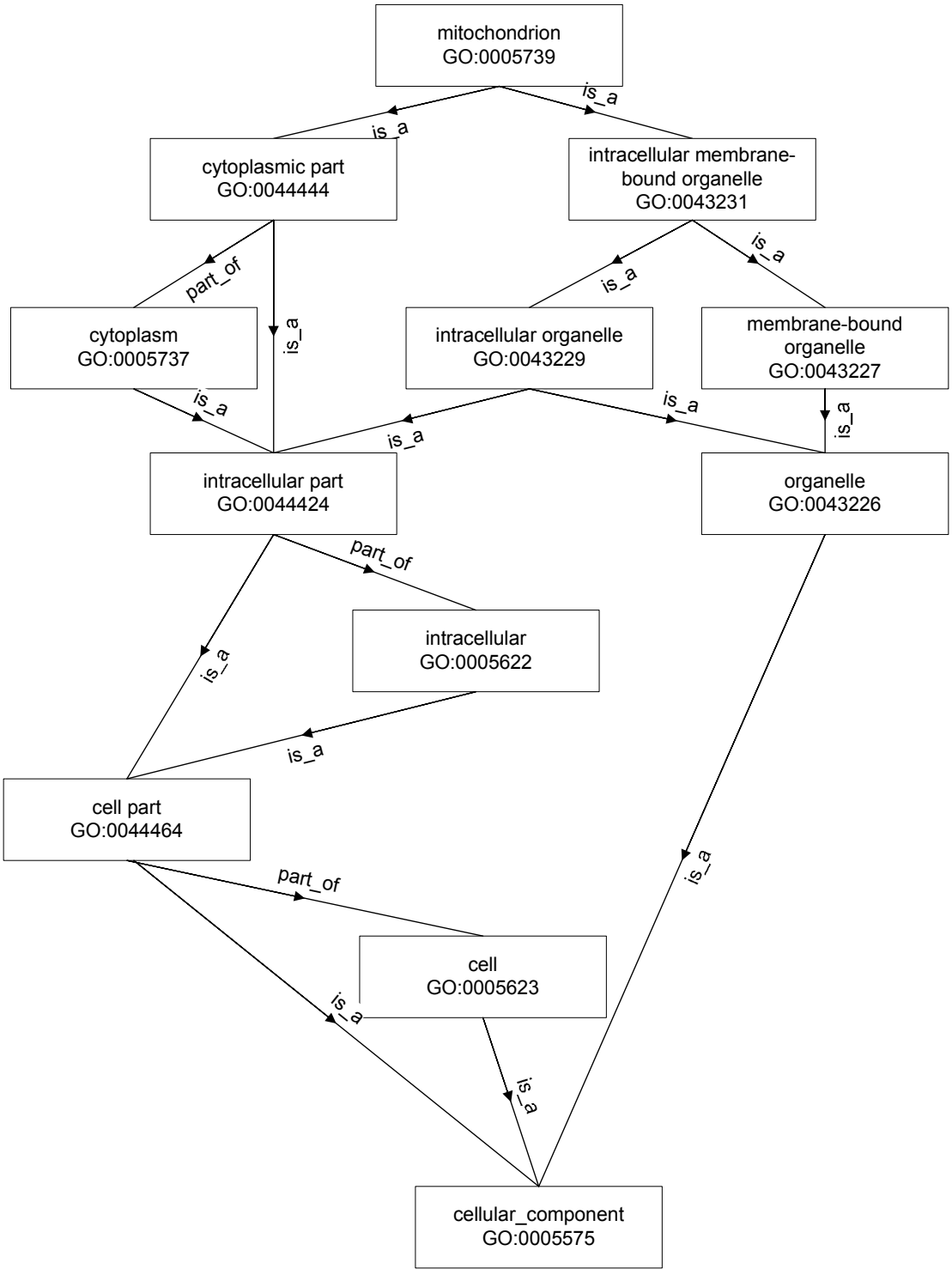
Εικόνα 3.3. Στιγμιότυπο της οντολογίας-άποψης της Βιολογικής Διαδικασίας

a) Στιγμιότυπο της οντολογίας-άποψης της Βιολογικής Διαδικασίας σε μορφή Κατευθυνόμενου Άκυκλου Γράφου (DAG). Όπως παρατηρείται, υπάρχουν κόμβοι που έχουν παραπάνω από έναν γονείς. Ρίζα της οντολογίας-άποψης είναι ο όρος: GO:0008150, biological_process.

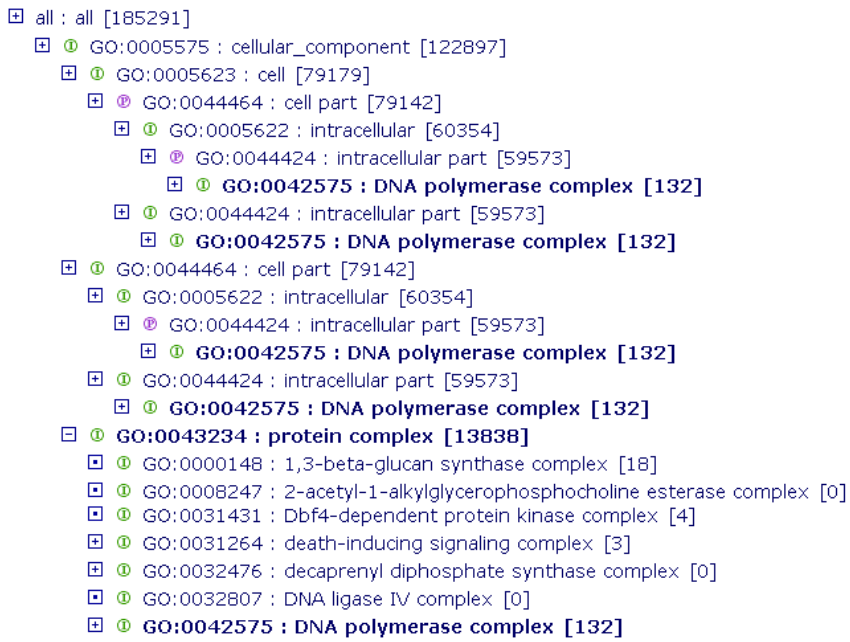
b) Στιγμιότυπο της οντολογίας-άποψης της Βιολογικής Διαδικασίας σε δενδρική μορφή. Ρίζα της οντολογίας-άποψης είναι ο όρος: GO:0008150, biological_process.

Η οντολογία-άποψη της Κυτταρικής Σύστασης (CC) περιγράφει τα μέρη στο κύτταρο όπου τα γονιδιακά προϊόντα εντοπίζονται ενεργά (Εικ. 3.4). Από εδώ και στο εξής, για λόγους συντομίας, θα αναφερόμαστε στην οντολογία-άποψη της Κυτταρικής Σύστασης (CC) με την ονομασία CC οντολογία.

Προσπαθώντας να καταστήσουμε σαφέστερη τη χρησιμότητα και τη συμπληρωματικότητα των τριών οντολογιών-απόψεων θα αναφέρουμε το παράδειγμα του γονιδιακού προϊόντος *κυτοχρώματος c*, το οποίο περιγράφεται από τον όρο της Μοριακής Λειτουργίας *oxidoreductase activity*, τους όρους Βιολογικής Διαδικασίας *oxidative phosphorylation* και *induction of cell death* και τους όρους Κυτταρικής Σύστασης *mitochondrial matrix* και *mitochondrial inner membrane*.


mitochondrion
GO:0005739
is_a
is_a
cytoplasmic part
GO:0044444
intracellular membrane-bound organelle
GO:0043231
part_of
is_a
is_a
is_a
cytoplasm
GO:0005737
intracellular organelle
GO:0043229
membrane-bound organelle
GO:0043227
is_a
is_a
is_a
is_a
intracellular part
GO:0044424
organelle
GO:0043226
part_of
is_a
intracellular
GO:0005622
is_a
cell part
GO:0044464
part_of
is_a
cell
GO:0005623
is_a
cellular_component
GO:0005575

a)



b)

Εικόνα 3.4. Στιγμιότυπο της οντολογίας-άποψης της Κυτταρικής Σύστασης

a) Στιγμιότυπο της οντολογίας-άποψης της Κυτταρικής Σύστασης σε μορφή Κατευθυνόμενου Άκυκλου Γράφου (DAG). Όπως παρατηρείται, υπάρχουν κόμβοι που έχουν παραπάνω από έναν γονείς. Ρίζα της οντολογίας-άποψης είναι ο όρος: GO:0005575, cellular_component.

b) Στιγμιότυπο της οντολογίας-άποψης της Κυτταρικής Σύστασης σε δενδρική μορφή. Ρίζα της οντολογίας-άποψης είναι ο όρος: GO:0005575, cellular_component.

Οι MF και CC οντολογίες απαντούν στο ερώτημα του τι κάνει ένα γονιδιακό προϊόν και σε ποια μέρη βρίσκεται ενεργό ενώ η BP οντολογία αποσαφηνίζει το βιολογικό σκοπό που επιτελεί ένα γονιδιακό προϊόν [23].

3.3.2 Δομή της Γονιδιακής Οντολογίας

Η δομή καθεμίας από τις τρεις οντολογίες-απόψεις φέρει τα εξής κύρια χαρακτηριστικά. Κατ' αρχήν, είναι μία δομή Κατευθυνόμενου Άκυκλου Γράφου που αντίθετα με τις δενδρικές δομές επιτρέπει την πολλαπλή γονεϊκότητα. Δεύτερον, αποτελείται αποκλειστικά από σχέσεις θεμελιώδους τύπου ("is-a", "part-of") και κάθε όρος έχει ένα μοναδικό αναγνωριστικό σημείο που χρησιμοποιείται σαν παραπομπή μεταξύ των συνεργαζόμενων βάσεων δεδομένων [23]. Είναι αξιοσημείωτο ότι ο αριθμός των σχέσεων (ακμών) και στις τρεις οντολογίες-απόψεις είναι περίπου μιάμιση φορά μεγαλύτερος από τον αριθμό των όρων (κόμβων). Η μόνη διαφορά

εντοπίζεται στην MF οντολογία όπου η αναλογία κόμβων προς ακμές είναι προσεγγιστικά ίση με ένα.

Όσον αφορά τον τύπο των σχέσεων της Γονιδιακής Οντολογίας ισχύουν τα ακόλουθα. Ο τύπος "είναι" ("is-a" type) αποτελεί μία απλή σχέση κλάσης – υποκλάσης. Μ' άλλα λόγια, δηλώνει μία σχέση κληρονομικότητας μεταξύ γονέα και παιδιού. Για παράδειγμα, ένα *πυρηνικό χρωμόσωμα (nuclear chromosome) είναι ένα (is-a) χρωμόσωμα (chromosome)* αφού φέρει όλες τις ιδιότητες του χρωμοσώματος σε συνδυασμό με τα όποια πρόσθετα γνωρίσματα του προσδίδει ο όρος *πυρηνικός*.

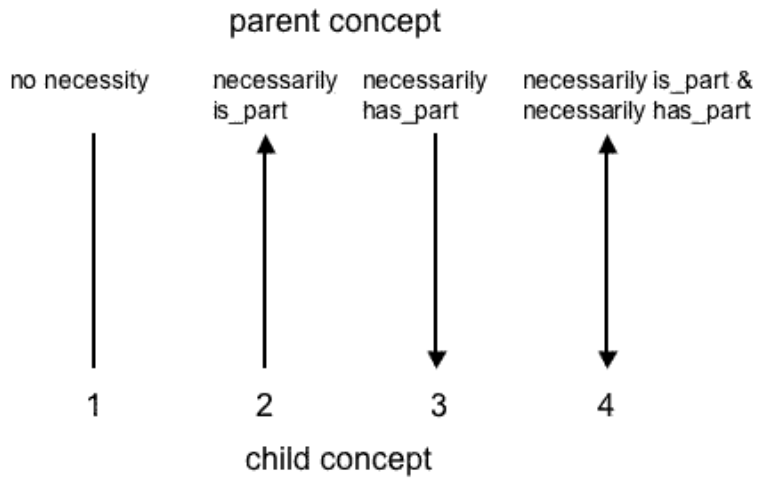
Από την άλλη, ο τύπος "μέρος-του" ("part-of" type) είναι πιο πολύπλοκος αφού υπάρχουν τέσσερα βασικά επίπεδα περιορισμού για τον τύπο αυτό: το επίπεδο όπου δεν υπάρχει κανένας περιορισμός (no restriction), το "απαραιτήτως είναι μέρος-του" ("necessarily is-part") επίπεδο, το "απαραιτήτως περιέχει" ("necessarily has-part") επίπεδο και το "απαραιτήτως είναι μέρος-του και περιέχει" ("necessarily is-part and has-part") επίπεδο.

Αναλυτικότερα, στο επίπεδο όπου δεν υπάρχει κανένας περιορισμός (no restriction) δεν μπορεί να εξαχθεί κανένα συμπέρασμα για τη σχέση μεταξύ γονέα και παιδιού εκτός βέβαια από το γεγονός ότι ο γονέας μπορεί να περιέχει το παιδί και το παιδί μπορεί να αποτελεί μέρος του γονέα.

Στο "απαραιτήτως είναι μέρος-του" ("necessarily is-part") επίπεδο η ύπαρξη του παιδιού δεν νοείται χωρίς αυτή του γονέα. Για παράδειγμα, η *διχάλα αντιγραφής (replication fork) είναι απαραίτητα μέρος του (necessarily is-part of) χρωμοσώματος (chromosome). Έτσι, όποτε αυτή συμβαίνει είναι πάντα μέρος του χρωμοσώματος. Από την άλλη, κάθε χρωμόσωμα δεν περιέχει απαραίτητα διχάλα αντιγραφής.*

Το "απαραιτήτως περιέχει" ("necessarily has-part") επίπεδο είναι το αντίστροφο του "απαραιτήτως είναι μέρος-του". Εδώ. όποτε ο γονέας υπάρχει, σημαίνει ότι περιέχει το παιδί. Παραδειγματικά, πάντα ο *πυρήνας (nucleus) περιέχει (necessarily has-part) χρωμόσωμα (chromosome). Αλλά όλα τα χρωμοσώματα δε βρίσκονται σε πυρήνες.*

Τέλος, το "απαραιτήτως είναι μέρος-του και περιέχει" ("necessarily is-part and has-part") επίπεδο αποτελεί ένα συνδυασμό των δύο παραπάνω επιπέδων. Για παράδειγμα, κάθε *πυρήνας (nucleus) περιέχει πυρηνική μεμβράνη (nucleus membrane) και κάθε πυρηνική μεμβράνη (nucleus membrane) αποτελεί μέρος του πυρήνα (nucleus)* [vii]. (Εικ. 3.5)



Εικόνα 3.5. Βασικοί τύποι (επίπεδα) part-of σχέσης

Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [xvi]

Συνεχίζοντας την καταγραφή γνωρισμάτων των δύο τύπων σχέσεων της Γονιδιακής Οντολογίας αξίζει να σημειώσουμε ότι ο αριθμός των “is-a” σχέσεων είναι κυρίαρχος αφού υπερτερεί κατά εφτά περίπου φορές αυτόν των “part-of” σχέσεων. Ενώ, η MF οντολογία αποτελείται σχεδόν αποκλειστικά από “is-a” ακμές (Πιν. IV). Τέλος, συνηθίζεται να αποφεύγεται όταν ένας όρος έχει “part-of” σχέσεις να συνδέεται ταυτόχρονα και με “is-a” ακμές ώστε να αποτραπεί η γενικότητα (abstractness), κάτι που δεν εμπίπτει στα πλάνα σχεδίασης της Γονιδιακής Οντολογίας [13].

Πίνακας IV. Στοιχεία για την κατανομή των όρων και των σχέσεων στη Γονιδιακή Οντολογία (έκδοση Απριλίου του 2007)

Οντολογία-άποψη	**# όρων**	**# σχέσεων (#ακμών)**		
		# is-a	**# part-of**	**σύνολο**
cellular component	1926	2902	799	3701
molecular function	7559	8823	1	8824
biological process	13413	20295	3872	24167
obsolete set	1099	-	-	-
Σύνολο	23997	32020	4672	36692

3.3.3 Αρχεία Επισημείωσης Όρων Γονιδιακής Οντολογίας

Η ανάγκη απόδοσης χαρακτηρισμών των λειτουργιών και των χώρων δραστηριότητας των γονιδιακών προϊόντων οδήγησε στην ίδρυση του Έργου Επισημείωσης Όρων Γονιδιακής Οντολογίας (GOA project, Gene Ontology Annotation project) και υποστηρίζεται από τις αντίστοιχες Βάσεις Δεδομένων

Επισημείωσης Όρων Γονιδιακής Οντολογίας (GOA Databases). Μερικές από αυτές είναι οι SwissProt, InterPro, Gramene, MGI, FlyBase, DictyBase, TAIR και άλλες. Το έργο αυτό ξεκίνησε για να επισημειώσει GO όρους σε γονιδιακά προϊόντα. Η επισημείωση αυτή ακολουθεί δύο βασικές αρχές· πρώτον, θα πρέπει να αναφέρεται η πηγή από όπου εξήχθη αυτή η επισημείωση και δεύτερον να παρατίθεται ο τύπος απόδειξης που τη στηρίζει. Γενικά, οι επισημειώσεις αναφέρονται σε γονιδιακά προϊόντα (και όχι στα ίδια τα γονίδια) γιατί ένα γονίδιο μπορεί να κωδικοποιεί τη δημιουργία πολλών διαφορετικών προϊόντων με πολύ διαφορετικές ιδιότητες για το καθένα από αυτά. Από εδώ και στο εξής θα αναφερόμαστε, για λόγους συντομίας, στα αρχεία Επισημείωσης Όρων Γονιδιακής Οντολογίας ως GOA αρχεία και στις βάσεις δεδομένων Επισημείωσης Όρων Γονιδιακής Οντολογίας ως GOA βάσεις δεδομένων.

3.3.3.1 Κανόνας του Ορθού Μονοπατιού

Επίσης, οι επισημειώσεις των GO όρων θα πρέπει να αναφέρονται με όσο το δυνατόν μεγαλύτερη λεπτομέρεια στην υπάρχουσα γνώση που αντιστοιχεί σε ένα γονιδιακό προϊόν, όπως και να ακολουθούν τον *Κανόνα του Ορθού Μονοπατιού (the True Path Rule)*, ο οποίος ουσιαστικά αξιώνει κάθε μονοπάτι από ένα όρο προς οποιοδήποτε πρόγονό του να είναι αληθές [vi]. Για παράδειγμα, ο όρος “alkali metal ion binding” μπορεί να θεωρείται ως “metal ion binding” (που αποτελεί ένα άμεσο γονέα του) όπως και ως “ion binding” (που αποτελεί ένα από τους προγόνους του) εξίσου. Με απλά λόγια, ένας όρος διατηρεί όλα τα χαρακτηριστικά των προγόνων του και μπορεί να θεωρηθεί ως ένας από αυτούς. Αυτός είναι και ο κύριος λόγος που στη Γονιδιακή Οντολογία παρατηρείται συνήθως ο τύπος “απαραιτήτως είναι μέρος-του” της σχέσης “μέρος-του” (αφού ένα παιδί όταν συνδέεται με “part-of” σχέση με το γονέα του πρέπει πάντα να αποτελεί μέρος του ώστε να διασφαλίζεται η εγκυρότητα του μονοπατιού) [vii].

3.3.3.2 Τύπος απόδειξης ως μέσο εξακρίβωσης της εγκυρότητας της επισημείωσης

Οι επισημειωμένες εγγραφές αναφέρονται είτε σε πολύ ειδικούς είτε σε πολύ γενικούς όρους. Όπως προαναφέρθηκε ένα βασικότατο πεδίο διαπίστωσης της

αξιοπιστίας μίας επισημείωσης είναι ο τύπος απόδειξης που τη στηρίζει. Σε γενικές γραμμές, παρέχονται ηλεκτρονικές και χειρωνακτικές επισημειώσεις μεταξύ GO όρων και γονιδιακών προϊόντων. Οι χειρωνακτικές επισημειώσεις βασίζονται σε δημοσιευμένες πληροφορίες που προέρχονται από επιστημονική βιβλιογραφία, πειράματα, βιολογική γνώση και το επίπεδο εμπιστοσύνης του ειδικού επόπτη. Επομένως, εμφανίζουν μία υποκειμενικότητα. Οι αυτοματοποιημένες διαδικασίες, από την άλλη, εκμεταλλεύονται αντιστοιχίες από άλλες βάσεις δεδομένων όπως για παράδειγμα οι διασταυρωμένες αναφορές σε βάσεις δεδομένων πρωτεϊνικών οικογενειών. Οι επισημειώσεις σε αυτές τις δευτερεύουσες βάσεις δεδομένων πραγματοποιούνται με μεθόδους βιοπληροφορικής όπως είναι η ομοιότητα αλληλουχίας ή η αυτόματη επεξεργασία επιστημονικών κειμένων. Αυτές οι αυτόματες τεχνικές είναι κατά πολύ γρηγορότερες από τις χειρωνακτικές μεθόδους. Επίσης, είναι πιο συνεπείς γιατί, σε αντίθεση με τις επισημειώσεις που πραγματοποιούνται από ειδικούς επόπτες, αυτές βασίζονται σε ακριβείς κανόνες. Ωστόσο, δεν είναι τόσο αξιόπιστες όσο αυτές που προκύπτουν κατόπιν εργασίας ανθρώπων-ειδικών εποπτών [20].

Οι χειρωνακτικές μέθοδοι παράγουν γενικά αξιόπιστα αποτελέσματα. Αποδίδουν GO όρους πιο ομοιόμορφα στις τρεις οντολογίες–απόψεις και παρέχουν βιβλιογραφικές αναφορές και πληροφορίες για τον τύπο των πειραμάτων που διεξήχθησαν και χρησιμοποιήθηκαν για την εξαγωγή της συγκεκριμένης επισημείωσης. Παρόλα αυτά, αυτή η μέθοδος δεν επαρκεί για την αντιμετώπιση του τεράστιου όγκου βιολογικών δεδομένων.

Αντίθετα, οι ηλεκτρονικές μέθοδοι δημιουργούν τη συντριπτική πλειοψηφία των επισημειώσεων. Από την άλλη όμως, είναι υπεύθυνες για μία πιο ανομοιόμορφη κατανομή των GO όρων κατά μήκος των τριών οντολογιών-απόψεων και αποδίδουν κυρίως γενικότερους GO όρους [2]. Έτσι, αναπόφευκτα υπάρχει ένα μεγάλο χάσμα μεταξύ του αριθμού των επισημειώσεων που αναφέρονται σε ένα γενικό GO όρο και σε ένα ειδικότερο απόγονό του (αφού οι ηλεκτρονικές μέθοδοι αναφέρονται κυρίως σε όρους που βρίσκονται ψηλά στη Γονιδιακή Οντολογία – επομένως γενικούς όρους).

Αποτέλεσμα των παραπάνω είναι να υπάρχουν διαφορετικοί τύποι αποδείξεων (evidence codes) που οφείλονται στον διαφορετικό τρόπο που έγινε μία επισημείωση μεταξύ γονιδιακού προϊόντος και GO όρου. Για την ακρίβεια, υπάρχουν 14 τύποι αποδείξεων εκ των οποίων ο ένας από αυτούς αναφέρεται σε ηλεκτρονικές μεθόδους (IEA: Inferred from Electronic Annotation). Οι υπόλοιποι είναι οι: IC (Inferred by Curator), IDA (Inferred from Direct Assay), IEP (Inferred from Expression Pattern),

IGC (Inferred from Genomic Context), IGI (Inferred from Genomic Context), IGI (Inferred from Genetic Interaction), IMP (Inferred from Mutant Phenotype), IPI (Inferred from Physical Interaction), ISS (Inferred from Sequence or Structural Similarity), NAS (Non-traceable Author Statement), ND (No biological Data available), RCA (inferred from Reviewed Computational Analysis), TAS (Traceable Author Statement) και NR (Not Recorded).

Παρακάτω θα αναφέρουμε συνοπτικά τη σημασία κάθε τύπου απόδειξης:

IC (Inferred by Curator): Χρησιμοποιείται όταν μία επισημείωση δεν στηρίζεται σε αποδείξεις αλλά μπορεί να εξαχθεί από ένα επόπτη βάσει άλλων επισημειώσεων. Για παράδειγμα, όταν ένας επόπτης επισημειώνει ένα ευκαρυωτικό γονιδιακό προϊόν και γνωρίζει ότι έχει τη λειτουργία transcription factor activity (δραστηριότητα μεταγραφικού παράγοντα) μπορεί να υποθέσει βάσιμα ότι βρίσκεται στην τοποθεσία nucleus (πυρήνας) του κυττάρου.

IDA (Inferred from Direct Assay): Χρησιμοποιείται όταν μία επισημείωση προέρχεται από απ' ευθείας ανάλυση ενζύμων ή ανοσοφθορισμού, ή από ανασύνθεση σε δοκιμαστικό σωλήνα (in-vitro reconstitution), ή από κυτταρικό τεμαχισμό ή τέλος από φυσική αλληλεπίδραση.

IEA (Inferred from Electronic Annotation): Χρησιμοποιείται όταν μία επισημείωση προέρχεται από υπολογιστικές διαδικασίες ή αυτόματη μεταφορά επισημειώσεων από βάσεις δεδομένων. Η διαφορά με άλλους τύπους αποδείξεων που χρησιμοποιούν στοιχεία ομοιότητας αλληλουχίας ή επισημειώσεις από βάσεις δεδομένων είναι ότι η IEA επισημείωση δεν ελέγχεται από κάποιο επόπτη.

IEP (Inferred from Expression Pattern): Χρησιμοποιείται όταν μία επισημείωση εξάγεται από τα επίπεδα έκφρασης ενός γονιδίου ή ενός γονιδιακού προϊόντος. Τα δεδομένα γονιδιακής έκφρασης είναι καταλληλότερα για χρησιμοποίηση επί επισημειώσεων που αφορούν την BP οντολογία και όχι τόσο της MF (οντολογίας). Έτσι, θα πρέπει να χρησιμοποιείται με σύνεση. Ο τύπος αυτός

χρησιμοποιείται κυρίως όταν ένας GO όρος προκύπτει από πειράματα μικροσυστοιχιών.

IGC (Inferred from Genomic Context): Χρησιμοποιείται όταν μία επισημείωση προκύπτει από το γονιδιακό περιεχόμενο του γονιδιακού προϊόντος. Αναφερόμενοι στο γονιδιακό περιεχόμενο εννοούμε τη γειτονιά του συγκεκριμένου γονιδιακού προϊόντος, τη δομή οπερονίου καθώς και άλλες φυλογενετικές ή γονιδιωματικές αναλύσεις.

IGI (Inferred from Genetic Interaction): Χρησιμοποιείται όταν μία επισημείωση εξάγεται από κάθε πιθανή αλλαγή στην αλληλουχία ή την έκφραση ενός γονιδίου ή γονιδιακού προϊόντος. Πιο συγκεκριμένα, αναφέρεται σε γονιδιακή αλληλεπίδραση, λειτουργική συμπληρωματικότητα ή φαινοτυπικές μεταλλάξεις σε διαφορετικό γονίδιο και καλύπτει πειράματα φαινοτυπικών μεταλλάξεων που λαμβάνουν χώρα σε μη-άγριου-τύπου περιβάλλον. Αυτός ο τύπος απόδειξης, επίσης, χρησιμοποιείται σε περιπτώσεις όπου μία μετάλλαξη σε ένα γονίδιο παρέχει πληροφορίες για τη λειτουργία, βιολογική διαδικασία ή τοποθεσία ενός άλλου γονιδίου.

IMP (Inferred from Mutant Phenotype): Χρησιμοποιείται όταν μία επισημείωση προκύπτει από αλλαγές, όπως μεταλλάξεις ή ασυνήθιστα επίπεδα γονιδιακής έκφρασης του (γονιδιακού) προϊόντος. Πιο συγκεκριμένα, αναφέρεται σε οποιαδήποτε γονιδιακή μετάλλαξη, υπερέκφραση άγριου-τύπου ή μεταλλαγμένων γονιδίων, RNAi πειράματα, καταστολείς εξειδικευμένων πρωτεϊνών και σε πολυμορφισμό. Ο IMP τύπος χρησιμοποιείται, επίσης, όταν η επισημείωση εξάγεται από πειράματα (όπως ανάλυση QTL) όπου οι παρατηρούμενες επιδράσεις οφείλονται σε φυσικές αλλαγές του γονιδίου. Ο IMP τύπος καλύπτει, επίσης, καταστάσεις φαινοτυπικής ομοιότητας· όταν ένας φαινότυπος είναι όμοιος με ένα άλλο ανεξάρτητο φαινότυπο για τον οποίο υπάρχει λεπτομερέστερη καταγραφή των λειτουργιών του.

Γενικά, ο IMP τύπος χρησιμοποιείται όταν παρατηρούνται μη φυσιολογικές καταστάσεις σε ένα κύτταρο ή οργανισμό που οφείλονται σε αλλαγές στην αλληλουχία ή την έκφραση ενός γονιδίου.

IPI (Inferred from Physical Interaction): Χρησιμοποιείται όταν μία επισημείωση εξάγεται από φυσικές αλληλεπιδράσεις μεταξύ του γονιδιακού προϊόντος και άλλου μορίου (π.χ. ιόντος ή συμπλέγματος). Πιο συγκεκριμένα, αναφέρεται σε 2-υβριδο επιδράσεις (2-hybrid interactions) συγκαθαρισμό (co-purification), συ-ανοσο-ιζηματοποίηση (co-immunoprecipitation) και πειράματα ιοντικής/πρωτεϊνικής σύνδεσης (ion/protein binding experiments).

ISS (Inferred from Sequence or Structural Similarity): Χρησιμοποιείται όταν μία επισημείωση εξάγεται από ανάλυση που βασίζεται σε συστοίχιση αλληλουχιών (sequence alignment), σύγκριση δομής, ή υπολογισμό χαρακτηριστικών αλληλουχίας όπως σύνθεση και έχει επικυρωθεί από ειδικό επόπτη. Καλύπτει περιπτώσεις αποτελεσμάτων που βασίζονται σε ομοιότητα αλληλουχίας και έχουν προκύψει από ανάλογες μεθόδους εντοπισμού (όπως BLAST) ή από αξιόπιστες δημοσιεύσεις έπειτα από επικύρωση ειδικού επόπτη.

NAS (Non-traceable Author Statement): Χρησιμοποιείται όταν μία επισημείωση εξάγεται από στοιχεία που εντοπίζονται σε γραπτά κείμενα (σύνοψη, εισαγωγή) ή βάσεις δεδομένων και δεν υπάρχει αναφορά σε ανάλογη δημοσίευση που να στηρίζει αυτήν την επισημείωση.

ND (No biological Data available): Χρησιμοποιείται όταν μία επισημείωση αναφέρεται σε μία άγνωστη βιολογική διαδικασία, μοριακή λειτουργία ή κυτταρική σύσταση του γονιδιακού προϊόντος. Με άλλα λόγια, την συγκεκριμένη χρονική στιγμή όπου έγινε η επισημείωση δεν υπήρχε ο κατάλληλος GO όρος της ανάλογης οντολογίας-άποψης που να μπορούσε να

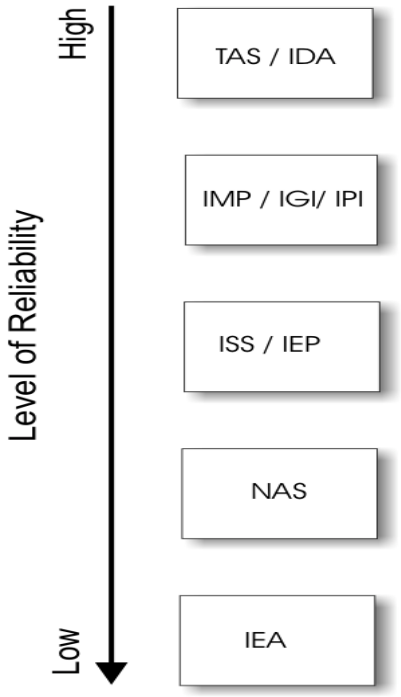
αποδοθεί από τον ειδικό επόπτη στο (γονιδιακό) προϊόν.

RCA (inferred from Reviewed Computational Analysis): Χρησιμοποιείται όταν μία επισημείωση εξάγεται από υπολογιστικές μεθόδους που δεν βασίζονται σε ομοιότητα αλληλουχίας και έχει επικυρωθεί από ειδικό επόπτη. Για παράδειγμα, προβλέψεις που στηρίζονται σε μεγάλης κλίμακας πειράματα ή σύνολα δεδομένων ή ακόμα και εξόρυξη κειμένων βιολογικού περιεχομένου εντάσσονται σε αυτή την κατηγορία. Επίσης, ο τύπος RCA χρησιμοποιείται όταν αποτελέσματα από μικροσυστοιχίες συνδυάζονται με αποτελέσματα από άλλου είδους μεγάλης κλίμακας πειράματα.

TAS (Traceable Author Statement): Χρησιμοποιείται όταν μία επισημείωση προκύπτει από επιστημονικά κείμενα και η αναφορά στην πρωτότυπη εργασία (από όπου και συνάγεται αυτή η επισημείωση) είναι διαθέσιμη.

NR (Not Recorded): Χρησιμοποιείται για τις επισημειώσεις που έλαβαν χώρα πριν οι ειδικοί επόπτες ξεκινήσουν τη διαδικασία εντοπισμού του τύπου απόδειξης για κάθε επισημείωση. Δεν χρησιμοποιείται πια για τις νέες επισημειώσεις. Στη θέση του χρησιμοποιείται ο TAS ή ο NAS τύπος.

Ουσιαστικά, ο τύπος της απόδειξης καθορίζει την εγκυρότητα της επισημείωσης. Για παράδειγμα, ο τύπος ISS, για παράδειγμα, αναφέρεται στην εξαγωγή μίας επισημείωσης με τη βοήθεια δομικής ομοιότητας μέσω σύγκρισης αλληλουχιών (που έχει όμως επιβεβαιωθεί και από ειδικό επόπτη). Γενικά, οι τύποι TAS και IDA θεωρούνται οι πιο αξιόπιστοι αφού αντιστοιχούν σε επισημειώσεις που αντλήθηκαν από επιστημονικά κείμενα, ενώ ο τύπος IEA είναι ο πιο αναξιόπιστος αφού αναφέρεται σε ομοιότητα αλληλουχίας που δεν έχει επικυρωθεί από κάποιον ειδικό επόπτη (curator) [viii]. (Εικ. 3.6)



Εικόνα 3.6. Δείκτης αξιοπιστίας των τύπων αποδείξεων
Η παραπάνω εικόνα έχει βασιστεί σε πληροφορία που παρέχεται από την τοποθεσία [viii]

Αναφερθήκαμε προηγουμένως εκτενώς στις επισημειώσεις GO όρων καθώς και στους τύπους αποδείξεων που τις στηρίζουν. Κρίνεται, έτσι, σκόπιμο να αναφερθούν και οι λόγοι δημιουργίας αυτού του επίδοξου έργου, δηλαδή του Έργου Καταχώρησης Όρων Γονιδιακής Οντολογίας. Με άλλα λόγια, είναι χρήσιμο να καταγραφούν τόσο τα πλεονεκτήματα όσο και τα μειονεκτήματα που προκύπτουν από τη χρήση επισημειώσεων όρων Γονιδιακής Οντολογίας (GO όρων).

3.3.3.3 Πλεονεκτήματα και μειονεκτήματα από τη χρήση επισημειώσεων GO όρων

Τα πλεονεκτήματα που θα μπορούσαμε να αναφέρουμε είναι τα παρακάτω:

i) Κατ' αρχήν, καθίσταται δυνατή η σύγκριση πρωτεϊνών ως προς τη λειτουργία τους με σκοπό την εύρεση λειτουργικά ισοδύναμων πρωτεϊνών ανεξάρτητα από την ομολογία τους. Συγκριτικές τεχνικές που βασίζονται μόνο στην ακολουθιακή ομοιότητα δεν παρέχουν πληροφορίες για τη λειτουργικότητα των πρωτεϊνών. Αυτό συμβαίνει γιατί είναι δυνατόν να

υπάρξουν πρωτεΐνες, που ενώ έχουν σημαντική ακολουθιακή ομοιότητα προερχόμενες από ένα κοινό πρόγονο, εμφανίζουν τελείως διαφορετικές λειτουργίες. Από την άλλη, είναι δυνατό να υπάρξουν πρωτεΐνες, που ενώ προέρχονται από διαφορετικούς προγόνους παρουσιάζοντας έτσι μία πολύ μικρή ακολουθιακή ομοιότητα, μπορούν να έχουν την ίδια λειτουργία. Οι μοντέρνες μέθοδοι εντοπισμού ορθολογίας αδυνατούν να βρουν ακολουθίες με παρόμοιες λειτουργίες που δεν έχουν υψηλή ακολουθιακή ομοιότητα. Τα παραπάνω προβλήματα επιστρατεύονται να λύσουν τεχνικές που είναι βασισμένες σε επισημειώσεις γονιδίων και των αντίστοιχων λειτουργιών τους (δηλ. GO όρων).

ii) Οι μέθοδοι που είναι βασισμένες σε επισημειώσεις μπορούν να βοηθήσουν να γίνει αντιληπτή πόσο διατηρημένη ή ποικιλόμορφη είναι η μοριακή βιολογία διαφορετικών ειδών οργανισμών και πως εξελίχθηκαν οι διαφορετικές μοριακές λειτουργίες και βιολογικές διαδικασίες στο πέρασμα του χρόνου. Για παράδειγμα, συγκρίνοντας τις κυτταρικές λειτουργίες παθογόνων και μη-παθογόνων στελεχών μπορούν να εξαχθούν πιο χρήσιμα συμπεράσματα για το μηχανισμό παθογένεσης από ότι αν συγκρίνονταν ακολουθίες.

iii) Η σύγκριση επισημειώσεων λειτουργιών μπορεί να εφαρμοστεί, επίσης, στην αναζήτηση νέων φαρμάκων-στόχων. Συγκρίνοντας την παθογόνο πρωτεΐνη με επισημειώσεις λειτουργιών στον άνθρωπο είναι δυνατό να αποκαλυφθούν επιπρόσθετες λειτουργίες της. Οι παθογόνες πρωτεΐνες που παρουσιάζουν αυτές τις λειτουργίες είναι υποψήφιοι στόχοι για τα φάρμακα.

iv) Η παραπάνω ανάλυση είναι δυνατόν να επεκταθεί σε πλήρεις ταξονομικές ομάδες εκτός των απλών οργανισμών, κάτι που επιτρέπει την εύρεση κοινών λειτουργιών σε μία ομάδα οργανισμών. Με αυτόν τον τρόπο, είναι δυνατή η εύρεση στόχων επιλεκτικά για ένα παθογόνο ή και ακόμα για μία ολόκληρη ομάδα οργανισμών [20].

Από την άλλη, τα μειονεκτήματα αυτών των μεθόδων είναι:

i) Πρώτον, οι υπάρχουσες επισημειώσεις δεν είναι πλήρεις. Σχεδόν για όλους τους χαρτογραφημένους οργανισμούς μόνο ένα υποσύνολο των γονιδίων είναι γνωστό, και ακόμη ένα μικρότερο υποσύνολο έχει τις λειτουργίες του επισημειωμένες. Όσο συγκεντρώνεται περισσότερη γνώση, νέα γονίδια (και οι αντίστοιχες επισημειώσεις τους) προστίθενται στις ανάλογες βάσεις δεδομένων. Αυτό πρακτικά σημαίνει ότι σε κάθε στιγμή, μία βάση δεδομένων επισημειώσεων GO όρων θα είναι ελλιπής, αφού περιέχει μόνο

ένα υποσύνολο των γονιδίων ενός δεδομένου οργανισμού και από αυτό το υποσύνολο έχει επισημειωθεί μόνο ένα μέρος των λειτουργιών τους.

ii) Επιπλέον, οι περισσότερες βάσεις δεδομένων καταχωρήσεων GO όρων ελέγχονται από ειδικούς επόπτες που χειροκίνητα εξετάζουν την υπάρχουσα βιβλιογραφία. Έτσι, είναι πιθανό ορισμένα δημοσιευμένα δεδομένα να αγνοηθούν κατά τη διάρκεια αυτής της διαδικασίας. Για παράδειγμα, το γονίδιο HMOX2 που είχε αποδειχθεί ότι συμμετείχε στη διαδικασία της *βιοσύνθεσης χρωστικής ουσίας (pigment biosynthesis)* το 1992 δεν είχε επισημειωθεί στις βάσεις της Γονιδιακής Οντολογίας έως το Μάιο του 2004.

iii) Επίσης, η πληροφορία που εξάγεται από τέτοιες βάσεις δεδομένων μπορεί να είναι ανακριβής, γιατί οι περισσότερες επισημειώσεις εξάγονται αποκλειστικά με ηλεκτρονικές μεθόδους χωρίς να έχει εκτιμηθεί η αξιοπιστία τους από άνθρωπο. Ακόμα και αν σε μερικές περιπτώσεις οι λανθασμένες επισημειώσεις είναι τόσο εμφανείς δεν έχει βρεθεί μία αυτόματη μέθοδος που να τις αναλύει, να τις εντοπίζει και να τις διορθώνει.

iv) Είναι πιθανό να μην εντοπίζονται όλοι οι σχετικοί όροι για ένα γονίδιο. Με άλλα λόγια, ένα γονίδιο ενώ μπορεί να είναι επισημειωμένο σε ένα GO όρο της MF οντολογίας λόγω μίας λειτουργίας του, υπάρχει η πιθανότητα να μην είναι επισημειωμένο στον αντίστοιχο GO όρο της BP οντολογίας. Για παράδειγμα, το γονίδιο SLC13A2 κωδικοποιεί τη σύνθεση του μεταφορέα κιτρικού άλατος συζευγμένου με ιόντα νατρίου (Na(+)-coupled citrate transporter). Έτσι, ενώ είναι επισημειωμένο στη δραστηριότητα μεταφορέα οργανικών ανιόντων (organic anion transporter activity) της MF οντολογίας, δεν είναι επισημειωμένο στον ανάλογο GO όρο της BP οντολογίας που είναι η μεταφορά οργανικών ανιόντων (organic anion transport) [7].

Επιχειρήσαμε, έτσι, μέχρι στιγμής να κάνουμε μία σύντομη ανάλυση της έννοιας της Οντολογίας. Στη συνέχεια, επεκταθήκαμε στη Γονιδιακή Οντολογία, συγκεκριμένα που αποτελεί το κύριο εργαλείο αυτής της διπλωματικής, και παραθέσαμε τις ιδιότητές της. Τέλος, περιγράψαμε τις Βάσεις Δεδομένων Καταχωρήσεων GO όρων τονίζοντας τους κανόνες που διέπουν τη δημιουργία και συντήρησή τους καθώς και όλα τα πιθανά οφέλη αλλά και ρίσκα που προκύπτουν από τη χρήση τους.

Στη συνέχεια, θα κάνουμε μία εισαγωγή στην Εξόρυξη Δεδομένων καθώς και στις μεθόδους που χρησιμοποιήθηκαν για τις ανάγκες της διπλωματικής.

4 Ανάλυση των χρησιμοποιούμενων μεθόδων

Σε αυτό το σημείο θα αναφερθούμε αναλυτικά στις μεθόδους που εφαρμόστηκαν σε αυτήν εδώ την εργασία. Πιο συγκεκριμένα, θα σταθούμε αρχικά στην αποσαφήνιση γενικότερων εννοιών όπως είναι η εξόρυξη δεδομένων και η απόκτηση γνώσης. Κατόπιν, θα προχωρήσουμε ένα βήμα παραπέρα αναλύοντας την έννοια και τους στόχους χρησιμοποίησης ενός σημαντικού "εργαλείου" εξόρυξης γνώσης, του φιλτραρίσματος γνωρισμάτων. Στη συνέχεια, θα παρουσιαστούν διεξοδικά τρεις αλγόριθμοι φιλτραρίσματος γνωρισμάτων που χρησιμοποιήθηκαν σε αυτή την εργασία, του Κέρδους σε Πληροφορία (Information Gain), Χι Τετράγωνο (Chi Square) και Relief-F. Τέλος, αφού έχουμε καλύψει ενδελεχώς το φάσμα μεθόδων στατιστικής εκμάθησης, θα συνεχίσουμε με την παρουσίαση της έννοιας της σημασιολογικής ομοιότητας καθώς και αντίστοιχων μεθόδων.

4.1 Εξόρυξη Δεδομένων

Αναφέραμε πριν ότι βασικός στόχος αυτής της διπλωματικής θέσης είναι η βελτίωση της ακρίβειας ταξινόμησης συνόλων μικροσυστοιχιών. Σε κάποιο σημείο επίσης αναφέρθηκε ότι επιχειρείται η εφαρμογή μεθόδων εξόρυξης γνώσης από τέτοια σύνολα δεδομένων ώστε να μπορούν να εξαχθούν χρήσιμα συμπεράσματα. Σε αυτό το σημείο κρίνεται σκόπιμο να περιγραφούν αναλυτικότερα αυτοί οι όροι αλλά και να καταγραφούν οι λόγοι που οδήγησαν στην ευρεία χρησιμοποίησή τους.

Ως *εξόρυξη δεδομένων (data mining)* εννοείται η εξαγωγή "κρυφής", προηγουμένως άγνωστης και συχνά ιδιαιτέρως χρήσιμης πληροφορίας από δεδομένα. Με άλλα λόγια, είναι η διαδικασία ανακάλυψης σχέσεων και προτύπων που υπάρχουν σε μεγάλες βάσεις δεδομένων. Η διαδικασία αυτή μπορεί να είναι αυτόματη ή συνηθέστερα ημι-αυτόματη [26]. Με τον όρο της εξόρυξης δεδομένων

αναφερόμαστε επίσης και στην επιστήμη εξαγωγής πληροφορίας από ογκώδη σύνολα δεδομένων ή μεγάλες βάσεις δεδομένων.

Όπως όμως κάθε επιστήμη έτσι και αυτή δημιουργήθηκε για την κάλυψη κάποιων αναγκών και με τη βοήθεια ορισμένων συγκυριών. Οι κυριότερες από αυτές τις ανάγκες ήταν η εκθετικώς αυξανόμενη παραγωγή δεδομένων τα τελευταία χρόνια. Είναι πια οικονομικά και χρονικά συμφέρον να αποθηκεύονται τεράστιες ποσότητες δεδομένων σε βάσεις. Υπολογίζεται ότι η ποσότητα των δεδομένων που αποθηκεύεται παγκοσμίως σε βάσεις δεδομένων διπλασιάζεται κάθε είκοσι μήνες. Ο λόγος που γίνεται αυτό είναι γιατί η κατανόηση της λειτουργίας ενός συστήματος και η εξαγωγή συμπερασμάτων και συνακολούθως κανόνων για τη βέλτιστη λειτουργία του επιτυγχάνεται με την όσο το δυνατόν καλύτερη περιγραφή του. Η ποιότητα της περιγραφής αυτής σχετίζεται άμεσα με το πλήθος των δεδομένων που αναφέρονται στο σύστημα αφού είναι εύλογο ότι η πληθώρα δεδομένων εξετάζει περισσότερες παραμέτρους και καλύπτει περισσότερες καταστάσεις.

Σε αυτά τα δεδομένα κρύβεται συχνά πολύ χρήσιμη πληροφορία, πληροφορία που είναι ικανή να οδηγήσει στη βελτίωση της λειτουργίας του συστήματος. Αυτή παρουσιάζεται συνήθως με τη μορφή προτύπων, με τη μορφή δηλαδή μίας σειράς γεγονότων που συμβαίνουν σε τέτοια συχνότητα ώστε να αποκαλύπτουν μία σχέση μεταξύ τους. Πιο φορμαλιστικά όμως ως *πρότυπο (pattern)* εννοείται μία μορφή, ένα μοντέλο ή ένα σύνολο κανόνων που χρησιμοποιούνται για την παραγωγή γεγονότων από ένα σύνολο δεδομένων.

Όπως όμως είναι λογικό η εύρεση και η επεξεργασία αυτών των προτύπων είναι μία δύσκολη διεργασία λόγω του τεράστιου όγκου των δεδομένων. Έτσι, όχι μόνο υπάρχουν δεδομένα που δε συνεισφέρουν στη διαδικασία της γνώσης αλλά λόγω της παρουσίας τους καθιστούν και δυσχερέστερη την προσπάθεια εντοπισμού χρήσιμης πληροφορίας. Αυτός είναι και ένας από τους λόγους που η διαδικασία ανακάλυψης σχέσεων από σύνολα δεδομένα είναι συνήθως ημι-αυτόματη. Επιχειρείται, έτσι, η *απόκτηση γνώσης (knowledge discovery)* μέσω του συνδυασμού της υπολογιστικής δύναμης και της ανθρώπινης ικανότητας. Με άλλα λόγια, με την απόκτηση γνώσης γίνεται η αποτίμηση των αποτελεσμάτων και σχέσεων που παράγονται αυτόματα με μεθόδους εξόρυξης δεδομένων.

Οι παραπάνω τεχνικές εφαρμόζονται ευρέως στη *εκμάθηση μηχανών (machine learning)*. Η εκμάθηση μηχανών αποτελεί ένα υποκλάδο της τεχνητής νοημοσύνης και ασχολείται με την ανάπτυξη αλγορίθμων και τεχνικών που επιτρέπουν την εκμάθηση μηχανών. Γενικά, υπάρχουν δύο είδη μάθησης: επαγωγικό και παραγωγικό. Οι επαγωγικές μέθοδοι εκμάθησης μηχανών ασχολούνται με την

εξαγωγή κανόνων και προτύπων από ογκώδη σύνολα δεδομένων. Ουσιαστικά, με τις επαγωγικές μεθόδους επιχειρείται η μετάβαση από το ειδικό στο γενικό· γίνεται δηλαδή η προσπάθεια να εξαχθούν κάποια συμπεράσματα από ένα υπάρχον σύνολο δεδομένων και να έχουν ισχύ και σε παρόμοια σύνολα δεδομένων που αναφέρονται στο ίδιο είδος προβλήματος.

Πολλές από τις μεθόδους που χρησιμοποιούνται είναι στατιστικές. Ουσιαστικά, η εκμάθηση μηχανών σε πολλά σημεία επικαλύπτεται με τη στατιστική. Επιχειρείται η αναγνώριση τάσεων, συσχετίσεων, μη κανονικοτήτων και άλλων χαρακτηριστικών για την πραγματοποίηση προβλέψεων και την εξαγωγή συμπερασμάτων από τα δεδομένα. Από την άλλη, στην εκμάθηση μηχανών δεν απαιτείται ειδική δομή των δεδομένων για την επικύρωση και την επαναληψιμότητα των σχεδίων υπόθεσης. Ενώ, η κύρια διαφορά έγκειται στο γεγονός ότι ενδιαφέρουν άμεσα οι υπολογιστικές ιδιότητες των χρησιμοποιούμενων στατιστικών μεθόδων. Για παράδειγμα, μία κρίσιμη παράμετρος είναι η υπολογιστική πολυπλοκότητα μίας μεθόδου. Με απλά λόγια, η συμπεριφορά του αλγορίθμου όσο το μέγεθος της εισόδου αυξάνει. Κατά πόσο, δηλαδή, αυξάνουν οι απαιτήσεις σε μνήμη και ο χρόνος εκτέλεσής του.

Η εκμάθηση μηχανών έχει ένα μεγάλο φάσμα εφαρμογών περιλαμβάνοντας την επεξεργασία φυσικής γλώσσας, την αναγνώριση φωνής και γραφικού χαρακτήρα, τις μηχανές αναζήτησης, την αναγνώριση αντικειμένων, την ανάλυση έρευνας αγοράς καθώς και ένα πλήθος εφαρμογών της Βιοπληροφορικής.

Μάλιστα, η εξόρυξη δεδομένων σε βιολογικά δεδομένα που σχετίζονται με προβλήματα Βιοπληροφορικής είναι ιδιαιτέρως ενδιαφέρουσα λόγω του τεράστιου όγκου τους και τις προφανείς ευεργετικές συνέπειες που έχει η εξαγωγή αξιόπιστων συμπερασμάτων από τέτοιου είδους δεδομένα.

Συνήθως, η διαδικασία απόκτησης γνώσης ακολουθεί δύο στάδια. Στο αρχικό, γίνεται προσπάθεια φιλτραρίσματος των σχετικών με το πρόβλημα γνωρισμάτων. Τα σύνολα δεδομένων αποτελούνται κυρίως από γνωρίσματα που δεν έχουν σχέση με το συγκεκριμένο πρόβλημα. Στην ουσία, δεν προσφέρουν καμιά πληροφορία για την εξαγωγή χρήσιμων συμπερασμάτων. Όπως θα δούμε παρακάτω, το πρόβλημα αυτό γίνεται εντονότερο σε δεδομένα βιολογικής προέλευσης εξαιτίας της παρουσίας ενός πολύ μεγάλου αριθμού άσχετων προς το βιολογικό πρόβλημα γνωρισμάτων και τη μεγάλη διαφορά μεταξύ του πλήθους των γνωρισμάτων και του πλήθους των παραδειγμάτων. Έτσι, επιχειρείται η απόρριψη αυτών των γνωρισμάτων (και των τιμών που αντιπροσωπεύουν) ώστε να μειωθεί σημαντικά το υπολογιστικό κόστος και να καταστεί ευκρινέστερη η διακριτική ισχύ των σχετικών προς το πρόβλημα γνωρισμάτων.

Στο τελικό στάδιο, επιχειρείται η εκμάθηση του εναπομείναντος συνόλου δεδομένων μέσω ενός αλγορίθμου εκμάθησης. Σε αυτό το στάδιο υπάρχουν αρκετές ποιοτικές παράμετροι που καθορίζουν την αξιοπιστία και τη γενικότητα των συμπερασμάτων. Επιγραμματικά, θα αναφέρουμε ότι η ακρίβεια (accuracy) αποτελεί το συνηθέστερο και απλούστερο μέτρο αξιοπιστίας. Με τον όρο *ακρίβεια* εννοούμε το λόγο του αριθμού των σωστών προβλέψεων που έγιναν στα παραδείγματα ενός συνόλου ελέγχου (test set) προς το συνολικό αριθμό παραδειγμάτων του συνόλου ελέγχου. Άλλες παράμετροι, που αναφέρονται σε ετικέτες της κλάσης (class labels), σε όλες δηλαδή τις δυνατές τιμές της κλάσης του προβλήματος, είναι οι: recall (ή αλλιώς TP rate), precision, FP rate και F-measure. Επίσης, μπορεί να αποτιμηθεί και το περιθώριο γενίκευσης ενός κανόνα. Θα μιλήσουμε όμως για αυτές τις παραμέτρους αναλυτικότερα παρακάτω.

4.2 Φιλτράρισμα γνωρισμάτων

Όπως προαναφέραμε, στο πρώτο στάδιο της διαδικασίας απόκτησης γνώσης γίνεται προσπάθεια φιλτραρίσματος των σχετικών με το πρόβλημα γνωρισμάτων. Η αναγνώριση των πιο πληροφοριακών γνωρισμάτων είναι εξέχουσας σημασίας αφού παρέχεται η δυνατότητα πιο αξιόπιστων αποτελεσμάτων και εμβάθυνσης στο υπόβαθρο του προβλήματος. Στην περίπτωση βιολογικών προβλημάτων, ο στόχος αφορά την καλύτερη κατανόηση των γονιδιακών λειτουργιών.

Έτσι, σε αυτό το σημείο, θα κάνουμε εισαγωγή στην έννοια του φιλτραρίσματος γνωρισμάτων. Πιο συγκεκριμένα, θα παραθέσουμε αρχικά τα πλεονεκτήματα από τη χρησιμοποίηση του φιλτραρίσματος γνωρισμάτων, τους διάφορους τύπους και μοντέλα που υπάρχουν, ενώ τέλος θα παρουσιάσουμε αναλυτικά τις μεθόδους φιλτραρίσματος που χρησιμοποιήθηκαν σε αυτήν τη διπλωματική εργασία.

4.2.1 Λόγοι χρησιμοποίησης φιλτραρίσματος γνωρισμάτων

Οι κυριότεροι λόγοι εφαρμογής του φιλτραρίσματος γνωρισμάτων είναι οι παρακάτω:

1. Βελτίωση της ακρίβειας ταξινόμησης (classification accuracy) στις περισσότερες των περιπτώσεων [25]. Κάτι που οφείλεται κυρίως στη μείωση της διαστατικότητας με τη διατήρηση των πιο πληροφοριακών γνωρισμάτων

και την ταυτόχρονη απόρριψη των πιο άσχετων και θορυβωδών [29]. Μειώνεται, έτσι, δραστικά η επίδραση της "κατάρας της διαστατικότητας" που οδηγεί κυρίως σε φαινόμενα υπερ-εκπαίδευσης, αφού επιστρατεύονται πάρα πολλά γνωρίσματα για την εξήγηση ενός προβλήματος με πολύ λίγα παραδείγματα. Ενώ, όσον αφορά τα άσχετα γνωρίσματα, η παρουσία τους όχι μόνο εμποδίζει την εμφάνιση της διακριτικής ισχύος των σχετικών (προς το πρόβλημα γνωρισμάτων) αλλά και την εξαγωγή συσχετίσεων μεταξύ τους [9]. Στο σημείο αυτό, τονίζοντας τη σημασία του φιλτραρίσματος γνωρισμάτων στην ακρίβεια ταξινόμησης να αναφέρουμε ότι έχει παρατηρηθεί ότι για ένα συγκεκριμένο μηχανισμό επιλογής γνωρισμάτων οι ακρίβειες διαφορετικών αλγορίθμων ταξινόμησης δεν διαφέρουν πολύ σε αντίθεση με τη σημαντική επίδραση που έχουν οι διαφορετικοί μηχανισμοί επιλογής γνωρισμάτων για ένα δεδομένο αλγόριθμο ταξινόμησης όσον αφορά την ακρίβεια ταξινόμησης [14].

2. Μείωση του υπολογιστικού κόστους.
3. Καλύτερη κατανόηση της διαδικασίας που παρήγαγε τα δεδομένα [9].
4. Σε προβλήματα μικροσυστοιχιών, τα επίλεκτα γνωρίσματα που αντιπροσωπεύουν γονίδια είναι γνωστά ως *γονίδια – δείκτες (marker genes)*. Με το φιλτράρισμα γνωρισμάτων, παρέχεται η δυνατότητα στους βιολόγους για περαιτέρω μελέτη των διαδράσεων μεταξύ των γονιδίων. Αυτό οφείλεται γιατί με την επιλογή ομάδων γνωρισμάτων (αντί για μεμονωμένα γνωρίσματα) ενσωματώνεται πληροφορία για τη διάδραση μεταξύ των γονιδίων και συνυπολογίζονται γονίδια που δρουν συμπληρωματικά για κάθε βιολογικό μονοπάτι [14].
5. Μελέτη της δομικής και λειτουργικής συμπεριφοράς γνωστών μαρκαρισμένων γονιδίων – δεικτών με σκοπό τον καθορισμό της λειτουργικότητας και άλλων γονιδίων [14].
6. Μείωση του κόστους της κλινικής επεξεργασίας των αποτελεσμάτων. Στη δεδομένη στιγμή, είναι αδύνατο να χρησιμοποιηθούν στα διαγνωστικά εργαστήρια μικροσυστοιχίες που περιέχουν χιλιάδες γονίδια. Έτσι, για να καταστεί αυτό δυνατό θα πρέπει να μειωθεί ο αριθμός των γονιδίων [25].

4.2.2 Παρουσίαση της ταξινόμησης φιλτραρίσματος γνωρισμάτων

Όπως θα έγινε αντιληπτό υπάρχουν πολλοί λόγοι για την εφαρμογή του φιλτραρίσματος γνωρισμάτων. Σε αυτό το σημείο, λοιπόν, θα αφιερώσουμε χώρο στην εκτενέστερη παρουσίαση των μεθόδων επιλογής γνωρισμάτων.

Γενικά, μία μέθοδος φιλτραρίσματος γνωρισμάτων επιλέγει υποψήφια γνωρίσματα (ή ομάδες γνωρισμάτων) βάσει μίας *διαδικασίας παραγωγής* και σύμφωνα με ένα *κριτήριο αξιολόγησης*.

Ανάλογα με το *κριτήριο αξιολόγησης* και τη *διαδικασία παραγωγής*, οι διάφορες μέθοδοι επιλογής γνωρισμάτων κατηγοριοποιούνται όπως φαίνεται στον Πιν. V ως εξής:

Πίνακας V. Ταξινόμηση μεθόδων φιλτραρίσματος γνωρισμάτων

Κριτήριο Αξιολόγησης		
Μοντέλο	Μέτρο	Παραδείγματα
Φιλτραρίσματος (Filter)	*Απόσταση (Distance)* – ο βαθμός διαχωρισμού μεταξύ των κλάσεων.	Κριτήριο του Fischer, στατιστική test και Sincich, Relief
	Συνοχή (Consistency) – βρίσκει ένα ελάχιστο αριθμό γνωρισμάτων που μπορούν να διακρίνουν τις κλάσεις.	Ρυθμός ασυνέχειας (Inconsistency rate)
	Συσχέτιση (Correlation) – μετράει την ικανότητα μία μεταβλητή να προβλέψει μία άλλη.	Συντελεστής συσχέτισης Pearson, κέρδος σε πληροφορία (information gain)
Περιτυλίγματος (Wrapper)	*Ταξινόμηση (Classification)* – η απόδοση ενός επαγωγικού αλγορίθμου ταξινόμησης.	Δένδρο απόφασης και naïve Bayes

Διαδικασία Παραγωγής

Τύπος	Αναζήτηση	Παραδείγματα
Κατάταξη Μεμονωμένη γνωρισμάτων (Individual Ranking)	Μετράει τη σχέση κάθε μεμονωμένου γνωρίσματος (με το συγκεκριμένο πρόβλημα).	Τα περισσότερα φίλτρα
Επιλογή Υποσυνόλου γνωρισμάτων (Subset Selection)	*Πλήρης* – εξετάζει όλες τις πιθανές λύσεις.	Προέκταση και Περιορισμός (Branch and Bound), καλύτερη πρώτη αναζήτηση (best-first search, BFS)
	Ευριστική. *Ντετερμινιστική* – χρησιμοποιεί μία “άπληστη” (greedy) τεχνική για την επιλογή των γνωρισμάτων βάσει τοπικών αλλαγών.	Διαδοχική προς τα εμπρός επιλογή (sequential forward selection, SFS), διαδοχική προς τα πίσω επιλογή (sequential backward selection, SBS), sequential floating forward selection (SFFS), sequential floating backward selection (SFBS)
	Μη ντετερμινιστική – προσπαθεί να βρει τη βέλτιστη λύση με ένα τυχαίο τρόπο.	Simulated annealing (SA), Las Vegas Filter (LVF), Genetic Algorithms (GA), Tabu Search (TS)

4.2.2.1 Κριτήρια αξιολόγησης

Αφού λοιπόν επιχειρήθηκε συνοπτικά μία ταξινόμηση των μεθόδων επιλογής γνωρισμάτων σύμφωνα με τη *διαδικασία παραγωγής* και το *κριτήριο αξιολόγησης*, κρίνεται αναγκαίο να αναλύσουμε περαιτέρω καθένα από αυτά. Θα ξεκινήσουμε με την εξέταση του κριτηρίου αξιολόγησης.

Ένα κριτήριο αξιολόγησης χρησιμοποιείται για να μετρήσει την διαχωριστική ικανότητα των υποψήφιων γνωρισμάτων. Γενικά, υπάρχουν τα *μοντέλα φιλτραρίσματος* και τα *μοντέλα περιτυλίγματος*. Το μοντέλο φιλτραρίσματος επιλέγει τα καλύτερα γνωρίσματα χωρίς να χρησιμοποιείται κάποιος αλγόριθμος εκμάθησης. Αντίθετα, ένα μοντέλο περιτυλίγματος χρησιμοποιεί ένα αλγόριθμο εκμάθησης σαν ένα “μαύρο” κουτί που το περιβάλλει γύρω από τη διαδικασία επιλογής με στόχο την αξιολόγηση υποσυνόλων γνωρισμάτων βάσει της προβλεπτικής τους ακρίβειας.

4.2.2.1.1 Μοντέλα φιλτραρίσματος

Τα μοντέλα φιλτραρίσματος επιλέγουν τα καλά γνωρίσματα βάσει ενός μέτρου που χρησιμοποιεί την εσωτερική πληροφορία του συνόλου δεδομένων. Αυτά τα μέτρα δείχνουν τη σχέση του γνωρίσματος με την κλάση του προβλήματος.

Σε γενικές γραμμές, τα μέτρα αυτά χωρίζονται σε μέτρα *απόστασης (distance)*, *συνοχής (consistency)* και *συσχέτισης (correlation)*.

- *Μέτρα απόστασης*

Τα μέτρα απόστασης ποσοτικοποιούν την ικανότητα ενός γνωρίσματος ή ενός συνόλου γνωρισμάτων να διαχωρίζουν τις διαφορετικές κλάσεις μεταξύ τους.

Ένα παράδειγμα μέτρου απόστασης είναι το κριτήριο του Fischer:

$$J(k) = \frac{(\mu_k{}^I - \mu_k{}^{II})^2}{(\sigma_k{}^I + \sigma_k{}^{II})^2}$$

,όπου $(\mu_k{}^I, \mu_k{}^{II})$ είναι ο μέσος όρος και $(\sigma_k{}^I, \sigma_k{}^{II})$ η τυπική απόκλιση όλων των γνωρισμάτων στα παραδείγματα που ανήκουν στην κλάση I και II, αντίστοιχα.

Άλλοι μέθοδοι επιλογής που χρησιμοποιούν μέτρα απόστασης είναι η στατιστική *test*, *F* και κ^2. Ένας από τους αλγορίθμους που χρησιμοποιήσαμε είναι ο Relief, που όπως θα δούμε και παρακάτω, χρησιμοποιεί ένα μέτρο απόστασης για

τον υπολογισμό της ικανότητας ενός γνωρίσματος να διαχωρίσει δύο παραδείγματα που είναι κοντά το ένα στο άλλο αλλά ανήκουν σε διαφορετικές κλάσεις.

- *Μέτρα συνοχής*

Τα μέτρα συνοχής βρίσκουν ένα ελάχιστο αριθμό γνωρισμάτων μου μπορούν να διαχωρίσουν τις κλάσεις στον ίδιο βαθμό με το πλήρες σύνολο γνωρισμάτων. Ως *ασυνέχεια* ορίζεται το γεγονός δύο παραδείγματα να έχουν το ίδιο πρότυπο, για παράδειγμα ίδιες τιμές γνωρισμάτων, αλλά διαφορετικές ετικέτες κλάσης.

Από τη στιγμή που το πλήρες σύνολο γνωρισμάτων έχει το χαμηλότερο ρυθμό ασυνέχειας, επιχειρείται να επιλεχθεί ένας αριθμός γνωρισμάτων με το μικρότερο δυνατό ρυθμό ασυνέχειας.

Αναλυτικότερα, αν η διαφορά στις τιμές ενός γνωρίσματος μεταξύ δύο παραδειγμάτων που ανήκουν σε διαφορετικές ετικέτες κλάσης είναι Δf_i, τότε σε ένα σύνολο n γνωρισμάτων υποθέτουμε (χωρίς βλάβη της γενικότητας) ότι οι τιμές ανάμεσα στα δύο παραδείγματα για τα πρώτα m γνωρίσματα είναι διαφορετικές. Με λίγα λόγια: $\Delta f_j \neq 0 \quad , j = \{1,...,m\}$.

Επομένως, ο ρυθμός ασυνέχειας για το πλήρες σύνολο γνωρισμάτων είναι $ir = \frac{\sum_{i=1}^{n} \Delta f_i}{n} = \frac{\sum_{i=1}^{m} \Delta f_i}{n}$ ενώ για το υποσύνολο με τον ελάχιστο ρυθμό ασυνέχειας έχουμε $ir = \frac{\sum_{i=1}^{m} \Delta f_i}{m}$ με $m < n$.

Παρόλα αυτά, το υπολογιστικό κόστος του βέλτιστου υποσυνόλου γνωρισμάτων βάσει μέτρων συνοχής είναι πολύ υψηλό.

- *Μέτρα συσχέτισης*

Όπως προαναφέρθηκε, τα μέτρα συσχέτισης μετράνε την ικανότητα πρόβλεψης μίας μεταβλητής από μία άλλη. Αυτά τα μέτρα στηρίζονται συνήθως στη γραμμική συσχέτιση ή στη θεωρία πληροφοριών.

Αναφορικά με τα μέτρα που σχετίζονται με τη γραμμική συσχέτιση, το πιο γνωστό είναι ο συντελεστής συσχέτισης του Pearson. Για δύο συνεχείς μεταβλητές X , Y , ο συντελεστής συσχέτισης ορίζεται ως:

$$r_{pearson} = \frac{E[(X-\overline{X})(Y-\overline{Y})]}{\sigma_X \sigma_Y}$$

$$\mu\varepsilon: \quad \overline{X} = E[X] = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$\overline{Y} = E[Y] = \sum_{i=1}^{n} \frac{y_i}{n}$$

$$E[(X-\overline{X})(Y-\overline{Y})] = \sum_{i=1}^{n} \frac{(x_i-\overline{X})}{n} \cdot \frac{(y_i-\overline{Y})}{n}$$

$$\sigma_X = \sqrt{E[(X-\overline{X})^2]} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2}{n}}$$

$$\sigma_Y = \sqrt{E[(Y-\overline{Y})^2]} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\overline{Y})^2}{n}}$$

Στη θεωρία πληροφοριών, η εντροπία είναι ένα μέτρο της αβεβαιότητας της τυχαίας μεταβλητής. Μία μέθοδος που σχετίζεται με τη θεωρία πληροφοριών, η οποία θα αναλυθεί σε βάθος στη συνέχεια, ονομάζεται *κέρδος σε πληροφορία* και υπολογίζει τη μείωση (κέρδος) σε εντροπία μίας μεταβλητής Y δεδομένου ότι είναι γνωστή μία άλλη μεταβλητή X.

4.2.2.1.2 Μοντέλα περιτυλίγματος

Σε αντίθεση με τα μοντέλα φιλτραρίσματος, τα μοντέλα περιτυλίγματος χρησιμοποιούν αλγορίθμους εκμάθησης για την κατάταξη των γνωρισμάτων. Για την ακρίβεια, η ακρίβεια ταξινόμησης αποτελεί ένα κριτήριο αξιολόγησης.

Για ένα σύνολο n γνωρισμάτων: $\{f_1, f_2, ..., f_n\}$ εξετάζονται 2^n διαφορετικά υποσύνολα: $\{ \quad \varnothing, \ \{f_1\}, \ \{f_2\}, \ ..., \ \{f_n\}, \ \{f_1, f_2\}, \ \{f_1, f_3\}, \ ..., \ \{f_1, f_n\}, \ ..., \ \{f_1, f_2, ..., f_n\} \quad \}$. Το υποσύνολο με τη μεγαλύτερη ακρίβεια ταξινόμησης θα αναγνωριστεί ως το βέλτιστο υποσύνολο.

Παρά τα πλεονεκτήματα των μοντέλων αυτών, οι δύο κυριότερες αδυναμίες τους εστιάζονται στο υψηλό υπολογιστικό κόστος και στη μικρή δυνατότητα γενίκευσης. Όπως κατέστη παραπάνω σαφές, η εξέταση των 2^n σεναρίων επιφέρει

μεγάλη χρονική πολυπλοκότητα. Τέλος, το βέλτιστο υποσύνολο είναι χαρακτηριστικό κάθε ταξινομητή. Με άλλα λόγια, με την εφαρμογή διαφορετικών ταξινομητών είναι δυνατό να εξαχθούν διαφορετικά βέλτιστα υποσύνολα.

4.2.2.2 Διαδικασίες παραγωγής

Αφού λοιπόν εξετάσαμε διεξοδικά τα κριτήρια υπολογισμού καλούμαστε να αναλύσουμε τις διαδικασίες παραγωγής και πιο συγκεκριμένα τους δύο βασικούς τύπους: *κατάταξη μεμονωμένων γνωρισμάτων* και *επιλογή υποσυνόλων γνωρισμάτων*.

4.2.2.2.1 Κατάταξη Μεμονωμένων Γνωρισμάτων (Individual Feature Ranking, IFR)

Στις μεθόδους κατάταξης μεμονωμένων γνωρισμάτων, κάθε γνώρισμα συγκρίνεται για τη σχέση του με την κλάση του προβλήματος. Τα γνωρίσματα έπειτα κατατάσσονται και επιλέγονται τα βέλτιστα. Τα περισσότερα φίλτρα που στοχεύουν μόνο στην απομάκρυνση των άσχετων γνωρισμάτων ανήκουν σε αυτή την κατηγορία.

Οι μέθοδοι αυτοί προτιμώνται για την απλότητα, κλιμάκωση και εμπειρική επιτυχία που παρουσιάζουν. Επίσης, η χρονική τους πολυπλοκότητά βρίσκεται σε ανεκτά επίπεδα αφού απαιτείται μόνο ο υπολογισμός και η κατάταξη των n γνωρισμάτων.

4.2.2.2.2 Επιλογή Υποσυνόλων Γνωρισμάτων (Feature Subset Selection, FSS)

Οι μέθοδοι επιλογής υποσυνόλων γνωρισμάτων αναζητούν ένα υποσύνολο γνωρισμάτων με την καλύτερη απόδοση όσον αφορά την ταξινόμηση του συνόλου δεδομένων. Το πλεονέκτημα αυτών των τεχνικών είναι δεν λαμβάνεται υπόψη μόνο η σχέση των γνωρισμάτων αλλά και η συσχέτιση μεταξύ των γνωρισμάτων. Από την άλλη, η αναζήτηση σε πολλούς διαφορετικούς συνδυασμούς υποσυνόλων γνωρισμάτων επιβάλλει ένα μεγάλο υπολογιστικό κόστος.

Οι δύο διαφορετικοί τρόποι προσέγγισης για την επίλυση τέτοιων προβλημάτων είναι: *πλήρης* και *ευριστική* αναζήτηση.

- *Πλήρης Αναζήτηση*

Για ένα σύνολο n γνωρισμάτων: $\{f_1, f_2, ..., f_n\}$ εξετάζονται 2^n διαφορετικά υποσύνολα: $\{ \varnothing, \{f_1\}, \{f_2\}, ..., \{f_n\}, \{f_1, f_2\}, \{f_1, f_3\}, ..., \{f_1, f_n\}, ..., \{f_1, f_2, ..., f_n\} \}$. Παρόλα αυτά, έχουν αναπτυχθεί και άλλες τεχνικές όπως οι *Προέκταση και Περιορισμός (Branch and Bound)* και *καλύτερη πρώτη αναζήτηση (best-first search, BFS)* που ελαχιστοποιούν το χρόνο αναζήτησης χωρίς να μειώνουν την πιθανότητα εύρεσης της βέλτιστης λύσης. Εδώ, παρόλο που η τάξη του χώρου αναζήτησης παραμένει $O(2^n)$ υπολογίζονται λιγότερα υποσύνολα. Σε περιπτώσεις όμως πολύ μεγάλου όγκου δεδομένων η ιδέα της πλήρους αναζήτησης καθίσταται μη πρακτική.

- *Ευριστική Αναζήτηση*

Άλλοι αλγόριθμοι αναζήτησης χρησιμοποιούν ευριστικές τεχνικές. Αυτές κατηγοριοποιούνται σε *ντετερμινιστικές ευριστικές αναζητήσεις* και σε *μη-ντετερμινιστικές ευριστικές αναζητήσεις*.

Οι μέθοδοι *ντετερμινιστικής ευριστικής αναζήτησης* βασίζονται σε μία λογική *βήμα προς βήμα*. Δηλαδή, σε κάθε βήμα της διαδικασίας αναζήτησης μόνο τα υπολειπόμενα γνωρίσματα που έχουν επιλεχθεί (απορριφθεί) είναι διαθέσιμα για επιλογή (απόρριψη). Τοπικές αλλαγές στην υπάρχουσα κατάσταση αποτελούν το κριτήριο επιλογής ή απόρριψης ενός συγκεκριμένου γνωρίσματος.

Σε γενικές γραμμές, οι μέθοδοι αυτές είναι υπολογιστικά συμφέρουσες και ανθεκτικές απέναντι στην υπερ-εκπαίδευση αλλά εξαιτίας της “άπληστης” στρατηγικής που υιοθετούν συχνά καταφέρνουν να εντοπίζουν μόνο τοπικά βέλτιστα.

Από την άλλη, οι μέθοδοι *μη- ντετερμινιστικής ευριστικής αναζήτησης* επιλέγουν το βέλτιστο ή (τοπικά) βέλτιστο υποσύνολο με ένα τυχαίο τρόπο και σε προκαθορισμένο αριθμό επαναλήψεων. Οι τεχνικές αυτές καλούνται επίσης ως *βέλτιστες αναζητήσεις* χάρη στην ικανότητά τους να βρίσκουν ολικά ή τοπικά βέλτιστες λύσεις [9].

4.2.3 Ανάλυση χρησιμοποιούμενων αλγορίθμων φιλτραρίσματος γνωρισμάτων

Στο σημείο αυτό, θα παρουσιάσουμε αναλυτικά τη βάση της θεωρίας για τους αλγορίθμους Κέρδος σε Πληροφορία (Information Gain), Χι-Τετράγωνο (Chi-Square) και Relief, όπου και χρησιμοποιήθηκαν στη διπλωματική αυτή θέση.

4.2.3.1 Κέρδος σε Πληροφορία (Information Gain)

Η μέθοδος Information Gain μας πληροφορεί για τον αριθμό των bits που εξοικονομούμε για την πρόβλεψη της κλάσης ενός προβλήματος εάν γνωρίζουμε την τιμή ενός γνωρίσματος.

Έστω ότι συμβολίζουμε με C την κλάση του προβλήματος και με $c_i, \quad i = \{1,2,...,k\}$ τις διαφορετικές ετικέτες της κλάσης και με F το γνώρισμα την επίδραση του οποίου θα εξετάσουμε πάνω στο συγκεκριμένο πρόβλημα.

Η εντροπία γενικά μίας τυχαίας μεταβλητής X ($H(X)$) συμβολίζει τον ελάχιστα απαιτούμενο αριθμό bits πληροφορίας για την πρόβλεψη της τιμής της μεταβλητής X και ισούται με το αρνητικό άθροισμα των πιθανοτήτων να υπάρξει μία συγκεκριμένη τιμή για τη μεταβλητή X πολλαπλασιασμένων με το λογάριθμο με βάση το 2 αυτών των πιθανοτήτων. Έτσι, αν η μεταβλητή X μπορεί να πάρει k διαφορετικές τιμές, τότε ισχύει:

$$H(X) = -\sum_{i=1}^{k} P(X = x_i) \log_2[P(X = x_i)]$$

Σε γενικές γραμμές, υψηλή εντροπία σημαίνει ομοιόμορφη κατανομή. Επομένως, μία κατανομή που δεν μας βοηθάει πολύ στην πρόβλεψη της τιμής της μεταβλητής X. Από την άλλη, χαμηλή εντροπία σηματοδοτεί μία “μη ομαλή” κατανομή που συνεισφέρει περισσότερο στην πρόβλεψη της τιμής της μεταβλητής X.

Στην περίπτωση των συνόλων δεδομένων, ως $H(C)$ ορίζουμε την εντροπία της κλάσης C χωρίς να έχουμε κανένα στοιχείο για κάποιο γνώρισμα του συνόλου. Επειδή ακριβώς δεν χρησιμοποιείται υπάρχουσα γνώση, με άλλα λόγια δεν

χρησιμοποιείται η πληροφορία που μας δίνουν τα γνωρίσματα που πολύ πιθανόν να καθορίζει τις ετικέτες της κλάσης είναι αναμενόμενο ότι αυτή η εντροπία θα είναι μεγαλύτερη από την αντίστοιχη όπου θα χρησιμοποιείτο εσωτερική πληροφορία του συνόλου δεδομένων.

$$H(C) = -\sum_{i=1}^{k} P(C = c_i) \log_2[P(C = c_i)] \quad (4.1)$$

Αυτό που μας ενδιαφέρει στην ουσία, στην επιλογή των καταλληλότερων γνωρισμάτων είναι ποια γνωρίσματα παρέχουν καλύτερη προβλεπτική ικανότητα για την κλάση του προβλήματος. Με απλά λόγια, με ποιο τρόπο επηρεάζεται η εντροπία της κλάσης C εάν κάποιο από τα γνωρίσματα είναι γνωστό. Αυτό φορμαλιστικά μεταφράζεται με το συμβολισμό: $H(C \mid F)$.

Πριν όμως προχωρήσουμε στον υπολογισμό της $H(C \mid F)$ κρίνεται ορθό να σταθούμε λιγάκι στον υπολογισμό της εντροπίας της κλάσης C εφόσον γνωρίζουμε μία συγκεκριμένη τιμή του γνωρίσματος F ($F = v$).

Αυτό μεταφράζεται ως εξής:

$$H(C \mid F = v) = -\sum_{i=1}^{k} P(C = c_i \mid F = v) \log_2[P(C = c_i \mid F = v)]$$

και αφορά τον υπολογισμό της εντροπίας μόνο στις περιπτώσεις όπου συμβαίνει ταυτόχρονα να ισχύει $F = v$.

Έτσι πια μπορούμε να διατυπώσουμε ότι η εντροπία της κλάσης C με δεδομένο το γνώρισμα F αποτελεί το μέσο όρο όλων των επιμέρους εντροπιών με δεδομένη κάθε δυνατή τιμή του γνωρίσματος F :

$$H(C \mid F) = \sum_{v \in V} P(F = v) H(C \mid F = v) \Leftrightarrow$$

$$H(C \mid F) = \sum_{v \in V} P(F = v)[-\sum_{i=1}^{k} P(C = c_i \mid F = v) \log_2[P(C = c_i \mid F = v)]] \Leftrightarrow$$

$$H(C \mid F) = -\sum_{v \in V} \sum_{i=1}^{k} P(F = v) P(C = c_i \mid F = v) \log_2[P(C = c_i \mid F = v)] \quad (4.2)$$

Έχουμε φτάσει στο σημείο όπου έχουμε ορίσει την εντροπία μίας μεταβλητής και την εντροπία υπό συνθήκη. Το τελευταίο πια που απομένει είναι ο ορισμός του

κέρδους σε πληροφορία (Information Gain). Συνεπώς, ως Information Gain ορίζουμε τη διαφορά $H(C \mid F)$ από την $H(C)$ ή με άλλα λόγια πόσα bits εξοικονομούμε για την πρόβλεψη της κλάσης C εάν γνωρίζουμε το γνώρισμα F :

$$IG(C \mid F) = H(C) - H(C \mid F) \overset{(4.1),(4.2)}{\Leftrightarrow}$$

$$IG(C \mid F) = -\sum_{i=1}^{k} P(C = c_i) \log_2[P(C = c_i)] + \sum_{v \in V} \sum_{i=1}^{k} P(F = v) P(C = c_i \mid F = v) \log_2[P(C = c_i \mid F = v)]$$

Αξίζει να σημειωθεί ότι αυτή η μέθοδος απαιτεί την διακριτοποίηση των αριθμητικών γνωρισμάτων [25].

4.2.3.2 Χι Τετράγωνο (Chi Square, Chi2)

Η μέθοδος Chi2 χρησιμοποιεί την στατιστική του χ^2 ώστε να διακριτοποιήσει αριθμητικά γνωρίσματα επαναλαμβανόμενα έως ότου βρεθούν ανακολουθίες στα δεδομένα και με αυτό τον τρόπο επιτυγχάνει την επιλογή γνωρισμάτων (μέσω της διακριτοποίησης). Με τον όρο *ανακολουθίες* εννοούμε όταν δύο πρότυπα είναι ακριβώς ίδια αλλά παρόλα αυτά ταξινομούνται σε διαφορετικές κατηγορίες.

Στον Πιν. VI παρουσιάζεται ο αλγόριθμος Chi2, ο οποίος χωρίζεται σε δύο φάσεις. Θα συνεχίσουμε περιγράφοντας αρχικά την πρώτη φάση.

Στην πρώτη φάση, ο αλγόριθμος ξεκινά με ένα υψηλό επίπεδο σημαντικότητας (`sigLevel`), για παράδειγμα 0.5, για τη διακριτοποίηση όλων των αριθμητικών γνωρισμάτων. Κάθε γνώρισμα ταξινομείται σύμφωνα με τις τιμές του (`sort(attribute, data)`).

Έπειτα, η ακόλουθη διαδικασία εκτελείται:

1. υπολογισμός της τιμής χ^2 (`chi-sq-calculation(attribute, data)`) με βάση την εξίσωση (4.3) για κάθε ζεύγος γειτονικών διαστημάτων (στην αρχή, κάθε πρότυπο τοποθετείται στο δικό του διάστημα που περιέχει μόνο μία τιμή γνωρίσματος).

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{(A_{ij}-E_{ij})^2}{E_{ij}} \quad (4.3)$$

k = *αριθμός των κλάσεων (ετικετών της κλάσης)*

A_{ij} = *αριθμός προτύπων στο* $i^{\text{οστό}}$ *διάστημα και στην* $j^{\text{οστή}}$ *κλάση*

$R_i = \sum_{j=1}^{k} A_{ij}$ = *αριθμός προτύπων στο* $i^{\text{οστό}}$ *διάστημα*

$C_j = \sum_{i=1}^{2} A_{ij}$ = *αριθμός προτύπων στην* $j^{\text{οστή}}$ *κλάση*

$N = \sum_{i=1}^{2} R_i$ = *συνολικός αριθμός προτύπων*

E_{ij} = *αναμενόμενη συχνότητα του* A_{ij} =

$$\frac{(αριθμος\ προτυπων\ στο\ i^{οστο}\ διαστημα)\cdot(αριθμος\ προτυπων\ στην\ j^{οστη}\ κλαση)}{συνολικος\ αριθμος\ προτυπων} = \frac{R_i \cdot C_j}{N}$$

- *Εάν κάποιο από τα* R_i *ή* C_j *ισούται με 0, τότε το* E_{ij} *τίθεται ίσο με 0.1.*
- *Ο βαθμός ελευθερίας της στατιστικής* χ^2 είναι κατά ένα μικρότερος από τον αριθμό των κλάσεων.

2. Συγχώνευση του ζεύγους γειτονικών διαστημάτων με τη μικρότερη τιμή χ^2. Η συγχώνευση συνεχίζει έως ότου (`while( merge(data) )`) όλα τα ζεύγη διαστημάτων έχουν τιμές χ^2 που ξεπερνούν την παράμετρο που καθορίζεται από το επίπεδο σημαντικότητας (`sigLevel`). Αρχικά, το επίπεδο σημαντικότητας (`sigLevel`) τίθεται στο 0.5 με αντίστοιχη τιμή χ^2 στο 0.455 εφόσον ο βαθμός ελευθερίας είναι 1.
Η παραπάνω διαδικασία επαναλαμβάνεται με ένα μειωμένο επίπεδο σημαντικότητας (`sigLevel = decreSigLevel(sigLevel)`) έως ότου τα διακριτοποιημένα δεδομένα υπερβούν ένα ρυθμό ανακολουθίας *δ* (`do while( InConsistency(data) < δ )`). Το κατώφλι του χ^2 αυξάνει αυτόματα με τη μείωση του επιπέδου

σημαντικότητας (`sigLevel`). Ενώ, ο έλεγχος `do while( InConsistency(data) < δ )` εισάγεται ως ένα κριτήριο τερματισμού ώστε να διασφαλιστεί ότι το διακριτοποιημένο σύνολο δεδομένων αναπαριστά ακριβώς το αρχικό σύνολο. Έτσι, με τα δύο αυτά χαρακτηριστικά, ο αλγόριθμος Chi2 καθορίζει αυτόματα το κατώφλι του χ^2 που είναι τέτοιο ώστε να διατηρείται η πιστότητα με τα αρχικά δεδομένα.

Η δεύτερη φάση είναι μία πιο κομψή διαδικασία. Ξεκινώντας με το `sigLevel_0` (που έχει λάβει ουσιαστικά την προτελευταία τιμή του επιπέδου σημαντικότητας (`sigLevel`) πριν την έξοδο από το βρόχο `do while( InConsistency(data) < δ )`) , κάθε γνώρισμα `i` συνδέεται με ένα επίπεδο σημαντικότητας (`sigLvl[i] = sigLevel_0 for attribute i`). Έπειτα, ακολουθεί το στάδιο της συγχώνευσης των γνωρισμάτων: Για κάθε γνώρισμα i αφού πρώτα υπολογιστεί η χ^2 τιμή του (`chi-sq-initialization(attribute, data)`) και υποστεί συγχώνευση (`while( merge(data) )`), ελέγχεται εάν έχει σημειωθεί υπέρβαση του ρυθμού ανακολουθίας (`if( InConsistency(data) < δ )`). Εάν αυτό δεν έχει συμβεί, υπάρχει τότε περιθώριο περαιτέρω μείωσης του επιπέδου σημαντικότητας του γνωρίσματος `i` (`sigLvl[i]`). Έτσι, το επίπεδο σημαντικότητας του γνωρίσματος `i` μειώνεται (`sigLvl[i] = decreSigLevel(sigLvl[i])`) για τον επόμενο γύρο συγχώνευσης του γνωρίσματος `i`. Διαφορετικά, το γνώρισμα `i` δεν συμπεριλαμβάνεται για περαιτέρω συγχώνευση (`attribute i cannot be merged`).

Αυτή η διαδικασία συνεχίζεται έως ότου οι τιμές όλων των γνωρισμάτων δεν μπορούν πια να συγχωνευθούν (`do until no-attribute-can-be-merged`). Στο τέλος της δεύτερης φάσης, αν κάποιο από τα γνωρίσματα έχει συγχωνευθεί μόνο σε μία τιμή, τότε αυτό το γνώρισμα δεν είναι σχετικό με την αναπαράσταση του αρχικού συνόλου δεδομένων. Αποτέλεσμα των παραπάνω, είναι ότι με το τέλος της διακριτοποίησης έχει επιτευχθεί παράλληλα και η επιλογή γνωρισμάτων [12].

Πίνακας VI. Αλγόριθμος Chi2

```
Phase 1:

set sigLevel = 0.5;
do while( InConsistency(data) < δ )
{
  for each numeric attribute
  {
    sort(attribute, data);
    chi-sq-initialization(attribute, data);
    do
    {
      chi-sq-calculation(attribute, data);
    } while( merge(data) )
  }
  sigLevel_0 = sigLevel;
  sigLevel   = decreSigLevel(sigLevel);
}

Phase 2:

set all sigLvl[i] = sigLevel_0 for attribute i;
do until no-attribute-can-be-merged
{
  for each attribute i that can be merged
  {
    sort(attribute, data);
    chi-sq-initialization(attribute, data);
    do
    {
      chi-sq-calculation(attribute, data);
    } while( merge(data) )
     if( InConsistency(data) < δ )
        sigLvl[i]   = decreSigLevel(sigLvl[i]);
     else
        attribute i cannot be merged
  }
```

```
}
```

4.2.3.3 Relief

Η κύρια ιδέα της μεθόδου Relief εντοπίζεται στον υπολογισμό γνωρισμάτων βάσει της διαχωριστικής ικανότητας των τιμών τους πάνω σε κοντινά παραδείγματα. Για αυτό το σκοπό, για ένα συγκεκριμένο παράδειγμα, η μέθοδος ψάχνει τους δύο κοντινότερους γείτονές του: έναν από την ίδια κλάση (ίδια ετικέτα κλάσης για την ακρίβεια) που ονομάζεται *κοντινότερη επιτυχία* (*nearest hit*) και τον άλλο από διαφορετική κλάση (διαφορετική ετικέτα κλάσης για την ακρίβεια) που ονομάζεται *κοντινότερη αποτυχία* (*nearest miss*).

Στην πραγματικότητα, ο υπολογισμός Relief ενός γνωρίσματος A είναι μία προσέγγιση της ακόλουθης διαφοράς μεταξύ των πιθανοτήτων $P(different\ value\ of\ A|nearest\ instance\ from\ different\ class)$ και $P(different\ value\ of\ A|nearest\ instance\ from\ same\ class)$:

$$W[A] = P(different\ value\ of\ A|nearest\ instance\ from\ different\ class) - P(different\ value\ of\ A|nearest\ instance\ from\ same\ class)$$

Το σκεπτικό είναι ότι ένα καλό γνώρισμα, ένα γνώρισμα δηλαδή με μεγάλη διακριτική ισχύ, πρέπει να λαμβάνει διαφορετικές τιμές μεταξύ παραδειγμάτων που ανήκουν σε διαφορετικές κλάσεις και πρέπει να έχει την ίδια τιμή για παραδείγματα της ίδιας κλάσης [8]. Παραδοχή, αρκετά λογική αν κάποιος σκεφτεί ότι ένα γνώρισμα σχετικό με το εξεταζόμενο είδος προβλήματος πρέπει να έχει διαφορετική συμπεριφορά (άρα και τιμή) για κάθε πτυχή (άρα και κλάση) του προβλήματος. Επομένως, μία μεγάλη τιμή $W[A]$ σηματοδοτεί ένα γνώρισμα με ισχυρή διακριτική ικανότητα αφού τουλάχιστον ένα από τα παρακάτω δύο θα ισχύει· θα υπάρχει μεγάλη πιθανότητα, για ένα συγκεκριμένο παράδειγμα, η κοντινότερη αποτυχία του (nearest miss) να εμφανίζει διαφορετική τιμή για το γνώρισμα A (*υψηλή* $P(different\ value\ of\ A|nearest\ instance\ from\ different\ class)$) ή θα υπάρχει μικρή πιθανότητα η κοντινότερη επιτυχία του (nearest hit) να εμφανίζει διαφορετική τιμή για το γνώρισμα A (*μικρή* $P(different\ value\ of\ A|nearest\ instance\ from\ same\ class)$).

Ο πρωτότυπος αλγόριθμος Relief (Πιν. VII) είχε σχεδιαστεί για διακριτά και συνεχή γνωρίσματα και μόνο για δυαδικά προβλήματα. Παρόλα αυτά, υπάρχει η δυνατότητα επέκτασης σε πολυ-κλασικά προβλήματα. Πριν όμως επεκταθούμε σε προβλήματα πολλών κλάσεων κρίνεται σκόπιμο να αναφέρουμε περισσότερα για την ίδια τη μέθοδο θεωρώντας δυαδικά προβλήματα.

Όπως γράψαμε παραπάνω το βάρος ενός γνωρίσματος ($W[A]$) δίνεται από τη σχέση:

$$W[A] = P(different\ value\ of\ A | nearest\ instance\ from\ different\ class) - P(different\ value\ of\ A | nearest\ instance\ from\ same\ class)$$

Έτσι, εάν θεωρήσουμε ότι έχουμε ένα σύνολο εκπαίδευσης και αντλούμε m φορές από αυτό παραδείγματα με επανάθεση (ένα παράδειγμα κάθε φορά), τότε για κάθε γνώρισμα A, η πιθανότητα να παρουσιάζει διαφορετική τιμή για την κοντινότερη αποτυχία (nearest miss) του τρέχοντος παραδείγματος ισούται με τον αριθμό των φορών που σημειώθηκε κάποια διαφορά στην τιμή του γνωρίσματος A ανάμεσα στο τρέχον παράδειγμα και στην κοντινότερη αποτυχία του προς το συνολικό αριθμό (m) των φορών που διεξήχθη το πείραμα.

Ομοίως, για κάθε γνώρισμα A, η πιθανότητα να παρουσιάζει διαφορετική τιμή για την κοντινότερη επιτυχία (nearest hit) του τρέχοντος παραδείγματος ισούται με τον αριθμό των φορών που σημειώθηκε κάποια διαφορά στην τιμή του γνωρίσματος A ανάμεσα στο τρέχον παράδειγμα και στην κοντινότερη επιτυχία του προς το συνολικό αριθμό (m) των φορών που διεξήχθη το πείραμα.

Διαισθητικά, είναι πολύ πιο εύκολο να κατανοήσουμε τον ορισμό της πιθανότητας για διακριτές μεταβλητές. Η διαφορά στις τιμές ενός διακριτού γνωρίσματος ανάμεσα σε δύο παραδείγματα είναι είτε 1 (αν οι τιμές είναι διαφορετικές) είτε 0 (αν οι τιμές είναι ίδιες).

Φορμαλιστικά αυτό μεταφράζεται ως εξής: Η διαφορά ($diff$) στις τιμές ενός διακριτού γνωρίσματος A ανάμεσα σε δύο παραδείγματα I_1 και I_2 είναι:

$$diff(A, I_1, I_2) = \begin{cases} 0 & : value(A, I_1) = value(A, I_2) \\ 1 & : otherwise \end{cases}$$

Για συνεχείς μεταβλητές, η πραγματική διαφορά κανονικοποιείται στο διάστημα [0, 1]. Είναι με άλλα λόγια, η διαφορά στις τιμές του συνεχούς γνωρίσματος ανάμεσα σε δύο παραδείγματα προς το εύρος τιμών του συγκεκριμένου γνωρίσματος κατά

μήκος όλων των παραδειγμάτων. Είναι ευνόητο, ότι όσο πλησιάζουμε το 1, η διαφορά μεγαλώνει ενώ όσο πλησιάζουμε το 0, η διαφορά μικραίνει.

Φορμαλιστικά αυτό συμβολίζεται ως εξής:

$$diff(A, I_1, I_2) = \frac{\mid value(A, I_1) - value(A, I_2) \mid}{\max(A) - \min(A)}$$

όπου: $\max(A)$ και $\min(A)$ είναι αντίστοιχα η μέγιστη και η ελάχιστη τιμή του γνωρίσματος A κατά μήκος όλων των παραδειγμάτων.

Συμβολίζοντας με R το *τυχαία* επιλεγόμενο κάθε φορά παράδειγμα, με H και M την κοντινότερη επιτυχία και αποτυχία του αντίστοιχα και με $diff(A, R, X)$ τη διαφορά στις τιμές του γνωρίσματος A ανάμεσα στα παραδείγματα R και X (όπου $X = \{H, M\}$), μπορούμε να πούμε τα εξής:

$$diff(A, R, X) \in \begin{cases} \{0,1\} & ,\gamma\iota\alpha\ \delta\iota\alpha\kappa\rho\iota\tau\alpha\ \gamma\nu\omega\rho\iota\sigma\mu\alpha\tau\alpha \\ [0,1] & ,\gamma\iota\alpha\ \sigma\upsilon\nu\varepsilon\chi\eta\ \gamma\nu\omega\rho\iota\sigma\mu\alpha\tau\alpha \end{cases}$$

Κλείνοντας την παρένθεση συνεχίζουμε από το σημείο όπου ορίσαμε την πιθανότητα λέγοντας ότι είναι ο μέσος όρος των διαφορών στις τιμές του γνωρίσματος A ανάμεσα σε δύο παραδείγματα στο σύνολο των m επαναλήψεων που έλαβε χώρα το πείραμα.

Αν το πείραμα επαναλαμβάνεται m φορές, τότε ορίζουμε ως:

$$P(different\ value\ of\ A | nearest\ instance\ from\ different\ class) = \frac{\sum_{i=1}^{m} diff(A, R, M)}{m}$$

και

$$P(different\ value\ of\ A | nearest\ instance\ from\ same\ class) = \frac{\sum_{i=1}^{m} diff(A, R, H)}{m}$$

Η κατώτερη τιμή των παραπάνω δύο πιθανοτήτων (0) συναντάται όταν η διαφορά στις τιμές του γνωρίσματος A ανάμεσα σε δύο παραδείγματα είναι μηδενική και για τις m επαναλήψεις, ενώ η ανώτερη (1) όταν πάντα υπάρχει διαφορά στις τιμές του γνωρίσματος A ανάμεσα σε δύο παραδείγματα. Εννοείται, ότι στην περίπτωση των συνεχών μεταβλητών αυτό συμβαίνει όταν και αυτή η διαφορά είναι η

μέγιστη δυνατή (και ίση με το εύρος τιμών του συγκεκριμένου γνωρίσματος κατά μήκος όλων των παραδειγμάτων).

Έτσι, το βάρος του γνωρίσματος A ($W[A]$) παίρνει τιμές από -1 έως 1 ($W[A] \in [-1,1]$). Την τιμή $W[A]=1$ την παίρνει όταν $P(different\ value\ of\ A|nearest\ instance\ from\ different\ class)=1$ και $P(different\ value\ of\ A|nearest\ instance\ from\ same\ class)=0$. Με άλλα λόγια, αυτή είναι η ιδανική περίπτωση όπου το γνώρισμα έχει τη μέγιστη διακριτική ισχύ. Οι τιμές του γνωρίσματος ανάμεσα σε παραδείγματα που είναι πολύ κοντά αλλά ανήκουν σε διαφορετικές κλάσεις είναι διαφορετικές, ενώ οι τιμές του γνωρίσματος ανάμεσα σε παραδείγματα που είναι πολύ κοντά και ανήκουν στην ίδια κλάση είναι ίδιες. Η τιμή $W[A]=-1$ συναντάται όταν $P(different\ value\ of\ A|nearest\ instance\ from\ different\ class)=0$ και $P(different\ value\ of\ A|nearest\ instance\ from\ same\ class)=1$. Σε αυτήν την περίπτωση, το γνώρισμα δεν έχει διακριτική ισχύ.

Προφανώς, όσο ο αριθμός των επαναλήψεων μεγαλώνει (m), τόσο οι πιθανότητες προσεγγίζονται καλύτερα. Παρόλα αυτά, ο m δεν πρέπει να ξεπεράσει τον αριθμό των διαθέσιμων παραδειγμάτων εκπαίδευσης.

Πίνακας VII. Πρωτότυπος αλγόριθμος Relief

```
set all weights W[A] = 0;
for i = 1 to m
{
  randomly select an instance Ri;
  find nearest Hit (Hi) and nearest Miss (Mi);
   // comment: for each possible pair of instances
   //          (Ri and one of the remaining instances
   //          of the training set (let's call it X) )
   //          calculate their sum of differences over
   //          all attributes. This is the total distance.
   //
   //          Namely:
   //
   //          for each pair Ri, X
   //          {
   //             total_distance(Ri, X) = 0;
   //             for A = 1 to all_attributes
   //             {
   //                 total_distance(Ri, X)  =
   //                         total_distance(Ri, X)  + diff(A, Ri, X);
   //             }
   //          }
   //
   //          H ≡ X with min{ total_distance(Ri, X) } AND
   //                      class(X) ≡ class(Ri)
   //          while
   //          M ≡ X with min{ total_distance(Ri, X) } AND
   //                      class(X)    class(Ri)


  for A = 1 to all_attributes
  {
    W[A] = W[A] + diff(A, Ri, Mi)/m - diff(A, Ri, Hi)/m;
  }
```

}

Όπως προαναφέρθηκε, ο αλγόριθμος λειτουργεί αρχικά με την εύρεση του κοντινότερου παραδείγματος με το υπό εξέταση παράδειγμα που ανήκει σε διαφορετική κλάση (nearest miss) και του κοντινότερου παραδείγματος (με το υπό εξέταση παράδειγμα) που ανήκει στην ίδια κλάση (nearest hit). Αυτή όμως η πρακτική μπορεί να οδηγήσει σε αναξιόπιστα αποτελέσματα, ειδικά αν στο σύνολο δεδομένων υπάρχουν περιττά (redundant) και θορυβώδη (noisy) δεδομένα (με την έννοια ότι είναι πιθανόν να "πολώσουν" την αναζήτηση). Έτσι, υπήρξε ανάγκη για μία επέκταση του αλγορίθμου Relief (Relief-A) που ψάχνει τις r κοντινότερες επιτυχίες/αποτυχίες (hits/misses) και υπολογίζει το μέσο όρο της συνεισφοράς τους.

Εδώ πια αντί να έχουμε τη σχέση:

$$W[A] = P(different\ value\ of\ A | nearest\ instance\ from\ different\ class) - P(different\ value\ of\ A | nearest\ instance\ from\ same\ class)$$

ισχύει:

$$W[A] = P(different\ value\ of\ A | different\ class) - P(different\ value\ of\ A | same\ class) \quad (4.4)$$

Εντοπίζονται δηλαδή, τα r κοντινότερα παραδείγματα που ανήκουν σε διαφορετική κλάση, έπειτα τα r κοντινότερα παραδείγματα που ανήκουν στην ίδια κλάση και εξετάζονται στη συνέχεια, οι τιμές του γνωρίσματος A σε αυτά.

Παρατηρώντας τον τύπο (4.4), μπορούμε να πούμε ότι η πιθανότητα να έχουμε διαφορετικές τιμές για το γνώρισμα A ανάμεσα σε δύο παραδείγματα είναι η συμπληρωματική της πιθανότητας να έχουμε ίδια τιμή για το γνώρισμα A ανάμεσα σε αυτά τα παραδείγματα.

Δηλαδή:

$$P(different\ value\ of\ A | different\ class) = 1 - P(equal\ value\ of\ A | different\ class)$$

$$P(different\ value\ of\ A | same\ class) = 1 - P(equal\ value\ of\ A | same\ class)$$

Επομένως, η (4.4) γίνεται:

$$W[A] = P(equal\ value\ of\ A | same\ class) - P(equal\ value\ of\ A | different\ class)$$

Πριν αναπτύξουμε περαιτέρω την παραπάνω σχέση, για την καλύτερη κατανόηση κρίνεται σκόπιμο να περιγράψουμε ξεχωριστά κάθε γεγονός.

Έτσι:

ως $equal\ value\ of\ A$ συμβολίζουμε το γεγονός οι τιμές του γνωρίσματος A για τα δύο παραδείγματα I_1 και I_2 να είναι ίδιες

ως $different\ value\ of\ A$ συμβολίζουμε το γεγονός οι τιμές του γνωρίσματος A για τα δύο παραδείγματα I_1 και I_2 να είναι διαφορετικές

ως $same\ class$ συμβολίζουμε το γεγονός τα δύο παραδείγματα I_1 και I_2 να ανήκουν στην ίδια κλάση

και ως $different\ class$ συμβολίζουμε το γεγονός τα δύο παραδείγματα I_1 και I_2 να ανήκουν σε διαφορετική κλάση.

Από τα παραπάνω, αναπτύσσοντας την $W[A] = P(equal\ value\ of\ A|same\ \ class) - P(equal\ value\ of\ A|different\ \ class)$ έχουμε (σύμφωνα και με τον τύπο: $P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$):

$$W[A] = \frac{P(\{equal\ value\ of\ A\} \cap \{same\ class\})}{P(same\ class)} - \frac{P(\{equal\ value\ of\ A\} \cap \{different\ class\})}{P(different\ class)}$$

Όμως:

$$P(X \cap Y) = P(Y \mid X) \cdot P(X) \quad \kappa\alpha\iota \quad P(different\ class) = 1 - P(same\ class)$$

Άρα:

$$W[A] = \frac{P(same\ class \mid equal\ value\ of\ A) \cdot P(equal\ value\ of\ A)}{P(same\ class)} - \frac{P(different\ class \mid equal\ value\ of\ A) \cdot P(equal\ value\ of\ A)}{1 - P(same\ class)}$$

Δεδομένου όμως ότι ανάμεσα σε δύο παραδείγματα I_1, I_2 το γνώρισμα A έχει την ίδια τιμή, η πιθανότητα αυτά τα δύο παραδείγματα να ανήκουν σε διαφορετικές κλάσεις είναι η συμπληρωματική της πιθανότητας αυτά τα δύο γνωρίσματα να ανήκουν στην ίδια κλάση:

$$P(different\ class \mid equal\ value\ of\ A) = 1 - P(same\ class \mid equal\ value\ of\ A)$$

Συνεπώς, έχουμε:

$$W[A] = \frac{P(same\ class \mid equal\ value\ of\ A) \cdot P(equal\ value\ of\ A)}{P(same\ class)} - \frac{(1 - P(same\ class \mid equal\ value\ of\ A)) \cdot P(equal\ value\ of\ A)}{1 - P(same\ class)}$$

Συμβολίζουμε με C την κλάση του προβλήματος και με $c_{i,}\ i = \{1,2,...,k\}$ τις διαφορετικές ετικέτες της ενώ με V το εύρος τιμών του γνωρίσματος A.

Έτσι, έχουμε:

$$P(I_1\ and\ I_2\ belong\ to\ class\ c_i) = P(\{I_1\ belongs\ to\ class\ c_i\} \cap \{I_2\ belongs\ to\ class\ c_i\}) \Leftrightarrow$$
$$P(I_1\ and\ I_2\ belong\ to\ class\ c_i) = P(I_1\ belongs\ to\ class\ c_i) \cdot P(I_2\ belongs\ to\ class\ c_i) \Leftrightarrow$$
$$P(I_1\ and\ I_2\ belong\ to\ class\ c_i) = P(C = c_i) \cdot P(C = c_i) = P(C = c_i)^2$$

Επομένως:

$$P(same\ class) = P(I_1\ and\ I_2\ belong\ to\ the\ same\ class) = P(\bigcup_{i=1}^{k}\{I_1\ and\ I_2\ belong\ to\ class\ c_i\}) \Leftrightarrow$$
$$P(same\ class) = \sum_{i=1}^{k} P(I_1\ and\ I_2\ belong\ to\ class\ c_i) = \sum_{i=1}^{k} P(C = c_i)^2 \qquad (4.5)$$

Έστω, τώρα, w μία δυνατή τιμή του γνωρίσματος A ($w \in V$). Η πιθανότητα το γνώρισμα A να έχει την ίδια τιμή w σε δύο παραδείγματα I_1, I_2 είναι:

$$P(A{=}\ w\ \ in\ I_1\ and\ I_2) = P(A = w) \cdot P(A = w) = P(A = w)^2$$

Επομένως, η πιθανότητα γενικά το γνώρισμα A να εμφανίζει ίδια τιμή σε δύο παραδείγματα είναι:

$$P(equal\,value\,of\ A) = P(\underset{u\in V}{\cup}\{A{=}\ u\ \ in\ I_1\ and\ I_2\}) = \sum_{u\in V} P(A=u)^2 \qquad (4.6)$$

Συνεπώς, η πιθανότητα τα δύο παραδείγματα να ανήκουν στην ίδια κλάση δεδομένου το γνώρισμα A έχει την ίδια τιμή είναι:

$$P(same\ class \mid equal\ value\ of\ A) =$$

$$\sum_{w\in V} P(A=w\ \ in\ I_1\ and\ I_2 \mid equal\ value\ of\ A)\cdot P(same\ class \mid A=w\ \ in\ I_1\ and\ I_2) \quad (4.7)$$

Όμως:

$$P(A=w\ \ in\ I_1\ and\ I_2 \mid equal\ value\ of\ A) = \frac{P(\{A=w\ \ in\ I_1\ and\ I_2\}\cap\{equal\ value\ of\ A\})}{P(equal\ value\ of\ A)} \Leftrightarrow$$

$$P(A=w\ \ in\ I_1\ and\ I_2 \mid equal\ value\ of\ A) = \frac{P(A=w\ \ in\ I_1\ and\ I_2)}{P(equal\ value\ of\ A)} \Leftrightarrow$$

$$P(A=w\ \ in\ I_1\ and\ I_2 \mid equal\ value\ of\ A) = \frac{P(A=w)^2}{\sum_{u\in V} P(A=u)^2}$$

Ενώ:

$$P(same\ class \mid A=w\ \ in\ I_1\ and\ I_2) = \sum_{i=1}^{k} P(C=c_i\ |A=w\ \ in\ I_1\ and\ I_2)^2$$

Άρα, η (4.7):

$$P(same\ class \mid equal\ value\ of\ A) =$$

$$\sum_{w\in V}\{[\frac{P(A=w)^2}{\sum_{u\in V} P(A=u)^2}]\cdot\sum_{i=1}^{k} P(C=c_i\ |A=w\ \ in\ I_1\ and\ I_2)^2\} \qquad (4.8)$$

Επομένως, το βάρος του γνωρίσματος A ($W[A]$) λαμβάνει την τιμή:

$$W[A] = \frac{P(equal\ value\ of\ A)\cdot(P(same\ class \mid equal\ value\ of\ A) - P(same\ class))}{P(same\ class)\cdot(1-P(same\ class))} \overset{(4.5),(4.6),(4.8)}{\Rightarrow}$$

$$W[A] = \frac{\sum_{u\in V} P(A=u)^2}{\sum_{i=1}^{k} P(C=c_i)^2 \cdot (1-\sum_{i=1}^{k} P(C=c_i)^2)} \cdot$$

$$(\sum_{w\in V}\{[\frac{P(A=w)^2}{\sum_{u\in V} P(A=u)^2}]\cdot\sum_{i=1}^{k} P(C=c_i\ |A=w\ \ in\ I_1\ and\ I_2)^2\} - \sum_{i=1}^{k} P(C=c_i)^2)$$

Αφού λοιπόν, έχουμε πια εξετάσει την περίπτωση δυαδικών προβλημάτων μπορούμε να συνεχίσουμε με την ανάλυση πολυ-κλασικών προβλημάτων (Relief-F). Ο αλγόριθμος αυτός, πέρα από το γεγονός ότι επεκτείνεται σε πολυ-κλασικά προβλήματα, παρουσιάζει μεγαλύτερη ευρωστία και μπορεί να διαχειριστεί θορυβώδη και μη πλήρη δεδομένα. Μία επισκόπηση του αλγορίθμου Relief-F παρουσιάζεται στον Πιν. VIII.

Πίνακας VIII. Αλγόριθμος Relief-F

```
set all weights W[A] = 0;
for i = 1 to m
{
  randomly select an instance Ri;
  find k nearest Hits Hj;
  for each class C ≠ class(Ri);
    from class C find k nearest misses Mj(C);

  for A = 1 to all_attributes
  {
    W[A] = W[A] + diff(A, Ri, Mj)/m - diff(A, Ri, Hj)/m;
    W[A] =
      W[A] − Σ_{j=1}^{k} diff(A, Ri, Hj) / (m · k) + Σ_{C≠class(Ri)} [ P(C) / (1 − P(class(Ri))) · Σ_{j=1}^{k} diff(A, Ri, Hj) / (m · k) ]
  }
}
```

Ουσιαστικά, το $\frac{\sum_{j=1}^{k} \mathrm{diff}(A, R_i, H_j)}{k}$ είναι η συνεισφορά των k κοντινότερων επιτυχιών στο βάρος του γνωρίσματος A ενώ το $\frac{\sum_{j=1}^{k} \mathrm{diff}(A, R_i, H_j)}{m \cdot k}$ είναι ο μέσος όρος αυτής της συνεισφοράς στο σύνολο των m πειραμάτων (που όπως προαναφέραμε προσεγγίζει καλύτερα την πιθανότητα $P(different\ value\ of\ A|same\ class)$ όσο το m μεγαλώνει).

Ανάλογα, το $\frac{\sum_{j=1}^{k} \mathrm{diff}(A, R_i, H_j)}{m \cdot k}$ παριστάνει την πιθανότητα $P(different\ value\ of\ A|different\ class\ with\ class(H_j) \neq class(R_i)\ for\ all\ k\ H_j)$, ενώ το $\frac{\mathrm{P(C)}}{1 - \mathrm{P(class(R_i))}}$ μας δίνει την πιθανότητα $P(class(H_j) = C\ for\ all\ k\ H_j | class(H_j) \neq class(R_i)\ for\ all\ k\ H_j)$.

Έτσι, το $\sum_{C \neq \mathrm{class(R_i)}} [\frac{\mathrm{P(C)}}{1 - \mathrm{P(class(R_i))}} \cdot \frac{\sum_{j=1}^{k} \mathrm{diff}(A, R_i, H_j)}{m \cdot k}]$ ισούται με την πιθανότητα $P(different\ value\ of\ A|different\ class)$. Τελειώνοντας, με την περιγραφή του αλγορίθμου Relief-F αξίζει να αναφέρουμε ότι μία καλή επιλογή του αριθμού των κοντινότερων επιτυχιών/αποτυχιών (k) είναι το 10 [19].

4.3 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVMs) ως αλγόριθμος εκπαίδευσης

Όσον αφορά τη διαδικασία εκμάθησης, χρησιμοποιήσαμε τις Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVMs) για την εξαγωγή των αποτελεσμάτων. Η μέθοδος αυτή προτιμάται για την εκμάθηση συνόλων δεδομένων μικροσυστοιχιών γιατί είναι υπολογιστικά αποτελεσματική, εύρωστη σε προβλήματα

που ανάγονται σε υψηλές διαστάσεις (όπως είναι αυτά των μικροσυστοιχιών) και βασισμένη σε στέρεες θεωρητικές αρχές. Επίσης, η ταξινόμηση των αντικειμένων γίνεται με υψηλό δείκτη εμπιστοσύνης ώστε να αποτραπούν φαινόμενα υπερεκπαίδευσης, ενώ δεν χρειάζεται να στηριχτεί σε δυνατές υποθέσεις για τη διαδικασία δημιουργίας των δεδομένων (όπως γίνεται για παράδειγμα με τις Bayesian μεθόδους).

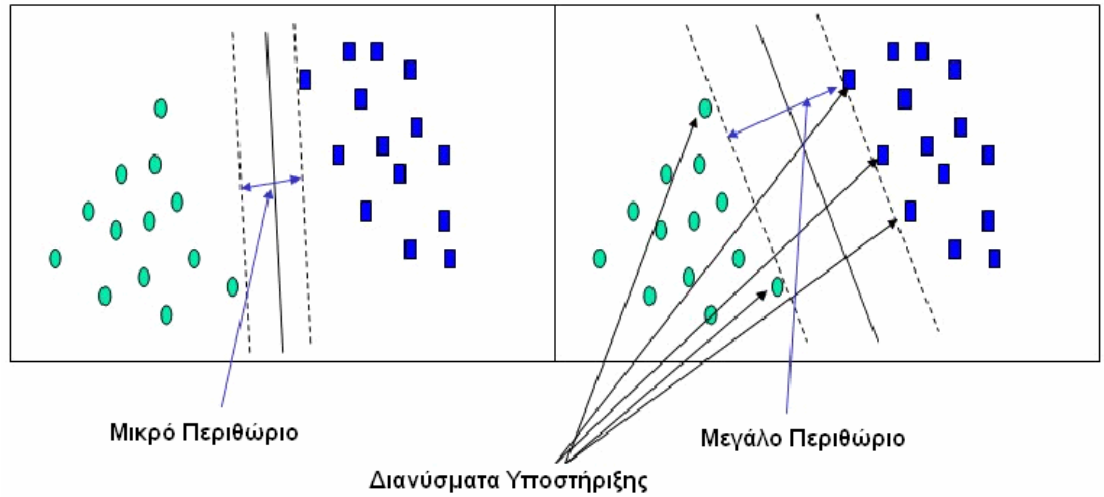
Για την πληρέστερη παρουσίαση, θα εξετάσουμε συνοπτικά τη μεθοδολογία των SVMs για γραμμικώς διαχωρίσιμα προβλήματα και έπειτα θα κάνουμε μία σύντομη αναφορά στη μεθοδολογία που εφαρμόζεται για μη-γραμμικώς διαχωρίσιμα προβλήματα.

Τα SVMs μετασχηματίζουν τον *χώρο εισόδου* τους (*input space*) σε έναν *χώρο γνωρισμάτων* (*feature space*) υψηλότερης διάστασης, στον οποίο κατασκευάζεται ένας *γραμμικός ταξινομητής* (*linear classifier*). Ο γραμμικός ταξινομητής βασίζεται σε μία *συνάρτηση απόφασης* (*decision function*) της μορφής:

$$f_{\overrightarrow{w},b}(\vec{x}) = \overrightarrow{w} \cdot \vec{x} + b$$

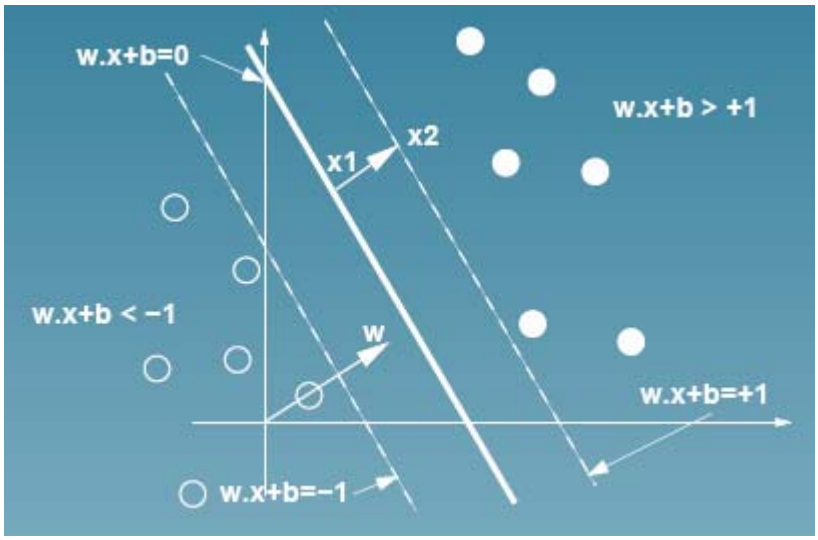
Ένα SVM εκτελεί την ταξινόμηση με την κατασκευή ενός n-διάστατου υπερ-επιπέδου που διαχωρίζει βέλτιστα τα στοιχεία σε δύο κατηγορίες (για δυαδικά προβλήματα). Συγκεκριμένα, επιχειρείται η εύρεση κατάλληλων τιμών για τα βάρη ($\overrightarrow{w}$) και την πόλωση (b), ώστε να επιτυγχάνεται ο μέγιστος διαχωρισμός.

Ένα σύνολο γνωρισμάτων που περιγράφει μια περίπτωση (δηλ., μια σειρά των τιμών των μεταβλητών εισόδου) καλείται *διάνυσμα (vector)*. Έτσι ο στόχος της διαμόρφωσης του SVM είναι να βρεθεί βέλτιστο υπερ-επίπεδο που χωρίζει τις “συστάδες” των διανυσμάτων κατά τέτοιο τρόπο ώστε οι περιπτώσεις που αντιστοιχούν στη μία ετικέτα κλάσης να βρίσκονται στη μια πλευρά του υπερ-επιπέδου και οι περιπτώσεις που αντιστοιχούν στην άλλη ετικέτα κλάσης να βρίσκονται στην άλλη πλευρά του υπερ-επιπέδου. Τα διανύσματα που βρίσκονται πιο κοντά στο υπερ-επίπεδο είναι τα *διανύσματα υποστήριξης (support vectors)*. Ένας αλγόριθμος SVM εντοπίζει το υπερ-επιπέδο που είναι έτσι προσανατολισμένο ώστε το περιθώριο (margin) μεταξύ των διανυσμάτων υποστήριξης να μεγιστοποιείται. Τα παραπάνω οπτικοποιούνται στην Εικ. 4.1.



Εικόνα 4.1. Απεικόνιση μη-βέλτιστου και βέλτιστου υπερ-επιπέδου
Η παραπάνω εικόνα αποτελεί τροποποίηση της πρωτότυπης που έχει ληφθεί από την τοποθεσία [x]

Το βέλτιστο υπερ-επίπεδο εντοπίζεται θεωρώντας ένα "διάδρομο" που ορίζεται από τις τιμές –1 και +1 της συνάρτησης απόφασης, όπως φαίνεται στην Εικ. 4.2.



Εικόνα 4.2. Εύρεση του βέλτιστου υπερ-επιπέδου από τις Μηχανές Διανυσμάτων Υποστήριξης (SVMs)
Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [xx]

Το μήκος κάθε "λωρίδας" του διαδρόμου είναι $\frac{1}{\| \vec{w} \|}$.

Τα παραδείγματα της μίας ετικέτας ($y_i = +1$) θα πρέπει να ικανοποιούν τη σχέση: $\vec{w} \cdot \overrightarrow{x_i} + b \geq 1$ και της άλλης ($y_i = -1$) τη σχέση: $\vec{w} \cdot \overrightarrow{x_i} + b \leq -1$.

Με άλλα λόγια, θα πρέπει να ικανοποιείται η σχέση: $y_i \cdot (\vec{w} \cdot \overrightarrow{x_i} + b) \geq 1$ με $i \in \{1,2,...,N\}$ (όπου N : το πλήθος των παραδειγμάτων).

Όπως προαναφέραμε το μήκος κάθε λωρίδας είναι $\frac{1}{\| \vec{w} \|}$. Δηλαδή, το μήκος του περιθωρίου είναι ίσο με $\frac{2}{\| \vec{w} \|}$. Επιθυμητή είναι η μεγιστοποίησή του. Επομένως, το βέλτιστο υπερ-επίπεδο βρίσκεται ελαχιστοποιώντας την παράσταση: $\frac{1}{2} \cdot \| \vec{w} \|^2$ υπό τους περιορισμούς $y_i \cdot (\vec{w} \cdot \overrightarrow{x_i} + b) - 1 \geq 0$ με $i \in \{1,2,...,N\}$.

(Ο συντελεστής $\frac{1}{2}$ στην παράσταση $\frac{1}{2} \cdot \| \vec{w} \|^2$ χρησιμοποιείται για την ευκολία της παρουσίασης, αφού για την εύρεση της βέλτιστης τιμής απαιτείται η παραγώγιση του $\| \vec{w} \|^2$. Κατά την παραγώγιση της παραπάνω παράστασης δημιουργείται ο συντελεστής 2 ο οποίος αναιρείται όταν πολλαπλασιάζεται με το $\frac{1}{2}$).

Θεωρώντας $\vec{w} = (w_1, w_2, ..., w_n)$ (όπου n ο αριθμός των γνωρισμάτων) η παράσταση $\frac{1}{2} \cdot \| \vec{w} \|^2$ ανάγεται σε μία συνάρτηση n μεταβλητών:

$$\frac{1}{2} \cdot \| \vec{w} \|^2 = h(\vec{w}) = h(w_1, w_2, ..., w_n)$$

Από τη θεωρία του λογισμού πολλών μεταβλητών, η παράσταση αυτή παρουσιάζει ελάχιστο όταν ισχύει:

$$\nabla h(\vec{w}) = \nabla h(w_1, w_2, ..., w_n) = \begin{pmatrix} \frac{\partial h}{\partial w_1} \\ \frac{\partial h}{\partial w_2} \\ \vdots \\ \frac{\partial h}{\partial w_n} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

υπό τους περιορισμούς $y_i \cdot (\vec{w} \cdot \overrightarrow{x_i} + b) - 1 \geq 0$ με $i \in \{1,2,...,N\}$.

Εναλλακτικά, αυτό το πρόβλημα μπορεί να λυθεί θεωρώντας τη μέθοδο των πολλαπλασιαστών Lagrange. Συγκεκριμένα, εισάγουμε μία δυαδική μεταβλητή $\alpha_i \geq 0$ με $i \in \{1,2,...,N\}$ για κάθε περιορισμό.

Έτσι, κατασκευάζουμε τη *συνάρτηση Lagrange* (*Lagrangian function*), η οποία έχει τη μορφή:

$$L(\vec{w},b,\vec{\alpha}) = \frac{1}{2}\cdot \| \vec{w} \|^2 - \sum_{i=1}^{N} \alpha_i \cdot [y_i \cdot (\vec{w}\cdot \overrightarrow{x_i} + b) - 1]$$

Η λύση στο πρόβλημα μεγιστοποίησης βρίσκεται στο σαγματικό σημείο (saddle point) της συνάρτησης Lagrange $L(\vec{w},b,\vec{\alpha})$, η οποία πρέπει να ελαχιστοποιηθεί ως προς $\vec{w}$, b ή με άλλα λόγια να μεγιστοποιηθεί ως προς $\vec{\alpha}$.

Έτσι, παραγωγίζοντας αυτή τη φορά τη συνάρτηση Lagrange $L(\vec{w},b,\vec{\alpha})$ ως προς $\vec{w}$ και b και θέτοντας τις παραγώγους ίσες με μηδέν (0), έχουμε:

$$\frac{\partial L(\vec{w},b,\vec{\alpha})}{\partial \vec{w}} = \vec{0}$$

και

$$\frac{\partial L(\vec{w},b,\vec{\alpha})}{\partial b} = 0$$

Η παράσταση $\frac{\partial L(\vec{w},b,\vec{\alpha})}{\partial \vec{w}} = \vec{0}$ μας οδηγεί στη σχέση:

$$\vec{w} = \sum_{i=1}^{N} \alpha_i \cdot y_i \cdot \overrightarrow{x_i} \quad (4.9)$$

, ενώ η παράσταση $\frac{\partial L(\vec{w},b,\vec{\alpha})}{\partial b} = 0$ μας οδηγεί στη σχέση:

$$\sum_{i=1}^{N} \alpha_i \cdot y_i = 0 \quad (4.10)$$

Από τα προηγούμενα, μπορούμε να κατασκευάσουμε το δυαδικό πρόβλημα (dual problem) του παραπάνω προβλήματος. Πιο συγκεκριμένα, ισχύει:

$$L(\vec{w},b,\vec{\alpha}) = \frac{1}{2}\cdot \| \vec{w} \|^2 \ -\sum_{i=1}^{N} \alpha_i \cdot [y_i \cdot (\vec{w} \cdot \overrightarrow{x_i} + b) - 1] \Leftrightarrow$$

$$L(\vec{w},b,\vec{\alpha}) = \frac{1}{2}\cdot \| \vec{w} \|^2 \ -\sum_{i=1}^{N} \alpha_i \cdot y_i \cdot \vec{w} \cdot \overrightarrow{x_i} - b \cdot \sum_{i=1}^{N} \alpha_i \cdot y_i + \sum_{i=1}^{N} \alpha_i$$

Λόγω όμως (4.9), η σχέση $\frac{1}{2}\cdot \| \vec{w} \|^2$ γίνεται:

$\frac{1}{2}\cdot \| \vec{w} \|^2 = \frac{1}{2}\cdot \vec{w}^{\mathrm{T}} \cdot \vec{w} \overset{(4.9)}{=} \frac{1}{2}\cdot \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \vec{w} \cdot \overrightarrow{x_i}^T \overrightarrow{x_j}$, ενώ η παράσταση $-\sum_{i=1}^{N} \alpha_i \cdot y_i \cdot \vec{w} \cdot \overrightarrow{x_i}$ παίρνει τη μορφή:

$$-\sum_{i=1}^{N} \alpha_i \cdot y_i \cdot \vec{w} \cdot \overrightarrow{x_i} = -\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \vec{w} \cdot \overrightarrow{x_i}^T \overrightarrow{x_j} \,.$$

Από τα παραπάνω και την (4.10), η $L(\vec{w},b,\vec{\alpha})$ λαμβάνει τη μορφή:

$$Q(\vec{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\cdot \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot \vec{w} \cdot \overrightarrow{x_i}^T \overrightarrow{x_j}$$

Η παραπάνω εξίσωση αποτελεί τη δυαδική της αρχικής ($L(\vec{w},b,\vec{\alpha})$), γι' αυτό και απαιτείται η μεγιστοποίησή της (αντί ελαχιστοποίησης), ενώ πρέπει να ικανοποιεί και τους παρακάτω περιορισμούς:

$$\sum_{i=1}^{N} \alpha_i \cdot y_i = 0$$

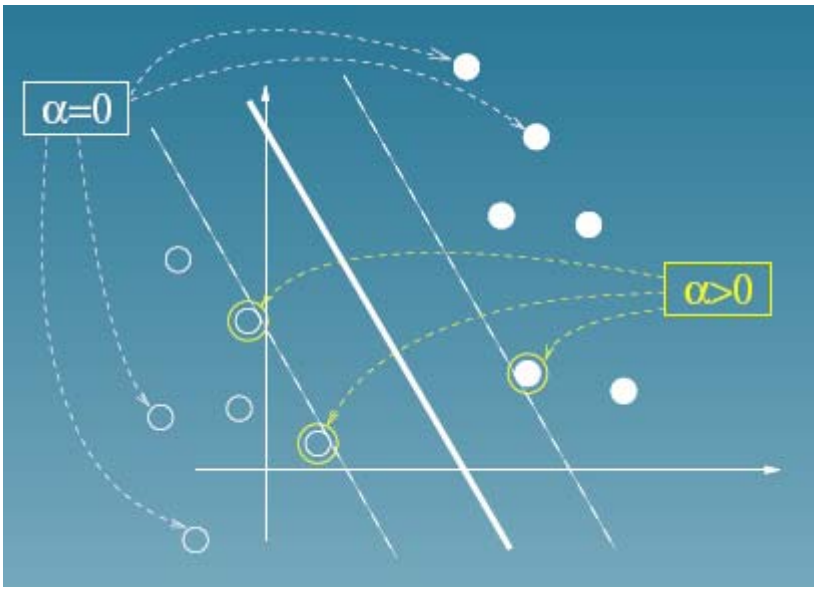
και

$$\alpha_i \geq 0 \text{ για } i \in \{1,2,...,N\}$$

Το παραπάνω (δυαδικό) πρόβλημα στηρίζεται εξολοκλήρου στα δεδομένα εκπαίδευσης $(\overrightarrow{x_i}, y_i)$ για $i \in \{1,2,...,N\}$.

Προσδιορίζοντας τους βέλτιστους πολλαπλασιαστές Lagrange ($\vec{\alpha}^* = (\alpha_1, \alpha_2, ..., \alpha_N)$), μπορούμε να υπολογίσουμε το βέλτιστο διάνυσμα βαρών από τη σχέση (4.9) ($\vec{w} = \sum_{i=1}^{N} \alpha_i \cdot y_i \cdot \overrightarrow{x_i}$), λαμβάνοντας: $\vec{w}^* = \sum_{i=1}^{N} \alpha_i^* \cdot y_i \cdot \overrightarrow{x_i}$.

Είναι αξιοσημείωτο ότι μόνο στα διανύσματα υποστήριξης αντιστοιχούν μη μηδενικοί πολλαπλασιαστές Lagrange ($\alpha_i > 0$), ακριβώς όπως απεικονίζεται στην Εικ. 4.3. Έτσι, για την πρόβλεψη νέων σημείων, αναγκαία είναι μόνο η χρήση των διανυσμάτων υποστήριξης.



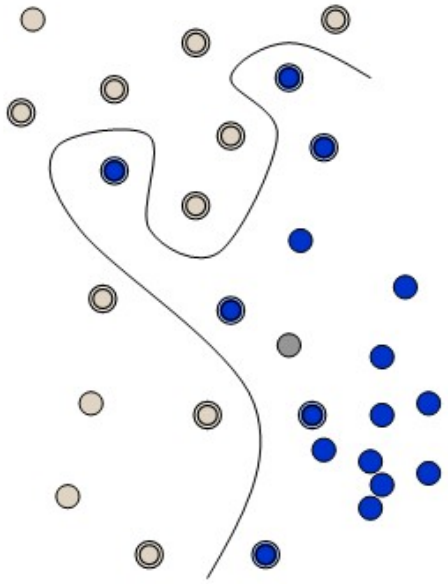
Εικόνα 4.3. Η σημασία των διανυσμάτων υποστήριξης στην πρόβλεψη νέων σημείων
Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [xx]

Τέλος, θεωρώντας ένα θετικό διάνυσμα υποστήριξης ($\overrightarrow{x_{SV+}}$), ένα διάνυσμα δηλαδή που αναφέρεται στην ετικέτα κλάσης με τιμή $y_i = +1$, αυτό θα ικανοποιεί τη σχέση:

$$\vec{w}^* \cdot \overrightarrow{x_{SV+}} + b^* = 1 \Leftrightarrow$$
$$b^* = 1 - \vec{w}^* \cdot \overrightarrow{x_{SV+}}$$

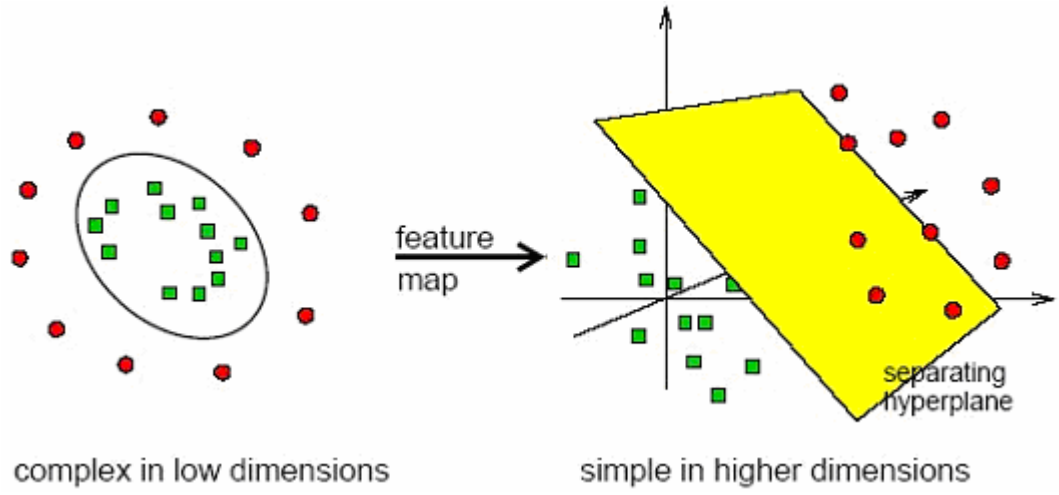
Με αυτό τον τρόπο υπολογίζουμε και τη βέλτιστη τιμή για την πόλωση [4].

Εν κατακλείδι, θα πούμε λίγα λόγια για την περίπτωση μη-γραμμικών προβλημάτων. Στα προβλήματα αυτά, τα σημεία του συνόλου δεδομένων που αντιστοιχούν σε διαφορετικές ετικέτες διαχωρίζονται με μία μη-γραμμική καμπύλη, όπως φαίνεται στην Εικ. 4.4.



Εικόνα 4.4. Μη-γραμμικώς διαχωρίσιμο πρόβλημα. Η καμπύλη που διαχωρίζει τις δύο ετικέτες (οι οποίες αναπαριστώνται με μπλε και άσπρο χρώμα) δεν είναι πλέον γραμμική. Επίσης, τα σημεία που περιέχουν δύο ομόκεντρους κύκλους αποτελούν τα διανύσματα υποστήριξης.
Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [xi]

Σε αυτή την περίπτωση, χρησιμοποιούνται οι λεγόμενες συναρτήσεις πυρήνα (kernel functions) που συμβολίζονται με $\vec{\Phi}$ και μετασχηματίζουν τα δεδομένα σε ένα χώρο υψηλότερης διάστασης όπου τα δεδομένα μπορούν να διαχωριστούν με υπερ-επίπεδο. Τα προαναφερθέντα οπτικοποιούνται στην Εικ. 4.5.



Εικόνα 4.5. Ο διαχωρισμός μη-γραμμικώς διαχωρίσιμων προβλημάτων μετά το μετασχηματισμό σε χώρο υψηλότερης διάστασης με χρήση συνάρτησης πυρήνα είναι απλούστερος.
Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [xiv]

Για μία δεδομένη αντιστοίχηση $\vec{\Phi}$ από το χώρο εισόδου σε ένα χώρο υψηλότερης διάστασης: $\vec{\Phi} : \vec{X} \rightarrow \vec{Y}$, ορίζεται ο πυρήνας δύο αντικειμένων $\vec{x}$ και $\vec{x}^{'}$ ($K(\vec{x}, \vec{x}^{'})$) ως το εσωτερικό γινόμενο των προβολών (των δύο αυτών αντικειμένων) στο χώρο υψηλότερης διάστασης.

Πιο φορμαλιστικά, ισχύει:

$$K(\vec{x}, \vec{x}^{'}) = \vec{\Phi}(\vec{x}) \cdot \vec{\Phi}(\vec{x}^{'})$$

Για την εκτέλεση της διαδικασίας σε μη-γραμμικά προβλήματα, δεν είναι απαραίτητος ο υπολογισμός της συνάρτησης πυρήνα $\vec{\Phi}(\vec{x})$, αρκεί ο υπολογισμός του πυρήνα $K(\vec{x}, \vec{x}^{'})$. Αυτό συμβαίνει γιατί η συνάρτηση απόφασης αποκτάει τη μορφή:

$$f(\vec{x}) = \sum_{i=1}^{N} a_i^{*} \cdot K(\overrightarrow{x_i}, \vec{x}) + b^{*}$$

με τις τιμές a_i^{*} για $i \in \{1, 2, ..., N\}$ να λαμβάνουν τέτοιες τιμές ώστε να μεγιστοποιείται η σχέση:

$$Q(\vec{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot K(\overrightarrow{x_i}, \overrightarrow{x_j})$$

υπό τους περιορισμούς:

$$\sum_{i=1}^{N} \alpha_i \cdot y_i = 0$$

και

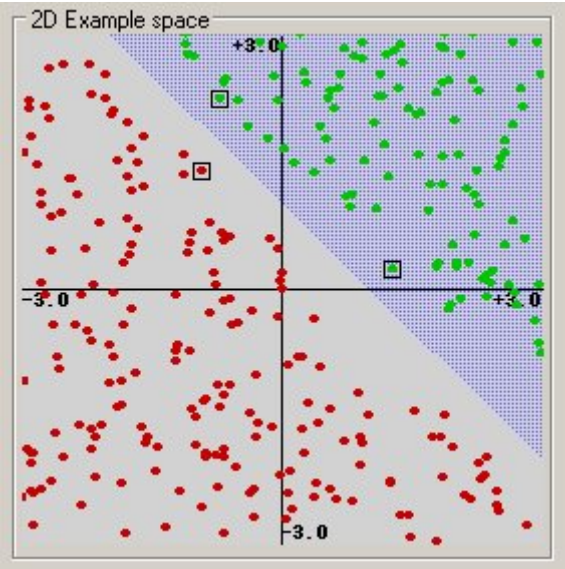
$$0 \leq \alpha_i \leq C \text{ για } i \in \{1, 2, ..., N\}$$

όπου C : μία θετική παράμετρος που ορίζεται από τον χρήστη

Μερικοί από τους συχνότερα χρησιμοποιούμενους πυρήνες είναι οι:

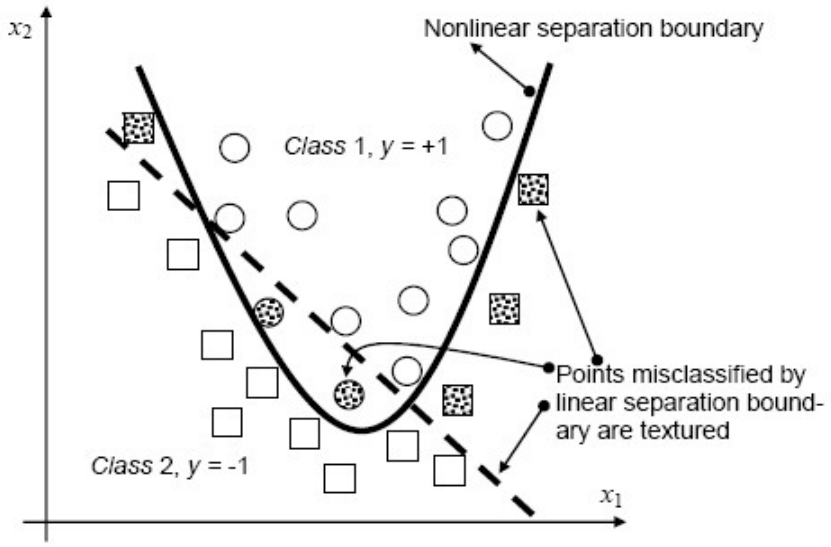
- Γραμμικοί (Linear): $K(\vec{x}, \vec{x}^{'}) = \vec{x} \cdot \vec{x}^{'} + \gamma$, όπου η παράμετρος γ καθορίζεται από τον χρήστη. Αυτός ο τύπος πυρήνα έχει εφαρμογή μόνο σε γραμμικώς

διαχωρίσιμα προβλήματα. Παράδειγμα ενός τέτοιου πυρήνα φαίνεται στην Εικ. 4.6.



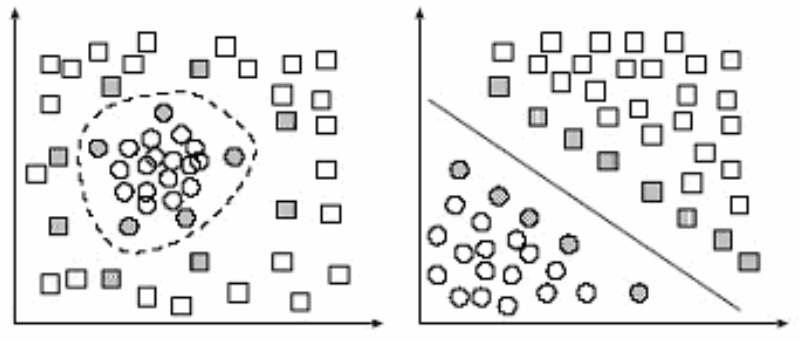
Εικόνα 4.6. Παράδειγμα γραμμικού πυρήνα. Τα σημεία που περικλείονται από τετραγωνικό πλαίσιο αποτελούν διανύσματα υποστήριξης.
Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [ix]

- Πολυωνυμικοί (Polynomial): $K(\vec{x}, \vec{x}') = (\vec{x} \cdot \vec{x}' + 1)^p$, όπου η παράμετρος p καθορίζεται από τον χρήστη. Στην Εικ. 4.7 παρουσιάζεται η περίπτωση εφαρμογής SVM γραμμικού πυρήνα και πολυωνυμικού πυρήνα σε μη-γραμμικό πρόβλημα. Είναι σαφής η μεγαλύτερη διαχωριστική ικανότητα που παρουσιάζει το SVM με πολυωνυμικό πυρήνα.



Εικόνα 4.7. Παράδειγμα διαχωρισμού μη-γραμμικού προβλήματος με χρήση SVMs γραμμικού (διακεκομμένη γραμμή) και πολυωνυμικού πυρήνα (συνεχής γραμμή). Είναι σαφής η μεγαλύτερη διαχωριστική ικανότητα του SVM πολυωνυμικού πυρήνα.
Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [xii]

- Ακτινωτοί (Radial Basis Functions): $K(\vec{x},\vec{x}') = e^{-\frac{\|\vec{x}-\vec{x}'\|}{2\sigma^2}}$, όπου η παράμετρος σ καθορίζεται από τον χρήστη. Παράδειγμα ενός τέτοιου πυρήνα φαίνεται στην Εικ. 4.8.



Εικόνα 4.8. Στην αριστερή υπο-εικόνα παρουσιάζεται ο τρόπος που θα φαινόταν ο διαχωρισμός με radial basis πυρήνα στον αρχικό χώρο (χώρο εισόδου), ενώ στη δεξιά έχει προηγηθεί ο μετασχηματισμός σε χώρο υψηλότερης διάστασης με τη βοήθεια radial basis πυρήνα όπου ο διαχωρισμός είναι πια γραμμικός. Τα σημεία που χρωματίζονται σε κλίμακα του γκρι αποτελούν διανύσματα υποστήριξης.
Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [xiii]

- Σιγμοειδείς (Sigmoid): $K(\vec{x}, \vec{x}^{'}) = \tanh(\kappa \cdot \vec{x} \cdot \vec{x}^{'} + \theta)$

Τέλος, αφού περιγράψαμε τα SVMs σε περιπτώσεις γραμμικών και μη-γραμμικών δυαδικών προβλημάτων, στο σημείο αυτό θα κάνουμε μία πολύ σύντομη αναφορά αντιμετώπισης πολυκλασικών προβλημάτων.

Οι δύο δημοφιλέστερες προσεγγίσεις είναι η "*μία εναντίον πολλών*" προσέγγιση ("*one against many*") και η "*μία εναντίον μίας*" ("*one against one*") προσέγγιση.

Στη "μία εναντίον πολλών" προσέγγιση κάθε κατηγορία (ετικέτα) διαχωρίζεται διαδοχικά από τις προηγούμενες και στο τέλος συγκεντρώνονται, ενώ στη "*μία εναντίον μίας*" προσέγγιση εξετάζονται ανά δύο όλες οι κατηγορίες (ετικέτες). Συνεπώς, αν έχουμε K κατηγορίες (ετικέτες), οι συνδυασμοί ανά δύο των K κατηγοριών είναι: $\binom{K}{2} = \frac{K!}{2! \cdot (K-2)!} = \frac{K \cdot (K-1)}{2}$. Αν και αυτή η προσέγγιση είναι αποδοτικότερη είναι υπολογιστική πιο δαπανηρή από την προηγούμενη.

Αφού πλέον έχουμε θέσει τις απαραίτητες βάσεις για την κατανόηση των παρακάτω χρησιμοποιούμενων εννοιών, μπορούμε να προχωρήσουμε στην αποτίμηση των αποτελεσμάτων.

4.4 Σημασιολογική ομοιότητα

Όπως προαναφέρθηκε μία σημαντική καινοτομία αυτής της διπλωματικής εργασίας είναι η εκμετάλλευση της υπάρχουσας βιολογικής γνώσης. Αυτό είναι δυνατό αντλώντας πληροφορία από τη Γονιδιακή Οντολογία. Εύλογα όμως κανείς διερωτάται πως ποσοτικοποιείται αυτή η πληροφορία ώστε να επιτραπεί η εφαρμογή εννοιών όπως η συσχέτιση και η ομοιότητα.

Η ποσοτικοποίηση αυτής της πληροφορίας καθίσταται δυνατή μέσω της σημασιολογικής ομοιότητας όπου μας παρέχεται η δυνατότητα ενσωμάτωσης της δομημένης πληροφορίας που περιέχει μία οντολογία.

Πιο συγκεκριμένα, ως *σημασιολογική ομοιότητα* εννοούμε την απόδοση μέτρου σε ένα σύνολο όρων βάσει της ομοιότητάς τους στο σημασιολογικό τους περιεχόμενο. Σε γενικές γραμμές, οι όροι, που αντιπροσωπεύονται ως κόμβοι στην οντολογία, τοποθετούνται τόσο πιο μακριά σε σχέση ο ένας με τον άλλο όσο μικρότερη είναι η ομοιότητά τους.

Εκτός από το προφανές μέτρο ποσοτικοποίησης της απόστασης μεταξύ δύο κόμβων, του αριθμού δηλαδή των ακμών του ελάχιστου μονοπατιού που συνδέει τους δύο αυτούς κόμβους υπάρχουν και άλλα μέτρα ποσοτικοποίησης. Ένα από τα πιο διαδεδομένα μέτρα που χρησιμοποιείται εκτενώς στις μεθόδους σημασιολογικής ομοιότητας που εφαρμόστηκαν σε αυτή τη διπλωματική εργασία είναι το *πληροφοριακό περιεχόμενο (information content)*.

Έτσι, πριν αναλύσουμε κάθε μέθοδο σημασιολογικής ομοιότητας χωριστά κρίνεται σκόπιμο να κάνουμε αναλυτική αναφορά στην έννοια του πληροφοριακού περιεχομένου ως το βασικό στοιχείο των περισσοτέρων από τις μεθόδους σημασιολογικής ομοιότητας.

4.4.1 Πληροφοριακό περιεχόμενο (Information content)

Η έννοια του πληροφοριακού περιεχομένου αφορά γενικότερα το επίπεδο δυσκολίας που απαιτείται για τον καθορισμό ή την περιγραφή ενός αντικειμένου. Με άλλα λόγια, το υψηλό πληροφοριακό περιεχόμενο υποδεικνύει μεγαλύτερη προσπάθεια για την επεξεργασία ενός αντικειμένου ή μεγαλύτερη εξειδίκευση αυτού του αντικειμένου. Το πληροφοριακό περιεχόμενο θα μπορούσε να εννοηθεί και ως δείκτης ακριβείας αφού όσο μεγαλύτερο είναι αυτό, τόσο πιο ακριβές ή με άλλα λόγια πιο ειδικό θα μπορούσε να θεωρηθεί.

Όσον αφορά το πληροφοριακό περιεχόμενο σε βιολογικούς όρους παριστάνει το δείκτη γενικότητας αυτών των όρων. Πιο συγκεκριμένα, όπως υπογραμμίστηκε η Γονιδιακή Οντολογία αποτελεί ένα δομημένο λεξιλόγιο. Υπάρχουν, έτσι, σχέσεις ιεραρχίας μεταξύ των όρων της. Άλλοι όροι είναι πιο γενικοί, ενώ άλλοι πιο ειδικοί. Διαισθητικά, όσο βαθύτερα βρισκόμαστε στην οντολογία, τόσο πιο ειδικός είναι ο όρος.

Όπως προαναφέραμε, ένα προφανές μέτρο της σχέσης μεταξύ δύο κόμβων ή της γενικότητας ενός κόμβου είναι η απόσταση μεταξύ τους ή η απόσταση μεταξύ αυτού και της ρίζας (βάθος κόμβου) αντίστοιχα. Ενώ η εννοιολογική προσέγγιση της παραπάνω μεθοδολογίας είναι απλή εμφανίζει το πολύ σημαντικό μειονέκτημα της αγνόησης της τοπικής πυκνότητας κάθε κόμβου. Με άλλα λόγια, εμφανίζει μειονεκτήματα σε περιπτώσεις γράφων (δένδρων) μη ομοιόμορφης πυκνότητας.

Επίσης, δεν λαμβάνεται υπόψη το γεγονός ότι ορισμένοι κλάδοι στην οντολογία είναι μακρύτεροι από άλλους κάτι που απαιτεί την κανονικοποίηση του πληροφοριακού περιεχομένου. Για παράδειγμα, οι όροι GO:0008218

"*bioluminescence*" (βάθους 3) και GO:0045994 "*positive regulation of translational initiation by iron*" (βάθους 8) αποτελούν αμφότεροι φύλλα της οντολογίας. Τα φύλλα περιγράφουν (στην προκειμένη περίπτωση) στον πληρέστερο δυνατό βαθμό μία συγκεκριμένη βιολογική διαδικασία για τη δεδομένη χρονική στιγμή. Έτσι, θα πρέπει αυτοί οι δύο όροι να εσωκλείουν το ίδιο πληροφοριακό περιεχόμενο [1].

Έγινε αναφορά προηγουμένως στην πυκνότητα ενός όρου η οποία αποτελεί κύριο δείκτη του πληροφοριακού του περιεχομένου. Το μόνο πια που απομένει είναι η φορμαλιστική περιγραφή των παραπάνω, ώστε να επιτραπεί η ποσοτικοποίηση της ομοιότητας μεταξύ διαφορετικών όρων. Είναι προφανές ότι όσες περισσότερες φορές εμφανίζεται ένας όρος να αποδίδεται σε γονιδιακά προϊόντα, τόσο πιο γενικός είναι αυτός. Για την ακρίβεια, αρκεί να αποδοθεί ένας οποιοσδήποτε απόγονος αυτού του όρου σε γονιδιακό προϊόν, ώστε να θεωρείται ότι αυτό το γονιδιακό προϊόν αντιστοιχεί επίσης και σε αυτόν τον όρο. Η θέση ενός όρου στην οντολογία έχει άμεση εξάρτηση με τον αριθμό των επισημειώσεων αυτού του όρου σε γονιδιακά προϊόντα. Σε γενικές γραμμές, όσο γενικότερος είναι ένας όρος, τόσο ψηλότερα θα εντοπίζεται στην ιεραρχία της οντολογίας. Αυτό όμως δεν αποτελεί πάντα τον κανόνα. Όπως φαίνεται και στον Πιν. ΙΧ όπου υπάρχουν όροι που ενώ συγκριτικά με άλλους βρίσκονται βαθύτερα στην οντολογία είναι γενικότεροι (μικρότερο πληροφοριακό περιεχόμενο).

Συνεχίζοντας από τα προηγούμενα, η σύμβαση του να αποδίδεται ένα γονιδιακό προϊόν εξίσου στους προγόνους ενός όρου εκτός του ότι είναι διαισθητικά προφανής βασίζεται και σε συγκεκριμένο λόγο. Όπως έχει τονιστεί, ο λεγόμενος *Κανόνας του Ορθού Μονοπατιού* επιτάσσει ότι το μονοπάτι ενός όρου προς τους προγόνους του πρέπει να είναι πάντα αληθές. Αυτό με άλλα λόγια, σημαίνει ότι κάθε όρος μπορεί και πρέπει να θεωρείται ως ένας από τους προγόνους του, ώστε η τοποθέτησή του μέσα στην οντολογία να είναι έγκυρη με το ίδιο ακριβώς σκεπτικό που μπορούμε να θεωρήσουμε ότι ένα αυτοκίνητο συγκεκριμένης μάρκας δεν παύει να είναι αυτοκίνητο. Για παράδειγμα, ο όρος "alkali metal ion binding" μπορεί να θεωρηθεί ως "metal ion binding" (ένας από τους άμεσους γονείς του) ή επίσης ως "ion binding" (ένας από τους προγόνους του). Με απλά λόγια, ένας όρος διατηρεί όλα τα χαρακτηριστικά των προγόνων του και έτσι μπορεί να θεωρηθεί ως ένας από αυτούς.

Τονίσαμε προηγουμένως ότι όσο περισσότερο εμφανίζεται ένας όρος να αποδίδεται σε γονιδιακά προϊόντα, τόσο πιο γενικός είναι. Όσο πιο γενικός, όμως, είναι ένας όρος τόσο μικρότερο πληροφοριακό περιεχόμενο έχει. Έτσι, με επαγωγική λογική, όσες περισσότερες φορές επισημειώνεται ένας όρος σε γονιδιακά προϊόντα,

τόσο μικρότερο πληροφοριακό περιεχόμενο θα έχει. Το πληροφοριακό περιεχόμενο ενός όρου σχετίζεται επομένως άμεσα με τη συχνότητα εμφάνισης του όρου ή ακόμα καλύτερα με την πιθανότητα εμφάνισης του.

Το πρώτο βήμα, λοιπόν, για τον καθορισμό του πληροφοριακού περιεχομένου ενός όρου είναι ο ορισμός της πιθανότητας εμφάνισης του. Ορίζουμε την πιθανότητα, $p(c)$, ενός όρου ως τον αριθμό των φορών που εμφανίζεται αυτός ο όρος (ή κάποιος από τους απογόνους του) προς τον αριθμό των φορών που εμφανίζεται ο οποιοσδήποτε όρος ή με άλλα λόγια, προς τον αριθμό των φορών που εμφανίζεται η ρίζα της οντολογίας (ή κάποιος από τους απογόνους της).

$$p(c) = \frac{n_c}{n_r}$$

n_c : Ο αριθμός των φορών που εμφανίζεται ο όρος (ή κάποιος από τους απογόνους του)

n_r : Ο αριθμός των φορών που εμφανίζεται η ρίζα της οντολογίας (ή κάποιος από τους απογόνους της)

Συνεπώς, το πληροφοριακό περιεχόμενο ενός όρου ορίζεται ως ο αρνητικός λογάριθμος της παραπάνω πιθανότητας:

$$IC(c) = -\log(p(c)) = -\log(\frac{n_c}{n_r})$$

Μάλιστα, επιχειρώντας την κανονικοποίηση του παραπάνω πληροφοριακού περιεχομένου επανερχόμαστε στην έννοια του πληροφοριακού περιεχομένου ενός φύλλου. Όπως προαναφέρθηκε, τα φύλλα περιγράφουν στον πληρέστερο δυνατό βαθμό μία συγκεκριμένη διαδικασία (λειτουργία) για τη δεδομένη χρονική στιγμή. Έτσι, εκφράζοντας την αναλυτικότερη μορφή αυτής της διαδικασίας (λειτουργίας) εσωκλείουν το υψηλότερο πληροφοριακό περιεχόμενο. Μάλιστα, η πιθανότητα εμφάνισης ενός φύλλου υπολογίζεται ως η μοναδική φορά που εμφανίζεται αυτό (αφού δεν έχει απογόνους) προς τον αριθμό των φορών που εμφανίζεται η ρίζα της οντολογίας (ή κάποιος από τους απογόνους της).

Πιο φορμαλιστικά:

$$p(leaf) = \frac{1}{n_r}$$

Ενώ:

$$IC(leaf) = -\log(p(leaf)) = -\log(\frac{1}{n_r})$$

Συνεπώς, για την κανονικοποίηση ενός όρου αρκεί η διαίρεση του πληροφοριακού του περιεχομένου με το πληροφοριακό περιεχόμενο ενός φύλλου.

$$IC_{normal}(c) = \frac{\log(p(c))}{\log(p(leaf))} = \frac{\log(\frac{n_c}{n_r})}{\log(\frac{1}{n_r})} = 1 - \frac{\log(n_c)}{\log(n_r)}$$

Έτσι, το ελάχιστο κανονικοποιημένο πληροφοριακό περιεχόμενο για έναν όρο είναι το 0 εάν αυτός ο όρος είναι η ρίζα της οντολογίας, ενώ το μέγιστο κανονικοποιημένο πληροφοριακό περιεχόμενο είναι 1 εάν αυτός ο όρος είναι φύλλο.

Πίνακας IX. Παρουσίαση ορισμένων όρων της BP οντολογίας κατά αύξουσα σειρά πληροφοριακού περιεχομένου. Ως ελάχιστη τιμή θεωρείται το μηδέν (0), ενώ ως μέγιστη το ένα (1). Όπως παρατηρείται, το πληροφοριακό περιεχόμενο ενός όρου εξαρτάται κυρίως από την πυκνότητά του (αριθμός απογόνων) παρά από το βάθος του (έκδοση Απριλίου του 2007)

Όρος		**Πληροφοριακό περιεχόμενο**	**Βάθος**
id	όνομα		
GO:0008150	biological_process	0.000	0
GO:0009987	cellular process	0.040	1
GO:0008152	metabolic process	0.099	1
GO:0044237	cellular metabolic process	0.103	2
GO:0044238	primary metabolic process	0.129	2
GO:0032502	developmental process	0.141	1
GO:0007275	multicellular organismal development	0.176	3
GO:0044248	cellular catabolic process	0.285	3
GO:0040011	locomotion	0.609	1
GO:0015979	photosynthesis	0.623	2
GO:0008218	bioluminescence	1.000	3

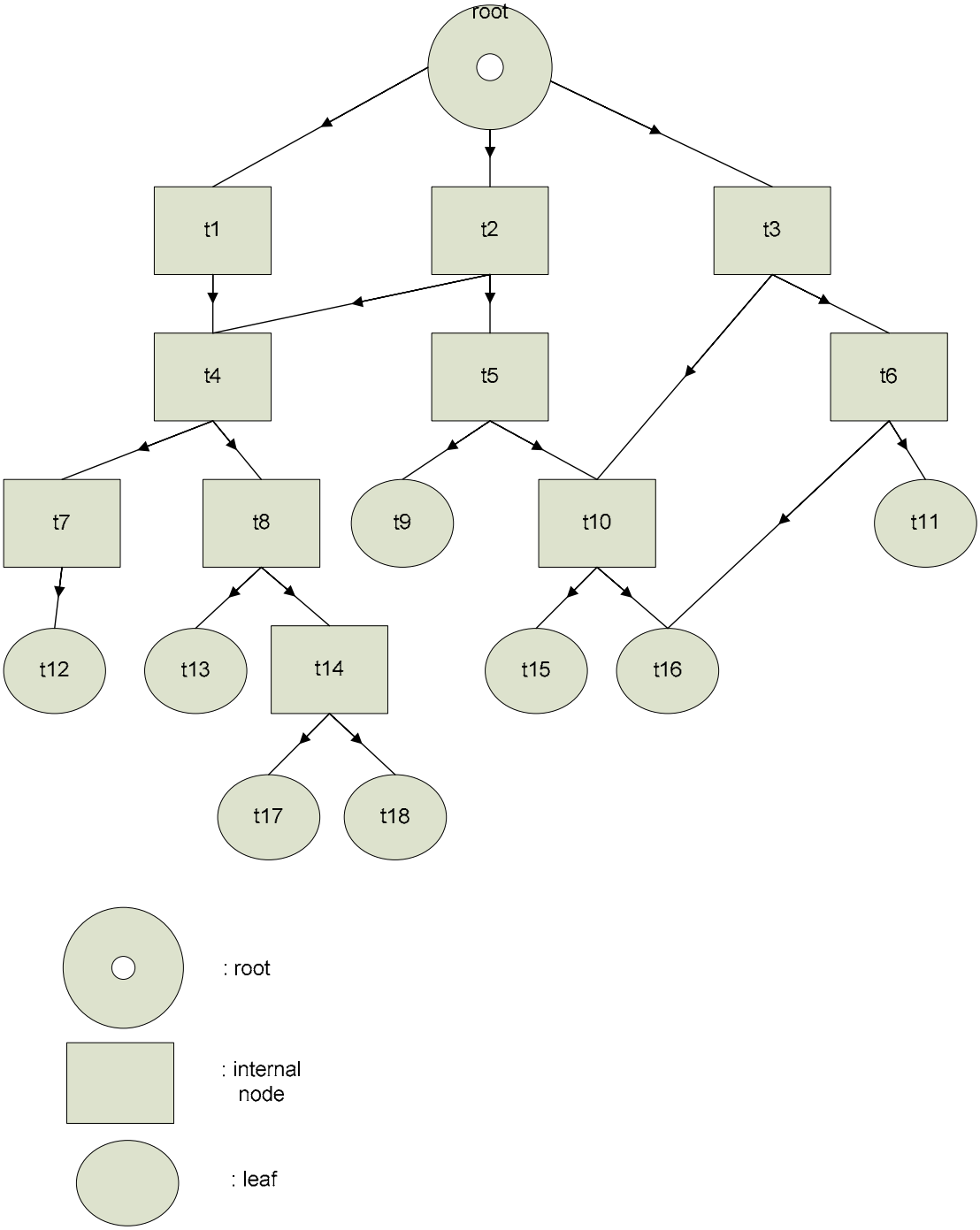
4.4.2 Μέθοδοι σημασιολογικής ομοιότητας

Αφού έχουμε κάνει μία αναλυτική περιγραφή της έννοιας του πληροφοριακού περιεχομένου απομένει να εξηγήσουμε τις μεθόδους σημασιολογικής ομοιότητας που χρησιμοποιήθηκαν και να αναπτύξουμε με παραδείγματα τη θεωρία όπου αυτό κρίνεται απαραίτητο. Για διευκρινιστικούς λόγους να σημειώσουμε ότι οι παρακάτω μέθοδοι αναφέρονται σε ζεύγη όρων (pairwise similarities).

Επίσης, πριν ξεκινήσουμε την παράθεση των διαφόρων μεθοδολογιών αξίζει να αναφέρουμε ότι αυτές γενικά ακολουθούν μία προσέγγιση βασισμένη σε κόμβους (node-based approach) ή μία προσέγγιση βασισμένη σε ακμές (edge-based approach) ή ακόμα και ένα συνδυασμό μεθόδων από τις δύο παραπάνω κατηγορίες.

Γενικά, η προσέγγιση βασισμένη σε κόμβους στηρίζεται στην πληροφορία που μας δίνει κάθε κόμβος της οντολογίας και πιο συγκεκριμένα πόσο κοινή θεωρείται η πληροφορία που αντιπροσωπεύει ένας κόμβος με την πληροφορία που αντιπροσωπεύει ένας άλλος κόμβος. Ως εκ τούτου, η έννοια του πληροφοριακού περιεχομένου που περιγράψαμε παραπάνω αποτελεί προσέγγιση βασισμένη σε κόμβους.

Τέλος, η προσέγγιση βασισμένη σε ακμές στηρίζεται στη μέτρηση της απόστασης (μήκους ακμών) μεταξύ των κόμβων των οποίων η ομοιότητα αναζητείται. Και οι δύο προσεγγίσεις στοχεύουν στον υπολογισμό της σημαντικής ομοιότητας από διαφορετικές όμως οπτικές γωνίες. Η βασισμένη σε ακμές προσέγγιση βρίσκεται πιο κοντά στην ανθρώπινη διαίσθηση, ενώ η βασισμένη σε κόμβους προσέγγιση είναι πιο στέρεη θεωρητικά [6].



Εικόνα 4.9. Παράδειγμα οντολογίας

4.4.2.1 Resnik

Η μέθοδος Resnik είναι η πιο απλή. Στηρίζεται στην έννοια του πληροφοριακού περιεχομένου και του ελαχίστου προγόνου (minimal subsumer). Ως εκ τούτου, μπορεί να θεωρηθεί ως προσέγγιση βασισμένη σε κόμβους. Ο *ελάχιστος πρόγονος (minimal subsumer)* αποτελεί τον πιο ειδικό κοινό πρόγονο των δύο όρων. Με άλλα

λόγια, είναι αυτός με το υψηλότερο πληροφοριακό περιεχόμενο. Συνήθως, βρίσκεται χαμηλότερα στην ιεραρχία της οντολογίας.

Για παράδειγμα, στην Εικ. 4.9, ο ελάχιστος πρόγονος των όρων $t17$ και $t18$ είναι ο $t14$, ενώ ο ελάχιστος πρόγονος των όρων $t15$ και $t17$ είναι ο $t2$.

Ουσιαστικά, η σημασιολογική ομοιότητα δύο όρων κατά Resnik ισούται με το πληροφοριακό περιεχόμενο του ελάχιστου προγόνου (αυτών των δύο όρων). Μάλιστα, συνηθίζεται να κανονικοποιείται η παραπάνω ομοιότητα διαιρώντας με το πληροφοριακό περιεχόμενο ενός φύλλου όπως ακριβώς κάναμε και προηγουμένως.

Έτσι, έχουμε:

$$sim_{norm}(t1,t2) = \frac{\max\limits_{t\in S(t1,t2)}[IC(t)]}{IC(leaf)} = \frac{\max\limits_{t\in S(t1,t2)}[-\log(p(t))]}{-\log(p(leaf))} = 1 - \frac{\min\limits_{t\in S(t1,t2)}[\log(n_t)]}{\log(n_r)}$$

όπου $S(t1,t2)$ είναι το σύνολο των όρων που είναι κοινοί πρόγονοι στους όρους $t1$ και $t2$, ενώ

n_t : Ο αριθμός των φορών που εμφανίζεται ο ελάχιστος πρόγονος (ή κάποιος από τους απογόνους του)

n_r : Ο αριθμός των φορών που εμφανίζεται η ρίζα της οντολογίας (ή κάποιος από τους απογόνους της)

Ενώ η ομοιότητα κατά Resnik είναι εννοιολογικά απλή και αποτελεσματική στις περισσότερες περιπτώσεις, εμφανίζει το μειονέκτημα της αγνόησης των πληροφοριακών περιεχομένων των όρων των οποίων η ομοιότητα αναζητείται. Με αυτό τον τρόπο, η ομοιότητα κατά Resnik διαφορετικών ζευγών όρων που έχουν τον ίδιο ελάχιστο πρόγονο είναι ίδια [18].

Για παράδειγμα, αναφερόμενοι στην Εικ. 4.9 η ομοιότητα κατά Resnik των όρων $t13$ και $t14$ ισούται με $sim_{norm}(t13,t14) = 1 - \frac{\log(n_{t8})}{\log(n_r)}$. Το ίδιο ισχύει και για την ομοιότητα μεταξύ των όρων $t13$ και $t17$ ($sim_{norm}(t13,t17) = 1 - \frac{\log(n_{t8})}{\log(n_r)}$). Στο παράδειγμά μας, έχουμε $n_r = 18+1 = 19$ ενώ $n_{t8} = 4+1 = 5$. Συνεπώς, $sim_{norm}(t13,t14) = sim_{norm}(t13,t17) = 1 - \frac{\log(5)}{\log(19)} = 0.453$. Αξίζει να

σημειωθεί ότι χρησιμοποιήσαμε χωρίς βλάβη της γενικότητας το φυσικό λογάριθμο (βάση *e = 2.7183*).

4.4.2.2 Lin

Η μέθοδος Lin μπορεί να θεωρηθεί το ίδιο απλή, ενώ σε αντίθεση με τη μέθοδο Resnik λαμβάνει υπόψη το πληροφοριακό περιεχόμενο των όρων των οποίων η ομοιότητα αναζητείται. Και η μέθοδος αυτή επίσης ακολουθεί τη βασισμένη σε κόμβους προσέγγιση.

Ουσιαστικά, η σημασιολογική ομοιότητα δύο όρων κατά Lin ισούται με το πληροφοριακό περιεχόμενο του ελάχιστου προγόνου (αυτών των δύο όρων) προς το άθροισμα των πληροφοριακών περιεχομένων των όρων των οποίων η σημασιολογική ομοιότητα αναζητείται.

Έτσι, έχουμε:

$$sim_{norm}(t1,t2) =$$

$$\frac{2 \cdot \max\limits_{t \in S(t1,t2)}[IC(t)]}{IC(t1)+IC(t2)} = \frac{2 \cdot \max\limits_{t \in S(t1,t2)}[-\log(p(t))]}{-[\log(p(t1))+\log(p(t2))]} = \frac{2 \cdot \max\limits_{t \in S(t1,t2)}[-\log(\frac{n_t}{n_r})]}{-[\log(\frac{n_{t1}}{n_r})+\log(\frac{n_{t2}}{n_r})]}$$

Αναπτύσσοντας την παραπάνω σχέση παίρνουμε:

$$sim_{norm}(t1,t2) = \frac{2 \cdot \min\limits_{t \in S(t1,t2)}[\log(n_t)] - 2 \cdot \log(n_r)}{\log(n_{t1})+\log(n_{t2}) - 2 \cdot \log(n_r)}$$

όπου $S(t1,t2)$ είναι το σύνολο των όρων που είναι κοινοί πρόγονοι στους όρους $t1$ και $t2$, ενώ

n_t : Ο αριθμός των φορών που εμφανίζεται ο ελάχιστος πρόγονος (ή κάποιος από τους απογόνους του)

n_r : Ο αριθμός των φορών που εμφανίζεται η ρίζα της οντολογίας (ή κάποιος από τους απογόνους της)

n_{t1} : Ο αριθμός των φορών που εμφανίζεται ο όρος $t1$ (ή κάποιος από τους απογόνους του)

n_{t2} : Ο αριθμός των φορών που εμφανίζεται ο όρος $t2$ (ή κάποιος από τους απογόνους του)

Όταν οι όροι $t1$ και $t2$ είναι φύλλα ($n_{t1} = n_{t2} = 1$) , η παραπάνω σχέση εκφυλίζεται στη σχέση που μας δίνει το μέτρο Resnik:

$$sim_{norm}(t1,t2) = 1 - \frac{\min\limits_{t \in S(t1,t2)}[\log(n_t)]}{\log(n_r)}$$

Η ελάχιστη ομοιότητα συναντάται όταν οι δύο όροι δεν έχουν κανένα κοινό μεταξύ τους και ισούται με μηδέν (0). Ουσιαστικά, σε αυτήν την περίπτωση ο μόνος κοινός πρόγονος που μπορεί να θεωρηθεί για αυτούς τους δύο όρους είναι η ίδια η ρίζα της οντολογίας.

Ενώ, το μέγιστο το συναντάμε όταν οι δύο όροι συμπίπτουν. Εδώ, θεωρείται ότι ο ελάχιστος κοινός πρόγονος των δύο όρων είναι ο ίδιος τους ο εαυτός. Έτσι, η ομοιότητα γίνεται 1 [11].

Για παράδειγμα, αναφερόμενοι στην Εικ. 4.9 η ομοιότητα κατά Lin των όρων $t13$ και $t14$ ισούται με $sim(t13,t14) = \frac{2\log(\frac{n_{t8}}{n_r})}{\log(\frac{n_{t13}}{n_r}) + \log(\frac{n_{t14}}{n_r})}$. Ενώ για την ομοιότητα μεταξύ των όρων $t13$ και $t17$ έχουμε $sim(t13,t17) = \frac{2\log(\frac{n_{t8}}{n_r})}{\log(\frac{n_{t13}}{n_r}) + \log(\frac{n_{t17}}{n_r})}$.

Στο παράδειγμά μας, έχουμε $n_r = 18+1 = 19$ ενώ $n_{t8} = 4+1 = 5$. Επίσης, ισχύει $n_{t13} = 0+1 = 1$, $n_{t14} = 2+1 = 3$ και $n_{t17} = 0+1 = 1$ Συνεπώς, $sim(t13,t14) = 0.557$ και $sim(t13,t17) = 0.453$. Αξίζει να σημειωθεί ότι χρησιμοποιήσαμε χωρίς βλάβη της γενικότητας το φυσικό λογάριθμο (βάση *e = 2.7183*). Όπως παρατηρείται, υπάρχει σαφής διαχωρισμός της σημασιολογικής ομοιότητας μεταξύ όρων που μοιράζονται τον ίδιο ελάχιστο πρόγονο (σε αντίθεση το μέτρο Resnik).

4.4.2.3 Jiang & Conrath

Η μέθοδος Jiang & Conrath αποτελεί συνδυασμό προσέγγισης βασισμένης σε κόμβους και προσέγγισης βασισμένης σε ακμές. Σε σχέση με τις μεθόδους Resnik και Lin λαμβάνει υπόψη πολλές περισσότερες παραμέτρους της δομής μίας οντολογίας. Όπως και στις δύο προηγούμενες μεθόδους, χρησιμοποιεί την έννοια του πληροφοριακού περιεχομένου προσεγγίζοντας έτσι το πρόβλημα με κόμβους. Εκτός, όμως, από το πληροφοριακό περιεχόμενο των κόμβων που εδώ λαμβάνει την έννοια της ισχύος μίας σύνδεσης εσωκλείει και πληροφορία όπως η πυκνότητα δικτύου, το βάθος κόμβου και ο τύπος της σύνδεσης που αποτελούν προσεγγίσεις βασισμένες σε ακμές.

Όλες αυτές οι πληροφορίες ενσωματώνονται στο βάρος μεταξύ δύο γειτονικών κόμβων:

$$wt(c, parent(c)) =$$

$$[\beta + (1-\beta)\frac{\overline{E}}{E(parent(c))}] \cdot [\frac{d(parent(c))+1}{d(parent(c))}]^{\alpha} \cdot [IC(c) - IC(parent(c))] \cdot T(c, parent(c))$$

όπου:

$\overline{E}$: η μέση πυκνότητα (ο αριθμός των ακμών που φεύγουν από όλους τους κόμβους προς το πλήθος των κόμβων)

$E(parent(c))$: η τοπική πυκνότητα (ο αριθμός των ακμών που φεύγουν από τον κόμβο–γονέα)

$d(parent(c))$: το βάθος του κόμβου–γονέα

$LS(c, parent(c)) = IC(c) - IC(parent(c))$: η ισχύς της συνδέσεως

$T(c, parent(c))$: ο τύπος σύνδεσης

β : ο παράγοντας συνεισφοράς πυκνότητας με $0 \leq \beta \leq 1$

α : ο παράγοντας συνεισφοράς βάθος κόμβου με $\alpha \geq 0$

Από τον παραπάνω τύπο διαπιστώνουμε ότι όσο αυξάνει η τοπική πυκνότητα ($E(parent(c))$), τόσο μειώνεται το βάρος μεταξύ του κόμβου και του γονέα του. Μέγιστο έχουμε όταν ο γονέας έχει μόνο ένα παιδί ($E(parent(c)) = 1$).

Ομοίως, όσο αυξάνει το βάθος του κόμβου–γονέα ($d(parent(c))$), τόσο μειώνεται το βάρος μεταξύ του κόμβου και του γονέα του. Μέγιστο έχουμε όταν ο γονέας συνδέεται άμεσα με τη ρίζα ($d(parent(c)) = 1$).

Επίσης, όσο μικρότερη είναι η ισχύς σύνδεσης ($LS(c, parent(c)) = IC(c) - IC(parent(c))$), τόσο μικρότερη είναι η διαφορά μεταξύ γονέα και παιδιού.

Τέλος, όσο οι παράγοντες $\alpha \to 0$ και $\beta \to 1$, τόσο η συνεισφορά της τοπικής πυκνότητας και του βάθους κόμβου γίνεται λιγότερο σημαντική.

Στην απλοποιημένη περίπτωση, θεωρούμε ότι οι δύο παραπάνω όροι δεν συνεισφέρουν καθόλου. Με άλλα λόγια, $\alpha = 0$ και $\beta = 1$.

Έτσι, ο τύπος γίνεται:

$$wt(c, parent(c)) = [IC(c) - IC(parent(c))] \cdot T(c, parent(c))$$

Κατόπιν της ανάλυσης του βάρους (απόστασης) μεταξύ δύο γειτονικών κόμβων (γονέα–παιδιού), είμαστε πια σε θέση να ορίσουμε την απόσταση μεταξύ δύο όρων.

Η απόσταση, λοιπόν, μεταξύ των όρων $t1$ και $t2$ ορίζεται ως εξής:

$$dist(t1, t2) = \sum_{c \in \{path(t1,t2) - minimal\_subsumer(t1,t2)\}} wt(c, parent(c))$$

όπου:

$path(t1, t2)$: το ελάχιστο μονοπάτι μεταξύ των $t1$ και $t2$

Παίρνοντας την απλοποιημένη περίπτωση έχουμε $wt(c, parent(c)) = [IC(c) - IC(parent(c))] \cdot T(c, parent(c))$. Όπως όμως είχαμε σημειώσει σε πρότερο σημείο, η Γονιδιακή Οντολογία αποτελείται κυρίως από “is-a” σχέσεις (ο αριθμός τους υπερτερεί κατά εφτά περίπου φορές αυτόν των “part-of” σχέσεων) ενώ υπάρχει μία τάση προς αύξηση αυτής της αναλογίας (μεταξύ “is-a” και “part-of” σχέσεων) [13]. Έτσι, μπορούμε να θεωρήσουμε ένα και μοναδικό τύπο σύνδεσης ο οποίος θα είναι κοινός σε όλες τις ακμές. Μ’ άλλα λόγια, $T(c, parent(c)) = 1$. Έτσι, παίρνουμε: $wt(c, parent(c)) = IC(c) - IC(parent(c))$.

Επομένως, η απόσταση μεταξύ των όρων $t1$ και $t2$ υπολογίζεται ως εξής:

$$dist(t1,t2) = \sum_{c\in\{path(t1,t2)-minimal\_subsumer(t1,t2)\}} wt(c, parent(c)) \Leftrightarrow$$

$$dist(t1,t2) = wt(t1, parent(t1)) + wt(parent(t1), parent(parent(t1))) + ... + wt(t', minimal\_subsumer(t1,t2)) + wt(t2, parent(t2)) + wt(parent(t2), parent(parent(t2))) + ... + wt(t'', minimal\_subsumer(t1,t2)) \Leftrightarrow$$

$$dist(t1,t2) = IC(t1) - \cancel{IC(parent(t1))} + \cancel{IC(parent(t1))} - \cancel{IC(parent(parent(t1)))} + ... + \cancel{IC(t')} - IC(minimal\_subsumer(t1,t2)) + IC(t2) - \cancel{IC(parent(t2))} + \cancel{IC(parent(t2))} - \cancel{IC(parent(parent(t2)))} + ... + \cancel{IC(t'')} - IC(minimal\_subsumer(t1,t2)) \Leftrightarrow$$

$$dist(t1,t2) = IC(t1) + IC(t2) - 2\cdot IC(minimal\_subsumer(t1,t2))$$

Καταλήγουμε έτσι στον πολύ κομψό τύπο ότι η απόσταση μεταξύ δύο όρων ισούται με το άθροισμα των πληροφοριακών τους περιεχομένων μείον της διπλάσιας ποσότητας του πληροφοριακού περιεχομένου του ελαχίστου προγόνου τους [6].

Ελάχιστο έχουμε όταν οι δύο όροι των οποίων η ομοιότητα αναζητείται συμπίπτουν. Έτσι: $t1 \equiv t2 \equiv minimal\_subsumer(t1,t2)$. Σε αυτήν την περίπτωση, η απόσταση μεταξύ των όρων είναι μηδενική (0). Έχουν συνεπώς μέγιστη ομοιότητα.

Μέγιστο έχουμε όταν οι δύο όροι αποτελούν φύλλα και είναι τόσο μακρινοί μεταξύ τους ώστε ο ελάχιστος τους πρόγονος είναι η ρίζα της οντολογίας ($IC(minimal\_subsumer(t1,t2)) = 0$). Έτσι, εάν συμβολίσουμε με n_r τον αριθμό των φορών που εμφανίζεται η ρίζα της οντολογίας (ή κάποιος από τους απογόνους της), έχουμε

$$dist(t1,t2) = -\log(\frac{1}{n_r}) - \log(\frac{1}{n_r}) = -2\cdot\log(\frac{1}{n_r}) = 2\cdot\log(n_r).$$

Τέλος, για την κανονικοποίηση της παραπάνω απόστασης αρκεί να διαιρέσουμε με τη μέγιστη τιμή ($2\cdot\log(n_r)$). Ενώ, λαμβάνουμε την ομοιότητα μεταξύ των όρων αν αφαιρέσουμε την κανονικοποιημένη απόσταση με το 1.

$$dist_{norm}(t1,t2) = \frac{dist(t1,t2)}{2\cdot\log(n_r)}$$

$$sim_{norm}(t1,t2) = 1 - dist_{norm}(t1,t2)$$

Για παράδειγμα, αναφερόμενοι στην Εικ. 4.9 η ομοιότητα κατά Jiang & Conrath των όρων $t13$ και $t14$ ισούται με $sim(t13,t14) = 1 - \frac{-\log(n_{t13}) - \log(n_{t14}) + 2\log(n_{t8})}{2\log(n_r)}$. Ενώ για την ομοιότητα μεταξύ των όρων $t13$ και $t17$ έχουμε $sim(t13,t17) = 1 - \frac{-\log(n_{t13}) - \log(n_{t17}) + 2\log(n_{t8})}{2\log(n_r)}$.

Στο παράδειγμά μας, έχουμε $n_r = 18+1 = 19$ ενώ $n_{t8} = 4+1 = 5$. Επίσης, ισχύει $n_{t13} = 0+1 = 1$, $n_{t14} = 2+1 = 3$ και $n_{t17} = 0+1 = 1$ Συνεπώς, $sim(t13,t14) = 0.64$ και $sim(t13,t17) = 0.453$. Αξίζει να σημειωθεί ότι χρησιμοποιήσαμε χωρίς βλάβη της γενικότητας το φυσικό λογάριθμο (βάση *e = 2.7183*). Όπως παρατηρείται, υπάρχει σαφής διαχωρισμός της σημασιολογικής ομοιότητας μεταξύ όρων που μοιράζονται τον ίδιο ελάχιστο πρόγονο (σε αντίθεση το μέτρο Resnik).

4.4.2.4 Cao

Η μέθοδος Cao υπολογίζει το σύνολο του πληροφοριακού περιεχομένου που περιέχουν οι κοινοί πρόγονοι των όρων των οποίων η ομοιότητα αναζητείται προς το σύνολο του πληροφοριακού περιεχομένου που περιέχουν όλοι οι πρόγονοι αυτών των όρων. Αποτελεί και αυτή προσέγγιση βασισμένη σε κόμβους.

Έτσι, έχουμε:

$$sim(t1,t2) = \frac{\sum_{t \in S(t1,t2)} IC(t)}{\sum_{t \in S(t1,t2)} IC(t) + \sum_{t' \notin S(t1,t2)} IC(t')}$$

όπου $S(t1,t2)$ είναι το σύνολο των κοινών προγόνων των $t1$ και $t2$, ενώ με $t' \notin S(t1,t2)$ εννοούμε κάθε πρόγονο των $t1$ ή $t2$ που δεν ανήκει στο σύνολο των κοινών προγόνων. Με άλλα λόγια, αν συμβολίσουμε με $S(t1)$ και $S(t2)$ τα σύνολα των προγόνων των όρων $t1$ και $t2$ αντίστοιχα, ο συμβολισμός $t' \notin S(t1,t2)$ ισοδυναμεί με το συμβολισμό $t' \in ((S(t1) \cup S(t2)) - S(t1,t2))$.

Στο σύνολο $S(t1)$ συμπεριλαμβάνεται και ο όρος $t1$. Ομοίως και στο σύνολο

$S(t2)$.

Η ελάχιστη ομοιότητα συναντάται όταν οι δύο όροι δεν έχουν κανένα κοινό μεταξύ τους και ισούται με μηδέν (0). Ουσιαστικά, σε αυτήν την περίπτωση ο μόνος κοινός πρόγονος που μπορεί να θεωρηθεί για αυτούς τους δύο όρους είναι η ίδια η ρίζα της οντολογίας.

Ενώ, το μέγιστο το συναντάμε όταν οι δύο όροι συμπίπτουν. Εδώ, θεωρείται ότι ο ελάχιστος κοινός πρόγονος των δύο όρων είναι ο ίδιος τους ο εαυτός. Έτσι, η ομοιότητα γίνεται 1.

Για παράδειγμα, αναφερόμενοι στην Εικ. 4.9 η ομοιότητα κατά Cao των όρων $t9$ και $t16$ ισούται με

$$sim(t9,t16) = \frac{IC(t5) + IC(t2) + IC(root)}{IC(t5) + IC(t2) + IC(root) + IC(t16) + IC(t10) + IC(t9) + IC(t3)}$$

Στο παράδειγμά μας, έχουμε $n_r = 18+1 = 19$ ενώ $n_{t5} = 4+1 = 5$, $n_{t2} = 13+1 = 14$, $n_{t16} = 0+1 = 1$, $n_{t10} = 2+1 = 3$, $n_{t9} = 0+1 = 1$ και $n_{t3} = 5+1 = 6$. Συνεπώς, $sim(t9,t16) = 0.156$. Αξίζει να σημειωθεί ότι χρησιμοποιήσαμε χωρίς βλάβη της γενικότητας το φυσικό λογάριθμο (βάση *e = 2.7183*).

4.4.2.5 oldZZL

Η μέθοδος oldZZL αρχικά εντοπίζει τα μονοπάτια του όρου $t1$ προς τη ρίζα που έχουν το μεγαλύτερο μήκος. Πράττει το ίδιο και για τον όρο $t2$. Έπειτα, υπολογίζει το μέγιστο υπο-μονοπάτι που είναι κοινό σε κάποιο από τα μέγιστα μονοπάτια του $t1$ και σε κάποιο από τα μέγιστα μονοπάτια του $t2$. Η μέθοδος oldZZL αποτελεί προσέγγιση βασισμένη σε ακμές.

Έτσι, η σημασιολογική ομοιότητα των $t1$ και $t2$ έχει τη μορφή:

$$sim(t1,t2) = 1 - (\frac{1}{2^{l_{common}}} - \frac{1}{2^{l_1+1}} - \frac{1}{2^{l_2+1}})$$

όπου:

$l_1 = \max[length(paths(t1))]$

$l_2 = \max[length(paths(t2))]$

$l_{common} = \max[length(A_i \cap B_j)] \;\; \mu\varepsilon \;\; A_i \in max\_paths(t1) \;\; \kappa\alpha\iota \;\; B_j \in max\_paths(t2)$

$paths(t1)$: το σύνολο των μονοπατιών του όρου $t1$ προς τη ρίζα

$paths(t2)$: το σύνολο των μονοπατιών του όρου $t2$ προς τη ρίζα

$max\_paths(t1) \subseteq paths(t1) \;\; \kappa\alpha\iota \;\; N(max\_paths(t1)) = l_1$

$max\_paths(t2) \subseteq paths(t2) \;\; \kappa\alpha\iota \;\; N(max\_paths(t2)) = l_2$

l_{common} : το μέγιστο υπο-μονοπάτι που είναι κοινό ανάμεσα σε κάποιο από τα μέγιστα μονοπάτια του όρου $t1$ ($max\_paths(t1)$) και σε κάποιο από τα μέγιστα μονοπάτια του όρου $t2$ ($max\_paths(t2)$)

Είναι φανερό ότι όσο πιο κοινή είναι η διαδρομή μεταξύ ενός από τα μεγαλύτερα μονοπάτια του όρου $t1$ και ενός από τα μεγαλύτερα μονοπάτια του όρου $t2$, τόσο οι όροι $t1$ και $t2$ θα παρουσιάζουν μεγαλύτερη ομοιότητα. Από την άλλη, όσο λιγότερο απέχουν οι όροι $t1$ και $t2$ από τη ρίζα, τόσο μεγαλύτερη θα είναι η ομοιότητα μεταξύ τους [30].

Για παράδειγμα, για τον υπολογισμό της ομοιότητας κατά oldZZL των όρων $t8$ και $t16$ (αναφερόμενοι στην Εικ. 4.9) έχουμε τα εξής:

$$paths(t8) = \{\{t8-t4-t1-root\},\{t8-t4-t2-root\}\}$$

$$paths(t16) = \{\{t16-t10-t5-t2-root\},\{t16-t10-t3-root\},\{t16-t6-t3-root\}\}$$

Άρα:

$$l_8 = \max[length(paths(t8))] = N(\{t8-t4-t1-root\}) = N(\{t8-t4-t2-root\}) = 3$$

$$l_{16} = \max[length(paths(t16))] = N(\{t16-t10-t5-t2-root\}) = 4$$

Ενώ:

$$l_{common} = \max[N(\{t8-t4-t1-root\} \cap \{t16-t10-t5-t2-root\}),$$
$$N(\{t8-t4-t2-root\} \cap \{t16-t10-t5-t2-root\})] \Leftrightarrow$$
$$l_{common} = \max[N(\{root\}), N(\{t2-root\})] = \max[0,1] = 1$$

Έτσι:

$$sim(t8,t16) = 1 - (\frac{1}{2^1} - \frac{1}{2^{4+1}} - \frac{1}{2^{3+1}}) = 0.594$$

4.4.2.6 ZZL

Η μέθοδος ZZL, που αποτελεί προσέγγιση βασισμένη σε ακμές, υπολογίζει τη σημασιολογική ομοιότητα μεταξύ των όρων $t1$ και $t2$ ως εξής:

$$sim(t1,t2) = 1 - (\frac{1}{2^{l_{ms}}} - \frac{1}{2^{l_1+1}} - \frac{1}{2^{l_2+1}})$$

όπου:

$l_1 = \max[length(paths'(t1))]$

$l_2 = \max[length(paths'(t2))]$

$l_{ms} = length(path(ms))$

$paths'(t1) = paths(t1,ms) \cup path(ms)$

$paths'(t2) = paths(t2,ms) \cup path(ms)$

$$length(path(ms)) = \max[length(paths'(t1))] - \max[length(paths(t1,ms))]$$
$$= \max[length(paths'(t2))] - \max[length(paths(t2,ms))]$$

και

ms : ο ελάχιστος πρόγονος των $t1$ και $t2$, ενώ ισχύει πάντα $l_1 \geq l_{ms}$ και $l_2 \geq l_{ms}$

Με $paths(t1)$ συμβολίζουμε όλα τα μονοπάτια του όρου $t1$ προς τη ρίζα, ενώ με $paths'(t1)$ όλα τα μονοπάτια του όρου $t1$ προς τη ρίζα που διέρχονται από τον ελάχιστο πρόγονο (ms) των $t1$ και $t2$ ($paths'(t1) \subseteq paths(t1)$). Το ίδιο ισχύει και για τον όρο $t2$. Επίσης, με $path(ms)$ συμβολίζουμε κάθε μονοπάτι του ελάχιστου προγόνου προς τη ρίζα που συναντάται τόσο στα μονοπάτια του $t1$ όσο και στα μονοπάτια του $t2$ προς τη ρίζα.

Αναλυτικότερα, εντοπίζεται αρχικά ο ελάχιστος πρόγονος. Έπειτα, όλα τα μονοπάτια από τον $t1$ προς τη ρίζα και από τον $t2$ προς τη ρίζα που διέρχονται

από τον ελάχιστο πρόγονο. Τέλος, επιλέγουμε το μονοπάτι με το μεγαλύτερο μήκος από τον όρο $t1$ προς τη ρίζα καθώς και το μονοπάτι με το μεγαλύτερο μήκος από τον όρο $t2$ προς τη ρίζα των οποίων η διαδρομή ταυτίζεται από τον ελάχιστο κοινό πρόγονο έως τη ρίζα.

Είναι φανερό ότι όσο ειδικότερος είναι ο ελάχιστος κοινός πρόγονος τόσο μεγαλύτερη θα είναι η ομοιότητα μεταξύ των δύο όρων. Με άλλα λόγια, όσο χαμηλότερα βρίσκεται στην ιεραρχία (μεγαλύτερο l_{ms}), οι όροι $t1$ και $t2$ θα παρουσιάζουν μεγαλύτερη ομοιότητα. Από την άλλη, όσο λιγότερο απέχουν οι όροι $t1$ και $t2$ από τη ρίζα, τόσο μεγαλύτερη θα είναι η ομοιότητα μεταξύ τους.

Για παράδειγμα, για τον υπολογισμό της ομοιότητας κατά ZZL των όρων $t8$ και $t16$ (αναφερόμενοι στην Εικ. 4.9) έχουμε τα εξής:

$$paths'(t8) = \{\{t8 - t4 - t2 - root\}\}$$

$$paths'(t16) = \{\{t16 - t10 - t5 - t2 - root\}\}$$

ενώ $ms = t2$

Το μεγαλύτερο μονοπάτι του όρου $t8$ του οποίου η διαδρομή από τον ελάχιστο πρόγονο προς τη ρίζα συμπίπτει με την αντίστοιχη διαδρομή του μεγαλύτερου μονοπατιού του όρου $t16$ είναι το $\{t8 - t4 - t2 - root\}$. Ενώ το μεγαλύτερο μονοπάτι του όρου $t16$ είναι το $\{t16 - t10 - t5 - t2 - root\}$. Η κοινή διαδρομή του ελάχιστου προγόνου προς τη ρίζα είναι $\{t2 - root\}$.

Άρα: $l_{8} = \max[3] = 3,\ l_{16} = \max[4] = 4,\ l_{ms} = 1$

Έτσι:

$$sim(t8, t16) = 1 - (\frac{1}{2^1} - \frac{1}{2^{4+1}} - \frac{1}{2^{3+1}}) = 0.594$$

4.4.2.7 Leacock & Chodorow

Η μέθοδος Leacock & Chodorow βασίζεται στα μήκη των μονοπατιών μεταξύ των όρων των οποίων η σημασιολογική ομοιότητα αναζητείται. Πιο συγκεκριμένα, εντοπίζεται κατ' αρχήν το ελάχιστο μονοπάτι μεταξύ των δύο όρων και έπειτα το συνολικό βάθος της οντολογίας (που ισούται με τα βάθος του πιο μακρινού φύλλου από τη ρίζα). Όπως οι μέθοδοι oldZZL και ZZL, έτσι και η μέθοδος Leacock & Chodorow ακολουθεί τη βασισμένη σε ακμές προσέγγιση.

Έτσι, η σημασιολογική ομοιότητα των $t1$ και $t2$ έχει τη μορφή:

$$sim(t1,t2) = -\log\frac{length(min\_path(t1,t2))}{2D}$$

όπου:

$min\_path(t1,t2)$: το ελάχιστο μονοπάτι μεταξύ των όρων $t1$ και $t2$

D : το συνολικό βάθος της οντολογίας (βάθος του πιο μακρινού φύλλου από τη ρίζα)

Ελάχιστη ομοιότητα έχουμε όταν οι όροι $t1$ και $t2$ είναι φύλλα και απέχουν μεταξύ τους απόσταση $2D$. Μ' άλλα λόγια, όταν κάθε ένας από αυτούς τους όρους απέχει από τη ρίζα απόσταση D. Η ομοιότητα αυτή ισούται με μηδέν (0).

Μέγιστη ομοιότητα έχουμε όταν οι όροι $t1$ και $t2$ συμπίπτουν. Επειδή σε αυτή την περίπτωση είναι $length(min\_path(t1,t2)) = 0$, έχουμε $sim(t1,t2) = \infty$ [16].

Έτσι, για να αποκτήσουμε την κανονικοποιημένη μορφή της παραπάνω ομοιότητας (εύρος [0,1]) πρέπει να κάνουμε μία μικρή τροποποίηση στον παραπάνω τύπο ως εξής:

$$sim(t1,t2) = \begin{cases} 1 & , length(min\_path(t1,t2)) = 0 \\ \dfrac{-\log\dfrac{length(min\_path(t1,t2))+0.1}{2D}}{-\log\dfrac{1}{2D}} & , length(min\_path(t1,t2)) \neq 0 \end{cases}$$

όπου:

$-\log\frac{1}{2D}$: η δεύτερη μεγαλύτερη τη τάξει ομοιότητα μεταξύ δύο όρων (αυτή μεταξύ γονέα – παιδιού)

Για παράδειγμα, για τον υπολογισμό της ομοιότητας κατά Leacock & Chodorow των όρων $t8$ και $t16$ (αναφερόμενοι στην Εικ. 4.9) έχουμε τα εξής:

$$paths(t8,t16) =$$
$$\{\{t8-t4-t2-t5-t10-t16\},\{t8-t4-t1-root-t2-t5-t10-t16\},$$
$$\{t8-t4-t1-root-t3-t10-t16\},\{t8-t4-t1-root-t3-t6-t16\}\}$$

Έτσι:

$$min\_path(t8, t16) = \{t8 - t4 - t2 - t5 - t10 - t16\}$$

και

$$length(min\_path(t8, t16)) = 5$$

Επίσης: $D = depth(t17) = depth(t18) = 5$

Έτσι:

$$sim(t8, t16) = \frac{-\ln \frac{5 + 0.1}{2 \cdot 5}}{-\ln \frac{1}{2 \cdot 5}} = 0.292$$

Αξίζει να σημειωθεί ότι χρησιμοποιήσαμε χωρίς βλάβη της γενικότητας το φυσικό λογάριθμο (βάση *e = 2.7183*).

4.4.2.8 Wu & Palmer

Η μέθοδος Wu & Palmer υπολογίζει τη σημασιολογική ομοιότητα μεταξύ των όρων $t1$ και $t2$ ως εξής:

$$sim(t1, t2) = \frac{2N_3}{N_1 + N_2 + 2N_3}$$

όπου:

$$N_1 = length(min\_path(t1, ms))$$
$$N_2 = length(min\_path(t2, ms))$$
$$N_3 = length(max\_path(ms, root))$$

και

ms : ο ελάχιστος κοινός πρόγονος των $t1$ και $t2$

Ελάχιστο έχουμε όταν οι δύο όροι των οποίων η ομοιότητα αναζητείται είναι τόσο μακρινοί ώστε ο κοινός τους πρόγονος να είναι η ρίζα της οντολογίας ($ms \equiv root$). Σε αυτή την περίπτωση έχουμε $N_3 = length(max\_path(root, root)) = 0$. Οπότε, η ομοιότητα ισούται με μηδέν (0).

Μέγιστο συναντούμε όταν οι όροι $t1$ και $t2$ συμπίπτουν. Έτσι, έχουμε: $t1 \equiv t2 \equiv ms$. Συνεπώς: $N_1 = N_2 = 0$. Καταλήγουμε λοιπόν ότι η ομοιότητα σε αυτή την περίπτωση ισούται με ένα (1) [27].

Όπως παρατηρούμε, η σχέση που δίνει την ομοιότητα κατά Wu & Palmer είναι αρκετά απλή και εύκολα μπορούμε να συνάγουμε τα εξής: Όσο βαθύτερα βρίσκεται ο ελάχιστος κοινός πρόγονος (ms) στην οντολογία, τόσο μεγαλύτερη είναι η ομοιότητα. Από την άλλη, όσο περισσότερο απέχουν οι δύο όροι από τον κοινό πρόγονο (μεγάλα N_1 και N_2), τόσο η ομοιότητα μεταξύ τους μειώνεται.

Για παράδειγμα, για τον υπολογισμό της ομοιότητας κατά Wu & Palmer των όρων $t8$ και $t16$ (αναφερόμενοι στην Εικ. 4.9) έχουμε τα εξής:

$ms(t8, t16) = t2$ και

$min\_path(t8, ms) = \{t8 - t4 - t2\}$

$min\_path(t16, ms) = \{t16 - t10 - t5 - t2\}$

$max\_path(ms, root) = \{t2 - root\}$

Άρα: $N_1 = 2, N_2 = 3, N_3 = 1$

Έτσι:

$$sim(t8, t16) = \frac{2 \cdot 1}{2 + 3 + 2 \cdot 1} = 0.286$$

4.4.2.9 Resnik GRaSM

Η μέθοδος Resnik GRaSM βασίζεται στην έννοια των κοινών διαχωριστικών προγόνων (common disjunctive ancestors) σε αντίθεση με τη μέθοδο Resnik που βασίζεται μόνο στον ελάχιστο κοινό πρόγονο. Πριν όμως προχωρήσουμε στην ανάλυση αυτής της μεθόδου είναι αναγκαίο να εξηγήσουμε την έννοια των *διαχωριστικών προγόνων (disjunctive ancestors)*. Δύο πρόγονοι ενός όρου θεωρούνται διαχωριστικοί αν υπάρχουν *ανεξάρτητα μονοπάτια* που οδηγούν από κάθε πρόγονο στον όρο. Με τον όρο *ανεξάρτητα μονοπάτια* εννοούμε να υπάρχει στο ένα μονοπάτι τουλάχιστον ένας όρος που δεν βρίσκεται στο άλλο μονοπάτι. Ουσιαστικά, δύο διαχωριστικοί πρόγονοι ενός όρου εκφράζουν δύο διακριτές υποστάσεις αυτού του όρου. Πριν, με τη χρησιμοποίηση μόνο του πιο

πληροφοριακού κοινού προγόνου (ελαχίστου προγόνου) λαμβανόταν υπόψη μόνο μία από τις πολλές υποστάσεις ενός όρου.

Πιο φορμαλιστικά, δύο όροι a_1, a_2 θεωρούνται διαχωριστικοί πρόγονοι ενός όρου t αν ισχύει:

$$DisjAnc(t) = \{(a_1, a_2) \mid \exists p : (p \in paths(a_1, t) \wedge (a_2 \notin p)) \wedge \exists p^{'} : (p^{'} \in paths(a_2, t) \wedge (a_1 \notin p^{'}))\}$$

Εάν $a_1 \notin Ancestors(a_2)$ και $a_2 \notin Ancestors(a_1)$, τότε οι a_1 και a_2 είναι διαχωριστικοί πρόγονοι του t. Από την άλλη, εάν $a_1 \in Ancestors(a_2)$ ή $a_2 \in Ancestors(a_1)$ είναι ακόμη πιθανό ότι οι a_1 και a_2 αντιπροσωπεύουν διαχωριστικούς προγόνους του t.

Επίσης μία ισοδύναμη μορφή της παραπάνω έκφρασης είναι:

$$DisjAnc(t) = \{(a_1, a_2) \mid IC((a_1) \leq IC(a_2) (\forall n, m, k (\mid paths(a_1, t) \mid = m \wedge \mid paths(a_2, t) \mid = n \wedge \mid paths(a_1, a_2) \mid = k) \Rightarrow (m > 0 \wedge n > 0 \wedge m > n \cdot k)\}$$

όπου:

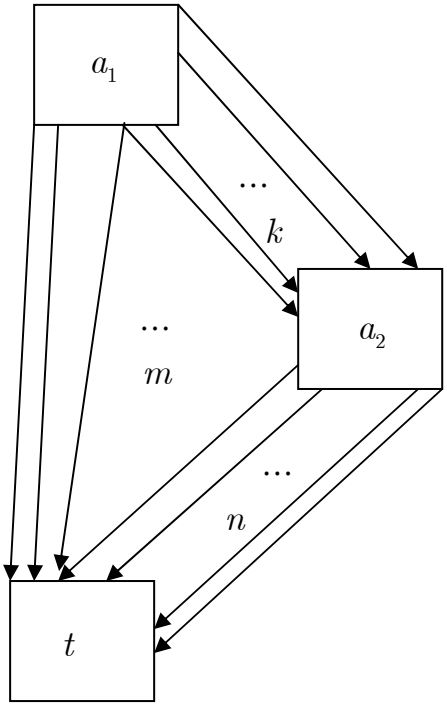
$\mid paths(a_1, t) \mid$: το πλήθος των μονοπατιών από τον όρο a_1 στον όρο t

$\mid paths(a_2, t) \mid$: το πλήθος των μονοπατιών από τον όρο a_2 στον όρο t

$\mid paths(a_1, a_2) \mid$: το πλήθος των μονοπατιών από τον όρο a_1 στον όρο a_2

Οι δύο μορφές είναι ισοδύναμες γιατί όταν $IC(a_1) \leq IC(a_2)$, ο όρος a_1 βρίσκεται συνήθως υψηλότερα στην ιεραρχία από τον όρο a_2. Έτσι, κανένα από τα n μονοπάτια που οδηγούν από τον όρο a_2 στον όρο t δεν θα περιέχει τον όρο a_1. Επίσης, κάθε μονοπάτι από τον όρο a_1 στον όρο t που διέρχεται από τον όρο a_2 θα ακολουθεί τη διαδρομή ενός από τα k μονοπάτια από τον όρο a_1 στον a_2 και τη διαδρομή ενός από τα n μονοπάτια από τον a_2 στον όρο t. Επομένως, υπάρχουν $n \cdot k$ διαφορετικοί συνδυασμοί που οδηγούν από τον a_1 στον t μέσω του a_2. Γενικά, ο a_1 οδηγεί στον t μέσω m μονοπατιών. Έτσι, για να υπάρχει έστω ένα μονοπάτι από τον a_1 στον t που δεν περιέχει τον a_2 θα πρέπει $m > n \cdot k$.

Τα παραπάνω οπτικοποιούνται για την καλύτερη κατανόηση στην ακόλουθη εικόνα:



Επιπλέον, αν υπάρχει ένα μονοπάτι από τον a_2 στον t που δεν περιέχει τον a_1 και αντίστοιχα ένα μονοπάτι από τον a_1 στον t που δεν περιέχει τον a_2 τότε οι a_1 και a_2 είναι διαχωριστικοί πρόγονοι του t, γιατί δεδομένου ότι $IC(a_1) \leq IC(a_2)$ οι διαδρομές του a_1 στον t θα περιέχουν όλα τα πιθανά μονοπάτια που διέρχονται από τον a_2. Έτσι, θα ισχύει: $m \geq n \cdot k + 1 > n \cdot k$ αφού υπάρχει ένα τουλάχιστον μονοπάτι από τον a_1 στον t που δεν περνάει από τον a_2.

Συνεπώς, οι κοινοί διαχωριστικοί πρόγονοι δύο όρων $t1$ και $t2$ δίνονται από τη σχέση:

$$CommonDisjAnc(t1, t2) =$$
$$\{a_1 \mid a_1 \in CommonAnc(t1, t2) \wedge$$
$$(\forall a_2 : [(a_2 \in CommonDisjAnc(t1, t2)) \wedge (IC(a_1) \leq IC(a_2)) \wedge (a_1 \neq a_2)]$$
$$\Rightarrow [(a_1, a_2) \in (DisjAnc(t1) \cup DisjAnc(t2))]\}$$

Με απλά λόγια, κοινός διαχωριστικός πρόγονος των $t1$ και $t2$ είναι κάθε κοινός πρόγονος των $t1$ και $t2$ που με κάθε έναν από τους ειδικότερους κοινούς διαχωριστικούς προγόνους που έχουν επισημειωθεί σε προηγούμενα στάδια ανήκουν στο σύνολο των διαχωριστικών προγόνων του $t1$ ($DisjAnc(t1)$) ή του $t2$ ($DisjAnc(t2)$). Επίσης, είναι σημαντικό να αναφέρουμε ότι αρχικοποιούμε το

σύνολο των κοινών διαχωριστικών προγόνων θέτοντάς το αρχικά ίσο με τον ειδικότερο κοινό πρόγονο των $t1$ και $t2$.

Έτσι, ενώ πριν η σημασιολογική ομοιότητα δύο όρων κατά Resnik δινόταν από το πληροφοριακό περιεχόμενο του ελάχιστου προγόνου (αυτών των δύο όρων), τώρα δίνεται από το μέσο όρο των πληροφοριακών περιεχομένων των κοινών διαχωριστικών προγόνων των δύο όρων.

$$sim(t1,t2) = Share_{GraSM}(t1,t2) = \frac{\sum_{a \in CommonDisjAnc(t1,t2)} IC(a)}{\mid CommonDisjAnc(t1,t2) \mid}$$

όπου:

$\mid CommonDisjAnc(t1,t2) \mid$: το πλήθος των κοινών διαχωριστικών προγόνων

Σε γενικές γραμμές, η χρησιμοποίηση κοινών διαχωριστικών προγόνων αντί του ελάχιστου κοινού προγόνου τείνει να μειώνει την ομοιότητα μεταξύ των όρων των οποίων η ομοιότητα αναζητείται, γιατί κάθε όρος εμφανίζει παραπάνω από μία υποστάσεις εκφράζοντας έτσι μικρότερη συγγένεια με τον άλλο όρο [3].

Για την κανονικοποίηση του παραπάνω μέτρου ισχύουν ακριβώς τα ίδια με τη μέθοδο Resnik.

4.4.2.10 Lin GRaSM

Η μέθοδος Lin GRaSM δίνεται από τον παρακάτω τύπο:

$$sim_{norm}(t1,t2) = \frac{2 \cdot Share_{GraSM}(t1,t2)}{IC(t1) + IC(t2)}$$

όπου:

$Share_{GraSM}(t1,t2)$ ορίστηκε λίγο πιο πάνω

Για την κανονικοποίηση του παραπάνω μέτρου ισχύουν ακριβώς τα ίδια με τη μέθοδο Lin.

4.4.2.11 Jiang & Conrath GRaSM

Η μέθοδος Jiang & Conrath GRaSM δίνεται από τον παρακάτω τύπο:

$$dist(t1, t2) = IC(t1) + IC(t2) - 2 \cdot Share_{GraSM}(t1, t2)$$

όπου:

$Share_{GraSM}(t1, t2)$ ορίστηκε λίγο πιο πάνω

Για την μετατροπή της απόστασης σε ομοιότητα και την κανονικοποίηση του παραπάνω μέτρου ισχύουν ακριβώς τα ίδια με τη μέθοδο Jiang & Conrath.

5 Η ανάπτυξη του εργαλείου SoFoCles

Το SoFoCles είναι ένα εργαλείο που ενσωματώνει την υπάρχουσα βιολογική γνώση σε δεδομένα μικροσυστοιχιών για την καλύτερη εκμάθηση των συνόλων δεδομένων με απώτερο σκοπό την βελτίωση της ακρίβειας πρόβλεψης (ταξινόμησης).

Παρέχει τη δυνατότητα εισαγωγής πολλών παραμέτρων καθιστώντας έτσι δυνατή την εξέταση πολλών διαφορετικών πιθανών σεναρίων. Ενώ, η εκτίμηση αυτών των διαφορετικών σεναρίων μπορεί να γίνει με έναν από τους υπάρχοντες αλγορίθμους εκμάθησης.

5.1 Λόγοι που οδήγησαν στην ανάπτυξη του SoFoCles

Όπως υπογραμμίσαμε στο προηγούμενο κεφάλαιο, οι μέθοδοι φιλτραρίσματος γνωρισμάτων εμφανίζουν πολλά πλεονεκτήματα, όμως όπως σε κάθε εφαρμοζόμενη μέθοδο έτσι και σε αυτές υπάρχει ένας συγκερασμός πλεονεκτημάτων και μειονεκτημάτων από τη χρήση τους.

Τα αρνητικά από τη χρησιμοποίηση των παραπάνω μεθόδων ώθησαν στην ανάγκη για την υλοποίηση ενός νέου εργαλείου όπως είναι το SoFoCles. Εύλογα όμως κάποιος διερωτάται ποιες είναι οι ατέλειες αυτών των μεθόδων.

Στο σημείο αυτό λοιπόν είναι η κατάλληλη στιγμή να τις παραθέσουμε:

- Κατ' αρχήν, με τη χρησιμοποίηση μεθόδων φιλτραρίσματος γνωρισμάτων υπάρχει πάντα ο κίνδυνος επιλογής γνωρισμάτων–γονιδίων που είναι πολύ συσχετισμένα μεταξύ τους [9]. Αυτό συχνά συμβαίνει γιατί γονίδια που ανήκουν σε κοινά (βιολογικά) μονοπάτια παρουσιάζουν όμοια γονιδιακή ρύθμιση και επομένως έχουν όμοια προφίλ έκφρασης. Έτσι, λαμβάνουν τα υψηλότερα σκορ [5]. Παρόλα αυτά, η χρησιμοποίηση υψηλά συσχετισμένων γονιδίων δεν

εγγυάται ότι η ομαδική τους δράση θα δώσει τα καλύτερα αποτελέσματα ταξινόμησης. Με απλά λόγια, δημιουργείται μία περίσσεια πληροφορίας (information redundancy).

- Επίσης, εάν υπάρχει κάποιο όριο στον αριθμό των επιλεγμένων γονιδίων, υπάρχει πιθανότητα επιλογής γονιδίων μόνο από το μονοπάτι που έχει την κύρια επίδραση [5]. Έτσι, συμπληρωματικά γονίδια που καθορίζουν άλλες ετικέτες κλάσης (μονοπάτια) μπορεί να μην επιλεχθούν επειδή λαμβάνουν χαμηλά σκορ [9].
- Τέλος, η επιλογή του αριθμού των επιλεγμένων γονιδίων κρύβει ανθρώπινη διαίσθηση. Έτσι, είναι γενικά δύσκολο να εντοπιστεί ο βέλτιστος αριθμός επιλεγμένων γονιδίων που θα μας δώσει και τα καλύτερα αποτελέσματα.

Με αυτό τον τρόπο είναι πιθανό να αποκλειστούν γονίδια που συμμετέχουν σε βιολογικά μονοπάτια τα οποία συνεισφέρουν στη βαθύτερη κατανόηση του προβλήματος. Το SoFoCles επιδιώκει ακριβώς αυτό, χρησιμοποιώντας δηλαδή υπάρχουσα βιολογική γνώση να επιτρέψει την επαναεισαγωγή γονιδίων που συμμετέχουν σε κρίσιμα βιολογικά μονοπάτια με απώτερο σκοπό την βελτίωση της ακρίβειας ταξινόμησης. Τον τρόπο με τον οποίο το επιτυγχάνει αυτό θα τον περιγράψουμε διεξοδικά παρακάτω.

5.2 Περιβάλλον ανάπτυξης του SoFoCles

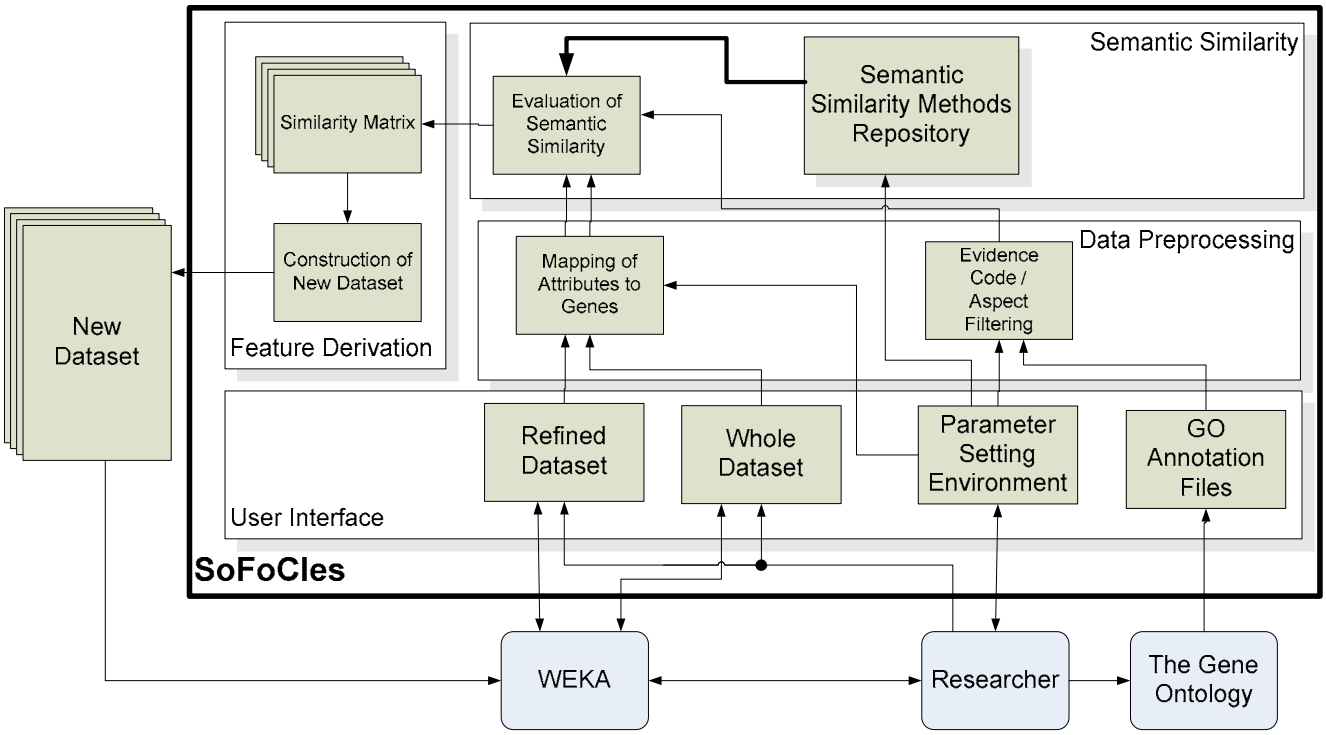
Κατ' αρχήν, το περιβάλλον ανάπτυξης του SoFoCles είναι η Java. Επιλέχθηκε η Java γιατί είναι μεταφέρσιμη (portable) τόσο σε επίπεδο της αριθμητικής των δομών δεδομένων όσο και σε επίπεδο του μεταγλωττισμένου κώδικα.

Πιο συγκεκριμένα, η Java σε αντίθεση με άλλες γλώσσες (όπως η C και η C++) ορίζει το μέγεθος των βασικών τύπων δεδομένων για όλες τις υλοποιήσεις. Έτσι, για παράδειγμα ένας ακέραιος έχει το ίδιο μέγεθος για όλα τα συστήματα. Επίσης, είναι μεταφέρσιμη όσον αφορά το μεταγλωττισμένο κώδικα. Αναλυτικότερα, ενώ οι περισσότεροι μεταγλωττιστές (compilers) γλωσσών προγραμματισμού παράγουν κώδικα που τρέχει βέλτιστα για ένα συγκεκριμένο σύστημα, ο μεταγλωττιστής (compiler) της Java μεταγλωττίζει τον κώδικα σε μία θεωρητική μηχανή· ο διερμηνευτής (interpreter) της Java εξομοιώνει αυτή τη μηχανή. Έτσι, ο κώδικας Java που δημιουργείται μπορεί να τρέχει σε οποιοδήποτε σύστημα με την ύπαρξη ενός μεταφραστή Java. Ωστόσο, αξίζει να σημειώσουμε ότι η παρουσία του διερμηνευτή

προσθέτει ένα επιπλέον χρόνο (overhead) κατά τη διαδικασία μετάφρασης κάτι που την κάνει βραδύτερη κατά 1.5 με 3 φορές σε σχέση με άλλες γλώσσες όπως η C και η C++.

5.3 Δομή του SoFoCles

Σε αυτό το σημείο θα παρουσιάσουμε τη δομή του SoFoCles. Με άλλα λόγια, θα αναφερθούμε στις βασικές δομές του εργαλείου και στις μονάδες περιβάλλοντος με τις οποίες επικοινωνεί. Το σχήμα του SoFoCles παρουσιάζεται στην Εικ. 5.1.



Εικόνα 5.1. Σχήμα του SoFoCles

Αρχικά από το περιβάλλον λογισμικού WEKA, εξάγεται το φιλτραρισμένο σύνολο δεδομένων (“Refined Dataset”) από το αρχικό (“Whole Dataset”). Όπως θα δούμε και στο υποκεφάλαιο 5.5, το αρχικό σύνολο θα πρέπει να βρίσκεται σε μορφή .arff για να είναι επεξεργάσιμο από το περιβάλλον λογισμικού WEKA.

Στη συνέχεια, μέσω της δομής “Διεπαφή Χρήστη” (“User Interface”) δίνεται η δυνατότητα στον χρήστη-ερευνητή (Researcher) της φόρτωσης των παραπάνω δύο αρχείων.

Επίσης, από το περιβάλλον της Γονιδιακής Οντολογίας (The Gene Ontology) παρέχεται η δυνατότητα της φόρτωσης των GOA αρχείων καθώς και του αρχείου της Γονιδιακής Οντολογίας.

Επιπλέον, η δομή "Διεπαφή Χρήστη" ("User Interface") παρέχει το περιβάλλον εισαγωγής παραμέτρων ("Parameter Setting Environment") που επικοινωνεί με άλλες δύο δομές του SoFoCles: την "Προεπεξεργασία Δεδομένων" ("Data Preprocessing") και τη "Σημασιολογική Ομοιότητα" ("Semantic Similarity"). Πιο συγκεκριμένα, με το περιβάλλον εισαγωγής παραμέτρων φορτώνονται τα αρχεία αντιστοίχησης γνωρισμάτων σε γονιδιακά σύμβολα, ids, ονόματα ή συνώνυμα ("Mapping of Attributes to Genes"), ενώ καθορίζονται ποιές οντολογίες και ποιοί τύποι αποδείξεων θα ληφθούν υπόψη κατά την υποβολή του ερωτήματος ("Evidence Code / Aspect Filtering"). Επίσης, με το περιβάλλον εισαγωγής παραμέτρων επιλέγεται και η μέθοδος σημασιολογικής ομοιότητας που θα χρησιμοποιηθεί.

Στη δομή "Προεπεξεργασία Δεδομένων" ("Data Preprocessing") εντοπίζονται τα γονιδιακά σύμβολα (ids, ονόματα ή συνώνυμα) που αντιστοιχούν στο αρχικό και στο φιλτραρισμένο σύνολο δεδομένων μέσω των αρχείων αντιστοίχησης.

Κατόπιν, μεταβαίνουμε στη δομή "Σημασιολογική Ομοιότητα" ("Semantic Similarity") όπου υπολογίζεται η σημασιολογική ομοιότητα μεταξύ των γονιδίων του αρχικού και του φιλτραρισμένου συνόλου δεδομένων βάσει της μεθόδου σημασιολογικής ομοιότητας ("Semantic Similarity Methods Repository") που έχει επιλεγεί.

Δημιουργείται, έτσι, ο πίνακας σημασιολογικής ομοιότητας ("Similarity Matrix") της δομής "Άντληση Γνωρισμάτων" ("Feature Derivation"). Μέσω αυτού του πίνακα εντοπίζονται τα πιο σημασιολογικώς όμοια γονίδια με αυτά του φιλτραρισμένου συνόλου και δημιουργείται το νέο σύνολο δεδομένων ("Construction of New Dataset") που θα περιέχει τόσο τα γονίδια–δείκτες που έχουν προκύψει από την εφαρμογή μεθόδων φιλτραρίσματος γνωρισμάτων όσο και γονίδια που είναι σημασιολογικώς συσχετισμένα με αυτά και έχουν προκύψει από τη χρησιμοποίηση μεθόδων σημασιολογικής ομοιότητας.

Τέλος, το σύνολο που παράχθηκε από το SoFoCles, όπως το αρχικό και το φιλτραρισμένο σύνολο, υποβάλλονται σε διαδικασία εκμάθησης μέσω της εφαρμογής κατάλληλου αλγορίθμου από το περιβάλλον WEKA. Αξίζει να σημειωθεί, ότι για τους σκοπούς της διπλωματικής, ως αλγόριθμος εκμάθησης επιλέχθηκε αυτός των Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines, SVMs).

5.4 Συνοπτική περιγραφή των λειτουργιών του SoFoCles

Οι λειτουργίες που προσφέρει το SoFoCles είναι οι παρακάτω:

5.4.1 Εισαγωγή δεδομένων

Το SoFoCles δίνει τη δυνατότητα εισαγωγής δεδομένων διαφορετικής φύσης. Όπως φαίνεται στην Εικ. 5.2, αυτά αφορούν τα σύνολα μικροσυστοιχιών (το αρχικό και αυτό που προκύπτει μετά το φιλτράρισμα γνωρισμάτων), τα αρχεία αντιστοιχήσεων γνωρισμάτων σε ετικέτες που είναι επεξεργάσιμες από τα GOA αρχεία (π.χ. γονιδιακά σύμβολα, γονιδιακά ονόματα), τα GOA αρχεία και το αρχείο της Γονιδιακής Οντολογίας.

Το αξιοσημείωτο είναι ότι τα δεδομένα μπορούν να εισάγονται σε μορφή αρχείων κειμένων (.txt files) ενισχύοντας έτσι την ευχρηστία του εργαλείου αφού δεν απαιτείται η εγκατάσταση επιπλέον προγραμμάτων για την εκτέλεση του. Επιπλέον, η παροχή πολλών από τα αρχεία εισόδου διατίθεται σε .txt μορφή από τον επίσημο φορέα της Γονιδιακής Οντολογίας: The Gene Ontology (http://www.geneontology.org).

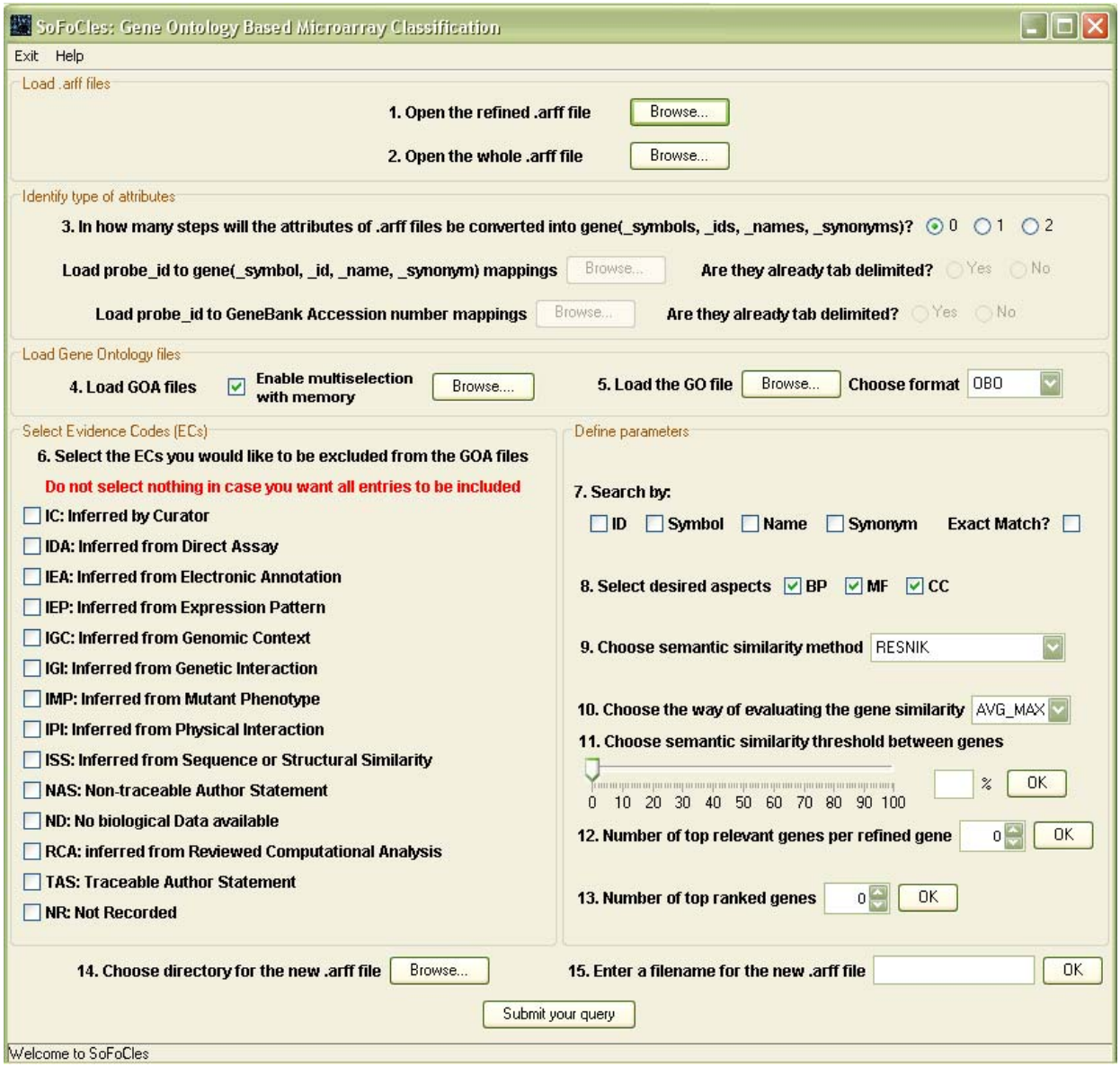
5.4.2 Εισαγωγή παραμέτρων

Παρέχεται επίσης η δυνατότητα εισαγωγής πολλών παραμέτρων για την εξέταση διαφορετικών σεναρίων.

Ο χρήστης μπορεί να επιλέξει τους τύπους των αποδείξεων που επιθυμεί να αποκλείσει από την αναζήτηση για να καθορίσει έτσι το επίπεδο αξιοπιστίας των αποτελεσμάτων. Επίσης, μπορεί να επιλέξει τα πεδία όπου θα επικεντρωθεί η αναζήτηση. Για παράδειγμα, αν θα αναζητηθούν γονίδια βάσει της ταυτότητας τους (id), του συμβόλου τους (symbol), του ονόματος τους (name) ή του συνωνύμου τους (synonym). Γίνεται επίσης η επιλογή πόσο αυστηρή θα είναι η αναζήτηση (αν το ταίριασμα θα είναι ακριβές ή μερικό).

Ακόμη, ο χρήστης μπορεί να αποκλείσει από την αναζήτηση τους όρους οντολογιών–απόψεων που δεν θεωρεί σημαντικούς για το πρόβλημά του. Επιπλέον, δίνεται η δυνατότητα επιλογής της μεθόδου σημασιολογικής ομοιότητας μεταξύ των όρων δύο γονιδίων καθώς και του τρόπου υπολογισμού της ομοιότητας μεταξύ των γονιδίων από τις παραπάνω ομοιότητες. Τέλος, επιλέγεται ένα κατώφλι ομοιότητας

πάνω από το οποίο προκρίνονται όλα τα γονίδια που είναι συσχετισμένα με τα γονίδια που είχαν επιλεγεί κατά το φιλτράρισμα γνωρισμάτων, ενώ εισάγεται και ο αριθμός των γνωρισμάτων που επιθυμείται να κρατηθεί τελικά.



Εικόνα 5.2. Οθόνη εισαγωγής δεδομένων του SoFoCles

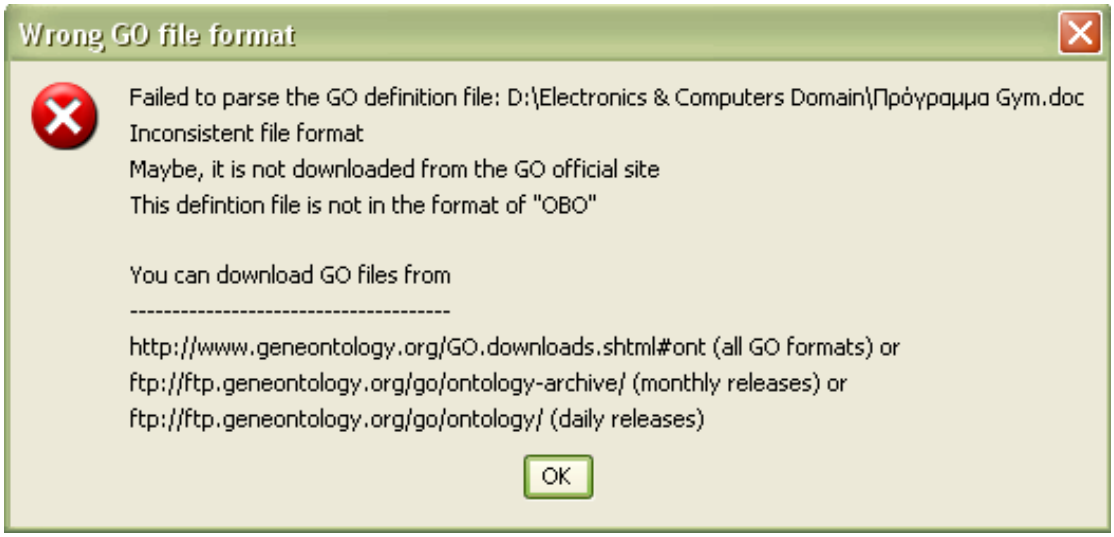
5.4.3 Παρουσίαση μηνυμάτων πληροφορίας

Κατά την εκτέλεση του προγράμματος παράγονται κάποια μηνύματα που καθοδηγούν το χρήστη για τη σωστή εκτέλεσή του.

Για παράδειγμα, όπως φαίνεται στην Εικ. 5.3, ο χρήστης ενημερώνεται ποιες διαδικασίες εκκρεμούν ώστε να είναι δυνατή η επεξεργασία του ερωτήματος από το πρόγραμμα.

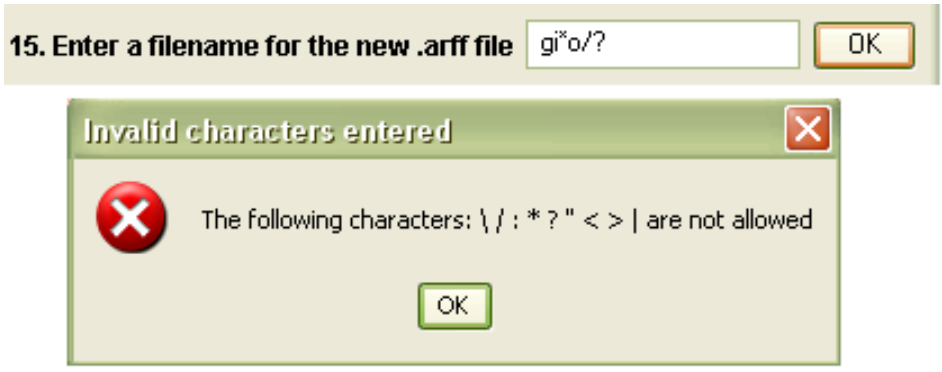
Εικόνα 5.3. Οθόνη ενημέρωσης του χρήστη για ατελές ερώτημα

Επίσης, εάν ο χρήστης φορτώσει αρχείο Γονιδιακής Οντολογίας (GO file) που δεν βρίσκεται σε ένα από τα αποδεκτά formats, εμφανίζεται το αντίστοιχο μήνυμα. Ένα τέτοιο παράδειγμα εμφανίζεται στην Εικ. 5.4.



Εικόνα 5.4. Οθόνη ενημέρωσης του χρήστη για εισαγωγή αρχείου Γονιδιακής Οντολογίας μη αποδεκτού format

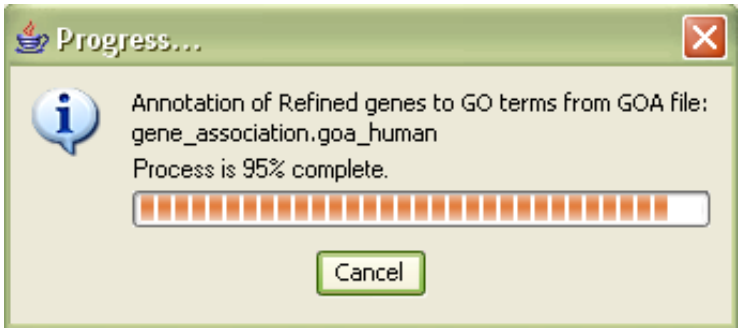
Επιπλέον, λαμβάνεται φροντίδα για την αποφυγή δημιουργίας εξαιρέσεων από μη έγκυρη ονομασία αρχείων. Για παράδειγμα, στην Εικ. 5.5 αποτρέπεται η ονομασία gi*o/? επειδή περιέχει έναν ή περισσότερους από τους μη αποδεκτούς χαρακτήρες \ / : * ? “ < > |



Εικόνα 5.5. Οθόνη ενημέρωσης του χρήστη για εισαγωγή μη αποδεκτών χαρακτήρων κατά την ονομασία αρχείων

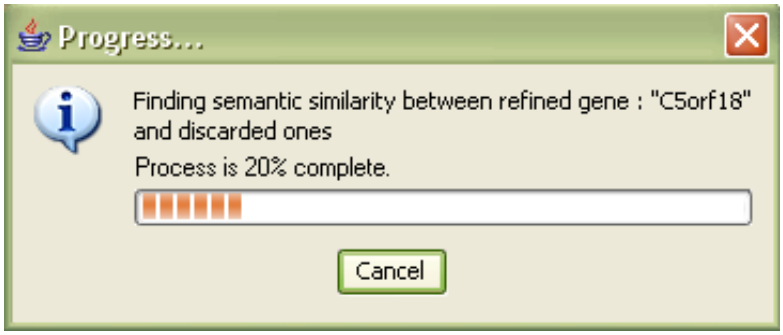
Πέρα όμως των μηνυμάτων λάθους εμφανίζονται και μηνύματα πληροφορίας που ενημερώνουν το χρήστη για το στάδιο της επεξεργασίας καθώς και για τη συνολική εικόνα που δημιουργείται από την υποβολή ενός συγκεκριμένου ερωτήματος.

Για παράδειγμα, όταν γίνεται προσπάθεια αντιστοίχησης των γονιδίων των συνόλων μικροσυστοιχιών στις λειτουργίες που αυτά έχουν (όροι Γονιδιακής Οντολογίας) μέσω των GOA αρχείων εμφανίζεται παράθυρο προόδου το οποίο ενημερώνει για την πορεία της διεργασίας. Ένα τέτοιο παράδειγμα παρουσιάζεται στην Εικ. 5.6.



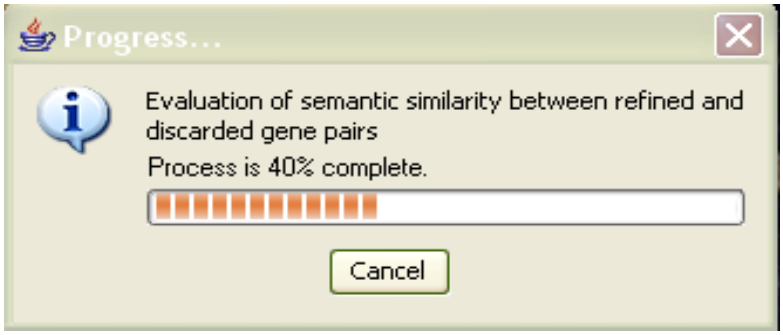
Εικόνα 5.6. Οθόνη ενημέρωσης του χρήστη για την πρόοδο αντιστοίχησης γονιδίων σε GO όρους

Το ίδιο συμβαίνει όταν επιχειρείται ο υπολογισμός της σημασιολογικής ομοιότητας μεταξύ των γονιδίων. Για παράδειγμα, στην Εικ. 5.7, παρουσιάζεται η πρόοδος που σημειώνει η διαδικασία υπολογισμού της σημασιολογικής ομοιότητας μεταξύ του γονιδίου “C5orf18” που είχε προκριθεί κατά το φιλτράρισμα γνωρισμάτων και αυτών που είχαν αποκλειστεί (από το φιλτράρισμα γνωρισμάτων).

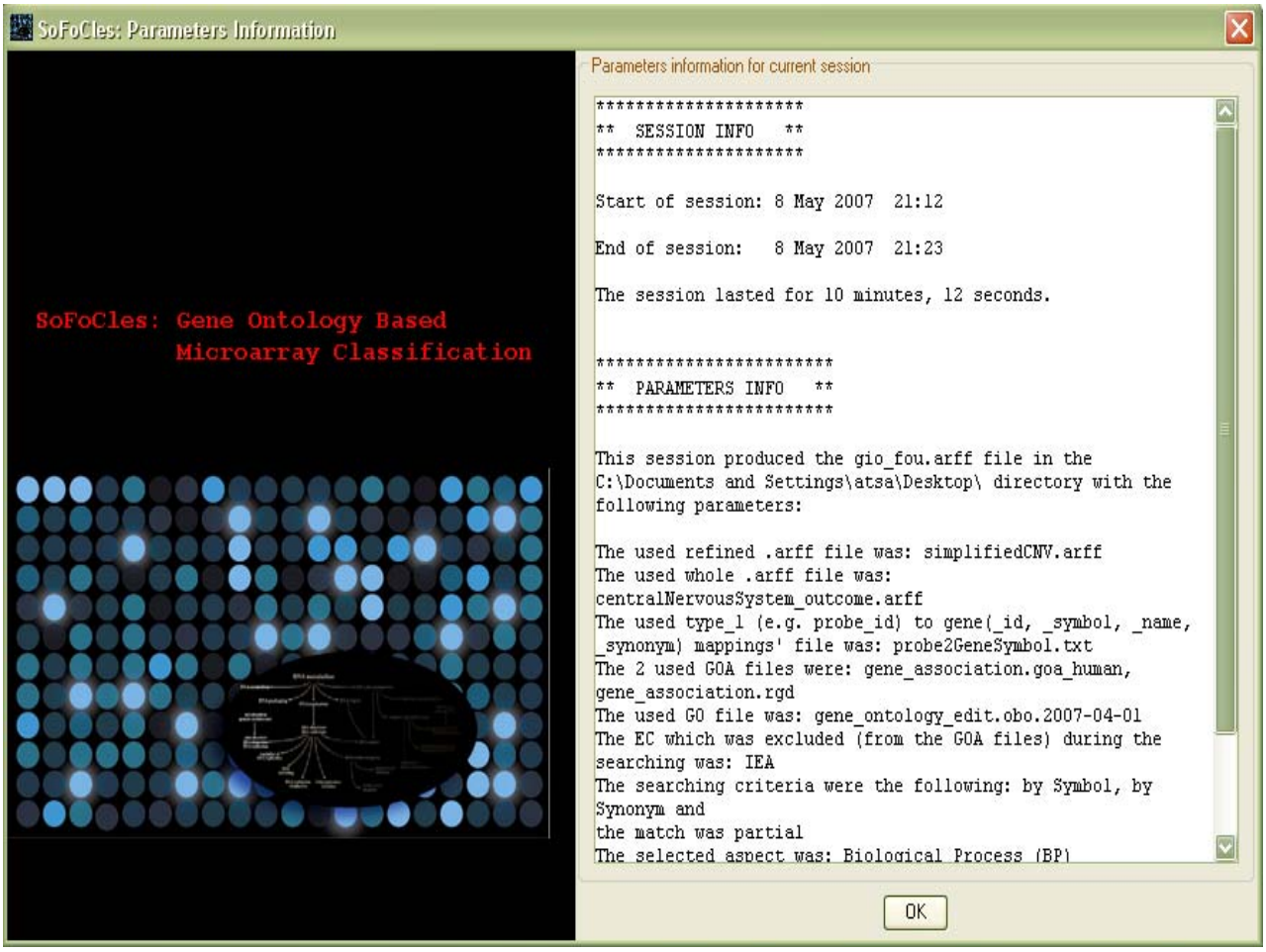


Εικόνα 5.7. Οθόνη ενημέρωσης του χρήστη για την πρόοδο υπολογισμού της σημασιολογικής ομοιότητας μεταξύ γονιδίου που προκρίθηκε κατά το φιλτράρισμα γνωρισμάτων και αυτών που αποκλείστηκαν

Ενώ, στην Εικ. 5.8 παρουσιάζεται η συνολική πρόοδος του υπολογισμού της σημασιολογικής ομοιότητας μεταξύ των γονιδίων που προκρίθηκαν κατά το φιλτράρισμα γνωρισμάτων και αυτών που αποκλείστηκαν.



Εικόνα 5.8. Οθόνη ενημέρωσης του χρήστη για τη συνολική πρόοδο υπολογισμού της σημασιολογικής ομοιότητας μεταξύ των γονιδίων που προκρίθηκαν κατά το φιλτράρισμα γνωρισμάτων και αυτών που αποκλείστηκαν

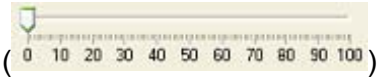


Εικόνα 5.9. Οθόνη ενημέρωσης του χρήστη για τη συνολική εικόνα που δημιουργείται μετά το πέρας του ερωτήματος

Τέλος, στην Εικ. 5.9 παρουσιάζεται η συνολική εικόνα που δημιουργείται από την υποβολή ενός συγκεκριμένου ερωτήματος. Με άλλα λόγια, παρουσιάζονται πληροφορίες για τη διάρκεια της εκτέλεσης του ερωτήματος καθώς και για τις παραμέτρους που χρησιμοποιήθηκαν.

5.4.4 Παραγωγή αποτελεσμάτων

Τα αποτελέσματα που παράγονται σε μορφή αρχείου .arff περιλαμβάνουν τόσο την πληροφορία από τα γονίδια–δείκτες όσο και από τα γονίδια που είναι υψηλά συσχετισμένα με αυτά (μετά την εφαρμογή των μεθόδων σημασιολογικής ομοιότητας). Έτσι, επιχειρείται η ενσωμάτωση υπάρχουσας βιολογικής γνώσης στο σύνολο μικροσυστοιχιών με σκοπό τη βελτίωση της ακρίβειας ταξινόμησης.

Εξάλλου, άλλα στοιχεία που εξυπηρετούν τη λειτουργικότητα του εργαλείου είναι φίλτρα αρχείων για την αποφυγή φόρτωσης αρχείων μη αποδεκτών μορφών και Java στοιχεία όπως sliders (

) και spinners (

) που αποτρέπουν την εισαγωγή μη αριθμητικών τιμών ή αριθμητικών τιμών εκτός των επιτρεπτών ορίων.

5.5 Προετοιμασία δεδομένων εισόδου

Στο σημείο αυτό θα μιλήσουμε για τη μορφή των δεδομένων εισόδου και τρόπων προ-επεξεργασίας τους.

5.5.1 Arff αρχεία

Κατ’ αρχήν, το αρχικό σύνολο μικροσυστοιχιών (whole.arff) θα πρέπει να βρίσκεται σε .arff format.

Κάθε αρχείο .arff αποτελείται από δύο τμήματα. Το τμήμα της Επικεφαλίδας (Header) και το τμήμα των Δεδομένων (Data).

Το τμήμα της Επικεφαλίδας περιέχει το όνομα της σχέσης (@relation), τα γνωρίσματα (@attribute) που εν προκειμένω αποτελούν γονίδια και τους τύπους

τους. Περιλαμβάνει, τέλος, την κλάση του προβλήματος (class) σε περίπτωση που είναι γνωστή.

Από την άλλη, το τμήμα των Δεδομένων περιλαμβάνει τις τιμές των γνωρισμάτων και της κλάσης για κάθε παράδειγμα. Στην περίπτωση μας, παράδειγμα μπορεί να είναι ένας ασθενής, ένας ιστός, ένα κύτταρο ή ένα δείγμα σε διαφορετικές χρονικές στιγμές.

Μία τυπική μορφή .arff format είναι η παρακάτω:

```
% 1. Title: Smoking Dataset
%
% 2. Sources:
%

@relation smoking dataset

@attribute AB011537   numeric
@attribute AB020630   numeric
@attribute NM_000691   numeric
@attribute NM_022003   numeric
@attribute W87466   numeric
@attribute class   {current-smoker, never-smoked, former-smoker}

@data
128.9, 9.3, 10.2, 725.3, 57.9, current-smoker
102.8, 16.9, 13.2, 1295.7, 62.9, current-smoker
88.7, 16.5, 15.9, 1326.2, 83.5, current-smoker
176.9, 6.3, 7.8, 1477.5, 21.1, never-smoked
155.6, 12.4, 62.6, 1898.8, 130, never-smoked
125, 14.8, 25.9, 2293.1, 201, never-smoked
209.2, 7.3, 7.6, 1217, 96.1, former-smoker
233.6, 9.6, 11.9, 1536.6, 61.7, former-smoker
97.6, 7.9, 6.2, 865, 54.1, former-smoker
```

Από το αρχικό σύνολο μικροσυστοιχιών παίρνουμε ένα μικρότερο σύνολο (refined.arff) μετά την εφαρμογή μεθόδου φιλτραρίσματος γνωρισμάτων. Υπάρχουν διάφορα πακέτα εξόρυξης δεδομένων σε Java που περιέχουν μεθόδους επιλογής γνωρισμάτων. Ένα ευρέως γνωστό πακέτο σε Java, που χρησιμοποιήθηκε σε αυτή τη διπλωματική εργασία, είναι το Weka που αναπτύχθηκε από το Πανεπιστήμιο του Waikato. Οι μέθοδοι επιλογής γνωρισμάτων που χρησιμοποιήθηκαν είναι οι Information Gain, Relief και Chi-2, οι οποίες εξετάστηκαν διεξοδικά στο κεφάλαιο 4.

Τα γνωρίσματα των συνόλων δεδομένων μπορεί να είναι απευθείας αναγνωρίσιμα από τα GOA αρχεία ή μπορεί να απαιτούν μία επιπρόσθετη διαδικασία μετατροπής τους σε μορφή επεξεργάσιμη από τα προαναφερθέντα αρχεία.

Αναλυτικότερα, τα GOA αρχεία αναγνωρίζουν το id ενός γονιδίου, το σύμβολό του, το όνομά του ή κάποιο συνώνυμό του (π.χ. δείκτης IPI). Στην περίπτωση που τα γνωρίσματα δεν βρίσκονται σε μορφή απευθείας επεξεργάσιμη από τα GOA αρχεία

πρέπει να εισαχθούν αρχεία αντιστοίχησης (mappings) γνωρισμάτων σε γονιδιακά σύμβολα, ονόματα, ids ή συνώνυμα.

5.5.2 Αρχεία αντιστοίχησης (mappings)

Γι' αυτό το σκοπό χρησιμοποιήθηκε το εργαλείο R που περιέχει έτοιμες βιβλιοθήκες μετατροπής από τον ένα τύπο γονιδίου στον άλλο. Τα αρχεία (mappings) που παράγονται με αυτό τον τρόπο έχουν την ακόλουθη μορφή:

```
"X222152_at"
"1"  "PDCD6"

"X204659_s_at"
"1"  "GFER"

"X210893_at"
"1"  NA

"X201871_s_at"
"1"  "LOC51035"

"X210132_at"
"1"  "EFNA3"
```

Παρόλα αυτά δίνεται η δυνατότητα χρησιμοποίησης ήδη μορφοποιημένων (οροθετημένων με στηλοθέτη) ζευγών γονιδίων από τον ένα τύπο στον άλλο:

```
X222152_at      PDCD6
X204659_s_at    GFER
X210893_at      NA
X201871_s_at    LOC51035
X210132_at      EFNA3
```

5.5.3 GOA αρχεία

Έχουμε αναφερθεί προηγουμένως στα GOA αρχεία. Κρίνεται αναγκαίο η περιγραφή μίας τέτοιας καταχώρησης για μία ολοκληρωμένη εικόνα του τρόπου χρησιμοποίησης ενός τέτοιου αρχείου.

Ένα GOA αρχείο αποτελείται από 15 πεδία. Αυτά παρουσιάζονται κατά σειρά στον Πιν. Χ:

Πίνακας Χ. Πεδία ενός GOA αρχείου.

Πεδίο	Περιγραφή	Σημαντικότητα
DB	Η βάση δεδομένων που παρείχε αυτή την επισημείωση	υποχρεωτικό
DB_Object_ID	Μοναδικός ταυτοποιητής (identifier) του αντικειμένου που επισημειώνεται	υποχρεωτικό
DB_Object_Symbol	Μοναδικό και έγκυρο σύμβολο με βιολογικό νόημα που αντιστοιχεί με μοναδικό τρόπο στο DB_Object_ID	υποχρεωτικό
Qualifier	Σημαίες που μεταβάλλουν τη σημασία μίας επισημείωσης	μη υποχρεωτικό
GO ID	Ο ταυτοποιητής (identifier) του GO όρου που αποδίδεται στο συγκεκριμένο DB_Object_ID	υποχρεωτικό
DB:Reference	Ένας ή περισσότεροι ταυτοποιητές που χρησιμοποιούνται ως αναφορά στην υπεύθυνη αρχή για την απόδοση του GO όρου στο DB_Object_ID	υποχρεωτικό
Evidence	Τύπος Απόδειξης. Ένας από τους: (IEA, ISS, IEP, IMP, IGC, IGI, IPI, IDA, RCA, TAS, NAS, IC, ND)	υποχρεωτικό
With (or) From	Ένα από τα: (DB:gene_symbol, DB:gene_symbol[allele_symbol], DB:gene_id, DB:protein_name, DB:sequence_id, GO:GO_id)	μη υποχρεωτικό
Aspect	Ένα από τα: (P (biological process), F (molecular function) ή C (cellular component))	υποχρεωτικό
DB_Object_Name	Όνομα του γονιδίου ή του γονιδιακού προϊόντος	μη υποχρεωτικό
Synonym	Γονιδιακό σύμβολο (ή άλλο κείμενο όπως δείκτης IPI). Μία κάθετος χρησιμοποιείται για το διαχωρισμό μεταξύ των διαφορετικών συνωνύμων	μη υποχρεωτικό

DB_Object_Type	Ο τύπος του αντικειμένου που επισημειώνεται (σε κάποιο GO όρο). Ένας από τους: (gene, transcript, protein, protein_structure, complex)	υποχρεωτικό
Taxon	Ταξινομικοί ταυτοποιητές	υποχρεωτικό
Date	Ημερομηνία στην οποία πραγματοποιήθηκε η επισημείωση. Η μορφή είναι YYYYMMDD	υποχρεωτικό
Assigned_by	Η βάση δεδομένων που έκανε την επισημείωση	υποχρεωτικό

Μία ενδεικτική καταχώρηση ενός τέτοιου αρχείου είναι η ακόλουθη:

```
UniProt Q99519 NEUR1_HUMAN GO:0005975 GOA:spkw IEA P
Sialidase-1 precursor IPI00029817 protein taxon:9606 20060223 UniProt
```

Η παραπάνω ένδειξη μας πληροφορεί ότι η πρωτεΐνη Sialidase–1 precursor (NEUR1_HUMAN ή IPI00029817) επισημειώθηκε στον GO όρο GO:0005975 (carbohydrate metabolic process), ο οποίος ανήκει στην BP οντολογία, στις 23 Ιουνίου 2003 από τη βάση δεδομένων UniProt.

5.5.4 Αρχείο Γονιδιακής Οντολογίας

Το αρχείο της Γονιδιακής Οντολογίας περιέχει πληροφορίες για τη δομημένη μορφή της Γονιδιακής Οντολογίας. Κάθε όρος εσωκλείεται σε αγκύλες: [Term] και είναι γνωστός ως stanza. Ο όρος περιγράφεται από τις επικεφαλίδες (tags). Οι υποχρεωτικές επικεφαλίδες είναι το **id** και το όνομα (**name**) του όρου. Λαμβάνουν μοναδικές τιμές για κάθε συγκεκριμένο όρο.

Μερικές από τις πιο ευρέως χρησιμοποιούμενες προαιρετικές επικεφαλίδες ([xxiii]) είναι οι ακόλουθες:

alt_id – Ορίζει ένα εναλλακτικό id για τον όρο. Ένας όρος μπορεί να έχει ένα οποιοδήποτε αριθμό εναλλακτικών ids. Για παράδειγμα, το εναλλακτικό id του όρου GO:0006214, thymidine catabolic process, που ανήκει στην BP οντολογία, είναι το GO:0006215.

def – Παρέχει τον ορισμό του όρου. Για παράδειγμα, ο ορισμός του όρου της μεταβολικής διαδικασίας (metabolic process) είναι ο εξής: ”*Διαδικασίες που προκαλούν χημικές αλλαγές στους ζώντες οργανισμούς συμπεριλαμβανομένου της διαδικασίας του αναβολισμού και του καταβολισμού. Οι μεταβολικές διαδικασίες συνήθως μετατρέπουν μικρά μόρια, αλλά είναι επίσης υπεύθυνες και για μακρομοριακές διαδικασίες όπως η επιδιόρθωση και η αντιγραφή του DNA και η πρωτεϊνοσύνθεση*”.

comment – Παρέχει ένα σχόλιο για το συγκεκριμένο όρο.

subset – Καταδεικνύει το υποσύνολο στο οποίο ανήκει ο όρος. Στην τρέχουσα έκδοση της Γονιδιακής Οντολογίας, τα υπάρχοντα υποσύνολα είναι: goslim_generic "Generic GO slim", goslim_goa "GOA and proteome slim", goslim_plant "Plant GO slim", goslim_yeast "Yeast GO slim", gosubset_prok "Prokaryotic GO subset". Για παράδειγμα, ο όρος αναπαραγωγή (GO:0000003, reproduction) ανήκει στα υποσύνολα: goslim_generic, goslim_plant και gosubset_prok.

synonym – Παρέχει συνώνυμα του όρου. Κάθε συνώνυμο μπορεί να λάβει έναν από τους 4 χαρακτηρισμούς: EXACT (ΑΚΡΙΒΕΣ), BROAD (ΓΕΝΙΚΟΤΕΡΟ), NARROW (ΕΙΔΙΚΟΤΕΡΟ) και RELATED (ΣΧΕΤΙΚΟ). Για παράδειγμα, το συνώνυμο του όρου GO:0043154, negative regulation of caspase activity είναι το down regulation of caspase activity και είναι ακριβές (exact).

is_a – Αυτή η επικεφαλίδα περιγράφει μία σχέση υποκλάσης μεταξύ ενός όρου και κάποιου άλλου. Η τιμή που παίρνει αυτή η επικεφαλίδα είναι το id του όρου στο οποίο ο συγκεκριμένος όρος αποτελεί υποκλάση. Μ’ άλλα λόγια, ως is_a χαρακτηρίζονται όλοι οι γονείς ενός όρου που συνδέονται με αυτόν με σχέση is_a. Ένας όρος μπορεί να έχει ένα οποιοδήποτε αριθμό is_a σχέσεων. Για παράδειγμα, η is_a επικεφαλίδα του όρου GO:0003676, nucleic acid binding είναι ο όρος GO:0005488, binding.

relationship – Περιγράφει μία σχέση μεταξύ δύο όρων. Για παράδειγμα, ο

όρος πυρηνικό περίβλημα (GO:0005635, nuclear envelope) είναι μέρος του (part_of) του ενδομεμβρανικού συστήματος (GO:0012505, endomembrane system).

is_obsolete – Πληροφορεί αν ένας όρος είναι ξεπερασμένος (outdated). Για παράδειγμα, ο όρος χρωματίδη (chromatid) είναι ξεπερασμένος (obsolete) επειδή δεν αποτελεί μία μοναδική υποκυτταρική δομή.

namespace – Πληροφορεί για την οντολογία–άποψη που ανήκει ο συγκεκριμένος όρος

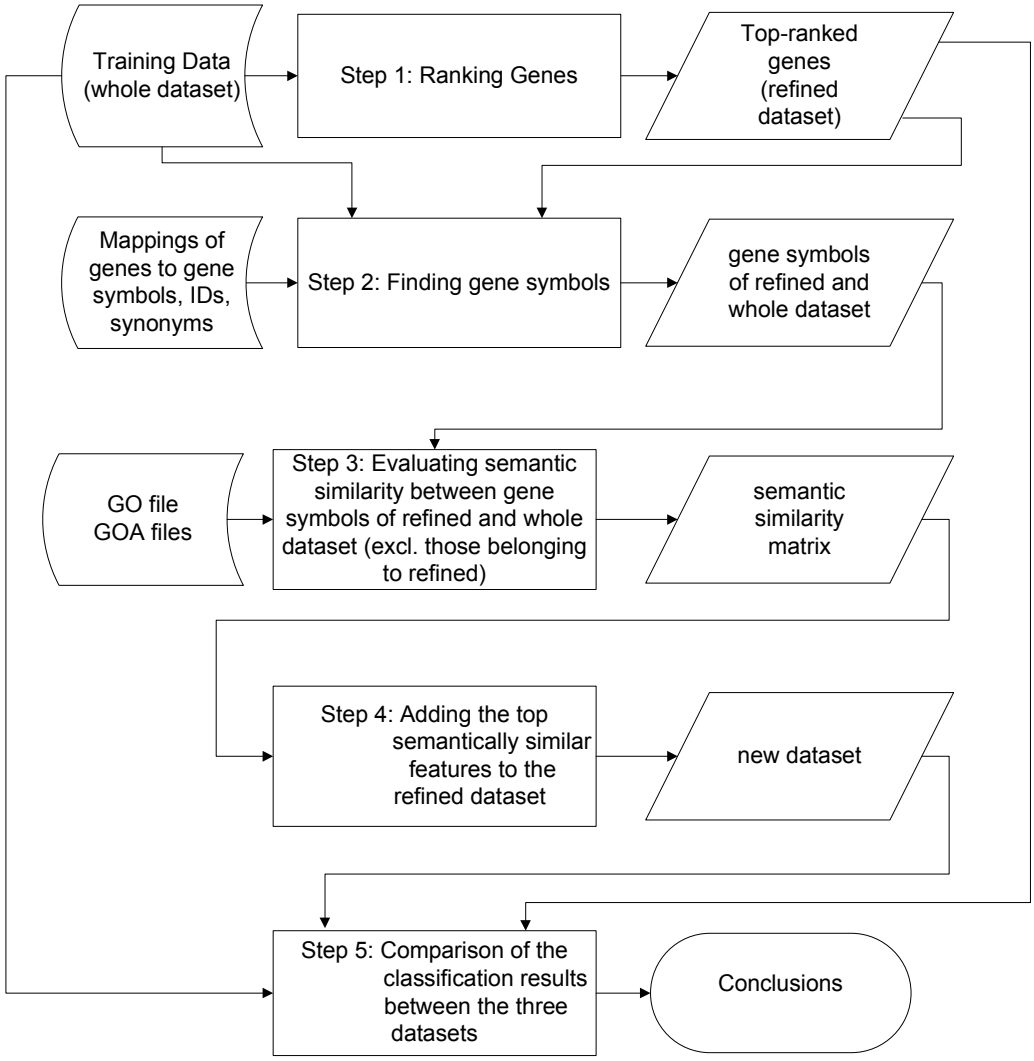
Μία ενδεικτική καταχώρηση ενός τέτοιου αρχείου είναι η ακόλουθη:

```
[Term]
id: GO:0000003
name: reproduction
namespace: biological_process
alt_id: GO:0019952
alt_id: GO:0050876
def: "The production by an organism of new individuals that contain some
     portion of their genetic material inherited from that organism."
     [GOC:go_curators, GOC:isa_complete, ISBN:0198506732 "Oxford
     Dictionary of Biochemistry and Molecular Biology"]
subset: goslim_generic
subset: goslim_plant
subset: gosubset_prok
synonym: "reproductive physiological process" EXACT []
is_a: GO:0008150 ! biological_process
```

Η παραπάνω ένδειξη μας πληροφορεί ότι ο GO όρος *αναπαραγωγή* (reproduction) έχει id το GO:0000003, εναλλακτικά ids τα GO:0019952 και GO:0050876, είναι γνωστός επίσης και με την ονομασία *αναπαραγωγική φυσιολογική διαδικασία* (*reproductive physiological* process), ανήκει στην BP οντολογία και στα υποσύνολα goslim_generic, goslim_plant και gosubset_prok, ενώ, τέλος, συνδέεται με is_a σχέση με τον GO όρο biological_process (GO:0008150).

5.6 Ο βασικός αλγόριθμος του SoFoCles

Σε αυτό το υποκεφάλαιο παρουσιάζεται ο βασικός αλγόριθμος του SoFoCles που φαίνεται στην Εικ. 5.10.



Εικόνα 5.10. Διάγραμμα ροής (flow chart) του SoFoCles

Ο αλγόριθμος περιλαμβάνει πέντε βήματα.

Στο πρώτο βήμα, γίνεται η κατάταξη (ranking) των γονιδίων βάσει της διαχωριστικής τους ικανότητας. Με άλλα λόγια, σε αυτό το βήμα πραγματοποιείται το φιλτράρισμα γνωρισμάτων με τη βοήθεια των μεθόδων που περιγράψαμε προηγουμένως (Chi-2, IG, Relief-F). Σε αυτό το στάδιο προκρίνονται τα γονίδια που βάσει των τιμών τους κατά μήκος όλων των παραδειγμάτων του συνόλου δεδομένων εμφανίζονται να έχουν περισσότερη πληροφορία για το συγκεκριμένο πρόβλημα. Στην έξοδο αυτού του βήματος παράγεται ένα νέο σύνολο δεδομένων, υποσύνολο

του αρχικού, που περιέχει τα πιο πληροφοριακά γονίδια, τα γνωστά και ως γονίδια – δείκτες (marker genes).

Στο δεύτερο βήμα, εντοπίζονται τα γονιδιακά σύμβολα, IDs, ονόματα ή συνώνυμα των γονιδίων που αποτελούν τα γνωρίσματα των συνόλων δεδομένων. Αυτό πραγματοποιείται με τη βοήθεια αρχείων αντιστοίχησης (mappings) όπου μετασχηματίζονται τα γνωρίσματα των συνόλων δεδομένων σε μία από τις παραπάνω τέσσερις μορφές που είναι αναγνωρίσιμες από τα GOA αρχεία.

Στη συνέχεια, προχωρούμε στο βήμα 3 όπου μέσω των GOA αρχείων αντιστοιχούμε GO όρους στα γονίδια. Στο βήμα αυτό γίνεται επίσης η επιλογή των επισημειώσεων που βασίζονται σε αποδεκτούς τύπους αποδείξεων (Evidence Codes), ενώ τελικά επιλέγονται μόνο οι GO όροι που ανήκουν στις επιθυμητές οντολογίες– απόψεις. Αφού έχει γίνει η αντιστοίχηση των γονιδίων του φιλτραρισμένου συνόλου γνωρισμάτων σε GO όρους, καθώς και των γονιδίων που είχαν αποκλειστεί από τη διαδικασία επιλογής γνωρισμάτων επιχειρείται ο υπολογισμός της σημασιολογικής ομοιότητας μεταξύ των γονιδίων που ανήκουν σε αυτές τις δύο ομάδες με τη βοήθεια του αρχείου της Γονιδιακής Οντολογίας.

Σε προηγουμένη ενότητα αναφερθήκαμε εκτενώς στη σημασιολογική ομοιότητα μεταξύ δύο GO όρων. Στο σημείο αυτό πρέπει να αναλύσουμε πως χειριζόμαστε τη σημασιολογική ομοιότητα μεταξύ δύο γονιδίων που ουσιαστικά περιγράφονται (το καθένα από αυτά) με μία σειρά GO όρων.

Ας υποθέσουμε ότι στο γονίδιο A αντιστοιχούν N_A και στο γονίδιο B N_B GO όροι. Αφού υπολογιστεί η σημασιολογική ομοιότητα κάθε πιθανού ζεύγους GO όρων που ανήκουν στο γονίδιο A και B (σύνολο: $N_A \cdot N_B$ συνδυασμοί) δημιουργείται ένας πίνακας (SSM) διαστάσεων $N_A \times N_B$:

$$SSM = \underbrace{\left.\begin{bmatrix} ss_{1,1} & ss_{1,2} & ss_{1,3} & \cdots & ss_{1,N_B} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ ss_{N_A,1} & ss_{N_A,2} & ss_{N_A,3} & \cdots & ss_{N_A,N_B} \end{bmatrix}\right\} gene\ A}_{gene\ B}$$

MAX: Ένας τρόπος υπολογισμού της σημασιολογικής ομοιότητας μεταξύ δύο γονιδίων είναι βρίσκοντας τη μέγιστη ομοιότητα ("*MAX*") μεταξύ δύο οποιωνδήποτε GO όρων που ανήκουν στα γονίδια.

Πιο φορμαλιστικά, αυτό αποδίδεται ως εξής:

$$SSG_{MAX} = \max\{ss_{i,j}\}$$

$$\forall\ i \in \{1,2,...,N_A)\ \text{και}\ \forall\ j \in \{1,2,...,N_B)$$

όπου:

SSG_{MAX}: Η σημασιολογική ομοιότητα μεταξύ δύο γονιδίων βάσει του τρόπου υπολογισμού *MAX*

Επειδή όμως συνήθως ένα γονιδιακό προϊόν λαμβάνει ταυτόχρονα όλους τους ρόλους που του αποδίδονται από τους επισημειωμένους σε αυτό GO όρους, ορίζεται και ένας άλλος τρόπος υπολογισμού της ομοιότητας μεταξύ δύο γονιδίων.

AVG: Ένας τρόπος υπολογισμού της σημασιολογικής ομοιότητας μεταξύ δύο γονιδίων είναι βρίσκοντας τη μέση ομοιότητα ("*AVG*") μεταξύ δύο οποιωνδήποτε GO όρων που ανήκουν στα γονίδια [13].

Πιο φορμαλιστικά, αυτό αποδίδεται ως εξής:

$$SSG_{AVG} = \frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N_B} ss_{i,j}}{N_A \cdot N_B}$$

όπου:

SSG_{AVG}: Η σημασιολογική ομοιότητα μεταξύ δύο γονιδίων βάσει του τρόπου υπολογισμού *AVG*

Τέλος, ένας τελευταίος τρόπος υπολογισμού της σημασιολογικής ομοιότητας μεταξύ δύο γονιδίων είναι ο *AVG_MAX* που αποτελεί μία μέθοδο που συνδυάζει τις δύο παραπάνω.

AVG_MAX: Για κάθε όρο ενός γονιδίου απομονώνουμε τη μέγιστη ομοιότητά του σε σχέση με όλους τους όρους του άλλου γονιδίου. Έτσι, βρίσκουμε την ομοιότητα των όρων ενός γονιδίου σε σχέση με τους όρους του άλλου γονιδίου. Σχηματίζουμε, έτσι, δύο πίνακες διαστάσεων $1 \times N_A$ και $1 \times N_B$ αντίστοιχα που περιέχουν τις ομοιότητες των όρων του γονιδίου A ως προς το γονίδιο B και τις ομοιότητες των όρων του γονιδίου B ως προς το γονίδιο A [3]. Θα συμβολίσουμε

τον πίνακα διαστάσεων $1 \times N_A$ που περιέχει τις ομοιότητες των όρων του γονιδίου A ως προς το γονίδιο B με $sim(A,B)$ και τον πίνακα διαστάσεων $1 \times N_B$ που περιέχει τις ομοιότητες των όρων του γονιδίου B ως προς το γονίδιο A με $sim(B,A)$.

Αυτό γράφεται ως εξής:

$$sim(A,B) = [\max_{j=\{1,2,\ldots,N_B\}}\{ss_{1,j}\} \quad \max_{j=\{1,2,\ldots,N_B\}}\{ss_{2,j}\} \quad \ldots \quad \max_{j=\{1,2,\ldots,N_B\}}\{ss_{N_A,j}\}]$$

και

$$sim(B,A) = [\max_{i=\{1,2,\ldots,N_A\}}\{ss_{i,1}\} \quad \max_{i=\{1,2,\ldots,N_A\}}\{ss_{i,2}\} \quad \ldots \quad \max_{i=\{1,2,\ldots,N_A\}}\{ss_{i,N_B}\}]$$

Τέλος, υπολογίζουμε τη μέση ομοιότητα του γονιδίου A προς το γονίδιο B ($\overline{sim(A,B)} = \frac{\sum_{i=1}^{N_A} \max_{j=\{1,2,\ldots,N_B\}}\{ss_{i,j}\}}{N_A}$) και του γονιδίου B προς το γονίδιο A ($\overline{sim(B,A)} = \frac{\sum_{j=1}^{N_B} \max_{i=\{1,2,\ldots,N_A\}}\{ss_{i,j}\}}{N_B}$).

Έτσι, η ομοιότητα μεταξύ των δύο γονιδίων είναι μέσος όρος των παραπάνω δύο μέσων ομοιοτήτων:

$$SSG_{AVG\_MAX} = \frac{\overline{sim(A,B)} + \overline{sim(B,A)}}{2}$$

όπου:

SSG_{AVG_MAX}: Η σημασιολογική ομοιότητα μεταξύ δύο γονιδίων βάσει του τρόπου υπολογισμού *AVG_MAX*

Στο τέταρτο βήμα, απομονώνουμε τον αριθμό των γονιδίων που εμφανίζουν την υψηλότερη σημασιολογική ομοιότητα με αυτά του φιλτραρισμένου συνόλου δεδομένων κατόπιν ορισμού των κατάλληλων παραμέτρων και δημιουργούμε ένα νέο σύνολο δεδομένων που περιέχει τόσο τα γονίδια–γνωρίσματα από το φιλτραρισμένο σύνολο δεδομένων όσο και τα γονίδια–γνωρίσματα που είναι σημασιολογικά συσχετισμένα με τα προηγούμενα. Συνδυάζεται, έτσι, γνώση από τη στατιστική ομοιότητα των γονιδίων εντός του συνόλου δεδομένων και από τη σημασιολογική ομοιότητα που προέρχεται από τη Γονιδιακή Οντολογία.

Στο τελευταίο βήμα (πέμπτο), πραγματοποιείται η εκμάθηση των παραπάνω συνόλων με στόχο την εξαγωγή και τη σύγκριση αποτελεσμάτων.

6 Αποτίμηση των αποτελεσμάτων

Στο σημείο αυτό θα αποτιμήσουμε την αξία των αποτελεσμάτων από το εργαλείο του SoFoCles. Για το σκοπό αυτό εργαστήκαμε με δύο σύνολα δεδομένων. Το ένα μας πληροφορεί για τις επιδράσεις του καπνίσματος στη μεταγραφή του γονιδιώματος των επιθηλιακών κυττάρων του ανθρώπινου αναπνευστικού συστήματος και το άλλο για τα αποτελέσματα της θεραπείας σε ασθενείς που έχουν αναπτύξει εμβρυϊκούς όγκους στο Κεντρικό Νευρικό Σύστημα (Central Nervous System, CNS). Από δω και στο εξής θα αναφερόμαστε στο σύνολο που εξετάζει τις επιδράσεις του καπνίσματος στη μεταγραφή του γονιδιώματος των επιθηλιακών κυττάρων του ανθρώπινου αναπνευστικού συστήματος ως D_S σύνολο δεδομένων και στο σύνολο που εξετάζει τα αποτελέσματα της θεραπείας σε ασθενείς που έχουν αναπτύξει εμβρυϊκούς όγκους στο Κεντρικό Νευρικό Σύστημα ως D_{CNV} σύνολο δεδομένων.

Για την αποτίμηση των αποτελεσμάτων χρησιμοποιείται κυρίως η ακρίβεια ταξινόμησης (classification accuracy) η οποία μας δίνει μία γενική εικόνα για την ποιότητα των αποτελεσμάτων. Παρόλα αυτά, όταν τα εξαγόμενα αποτελέσματα αρκούνται μόνο σε αυτό το μέτρο είναι ατελή και συχνά μπορεί να οδηγήσουν σε παρερμηνείες.

Εξάλλου, η αξία κάθε πρόβλεψης δεν είναι ίδια σε σχέση με μία άλλη. Με άλλα λόγια, δεν έχει γίνει αναφορά ακόμα στο κόστος μία λανθασμένης απόφασης ή ταξινόμησης. Η βελτιστοποίηση της ακρίβειας ταξινόμησης χωρίς να λαμβάνεται υπόψη η παράλληλη μεταβολή του κόστους των λανθασμένων προβλέψεων συχνά οδηγεί σε παραπλανητικά αποτελέσματα. Επί παραδείγματι, αν ένα σύνολο δεδομένων αναφέρεται σε ασθενείς που έχουν αναπτύξει κάποιο είδος καρκίνου και σε υγιείς, οι συνέπειες είναι πολύ διαφορετικές από το να ταξινομηθεί ένας υγιής ασθενής ως καρκινοπαθής σε σχέση με την ταξινόμηση ενός καρκινοπαθούς ασθενούς ως υγιούς. Στην πρώτη περίπτωση, ίσως να υπάρξουν παρενέργειες από τη χορήγηση κάποιας φαρμακευτικής αγωγής (χρήση φαρμάκων, υποβολή σε

χημειοθεραπείες κ.α.), αλλά στη δεύτερη περίπτωση ίσως ο θάνατος του ασθενούς να είναι αναπόφευκτος αφού δεν θα κινηθούν οι κατάλληλες διαδικασίες για την αποθεραπεία του.

Επιπλέον, η εξαγωγή κανόνων από ένα μοντέλο εκμάθησης με υψηλό δείκτη λανθασμένων προβλέψεων είναι αρκετά επισφαλής και μπορεί να οδηγήσει σε ενέργειες που ενδέχεται να έχουν αντίθετες συνέπειες στο δείγμα από τις προσδοκώμενες. Για παράδειγμα, αν η υψηλή έκφραση δύο γονιδίων καταδεικνύει την ύπαρξη καρκίνου με υψηλό όμως βαθμό λανθασμένων προβλέψεων, η προσπάθεια καταστολής τους μπορεί να έχει τελείως αντίθετα αποτελέσματα από τα επιθυμητά.

Έτσι, για αυτούς τους λόγους έχουν οριστεί μέτρα που αντιστοιχούν σε κάθε ετικέτα της κλάσης ξεχωριστά και όχι όπως πριν σε όλο το σύνολο δεδομένων.

6.1 Μέτρα ελέγχου της αξιοπιστίας των αποτελεσμάτων

Έχοντας μία ετικέτα κλάσης, έστω A, ορίζουμε για αυτήν τέσσερα μεγέθη, τα οποία είναι τα εξής:

TP: True Positive (Πραγματικά Θετικός) και χαρακτηρίζει ένα παράδειγμα όταν αποδίδεται στην ετικέτα A και ανήκει πράγματι σε αυτήν

FP: False Positive (Λανθασμένα Θετικός) και χαρακτηρίζει ένα παράδειγμα όταν αποδίδεται στην ετικέτα A αλλά στην πραγματικότητα δεν ανήκει σε αυτήν

TN: True Negative (Πραγματικά Αρνητικός) και χαρακτηρίζει ένα παράδειγμα όταν δεν αποδίδεται στην ετικέτα A και δεν ανήκει πράγματι σε αυτήν

FN: False Negative (Λανθασμένα Αρνητικός) και χαρακτηρίζει ένα παράδειγμα όταν δεν αποδίδεται στην ετικέτα A αλλά στην πραγματικότητα ανήκει σε αυτήν

Στηριζόμενα στα παραπάνω μεγέθη έχουν οριστεί τα εξής μέτρα αξιοπιστίας για κάθε ετικέτα κλάσης (για λόγους συμβολισμού θα αποδώσουμε σε μία ετικέτα κλάσης το σύμβολο A):

Ρυθμός Πραγματικών Θετικών (**TP Rate** ή **Recall**): Είναι ο ρυθμός των Πραγματικών Θετικών της κλάσης A. Μ' άλλα λόγια, μας πληροφορεί για το

ποσοστό των παραδειγμάτων που ταξινομήθηκαν σωστά στην κλάση A προς το σύνολο των παραδειγμάτων που πράγματι ανήκουν στην κλάση A.

$$TP\ Rate = Recall = \frac{\#\ of\ instances\ that\ were\ correctly\ classified\ to\ class\ label\ A}{\#\ of\ instances\ that\ belong\ to\ class\ label\ A} \Leftrightarrow$$

$$TP\ Rate = Recall = \frac{TP}{TP + FN}$$

Ο $TP\ Rate$ αυξάνει με την αύξηση του TP και μειώνεται με την αύξηση του FN. Με απλά λόγια, όσο περισσότερα παραδείγματα που πράγματι ανήκουν στην κλάση A ταξινομούνται σε αυτήν, τόσο ο $TP\ Rate$ αυξάνει.

Η ελάχιστη τιμή που μπορεί να λάβει αυτό το μέτρο είναι μηδέν (0) και προκύπτει για $TP = 0$ και $FN = \#\ of\ instances\ that\ belong\ to\ class\ label\ A$, ενώ η μέγιστη είναι ένα (1) για $TP = \#\ of\ instances\ that\ belong\ to\ class\ label\ A$ και $FN = 0$.

Ρυθμός Λανθασμένων Θετικών (**FP Rate**): Είναι ο ρυθμός των Λανθασμένων Θετικών της κλάσης A. Μ' άλλα λόγια, μας πληροφορεί για το ποσοστό των παραδειγμάτων που ταξινομήθηκαν λανθασμένα στην κλάση A προς το σύνολο των παραδειγμάτων που στην πραγματικότητα δεν ανήκουν στην κλάση A.

$$FP\ Rate = \frac{\#\ of\ instances\ that\ were\ incorrectly\ classified\ to\ class\ label\ A}{\#\ of\ instances\ that\ don't\ belong\ to\ class\ label\ A} \Leftrightarrow$$

$$FP\ Rate = \frac{FP}{FP + TN}$$

Ο $FP\ Rate$ μειώνεται με τη μείωση του FP και αυξάνει με τη μείωση του TN. Με απλά λόγια, όσο περισσότερα παραδείγματα που στην πραγματικότητα δεν ανήκουν στην κλάση A δεν ταξινομούνται σε αυτήν, τόσο ο $FP\ Rate$ μειώνεται.

Η ελάχιστη τιμή που μπορεί να λάβει αυτό το μέτρο είναι μηδέν (0) και προκύπτει για $FP = 0$ και $TN = \#\ of\ instances\ that\ don't\ belong\ to\ class\ label\ A$, ενώ η μέγιστη είναι ένα (1) για $FP = \#\ of\ instances\ that\ don't\ belong\ to\ class\ label\ A$ και $TN = 0$.

Precision: Μας πληροφορεί για το ποσοστό των παραδειγμάτων που ταξινομήθηκαν σωστά στην κλάση A προς το σύνολο των παραδειγμάτων που ταξινομήθηκαν (σωστά ή λανθασμένα) στην κλάση A.

$$Precision = \frac{\#\ of\ instances\ that\ were\ correctly\ classified\ to\ class\ label\ A}{\#\ of\ instances\ that\ were\ (correctly\ or\ incorrectly)\ classified\ to\ class\ label\ A} \Leftrightarrow$$

$$Precision = \frac{TP}{TP + FP}$$

Το $Precision$ αυξάνει με την αύξηση του TP και μειώνεται με την αύξηση του FP. Με απλά λόγια, όσο περισσότερα παραδείγματα ταξινομούνται σωστά στην κλάση A, τόσο το $Precision$ αυξάνει.

Η ελάχιστη τιμή που μπορεί να λάβει αυτό το μέτρο είναι μηδέν (0) και προκύπτει για $TP = 0$ και $FP = \#\ of\ instances\ that\ were\ (correctly\ or\ incorrectly)\ classified\ to\ class\ label\ A$, ενώ η μέγιστη είναι ένα (1) για $TP = \#\ of\ instances\ that\ were\ (correctly\ or\ incorrectly)\ classified\ to\ class\ label\ A$ και $FP = 0$.

F - Measure: Ορίζεται ως

$$F\text{-}Measure = \frac{2 \cdot TP_{Rate} \cdot Precision}{TP_{Rate} + Precision}$$

Επειδή ισχύει $\frac{\partial(F\text{-}Measure)}{\partial TP_{Rate}} = \frac{2 \cdot Precision^2}{(TP_{Rate} + Precision)^2} > 0$ και $\frac{\partial(F\text{-}Measure)}{\partial Precision} = \frac{2 \cdot {TP_{Rate}}^2}{(TP_{Rate} + Precision)^2} > 0$, το $F\text{-}Measure$ αυξάνει με την αύξηση του TP_{Rate} και του $Precision$.

Η ελάχιστη τιμή που μπορεί να πάρει είναι μηδέν (0) για $TP_{Rate} = 0$ ή $Precision = 0$ και η μέγιστη ένα (1) για $TP_{Rate} = 1$ και $Precision = 1$ [26].

Τέλος, η **ακρίβεια ταξινόμησης** (**classification accuracy**) ορίζεται ως ο αριθμός των σωστά ταξινομημένων παραδειγμάτων προς το σύνολο των παραδειγμάτων:

$$Classification\_accuracy = \frac{\#\ of\ correctly\ classsified\ instances}{total\ \#\ of\ instances}$$

Ο ορισμός έχει αρκετά κοινά στοιχεία με τον ορισμό του μέτρου $Precision$. Προφανώς συνδέονται με κάποια σχέση. Θα προσπαθήσουμε σε αυτό το σημείο να βρούμε τη σχέση που συνδέει τους δύο τύπους.

Έστω ότι έχουμε μία κλάση με K ετικέτες. Το Πραγματικά Θετικό (TP) μέγεθος της ετικέτας i, συμβολίζεται με TP_i και δηλώνει τον αριθμό των παραδειγμάτων που έχουν ταξινομηθεί σωστά στην ετικέτα i, ενώ το $Precision_i$ δηλώνει το ποσοστό των παραδειγμάτων που ταξινομήθηκαν σωστά στην ετικέτα i.

Ο συνολικός αριθμός των παραδειγμάτων ισούται με το άθροισμα των παραδειγμάτων που ταξινομούνται σε κάθε ετικέτα της κλάσης.

Πιο αυστηρά, αυτό αποδίδεται ως εξής:

$$total\ \#\ of\ instances= \sum_{i=1}^{K}(\#\ of\ instances\ that\ were\ (correctly\ or\ incorrectly)\ classified\ to\ class\ label\ i)$$

Ισχύει:

$$\#\ of\ instances\ that\ were\ (correctly\ or\ incorrectly)\ classified\ to\ class\ label\ i= \frac{TP_i}{Precision_i}$$

Ενώ ο αριθμός των σωστά ταξινομημένων παραδειγμάτων ισούται με το άθροισμα των σωστά ταξινομημένων παραδειγμάτων σε κάθε ετικέτα της κλάσης.

Έτσι, η ακρίβεια ταξινόμησης παίρνει τη μορφή:

$$Classification\_accuracy = \frac{\sum_{i=1}^{K} TP_i}{\sum_{i=1}^{K} \frac{TP_i}{Precision_i}}$$

Η ακρίβεια ταξινόμησης εμφανίζει ελάχιστο για $\#\ of\ correctly\ classsified\ instances = 0$ το μηδέν (0) και μέγιστο για $\#\ of\ correctly\ classsified\ instances = total\ \#\ of\ instances$ το ένα (1).

Έτσι, για την αποτίμηση των αποτελεσμάτων μας θα στηριχτούμε σε πέντε (5) μέτρα:

a) την ακρίβεια ταξινόμησης
b) το TP Rate
c) το FP Rate
d) το Precision και
e) το F – Measure

6.2 D_S σύνολο δεδομένων

6.2.1 Παρουσίαση του D_S συνόλου δεδομένων και των αποτελεσμάτων του

Το D_S σύνολο εξετάζει τις αλλαγές που προκαλούνται στη γονιδιακή έκφραση από το κάπνισμα υπό το πρίσμα πολλών διαφορετικών παραμέτρων όπως η συνεχής έκθεση (στο κάπνισμα), η ηλικία, το φύλο και η φυλή. Ο απώτερος σκοπός αυτού του πειράματος είναι η έρευνα του βιολογικού υποβάθρου ενός συγκεκριμένου υποσυνόλου των κυττάρων πολλών ατόμων για τον εντοπισμό αναστρέψιμων και μη γενετικών επιδράσεων του καπνίσματος στα επιθηλιακά κύτταρα του ανθρώπινου αναπνευστικού συστήματος.

Το σύνολο δεδομένων αποτελείται από 22216 γνωρίσματα (22215 γονίδια–γνωρίσματα και μία κλάση). Η κλάση χωρίζεται σε τρεις ετικέτες: {current-smoker (καπνιστής), never-smoked (μη-καπνιστής), former-smoker (πρώην-καπνιστής)}. Το σύνολο περιέχει 74 παραδείγματα (άτομα), τα 33 εκ των οποίων είναι καπνιστές, τα 23 είναι μη-καπνιστές και τα 18 είναι πρώην-καπνιστές.

Τα γνωρίσματα αντιπροσωπεύουν GeneBank accession numbers και επομένως για τη μετατροπή τους χρειάστηκαν τα αρχεία αντιστοίχησης (από GeneBank accession numbers σε γονιδιακά σύμβολα) NetAffx HG-U133A, η επεξεργασία των οποίων έγινε με το πρόγραμμα R [22].

Για τα πειράματά μας χρησιμοποιήσαμε το αρχικό σύνολο δεδομένων και μία μικρότερη έκδοσή του μετά την εφαρμογή της Relief-F μεθόδου φιλτραρίσματος γνωρισμάτων. Το μικρότερο σύνολο δεδομένων αποτελείται από 40 γνωρίσματα. Χρησιμοποιήθηκαν επίσης τα εξής GOA αρχεία: του ανθρώπου (gene_association.goa_human), του ποντικού (gene_association.mgi) και του αρουραίου (gene_association.rgd). Τα δύο τελευταία χρησιμοποιήθηκαν εξαιτίας της μεγάλης ομοιότητας που παρουσιάζουν οι δύο αυτοί οργανισμοί με τον άνθρωπο. Επιπλέον, έγινε χρήση του πιο πρόσφατου αρχείου Γονιδιακής Οντολογίας (GO file) αυτή του Μαΐου 2007. Από την αναζήτηση αποκλείστηκαν οι όροι που ήταν βασισμένοι σε IEA τύπους αποδείξεων εξαιτίας της μικρής αξιοπιστίας αυτών των επισημειώσεων, ενώ η αναζήτηση, που βασίστηκε στο ταίριασμα μεταξύ IDs, συμβόλων ή συνωνύμων, επιλέχτηκε να είναι μερική (και όχι ακριβής) για να είναι πιο ευέλικτη. Για παράδειγμα, εάν επιχειρείται η αναζήτηση των όρων του γονιδίου SLIT1, τότε με την επιλογή του μερικού ταιριάσματος επιτυγχάνεται και η προσθήκη όρων του γονιδίου SLIT1 στον άνθρωπο που συμβολίζεται ως SLIT1_HUMAN.

Εξάλλου, ο υπολογισμός της σημασιολογικής ομοιότητας μεταξύ των γονιδίων στηρίχτηκε μόνο σε όρους από την BP οντολογία καθώς αυτοί οι όροι περιγράφουν πλήρεις βιολογικές διαδρομές (διαδικασίες) παρά μεμονωμένες λειτουργίες. Τέλος, επιλέχτηκε ως ελάχιστο κατώφλι σημασιολογικής ομοιότητας το 50%, ενώ ο μέγιστος αριθμός των πιο συσχετισμένων γονιδίων για κάθε γονίδιο του συνόλου δεδομένων (μετά την εφαρμογή του φιλτραρίσματος γνωρισμάτων) επιλέχτηκε να είναι το 5 και ο μέγιστος αριθμός των τελικά προκρινόμενων γονιδίων επιλέχτηκε να είναι ο 20 (ο μισός του αριθμού γνωρισμάτων του μικρότερου συνόλου δεδομένων).

Πειραματιστήκαμε μεταξύ διαφόρων μεθόδων σημασιολογικής ομοιότητας (Resnik, Lin, Jiang & Conrath, Cao, oldZZL) και τρόπων υπολογισμού της ομοιότητας μεταξύ δύο γονιδίων (AVG, MAX, AVG_MAX) δημιουργώντας έτσι $5 \cdot 3 = 15$ σύνολα δεδομένων.

Για μία πιο ολοκληρωμένη προσπάθεια αποτίμησης των αποτελεσμάτων δημιουργήσαμε 3 επιπλέον σύνολα δεδομένων με αριθμό γνωρισμάτων ίσο με το μέσο όρο του αριθμού γνωρισμάτων των συνόλων δεδομένων που προκύπτουν μετά την εφαρμογή του εργαλείου SoFoCles. Το πρώτο αποτελείται από 74 γνωρίσματα μετά την εφαρμογή της μεθόδου Relief-F (όπως είχε γίνει και για το σύνολο των 40 γνωρισμάτων), το δεύτερο αποτελείται από τα 40 γνωρίσματα του αρχικού μικρού συνόλου συν 34 τυχαίως εξαγόμενα γνωρίσματα από το αρχικό πλήρες σύνολο και το τρίτο από 74 τυχαίως εξαγόμενα γνωρίσματα από το αρχικό πλήρες σύνολο.

Τα αποτελέσματά μας συγκεντρώνονται στους παρακάτω πίνακες:

Πίνακας XI. Παρουσίαση των ποιοτικών χαρακτηριστικών για το D_S σύνολο δεδομένων

	Total Accuracy	Current Smoker			
		TP Rate	FP Rate	Precision	F-Measure
whole	64.86	0.788	0.244	0.722	0.754
refined (40 att)	77.03	0.879	0.220	0.763	0.817
refined (74 att)	79.73	0.879	0.146	0.829	0.853
refined (40 att + 34 rand)	70.27	0.758	0.195	0.758	0.758
refined (74 rand)	50.00	0.727	0.390	0.600	0.658
Resnik.AVG	72.97	0.788	0.122	0.839	0.813
Lin.AVG	74.32	0.879	0.171	0.806	0.841
JiangConrath.AVG	75.68	0.818	0.146	0.818	0.818
Cao.AVG	71.62	0.879	0.195	0.784	0.829
oldZZL.AVG	77.03	0.818	0.146	0.818	0.818
Resnik.MAX	81.08	0.939	0.171	0.816	0.873
Lin.MAX	70.27	0.758	0.146	0.806	0.781
JiangConrath.MAX	85.14	0.939	0.146	0.838	0.886
Cao.MAX	71.62	0.818	0.171	0.794	0.806
oldZZL.MAX	82.43	0.909	0.122	0.857	0.882
Resnik.AVG_MAX	74.32	0.758	0.146	0.806	0.781
Lin.AVG_MAX	77.03	0.848	0.146	0.824	0.836
JiangConrath.AVG_MAX	78.38	0.879	0.171	0.806	0.841
Cao.AVG_MAX	78.38	0.848	0.122	0.848	0.848
oldZZL.AVG_MAX	81.08	0.909	0.146	0.833	0.870

	Never Smoked			
	TP Rate	FP Rate	Precision	F-Measure
whole	0.696	0.118	0.727	0.711
refined (40 att)	0.913	0.039	0.913	0.913
refined (74 att)	0.913	0.059	0.875	0.894
refined (40 att + 34 rand)	0.826	0.059	0.864	0.844
refined (74 rand)	0.304	0.157	0.467	0.368
Resnik.AVG	0.826	0.078	0.826	0.826
Lin.AVG	0.783	0.059	0.857	0.818
JiangConrath.AVG	0.870	0.078	0.833	0.851
Cao.AVG	0.739	0.078	0.810	0.773
oldZZL.AVG	0.870	0.059	0.870	0.870
Resnik.MAX	0.870	0.059	0.870	0.870
Lin.MAX	0.826	0.078	0.826	0.826
JiangConrath.MAX	1.000	0.059	0.885	0.939
Cao.MAX	0.826	0.078	0.826	0.826
oldZZL.MAX	0.957	0.078	0.846	0.898
Resnik.AVG_MAX	0.913	0.059	0.875	0.894

Lin.AVG_MAX	0.826	0.078	0.826	0.826
JiangConrath.AVG_MAX	0.870	0.059	0.870	0.870
Cao.AVG_MAX	0.870	0.078	0.833	0.851
oldZZL.AVG_MAX	0.913	0.059	0.875	0.894

	Former TP Rate	**Smoker FP Rate**	**Precision**	**F-Measure**
whole	0.333	0.179	0.375	0.353
refined (40 att)	0.389	0.107	0.538	0.452
refined (74 att)	0.500	0.107	0.600	0.545
refined (40 att + 34 rand)	0.444	0.196	0.421	0.432
refined (74 rand)	0.333	0.232	0.316	0.324
Resnik.AVG	0.500	0.196	0.450	0.474
Lin.AVG	0.444	0.161	0.471	0.457
JiangConrath.AVG	0.500	0.143	0.529	0.514
Cao.AVG	0.389	0.161	0.438	0.412
oldZZL.AVG	0.556	0.143	0.556	0.556
Resnik.MAX	0.500	0.071	0.692	0.581
Lin.MAX	0.444	0.214	0.400	0.421
JiangConrath.MAX	0.500	0.036	0.818	0.621
Cao.MAX	0.389	0.179	0.412	0.400
oldZZL.MAX	0.500	0.071	0.692	0.581
Resnik.AVG_MAX	0.500	0.179	0.474	0.486
Lin.AVG_MAX	0.556	0.125	0.588	0.571
JiangConrath.AVG_MAX	0.500	0.107	0.600	0.545
Cao.AVG_MAX	0.556	0.125	0.588	0.571
oldZZL.AVG_MAX	0.500	0.089	0.643	0.563

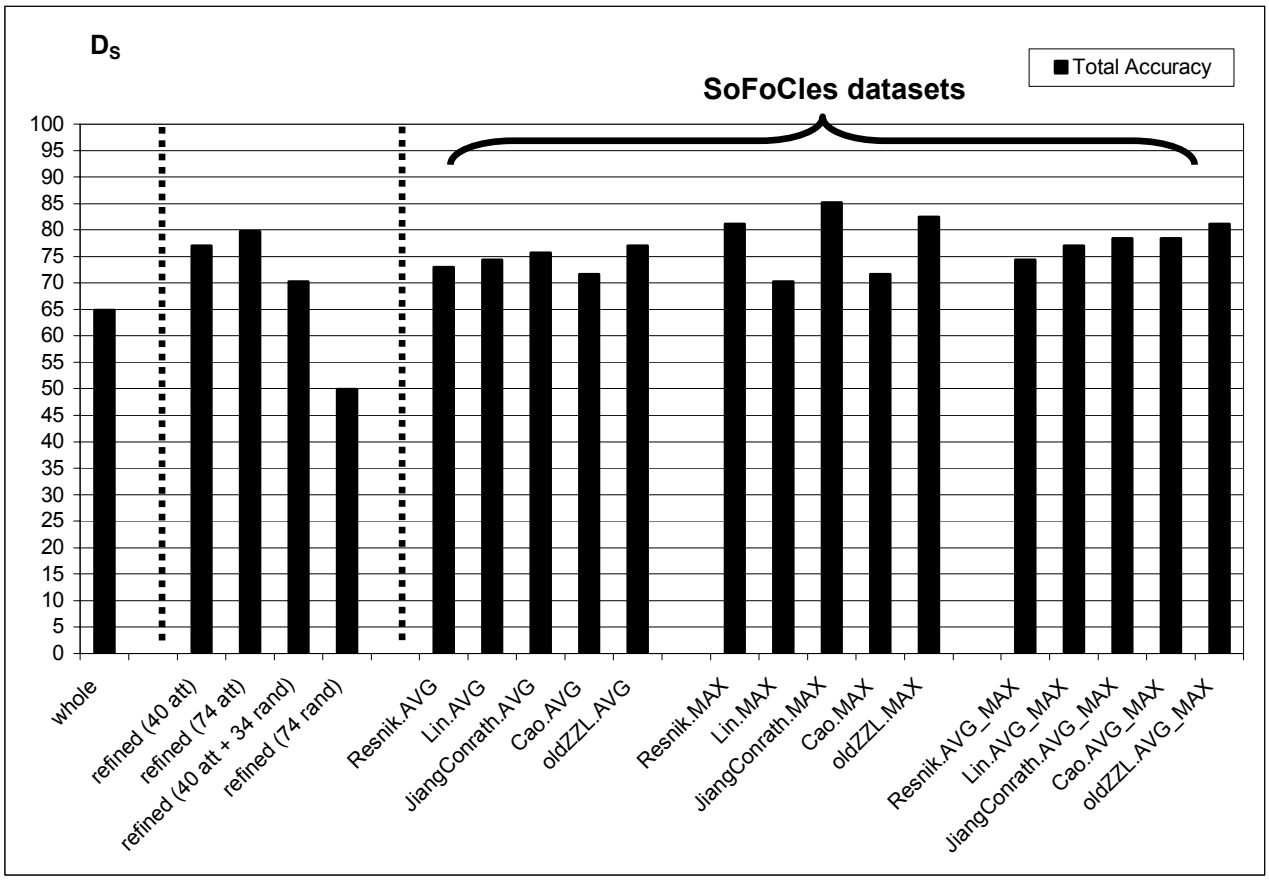
Πίνακας XII. Confusion-Matrix για το D_S σύνολο δεδομένων

	Current-Smoker	**Never-Smoked**	**Former-Smoker**	**← *classified as***	**# of instances**
whole	26	1	6	**Current-Smoker**	**33**
	3	16	4	**Never-Smoked**	**23**
	7	5	6	**Former-Smoker**	**18**
refined (40 att)	29	0	4	**Current-Smoker**	**33**
	0	21	2	**Never-Smoked**	**23**
	9	2	7	**Former-Smoker**	**18**
refined (74 att)	29	0	4	**Current-Smoker**	**33**

	0	21	2	**Never-Smoked**	**23**
	6	3	9	**Former-Smoker**	**18**
refined (40 att + 34 rand)	25	0	8	**Current-Smoker**	**33**
	1	19	3	**Never-Smoked**	**23**
	7	3	8	**Former-Smoker**	**18**
refined (74 rand)	24	4	5	**Current-Smoker**	**33**
	8	7	8	**Never-Smoked**	**23**
	8	4	6	**Former-Smoker**	**18**
Resnik.AVG	26	0	7	**Current-Smoker**	**33**
	0	19	4	**Never-Smoked**	**23**
	5	4	9	**Former-Smoker**	**18**
Lin.AVG	29	0	4	**Current-Smoker**	**33**
	0	18	5	**Never-Smoked**	**23**
	7	3	8	**Former-Smoker**	**18**
JiangConrath.AVG	27	0	6	**Current-Smoker**	**33**
	1	20	2	**Never-Smoked**	**23**
	5	4	9	**Former-Smoker**	**18**
Cao.AVG	29	0	4	**Current-Smoker**	**33**
	1	17	5	**Never-Smoked**	**23**
	7	4	7	**Former-Smoker**	**18**
oldZZL.AVG	27	0	6	**Current-Smoker**	**33**
	1	20	2	**Never-Smoked**	**23**
	5	3	10	**Former-Smoker**	**18**
Resnik.MAX	31	0	2	**Current-Smoker**	**33**
	1	20	2	**Never-Smoked**	**23**
	6	3	9	**Former-Smoker**	**18**
Lin.MAX	25	0	8	**Current-Smoker**	**33**
	0	19	4	**Never-Smoked**	**23**

	6	4	8	**Former-Smoker**	**18**
JiangConrath.MAX	31	0	2	**Current-Smoker**	**33**
	0	23	0	**Never-Smoked**	**23**
	6	3	9	**Former-Smoker**	**18**
Cao.MAX	27	0	6	**Current-Smoker**	**33**
	0	19	4	**Never-Smoked**	**23**
	7	4	7	**Former-Smoker**	**18**
oldZZL.MAX	30	0	3	**Current-Smoker**	**33**
	0	22	1	**Never-Smoked**	**23**
	5	4	9	**Former-Smoker**	**18**
Resnik.AVG_MAX	25	0	8	**Current-Smoker**	**33**
	0	21	2	**Never-Smoked**	**23**
	6	3	9	**Former-Smoker**	**18**
Lin.AVG_MAX	28	1	4	**Current-Smoker**	**33**
	1	19	3	**Never-Smoked**	**23**
	5	3	10	**Former-Smoker**	**18**
JiangConrath.AVG_MAX	29	0	4	**Current-Smoker**	**33**
	1	20	2	**Never-Smoked**	**23**
	6	3	9	**Former-Smoker**	**18**
Cao.AVG_MAX	28	1	4	**Current-Smoker**	**33**
	0	20	3	**Never-Smoked**	**23**
	5	3	10	**Former-Smoker**	**18**
oldZZL.AVG_MAX	30	0	3	**Current-Smoker**	**33**
	0	21	2	**Never-Smoked**	**23**
	6	3	9	**Former-Smoker**	**18**

Τα αποτελέσματα του Πιν. XI οπτικοποιούνται στα παρακάτω διαγράμματα:



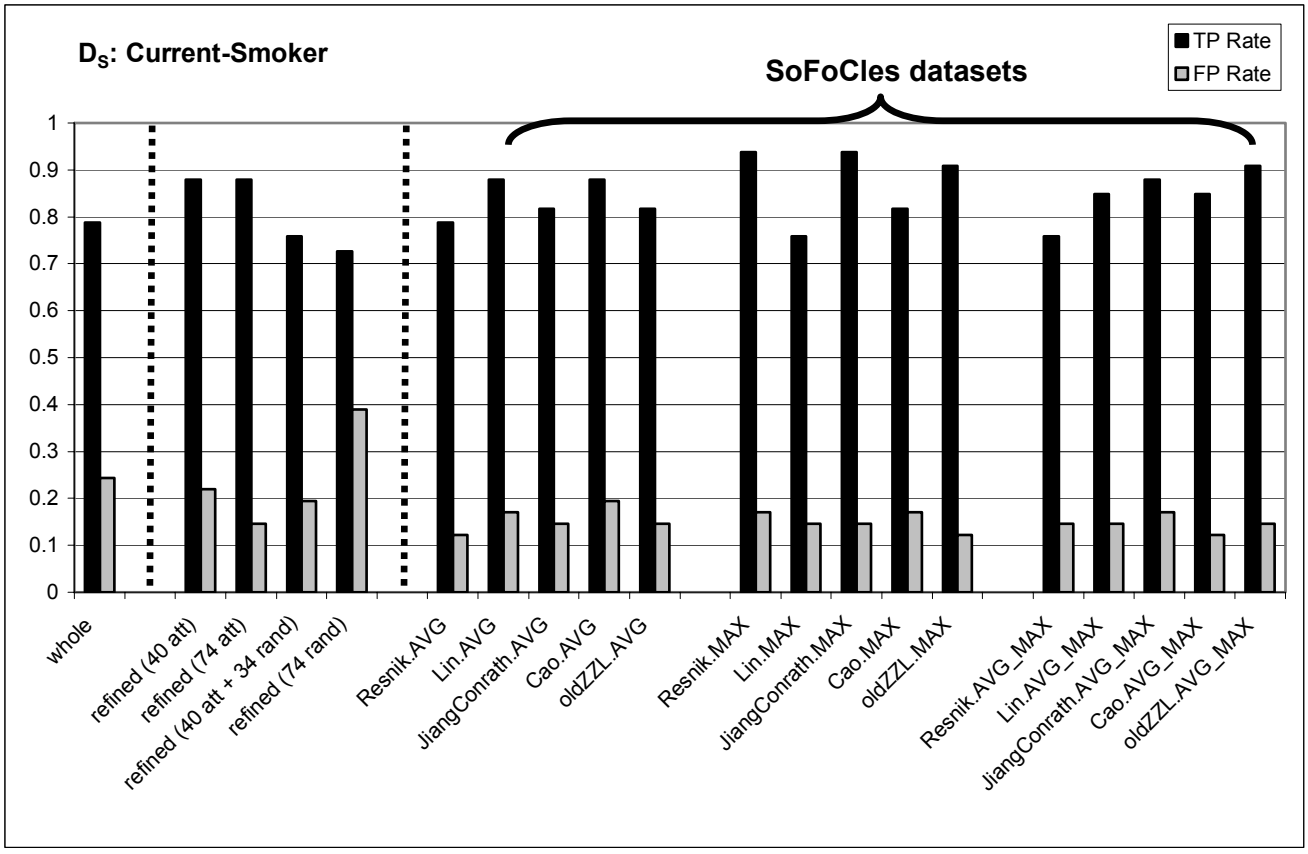
Εικόνα 6.1. Ακρίβεια ταξινόμησης για το D_S σύνολο δεδομένων

Από το αρχικό φιλτραρισμένο σύνολο δεδομένων (αυτό των 40 γνωρισμάτων) έξι σύνολα δεδομένων μετά την εφαρμογή του εργαλείου SoFoCles παρουσιάζουν μεγαλύτερη ακρίβεια ταξινόμησης ενώ δύο έχουν ίση ακρίβεια ταξινόμησης. Με άλλα λόγια, περισσότερα από τα μισά σύνολα δεδομένων που δημιουργήθηκαν μετά την εφαρμογή του εργαλείου SoFoCles εμφανίζονται να έχουν ίση ή μεγαλύτερη ακρίβεια.

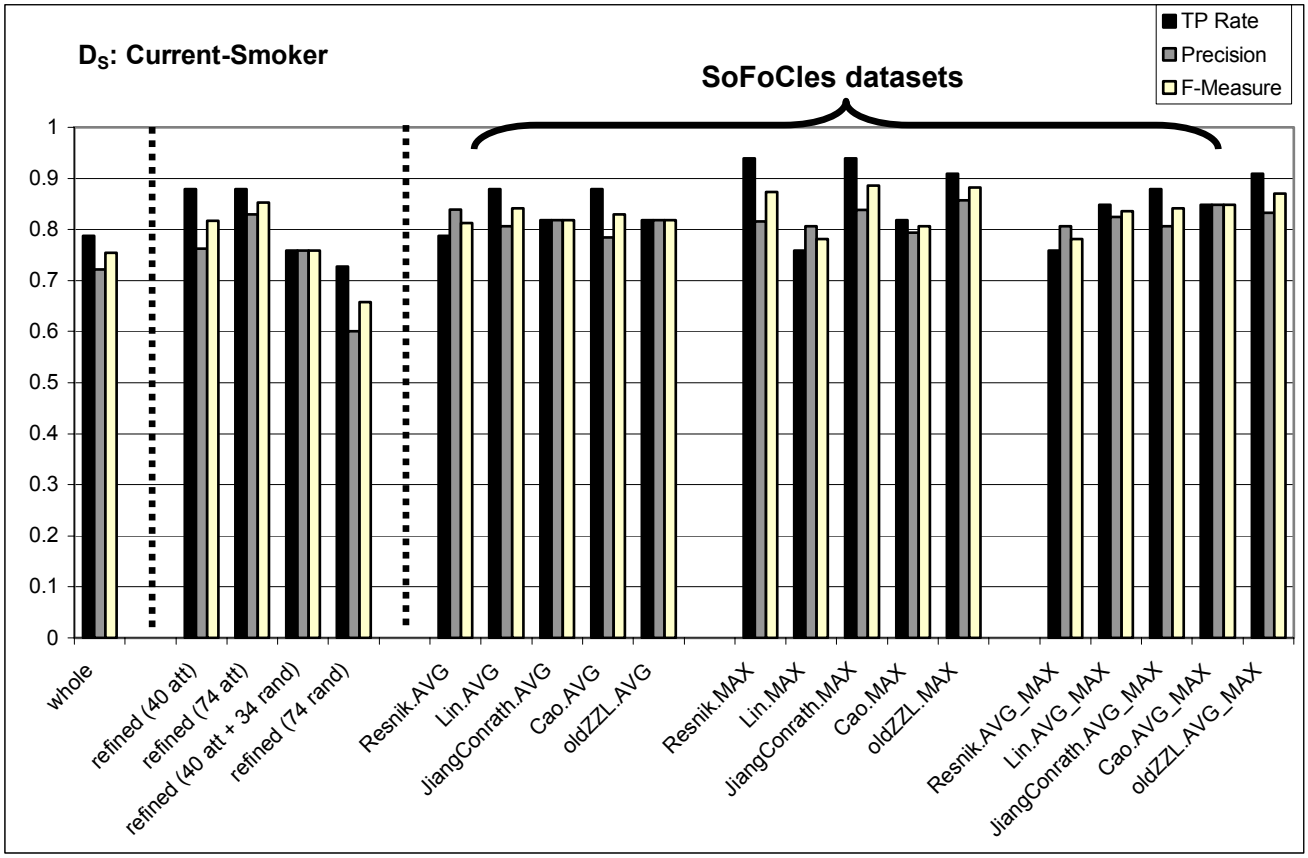
Από το φιλτραρισμένο σύνολο δεδομένων που αριθμεί 74 γνωρίσματα, τέσσερα σύνολα δεδομένων του SoFoCles εμφανίζουν μεγαλύτερη ακρίβεια ταξινόμησης.

Επίσης, όπως αναμενόταν, τα μερικώς και αμιγώς τυχαία σύνολα δεδομένων εμφανίζουν πολύ μικρότερη ακρίβεια ταξινόμησης. Για το αμιγώς τυχαίο σύνολο δεδομένων (αυτό των 74 τυχαίως εξαγόμενων γνωρισμάτων), η ακρίβεια ταξινόμησης είναι ακριβώς 50% κάτι που έρχεται σε πλήρη συμφωνία με τη φύση ενός τέτοιου συνόλου αφού η πρόβλεψη της ετικέτας ενός νέου παραδείγματος βασίζεται τελείως στην τύχη.

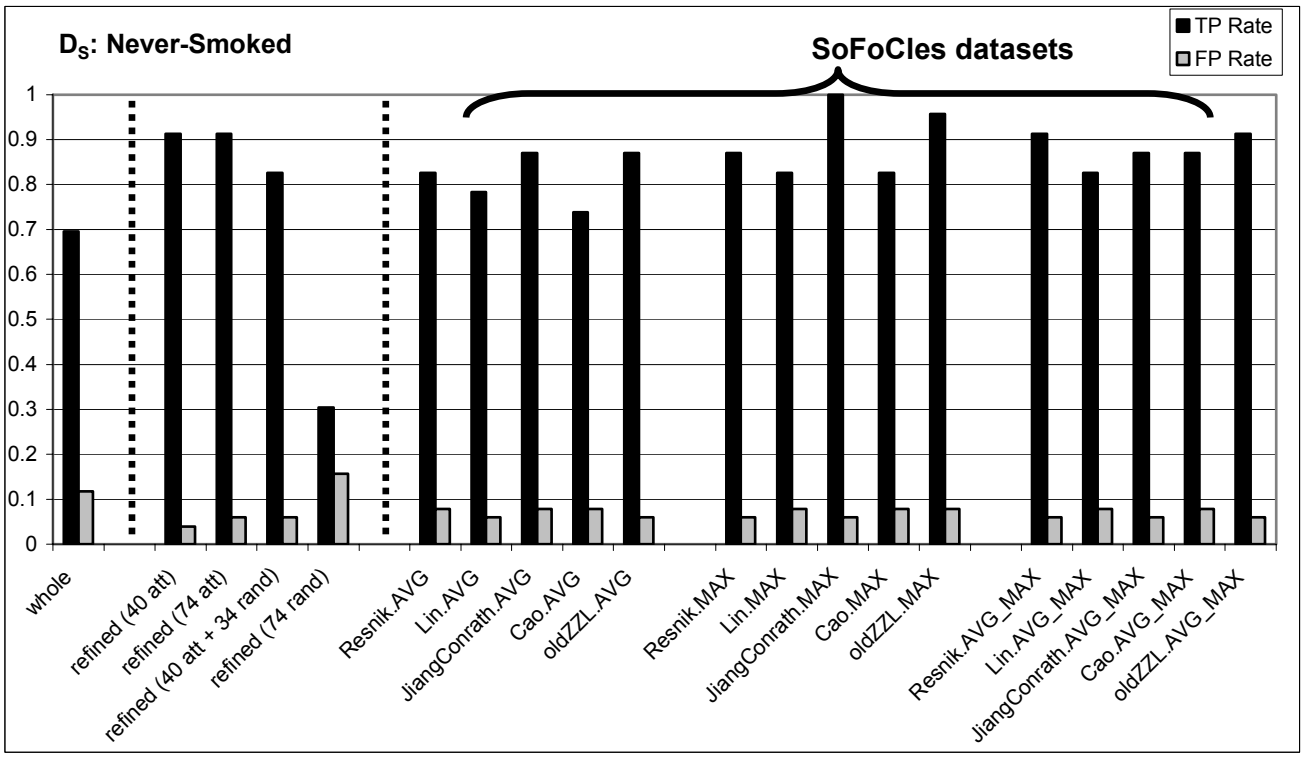
Παρακάτω θα παρουσιάσουμε τα ποιοτικά χαρακτηριστικά για κάθε ετικέτα της κλάσης ξεχωριστά:



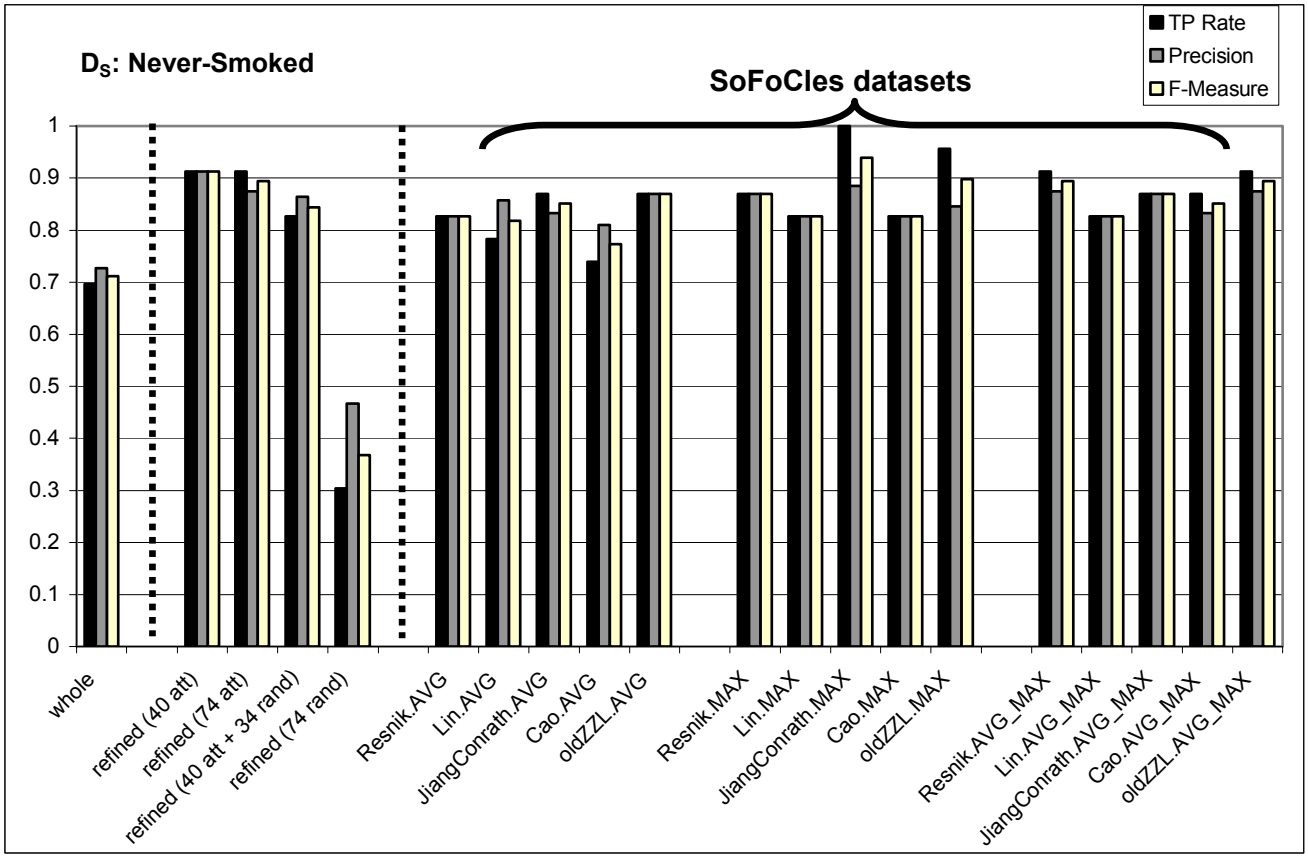
Εικόνα 6.2. Ρυθμοί Πραγματικά Θετικών (TP Rates) και Λανθασμένα Θετικών (FP Rates) της ετικέτας Current-Smoker για το D_S σύνολο δεδομένων



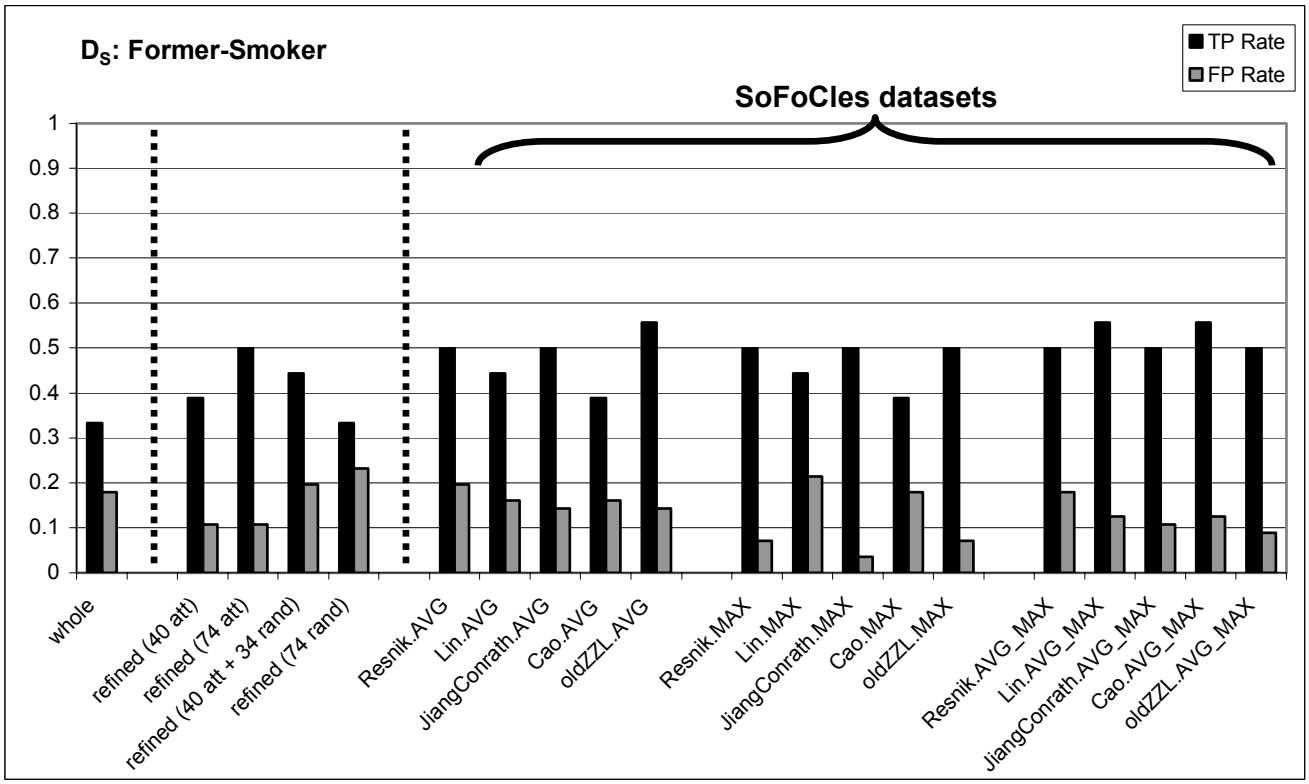
Εικόνα 6.3. Ρυθμοί Πραγματικά Θετικών (TP Rates), Precisions και F-Measures της ετικέτας Current-Smoker για το D_S σύνολο δεδομένων



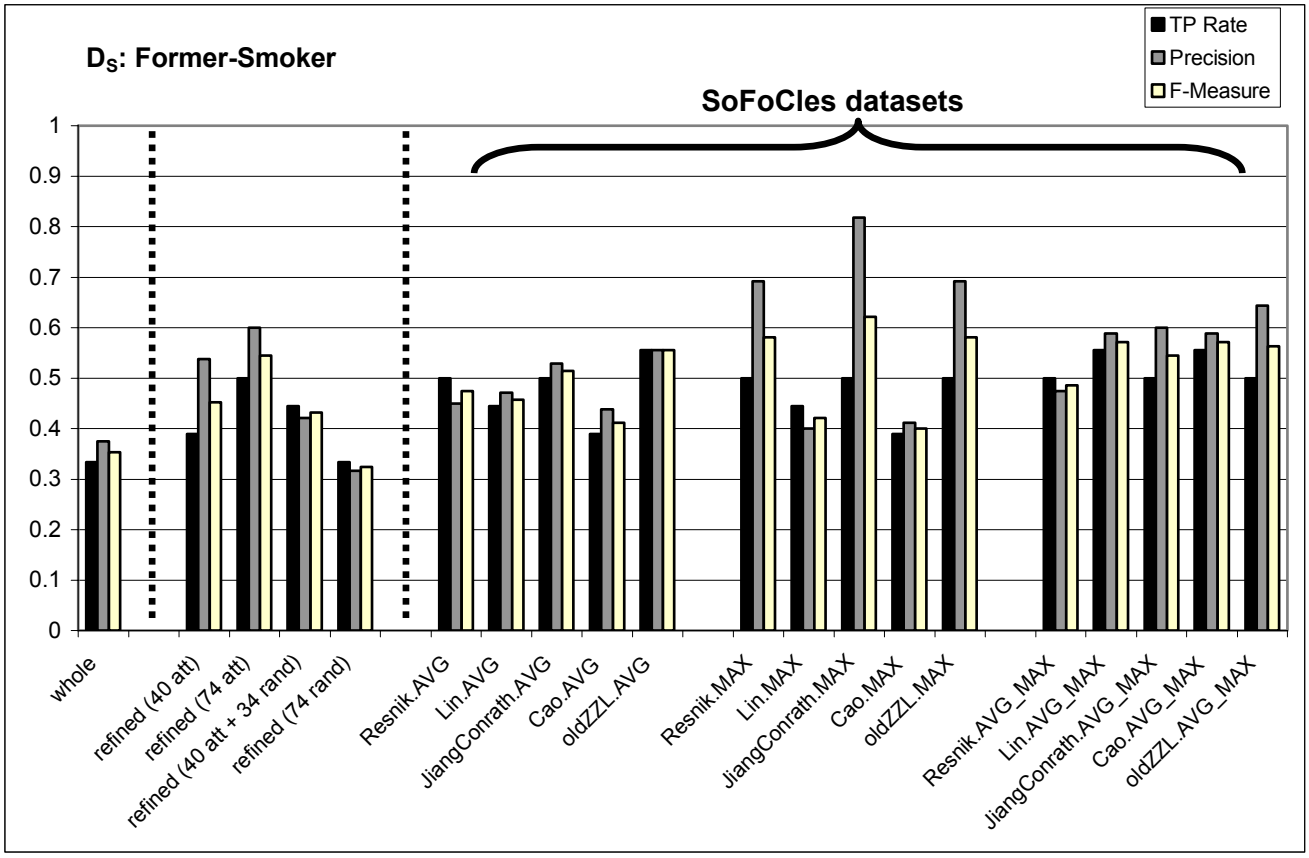
Εικόνα 6.4. Ρυθμοί Πραγματικά Θετικών (TP Rates) και Λανθασμένα Θετικών (FP Rates) της ετικέτας Never-Smoked για το D_S σύνολο δεδομένων



Εικόνα 6.5. Ρυθμοί Πραγματικά Θετικών (TP Rates), Precisions και F-Measures της ετικέτας Never-Smoked για το D_S σύνολο δεδομένων



Εικόνα 6.6. Ρυθμοί Πραγματικά Θετικών (TP Rates) και Λανθασμένα Θετικών (FP Rates) της ετικέτας Former-Smoker για το D_S σύνολο δεδομένων



Εικόνα 6.7. Ρυθμοί Πραγματικά Θετικών (TP Rates), Precisions και F-Measures της ετικέτας Former-Smoker για το D_S σύνολο δεδομένων

6.2.2 Ανάλυση των αποτελεσμάτων του D_S συνόλου δεδομένων

Τα σύνολα δεδομένων Resnik.MAX και JiangConrath.MAX εκτός του ότι παρουσιάζουν μεγαλύτερη ακρίβεια ταξινόμησης, μας δίνουν μεγαλύτερους ρυθμούς Πραγματικά Θετικών (TP Rates) και μικρότερους ρυθμούς Λανθασμένα Θετικών (FP Rates) για την ετικέτα Current-Smoker, αντίστοιχα. Επίσης, τέσσερα επιπλέον σύνολα δεδομένων που προκύπτουν από το SoFoCles εμφανίζουν ίσους ή μεγαλύτερους ρυθμούς Πραγματικά Θετικών από τα αντίστοιχα φιλτραρισμένα σύνολα γνωρισμάτων των 40 και 74 γνωρισμάτων (για την ετικέτα Current-Smoker). Είναι πολύ σημαντικό να μπορούμε με μεγάλη βεβαιότητα να αποφανθούμε για ένα καπνιστή ώστε να είναι δυνατή η χορήγηση της κατάλληλης αγωγής.

Από την άλλη, επιθυμούμε ο ρυθμός Πραγματικά Αρνητικών για ένα καπνιστή (FP Rate) να είναι όσο το δυνατόν μικρότερος ώστε να μην θεωρηθεί εσφαλμένα ότι

συγκεκριμένα επίπεδα γονιδιακής έκφρασης σχετίζονται με το κάπνισμα ενώ στην πραγματικότητα δεν συμβαίνει αυτό.

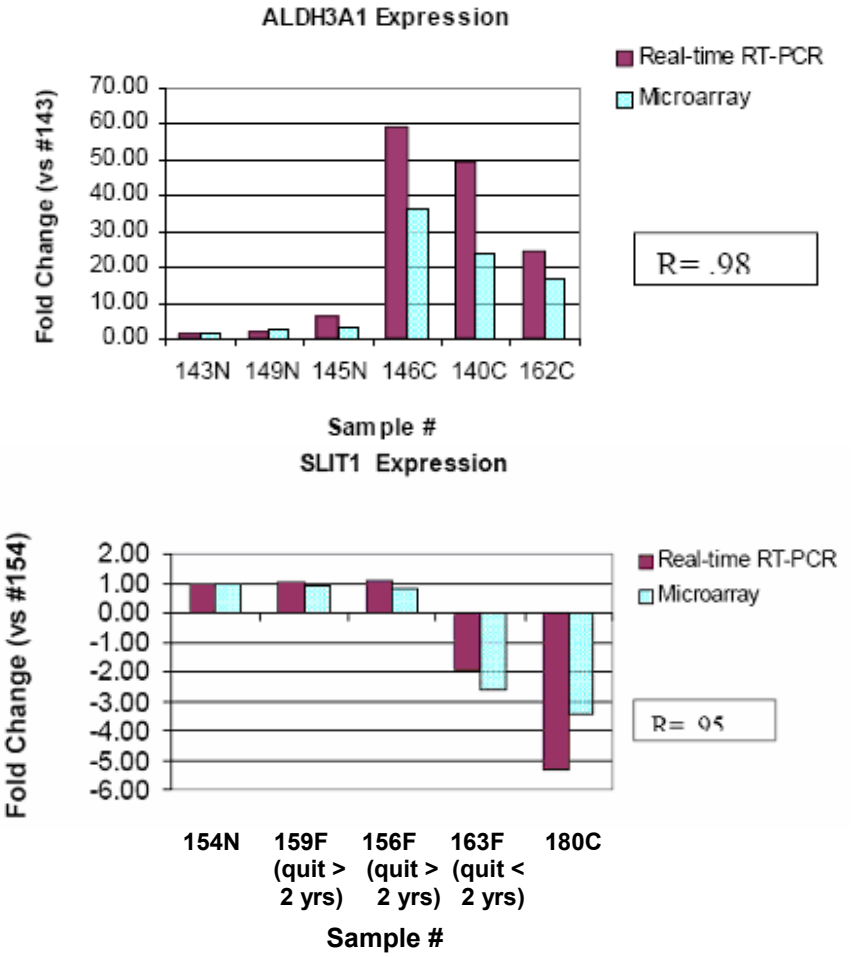
Επιπλέον, η πιθανότητα μία ταξινόμηση για ένα καπνιστή να είναι σωστή είναι μεγαλύτερη στα σύνολα JiangConrath.MAX, oldZZL.MAX και Cao.AVG_MAX από τα αντίστοιχα φιλτραρισμένα αφού το precision είναι μεγαλύτερο.

Όσον αφορά τους μη-καπνιστές (Never-Smoked), ο ρυθμός των Πραγματικά Θετικών (TP Rate) είναι ήδη πολύ μικρός από τα φιλτραρισμένα σύνολα δεδομένων των 40 και 74 γνωρισμάτων. Παρόλα αυτά, ο ρυθμός αυτός μεγαλώνει ακόμη περισσότερο στα σύνολα oldZZL.MAX και JiangConrath.MAX και μάλιστα στο τελευταίο φτάνει στη μέγιστη τιμή του (1). Μ' άλλα λόγια, το σύνολο JiangConrath.MAX μπορούμε να πούμε ότι περιέχει όλες τις απαραίτητες πληροφορίες για να ταξινομήσει σωστά έναν μη-καπνιστή. Επίσης, οι ρυθμοί των Λανθασμένα Θετικών (FP Rates) διατηρούνται σε χαμηλά επίπεδα με τιμές κοντά σε αυτά των φιλτραρισμένων συνόλων (Lin.AVG, oldZZL.AVG, Resnik.MAX, JiangConrath.MAX, Resnik.AVG_MAX, JiangConrath.AVG_MAX και oldZZL.AVG_MAX). Και σε αυτήν την περίπτωση είναι πολύ σημαντικό να χαρακτηρίζουμε με πολύ μεγάλη βεβαιότητα έναν μη-καπνιστή αφού έτσι ελαχιστοποιούμε την πιθανότητα το παράδειγμά μας να ανήκει στην πραγματικότητα σε άλλη ετικέτα (π.χ. καπνιστής) με συνέπεια την μη χορήγηση κατάλληλης αγωγής.

Ωστόσο, στην περίπτωση πρώην-καπνιστών (Former-Smoker) η κατάσταση περιπλέκεται κάπως. Οι ρυθμοί Πραγματικά Θετικών (TP Rates) παρόλο που είναι υψηλότεροι του φιλτραρισμένου συνόλου των 40 γνωρισμάτων (σε πολλά από τα σύνολα του SoFoCles) δεν είναι ικανοποιητικοί (σχεδόν σε όλα τα σύνολα δεδομένων) ενώ πέφτουν στα επίπεδα του φιλτραρισμένου συνόλου των 40 γνωρισμάτων μόνο σε δύο περιπτώσεις (Cao.AVG και Cao.MAX). Επίσης, οι ρυθμοί των Λανθασμένα Θετικών (FP Rates) παραμένουν υψηλοί. Βέβαια, σε τέσσερις περιπτώσεις (Resnik.MAX, JiangConrath.MAX, oldZZL.MAX και oldZZL.AVG_MAX) τα σύνολα δεδομένων που παράγονται από το εργαλείο του SoFoCles παρουσιάζουν μικρότερους ρυθμούς από τους αντίστοιχους των φιλτραρισμένων συνόλων (των 40 και 74 γνωρισμάτων). Μάλιστα, στην περίπτωση του συνόλου JiangConrath.MAX ο ρυθμός αυτός μειώνεται στο ένα τρίτο (1/3) σε σχέση με το ρυθμό Λανθασμένα Θετικών (FP Rate) του φιλτραρισμένου συνόλου των 40 γνωρισμάτων. Επίσης, για το ίδιο σύνολο (JiangConrath.MAX) η πιθανότητα η πρόβλεψη ενός πρώην-καπνιστή να είναι σωστή είναι αρκετά μεγάλη, αφού το precision είναι αρκετά υψηλό (0.818).

Εξάλλου, το γεγονός των μη ικανοποιητικών TP Rates μπορεί να εξηγηθεί και από την ίδια τη φύση των παραδειγμάτων των πρώην-καπνιστών. Όπως αναφέρεται χαρακτηριστικά στη σχετική δημοσίευση ([22]), το γονιδιακό προφίλ των πρώην-καπνιστών που είχαν διακόψει το κάπνισμα για χρονικό διάστημα μικρότερο των δύο (2) χρόνων έμοιαζε περισσότερο με αυτό των καπνιστών, ενώ το προφίλ αυτών που είχαν διακόψει το κάπνισμα περισσότερα από δύο (2) χρόνια έμοιαζε περισσότερο με αυτό των μη-καπνιστών. Συγκεκριμένα, στα σύνολα με τα οποία πειραματιστήκαμε, ο αριθμός των πρώην-καπνιστών που ταξινομήθηκαν ως καπνιστές ήταν πάντα μεγαλύτερος από τον αριθμό των πρώην-καπνιστών που ταξινομήθηκαν ως μη-καπνιστές.

Πιο συγκεκριμένα, τα επίπεδα έκφρασης των γονιδίων που ενεργοποιούνται από το κάπνισμα προσομοίαζαν αυτά των μη-καπνιστών μετά από δύο έτη αποχής από το κάπνισμα. Αυτά τα γονίδια, που επανέρχονταν στα φυσιολογικά επίπεδα, φαίνεται να εξυπηρετούσαν αντι-οξειδωτικές και μεταβολικές διαδικασίες. Επί παραδείγματι, γονίδια των οποίων η λειτουργία επηρεάζεται από το κάπνισμα είναι τα: PIR (pirin) που συμμετέχει στο μεταβολισμό νουκλεοτιδίων και νουκλεϊκών οξέων, GPX2 (Glutathione peroxidase 2) και ALDH3A1 (Aldehyde dehydrogenase 3) που συμμετέχουν σε μεταβολικές διαδικασίες, CLDN10 (Claudin 10) που συμβάλλει στην κυτταρική ανάπτυξη και συντήρηση, SLIT1 (Slit 1), SLIT2 (Slit 2) και CX3CL1 Chemokine (CX3C motif ligand 1) που συμμετέχουν στην κυτταρική επικοινωνία καθώς και το MMP10 (Matrix metalloproteinase 10) που συμβάλλει στο μεταβολισμό των πρωτεϊνών. Ένα παράδειγμα δύο τέτοιων γονιδίων δίνεται στην Εικ. 6.8.



Εικόνα 6.8. Μεταβολή της έκφρασης γονιδίων σε σχέση με την ετικέτα του ασθενούς (μη-καπνιστής, καπνιστής, πρώην-καπνιστής) για το D_S σύνολο δεδομένων. Όπως παρατηρούμε η έκφραση του γονιδίου ALDH3A1παραμένει χαμηλή και σε παρόμοια επίπεδα στους μη-καπνιστές (143N, 149N, 145N), ενώ αυξάνεται σημαντικά στους καπνιστές (146C, 140C, 162C). Όσον αφορά, το γονίδιο SLIT1, τα επίπεδα έκφρασής του μειώνονται σημαντικά στους καπνιστές. Αξιοσημείωτο είναι επίσης, το παρόμοιο γονιδιακό προφίλ των πρώην-καπνιστών που έχουν διακόψει το κάπνισμα για χρονικό διάστημα μεγαλύτερο των δύο ετών (159F, 156F) με αυτό των μη-καπνιστών (154N) και των πρώην-καπνιστών που έχουν διακόψει το κάπνισμα για χρονικό διάστημα μικρότερο των δύο ετών (163F) με αυτό των μη-καπνιστών (180C).
Η παραπάνω εικόνα αποτελεί τροποποίηση της αρχικής και έχει ληφθεί από την τοποθεσία [xix]

Από την άλλη, όμως, υπήρχαν συγκεκριμένα ογκογονίδια και γονίδια καταστολής όγκων (ογκοκατασταλτικά γονίδια) που αποτύγχαναν να επανέλθουν στη φυσιολογική τους κατάσταση κάτι που ίσως εξηγεί και τους αυξημένους κινδύνους ανάπτυξης καρκίνου του πνεύμονα έστω και μετά από μεγάλο διάστημα διακοπής του καπνίσματος. Αυτή η περίπλοκη κατάσταση του γονιδιακού προφίλ στους πρώην-καπνιστές ίσως να καθιστά και την προσπάθεια ταξινόμησής τους ιδιαίτερα δύσκολη διαδικασία.

Επιχειρώντας να εντοπίσουμε τα βιολογικά αίτια των καλύτερων αποτελεσμάτων που λαμβάνουμε έχουμε να σημειώσουμε τα εξής. Κατ' αρχήν, η

εμφανής βελτίωση της ακρίβειας ταξινόμησης (περίπου κατά 14%) στα φιλτραρισμένα σύνολα δεδομένων (των 40 και 74 γνωρισμάτων) σε σχέση με το πλήρες ίσως δικαιολογείται από το γεγονός ότι αρκετά από τα γονίδια που βρέθηκαν πειραματικά να σχετίζονται με τον καρκίνο του πνεύμονα περιλαμβάνονταν ως γνωρίσματα στα φιλτραρισμένα σύνολα.

Πιο συγκεκριμένα, στο [22], αναφέρεται ότι τα αντι-οξειδωτικά γονίδια GPX2 (3-πτυχο), ALDH3A1 (6-πτυχο) και CLDN10 (3-πτυχο) εμφανίζουν (στους καπνιστές) μεγάλες αυξήσεις στη γονιδιακή τους έκφραση. Στο φιλτραρισμένο σύνολο (των 40 γνωρισμάτων) περιλαμβάνονται τα γνωρίσματα NM_002083, NM_000691 και NM_006984 τα οποία αποτελούν GeneBank accession numbers και αντιστοιχούν κατά σειρά στα γονίδια GPX2, ALDH3A1 και CLDN10.

Παρατηρείται επίσης αύξηση της έκφρασης του ογκογονιδίου PIR, το οποίο αντιστοιχεί στο GeneBank accession number: NM_003662 και περιέχεται στα φιλτραρισμένα σύνολα δεδομένων (των 40 και 74 γνωρισμάτων).

Επίσης, τα επίπεδα έκφρασης των ογκοκατασταλτικών γονιδίων MMP10, SLIT1 και SLIT2 που αντιστοιχούν κατά σειρά στα GeneBank accession numbers: NM_002425, AB011537, AF055585 μειώνονται. Το ίδιο συμβαίνει και για το ογκοκατασταλτικό γονίδιο TU3A που αντιστοιχεί στα GeneBank accession numbers AL050264 και NM_007177. Τα παραπάνω γονίδια, που σχετίζονται με την καρκινογένεση εμπεριέχονται στα φιλτραρισμένα σύνολα δεδομένων.

Επίσης, από το Υλικό Υποστήριξης (Supporting Information) του [22] και συγκεκριμένα στον Πιν. 9 (http://www.pnas.org/cgi/data/0401422101/DC1/13) όπου αναφέρονται κατά φθίνουσα σειρά τιμής P τα γονίδια που υπέστησαν ανεπανόρθωτες αλλαγές λόγω καπνίσματος εντοπίστηκαν γονίδια που περιλαμβάνονται στα φιλτραρισμένα σύνολα δεδομένων (των 40 και 74 γνωρισμάτων). Αυτά είναι τα: LOC92689 (υποθετική πρωτεΐνη BC001096) που αντιστοιχεί στο GeneBank accession number W87466, ME1 (Malic enzyme 1) που συμμετέχει σε μεταβολικές διαδικασίες και αντιστοιχεί στο GeneBank accession number NM_002395, MT1F (Metallothionein 1F) που αναλαμβάνει μεταφορικές διεργασίες και αντιστοιχεί στο GeneBank accession number BF246115 και τέλος το EPAS (MOP2) που συμμετέχει στο μεταβολισμό νουκλεοτιδίων και νουκλεϊκών οξέων και αντιστοιχεί στα GeneBank accession numbers AF052094 και NM_001430.

Το SoFoCles δημιουργεί σύνολα δεδομένων που περιέχουν όλα τα παραπάνω γονίδια καθώς και άλλα (γονίδια) που ίσως να σχετίζονται με την ανάπτυξη του καρκίνου του πνεύμονα. Για παράδειγμα, τα σύνολα Cao.AVG, Cao.AVG_MAX, Cao.MAX, Lin.AVG, Lin.AVG_MAX και Lin.MAX ενσωματώνουν τον GeneBank

accession number NM_0144434 που αντιστοιχεί στο γονίδιο NDOR1 και το οποίο εμφανίζει υψηλή σημασιολογική ομοιότητα με το αντι-οξειδωτικό ALDH3A1. Όπως φαίνεται από τον Πιν. Χ, τα σύνολα Cao.AVG_MAX και Lin.AVG_MAX εμφανίζουν το μεγαλύτερο ρυθμό Πραγματικά Θετικών (TP Rate) για τους πρώην-καπνιστές και από τους μεγαλύτερους ρυθμούς Πραγματικά Θετικών (TP Rates) για τους καπνιστές και τους μη-καπνιστές. Με κάποια επιφύλαξη μπορούμε να πούμε ότι η εισαγωγή αυτού του γονιδίου συνεισφέρει στη διαδικασία εκμάθησης.

Επίσης, τα σύνολα Resnik.AVG, Lin.AVG, JiangConrath.AVG, Cao.AVG και oldZZL.AVG ενσωματώνουν τον GeneBank accession number U58856 που αντιστοιχεί στο γονίδιο MRC2 και το οποίο εμφανίζει υψηλή σημασιολογική ομοιότητα με το αντι-οξειδωτικό ALDH3A1. Παρόλα αυτά, δεν μπορούμε να συνάγουμε με βεβαιότητα τη χρησιμότητα αυτού του γονιδίου.

Εξάλλου, στα σύνολα δεδομένων Cao.AVG_MAX, JiangConrath.AVG_MAX και oldZZL.AVG_MAX περιλαμβάνεται ο GeneBank accession number NM_002996 που αντιστοιχεί στο ογκοκατασταλτικό γονίδιο CX3CL1. Το γονίδιο αυτό, όπως περιγράφεται και από το [22], εμφανίζει μειωμένα επίπεδα έκφρασης στους καπνιστές. Αξιοσημείωτο είναι επίσης ότι αυτό το γονίδιο δεν ενσωματώθηκε στα φιλτραρισμένα σύνολα δεδομένων μετά την εφαρμογή μεθόδων επιλογής γνωρισμάτων. Με άλλα λόγια, το γεγονός ότι συναντάται αποκλειστικά σε σύνολα που έχουν προκύψει από το εργαλείο SoFoCles να δικαιολογεί και τις μεγάλες τιμές της ακρίβειας ταξινόμησης (για αυτά τα σύνολα).

Τέλος, η παρουσία του γονιδίου SLIT1 ενισχύεται με την ενσωμάτωση του GeneBank accession number: NM_003061 στο σύνολο Resnik.MAX που εμφανίζει μεγάλη τιμή ακρίβειας ταξινόμησης.

6.3 D_{CNV} σύνολο δεδομένων

6.3.1 Παρουσίαση του D_{CNV} συνόλου δεδομένων και των αποτελεσμάτων του

Το D_{CNV} σύνολο εξετάζει το αποτέλεσμα της ανάπτυξης εμβρυϊκών όγκων στο Κεντρικό Νευρικό Σύστημα μετά από θεραπεία. Το σύνολο δεδομένων αποτελείται από 7130 (7129 γονίδια–γνωρίσματα και μία κλάση). Η κλάση χωρίζεται σε δύο

ετικέτες: {Class1 (επιζώντες), Class0 (αποτυχία)}. Το σύνολο περιέχει 60 παραδείγματα (άτομα), εκ των οποίων 21 ανήκουν στην Class1 και 39 στην Class0 [17].

Για την αντιστοίχηση probe_ids σε γονιδιακά σύμβολα χρησιμοποιήθηκαν τα αρχεία NetAffx HG-U6800, η επεξεργασία των οποίων έγινε με το πρόγραμμα R. Για τα πειράματά μας χρησιμοποιήσαμε το αρχικό σύνολο δεδομένων και μία μικρότερη έκδοσή του μετά την εφαρμογή της Information Gain (IG) μεθόδου φιλτραρίσματος γνωρισμάτων. Το μικρότερο σύνολο δεδομένων αποτελείται από 20 γνωρίσματα. Χρησιμοποιήθηκαν επίσης τα εξής GOA αρχεία: του ανθρώπου (gene_association.goa_human), του ποντικού (gene_association.mgi) και του αρουραίου (gene_association.rgd). Τα δύο τελευταία χρησιμοποιήθηκαν εξαιτίας της μεγάλης ομοιότητας που παρουσιάζουν οι δύο αυτοί οργανισμοί με τον άνθρωπο. Επιπλέον, έγινε χρήση του πιο πρόσφατου αρχείου Γονιδιακής Οντολογίας (GO file) αυτή του Μαΐου 2007. Από την αναζήτηση αποκλείστηκαν οι όροι που ήταν βασισμένοι σε IEA τύπους αποδείξεων εξαιτίας της μικρής αξιοπιστίας αυτών των επισημειώσεων, ενώ η αναζήτηση, που βασίστηκε στο ταίριασμα μεταξύ IDs, συμβόλων ή συνωνύμων, επιλέχτηκε να είναι μερική (και όχι ακριβής) για να είναι πιο ευέλικτη.

Εξάλλου, ο υπολογισμός της σημασιολογικής ομοιότητας μεταξύ των γονιδίων στηρίχτηκε μόνο σε όρους από την BP οντολογία καθώς αυτοί οι όροι περιγράφουν πλήρεις βιολογικές διαδρομές (διαδικασίες) παρά μεμονωμένες λειτουργίες. Τέλος, επιλέχτηκε ως ελάχιστο κατώφλι σημασιολογικής ομοιότητας το 50%, ενώ ο μέγιστος αριθμός των πιο συσχετισμένων γονιδίων για κάθε γονίδιο του συνόλου δεδομένων (μετά την εφαρμογή του φιλτραρίσματος γνωρισμάτων) επιλέχτηκε να είναι το 5 και ο μέγιστος αριθμός των τελικά προκρινόμενων γονιδίων επιλέχτηκε να είναι ο 10 (ο μισός του αριθμού γνωρισμάτων του μικρότερου συνόλου δεδομένων).

Πειραματιστήκαμε μεταξύ διαφόρων μεθόδων σημασιολογικής ομοιότητας (Resnik, Lin, Jiang & Conrath, Cao, oldZZL) και τρόπων υπολογισμού της ομοιότητας μεταξύ δύο γονιδίων (AVG, MAX, AVG_MAX) δημιουργώντας έτσι $5 \cdot 3 = 15$ σύνολα δεδομένων.

Για μία πιο ολοκληρωμένη προσπάθεια αποτίμησης των αποτελεσμάτων δημιουργήσαμε 3 επιπλέον σύνολα δεδομένων με αριθμό γνωρισμάτων ίσο με το μέσο όρο του αριθμού γνωρισμάτων των συνόλων δεδομένων που προκύπτουν μετά την εφαρμογή του εργαλείου SoFoCles. Το πρώτο αποτελείται από 34 γνωρίσματα μετά την εφαρμογή της μεθόδου Information Gain (IG) (όπως είχε γίνει και για το σύνολο των 20 γνωρισμάτων), το δεύτερο αποτελείται από τα 20 γνωρίσματα του

αρχικού μικρού συνόλου συν 14 τυχαίως εξαγόμενα γνωρίσματα από το αρχικό πλήρες σύνολο και το τρίτο από 34 τυχαίως εξαγόμενα γνωρίσματα από το αρχικό πλήρες σύνολο.

Τα αποτελέσματά μας συγκεντρώνονται στους παρακάτω πίνακες:

Πίνακας XIII. Παρουσίαση των ποιοτικών χαρακτηριστικών για το D_{CNV} σύνολο δεδομένων

	Total Accuracy	**Class1**			
		TP Rate	**FP Rate**	**Precision**	**F-Measure**
whole	68.33	0.476	0.205	0.556	0.513
refined (20 att)	81.67	0.762	0.154	0.727	0.744
refined (34 att)	80.00	0.762	0.179	0.696	0.727
refined (20 att + 14 rand)	81.67	0.714	0.128	0.750	0.732
refined (34 rand)	63.33	0.048	0.051	0.333	0.083
Resnik.AVG	76.67	0.667	0.179	0.667	0.667
Lin.AVG	75.00	0.619	0.179	0.650	0.634
JiangConrath.AVG	80.00	0.714	0.154	0.714	0.714
Cao.AVG	85.00	0.762	0.103	0.800	0.780
oldZZL.AVG	83.33	0.714	0.103	0.789	0.750
Resnik.MAX	88.33	0.810	0.077	0.850	0.829
Lin.MAX	86.67	0.810	0.103	0.810	0.810
JiangConrath.MAX	81.67	0.762	0.154	0.727	0.744
Cao.MAX	81.67	0.762	0.154	0.727	0.744
oldZZL.MAX	85.00	0.810	0.128	0.773	0.791
Resnik.AVG_MAX	75.00	0.619	0.179	0.650	0.634
Lin.AVG_MAX	91.67	0.952	0.103	0.833	0.889
JiangConrath.AVG_MAX	91.67	0.952	0.103	0.833	0.889
Cao.AVG_MAX	91.67	0.952	0.103	0.833	0.889
oldZZL.AVG_MAX	91.67	0.952	0.103	0.833	0.889

Class0			
TP Rate	**FP Rate**	**Precision**	**F-Measure**
0.795	0.524	0.738	0.765
0.846	0.238	0.868	0.857
0.821	0.238	0.865	0.842
0.872	0.286	0.850	0.861
0.949	0.952	0.649	0.771
0.821	0.333	0.821	0.821
0.821	0.381	0.800	0.810
0.846	0.286	0.846	0.846
0.897	0.238	0.875	0.886
0.897	0.286	0.854	0.875
0.923	0.190	0.900	0.911
0.897	0.190	0.897	0.897

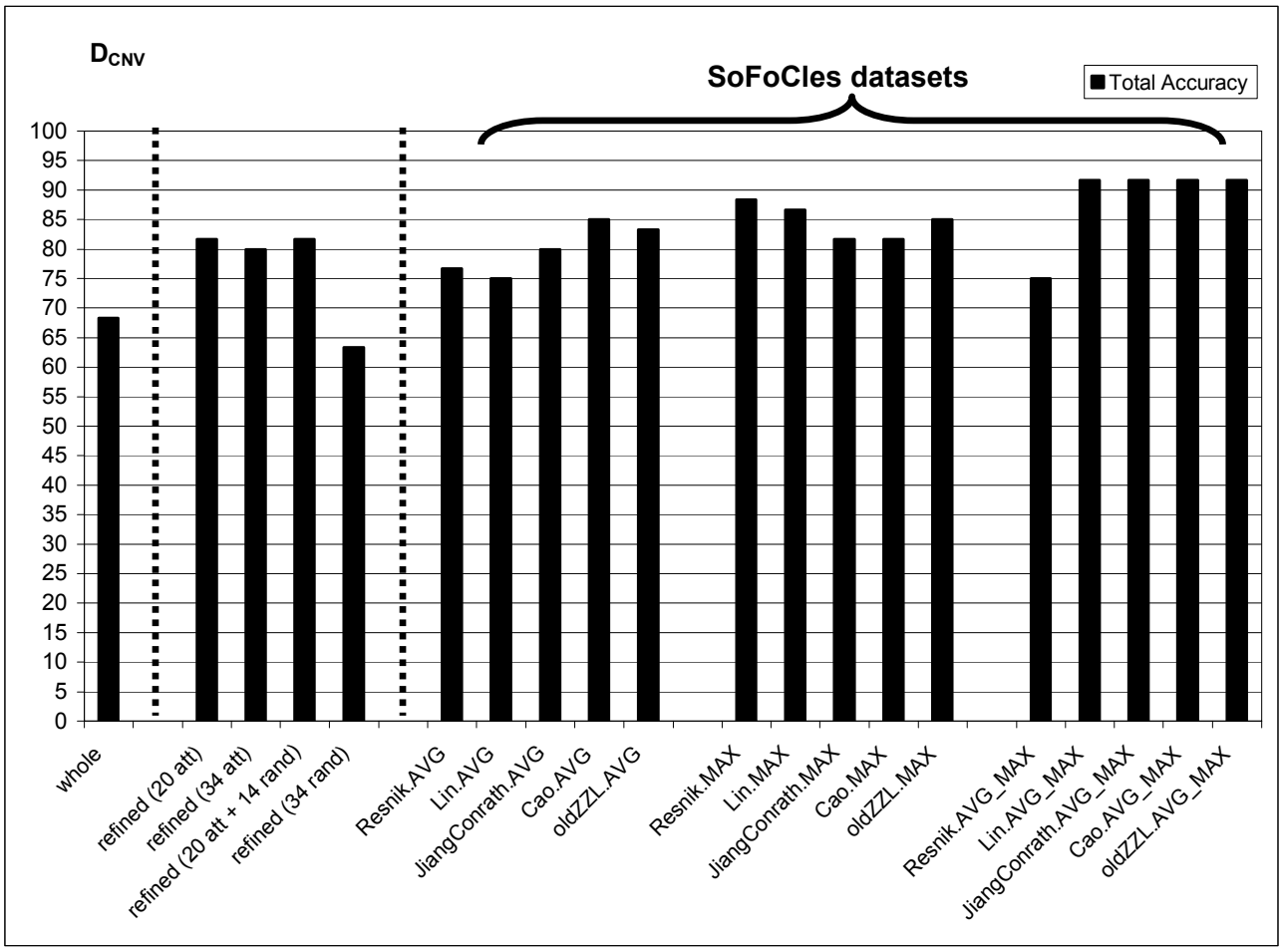
0.846	0.238	0.868	0.857
0.846	0.238	0.868	0.857
0.872	0.190	0.895	0.883
0.821	0.381	0.800	0.810
0.897	0.048	0.972	0.933
0.897	0.048	0.972	0.933
0.897	0.048	0.972	0.933
0.897	0.048	0.972	0.933

Πίνακας XIV. Confusion-Matrix για το D_{CNV} σύνολο δεδομένων

	Class1	**Class0**	**← *classified as***	**# of instances**
whole	10	11	**Class1**	**21**
	8	31	**Class0**	**39**
refined (20 att)	16	5	**Class1**	**21**
	6	33	**Class0**	**39**
refined (34 att)	16	5	**Class1**	**21**
	7	32	**Class0**	**39**
refined (20 att + 14 rand)	15	6	**Class1**	**21**
	5	34	**Class0**	**39**
refined (34 rand)	1	20	**Class1**	**21**
	2	37	**Class0**	**39**
Resnik.AVG	14	7	**Class1**	**21**
	7	32	**Class0**	**39**
Lin.AVG	13	8	**Class1**	**21**
	7	32	**Class0**	**39**
JiangConrath.AVG	15	6	**Class1**	**21**
	6	33	**Class0**	**39**

Cao.AVG	16	5	**Class1**	**21**
	4	35	**Class0**	**39**
oldZZL.AVG	15	6	**Class1**	**21**
	4	35	**Class0**	**39**
Resnik.MAX	17	4	**Class1**	**21**
	3	36	**Class0**	**39**
Lin.MAX	17	4	**Class1**	**21**
	4	35	**Class0**	**39**
JiangConrath.MAX	16	5	**Class1**	**21**
	6	33	**Class0**	**39**
Cao.MAX	16	5	**Class1**	**21**
	6	33	**Class0**	**39**
oldZZL.MAX	17	4	**Class1**	**21**
	5	34	**Class0**	**39**
Resnik.AVG_MAX	13	8	**Class1**	**21**
	7	32	**Class0**	**39**
Lin.AVG_MAX	20	1	**Class1**	**21**
	4	35	**Class0**	**39**
JiangConrath.AVG_MAX	20	1	**Class1**	**21**
	4	35	**Class0**	**39**
Cao.AVG_MAX	20	1	**Class1**	**21**
	4	35	**Class0**	**39**
oldZZL.AVG_MAX	20	1	**Class1**	**21**
	4	35	**Class0**	**39**

Τα αποτελέσματα του Πιν. XIII οπτικοποιούνται στα παρακάτω διαγράμματα:



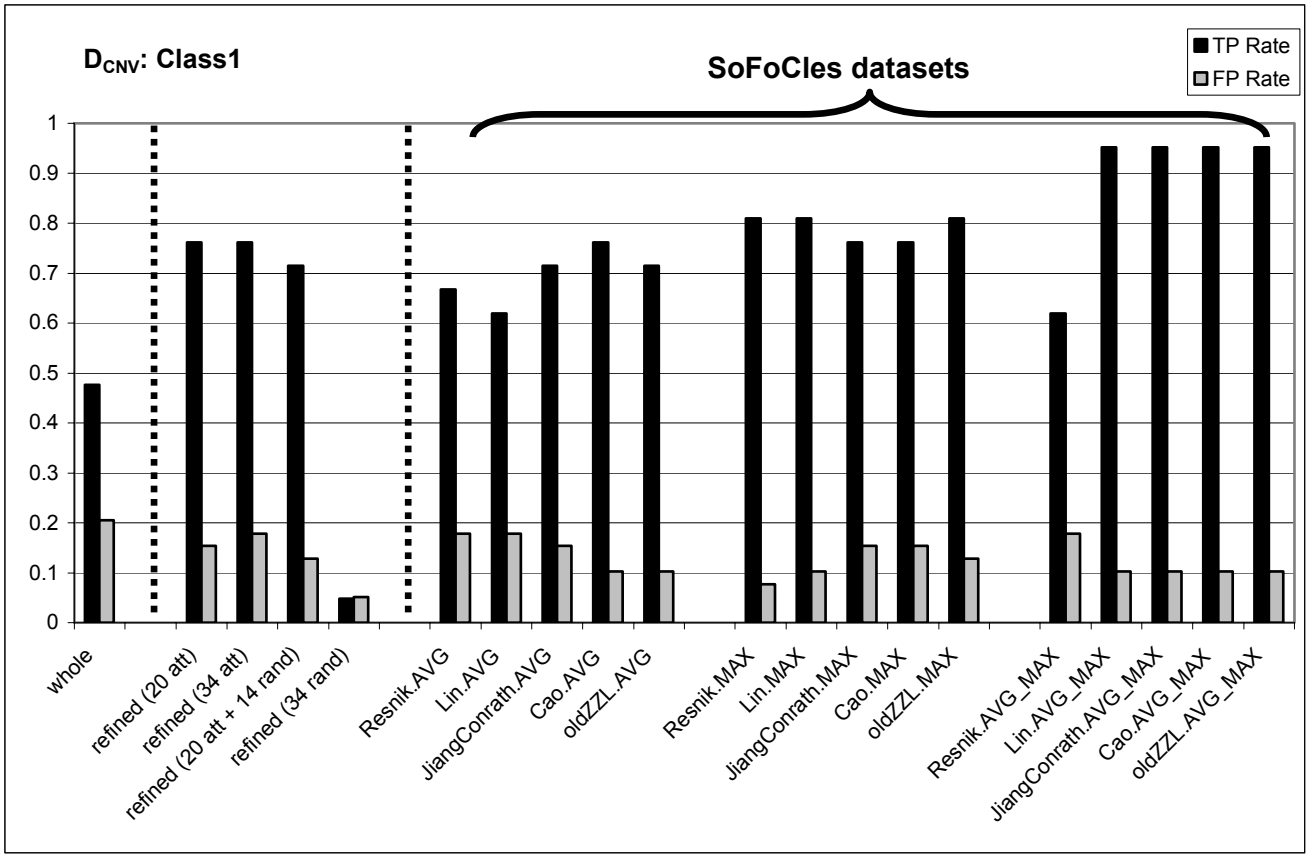
Εικόνα 6.9. Ακρίβεια ταξινόμησης για το D_{CNV} σύνολο δεδομένων

Από το αρχικό φιλτραρισμένο σύνολο δεδομένων (αυτό των 20 γνωρισμάτων) εννιά σύνολα δεδομένων μετά την εφαρμογή του εργαλείου SoFoCles παρουσιάζουν μεγαλύτερη ακρίβεια ταξινόμησης ενώ δύο έχουν ίση ακρίβεια ταξινόμησης. Με άλλα λόγια, σχεδόν τα δύο τρίτα των συνόλων δεδομένων που δημιουργήθηκαν μετά την εφαρμογή του εργαλείου SoFoCles εμφανίζονται να έχουν ίση ή μεγαλύτερη ακρίβεια.

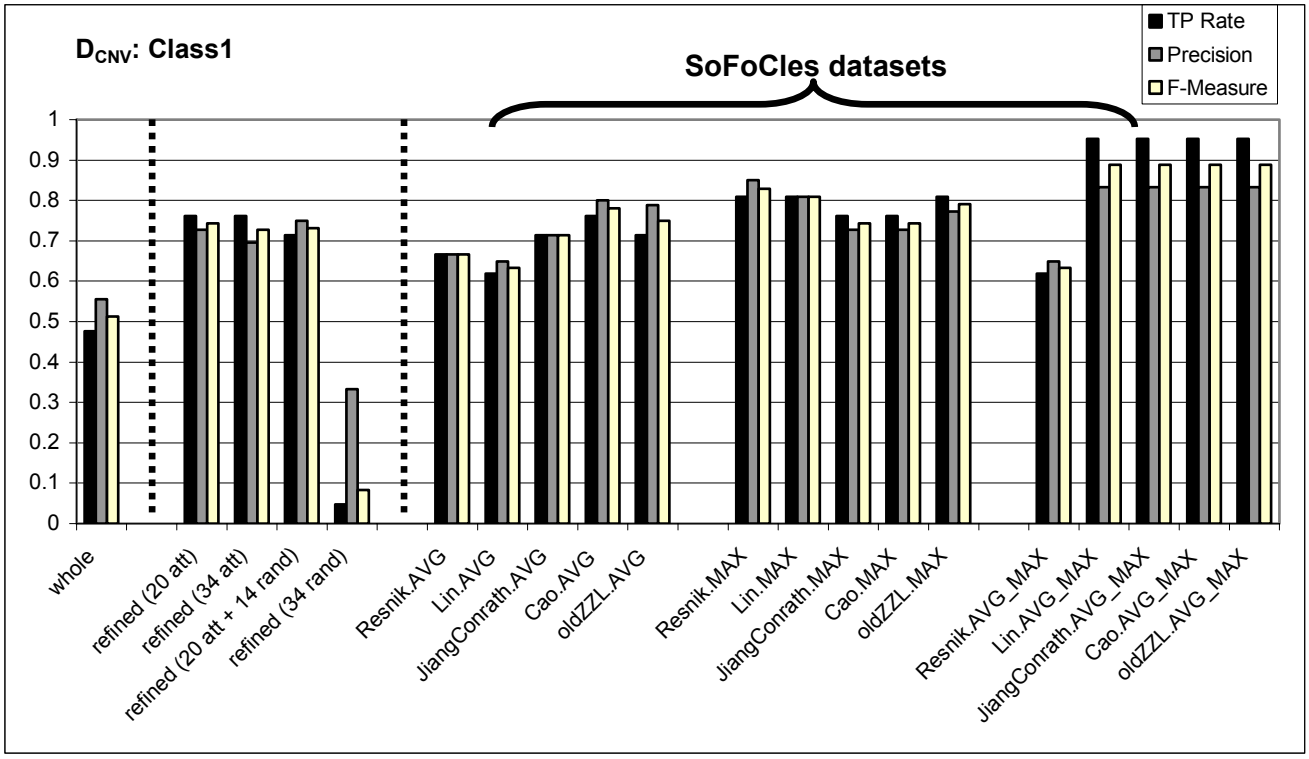
Από το φιλτραρισμένο σύνολο δεδομένων που αριθμεί 34 γνωρίσματα, έντεκα σύνολα δεδομένων του SoFoCles εμφανίζουν μεγαλύτερη ακρίβεια ταξινόμησης.

Το μερικώς τυχαίο σύνολο δεδομένων (20 γνωρίσματα του φιλτραρισμένου συν 14 τυχαία) εμφανίζει ίδια ακρίβεια ταξινόμησης με το αρχικό, η οποία είναι ήδη πολύ μεγάλη (81.67%). Αυτό ίσως σημαίνει ότι τα επιπρόσθετα γνωρίσματα-γονίδια δεν επηρεάζουν το μοντέλο εκμάθησης. Επίσης, όπως αναμενόταν, το αμιγώς τυχαίο σύνολο δεδομένων εμφανίζει πολύ μικρή ακρίβεια ταξινόμησης (63.33%).

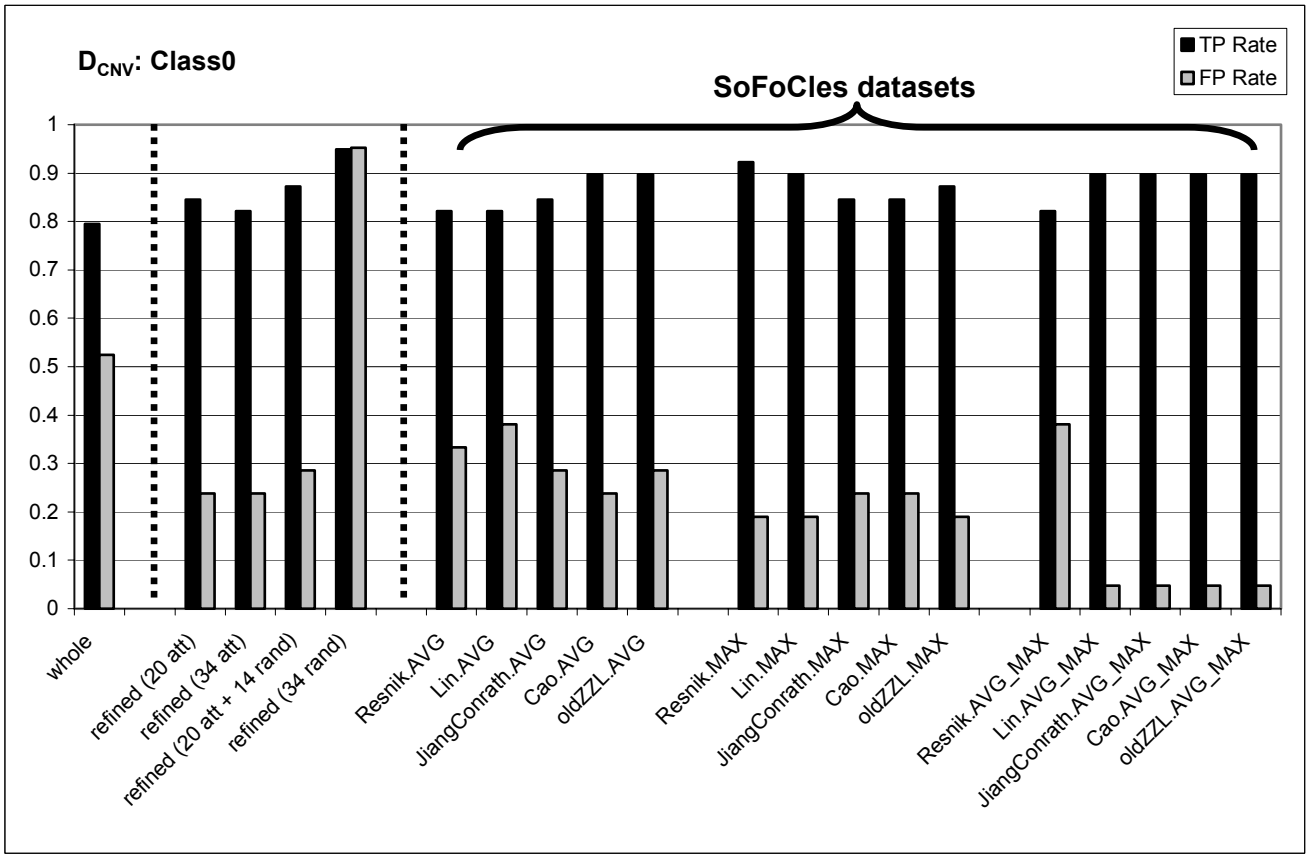
Παρακάτω θα παρουσιάσουμε τα ποιοτικά χαρακτηριστικά για κάθε ετικέτα της κλάσης ξεχωριστά:



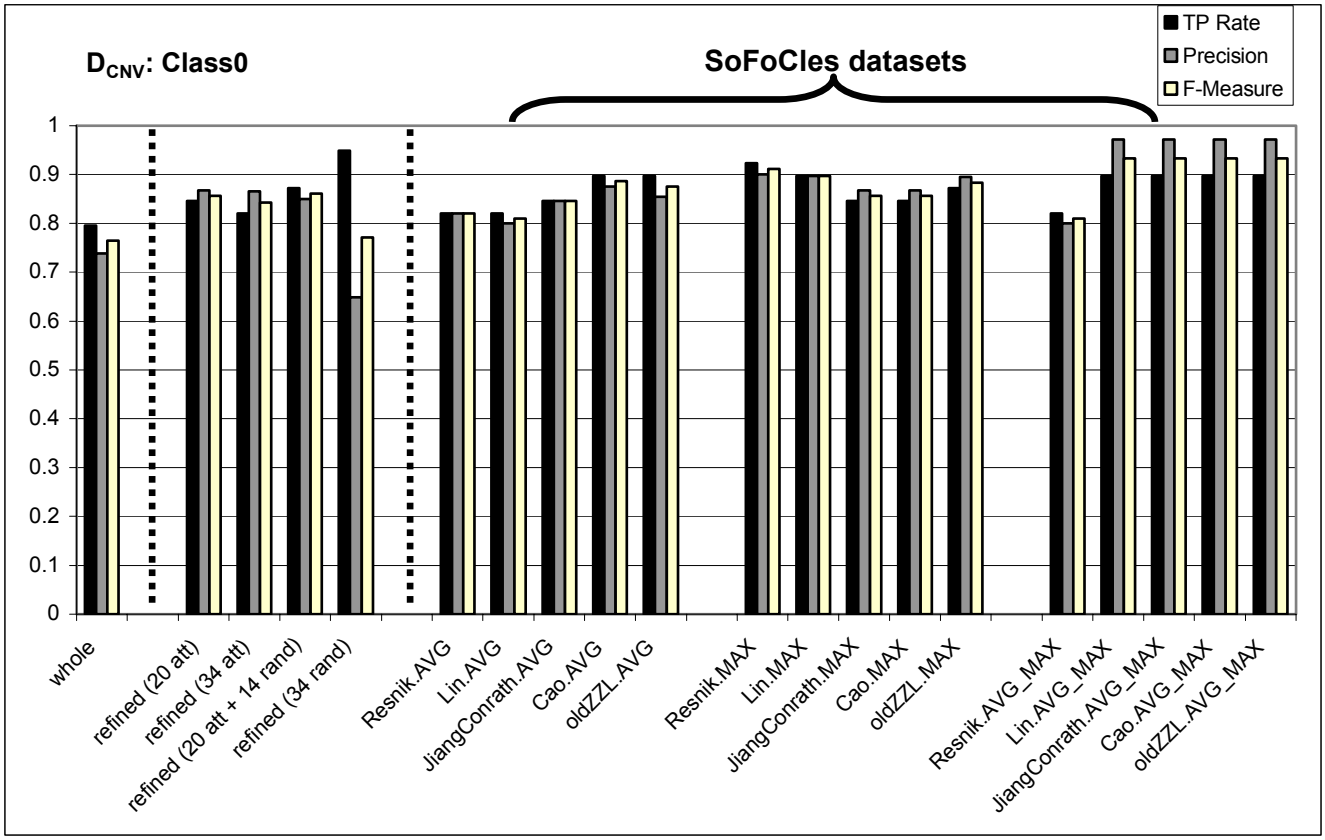
Εικόνα 6.10. Ρυθμοί Πραγματικά Θετικών (TP Rates) και Λανθασμένα Θετικών (FP Rates) της ετικέτας Class1 (επιζώντες) για το D_{CNV} σύνολο δεδομένων



Εικόνα 6.11. Ρυθμοί Πραγματικά Θετικών (TP Rates), Precisions και F-Measures της ετικέτας Class1 (επιζώντες) για το D_{CNV} σύνολο δεδομένων



Εικόνα 6.12. Ρυθμοί Πραγματικά Θετικών (TP Rates) και Λανθασμένα Θετικών (FP Rates) της ετικέτας Class0 (αποτυχία) για το D_{CNV} σύνολο δεδομένων



Εικόνα 6.13. Ρυθμοί Πραγματικά Θετικών (TP Rates), Precisions και F-Measures της ετικέτας Class0 (αποτυχία) για το D_{CNV} σύνολο δεδομένων

6.3.2 Ανάλυση των αποτελεσμάτων του D_{CNV} συνόλου δεδομένων

Για την ετικέτα Class1 (επιζώντες) επτά από τα δεκαπέντε σύνολα δεδομένων που παράγονται από το SoFoCles εμφανίζουν μεγαλύτερους ρυθμούς Πραγματικά Θετικών (TP rates) ενώ εννιά από (από τα δεκαπέντε σύνολα δεδομένων) παρουσιάζουν μικρότερους ρυθμούς Λανθασμένα Θετικών (FP rates) σε σχέση με το αρχικό φιλτραρισμένο σύνολο (των 20 γνωρισμάτων). Τα σύνολα Lin.AVG_MAX, JiangConrath.AVG_MAX, Cao.AVG_MAX και oldZZL.AVG_MAX εντοπίζουν με απόλυτη σχεδόν ακρίβεια τους επιζώντες (μόνο ένας ταξινομείται ως αποθανών (Class0)). Επίσης, οι ρυθμοί Λανθασμένα Θετικών (FP rates) μειώνονται περίπου κατά 33% με μέγιστη τιμή το 50% (Resnik.MAX). Είναι πολύ σημαντικό να μην ταξινομείται ένας ασθενής ως επιζώντας ενώ στην πραγματικότητα δεν είναι ή δεν

πρόκειται να επιβιώσει. Με αυτό τον τρόπο ελαχιστοποιείται η πιθανότητα μη χορήγησης κατάλληλης αγωγής σε κάποιον που χρειάζεται θεραπεία.

Από την άλλη, εννιά από τα δεκαπέντε σύνολα δεδομένων εμφανίζουν μεγαλύτερους ρυθμούς Πραγματικά Θετικών (TP rates) και μικρότερους ρυθμούς Λανθασμένα Θετικών (FP rates) όσον αφορά την ετικέτα Class0. Το μεγαλύτερο TP rate το παρουσιάζει το σύνολο Resnik.MAX (0.923) κάτι που είναι πολύ σημαντικό αφού έτσι διασφαλίζεται η μεγάλη πιθανότητα εντοπισμού όλων των καρκινοπαθών. Τέλος, σχεδόν μηδαμινούς ρυθμούς Λανθασμένα Θετικών (FP rates) παρουσιάζουν τα σύνολα Lin.AVG_MAX, JiangConrath.AVG_MAX, Cao.AVG_MAX και oldZZL.AVG_MAX, αφού μόλις ένας ασθενής ταξινομείται λανθασμένα ως καρκινοπαθής.

Μπορούμε να πούμε ότι τα σύνολα Resnik.MAX, Lin.AVG_MAX, JiangConrath.AVG_MAX, Cao.AVG_MAX και oldZZL.AVG_MAX βελτιώνουν κατά πολύ την ποιότητα των αποτελεσμάτων σε σχέση με αυτά των φιλτραρισμένων συνόλων (των 20 και 34 γνωρισμάτων).

Επιχειρώντας να εξηγήσουμε τα αποτελέσματα θα προσπαθήσουμε να εντρυφήσουμε στη βιολογική φύση των γνωρισμάτων. Κατ' αρχήν, η αύξηση της ακρίβειας ταξινόμησης στα φιλτραρισμένα σύνολα δεδομένων κατά 13% ως προς το αρχικό ίσως να οφείλεται στην ενσωμάτωση γονιδίων που σχετίζονται με το συγκεκριμένο πρόβλημα.

Αναλυτικότερα, στο [17] αναγράφονται τα γονίδια που έχουν τη μεγαλύτερη διαχωριστική ισχύ. Μερικά από αυτά εμπεριέχονται στα φιλτραρισμένα σύνολα δεδομένων. Πιο συγκεκριμένα, πρόκειται για τα γονίδια: LOC440345 που αντιστοιχεί στο γνώρισμα D86974_at, AP3B2 (Adaptor related protein complex 3, beta-2 subunit) που αναλαμβάνει μεταφορικές διεργασίες και αντιστοιχεί στο γνώρισμα U37673_at, GPS2 (G protein pathway suppressor 2) που συμμετέχει στην κυτταρική επικοινωνία και αντιστοιχεί στο γνώρισμα U28963_at, AMT (Aminomethyl transferase) που συμμετέχει στο μεταβολισμό και αντιστοιχεί στο γνώρισμα D14686_at, NHLH1 (Nescient helix loop helix 1) που συμμετέχει στο μεταβολισμό νουκλεοτιδίων και νουκλεϊκών οξέων και αντιστοιχεί στο γνώρισμα M96739_at, ACADVL (Acyl CoA dehydrogenase, very long chain) που συμμετέχει στο μεταβολισμό και αντιστοιχεί στο γνώρισμα D43682_at και C5orf18 που αντιστοιχεί στο γνώρισμα M73547_at. Τα LOC440345 και AP3B2 είναι υπεύθυνα για την Class1 (επιζώντες) ενώ όλα τα υπόλοιπα τόσο για την Class1 (επιζώντες) όσο και για την Class0 (αποτυχία). Επίσης, στα φιλτραρισμένα σύνολα συναντάται το γνώρισμα Z14978_at. Το γνώρισμα αυτό αντιστοιχεί στο γονίδιο ACTR1A (Actin related protein

1A) το οποίο είναι υψηλά συσχετισμένο με το γονίδιο AP3B2 που όπως προαναφέρθηκε είναι υπεύθυνο για την Class1 (επιζώντες). Το γονίδιο ACTR1A συμμετέχει στην κυτταρική ανάπτυξη.

Όλα τα παραπάνω γνωρίσματα περιλαμβάνονται επίσης στα σύνολα δεδομένων που παράγονται από το SoFoCles. Επιπλέον, περιέχονται και άλλα γονίδια που φαίνεται να έχουν σημασία για το συγκεκριμένο πρόβλημα. Για παράδειγμα, τα σύνολα oldZZL.MAX και Resnik.AVG_MAX περιέχουν το γνώρισμα U16660_at το οποίο αντιστοιχεί στο γονίδιο ECH1 (Enoyl CoA hydratase, peroxisomal). Το γονίδιο ECH1 εκτός του ότι είναι υψηλά συσχετισμένο με το γονίδιο ACADVL περιγράφεται στο [17] ως ένα από τα γονίδια που είναι υπεύθυνα τόσο για την Class1 (επιζώντες) όσο και για την Class0 (αποτυχία). Το γονίδιο ECH1 συμμετέχει σε μεταβολικές διαδικασίες.

Εξάλλου, το σύνολο JiangConrath.MAX περιέχει τα γνωρίσματα D13900_at και L24774_at που αντιστοιχούν κατά σειρά στα γονίδια ECHS1 (Enoyl CoA hydratase 1) και DCI (Dodecenoyl CoA delta isomerase). Τα γονίδια αυτά εμφανίζουν υψηλή συσχέτιση με το γονίδιο ACADVL, ενώ συμμετέχουν και τα δύο σε μεταβολικές διαδικασίες. Παρόλα αυτά, στο σύνολο αυτό δεν επέρχεται κάποια βελτίωση στην ακρίβεια ταξινόμησης.

Επίσης, στο σύνολο JiangConrath.AVG περιλαμβάνεται το γνώρισμα U57911_at που αντιστοιχεί στο γονίδιο C11orf8 (Fetal brain protein 239). Το γονίδιο αυτό είναι υψηλά συσχετισμένο με το γονίδιο NHLH1 και συμμετέχει σε μεταβολικές διαδικασίες. Ωστόσο, και σε αυτή την περίπτωση δεν επέρχεται κάποια βελτίωση στην ακρίβεια ταξινόμησης.

Το ίδιο συμβαίνει και για το σύνολο Resnik.AVG_MAX που παρότι ενσωματώνει το γονίδιο DUSP9 (Dual specificity phosphatase 9) το οποίο είναι υψηλά συσχετισμένο με το γονίδιο GPS2 δεν παρατηρείται κάποια βελτίωση της ακρίβειας ταξινόμησης.

Τέλος, στα σύνολα Cao.AVG_MAX, JiangConrath.AVG_MAX, Lin.AVG_MAX και oldZZL.AVG_MAX όπου εμφανίζεται βελτίωση της ακρίβειας ταξινόμησης κατά 10% σε σχέση με τα φιλτραρισμένα σύνολα γνωρισμάτων συναντώνται τα γονίδια DSCR1L1 (Calcipressin 2) και FMR1 (Cytoplasmic FMR1 interacting protein 1) που έχουν υψηλή συσχέτιση με το γονίδιο NHLH1 καθώς και το γονίδιο ACTR1A που έχει υψηλή συσχέτιση με το γονίδιο AP3B2. Τα γονίδια DSCR1L1 και FMR1 αντιστοιχούν κατά σειρά στα γνωρίσματα D83407_at και X69962_s_at και συμβάλλουν στην κυτταρική επικοινωνία. Εξάλλου, το γνώρισμα που αντιστοιχεί στο γονίδιο ACTR1A είναι το X82206_at και ενισχύει την διαχωριστική του ισχύ.

7 Συμπεράσματα και προοπτικές

7.1 Συμπεράσματα

Στην παρούσα εργασία παρουσιάστηκε το εργαλείο SoFoCles. Το εργαλείο SoFoCles παρέχει τη δυνατότητα ενσωμάτωσης βιολογικής γνώσης σε σύνολα δεδομένων μικροσυστοιχιών με σκοπό τη βελτίωση της ποιότητας των αποτελεσμάτων. Πιο συγκεκριμένα, χρησιμοποιείται η Γονιδιακή Οντολογία για την εξαγωγή βιολογικής πληροφορίας για τα γονίδια–γνωρίσματα του αρχικού συνόλου δεδομένων. Με αυτό τον τρόπο εντοπίζονται γονίδια που εμφανίζουν υψηλή λειτουργική ομοιότητα αυξάνοντας έτσι την πιθανότητα εύρεσης γονιδίων που συμμετέχουν σε κοινά βιολογικά μονοπάτια που αφορούν ένα συγκεκριμένο πρόβλημα. Με αυτό τον τρόπο, πέρα από την εκμετάλλευση της εσωτερικής πληροφορίας του συνόλου δεδομένων μικροσυστοιχίας που επιτυγχάνεται με την εφαρμογή μεθόδων φιλτραρίσματος γνωρισμάτων προστίθεται και γνώση ανεξάρτητη από τα εσωτερικά χαρακτηριστικά του συνόλου δεδομένων, η οποία στην πλειοψηφία των περιπτώσεων είναι επιστημονικά τεκμηριωμένη και πιστοποιημένη. Η έξοδος, λοιπόν, του SoFoCles είναι ένα νέο σύνολο δεδομένων που περιέχει συνδυασμό πληροφοριών που αντλούνται τόσο από το φιλτράρισμα γνωρισμάτων όσο και από την εκμετάλλευση της γνώσης από τη Γονιδιακή Οντολογία.

Εξάλλου, όσον αφορά το περιβάλλον διεπαφής, αυτό είναι ιδιαίτερα φιλικό προς το χρήστη, αφού περιέχει πολλά στοιχεία που διευκολύνουν τη σωστή χρήση του εργαλείου ενώ παρέχονται και πολλές επεξηγήσεις για τη σημασία κάθε παραμέτρου. Εμφανίζονται ακόμη μηνύματα που ενημερώνουν το χρήστη για τυχόν παραλείψεις καθώς και μηνύματα που δίνουν πληροφορίες για την πρόοδο της εκτέλεσης μίας εργασίας καθώς και συνολικά για τα αποτελέσματα μίας συνόδου.

Το εργαλείο αυτό μπορεί να χρησιμοποιηθεί (ίσως με μερικές υποτυπώδεις τροποποιήσεις) και σε σύνολα δεδομένων που διαθέτουν πρωτεΐνες αντί γονιδίων.

7.2 Προοπτικές

Παρά τις καινοτομίες και τα οφέλη από τη χρήση αυτού του εργαλείου υπάρχει ασφαλώς περιθώριο βελτίωσής του. Για παράδειγμα, μόνο ένα ποσοστό από τα υπάρχοντα γονίδια έχουν καταγεγραμμένες λειτουργίες μέσω της Γονιδιακής Οντολογίας. Μ' άλλα λόγια, μόνο ένα μέρος από τα υπάρχοντα γονίδια έχει επισημειωθεί μέσω των GOA αρχείων σε GO όρους. Επίσης, πρέπει να λαμβάνεται ιδιαίτερη φροντίδα για αποφυγή χρησιμοποίησης επισημειώσεων με χαμηλό επίπεδο αξιοπιστίας, όπως είναι για παράδειγμα οι επισημειώσεις που βασίζονται σε IEA τύπο αποδείξεων. Επιπλέον, υπάρχει ο κίνδυνος εύρεσης πολλών γονιδίων που ανήκουν σε ένα μόνο μονοπάτι και από την άλλη μη δυνατότητα εντοπισμού άλλων που ανήκουν σε άλλα μονοπάτια, κάτι που μπορεί να οδηγήσει σε περίσσεια πληροφορίας.

Αυτό που θα μπορούσε να βελτιωθεί σε πρώτο επίπεδο είναι η ομαδοποίηση (clustering) γονιδίων του αρχικού συνόλου δεδομένων με σκοπό την εύρεση ομάδων γονιδίων που δρουν πιθανώς σε κοινά ή διαφορετικά μονοπάτια για την συμπερίληψη περισσοτέρων μονοπατιών στο φιλτραρισμένο σύνολο δεδομένων και εν συνεχεία η εφαρμογή μεθόδων στατιστικής ομοιότητας (όπως ο συντελεστής Pearson) στα γονίδια που προκύπτουν από το εργαλείο του SoFoCles με σκοπό την αποφυγή της περίσσειας πληροφορίας.

Εν κατακλείδι, η απόδοση του εργαλείου του SoFoCles αποδείχτηκε σε πολλές περιπτώσεις καλύτερη όχι μόνο ως προς το αρχικό σύνολο δεδομένων αλλά και σε σχέση με το φιλτραρισμένο και για τα δύο σύνολα δεδομένων που εξετάστηκαν. Η δυνατότητα εξαγωγής ασφαλών συμπερασμάτων βρίσκεται ακόμη σε ερευνητικό στάδιο, ενώ με κάποιες ενδεχόμενες προσθήκες είναι δυνατή η περαιτέρω βελτίωση τόσο της απόδοσης όσο και της ευρωστίας του. Σε κάθε περίπτωση, ευελπιστούμε η αρχική έκδοση αυτού του εργαλείου να είναι η απαρχή της ένωσης των μεθόδων στατιστικής και σημασιολογικής ομοιότητας με στόχο τη βελτίωση της ακρίβειας των αποτελεσμάτων και την απόδοση μεγαλύτερης βιολογικής ερμηνείας σε αυτά.

8 Βιβλιογραφία

8.1 Βιβλία και Δημοσιεύσεις

[1] Bérard Sèverine, Tichit Laurent and Herrmann Carl. "*ClusterInspector: a tool to visualize ontology-based relationships between biological entities*". Journees Ouvertes Biologie Informatique Mathematiques (JOBIM), Lyon, 6–8 July 2005, pp. 1–12.

[2] Camon Evelyn et al. "*The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro*". Genome Research, Volume 13, Issue 4, April 2003, pp.662–672.

[3] Couto Francisco, Silva Mário and Coutinho Pedro. "*Measuring Semantic Similarity between Gene Ontology Terms*". Data & Knowledge Engineering, Volume 61, Issue 1, April 2007, pp. 137–152.

[4] Haykin Simon. "*Neural Networks: A comprehensive foundation (2nd Edition)*", International Edition, Prentice Hall, New Jersey, NJ, USA, 1999.

[5] Jaeger J., Sengupta R. and Ruzzo W.L. "*Improved Gene Selection for Classification of Microarrays*". In Pacific Symposium on Biocomputing, Volume 8, 2003, pp. 53–64.

[6] Jiang Jay, Conrath David. "*Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*". In Proceedings of International Conference on Research in Computational Linguistics, Taiwan, 1997, pp. 1−15.

[7] Khatri Purvesh, Done Bogdan, Rao Archana, Done Arina and Draghici Sorin. "*A Semantic Analysis of the Annotations of the Human Genome*". Bioinformatics, Volume 21, No 16, 2005, pp. 3416–3421.

[8] Kononenko Igor. "*Estimating Attributes: Analysis and Extensions of Relief*". In Proceedings of the European conference on Machine Learning, Catana, Italy, 1994, Springer-Verlag, New York, pp. 171–182.

[9] Li Jiexun, Su Hua, Chen Hsinchun. "*Identification of Marker Genes from High-dimensional Microarray Data for Cancer Classification*". Knowledge Discovery in Bioinformatics, Chapter 5, Wiley STM.

[10] Li Tao, Zhang Chengliang and Ogihara Mitsunori. "*A Comparative Study of Feature Selection and Multiclass Classification Methods for Tissue Classification Based on Gene Expression*". Bioinformatics, Volume 20, No 15, 2004, pp. 2429–2437.

[11] Lin Dekang. "*An Information - Theoretic Definition of Similarity*". In Proceedings of 15th International Conference on Machine Learning, San Francisco US, 1998, pp. 296−304.

[12] Liu Huan and Setiono Rudy. "*Chi2: Feature Selection and Discretization of Numeric Attributes*". In Proceedings of the 7th International Conference on Tools with Artificial Intelligence, Herndon, VA, 1995, pp. 388–391.

[13] Lord P. , Stevens R. , Brass A. and Goble C. "*Investigating Semantic Similarity Measures across the Gene Ontology: the Relationship between Sequence and Annotation*". Bioinformatics, Volume 19, No 10, 2003, pp. 1275–1283.

[14] Lu Ying and Han Jiawei. "*Cancer Classification Using Gene Expression Data*". Special issue: Data management in bioinformatics, Volume 28, Issue 4, June 2003, pp. 243–278.

[15] NIH Working Definition of Bioinformatics and Computational Biology. July 17, 2000.

[16] Patwardhan Siddharth, Banerjee Satanjeev and Pedersen Ted. "*Using Measures of Semantic Relatedness for Word Sense Disambiguation*". In Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, 2003, pp. 1–17.

[17] Pomeroy SL et al. "*Prediction of central nervous system embryonal tumour outcome based on gene expression*". Nature, 415 (6870), 24 Jan 2002, pp. 436–442.

[18] Resnik Philip. "*Using Information Content to Evaluate Semantic Similarity in a Taxonomy*". In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal Canada, August 1995, pp. 448−453.

[19] Robnik-Šikonja Marko and Kononenko Igor. "*Theoretical and Empirical Analysis of ReliefF and RReliefF*". Machine Learning Journal, Volume 53, October–November 2003, pp. 23–69.

[20] Schlicker Andreas. "*A Global Approach to Comparative Genomics: Comparison of Functional Annotation over the Taxonomic Tree*". Master Thesis, Center for

Bioinformatics of Saarland University, Saarbrücken, Germany, August 30, 2005, pp. 1–82.

[21] Smith Barry, Ceusters Werner, Klagges Bert, Köhler Jacob, Kumar Anand, Lomax Jane, Mungall Chris, Neuhaus Fabian, Rector Alan and Rosse Cornelius. "*Relations in Biomedical Ontologies*". Genome Biology, Volume 6, Issue 5, Article R46, 2005, pp. 1−15.

[22] Spira Avrum, Beane Jennifer, Shah Vishal, Liu Gang, Schembri Frank, Yang Xuemei, Palma John and Brody Jerome. "*Effects of Cigarette Smoke on the Human Airway Epithelial Cell Transcriptome*". PNAS, Volume 101, No 27, July 6, 2004, pp. 10143−10148.

[23] The Gene Ontology Consortium. "*Creating the Gene Ontology Resource: Design and Implementation*". Genome Research, Volume 11, August 2001, pp. 1425–1433.

[24] The Gene Ontology Consortium. "*Gene ontology: Tool for the Unification of Biology*". Nature Genetics, Volume 25, May 2000, pp. 25−29.

[25] Wang Yuhang, Makedon Fillia, Ford James and Pearlman Justin. "*HykGene: a Hybrid Approach for selecting Marker Genes for Phenotype Classification using Microarray Gene Expression Data*". Bioinformatics, Volume 21, No 8, 2005, pp. 1530–1537.

[26] Witten Ian and Frank Eibe. "*Data Mining: Practical Machine Learning Tools with Java Implementations*". Morgan Kaufmann, San Francisco, CA, USA, 2000.

[27] Wu Zhibiao and Palmer Martha. "*Verb Semantics and Lexical Selection*". In Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics, Las Cruces, New Mexico, June 1994, pp. 133–138.

[28] Xing Eric, Jordan Michael and Karp Richard. "*Feature Selection for High-Dimensional Genomic Microarray Data*". In Proceedings of the 18th International Conference on Machine Learning, 2001, pp. 601–608.

[29] Xiong Fei, Huang Heng, Ford James, Makedon Fillia, Pearlman Justin. "*A New Test System for Stability Measurement of Marker Gene Selection in DNA Microarray Data Analysis*". In Proceedings of the 10th Panhellenic Conference on Informatics, Volos, Greece, November 2005, pp. 437−447.

[30] Zhong Jiwei, Zhu Haiping, Li Jianming and Yu Yong. "*Conceptual Graph Matching for Semantic Search*". In Proceedings of the 10th International Conference on Conceptual Structures: Integration and Interfaces, 2002, pp. 92−106.

8.2 Διαδικτυακές Αναφορές

[i] An Introduction to the Gene Ontology. http://www.geneontology.org/GO.doc.shtml?all

[ii] Codon. http://www.genome.gov/Pages/Hyperion/DIR/VIP/Glossary/Illustration/Images/codon.gif

[iii] Detailed cDNA Microarray Technology Scheme. http://plasticdog.cheme.columbia.edu/undergraduate_research/projects/sahil_mehta_project/images/cDNA-array.jpg

[iv] DNA Microarrays. http://www.cse.lehigh.edu/~lopresti/Courses/2003-04/CSE397-497/lecture_20.ppt

[v] DNA Structure: Nucleic Acids. http://www.cartage.org.lb/en/themes/Sciences/LifeScience/PhysicalAnthropology/HeredityBeyond/DNA/DNAStructure/NucleicAcids/pictures/ATBond.jpg

[vi] GO Annotation Guide. http://www.geneontology.org/GO.annotation.shtml?all

[vii] GO Editorial Style Guide. http://www.geneontology.org/GO.usage.shtml?all

[viii] Guide to GO Evidence Codes. http://www.geneontology.org/GO.evidence.shtml?all

[ix] Introduction to Support Vector Machine (SVM) Models: Linear Kernel. http://www.dtreg.com/SvmLinear.jpg

[x] Introduction to Support Vector Machine (SVM) Models: Margin. http://www.dtreg.com/SvmMargin2.jpg

[xi] Introduction to Support Vector Machine (SVM) Models: Non linear problem. http://www.dtreg.com/SvmWiggle.jpg

[xii] Introduction to Support Vector Machine (SVM) Models: Polynomial Kernel. http://www.dtreg.com/SvmPolynomial.jpg

[xiii] Introduction to Support Vector Machine (SVM) Models: Radial Basis Kernel. http://www.dtreg.com/SvmRBF.jpg

[xiv] Introduction to Support Vector Machine (SVM) Models: Separation in higher dimensions. http://www.dtreg.com/SvmDimensionMap.jpg

[xv] Loyola University of Chicago. History Department. Visual Arts. Sophocles. http://www.luc.edu/depts/history/dennis/Visual_Arts/101Images/07_5.05-97_Sophocles.jpg

[xvi] Part-of types. http://www.geneontology.org/images/diag-partof-types.gif

[xvii] Reviewing DNA.
http://evolution.berkeley.edu/evosite/evo101/images/dna_bases.gif

[xviii] Structure and Function of the Genome.
http://web.mit.edu/6.874/www/lectures/lecture-1.pdf

[xix] Supplementary Information of "Effects of Cigarette Smoke on the Human Airway Epithelial Cell Transcriptome".
http://www.pnas.org/cgi/data/0401422101/DC1/9

[xx] Support Vector Machines (SVM) in bioinformatics.
http://cg.ensmp.fr/~vert/talks/020717todai1/todai1.pdf

[xxi] The Gene Ontology. http://www.geneontology.org/images/GO-head.gif

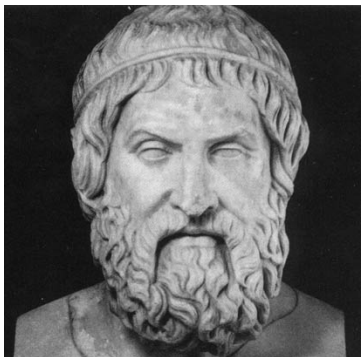
[xxii] The Molecules of Cells. http://nash.cbs.umn.edu/bs101/pix/nucleotide.gif

[xxiii] The OBO Flat File Format Specification, version 1.2.
http://www.geneontology.org/GO.format.obo-1_2.shtml

[xxiv] University of Illinois at Urbana-Champaign Digital Libraries Initiative. Glossary. Ontology definition. http://dli.grainger.uiuc.edu/glossary.htm

[xxv] Wikimedia: Gene.
http://upload.wikimedia.org/wikipedia/commons/0/07/Gene.png

Παράρτημα Α – Λίγα λόγια για τον Σοφοκλή και τους λόγους επιλογής αυτού του ονόματος για το εργαλείο



Σοφοκλής
Η παραπάνω εικόνα έχει ληφθεί από την τοποθεσία [xv]

Ο Σοφοκλής (495 π.Χ. – 406 π.Χ.), γιος του Αθηναίου κατασκευαστή όπλων Σόφιλλου, γεννήθηκε το 495 στον Κολωνό της Αττικής και έζησε όλα του τα χρόνια στην Αθήνα. Ήταν ένας από τους τρεις μεγαλύτερους τραγικούς ποιητές μαζί με τους Αισχύλο και Ευριπίδη [a].

Ο Σοφοκλής συνδύαζε την ευφυΐα με τη χάρη και την ομορφιά. Χαρακτηριστικά αναφέρουμε ότι λόγω του εντυπωσιακού παραστήματος, επιλέχθηκε, μετά την ναυμαχία της Σαλαμίνας, σε ηλικία μόλις 16 χρονών να γίνει ο κορυφαίος του χορού των εφήβων (παιάνας) στις δημόσιες τελετές για τον εορτασμό της νίκης [b, c].

Πιστεύεται ότι έγραψε περισσότερα από 123 έργα ενώ όταν συμμετείχε σε αγώνες πάντα ελάμβανε το πρώτο ή το δεύτερο βραβείο. Την πρώτη φορά που πήρε μέρος στους δραματικούς αγώνες το 468 π.Χ., σε ηλικία μόλις 28 χρονών, παίρνει το πρώτο βραβείο με αντίπαλο τον Αισχύλο. Στα επόμενα χρόνια παίρνει 24 πρώτες νίκες (18 φορές στους Διονύσιους και 6 στους Ληναίους αγώνες) [b].

Τα τρία πιο γνωστά έργα του είναι ο *Οιδίπους Τύραννος,* ο *Οιδίπους επί Κολωνώ* και η *Αντιγόνη*.

Ο Σοφοκλής έφερε πολλές καινοτομίες στο θέατρο. Οι κυριότερες από αυτές ήταν οι εξής:

- Εισήγαγε και τρίτο υποκριτή μειώνοντας έτσι το ρόλο του χορού στην παρουσίαση του έργου και δίνοντας έτσι τη δυνατότητα για ανάπτυξη των χαρακτήρων σε μεγαλύτερο βαθμό σε σχέση με προγενέστερους ποιητές όπως ο Αισχύλος [a].

- Χρησιμοποιούσε επίσης θηλυκούς χαρακτήρες στα έργα του [a].
- Έδωσε μεγάλη σημασία στην κατασκευή και διευθέτηση των σκηνικών [c].
- Κατάργησε την *τριλογία* (τρεις τραγωδίες) ως μορφή για την αφήγηση μίας και μοναδικής ιστορίας. Ο Σοφοκλής επέλεξε να δώσει ανεξάρτητη και ολοκληρωμένη οντότητα σε κάθε τραγωδία του ενισχύοντας έτσι τη δραματικότητα [d].
- Τέλος, αύξησε τα μέλη του χορού σε 15 ώστε να πετύχει καλύτερη συμμετρία: δύο ημιχόρια των 7 και ο κορυφαίος του χορού [b].

Κανείς μπορεί εύλογα να αναρωτηθεί πώς προέκυψε το όνομα Σοφοκλής (SoFoCles) γι' αυτό το εργαλείο. Η επιλογή αυτού του ονόματος αποτελεί κατ' αρχήν αναφορά στους τρεις συνεισφέροντες στη σύλληψη, σχεδιασμό και υλοποίηση του:

Παπαχριστούδης Γεώργιος με προσωνύμιο Fou
Διπλάρης Σωτήρης (Sotiris)
Μήτκας Περικλής (PeriCles)

Έτσι από την παρακάτω ένωση: SotirisFouPeriCles προκύπτει η συντομογραφία SoFoCles.

Πέρα όμως από την προφανή πρόθεση χρήσης αυτού του ονόματος, θα μπορούσε επίσης να ειπωθεί ότι όπως ακριβώς ο Σοφοκλής επιχείρησε να εισάγει κάποιες καινοτομίες στο δράμα, έτσι και το SoFoCles προσπαθεί να λειτουργήσει παρόμοια. Με άλλα λόγια, η προσπάθεια του Σοφοκλή να δώσει ανεξαρτησία στα έργα του και να αναπτύξει σε μεγαλύτερο βάθος τους χαρακτήρες του επιχειρείται και από το SoFoCles σε μία άλλη βεβαίως διάσταση.

Πιο συγκεκριμένα, με τη χρησιμοποίηση της Γονιδιακής Οντολογίας επιχειρείται η εξαγωγή συμπερασμάτων από σύνολα μικροσυστοιχιών που δεν είναι τόσο προσκολλημένα στην εσωτερική φύση των δεδομένων και επομένως διατηρούν μία σχετική ανεξαρτησία από αυτά. Ενώ από την άλλη, με την αναφορά σε βιολογικά δεδομένα, η αξιοπιστία των περισσοτέρων από τα οποία έχει επικυρωθεί, επιχειρείται η απόδοση βιολογικής σημασίας στα αποτελέσματα και κατά ένα τρόπο, έτσι, επιτυγχάνεται η μεγαλύτερη εμβάθυνση στο υπόβαθρο του προβλήματος.

τὸ δὲ ζητούμενον ἁλωτόν, ἐκφεύγειν δὲ τ' ἀμελούμενον.

αυτός που ψάχνει θα βρει αυτό που ζητάει, αυτός που το αμελεί θα του διαφύγει.
(στ. 110-111, Οιδίποδας Τύραννος, Σοφοκλής)

[a] Wikipedia, the free encyclopedia - Sophocles (http://en.wikipedia.org/wiki/Sophocles)

[b] TheaterInfo - Σοφοκλής (http://www.theaterinfo.gr/abouttheatre/ancientgreektheatre/sofoklis/index.html)

[c] TheatreHistory - Sophocles and his Tragedies (http://www.theatrehistory.com/ancient/sophocles001.html)

[d] The Literature Network - Sophocles (http://www.online-literature.com/sophocles/)

Παράρτημα Β – Εργαλεία και πακέτα που χρησιμοποιήθηκαν

Β.1 Εργαλεία

Β.1.1 JBuilder® 2006 Enterprise

Επιλέχτηκε η χρήση του περιβάλλοντος ανάπτυξης του JBuilder® 2006 Enterprise (http://www.borland.com/) καθώς αποτελεί ένα από τα ευρέως χρησιμοποιούμενα περιβάλλοντα ανάπτυξης λογισμικού σε Java. Μπορεί ο προγραμματιστής να ολοκληρώσει ένα έργο γράφοντας κώδικα υποβοηθούμενος από πολλά ενσωματωμένα στοιχεία του JBuilder® είτε παράγοντας αυτόματα χρησιμοποιώντας μερικά από τα αυτοματοποιημένα εργαλεία (για παράδειγμα, ο σχεδιαστής Γραφικού Περιβάλλοντος, GUI Designer).

Πιο συγκεκριμένα, ο JBuilder® 2006 Enterprise προσφέρει τη δυνατότητα ευκολότερης παρακολούθησης των αλλαγών μέσω της λειτουργίας του: Active Difference Editing. Επιπλέον, υποστηρίζει τα πιο πρόσφατα πρότυπα συμπεριλαμβανομένων των Java 2, Java 2 JFC/Swing, XML, Java2D, Message Queue, Java collections, Accessibility API, Speech API, Java™ 2 Platform και άλλων. Υποστηρίζει, επίσης, όλα τα στοιχεία των J2SE 5.0.

Τέλος, ένα από τα σημαντικότερα οφέλη του JBuilder® 2006 Enterprise είναι το γραφικό περιβάλλον του αποσφαλματωτή (debugger). Ενώ το νέο χαρακτηριστικό του υπολογισμού / τροποποίησης (evaluation / modification) παραστάσεων μεταβλητών συνεισφέρει σημαντικά στη διαδικασία αποσφαλμάτωσης. Τέλος, επιπρόσθετες επιλογές αναζήτησης και το βελτιωμένο περιβάλλον καθοδήγησης διόρθωσης των σφαλμάτων επιταχύνουν σημαντικά το χρόνο ανάπτυξης.

B.1.2 The R Project for Statistical Computing

Το εργαλείο R (http://www.r-project.org/) χρησιμοποιήθηκε για τη δημιουργία αρχείων αντιστοιχήσεων (mappings) γνωρισμάτων συνόλων δεδομένων σε γονιδιακά σύμβολα. Επιλέχθηκε γιατί περιέχει πολύ καλά ενημερωμένες βιβλιοθήκες Βιοπληροφορικής, χειρίζεται πολύ αποδοτικά τα δεδομένα και τέλος γιατί οι τελεστές (operators) είναι κατάλληλα ορισμένοι για πράξεις μεταξύ πινάκων ενώ η γλώσσα που χρησιμοποιείται είναι απλή.

B.2 Πακέτα

B.2.1 GO4J

Το GO4J (http://www.bioinformatics.org/GO4J/) είναι το βασικό πακέτο που χρησιμοποιήθηκε σε αυτή τη διπλωματική. Είναι λογισμικό ανοιχτού κώδικα (Open Source software) και αποτελείται από ένα σύνολο διεπαφών (Java Application Programming Interfaces, Java APIs) που χειρίζονται τη Γονιδιακή Οντολογία.

B.2.2 Spring Framework

Το Spring Framework (http://www.springframework.org/) είναι ένα πλαίσιο εφαρμογών σε Java που προσφέρει πολλά οφέλη διευκολύνοντας κυρίως την ολοκλήρωση εργασιών και μειώνοντας το χρόνο εκτέλεσης. Στη διπλωματική χρησιμοποιήθηκε κυρίως για το χειρισμό αρχείων και δομών δεδομένων.

B.2.3 Weka

Το Weka (http://www.cs.waikato.ac.nz/ml/weka/) αποτελεί ένα ισχυρό πακέτο αλγορίθμων μηχανικής εκμάθησης ανεπτυγμένο σε Java. Περιλαμβάνει εργαλεία για προ-επεξεργασία δεδομένων, ταξινόμηση, παλινδρόμηση, ομαδοποίηση, παραγωγή κανόνων συσχέτισης καθώς και εργαλεία για την οπτικοποίηση των αποτελεσμάτων. Το αξιοσημείωτο είναι ότι αποτελεί και αυτό λογισμικό ανοιχτού κώδικα (Open Source software). Τέλος, όσον αφορά τη διπλωματική χρησιμοποιήθηκε τόσο για το χειρισμό και τη δημιουργία .arff αρχείων όσο και για την αποτίμηση των αποτελεσμάτων.

Ευρετήριο

Index

S

T

W