ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανακάλυψη Γνώσης από Βιολογικά Δεδομένα

Διδακτορική Διατριβή

Γεώργιος Τζανής

Πτυχιούχος Τμήματος Πληροφορικής Α.Π.Θ.

ΘΕΣΣΑΛΟΝΙΚΗ 2011

Γεώργιος Τζανής

Πτυχιούχος Τμήματος Πληροφορικής Α.Π.Θ.

Ανακάλυψη Γνώσης από Βιολογικά Δεδομένα

Διδακτορική Διατριβή

Υποβλήθηκε στο Τμήμα Πληροφορικής
του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης
Ημερομηνία Προφορικής Εξέτασης: 20/06/2011

Συμβουλευτική Επιτροπή:
Βλαχάβας Ιωάννης - Καθηγητής (επιβλέπων)
Βακάλη Αθηνά - Αναπληρώτρια Καθηγήτρια
Βασιλειάδης Νικόλαος - Αναπληρωτής Καθηγητής

Εξεταστική Επιτροπή:
Βλαχάβας Ιωάννης - Καθηγητής
Βακάλη Αθηνά - Αναπληρώτρια Καθηγήτρια
Βασιλειάδης Νικόλαος - Αναπληρωτής Καθηγητής
Λάσκαρης Νικόλαος - Επίκουρος Καθηγητής
Ποταμιάς Γεώργιος - Ερευνητής Β΄
Τσαμαρδίνος Ιωάννης - Επίκουρος Καθηγητής
Τσουμάκας Γρηγόριος - Λέκτορας

Ευχαριστίες

Η εκπόνηση της διατριβής πραγματοποιήθηκε στο Εργαστήριο Γλωσσών Προγραμματισμού και Τεχνολογίας Λογισμικού του Τμήματος Πληροφορικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, σε συνεργασία με την ομάδα Μηχανικής Μάθησης και Ανακάλυψης Γνώσης.

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα της διατριβής μου Καθηγητή κ. Ιωάννη Βλαχάβα, για την εμπιστοσύνη που μου έδειξε, για τον απεριόριστο χρόνο που αφιέρωσε στην καθοδήγησή μου και για την εξασφάλιση υλικών και τεχνικών υποδομών κατά τη διάρκεια της εκπόνησης της διατριβής. Επίσης, θα ήθελα να τον ευχαριστήσω, γιατί μου έδωσε τη δυνατότητα συμμετοχής σε ερευνητικά και αναπτυξιακά προγράμματα, τα οποία αποτέλεσαν πηγή σημαντικής επαγγελματικής εμπειρίας και γνώσης.

Θα ήθελα ακόμη να ευχαριστήσω τον Αναπληρωτή Καθηγητή κ. Νικόλαο Βασιλειάδη και την Αναπληρώτρια Καθηγήτρια κ. Αθηνά Βακάλη, που διατέλεσαν μέλη της συμβουλευτικής μου επιτροπής και συνέβαλλαν σημαντικά με την πολύτιμη καθοδήγηση και υποστήριξή τους.

Θα ήθελα επίσης να ευχαριστήσω το διδάκτορα κ. Χρήστο Μπερμπερίδη, το Λέκτορα κ. Γρηγόριο Τσουμάκα και τον υποψήφιο διδάκτορα κ. Ιωάννη Καβακιώτη, για τις εποικοδομητικές ανταλλαγές απόψεων και την άριστη συνεργασία τους κατά την προετοιμασία και συγγραφή πολλών από τα επιστημονικά άρθρα που προέκυψαν ως αποτέλεσμα της εκπόνησης της διατριβής. Επιπλέον, θα ήθελα να ευχαριστήσω την Καθηγήτρια κ. Άρτεμις Χατζηγεωργίου και την Καθηγήτρια κ. Φίλια Μακεδών για την πρόθυμη και πολύτιμη συνεισφορά τους με τις γνώσεις και την εμπειρία τους.

Επίσης, οφείλω να εκφράσω τις ευχαριστίες μου στους φίλους και συναδέλφους, διδάκτορες κκ. Ιωάννη Κατάκη, Ιωάννη Παρτάλα, Ευστράτιο Κοντόπουλο και Γεώργιο Μεδίτσκο για την αρμονική συνύπαρξη και την άριστη συνεργασία τους κατά τη διάρκεια της εκπόνησης της διατριβής.

Θα ήθελα ακόμη να ευχαριστήσω και τους υπόλοιπους συνεργάτες του εργαστηρίου, το Λέκτορα κ. Δημήτριο Βράκα, τον Καθηγητή Εφαρμογών κ. Φώτιο Κόκκορα και το Λέκτορα κ. Ηλία Σακελλαρίου, καθώς και τα νεότερα μέλη του εργαστηρίου, τους υποψήφιους διδάκτορες κκ. Καλλιόπη Κράβαρη, Αθανάσιο Σταυρόπουλο και Ελευθέριο Σπυρομήτρο-Ξιούφη.

Τέλος, αισθάνομαι την ανάγκη να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την υπομονή και την υποστήριξή τους καθ' όλη τη διάρκεια εκπόνησης της διατριβής.

Γεώργιος Τζανής

Θεσσαλονίκη, Ιούνιος 2011

Περίληψη

Κατά τη διάρκεια των τελευταίων δεκαετιών, οι σημαντικές εξελίξεις των επιστημών της πληροφορικής και της βιολογίας έχουν προσελκύσει το ενδιαφέρον όλης της ανθρωπότητας. Η ανάπτυξη του Παγκόσμιου Ιστού και η αποκρυπτογράφηση του ανθρώπινου γονιδιώματος αντανακλούν την έκταση της προόδου αυτών των επιστημών. Μάλιστα, η ραγδαία ανάπτυξη της τεχνολογίας έχει οδηγήσει στην αύξηση του ρυθμού παραγωγής βιολογικών δεδομένων, γεγονός που κατέστησε επιτακτική την ανάγκη στενής συνεργασίας μεταξύ των επιστημόνων της βιολογίας και της πληροφορικής. Αυτή η συνεργασία οδήγησε στην εμφάνιση της βιοπληροφορικής, μιας νέας διεπιστημονικής ερευνητικής περιοχής, που αποτελεί την τομή των επιστημών της βιολογίας και της πληροφορικής. Σκοπός της βιοπληροφορικής είναι η ανάπτυξη και η χρήση εργαλείων για τη διαχείριση και ανάλυση του τεράστιου όγκου των βιολογικών δεδομένων.

Παράλληλα με τη βιοπληροφορική εξελίχθηκε και ένα άλλο ερευνητικό πεδίο ως απότοκος της ανάγκης αποδοτικής και αποτελεσματικής διαχείρισης μεγάλων όγκων δεδομένων. Η ανακάλυψη γνώσης από βάσεις δεδομένων παρέχει ένα σύνολο ισχυρών εργαλείων για την ανάλυση νέων τύπων δεδομένων με σκοπό την αποδοτική και αποτελεσματική εξαγωγή γνώσης. Αυτά τα εργαλεία συνδυάζουν τεχνικές από διάφορες περιοχές, όπως είναι η στατιστική, η τεχνητή νοημοσύνη, η αναγνώριση προτύπων και η τεχνολογία βάσεων δεδομένων, με σκοπό την αντιμετώπιση των εμποδίων και των περιορισμών που τίθενται από τις παραδοσιακές μεθόδους ανάλυσης δεδομένων.

Η παρούσα διατριβή έχει ως αντικείμενο μελέτης την ανακάλυψη γνώσης από βιολογικά δεδομένα και εντάσσεται στα πλαίσια των ερευνητικών περιοχών της ανακάλυψης γνώσης από βάσεις δεδομένων και της βιοπληροφορικής. Στους στόχους της διατριβής περιλαμβάνεται η επινόηση νέων και η τροποποίηση ή επέκταση υπαρχουσών μεθόδων ανακάλυψης γνώσης, καθώς και η εφαρμογή τους για την ανάλυση βιολογικών δεδομένων.

Στα πλαίσια της διατριβής ορίζεται ένα νέο είδος προβλήματος της ανακάλυψης γνώσης από βάσεις δεδομένων, που περιλαμβάνει την ανακάλυψη αμοιβαίως αποκλειόμενων αντικειμένων. Προτείνεται ένας αλγόριθμος ανακάλυψής τους από μια βάση δεδομένων συναλλαγών, ο οποίος εφαρμόζεται σε βιολογικά δεδομένα που αφορούν την έκφραση γονιδίων. Επιπλέον, προτείνεται μια μέθοδος για την αποτελεσματική ταξινόμηση δεδομένων γονιδιακής έκφρασης με απώτερο σκοπό την πρόβλεψη φαινοτύπου. Επίσης, προτείνονται προσεγγίσεις που εντάσσονται στα πλαίσια της ανάλυσης βιολογικών αλληλουχιών. Συγκεκριμένα, αντιμετωπίζονται δύο ξεχωριστά προβλήματα που προσελκύουν το ενδιαφέρον των ερευνητών, η πρόβλεψη του σημείου έναρξης της μετάφρασης και η πρόβλεψη του σημείου αποκοπής και πολυαδενυλίωσης.

Abstract

Title of PhD Thesis: Knowledge Discovery from Biological Data

During the last decades the major advances in informatics and biology have attracted the interest of all humanity. The growth of World Wide Web and the Human Genome Project reflect the extent of the development of these two scientific areas. Moreover, the rapid technological development has resulted in an increased rate of biological data accumulation, necessitating the collaboration between scientists of biology and informatics. This collaboration has lead to the emergence of bioinformatics, a new interdisciplinary research area, which is the intersection of biology and informatics. The aim of bioinformatics is the development and use of tools for managing and analyzing the vast amounts of biological data.

Another research area has been grown along with bioinformatics, in order to deal with large-volume data management efficiently and effectively. Knowledge discovery in databases provides a set of tools to analyze new data types in order to extract various types of knowledge efficiently and effectively. These tools combine powerful techniques from different areas, such as statistics, artificial intelligence, pattern recognition and database technology. This fusion of technologies aims to overcome the obstacles and constraints posed by the traditional data analysis methods.

The scope of this thesis includes knowledge discovery from biological data and is strongly related to the research areas of knowledge discovery in data bases and bioinformatics. This thesis targets to the invention of new knowledge discovery methods and the modification or extension of existing ones, as well as the application of these methods for biological data analysis.

In the framework of this thesis a new knowledge discovery problem, which deals with the discovery of mutually exclusive items, is defined. An algorithm for mining mutually exclusive items from transactional data is provided and is applied on gene expression data. Moreover, an approach for the effective classification of gene expression data and phenotype prediction is proposed. Furthermore, a number of biological sequence analysis approaches are proposed. In particular, two interesting and distinct sequence analysis problems are studied, namely translation initiation site prediction and cleavage-polyadenylation site prediction.

Περιεχόμενα

Ευχαριστίες i

Περίληψη iii

Abstract v

Περιεχόμενα vii

Κατάλογος Εικόνων xiii

Κατάλογος Πινάκων xvii

1 Εισαγωγή 1

1.1 Κίνητρο Διατριβής και Συναφή Ερευνητικά Πεδία 1

1.1.1 Τάσεις της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων 3

1.1.2 Τάσεις της Βιοπληροφορικής 4

1.2 Συμβολή και Δομή της Διατριβής 5

2 Ανακάλυψη Γνώσης από Βάσεις Δεδομένων 9

2.1 Η Διαδικασία Ανακάλυψης Γνώσης από Βάσεις Δεδομένων 10

2.1.1 Στάδια της Διαδικασίας Ανακάλυψης Γνώσης από Βάσεις Δεδομένων 11

2.2 Μηχανική Μάθηση 13

2.2.1 Κατηγορίες Μεθόδων Μηχανικής Μάθησης 14

2.2.2 Σχέση Μηχανικής Μάθησης και Ανακάλυψης Γνώσης 15

2.3 Μέθοδοι Εξόρυξης από Δεδομένα 15

2.3.1 Ταξινόμηση και Παλινδρόμηση 15

2.3.2 Εξόρυξη Κανόνων Συσχέτισης 20

2.3.3 Ομαδοποίηση 23

2.4 Μέθοδοι Προεπεξεργασίας και Μεταεπεξεργασίας 24

2.4.1 Μείωση Διαστάσεων 24

2.4.2 Διακριτοποίηση 25

2.4.3 Αξιολόγηση Ταξινομητών 27

2.5 Αποθήκες Δεδομένων 29

2.6 Προβλήματα της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων 30

2.6.1 Ποιότητα Δεδομένων 31

2.6.2 Όγκος Δεδομένων 32

2.6.3 Διαχείριση Αποτελεσμάτων 32

2.6.4 Προβλήματα Μεθοδολογίας 33

2.6.5 Προβλήματα Απόδοσης 34

2.7 Εφαρμογές της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων 35

2.8 Σύνοψη 35

3 Μοριακή Βιολογία και Βιοπληροφορική 37

3.1 Οργανισμοί και Κύτταρα 38

3.2 Βιομόρια 44
3.2.1 Μικρά Μόρια 44
3.2.2 Πρωτεΐνες 45
3.2.3 DNA 47
3.2.4 RNA 50

3.3 Γονίδια και Γονιδίωμα 51

3.4 Μεταλλάξεις και Πολυμορφισμοί 53

3.5 Γονιδιακή Έκφραση 55
3.5.1 Μεταγραφή 55
3.5.2 Μάτισμα 55
3.5.3 Προσθήκη Καλύπτρας και Πολυαδενυλίωση 56
3.5.4 Μετάφραση 56

3.6 Γονιδιακή Ρύθμιση 59

3.7 Η Ροή της Βιολογικής Πληροφορίας 59

3.8 Βιοπληροφορική 60
3.8.1 Οι Στόχοι της Βιοπληροφορικής 61

3.9 Εύρεση και Ανάλυση Βιολογικών Αλληλουχιών 61
3.9.1 Αλληλούχιση DNA 61
3.9.2 Πρόβλεψη Γονιδίων 61
3.9.3 Στοίχιση Αλληλουχιών 62

3.10 Μέτρηση και Ανάλυση της Γονιδιακής Έκφρασης 64
3.10.1 Μικροσυστοιχίες 64
3.10.2 Η Τεχνική SAGE 65

3.11 Νεολογισμοί της Βιοπληροφορικής 66
3.11.1 Γονιδιωματική 66
3.11.2 Συγκριτική Γονιδιωματική 66
3.11.3 Πρωτεωμική 67

3.12 Βιολογικές Βάσεις Δεδομένων 67
3.12.1 Γενικευμένες ή Αρχειακές Βιολογικές Βάσεις Δεδομένων 68
3.12.2 Δευτερογενείς Βιολογικές Βάσεις Δεδομένων 69
3.12.3 Εξειδικευμένες Βιολογικές Βάσεις Δεδομένων 69
3.12.4 Βιβλιογραφικές Βιολογικές Βάσεις Δεδομένων 69
3.12.5 Βιολογικές Βάσεις Δεδομένων Ιστοσελίδων 70

3.13 Η Ανακάλυψη Γνώσης στη Βιοπληροφορική 70
3.13.1 Ανακάλυψη Γνώσης από Βιολογικές Αλληλουχίες 70
3.13.2 Ανακάλυψη Γνώσης από Δεδομένα Γονιδιακής Έκφρασης 71
3.13.3 Ανακάλυψη Γνώσης στη Δομική Βιοπληροφορική 71
3.13.4 Ανακάλυψη Γνώσης από Βιολογικά Κείμενα 72

3.14 Ερευνητικά Κέντρα Βιοπληροφορικής 73
3.15 Εργαλεία Βιοπληροφορικής 73
3.16 Εφαρμογές και Τάσεις της Βιοπληροφορικής 75
3.17 Σύνοψη 77

4 Ανακάλυψη Αμοιβαίως Αποκλειόμενων Αντικειμένων 79
4.1 Εισαγωγή 79
4.2 Βασικές Έννοιες 80
4.3 Σχετικές Εργασίες 81
4.4 Εξόρυξη Αμοιβαίως Αποκλειόμενων Αντικειμένων 83
4.4.1 Ορισμός του Προβλήματος 83
4.4.2 Εξόρυξη Εκτεταμένων Συχνών Συνόλων Αντικειμένων 84
4.4.3 Εξόρυξη Ζευγών Αμοιβαίως Αποκλειόμενων Αντικειμένων 85
4.4.4 Εξόρυξη Τριάδων Αμοιβαίως Αποκλειόμενων Αντικειμένων 88
4.4.5 Παράδειγμα Εκτέλεσης του Αλγορίθμου 89
4.4.6 Πειράματα 90
4.4.7 Αμοιβαίως Αποκλειόμενα Αντικείμενα και Ταξονομίες 91
4.5 Ανακάλυψη Αμοιβαίως Αποκλειόμενων Γονιδίων 93
4.5.1 Η Τεχνική SAGE 93
4.5.2 Τα Δεδομένα SAGE 94
4.5.3 Διακριτοποίηση των Δεδομένων 94
4.5.4 Πειράματα 95
4.6 Συμπεράσματα 96

5 Πρόβλεψη Φαινοτύπου από Δεδομένα Γονιδιακής Έκφρασης 99
5.1 Εισαγωγή 99
5.2 Βασικές Έννοιες 99
5.3 Σχετικές Εργασίες 101
5.4 Η Προτεινόμενη Προσέγγιση 103
5.4.1 Δομή Δεδομένων Εισόδου 103
5.4.2 Διακριτοποίηση 104
5.4.3 Εξόρυξη Συχνών Συνόλων Αντικειμένων 105
5.4.4 Επιλογή Χαρακτηριστικών 106
5.4.5 Ταξινόμηση 106
5.5 Πειραματική Αξιολόγηση 107
5.5.1 Σύνολα Δεδομένων 107
5.5.2 Εξόρυξη Συχνών Συνόλων Αντικειμένων 108
5.5.3 Επιλογή Χαρακτηριστικών 108
5.5.4 Ταξινόμηση 109
5.5.5 Μέθοδος Επικύρωσης 109
5.5.6 Αποτελέσματα 109
5.6 Συμπεράσματα 115

6 Πρόβλεψη του Σημείου Έναρξης της Μετάφρασης 117
6.1 Εισαγωγή 117
6.2 Βασικές Έννοιες 118
6.3 Σχετικές Εργασίες 119
6.4 Η Μέθοδος Πλειοψηφίας 120
6.4.1 Χαρακτηριστικά 121
6.4.2 Ταξινομητές 124
6.4.3 Πειραματική Αξιολόγηση 124
6.5 Η Μέθοδος StackTIS 128
6.5.1 Το Στοιχείο Κωδικής Περιοχής 129
6.5.2 Το Στοιχείο Προτύπου Συναίνεσης 131
6.5.3 Το Στοιχείο Απόστασης του ATG 133
6.5.4 Σύντηξη των Επιμέρους Προβλέψεων 134
6.5.5 Τελική Πρόβλεψη του Σημείου Έναρξης της Μετάφρασης 135
6.5.6 Πειραματική Αξιολόγηση 136
6.6 Συμπεράσματα 145

7 Πρόβλεψη του Σημείου Αποκοπής και Πολυαδενυλίωσης 147
7.1 Εισαγωγή 147
7.2 Βασικές Έννοιες 147
7.2.1 Αποκοπή και Πολυαδενυλίωση 147
7.2.2 Ταξινόμηση με Αναδυόμενα Πρότυπα 150
7.3 Σχετικές Εργασίες 152
7.4 Η Μέθοδος PolyA-iEP 153
7.4.1 Εξαγωγή Στοιχείων 154
7.4.2 Εξαγωγή των Προτύπων Νουκλεοτιδίων 154
7.4.3 Διακριτοποίηση 155
7.4.4 Εξόρυξη Χι Αναδυόμενων Προτύπων 156
7.4.5 Βαθμολόγηση με Βάση τα Χι Αναδυόμενα Πρότυπα 156
7.4.6 Βαθμολόγηση με Βάση την Απόσταση 157
7.4.7 Ταξινόμηση 158
7.5 Πειραματική Αξιολόγηση 159
7.5.1 Δεδομένα 159
7.5.2 Εξερευνητική Ανάλυση 160
7.5.3 Συνεισφορά των Χαρακτήρων Διττής Σημασίας 162
7.5.4 Αξιολόγηση της Βαθμολόγησης με Βάση τα Χι Αναδυόμενα Πρότυπα 164
7.5.5 Αξιολόγηση της Βαθμολόγησης με Βάση την Απόσταση 164
7.5.6 Αξιολόγηση του Ταξινομητή 165
7.6 Συμπεράσματα 169

8 Επίλογος 171
8.1 Συμπεράσματα 171
8.2 Μελλοντικές Επεκτάσεις 173

Βιβλιογραφία 175

Λίστα Δημοσιεύσεων 187

Λίστα Ετεροαναφορών 189

Κατάλογος Εικόνων

Εικόνα 1.1: Εξέλιξη του μεγέθους της βάσης δεδομένων GenBank (1982-2010). 2

Εικόνα 2.1: Τα βασικά στάδια της διαδικασίας ανακάλυψης γνώσης. 11

Εικόνα 2.2: Διαδικασία δημιουργίας και χρήσης αποθήκης δεδομένων. 30

Εικόνα 3.1: Η ταυτότητα του σύγχρονου ανθρώπου. 40

Εικόνα 3.2: Τυπικά μεγέθη κυττάρων, μικροοργανισμών και μορίων. 41

Εικόνα 3.3: Τυπικό ζωικό κύτταρο. 42

Εικόνα 3.4: Τυπικό φυτικό κύτταρο. 43

Εικόνα 3.5: Τυπικό βακτηριακό κύτταρο. 44

Εικόνα 3.6: Τα τέσσερα επίπεδα διαμόρφωσης των πρωτεϊνών. 47

Εικόνα 3.7: Η διπλή έλικα του μορίου του DNA. 49

Εικόνα 3.8: Αναλυτικότερη άποψη του δίκλωνου μορίου του DNA. 49

Εικόνα 3.9: Η δομή του χρωμοσώματος σε διάφορα επίπεδα μεγέθυνσης. 52

Εικόνα 3.10: Η διαδικασία της μετάφρασης. 58

Εικόνα 3.11: Το κεντρικό δόγμα της μοριακής βιολογίας. 59

Εικόνα 3.12: Η ροή της βιολογικής πληροφορίας. 60

Εικόνα 3.13: Η συστοιχία GeneChip® της εταιρείας Affymetrix. 64

Εικόνα 3.14: Η τεχνική SAGE. 65

Εικόνα 4.1: Τα τρία βήματα του αλγορίθμου εξόρυξης. 84

Εικόνα 4.2: Το τοπικό μέτρο αμοιβαίου αποκλεισμού (LM) σε σύγκριση με την τοπική υποστήριξη των αντικειμένων *A* και *B* και του συνόλου {*A*, *B*} στον *F*-υποχώρο. 86

Εικόνα 4.3: Διαγράμματα Venn ζεύγους υποψήφιων αμοιβαίως αποκλειόμενων αντικειμένων. 87

Εικόνα 4.4: Χρόνοι εκτέλεσης του αλγορίθμου εξόρυξης. 91

Εικόνα 4.5: Μια ταξονομία προϊόντων καφέ. 92

Εικόνα 4.6: Μέσο μέγεθος συναλλαγής σε σχέση με τη μέθοδο διακριτοποίησης. 96

Εικόνα 4.7: Πλήθος αμοιβαίως αποκλειόμενων γονιδίων σε σχέση με το κατώφλι ελάχιστης τοπικής υποστήριξης (min_gSup = 0,25, διακριτοποίηση mid-range-based cutoff). 96

Εικόνα 5.1: Η αρχιτεκτονική της προτεινόμενης μεθόδου. 103

Εικόνα 5.2: Ακρίβεια των ταξινομητών στο σύνολο δεδομένων 99×27679. Ως μέτρο αξιολόγησης των χαρακτηριστικών χρησιμοποιήθηκε το στατιστικό X^2. 110

Εικόνα 5.3: Ακρίβεια των ταξινομητών στο σύνολο δεδομένων 99×27679. Ως μέτρο αξιολόγησης των χαρακτηριστικών χρησιμοποιήθηκε το κέρδος πληροφορίας. 111

Εικόνα 5.4: Ακρίβεια των ταξινομητών στο σύνολο δεδομένων 99×27679. Ως μέτρο αξιολόγησης των χαρακτηριστικών χρησιμοποιήθηκε το μέτρο Relief F. 112

Εικόνα 5.5: Ακρίβεια των ταξινομητών στο σύνολο δεδομένων 74×822. Ως μέτρο αξιολόγησης των χαρακτηριστικών χρησιμοποιήθηκε το στατιστικό X^2. 114

Εικόνα 5.6: Ακρίβεια των ταξινομητών στο σύνολο δεδομένων 74×822. Ως μέτρο αξιολόγησης των χαρακτηριστικών χρησιμοποιήθηκε το κέρδος πληροφορίας. 114

Εικόνα 5.7: Ακρίβεια των ταξινομητών στο σύνολο δεδομένων 74×822. Ως μέτρο αξιολόγησης των χαρακτηριστικών χρησιμοποιήθηκε το μέτρο Relief F. 114

Εικόνα 6.1: Έναρξη της μετάφρασης σε αλληλουχία mRNA. 119

Εικόνα 6.2: Εξαγωγή διανύσματος χαρακτηριστικών από αλληλουχία cDNA. 121

Εικόνα 6.3: Αρίθμηση των θέσεων των νουκλεοτιδίων μέσα στα κωδικόνια που βρίσκονται στο ίδιο πλαίσιο ανάγνωσης με το κωδικόνιο ATG που αποτελεί το σημείο αναφοράς. 123

Εικόνα 6.4: Αρίθμηση των νουκλεοτιδίων μιας αλληλουχίας ως προς το κωδικόνιο αναφοράς. 123

Εικόνα 6.5: Σύγκριση μεταξύ ομάδων ταξινομητών που δημιουργήθηκαν χρησιμοποιώντας τα χαρακτηριστικά της προτεινόμενης μεθόδου ή τα χαρακτηριστικά της μελέτης [Liu et al., 2004] (ράβδοι σφάλματος: ±τυπική απόκλιση). 126

Εικόνα 6.6: Σύγκριση της προτεινόμενης προσέγγισης με τις δύο παραλλαγές της. 127

Εικόνα 6.7: Λειτουργία της μεθόδου StackTIS. 129

Εικόνα 6.8: Εξαγωγή διανυσμάτων συχνοτήτων κωδικονίων για το στοιχείο κωδικής περιοχής. 130

Εικόνα 6.9: Αξιολόγηση των ταξινομητών του στοιχείου κωδικής περιοχής. 139

Εικόνα 6.10: Αξιολόγηση των ταξινομητών του στοιχείου προτύπου συναίνεσης (ράβδοι σφάλματος: ±τυπική απόκλιση). 140

Εικόνα 7.1: Η διαδικασία της αποκοπής και πολυαδενυλίωσης του mRNA. 149

Εικόνα 7.2: Η αρχιτεκτονική της μεθόδου PolyA-iEP. 153

Εικόνα 7.3: Τα στοιχεία γύρω από το σημείο πολυαδενυλίωσης. 154

Εικόνα 7.4: Κατανομή νουκλεοτιδίων στις θέσεις -200 έως +100 γύρω από το σημείο πολύ(A) στο 3΄ άκρο των αλληλουχιών mRNA. 161

Εικόνα 7.5: Κατανομή νουκλεοτιδίων στις αλληλουχίες 5΄ UTR. 162

Εικόνα 7.6: Κατανομή νουκλεοτιδίων στις κωδικές αλληλουχίες. 162

Εικόνα 7.7: Κατανομή νουκλεοτιδίων στα ιντρόνια. 162

Εικόνα 7.8: Συνεισφορά των χαρακτήρων διττής σημασίας στην ισχύ των χι αναδυόμενων προτύπων. 163

Εικόνα 7.9: Απόδοση ταξινόμησης με χρήση των ισχυρότερων χι αναδυόμενων προτύπων. 164

Εικόνα 7.10: Μέση απόσταση αλληλουχιών της θετικής τάξης από κάθε τάξη και υποτάξη. 164

Εικόνα 7.11: Προσαρμοσμένη ακρίβεια των ταξινομητών. 166

Εικόνα 7.12: Ευαισθησία και ειδικότητα των ταξινομητών. 168

Κατάλογος Πινάκων

Πίνακας 2.1: Πίνακας σύγχυσης (confusion matrix). 27

Πίνακας 3.1: Οι προτεινόμενες ταξινομήσεις των έμβιων όντων. 40

Πίνακας 3.2: Τα 20 αμινοξέα. 45

Πίνακας 3.3: Μεγέθη των γονιδιωμάτων και του αριθμού των γονιδίων οργανισμών. 53

Πίνακας 3.4: Ο γενετικός κώδικας. 57

Πίνακας 4.1: Ο αλγόριθμος εξόρυξης αμοιβαίως αποκλειόμενων αντικειμένων. 89

Πίνακας 4.2: Ένα σύνολο δεδομένων καλαθιού αγορών. 90

Πίνακας 4.3: Οι κανόνες συσχέτισης του συνόλου δεδομένων (Πίνακας 4.2). 90

Πίνακας 4.4: Αμοιβαίως αποκλειόμενα αντικείμενα, τα συχνά σύνολα που τα υποστηρίζουν και τα αντίστοιχα μέτρα αμοιβαίου αποκλεισμού. 90

Πίνακας 4.5: Γενικευμένοι κανόνες συσχέτισης βάσει της ταξονομίας στην Εικόνα 4.5. 92

Πίνακας 4.6: Σύνολο δεδομένων SAGE. 94

Πίνακας 5.1: Δομή πίνακα δεδομένων SAGE. 103

Πίνακας 5.2: Ο αλγόριθμος για την υλοποίηση της προτεινόμενης προσέγγισης. 107

Πίνακας 5.3: Αποτελέσματα μελετών για τα σύνολα δεδομένων του κεφαλαίου. 115

Πίνακας 6.1: Σύνολο χαρακτηριστικών. 122

Πίνακας 6.2: Τα 10 πρώτα χαρακτηριστικά της κατάταξης βάσει των δύο μέτρων. 125

Πίνακας 6.3: Αποτελέσματα μελετών για το σύνολο δεδομένων «Verebrates». 128

Πίνακας 6.4: Ο αλγόριθμος που υλοποιεί τη μέθοδο StackTIS. 135

Πίνακας 6.5: Σύγκριση των μεθόδων StackTIS, MANTIS και της τυπικής προσέγγισης με ταξινομητή επιπέδου-1 τον ταξινομητή MLR. 144

Πίνακας 6.6: Σύγκριση των μεθόδων StackTIS, MANTIS και της τυπικής προσέγγισης με ταξινομητή επιπέδου-1 τον ταξινομητή M5΄. 144

Πίνακας 6.7: Στατιστική σύγκριση της μεθόδου StackTIS με τη μέθοδο MANTIS. 145

Πίνακας 6.8: Αποτελέσματα μελετών για το σύνολο δεδομένων «H. sapiens 1». 145

Πίνακας 7.2: Το αλφάβητο των προτύπων. 155

Πίνακας 7.3: Παράδειγμα πίνακα με τις συχνότητες των νουκλεοτιδίων. 158

Πίνακας 7.4: Παράδειγμα πίνακα με τις κατατάξεις των νουκλεοτιδίων. 158

Πίνακας 7.4: Ο αλγόριθμος που υλοποιεί τη μέθοδο PolyA-iEP. 159

Πίνακας 7.5: Σύνολα Αλληλουχιών. 160

Πίνακας 7.6: Απόδοση των ταξινομητών της μεθόδου PolyA-iEP. 167

Πίνακας 7.7: Στατιστική σύγκριση των ταξινομητών για την ταξινόμηση θετικές/σύνολο αρνητικών (+/-/0: ο ταξινομητής της γραμμής υπερέχει/υστερεί/δε διαφέρει). 168

Εισαγωγή

1

Η παρούσα διατριβή έχει ως αντικείμενο μελέτης την ανακάλυψη γνώσης από βιολογικά δεδομένα και εντάσσεται στα πλαίσια των ερευνητικών περιοχών της ανακάλυψης γνώσης από βάσεις δεδομένων και της βιοπληροφορικής. Οι βασικοί στόχοι της διατριβής και οι άξονες πάνω στους οποίους κινείται είναι οι εξής:

- Η επινόηση νέων και η τροποποίηση ή επέκταση υπαρχουσών μεθόδων ανακάλυψης γνώσης για την ανάλυση βιολογικών δεδομένων.
- Η εφαρμογή νέων ή υπαρχουσών μεθόδων ανακάλυψης γνώσης για την ανάλυση βιολογικών δεδομένων.

Στις επόμενες παραγράφους παρέχεται μια σύντομη εισαγωγή στα ερευνητικά πεδία που είναι συναφή με το αντικείμενο της διατριβής και παρουσιάζονται τα βασικά ανοιχτά προβλήματα που αποτέλεσαν το κίνητρο για την εκπόνηση της. Επίσης, περιγράφεται η συμβολή και η δομή της διατριβής.

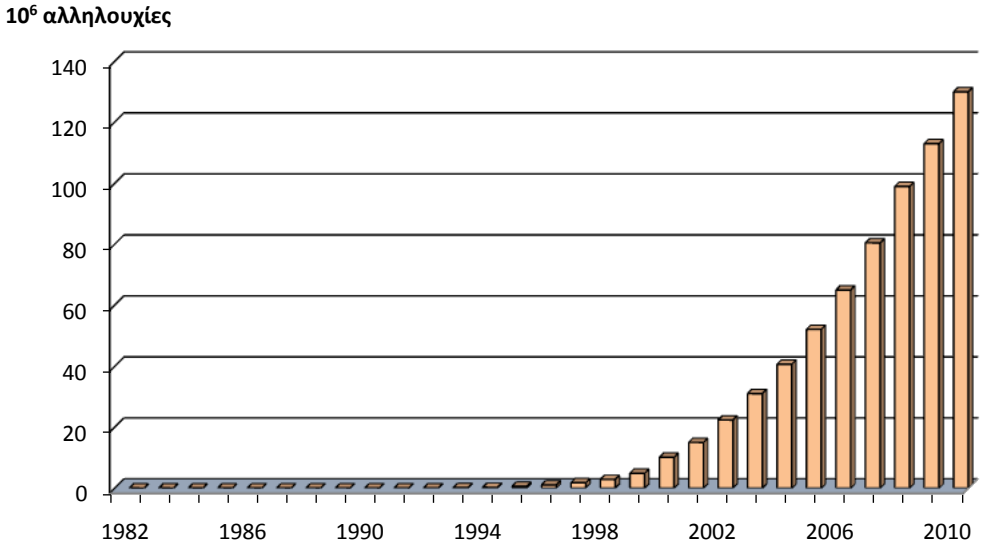
1.1 Κίνητρο Διατριβής και Συναφή Ερευνητικά Πεδία

Κατά τη διάρκεια των τελευταίων δεκαετιών, δύο ξεχωριστές επιστήμες, η πληροφορική και η βιολογία, έχουν χαρακτηριστεί από σημαντικές εξελίξεις, οι οποίες προσέλκυσαν το ενδιαφέρον όλης της ανθρωπότητας. Η ανάπτυξη του Παγκόσμιου Ιστού και η επιτυχής διεξαγωγή του προγράμματος αποκρυπτογράφησης του ανθρώπινου γονιδιώματος αποτελούν δύο αντιπροσωπευτικά επιτεύγματα που αντανακλούν την έκταση της προόδου αυτών των επιστημών. Ωστόσο, οι πορείες της εξέλιξης αυτών των δύο επιστημών τα τελευταία χρόνια δεν είναι ανεξάρτητες. Η ανάγκη συνεργασίας μεταξύ των επιστημόνων της βιολογίας και της πληροφορικής αυξάνεται διαρκώς, καθώς οι δύο επιστήμες εξελίσσονται και καθώς ανακύπτουν νέα επιστημονικά ερωτήματα. Αυτή η συνεργασία έχει ως αποτέλεσμα μια αμφίδρομη επίδραση στην εξέλιξη της κάθε

επιστήμης. Η πρόοδος της πληροφορικής, μεν, προσφέρει στους βιολόγους τα μέσα για τη διαχείριση και ανάλυση των βιολογικών δεδομένων με αποτέλεσμα την πρόοδο της βιολογίας. Η εξέλιξη της βιολογίας, δε, παρέχει νέα ερωτήματα που απαιτούν την εξέλιξη της πληροφορικής για να απαντηθούν.

Η βιοπληροφορική είναι μια νέα διεπιστημονική ερευνητική περιοχή, που έχει εμφανιστεί για να καλύψει την ανάγκη συνεργασίας των επιστημόνων της βιολογίας και της πληροφορικής. Είναι ένα πολλά υποσχόμενο πεδίο και έχει σκοπό να εξασφαλίσει τα μέσα για την ανάλυση και ερμηνεία των τεράστιων όγκων βιολογικών δεδομένων, συνεισφέροντας με αυτόν τον τρόπο στην ανάπτυξη και άλλων συγγενικών επιστημών, όπως είναι η ιατρική.

Αποτέλεσμα της αλματώδους εξέλιξης των επιστημών της πληροφορικής και της βιολογίας είναι η συλλογή τεράστιων όγκων βιολογικών δεδομένων. Η συμβολή της τεχνολογίας σε αυτήν την εκρηκτική αύξηση των συσσωρευμένων δεδομένων είναι διττή και καθορίζεται μέσω της αμφίδρομης σχέσης της με τη βιολογία. Πρώτον, η βελτίωση των τεχνολογιών έχει οδηγήσει στην επίσπευση των διαδικασιών παραγωγής και συλλογής δεδομένων. Δεύτερον, η δημιουργία νέων βιολογικών ερωτημάτων, που οδηγούν στην ανακάλυψη και εφαρμογή νέων τεχνολογιών, έχει ανοίξει το δρόμο σε διαδικασίες παραγωγής και συλλογής νέων μορφών δεδομένων. Αντιπροσωπευτικό παράδειγμα της ταχείας συσσώρευσης βιολογικών δεδομένων αποτελεί η εκθετική αύξηση του αριθμού των αλληλουχιών που διατηρούνται στη δημοφιλή βάση βιολογικών δεδομένων GenBank (Εικόνα 1.1).



Εικόνα 1.1: Εξέλιξη του μεγέθους της βάσης δεδομένων GenBank (1982-2010)[1].

Οι μεγάλοι όγκοι και οι νέες μορφές των δεδομένων, καθώς και οι ιδιαιτερότητες που μπορεί να παρουσιάζουν, όπως η παρουσία θορύβου, ο μεγάλος αριθμός διαστά-

[1] http://www.ncbi.nlm.nih.gov/Sitemap/Summary/statistics.html (προσπελάστηκε: 17/05/2011).

σεων, κ.α., προβάλλουν νέες προκλήσεις που σχετίζονται με τη διαχείριση και την ανάλυσή τους. Οι παραδοσιακές τεχνικές συχνά αδυνατούν να χειριστούν αποδοτικά και αποτελεσματικά τα δεδομένα αυτά. Ωστόσο, για να επιτευχθεί η επιστημονική πρόοδος, δεν αρκεί μόνο η συλλογή των δεδομένων. Τα συγκεντρωμένα δεδομένα πρέπει να αναλυθούν με σκοπό την ανακάλυψη της γνώσης που πιθανώς κρύβουν. Μάλιστα, οι μέθοδοι ανάλυσης που θα χρησιμοποιηθούν πρέπει να είναι αποδοτικές, καθώς αυτό απαιτεί ο μεγάλος όγκος των δεδομένων και φυσικά πρέπει να είναι αποτελεσματικές, ώστε να παράγεται χρήσιμη γνώση, η οποία μπορεί να αξιοποιηθεί κατάλληλα στη συνέχεια. Μια ερευνητική περιοχή της πληροφορικής, η ανακάλυψη γνώσης από βάσεις δεδομένων, προσφέρει στους βιολόγους, αλλά και σε άλλους επιστήμονες, ένα σύνολο ισχυρών εργαλείων για την ανάλυση νέων τύπων δεδομένων με σκοπό την αποδοτική και αποτελεσματική εξαγωγή γνώσης. Αυτά τα εργαλεία συνδυάζουν τεχνικές από διάφορες περιοχές, όπως είναι τα μαθηματικά, η στατιστική, η τεχνητή νοημοσύνη και η τεχνολογία βάσεων δεδομένων. Απώτερος σκοπός της σύντηξης όλων αυτών των τεχνολογιών είναι η αντιμετώπιση των εμποδίων και των περιορισμών που τίθενται από τις παραδοσιακές μεθόδους ανάλυσης δεδομένων.

1.1.1 Τάσεις της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων

Στα τέλη του 2005 οι Qiang Yang και Xindong Wu ανέλαβαν την πρωτοβουλία να προσδιορίσουν τα δέκα σημαντικότερα ανοιχτά προβλήματα της ανακάλυψης γνώσης από βάσεις δεδομένων [Yang & Wu, 2006]. Στην προσπάθεια αυτή ζήτησαν τη συνεισφορά των μελών των οργανωτικών επιτροπών των συνεδρίων IEEE International Conference on Data Mining (ICDM) και ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). Από την έρευνα αυτή, στην οποία ανταποκρίθηκαν 14 ερευνητές[2], προσδιορίστηκαν τα ακόλουθα δέκα προβλήματα ως τα σημαντικότερα:

1. Ανάπτυξη μιας ενιαίας θεωρίας ανακάλυψης γνώσης.
2. Κλιμάκωση για πολυδιάστατα δεδομένα και ροές δεδομένων υψηλής ταχύτητας.
3. Ανακάλυψη γνώσης από δεδομένα ακολουθιών και χρονοσειρών.
4. Ανακάλυψη πολύπλοκης γνώσης από πολύπλοκα δεδομένα, όπως για παράδειγμα δεδομένα από πολλαπλούς πίνακες (multi-relational), δεδομένα γράφων και μη σχεσιακά δεδομένα (κείμενο, ιστοσελίδες, εικόνες, κλπ.).
5. Ανακάλυψη γνώσης σε δικτυακό περιβάλλον.
6. Ανακάλυψη γνώσης από κατανεμημένα δεδομένα και δεδομένα πολλαπλών πρακτόρων.
7. Ανακάλυψη γνώσης για βιολογικά και περιβαλλοντικά προβλήματα.

[2] Pedro Domingos, Charles Elkan, Johannes Gehrke, Jiawei Han, David Heckerman, Daniel Keim, Jiming Liu, David Madigan, Gregory Piatetsky-Shapiro, Vijay V. Raghavan, Rajeev Rastogi, Salvatore J. Stolfo, Alexander Tuzhilin και Benjamin W. Wah.

8. Προβλήματα σχετικά με τις διαδικασίες ανακάλυψης γνώσης.
9. Ασφάλεια, προστασία απορρήτου και ακεραιότητα δεδομένων.
10. Χειρισμός δεδομένων τα οποία δεν είναι στατικά, ή έχουν άνιση κατανομή των τάξεων, ή έχουν διαφορετικό κόστος λανθασμένης ταξινόμησης μεταξύ διαφορετικών τάξεων.

Το αντικείμενο της παρούσας διατριβής κινείται στα πλαίσια του έβδομου ερευνητικού προβλήματος από αυτά που καταγράφονται στην παραπάνω λίστα. Ωστόσο, η έρευνα που πραγματοποιήθηκε στα πλαίσια της διατριβής συχνά ήρθε αντιμέτωπη και με κάποια από τα άλλα ερευνητικά προβλήματα που προσδιορίστηκαν παραπάνω. Για παράδειγμα, τα δεδομένα γονιδιακής έκφρασης, που αποτελούν ένα από τα αντικείμενα μελέτης της διατριβής, είναι πολυδιάστατα δεδομένα και συνεπώς η αντιμετώπισή τους υπάγεται και στα πλαίσια του δεύτερου προβλήματος. Επίσης, η ανάλυση βιολογικών γονιδιακών αλληλουχιών, που και αυτή αποτελεί αντικείμενο έρευνας της διατριβής, μπορεί να ενταχθεί στα πλαίσια του τρίτου προβλήματος, λόγω της ακολουθιακής φύσης των δεδομένων. Επιπλέον, κατά την εκπόνηση της διατριβής εμφανίστηκαν προβλήματα σχετικά με τις διαδικασίες ανακάλυψης γνώσης (π.χ. καθαρισμός των δεδομένων γονιδιακής έκφρασης), που αποτελούν μέρος του όγδοου προβλήματος της παραπάνω λίστας. Τέλος, το δέκατο πρόβλημα σχετίζεται και αυτό με την παρούσα διατριβή, καθώς σε πολλές από τις προσεγγίσεις ταξινόμησης υπήρχε σημαντική διαφορά ανάμεσα στις συχνότητες εμφάνισης των παραδειγμάτων διαφορετικών τάξεων (ανισοκατανομή των τάξεων).

1.1.2 Τάσεις της Βιοπληροφορικής

Η βιοπληροφορική διαδραματίζει σημαντικό ρόλο τόσο στην αποκρυπτογράφηση των δεδομένων που παράγονται από τις σύγχρονες τεχνολογίες της γονιδιωματικής, της πρωτεωμικής και της ανάλυσης γονιδιακής έκφρασης, όσο και στη διαχείριση και ανάλυση των δεδομένων που παράγονται από τις παραδοσιακές μεθόδους της βιολογίας και της ιατρικής. Η εξέλιξη της βιοπληροφορικής ξεκίνησε από την ανάλυση μεμονωμένων αλληλουχιών και πλέον κινείται προς την κατεύθυνση της ανάλυσης ολόκληρου του γονιδιώματος (genome), του μεταγραφώματος (trascriptome) και του πρωτεώματος (proteome), που αποτελούν αντίστοιχα τα σύνολα των γονιδιακών αλληλουχιών, των μεταφρασμένων αλληλουχιών (RNA) και των πρωτεϊνών ενός οργανισμού. Υπαίτιος αυτής της εξέλιξης, όπως έχει ήδη αναφερθεί, είναι η ραγδαία ανάπτυξη της τεχνολογίας. Η εξέλιξη της βιοπληροφορικής δημιούργησε νέες προοπτικές για εφαρμογές με σημαντικό αντίκτυπο στην κοινωνία, όπως είναι η μοριακή και η εξατομικευμένη ιατρική. Μάλιστα, αυτές οι νέες εφαρμογές χαράσσουν το δρόμο στη βιοπληροφορική και την οδηγούν προς την κατεύθυνση της συστημικής βιολογίας (systems biology). Η προσέγγιση της συστημικής βιολογίας αποτελεί ένα διεπιστημονικό πεδίο έρευνας που εστιάζει στη μελέτη των πολύπλοκων αλληλεπιδράσεων μεταξύ των βιολογικών συστη-

μάτων, υπό το πρίσμα μιας ολιστικής θεώρησης των πραγμάτων. Αυτές οι νέες προοπτικές εξέλιξης της βιοπληροφορικής απαιτούν την ενοποίηση και ανάλυση γονιδιακών, μοριακών, κυτταρικών και κλινικών δεδομένων, θέτοντας μια σειρά από προκλήσεις στους ερευνητές.

Η πρόβλεψη της δομής των πρωτεϊνών, η αναζήτηση ομολογιών, η πολλαπλή στοίχιση αλληλουχιών, η κατασκευή φυλογενετικών δένδρων, η ανάλυση μεταβολικών μονοπατιών, η ανάλυση και ο υπομνηματισμός (annotation) γονιδιωμάτων, η πρόβλεψη γονιδίων και η ανάλυση των επιπέδων γονιδιακής έκφρασης αποτελούν μερικά από τα πλέον σημαντικά ερευνητικά θέματα που απασχολούν τους ερευνητές της βιοπληροφορικής. Οι προκλήσεις που τίθενται από τα παραπάνω ερευνητικά προβλήματα έθεσαν το κίνητρο για την εκπόνηση της παρούσας διατριβής. Ένα από τα αντικείμενα της διατριβής είναι η ανάλυση βιολογικών αλληλουχιών (πρόβλεψη σημείου έναρξης της μετάφρασης και πρόβλεψη σημείου αποκοπής και πολυαδενυλίωσης), η οποία σχετίζεται με την ανάλυση και τον υπομνηματισμό γονιδιωμάτων, καθώς και με την πρόβλεψη γονιδίων. Επιπλέον, αντικείμενο της διατριβής αποτελεί και η ανάλυση των επιπέδων γονιδιακής έκφρασης.

1.2 Συμβολή και Δομή της Διατριβής

Το κείμενο της διατριβής είναι δομημένο σε οκτώ κεφάλαια, συμπεριλαμβανομένου του παρόντος εισαγωγικού κεφαλαίου. Τα επόμενα δύο κεφάλαια (κεφάλαια 2 και 3) παρέχουν το βασικό επιστημονικό και τεχνικό υπόβαθρο των ερευνητικών περιοχών στις οποίες εντάσσεται η διατριβή. Δόθηκε έμφαση τόσο στην περιγραφή των βασικών εννοιών για την κατανόηση της ερευνητικής εργασίας που πραγματοποιήθηκε, όσο και στην κατά το δυνατόν πληρέστερη, πλην όμως συνοπτική, παρουσίαση των συγκεκριμένων ερευνητικών πεδίων. Τα κεφάλαια 4, 5, 6 και 7 παρουσιάζουν τις τέσσερεις προσεγγίσεις που προτείνονται στα πλαίσια της διατριβής και το κεφάλαιο 8 αποτελεί τον επίλογο. Στις επόμενες παραγράφους περιγράφονται με περισσότερες λεπτομέρειες τα περιεχόμενα κάθε κεφαλαίου.

Στο δεύτερο κεφάλαιο, με τίτλο «Ανακάλυψη Γνώσης από Βάσεις Δεδομένων», παρουσιάζονται τα στάδια που περιλαμβάνει η διαδικασία ανακάλυψης γνώσης από βάσεις δεδομένων καθώς και κάποιοι από τους βασικότερους αλγορίθμους που χρησιμοποιούνται σε καθένα από τα στάδια. Η πλειοψηφία των αλγορίθμων αυτών εφαρμόστηκαν, τροποποιήθηκαν, ή επεκτάθηκαν κατά την εκπόνηση της παρούσας διατριβής. Επίσης, περιγράφηκε η σχέση της ανακάλυψης γνώσης με τη συγγενική περιοχή της μηχανικής μάθησης καθώς και κάποια προβλήματα που αντιμετωπίζονται κατά την εφαρμογή της διαδικασίας. Τέλος, παρουσιάζονται μερικά από τα βασικότερα πεδία εφαρμογής της διαδικασίας.

Στο τρίτο κεφάλαιο, με τίτλο «Μοριακή Βιολογία και Βιοπληροφορική», παρουσιάζονται τα επιστημονικά πεδία της μοριακής βιολογίας και της βιοπληροφορικής. Γίνεται αναφορά σε κάποιες βασικές έννοιες που συναντώνται στη μοριακή βιολογία και

παρουσιάζονται οι σημαντικότερες τεχνολογίες που έχουν αναπτυχθεί και εφαρμόζονται στα πλαίσια της βιοπληροφορικής. Επιπλέον, περιγράφονται διάφορες κατηγορίες βιολογικών βάσεων δεδομένων, ενώ δίνεται έμφαση στην παρουσίαση των δημοφιλέστερων εφαρμογών της ανακάλυψης γνώσης σε διάφορα είδη βιολογικών δεδομένων.

Στο τέταρτο κεφάλαιο, με τίτλο «Ανακάλυψη Αμοιβαίως Αποκλειόμενων Αντικειμένων», παρουσιάζεται ένα νέο είδος προβλήματος στα πλαίσια της περιοχής της ανακάλυψης γνώσης από βάσεις δεδομένων, που πραγματεύεται την ανακάλυψη αμοιβαίως αποκλειόμενων αντικειμένων. Τα αμοιβαίως αποκλειόμενα αντικείμενα αποτελούν ένα ενδιαφέρον είδος συσχετίσεων που είναι δυνατό να οδηγήσουν στην ανακάλυψη ιεραρχικών σχέσεων. Στο κεφάλαιο αυτό, ορίζεται το πρόβλημα την ανακάλυψης αμοιβαίως αποκλειόμενων αντικειμένων, προτείνεται ένας αλγόριθμος ανακάλυψης των αμοιβαίως αποκλειόμενων αντικειμένων από μια βάση δεδομένων συναλλαγών και παρουσιάζεται η εφαρμογή του σε βιολογικά δεδομένα που αφορούν την έκφραση γονιδίων.

Στο πέμπτο κεφάλαιο, με τίτλο «Πρόβλεψη Φαινοτύπου από Δεδομένα Γονιδιακής Έκφρασης», παρουσιάζεται μια προσέγγιση που αναπτύχθηκε στα πλαίσια της διατριβής για την αποτελεσματική ταξινόμηση δεδομένων γονιδιακής έκφρασης. Το βασικό χαρακτηριστικό των δεδομένων γονιδιακής έκφρασης είναι ο μεγάλος αριθμός των χαρακτηριστικών και ο σχετικά μικρός αριθμός παραδειγμάτων. Αυτή η ιδιαιτερότητα των δεδομένων γονιδιακής έκφρασης προβάλλει σημαντικές προκλήσεις στους επιστήμονες της ανακάλυψης γνώσης, οι οποίοι καλούνται να αναπτύξουν μεθόδους που διατηρούν υψηλά επίπεδα αποτελεσματικότητας, μολονότι έχουν στη διάθεσή τους ελάχιστα παραδείγματα και πολλά χαρακτηριστικά. Στην προτεινόμενη προσέγγιση αξιοποιείται η πληροφορία που παρέχεται από τις συσχετίσεις μεταξύ των γονιδιακών εκφράσεων με σκοπό τη βελτίωση της απόδοσης των ταξινομητών. Τα αποτελέσματα των πειραμάτων που διεξήχθησαν έδειξαν ότι η προτεινόμενη μέθοδος παρέχει σημαντική βελτίωση της ακρίβειας ταξινόμησης σε σχέση με άλλες προσεγγίσεις, η οποία σε κάποιες περιπτώσεις φτάνει στο 99%.

Στο έκτο κεφάλαιο, με τίτλο «Πρόβλεψη του Σημείου Έναρξης της Μετάφρασης», παρουσιάζονται μέθοδοι που υλοποιήθηκαν στα πλαίσια της διατριβής για την πρόβλεψη του σημείου έναρξης της μετάφρασης (translation initiation site). Παρουσιάζονται διάφορες προσεγγίσεις που αντιμετωπίζουν το συγκεκριμένο πρόβλημα με πιο αντιπροσωπευτική τη μέθοδο StackTIS, η οποία είναι η πιο ολοκληρωμένη και αποτελεσματική από τις προταθείσες προσεγγίσεις. Η μέθοδος StackTIS, αποτελείται από τρία βασικά στοιχεία πρόβλεψης, καθένα από τα οποία είναι προσανατολισμένο σε ένα διαφορετικό μέρος του προβλήματος. Τα τρία επιμέρους στοιχεία συνδυάζονται με μια ιδιαίτερα αποτελεσματική μέθοδο συνδυασμού ταξινομητών, τη στοιβαγμένη γενίκευση. Εκτενή πειράματα έδειξαν ότι η μέθοδος StackTIS παρουσιάζει ιδιαίτερα υψηλή αποτελεσματικότητα η οποία βάσει των μέτρων της ακρίβειας και της προσαρμοσμένης

ακρίβειας αποδεικνύεται πως είναι στατιστικά σημαντική σε σχέση με άλλες δημοφιλείς προσεγγίσεις.

Στο έβδομο κεφάλαιο, με τίτλο «Πρόβλεψη του Σημείου Αποκοπής και Πολυαδενυλίωσης», παρουσιάζεται η μέθοδος PolyA-iEP, η οποία αναπτύχθηκε στα πλαίσια της διατριβής για την πρόβλεψη του σημείου αποκοπής και πολυαδενυλίωσης (cleavage-polyadenylation site). Το συγκεκριμένο πρόβλημα αποτελεί ένα σημαντικό πεδίο έρευνας, που προσελκύει την προσοχή ερευνητών της βιοπληροφορικής, της βιολογίας και της ιατρικής. Το ενδιαφέρον εστιάζεται στην αποτελεσματική πρόβλεψη του σημείου πολυαδενυλίωσης και στην ανακάλυψη νέων προτύπων και σημάτων γύρω από αυτό. Η μέθοδος που παρουσιάζεται κινείται και προς τις δύο κατευθύνσεις. Πρόσφατη μελέτη [Mayr & Bartel, 2009] συσχετίζει την πολυαδενυλίωση με τον καρκίνο, καθιστώντας σαφές ότι η έρευνα του συγκεκριμένου προβλήματος αναμένεται στο μέλλον να αποτελέσει ένα από τα πιο ενδιαφέροντα πεδία της βιοπληροφορικής. Η αξιολόγηση της μεθόδου PolyA-iEP έδειξε ότι η αποτελεσματικότητα της είναι πολύ μεγάλη αγγίζοντας το 93,7% σε ευαισθησία (sensitivity) και το 88,2% σε ειδικότητα (specificity).

Τέλος, στο όγδοο κεφάλαιο, με τίτλο «Επίλογος», παρουσιάζονται συμπεράσματα σχετικά με τα πεπραγμένα της διατριβής, καθώς και ιδέες και σκέψεις για μελλοντικές επεκτάσεις.

Ανακάλυψη Γνώσης από Βάσεις Δεδομένων

2

Δύο έννοιες που απασχολούσαν τον άνθρωπο από την αρχαιότητα και υπήρξαν αντικείμενο έρευνας και φιλοσοφικού στοχασμού είναι η «μάθηση» και η «γνώση». Σήμερα δύο συγγενικές ερευνητικές περιοχές, η μηχανική μάθηση και η ανακάλυψη γνώσης από βάσεις δεδομένων προσπαθούν να επιλύσουν σύγχρονα προβλήματα, που αφορούν τη δυνατότητα μάθησης των μη έμβιων όντων, δηλαδή των μηχανών και την αποδοτική ανακάλυψη γνώσης από μεγάλες συλλογές δεδομένων.

Η *μηχανική μάθηση* (*machine learning*) είναι μια από τις παλαιότερες ερευνητικές περιοχές της *τεχνητής νοημοσύνης* (*artificial intelligence*). Σκοπός της είναι η κατασκευή υπολογιστικών συστημάτων που μπορούν να προσαρμοστούν και να μάθουν από την εμπειρία τους [Dietterich, 1999]. Η εμφάνιση της μηχανικής μάθησης τοποθετείται στα μέσα του περασμένου αιώνα. Όμως η ραγδαία αναπτυξή της, που οδήγησε στην εμφάνιση πολλών νέων ερευνητικών κατευθύνσεων, πραγματοποιήθηκε κατά τη δεκαετία του 1980.

Στα τέλη της δεκαετίας του 1980, μια νέα έννοια με την ονομασία *εξόρυξη από δεδομένα* (*data mining*)[3] άρχισε να κάνει την εμφάνισή της. Σκοπός της νέας αυτής επιστημονικής περιοχής ήταν η εξαγωγή γνώσης από μεγάλους όγκους δεδομένων με τη βοήθεια ηλεκτρονικών υπολογιστών. Η συγκεκριμένη ερευνητική περιοχή έχει λάβει στοιχεία από πολλές επιστημονικές περιοχές, όπως τη στατιστική, τη μηχανική μάθηση,

[3] Ο όρος *data mining* (στα ελληνικά κατά λέξη μεταφράζεται ως *εξόρυξη δεδομένων*) οφείλει το όνομά του στις ομοιότητες μεταξύ της αναζήτησης πολύτιμων πληροφοριών σε μια μεγάλη βάση δεδομένων και της αναζήτησης πολύτιμων μεταλλευμάτων σε πετρώματα. Και στις δυο περιπτώσεις γίνεται επιλογή από μεγάλες ποσότητες υλικού. Ωστόσο, η χρήση της έννοιας data mining δεν είναι απόλυτα ορθή. Η αναζήτηση για χρυσό στα πετρώματα ορθώς καλείται εξόρυξη χρυσού (gold mining) και όχι εξόρυξη πετρωμάτων (rock mining). Συνεπώς, αναλογικά, η εξόρυξη δεδομένων θα έπρεπε να είχε ονομαστεί εξόρυξη γνώσης (knowledge mining). Εντούτοις, ο όρος data mining έχει γίνει αποδεκτός και χρησιμοποιείται ευρέως. Στην παρούσα διατριβή η απόδοση του όρου στα ελληνικά είναι *εξόρυξη από δεδομένα* και όχι *εξόρυξη δεδομένων*.

την αναγνώριση προτύπων, την τεχνολογία βάσεων δεδομένων, την ανάκτηση πληροφορίας, την οπτική αναπαράσταση δεδομένων, κ.α. Αρχικά, η έννοια του όρου «εξόρυξη από δεδομένα» είχε ταυτιστεί με την έννοια του όρου «ανακάλυψη γνώσης από βάσεις δεδομένων» ("knowledge discovery in databases" - *KDD*). Στα μέσα της δεκαετίας του 1990 έγινε σαφής ο διαχωρισμός των δύο εννοιών, ώστε η ανακάλυψη γνώσης από βάσεις δεδομένων να αναφέρεται σε μια διαδικασία που αποτελείται από κάποια στάδια, ένα εκ των οποίων είναι η εξόρυξη από δεδομένα (Εδάφιο 2.1.1). Ωστόσο, ακόμη και σήμερα οι δύο έννοιες χρησιμοποιούνται πολλές φορές ως ταυτόσημες.

2.1 Η Διαδικασία Ανακάλυψης Γνώσης από Βάσεις Δεδομένων

Η ανακάλυψη γνώσης από βάσεις δεδομένων αναφέρεται σε μια διαδικασία που αποτελείται από κάποια στάδια και σύμφωνα με τους Fayyad et al. [1996] ορίζεται ως εξής:

> *Η ανακάλυψη γνώσης από βάσεις δεδομένων είναι η μη τετριμμένη διαδικασία αναγνώρισης έγκυρων, πρωτότυπων, δυνητικά χρήσιμων και απόλυτα κατανοητών προτύπων στα δεδομένα.*

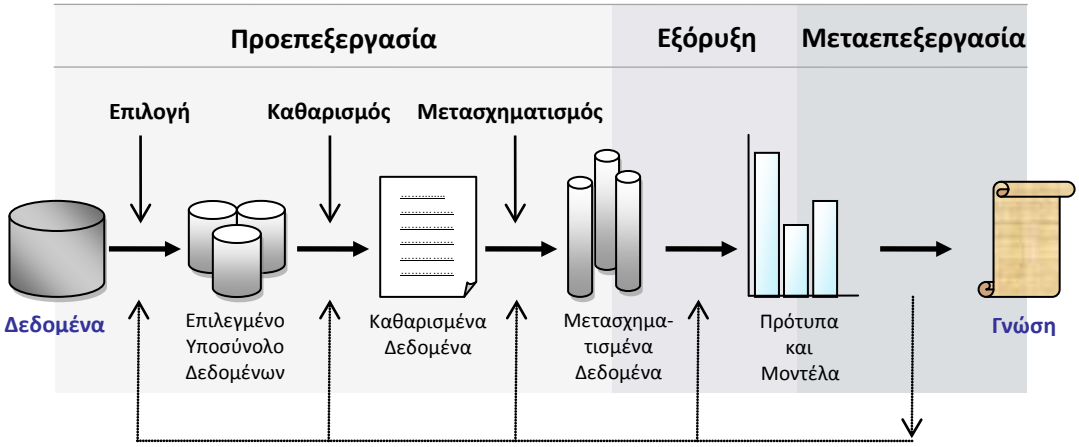
Σύμφωνα με τον παραπάνω ορισμό, η εφαρμογή της διαδικασίας ανακάλυψης γνώσης από βάσεις δεδομένων έχει ως αποτέλεσμα την παραγωγή *προτύπων* (*patterns*). Ωστόσο, η διαδικασία της ανακάλυψης γνώσης περιλαμβάνει και τη δημιουργία *μοντέλων* (*models*), η οποία δεν αναφέρεται άμεσα στον ορισμό, αλλά υπονοείται, καθώς ένα μοντέλο μπορεί να θεωρηθεί μια αφηρημένη αναπαράσταση πολλών προτύπων. Για παράδειγμα, αν η συνάρτηση $f(x) = 5x^2 + 2x$ είναι ένα πρότυπο, τότε η συνάρτηση $g(x) = ax^2 + \beta x$ είναι το μοντέλο από το οποίο μπορεί να προκύψει το πρότυπο f.

Τα πρότυπα που προκύπτουν από τη διαδικασία ανακάλυψης γνώσης από βάσεις δεδομένων πρέπει να είναι έγκυρα, ώστε να ισχύουν και σε νέα δεδομένα με κάποιο βαθμό βεβαιότητας. Η βεβαιότητα μπορεί να μετρηθεί με μια συνάρτηση η οποία αντιστοιχεί σε κάθε πρότυπο μια τιμή βεβαιότητας. Η πρωτοτυπία των προτύπων μπορεί να μετρηθεί σε σχέση με τις αλλαγές στα δεδομένα (συγκρίνοντας τρέχουσες με προηγούμενες ή αναμενόμενες τιμές) ή τη γνώση (κατά πόσο ένα νέο εύρημα σχετίζεται με τα προηγούμενα) χρησιμοποιώντας μια συνάρτηση που προσδιορίζει το βαθμό πρωτοτυπίας ενός προτύπου. Η χρησιμότητα των προτύπων, επίσης, εκτιμάται με μια συνάρτηση χρησιμότητας. Τα πρότυπα θα πρέπει να είναι αρκετά κατανοητά στους ανθρώπους, ώστε να διευκολύνεται η κατανόηση των δεδομένων που περιγράφονται από αυτά. Καθώς είναι δύσκολη μια τέτοια μέτρηση, συνήθως μετράται η απλότητα, που μπορεί να ποικίλλει από καθαρά συντακτική (π.χ. το μέγεθος του προτύπου σε bits), έως σημασιολογική (π.χ. ευκολία αντίληψης του προτύπου από τον άνθρωπο σε δεδομένη κατάσταση).

Η συνολική εκτίμηση του βαθμού στον οποίο ένα πρότυπο είναι ενδιαφέρον (*interestingness*) μπορεί να υπολογιστεί με μια συνάρτηση η οποία λαμβάνει υπόψη τις μετρήσεις για όλα τα προαναφερθέντα επιθυμητά χαρακτηριστικά του προτύπου. Ένα πρότυπο είναι *γνώση* αν η τιμή της συνάρτησης ενδιαφέροντος ξεπερνά το κατώφλι που ορίζει ο χρήστης.

2.1.1 Στάδια της Διαδικασίας Ανακάλυψης Γνώσης από Βάσεις Δεδομένων

Πριν από την εφαρμογή της διαδικασίας ανακάλυψης γνώσης από βάσεις δεδομένων είναι απαραίτητος ο καθορισμός του προβλήματος και η κατανόηση του τομέα στον οποίο θα εφαρμοστεί. Απαιτείται να καθοριστούν οι στόχοι και να ληφθούν υπόψη όλοι οι διαθέσιμοι πόροι, οικονομικοί και ανθρώπινοι, καθώς και οι περιορισμοί που ενδεχομένως υπάρχουν. Η συνεργασία του ειδικού επί θεμάτων ανακάλυψης γνώσης με τον ειδικό του τομέα του προβλήματος είναι απαραίτητη προϋπόθεση για τον καθορισμό των παραπάνω λεπτομερειών και για τον προσδιορισμό του βαθμού που οι στόχοι είναι εφικτοί και αξιοποιήσιμοι. Συχνά κάποια από τα βήματα της ανακάλυψης γνώσης είναι αναγκαίο να επαναληφθούν, λόγω προβλημάτων που προκύπτουν εξαιτίας των αρχικών επιλογών. Τα επιμέρους στάδια της διαδικασίας ανακάλυψης γνώσης απεικονίζονται στην Εικόνα 2.1 και περιγράφονται παρακάτω σύμφωνα με τα [Dunham, 2002; Βλαχάβας *κ.α.*, 2006].



Εικόνα 2.1: Τα βασικά στάδια της διαδικασίας ανακάλυψης γνώσης.

Επιλογή Δεδομένων

Τα δεδομένα μπορούν να ληφθούν από διαφορετικές και ετερογενείς πηγές. Συνήθως προέρχονται από σχεσιακές βάσεις δεδομένων. Επειδή τις περισσότερες φορές δεν είναι οργανωμένα με τρόπο που να διευκολύνει τη διαδικασία ανακάλυψης γνώσης, απαιτείται η εξαγωγή των δεδομένων και η οργάνωσή τους σε απλούστερες δομές. Η επιλογή των δεδομένων μπορεί να διευκολυνθεί αν τα δεδομένα βρίσκονται σε μια *αποθήκη δεδομένων (data warehouse*), η οποία παρέχει στους αλγόριθμους ανακάλυψης γνώσης μια ευκολότερα προσβάσιμη όψη (view) των δεδομένων. Εκτός από τα τεχνικής φύσεως προβλήματα στο στάδιο αυτό μπορούν να αντιμετωπισθούν και

νομικής φύσεως προβλήματα, όπως για παράδειγμα κωλύματα που αφορούν την προστασία προσωπικών δεδομένων.

Καθαρισμός Δεδομένων

Τα δεδομένα ενδέχεται να είναι ελλιπή ή λανθασμένα. Συχνά υπάρχουν διαφορετικοί τύποι δεδομένων και διαφορετικές μονάδες μέτρησης για το ίδιο είδος πληροφορίας. Συνεπώς είναι αναγκαία η διόρθωση ή η απομάκρυνση των λανθασμένων δεδομένων και η απόκτηση ή η πρόβλεψη των δεδομένων που λείπουν. Το στάδιο αυτό, λόγω της φύσης των εργασιών που λαμβάνουν χώρα, ονομάζεται *στάδιο καθαρισμού των δεδομένων* (*data cleaning*). Στη βιβλιογραφία συναντάται και ως στάδιο *προεπεξεργασίας δεδομένων* (*data preprocessing*), ωστόσο, η χρήση του τελευταίου όρου είναι καταχρηστική, καθώς η προεπεξεργασία των δεδομένων ουσιαστικά περιλαμβάνει και άλλα στάδια επιπλέον του καθαρισμού.

Μετασχηματισμός Δεδομένων

Τα δεδομένα μετασχηματίζονται με σκοπό να διευκολύνουν την εξόρυξη από τα δεδομένα και να παρέχουν πιο κατανοητά αποτελέσματα. Δεδομένα από διαφορετικές πηγές μετατρέπονται σε μια κοινή μορφή που επιτρέπει την επεξεργασία τους. Επιπλέον, πολλοί αλγόριθμοι εξόρυξης απαιτούν συγκεκριμένες δομές δεδομένων με αποτέλεσμα να επιβάλλεται η προσαρμογή των αρχικών δεδομένων σε αυτές τις δομές. Παραδείγματα μετασχηματισμών που μπορούν να γίνουν είναι τα εξής:

- Επιλογή ή συγχώνευση χαρακτηριστικών, ώστε να μειωθεί η πολυπλοκότητα των δεδομένων.
- Αντικατάσταση ενός χαρακτηριστικού από άλλο.
- Μετατροπή συνεχών τιμών σε διακριτές τιμές (διακριτοποίηση).
- Απομάκρυνση σπάνια εμφανιζόμενων ακραίων τιμών, όπως είναι τα παράτυπα σημεία (outliers)[4].
- Μετασχηματισμός με εφαρμογή κάποιας συνάρτησης (π.χ. λογαριθμικής) στις τιμές ενός χαρακτηριστικού.

Όλες οι παραπάνω τεχνικές διευκολύνουν τη διαδικασία εξόρυξης, είτε μειώνοντας τον αριθμό των χαρακτηριστικών (*dimensionality reduction*), είτε μειώνοντας τον αριθμό των τιμών που παίρνει ένα χαρακτηριστικό (*variability reduction*).

Εξόρυξη από Δεδομένα

Σε αυτό το βήμα, εφαρμόζεται ο κατάλληλος αλγόριθμος πάνω στα επεξεργασμένα δεδομένα. Η επιλογή του αλγορίθμου εξαρτάται από το είδος της γνώσης που αναζη-

[4] Αξίζει να σημειωθεί ότι ενώ στις περισσότερες εφαρμογές τα παράτυπα σημεία θεωρούνται θόρυβος και απορρίπτονται, σε άλλες εφαρμογές (π.χ. ανίχνευση απάτης) μπορούν να αποκαλύψουν σημαντική γνώση. Η διαδικασία ανακάλυψης παράτυπων σημείων καλείται *εξόρυξη παράτυπων σημείων* (*outlier mining*).

τείται. Συχνά, χρησιμοποιούνται και άλλα κριτίρια επιλογής, όπως είναι η ευκολία ερμηνείας της παραγόμενης γνώσης και η δυνατότητα χρήσης της παραγόμενης γνώσης για την πρόβλεψη των χαρακτηριστικών νέων δεδομένων. Μάλιστα, πολλές φορές αλγόριθμοι που παράγουν εύκολα ερμηνεύσιμη γνώση δεν προσφέρουν υψηλης ακρίβειας πρόβλεψη και αντίστροφα. Συνεπώς, η επιλογή του αλγορίθμου θα πρέπει να γίνει με προσοχή, ώστε να υπάρχει μια εξισορρόπηση μεταξύ των διάφορων κριτηρίων σύμφωνα με τους στόχους που έχουν τεθεί.

Μεταεπεξεργασία

Σε αυτό το βήμα λαμβάνουν χώρα διαδικασίες όπως η ερμηνεία, η αξιολόγηση και γενικότερα η διαχείριση των αποτελεσμάτων. Διάφορες τεχνικές *οπτικής αναπαράστασης* (*visualization*) μπορούν να χρησιμοποιηθούν για την παρουσίαση των δεδομένων. Η κατανόηση της χρησιμότητας των αποτελεσμάτων εξαρτάται σε μεγάλο βαθμό από τον τρόπο παρουσίασής τους. Οι παραπάνω τεχνικές, δίνουν στο χρήστη τη δυνατότητα να συνοψίζει και να εξάγει πιο πολύπλοκα συμπεράσματα από ότι με μαθηματικές ή βασισμένες σε κείμενο περιγραφές των αποτελεσμάτων. Πολλές φορές η γνώση που προκύπτει καταγράφεται σε μια *βάση γνώσης* (*knowledge base*), οπότε ενδέχεται να παρουσιαστεί η ανάγκη επίλυσης συγκρούσεων με προϋπάρχουσα γνώση. Συνήθως τα αποτελέσματα μετά από ένα κύκλο της διαδικασίας δίνουν ερέθισμα για νέες αναζητήσεις με αποτέλεσμα την επανάληψη ολόκληρης της διαδικασίας.

2.2 Μηχανική Μάθηση

Έχουν διατυπωθεί πολλοί ορισμοί κατά καιρούς για την έννοια της μηχανικής μάθησης. Ωστόσο, ένας από τους πιο αυστηρά διατυπωμένους είναι ο εξής [Mitchell, 1997]:

> *Ένα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από την εμπειρία E σε σχέση με μια κατηγορία εργασιών T και μια μετρική απόδοσης P, αν η απόδοση του σε εργασίες της T, όπως μετριούνται από την P, βελτιώνονται με την εμπειρία E.*

Η μηχανική μάθηση βασίζεται στην *επαγωγή*, μια από τις τρεις βασικές συλλογιστικές που περιγράφονται παρακάτω:

- Η *συνεπαγωγή* (*deduction*) περιλαμβάνει την εξαγωγή ορθών συμπερασμάτων. Τα συμπεράσματα που προκύπτουν είναι ορθά, διότι η διαδικασία εξαγωγής συμπεράσματος κατευθύνεται από το γενικό προς το ειδικό. Για παράδειγμα, αν ισχύει $A \rightarrow B$ και A, τότε εξάγεται το ορθό συμπέρασμα ότι ισχύει το B.
- Η *επαγωγή* (*induction*) περιλαμβάνει την εξαγωγή γενικών συμπερασμάτων από ένα σύνολο παραδειγμάτων. Η διαδικασία εξαγωγής συμπεράσματος στην περίπτωση αυτή κατευθύνεται από το ειδικό προς το γενικό, γεγονός που εισάγει κάποιο βαθμό αβεβαιότητας για το εξαγόμενο συμπέρασμα. Για παράδειγμα, αν ισχύουν τα $A_i \rightarrow B_i$, όπου $i = \{1, 2, \ldots, n\}$, τότε εξάγεται το συμπέρασμα ότι ισχύει $A \rightarrow B$. Τα A και B είναι γενικεύσεις των προτάσεων A_i και B_i αντίστοιχα. Αυτό το

είδος συλλογιστικής χρησιμοποιείται εδώ και πολλούς αιώνες για την παραγωγή επιστημονικής γνώσης.

- Η *απαγωγή* (*abduction*) περιλαμβάνει την εξαγωγή συμπερασμάτων, με βάση την παρατήρηση κάποιου γεγονότος που αποτελεί συμπέρασμα ενός κανόνα. Για παράδειγμα, αν ισχύει $A \rightarrow B$ και B, τότε εξάγεται το συμπέρασμα ότι είναι πιθανό να ισχύει το A. Η συλλογιστική που βασίζεται στην απαγωγή δεν είναι μαθηματικά ακριβής, άλλα ουσιαστικά εκφράζει κάποια πιθανότητα.

2.2.1 Κατηγορίες Μεθόδων Μηχανικής Μάθησης

Υπάρχουν διάφοροι τρόποι με τους οποίους μπορούν να διαχωριστούν οι μέθοδοι της μηχανικής μάθησης. Ο πιο διαδεδομένος διαχωρισμός διακρίνει δύο είδη μάθησης:

- *Μάθηση με επίβλεψη* (*supervised learning*). Η μάθηση με επίβλεψη περιλαμβάνει τη μάθηση από παραδείγματα που ανήκουν σε καθορισμένες τάξεις. Το σύστημα καλείται να ανακαλύψει τις κοινές ιδιότητες των αντικειμένων κάθε τάξης.
- *Μάθηση χωρίς επίβλεψη* (*unsupervised learning*). Στη μάθηση χωρίς επίβλεψη το σύστημα πρέπει να ανακαλύψει τις τάξεις από μόνο του βασιζόμενο στις κοινές ιδιότητες των αντικειμένων.

Ένας άλλος διαχωρισμός των μεθόδων της μηχανικής μάθησης είναι ο εξής:

- *Μάθηση συνάρτησης*. Στόχος είναι η προσέγγιση μιας συνάρτησης μελετώντας ένα σύνολο δεδομένων με ζευγάρια τιμών εισόδου και εξόδου της συνάρτησης αυτής. Αν η συνάρτηση έχει διακριτό και πεπερασμένο σύνολο τιμών, τότε το πρόβλημα ονομάζεται *ταξινόμηση* (*classification*), ενώ αν έχει συνεχές ή μη πεπερασμένο σύνολο τιμών ονομάζεται *παλινδρόμηση* (*regression*).
- *Εξόρυξη κανόνων συσχέτισης* (*association rule mining*). Στόχος είναι η εύρεση συσχετίσεων μεταξύ αντικειμένων σε συναλλακτικές (transactional) βάσεις δεδομένων.
- *Ομαδοποίηση* (*clustering*). Στόχος είναι η ανακάλυψη φυσικών οργανώσεων των δεδομένων σε ομάδες.
- *Ενισχυτική μάθηση* (*reinforcement learning*). Στόχος είναι η εύρεση της βέλτιστης συμπεριφοράς ενός πράκτορα σε κάποιο περιβάλλον, με βάση την ανταμοιβή που παίρνει σε μια τελική κατάσταση, έχοντας ξεκινήσει από μια αρχική κατάσταση και ακολουθώντας μιας σειρά από ενέργειες και ενδιάμεσες καταστάσεις.

Η μάθηση με επίβλεψη ταυτίζεται με την πρώτη κατηγορία προβλημάτων, δηλαδή της ταξινόμησης και της παλινδρόμησης. Το όνομα της υπαινίσσεται την ύπαρξη κάποιου «επιβλέποντος», ο οποίος παρέχει την τιμή εξόδου της συνάρτησης για τα δεδομένα που εξετάζονται. Η μάθηση χωρίς επίβλεψη παραδοσιακά ταυτιζόταν με το πρόβλημα

της ομαδοποίησης. Ο λόγος είναι ότι στην ομαδοποίηση δεν υπάρχει κάποιος «επιβλέπων», καθώς είναι άγνωστος ο αριθμός και η περιγραφή των ομάδων που ενδεχομένως υπάρχουν στα δεδομένα. Η εξόρυξη κανόνων συσχέτισης εμφανίστηκε αρκετά αργότερα από την μηχανική μάθηση και έχει περισσότερες επιρροές από την περιοχή των βάσεων δεδομένων. Ωστόσο, μπορεί να ενταχθεί στη μάθηση χωρίς επίβλεψη, καθώς ούτε σε αυτήν την περίπτωση είναι γνωστές οι συσχετίσεις που ενδεχομένως υπάρχουν στα δεδομένα. Τέλος, η ενισχυτική μάθηση έχει επιρροές και από τις δύο κατηγορίες μάθησης. Η αριθμητική ανταμοιβή για κάθε ενέργεια του πράκτορα δίνεται από κάποιον «επιβλέποντα» (το περιβάλλον). Ωστόσο, η συμπεριφορά του περιβάλλοντος είναι άγνωστη στον πράκτορα, ο οποίος προσπαθεί να την ανακαλύψει μέσω δοκιμής και αποτυχίας, διαδικασία που παραπέμπει στη μάθηση χωρίς επίβλεψη.

Αντίστοιχα με τη διαδικασία ανακάλυψης γνώσης, η εφαρμογή της μηχανικής μάθησης σε ένα σύνολο δεδομένων έχει ως αποτέλεσμα τη δημιουργία προτύπων ή μοντέλων. Το πρότυπο αποτελεί γνώση που σχετίζεται με ένα τοπικό τμήμα των δεδομένων (π.χ. ένας κανόνας συσχέτισης, ένας κανόνας ταξινόμησης ή μια ομάδα), ενώ το μοντέλο αποτελεί γνώση που σχετίζεται με το σύνολο των δεδομένων (π.χ. ένα δένδρο απόφασης, ένα νευρωνικό δίκτυο ή μια γραμμική συνάρτηση).

2.2.2 Σχέση Μηχανικής Μάθησης και Ανακάλυψης Γνώσης

Η μηχανική μάθηση και η ανακάλυψη γνώσης από βάσεις δεδομένων έχουν πολλά κοινά σημεία που κάνουν δυσδιάκριτες τις μεταξύ τους διαφορές. Τυπικά, η μηχανική μάθηση σχετίζεται περισσότερο με το βήμα της εξόρυξης από δεδομένα στη διαδικασία της ανακάλυψης γνώσης και έχει περισσότερες επιρροές από την περιοχή της τεχνητής νοημοσύνης. Η ανακάλυψη γνώσης έχει περισσότερες επιρροές από την περιοχή των βάσεων δεδομένων και είναι περισσότερο προσανατολισμένη στην αποδοτική εκτέλεση των αλγορίθμων, καθώς και στη γενικότερη διαχείριση και ανάλυση δεδομένων που βρίσκονται αποθηκευμένα σε βάσεις δεδομένων.

2.3 Μέθοδοι Εξόρυξης από Δεδομένα

Στην ενότητα αυτή περιγράφονται οι βασικότερες μέθοδοι εξόρυξης από δεδομένα. Συγκεκριμένα, παρουσιάζονται μέθοδοι που αφορούν τη μάθηση συνάρτησης (δηλ. την ταξινόμηση και την παλινδρόμηση), την εξόρυξη κανόνων συσχέτισης και την ομαδοποίηση.

2.3.1 Ταξινόμηση και Παλινδρόμηση

Κατά τη μάθηση συνάρτησης επιδιώκεται ο προσδιορισμός μιας άγνωστης συνάρτησης, που καλείται *συνάρτηση στόχος* (*target function*), από γνωστά ζεύγη τιμών εισόδου και εξόδου της υπό μάθηση συνάρτησης. Ουσιαστικά, η συνάρτηση στόχος χρησιμοποιείται για την πρόβλεψη της τιμής μιας εξαρτημένης μεταβλητής που ονομάζεται *μεταβλητή στόχος* με βάση τις τιμές ενός συνόλου από ανεξάρτητες μεταβλητές που ονομά-

ζονται *χαρακτηριστικά* (*attributes* ή *features*). Το σύνολο των διαφορετικών δυνατών τιμών εισόδου της συνάρτησης στόχου, δηλαδή το πεδίο ορισμού της, ονομάζεται *σύνολο περιπτώσεων* ή *στιγμιότυπων* (*instances*). Ένα σύστημα μάθησης συνάρτησης, κατά την εκπαίδευση του, έχει ως είσοδο ένα υποσύνολο του συνόλου των περιπτώσεων, για το οποίο είναι γνωστή η τιμή της μεταβλητής στόχου. Αυτό το σύνολο ονομάζεται *σύνολο εκπαίδευσης* (*training set*). Ένα στοιχείο του συνόλου εκπαίδευσης ονομάζεται *παράδειγμα* (*example*). Το σύστημα μάθησης για να προσεγγίσει όσο καλύτερα μπορεί τη συνάρτηση στόχο δεδομένου ενός συνόλου εκπαίδευσης, εξετάζει διάφορες εναλλακτικές συναρτήσεις, οι οποίες ονομάζονται *υποθέσεις* (*hypotheses*). Το σύνολο όλων των δυνατών διαφορετικών υποθέσεων που το σύστημα μάθησης ενδέχεται να εξετάσει ονομάζεται *χώρος υποθέσεων* (*hypotheses space*). Παραδείγματα μάθησης συνάρτησης είναι η παλινδρόμηση και η ταξινόμηση.

Η *παλινδρόμηση* (*regression*) περιλαμβάνει τη μάθηση συναρτήσεων, με συνεχές πεδίο αριθμητικών τιμών. Οι κυριότερη μέθοδος παλινδρόμησης είναι η *γραμμική παλινδρόμηση* (*linear regression*), όπου η συνάρτηση μοντελοποιείται με μια απλή γραμμική εξίσωση των παραμέτρων εισόδου.

Η *ταξινόμηση*[5] (*classification*) είναι μέθοδος μάθησης συνάρτησης με κατηγορικό (categorical/nominal – π.χ. ομάδα αίματος) πεδίο τιμών. Είναι δυνατό οι πιθανές κατηγορίες να είναι διατεταγμένες διακριτές τιμές (ordinal – π.χ. υψηλό, μεσαίο, χαμηλό). Στην περίπτωση της ταξινόμησης η μεταβλητή στόχος καλείται *τάξη* ή *κλάση* (*class*). Ένα σύστημα που εκπαιδεύεται με κάποιο σύνολο εκπαίδευσης χρησιμοποιώντας κάποιον αλγόριθμο ταξινόμησης καλείται *ταξινομητής* (*classifier*) και μπορεί να χρησιμοποιηθεί για την ταξινόμηση (πρόβλεψη της τάξης) νέων στιγμιοτύπων. Οι πιο κοινές μέθοδοι ταξινόμησης παρουσιάζονται παρακάτω.

Δένδρα Ταξινόμησης

Τα *δένδρα ταξινόμησης* (*classification trees*) ή *δένδρα απόφασης* (*decision trees*) είναι δενδροειδείς δομές που αναπαριστούν με γραφικό τρόπο τις συσχετίσεις στα δεδομένα εκπαίδευσης. Ένα δένδρο ταξινόμησης μπορεί πολύ εύκολα να αναπαρασταθεί και ως ένα σύνολο κανόνων. Το μεγαλύτερο πλεονέκτημα των δενδροειδών μοντέλων ταξινόμησης είναι η εύκολη ερμηνεία τους.

Ένας γενικός αλγόριθμος δημιουργίας δένδρου ταξινόμησης περιλαμβάνει τα εξής βήματα:

1. Εύρεση της ανεξάρτητης μεταβλητής η οποία αν χρησιμοποιηθεί ως κριτήριο διαχωρισμού των δεδομένων θα οδηγήσει σε όσο το δυνατόν διαφορετικούς μεταξύ τους κόμβους σε σχέση με την εξαρτημένη μεταβλητή.

2. Πραγματοποίηση διαχωρισμού βάσει της ανεξάρτητης μεταβλητής που βρέθηκε.

[5] Η μετάφραση του όρου «classification» συναντάται και ως «κατηγοριοποίηση» στην ελληνική βιβλιογραφία.

3. Επανάληψη της διαδικασίας διαχωρισμού στους κόμβους που προέκυψαν έως ότου να μην υπάρχει δυνατός διαχωρισμός.

Οι πιο γνωστοί αλγόριθμοι κατασκευής δένδρων ταξινόμησης είναι οι εξής:

- Ο αλγόριθμος *ID3*, που είναι γνωστός και ως *αλγόριθμος κατασκευής δένδρων απόφασης με επαγωγή* (*decision tree induction algorithm*). Ο αλγόριθμος αυτός χρησιμοποιεί τα μεγέθη της *εντροπίας* και του *κέρδους πληροφορίας* για την επιλογή του επόμενου καλύτερου διαχωρισμού. Η *εντροπία* (*entropy*) μετρά την ανομοιογένεια του συνόλου των παραδειγμάτων που χρησιμοποιούνται για εκπαίδευση και το *κέρδοςπληροφορίας* (*information gain*) μετρά την αναμενόμενη μείωση της εντροπίας εξαιτίας του διαχωρισμού σε συγκεκριμένη ανεξάρτητη μεταβλητή (Εδάφιο 2.4.2).
- Ο *CART* (*Classification and Regression Trees*), ο οποίος δημιουργεί δυαδικά δένδρα απόφασης και υποστηρίζει την ταξινόμηση αλλά και την παλινδρόμηση.
- Ο *C4.5*, ο οποίος αποτελεί βελτίωση του *ID3*, και ο *C5.0,* που χρησιμοποιεί ειδικές τεχνικές (boosting) για να ξεπεράσει καταστάσεις χαμηλής βελτιστοποίησης, που οφείλονται στην άπληστη συμπεριφορά των αλγορίθμων κατασκευής δένδρων απόφασης.
- Ο *CHAID* (*Chi-Squared Automatic Interaction detection*), ο οποίος χρησιμοποιεί τη δοκιμασία X^2 για την επιλογή του επόμενου καλύτερου διαχωρισμού.

Κανόνες Ταξινόμησης

Μια από τις πιο εκφραστικές και κατανοητές για τον άνθρωπο αναπαραστάσεις για μάθηση υποθέσεων είναι τα σύνολα κανόνων της μορφής «εάν *X* τότε *Y*». Δύο σημαντικές κατηγορίες κανόνων ταξινόμησης είναι οι παρακάτω:

- Οι *προτασιακοί κανόνες* (*propositional rules*). Οι προτασιακοί κανόνες δεν περιέχουν μεταβλητές και έτσι δεν μπορούν να αναπαρασταθούν γενικές σχέσεις μεταξύ των χαρακτηριστικών. Μπορούν να προκύψουν από άλλες μορφές αναπαράστασης (π.χ. δένδρα, γενετικοί αλγόριθμοι) ή από απ' ευθείας μάθηση. Ένας γενικός αλγόριθμος μάθησης τέτοιων κανόνων είναι ο *αλγόριθμος σειριακής κάλυψης* (*serial covering algorithms*), που δημιουργεί ένα σύνολο προτασιακών κανόνων επαυξητικά (incrementally), μαθαίνοντας έναν κανόνα κάθε φορά.
- Οι *κανόνες πρώτης τάξης* (*first order rules*). Οι κανόνες πρώτης τάξης, αντίθετα με τους προτασιακούς, περιέχουν μεταβλητές. Επίσης, έχουν μεγάλη εκφραστική ικανότητα και προκύπτουν από απ' ευθείας μάθηση. Η μάθηση των κανόνων πρώτης τάξης ονομάζεται και *Λογικός Επαγωγικός Προγραμματισμός* (*Inductive Logic Programming - ILP*). Ο πιο γνωστός αλγόριθμος της κατηγορίας είναι ο FOIL του Quinlan [1990].

Ταξινόμηση Βάσει Περιπτώσεων

Στην *ταξινόμηση* (ή *μάθηση*) *βάσει περιπτώσεων* (*instance-based learning*) τα δεδομένα εκπαίδευσης χρησιμοποιούνται αυτούσια. Όταν ένα σύστημα κληθεί να αποφασίσει για μια νέα περίπτωση, εξετάζει τη σχέση της με τις ήδη αποθηκευμένες περιπτώσεις.

Χαρακτηριστικός αλγόριθμος αυτής της κατηγορίας είναι ο αλγόριθμος των *k-κοντινότερων γειτόνων* (*k-nearest neighbors*). Σύμφωνα με τον αλγόριθμο αυτό όταν πρόκειται να ταξινομηθεί κάποιο νέο παράδειγμα υπολογίζονται οι αποστάσεις του παραδείγματος από κάθε αντικείμενο του συνόλου εκπαίδευσης. Στη συνέχεια λαμβάνοντας υπόψη μόνο τα *k* κοντινότερα παραδείγματα, η νέα περίπτωση τοποθετείται στην τάξη που ανήκουν τα περισσότερα από αυτά.

Ταξινόμηση Βάσει της Θεωρίας του Bayes

Η συλλογιστική κατά Bayes αποτελεί τη θεωρητική βάση για αλγορίθμους ταξινόμησης που διαχειρίζονται πιθανότητες, καθώς και ένα πλαίσιο για το συγκριτικό έλεγχο της απόδοσης άλλων αλγορίθμων.

Στην *ταξινόμηση* (ή *μάθηση*) *βάσει της θεωρίας του Bayes* (*Bayesian learning*) κάθε παράδειγμα εκπαίδευσης μπορεί σταδιακά να μειώσει ή να αυξήσει την πιθανότητα να είναι σωστή μια υπόθεση. Αυτό δίνει μεγάλη ευελιξία στους αλγορίθμους, καθώς δεν απορρίπτουν αμέσως μια υπόθεση που φαίνεται να μην είναι σε απόλυτη συμφωνία με τα παραδείγματα εκπαίδευσης. Επιπλέον, προϋπάρχουσα γνώση μπορεί να συνδυαστεί με τα δεδομένα εκπαίδευσης με τη μορφή αρχικών τιμών πιθανότητας για τις υπό εξέταση υποθέσεις. Μια πρακτική δυσκολία στην εφαρμογή της μάθησης με βάση τη θεωρία του Bayes είναι η απαίτηση για τη γνώση πολλών πιθανοτήτων.

Η παραπάνω δυσκολία έχει δώσει μεγάλη πρακτική αξία σε μια απλουστευμένη μορφή της ταξινόμησης βάσει της θεωρίας του Bayes, τον *αφελή ταξινομητή Bayes* (*naïve Bayes classifier*). Ο ταξινομητής αυτός βασίζεται στην παραδοχή ότι οι τιμές όλων των χαρακτηριστικών, δεδομένης μιας τάξης, είναι ανεξάρτητες μεταξύ τους. Το όφελος που προκύπτει από αυτήν την απλούστευση είναι η απαίτηση υπολογισμού λιγότερων πιθανοτήτων με συνέπεια τη μείωση της υπολογιστική πολυπλοκότητας του αλγορίθμου.

Μηχανές Διανυσμάτων Υποστήριξης

Οι *μηχανές διανυσμάτων υποστήριξης* ή *μηχανές εδραίων διανυσμάτων* (*Support Vector Machines – SVMs*) επινοήθηκαν από το Vladimir Vapnik και υλοποιήθηκαν από τους Cortes & Vapnik [1995]. Ο συγκεκριμένος αλγόριθμος ταξινόμησης στοχεύει στην εύρεση ενός υπερεπιπέδου στο χώρο των χαρακτηριστικών το οποίο διαχωρίζει τα θετικά από τα αρνητικά παραδείγματα. Το επίπεδο αυτό επιλέγεται με τέτοιο τρόπο, ώστε να απέχει όσο το δυνατόν περισσότερο από τα κοντινότερα θετικά και αρνητικά παραδείγματα (maximum margin hyperplane). Τα παραδείγματα με τη μικρότερη απόσταση από το υπερεπίπεδο αυτό ονομάζονται διανύσματα υποστήριξης (support vectors).

Η μέθοδος που περιγράφηκε μπορεί να εφαρμοστεί και σε περιπτώσεις που τα θετικά και τα αρνητικά παραδείγματα δεν είναι γραμμικώς διαχωρίσιμα. Σε αυτές τις περιπτώσεις χρησιμοποιούνται οι λεγόμενες συναρτήσεις πυρήνα (kernel functions) για τον ορισμό του υπερεπιπέδου σε ένα νέο μετασχηματισμένο χώρο χαρακτηριστικών.

Νευρωνικά Δίκτυα

Ένα *νευρωνικό δίκτυο* (*neural network*) είναι ένα δίκτυο μονάδων όπου οι μονάδες εισόδου (input units) αντιπροσωπεύουν τα χαρακτηριστικά και οι μονάδες εξόδου (output units) τις τάξεις. Τα βάρη στις συνδέσεις (edges) μεταξύ των εισόδων και των εξόδων αναπαριστούν τις συσχετίσεις μεταξύ τους. Για την ταξινόμηση ενός παραδείγματος, τα βάρη των χαρακτηριστικών εισάγονται στις εισόδους του δικτύου. Η ενεργοποίηση αυτών των μονάδων προωθείται μέσω του δικτύου και η τελική τιμή της μονάδας εξόδου καθορίζει την απόφαση.

Η απλούστερη μορφή νευρωνικού δικτύου είναι το perceptron [Rosenblatt, 1958] που αποτελεί έναν γραμμικό ταξινομητή. Ένα μη-γραμμικό νευρωνικό δίκτυο μπορεί να κατασκευαστεί έχοντας επιπλέον «κρυφά» επίπεδα μονάδων.

Μία τυπική μέθοδος για την εκπαίδευση νευρωνικών δικτύων είναι η *ανάστροφη μετάδοση σφάλματος* (*backpropagation*) όπου σε περίπτωση σφάλματος εξόδου το σφάλμα επιμερίζεται στις κρυφές συνδέσεις.

Ομάδες Ταξινομητών

Στις *ομάδες ταξινομητών* (*classifier committees* ή *ensemble classifiers* ή *multiple classifier systems*) οι αποφάσεις *k* διαφορετικών ταξινομητών συνδυάζονται με στόχο την αύξησης της απόδοσης. Δύο σημαντικοί παράγοντες για την ανάπτυξη μίας ομάδας ταξινομητών είναι α) η επιλογή των *k* ταξινομητών που θα αποτελέσουν τα μέλη της ομάδας και β) ο τρόπος συνδυασμού των αποφάσεων τους.

Για το πρώτο θέμα, είναι γνωστό ότι για να επιτευχθεί καλή απόδοση σε μία ομάδα ταξινομητών τα μέλη της ομάδας πρέπει να έχουν όσο το δυνατόν μεγαλύτερη ανεξαρτησία μεταξύ τους (diversity ή independency) [Tumer and Ghosh, 1996].

Για το δεύτερο θέμα, έχουν προταθεί διάφορες προσεγγίσεις:

- Μία απλή προσέγγιση είναι η μέθοδος πλειοψηφίας (majority voting), όπου μία τάξη αποτελεί πρόβλεψη μόνο αν την έχουν προβλέψει («ψηφίσει») τουλάχιστον (k + 1)/2 ταξινομητές. Εναλλακτικά, στην ψήφο κάθε ταξινομητή μπορεί να προσμετράται και η εμπιστοσύνη που αυτός έχει για τη συγκεκριμένη τάξη (μέθοδος σταθμισμένης πλειοψηφίας – weighted majority voting).

- Στη μέθοδο στοιβαγμένης γενίκευσης (stacked generalization) ή στοιβάγματος (stacking) υπάρχει ένας μέτα-ταξινομητής ο οποίος «μαθαίνει» να συνδυάζει τις αποφάσεις των μελών της ομάδας, καθώς εκπαιδεύεται με τα δεδομένα εξόδου των μελών της ομάδας [Wolpert, 1992].

2.3.2 Εξόρυξη Κανόνων Συσχέτισης

Η *εξόρυξη κανόνων συσχέτισης* (*association rule mining*) περιλαμβάνει την αναζήτηση κοινών εμφανίσεων αντικειμένων σε μια βάση δεδομένων συναλλαγών. Η κοινή παρουσία αντικειμένων σε ένα πλήθος συναλλαγών είναι ένδειξη κάποιας πιθανής σχέσης μεταξύ αυτών των αντικειμένων.

Κλασικό παράδειγμα εφαρμογής των κανόνων συσχέτισης αποτελεί η *ανάλυση καλαθιού αγορών* (*market basket analysis*), δηλαδή η ανάλυση των προϊόντων που αγοράζονται από πελάτες σε καταστήματα. Παράδειγμα ενός τέτοιου κανόνα είναι το εξής: «Αν ο πελάτης αγοράσει γάλα, τότε θα αγοράσει και ψωμί». Οι κανόνες συσχέτισης βρίσκουν εφαρμογή και σε βιολογικά δεδομένα. Παράδειγμα κανόνα συσχέτισης στο συγκεκριμένο πεδίο εφαρμογής είναι το εξής: «Αν το γονίδιο *A* εκφράζεται, τότε και το γονίδιο *B* εκφράζεται». Στην παράγραφο που ακολουθεί ορίζονται οι όροι που χρησιμοποιούνται στην περιοχή της εξόρυξης κανόνων συσχέτισης.

Έστω ότι $I = \{i_1, i_2, \ldots, i_N\}$ είναι ένα πεπερασμένο σύνολο δυαδικών χαρακτηριστικών, τα οποία καλούνται *αντικείμενα* (*items*) και D είναι ένα πεπερασμένο πολυσύνολο συναλλαγών, το οποίο καλείται *σύνολο δεδομένων* (*dataset*). Κάθε *συναλλαγή* $T \in D$ είναι ένα *σύνολο αντικειμένων* (*itemset*) του I ($T \subseteq I$). Ο αριθμός k των αντικειμένων που περιλαμβάνει ένα σύνολο αντικειμένων ονομάζεται *μήκος* ή *μέγεθος* και το σύνολο αντικειμένων προσδιορίζεται ως *k-σύνολο αντικειμένων*. Μια συναλλαγή $T \in D$ λέγεται ότι περιέχει ένα σύνολο αντικειμένων $X \subseteq I$, αν $X \subseteq T$. Η *υποστήριξη* (*support*) ενός συνόλου αντικειμένων X ορίζεται ως το κλάσμα του αριθμού των συναλλαγών του D που περιέχουν το X προς το συνολικό αριθμό των συναλλαγών του D:

$$supp_D(X) = \frac{\left|\{T \in D \mid T \supseteq X\}\right|}{|D|} \qquad 2.1$$

Δεδομένου ενός κατωφλίου ελάχιστης υποστήριξης $\sigma \in (0,1]$, ένα σύνολο αντικειμένων X καλείται *σ-συχνό* (*σ-frequent*) ή απλά *συχνό* (*frequent*) στο D, αν $supp_D(X) \geq \sigma$.

Κανόνας συσχέτισης είναι ένα πρότυπο της μορφής $X \Rightarrow Y$, όπου $X \subset I$, $Y \subset I$, και $X \cap Y = \emptyset$. Το X ονομάζεται *προϋπόθεση* (*antecedent*) ή *αριστερό μέρος* και το Y *συμπέρασμα* (*consequent*) ή *δεξιό μέρος* του κανόνα.

Η απλή αναφορά ενός κανόνα συσχέτισης δεν έχει μεγάλη αξία αν δε συνοδεύεται από κάποια ποσοτικά μεγέθη που αξιολογούν τον κανόνα και του προσδίδουν βαρύτητα. Τα δύο δημοφιλέστερα μέτρα αξιολόγησης των κανόνων συσχέτισης είναι η *υποστήριξη* (*support*) και η *εμπιστοσύνη* (*confidence*). Η υποστήριξη του κανόνα $X \Rightarrow Y$ ισούται με την υποστήριξη του συνόλου αντικειμένων $X \cup Y$ και υπολογίζεται από την Εξίσωση 2.1. Η εμπιστοσύνη του κανόνα $X \Rightarrow Y$ ορίζεται ως το κλάσμα του αριθμού των συναλλαγών του D που περιέχουν το $X \cup Y$ προς τον αριθμό των συναλλαγών του D που περιέχουν το X και άρα ισούται με την υποστήριξη του $X \cup Y$ στο D προς την υποστήριξη του X στο D:

$$conf_D(X \Rightarrow Y) = \frac{supp_D(X \cup Y)}{supp_D(X)} \qquad 2.2$$

Το πρόβλημα της εξόρυξης κανόνων συσχέτισης μπορεί να διαιρεθεί στα εξής δύο υποπροβλήματα:

- Εύρεση όλων των συνόλων αντικειμένων που έχουν τιμή υποστήριξης μεγαλύτερη ή ίση με το κατώφλι ελάχιστης υποστήριξης που έχει οριστεί (συχνά σύνολα αντικειμένων).
- Εύρεση όλων των κανόνων που προκύπτουν από τα συχνά σύνολα αντικειμένων, οι οποίοι ικανοποιούν το κατώφλι ελάχιστης εμπιστοσύνης που έχει οριστεί.

Από τα δύο παραπάνω υποπροβλήματα, το πρώτο παρουσιάζει μεγάλη υπολογιστική πολυπλοκότητα και έχει συγκεντρώσει έντονο ερευνητικό ενδιαφέρον.

Αλγόριθμος Apriori

Οι περισσότεροι αλγόριθμοι για την ανακάλυψη των κανόνων συσχέτισης βασίζονται στην ιδιότητα της *κλειστότητας προς τα κάτω* (*downward closure*), σύμφωνα με την οποία αν ένα σύνολο αντικειμένων *S* είναι συχνό, τότε όλα τα μη-κενά υποσύνολα του *S* είναι επίσης συχνά. Αντιπροσωπευτικός αλγόριθμος της κατηγορίας είναι ο αλγόριθμος A-priori [Agrawal & Srikant, 1994], ο οποίος δημιουργεί υποψήφια συχνά σύνολα αντικειμένων για τα οποία στη συνέχεια ελέγχεται με περάσματα στη βάση δεδομένων η ικανοποίηση του περιορισμού της υποστήριξης. Ο A-priori λειτουργεί σταδιακά. Στο πρώτο του πέρασμα βρίσκει τα συχνά αντικείμενα, στο δεύτερο τα συχνά ζεύγη, στο τρίτο τις συχνές τριάδες κ.ο.κ. Η διαδικασία τερματίζεται σε συγκεκριμένο αριθμό περασμάτων ή όταν πάψουν να υπάρχουν υποψήφια συχνά σύνολα αντικειμένων.

Αλγόριθμος FP-Growth

Έχουν προταθεί και κάποιοι αλγόριθμοι, που δε βασίζονται στη δημιουργία υποψήφιων συχνών συνόλων αντικειμένων, με κυριότερο αντιπρόσωπο τον FP-Growth [Han *et al.*, 2000]. Ο συγκεκριμένος αλγόριθμος χρησιμοποιεί μια δενδρική δομή, το FP-Tree, στην οποία καταχωρούνται όλα τα δεδομένα. Η συγκεκριμένη δομή μπορεί να συμπιέσει έως 200 φορές τα δεδομένα και αποθηκεύεται στη μνήμη του υπολογιστή, οδηγώντας σε αποδοτικότερη εξόρυξη των συχνών συνόλων αντικειμένων.

Ο FP-growth ακολουθεί μια «*διαίρει και βασίλευε*» προσέγγιση με σκοπό την ανάλυση σε απλούστερα μέρη τόσο των διαδικασιών εξόρυξης όσο και της βάσης δεδομένων. Η δημιουργία και ο έλεγχος των υποψήφιων συχνών προτύπων (συνόλων αντικειμένων) αντικαθίσταται από μια διαδικασία σταδιακής αύξησης των συχνων προτύπων (εξού και το όνομα του αλγορίθμου: frequent pattern growth).

Αναδυόμενα Πρότυπα

Τα *αναδυόμενα πρότυπα* (*emerging patterns*) [Dong & Li, 1999] είναι σύνολα αντικειμένων των οποίων η υποστήριξη διαφέρει σημαντικά μεταξύ δύο συνόλων δεδομένων.

Δοθέντων δύο συνόλων δεδομένων D_1 και D_2, ο *ρυθμός ανάπτυξης* (*growth rate*) ενός συνόλου αντικειμένων X από το D_1 στο D_2 ορίζεται ως εξής (οι δείκτες 1 και 2 παρακάτω χρησιμοποιούνται αντί των D_1 και D_2):

$$gr_{1\to 2}(X)=\begin{cases}0, & \text{αν } supp_1(X)=0 \text{ και } supp_2(X)=0\\ \infty, & \text{αν } supp_1(X)=0 \text{ και } supp_2(X)>0\\ \dfrac{supp_2(X)}{supp_1(X)}, & \text{αλλιώς}\end{cases} \qquad 2.3$$

Δεδομένου ενός κατωφλίου ελάχιστου ρυθμού ανάπτυξης $\rho > 1$, ένα σύνολο αντικειμένων X καλείται *ρ-αναδυόμενο πρότυπο* (*ρ-emerging pattern*), ή απλά *αναδυόμενο πρότυπο* (*emerging pattern*), από το D_1 στο D_2, αν $gr_{1\to 2}(X)\geq\rho$. Το D_1 καλείται *σύνολο δεδομένων-υπόβαθρο* (*background dataset*), ενώ το D_2 καλείται *σύνολο δεδομένων-στόχος* (*target dataset*).

Η *ισχύς* (*strength*) ενός αναδυόμενου προτύπου X από το D_1 στο D_2 ορίζεται ως εξής:

$$strength_{1\to 2}(X)=\begin{cases}supp_2(X), & \text{αν } gr_{1\to 2}(X)=\infty\\ supp_2(X)\dfrac{gr_{1\to 2}(X)}{gr_{1\to 2}(X)+1}, & \text{αλλιώς}\end{cases} \qquad 2.4$$

Ταξινόμηση με Αναδυόμενα Πρότυπα

Τα αναδυόμενα πρότυπα μπορούν να χρησιμοποιηθούν για ταξινόμηση. Δεδομένων δύο συνόλων, του D_+ που περιέχει θετικά και του D_- που περιέχει αρνητικά παραδείγματα[6], μπορούν να εξαχθούν δύο σύνολα αναδυόμενων προτύπων E_+ και E_-, αντίστοιχα. Για την εξόρυξη των προτύπων του συνόλου E_+ χρησιμοποιείται το σύνολο παραδειγμάτων D_- ως σύνολο δεδομένων-υπόβαθρο και το D_+ ως σύνολο δεδομένων-στόχος. Αντίστοιχα, για την εξόρυξη των προτύπων του συνόλου E_- χρησιμοποιείται το σύνολο παραδειγμάτων D_+ ως σύνολο δεδομένων-υπόβαθρο και το D_- ως σύνολο δεδομένων-στόχος.

Όταν ένα νέο στιγμιότυπο (συναλλαγή) T καλείται να ταξινομηθεί, τότε υπολογίζονται οι τιμές των εξισώσεων 2.5 και 2.6. Η τιμή της Εξίσωσης 2.5 μετρά το βαθμό στον οποίο το παράδειγμα T ανήκει στη θετική τάξη, ενώ η τιμή της Εξίσωσης 2.6 μετρά το βαθμό στον οποίο το παράδειγμα T ανήκει στην αρνητική τάξη. Το παράδειγμα τελικά ταξινομείται στην τάξη για την οποία έχει υπολογιστεί η μεγαλύτερη τιμή.

[6] Τα παραδείγματα σε αυτήν την περίπτωση έχουν τη μορφή συναλλαγής.

$$score(T,+) = \sum_{e \subseteq T, e \in E_+} strength_{-\to+}(e) \qquad 2.5$$

$$score(T,-) = \sum_{e \subseteq T, e \in E_-} strength_{+\to-}(e) \qquad 2.6$$

2.3.3 Ομαδοποίηση

Οι *ομάδες*[7] (*clusters*) είναι πρότυπα που προκύπτουν από την *ομαδοποίηση* (*clustering*) των παραδειγμάτων, έτσι ώστε τα μέλη των ομάδων να έχουν κοινά χαρακτηριστικά. Ένας από τους ορισμούς που έχουν προταθεί για τις ομάδες σύμφωνα με τη Dunham [2002] είναι ο εξής:

> *Ομάδα είναι ένα σύνολο σημείων για τα οποία ισχύει ότι η απόσταση ανάμεσα σε σημεία που ανήκουν στην ομάδα είναι μικρότερη από την απόσταση μεταξύ ενός σημείου της ομάδας και οποιουδήποτε σημείου εκτός αυτής.*

Οι ομάδες που προκύπτουν συνήθως περιγράφονται από κάποιας μορφής μαθηματική σχέση, η οποία απεικονίζει τα παραδείγματα στις ομάδες. Υπάρχουν αρκετές κατηγορίες αλγορίθμων ομαδοποίησης, οι βασικότερες από τις οποίες περιγράφονται παρακάτω [Βαζιριγιάννης & Χαλκίδη, 2003]:

- *Βασισμένοι σε διαχωρισμούς* (*partitional*), που προσπαθούν να βρουν τον καλύτερο διαχωρισμό ενός συνόλου δεδομένων σε ένα συγκεκριμένο αριθμό ομάδων. Ένας πολύ γνωστός αλγόριθμος της κατηγορίας αυτής είναι ο αλγόριθμος των *Κ-μέσων* (*K-means*), ο οποίος αρχικά επιλέγει *Κ* τυχαία σημεία ως κέντρα των *Κ* ομάδων και στη συνέχεια αναθέτει κάθε σημείο στην ομάδα της οποίας το κέντρο βρίσκεται πιο κοντά. Έπειτα, υπολογίζονται τα μέσα διανύσματα των σημείων κάθε ομάδας και ορίζονται ως νέα κέντρα των ομάδων. Η διαδικασία επαναλαμβάνεται για συγκεκριμένο αριθμό βημάτων.
- *Ιεραρχικοί* (*hierarchical*), που προσπαθούν με ιεραρχικό τρόπο να ανακαλύψουν τη δομή των ομάδων. Το αποτέλεσμα αυτών των αλγορίθμων είναι μια ιεραρχία από διαφορετικές ομαδοποιήσεις των δεδομένων, στο ένα άκρο της οποίας βρίσκεται μία μόνο ομάδα με όλα τα δεδομένα και στο άλλο τόσες ομάδες όσα είναι τα δεδομένα. Με βάση την κατεύθυνση ανάπτυξης της ιεραρχίας που ακολουθούν οι ιεραρχικοί αλγόριθμοι ομαδοποίησης διακρίνονται σε:
    - *Συγχωνευτικούς* (*agglomerative*)
    - *Διαιρετικούς* (*divisive*)

[7] Η μετάφραση του όρου «cluster» συναντάται και ως «συστάδα» στην ελληνική βιβλιογραφία. Αντίστοιχα, η μετάφραση του όρου «clustering» συναντάται και ως «συσταδοποίηση».

- *Ασαφείς* (*fuzzy*), οι οποίοι ομαδοποιούν τα δεδομένα με χρήση τεχνικών ασαφούς λογικής. Οι αλγόριθμοι της κατηγορίας αυτής θεωρούν ότι ένα αντικείμενο μπορεί να ανήκει σε περισσότερες από μία ομάδες. Ο σημαντικότερος ασαφής αλγόριθμος ομαδοποίησης είναι ο *Fuzzy C-Means* [Bezdek *et al*., 1984].
- *Βασισμένοι στην πυκνότητα* (*density-based*), οι οποίοι οργανώνουν τα γειτονικά αντικείμενα ενός συνόλου δεδομένων σε ομάδες βασισμένοι σε κριτήρια πυκνότητας. Αντιπροσωπευτικός αλγόριθμος της κατηγορίας αυτής είναι ο DBSCAN [Ester *et al.*, 1996].
- *Βασισμένοι σε πλέγμα* (*grid-based*). Αυτοί οι αλγόριθμοι χρησιμοποιούνται κυρίως για την ανάλυση χωρικών δεδομένων. Το βασικό χαρακτηριστικό τους είναι ο διαχωρισμός του χώρου σε ένα πεπερασμένο αριθμό κελιών πριν από την εφαρμογή των βασικών διαδικασιών.
- *Βασισμένοι σε δίκτυα Kohonen* (*Kohonen net*), οι οποίοι βασίζονται στα νευρωνικά δίκτυα. Το δίκτυο Kohonen έχει κόμβους εισόδου και εξόδου. Ένας κόμβος εισόδου αντιστοιχεί με ένα χαρακτηριστικό των παραδειγμάτων και συνδέεται με κάθε κόμβο εξόδου. Κάθε σύνδεση σχετίζεται με ένα βάρος. Ο αλγόριθμος αλλάζει κατάλληλα τα βάρη και οι κόμβοι εξόδου τείνουν να σχηματίζουν ομάδες.
- *Υποχώρων* (*subspace*), οι οποίοι προσπαθούν να βρουν τα υποσύνολα του αρχικού χώρου που βελτιστοποιούν τα αποτελέσματα ομαδοποίησης. Ως αποτέλεσμα αυτών των αλγορίθμων μπορούν να προκύψουν ομάδες παραδειγμάτων, για συγκεκριμένους υποχώρους των χαρακτηριστικών.

2.4 Μέθοδοι Προεπεξεργασίας και Μεταεπεξεργασίας

Στις παραγράφους που ακολουθούν παρουσιάζονται κάποιες βασικές μέθοδοι προεπεξεργασίας δεδομένων, οι οποίες αφορούν τη μείωση διαστάσεων και τη διακριτοποίηση και θώς και μέθοδοι μεταεπεξεργασίας που αφορούν την αξιολόγηση μοντέλων και προτύπων.

2.4.1 Μείωση Διαστάσεων

Ο μεγάλος αριθμός διαστάσεων (χαρακτηριστικών) μπορεί να προκαλέσει προβλήματα αποδοτικότητας και αποτελεσματικότητας στις μεθόδους ταξινόμησης ή παλινδρόμησης. Πολλές μέθοδοι δεν κλιμακώνονται καλά καθώς αυξάνεται ο αριθμός των χαρακτηριστικών. Επίσης, η παρουσία ακατάλληλων χαρακτηριστικών επηρεάζει αρνητικά την αποτελεσματικότητα των μεθόδων. Το μέγεθος του προβλήματος που μπορεί να δημιουργήσει το πλήθος των διαστάσεων, περιγράφεται χαρακτηριστικά από την έκφραση «κατάρα των πολλών διαστάσεων» (“curse of dimensionality”). Για το λόγο αυτό πριν από την εφαρμογή κάποιου αλγορίθμου ταξινόμησης ή παλινδρόμησης συνήθως προηγείται ένα στάδιο μείωσης διαστάσεων (dimensionality reduction).

Για τη μείωση των διαστάσεων έχουν προταθεί αρκετές μέθοδοι προερχόμενες κυρίως από τις περιοχές της θεωρίας της πληροφορίας και της γραμμικής άλγεβρας. Οι μέθοδοι αυτές μπορούν να διακριθούν σε δύο κατηγορίες, την *επιλογή χαρακτηριστικών* (*feature selection*) και την *εξαγωγή χαρακτηριστικών* (*feature extraction*).

Επιλογή Χαρακτηριστικών

Με τις μεθόδους επιλογής χαρακτηριστικών επιλέγεται ένα υποσύνολο των αρχικών χαρακτηριστικών. Οι περισσότερες μέθοδοι πραγματοποιούν αναζήτηση στο χώρο των χαρακτηριστικών με σκοπό την εύρεση του υποσυνόλου που παρέχει αποτελεσματικότερα μοντέλα. Η αναζήτηση μπορεί να γίνει με δύο τρόπους:

- Η *επιλογή προς τα εμπρός* (*forward selection*) ξεκινά με ένα κενό σύνολο επιλεγμένων χαρακτηριστικών και σε κάθε βήμα προσθέτει ένα χαρακτηριστικό στο σύνολο.
- Η *απαλοιφή προς τα πίσω* (*backward elimination*) ξεκινά με το σύνολο όλων των χαρακτηριστικών και σε κάθε βήμα αφαιρεί ένα χαρακτηριστικό από το σύνολο.

Η αναζήτηση σταματά όταν δεν υπάρχουν χαρακτηριστικά που να ικανοποιούν τις προϋποθέσεις για επιλογή ή απαλοιφή, αντίστοιχα[8].

Εξαγωγή Χαρακτηριστικών

Με τις μεθόδους εξαγωγής χαρακτηριστικών δημιουργείται ένα νέο σύνολο μικρότερου μεγέθους από το αρχικό σύνολο των χαρακτηριστικών. Τα χαρακτηριστικά που περιλαμβάνονται στο νέο σύνολο είναι συνδυασμοί των αρχικών χαρακτηριστικών. Μια από τις πιο γνωστές και δημοφιλείς μεθόδους εξαγωγής χαρακτηριστικών είναι η *Ανάλυση Κυρίων Συνιστωσών* (*Principal Components Analysis - PCA*).

2.4.2 Διακριτοποίηση

Η διακριτοποίηση είναι η διαδικασία κατά την οποία ένα συνεχές διάστημα τιμών μετατρέπεται σε έναν αριθμό διακριτών τιμών. Παρακάτω παρουσιάζονται μερικές από τις δημοφιλέστερες μεθόδους διακριτοποίησης.

Διαχωρισμός Ίσων Διαστημάτων

Ένας προφανής τρόπος διακριτοποίησης είναι να διαιρεθεί το διάστημα τιμών ενός αριθμητικού χαρακτηριστικού σε έναν προκαθορισμένο αριθμό ίσων διαστημάτων και να ανατεθεί μια βαθμωτή (nominal) τιμή σε καθένα από αυτά τα διαστήματα. Η μέθοδος αυτή ονομάζεται *διαχωρισμός ίσων διαστημάτων* (*equal interval binning*) και έχει δυο βασικά μειονεκτήματα:

[8] Η προϋπόθεση επιλογής ή απαλοιφής ενός χαρακτηριστικού συνήθως εξαρτάται από τη βελτίωση που προσφέρει το συγκεκριμένο χαρακτηριστικό στην αποτελεσματικότητα του μοντέλου. Συνήθως χρησιμοποιείται ένα κατώφλι (threshold) βελτίωσης, το οποίο έχει ως στόχο τη διατήρηση ισορροπίας μεταξύ αποτελεσματικότητας και πολυπλοκότητας.

- Τα διαστήματα ενδέχεται να έχουν μεγάλες διαφορές σε σχέση τον αριθμό των παραδειγμάτων που περιέχουν.
- Το εύρος των διαστημάτων που επιλέγεται μπορεί να είναι είτε αρκετά μεγάλο είτε αρκετά μικρό, ώστε να επηρεάζει αρνητικά την ακρίβεια του μοντέλου.

Διαχωρισμός Ίσης Συχνότητας

Μια άλλη μέθοδος που χρησιμοποιείται ευρέως και ξεπερνά τα μειονεκτήματα της προηγούμενης είναι η μέθοδος *διαχωρισμού ίσης συχνότητας* (*equal frequency binning*). Σε αυτήν την περίπτωση τα διαστήματα ορίζονται έτσι ώστε όλα να έχουν τον ίδιο αριθμό παραδειγμάτων, χωρίς να λαμβάνεται υπόψη το εύρος τους. Η συγκεκριμένη μέθοδος ονομαζεται και *ισοστάθμιση ιστογράμματος* (*histogram equalization*).

Διαχωρισμός Βάσει της Εντοπίας

Οι προηγούμενες δύο μέθοδοι διακριτοποίησης δε λαμβάνουν υπόψη την τάξη των παραδειγμάτων (χωρίς επίβλεψη). Ωστόσο, η μέθοδος διαχωρισμού βάσει της εντροπίας που παρουσιάζεται παρακάτω λαμβάνει υπόψη την τάξη των παραδειγμάτων (με επίβλεψη).

Έστω ένα σύνολο παραδειγμάτων εκπαίδευσης *S*. Τα παραδείγματα του συνόλου *S* μπορούν να ανήκουν σε μια από τις *c* διαφορετικές τάξεις, οι οποίες συμβολίζονται με τους αριθμούς 1 έως *c*. Δεδομένων των παραπάνω η *εντροπία* (*entropy*) *E* του συνόλου *S* ορίζεται από την Εξίσωση 2.7.

$$E(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \qquad 2.7$$

Το *c* είναι ο αριθμός των διαφορετικών τάξεων και το p_i είναι ο λόγος των παραδειγμάτων του *S* που ανήκουν στην τάξη *i*. Εξ' ορισμού ισχύει ότι αν $p_i = 0$, τότε $p_i \log_2(p_i) = 0$.

Η εντροπία του *S* μειώνεται όταν αυξάνεται η ομοιογένεια του *S* ως προς την τάξη στην οποία ανήκουν τα παραδείγματα που περιέχει. Για παράδειγμα, η εντροπία είναι 0 αν όλα τα μέλη του *S* ανήκουν στην ίδια τάξη. Αντίθετα, η εντροπία αυξάνεται όταν μειώνεται η ομοιογένεια των παραδειγμάτων που περιέχονται στο *S*.

Έστω ένα διατεταγμένο σύνολο από *N* σημεία διαχωρισμου $T=\{t_1, \ldots, t_N\}$ για τις τιμές ενός χαρακτηριστικού *A*, τα οποία χωρίζουν το σύνολο των παραδειγμάτων σε *N*+1 υποσύνολα (S_1, …, S_{N+1}). Δεδομένων των παραπάνω, το *κέρδος πληροφορίας* (*information gain*) *G* ορίζεται από την Εξίσωση 2.8.

$$G(S;A,T) = E(S) - \sum_{i=1}^{N+1} \frac{|S_i|}{|S|} E(S_i) \qquad 2.8$$

Το κέρδος πληροφορίας μετρά τη μείωση της εντροπίας που προκαλείται αν διαχωριστούν τα παραδείγματα του συνόλου *S* με βάση τα σημεία διαχωρισμού *T* του χαρακτηριστικού *A*. Απαραίτητη προϋπόθεση για έναν σωστό διαχωρισμό είναι η ελαχιστο-

ποίηση της ανομοιογένειας μεταξύ των παραδειγμάτων που περιέχονται σε κάθε υποσύνολο. Αυτό σημαίνει ότι ανάμεσα στους υποψήφιους διαχωρισμούς, πρέπει να επιλεγεί εκείνος που ελαχιστοποιεί την εντροπία, δηλαδή εκείνος που μεγιστοποιεί το κέρδος πληροφορίας.

2.4.3 Αξιολόγηση Ταξινομητών

Η αξιολόγηση των ταξινομητών απαιτεί της χρήση μέτρων και διαδικασιών αξιολόγησης. Σε αυτό το εδάφιο παρουσιάζονται κάποια δημοφιλή μέτρα της αποτελεσματικότητας των ταξινομητών, καθώς και οι σημαντικότερες διαδικασίες αξιολόγησης.

Μέτρα Αποτελεσματικότητας Ταξινομητών δύο Τάξεων

Τα αποτελέσματα της πρόβλεψης ενός συνόλου στιγμιότυπων από έναν ταξινομητή δύο τάξεων (θετική και αρνητική) μπορούν να χωριστούν σε τέσσερεις κατηγορίες, όπως δείχνει ο Πίνακας 2.1.

Πίνακας 2.1: Πίνακας σύγχυσης (confusion matrix).

		ΠΡΟΒΛΕΨΗ ΤΑΞΗΣ	
		Θετική	**Αρνητική**
ΠΡΑΓΜΑΤΙΚΗ ΤΑΞΗ	**Θετική**	Αληθή Θετικά (ΑΘ)	Ψευδή Αρνητικά (ΨΑ)
	Αρνητική	Ψευδή Θετικά (ΨΘ)	Αληθή Αρνητικά (ΑΑ)

Σύμφωνα με τις τέσσερεις κατηγορίες προβλέψεων που παρουσιάζει ο Πίνακας 2.1 μπορούν να οριστούν τα μέτρα της *ευαισθησίας* (*sensitivity*) και της *ειδικότητας* (*specificity*), που παρουσιάζονται στη 2.9 και στη 2.10, αντίστοιχα. Η ευαισθησία ονομάζεται και *ρυθμός αληθών θετικών* (*true positive rate*), ενώ η ειδικότητα ονομάζεται και *ρυθμός αληθών αρνητικών* (*false positive rate*), καθώς περιγράφουν την ικανότητα του ταξινομητή να ταξινομεί σωστά τα θετικά και τα αρνητικά στιγμιότυπα, αντίστοιχα[9].

$$Ευαισθησία = \frac{AΘ}{AΘ + ΨA} \qquad 2.9$$

$$Ιδιαιτερότητα = \frac{AA}{AA + ΨΘ} \qquad 2.10$$

Η συνολική αποτελεσματικότητα ενός ταξινομητή συνήθως μετράται από την *ακρίβεια* (*accuracy*) και την *προσαρμοσμένη ακρίβεια* (*adjusted accuracy*), που παρουσιάζονται στη 2.11 και στη 2.12, αντίστοιχα. Η προσαρμοσμένη ακρίβεια είναι ένα μέτρο που προτιμάται σε περιπτώσεις δεδομένων με άνιση κατανομή τάξεων.

[9] Η ευαισθησία και η ειδικότητα είναι ισοδύναμες με την *ανάκληση* (*recall*) ως προς τη θετική τάξη και την ανάκληση ως προς τη αρνητική τάξη, αντίστοιχα. Η ανάκληση είναι ιδιαίτερα δημοφιλής στην περιοχή της *ανάκτησης πληροφορίας* (*information retrieval*).

$$Ακρίβεια = \frac{ΑΘ + ΑΑ}{ΑΘ + ΑΑ + ΨΘ + ΨΑ} \quad 2.11$$

$$Προσαρμοσμένη\ Ακρίβεια = \frac{Ευαισθησία +\ Ειδικότητα}{2} \quad 2.12$$

Ένα άλλο μέτρο αξιολόγησης ταξινομητών είναι η *καμπύλη Λειτουργικού Χαρακτηριστικού Δέκτη* ή *καμπύλη ROC* (*Receiver Operating Characteristic*), η οποία αποτελεί τη γραφική παράσταση της ευαισθησίας (ρυθμού αληθών θετικών), ως προς την ποσότητα *1-ειδικότητα* (ρυθμό ψευδών θετικών - false positive rate), καθώς μεταβάλλεται το κατώφλι απόφασης. Κατώφλι απόφασης είναι μια παράμετρος του ταξινομητή η οποία καθορίζει την αναλογία μεταξύ αληθών και ψευδών θετικών ταξινομημένων στιγμιότυπων. Τέλος, ένα άλλο μέτρο αξιολόγησης που χρησιμοποιείται συχνά είναι το *εμβαδό κάτω από την καμπύλη ROC* (*Area Under ROC Curve – AUC*). Όταν το εμβαδό είναι ίσο με 1 τότε ο ταξινομητής επιτυγχάνει τέλεια ακρίβεια. Αντίθετα, ένας ταξινομητής που ταξινομεί τυχαία τα δεδομένα έχει AUC που ισούται με 0,5.

Διαδικασίες Αξιολόγησης Ταξινομητών

Η εκτίμηση της αποτελεσματικότητας ενός ταξινομητή σε νέα/άγνωστα δεδομένα προϋποθέτει την εφαρμογή μιας διαδικασίας αξιολόγησης και τη χρήση ενός συνόλου γνωστών δεδομένων (παραδειγμάτων με γνωστή τάξη). Υπάρχουν αρκετές προσεγγίσεις της αξιολόγησης ενός ταξινομητή, η απλούστερη των οποίων περιλαμβάνει την εξέταση της απόδοσης του ταξινομητή στο σύνολο εκπαίδευσης. Η αξιολόγηση ενός ταξινομητή στο σύνολο εκπαίδευσης παρέχει μια ιδιαίτερα «αισιόδοξη» εκτίμηση της απόδοσης του σε νέα στιγμιότυπα, επειδή η εκτίμηση αυτή πραγματοποιείται από τα ίδια δεδομένα που χρησιμοποιήθηκαν στην εκπαίδευση του μοντέλου. Επομένως, αυτή η προσέγγιση δεν είναι αρκετά αξιόπιστη.

Η αξιολόγηση ενός ταξινομητή μπορεί να γίνει πιο αξιόπιστη, αν χρησιμοποιηθεί ένα νέο σύνολο δεδομένων, το οποίο δε θα έχει συμμετάσχει σε κανένα στάδιο της εκπαίδευσης του ταξινομητή. Αυτό το σύνολο δεδομένων ονομάζεται *σύνολο δοκιμής* ή *σύνολο αξιολόγησης* (*test set* ή *evaluation set*). Σε πολλές περιπτώσεις χρησιμοποιείται και ένα επιπλέον σύνολο δεδομένων, το *σύνολο επικύρωσης* (*validation set*), που χρησιμοποιείται για τη βελτιστοποίηση (π.χ. επιλογή κατάλληλων παραμέτρων) του μοντέλου. Αυτή η προσέγγιση είναι κατάλληλη όταν υπάρχει πληθώρα δεδομένων, οπότε μπορούν να χρησιμοποιηθούν αρκετά μεγάλα σύνολα εκπαίδευσης και δοκιμής, ώστε να αποτελούν αντιπροσωπευτικά δείγματα του πληθυσμού του προβλήματος.

Πολλές φορές τα δεδομένα δεν είναι αρκετά, ώστε να χρησιμοποιηθούν αρκετά μεγάλα σύνολα εκπαίδευσης και δοκιμής. Σε αυτές τις περιπτώσεις ακολουθούνται διαφορετικές προσεγγίσεις με πιο διαδεδομένη και γενικά αποδεκτή τη μέθοδο της *σταυρωτής επικύρωσης* (*cross validation*). Η λειτουργία της μεθόδου στηρίζεται στη χρήση ενός συνόλου με γνωστά δεδομένα. Αρχικά, το σύνολο δεδομένων χωρίζεται σε

κ (*κ-πλή σταυρωτή επικύρωση*) ίσα υποσύνολα (τα υποσύνολα μπορεί να μην είναι ακριβώς ίσα, αλλά ενδέχεται κάποια να διαφέρουν μεταξύ τους κατά ένα παράδειγμα, λόγω μη ακέραιης διαίρεσης του συνολικού αριθμού παραδειγμάτων με το *κ*). Αφού γίνει ο διαχωρισμός του συνόλου, ξεκινά μια επαναλαμβανόμενη διαδικασία εκπαίδευσης και αξιολόγησης του ταξινομητή. Η διαδικασία λειτουργεί ως εξής: χρησιμοποιούνται τα *κ*-1 υποσύνολα (σύνολο εκπαίδευσης) για εκπαίδευση του ταξινομητή και το απομένων υποσύνολο (σύνολο αξιολόγησης) χρησιμοποιείται για την αξιολόγηση του εκπαιδευμένου ταξινομητή. Για την αξιολόγηση μπορεί να χρησιμοποιηθεί οποιοδήποτε από τα μέτρα που παρουσιάστηκαν παραπάνω, όπως π.χ. η ακρίβεια. Η διαδικασία επαναλαμβάνεται *κ* φορές, χρησιμοποιώντας διαφορετικό σύνολο αξιολόγησης κάθε φορά. Ο σταθμισμένος μέσος όρος της τιμής του μέτρου (π.χ. ακρίβειας) που υπολογίστηκε σε κάθε επανάληψη αποτελεί την εκτίμηση της αποτελεσματικότητας του ταξινομητή σε νέα/άγνωστα στιγμιότυπα.

Η αξιοπιστία της αξιολόγησης των ταξινομητών αυξάνεται όταν εφαρμόζεται *διαστρωμάτωση* (*stratification*) στα σύνολα εκπαίδευσης και αξιολόγησης. Η έννοια της διαστρωμάτωσης έγκειται στο ότι η σχετική αναλογία μεταξύ των τάξεων του αρχικού συνόλου δεδομένων διατηρείται και στα σύνολα εκπαίδευσης και αξιολόγησης. Έχει βρεθεί ότι η διαστρωματωμένη 10-πλή σταυρωτή επικύρωση (stratified 10-fold cross validation) αποτελεί μια από τις πιο αξιόπιστες προσεγγίσεις για την αξιολόγηση μοντέλων [Kohavi, 1995a].

2.5 Αποθήκες Δεδομένων

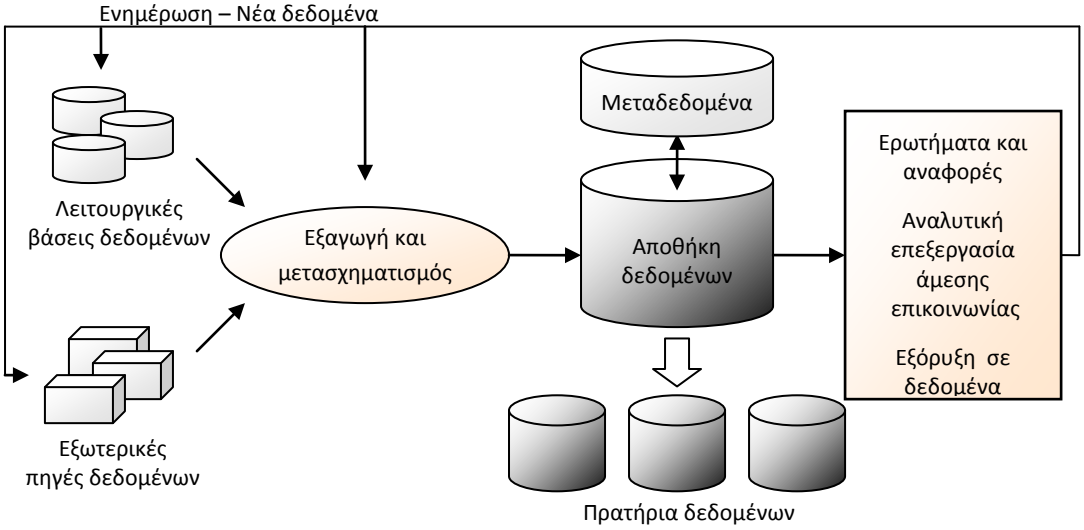
Οι μεγάλοι οργανισμοί αποτελούνται από πολλά τμήματα και τα δεδομένα που παράγονται συχνά βρίσκονται αποθηκευμένα σε διαφορετικές τοποθεσίες, διαφορετικές λειτουργικές βάσεις δεδομένων και κάτω από διαφορετικά σχήματα. Επιπλέον, οι διάφορες πηγές δεδομένων μπορούν να διαθέσουν άμεσα μόνο τα τρέχοντα δεδομένα. Ωστόσο, η λήψη αποφάσεων απαιτεί μια ενοποιημένη όψη όλων των δεδομένων, στα οποία συμπεριλαμβάνονται και τα ιστορικά δεδομένα. Αυτό ακριβώς το πρόβλημα αντιμετωπίζει μια *αποθήκη δεδομένων*.

Η *αποθήκη δεδομένων* (*data warehouse*) είναι μια βάση δεδομένων με εργαλεία αναφορών και ερωτημάτων στην οποία αποθηκεύονται τρέχοντα και ιστορικά δεδομένα που προέρχονται από διάφορα λειτουργικά (επεξεργασίας συναλλαγών) συστήματα και συνενώνονται με σκοπό την παραγωγή αναφορών και αναλύσεων για τη διοίκηση (management) [Laudon & Laudon, 2001]. Τα δεδομένα είναι συγκεντρωμένα κάτω από ένα ενοποιημένο σχήμα και σε μία τοποθεσία. Η συλλογή των δεδομένων στις αποθήκες δεδομένων συνήθως γίνεται με δύο τρόπους:

- *Οδηγούμενη από την πηγή* (*source driven*), όπου οι πηγές μεταφέρουν δεδομένα στην αποθήκη, είτε συνεχώς είτε περιοδικά (π.χ. τη νύχτα).

- *Οδηγούμενη από τον προορισμό* (*destination driven*), όπου η αποθήκη περιοδικά ζητά δεδομένα από τις πηγές.

Η διαδικασία δημιουργίας και χρήσης αποθηκών δεδομένων (*data warehousing*), περιγράφεται στην Εικόνα 2.2.



Εικόνα 2.2: Διαδικασία δημιουργίας και χρήσης αποθήκης δεδομένων.

Τα δεδομένα εξάγονται από τις διάφορες πηγές δεδομένων. Αυτά αναδιοργανώνονται σε μια κεντρική βάση δεδομένων για την παραγωγή αναφορών και αναλύσεων, καθώς και για τη διευκόλυνση της εξόρυξης από τα δεδομένα. Τα *μεταδεδομένα* (*metadata*) περιέχουν πληροφορίες σχετικές με τα αποθηκευμένα δεδομένα και τον τρόπο οργάνωσής τους. Τα δεδομένα ανανεώνονται περιοδικά αντικατοπτρίζοντας τις αλλαγές στις πηγές και εκείνα που είναι πολύ παλιά απομακρύνονται από την αποθήκη (ενδεχομένως σε αρχειακά μέσα). Νέες πληροφορίες, όπως κανόνες, που δημιουργούνται από τις διάφορες διεργασίες πάνω στα δεδομένα (εξόρυξη, OLAP κλπ.), μπορούν να αποθηκευτούν στην αποθήκη δεδομένων. Οι οργανισμοί μπορούν να δημιουργήσουν μικρότερες και αποκεντρωμένες αποθήκες που αποτελούν υποσύνολο μιας αποθήκης δεδομένων. Στις αποθήκες αυτές, που ονομάζονται *πρατήρια δεδομένων* (*data marts*), αποθηκεύεται συνοπτικό ή συγκεκριμένου ενδιαφέροντος τμήμα των δεδομένων και απευθύνεται σε συγκεκριμένο πληθυσμό χρηστών. Η *αναλυτική επεξεργασία άμεσης επικοινωνίας* (*on-line analytical processing – OLAP*) παρέχει τη δυνατότητα χειρισμού και πολυδιάστατης ανάλυσης μεγάλων όγκων δεδομένων.

2.6 Προβλήματα της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων

Κατά την εφαρμογή της διαδικασίας ανακάλυψης γνώσης από βάσεις δεδομένων παρουσιάζονται διάφορα προβλήματα που πρέπει να αντιμετωπιστούν και πολλές φορές προβάλλουν μεγάλες προκλήσεις και δίνουν το έναυσμα για περαιτέρω έρευνα.

Τα προβλήματα αυτά που σχετίζονται με την ποιότητα και τον όγκο των δεδομένων, τη διαχείριση των αποτελεσμάτων, τη μεθοδολογία της εξόρυξης από δεδομένα και την απόδοση των αλγορίθμων εξόρυξης από δεδομένα παρουσιάζονται παρακάτω σύμφωνα με τους Holsheimer & Siebes [1994], Han & Kamber [2001] και Βλαχάβας, κ.α. [2006].

2.6.1 Ποιότητα Δεδομένων

Προβλήματα που σχετίζονται με την ποιότητα των δεδομένων παρουσιάζονται συνήθως λόγω της έλλειψης, της πολυπλοκότητας και της ετερογένειας των δεδομένων, καθώς και λόγω της παρουσίας θορύβου σε αυτά.

Περιορισμένη Πληροφορία

Συχνά κάποια χαρακτηριστικά που θα απλοποιούσαν τη διαδικασία ανακάλυψης γνώσης απουσιάζουν ή είναι αδύνατο να συλλεχθούν. Γι' αυτό το λόγο δεν είναι πάντοτε εφικτή η κατασκευή αποτελεσματικών μοντέλων βάσει των διαθέσιμων μόνο χαρακτηριστικών.

Αραιά Δεδομένα

Συνήθως ένας ταξινομητής πρέπει να ανακαλύψει τα όρια των τάξεων. Η ακριβής θέση των ορίων μπορεί να βρεθεί μόνο αν υπάρχουν αρκετά παραδείγματα που εντοπίζονται κοντά στα όρια των τάξεων. Σε αντίθετη περίπτωση (αραιά δεδομένα) τα όρια των τάξεων δεν είναι δυνατό να καθοριστούν με ακρίβεια.

Χαρακτηριστικά χωρίς Τιμή

Ένα πρόβλημα που σχετίζεται με τη χρήση των βάσεων δεδομένων είναι η ενδεχόμενη απουσία τιμής κάποιων χαρακτηριστικών. Ο χειρισμός τέτοιων δεδομένων ποικίλλει και οι μεθοδολογίες που έχουν αναπτυχθεί ενδεικτικά περιλαμβάνουν την απόρριψη των εγγραφών με τα πεδία που δεν έχουν τιμή, τον υπολογισμό των τιμών που λείπουν με βάση άλλες εγγραφές, τον καθορισμό κάποιας ειδικής τιμής που να αντιστοιχεί στις τιμές που δεν υπάρχουν, κλπ.

Θόρυβος

Τα δεδομένα μιας βάσης δεδομένων μπορεί να προέρχονται από μετρήσεις ή από υποκειμενικές κρίσεις με αποτέλεσμα να υπάρχουν λανθασμένες καταχωρήσεις. Μια τέτοια περίπτωση μπορεί να οδηγήσει σε ανακριβή πρότυπα και μοντέλα. Τέτοια λάθη στις τιμές των χαρακτηριστικών ή στην περιγραφή των τάξεων είναι γνωστά ως *θόρυβος* (*noise*). Η απαλοιφή του θορύβου από τα δεδομένα θεωρείται απαραίτητη. Ωστόσο, πειράματα σε κάποια συστήματα έχουν δείξει ότι η προσθήκη θορύβου στα δεδομένα οδηγεί σε ικανοποιητική αποτελεσματικότητα ταξινόμησης των παραδειγμάτων, ακόμη και αν ο θόρυβος είναι σημαντικός. Ιδιαίτερο ενδιαφέρον, όπως περιγράφει ο J.R. Quinlan [1986a; 1986b], παρουσιάζει το φαινόμενο κατά το οποίο μοντέλα που παράγονται από ένα σύνολο εκπαίδευσης με θόρυβο, λειτουργούν καλύτερα όταν εφαρμόζονται σε δεδομένα με θόρυβο εν συγκρίσει με μοντέλα που έχουν προκύψει από το ίδιο «αθόρυβο» σύνολο δεδομένων. Το συμπέρασμα που προκύπτει είναι ότι

είναι άσκοπο να καταβάλλεται προσπάθεια για την εξάλειψη του θορύβου από το σύνολο εκπαίδευσης, αν το μοντέλο που θα παραχθεί εφαρμόζεται σε δεδομένα με σημαντική ποσότητα θορύβου.

Διαχείριση Πολύπλοκων Τύπων Δεδομένων

Η δημιουργία ενός συστήματος που θα διαχειρίζεται όλων των ειδών τα δεδομένα και θα εφαρμόζει σε αυτά οποιαδήποτε μέθοδο εξόρυξης από δεδομένα αποδοτικά και αποτελεσματικά είναι αδύνατη. Ωστόσο, η ύπαρξη βάσεων δεδομένων, που περιέχουν δεδομένα διαφόρων τύπων, από τα πιο απλά μέχρι τα πιο σύνθετα, όπως αντικείμενα, πολυμέσα, υπερκείμενο, χωρικά, χρονικά, κ.α. επιτάσσει την ανάγκη ανάπτυξης εξειδικευμένων συστημάτων.

Ανακάλυψη Γνώσης από Ετερογενείς Βάσεις Δεδομένων

Ιδιαίτερα ενδιαφέρουσα είναι η ανακάλυψη γνώσης από μεγάλες, ετερογενείς βάσεις δεδομένων. Η εξόρυξη από διαφορετικές πηγές δομημένων, ημιδομημένων και αδόμητων δεδομένων που συνήθως ποικίλουν και σε σημασιολογικό επίπεδο, αποτελεί μια μεγάλη πρόκληση. Η μέθοδοι εξόρυξης από δεδομένα μπορούν να συντελέσουν στην ανακάλυψη κανονικοτήτων και συσχετίσεων υψηλού επιπέδου, που θα ήταν αδύνατον να βρεθούν με τα κλασικά εργαλεία.

2.6.2 Όγκος Δεδομένων

Τα προβλήματα αυτής της κατηγορίας σχετίζονται με το μέγεθος της βάσης δεδομένων και αφορούν το πλήθος των χαρακτηριστικών και το πλήθος των εγγραφών.

Δειγματοληψία

Οι βάσεις δεδομένων συνήθως έχουν τεράστιο αριθμό εγγραφών και πολλοί αλγόριθμοι ανακάλυψης γνώσης, λόγω προβλημάτων αποδοτικότητας, είναι αδύνατο να εφαρμοστούν σε όλα τα δεδομένα. Γι’ αυτόν το λόγο είναι απαραίτητη η λήψη ενός δείγματος από τη βάση δεδομένων. Κατά την επιλογή του δείγματος απαιτείται ιδιαίτερα προσεκτική εφαρμογή στατιστικών τεχνικών, ώστε αυτό να αντιπροσωπεύει ικανοποιητικά τη βάση δεδομένων.

Πλήθος Χαρακτηριστικών

Ο μεγάλος αριθμός χαρακτηριστικών που περιγράφουν ένα στιγμιότυπο είναι ένα πλεονέκτημα, αφού είναι πιθανότερη η εμφάνιση πραγματικών συσχετίσεων. Ωστόσο, μπορεί να δημιουργήσει προβλήματα αποδοτικότητας και αποτελεσματικότητας. Χαρακτηριστικό παράδειγμα δεδομένων με πολύ μεγάλο αριθμό πεδίων είναι τα δεδομένα γονιδιακής έκφρασης. Λύση στο πρόβλημα παρέχουν οι τεχνικές μείωσης διαστάσεων (Εδάφιο 2.4.1).

2.6.3 Διαχείριση Αποτελεσμάτων

Τα προβλήματα αυτής της κατηγορίας σχετίζονται με την παρουσίαση και την αξιολόγηση των αποτελεσμάτων της ανακάλυψης γνώσης από βάσεις δεδομένων.

Παρουσίαση και Οπτικοποίηση των Αποτελεσμάτων

Η γνώση που ανακαλύπτεται πρέπει να μπορεί να εκφραστεί με γλώσσες υψηλού επιπέδου, οπτικές αναπαραστάσεις (δένδρα, γραφήματα, καμπύλες, πίνακες, κλπ.) ή άλλες μορφές έκφρασης, έτσι ώστε να είναι όσο πιο κατανοητή γίνεται από τον άνθρωπο. Αυτό γίνεται ακόμη πιο αναγκαίο, όταν το σύστημα ανακάλυψης γνώσης είναι αλληλεπιδραστικό.

Αξιολόγηση Μοντέλων και Προτύπων

Ένα σύστημα ανακάλυψης γνώσης από δεδομένα μπορεί να ανακαλύψει έναν μεγάλο αριθμό προτύπων. Ωστόσο, λίγα από αυτά είναι πραγματικά ενδιαφέροντα και χρήσιμα, ενώ τα περισσότερα είναι ακατανόητα, εκφράζουν κοινή λογική ή δεν είναι πρωτότυπα. Ένα ιδιαίτερα σημαντικό αντικείμενο έρευνας είναι η ανάπτυξη τεχνικών αξιολόγησης της παραγόμενης γνώσης, οι οποίες πρέπει να λαμβάνουν υπόψη όχι μόνο στατιστικούς αλλά και υποκειμενικούς παράγοντες που έχουν να κάνουν με τις αντιλήψεις, τα ενδιαφέροντα και τις προσδοκίες του χρήστη.

2.6.4 Προβλήματα Μεθοδολογίας

Τα προβλήματα της διαδικασίας ανακάλυψης γνώσης που αφορούν τη μεθοδολογία και την αλληλεπίδραση με το χρήστη παρουσιάζονται παρακάτω.

Εξόρυξη από Δεδομένα με Προστασία Απορρήτου

Η προστασία του απορρήτου των δεδομένων είναι ένα από τα προβλήματα της εξόρυξης από δεδομένα με το μεγαλύτερο ενδεχομένως κοινωνικό αντίκτυπο. Η διαδικασία ανακάλυψης γνώσης από βάσεις δεδομένων, που έχει πρόσβαση σε ευαίσθητα δεδομένα ενός ατόμου, οφείλει να λαμβάνει όλα εκείνα τα μέτρα που απαιτούνται για την προστασία του από αθέμιτη χρήση. Μάλιστα, ορισμένες φορές τα προς επεξεργασία δεδομένα είναι σκόπιμα αλλοιωμένα, ώστε να αποφευχθεί κάτι τέτοιο. Συνεπώς, υφίσταται το πρόβλημα της αξιολόγησης της ακεραιότητας όχι μόνο των δεδομένων αλλά και της παραγόμενης γνώσης.

Αλληλεπιδραστική Εξόρυξη από Δεδομένα σε Διάφορα Ιεραρχικά Επίπεδα

Η διαδικασία εξόρυξης από δεδομένα πρέπει να είναι αλληλεπιδραστική, καθώς δεν είναι εκ των προτέρων γνωστό τι ακριβώς θα ανακαλύψει. Ο χρήστης πρέπει να έχει τη δυνατότητα να παρέμβει στη διαδικασία, εστιάζοντας στα πρότυπα που θεωρεί χρήσιμα και ενδιαφέροντα.

Ενσωμάτωση Υπάρχουσας Γνώσης

Για την καθοδήγηση της διαδικασίας ανακάλυψης γνώσης είναι δυνατόν να χρησιμοποιηθεί η υπάρχουσα γνώση με διάφορους τρόπους, όπως για την αξιολόγηση των αποτελεσμάτων, τον προσανατολισμό της διαδικασίας στην εξαγωγή προτύπων διαφορετικών επιπέδων και την επιτάχυνση της διαδικασίας για αποφυγή της άσκοπης αναζήτησης.

Γλώσσες Ερωτημάτων Εξόρυξης από Δεδομένα

Οι γλώσσες ερωτημάτων, όπως η SQL, επιτρέπουν στους χρήστες να υποβάλλουν ερωτήματα στις σχεσιακές βάσεις δεδομένων για την ανάκτηση δεδομένων που περιέχονται σε αυτές. Ομοίως, η ανάπτυξη υψηλού επιπέδου γλωσσών ερωτημάτων εξόρυξης από δεδομένα είναι ιδιαίτερα χρήσιμη. Οι γλώσσες αυτές πρέπει να υλοποιούν μεθόδους εξόρυξης, έτσι ώστε να ενσωματώνουν την προηγούμενη γνώση, να μπορούν να παραμετροποιηθούν κατάλληλα και να θέτουν περιορισμούς και κανόνες σχετικά με την αναζητούμενη γνώση.

2.6.5 Προβλήματα Απόδοσης

Στη συγκεκριμένη κατηγορία ανήκουν προβλήματα που σχετίζονται κυρίως με την αποδοτικότητα, άλλα και με την αποτελεσματικότητα των αλγορίθμων εξόρυξης από δεδομένα.

Αποδοτικότητα και Κλιμάκωση των Αλγορίθμων

Η κλιμάκωση (scalability) είναι ένα χαρακτηριστικό που πρέπει να έχουν οι αλγόριθμοι εξόρυξης από δεδομένα προκειμένου να μπορούν να διαχειριστούν δεδομένα των οποίων το πλήθος μπορεί να κλιμακώνεται σε διαφορετικές τάξεις μεγέθους. Γενικότερα, οι αλγόριθμοι πρέπει να είναι αποδοτικοί, ώστε να έχουν προβλέψιμο χρόνο εκτέλεσης.

Παράλληλοι και Κατανεμημένοι Αλγόριθμοι

Το τεράστιο μέγεθος πολλών βάσεων δεδομένων, το εύρος κατανομής των δεδομένων τους και η υπολογιστική πολυπλοκότητα ορισμένων αλγορίθμων εξόρυξης από δεδομένα είναι τα βασικότερα κίνητρα για την ανάπτυξη παράλληλων και κατανεμημένων αλγορίθμων. Οι αλγόριθμοι αυτοί διαιρούν τα δεδομένα σε τμήματα, και στη συνέχεια τα επεξεργάζονται παράλληλα. Τα επιμέρους αποτελέσματα τελικά συνενώνονται κατάλληλα.

Επαυξητικές Μέθοδοι

Οι βάσεις δεδομένων ενημερώνονται συχνά. Νέα δεδομένα προστίθενται, κάποια τροποποιούνται και κάποια άλλα απομακρύνονται. Οποιαδήποτε γνώση είχε εξαχθεί από μια μη ενημερωμένη βάση, ενδέχεται να μην είναι συνεπής με τα νέα δεδομένα. Είναι προφανές ότι ένα σύστημα ανακάλυψης γνώσης θα πρέπει να προσαρμόζεται σε τέτοιου είδους αλλαγές. Επιπλέον, είναι σημαντική η διατήρηση της συνέπειας των μοντέλων με τα πιο πρόσφατα δεδομένα, καθώς τα χαρακτηριστικά των αντικειμένων μπορεί να αλλάζουν με το χρόνο εξαιτίας τάσεων που κυριαρχούν και διεργασιών που εκτελούνται στο περιβάλλον. Η ανακατασκευή ενός μοντέλου μπορεί να γίνει από την αρχή, όμως είναι πιο βολικό να χρησιμοποιηθεί μια *επαυξητική* (*incremental*) μέθοδος εξόρυξης από δεδομένα. Κατά τη διαδικασία αυτή χρησιμοποιείται γνώση που έχει παραχθεί προηγουμένως για την ανακατασκευή του μοντέλου.

2.7 Εφαρμογές της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων

Η ανακάλυψη γνώσης από βάσεις δεδομένων βρίσκει εφαρμογή σε πολλούς τομείς. Ουσιαστικά, οποιοσδήποτε τομέας έχει τη δυνατότητα να παρέχει μεγάλο πλήθος δεδομένων είναι κατάλληλο πεδίο εφαρμογής της ανακάλυψης γνώσης. Στην παρούσα ενότητα παρουσιάζονται μερικές από τις δημοφιλέστερες εφαρμογές της διαδικασίας ανακάλυψης γνώσης από βάσεις δεδομένων. Οι εφαρμογές της διαδικασίας στους τομείς της βιολογίας και της βιοπληροφορικής παρουσιάζονται εκτενώς στο επόμενο κεφάλαιο.

Η ανακάλυψη γνώσης από βάσεις δεδομένων βρίσκει εφαρμογή στη διαχείριση πελατειακών σχέσεων (customer relationship management) και συγκεκριμένα μπορεί να συνεισφέρει στην απόκτηση νέων πελατών, στην αύξηση των κερδών από τους υπάρχοντες πελάτες, στη διατήρηση των καλών πελατών, στη στοχευμένη διαφήμιση, κ.α.

Άλλη δημοφιλής εφαρμογή είναι η ανάλυση κινδύνου (risk analysis), που περιλαμβάνει προβλήματα λήψης αποφάσεων με κάποιο βαθμό κινδύνου, όπως είναι η έκδοση δανείου ή πιστωτικής κάρτας σε πελάτες τραπεζών.

Η ανίχνευση απάτης (fraud detection) αποτελεί, επίσης, μια πολύ δημοφιλή εφαρμογή της ανακάλυψης γνώσης με άμεσα ενδιαφερόμενους φορείς τις εταιρείες τηλεπικοινωνιών, τις τράπεζες και τις ασφαλιστικές εταιρείες.

Ο καθορισμός των χαρακτηριστικών της αγοράς και της βιομηχανίας, η τμηματοποίηση της αγοράς (market segmentation) και η διαχείριση αποθεμάτων (stock management) αποτελούν μερικές ακόμη εφαρμογές της ανακάλυψης γνώσης που ενδιαφέρουν πολλές επιχειρήσεις με δραστηριότητα στον οικονομικό τομέα.

Η ιατρική είναι ακόμη μια πρόσφορη περιοχή για την εφαρμογή της ανακάλυψης γνώσης, με χαρακτηριστικό παράδειγμα την πρόβλεψη της αποτελεσματικότητας των χειρουργικών διαδικασιών, των ιατρικών εξετάσεων ή των φαρμάκων. [Edelstein, 1999].

Η ανακάλυψη γνώσης από βάσεις δεδομένων εφαρμόζεται και στον τομέα του περιβάλλοντος με κύριο στόχο την πρόβλεψη μεταβλητών που προσδιορίζουν τη ρύπανση της ατμόσφαιρας, των υδάτων, του εδάφους και του υπεδάφους, καθώς και την εξέλιξη των κλιματικών συνθηκών. Παράδειγμα τέτοιας εφαρμογής είναι η πρόβλεψη μελλοντικών τιμών μεταβλητών που προσδιορίζουν την ποιότητα του νερού, βάσει μετρήσεων από υποβρύχιους αισθητήρες [Hatzikos et al., 2008].

2.8 Σύνοψη

Στο κεφάλαιο αυτό παρουσιάστηκε το πεδίο της ανακάλυψης γνώσης από βάσεις δεδομένων. Παρουσιάστηκαν τα στάδια που περιλαμβάνει η διαδικασία ανακάλυψης γνώσης από βάσεις δεδομένων καθώς και κάποιοι από τους βασικότερους αλγορίθμους που χρησιμοποιούνται σε καθένα από τα στάδια. Η πλειοψηφία των αλγορίθμων αυτών εφαρμόστηκαν, τροποποιήθηκαν, ή επεκτάθηκαν κατά την εκπόνηση της πα-

ρούσας διατριβής. Επίσης, περιγράφηκε η σχέση της ανακάλυψης γνώσης με τη συγγενική περιοχή της μηχανικής μάθησης καθώς και κάποια προβλήματα που αντιμετωπίζονται κατά την εφαρμογή της διαδικασίας. Τέλος, παρουσιάστηκαν μερικά από τα βασικότερα πεδία εφαρμογής της διαδικασίας.

Μοριακή Βιολογία και Βιοπληροφορική

3

Το αντικείμενο της επιστήμης της βιολογίας είναι η μελέτη του φαινομένου της ζωής. Η ζωή είναι εξαιρετικά σύνθετη και παρουσιάζει τρία βασικά χαρακτηριστικά, την *ποικιλότητα* (*diversity*), την *ενότητα* (*unity*) και την *εξέλιξη* (*evolution*) [Hunter, 1993; 2004].

Τα διάφορα είδη (species) των οργανισμών παρουσιάζουν πολλές και μεγάλες διαφορές μεταξύ τους. Ακόμη και άτομα που ανήκουν στο ίδιο είδος διαφέρουν αρκετά. Αυτή η ποικιλότητα στα χαρακτηριστικά και τις δραστηριότητες των οργανισμών φτάνει ως το μοριακό επίπεδο. Δεδομένης αυτής της ποικιλότητας είναι εξαιρετικά εκπληκτική η ανακάλυψη της σχεδόν καθολικής ομοιότητας των οργανισμών στη μοριακή τους λεπτομέρεια. Όλοι οι ζωντανοί οργανισμοί βασίζονται στη δραστηριότητα μιας οικογένειας μορίων, που ονομάζονται *πρωτεΐνες* (*proteins*). Ωστόσο, η ενότητα μεταξύ των οργανισμών δεν οφείλεται μόνο στο γεγονός ότι οι πρωτεΐνες επιτελούν την περισσότερη βιοχημική εργασία που απαιτεί η ζωή. Η κύρια παρατήρηση που στηρίζει την ενότητα αυτή είναι ότι σύνολα πρωτεϊνών που μοιάζουν πολύ μεταξύ τους και έχουν παρόμοια λειτουργία εμφανίζονται σε εξαιρετικά διαφορετικούς οργανισμούς.

Επίσης, ένα ακόμη κοινό χαρακτηριστικό όλων των οργανισμών είναι η παρουσία μιας άλλης οικογένειας μορίων, των *νουκλεϊκών οξέων* (*nucleic acids*), που έχουν το ρόλο της διατήρησης και μεταβίβασης της πληροφορίας που συνθέτει τη ζωή. Οι πρωτεΐνες και τα νουκλεϊκά οξέα είναι μόρια υψηλού μοριακού βάρους και ως εκ τούτου καλούνται βιολογικά *μακρομόρια* (*macromolecules*). Επιπλέον, χαρακτηρίζονται και ως *πολυμερή* (*polymers*), καθώς αποτελούνται από επαναλαμβανόμενα μικρότερα μόρια, που καλούνται *μονομερή* (*monomers*).

Μία από τις σημαντικότερες έννοιες της βιολογίας είναι η εξέλιξη. Όλοι οι οργανισμοί είναι μέρη μιας συνεχούς γραμμής προγόνων και απογόνων και κάθε ζεύγος οργανισμών έχει έναν κοινό πρόγονο. Η παρουσία κοινών προγόνων αποτελεί σημαντι-

κό μέρος της εξήγησης της ομοιότητας που παρατηρείται μέσα στις οικογένειες των οργανισμών. Τρία βασικά φαινόμενα συνθέτουν την εξέλιξη [Hunter, 1993; 2004]:

- Η *κληρονομικότητα* (*inheritance*), που είναι η μεταβίβαση των χαρακτηριστικών των γονέων στους απογόνους.
- Η *μεταβλητότητα* (*variation*), που αναφέρεται στη διαφοροποίηση των απογόνων, ώστε να μην αποτελούν ακριβή αντίγραφα των γονέων τους.
- Η *επιλογή* (*selection*), που είναι η διαδικασία που ευνοεί τη δυνατότητα αναπαραγωγής κάποιων οργανισμών σε σχέση με άλλους.

Η συνισταμένη των τριών παραπάνω φαινομένων, δηλαδή η εξέλιξη, είναι η δύναμη που δημιούργησε την ενότητα και την ποικιλότητα των ζωντανών οργανισμών.

Σημαντική προσπάθεια για την κατανόηση της ζωής καταβάλλεται με τη μελέτη της δομής και της λειτουργίας των βιολογικών μακρομορίων. Ο κλάδος της βιολογίας που ασχολείται με αυτή τη μελέτη ονομάζεται *μοριακή βιολογία* (*molecular biology*).

3.1 Οργανισμοί και Κύτταρα

Η θεμελιώδης δομική και λειτουργική μονάδα κάθε ζωντανού οργανισμού είναι το *κύτταρο* (*cell*). Τα περισσότερα κύτταρα είναι πολύ μικρα και δεν μπορούν να φανούν με γυμνό μάτι, αλλά είναι επαρκώς μεγάλα για να παρατηρηθούν με ένα οπτικό μικροσκόπιο. Κάθε κύτταρο είναι ένα σύνθετο σύστημα που αποτελείται από πολλές διαφορετικές δομικές μονάδες που περικλείονται από την *πλασματική μεμβράνη* (*plasma membrane*). Η πλασματική μεμβράνη είναι μια σύνθετη δομή που οριοθετεί το κύτταρο από το περιβάλλον του και ρυθμίζει τη ροή των ουσιών, της ενέργειας και των πληροφοριών από και προς το κύτταρο. Υπάρχουν *μονοκύτταροι* (*unicellular*) οργανισμοί, που αποτελούνται από ένα μόνο κύτταρο και *πολυκύτταροι* (*multicellular*), που αποτελούνται από πολλά. Τα βακτήρια και η μαγιά είναι παραδείγματα μονοκύτταρων οργανισμών. Οποιοδήποτε κύτταρο είναι σε θέση να επιζήσει και να πολλαπλασιαστεί ανεξάρτητα στο κατάλληλο περιβάλλον.

Εκτιμάται ότι ο αριθμός των κυττάρων στον ανθρώπινο οργανισμό είναι της τάξης των 10^{14}. Τα κύταρρα αυτά ταξινομούνται σε εκατοντάδες διαφορετικούς τύπους. Ωστόσο, η ταξινόμηση των κυττάρων δεν γίνεται με μοναδικό τρόπο και εξαρτάται από το επίπεδο λεπτομέρειας που επιλέγεται να χρησιμοποιηθεί για τη διάκριση των κυτταρικών τύπων. Για παράδειγμα, είναι αδύνατο να βρεθούν δύο ίδια κύτταρα σε έναν οργανισμό εάν χρησιμοποιηθεί ως κριτήριο σύγκρισης ο αριθμός των μορίων τους.

Πολύτιμα εργαλεία για την προσέγγιση της ζωής και την αναγνώριση των μορφών της αποτελούν η *ταξινόμηση* (*biological classification*), η επιστήμη δηλαδή που έχει ως αντικείμενο την κατάταξη των έμβιων όντων σε ομάδες και η *συστηματική* (*systematics*), που περιλαμβάνει τη μελέτη των σχέσεων μεταξύ των οργανισμών. Οι

παραπάνω επιστημονικοί κλάδοι συνδέονται άμεσα μεταξύ τους, καθώς η ταξινομική κατάταξη των οργανισμών αντανακλά τη φυλογενετική τους εξέλιξη στο χρόνο. Ήδη από τον 4ο αιώνα π.Χ. (Αριστοτέλης, 350 π.Χ. - ταξινόμηση ζώων και Θεόφραστος, 320 π.Χ. - ταξινόμηση φυτών) γίνονται απόπειρες ταξινόμησης των τότε γνωστών οργανισμών σε υποομάδες των δύο παραδοσιακά αποδεκτών κατηγοριών, των φυτών και των ζώων.

Οι διάφορες μορφές ζωής, σύγχρονες και παλαιότερες, κατατάσσονται με βάση ανατομικά, μορφολογικά, βιοχημικά, φυσιολογικά, μοριακά, γενετικά και φυλογενετικά κριτήρια. Το 1969 ο αμερικανός οικολόγος Robert Whittaker [Whittaker, 1969] καθιέρωσε το σύστημα ταξινόμησης των πέντε βασιλείων που χρησιμοποιείται έως σήμερα, μολονότι πολλοί επιστήμονες το θεωρούν παρωχημένο. Ο Πίνακας 3.1 παρουσιάζει τις σημαντικότερες ταξινομήσεις που έχουν προταθεί τους τελευταίους αιώνες. Η τελευταία πρόταση ταξινόμησης των ειδών [Woese *et al.*, 1990] κατατάσσει τους οργανισμούς σε τρεις *επικράτειες* (*domains*):

- Τα *βακτήρια* (*bacteria*), τα οποία είναι μονοκύτταροι οργανισμοί και δε διαθέτουν πυρήνα. Το μήκος τους μπορεί να φτάσει τα μερικά μικρόμετρα (μm) και συναντώνται σε διάφορα σχήματα, κυρίως σφαιρικά (κόκκοι), ραβδοειδή (βάκιλοι) και σπειροειδή.

- Τα *αρχαία* (*archaea*), τα οποία όπως και τα βακτήρια, είναι μονοκύτταροι οργανισμού και δε διαθέτουν πυρήνα. Παρά τις ομοιότητες τους με τα βακτήρια ταξινομούνται σε διαφορετική επικράτεια, καθώς διαφέρουν στην πορεία εξέλιξής τους.

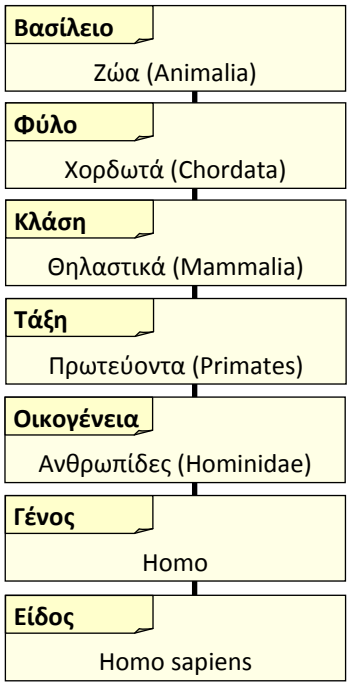
- Τα *ευκάρυα* (*eukarya*), στα οποία ανήκουν οι περισσότεροι οργανισμοί τους οποίους μπορούμε να δούμε, όπως ο άνθρωπος, τα ζώα, τα φυτά, οι μύκητες. Το κύριο μορφολογικό χαρακτηριστικό των ευκαρυωτικών κυττάρων είναι η ύπαρξη ενός σχηματισμού που βρίσκεται συνήθως στο κέντρο τους, τον *πυρήνα* (*nucleus*), ο οποίος περιβάλλεται από μεμβράνη και περικλείει το γενετικό τους υλικό. Αυτός είναι και ο λόγος για τον οποίο ονομάζονται ευκάρυα (ευ + κάρυον), όπου το κάρυον υποδηλώνει τον πυρήνα.

Αρκετά συχνά οι δύο πρώτες επικράτειες (βακτήρια και αρχαία) αναφέρονται και ως *προκαρυώτες* (*prokaryotes*), λόγω της απουσίας πυρήνα, ενώ τα ευκάρυα, αναφέρονται και ως *ευκαρυώτες* (*eukaryotes*).

Για τη συστηματική ταξινόμηση των ειδών ακολουθείται μια κοινή διεθνώς ονοματολογία, η οποία περιλαμβάνει τις εξής βασικές κατηγορίες της ταξινομικής ιεραρχίας: το *βασίλειο* (*kingdom*), το *φύλο* (*phylum*), την *κλάση* (*class*), την *τάξη* (*order*), την *οικογένεια* (*family*), το *γένος* (*genus*) και το *είδος* (*species*). Η ταυτότητα του σύγχρονου ανθρώπου σύμφωνα με αυτή την πρότυπη ταξινομική ιεραρχία απεικονίζεται στην Εικόνα 3.1.

Πίνακας 3.1: Οι προτεινόμενες ταξινομήσεις των έμβιων όντων.

1735	1866	1937	1956	1969	1977	1990
Linnaeus	**Haeckel**	**Chatton**	**Copeland**	**Whittaker**	**Balch *et al*.**	**Woese *et al*.**
	Πρώτιστα	Προκάρυα	Μονήρη	Μονήρη	Ευβακτήρια	Βακτήρια
					Αρχαιοβακτήρια	Αρχαία
		Ευκάρυα	Πρώτιστα	Πρώτιστα	Πρώτιστα	Ευκάρυα
Φυτά	Φυτά			Μύκητες	Μύκητες	
			Φυτά	Φυτά	Φυτά	
Ζώα	Ζώα		Ζώα	Ζώα	Ζώα	



Εικόνα 3.1: Η ταυτότητα του σύγχρονου ανθρώπου.

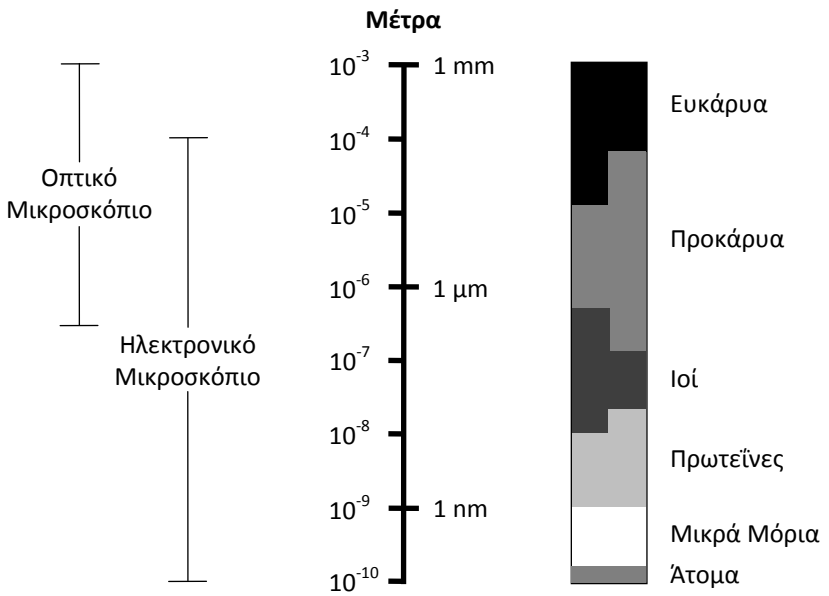
Εκτός από το σύνολο των οργανισμών που περιγράφηκαν παραπάνω, υπάρχει και μια σειρά από «κατώτερους» οργανισμούς, που είναι ενδοκυτταρικά παράσιτα:

- Οι *ιοί* (*viruses*), που δε μοιάζουν με ζωντανούς οργανισμούς, αλλά όταν βρεθούν μέσα σε ένα ζωντανό *κύτταρο-ξενιστή* (*host cell*) παρουσιάζουν μερικά χαρακτηριστικά γνωρίσματα ενός ζωντανού οργανισμού. Μάλιστα, υπάρχει διχογνωμία στην κοινότητα των βιολόγων σχετικά με το αν ένας ιός είναι ζωντανός οργανισμός ή όχι.
- Τα *ιοειδή* (*viroids*), τα οποία αποτελούνται μόνο από ένα μικρό κυκλικό μόριο RNA (Εδάφιο 3.2.4). Βρίσκονται στον πυρήνα των κυττάρων που μολύνουν και δε

συναντώνται ελεύθερα. Όσα έχουν αναγνωριστεί μέχρι σήμερα σχετίζονται με ασθένειες των φυτών.

- Τα *prions*, τα οποία είναι μολυσματικοί φορείς που αποτελούνται μόνο από πρωτεΐνη η οποία έχει λανθασμένη στεροδιάταξη. Σε αυτά οφείλεται ένας αριθμός από ασθένειες που προσβάλλουν διάφορα θηλαστικά συμπεριλαμβανομένης της νόσου Creutzfeldt-Jakob, που προσβάλλει τον άνθρωπο, της σπογγώδους εγκεφαλοπάθειας, που προσβάλλει τα βοοειδή και της νόσου Scrapie, που προσβάλλει τα αιγοπρόβατα.

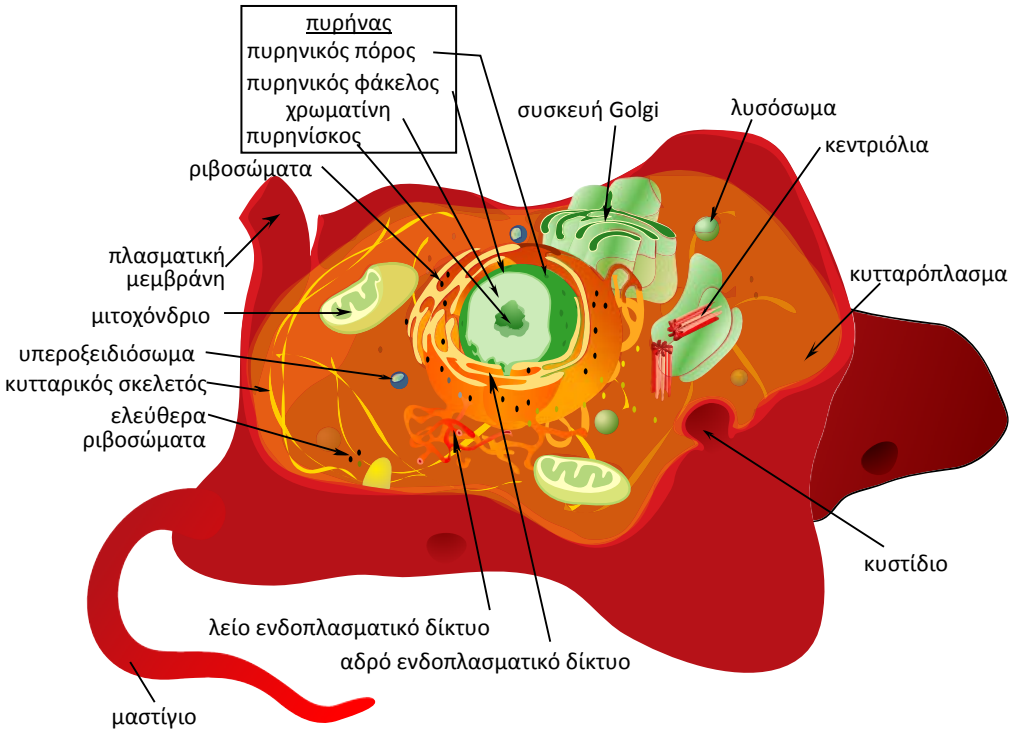
Τα προκαρυωτικά κύτταρα (δηλαδή τα βακτήρια και τα αρχαία) είναι μικρότερα από τα ευκαρυωτικά και έχουν απλούστερη δομή. Για παράδειγμα, δεν έχουν καθόλου εσωτερικές κυτταρικές μεμβράνες, οι οποίες είναι πάντα παρούσες στα ευκαρυωτικά κύτταρα και τα διαχωρίζουν σε εσωτερικά διαμερίσματα. Επίσης, όπως ήδη αναφέρθηκε, όλοι οι προκαρυώτες είναι μονοκύτταροι. Ωστόσο, αυτό δε σημαίνει ότι ένας μονοκύτταρος οργανισμός είναι πάντα προκαρυώτης. Το γεγονός ότι οι προκαρυώτες έχουν μικρότερο μέγεθος από τους ευκαρυώτες δεν τους κάνει λιγότερο σημαντικούς. Οι προκαρυώτες είναι πολλές φορές γνωστοί και ως μικρόβια, εξαιτίας του μικρού τους μεγέθους. Ακόμη μικρότερο είναι το μέγεθος των ιών, των ιοειδών και των prions, τα οποία δεν είναι εμφανή με το οπτικό μικροσκόπιο, αλλά μπορούν να αποκαλύψουν τη δομή τους στο ηλεκτρονικό μικροσκόπιο. Τα τυπικά μεγέθη διάφορων κυττάρων, μικροοργανισμών και μορίων παρουσιάζονται στην Εικόνα 3.2.



Εικόνα 3.2: Τυπικά μεγέθη κυττάρων, μικροοργανισμών και μορίων.

Ένα τυπικό ζωικό ευκαρυωτικό κύτταρο (Εικόνα 3.3) περιλαμβάνει τον *πυρήνα* (*nucleus*), ο οποίος διαχωρίζεται από το υπόλοιπο του κυττάρου με μια διπλή μεμβρά-

νη, τον *πυρηνικό φάκελο* (*nuclear envelope*). Ο πυρήνας περιέχει το γενετικό υλικό. Μέσα στα ευκαρυωτικά κύτταρα υπάρχουν εσωτερικά διαμερίσματα, που περικλείονται από μεμβράνες, τα οποία καλούνται *οργανίδια* (*organelles*), όπως είναι τα *λυσοσώματα* (*lysosomes*), η *συσκευή Golgi* (*Golgi apparatus*), τα *μιτοχόνδρια* (*mitochondria*), κ.α. Τα μιτοχόνδρια βρίσκονται σε όλα σχεδόν τα ευκαρυωτικά κύτταρα και εξειδικεύονται στην παραγωγή ενέργειας (κυτταρική αναπνοή). Η περιοχή του κυττάρου έξω από τον πυρήνα μαζί με τα οργανίδια καλείται *κυτταρόπλασμα* (*cytoplasm*). Τα μιτοχόνδρια περιέχουν δικό τους γενετικό υλικό (DNA). Μάλιστα, υπάρχει μια θεωρία που πρεσβεύει ότι τα μιτοχόνδρια είναι απόγονοι προκαρυωτικών οργανισμών που ζουν μέσα στα ευκαρυωτικά κύτταρα.

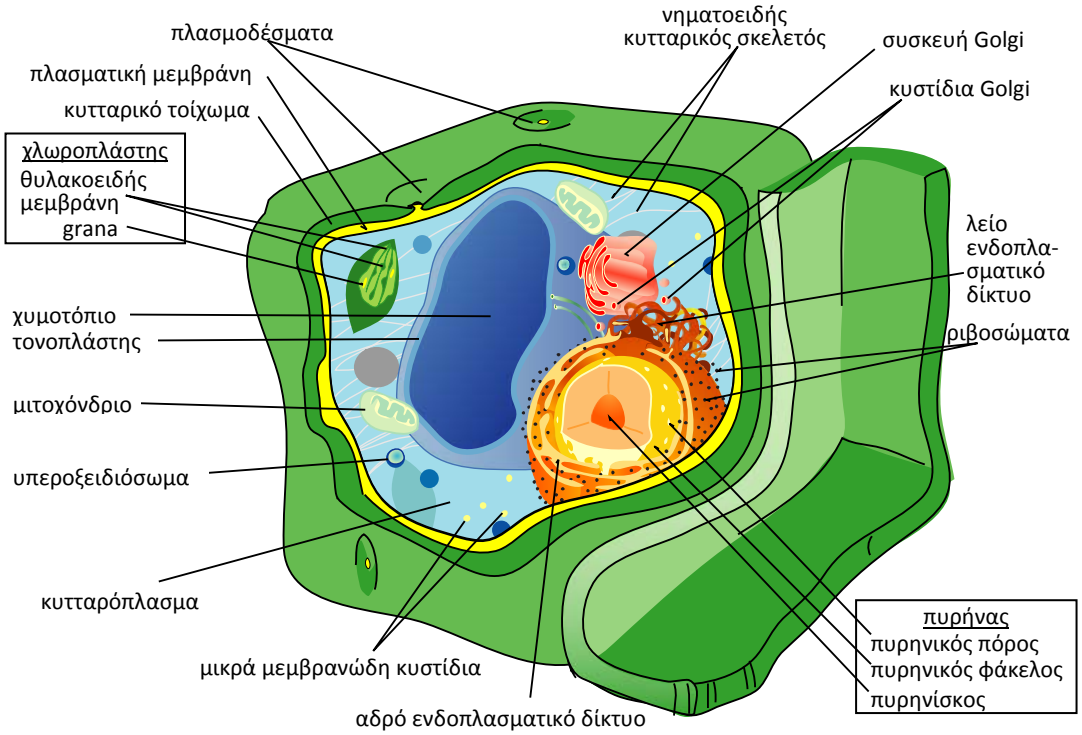


Εικόνα 3.3: Τυπικό ζωικό κύτταρο.

Τα φυτικά κύτταρα (Εικόνα 3.4) διαφέρουν σε πολλά χαρακτηριστικά από τα ζωικά, τα κυριότερα από τα οποία είναι τα εξής [Καψάλης *κ.α.*, 2003]:

- Τα φυτικά κύτταρα περιβάλλονται από ένα επιπλέον *κυτταρικό τοίχωμα* (*cell wall*), που τα προφυλάσσει από περιβαλλοντικές επιδράσεις αλλά και από μεταβολές στο σχήμα τους.
- Τα φυτικά κύτταρα έχουν κάποια επιπλέον οργανίδια, όπως είναι οι *χλωροπλάστες* (*chloroplasts*), που περιέχουν τη *χλωροφύλλη* (*chlorophyll*), η οποία είναι απαραίτητη για τη δέσμευση της ηλιακής ενέργειας κατά τη φωτοσύνθεση. Οι χλωροπλάστες όπως και τα μιτοχόνδρια περιέχουν το δικό τους DNA.

- Τα περισσότερα φυτικά κύτταρα έχουν ένα μεγάλο κεντρικό *χυμοτόπιο* (*vacuole*) και αρκετά μικρότερα γύρω του, τα οποία χρησιμεύουν στη διακίνηση και αποθήκευση θρεπτικών υλικών, νερού, αλάτων, καθώς και άχρηστων ουσιών.
- Ορισμένα οργανίδια, όπως το κεντροσωμάτιο και τα λυσοσώματα δεν υπάρχουν στα κύτταρα των ανώτερων φυτών.
- Τα φυτικά κύτταρα συνήθως έχουν μια αρκετά κανονική μορφή, σε αντίθεση με τα ζωικά κύτταρα που συχνά έχουν αρκετά ακανόνιστα σχήματα.

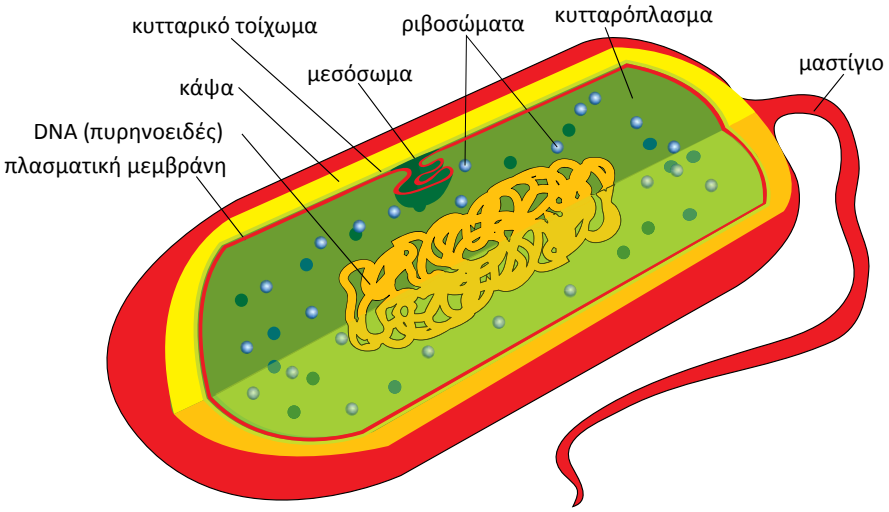


Εικόνα 3.4: Τυπικό φυτικό κύτταρο.

Τα βακτήρια έχουν αρκετά απλή κυτταρική δομή, χωρίς πυρήνα και σύνθετα οργανίδια. Το DNA τους οργανώνεται σε πυρηνοειδές, χωρίς πυρηνική μεμβράνη και χωρίς σχηματισμό χρωμοσωμάτων. Δε διαθέτουν μιτοχόνδρια, χλωροπλάστες και άλλα οργανίδια που υπάρχουν στα ευκαρυωτικά κύτταρα (Εικόνα 3.5).

Ένα ουσιαστικό χαρακτηριστικό γνώρισμα των περισσότερων ζωντανών κυττάρων είναι η ικανότητά τους να αναπτύσσονται σε ένα κατάλληλο περιβάλλον και να υπόκεινται στην κυτταρική διαίρεση. Η ανάπτυξη ενός κυττάρου και η επακόλουθη διαίρεσή του αποτελούν τον *κυτταρικό κύκλο* (*cell cycle*). Ωστόσο, δεν ακολουθούν την πορεία αυτή όλα τα κύτταρα. Για παράδειγμα, τα νευρικά κύτταρα υποβάλλονται μόνο σε μια αρχική φάση ανάπτυξης. Οι προκαρυώτες, ιδιαίτερα τα βακτήρια, είναι ικανοί να αναπτύσσονται και να διαιρούνται πολύ γρήγορα. Οι πολυκύτταροι οργανισμοί αρχίζουν τη ζωή ως ένα κύτταρο, συνήθως αποτέλεσμα της ένωσης ενός αρσενικού και ενός θηλυκού φυλετικού κυττάρου (γαμέτες). Το μοναδικό κύτταρο πρέπει να αναπτυχθεί,

να διαιρεθεί και να διαφοροποιηθεί στους διαφορετικούς τύπους κυττάρων για την παραγωγή των ιστών και των οργάνων. Η κυτταρική διαίρεση και η *διαφοροποίηση* (*differentiation*) πρέπει να ελεγχθούν. Η μη ελεγχόμενη κυτταρική διαίρεση μπορεί να οδηγήσει σε σοβαρή διαταραχή ενός οργανισμού, όπως για παράδειγμα συμβαίνει στην περίπτωση των καρκινικών κυττάρων, τα οποία αυξάνονται και συσσωρεύονται διαμορφώνοντας όγκους.



Εικόνα 3.5: Τυπικό βακτηριακό κύτταρο.

3.2 Βιομόρια

Όλα τα μόρια που σχετίζονται με το φαινόμενο της ζωής καλούνται *βιομόρια* (*biomolecules*). Τα βιομόρια μπορούν να διακριθούν σε *μικρά μόρια (small molecules)*, και *μακρομόρια* (*macromolecules*). Όπως αναφέρθηκε στην αρχή του κεφαλαίου, οι πρωτεΐνες, και τα νουκλεϊκά οξέα είναι μακρομόρια. Στην κατηγορία των μακρομορίων, επίσης, ανήκουν και οι *πολυσακχαρίτες* (*polysaccharides*), όπως για παράδειγμα η κυτταρίνη, το άμυλο και το γλυκογόνο που αποτελούν μεγάλα πολυμερή της γλυκόζης και τα *λιπίδια* (*lipids*), τα οποία αποτελούν τα βασικά δομικά υλικά των βιολογικών μεμβρανών.

3.2.1 Μικρά Μόρια

Τα μικρά μόρια μπορεί να είναι οι δομικές μονάδες των μακρομορίων, να έχουν ανεξάρτητους ρόλους, όπως η μετάδοση σημάτων ή να αποτελούν την πηγή ενέργειας ή το υλικό ενός κυττάρου. Μερικά σημαντικά παραδείγματα εκτός από το νερό είναι οι *μονοσακχαρίτες* (*monosaccharides*), τα *λιπαρά οξέα* (*fatty acids*), τα *αμινοξέα* (*amino acids*) και τα *νουκλεοτίδια* (*nucleotides*).

Αμινοξέα

Υπάρχουν 20 διαφορετικά μόρια αμινοξέων, τα οποία είναι οι δομικές μονάδες των πρωτεϊνών. Για την ακρίβεια, υπάρχουν 19 αμινοξέα και η *προλίνη*, που έχει μια ελαφρώς διαφορετική δομή (περιέχει την ιμινομάδα NH και όχι την αμινομάδα NH_2) και

επομένως είναι πιο ορθό να χαρακτηριστεί ως *ιμινοξύ* (*imino acid*). Ωστόσο, τις περισσότερες φορές για να περιγράψει και την προλίνη. Ο Πίνακας 3.2 παρουσιάζει τα 20 αμινοξέα.

Πίνακας 3.2: Τα 20 αμινοξέα.

Ονομασία	Συμβολισμοί		Ονομασία	Συμβολισμοί	
Αλανίνη (Alanine)	ALA	A	Ιστιδίνη (Histidine)	HIS	H
Αργινίνη (Arginine)	ARG	R	Κυστεΐνη (Cysteine)	CYS	C
Ασπαραγίνη (Asparagine)	ASN	N	Λευκίνη (Leucine)	LEU	L
Ασπαρτικό οξύ (Aspartic acid)	ASP	D	Λυσίνη (Lysine)	LYS	K
Βαλίνη (Valine)	VAL	V	Μεθειονίνη (Methionine)	MET	M
Γλουταμικό οξύ (Glutamic acid)	GLU	E	Προλίνη (Proline)	PRO	P
Γλουταμίνη (Glutamine)	GLN	Q	Σερίνη (Serine)	SER	S
Γλυκίνη (Glycine)	GLY	G	Τρυπτοφάνη (Tryptophan)	TRP	W
Θρεονίνη (Threonine)	THR	T	Τυροσίνη (Tyrosine)	TYR	Y
Ισολευκίνη (Isoleucine)	ILE	I	Φαινυλαλανίνη (Phenylalanine)	PHE	F

Νουκλεοτίδια

Τα νουκλεοτίδια αποτελούνται από μια πεντόζη (σάκχαρο με πέντε άτομα άνθρακα), η οποία ενώνεται με μια φωσφορική ομάδα και μια αζωτούχο βάση. Η πεντόζη των νουκλεοτιδίων του DNA είναι η *δεοξυριβόζη* (*deoxyribose*), ενώ του RNA η *ριβόζη* (*ribose*). Η διαφορετικότητα των νουκλεοτιδίων μέσα στα μόρια του DNA και του RNA καθορίζεται από τις πέντε διαφορετικές αζωτούχες βάσεις. Οι αζωτούχες βάσεις είναι η *αδενίνη* (*adenine*) και η *γουανίνη* (*guanine*), που ανήκουν στις *πουρίνες* (*purines*), η *κυτοσίνη* (*cytosine*), η *θυμίνη* (*thymine*) και *ουρακίλη* (*uracil*), που ανήκουν στις *πυριμιδίνες* (*pyrimidines*). Οι παραπάνω βάσεις συμβολίζονται με τα αρχικά τους γράμματα *A*, *G*, *C*, *T* και *U* αντίστοιχα.

3.2.2 Πρωτεΐνες

Οι πρωτεΐνες είναι τα κύρια δομικά και λειτουργικά μόρια του κυττάρου, που καταλαμβάνουν σχεδόν το 20% της μάζας ενός ευκαρυωτικού κυττάρου, αποτελώντας τη μεγαλύτερη συμβολή μετά από το νερό (70%). Οι βασικότερες κατηγορίες πρωτεϊνών βάσει της λειτουργίας τους είναι οι εξής:

- Δομικές πρωτεΐνες, οι οποίες μπορούν να θεωρηθούν ως οι βασικές δομικές μονάδες του οργανισμού. Ένα παράδειγμα είναι το κολλαγόνο, το οποίο αποτελεί τη σημαντικότερη δομική πρωτεΐνη του συνδετικού ιστού και των οστών.
- Ένζυμα, τα οποία καταλύουν ένα πλήθος βιοχημικών αντιδράσεων. Το σύνολο των αντιδράσεων αυτών καλείται *μεταβολισμός* (*metabolism*). Συνήθως τα ένζυ-

μα είναι πολύ εξειδικευμένα και καταλύουν μόνο έναν τύπο αντίδρασης, ωστόσο το ίδιο ένζυμο μπορεί να συμβάλλει σε περισσότερες από μια *μεταβολικές οδούς* (*metabolic pathways*).

- Οι πρωτεΐνες της μεμβράνης είναι βασικές στη συντήρηση του κυτταρικού περιβάλλοντος, ρυθμίζοντας τον όγκο του κυττάρου, την εξαγωγή και τη συγκέντρωση των μικρών μορίων από το εξωκυττάριο περιβάλλον και την παραγωγή των βαδμιδώσεων ιόντων, που είναι σημαντικές για τη λειτουργία των μυϊκών και νευρικών κυττάρων. Ένα παράδειγμα είναι η αντλία Na^+/K^+, η οποία διατηρεί τις ηλεκτροχημικές βαθμιδώσεις των ιόντων Na^+ και K^+ εκατέρωθεν της κυτταρικής μεμβράνης.

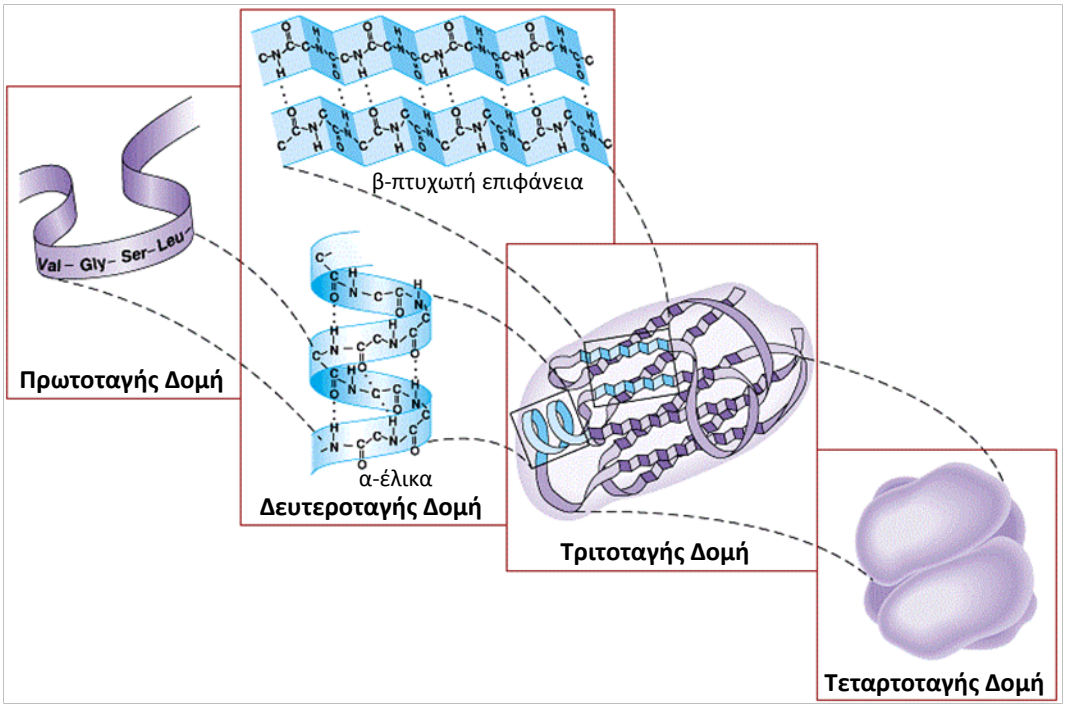
Επίπεδα Διαμόρφωσης των Πρωτεϊνών

Οι πρωτεΐνες λμβάνουν συγκεκριμένη δομή στο χώρο. Αυτή η δομή οφείλεται σε διάφορες χημικές αλληλεπιδράσεις των δομικών τους μερών. Υπάρχουν τέσσερα επίπεδα *διαμόρφωσης* (*conformation*) των πρωτεϊνών, τα οποία παρουσιάζονται παρακάτω:

- Οι πρωτεΐνες είναι αλυσίδες των 20 διαφορετικών αμινοξέων, τα οποία μπορούν να ενωθούν σε οποιαδήποτε γραμμική διάταξη. Οι αλυσίδες αυτές καλούνται *πολυπεπτίδια* (*polypeptides*) ή *πολυπεπτιδικές αλυσίδες* (*polypeptide chains*). Αυτή η αλληλουχία των αμινοξέων στην πολυπεπτιδική αλυσίδα αποτελεί την *πρωτοταγή δομή* (*primary structure*) και μπορεί να αναπαρασταθεί ως συμβολοσειρά από το συνδυασμό 20 διαφορετικών συμβόλων (Πίνακας 3.2). Τα δύο άκρα μιας πολυπεπτιδικής αλυσίδας είναι διαφορετικά από χημική άποψη και καλούνται *αμινοτελικό άκρο* (*N-τελικό άκρο*) και *καρβοξυτελικό άκρο* (*C-τελικό άκρο*).
- Η *δευτερο*τοταγής *δομή* (*secondary structure*) οφείλεται στην αναδίπλωση της πολυπεπτιδικής αλυσίδας στο χώρο. Υπάρχουν δύο δομές που εμφανίζονται συχνά μέσα στις αναδιπλωμένες αλυσίδες, οι *α-έλικες* (*α-helices*) και οι *β-πτυχωτές επιφάνειες* (*β-sheets*). Μια πρωτεΐνη μπορεί να περιέχει αρκετές α-έλικες και β-πτυχωτές επιφάνειες. Οι δομές αυτές ενώνονται μεταξύ τους με πιο ακανόνιστες δομές, που ονομάζονται *βρόχοι* (*loops*).
- Ως αποτέλεσμα της αναδίπλωσης, τα μέρη του πρωτεϊνικού μορίου πλησιάζουν μεταξύ τους και οι διάφορες ελκτικές ή απωστικές δυνάμεις (δεσμοί υδρογόνου, δισουλφιδικοί δεσμοί, έλξεις μεταξύ θετικών και αρνητικών φορτίων, υδρόφοβες και υδρόφιλες δυνάμεις) μεταξύ αυτών των μερών αναγκάζουν το μόριο να αποκτήσει μια σχετικά σταθερή τρισδιάστατη δομή. Αυτή είναι η *τριτοταγής δομή* (*tertiary structure*) της πρωτεΐνης. Σε πολλές περιπτώσεις αυτή η τρισδιάστατη δομή είναι αρκετά συμπαγής και μπορεί να είναι η τελική διαμόρφωση της πρωτεΐνης.

– Η *τεταρτοτοταγής δομή* (*quaternary structure*) είναι η τελική τρισδιάστατη μορφή που λαμβάνουν οι πρωτεΐνες που αποτελείται από περισσότερες της μίας πολυπεπτιδικές αλυσίδες.

Τα τέσσερα επίπεδα διαμόρφωσης καθορίζονται τόσο από την πρωτοταγή δομή όσο και από το φυσικοχημικό περιβάλλον στο οποίο βρίσκεται το μόριο της πρωτεΐνης. Η πρόβλεψη της δομής της πρωτεΐνης από την αλληλουχία των αμινοξέων είναι ένα από τα σημαντικότερα προβλήματα που αντιμετωπίζει η βιοπληροφορική. Στην Εικόνα 3.6 παρουσιάζονται τα τέσσερα επίπεδα της διαμόρφωσης των πρωτεϊνών.



Εικόνα 3.6: Τα τέσσερα επίπεδα διαμόρφωσης των πρωτεϊνών.

3.2.3 DNA

Το *δεοξυριβονουκλεϊκό οξύ* (*deoxyribonucleic acid*), ή συντομότερα *DNA* ανήκει στην κατηγορία των νουκλεϊκών οξέων και είναι ο βασικός φορέας πληροφορίας σε ένα κύτταρο. Το DNA μπορεί να είναι μονόκλωνο ή δίκλωνο. Ένα μονόκλωνο μόριο DNA, αλλιώς *πολυνουκλεοτίδιο* (*polynucleotide*), είναι μια αλυσίδα (*πολυνουκλεοτιδική αλυσίδα*) από νουκλεοτίδια. Τα νουκλεοτίδια του DNA περιέχουν το σάκχαρο δεοξυριβόζη και γι' αυτό καλούνται *δεοξυριβονουκλεοτίδια* (*deoxyribonuceotides*). Υπάρχουν τέσσερα διαφορετικά δεοξυριβονουκλεοτίδια, που καθορίζονται από τις βάσεις αδενίνη, γουανίνη, κυτοσίνη και θυμίνη. Ένα νουκλεοτίδιο συνδέεται με το επόμενο στην αλυσίδα νουκλεοτίδιο με ένα φωσφοδιεστερικό δεσμό.

Ένα πολυνουκλεοτίδιο μπορεί να αναπαρασταθεί ως συμβολοσειρά από το συνδυασμό των τεσσάρων διαφορετικών βάσεων. Για παράδειγμα:

A G T C C G T

Τα δύο άκρα ενός τέτοιου μορίου είναι διαφορετικά από χημική άποψη και καλούνται *5΄ άκρο* και *3΄ άκρο*. Άρα, ένα πολυνουκλεοτίδιο είναι κατευθυνόμενο και μάλιστα έχει προσανατολισμό 5΄→3΄:

5΄ A → G → T → C → C → G → T 3΄

Το παραπάνω πολυνουκλεοτίδιο είναι διαφορετικό από το αντίστοιχο με την αντίστροφη σειρά:

T G C C T G A

Ο προσανατολισμός του, οποίου είναι:

5΄ T → G → C → C → T → G → A 3΄

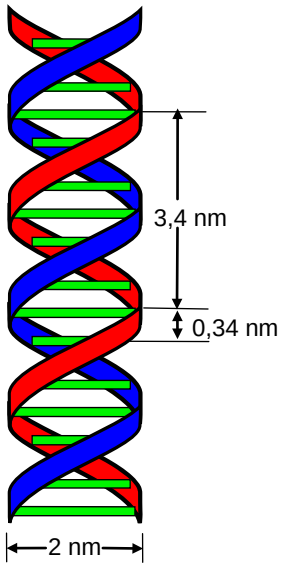
Κατά σύμβαση το DNA γράφεται με το 5΄ άκρο αριστερά και το 3΄ άκρο δεξιά και μάλιστα στην περίπτωση του δίκλωνου DNA η *κωδική αλυσίδα* γράφεται από πάνω. Δύο αλυσίδες καλούνται *συμπληρωματικές* (*complementary*), αν η μία μπορεί να προκύψει από την άλλη με αμοιβαία ανταλλαγή της A με την T και της G με τη C, και ταυτόχρονη αλλαγή του προσανατολισμού του μορίου. Επίσης δύο συμπληρωματικές αλυσίδες καλούνται *αντιπαράλληλες* (*antiparallel*), λόγω των αντίθετων και συνάμα παράλληλων κατευθύνσεών τους. Για παράδειγμα, οι δύο παρακάτω πολυνουκλεοτιδικές αλυσίδες είναι συμπληρωματικές και αντιπαράλληλες:

5΄ A T T G G A T C 3΄

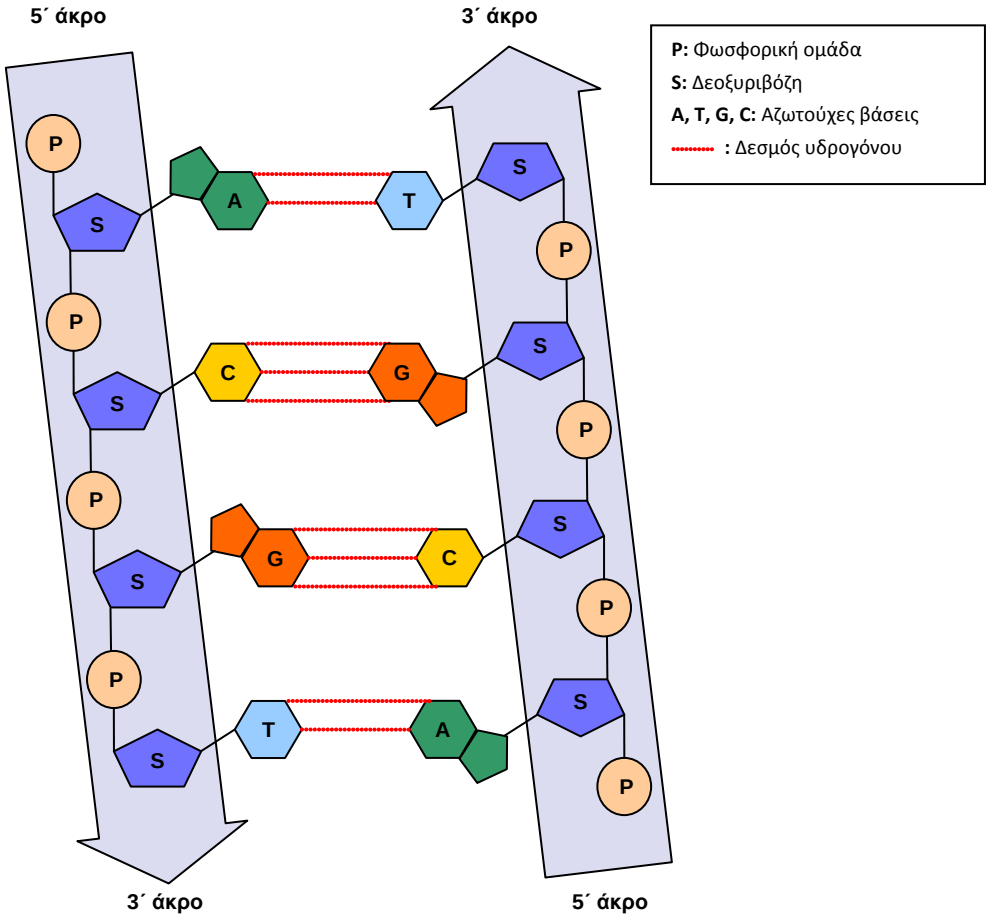
3΄ T A A C C T A G 5΄

Κάποια ζεύγη νουκλεοτιδίων μπορούν να σχηματίσουν μεταξύ τους ασθενείς δεσμούς. Συγκεκριμένα, δύο *δεσμοί υδρογόνου* (*hydrogen bonds*) μπορούν να σχηματιστούν μεταξύ A και T και τρεις δεσμοί υδρογόνου μεταξύ G και C. Παρότι αυτές οι δυνάμεις είναι ασθενείς, όταν δύο μεγάλες συμπληρωματικές αλυσίδες συναντώνται συνενώνονται, λόγω της άθροισης πολλών ασθενών δυνάμεων. Τα ζεύγη A–T και G–C καλούνται *ζεύγη βάσεων* (*base pairs*). Το μήκος ενός μορίου DNA συνήθως μετράται σε ζεύγη βάσεων (*bp*) ή σε νουκλεοτίδια (*nt*). Δύο συμπληρωματικές πολυνουκλεοτιδικές αλυσίδες σχηματίζουν μια σταθερή δομή, η οποία μοιάζει με έλικα και είναι γνωστή ως η *διπλή έλικα του DNA* (*DNA double helix*). Το μοντέλο της διπλής έλικας του DNA διατυπώθηκε το 1953 στο Cambridge από τους Francis Crick και James Watson με τη συμβολή των Maurice Wilkins και Rosalind Franklin. Αξίζει να σημειωθεί ότι η συμπληρωματικότητα των δύο αλυσίδων σημαίνει ότι μπορεί να προσδιοριστεί επακριβώς η μία από τη γνώση της άλλης. Στην Εικόνα 3.7 παρουσιάζεται η διπλή έλικα του DNA, καθώς και οι σχετικές αποστάσεις μεταξύ μιας στροφής της αλυσίδας, μεταξύ δύο

διαδοχικών βάσεων και μεταξύ των δύο αλυσίδων. Η Εικόνα 3.8 παρουσιάζει αναλυτικότερα ένα τμήμα της διπλής έλικας.



Εικόνα 3.7: Η διπλή έλικα του μορίου του DNA.



Εικόνα 3.8: Αναλυτικότερη άποψη του δίκλωνου μορίου του DNA.

Αντιγραφή του DNA

Η συμπληρωματικότητα των δύο κλώνων του DNA συμβάλλει στη δυνατότητα δημιουργίας αντιγράφων του DNA, με τη διαδικασία της *αντιγραφής* (*replication*). Αρχικά προτάθηκαν τρία μοντέλα που αφορούν το μηχανισμό της αντιγραφής του DNA:

- Ο *συντηρητικός* (*conservative*) διπλασιασμός του DNA, σύμφωνα με τον οποίο το μητρικό μόριο διατηρεί την αρχική του μορφή και το θυγατρικό αποτελείται από δύο νέες αλυσίδες.
- Ο *ημισυντηρητικός* (*semiconservative*) διπλασιασμός του DNA, σύμφωνα με τον οποίο τα δύο θυγατρικά μόρια που προκύπτουν είναι πανομοιότυπα με το μητρικό και καθένα αποτελείται από μια παλιά και μια καινούρια αλυσίδα.
- Ο *διασπαρτικός* (*dispersive*) διπλασιασμός του DNA, σύμφωνα με τον οποίο τα δύο θυγατρικά μόρια περιέχουν τυχαίες περιοχές και των παλιών και των νέων αλυσίδων.

Οι Meselson & Stahl [1958] απέδειξαν ότι ο μηχανισμός διπλασιασμού του DNA είναι ημισυντηρητικός, αποκλείοντας το συντηρητικό και το διασπαρτικό. Κατά τη διαδικασία της αντιγραφής ξετυλίγονται οι έλικες και σπάζουν οι δεσμοί υδρογόνου μεταξύ των συμπληρωματικών βάσεων με τη βοήθεια ειδικών ενζύμων που ονομάζονται *DNA ελικάσες* (*DNA helicases*). Όταν ανοίξει η διπλή έλικα, συντίθενται νέες συμπληρωματικές αλυσίδες χρησιμοποιώντας ως καλούπι τους δύο αρχικούς κλώνους με τη συμμετοχή ενζύμων που καλούνται *DNA πολυμεράσες* (*DNA polymerases*). Η σύνθεση των νέων αλυσίδων έχει την κατεύθυνση 5′→3′. Οι νέες αλυσίδες δε δημιουργούνται από ένα μόνο σημείο εκκίνησης, αλλά πολλά τμήματα σε διαφορετικά σημεία της ίδιας αλυσίδας συντίθενται ταυτόχρονα. Ειδικά ένζυμα, οι *DNA δεσμάσες* (*DNA ligases*), αναλαμβάνουν να ενώσουν τα επιμέρους τμήματα της αλυσίδας σε ένα. Έτσι το αρχικό μόριο DNA διπλασιάζεται.

3.2.4 RNA

Το *ριβονουκλεϊκό οξύ* (*ribonucleic acid*), ή συντομότερα *RNA* ανήκει, όπως και το DNA, στην κατηγορία των νουκλεϊκών οξέων και αποτελείται από νουκλεοτίδια. Ωστόσο, τα νουκλεοτίδια που συνθέτουν το RNA, σε αντίθεση με αυτά του DNA, αντί της πεντόζης δεοξυριβόζης περιέχουν την πεντόζη *ριβόζη* (*ribose*) και γι' αυτό καλούνται *ριβονουκλεοτίδια* (*ribonucleotides*). Ακόμη, στο μόριο του RNA αντί της αζωτούχου βάσης θυμίνης (T) υπάρχει η ουρακίλη (U). Μια ακόμη διαφορά με το DNA είναι ότι το RNA δε σχηματίζει διπλή έλικα, αλλά είναι μονόκλωνο. Μπορεί όμως να έχει μια πολύπλοκη δομή στο χώρο εξαιτίας των συμπληρωματικών τμημάτων της ίδιας αλυσίδας. Το RNA μπορεί να προσδεθεί σε ένα συμπληρωματικό μονόκλωνο μόριο DNA, παρά την αντικατάσταση της T από την U.

Το RNA έχει πολλές λειτουργίες μέσα σε ένα κύτταρο, οι σημαντικότερες από τις οποίες σχετίζονται με τη διαδικασία της πρωτεϊνοσύνθεσης. Επίσης, το RNA αποτελεί το γενετικό υλικό κάποιων ιών (RNA ιοί), όπως για παράδειγμα του ιού HIV.

Παρακάτω περιγράφονται τέσσερα βασικά είδη RNA που σχετίζονται με τη διαδικασία της πρωτεϊνοσύνθεσης, η οποία θα περιγραφεί σε επόμενη ενότητα (3.5):

- *Αγγελιαφόρο RNA* (*messenger RNA – mRNA*), το οποίο μεταφέρει την πληροφορία του DNA για τη σύνθεση μιας πολυπεπτιδικής αλυσίδας.
- *Ριβοσωμικό RNA* (*ribosomal RNA – rRNA*), το οποίο συνδέεται με πρωτεΐνες και σχηματίζει το *ριβόσωμα* (*ribosome*), που είναι απαραίτητο για τη διαδικασία της πρωτεϊνοσύνθεσης.
- *Μεταφορικό RNA* (*transfer RNA – tRNA*), το οποίο μεταφέρει τα αμινοξέα στην περιοχή της πρωτεϊνοσύνθεσης.
- *Μικρό πυρηνικό RNA* (*small nuclear RNA – snRNA*), το οποίο συνδέεται με πρωτεΐνες και σχηματίζει ριβονουκλεοπρωτεϊνικά σωματίδια, τα οποία καταλύουν τη διαδικασία του ματίσματος του mRNA, που θα περιγραφεί στο Εδάφιο 3.5.2.

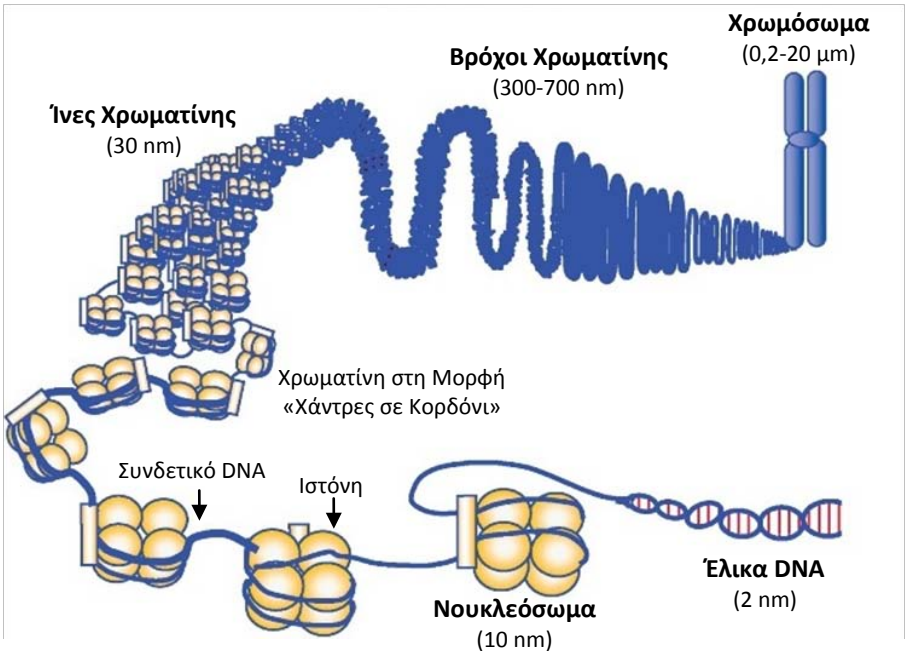
3.3 Γονίδια και Γονιδίωμα

Το γενετικό υλικό των προκαρυωτικών κυττάρων είναι ένα δίκλωνο κυκλικό μόριο DNA. Αντίθετα με τα προκαρυωτικά, τα ευκαρυωτικά κύτταρα περιέχουν πολλά γραμμικά μόρια DNA, ο αριθμός και το μήκος των οποίων είναι χαρακτηριστικά για τα διάφορα είδη των οργανισμών. Τα μόρια αυτά καλούνται *χρωμοσώματα* (*chromosomes*). Οργανισμοί των οποίων κάθε χρωμόσωμα περιέχεται σε πολλαπλά (> 2) αντίγραφα καλούνται *πολλαπλοειδείς* (*polyploids*), π.χ. τριπλοειδείς, τετραπλοειδείς, κλπ.. Οργανισμοί που περιέχουν σε δύο αντίγραφα τα χρωμοσώματά τους καλούνται *διπλοειδείς* (*diploids*), ενώ οργανισμοί με ένα αντίγραφο κάθε χρωμοσώματος, καλούνται *απλοειδείς* (*haploids*). Ο άνθρωπος, όπως και οι περισσότεροι οργανισμοί, είναι διπλοειδής οργανισμός και έχει 23 ζεύγη χρωμοσωμάτων. Οι διπλοειδείς οργανισμοί κληρονομούν το ένα σύνολο χρομοσωμάτων από τη μητέρα τους και το άλλο από τον πατέρα τους. Τα δύο αντίγραφα του ίδιου χρωμοσώματος καλούνται *ομόλογα χρωμοσώματα* (*homologous chromosomes*). Οι προκαρυωτικοί οργανισμοί είναι απλοειδείς. Τα χρωμοσώματα που είναι μορφολογικά ίδια στο αρσενικό και το θηλυκό άτομο καλούνται *αυτοσωμικά* (*autosomes*). Τα χρωμοσώματα που καθορίζουν το φύλο ενός οργανισμού καλούνται *φυλετικά χρωμοσώματα* (*sex chromosomes*). Για παράδειγμα, στον άνθρωπο υπάρχουν 22 ζεύγη αυτοσωμικών χρωμοσωμάτων και δύο χρωμοσώματα τα X και Y, που καθορίζουν το θηλυκό και το αρσενικό άτομο αντίστοιχα. Τα θηλυκά άτομα περιλαμβάνουν το ζεύγος XX, ενώ τα αρσενικά το ζεύγος XY. Στην Εικόνα 3.9 παρουσιάζεται η δομή του χρωμοσώματος σε διάφορα επίπεδα μεγέθυνσης.

Στα ευκαρυωτικά κύτταρα το γενετικό υλικό κατανέμεται στον πυρήνα, στα μιτοχόνδρια και στους χλωροπλάστες. Το γενετικό υλικό σε όλα τα κύτταρα ενός οργανισμού είναι το ίδιο με ελάχιστες ειδικές εξαιρέσεις. Το σύνολο του γενετικού υλικού ενός οργανισμού καλείται *γονιδίωμα* ή *γένωμα* (*genome*). Το *γονίδιο* (*gene*) είναι ένα συνεχές τμήμα μορίου χρωμοσωμικού DNA στο οποίο περιέχονται πληροφορίες που καθορίζουν τη σύνθεση ενός συγκεκριμένου τύπου πρωτεΐνης ή ενός μικρού αριθμού διαφορετικών πρωτεϊνών. Υπάρχουν, επίσης, γονίδια που περιέχουν πληροφορίες μόνο για τη σύνθεση μορίων RNA. Το γονίδιο είναι η βασική μονάδα κληρονομικότητας. Ο Πίνακας 3.3 παρουσιάζει τα μεγέθη των γονιδιωμάτων και τον αριθμό των γονιδίων κάποιων από τους οργανισμούς που έχουν μελετηθεί.

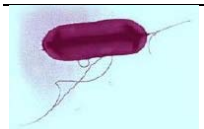
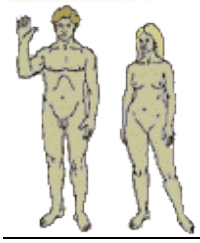
Σημαντικές στη μοριακή βιολογία και τη γενετική (genetics) είναι οι έννοιες του *γονότυπου* (*genotype*) και του *φαινότυπου* (*phenotype*). Ο γονότυπος είναι η συγκεκριμένη γενετική σύσταση ενός ατόμου, ενώ ο φαινότυπος είναι, είτε η συνολική φυσική του εμφάνιση, είτε η συγκεκριμένη εκδήλωση ενός χαρακτηριστικού (π.χ. χρώμα ματιών). Συνήθως η έννοια του γονότυπου χρησιμοποιείται για να περιγράψει ένα συγκεκριμένο γονίδιο που συγκεντρώνει το ενδιαφέρον ή έναν αριθμό γονιδίων. Στην περίπτωση των διπλοειδών οργανισμών ο γονότυπος αναφέρεται στα γονίδια που βρίσκονται στην ίδια θέση των ομόλογων χρωμοσωμάτων και ελέγχουν την ίδια ιδιότητα, με τον ίδιο ή διαφορετικό ενδεχομένως τρόπο (*αλληλόμορφα γονίδια – alleles*). Η σχέση μεταξύ γονότυπου και φαινότυπου δεν είναι άμεση. Κάποια γονίδια μπορεί να εκφράσουν ένα συγκεκριμένο φαινότυπο αν βρεθούν στις κατάλληλες περιβαλλοντικές συνθήκες. Συνεπώς, η σχέση γονότυπου και φαινοτύπου μπορεί να περιγραφεί ως εξής:

γονότυπος + περιβάλλον → φαινότυπος



Εικόνα 3.9: Η δομή του χρωμοσώματος σε διάφορα επίπεδα μεγέθυνσης.

Πίνακας 3.3: Μεγέθη των γονιδιωμάτων και του αριθμού των γονιδίων οργανισμών.

	Οργανισμός	Μήκος DNA (Mbp)	Αριθμός γονιδίων
	Escherichia coli K12 (βακτήριο)	4,6	4.289
	Caenorhabditis elegans (βακτήριο)	97	20.443
	Saccharomyces cerevisiae (μύκητας)	12	6.802
	Drosophila melanogaster (έντομο)	120	13.871
	Homo sapiens (άνθρωπος)	3200	20.000 – 25.000

3.4 Μεταλλάξεις και Πολυμορφισμοί

Η αλλαγή στις βάσεις ενός γονιδίου ή η μεταβολή ενός χρωμοσώματος, καλείται *μετάλλαξη* (*mutation*) ή *πολυμορφισμός* (*polymorphism*). Οι πολυμορφισμοί είναι συνηθισμένες διαφοροποιήσεις στο γενετικό υλικό, που συμβαίνουν τουλάχιστο στο 1% του πληθυσμού. Οι μεταλλάξεις είναι λιγότερο συνηθισμένες και συμβαίνουν σε ένα ποσοστό του πληθυσμού που δεν ξεπερνά το 1%. Μάλιστα, οι δύο έννοιες δεν είναι αυστηρά διακεκριμένες. Μία τέτοια διαφοροποίηση σε κάποιο σημείο του πλανήτη μπορεί να θεωρείται πολυμορφισμός, ενώ σε κάποιο άλλο μετάλλαξη. Οι περισσότερες από τις ποικιλομορφίες στο DNA είναι ουδέτερες, δηλαδή δεν έχουν καμιά συνέπεια στο άτομο. Πολλές φορές όμως είναι επιβλαβείς. Αλλαγές στην αλληλουχία των γονιδίων μπορούν να οδηγήσουν σε πρωτεΐνες που δε λειτουργούν φυσιολογικά ή δε λειτουργούν καθόλου. Κάποιες από αυτές τις αλλαγές μπορούν να συμβάλλουν στην εμφάνιση ασθενειών ή να επηρεάσουν τον τρόπο αντίδρασης του οργανισμού στις συνθήκες του περιβάλλοντος. Οι μεταλλάξεις μπορούν να συμβούν κατά τη σύλληψη του οργανισμού ή κατά τη διάρκεια της ζωής του. Μπορούν να συμβούν σε οποιοδήποτε σωματικό ή γεννητικό κύτταρο. Ωστόσο, μόνο οι μεταλλάξεις που συμβαίνουν στα γεννητικά κύτταρα μπορούν να μεταφερθούν και στους απογόνους. Οι μεταλλάξεις είναι δυνατό να προκύψουν κατά τη διάρκεια φυσιολογικών κυτταρικών λειτουργιών, όπως η κυτταρική διαίρεση, ή κατά την αντίδραση σε περιβαλλοντικούς παράγοντες όπως τοξικές ουσίες και ακτινοβολία. Μάλιστα, οι οργανισμοί διαθέτουν ένα σύστημα από *επιδιορθωτικά*

ένζυμα (*repair enzymes*), που βρίσκουν και διορθώνουν τα περισσότερα λάθη που συμβαίνουν στην αλληλουχία του DNA. Όμως, καθώς οι οργανισμοί αλλάζουν με την ηλικία, τις ασθένειες και άλλους παράγοντες, τα συστήματα αυτά γίνονται λιγότερο αποτελεσματικά. Οι μεταλλάξεις που δεν έχουν διορθωθεί συσσωρεύονται, έχοντας ως αποτέλεσμα ασθένειες όπως ο καρκίνος.

Με βάση την έκταση και τη μορφή της αλλαγής των νουκλεοτιδικών βάσεων διακρίνονται οι παρακάτω τύποι μεταλλάξεων [Connor & Ferguson Smith, 1997; Αλεπόρου-Μαρίνου *κ.α.*, 2003]:

- Οι *σημειακές μεταλλάξεις* (*point mutations*) ή *μονονουκλεοτιδικοί πολυμορφισμοί* (*single nucleotide polymorphisms – SNPs*), που οφείλονται στην αντικατάσταση μιας μόνο νουκλεοτιδικής βάσης από μια άλλη. Αυτός ο τύπος μεταλλάξεων είναι υπεύθυνος για την ποικιλομορφία που παρουσιάζεται μεταξύ των ειδών και των ατόμων. Μια τέτοια μετάλλαξη μπορεί να είναι ουδέτερη, για παράδειγμα αν η βάση που θα αλλάξει δεν οδηγήσει στην παραγωγή διαφορετικού αμινοξέος, λόγω εκφυλισμού του γενετικού κώδικα (Εδάφιο 3.5.4). Ωστόσο, πολλές φορές συμβαίνει το αντίθετο με αποτέλεσμα να προκαλούνται ασθένειες, όπως για παράδειγμα η *δρεπανοκυτταρική αναιμία* (*sickle cell anemia*).
- Οι *προσθήκες* (*insertions*) ή *διαγραφές* (*deletions*) βάσεων είναι μια άλλη κατηγορία μεταλλάξεων. Αν ο αριθμός των συνεχόμενων βάσεων που προστίθενται είναι πολλαπλάσιος του τρία, τότε υπάρχουν δύο δυνατές περιπτώσεις. Είτε δε θα συμβεί καμιά αλλαγή στην τελική πρωτεΐνη, αν οι πλεονάζουσες τριάδες αφαιρεθούν κατά τη διάρκεια απομάκρυνσης των ιντρονίων, είτε θα προστεθούν στην πολυπεπτιδική αλυσίδα τόσα αμινοξέα όσες είναι και οι τριάδες των πρόσθετων βάσεων. Στην περίπτωση της έλλειψης ενός πολλαπλάσιου του τρία αριθμού βάσεων θα απουσιάζουν από την πολυνουκλεοτιδική αλυσίδα τόσα αμινοξέα όσες είναι και οι τριάδες των βάσεων που λείπουν. Αν ο αριθμός των βάσεων δεν είναι πολλαπλάσιος του τρία, τότε η αλληλουχία των αμινοξέων θα είναι πολύ διαφορετική από την αρχική.
- Οι *χρωμοσωμικές ανωμαλίες* (*chromosomal mutations*) είναι μεγάλες σε έκταση αλλαγές που περιλαμβάνουν τμήματα χρωμοσωμάτων ή ακόμη και ολόκληρα χρωμοσώματα. Οι αλλαγές στον αριθμό των χρωμοσωμάτων καλούνται *αριθμητικές χρωμοσωμικές ανωμαλίες*, ενώ οι αλλαγές στη δομή αποτελούν τις *δομικές χρωμοσωμικές ανωμαλίες*. Οι αλλαγές αυτές έχουν συνήθως ως αποτέλεσμα την τροποποίηση του φαινοτύπου του ατόμου. Ανάλογα με τον τύπο της αλλαγής διακρίνονται διάφορα είδη δομικών χρωμοσωμικών ανωμαλιών:
    - Η *έλλειψη* (*deletion*) οφείλεται στην απουσία ενός τμήματος του χρωμοσώματος.

- Ο *διπλασιασμός* (*duplication*) είναι η επανάληψη ενός χρωμοσωμικού τμήματος στο χρωμόσωμα.
- Η *αναστροφή* (*inversion*) προκαλείται από θραύσεις του χρωμοσώματος σε δύο σημεία και επανένωση του τμήματος ύστερα από αναστροφή. Αποτέλεσμα της αναστροφής είναι η αλλαγή της διάταξης των γονιδίων στο χρωμόσωμα.
- Η *μετατόπιση* (*translocation*) είναι το αποτέλεσμα της θραύσης ενός τμήματος του χρωμοσώματος και στη συνέχεια ένωσής του σε άλλο διαφορετικό χρωμόσωμα. Κατά την αμοιβαία μετατόπιση υπάρχει ανταλλαγή χρωμοσωμικών τμημάτων ανάμεσα σε διαφορετικά χρωμοσώματα.

Αλλαγές στην αλληλουχία των βάσεων μπορούν να παρουσιαστούν και σε περιοχές του DNA που δε μεταγράφονται. Αυτές οι αλλαγές, συνήθως, δεν έχουν αρνητικές επιπτώσεις στον οργανισμό.

3.5 Γονιδιακή Έκφραση

Γονιδιακή έκφραση (*gene expression*) καλείται η διαδικασία κατά την οποία η κωδικοποιημένη πληροφορία ενός γονιδίου μετατρέπεται σε δομές που παρουσιάζονται και λειτουργούν σε ένα κύτταρο. Πιο συγκεκριμένα, εκφρασμένα είναι τα γονίδια που είτε έχουν μεταγραφεί σε mRNA και στη συνέχεια έχουν μεταφραστεί σε πρωτεΐνες, είτε έχουν μεταγραφεί σε άλλα είδη RNA και δεν έχουν μεταφραστεί σε πρωτεΐνες (π.χ. tRNA και rRNA).

3.5.1 Μεταγραφή

Στο στάδιο της *μεταγραφής* (*transcription*) η μια αλυσίδα του μορίου του DNA (του γονιδίου) αντιγράφεται σε ένα συμπληρωματικό τμήμα RNA, που καλείται *πρόδρομο* (*primary*) *mRNA* ή *ετερογενές πυρηνικό* (*heterogeneous nuclear*) *RNA* (*hnRNA*). Η μεταγραφή καταλύεται από το ένζυμο *RNA πολυμεράση* (*RNA polymerase*), το οποίο προσκολλάται στο σημείο του DNA από το οποίο θα αρχίσει η μεταγραφή. Στη συνέχεια, ξετυλίγει τοπικά την έλικα του DNA και τοποθετεί συμπληρωματικά ριβονουκλεοτίδια απέναντι από τα δεοξυριβονουκλεοτίδια της μιας αλυσίδας. Η μεταγραφή σταματά όταν εντοπιστούν λειτουργικά σήματα τερματισμού, τα οποία συνήθως βρίσκονται στο νεοσυντιθέμενο mRNA.

3.5.2 Μάτισμα

Η παραγωγή του ευκαρυωτικού mRNA περιλαμβάνει επιπρόσθετα στάδια μετά την μεταγραφή με μια διαδικασία που αποκαλείται ωρίμανση του mRNA. Το αποτέλεσμα της διαδικασίας είναι το *ώριμο* (*mature*) mRNA. Η διαδικασία της ωρίμανσης περιλαμβάνει το *μάτισμα* ή *συρραφή* του mRNA καθώς και την τροποποίηση των άκρων του, η οποία περιγράφεται στο Εδάφιο 3.5.3.

Μετά από τη μεταγραφή των ευκαρυωτικών γονιδίων, απομακρύνονται από το πρόδρομο mRNA κάποια ενδιάμεσα τμήματα. Τα τμήματα αυτά ονομάζονται *ιντρόνια* (*introns*, δηλαδή *intervening sequences*) και είναι αλληλουχίες που δε μεταφράζονται σε αμινοξέα. Οι αλληλουχίες που παραμένουν, μεταφράζονται σε αμινοξέα και ονομάζονται *εξόνια* (*exons*, δηλαδή *expressed regions*). Η διαδικασία της απομάκρυνσης των ιντρονίων αποδίδεται με τον αγγλικό όρο *splicing*. Δεν υπάρχει κοινά αποδεκτός όρος στα ελληνικά. Μπορεί να χρησιμοποιηθεί ο όρος *μάτισμα* ή ο όρος *συρραφή*. Κάποια γονίδια ευκαρυωτικών οργανισμών είναι γνωστό ότι μπορούν να υποστούν *εναλλακτικό μάτισμα* (*alternative splicing*) διατηρώντας διαφορετικούς συνδυασμούς εξονίων και να οδηγήσουν σε περισσότερα του ενός mRNA.

3.5.3 Προσθήκη Καλύπτρας και Πολυαδενυλίωση

Και τα δύο άκρα του μεταγράφου τροποποιούνται με την προσθήκη επιπλέον νουκλεοτιδίων. Το 5΄ άκρο του RNA τροποποιείται αμέσως μετά την εμφάνισή του με την προσθήκη μίας «καλύπτρας». Η τριφωσφορική ομάδα του αρχικού μεταγράφου αντικαθίσταται από ένα νουκλεοτίδιο που προστίθεται σε αντίθετο προσανατολισμό (3΄→5΄), «σφραγίζοντας» με αυτόν τον τρόπο το άκρο. Το 3΄ άκρο τροποποιείται εξαιτίας της προσθήκης μιας σειράς νουκλεοτιδίων αδενίνης (πολυ(A) ή poly(A)) αμέσως μετά την αποκοπή του. Μόνο μετά την ολοκλήρωση όλων των τροποποιήσεων μπορεί το mRNA να εξαχθεί από τον πυρήνα στο κυτταρόπλασμα. Μόλις το mRNA εισέλθει στο κυτταρόπλασμα, αναγνωρίζεται από τα ριβοσώματα και μεταφράζεται.

Η 3΄ τερματική αλληλουχία από κατάλοιπα αδενίνης συχνά περιγράφεται ως *ουρά πολυ(A),* ενώ το mRNA που έχει αυτό το χαρακτηριστικό δηλώνεται ως *πολυ(A)*$^+$ (*poly(A)*$^+$). Όπως προαναφέρθηκε, η αλληλουχία πολυ(A) δεν κωδικοποιείται στο DNA αλλά προστίθεται στο RNA, μέσα στον πυρήνα, μετά τη μεταγραφή. Η προσθήκη της ουράς πολυ(A) καταλύεται από το ένζυμο *πολυ(A) πολυμεράση* (*poly(A) polymerase*), η οποία προσθέτει περίπου 200 κατάλοιπα A στο ελεύθερο 3΄-OH άκρο του mRNA. Η αλληλουχία πολυ(A), τόσο του πυρηνικού RNA όσο και του mRNA είναι συνδεδεμένη με μια πρωτεΐνη που ονομάζεται *πολυ(A) συνδεόμενη πρωτεΐνη (PABP, Poly(A)-Binding Protein).* Κάποιες από τις επιδράσεις της ουράς πολυ(A) στις ιδιότητες του mRNA, όπως η σταθερότητα ή η ικανότητα να προστατεύεται από την αποικοδόμηση, σχετίζονται με την σύνδεση της πολυ(A) με την PABP.

3.5.4 Μετάφραση

Στο στάδιο της *μετάφρασης* (*translation*) γίνεται η αντιστοίχιση τριπλετών (τριών διαδοχικών νουκλεοτιδίων) σε αμινοξέα και η διαδοχική σύνδεση των αμινοξέων σε πολυπεπτιδική αλυσίδα. Αυτές οι τριπλέτες καλούνται *κωδικόνια* (*codons*). Ο κώδικας αντιστοίχισης των νουκλεοτιδίων του mRNA σε αμινοξέα ονομάζεται *γενετικός κώδικας* (*genetic code*) ή *κώδικας τριπλέτας* (*triplet code*) (Πίνακας 3.4). Υπάρχουν αμινοξέα που κωδικοποιούνται από περισσότερα του ενός κωδικόνια, καθώς υπάρχουν 4^3 = 64 διαφορετικά κωδικόνια και μόνο 20 αμινοξέα. Για το λόγο αυτό ο γενετικός κώδικας

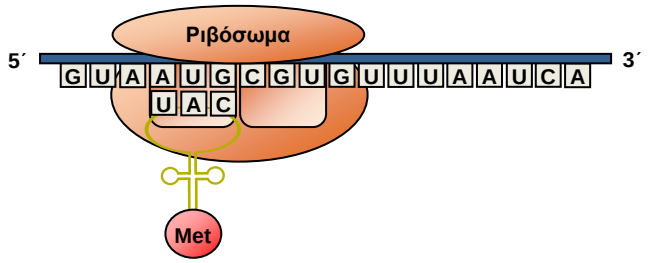
χαρακτηρίζεται ως *εκφυλισμένος* (*degenerate*), ενώ τα κωδικόνια που κωδικοποιούν το ίδιο αμινοξύ καλούνται *συνώνυμα*. Μάλιστα, υπάρχουν και τρία *κωδικόνια λήξης* (*stop codons*), τα UAA, UAG και UGA, η παρουσία των οποίων σηματοδοτεί το τέλος της αλληλουχίας των αμινοξέων, καθώς και ένα *κωδικόνιο έναρξης* (*start codon*), το AUG, που κωδικοποιεί το αμινοξύ *μεθειονίνη* (*methionine*). Επιπλέον, ο γενετικός κώδικας είναι σχεδόν *καθολικός* (*universal*), αφού ισχύει για όλους σχεδόν τους οργανισμούς.

Η διαδικασία της μετάφρασης πραγματοποιείται από τα ριβοσώματα, τα οποία είναι μεγάλα συμπλέγματα πρωτεϊνών και μορίων RNA αποτελούμενα από δύο υπομονάδες. Τα ριβοσώματα περιέχουν μια θέση πρόσδεσης του μορίου mRNA και τρεις θέσεις πρόσδεσης μορίων tRNA, γνωστές ως θέση A, θέση P και θέση E. Στη θέση A και τη θέση P τα μόρια tRNA προσδένονται με μεγάλη σταθερότητα, εφόσον φέρουν ένα *αντικωδικόνιο* (*anticodon*) που είναι συμπληρωματικό με το αντίστοιχο κωδικόνιο του mRNA. Αντίθετα, ένα μόριο tRNA που καταλαμβάνει τη θέση E διίσταται και απομακρύνεται. Ένα μόριο tRNA συνδέεται με ένα αμινοξύ και το μεταφέρει στο ριβόσωμα, καταλαμβάνοντας τη θέση A. Το αμινοξύ προστίθεται στην αναπτυσσόμενη πρωτεΐνη (θέση P) και στη συνέχεια το μόριο tRNA απομακρύνεται (θέση E). Η Εικόνα 3.10 παρέχει ένα υποτυπώδες παράδειγμα μετάφρασης περιγράφοντας αναλυτικά τη διαδικασία. Για λόγους απλότητας δεν απεικονίζεται η θέση E.

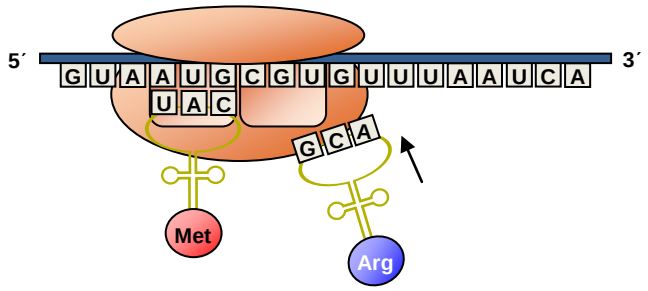
Οι πρωτεΐνες μετά το πέρας της μετάφρασης μπορούν να υποστούν *μετα-μεταφραστικές τροποποιήσεις* (*post-translational modifications*), γεγονός που επηρεάζει τη λειτουργία τους. Εξαιτίας του εναλλακτικού ματίσματος και των μετα-μεταφραστικών τροποποιήσεων ένα γονίδιο μπορεί να παράγει περισσότερες από μία πρωτεΐνες.

Πίνακας 3.4: Ο γενετικός κώδικας.

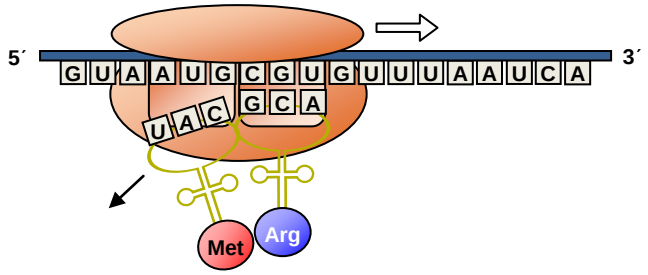
	U		**C**		**A**		**G**		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	**U**
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	**C**
	UUA	Leu	UCA	Ser	UAA	Λήξη	UGA	Λήξη	**A**
	UUG	Leu	UCG	Ser	UAG	Λήξη	UGG	Trp	**G**
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	**U**
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	**C**
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	**A**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



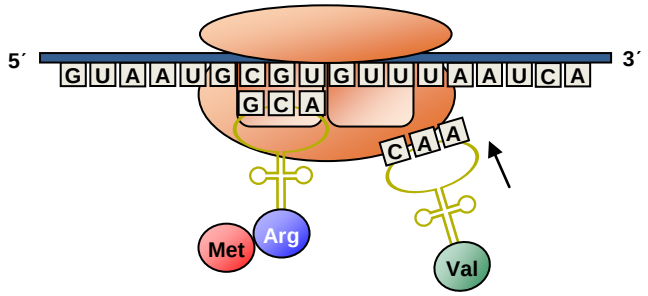
Το ριβόσωμα προσδένεται στο mRNA και βρίσκει το σημείο έναρξης της μετάφρασης, δηλαδή το κωδικόνιο AUG. Το αντικωδικόνιο (anticodon) 3΄-UAC-5΄ του tRNA που φέρει το αμινοξύ μεθειονίνη συνδέεται με το αντίστοιχο κωδικόνιο.



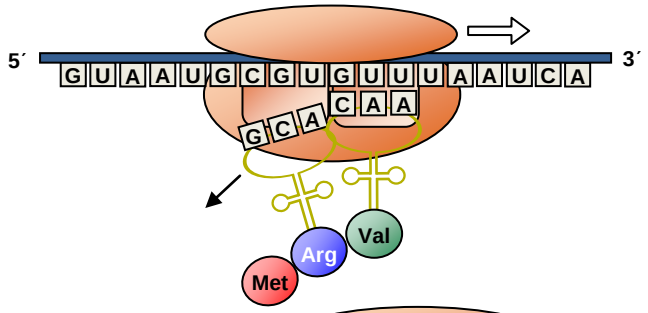
Το αντικωδικόνιο 3΄-GCA-5΄ του tRNA που φέρει το αμινοξύ αργινίνη συνδέεται με το αντίστοιχο κωδικόνιο CGU.



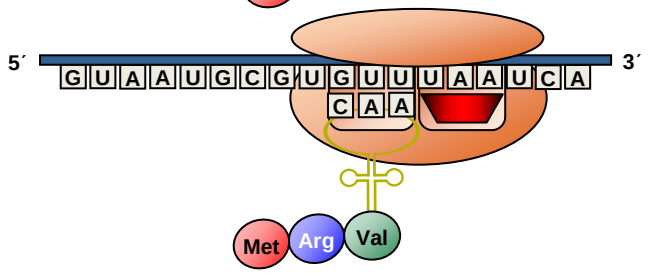
Σχηματίζεται πεπτιδικός δεσμός μεταξύ της μεθειονίνης και της αργινίνης και το ελεύθερο tRNA απομακρύνεται. Έπειτα το ριβόσωμα μετακινείται κατά μια τριπλέτα με κατεύθυνση 5΄ → 3΄.



Το αντικωδικόνιο 3΄-CAA-5΄ του tRNA που φέρει το αμινοξύ βαλίνη συνδέεται με το αντίστοιχο κωδικόνιο GUU.



Σχηματίζεται πεπτιδικός δεσμός μεταξύ της αργινίνης και της βαλίνης και το ελεύθερο tRNA απομακρύνεται. Έπειτα το ριβόσωμα μετακινείται κατά μια τριπλέτα με κατεύθυνση 5΄ → 3΄.



Το κωδικόνιο UAA κωδικοποιεί τη λήξη της μετάφρασης. Ο παράγοντας απελευθέρωσης (release factor) συνδέεται με το κωδικόνιο λήξης, τερματίζοντας τη μετάφραση και απελευθερώνοντας την πολυπεπτιδική αλυσίδα.

Εικόνα 3.10: Η διαδικασία της μετάφρασης.

3.6 Γονιδιακή Ρύθμιση

Ο όρος *γονιδιακή ρύθμιση* (*gene regulation*) αναφέρεται σε όλες τις διαδικασίες με της οποίες πραγματοποιείται η ρύθμιση της έκφρασης των γονιδίων. Εφόσον όλα τα κύτταρα σε έναν οργανισμό έχουν το ίδιο γενετικό υλικό, η κυτταρική διαφοροποίηση οφείλεται σε διαφορές της γονιδιακής έκφρασης και όχι σε διαφορές του περιεχομένου του γονιδιώματος. Η γονιδιακή ρύθμιση μπορεί να επιμεριστεί στη ρύθμιση της διαδικασίας της μεταγραφής, και στη ρύθμιση της διαδικασίας της μετάφρασης. Ωστόσο, η ρύθμιση της μεταγραφής είναι εκείνη που σχεδόν αποκλειστικά ρυθμίζει την έκφραση των γονιδίων.

Σημαντικό ρόλο στη γονιδιακή ρύθμιση έχει ένα είδος πρωτεϊνών, που ονομάζονται *μεταγραφικοί παράγοντες* (*transcription factors*). Οι υποκινητές μπορούν να προσδεθούν σε κάποια ειδικά τμήματα του DNA που ονομάζονται *θέσεις δέσμευσης* (*binding sites*) των μεταγραφικών παραγόντων, οι οποίες βρίσκονται στις αποκαλούμενες περιοχές *υποκινητή* (*promoter*). Συγκεκριμένοι υποκινητές σχετίζονται με συγκεκριμένα γονίδια και οι αλληλουχίες τους βρίσκονται κοντά στις αλληλουχίες των αντίστοιχων γονιδίων. Οι μεταγραφικοί παράγοντες ελέγχουν τη γονιδιακή έκφραση δεσμεύοντας τον υποκινητή του γονιδίου και είτε ενεργοποιούν τη μεταγραφή του γονιδίου είτε την καταστέλλουν.

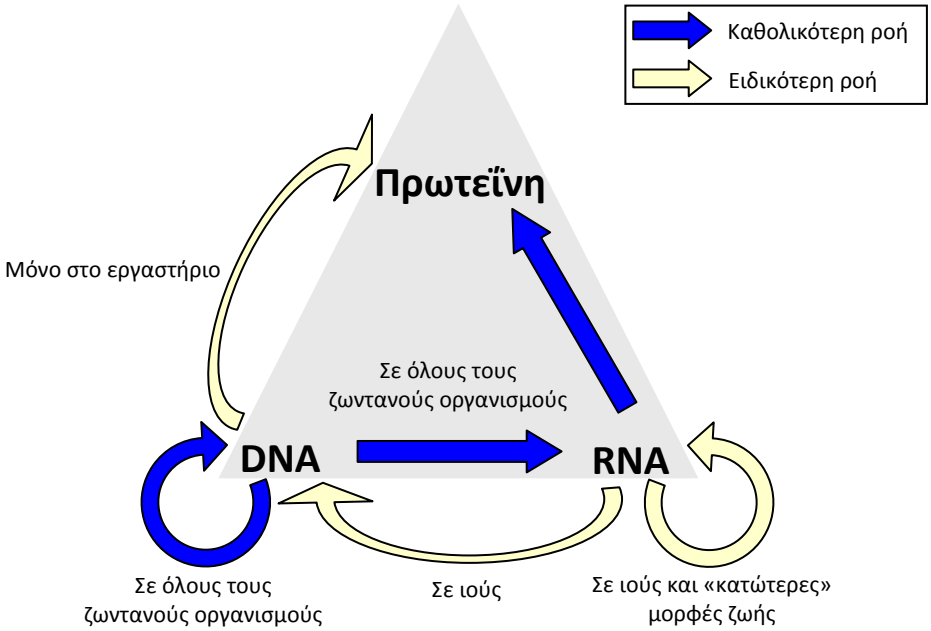
Οι μεταγραφικοί παράγοντες μπορούν να ελέγχουν πολλά γονίδια, ενώ κάποια γονίδια ελέγχονται από συνδυασμό μεταγραφικών παραγόντων. Βρόχοι ανάδρασης είναι πιθανοί. Μπορεί, λοιπόν, να γίνει λόγος για *γονιδιακά ρυθμιστικά δίκτυα* (*gene regulatory networks*). Η κατανόηση, η περιγραφή και η μοντελοποίηση τέτοιων δικτύων είναι ένα πολύ ενδιαφέρον πρόβλημα που απασχολεί τα τελευταία χρόνια τους επιστήμονες.

3.7 Η Ροή της Βιολογικής Πληροφορίας

Η ροή της γενετικής πληροφορίας από τα νουκλεϊκά οξέα προς τις πρωτεΐνες, δηλαδή η διαδικασία της πρωτεϊνοσύνθεσης που περιγράφηκε παραπάνω, αλλά και η μεταβίβαση της πληροφορίας από τους γονείς στους απογόνους, περιγράφεται με το *κεντρικό δόγμα* (*central dogma*) της μοριακής βιολογίας. Το κεντρικό δόγμα διατυπώθηκε το 1958 από τον Francis Crick. Σύμφωνα με αυτή τη διατύπωση του δόγματος (Εικόνα 3.11), το DNA μεταγράφεται σε RNA, το οποίο στη συνέχεια μεταφράζεται σε πρωτεΐνες. Το κυκλικό βέλος γύρω από το DNA υποδηλώνει τη δυνατότητα διπλασιασμού του.

Εικόνα 3.11: Το κεντρικό δόγμα της μοριακής βιολογίας.

Ωστόσο, σήμερα είναι γνωστό ότι στους ρετροϊούς, όπως για παράδειγμα ο HIV, που προκαλεί το AIDS, το RNA με τη διαδικασία της αντίστροφης μεταγραφής μπορεί να μετατραπεί σε DNA. Επιπλέον, σε κάποιους ιούς και άλλους «κατώτερους» οργανισμούς είναι δυνατός ο αυτοδιπλασιασμός του RNA. Η ροή της βιολογικής πληροφορίας περιγράφεται πλέον όπως στην Εικόνα 3.12.



Εικόνα 3.12: Η ροή της βιολογικής πληροφορίας.

3.8 Βιοπληροφορική

Η μεγάλη συσσώρευση βιολογικών δεδομένων οδήγησε στη γέννηση μιας νέας επιστημονικής περιοχής η οποία είναι αφοσιωμένη στη διαχείριση και την ανάλυση αυτών των δεδομένων με τη χρήση προηγμένων υπολογιστικών τεχνικών. Η νέα αυτή επιστημονική περιοχή, ονομάζεται *βιοπληροφορική* (*bioinformatics*). Έχουν υπάρξει αρκετές συζητήσεις σχετικά με τον ορισμό της βιοπληροφορικής και το διαχωρισμό της από άλλα συναφή πεδία, όπως είναι η *υπολογιστική βιολογία* (*computational biology*). Η υπολογιστική βιολογία ασχολείται με την ανάπτυξη και εφαρμογή μαθηματικών και πληροφοριακών θεωριών για την μελέτη βιολογικών συστημάτων και την προσέγγιση ή ακόμα και απάντηση θεμελιωδών θεωρητικών και πειραματικών προβλημάτων της επιστήμης της βιολογίας. Μολονότι τα πεδία της βιοπληροφορικής και της υπολογιστικής βιολογίας εννοιολογικά είναι σαφώς διαχωρισμένα, πρακτικά στον τομέα της έρευνας, υπάρχει σε μεγάλο βαθμό αλληλοεπικάλυψη. Το γεγονός αυτό έχει ως αποτέλεσμα να χρησιμοποιούνται αυτοί οι δύο όροι εναλλάξ. Άλλοι όροι που σπανιότερα χρησιμοποιούνται αντί της βιοπληροφορικής είναι η *θεωρητική βιολογία* (*theoretical biology*) και η *βιοϋπολογιστική* (*biocomputing*) [Glasgow *et al.*, 2004]. Αυτή η σύγχυση που επικρατεί οφείλεται στο γεγονός ότι η βιοπληροφορική είναι μια διεπιστημονική περιοχή, η οποία αποτελεί τη διασταύρωση της βιολογίας, με την επιστήμη των υπολο-

γιστών και την τεχνολογία πληροφοριών. Οι διαφορετικοί αντιπρόσωποι του πεδίου έχουν και διαφορετικές απόψεις για το σκοπό και το ρόλο της βιοπληροφορικής. Στις παραγράφους που ακολουθούν θα δοθεί μια γενική περιγραφή του πεδίου, θα καταγραφούν τα σημαντικότερα προβλήματα με τα οποία ασχολείται και θα παρουσιαστούν οι βασικότερες εφαρμογές της ανακάλυψης γνώσης στη βιοπληροφορική [Tzanis et al., 2005b; Tzanis et al., 2005c; Tzanis et al., 2009].

3.8.1 Οι Στόχοι της Βιοπληροφορικής

Οι βασικοί στόχοι της βιοπληροφορικής είναι:

- Η οργάνωση των δεδομένων με τρόπο που να επιτρέπει στους ερευνητές την προσπέλαση της διαθέσιμης πληροφορίας, καθώς και την εισαγωγή νέων δεδομένων.
- Η ανάπτυξη εργαλείων και μέσων που βοηθούν την ανάλυση των δεδομένων.
- Η χρήση των εργαλείων αυτών για την ανάλυση των δεδομένων και την ερμηνεία των αποτελεσμάτων, ώστε να προκύψει βιολογικά σημαντική γνώση.

3.9 Εύρεση και Ανάλυση Βιολογικών Αλληλουχιών

Ένα πολύ μεγάλο τμήμα των βιολογικών δεδομένων που καλούνται να οργανώσουν και να αναλύσουν οι επιστήμονες είναι αλληλουχίες. Παρακάτω παρουσιάζονται οι βασικότερες εφαρμογές της βιοπληροφορικής σε βιολογικές αλληλουχίες.

3.9.1 Αλληλούχιση DNA

Η διαδικασία εύρεσης της αλληλουχίας των νουκλεοτιδικών βάσεων σε ένα τμήμα μορίου DNA, είναι γνωστή ως *αλληλούχιση DNA* (*DNA sequencing*). Η πρώτη αλληλούχιση ενός πλήρους γονιδιώματος οργανισμού και συγκεκριμένα βακτηρίου έγινε το 1995, ενώ η αποκρυπτογράφηση του ανθρώπινου γονιδιώματος ολοκληρώθηκε το 2003. Η αλληλούχιση DNA σε σχετικά μικρά βακτηριακά γονιδιώματα δεν προβάλει ιδιαίτερες προκλήσεις. Το κύριο πρόβλημα είναι η ελαχιστοποίηση του κόστους ανά νουκλεοτίδιο και η μεγιστοποίηση της ταχύτητας, με ταυτόχρονη διατήρηση της ποιότητας. Η διαδικασία της αλληλούχισης περιλαμβάνει την εύρεση αλληλουχιών μικρών σχετικά τμημάτων DNA, τα οποία στη συνέχεια συναρμολογούνται από υπολογιστές με τη βοήθεια ειδικών αλγορίθμων. Η μεγάλη δυσκολία έγκειται στην πολυπλοκότητα των γονιδιωμάτων των ανώτερων ευκαρυωτικών οργανισμών, τα οποία παρουσιάζουν πολλές επαναλαμβανόμενες υποαλληλουχίες, γεγονός που καθιστά τη διαδικασία συναρμολόγησης δύσκολη. Η μεσολάβηση του ανθρώπου στο τελικό στάδιο της διαδικασίας είναι απαραίτητη.

3.9.2 Πρόβλεψη Γονιδίων

Η πρόβλεψη γονιδίων περιλαμβάνει την αναγνώριση τμημάτων του DNA που έχουν κάποιον ενεργό βιολογικό ρόλο. Αυτό είναι και το επόμενο βήμα μετά την αλληλούχιση

και θεωρείται ιδιαίτερα σημαντικό για τον *υπομνηματισμό* (*annotation*) του γονιδιώματος ενός οργανισμού, δηλαδή την σύνδεση βιολογικών λειτουργιών με τις αλληλουχίες του. Ωστόσο, η ακρίβεια μιας τέτοιας πρόβλεψης δεν είναι πολύ μεγάλη. Η περισσότερη γνώση που απαιτείται για τέτοιου είδους προβλέψεις προέρχεται από πειραματικά προσδιορισμένα γονίδια. Η πρόβλεψη γονιδίων αποτελεί πεδίο έρευνας της βιοπληροφορικής. Υπάρχουν τρεις βασικές προσεγγίσεις για την πρόβλεψη γονιδίων μέσα στην αλληλουχία του DNA [Τσακαλίδης κ.α., 2002]:

- Αναζήτηση βάσει σήματος (signal-based), κατά την οποία τα γονίδια εντοπίζονται «έμμεσα» με την εύρεση συγκεκριμένων σημάτων, δηλαδή συγκεκριμένων περιοχών του DNA, που πραγματοποιούν κάποια ιδιαίτερη λειτουργία, όπως π.χ. τη σύνδεση ενός ενζύμου.
- Αναζήτηση βάσει περιεχομένου (content-based), κατά την οποία τα γονίδια εντοπίζονται με την εύρεση τμημάτων της αλληλουχίας του DNA, μοτίβων (motifs), τα οποία έχουν τα γενικά χαρακτηριστικά των κωδικών περιοχών. Η αναζήτηση βάσει περιεχομένου εκμεταλλεύεται τη γνώση της διαφοράς των στατιστικών σε κωδικές και μη κωδικές περιοχές.
- Αναζήτηση βάσει ομοιότητας (similarity-based), κατά την οποία τα γονίδια εντοπίζονται με βάση την ομοιότητα με άλλα γνωστά γονίδια.

Στην αναζήτηση βάση περιεχομένου ο εντοπισμός των γονιδίων γίνεται «άμεσα», που σημαίνει ότι είναι γνωστή η αλληλουχία των βάσεων που απαρτίζουν το συγκεκριμένο γονίδιο και αναζητείται στην αλληλουχία του μορίου του DNA. Από την άλλη, όταν δεν υπάρχουν στοιχεία για το γονίδιο, η αναζήτηση γίνεται σε περιοχές που είναι πιθανό να εμφανίζεται. Τέτοιες περιοχές είναι αυτές που βρίσκονται κοντά στα σήματα.

Η αναγνώριση τέτοιων αλληλουχιών (σημάτων ή μοτίβων) είναι ένα από τα προβλήματα με τα οποία ασχολούνται οι επιστήμονες της βιοπληροφορικής. Μια τέτοια αλληλουχία είναι το *σημείο έναρξης μεταγραφής* (*transcription start site*). Η αλληλουχία αυτή, όπως φανερώνει και το όνομά της, αποτελεί την περιοχή του DNA στην οποία αρχίζει η διαδικασία της μεταγραφής. Μια άλλη βιολογικά σημαντική αλληλουχία είναι το *σημείο έναρξης μετάφρασης* (*translation initiation site*). Παραδείγματα άλλων τέτοιων αλληλουχιών είναι οι υποκινητές, τα ιντρόνια και τα εξόνια.

3.9.3 Στοίχιση Αλληλουχιών

Η *στοίχιση αλληλουχιών* (*sequence alignment*) είναι η συγκριτική μέθοδος που χρησιμοποιείται για την εύρεση δομικών, λειτουργικών και εξελικτικών σχέσεων μεταξύ βιολογικών αλληλουχιών. Υπάρχουν δύο βασικοί τύποι στοίχισης αλληλουχιών:

- Η *κατά ζεύγη* (*pairwise*) στοίχιση αλληλουχιών, επιδιώκει την εύρεση της καλύτερης στοίχισης μεταξύ δύο πρωτεϊνικών ή νουκλεοτιδικών αλληλουχιών. Σκοπός αυτού του είδους στοίχισης είναι η εύρεση ομολόγων ενός γονιδίου ή πρω-

τεΐνης σε μια βάση δεδομένων από γνωστά παραδείγματα. Για να εξαχθεί το συμπέρασμα της ομολογίας μεταξύ δύο γονιδίων πρέπει να καθοριστεί αν τα συγκεκριμένα γονίδια παρουσιάζουν αρκετή ομοιότητα. Η *ομοιότητα* (*similarity*) είναι η παρατηρούμενη ποσότητα, η οποία μπορεί να εκφραστεί ως ποσοστό ταύτισης ή ως κάποιο άλλο κατάλληλο μέτρο. Η *ομολογία* (*homology*) αναφέρεται στο συμπέρασμα που προκύπτει από τα δεδομένα ομοιότητας και μπορεί να σημαίνει ότι δύο γονίδια έχουν κοινή εξελικτική προέλευση. Τα γονίδια είτε είναι, είτε δεν είναι ομόλογα. Δεν υπάρχουν βαθμοί ομολογίας, όπως γίνεται με την περίπτωση της ομοιότητας [Schuler, 1998]. Οι πληροφορίες αυτές χρησιμεύουν στην απάντηση μιας ποικιλίας βιολογικών ερωτημάτων. Η σημαντικότερη εφαρμογή της κατά ζεύγη στοίχισης είναι η αναγνώριση αλληλουχιών άγνωστης δομής ή/και λειτουργίας. Διακρίνονται δύο μέθοδοι στοίχισης μεταξύ δύο αλληλουχιών, η *τοπική* (*local*) και η *ολική* (*global*) στοίχιση. Στις μεθόδους ολικής στοίχισης μεταξύ δύο αλληλουχιών συμμετέχουν όλοι οι χαρακτήρες που αναπαριστούν τις δύο αλληλουχίες. Οι ολικές στοιχίσεις είναι περισσότερο χρήσιμες στην εύρεση στενά συγγενικών αλληλουχιών. Ωστόσο, αυτές οι αλληλουχίες μπορούν να εντοπιστούν και με τις μεθόδους τοπικής στοίχισης. Οι μέθοδοι τοπικής στοίχισης βρίσκουν συγγενικές περιοχές μέσα στις αλληλουχίες, δηλαδή μπορούν να αποτελούνται από ένα υποσύνολο των χαρακτήρων που συνθέτουν κάθε αλληλουχία. Για παράδειγμα οι θέσεις 10-30 της μίας αλληλουχίας μπορεί να ταιριάζουν με τις θέσεις 40-60 της άλλης αλληλουχίας. Η μέθοδοι τοπικής στοίχισης είναι πιο ευέλικτες από της μεθόδους ολικής στοίχισης και έχουν το πλεονέκτημα της εύρεσης συγγενικών περιοχών που εμφανίζονται σε διαφορετική σειρά μέσα σε κάθε αλληλουχία.

- Η *πολλαπλή* (*multiple*) στοίχιση αλληλουχιών επεκτείνει την κατά ζεύγη στοίχιση, για να συμπεριλάβει περισσότερες αλληλουχίες. Οι μέθοδοι πολλαπλής στοίχισης δεν αναζητούν αλληλουχίες σε μια βάση δεδομένων, αλλά λαμβάνουν μερικές αλληλουχίες και βρίσκουν κοινές περιοχές μεταξύ όλων αυτών. Μια εφαρμογή της πολλαπλής στοίχισης είναι στην *κλαδιστική* (*cladistics*), τον κλάδο της βιολογίας που μελετά τις εξελικτικές σχέσεις μεταξύ των οργανισμών, ως μέθοδο δημιουργίας φυλογενετικών δένδρων, καθώς και στη δημιουργία πρότυπων αλληλουχιών, που μπορούν να χρησιμοποιηθούν για την αναζήτηση σε βάσεις δεδομένων λιγότερο στενών συγγενειών.

Δύο από τους πιο γνωστούς αλγορίθμους που χρησιμοποιούνται για τη στοίχιση αλληλουχιών είναι ο αλγόριθμος των Needleman-Wunsch [Needleman & Wunch, 1970], που χρησιμοποιείται για ολική κατά ζεύγη στοίχιση και ο αλγόριθμος των Smith-Waterman [Smith & Waterman, 1981], που χρησιμοποιείται για ολική ή τοπική κατά ζεύγη στοίχιση. Εργαλεία όπως το Clustal (Ενότητα 3.15) χρησιμοποιούνται για πολλαπλή στοίχιση αλληλουχιών.

3.10 Μέτρηση και Ανάλυση της Γονιδιακής Έκφρασης

Στα επόμενα εδάφια παρουσιάζονται οι δύο πιο σημαντικές μέθοδοι μέτρησης των επιπέδων της γονιδιακής έκφρασης, οι μικροσυστοιχίες και η τεχνική SAGE.

3.10.1 Μικροσυστοιχίες

Οι *μικροσυστοιχίες* (*microarrays*) [Schena *et al.*, 1995] χρησιμοποιούν τις αλληλουχίες που έχουν δημιουργηθεί από προγράμματα χαρτογράφησης γονιδιωμάτων καθώς και από άλλα προγράμματα αλληλούχισης DNA, με σκοπό τον προσδιορισμό των γονιδίων που εκφράζονται σε κάποιο συγκεκριμένο τύπο κυττάρου ενός οργανισμού, σε συγκεκριμένη χρονική στιγμή και κάτω από συγκεκριμένες συνθήκες. Υπάρχουν αρκετά εναλλακτικά ονόματα αυτής της τεχνολογίας, όπως *μικροσυστοιχία DNA* (*DNA microarray*), *συστοιχία DNA* (*DNA array*), *τσιπ DNA* (*DNA chip*), *τσιπ γονιδίων* (*gene chip*) και άλλες. Στην Εικόνα 3.13 παρουσιάζεται μια τέτοια συσκευή.

Εικόνα 3.13: Η συστοιχία GeneChip® της εταιρείας Affymetrix.

Υπάρχουν διάφοροι τρόποι με τους οποίους μία μικροσυστοιχία μπορεί να χρησιμοποιηθεί για να μετρήσει τα επίπεδα της γονιδιακής έκφρασης. Μία από τις πιο δημοφιλείς εφαρμογές των μικροσυστοιχιών επιτρέπει τη σύγκριση των επιπέδων γονιδιακής έκφρασης σε δύο διαφορετικά δείγματα, για παράδειγμα τον ίδιο τύπο κυττάρου σε υγιή και ασθενή κατάσταση.

Εφαρμογές των Μικροσυστοιχιών

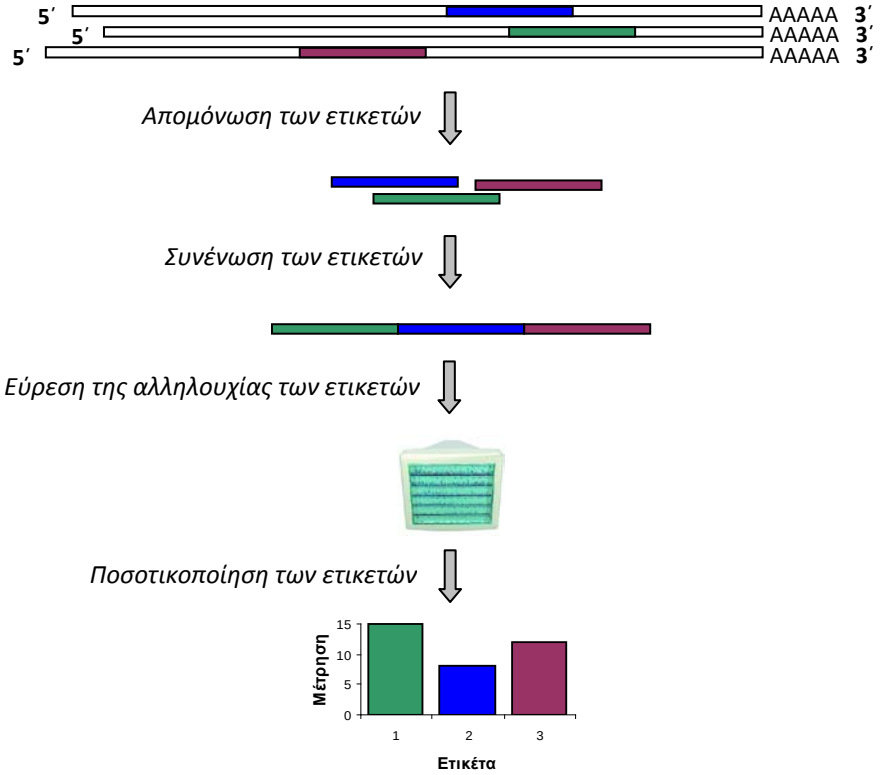
Από το 1995 οι μικροσυστοιχίες έχουν χρησιμοποιηθεί ευρέως για πολλές εφαρμογές οι σημαντικότερες από τις οποίες παρουσιάζονται παρακάτω:

- *Πρόβλεψη γονιδίων*, δηλαδή τη διαδικασία ανακάλυψης αλληλουχιών που αντιστοιχούν σε μια συγκεκριμένη φυσιολογική λειτουργία.
- *Διάγνωση ασθενειών*, εντοπίζοντας μεταλλάξεις σε γονίδια που προκαλούν ασθένειες.
- *Ανακάλυψη νέων φαρμάκων*, μέσω της εύρεσης συσχετίσεων μεταξύ των θεραπευτικών αποκρίσεων σε φάρμακα και των γενετικών προφίλ ασθενών.

- *Τοξικολογική έρευνα*, μέσω της εύρεσης συσχετίσεων μεταξύ των αποκρίσεων σε τοξικές ουσίες και των αλλαγών στα γενετικά προφίλ των ατόμων που εκτίθενται σε τέτοιες ουσίες [Shi, 2001].

3.10.2 Η Τεχνική SAGE

Η τεχνική *SAGE* (Serial Analysis of Gene Expression) είναι μια μέθοδος που επινοήθηκε στο πανεπιστήμιο Johns Hopkins, στη Βαλτιμόρη των Η.Π.Α. [Velculescu et al., 1995], με σκοπό να δώσει στους επιστήμονες τη δυνατότητα της εποπτικής παρακολούθησης του συνόλου της γονιδιακής δραστηριότητας ενός κυττάρου. Η διαδικασία μετρά τον αριθμό των *μεταγράφων mRNA* (*mRNA transcripts*) που υπάρχουν σε μια δεδομένη στιγμή σε ένα κύτταρο. Τα συγκεκριμένα mRNA χαρακτηρίζονται από μια μικρή αλληλουχία δέκα περίπου βάσεων, η οποία καλείται *ετικέτα* (*tag*). Το σύνολο των ετικετών που παράγονται σε ένα πείραμα σχηματίζουν μια εγγραφή που καλείται *βιβλιοθήκη* (*library*). Η εφαρμογή της τεχνικής SAGE είναι ακριβότερη από τη διεξαγωγή πειραμάτων με μικροσυστοιχίες[10]. Η τεχνική SAGE λειτουργεί με τον τρόπου που παρουσιάζει η Εικόνα 3.14. Αρχικά, απομονώνονται οι ετικέτες των αλληλουχιών, συνενώνονται και πραγματοποιείται αλληλούχισή τους. Στο τέλος καταγράφεται ο αριθμός εμφανίσεων κάθε ετικέτας.



Εικόνα 3.14: Η τεχνική SAGE.

[10] Αυτός είναι και κύριος λόγος που υπάρχουν λιγότερα δημοσίως διαθέσιμα SAGE δεδομένα σε σύγκριση με τα δεδομένα μικροσυστοιχιών.

3.11 Νεολογισμοί της Βιοπληροφορικής

Η κατάληξη «-ωμική» αποτελεί νεολογισμό που αναφέρεται σε μια ευρεία περιοχή της επιστήμης, που μελετά και αναλύει τις αλληλεπιδράσεις αντικειμένων βιολογικής πληροφορίας, που διακρίνονται από την κατάληξη «-ωμα». Για παράδειγμα, η *γενωμική* ή *γονιδιωματική* περιλαμβάνει τη μελέτη του *γενώματος* ή *γονιδιώματος*, η *πρωτεωμική* περιλαμβάνει τη μελέτη του *πρωτεώματος* (το σύνολο των πρωτεϊνών που εκφράζουν ένα γένωμα). Μάλιστα, είναι αρκετά δύσκολο να βρεθεί ένας δόκιμος όρος για πολλούς από αυτούς τους νεολογισμούς, όπως για παράδειγμα για το «textome» (Εδάφιο 3.13.4). Γενικά, υπάρχει ένας μεγάλος αριθμός «-ωμικών» και υποπεριοχών τους. Παρακάτω περιγράφονται οι δύο περισσότερο διαδεδομένες «-ωμικές».

3.11.1 Γονιδιωματική

Γονιδιωματική ή γενωμική (*genomics*) είναι η μελέτη των γονιδιωμάτων, η οποία περιλαμβάνει τη χαρτογράφηση γονιδιώματος, την αλληλούχιση του DNA και τον προσδιορισμό της λειτουργίας γονιδίων.

3.11.2 Συγκριτική Γονιδιωματική

Η *συγκριτική γονιδιωματική* (*comparative genomics*) ασχολείται με τις συγκρίσεις αλληλουχιών του γονιδιώματος. Τα προγράμματα χαρτογράφησης γονιδιωμάτων έχουν δείξει ότι τα γονιδιώματα φαινομενικά διαφορετικών οργανισμών μπορεί να έχουν αρκετές ομοιότητες. Αυτές οι ομοιότητες υποδηλώνουν στενές εξελικτικές σχέσεις μεταξύ αυτών των οργανισμών. Είναι δυνατό να κατασκευαστεί ένα *φυλογενετικό δένδρο* (*phylogenetic tree*) της εξέλιξης των πρωτεϊνών, των γονιδίων και των οργανισμών βασισμένο σε συγκρίσεις αλληλουχιών.

Λειτουργική Γονιδιωματική

Η *λειτουργική γονιδιωματική* (*functional genomics*) χρησιμοποιεί τη γνώση που προκύπτει από τη μελέτη των γονιδιωμάτων με σκοπό την κατανόηση του ρόλου των γονιδίων, της λειτουργίας και αλληλεπίδρασης των προϊόντων τους και κυρίως του τρόπου με τον οποίο όλα αυτά κάνουν τους οργανισμούς να λειτουργούν με συγκεκριμένο τρόπο.

Οι πρωτεΐνες επιτελούν ένα μεγάλο πλήθος από λειτουργίες μέσα σε ένα κύτταρο. Οι βιολόγοι πιστεύουν ότι υπάρχει ένας περιορισμένος αριθμός γονιδίων και αντίστοιχων πρωτεϊνών, παρατηρώντας τα από λειτουργική σκοπιά, πολλά από τα οποία μάλιστα εμφανίζονται στα περισσότερα γονιδιώματα. Για παράδειγμα, ένα ποσοστό γονιδίων σκουληκιών κωδικοποιούν πρωτεΐνες των οποίων ο βιολογικός ρόλος μπορεί να προβλεφθεί από την ομοιότητά των γονιδίων αυτών με τα γονίδια της ζύμης και αντίστροφα. Η αναγνώριση αυτού του περιορισμένου συνόλου γονιδίων έχει οδηγήσει στην ενοποίηση της βιολογίας κάτω από το πρίσμα της *γονιδιακής οντολογίας* (*Gene Ontology – GO*) [The Gene Ontology Consortium, 2000]. Σκοπός του προγράμματος αυτού είναι η ανάπτυξη ενός δυναμικού ελεγχόμενου λεξιλογίου για την περιγραφή της λειτουργίας και της θέσης των γονιδίων. Η χρήση των όρων της γονιδιακής οντολογίας

επιτρέπει την αυτοματοποιημένη αναζήτηση γονιδιακών λειτουργιών σε βάσεις δεδομένων.

Η συγκέντρωση των πρωτεϊνών εξαρτάται από πολλούς παράγοντες, όπως την έκφραση ή μη-έκφραση του αντίστοιχου γονιδίου, τη συχνότητα έκφρασης, το χρόνο και την ταχύτητα ωρίμανσης, μετάφρασης και τροποποίησης, το χρόνο ημιζωής του mRNA και των πρωτεϊνών και την απώλεια της λειτουργικότητάς τους σε κάποιες στιγμές. Άμεσες πειραματικές μελέτες είναι τεχνικά δύσκολες. Ωστόσο, χάρη στις μεθόδους καταγραφής της γονιδιακής έκφρασης, π.χ. την τεχνολογία των μικροσυστοιχιών, είναι δυνατό να μετρηθεί με ένα πείραμα η συγκέντρωση του mRNA για δεκάδες χιλιάδες γονίδια ταυτόχρονα. Δεν υπάρχει άμεση συσχέτιση μεταξύ της γονιδιακής έκφρασης και της παρουσίας των αντίστοιχων πρωτεϊνών στο κύτταρο, όμως σε πολλές περιπτώσεις μπορούν να γίνουν κάποιες εκτιμήσεις. Ένα σχετικό θέμα με τη συγκέντρωση των πρωτεϊνών στο κύτταρο είναι οι αλληλεπιδράσεις μεταξύ πρωτεϊνών σε κάποια στιγμή της ζωής του κυττάρου. Μπορεί, λοιπόν να γίνει λόγος για δίκτυα *αλληλεπίδρασης πρωτεΐνης με πρωτεΐνη* (*protein-protein interaction*).

3.11.3 Πρωτεωμική

Η *Πρωτεωμική* (*Proteomics*) ασχολείται με την ευρείας κλίμακας μελέτη των πρωτεϊνών, ιδιαιτέρως της δομής και της λειτουργίας τους.

Σε έναν οργανισμό υπάρχουν πολύ περισσότερες πρωτεΐνες από γονίδια. Αυτό συμβαίνει λόγω της δυνατότητας εναλλακτικού ματίσματος και των μετα-μεταφραστικών τροποποιήσεων. Ωστόσο, υπάρχουν κάποια χαρακτηριστικά των πρωτεϊνών, όπως τα *μοτίβα* (*motifs*) και οι *πτυχώσεις* (*folds*), που επιτρέπουν την ταξινόμησή τους σε ομάδες και οικογένειες πρωτεϊνών, κάνοντας τη μελέτη τους ευκολότερη.

Υπάρχει ένα ευρύ φάσμα τεχνολογιών που χρησιμοποιούνται στην πρωτεωμική, ωστόσο η χρήση της ηλεκτροφόρησης γέλης δυο διαστάσεων (*2D gel electrophoresis)* και *φασματομετρία μάζας* (*mass spectrometry*) είναι η διαδικασίες που χρησιμοποιούνται συνήθως. Η ηλεκτροφόρηση γέλης δυο διαστάσεων χρησιμοποιείται για το διαχωρισμό των πρωτεϊνών με βάση το ισοηλεκτρικό τους σημείο (το pH στο οποίο το καθαρό φορτίο της πρωτεϊνης είναι μηδενικό) και στη συνέχεια με βάση το μέγεθός τους. Η φασματομετρία μάζας χρησιμοποιείται για τον προσδιορισμό της ταυτότητας και των χαρακτηριστικών της κάθε πρωτεΐνης.

3.12 Βιολογικές Βάσεις Δεδομένων

Οι βιολογικές βάσεις δεδομένων περιέχουν δεδομένα που συλλέγονται από μεγάλα ερευνητικά προγράμματα (π.χ. χαρτογραφήσεις γονιδιωμάτων), επιστημονικά πειράματα, τη δημοσιευμένη βιβλιογραφία, κ.α. Τα δεδομένα αυτά αποτελούν την κύρια πηγή πληροφοριών των επιστημόνων στην προσπάθειά τους να εξηγήσουν τα διάφορα βιολογικά φαινόμενα.

Ανάλογα με το είδος των δεδομένων που περιέχουν οι βιολογικές βάσεις δεδομένων διακρίνονται στις κατηγορίες που περιγράφονται παρακάτω σύμφωνα με τα [Μπίκης, κ.α., 2008; Ramanathan & Davison, 2002].

3.12.1 Γενικευμένες ή Αρχειακές Βιολογικές Βάσεις Δεδομένων

Οι γενικευμένες (generalized) ή αρχειακές (archival) βιολογικές βάσεις δεδομένων περιλαμβάνουν πρωτογενή βιολογική πληροφορία και διακρίνονται σε:

- Πρωτογενείς βάσεις δεδομένων αλληλουχιών (primary sequence databases) που περιέχουν αλληλουχίες DNA και πρωτεϊνών από οργανισμούς των οποίων τα γονιδιώματα έχουν αποκρυπτογραφηθεί πλήρως ή μερικώς.
- Βάσεις δεδομένων που περιέχουν τρισδιάστατες δομές νουκλεϊκών οξέων και πρωτεϊνών.

Στις αρχές της δεκαετίας του 1980 ξεκίνησαν αρκετά προγράμματα δημιουργίας πρωτογενών βάσεων δεδομένων αλληλουχιών DNA και πρωτεϊνών. Η συλλογή των δεδομένων ήταν αρχικά μικρή, ωστόσο μέσα σε λίγα χρόνια αυξήθηκε εκθετικά.

Οι τρεις μεγάλες πρωτογενείς βάσεις δεδομένων DNA είναι η Genbank (Η.Π.Α.), η EMBL (Ηνωμένο Βασίλειο) και η DDBJ (Ιαπωνία) και είναι αποτέλεσμα μιας διεθνούς συνεργασίας (International Nucleotide Sequence Databases – INSD) που αναπτύχθηκε στα τέλη της δεκαετίας του 1980. Αποστολή τους είναι η συλλογή και διάθεση αλληλουχιών DNA και RNA από πλήρως ή μερικώς αποκρυπτογραφημένα γονιδιώματα οργανισμών. Τα δεδομένα υποβάλλονται από διάφορες επιστημονικές ομάδες σε μια από τις τρεις βάσεις δεδομένων και μέσω της καθημερινής αλληλεπίδρασής τους ενημερώνονται και οι υπόλοιπες.

Μέχρι το 2002 υπήρχαν τρεις μεγάλες πρωτογενείς βάσεις δεδομένων πρωτεϊνικών αλληλουχιών. Η Swiss-Prot δημιουργήθηκε το 1987 στο Τμήμα Ιατρικής Βιοχημείας του Πανεπιστημίου της Γενεύης. Η ποιότητα των δεδομένων της Swiss-Prot ήταν υψηλή, ωστόσο η διατήρηση της ποιότητας απαιτούσε χρονοβόρες διαδικασίες γεγονός που οδήγησε τους επιστήμονες στη δημιουργία της TrEMBL (Translation of EMBL nucleotide sequence database) το 1996. Η TrEMBL προέκυψε από την ηλεκτρονική μετάφραση όλων των κωδικών αλληλουχιών της πρωτογενούς βάσης δεδομένων νουκλεοτιδικών αλληλουχιών EMBL, πλην εκείνων που υπήρχαν ήδη στη Swiss-Prot. Ωστόσο, η ηλεκτρονική μετάφραση των αλληλουχιών είχε ως αντίκτυπο τη μειωμένη αξιοπιστία των δεδομένων της TrEMBL. Η τρίτη πρωτογενής βάση πρωτεϊνικών αλληλουχιών, η PIR-PSD (Protein Information Resource – Protein Sequence Database), δημιουργήθηκε το 1984 από το National Biomedical Research Foundation (NBRF). Το 2002, οι τρεις αυτές μεγάλες βάσεις δεδομένων ενώθηκαν για να δημιουργήσουν μια μεγάλη καθολική βάση δεδομένων, τη UniProt. Η UniProt (Universal Protein Resource) αποτελεί την πιο αναλυτική συλλογή δομικών, λειτουργικών και άλλων πληροφοριών για τις αλληλουχίες πρωτεϊνών.

3.12.2 Δευτερογενείς Βιολογικές Βάσεις Δεδομένων

Οι δευτερογενείς (secondary) βιολογικές βάσεις δεδομένων προέρχονται από ανάλυση των δεδομένων που είναι αποθηκευμένα στις αρχειακές βιολογικές βάσεις δεδομένων και διακρίνονται σε:

- Δευτερογενείς βάσεις δεδομένων αλληλουχιών DNA και πρωτεϊνών, οι οποίες προκύπτουν από τις πρωτογενείς βάσεις δεδομένων αλληλουχιών με περαιτέρω επεξεργασία των δεδομένων τους και περιλαμβάνουν:
    - Βάσεις δεδομένων αλληλουχιών στις οποίες έχουν απομακρυνθεί οι αλληλουχίες που έχουν καταχωρηθεί περισσότερες από μία φορά (non-redundant databases).
    - Βάσεις δεδομένων που καταγράφουν μεταλλάξεις ή πολυμορφισμούς στις αλληλουχίες DNA και πρωτεϊνών.
    - Γονιδιωματικές βάσεις δεδομένων που είτε ομαδοποιούν συγγενή ή μερικώς αποκρυπτογραφημένα γονιδιώματα, είτε σχετίζονται με γονιδιώματα οργανισμών μοντέλων (Saccharomyces cerevisiae, Arabidopsis thaliana, Drosophila melanogaster, Homo sapiens κ.α.).
- Βάσεις δεδομένων που σχετίζονται με τις ιεραρχήσεις ή συσχετίσεις μεταξύ βιομορίων (π.χ. οικογένειες πρωτεϊνών, κοινές δομές πρωτεϊνών, κοινά μοτίβα αλληλουχιών DNA και πρωτεϊνών).

3.12.3 Εξειδικευμένες Βιολογικές Βάσεις Δεδομένων

Στην κατηγορία των εξειδικευμένων βιολογικών βάσεων δεδομένων (specialised databases) ανήκουν:

- Βάσεις δεδομένων μικροσυστοιχιών, που περιλαμβάνουν πληροφορίες για την έκφραση γονιδίων και πρωτεϊνών.
- Βάσεις δεδομένων μεταβολικών μονοπατιών, που περιέχουν πληροφορίες για τις χημικές αναδράσεις που πραγματοποιούνται στο κύτταρο.

3.12.4 Βιβλιογραφικές Βιολογικές Βάσεις Δεδομένων

Η ραγδαία πρόοδος της βιολογικής και ιατρικής έρευνας οδήγησε στη δραματική αύξηση του όγκου της διαθέσιμης δημοσιευμένης πληροφορίας σε άρθρα, περιοδικά, βιβλία και πρακτικά συνεδρίων. Σήμερα, η διαδικτυακή βιβλιοθήκη PubMed παρέχει πρόσβαση σε δεκάδες εκατομμύρια αναφορές από τη Medline και άλλες συλλογές περιοδικών ανθρωπιστικών επιστημών. Συνολικά, στην PubMed υπάρχουν καταγεγραμμένα περισσότερα από 20 εκατομμύρια[11] άρθρα και μάλιστα κάθε χρόνο προστίθενται περίπου 500.000 νέες εγγραφές. Παρ' όλο που η PubMed αποτελεί τη σημαντι-

[11] Στις 28/2/2011 οι εγγραφές της PubMed ήταν περισσότερες από 20,7 εκατομμύρια.

κότερη και πλουσιότερη βάση δεδομένων άρθρων, περιλήψεων και αναφορών, υπάρχουν και άλλες σημαντικές βιβλιοθήκες, όπως το NCBI BookShelf, με ηλεκτρονικά βιβλία και μια πληθώρα διαδικτυακών πηγών επιστημονικών δημοσιεύσεων στη βιολογία.

3.12.5 Βιολογικές Βάσεις Δεδομένων Ιστοσελίδων

Οι βιολογικές βάσεις δεδομένων ιστοσελίδων περιλαμβάνουν:

- Βάσεις δεδομένων που περιέχουν ως εγγραφές βιολογικές βάσεις δεδομένων.
- Συνδέσμους μεταξύ των βιολογικών βάσεων δεδομένων.

3.13 Η Ανακάλυψη Γνώσης στη Βιοπληροφορική

Η διαδικασία της ανακάλυψης γνώσης σε βάσεις δεδομένων εφαρμόζεται στα πλαίσια του επιστημονικού πεδίου της βιοπληροφορικής για την ανάλυση διάφορων ειδών βιολογικών δεδομένων. Οι σημαντικότερες από αυτές τις εφαρμογές περιγράφονται στις παραγράφους που ακολουθούν.

3.13.1 Ανακάλυψη Γνώσης από Βιολογικές Αλληλουχίες

Οι σημαντικότερες διαδικασίες ανάλυσης αλληλουχιών που αντιμετωπίζονται με αλγορίθμους της ανακάλυψης γνώσης είναι οι εξής:

- Πρόβλεψη *ρυθμιστικών περιοχών* (regulatory regions), όπως υποκινητές (promoters) και ενισχυτές (enhancers), που αποτελούν τμήματα του DNA όπου προσδένονται ρυθμιστικές πρωτεΐνες με σκοπό να ελέγχουν την έκφραση των γονιδίων και κατά συνέπεια και την παραγωγή των πρωτεϊνών.
- Πρόβλεψη του σημείου έναρξης της μεταγραφής και του σημείου έναρξης της μετάφρασης.
- Πρόβλεψη των σημείων ματίσματος, για τον καθορισμό των εξονίων και των ιντρονίων. Το φαινόμενο του εναλλακτικού ματίσματος καθιστά το πρόβλημα ακόμη πιο πολύπλοκο.
- Πρόβλεψη των σημείων πολυαδενυλίωσης, όπου η πολύ(Α) ουρά προστίθεται στην αλληλουχία του mRNA. Το φαινόμενο της εναλλακτικής πολυαδενυλίωσης αυξάνει την πολυπλοκότητα του προβλήματος.
- Πρόβλεψη των περιοχών κωδικοποίησης σε τμήματα αλληλουχιών μεταγεγραμμένων ματισμένων μορίων mRNA, που ονομάζονται *ετικέτες εκφρασμένων αλληλουχιών* (*Expressed Sequence Tags – ESTs*).
- Σύγκριση μιας άγνωστης αλληλουχίας με γνωστές αλληλουχίες σε μια βάση δεδομένων, με στόχο την εύρεση πιθανών ομολογιών (π.χ. στενές εξελικτικές σχέσεις) και την ομαδοποίηση δομικά συγγενών αλληλουχιών.

Μέχρι σήμερα, έχουν χρησιμοποιηθεί αρκετές τεχνικές ανακάλυψης γνώσης και μηχανικής μάθησης για τα παραπάνω προβλήματα. Συνήθως, οι περισσότερες από αυτές τις τεχνικές τροποποιούνται και προσαρμόζονται ανάλογα με το πρόβλημα, ώστε να μπορούν να εφαρμοστούν με βέλτιστο τρόπο στις ιδιαίτερες απαιτήσεις καθενός από αυτά. Οι πλέον συνηθισμένες τεχνικές είναι τα νευρωνικά δίκτυα, οι ταξινομητές Bayes, τα δένδρα ταξινόμησης και οι μηχανές διανυσμάτων υποστήριξης [Ma & Wang, 1999; Hirsh & Noordewier, 1994; Zien et al., 2000].

3.13.2 Ανακάλυψη Γνώσης από Δεδομένα Γονιδιακής Έκφρασης

Το σημαντικότερο πρόβλημα της ανακάλυψης γνώσης από δεδομένα γονιδιακής έκφρασης είναι ότι περιλαμβάνουν έναν μικρό αριθμό παραδειγμάτων (συνήθως μικρότερο από 100) και έναν μεγάλο αριθμό χαρακτηριστικών (γονιδίων), που συνήθως είναι αρκετές χιλιάδες. Προκειμένου να αντιμετωπιστεί το πρόβλημα αυτό έχουν επιστρατευθεί αρκετές τεχνικές επιλογής χαρακτηριστικών, ώστε να μειωθεί δραστικά ο αριθμός των διαστάσεων των δεδομένων, επιλέγοντας έναν μικρό αριθμό γονιδίων [Xing *et al*., 2001].

Η πλέον δημοφιλής τεχνική στην ανάλυση γονιδιακής έκφρασης είναι η ομαδοποίηση [Tibshirani *et al.*, 1999]. Οι τεχνικές ομαδοποίησης χρησιμοποιούνται προκειμένου να ομαδοποιηθούν γονίδια με παρόμοια συμπεριφορά ή δείγματα με παρόμοια γονιδιακή έκφραση. Αλγόριθμοι ομαδοποίησης υποχώρων (στη βιβλιογραφία αναφέρονται και ως αλγόριθμοι *αμφι-ομαδοποίησης* - *bi-clustering*), χρησιμοποιούνται συχνά για την ταυτόχρονη ομαδοποίηση γονιδίων και δειγμάτων. Αρκετά δημοφιλής τεχνική ομαδοποίησης για την ανάλυση δεδομένων γονιδιακής έκφρασης είναι και η ιεραρχική ομαδοποίηση [Ng *et al*., 2001]. Ένα σημαντικό ζήτημα που ανακύπτει κατά τη ομαδοποίηση δεδομένων γονιδιακής έκφρασης είναι η αξιολόγηση της ποιότητας των ομάδων. Διάφορες τεχνικές έχουν προταθεί, όπως οι επαναλαμβανόμενες μετρήσεις, bootsrap καθώς και προσεγγίσεις βασισμένες σε μοντέλα μίξης (mixture model-based) και η υπο-δειγματοληψία [Kerr & Churchill, 2001; Yeung *et al*., 2003; Ghosh & Chinnaiyan, 2002; Smolkin & Ghosh, 2003].

3.13.3 Ανακάλυψη Γνώσης στη Δομική Βιοπληροφορική

Η *δομική βιοπληροφορική* (*structural bioinformatics*) είναι το υποπεδίο της βιοπληροφορικής που σχετίζεται με την ανάλυση και την πρόβλεψη της τρισδιάστατης δομής των βιολογικών μακρομορίων και κυρίως των πρωτεϊνών. Η εφαρμογή της μηχανικής μάθησης και της ανακάλυψης γνώσης στη δομική βιοπληροφορική έχει ιδιαίτερο ενδιαφέρον καθώς τα δομικά δεδομένα δεν είναι γραμμικά. Επιπλέον, ο χώρος αναζήτησης για τα περισσότερα από τα δομικά προβλήματα είναι συνεχής και άπειρος και απαιτεί ιδιαίτερα ισχυρούς ευριστικούς αλγορίθμους. Η πρόβλεψη διάφορων χαρακτηριστικών των πρωτεϊνών, όπως τα ενεργά κέντρα (active sites) των ενζύμων, τα σημεία τροποποίησης (modification sites), το σχήμα, τα αυτοτελή δομικά στοιχεία (domains), η

δευτεροταγής δομή και η αλληλεπίδραση με άλλες πρωτεΐνες [Whishart, 2002] είναι προβλήματα με ιδιαίτερο ενδιαφέρον.

Διάφορες μέθοδοι εξόρυξης από δεδομένα έχουν εφαρμοστεί και για την πρόβλεψη της δευτεροταγούς δομής των πρωτεϊνών. Το πρόβλημα αυτό μελετάται εδώ και αρκετές δεκαετίες, με αποτέλεσμα να έχουν αναπτυχθεί πολλές τεχνικές. Οι πρώτες προσεγγίσεις ήταν στατιστικές. Αργότερα, προτάθηκαν πιο ακριβείς τεχνικές, που βασίζονταν στη θεωρία πληροφορίας, τη θεωρία του Bayes, την ταξινόμηση κοντινότερου γείτονα και τα νευρωνικά δίκτυα. Ένας δημοφιλής τρόπος για τη βελτίωση της ακρίβειας πρόβλεψης είναι η χρήση συνδυασμένων μεθόδων, όπως οι ολοκληρωμένες πολλαπλές στοιχίσεις με μεθόδους ταξινόμησης.

Μερικά ακόμη σημαντικά προβλήματα της δομικής βιοπληροφορικής που αντιμετωπίζονται με μεθόδους ανακάλυψης γνώσης είναι η πρόβλεψη της δευτεροταγούς δομής του RNA, η διαδικασία εύρεσης της λειτουργίας μιας πρωτεΐνης από τη δομή της, η αναγνώριση της αλληλεπίδρασης των πρωτεϊνών μεταξύ τους και ο αποτελεσματικός σχεδιασμός φαρμάκων βάσει της δομικής πληροφορίας του στόχου τους.

3.13.4 Ανακάλυψη Γνώσης από Βιολογικά Κείμενα

Η εξόρυξη από κείμενα στη μοριακή βιολογία ορίζεται ως η αυτόματη εξαγωγή γνώσης σχετικής με τα γονίδια, τις πρωτεΐνες και τις λειτουργικές τους σχέσεις, από αρχεία κειμένου [Krallinger & Valencia, 2005]. Ο σχετικά νέος αυτός κλάδος έρευνας και εφαρμογής αντλεί στοιχεία από την πληροφορική, τη βιοπληροφορική και την υπολογιστική γλωσσολογία.

Για τους ερευνητές της βιολογίας και της βιοπληροφορικής είναι ιδιαίτερα σημαντικό να έχουν πρόσβαση στην πλέον πρόσφατη γνώση του πεδίου έρευνάς τους. Η σύγχρονη ερευνητική πρακτική περιλαμβάνει την αναζήτηση σε διαδικτυακές πηγές για πληροφορίες που έχουν να κάνουν με το εκάστοτε αντικείμενο έρευνας, χρησιμοποιώντας τις πιο πρόσφατες τεχνολογίες ανάκτησης πληροφοριών, του σημασιολογικού διαδικτύου (semantic web) και της εξόρυξης από κείμενα. Ένας νεολογισμός που χρησιμοποιείται τελευταία από τους επιστήμονες της βιοπληροφορικής όταν γίνεται αναφορά στα κείμενα εκείνα από τα οποία μπορούν να εξάγουν πληροφορία είναι ο αγγλικός όρος «textome», που θα μπορούσε να αποδοθεί από τον μάλλον αδόκιμο όρο «κειμένωμα». Γενικά, μπορεί να περιλαμβάνει όλα τα επιστημονικά κείμενα που μπορούν να υποστούν υπολογιστική επεξεργασία και συμβολομετάφραση (parsing).

Η ανάγκη των ερευνητών να εκμεταλλευθούν όλη αυτή τη διαθέσιμη πληροφορία σε συνδυασμό με τη δυνατότητα χρήσης των αποτελεσματικών εργαλείων που προσφέρουν οι περιοχές της ανακάλυψης γνώσης, της επεξεργασίας φυσικής γλώσσας, της ανάκτησης πληροφορίας και του σημασιολογικού διαδικτύου, οδήγησε στη δημιουργία ενός νέου πεδίου έρευνας και εφαρμογής, της *εξόρυξης από βιολογικά κείμενα* (*biological text mining*). Στη βιβλιογραφία απαντώνται και οι όροι *εξόρυξη από βιοϊα-*

τρικά κείμενα (*biomedical text mining*) και *εξόρυξη από κείμενα βιοπληροφορικής* (*bioinformatics text mining*).

Από τη σκοπιά της ανακάλυψης γνώσης από βάσεις δεδομένων, η βιοϊατρική βιβλιογραφία διαθέτει κάποια χαρακτηριστικά που απαιτούν ιδιαίτερη προσοχή, όπως η χρήση συγκεκριμένης ορολογίας, η πολυσημία των λέξεων, η χαμηλή συχνότητα λέξεων (αραιά δεδομένα), η δημιουργία νέων ονομάτων και όρων καθώς επίσης και διαφορετικά στυλ γραφής.

3.14 Ερευνητικά Κέντρα Βιοπληροφορικής

Το ενδιαφέρον της επιστημονικής κοινότητας για τη βιοπληροφορική είναι πολύ μεγάλο. Ήδη υπάρχουν αρκετοί μεγάλοι οργανισμοί που ασχολούνται με την παροχή υπηρεσιών και την έρευνα στον τομέα της βιοπληροφορικής. Ένας από αυτούς τους οργανισμούς είναι το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (European Bioinformatics Institute – EBI), το οποίο εδρεύει στο Χίνκστον του Ηνωμένου Βασιλείου. Το EBI είναι ένας μη κερδοσκοπικός ακαδημαϊκός οργανισμός και αποτελεί μέρος του Ευρωπαϊκού Εργαστηρίου Μοριακής Βιολογίας (European Molecular Biology Laboratory – EMBL), που εδρεύει στη Χαϊδελβέργη της Γερμανίας. Το EBI είναι ένα κέντρο έρευνας και υπηρεσιών για τη βιοπληροφορική. Διαχειρίζεται βάσεις βιολογικών δεδομένων. Σκοπός του EBI είναι η εξασφάλιση της δημόσιας και ελεύθερης προσπέλασης των διαρκώς αυξανόμενων πληροφοριών της μοριακής βιολογίας και της έρευνας των γονιδιωμάτων με τρόπους που να προωθούν την επιστημονική πρόοδο.

Ένας άλλος παρόμοιος οργανισμός είναι το Εθνικό Κέντρο Πληροφοριών Βιοτεχνολογίας (National Center for Biotechnology Information – NCBI) των Η.Π.Α. Το NCBI ιδρύθηκε το 1988 ως τμήμα της Εθνικής Βιβλιοθήκης Ιατρικής (National Library of Medicine – NLM) του Εθνικού Ινστιτούτου Υγείας (National Institutes of Health – NIH), το οποίο εδρεύει στη Βηθεσδά του Μέριλαντ των Η.Π.Α. Σκοπός του NCBI είναι η ανάπτυξη νέων τεχνολογιών, που θα βοηθήσουν στην κατανόηση των θεμελιωδών μοριακών και γενετικών διεργασιών, οι οποίες ελέγχουν την υγεία και την ασθένεια. Στα πλαίσια αυτής της προσπάθειας το NCBI δημιουργεί δημόσιες βάσεις δεδομένων, πραγματοποιεί έρευνες στον τομέα της βιοπληροφορικής, αναπτύσσει εργαλεία λογισμικού για την ανάλυση βιολογικών δεδομένων και διαδίδει τη βιοϊατρική πληροφορία.

3.15 Εργαλεία Βιοπληροφορικής

Τα εργαλεία που χρησιμοποιούνται στη βιοπληροφορική είναι προγράμματα λογισμικού σχεδιασμένα για την εξαγωγή πολύτιμης πληροφορίας από την πληθώρα των βιολογικών δεδομένων. Υπάρχουν εργαλεία γενικής χρήσης και εργαλεία που είναι προσαρμοσμένα στις απαιτήσεις μιας συγκεκριμένης εργασίας. Τα εργαλεία αυτά μπορούν να ταξινομηθούν στις παρακάτω κατηγορίες:

- *Εργαλεία ομολογίας (homology) και ομοιότητας (similarity).* Στην κατηγορία αυτή ανήκουν τα εργαλεία που χρησιμοποιούνται για την αναγνώριση ομοιοτήτων μεταξύ νέων αλληλουχιών άγνωστης δομής και λειτουργίας και αλληλουχιών με γνωστή δομή και λειτουργία. Οι νέες αλληλουχίες υποβάλλονται ως ερωτήματα προς τις βάσεις δεδομένων στις οποίες βρίσκονται αποθηκευμένες οι γνωστές.
- *Εργαλεία ανάλυσης της λειτουργίας πρωτεϊνών.* Τα εργαλεία αυτής της κατηγορίας επιτρέπουν τη σύγκριση πρωτεϊνικών αλληλουχιών με πληροφορίες που περιέχονται σε δευτεροταγείς κυρίως βασεις βιολογικών δεδομένων. Σε περίπτωση που το ποσοστό επιτυχίας της αναζήτησης είναι αρκετά υψηλό, υπάρχει η δυνατότητα προσέγγισης της βιοχημικής λειτουργίας της άγνωστης πρωτεΐνης.
- *Εργαλεία ανάλυσης δομών.* Η λειτουργία μιας πρωτεΐνης μπορεί να προσδιοριστεί πιο άμεσα από τη δισδιάστατη ή τρισδιάστατη δομή της παρά από την αλληλουχία των αμινοξέων που την αποτελούν. Με τα εργαλεία της κατηγορίας αυτής δίνεται η δυνατότητα σύγκρισης άγνωστων πρωτεϊνικών δομών με γνωστές δομές που βρίσκονται σε βάσεις δεδομένων.
- *Εργαλεία ανάλυσης αλληλουχιών.* Τα εργαλεία αυτά επιτρέπουν την λεπτομερή ανάλυση των βιολογικών αλληλουχιών, όπως για παράδειγμα την εξελικτική ανάλυση και την αναγνώριση μεταλλάξεων.
- *Διάφορα εργαλεία.* Στην κατηγορία αυτή ανήκουν εργαλεία που δεν ανήκουν σε κάποια από τις προηγούμενες κατηγορίες. Τέτοια μπορεί να είναι τα εργαλεία ανάλυσης γονιδιακής έκφρασης.

Μερικά από τα πιο δημοφιλή εργαλεία περιγράφονται παρακάτω:

- Ο αλγόριθμος *BLAST* (*Basic Local Alignment Search Tool*) [Altschul *et al.*, 1990] ανήκει στην κατηγορία των εργαλείων ομολογίας και ομοιότητας. Είναι σχεδιασμένος έτσι ώστε να δέχεται ως είσοδο μια πρωτεϊνική ή νουκλεοτιδική αλληλουχία, την οποία συγκρίνει με όλες τις αλληλουχίες που βρίσκονται σε μια μεγάλη βάση δεδομένων. Η σύγκριση γίνεται με τη στοίχιση της νέας αλληλουχίας με τις αλληλουχίες γνωστών γονιδίων και στην έξοδο εμφανίζονται οι αλληλουχίες που παρουσιάζουν τη μεγαλύτερη ομοιότητα.
- Ο αλγόριθμος *FASTA* (*FAST All*), όπως και ο BLAST, ανήκει στην πρώτη κατηγορία των εργαλείων. Υλοποιήθηκε από τους Lipman & Pearson [1985] και βελτιώθηκε αργότερα από τους ίδιους [Lipman & Pearson, 1988]. Όπως δηλώνει το όνομά του, μπορεί να χρησιμοποιηθεί για γρήγορες συγκρίσεις πρωτεϊνικών ή νουκλεοτιδικών αλληλουχιών. Το εργαλείο αυτό επιτυγχάνει υψηλά επίπεδα ευαισθησίας για αναζητήσεις ομοιότητας με μεγάλη ταχύτητα.
- Το *EMBOSS* (*European Molecular Biology Open Software Suite*) είναι ένα υψηλής ποιότητας, ανοιχτού κώδικα, πακέτο λογισμικού. Παρέχει ένα σύνολο από πακέ-

τα και εργαλεία ανάλυσης αλληλουχιών, κάτω από ένα ολοκληρωμένο περιβάλλον [Rice *et al*., 2000]. Μπορεί να δουλέψει με διάφορες μορφές δεδομένων και να ανακτήσει αλληλουχίες από το Διαδίκτυο με διαφανή τρόπο. Με το πακέτο παρέχονται βιβλιοθήκες επέκτασης που επιτρέπουν στους επιστήμονες να εκδίδουν το λογισμικό τους ως λογισμικό ανοιχτού κώδικα.

- Το *Clustal* [Chenna et al., 2003] είναι ένα εργαλείο που χρησιμοποιείται ευρέως για την πολλαπλή στοίχιση πρωτεϊνικών ή νουκλεοτιδικών αλληλουχιών με σκοπό την αποκάλυψη της συγγένειας και της εξελικτικής τους προέλευσης. Υπάρχουν δυο βασικές μορφές του εργαλείου, η *ClustalW*, που παρέχει μια απλή διεπαφή γραμμής εντολών και η *ClustalX*, που παρέχει ένα γραφικό περιβάλλον.
- Το *RasMol* είναι ένα ελεύθερο πρόγραμμα λογισμικού, που αναπτύχθηκε από τον R. A. Sayle (Πανεπιστήμιο του Εδιμβούργου και Glaxo R&D). Το εργαλείο αυτό χρησιμοποιείται για την τρισδιάστατη αναπαράσταση των δομών του DNA, των πρωτεϊνών και άλλων μικρότερων μορίων. Ένα πιο εύχρηστο πρόγραμμα, που προέρχεται από το RasMol είναι το *Protein Explorer*.

3.16 Εφαρμογές και Τάσεις της Βιοπληροφορικής

Λόγω των ιδιαίτερων χαρακτηριστικών των βιολογικών δεδομένων, την ποικιλία των νέων προβλημάτων και την ιδιαίτερη σημασία της βιολογικής έρευνας, υπάρχει ένας μεγάλος αριθμός κρίσιμων προβλημάτων που αναζητούν λύση. Οι νέες τεχνολογίες επιτρέπουν την ταχύτερη και οικονομικότερη εκτέλεση πειραμάτων μεγάλης κλίμακας με αποτέλεσμα να δημιουργούνται συνεχώς νέα ερωτήματα για νέα δεδομένα. Οι εφαρμογές της βιοπληροφορικής είναι πολλές. Οι σημαντικότερες από αυτές παρουσιάζονται παρακάτω.

Η αποκρυπτογράφηση του ανθρώπινου γονιδιώματος ανοίγει το δρόμο στη *βιοϊατρική έρευνα* και στην *κλινική ιατρική*. Η γνώση των μοριακών μηχανισμών των ασθενειών παρέχει τη δυνατότητα καλύτερων θεραπειών και προληπτικών ελέγχων. Καθίσταται δυνατή η ανάπτυξη διαγνωστικών ελέγχων για τη μέτρηση της ευαισθησίας ενός ατόμου σε διάφορες ασθένειες. Έτσι, με την εφαρμογή μεθόδων της προληπτικής ιατρικής, όπως η αλλαγή του τρόπου ζωής ή η αντιμετώπιση της ασθένειας στα πλέον αρχικά στάδια, μπορούν να προαχθούν σημαντικά οι προσπάθειες καταπολέμησης των ασθενειών. Η πρόοδος στο πεδίο της φαρμακογενωμικής δημιουργεί νέες προκλήσεις στα πλαίσια της εφαρμογής της *εξατομικευμένης ιατρικής*. Κάποια φάρμακα αποτυγχάνουν και πολλές φορές αποσύρονται από την αγορά, καθώς ευθύνονται για την εμφάνιση σοβαρών παρενεργειών σε ασθενείς που παρουσιάζουν συγκεκριμένες παραλλαγές στην αλληλουχία του γενετικού τους υλικού. Ωστόσο, τα ίδια φάρμακα μπορούν να σώσουν ζωές σε κάποιες άλλες περιπτώσεις. Η μελέτη του γενετικού προφίλ του ασθενούς είναι δυνατό να καθοδηγήσει το γιατρό στη λήψη αποφάσεων σχετικά με τη χρήση της καλύτερης φαρμακευτικής αγωγής. Η παρασκευή φαρμάκων που θα επι-

δρούν πάνω στην αιτία και όχι στα συμπτώματα της ασθένειας υπόσχεται τη μείωση των παρενεργειών που έχουν πολλές από τις κλασικές φαρμακευτικές αγωγές. Μια άλλη σημαντική εφαρμογή είναι η *γονιδιακή θεραπεία*, η οποία αποτελεί την προσέγγιση που χρησιμοποιείται για την θεραπεία ή πρόληψη ασθενειών με τη μεταβολή της έκφρασης των γονιδίων ενός ατόμου.

Μια άλλη εφαρμογή της βιοπληροφορικής είναι η *οικολογία*. Για παράδειγμα, το Τμήμα Ενέργειας (Department of Energy – DOE) των Η.Π.Α. διεξάγει προγράμματα για την ελάττωση των επιπέδων του διοξειδίου του άνθρακα στην ατμόσφαιρα. Μια μέθοδος που εφαρμόζεται είναι η μελέτη του γονιδιώματος μικροοργανισμών που χρησιμοποιούν το συγκεκριμένο ρύπο ως μοναδική πηγή άνθρακα. Οι επιστήμονες φιλοδοξούν στην εύρεση εναλλακτικών πηγών ενέργειας μελετώντας το γονιδίωμα του βακτηρίου Chlorobium tepidum, το οποίο έχει την ασυνήθιστη δυνατότητα να παράγει ενέργεια από το φως. Μια ακόμη εφαρμογή της βιοπληροφορικής μπορεί να αποτελέσει η *κατασκευή βιολογικών όπλων*. Πρόσφατα μάλιστα οι επιστήμονες δημιούργησαν τον ιό της πολιομυελίτιδας (poliomyelitis), χρησιμοποιώντας εξ' ολοκλήρου τεχνητά μέσα. Η χαρτογράφηση των γονιδιωμάτων επιτρέπει τη μελέτη της εξέλιξης των οργανισμών με σκοπό την κατασκευή του δένδρου της ζωής και την εύρεση του τελευταίου καθολικού κοινού προγόνου.

Οι πληροφορίες που προκύπτουν από τη μελέτη φυτικών οργανισμών-μοντέλων (π.χ. το Arabidopsis thaliana) μπορούν να χρησιμοποιηθούν για να προτείνουν βελτιώσεις σε άλλες φυτικές καλλιέργειες. Η *βελτίωση της θρεπτικής αξίας* ενός φυτικού είδους μπορεί να επιτευχθεί με την εισαγωγή γονιδίων στο συγκεκριμένο είδος. Για παράδειγμα, οι επιστήμονες έχουν καταφέρει να εισάγουν γονίδια στο ρύζι με σκοπό την αύξηση των επιπέδων της βιταμίνης Α, του σιδήρου και άλλων ιχνοστοιχείων. Σκοπός αυτής της προσπάθειας είναι η μείωση των περιπτώσεων τύφλωσης και αναιμίας που οφείλονται σε έλλειψη της βιταμίνης Α και του σιδήρου αντίστοιχα.

Ένας άλλος τομέας που ωφελείται από τις προόδους της βιοπληροφορικής είναι η *γεωργία*. Στόχος είναι η δημιουργία φυτικών ποικιλιών που παρουσιάζουν ανθεκτικότητα σε περιβαλλοντικές καταστάσεις όπως η ξηρασία, η αλκαλικότητα του χώματος, η έλλειψη νερού. Παρόμοιες εξελίξεις υπάρχουν και στον τομέα της *κτηνοτροφίας*. Η χαρτογράφηση και η ανάλυση των γονιδιωμάτων οικόσιτων ζώων μπορεί να οδηγήσει στη βελτίωση της παραγωγής και της υγείας αυτών των ειδών, παρέχοντας σημαντικά πλεονεκτήματα στη διατροφή του ανθρώπου.

Η ανάλυση και η σύγκριση του γενετικού υλικού διαφορετικών οργανισμών είναι μια σημαντική μέθοδος μελέτης της λειτουργίας των γονιδίων, των μηχανισμών που περιγράφουν τις κληρονομικές ασθένειες και της εξέλιξης των ειδών. Οργανισμοί που είναι κατάλληλοι για χρήση στην πειραματική έρευνα καλούνται *οργανισμοί-μοντέλα* (*model organisms*). Οι οργανισμοί αυτοί έχουν ένα σύνολο ιδιοτήτων που τους καθιστά ιδανικούς για ερευνητικούς σκοπούς. Τέτοιες ιδιότητες είναι η μικρή διάρκεια ζωής, η ταχεία αναπαραγωγή, η ευκολία διαχείρισης, το χαμηλό κόστος και η δυνατότητα

χειρισμού σε γενετικό επίπεδο. Ένα παράδειγμα οργανισμού-μοντέλου για τον άνθρωπο είναι το ποντίκι. Ο χειρισμός του ποντικιού σε γενετικό επίπεδο και οι συγκρίσεις των γονιδιωμάτων μεταξύ των δύο ειδών μπορούν να αποκαλύψουν λεπτομερείς πληροφορίες σχετικά με τις λειτουργίες των γονιδίων του ανθρώπου, την εξελικτική σχέση των δύο οργανισμών και τους μοριακούς μηχανισμούς πολλών ανθρώπινων ασθενειών.

Η έρευνα στη μηχανική μάθηση και την ανακάλυψη γνώσης από βιολογικά δεδομένα διαρκώς παράγει νέα εργαλεία και βελτιώνει τα υπάρχοντα, προκειμένου να αντιμετωπίσει τα διάφορα προβλήματα που τίθενται από τη βιολογική έρευνα με τρόπο αποτελεσματικό και αποδοτικό. Όπως αναφέρουν οι Houle *et al*. [2000], οι βελτιώσεις αυτές θα επιτρέψουν την πρόβλεψη της λειτουργίας των πρωτεϊνών στο πλαίσιο διαδικασιών υψηλότερης τάξης, όπως η ρύθμιση της γονιδιακής έκφρασης, οι μεταβολικές οδοί και οι σηματοδοτικές ακολουθίες (signaling cascades), που είναι αλληλουχίες αντιδράσεων που λαμβάνουν χώρα ως αποτέλεσμα μιας μοναδικής διέγερσης. Ο τελικός σκοπός της έρευνας που επιτελείται στα πλαίσια της βιολογίας και της βιοπληροφορικής είναι να ανακαλυφθεί ο τρόπος μετάβασης από το γονότυπο στο φαινότυπο και οι μοριακές και κυτταρικές λεπτομέρειες που διέπουν τη ζωή.

3.17 Σύνοψη

Στο κεφάλαιο αυτό παρουσιάστηκαν τα επιστημονικά πεδία της μοριακής βιολογίας και της βιοπληροφορικής. Περιγράφηκαν κάποιες βασικές έννοιες που συναντώνται στη μοριακή βιολογία, οι σημαντικότερες τεχνολογίες που έχουν αναπτυχθεί και εφαρμόζονται στα πλαίσια της βιοπληροφορικής, διάφορες κατηγορίες βιολογικών βάσεων δεδομένων, καθώς και οι δημοφιλέστερες εφαρμογές της διαδικασίας ανακάλυψης γνώσης σε βιολογικά δεδομένα.

Ανακάλυψη Αμοιβαίως Αποκλειόμενων Αντικειμένων

4

4.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζεται ένα νέο είδος προβλήματος στα πλαίσια της περιοχής της ανακάλυψης γνώσης από βάσεις δεδομένων, το οποίο περιλαμβάνει την ανακάλυψη αμοιβαίως αποκλειόμενων αντικειμένων (mutually exclusive items). Τα αμοιβαίως αποκλειόμενα αντικείμενα αποτελούν ένα ενδιαφέρον είδος συσχετίσεων που είναι δυνατό να οδηγήσουν στην ανακάλυψη ιεραρχικών σχέσεων. Ορίζεται το πρόβλημα της ανακάλυψης αμοιβαίως αποκλειόμενων αντικειμένων και προτείνεται ένας αλγόριθμος ανακάλυψης των αμοιβαίως αποκλειόμενων αντικειμένων από μια βάση δεδομένων συναλλαγών.

Οι κανόνες συσχέτισης (Εδάφιο 2.3.2) αποτελούν εκφράσεις που περιγράφουν ένα υποσύνολο μιας βάσης δεδομένων συναλλαγών. Συχνά, κατά την εξόρυξή τους το αποτέλεσμα είναι ένα μεγάλο πλήθος κανόνων, πολλοί από τους οποίους είναι υπερβολικά εξειδικευμένοι και όχι ιδιαίτερα ενδιαφέροντες. Ένας κανόνας που συσχετίζει δύο συγκεκριμένα αντικείμενα (π.χ. προϊόντα) σε μια βάση δεδομένων είναι δύσκολο να είναι αρκετά ισχυρός συγκριτικά με έναν κανόνα που συσχετίζει δυο οικογένειες αντικειμένων. Σε αρκετές εφαρμογές τα διάφορα αντικείμενα διέπονται από ιεραρχικές σχέσεις, οι οποίες είναι δυνατόν να χρησιμοποιηθούν, ώστε να ομαδοποιηθεί ένα μέρος των εξειδικευμένων και αδύναμων κανόνων σε γενικότερους και ισχυρότερους. Το αποτέλεσμα αυτής της διαδικασίας είναι ιεραρχικοί (hierarchical) ή πολυεπίπεδοι (multiple level) ή γενικευμένοι (generalized) κανόνες συσχέτισης. Ωστόσο, η πληροφορία αυτή δεν είναι πάντοτε διαθέσιμη, παρότι μπορεί να υφίσταται.

Οι *εννοιολογικές ιεραρχίες* (*concept hierarchies*) ονομάζονται *ταξονομίες* (*taxonomies*) και υλοποιούνται με σχέσεις «είναι ένα» (“is-a” relationships). Η ανακάλυψή τους είναι ένα δύσκολο πρόβλημα με αμφίβολα αποτελέσματα, το οποίο δεν έχει

προσεγγισθεί μέχρι τώρα με κάποιον τρόπο. Η ανακάλυψη τέτοιων σχέσεων θα προϋπέθετε την πλήρη περιγραφή και προσδιορισμό των εννοιών που είναι γενικότερες από κάποιες άλλες σε μια βάση δεδομένων. Η ανακάλυψη ταξονομιών αναφέρεται στην αυτόματη ανακάλυψη γενικευμένων εννοιών από τα δεδομένα, με τη χρήση ενός αλγορίθμου, κάτι που δεν πραγματοποιείται με τις υπάρχουσες μεθόδους. Ωστόσο, κατά την εξόρυξη διαφόρων τύπων συσχετίσεων, μπορεί κανείς να εξάγει μερικές σοβαρές ενδείξεις για την ύπαρξη πιθανών ιεραρχικών σχέσεων.

Ένα ενδιαφέρον είδος συσχετίσεων που θα μπορούσε να οδηγήσει στην ανακάλυψη ταξονομικών σχέσεων είναι τα *αμοιβαίως αποκλειόμενα αντικείμενα* (*mutually exclusive items*). Έστω, για παράδειγμα, ότι ο πελάτης ενός καταστήματος που πουλά γαλακτοκομικά ενδιαφέρεται να αγοράσει γάλα. Στην περίπτωση αυτή είναι μάλλον απίθανο να αγοράσει ταυτόχρονα πλήρες γάλα και άπαχο γάλα. Αυτά τα δύο προϊόντα ικανοποιούν ακριβώς την ίδια ανάγκη, την κατανάλωση γάλακτος, αλλά διαφέρουν ως προς την περιεκτικότητά τους σε λιπαρά. Συνεπώς, όποιος ενδιαφέρεται να καταναλώσει γάλα θα επιλέξει ένα από τα δύο προϊόντα. Ως εκ τούτου, το πλήρες και το άπαχο γάλα είναι δύο αμοιβαίως αποκλειόμενα προϊόντα, καθώς η κατανάλωση και κατ' επέκταση η αγορά του ενός αποκλείει την αγορά του άλλου. Παρατηρήσεις όπως αυτή ενέπνευσαν την ανάπτυξη μιας μεθόδου για την εξόρυξη αμοιβαίως αποκλειόμενων αντικειμένων. Η πληροφορία αυτή θα μπορούσε να είναι ενδιαφέρουσα, ανεξάρτητα από το αν θα χρησιμοποιηθεί τελικά για την ανακάλυψη ταξονομιών.

4.2 Βασικές Έννοιες

Η εξόρυξη κανόνων συσχέτισης (Εδάφιο 2.3.2) περιλαμβάνει την αναζήτηση κοινών εμφανίσεων αντικειμένων σε βάσεις δεδομένων συναλλαγών. Αυτές οι κοινές εμφανίσεις ενδέχεται να υπονοούν ή να οφείλονται σε κάποια ιδιαίτερη σχέση ανάμεσα στα αντικείμενα που περιλαμβάνουν και αν αναλυθούν περαιτέρω, μπορεί να αποκαλύψουν σχέσεις διαφόρων ειδών, όπως χρονική ή αιτιακή. Ο ορισμός του προβλήματος της εξόρυξης κανόνων συσχέτισης δόθηκε από τους [Agrawal et al., 1993]. Έχοντας μια βάση δεδομένων συναλλαγών *D*, το ζητούμενο είναι η εξόρυξη όλων των κανόνων συσχέτισης που έχουν υποστήριξη (support) και εμπιστοσύνη (confidence) ίση ή μεγαλύτερη από τα αντίστοιχα κατώτατα όρια *min_sup* και *min_conf* που έχει θέσει ο χρήστης. Το πρόβλημα αυτό μπορεί να διαιρεθεί σε δύο υπο-προβλήματα:

1. Στην ανακάλυψη όλων των συνόλων αντικειμένων που έχουν υποστήριξη ίση ή μεγαλύτερη από *min_sup*. Αυτά τα σύνολα ονομάζονται *συχνά.*

2. Στη δημιουργία όλων των κανόνων από τα συχνά σύνολα αντικειμένων. Για κάθε συχνό σύνολο αντικειμένων *F*, δημιουργούνται όλα τα μη κενά υποσύνολα του *F*. Για κάθε $S \subseteq F$ δημιουργείται ένας κανόνας $S \Rightarrow F\text{-}S$, εφόσον η εμπιστοσύνη του είναι ίση ή μεγαλύτερη από *min_conf.*

Ένας άλλος τρόπος για την εξαγωγή ισχυρών κανόνων είναι η χρήση των ταξονομιών που εμπεριέχονται σε ορισμένα σύνολα δεδομένων. Μια ταξονομία είναι ένα δένδρο εννοιών, όπου οι ακμές αναπαριστούν μια σχέση «είναι ένα», από τον κόμβο-παιδί προς τον κόμβο-γονέα. Ένα τέτοιο παράδειγμα είναι το εξής: «Το Cheddar *είναι ένα* Τυρί *είναι ένα* Γαλακτοκομικό *είναι ένα* Τρόφιμο *είναι ένα* Προϊόν». Στην περίπτωση κατά την οποία μια ταξονομία ενός πεδίου εφαρμογής είναι διαθέσιμη, ένας αριθμός κανόνων υψηλής εμπιστοσύνης αλλά χαμηλής υποστήριξης είναι δυνατόν να συγχωνευθούν, δημιουργώντας έναν κανόνα που θα συγκεντρώνει την υποστήριξη των επιμέρους κανόνων σε ένα υψηλότερο αφαιρετικό επίπεδο. Με άλλα λόγια, είναι δυνατόν να συνδυαστεί πληροφορία από αδύναμους κανόνες που συσχετίζουν αντικείμενα στα χαμηλότερα επίπεδα μιας ταξονομίας και να δημιουργηθούν ισχυρές συσχετίσεις στα υψηλότερα επίπεδα. Έτσι, για παράδειγμα, ο κανόνας «αν ένας πελάτης αγοράσει ½ κιλό ψωμί ολικής αλέσεως, τότε θα αγοράσει και 1 λίτρο φρέσκο παστεριωμένο γάλα» είναι πιθανό να έχει χαμηλή υποστήριξη, ενώ ένας κανόνας όπως ο «αν ένας πελάτης αγοράσει ψωμί, τότε θα αγοράσει και γάλα» είναι πιθανότερο να έχει υψηλότερη υποστήριξη, αφού θα συγκεντρώνει πληροφορία για όλα τα είδη, συσκευασίες και μάρκες ψωμιού και γάλακτος που αγοράζουν οι πελάτες του καταστήματος.

4.3 Σχετικές Εργασίες

Οι κανόνες συσχέτισης προτάθηκαν αρχικά από τους Agrawal et al. [1993] ώς ένα εργαλείο *ανάλυσης καλαθιού αγορών* (*market basket analysis*). Αργότερα, οι Agrawal & Srikant [1994] πρότειναν τον Apriori, έναν αλγόριθμο που εκτελείται σε επίπεδα και βασίζεται στη δημιουργία υποψήφιων συχνών συνόλων αντικειμένων, των οποίων η συχνότητα κάθε φορά μετράται με ένα πέρασμα στη βάση δεδομένων. Ο Apriori εκμεταλλεύεται το γεγονός ότι όλα τα μη-κενά υποσύνολα ενός συχνού συνόλου αντικειμένων θα είναι και εκείνα συχνά, βάσει του ίδιου κατώτατου κριτηρίου υποστήριξης. Έτσι, σε κάθε επίπεδο, τα υποψήφια συχνά σύνολα δημιουργούνται από την ένωση μόνο των συχνών συνόλων αντικειμένων του προηγούμενου επιπέδου. Την ίδια περίοδο οι Mannila et al. [1994] εκμεταλλευόμενοι την ίδια ιδιότητα, πρότειναν μια παραλλαγή του Apriori, τον αλγόριθμο OCD. Αργότερα, οι συγγραφείς των δύο αλγορίθμων δημοσίευσαν από κοινού ένα άρθρο που συνδυάζει τις εργασίες και των δύο [Agrawal et al., 1996]. Στη συνέχεια, προτάθηκαν διάφοροι αλγόριθμοι ανακάλυψης κανόνων συσχέτισης, άλλοτε με στόχο τη βελτίωση της απόδοσης, όπως ο FP-Growth [Han et al., 2000] και άλλοτε με στόχο την αντιμετώπιση ειδικών προβλημάτων και εφαρμογών, όπως χωρικοί κανόνες συσχέτισης [Koperski & Han, 1995], χρονικοί [Chen & Petrounias, 2000] και διασυναλλακτικοί κανόνες συσχέτισης [Tung, et al., 2003].

Ένα από τα σημαντικότερα προβλήματα στην εξόρυξη κανόνων συσχέτισης είναι ο μεγάλος αριθμός κανόνων, πολλοί από τους οποίους μάλιστα μπορεί να μην είναι και ενδιαφέροντες. Οι Srikant & Agrawal το 1995 πρότειναν για πρώτη φορά την ιδέα της

εξόρυξης *γενικευμένων κανόνων συσχέτισης* (generalized association rules). Έτσι, για παράδειγμα, γνωρίζοντας ότι «το κασέρι είναι ένα τυρί» και «το Edam είναι ένα τυρί», είναι δυνατό να εξαχθούν κανόνες όπως «αν κάποιος πελάτης αγοράσει ψωμί, τότε θα αγοράσει και *τυρί*», με υποστήριξη μεγαλύτερη από έναν κανόνα που θα αφορούσε ένα συγκεκριμένο είδος τυριού (π.χ. Edam). Στην ίδια εργασία, οι συγγραφείς αρχικά προτείνουν έναν βασικό αλγόριθμο και στη συνέχεια τρεις βελτιωμένους αλγορίθμους που βασίζονται σε ένα νέο μέτρο ενδιαφέροντος κανόνων, που χρησιμοποιεί πληροφορία από την ταξονομία. Οι Thomas & Sarawagi [1998] προτείνουν μια τεχνική που στηρίζεται σε ερωτήματα SQL για την εξόρυξη γενικευμένων κανόνων συσχέτισης. Οι Han & Fu [1995] περιγράφουν το πρόβλημα της εξόρυξης κανόνων συσχέτισης *πολλαπλών επιπέδων*, βασιζόμενοι και εκείνοι σε ταξονομίες και προτείνουν ένα σύνολο αλγορίθμων βαθμιαίας εμβάθυνσης (progressive deepening).

Η Teng [2002] προτείνει έναν τύπο ενισχυμένων συσχετίσεων, οι οποίες χρησιμοποιούν και αρνητική πληροφορία και ονομάζονται αντιδιαστολές (dissociations). Μια αντιδιαστολή είναι μια σχέση της μορφής «*X* δε συνεπάγεται *Y*», αλλά μπορεί να ισχύει «όταν το *X* εμφανίζεται μαζί με το *Z*, αυτό συνεπάγεται *Y*». Ένα άλλο είδος κανόνων συσχέτισης είναι οι αρνητικοί κανόνες συσχέτισης (negative association rules). Οι Savasere et al. [1998] εισήγαγαν το πρόβλημα της εξόρυξης αρνητικών συσχετίσεων, δηλαδή κανόνων που υποδηλώνουν ποια αντικείμενα δεν είναι πιθανό να αγοραστούν, όταν αγοράζεται ένα συγκεκριμένο σύνολο αντικειμένων. Η προσέγγιση των Savasere et al. προϋποθέτει την ύπαρξη ταξονομίας και βασίζεται στην υπόθεση ότι τα αντικείμενα που ανήκουν στον ίδιο κόμβο-γονέα στο δένδρο της ιεραρχίας, είναι αναμενόμενο να έχουν και παρόμοιες συσχετίσεις. Εφόσον λαμβάνονται υπόψη μόνο εκείνες οι περιπτώσεις που η υποστήριξη μπορεί να υπολογιστεί βάσει ταξονομίας, μόνο ένα υποσύνολο των αρνητικών κανόνων μπορεί να ανακαλυφθεί. Στη συγκεκριμένη εργασία προτείνουν έναν αφελή και έναν βελτιωμένο αλγόριθμο εξόρυξης, καθώς και ένα νέο μέτρο ενδιαφέροντος. Πιο πρόσφατα, οι Wu et al. [2004] πρότειναν μια αποτελεσματική μέθοδο για την εξόρυξη θετικών και αρνητικών συσχετίσεων και προτείνουν μια στρατηγική κλαδέματος και ένα μέτρο ενδιαφέροντος. Η μέθοδός τους επεκτείνει τους παραδοσιακούς κανόνες συσχέτισης ($A \Rightarrow B$) ώστε να περιλαμβάνει και κανόνες της μορφής $A \Rightarrow -B$, $-A \Rightarrow B$, και $-A \Rightarrow -B$. Οι τρεις τελευταίοι κανόνες υποδηλώνουν αρνητική συσχέτιση ανάμεσα στα σύνολα αντικειμένων *A* και *B*. Ωστόσο, ένας τέτοιος κανόνας δε μπορεί να υποδηλώσει αμοιβαίο αποκλεισμό δύο αντικειμένων. Αν τα αντικείμενα *α* και *β* είναι αμοιβαίως αποκλειόμενα, τότε ισχύει ταυτόχρονα $\{\alpha\} \Rightarrow -\{\beta\}$ και $\{\beta\} \Rightarrow -\{\alpha\}$, κάτι που είναι διαφορετικό από το $-\{\alpha\} \Rightarrow -\{\beta\}$.

Οι Han & Fu [1994] προτείνουν αλγορίθμους για τη δυναμική δημιουργία και βελτίωση εννοιολογικών ιεραρχιών. Η δημιουργία των εννοιολογικών ιεραρχιών αφορά μόνο αριθμητικά χαρακτηριστικά και βασίζεται στην κατανομή των δεδομένων. Η δυναμική βελτίωση μιας ιεραρχίας στηρίζεται σε ένα συγκεκριμένο αίτημα μάθησης, τα σχετικά μόνο δεδομένα και ορισμένα στατιστικά της βάσης δεδομένων.

4.4 Εξόρυξη Αμοιβαίως Αποκλειόμενων Αντικειμένων

Στην ενότητα αυτή, ορίζεται το πρόβλημα την ανακάλυψης αμοιβαίως αποκλειόμενων αντικειμένων, προτείνονται κάποια μέτρα αμοιβαίου αποκλεισμού και παρουσιάζεται ένας αλγόριθμος ανακάλυψης αμοιβαίως αποκλειόμενων αντικειμένων από μια βάση δεδομένων συναλλαγών. Τέλος, παρουσιάζονται κάποιες ιδέες σχετικά με την αξιοποίηση της πληροφορίας που παρέχουν τα αμοιβαίως αποκλειόμενα αντικείμενα για τη δημιουργία ή τροποποίηση ταξονομιών.

4.4.1 Ορισμός του Προβλήματος

Ορισμός 4.1. Έστω D ένα πεπερασμένο πολυσύνολο συναλλαγών και I ένα πεπερασμένο σύνολο αντικειμένων. Κάθε συναλλαγή $T \in D$ είναι ένα σύνολο αντικειμένων τέτοιο, ώστε $T \subseteq I$. Αν δύο αντικείμενα $i_1 \in I$ και $i_2 \in I$ είναι αμοιβαίως αποκλειόμενα, τότε δεν υπάρχει καμία συναλλαγή $T \in D$ τέτοια, ώστε $\{i_1, i_2\} \subseteq T$.

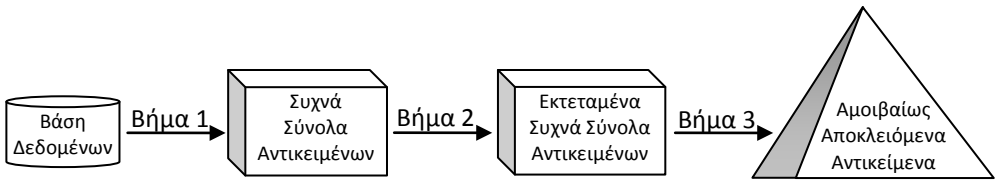
Ο παραπάνω ορισμός είναι ιδιαίτερα αυστηρός. Ωστόσο, το αντίστροφό του γενικά δεν ισχύει και συνεπώς δε μπορεί να χρησιμοποιηθεί για την αναγνώριση αμοιβαίως αποκλειόμενων αντικειμένων. Θεωρώντας μια βάση δεδομένων με περισσότερα από 10.000 αντικείμενα και μερικά εκατομμύρια συναλλαγές, είναι πιθανό να υπάρχει ένας μεγάλος αριθμός ζευγών αντικειμένων που δεν αγοράζονται μαζί. Σύμφωνα με όσα ορίζει ο Ορισμός 4.1, τα ζεύγη αυτά είναι πιθανό να αποτελούνται από αμοιβαίως αποκλειόμενα αντικείμενα. Στην πραγματικότητα όμως, ένα πολύ μικρό ποσοστό από αυτά μπορεί να είναι πράγματι αμοιβαίως αποκλειόμενα.

Με σκοπό την επίλυση του παραπάνων προβλήματος προτείνεται ένας αλγόριθμος για την εξόρυξη αμοιβαίως αποκλειόμενων αντικειμένων από μια βάση δεδομένων συναλλαγών. Ο αμοιβαίος αποκλεισμός δύο αντικειμένων μπορεί να αποτελέσει ένα εναλλακτικό είδος συσχέτισης. Φυσικά, αυτό είναι κάτι που πρέπει να επιβεβαιώνεται από τον ειδικό του εκάστοτε πεδίου έτσι, ώστε να μπορεί τελικά να αξιοποιηθεί. Επίσης, ο αμοιβαίος αποκλεισμός μπορεί να συνιστά αιτιώδη συσχέτιση μεταξύ δύο αντικειμένων, υπό την έννοια ότι η εμφάνιση ενός αντικειμένου τελικά ενδέχεται να «αποτρέπει» ή να «προκαλεί» τη μη εμφάνιση ενός άλλου.

Η εξόρυξη αμοιβαίως αποκλειόμενων αντικειμένων σε μια βάση δεδομένων που μπορεί να περιέχει μερικές χιλιάδες διαφορετικά αντικείμενα, περιλαμβάνει την αναζήτηση σε έναν χώρο που αποτελείται από όλα τα πιθανά ζεύγη αντικειμένων. Ωστόσο, η απλοϊκή αυτή προσέγγιση θα παρήγαγε έναν μεγάλο αριθμό υποψήφιων αμοιβαίως αποκλειόμενων ζευγών, πολλά από τα οποία δε θα ήταν αμοιβαίως αποκλειόμενα, για τους λόγους που εξηγήθηκαν στην αρχή της ενότητας. Η λύση που προτείνεται στην παρούσα διατριβή στηρίζεται στην παρατήρηση ότι κάθε συχνό σύνολο αντικειμένων εκφράζει μια συγκεκριμένη συμπεριφορά μιας ομάδας πελατών και ως εκ τούτου μπορεί να χρησιμοποιηθεί για να κατευθύνει την αναζήτηση. Τα αντικείμενα που εμφανίζονται με υψηλή συχνότητα στον υποχώρο ενός συχνού συνόλου αντικειμένων είναι πιο πιθανό να αλληλο-αποκλείονται συστηματικά, λόγω του ότι ακολουθούν μια

συγκεκριμένη αγοραστική συμπεριφορά και όχι λόγω τυχαίων ή ασυνήθιστων περιπτώσεων [Berberidis et al., 2005].

Ο προτεινόμενος αλγόριθμος αποτελείται από τρία βήματα (Εικόνα 4.1). Στο πρώτο βήμα, εξάγονται όλα τα συχνά σύνολα αντικειμένων. Στο βήμα αυτό μπορεί να χρησιμοποιηθεί οποιοσδήποτε αλγόριθμος εξόρυξης συχνών συνόλων αντικειμένων. Στη συνέχεια, τα συχνά σύνολα αντικειμένων χρησιμοποιούνται για την εξόρυξη *εκτεταμένων συχνών συνόλων αντικειμένων* (σύνολα που μοιράζονται ένα κοινό υποσύνολο), παράγοντας τις *επεκτάσεις* που θα χρησιμοποιηθούν στο επόμενο βήμα, ως υποψήφια αμοιβαίως αποκλειόμενα αντικείμενα. Το δεύτερο αυτό βήμα εκτελείται σε επίπεδα, όπως ο Apriori, και απαιτεί έναν αριθμό περασμάτων στη βάση δεδομένων ανάλογο με το μέγεθος των επεκτάσεων που πρόκειται να ανακαλυφθούν [Tzanis & Berberidis, 2007]. Στην περίπτωση του κλασικού αλγορίθμου Apriori, ο αριθμός των περασμάτων στη βάση είναι ίσος με το μέγεθος των συνόλων αντικειμένων, ωστόσο υπάρχουν ορισμένες βελτιώσεις του που απαιτούν λιγότερα περάσματα και συνεπώς μικρότερο χρόνο εκτέλεσης.



Εικόνα 4.1: Τα τρία βήματα του αλγορίθμου εξόρυξης.

4.4.2 Εξόρυξη Εκτεταμένων Συχνών Συνόλων Αντικειμένων

Έστω D ένα πεπερασμένο σύνολο συναλλαγών και I ένα πεπερασμένο σύνολο αντικειμένων. Κάθε συναλλαγή $T \in D$ είναι ένα σύνολο αντικειμένων τέτοιο, ώστε $T \subseteq I$. Η εξόρυξη των συχνών k-συνόλων αντικειμένων στο I απαιτεί την αναζήτηση σε έναν χώρο που αποτελείται από όλους τους πιθανούς συνδυασμούς μήκους k όλων των αντικειμένων του I. Κάθε συχνό σύνολο αντικειμένων $F \subseteq I$ διαιρεί το χώρο αναζήτησης σε δύο διακριτούς υποχώρους: ο πρώτος αποτελείται από τις συναλλαγές που περιέχουν το F και εφεξής θα αναφέρεται ως ο *F-υποχώρος* και ο δεύτερος από όλες τις άλλες συναλλαγές.

Ορισμός 4.2. Έστω $F \subseteq I$ ένα συχνό σύνολο αντικειμένων στο D, βάσει ενός κατώτατου ορίου υποστήριξης πρώτου επιπέδου ($supp_D(F) \geq min_gsup$) και $E \subseteq I$ ένα άλλο σύνολο αντικειμένων, τέτοιο ώστε $F \cap E = \varnothing$. Το σύνολο $F \cup E$ αποτελεί *εκτεταμένο* (*extended*) *συχνό σύνολο αντικειμένων*, αν το E είναι συχνό στον F-υποχώρο, βάσει ενός ορίου υποστήριξης δεύτερου επιπέδου ($supp_{F-\text{υποχώρος}}(E) \geq min_lsup$).

Το σύνολο αντικειμένων E αποτελεί την τοπικά συχνή επέκταση του F. Ο όρος «τοπικά» χρησιμοποιείται επειδή το E μπορεί να μην είναι συχνό στο σύνολο όλων των συναλλαγών. Προκειμένου να αποφευχθεί η σύγχυση, στο εξής θα χρησιμοποιούνται οι όροι τοπικός και τοπικά, όταν θα γίνεται αναφορά σε ένα υποσύνολο του D, ενώ θα

γίνεται χρήση των όρων καθολικός και καθολικά, όταν θα γίνεται αναφορά σε ολόκληρο το *D*. Για παράδειγμα, η υποστήριξη πρώτου επιπέδου ονομάζεται *καθολική υποστήριξη* (*gsup*) και η υποστήριξη δεύτερου επιπέδου καλείται *τοπική υποστήριξη* (*lsup*). Ένα σύνολο αντικειμένων *F* που πληροί το κριτήριο της ελάχιστης καθολικής υποστήριξης (*min_gsup*) θεωρείται καθολικώς συχνό, ενώ ένα σύνολο αντικειμένων *E* που είναι συχνό στον *F*-υποχώρο, ικανοποιώντας το κριτήριο ελάχιστης τοπικής υποστήριξης (*min_lsup*), θεωρείται τοπικά συχνό. Η τοπική υποστήριξη ενός αντικειμένου *E* στον *F*-υποχώρο υπολογίζεται βάσει της Εξίσωσης 4.1.

$$supp_{F-\text{υποχώρος}}(E) = supp_{T \in D | T \supseteq F}(E) = \frac{supp_D(E \cup F)}{supp_D(F)} \qquad 4.1$$

Το κατώτατο όριο τοπικής υποστήριξης ορίζεται από το χρήστη – ειδικό του πεδίου εφαρμογής. Τα εκτεταμένα συχνά σύνολα αντικειμένων που περιέχουν τοπικά συχνές επεκτάσεις μήκους *k* ονομάζονται *k-εκτεταμένα συχνά* σύνολα αντικειμένων (*k-extended frequent itemsets*).

Δοθέντος ενός πεπερασμένου συνόλου συναλλαγών *D*, το πρόβλημα της εξόρυξης των εκτεταμένων συχνών συνόλων αντικειμένων ταυτίζεται με τη δημιουργία όλων των συνόλων αντικειμένων $F \cup E$, που αποτελούνται από ένα σύνολο αντικειμένων *F*, που έχει καθολική υποστήριξη τουλάχιστον ίση με το κατώτατο όριο και μια επέκταση *E*, που έχει τοπική υποστήριξη τουλάχιστον ίση με το αντίστοιχο κατώτατο όριο.

4.4.3 Εξόρυξη Ζευγών Αμοιβαίως Αποκλειόμενων Αντικειμένων

Η ανακάλυψη των αμοιβαίως αποκλειόμενων αντικειμένων προϋποθέτει τον καθορισμό ενός μέτρου αμοιβαίου αποκλεισμού βάσει του οποίου θα αξιολογείται ο βαθμός κατά τον οποίο δύο αντικείμενα θεωρούνται αμοιβαίως αποκλειόμενα. Ωστόσο, το μέτρο του αμοιβαίου αποκλεισμού πρέπει να υπολογίζεται σε δύο φάσεις. Αρχικά, πρέπει να υπολογιστεί ένα μέτρο αμοιβαίου αποκλεισμού δύο αντικειμένων μέσα στον υποχώρο ενός συχνού συνόλου αντικειμένων (τοπικά). Στη συνέχεια, πρέπει να να υπολογιστεί καθολικά, όπου όλα τα συχνά σύνολα αντικειμένων που υποστηρίζουν το συγκεκριμένο ζεύγος θα συνεισφέρουν αναλόγως. Για το σκοπό αυτό ορίζονται δύο μέτρα αμοιβαίου αποκλεισμού (Mutual Exclusion Measures - MEM), ένα τοπικό και ένα καθολικό.

Τοπικό Μέτρο Αμοιβαίου Αποκλεισμού

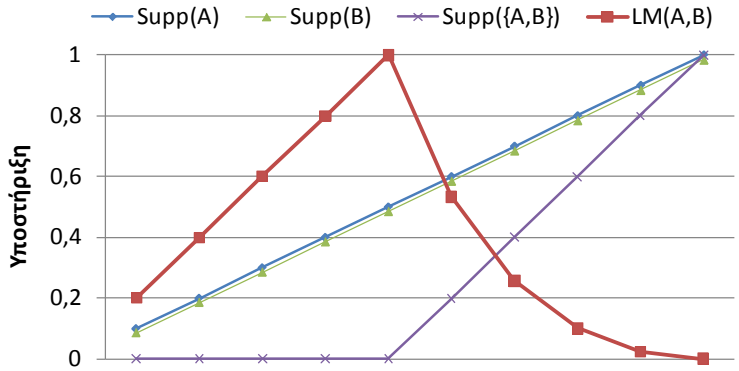
Το τοπικό μέτρο αμοιβαίου αποκλεισμού (*Local MEM – LM*) ορίζεται από την Εξίσωση 4.2 για την αξιολόγηση ενός υποψήφιου ζεύγους αμοιβαίως αποκλειόμενων αντικειμένων *A* και *B*, που υποστηρίζονται από ένα συχνό σύνολο αντικειμένων *F*. Το μέτρο αυτό προσδιορίζει το βαθμό στον οποίο τα αντικείμενα *A* και *B* είναι τοπικά αμοιβαίως αποκλειόμενα μέσα στον υποχώρο του συχνού συνόλου αντικειμένων *F*. Το τοπικό μέτρο αμοιβαίου αποκλεισμού *LM* λαμβάνει τιμές στο διάστημα [0, 1].

$$LM_F(A,B) = [P(A\text{-}B) + P(B\text{-}A)] \times min[P(A\text{-}B|A),\ P(B\text{-}A|B)]$$

$$\text{ή } LM_F(A,B) = \left(S_A - S_{\{A,B\}} + S_B - S_{\{A,B\}}\right) min\left(\frac{S_A - S_{\{A,B\}}}{S_A}, \frac{S_B - S_{\{A,B\}}}{S_B}\right) \qquad 4.2$$

$$\text{ή } LM_F(A,B) = \left(S_A + S_B - 2S_{\{A,B\}}\right)\left(1 - \frac{S_{\{A,B\}}}{min(S_A,\ S_B)}\right)$$

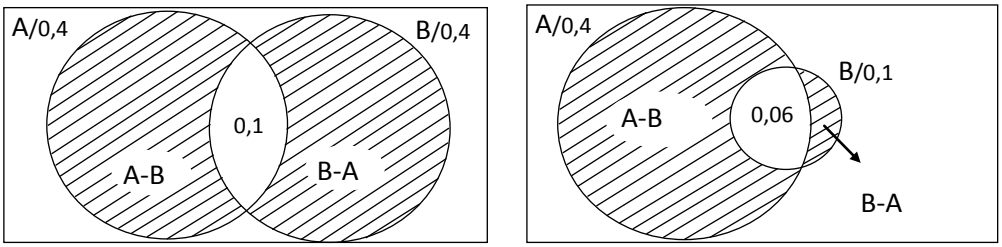
Στην Εξίσωση 4.2 ισχύει *P(F)* = 1. Το S_X είναι το κλάσμα των συναλλαγών που υποστηρίζουν το *F* και περιέχουν το αντικείμενο (ή σύνολο αντικειμένων) *X*, δηλαδή η τοπική υποστήριξη του *X* ($supp_{F\text{-υποχώρος}}(X)$). Η Εικόνα 4.2 δείχνει ότι η *LM* αυξάνει ανάλογα με την αύξηση της υποστήριξης των *A* και *B* και για όσο δεν υπάρχει αλληλεπικάλυψη (δηλαδή ισχύει S_A<0,5, S_B<0,5 και $S_{\{A,B\}}$=0), μέχρις ότου φτάσει στη μέγιστη τιμή της που είναι 1. Για λόγους απλότητας θεωρείται ότι S_A=S_B, χωρίς να επηρεάζεται η γενικότητα. Όταν αρχίζει η αλληλεπικάλυψη (S_A>0,5 και S_B>0,5), η *LM* ελαττώνεται απότομα έως ότου μηδενίζεται. Αυτή η ιδιότητα είναι επιθυμητή, αφού ακόμα και ένας μικρός βαθμός αλληλεπικάλυψης αποτελεί σοβαρή ένδειξη εις βάρος του γεγονότος ότι τα δύο αντικείμενα είναι αμοιβαίως αποκλειόμενα. Επιπλέον, η τιμή της *LM* διατηρείται χαμηλή, όταν οι τοπικές υποστηρίξεις των δύο αντικειμένων (S_A και S_B) είναι μικρές, ακόμη και αν δεν υπάρχει καθόλου αλληλεπικάλυψη των συναλλαγών που τα υποστηρίζουν. Αυτό είναι επίσης μια επιθυμητή ιδιότητα, καθώς όταν δύο αντικείμενα έχουν πολύ μικρή τοπική υποστήριξη είναι πιθανό να παρουσιάζονται ψευδώς ως αμοιβαίως αποκλειόμενα καθώς ούτως ή άλλως η πιθανότητα να εμφανιστούν μαζί είναι μικρή.



Εικόνα 4.2: Το τοπικό μέτρο αμοιβαίου αποκλεισμού (LM) σε σύγκριση με την τοπική υποστήριξη των αντικειμένων *A* και *B* και του συνόλου {*A*, *B*} στον *F*-υποχώρο.

Στην Εικόνα 4.3 παρουσιάζεται ένα χαρακτηριστικό παράδειγμα με διαγράμματα Venn. Ο χώρος των διαγραμμάτων (α) και (β) αποτελείται από το σύνολο των συναλλαγών που υποστηρίζουν το συχνό σύνολο *F*. Οι γραμμοσκιασμένες περιοχές αναπαριστούν τις συναλλαγές που υποστηρίζουν το *A* αλλά όχι το *B* (ή *A-B*) και το *B* αλλά όχι το *A* (ή *B-A*), που είναι ο πρώτος παράγοντας στην Εξίσωση 4.2. Ο δεύτερος παράγοντας

της Εξίσωσης 4.2 εκφράζεται από το ελάχιστο των (*A*–*B*)/*A* και (*B*–*A*)/*B*. Αυτός ο παράγοντας λαμβάνεται υπόψη, καθώς δεν είναι αρκετό να είναι πολλές οι συναλλαγές που υποστηρίζουν μόνο το ένα από τα δύο αντικείμενα (γραμμοσκιασμένες περιοχές), αλλα πρέπει και η πιθανή αλληλεπικάλυψη να αποτελεί μικρό ποσοστό καθενός εκ των δύο αντικειμένων. Για παράδειγμα, τα δύο αντικείμενα στο διάγραμμα (β) (Εικόνα 4.3) έχουν αθροιστικά μεγάλο πλήθος συναλλαγών που υποστηρίζουν μόνο το ένα από τα δύο αντικείμενα, και ο πρώτος παράγοντας της Εξίσωσης 4.2 θα είναι αρκετά μεγάλος. Ωστόσο, ο δεύτερος παράγοντας θα είναι μικρός, με αποτέλεσμα τα αντικείμενα του διαγράμματος (β) να είναι λιγότερο πιθανό να είναι αμοιβαίως αποκλειόμενα απ' ότι του διαγράμματος (α). Σύμφωνα με τα αριθμητικά στοιχεία των διαγραμμάτων, η *LM* για το διάγραμμα (α) έχει τιμή 0,45, ενώ για το διάγραμμα (β) έχει τιμή 0,152.



(α) **(β)**

Εικόνα 4.3: Διαγράμματα Venn ζεύγους υποψήφιων αμοιβαίως αποκλειόμενων αντικειμένων.

Καθολικό Μέτρο Αμοιβαίου Αποκλεισμού

Το καθολικό μέτρο αμοιβαίου αποκλεισμού (*Global MEM – GM*) υπολογίζεται από την Εξίσωση 4.3 για την αξιολόγηση του αμοιβαίου αποκλεισμού ενός υποψήφιου ζεύγους αμοιβαίως αποκλειόμενων αντικειμένων, που υποστηρίζεται από ένα σύνολο *FS* από συχνά σύνολα αντικειμένων.

$$GM_1(A,B) = IIF\left(\sum_{F \in FS} supp_D(F) LM_F(A,B)\right) \qquad 4.3$$

Το *IIF* (*Itemset Independence Factor)* είναι ο *συντελεστής ανεξαρτησίας* συνόλων αντικειμένων και είναι ο λόγος του αριθμού των διακριτών αντικειμένων που περιέχονται σε όλα τα σύνολα αντικειμένων που υποστηρίζουν ένα υποψήφιο ζεύγος, προς το συνολικό αριθμό των εμφανίσεων των αντικειμένων που περιέχονται στα σύνολα αυτά. Για παράδειγμα, ο συντελεστής *IIF* των συνόλων {*A, B, C*} και {*A, D*} είναι 0,8 επειδή υπάρχουν 4 διακριτά αντικείμενα (*A, B, C* και *D*) σε ένα σύνολο 5 εμφανίσεων αντικειμένων (*A, B, C, A* και *D*). Ο συντελεστής *IIF* χρησιμοποιείται στον υπολογισμό του καθολικού μέτρου προκειμένου να ελαττωθεί η πόλωση που υπεισέρχεται λόγω της αλληλεπικάλυψης των συχνών συνόλων αντικειμένων που υποστηρίζουν τα υποψήφια αμοιβαίως αποκλειόμενα αντικείμενα. Εναλλακτικά, για τον υπολογισμό της *GM* μπορεί να χρησιμοποιηθεί και η Εξίσωση 4.4, που προκύπτει από κανονικοποίηση της 4.3. Το πεδίο τιμών της GM_1 είναι το [0, +∞), ενώ της GM_2 είναι το [0, 1].

$$GM_2(A,B) = IIF \frac{\sum_{F \in FS} supp_D(F) LM_F(A,B)}{\max_{FS \in FSS} \left(\sum_{F \in FS} supp_D(F) \right)} \qquad 4.4$$

Το *FSS* είναι το σύνολο που περιέχει όλα τα συχνά σύνολα αντικειμένων που υποστηρίζουν κάποιο ζεύγος υποψήφιων αμοιβαίως αποκλειόμενων αντικειμένων της βάσης δεδομένων. Η διαφορά ανάμεσα στις εξισώσεις 4.3 και 4.4 είναι ο παρονομαστής της 4.4. Ο συγκεκριμένος όρος χρησιμοποιείται προκειμένου να κανονικοποιηθεί το μέτρο, αφού η τιμή που παίρνει ο αριθμητής είναι πάντοτε μικρότερη ή ίση με την τιμή του παρονομαστή. Με άλλα λόγια, για όλα τα υποψήφια ζεύγη υπολογίζεται το άθροισμα των υποστηρίξεων των συχνών συνόλων αντικειμένων που τα υποστηρίζουν. Το μεγαλύτερο από τα αθροίσματα χρησιμοποιείται για την κανονικοποίηση των καθολικών μέτρων, κάτι που τελικά είναι χρήσιμο για το σωστό ορισμό του κατώτατου ορίου από το χρήστη. Ωστόσο, αυτό δεν είναι επιθυμητό όταν ο χρήστης θέλει να συγκρίνει τα αποτελέσματα του αλγορίθμου σε διαφορετικά σύνολα δεδομένων ή τα αποτελέσματα από εκτελέσεις του αλγορίθμου με διαφορετικό σύνολο παραμέτρων, οπότε προτιμάται η μη κανονικοποιημένη μορφή του καθολικού μέτρου (GM_1).

Τελικά, η τιμή του καθολικού μέτρου χρησιμοποιείται για να προσδιορίσει κατά πόσο δύο αντικείμενα της βάσης δεδομένων είναι αμοιβαίως αποκλειόμενα.

4.4.4 Εξόρυξη Τριάδων Αμοιβαίως Αποκλειόμενων Αντικειμένων

Με επέκταση του αλγορίθμου εξόρυξης ζευγών αμοιβαίως αποκλειόμενων αντικειμένων είναι δυνατό να ανακαλυφθούν μεγαλύτερου μεγέθους σύνολα από αμοιβαίως αποκλειόμενα αντικειμένα. Στο συγκεκριμένο εδάφιο παρουσιάζεται η εξόρυξη τριάδων αμοιβαίως αποκλειόμενων αντικειμένων [Tzanis & Vlahavas, 2010].

Για τη δημιουργία των αμοιβαίως αποκλειόμενων τριάδων ακολουθείται η εξής διαδικασία. Αρχικά, για κάθε συχνό σύνολο εξάγονται σύμφωνα με όσα περιγράφηκαν στις προηγούμενες παραγράφους τα υποψήφια αμοιβαίως αποκλειόμενα ζεύγη αντικειμένων. Στη συνέχεια, από τα ζεύγη αυτά δημιουργούνται όλες οι υποψήφιες αμοιβαίως αποκλειόμενες τριάδες και υπολογίζονται τα τοπικά και καθολικά μέτρα τους. Η δημιουργία των αμοιβαίως αποκλειόμενων τριάδων γίνεται με την ταξινόμηση όλων των αμοιβαίως αποκλειόμενων ζεύγών και την ένωση εκείνων που έχουν κοινό το πρώτο στοιχείο τους. Το τοπικό μέτρο αμοιβαίου αποκλεισμού μιας τριάδας αντικειμένων ισούται με το ελάχιστο από τα αντίστοιχα τοπικά μέτρα αμοιβαίου αποκλεισμού των υποσυνόλων της με πλήθος δύο αντικειμένων. Το καθολικό μέτρο αμοιβαίου αποκλεισμού μιας τριάδας αντικειμένων υπολογίζεται με τον ίδιο τρόπο που υπολογίζεται το καθολικό μέτρο για ζεύγη αντικειμένων (εξισώσεις 4.3 και 4.4). Τέλος, με βάση το όριο του καθολικού μέτρου δημιουργείται το τελικό σύνολο των αμοιβαίως αποκλειόμενων τριάδων.

Ο Πίνακας 4.1 παρουσιάζει τον αλγόριθμο εξόρυξης αμοιβαίως αποκλειόμενων αντικειμένων. Στο πρώτο βήμα, κατά το οποίο εξάγονται τα συχνά σύνολα αντικειμένων, μπορεί να χρησιμοποιηθεί οποιοσδήποτε γνωστός αλγόριθμος.

Πίνακας 4.1: Ο αλγόριθμος εξόρυξης αμοιβαίως αποκλειόμενων αντικειμένων.

Είσοδος: Ένα σύνολο συναλλαγών *D*, μια μέθοδος εξόρυξης συχνών συνόλων αντικειμένων (*FreqItemsetsMiningMethod*), ένα κατώτατο όριο καθολικής υποστήριξης (*min_gSup*), ένα κατώτατο όριο τοπικής υποστήριξης (*min_lSup*) και ένα κατώτατο όριο καθολικού μέτρου αμοιβαίου αποκλεισμού (*min_GM*).

Έξοδος: Αμοιβαίως αποκλειόμενα ζεύγη και τριάδες αντικειμένων.

```
FI ← mineFrequentItemsets(D, FreqItemsetsMiningMethod, min_gSup)
for each (T ∈ D)
  for each (I ∈ FI)
    if (I ⊆ T)
      for each (E ∈ T-I)
        Extensions(I) ← Extensions(I) ∪ E
        Count[I][E] ← Count[I][E] + 1
for each (I ∈ FI)
  for each (E ∈ Extensions(I))
    if (Count[I][E] < min_lSup)
      Extensions(I) ← Extensions(I) - E
for each (I ∈ FI)
  for each (E1, E2 ∈ Extensions(I))
    CandidatePairs(I) ← CandidatePairs(I) ∪ {E1, E2}
    LM[I][{E1, E2}] ← calculateLM(I, {E1, E2})
  for each ({E1, E2}, {E1, E3} ∈ CandidatePairs(I))
    CandidateTriples(I) ← CandidateTriples(I) ∪ {E1, E2, E3}
    LM[I][{E1, E2, E3}] ← min(LM[I][{E1, E2}], LM[I][{E1, E3}],
                              LM[I][{E2, E3}])
for each (I ∈ FI)
  for each (EP ∈ CandidatePairs(I))
    GlobCandidatePairs ← GlobCandidatePairs ∪ EP
  for each (ET ∈ CandidateTriples(I))
    GlobCandidateTriples ← GlobCandidateTriples ∪ ET
for each (EP ∈ GlobCandidatePairs)
  if (calculateGM(EP) ≥ min_GM)
    MutuallyExclusivePairs ← MutuallyExclusivePairs ∪ EP
for each (ET ∈ GlobCandidateTriples)
  if (calculateGM(ET) ≥ min_GM)
    MutuallyExclusiveTriples ← MutuallyExclusiveTriples ∪ EP
return MutuallyExclusivePairs ∪ MutuallyExclusiveTriples
```

4.4.5 Παράδειγμα Εκτέλεσης του Αλγορίθμου

Στο εδάφιο αυτό περιγράφεται ένα παράδειγμα εφαρμογής προτεινόμενου αλγορίθμου. Ο Πίνακας 4.2 περιέχει ένα σύνολο δεδομένων καλαθιού αγορών και ο Πίνακας 4.3 τους κανόνες συσχέτισης που εξάγονται με κατώτατο όριο υποστήριξης 2/9.

Η εκτέλεση του αλγορίθμου παράγει τα αμοιβαίως αποκλειόμενα ζεύγη αντικειμένων που περιέχει ο Πίνακας 4.4, μαζί με τα σχετικά μέτρα αμοιβαίου αποκλεισμού και τα συχνά σύνολα αντικειμένων που τα υποστηρίζουν.

Πίνακας 4.2: Ένα σύνολο δεδομένων καλαθιού αγορών.

TID	Αντικείμενα
1	espresso, sugar, newspaper
2	espresso, sugar, cola
3	espresso, sugar
4	cappuccino, cigarettes
5	cappuccino, sugar
6	cappuccino, sugar, sweets
7	decaf, sugar, chewing_gums
8	decaf, soda, vinegar
9	decaf, sugar, cigarettes

Πίνακας 4.3: Οι κανόνες συσχέτισης του συνόλου δεδομένων (Πίνακας 4.2).

Κανόνες Συσχέτισης	Υποστήριξη	Εμπιστοσύνη
espresso ⇒ sugar	3/9	1
decaf ⇒ sugar	2/9	2/3
cappuccino ⇒ sugar	2/9	2/3

Πίνακας 4.4: Αμοιβαίως αποκλειόμενα αντικείμενα, τα συχνά σύνολα που τα υποστηρίζουν και τα αντίστοιχα μέτρα αμοιβαίου αποκλεισμού.

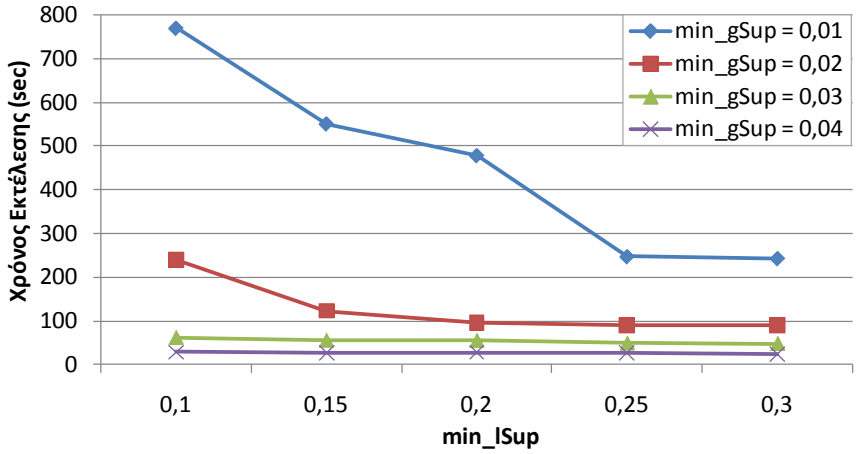
Αμοιβαίως Αποκλειόμενα Ζεύγη: (*GM*, Καθολική Υποστήριξη)	Συχνά Σύνολα Αντικειμένων	
	Καθολική Υποστήριξη	*Τοπικό Μέτρο (LM)*
{espresso, cappuccino}: 5/9, 0	{sugar}: 7/9	0,714
{espresso, decaf}: 5/9, 0	{sugar}: 7/9	0,714
{cappuccino, decaf}: 4/9, 0	{sugar}: 7/9	0,571

4.4.6 Πειράματα

Για την αξιολόγηση της απόδοσης του αλγορίθμου, εκτελέστηκε ένας αριθμός πειραμάτων σε ένα συνθετικό σύνολο δεδομένων καλαθιού αγορών της IBM. Το σύνολο αυτό περιέχει 100.000 συναλλαγές. Το γράφημα στην Εικόνα 4.4 δείχνει την απόδοση του αλγορίθμου εξόρυξης με όρους χρόνων εκτέλεσης (δευτερόλεπτα), για μεταβαλλόμενο όριο τοπικής υποστήριξης (από 0,1 μέχρι 0,3) και μεταβαλλόμενο όριο καθολικής υποστήριξης (από 0,01 μέχρι 0,04).

Όπως ήταν αναμενόμενο, ο χρόνος εκτέλεσης αυξάνει καθώς μειώνονται τα κατώτατα όρια υποστήριξης. Μάλιστα, η απόδοση φθίνει σημαντικά για πολύ χαμηλά όρια. Ωστόσο, το όριο καθολικής υποστήριξης (min_gsup) επηρεάζει περισσότερο την απόδοση, επειδή όσο χαμηλότερο είναι τόσο περισσότερα συχνά σύνολα αντικειμένων θα

εξορυχθούν στο πρώτο βήμα και κατ' επέκταση, τόσο περισσότεροι υποχώροι συχνών συνόλων αντικειμένων θα πρέπει να ελεγχθούν.



Εικόνα 4.4: Χρόνοι εκτέλεσης του αλγορίθμου εξόρυξης.

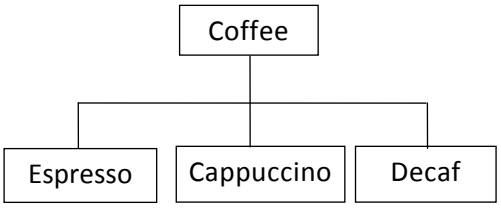
4.4.7 Αμοιβαίως Αποκλειόμενα Αντικείμενα και Ταξονομίες

Η λογική πίσω από την προσέγγιση της ανακάλυψης αμοιβαίως αποκλειόμενων αντικείμενων που παρουσιάστηκε είναι διττή. Πρώτον, οι επεκτάσεις που είναι συχνές στον υποχώρο ενός συχνού συνόλου αντικειμένων, ενδέχεται να αποτελούν ενδιαφέρουσα πληροφορία που χαρακτηρίζει το σύνολο αυτό. Δεύτερον, αν ένας μεγάλος αριθμός συνόλων αντικειμένων μοιράζονται τις ίδιες επεκτάσεις και οι κοινές αυτές επεκτάσεις είναι συχνές στους αντίστοιχους υποχώρους, τότε είναι πιθανό να είναι αμοιβαίως αποκλειόμενες και να ανήκουν στην ίδια κατηγορία και στο ίδιο επίπεδο μιας υποκείμενης ταξονομίας. Στις περιπτώσεις αυτές, η συνολική υποστήριξη του κόμβου-γονέα στην ταξονομία διαιρείται σε επιμέρους υποστηρίξεις χαμηλότερων επιπέδων, οι οποίες δεν είναι αρκετά μεγάλες, ώστε να ικανοποιούν το κριτήριο ελάχιστης υποστήριξης. Αυτό οδηγεί σε απώλεια πληροφορίας. Στις περιπτώσεις που δεν υπάρχει διαθέσιμη πληροφορία σχετικά με υποκείμενη ταξονομία, η πληροφορία που συγκεντρώνεται κατά τη διαδικασία που περιγράφηκε, μπορεί να αποτελέσει σημαντική ένδειξη για την ύπαρξη ταξονομίας.

Η γνώση ότι δύο αντικείμενα είναι αμοιβαίως αποκλειόμενα μπορεί να αναλυθεί περαιτέρω, προκειμένου να αποφασιστεί αν τα συγκεκριμένα αντικείμενα μπορούν να ταξινομηθούν στο ίδιο επίπεδο της ταξονομίας, κάτω από τον ίδιο κόμβο-γονέα. Η «τυφλή» αναζήτηση θα παρήγαγε έναν τεράστιο αριθμό υποψήφιων ζευγών, τα περισσότερα από τα οποία πιθανότατα δε θα είχαν κάποιο ενδιαφέρον. Η λογική στην οποία βασίζεται η αναζήτηση αμοιβαίως αποκλειόμενων αντικειμένων ανάμεσα στις επεκτάσεις των συχνών συνόλων αντικειμένων έχει δύο σκέλη. Πρώτον, ο χώρος αναζήτησης μειώνεται δραματικά και δεύτερον, ένα συχνό σύνολο αντικειμένων εκφράζει μια κατηγορία συμπεριφοράς με συγκεκριμένα χαρακτηριστικά, π.χ. καταναλωτές που έχουν συγκεκριμένες προτιμήσεις. Έτσι, τα αμοιβαίως αποκλειόμενα αντικείμενα που

βρίσκονται στον υποχώρο ενός συχνού συνόλου αντικειμένων έχουν ένα επιπλέον ενδιαφέρον. Ας υποτεθεί ότι σε ένα μεγάλο κατάστημα ειδών ένδυσης πωλείται το σύνολο αντικειμένων A = {Κάλτσες, Αθλητικά Παπούτσια, Ρακέτα}, που προκύπτει από αθλητές του τένις και το σύνολο B = {Κάλτσες, Αθλητικά Παπούτσια, Μπάλα Πετοσφαίρισης}, που προκύπτει από αθλητές πετοσφαίρισης. Ενώ τα αντικείμενα Α και Β ενδέχεται να μην έχουν μεγάλη υποστήριξη, το $C = A \cap B$ = {Κάλτσες, Αθλητικά Παπούτσια} είναι πολύ πιθανόν να έχει μεγαλύτερη υποστήριξη από αυτά, αφού προκύπτει από όλους σχεδόν τους αθλητές. Ωστόσο, τα αντικείμενα «Ρακέτα» και «Μπάλα Πετοσφαίρισης» είναι λογικό να είναι τοπικά συχνά στο υποσύνολο των συναλλαγών που περιέχουν το C. Οι παραδοσιακοί κανόνες συσχέτισης δεν παρέχουν τη δυνατότητα να εξερευνήσει κανείς τον υποχώρο του C με αποτέλεσμα πιθανώς πολύτιμη γνώση να παραμένει ανεξερεύνητη. Με τον προτεινόμενο αλγόριθμο, τα αντικείμενα «Ρακέτα» και «Μπάλα Πετοσφαίρισης» θα ανακαλυφθούν ως αμοιβαίως αποκλειόμενα, υποδεικνύοντας μάλιστα ότι ενδέχεται να συνδέονται με κάποια ταξονομική σχέση [Tzanis et al., 2006a].

Επιστρέφοντας στο παράδειγμα του Εδαφίου 4.4.5, τα αμοιβαίως αποκλειόμενα ζεύγη που περιέχει ο Πίνακας 4.4 είναι όλα διαφορετικά είδη καφέ (espresso, cappuccino και decaf). Θεωρώντας ότι ανήκουν στο ίδιο επίπεδο της ταξονομίας (Εικόνα 4.5), αν αντικατασταθούν όλα με ένα άλλο αντικείμενο και συγκεκριμένα με τον κόμβο-γονέα «Coffee», τότε είναι δυνατό να αυξηθεί το κατώτατο όριο υποστήριξης με αποτέλεσμα να ανακαλυφθούν ισχυρότεροι κανόνες συσχέτισης (Πίνακας 4.5).



Εικόνα 4.5: Μια ταξονομία προϊόντων καφέ.

Πίνακας 4.5: Γενικευμένοι κανόνες συσχέτισης βάσει της ταξονομίας στην Εικόνα 4.5.

Κανόνες συσχέτισης	Υποστήριξη	Εμπιστοσύνη
sugar ⇒ coffee	7/9	1
coffee ⇒ sugar	7/9	7/9

Η δημιουργία ταξονομιών είναι χρήσιμη σε πολλά πεδία εφαρμογής. Ακόμα και στις περιπτώσεις όπου η ταξονομία είναι εκ των προτέρων γνωστή, συχνά χρειάζονται διορθώσεις και αναθεωρήσεις, είτε λόγω αλλαγών στις προτιμήσεις των πελατών, είτε λόγω της εισαγωγής νέων προϊόντων, που μπορεί να επηρεάζουν τελικά και την ίδια την ταξονομία αλλά και τις επιλογές των πελατών. Επίσης, οι προτιμήσεις των πελατών μπορεί να διαφέρουν σε διαφορετικές γεωγραφικές τοποθεσίες και αυτό να συντελεί

σε διαφορετικές ταξονομίες. Για παράδειγμα, μια ταξονομία που χρησιμοποιείται σε ένα κατάστημα ενδυμάτων σε μια ορεινή περιοχή πιθανότατα θα είναι ακατάλληλη για κάποιο άλλο παραθαλάσσιο. Στο παράδειγμα με τα αθλητικά είδη, μπορεί να μην υπάρχει κάποια ταξονομία που να περιέχει ρακέτες και μπάλες πετοσφαίρισης στο ίδιο επίπεδο και κάτω από τον ίδιο κόμβο-γονέα. Πιθανόν όμως να πρέπει να οριστεί μια τέτοια σχέση και να εισαχθεί στην εννοιολογική ιεραρχία ή να αναθεωρηθεί κάποια από τις υπάρχουσες ώστε να συμπεριλάβει τη νέα γνώση.

4.5 Ανακάλυψη Αμοιβαίως Αποκλειόμενων Γονιδίων

Στην ενότητα αυτή παρουσιάζεται η εφαρμογή του αλγορίθμου ανακάλυψης αμοιβαίως αποκλειόμενων αντικειμένων στον τομέα της βιοπληροφορικής και συγκεκριμένα στην ανάλυση της γονιδιακής έκφρασης [Tzanis & Vlahavas, 2010]. Ο σκοπός της συγκεκριμένης εφαρμογής είναι η ανακάλυψη γονιδίων που η έκφραση του ενός αποκλείει την έκφραση του άλλου.

4.5.1 Η Τεχνική SAGE

Τα γονίδια περιέχουν τις πληροφορίες για τη δημιουργία των μορίων RNA, τα οποία χρησιμοποιούνται στη συνέχεια για τη δημιουργία των πρωτεϊνών, οι οποίες είναι υπεύθυνες για τις λειτουργίες του κυττάρου και παίζουν κύριο ρόλο στη συμπεριφορά του. Ενώ τα ανθρώπινα κύτταρα περιέχουν περίπου 20.000-25.000 γονίδια, μόνο το 20% περίπου αυτών συνήθως είναι ενεργοποιημένο κάθε χρονική στιγμή. Όταν αλλάζουν οι συνθήκες του περιβάλλοντος του κυττάρου, με αποτέλεσμα την αλλαγή του σχήματος ή της συμπεριφοράς του ή όταν υπάρξει παρουσία μολυσματικού παράγοντα, κάποια από τα γονίδια του κυττάρου θα σταματήσουν να εκφράζονται και κάποια άλλα θα ενεργοποιηθούν. Το πλήθος των μορίων RNA που θα παράγει κάθε γονίδιο μπορεί επίσης να τροποποιηθεί.

Η τεχνική SAGE (Εδάφιο 3.10.2) είναι μια μέθοδος που παρέχει τη δυνατότητα της εποπτικής παρακολούθησης του συνόλου της γονιδιακής δραστηριότητας ενός κυττάρου. Ένα από τα πιο ενδιαφέροντα χαρακτηριστικά της τεχνικής SAGE είναι ότι μπορεί να χρησιμοποιηθεί για την ανακάλυψη νέων γονιδίων σε οργανισμούς για τους οποίους δεν έχει βρεθεί το πλήρες γονιδίωμα. Αν μια αλληλουχία δεν ταιριάζει με κάποιο γνωστό γονίδιο, τότε προφανώς προέρχεται από κάποιο γονίδιο το οποίο δεν έχει ανακαλυφθεί. Τα προφίλ διαφορετικών τύπων κυττάρων συνήθως είναι αρκετά διαφορετικά. Το προφίλ ενός καρκινικού ή ενός μολυσμένου κυττάρου, επίσης, είναι αναμενόμενο να παρεκκλίνει από το αντίστοιχο προφίλ του φυσιολογικού κυττάρου. Με την καταγραφή της ολοκληρωμένης δραστηριότητας του γονιδιώματος, η τεχνική SAGE είναι ικανή να αποκαλύψει σημαντική γνώση για τα πρότυπα της δραστηριότητας των γονιδίων που συνεισφέρουν σε μία συγκεκριμένη ασθένεια. Η γνώση αυτή είναι ζωτική για την ανάπτυξη φαρμάκων τα οποία στοχεύουν στα αδύναμα σημεία των κυττάρων.

4.5.2 Τα Δεδομένα SAGE

Ο αλγόριθμος ανακάλυψης αμοιβαίως αποκλειόμενων αντικειμένων εφαρμόστηκε σε σύνολο δεδομένων γονιδιακής έκφρασης, τα οποία έχουν ληφθεί με τη μέθοδο SAGE. Τα δεδομένα αυτά έχουν συγκεντρωθεί το Δεκέμβριο του 2002 από την ερευνητική ομάδα του Dr. Olivier Gandrillon (Centre génétique Moléculaire et Cellulaire de Lyon). Το συγκεκριμένο σύνολο δεδομένων έχει παρουσιαστεί και μελετηθεί στα ECML/PKDD Discovery Challenge Workshops κατά τα έτη 2004 και 2005 [Gandrillon, 2004]. Το σύνολο δεδομένων έχει τη μορφή πίνακα σχεσιακής βάσης δεδομένων και αποτελείται από 27.679 χαρακτηριστικά (ετικέτες) και 90 παραδείγματα (βιβλιοθήκες), στη μορφή που παρουσιάζει ο Πίνακας 4.6.

Πίνακας 4.6: Σύνολο δεδομένων SAGE.

	Ετικέτα 1	Ετικέτα 2	…	Ετικέτα 27679
Βιβλιοθήκη 1	6	10	…	8
Βιβλιοθήκη 2	23	13	…	15
…	…	…	…	…
Βιβλιοθήκη 90	43	4	…	4

Τα δεδομένα συνοδεύει ένας μεγάλος αριθμός χαρακτηριστικών, που περιγράφουν κάθε βιβλιοθήκη. Ανάμεσα σε αυτά είναι ο αριθμός, το όνομα και η περιγραφή της βιβλιοθήκης, το φύλο του ατόμου από το οποίο ελήφθει, το στάδιο ανάπτυξης, ο τύπος του κυττάρου, ο προμηθευτής του ιστού, το εργαστήριο στο οποίο έχει παραχθεί, ο τύπος ιστού ή οργάνου (νεφρός, μαστός, λέμφος, προστάτης, ωοθήκη, εγκέφαλος, έντερο, πάγκρεας, παρεγκεφαλίδα, αγγείο, δέρμα, θάλαμος εγκεφάλου, περιτόναιο) και η βιολογική κατάσταση του δείγματος (φυσιολογικό, καρκινικό). Από το σύνολο των χαρακτηριστικών που περιγράφουν τα δεδομένα, μόνο τα δύο τελευταία ελήφθησαν υπόψη κατά την εφαρμογή του αλγορίθμου ανακάλυψης αμοιβαίως αποκλειόμενων αντικειμένων και την ερμηνεία των αποτελεσμάτων.

4.5.3 Διακριτοποίηση των Δεδομένων

Οι μετρήσεις των ετικετών που εμφανίζονται στα κελιά του πίνακα των δεδομένων θεωρητικά μπορούν να πάρουν οποιαδήποτε μη αρνητική ακέραια τιμή. Πριν από την εφαρμογή του αλγορίθμου ανακάλυψης αμοιβαίως αποκλειόμενων αντικειμένων είναι απαραίτητη η φάση της διακριτοποίησης κατά την οποία το σύνολο τιμών των χαρακτηριστικών αντιστοιχίζεται σε έναν πεπερασμένο αριθμό κατηγοριών. Στα πλαίσια του συγκεκριμένου προβλήματος χρησιμοποιούνται δύο κατηγορίες, η μια δηλώνεται με την τιμή 0 και η άλλη με την τιμή 1. Η τιμή 0 σημαίνει ότι η έκφραση του γονιδίου είναι σχετικά μικρή, δηλ. το γονίδιο «υποεκφράζεται», ενώ αντίθετα η τιμή 1 σημαίνει ότι το γονίδιο «υπερεκφράζεται». Μετά τη φάση της διακριτοποίησης το σύνολο δεδομένων περιέχει μόνο τις τιμές 0 και 1. Το διακριτοποιημένο σύνολο δεδομένων μπορεί να

θεωρηθεί ότι είναι ένας πίνακας ο οποίος σε κάθε κελί του περιέχει τη λογική τιμή (true/1 ή false/0) που χαρακτηρίζει την ακόλουθη συνθήκη: «το γονίδιο που αντιστοιχεί στη συγκεκριμένη ετικέτα και στη συγκεκριμένη βιβλιοθήκη υπερεκφράζεται».

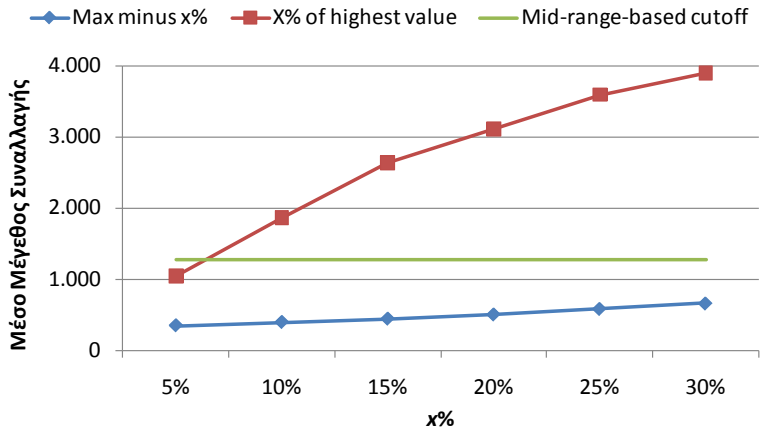
Για τη διακριτοποίηση των δεδομένων χρησιμοποιήθηκαν τρεις διαφορετικές μέθοδοι [Becquet et al., 2002]:

- Η πρώτη μέθοδος διακριτοποίησης ονομάζεται *max minus x%*. Για την εφαρμογή της, αρχικά εντοπίζεται η μεγαλύτερη τιμή έκφρασης, *max*, για κάθε ετικέτα. Στη συνέχεια τα κελιά που η τιμή τους ξεπερνά την τιμή (*max* – *x*)/100 λαμβάνουν την τιμή 1, ενώ τα υπόλοιπα την τιμή 0.
- Η δεύτερη μέθοδος διακριτοποίησης ονομάζεται *mid–range–based cutoff*. Για την εφαρμογή της, αρχικά αναγνωρίζεται η μεγαλύτερη και η μικρότερη τιμή έκφρασης για κάθε ετικέτα και υπολογίζεται ο αριθμητικός τους μέσος. Στη συνέχεια τα κελιά των οποίων η τιμή ξεπερνά τον υπολογισμένο αριθμητικό μέσο λαμβάνουν την τιμή 1, ενώ τα υπόλοιπα την τιμή 0.
- Η τρίτη μέθοδος διακριτοποίησης ονομάζεται *x% of highest value*. Σύμφωνα με τη μέθοδο αυτή λαμβάνεται το *x*% των υψηλότερων τιμών και αντικαθίστανται από την τιμή 1, ενώ οι υπόλοιπες λαμβάνουν την τιμή 0.

Η χρήση αυτών των τεχνικών διακριτοποίησης έχει όχι μόνο τεχνικό ενδιαφέρον, άλλα είναι και βιολογικά ορθή. Το κύριο τεχνικό ενδιαφέρον έγκειται στην απαλοιφή των πολύ χαμηλών τιμών των ετικετών, οι οποίες μπορεί να είναι αποτέλεσμα σφαλμάτων αλληλούχισης. Από βιολογικής άποψης, η διακριτικοποίηση μπορεί να αντιμετωπισθεί σαν ένα πείραμα προσδιορισμού γονιδίων τα οποία υπερεκφράζονται. Ωστόσο, οι διαφορετικοί δυαδικοί πίνακες που παράγονται λόγω των διαφορετικών τεχνικών διακριτοποίησης, θέτουν το πρόβλημα του προσδιορισμού του καλύτερου μαθηματικού ορισμού για την υπερέκφραση των γονιδίων.

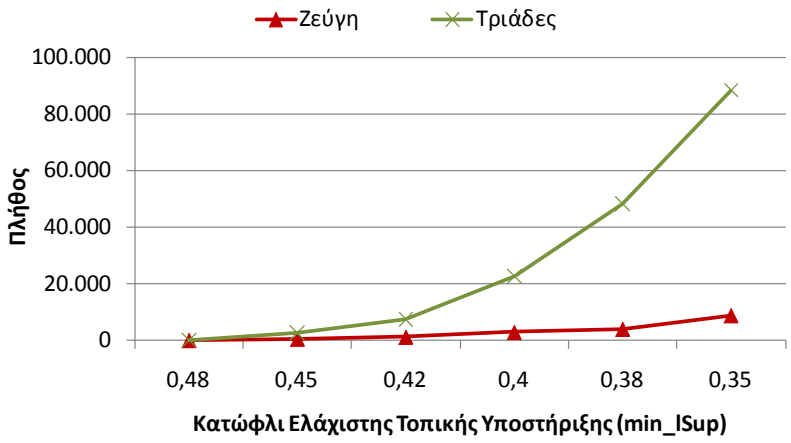
4.5.4 Πειράματα

Η Εικόνα 4.6 παρουσιάζει το μέσο μέγεθος συναλλαγής των συνόλων δεδομένων που προκύπτουν από την εφαρμογή διάφορων παραμετροποιήσεων των τριών μεθόδων διακριτοποίησης που περιγράφηκαν στο Εδάφιο 4.5.3. Όπως φαίνεται οι διάφορες μέθοδοι διακριτοποίησης έχουν διαφορετική ευαισθησία σε αντίστοιχες αλλαγές των παραμέτρων τους. Μια ενδιαφέρουσα παρατήρηση είναι ότι οι μετρήσεις που καταγράφονται για κάθε ετικέτα γονιδίου δεν είναι ομοιόμορφα κατανεμημένες. Για το λόγο αυτό, για παράδειγμα, η μέθοδος *max minus 5%* παρέχει μέσο μέγεθος συναλλαγής 344 ετικετών, ενώ η μέθοδος *5% of highest value* παρέχει μέσο μέγεθος συναλλαγής 1047 ετικετών. Αυτό σημαίνει ότι οι ετικέτες που έχουν το πολύ 5% μικρότερη μέτρηση από τη μέγιστη είναι πολύ λιγότερες από το 5% των ετικετών με τη μεγαλύτερη μέτρηση.



Εικόνα 4.6: Μέσο μέγεθος συναλλαγής σε σχέση με τη μέθοδο διακριτοποίησης.

Η Εικόνα 4.7 παρουσιάζει τον αριθμό των αμοιβαίως αποκλειόμενων ζευγών και τριάδων ετικετών γονιδίων σε σχέση με το κατώφλι ελάχιστης τοπικής υποστήριξης, όταν το κατώφλι ελάχιστης καθολικής υποστήριξης είναι 0,25 και η μέθοδος διακριτοποίησης η mid-range-based cutoff. Όπως φαίνεται στο γράφημα ο αριθμός των αμοιβαίως αποκλειόμενων γονιδίων αυξάνει εκθετικά σε σχέση με το κατώφλι ελάχιστης τοπικής υποστήριξης. Επίσης, ο αριθμός των τριάδων που ανακαλύπτονται είναι σχεδόν πάντοτε μεγαλύτερος από των αριθμό των ζευγών. Αντίστοιχο γράφημα προκύπτει και στην περίπτωση που το κατώφλι ελάχιστης τοπικής υποστήριξης έχει σταθερή τιμή και μεταβάλλεται το κατώφλι ελάχιστης καθολικής υποστήριξης.



Εικόνα 4.7: Πλήθος αμοιβαίως αποκλειόμενων γονιδίων σε σχέση με το κατώφλι ελάχιστης τοπικής υποστήριξης (min_gSup = 0,25, διακριτοποίηση mid-range-based cutoff).

4.6 Συμπεράσματα

Στο κεφάλαιο αυτό προτάθηκε ένα νέο πρόβλημα, που σχετίζεται με την ανακάλυψη αμοιβαίως αποκλειόμενων αντικειμένων ως επέκταση του παραδοσιακού προβλήματος της ανακάλυψης κανόνων συσχέτισης. Η γνώση ότι δύο αντικείμενα είναι αμοιβαίως αποκλειόμενα, μπορεί να αποτελέσει σημαντική πληροφορία κατά την αναζήτηση

άγνωστων ταξονομικών σχέσεων μεταξύ τους. Σε μια τέτοια περίπτωση, αυτό αποτελεί ενδιαφέρουσα γνώση για τον ειδικό του εκάστοτε πεδίου εφαρμογής αλλά και πολύτιμη πληροφορία για την εξόρυξη ιεραρχικών και γενικευμένων κανόνων συσχέτισης. Η συμβολή της παρούσας διατριβής στο αντικείμενο αυτό είναι ο ορισμός του προβλήματος, ο ορισμός μέτρων αμοιβαίου αποκλεισμού και ένας αλγόριθμος ανακάλυψης γνώσης, ο οποίος εφαρμόστηκε σε δεδομένα γονιδιακής έκφρασης.

Πρόβλεψη Φαινοτύπου από Δεδομένα Γονιδιακής Έκφρασης

5

5.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζεται μια μέθοδος που αναπτύχθηκε για την αποτελεσματική ταξινόμηση δεδομένων γονιδιακής έκφρασης. Τα βασικό χαρακτηριστικό των δεδομένων γονιδιακής έκφρασης είναι το μεγάλο πλήθος των χαρακτηριστικών και το σχετικά μικρό πλήθος παραδειγμάτων. Το πλήθος των χαρακτηριστικών, τα οποία στην προκειμένη περίπτωση είναι γονίδια, συνήθως είναι της τάξης των χιλιάδων ή δεκάδων χιλιάδων. Αντίθετα, το πλήθος των παραδειγμάτων συνήθως δεν ξεπερνά τις μερικές εκατοντάδες. Αυτή η ιδιαιτερότητα των δεδομένων γονιδιακής έκφρασης προβάλλει σημαντικές προκλήσεις στους επιστήμονες της ανακάλυψης γνώσης, οι οποίοι καλούνται να αναπτύξουν μεθόδους που διατηρούν υψηλά επίπεδα αποτελεσματικότητας, μολονότι έχουν στη διάθεσή τους ελάχιστα παραδείγματα και πολλά χαρακτηριστικά. Μάλιστα, η πλειοψηφία των χαρακτηριστικών συνήθως δε σχετίζεται με το πρόβλημα που αντιμετωπίζεται, γεγονός που καθιστά επιτακτική την ανάγκη απομάκρυνσης τους.

Στην προτεινόμενη προσέγγιση αξιοποιείται η πληροφορία που παρέχεται από τις συσχετίσεις μεταξύ των γονιδιακών εκφράσεων με σκοπό τη βελτίωση της απόδοσης των ταξινομητών. Η μέθοδος αυτή αξιολογήθηκε με πειράματα που πραγματοποιήθηκαν σε δύο σύνολα δεδομένων. Τα αποτελέσματα των πειραμάτων έδειξαν ότι η μέθοδος παρέχει σημαντική βελτίωση της ακρίβειας ταξινόμησης σε σχέση με άλλες προσεγγίσεις, η οποία σε κάποιες περιπτώσεις προσεγγίζει το 99%.

5.2 Βασικές Έννοιες

Σχεδόν όλα τα κύτταρα ενός οργανισμού περιέχουν το ίδιο γενετικό υλικό και άρα τα ίδια γονίδια. Ωστόσο, διαφορετικά κύτταρα μπορούν να έχουν πολύ διαφορετικές ιδιότητες και λειτουργίες. Αυτό οφείλεται στις διαφορές της συγκέντρωσης των πρωτε-

ϊνών που εμφανίζονται σε κάθε κύτταρο. Η συγκέντρωση μιας πρωτεΐνης που εντοπίζεται σε ένα κύτταρο εξαρτάται σε μεγάλο βαθμό από τον αριθμό των μορίων mRNA τα οποία κωδικοποιούν τη σύνθεση της συγκεκριμένης πρωτεΐνης. Το πλήθος των μορίων μιας mRNA αλληλουχίας με τη σειρά του καθορίζεται από τα επίπεδα έκφρασης του αντίστοιχου γονιδίου. Από τα παραπάνω γίνεται φανερό ότι οι αλλαγές στη γονιδιακή έκφραση αποτελούν τη βάση για πάρα πολλά βιολογικά φαινόμενα. Συνεπώς, η μελέτη της γονιδιακής έκφρασης μπορεί να οδηγήσει σε σημαντικά ευρήματα σχετικά με τη φυσιολογία και την παθολογία των οργανισμών. Ένας από τους βασικούς σκοπούς αυτής της μελέτης είναι η ανακάλυψη διαφορών ανάμεσα στα πρότυπα γονιδιακής έκφρασης από ιστούς ή κύτταρα που παρουσιάζουν φαινοτυπικές διαφορές (π.χ. υγιής ή ασθενής) και η χρήση αυτής της γνώσης για την πρόβλεψη της κατάστασης νέων παραδειγμάτων.

Τα τελευταία χρόνια έχουν αναπτυχθεί δύο βασικές οικογένειες που επιτρέπουν τη μέτρηση των επιπέδων γονιδιακής έκφρασης. Η πρώτη περιλαμβάνει τεχνικές που βασίζονται στην αλληλούχιση, π.χ. η τεχνική SAGE [Velculescu et al., 1995] και η δεύτερη περιλαμβάνει τεχνικές που βασίζονται στην υβριδοποίηση, π.χ. μικροσυστοιχίες [Schena et al., 1995]. Περισσότερες λεπτομέρειες για αυτές τις τεχνικές αναφέρονται στην ενότητα 3.10. Στο παρόν κεφάλαιο θα χρησιμοποιηθούν δεδομένα που προέρχονται από την τεχνική SAGE (Εδάφιο 3.10.2) με την οποία καταγράφεται ο αριθμός των μεταγράφων mRNA που υπάρχουν σε μια δεδομένη στιγμή σε ένα κύτταρο. Κάθε μετάγραφο mRNA χαρακτηρίζεται από μια μικρή αλληλουχία που καλείται *ετικέτα* (*tag*). Το σύνολο των ετικετών που παράγονται σχηματίζουν μια εγγραφή που καλείται *βιβλιοθήκη* (*library*).

Η τεχνική SAGE είναι ακριβότερη σε σχέση με τις μικροσυστοιχίες, ωστόσο παρέχει τη δυνατότητα ευρύτερης μελέτης του γονιδιώματος. Σχετικά πρόσφατα, (2004 και 2005) τα δεδομένα SAGE αποτέλεσαν αντικείμενο μελέτης των ECML/PKDD Discovery Challenge Workshops, γεγονός που έδωσε κάποια ώθηση στη έρευνα αυτών των δεδομένων. Ωστόσο, απόδοση των προσεγγίσεων που έχουν προταθεί παρέμεινε μέτρια. Αυτό αποτέλεσε ένα ισχυρό κίνητρο για τη μελέτη του προβλήματος και την ανάπτυξη της μεθόδου που παρουσιάζεται στο κεφάλαιο αυτό.

Ένα επιπλέον πλεονέκτημα που παρέχει η τεχνική SAGE είναι ότι δε χρειάζεται να γίνει επιλογή συγκεκριμένων αλληλουχιών mRNA προς καταμέτρηση. Αυτό παρέχει μια σημαντική ευελιξία στη μελέτη αντικειμένων, όπως ο καρκίνος, για τα οποία είναι λίγα πράγματα γνωστά εκ των προτέρων. Αντίθετα, στη μέθοδο των μικροσυστοιχιών, πρέπει να γίνει επιλογή των αλληλουχιών mRNA, γεγονός που εισάγει κάποιου βαθμού πόλωση στην πειραματική διαδικασία. Αυτό το χαρακτηριστικό της τεχνικής SAGE, την κάνει μια υποσχόμενη μέθοδο για τη μελέτη της φυσιολογίας και της παθολογίας των οργανισμών ειδικά για περιπτώσεις στις οποίες δεν υπάρχει σημαντικό υπόβαθρο γνώσης, όπως είναι ο καρκίνος. Τα δεδομένα που χρησιμοποιούνται στο παρόν κεφάλαιο της διατριβής αφορούν διάφορες μορφές καρκίνου.

5.3 Σχετικές Εργασίες

Στην εργασία [Gamberoni & Storari, 2004] μέθοδοι μηχανικής μάθησης με επίβλεψη και χωρίς επίβλεψη χρησιμοποιήθηκαν με σκοπό την ανάλυση των δεδομένων γονιδιακής έκφρασης που λήφθηκαν με την τεχνική SAGE. Στη συγκεκριμένη εργασία χρησιμοποίησαν το σύνολο δεδομένων 90×27679, που έχει χρησιμοποιηθεί και στην παρούσα διατριβή (Εδάφιο 5.5.1). Συγκεκριμένα, χρησιμοποίησαν δένδρα ταξινόμησης (αλγόριθμος C4.5) και μηχανές διανυσμάτων υποστήριξης για την ταξινόμηση των δεδομένων βάσει της κυτταρικής κατάστασης του δείγματος (φυσιολογικό ή καρκινικό) και του οργάνου (κόλον, εγκέφαλος, ωοθήκη, κλπ.). Στην πρώτη περίπτωση η ακρίβεια της ταξινόμησης που επετεύχθη είναι 74,4% και 82,2% για το δένδρο ταξινόμησης και τη μηχανή διανυσμάτων υποστήριξης, αντίστοιχα. Στην περίπτωση ταξινόμησης βάσει του οργάνου η ακρίβεια της ταξινόμησης που επετεύχθη είναι αρκετά χειρότερη (48,9% και 71,1% για το δένδρο ταξινόμησης και τη μηχανή διανυσμάτων υποστήριξης, αντίστοιχα). Τέλος, οι συγγραφείς της εργασίας μελέτησαν την εφαρμογή μεθόδων ιεραρχικής ομαδοποίησης για την αναγνώριση διαφόρων τύπων καρκινικών και φυσιολογικών ιστών.

Ιεραρχικές μέθοδοι ομαδοποίησης χρησιμοποιήθηκαν και στην εργασία [Ng et al., 2001] για τον εντοπισμό ομοιοτήτων και διαφορών μεταξύ των διαφορετικών τύπων καρκίνου. Οι ερευνητές χρησιμοποίησαν διάφορες τεχνικές προεπεξεργασίας, συμπεριλαμβανομένων της διαγραφής σφαλμάτων, της κανονικοποίησης, της συμπλήρωσης ελλιπών τιμών και της επιλογής υποχώρου που βασίζεται στη δοκιμασία Wilcoxon. Τα αποτελέσματα έδειξαν ότι τα δείγματα (βιβλιοθήκες SAGE) ομαδοποιούνται βάσει του οργάνου και της κυτταρικής κατάστασης. Επιπλέον, αποκάλυψαν μια πιθανή συσχέτιση μεταξύ καρκίνου του εγκεφάλου και των ωοθηκών.

Στην εργασία [Rioult, 2004] παρουσιάστηκε μια μέθοδος για την εξόρυξη ισχυρών αναδυόμενων προτύπων από δεδομένα γονιδιακής έκφρασης που ελήφθησαν με την τεχνική SAGE. Η προσέγγιση αυτή βασίστηκε στην αντιμετάθεση του πίνακα δεδομένων και τη χρήση της σύνδεσης Galois για την παραγωγή κλειστών συνόλων του αρχικού πίνακα και τελικά, την εξαγωγή ισχυρών αναδυόμενων προτύπων. Τα σύνολα δεδομένων που χρησιμοποιήθηκαν σε αυτήν την εργασία (90×27679 και 74×822) χρησιμοποιούνται και στη διατριβή (Εδάφιο 5.5.1).

Η εργασία των Becquet et al. [2002] περιλαμβάνει την εξόρυξη κανόνων συσχέτισης από δεδομένα SAGE. Μελετήθηκαν διάφορες μέθοδοι διακριτοποίησης, ο σκοπός των οποίων είναι ο μετασχηματισμός του πίνακα δεδομένων σε δυαδικές τιμές, ώστε να καταστεί δυνατή η εξόρυξη των κανόνων συσχέτισης. Ανάλογα με τη μέθοδο διακριτοποίησης που επιλέγεται να χρησιμοποιηθεί, δίνεται βάρος σε διαφορετικές ιδιότητες των δεδομένων. Σε κάθε περίπτωση, τα εξαγόμενα σύνολα κανόνων συσχέτισης αποκαλύπτουν μια πολύ ισχυρή συσχέτιση της γονιδιακής ρύθμισης μορίων mRNA που κωδικοποιούν ριβοσωματικές πρωτεΐνες. Άλλες σχετικές μέθοδοι που περιλαμβάνουν

την εξόρυξη συχνών κλειστών συνόλων αντικειμένων [Gasmi et al., 2005] και δ-ισχυρών (δ-strong) κανόνων χαρακτηρισμού [Hebert et al., 2005] προτάθηκαν και εφαρμόστηκαν σε δεδομένα γονιδιακής έκφρασης που έχουν ληφθεί με την τεχνική SAGE (σύνολο δεδομένων 90×27679).

Στην εργασία [Alves et al., 2005] μελετήθηκαν διάφορες μέθοδοι μείωσης διαστάσεων χρησιμοποιώντας το σύνολο δεδομένων SAGE 74×822. Οι συγγραφείς συμπέραναν ότι οι τυπικές μέθοδοι διήθησης (filtering) επηρεάζουν αρνητικά την αποτελεσματικότητα των ταξινομητών. Επίσης, παρατήρησαν ότι πολλά γονίδια τα οποία δεν παρουσιάζουν σημαντική διαφοροποίηση της έκφρασής τους μεταξύ των παραδειγμάτων διαφορετικών τάξεων, μπορούν να συνεισφέρουν σημαντικά στην ταξινόμηση. Χρησιμοποιώντας τον αλγόριθμο GRAD [Zagoruiko et al., 2005] για επιλογή χαρακτηριστικών και τον C4.5 για κατασκευή δένδρου ταξινόμησης κατασκεύασαν ταξινομητή η ακρίβεια του οποίου φτάνει το 86,18%.

Οι Martinez et al. [2005] μελέτησαν μεθόδους καθαρισμού δεδομένων και σχολίασαν τη διαδικασία της απόδοσης μιας ετικέτας σε ένα γονίδιο. Εφάρμοσαν τη μέθοδο ανάλυσης κυρίων συνιστωσών (PCA) και μεθόδους ιεραρχικής ομαδοποίησης στο σύνολο δεδομένων 74×822. Το βασικό συμπέρασμά τους είναι ότι η σύγκριση καρκινικών όγκων προερχόμενων από διάφορους τύπους ιστών είναι εξαιρετικά δύσκολη διαδικασία, καθώς τα δείγματα ομαδοποιούνται κυρίως βάσει της προέλευσης του ιστού και όχι βάσει της κατάστασης των κυττάρων του (φυσιολογικά ή καρκινικά).

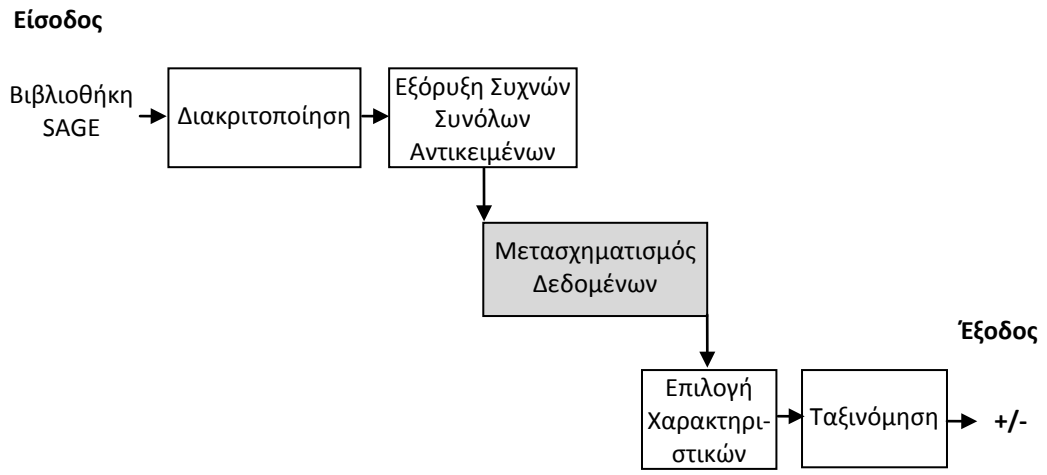
Η εργασία [Lin & Li, 2005] παρουσιάζει μια ανάλυση των δεδομένων SAGE (σύνολο δεδομένων 90×27679) χρησιμοποιώντας διάφορες μεθόδους αξιολόγησης χαρακτηριστικών, ταξινόμησης και υπολογισμού σφάλματος. Τα αποτελέσματά τους έδειξαν ότι οι μηχανές διανυσμάτων υποστήριξης παρουσιάζουν καλή απόδοση στα συγκεκριμένα δεδομένα (14,4% σφάλμα με 10-πλή σταυρωτή επικύρωση και 16,1% σφάλμα με σταυρωτή επικύρωση με μια παράλειψη). Επιπλέον, κατέληξαν στο συμπέρασμα ότι περίπου 500 με 1.000 χαρακτηριστικά είναι αρκετά για την πρόβλεψη της τάξης (καρκινικό ή φυσιολογικό) ενός στιγμιότυπου.

Ένα υβριδικό σύστημα βασισμένο σε γενετικούς αλγορίθμους και νευρωνικά δίκτυα προτάθηκε για την ταξινόμηση βιβλιοθηκών SAGE [Esseghir et al., 2005]. Το σύστημα βασίζεται στην επιλογή ενός συμπαγούς συνόλου γονιδίων που προβλέπουν την τάξη των δεδομένων.

Τέλος, μέθοδοι ιεραρχικής ομαδοποίησης και ομαδοποίησης βασισμένης σε διαχωρισμούς (αλγόριθμος *K*-Means), καθώς και διάφορα κριτήρια επικύρωσης ομάδων μελετήθηκαν στα πλαίσια της παρούσας διατριβής και τα συμπεράσματα παρουσιάστηκαν στην εργασία [Tzanis & Vlahavas, 2007a]. Τα δύο σύνολα δεδομένων που χρησιμοποιήθηκαν στη συγκεκριμένη μελέτη (90×27679 και 74×822) παρουσιάζονται αναλυτικά στο Εδάφιο 5.5.1.

5.4 Η Προτεινόμενη Προσέγγιση

Στην ενότητα αυτή περιγράφεται η προτεινόμενη προσέγγιση, η οποία αποτελείται από τέσσερα στάδια, τη διακριτοποίηση, την εξόρυξη συχνών προτύπων (δηλ. συχνών συνόλων αντικειμένων), την επιλογή χαρακτηριστικών και την ταξινόμηση (Εικόνα 5.1). Για την καλύτερη κατανόηση της λειτουργίας της προτεινόμενης μεθόδου είναι απαραίτητο πριν από την παρουσίαση των σταδίων που την αποτελούν να περιγραφεί η δομή των δεδομένων εισόδου.



Εικόνα 5.1: Η αρχιτεκτονική της προτεινόμενης μεθόδου.

5.4.1 Δομή Δεδομένων Εισόδου

Τα δεδομένα αναπαρίστανται από έναν πίνακα *A* διαστάσεων *M*×*N*. Οι στήλες του πίνακα αντιπροσωπεύουν τις ετικέτες (tags) των *N* γονιδίων, ενώ οι γραμμές αντιπροσωπεύουν τα *M* δείγματα (βιβλιοθήκες SAGE). Η τιμή a_{ij} που περιλαμβάνεται στο κελί του πίνακα το οποίο αποτελεί την τομή της γραμμής *i* και της στήλης *j* αντιπροσωπεύει το επίπεδο έκφρασης του γονιδίου *j* στο δείγμα *i*. Κάθε δείγμα *i* σχετίζεται με μια ετικέτα (label) c_i, η οποία προσδιορίζει την τάξη στην οποία ανήκει το δείγμα. Στα πλαίσια της διατριβής $c_i \in \{+, -\}$, όπου το «+» δηλώνει την καρκινική κυτταρική κατάσταση, ενώ το «-» δηλώνει τη φυσιολογική. Επιπλέον, οι τιμές a_{ij} ανήκουν στο σύνολο των φυσικών αριθμών καθώς αποτελούν καταμετρήσεις των αντίστοιχων ετικετών. Τη δομή του πίνακα των δεδομένων παρουσιάζει ο Πίνακας 5.1.

Πίνακας 5.1: Δομή πίνακα δεδομένων SAGE.

	Ετικέτα 1	Ετικέτα 2	…	Ετικέτα *N*	Τάξη
Βιβλιοθήκη 1	a_{11}	a_{12}	…	a_{1N}	c_1
Βιβλιοθήκη 2	a_{21}	a_{22}	…	a_{2N}	c_2
…	…	…	…	…	…
Βιβλιοθήκη *M*	a_{M1}	a_{M2}	…	a_{MN}	c_M

5.4.2 Διακριτοποίηση

Στο στάδιο αυτό πραγματοποιείται η διακριτοποίηση των δεδομένων, η οποία είναι σημαντική για τους δύο παρακάτω λόγους:

- Για τον εντοπισμό ισχυρών γονιδιακών υπό-εκφράσεων (επίπεδα γονιδιακών εκφράσεων που είναι σημαντικά χαμηλότερα από τη μέση γονιδιακή έκφραση) και υπερ-εκφράσεων (επίπεδα γονιδιακών εκφράσεων που είναι σημαντικά υψηλότερα από τη μέση γονιδιακή έκφραση).
- Για το μετασχηματισμό των δεδομένων, ώστε να είναι δυνατή η εφαρμογή ενός αλγορίθμου εξόρυξης συχνών συνόλων αντικειμένων.

Η διαδικασία της διακριτοποίησης εφαρμόζεται ως εξής. Αρχικά, ο πίνακας δεδομένων A διαιρείται σε δύο υποπίνακες A_+ ($M_1 \times N$) και A_- ($M_2 \times N$) που περιέχουν τα δείγματα των καρκινικών και φυσιολογικών κυτταρικών καταστάσεων, αντίστοιχα ($M_1 + M_2 = M$). Στη συνέχεια, για κάθε υποπίνακα υπολογίζεται ένα 99% διάστημα εμπιστοσύνης για τα επίπεδα έκφρασης του κάθε γονιδίου. Συγκεκριμένα, για κάθε γονίδιο j υπολογίζονται δύο διαστήματα εμπιστοσύνης $[L(j, +), R(j, +)]$ και $[L(j, -), R(j, -)]$ για την καρκινική και τη φυσιολογική κυτταρική κατάσταση, αντίστοιχα. Τέλος, δημιουργούνται δύο νέοι πίνακες A'_+ και A'_-, ίδιων διαστάσεων με τους A_+ και A_-, όπου $a'_{+,ij}, a'_{-,ij} \in \{-1, 0, +1\}$. Οι τιμές αυτές ανατίθενται ως εξής:

- $a'_{c_i,ij} = -1$, αν $a_{ij} \in (0, L(j, c_i))$.
- $a'_{c_i,ij} = 0$, αν $a_{ij} \in [L(j, c_i), R(j, c_i)]$.
- $a'_{c_i,ij} = +1$, αν $a_{ij} \in (R(j, c_i), +\infty)$.

Η ανάθεση της τιμής -1 στο $a'_{c_i,ij}$ δηλώνει ότι το γονίδιο j υπο-εκφράζεται σημαντικά στο δείγμα i σε σχέση με τα επίπεδα έκφρασης του j στα δείγματα της τάξης c_i. Αντίστοιχα, η ανάθεση της τιμής +1 στο $a'_{c_i,ij}$ δηλώνει ότι το γονίδιο j υπερ-εκφράζεται σημαντικά στο δείγμα i σε σχέση με τα επίπεδα έκφρασης του j στα δείγματα της τάξης c_i. Η τιμή 0 δηλώνει ότι δεν υπάρχει σημαντική υπο-έκφραση ή υπερ-έκφραση του συγκεκριμένου γονιδίου.

Αξίζει να σημειωθεί ότι οι μηδενικές τιμές έκφρασης δεν υπολογίζονται ως υποεκφράσεις. Αυτό γίνεται για δύο λόγους. Πρώτον, μια μηδενική τιμή μπορεί να σημαίνει ότι η ετικέτα δε βρέθηκε στο δείγμα, οπότε το αντίστοιχο γονίδιο όχι απλά υποεκφράζεται, αλλά δεν εκφράζεται καθόλου. Όπως αναφέρεται και στην εργασία [Ng et al., 2001], οι βιοχημικοί υποστηρίζουν ότι η μεγάλη πλειοψηφία των γονιδίων του ανθρώπου εκφράζονται μόνο σε έναν τύπο κυττάρου και μόνο κάποια διαχειριστικά γονίδια (housekeeping genes) εκφράζονται σε όλα τα κύτταρα. Σύμφωνα με αυτήν την άποψη είναι πολύ πιθανό ότι ένα γονίδιο που παρουσιάζει μηδενική έκφραση σε ένα συγκεκριμένο δείγμα να μην εκφράζεται στο συγκεκριμένο τύπο ιστού ή κυττάρου από

τον οποίο ελήφθη το δείγμα. Επομένως, θα ήταν ανακριβές να θεωρηθεί στη συγκεκριμένη περίπτωση το γονίδιο ως υπο-εκφραζόμενο. Δεύτερον, υπάρχουν πολλές μηδενικές τιμές σε πίνακες δεδομένων γονιδιακής έκφρασης. Για παράδειγμα, στο σύνολο δεδομένων 99×27679 το 45% των τιμών είναι μηδενικές. Η παρουσία πολλών μηδενικών γονιδιακών εκφράσεων προφανώς οφείλεται και στη μη έκφραση όλων των γονιδίων σε όλους τους ιστούς. Αν οι μηδενικές τιμές εκλαμβάνονταν ως υπο-εκφράσεις, εκτός από την ανακρίβεια που θα υπεισερχόταν, θα αυξάνονταν σημαντικά και οι υπολογιστικές απαιτήσεις. Επιπλέον, οι πραγματικές υπο-εκφράσεις, που ενδεχομένως κρύβουν πολύτιμη πληροφορία, θα χάνονταν ανάμεσα στη μεγάλη πλειοψηφία των ανακριβών μηδενικών υπο-εκφράσεων.

Οι πίνακες A'_+ και A'_- αποτελούν την είσοδο του επόμενου σταδίου που περιλαμβάνει την εξόρυξη των συχνών συνόλων αντικειμένων (εκφράσεις γονιδίων).

5.4.3 Εξόρυξη Συχνών Συνόλων Αντικειμένων

Στο στάδιο αυτό χρησιμοποιούνται οι διακριτοποιημένοι πίνακες δεδομένων με σκοπό την ανακάλυψη συχνών συνόλων αντικειμένων για κάθε τάξη (κυτταρική κατάσταση) ξεχωριστά. Στο στάδιο αυτό μπορεί να γίνει χρήση οποιουδήποτε αλγορίθμου εξόρυξης συχνών συνόλων αντικειμένων, όπως για παράδειγμα οι αλγόριθμοι Apriori και FPGrowth (Εδάφιο 2.3.2). Έπειτα από την εφαρμογή του αλγορίθμου εξόρυξης συχνών συνόλων αντικειμένων προκύπτουν δύο σύνολα από συχνά σύνολα αντικειμένων, τα σύνολα F_+ και F_-, που αντιστοιχούν στους πίνακες A'_+ και A'_-. Από αυτά τα δύο σύνολα δημιουργείται ένα νέο σύνολο από συχνά σύνολα αντικειμένων F, τέτοιο ώστε να ισχύει $F = F_+ \cup F_- - F_+ \cap F_-$. Το σύνολο F περιλαμβάνει εκείνα τα συχνά σύνολα αντικειμένων που είναι συχνά μόνο στη μία εκ των δύο τάξεων. Η λογική του αποκλεισμού των συχνών συνόλων αντικειμένων που είναι κοινά και για τις δύο τάξεις είναι ότι τα συγκεκριμένα συχνά σύνολα προφανώς στερούνται σημαντικής ισχύος διαχωρισμού των δειγμάτων που ανήκουν σε διαφορετικές τάξεις.

Το σύνολο F αποτελεί το νέο σύνολο χαρακτηριστικών που θα χρησιμοποιηθεί για την περιγραφή των δεδομένων. Αυτό σημαίνει ότι οι αρχικές ετικέτες των γονιδίων θα αντικατασταθούν από συχνά σύνολα υπό-εκφραζόμενων και/ή υπερ-εκφραζόμενων γονιδίων. Αυτό επιτυγχάνεται με τη δημιουργία ενός νέου πίνακα δεδομένων A' διαστάσεων $M \times |F|$. Οι στήλες του πίνακα A' αναπαριστούν τα συχνά σύνολα αντικειμένων, τα οποία είναι στοιχεία του συνόλου F και οι γραμμές αναπαριστούν τα διάφορα δείγματα και των δύο τάξεων. Το κελί που βρίσκεται στην τομή της γραμμής i και της στήλης j, δηλαδή το στοιχείο $a'_{ij} \in \{0, 1\}$, δηλώνει την παρουσία (1) ή την απουσία (0) του συχνού συνόλου αντικειμένων $j \in F$ στο δείγμα i.

Σε αυτό το σημείο πρέπει να σημειωθεί ότι μια ετικέτα που εμφανίζεται με συχνότητα 1 σε μια βιβλιοθήκη SAGE και με συχνότητα 0 σε όλες τις υπόλοιπες, είναι κατά πάσα πιθανότητα το προϊόν σφάλματος κατά τη διαδικασία αλληλούχισης. Λόγω αυτού

του σφάλματος δεν προσδιορίζεται η πραγματική ετικέτα, άλλα μια διαφορετική, που πιθανώς μοιάζει πολύ με τη σωστή και δεν ταυτίζεται με καμιά άλλη ετικέτα. Αντίστοιχο παράδειγμα αποτελούν τα ορθογραφικά λάθη που εμφανίζονται σε κείμενα και τις περισσότερες φορές δεν ταυτίζονται με καμία γνωστή λέξη. Τέτοιου είδους ετικέτες δεν προσφέρουν καμιά πληροφορία σχετικά με το πρόβλημα που μελετάται και πρέπει να αποκλείονται. Συνεπώς, αυτές οι ετικέτες δεν θα πρέπει να σχετισθούν με υπό-εκφράσεις ή υπερ-εκφράσεις κάποιου γονιδίου. Η πρώτη περίπτωση δεν μπορεί ποτέ να συμβεί, καθώς οι μηδενικές εκφράσεις δεν θεωρούνται υπο-εκφράσεις, ενώ η γονιδιακή έκφραση με τιμή 1, δεδομένου ότι οι εκφράσεις του γονιδίου στα υπόλοιπα δείγματα είναι μηδενικές, επίσης, δεν μπορεί να εκληφθεί ως υπό-έκφραση. Τέλος, το στάδιο εξόρυξης συχνών συνόλων αντικειμένων εξασφαλίζει ότι μια τέτοια ετικέτα δε θα θεωρηθεί ότι αντιστοιχεί σε υπερ-έκφραση γονιδίου, καθώς η θεωρούμενη υπερ-έκφραση απαιτείται να είναι συχνή, δηλαδή να εμφανίζεται σε αρκετά δείγματα.

5.4.4 Επιλογή Χαρακτηριστικών

Σε αυτό στο στάδιο πραγματοποιείται η επιλογή των χαρακτηριστικών. Από τα δεδομένα του πίνακα *A΄* επιλέγονται εκείνα τα χαρακτηριστικά (συχνά σύνολα υπο-εκφράσεων και/η υπερ-εκφράσεων) που παρουσιάζουν καλύτερη ικανότητα διαχωρισμού των δειγμάτων των δύο τάξεων (φυσιολογική ή καρκινική κυτταρική κατάσταση). Πρέπει να σημειωθεί ότι η εφαρμογή της μεθόδου επιλογής χαρακτηριστικών στο μετασχηματισμένο πίνακα *A΄* απαιτεί την κατανάλωση πολύ λιγότερων πόρων και χρόνου από ότι η εφαρμογή του στον αρχικό πίνακα δεδομένων *A*. Αυτό οφείλεται στο γεγονός ότι συνήθως ο αριθμός των ετικετών είναι της τάξης των δεκάδων ή ακόμη και εκατοντάδων χιλιάδων, ενώ ο αριθμός των συχνών συνόλων είναι τυπικά της τάξης των εκατοντάδων ή χιλιάδων. Βέβαια, πρέπει να σημειωθεί ότι και ο αριθμός των συχνών συνόλων θα μπορούσε να είναι εξαιρετικά μεγάλος στην περίπτωση που χρησιμοποιηθεί ένα ιδιαίτερα χαμηλό κατώφλι ελάχιστης υποστήριξης. Ωστόσο, συνήθως δεν υπάρχει λόγος να γίνει κάτι τέτοιο, καθώς μερικές εκατοντάδες ή χιλιάδες από τα συχνότερα σύνολα αντικειμένων είναι αρκετά.

5.4.5 Ταξινόμηση

Κατά τη διάρκεια του τελευταίου σταδίου πραγματοποιείται δημιουργία ταξινομητή που εκπαιδεύεται με τα δεδομένα που έχουν προκύψει από τα προηγούμενα στάδια. Εννοείται ότι και τα στιγμιότυπα που πρόκειται να ταξινομηθούν υπόκεινται ακριβώς την ίδια διαδικασία επεξεργασίας πριν εισέλθουν ως είσοδος στον ταξινομητή. Ο σκοπός του ταξινομητή είναι να κατατάξει κάθε εισερχόμενο δείγμα σε μια από της δύο τάξεις, οι οποίες είναι η φυσιολογική και η καρκινική κυτταρική κατάσταση.

Ο Πίνακας 5.2 παρουσιάζει τον αλγόριθμο βάσει του οποίου υλοποιείται η προτεινόμενη προσέγγιση, συνοψίζοντας όλα όσα περιγράφηκαν στις προηγούμενες παραγράφους.

Πίνακας 5.2: Ο αλγόριθμος για την υλοποίηση της προτεινόμενης προσέγγισης.

Είσοδος: Ένας πίνακας δεδομένων γονιδιακής έκφρασης *A* διαστάσεων *M*×*N*, μια μέθοδος εξόρυξης συχνών συνόλων αντικειμένων (*FreqItemsetsMiningMethod*), ένα κατώτατο όριο υποστήριξης (*min_sup*), μια μέθοδος επιλογής χαρακτηριστικών (*FeatureSelectionMethod*) και μια μέθοδος ταξινόμησης (*ClassificationMethod*).

Έξοδος: Οι παράμετροι που προσδιορίζουν το σύστημα πρόβλεψης.

```
A+<M+×N>, A-<M-×N> ← classBasedDivisionOfSamples(A)
for each (c ∈{+,-})
  for (j = 1:N)
    [L_{c,j}, H_{c,j}] ← calculateConfidenceInterval(A_c[:,j], 99%)
for each (c ∈ {+,-})
  for (i = 1:M_c)
    for (j = 1:N)
      if (A_c[i,j] < L_{c,j})
        A´_c[i,j] = -1
      else if (A_c[i,j] > H_{c,j})
        A´_c[i,j] = +1
      else
        A´_c[i,j] = 0
for each (c ∈ {+,-})
  F_c ← mineFrequentItemsets(A´_c, FreqItemsetsMiningMethod, min_sup)
F ← F+ ∪ F- - F+ ∩ F-
k, l ← 0
for each (c ∈ {+,-})
  for (i = l+1:l+M_c)
    k ← k + 1
    for (j = 1:|F|)
      if (F[j] ⊆ A´_c[i,:])
        A´[i,j] = 1
      else
        A´[i,j] = 0
  l ← l + k
F´ ← selectFeatures(A´, FeatureSelectionMethod)
A´´<M×|F´|> ← selectFeatures(A´, F´)
C ← buildClassifier(A´´, ClassificationMethod)
return < ⋃_{j=1}^{N} [L_{+,j}, H_{+,j}], ⋃_{j=1}^{N} [L_{-,j}, H_{-,j}], F´, C >
```

5.5 Πειραματική Αξιολόγηση

Στην ενότητα αυτή παρουσιάζονται τα πειράματα που διεξήχθησαν με σκοπό την αξιολόγηση της προτεινόμενης προσέγγισης. Στα πλαίσια της διεξαγωγής των πειραμάτων χρησιμοποιήθηκαν κάποια σύνολα δεδομένων γονιδιακής έκφρασης που ελήφθησαν με την τεχνική SAGE, τα οποία, επίσης, παρουσιάζονται σε αυτήν την ενότητα.

5.5.1 Σύνολα Δεδομένων

Στα πλαίσια της διατριβής χρησιμοποιήθηκαν δύο σύνολα δεδομένων γονιδιακής έκφρασης τα οποία έχουν ληφθεί με την τεχνική SAGE. Το πρώτο σύνολο δεδομένων

(90×27679) περιλαμβάνει 90 παραδείγματα (βιβλιοθήκες SAGE) και 27.679 χαρακτηριστικά (ετικέτες). Το δεύτερο σύνολο δεδομένων (74×822) περιλαμβάνει 74 βιβλιοθήκες SAGE και 822 ετικέτες. Τα δεδομένα αυτά έχουν συγκεντρωθεί το Δεκέμβριο του 2002 από την ερευνητική ομάδα του Dr. Olivier Gandrillon (Centre de Génétique Moléculaire et Cellulaire de Lyon) και έχουν παρουσιαστεί και μελετηθεί στα ECML/PKDD Discovery Challenge Workshops κατά τα έτη 2004 και 2005 [Gandrillon, 2004]. Τα δεδομένα και των δύο συνόλων, όπως έχει ήδη περιγραφεί στο Εδάφιο 4.5.2, συνοδεύει ένας μεγάλος αριθμός χαρακτηριστικών. Από τα χαρακτηριστικά αυτά το μεγαλύτερο ενδιαφέρον συγκεντρώνουν το όργανο από τον οποίο ελήφθη το δείγμα (νεφρός, μαστός, λέμφος, προστάτης, ωοθήκης, εγκέφαλος, έντερο, πάγκρεας, παρεγκεφαλίδα, αγγείο, δέρμα, θάλαμος εγκεφάλου, περιτόναιο) και η βιολογική κατάσταση του δείγματος (φυσιολογικό, καρκινικό).

5.5.2 Εξόρυξη Συχνών Συνόλων Αντικειμένων

Για την εξόρυξη των συχνών συνόλων αντικειμένων μπορεί να χρησιμοποιηθεί οποιοσδήποτε αλγόριθμος εξόρυξης συχνών συνόλων αντικειμένων, π.χ. Apriori ή FPGrowth (Εδάφιο 2.3.2). Όλοι οι σχετικοί αλγόριθμοι παράγουν τα ίδια αποτελέσματα, εφόσον χρησιμοποιηθούν οι ίδιες παράμετροι. Η διαφορά τους παρουσιάζεται στην αποδοτικότητά (δηλ. το χρόνο εκτέλεσης) τους, η οποία δεν αποτελεί σημαντικό παράγοντα για τα δεδομένα που μελετώνται στο κεφάλαιο αυτό. Ο αλγόριθμος που χρησιμοποιήθηκε για τη διεξαγωγή των πειραμάτων είναι ο FPGrowth. Καθώς το ενδιαφέρον στην παρούσα μελέτη είναι στραμμένο στην αποτελεσματικότητα διάφορων ταξινομητών που εκπαιδεύονται με τη χρήση διαφορετικών αριθμών χαρακτηριστικών (δηλ. συνόλων αντικειμένων), η παράμετρος της υποστήριξης επιλέχθηκε με σκοπό την εξαγωγή ικανού αριθμού συχνών συνόλων αντικειμένων. Οι τιμές της υποστήριξης που χρησιμοποιήθηκαν για την εξόρυξη των συχνών συνόλων αντικειμένων σε όλες της περιπτώσεις κυμαίνεται από 0,35 έως 0,45.

5.5.3 Επιλογή Χαρακτηριστικών

Για τη διεξαγωγή των πειραμάτων χρησιμοποιήθηκε η βιβλιοθήκη αλγορίθμων μηχανικής μάθησης Weka [Witten & Frank, 2005]. Συγκεκριμένα, για την επιλογή των καταλληλότερων χαρακτηριστικών μελετήθηκαν τρία δημοφιλή μέτρα αξιολόγησης της πληροφορίας. Τα μέτρα αυτά παρέχουν μια κατάταξη των χαρακτηριστικών βάσει της πληροφορίας που τα χαρακτηριστικά παρέχουν για το διαχωρισμό των παραδειγμάτων των δύο τάξεων. Τα μέτρα αυτά είναι τα εξής:

- Μέτρο X^2. Αξιολογεί ένα χαρακτηριστικό υπολογίζοντας την τιμή του στατιστικού X^2 σε σχέση με την τάξη των παραδειγμάτων.
- Μέτρο κέρδους πληροφορίας. Αξιολογεί ένα χαρακτηριστικό υπολογίζοντας το κέρδος πληροφορίας (information gain) σε σχέση με την τάξη των παραδειγμάτων.

- Μέτρο *Relief F*. Αξιολογεί ένα χαρακτηριστικό πραγματοποιώντας επαναλαμβανόμενες δειγματοληψίες ενός παραδείγματος και λαμβάνοντας υπόψη τις τιμές του χαρακτηριστικού για το κοντινότερο παράδειγμα της ίδιας και της αντίθετης τάξης [Kira, 1992; Kononenko, 1994].

5.5.4 Ταξινόμηση

Στα πλαίσια της πειραματικής αξιολόγησης χρησιμοποιήθηκαν 5 αλγόριθμοι ταξινόμησης για την κατασκευή των ταξινομητών που περιγράφονται παρακάτω (οι τέσσερεις πρώτοι υλοποιήθηκαν από αλγορίθμους που περιλαμβάνονται στη βιβλιοθήκη Weka). Για τη δημιουργία των ταξινομητών χρησιμοποιήθηκαν οι προκαθορισμένες παράμετροι της βιβλιοθήκης Weka, έκτος αν αναφέρεται κάτι διαφορετικό:

- Δένδρο ταξινόμησης (αλγόριθμος C4.5) [Quinlan, 1993].
- Ταξινομητής *k*-Κοντινότερων γειτόνων (*k*-Nearest Neighbours – *k*NN) [Aha et al., 1991]. Ο αριθμός των κοντινότερων γειτόνων ορίστηκε *k*=1.
- Μηχανή διανυσμάτων υποστήριξης (SVM). Για τη δημιουργία αυτού του ταξινομητή χρησιμοποιήθηκε ο αλγόριθμος που παρέχει η βιβλιοθήκη Weka (Sequential Minimal Optimization – SMO [Platt, 1998; Keerthi et al., 2001]).
- Ταξινομητής προτασιακών κανόνων ταξινόμησης που δημιουργήθηκε με τον αλγόριθμο RIPPER (Repeated Incremental Pruning to Produce Error Reduction) [Cohen, 1995].
- Ταξινομητής πλειοψηφίας, ο οποίος ταξινομεί κάθε στιγμιότυπο στην τάξη που ανήκει η πλειοψηφία των παραδειγμάτων του συνόλου εκπαίδευσης. Ο ταξινομητής αυτός αποτελεί το μέτρο σύγκρισης για τους υπόλοιπους ταξινομητές.

5.5.5 Μέθοδος Επικύρωσης

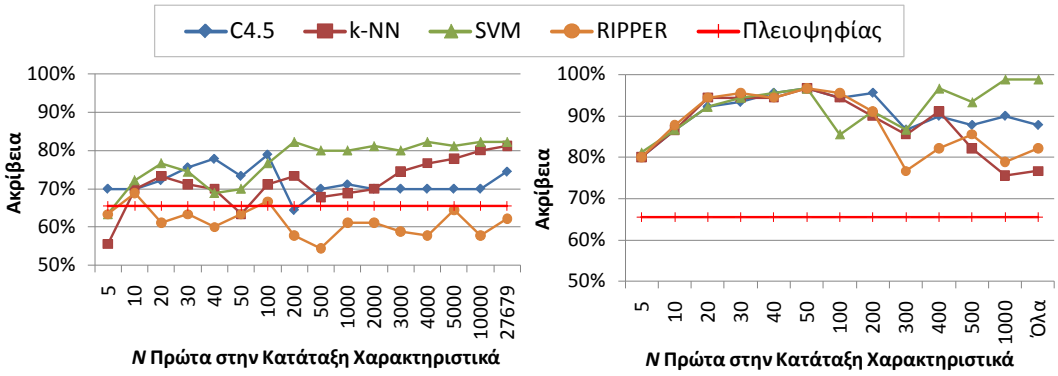
Λόγω του μικρού πλήθους παραδειγμάτων που περιέχουν τα σύνολα δεδομένων προτιμήθηκε η εφαρμογή της διαδικασίας σταυρωτής επικύρωσης με μία παράλειψη (leave-one-out cross-validation) παρά της 10-πλής σταυρωτής επικύρωσης (10-fold cross-validation). Πρέπει να τονιστεί ότι όλα τα στάδια της διαδικασίας ανακάλυψης γνώσης (διακριτοποίηση, εξόρυξη συχνών συνόλων, επιλογή χαρακτηριστικών, εκπαίδευση και αξιολόγηση ταξινομητών) ενσωματώθηκαν στη διαδικασία επικύρωσης, καθώς επίσης και ότι όλες οι επαναλήψεις πραγματοποιήθηκαν με τις ίδιες παραμέτρους και τα ίδια δεδομένα για κάθε συγκρινόμενη μέθοδο. Όλα τα παραπάνω εξασφαλίζουν μια όσο το δυνατόν πιο δίκαιη αξιολόγηση για όλες τις συγκρινόμενες μεθόδους.

5.5.6 Αποτελέσματα

Σε αυτό το εδάφιο παρουσιάζονται τα αποτελέσματα των πειραμάτων που διεξήχθησαν με σκοπό την αξιολόγηση της προτεινόμενης προσέγγισης, χρησιμοποιώντας τις μεθό-

δους επιλογής χαρακτηριστικών και ταξινόμησης που περιγράφηκαν παραπάνω. Για τους σκοπούς της αξιολόγησης η προτεινόμενη προσέγγιση συγκρίθηκε με μια τυπική προσέγγιση η οποία περιλαμβάνει την εφαρμογή των σταδίων της επιλογής χαρακτηριστικών και ταξινόμησης στα αρχικά σύνολα δεδομένων. Η συγκεκριμένη προσέγγιση που χρησιμοποιείται ως μέτρο σύγκρισης στα πλαίσια της διατριβής έχει μελετηθεί σε προηγούμενες εργασίες [Gamberoni & Storari, 2004; Lin & Li, 2005].

Τα γραφήματα που απεικονίζονται στην Εικόνα 5.2 παρουσιάζουν την ακρίβεια ταξινόμησης των πέντε ταξινομητών που περιγράφηκαν παραπάνω στο σύνολο δεδομένων 90×27679. Ως μέτρο αξιολόγησης των χαρακτηριστικών χρησιμοποιήθηκε το στατιστικό X^2. Ο κατακόρυφος άξονας αναπαριστά την ακρίβεια ταξινόμησης, ενώ ο οριζόντιος άξονας αναπαριστά τον αριθμό των πρώτων στην κατάταξη (βάσει του στατιστικού X^2) χαρακτηριστικών που επιλέχθηκαν για την εκπαίδευση και αξιολόγηση των ταξινομητών. Το πρώτο γράφημα απεικονίζει τα αποτελέσματα της τυπικής προσέγγισης, ενώ το δεύτερο τα αποτελέσματα της προτεινόμενης.



(α) τυπική προσέγγιση (β) προτεινόμενη προσέγγιση

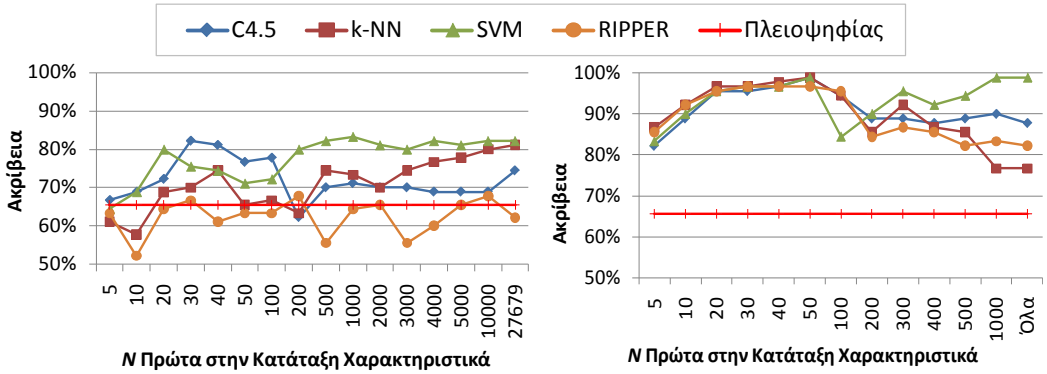
Εικόνα 5.2: Ακρίβεια των ταξινομητών στο σύνολο δεδομένων 99×27679. Ως μέτρο αξιολόγησης των χαρακτηριστικών χρησιμοποιήθηκε το στατιστικό X^2.

Όπως γίνεται φανερό από τα δύο γραφήματα, οι ταξινομητές της τυπικής προσέγγισης παρουσιάζουν αρκετά χαμηλότερη ακρίβεια ταξινόμησης από τους αντίστοιχους της προτεινόμενης προσέγγισης. Σε αρκετές περιπτώσεις η ακρίβεια των ταξινομητών της τυπικής προσέγγισης προσεγγίζει την ακρίβεια του ταξινομητή πλειοψηφίας και μάλιστα ένας από τους ταξινομητές (RIPPER) επιτυγχάνει σχεδόν πάντα χειρότερη ακρίβεια από τον ταξινομητή πλειοψηφίας. Η κορυφαία απόδοση από τους ταξινομητές της τυπικής προσέγγισης επιτυγχάνεται από τον ταξινομητή μηχανής διανυσμάτων υποστήριξης (SVM) με επιλεγμένα τα 200 πρώτα στην κατάταξη χαρακτηριστικά και είναι 82,22%.

Σε αντίθεση με την τυπική προσέγγιση, οι ταξινομητές της προτεινόμενης προσέγγισης επιτυγχάνουν ακρίβειες που είναι μακράν καλύτερες από την ακρίβεια του ταξινομητή πλειοψηφίας. Συγκεκριμένα, η κορυφαία απόδοση επιτυγχάνεται από τον ταξινομητή μηχανής διανυσμάτων υποστήριξης (SVM), όταν επιλέγονται τουλάχιστον τα 1.000

πρώτα στην κατάταξη χαρακτηριστικά και είναι 98,89%. Ο ταξινομητής *k*-Κοντινότερων γειτόνων με επιλεγμένα τα 1.000 πρώτα στην κατάταξη χαρακτηριστικά επιτυγχάνει τη χειρότερη απόδοση, η οποία ωστόσο είναι κατά 11 ποσοστιαίες μονάδες ανώτερη από την ακρίβεια του ταξινομητή πλειοψηφίας.

Τα γραφήματα που απεικονίζονται στην Εικόνα 5.3 παρουσιάζουν την ακρίβεια ταξινόμησης των πέντε ταξινομητών που αξιολογήθηκαν στο σύνολο δεδομένων 90×27679. Ως μέτρο αξιολόγησης των χαρακτηριστικών στην περίπτωση αυτή χρησιμοποιήθηκε κέρδος πληροφορίας. Τα αποτελέσματα των συγκεκριμένων πειραμάτων εμφανίζουν σημαντικές ομοιότητες με τα αποτελέσματα που παρουσιάστηκαν παραπάνω για το στατιστικό X^2 (Εικόνα 5.2). Την κορυφαία απόδοση για την τυπική μεθοδολογία πετυχαίνει ο ταξινομητής της μηχανής διανυσμάτων υποστήριξης με επιλεγμένα τα 1.000 πρώτα στην κατάταξη χαρακτηριστικά, η ακρίβεια του οποίου φτάνει το 83,33%. Την κορυφαία απόδοση της προτεινόμενης προσέγγισης επιτυγχάνουν τρεις ταξινομητές, το δένδρο ταξινόμησης (C4.5), ο ταξινομητής *k*-Κοντινότερων γειτόνων (*k*-NN)και ο ταξινομητής μηχανής διανυσμάτων υποστήριξης (SVM) με τα 50 πρώτα στην κατάταξη χαρακτηριστικά. Η ακρίβειά τους είναι 98,89%. Αξίζει να σημειωθεί ότι η χειρότερη ακρίβεια ταξινόμησης της προτεινόμενης προσέγγισης (ταξινομητής *k*-Κοντινότερων γειτόνων με περισσότερα από 1.000 χαρακτηριστικά) είναι 11,11 ποσοστιαίες μονάδες υψηλότερη από την ακρίβεια του ταξινομητή πλειοψηφίας.

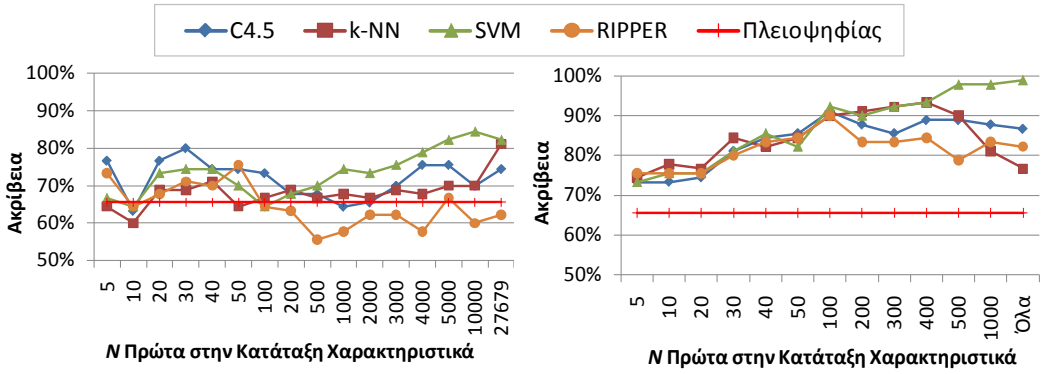


(α) τυπική προσέγγιση (β) προτεινόμενη προσέγγιση

Εικόνα 5.3: Ακρίβεια των ταξινομητών στο σύνολο δεδομένων 99×27679. Ως μέτρο αξιολόγησης των χαρακτηριστικών χρησιμοποιήθηκε το κέρδος πληροφορίας.

Τα γραφήματα που απεικονίζονται στην Εικόνα 5.4 παρουσιάζουν την ακρίβεια ταξινόμησης των πέντε ταξινομητών που αξιολογήθηκαν στο σύνολο δεδομένων 90×27679 με μέτρο αξιολόγησης των χαρακτηριστικών το μέτρο Relief F. Οι ακρίβειες των ταξινομητών, μολονότι παρουσιάζουν μικρές διαφορές με αυτές που παρουσιάστηκαν στα προηγούμενα γραφήματα, εμφανίζουν παρόμοια συμπεριφορά και τα συμπεράσματα που διατυπώθηκαν για τις προηγούμενες δύο περιπτώσεις ισχύουν και σε αυτήν την περίπτωση. Την κορυφαία απόδοση για την τυπική μεθοδολογία πετυχαίνει ο ταξινομητής της μηχανής διανυσμάτων υποστήριξης με τα 1.000 πρώτα στην

κατάταξη χαρακτηριστικά, η ακρίβεια του οποίου φτάνει το 84,44%. Την κορυφαία απόδοση της προτεινόμενης προσέγγισης πετυχαίνει ο ταξινομητής της μηχανής διανυσμάτων υποστήριξης με περισσότερα από τα 1.000 πρώτα στην κατάταξη χαρακτηριστικά, η ακρίβεια του οποίου είναι 98,89%.



(α) τυπική προσέγγιση (β) προτεινόμενη προσέγγιση

Εικόνα 5.4: Ακρίβεια των ταξινομητών στο σύνολο δεδομένων 99×27679. Ως μέτρο αξιολόγησης των χαρακτηριστικών χρησιμοποιήθηκε το μέτρο Relief F.

Παρατηρώντας τα γραφήματα που παρουσιάζουν την απόδοση της προτεινόμενης προσέγγισης στις τρεις παραπάνω εικόνες μπορεί να εξαχθεί ένα σημαντικό συμπέρασμα για την ακρίβεια των ταξινομητών. Όταν για την αξιολόγηση των χαρακτηριστικών χρησιμοποιούνται το στατιστικό X^2 και το κέρδος πληροφορίας, τότε όλοι οι ταξινομητές (πλην του ταξινομητή πλειοψηφίας που χρησιμοποιείται ως μέτρο σύγκρισης) επιτυγχάνουν πολύ υψηλή και σχεδόν σταθερή ακρίβεια (περίπου 95%) με την επιλογή μόνο των πρώτων 20 έως 50 στην κατάταξη χαρακτηριστικών. Στην περίπτωση που για την αξιολόγηση των χαρακτηριστικών χρησιμοποιείται το μέτρο Relief F τα αποτελέσματα εμφανίζονται ελαφρώς διαφορετικά, αφού η ακρίβεια των ταξινομητών αυξάνεται όταν με εκκίνηση τα 20 χαρακτηριστικά επιλέγονται όλο και περισσότερα έως τα 100 χαρακτηριστικά, όπου η ακρίβεια φτάνει το 91%. Η παραπάνω παρατήρηση είναι πολύ σημαντική, επειδή όπως γίνεται αντιληπτό είναι αρκετός ένας πολύ μικρός αριθμός από χαρακτηριστικά για να δημιουργηθούν ιδιαίτερα αποτελεσματικοί ταξινομητές. Αυτό σημαίνει ότι το υπολογιστικό κόστος μπορεί να μειωθεί σημαντικά. Όμως το μεγαλύτερο κέρδος προκύπτει από το γεγονός ότι μειώνεται σημαντικά ο αριθμός των γονιδίων που πρέπει να ληφθούν υπόψη στη μελέτη του συγκεκριμένου προβλήματος. Υπενθυμίζεται ότι τα χαρακτηριστικά του τελικού πίνακα δεδομένων δεν είναι γονίδια αλλά συχνά σύνολα γονιδιακών εκφράσεων. Ωστόσο, ακόμη και έτσι, ο αριθμός των διαφορετικών γονιδίων που σχετίζονται με τα επιλεγμένα συχνά σύνολα είναι εξαιρετικά μικρός (της τάξης των δεκάδων ή λίγων εκατοντάδων).

Από πειράματα στατιστικής σύγκρισης που διεξήχθησαν προκύπτει ότι η απόδοση όλων των ταξινομητών που αξιολογήθηκαν με το σύνολο δεδομένων 90×27679 και με όλα τα μέτρα αξιολόγησης χαρακτηριστικών είναι στατιστικά σημαντικά καλύτερη (με

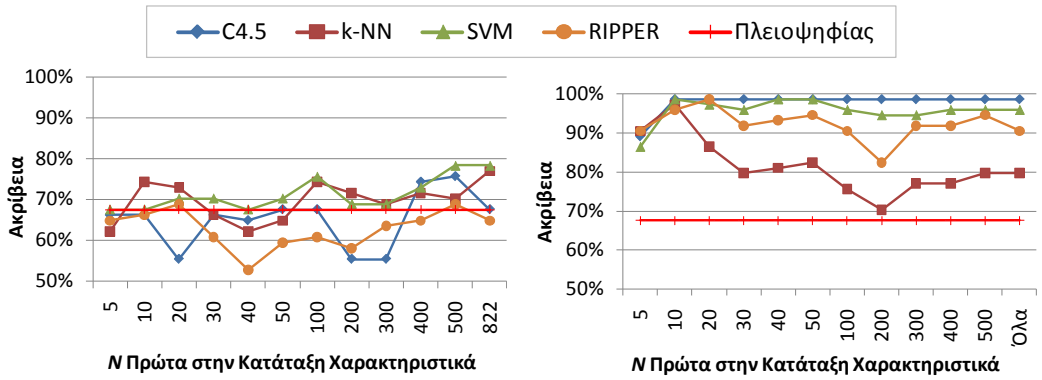
επίπεδο εμπιστοσύνης 95%) από την απόδοση του ταξινομητή πλειοψηφίας. Εξαίρεση αποτελούν κάποιοι ταξινομητές που παρουσιάζουν χαμηλή ακρίβεια, όταν χρησιμοποιείται το μέτρο Relief F και επιλέγονται μόνο τα 5 έως 20 πρώτα στην κατάταξη χαρακτηριστικά. Αντίθετα, στις περισσότερες περιπτώσεις ταξινομητών της τυπικής προσέγγισης δεν υπάρχει στατιστικά σημαντική υπεροχή της ακρίβειας από την ακρίβεια του ταξινομητή πλειοψηφίας.

Από τα γραφήματα που παρουσιάζουν την απόδοση της τυπικής προσέγγισης γίνεται φανερό ότι η ακρίβεια των ταξινομητών τείνει να αυξάνεται καθώς αυξάνεται και ο αριθμός των πρώτων στην κατάταξη χαρακτηριστικών που επιλέγονται. Η ακρίβεια μοιάζει να σταθεροποιείται όταν επιλέγονται περισσότερα από 500 χαρακτηριστικά, κυρίως όταν χρησιμοποιούνται τα μέτρα X^2 και κέρδος πληροφορίας. Βέβαια, κάποιοι ταξινομητές αυξάνουν σημαντικότερα την ακρίβειά τους όταν επιλέγονται ακόμη περισσότερα χαρακτηριστικά. Όταν χρησιμοποιείται το μέτρο Relief F για την αξιολόγηση των χαρακτηριστικών, απαιτείται η επιλογή μεγαλύτερου πλήθους χαρακτηριστικών για την επίτευξη μεγαλύτερης ακρίβειας. Το ίδιο παρατηρείται και στο αντίστοιχο γράφημα της προτεινόμενης προσέγγισης. Αυτό σημαίνει ότι ο Relief F δεν ενδείκνυται στην περίπτωση που πρέπει να επιλεγεί ένα πολύ μικρό σύνολο χαρακτηριστικών στα πλαίσια του προβλήματος που μελετάται στο κεφάλαιο αυτό.

Ο μεγαλύτερος αριθμός χαρακτηριστικών που απαιτούνται για την επίτευξη καλύτερης ακρίβειας ταξινόμησης στην περίπτωση της τυπικής προσέγγισης αυξάνει το υπολογιστικό κόστος. Αυτό το φαινόμενο έχει περιγραφεί και σε προηγούμενες εργασίες όπως για παράδειγμα στην [Lin & Li, 2005]. Η εξήγηση στηρίζεται στο γεγονός ότι υπάρχει ένας σημαντικός αριθμός γονιδίων (περίπου 500-1.000) που είναι ιδιαίτερα σημαντικά για την αποτελεσματική ταξινόμηση των δεδομένων στις δύο τάξεις. Ο αριθμός των απαιτούμενων γονιδίων αυξάνεται όταν αυξάνεται και ο αριθμός των διαφορετικών ιστών και οργάνων από τους οποίους προέρχονται τα δεδομένα. Αυτό συμβαίνει γιατί, όπως έχει ήδη αναφερθεί, σε διαφορετικούς ιστούς εκφράζονται και διαφορετικά γονίδια, πλην κάποιων διαχειριστικών γονιδίων που εκφράζονται σε όλους τους ιστούς και τους κυτταρικούς τύπους. Τα υπόλοιπα γονίδια είναι συνήθως άσχετα με το πρόβλημα που μελετάται, αλλά συνήθως δεν επηρεάζουν αρνητικά την αποτελεσματικότητα των ταξινομητών παρά μόνον αυξάνουν το υπολογιστικό κόστος. Στην προτεινόμενη προσέγγιση ο μετασχηματισμός του χώρου των χαρακτηριστικών από γονίδια σε συχνά σύνολα γονιδιακών εκφράσεων μειώνει σημαντικά των απαιτούμενο αριθμό χαρακτηριστικών, αλλά ταυτόχρονα αυξάνει σημαντικά και την αποτελεσματικότητα των ταξινομητών, οι οποίοι επωφελούνται από την πληροφορία που παρέχουν οι συσχετίσεις μεταξύ γονιδιακών εκφράσεων.

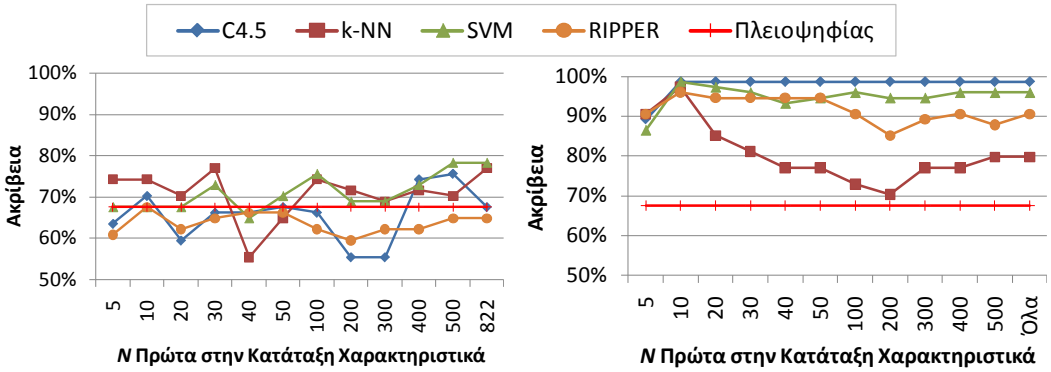
Τα αποτελέσματα των πειραμάτων που διεξήχθησαν με το σύνολο δεδομένων 74×822 σε γενικές γραμμές παρουσιάζουν τα ίδια χαρακτηριστικά με τα αποτελέσματα του συνόλου δεδομένων 99×27679. Η Εικόνα 5.5 παρουσιάζει την ακρίβεια των ταξινομητών που προέκυψαν με χρήση του στατιστικού X^2 ως μέτρου αξιολόγησης των χαρα-

κτηριστικών. Τα αντίστοιχα αποτελέσματα για το κέρδος πληροφορίας παρουσιάζονται στην Εικόνα 5.6 και τα αποτελέσματα για το μέτρο Relief F στην Εικόνα 5.7.



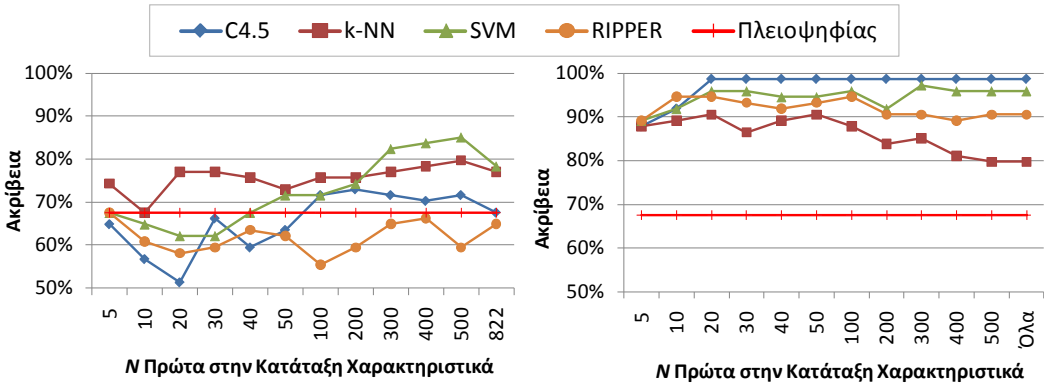
(α) τυπική προσέγγιση (β) προτεινόμενη προσέγγιση

Εικόνα 5.5: Ακρίβεια των ταξινομητών στο σύνολο δεδομένων 74×822. Ως μέτρο αξιολόγησης των χαρακτηριστικών χρησιμοποιήθηκε το στατιστικό X^2.



(α) τυπική προσέγγιση (β) προτεινόμενη προσέγγιση

Εικόνα 5.6: Ακρίβεια των ταξινομητών στο σύνολο δεδομένων 74×822. Ως μέτρο αξιολόγησης των χαρακτηριστικών χρησιμοποιήθηκε το κέρδος πληροφορίας.



(α) τυπική προσέγγιση (β) προτεινόμενη προσέγγιση

Εικόνα 5.7: Ακρίβεια των ταξινομητών στο σύνολο δεδομένων 74×822. Ως μέτρο αξιολόγησης των χαρακτηριστικών χρησιμοποιήθηκε το μέτρο Relief F.

Την υψηλότερη ακρίβεια (85,14%) για την τυπική προσέγγιση πέτυχε ο ταξινομητής μηχανής διανυσμάτων υποστήριξης με τη χρήση του μέτρου Relief F και την επιλογή των 500 πρώτων στην κατάταξη χαρακτηριστικών. Η υψηλότερη ακρίβεια που έχει επιτευχθεί για την προτεινόμενη προσέγγιση είναι 98,65% και την έχουν πετύχει αρκετοί ταξινομητές. Αξιοσημείωτη είναι η πτώση της απόδοσης του ταξινομητή των k-Κοντινότερων γειτόνων στην προτεινόμενη προσέγγιση όταν χρησιμοποιείται μεγαλύτερος αριθμός χαρακτηριστικών. Μια πιθανή εξήγηση για αυτήν τη συμπεριφορά είναι ότι τα δεδομένα εισόδου στους αλγορίθμους ταξινόμησης της προτεινόμενης προσέγγισης έχουν δυαδικές τιμές, καθώς αναπαριστούν την παρουσία ή την απουσία συχνών συνόλων στα παραδείγματα. Ωστόσο, οι μέθοδοι υπολογισμού των αποστάσεων που χρησιμοποιούνται από τον αλγόριθμο *k*-Κοντινότερων γειτόνων συχνά δεν είναι κατάλληλες για την ταξινόμηση παραδειγμάτων με δυαδικές τιμές και το πρόβλημα γίνεται ακόμη μεγαλύτερο όταν αυξάνονται οι διαστάσεις.

Από πειράματα στατιστικής σύγκρισης που διεξήχθησαν προκύπτει ότι η απόδοση όλων των ταξινομητών που αξιολογήθηκαν με το σύνολο δεδομένων 74×822 και με όλα τα μέτρα αξιολόγησης χαρακτηριστικών είναι στατιστικά σημαντικά καλύτερη (με επίπεδο εμπιστοσύνης 95%) από την απόδοση του ταξινομητή πλειοψηφίας. Εξαίρεση αποτελούν κάποιοι ταξινομητές *k*-Κοντινότερων γειτόνων με πολλά επιλεγμένα χαρακτηριστικά, οι οποίοι παρουσιάζουν αρκετά χαμηλότερη ακρίβεια σε σχέση με τους υπόλοιπους ταξινομητές. Αντίθετα, στις περισσότερες περιπτώσεις ταξινομητών της τυπικής προσέγγισης δεν υπάρχει στατιστικά σημαντική υπεροχή της ακρίβειας από την ακρίβεια του ταξινομητή πλειοψηφίας.

Ο Πίνακας 5.3 παρουσιάζει τις υψηλότερες ακρίβειες ταξινόμησης που έχουν αναφερθεί σε προγενέστερες μελέτες καθώς και τις ακρίβειες της προτεινόμενης προσέγγισης.

Πίνακας 5.3: Αποτελέσματα μελετών για τα σύνολα δεδομένων του κεφαλαίου.

Προσέγγιση	Σύνολο Δεδομένων	Ακρίβεια
Gamberoni and Storari [2004]	90×27679	82,20%
Lin & Li [2005]	90×27679	85,60%
Alves et al. [2005]	74×822	86,18%
Tzanis & Vlahavas [2007b]	90×27679	98,89%
	74×822	98,65%

5.6 Συμπεράσματα

Στο κεφάλαιο αυτό παρουσιάστηκε μια προσέγγιση για την αποτελεσματική και αποδοτική ταξινόμηση δεδομένων γονιδιακής έκφρασης, τα οποία έχουν συλλεχθεί με την τεχνική SAGE. Η προτεινόμενη προσέγγιση αξιοποιεί την πληροφορία που παρέχουν οι συσχετίσεις μεταξύ γονιδιακών εκφράσεων και επιτυγχάνει σημαντική βελτίωση της αποτελεσματικότητας των ταξινομητών. Τα πειράματα που διεξήχθησαν έδειξαν ότι ένα μικρό πλήθος χαρακτηριστικών, τα οποία είναι συχνά σύνολα γονιδιακών εκφρά-

σεων, είναι ικανό να προσφέρει πολύ αποτελεσματικούς ταξινομητές με ακρίβεια που κυμαίνεται στο 95%. Ο αριθμός των απαιτούμενων χαρακτηριστικών μπορεί να είναι μικρότερος από 100. Το όφελος σε αυτήν την περίπτωση είναι διπλό. Πρώτον, βελτιώνεται σημαντικά η αποτελεσματικότητα των ταξινομητών. Δεύτερον, μπορούν να μειωθούν σημαντικά οι διαστάσεις των δεδομένων και κατ' επέκταση το υπολογιστικό κόστος, χωρίς να θυσιαστεί σημαντικά η αποτελεσματικότητα. Τα παραπάνω επιτυγχάνονται με τη χρήση ενός αλγόριθμου εξόρυξης συχνών συνόλων αντικειμένων η αποδοτικότητα του οποίου είναι εξαιρετικά μεγάλη σε σύνολα δεδομένων γονιδιακής έκφρασης, εφόσον τέτοιου είδους αλγόριθμοι ειδικεύονται στην εξόρυξη συχνών συνόλων από δεδομένα με εκατομμύρια παραδειγμάτων και χιλιάδες χαρακτηριστικών.

Πρόβλεψη του Σημείου Έναρξης της Μετάφρασης

6

6.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζονται μέθοδοι που υλοποιήθηκαν στα πλαίσια της διατριβής για την πρόβλεψη του σημείου έναρξης της μετάφρασης (Translation Initiation Site - TIS). Η πρόβλεψη του σημείου έναρξης της μετάφρασης γίνεται σε αλληλουχίες mRNA ή σε αλληλουχίες συμπληρωματικού DNA (complementary DNA - cDNA). Οι αλληλουχίες cDNA συντίθενται χρησιμοποιώντας ένα ώριμο mRNA ως πρότυπο μέσω της διαδικασίας της αντίστροφης μεταγραφής, που καταλύεται από το ένζυμο αντίστροφη μεταγραφάση (reverse transcriptase). Για κάποιον οργανισμό μπορεί να κατασκευαστεί μια βιβλιοθήκη cDNA, η οποία είναι μια συλλογή από κλώνους που περιέχουν cDNA και αναπαριστούν τα γονίδια που εκφράζονται σε ένα κύτταρο ή ιστό σε ένα δεδομένο χρονικό διάστημα. Μια από τις πιο συνηθισμένες χρήσεις των βιβλιοθηκών cDNA είναι η κλωνοποίηση, δηλαδή αντιγραφή και εισαγωγή, ευκαρυωτικών γονιδίων σε γονιδιώματα προκαρυωτικών οργανισμών. Οι βασικότεροι λόγοι για τους οποίους είναι σημαντική η χρήση των βιβλιοθηκών cDNA είναι οι εξής:

- Συνήθως είναι ευκολότερη η απομόνωση ενός μορίου mRNA απ' ότι του ίδιου του γονιδίου.
- Μόνο ένα μικρό ποσοστό του συνολικού DNA πολλών οργανισμών εκφράζεται. Αυτό σημαίνει ότι μια βιβλιοθήκη cDNA που θα περιλαμβάνει μόνο τα ενεργά γονίδια θα έχει πολύ μικρότερο μέγεθος και θα είναι ευκολότερη η κατασκευή και η διαχείρισή της.
- Οι αλληλουχίες cDNA, εφόσον προέρχονται από τις αλληλουχίες mRNA, δεν περιλαμβάνουν παρεμβαλλόμενα μη-κωδικά τμήματα (ιντρόνια).

- Η κλωνοποίηση του γενετικού υλικού των RNA ιών μπορεί να γίνει μόνο με cDNA αλληλουχίες.

Οι προτεινόμενες μέθοδοι υλοποιήθηκαν στη γλώσσα προγραμματισμού JAVA και χρησιμοποιήθηκε η βιβλιοθήκη αλγορίθμων Weka. Η βιβλιοθήκη αλγορίθμων Weka[12] [Witten & Frank, 2005] είναι μια συλλογή από αλγορίθμους μηχανικής μάθησης. Αναπτύχθηκε το 1999 με σκοπό να καλύψει τις πειραματικές ανάγκες του εργαστηρίου Μηχανικής Μάθησης του Πανεπιστημίου του Waikato στη Νέα Ζηλανδία. Περιλαμβάνει αλγορίθμους για προεπεξεργασία δεδομένων, ταξινόμηση, παλινδρόμηση, ομαδοποίηση, κανόνες συσχέτισης και οπτική αναπαράσταση. Οι αλγόριθμοι μπορούν να εφαρμοστούν είτε σε ένα σύνολο δεδομένων μέσω ενός εύχρηστου γραφικού περιβάλλοντος διασύνδεσης (GUI), είτε να εκτελεστούν μέσω ενός προγράμματος γραμμένου στη γλώσσα προγραμματισμού JAVA. Επίσης, παρέχει τη δυνατότητα επέκτασης και ανάπτυξης νέων σχημάτων μηχανικής μάθησης. Η βιβλιοθήκη Weka είναι λογισμικό ανοιχτού κώδικα και διανέμεται κάτω από τους όρους της Γενικής Άδειας Δημόσιας Χρήσης GNU. Η υλοποίηση ενός μεγάλου αριθμού αλγορίθμων, η ευχρηστία του γραφικού περιβάλλοντος και η ελεύθερη διανομή της βιβλιοθήκης, την έχουν καταστήσει ως ένα πολύ δημοφιλές εργαλείο μηχανικής μάθησης και χρησιμοποιείται ευρέως στην ερεύνα, στην εκπαίδευση αλλά και σε εφαρμογές.

Επιπλέον, στην πιο πρόσφατη από τις προτεινόμενες προσεγγίσεις (μέθοδος StackTIS) έγινε χρήση και της βιβλιοθήκης WLSVM [El-Manzalawy & Honavar, 2005], η οποία παρέχει διάφορες μεθόδους κατασκευής μηχανών διανυσμάτων υποστήριξης (SVM), που είναι αποδοτικότερες από τον αντίστοιχο αλγόριθμο (Sequential Minimal Optimization – SMO [Platt, 1998; Keerthi, et al., 2001]) της βιβλιοθήκης Weka. Η βιβλιοθήκη WLSVM ουσιαστικά αποτελεί μια υλοποίηση της βιβλιοθήκης LibSVM [Chang & Lin, 2005], η οποία είναι απόλυτα συμβατή με το περιβάλλον της βιβλιοθήκης Weka.

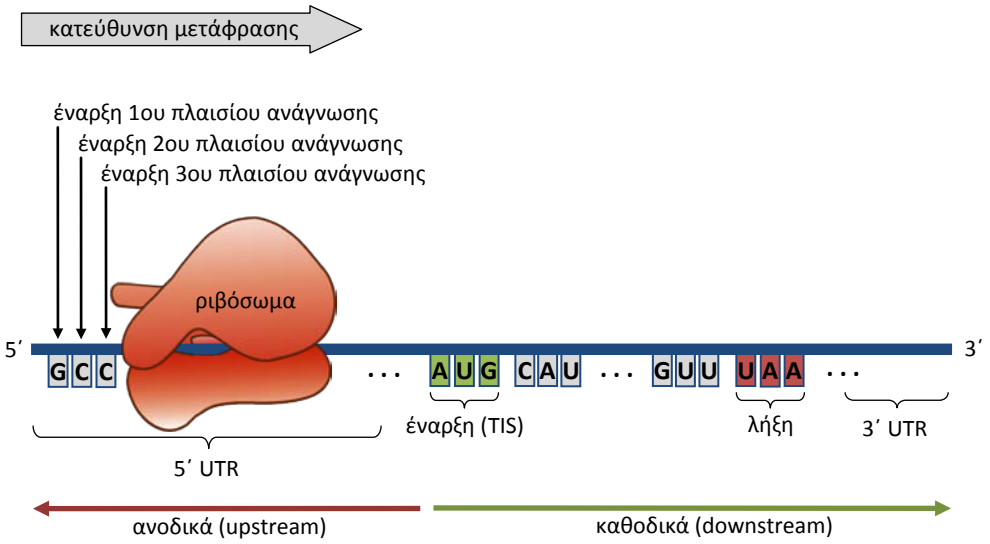
6.2 Βασικές Έννοιες

Ένα από τα βασικά στάδια της πρόβλεψης γονιδίων (Εδάφιο 3.9.2) περιλαμβάνει την εύρεση του σημείου έναρξης της μετάφρασης, δηλαδή του κωδικονίου στο οποίο αρχίζει η διαδικασία της μετάφρασης μιας αλληλουχίας mRNA. Στην Εικόνα 6.1 παρουσιάζεται η διαδικασία έναρξης της μετάφρασης. Το ριβόσωμα προσδένεται στο 5΄ άκρο της αλληλουχίας του mRNA και σαρώνει την αλληλουχία κατευθυνόμενο προς το 3΄ άκρο. Όταν εντοπιστεί το κατάλληλο κωδικόνιο AUG, που αποτελεί το σημείο έναρξης της μετάφρασης, τότε αρχίζει η διαδικασία μετάφρασης. Η μετάφραση ολοκληρώνεται όταν εντοπιστεί το πρώτο κωδικόνιο λήξης (UAA, UAG ή UGA) που είναι τοποθετημένο στο ίδιο πλαίσιο ανάγνωσης με το κωδικόνιο έναρξης. Οι περιοχές της αλληλουχίας mRNA πριν το σημείο έναρξης της μετάφρασης και έπειτα από το σημείο λήξης αυτής

[12] http://www.cs.waikato.ac.nz/ml/weka (προσπελάστηκε: 17/05/2011).

ονομάζονται 5΄ και 3΄ αμετάφραστες περιοχές (Untranslated Regions - UTRs) αντίστοιχα.

Λόγω του γεγονότος ότι η ανάγνωση και μετάφραση του mRNA γίνεται ανά κωδικόνιο-τριπλέτα, υπάρχουν τρεις διαφορετικοί τρόποι ανάγνωσης της αλληλουχίας mRNA προς κάθε κατεύθυνση. Οι τρεις αυτοί τρόποι ανάγνωσης καλούνται *πλαίσια ανάγνωσης* (*reading frames*). Μάλιστα, ένα τμήμα σε κάποιο πλαίσιο ανάγνωσης της αλληλουχίας που δεν περιλαμβάνει κωδικόνια λήξης καλείται *ανοικτό πλαίσιο ανάγνωσης* (*Open Reading Frame – ORF*). Η έναρξη της μετάφρασης συνήθως γίνεται στο κωδικόνιο AUG που βρίσκεται πιο κοντά στο 5΄ άκρο της αλληλουχίας του mRNA. Ωστόσο, δε συμβαίνει πάντοτε αυτό, καθώς υπάρχουν κάποιοι μηχανισμοί που επιτρέπουν την έναρξη της μετάφρασης σε επόμενα κωδικόνια AUG, τα οποία συνήθως βρίσκονται και αυτά κοντά στο 5΄ άκρο. Η ύπαρξη αυτών των μηχανισμών δυσκολεύει την πρόβλεψη του σημείου έναρξης της μετάφρασης. Για την αναφορά σε ένα τμήμα της αλληλουχίας mRNA από κάποιο σημείο αναφοράς (π.χ. στο πρόβλημα που μελετάται στο παρόν κεφάλαιο το σημείο αναφοράς είναι το σημείο έναρξης της μετάφρασης) προς το 5΄ άκρο χρησιμοποιείται ο όρος *ανοδικά* (*upstream*), ενώ για την αναφορά σε ένα τμήμα της αλληλουχίας mRNA από το σημείο αναφοράς προς το 3΄ άκρο χρησιμοποιείται ο όρος *καθοδικά* (*downstream*). Όλα τα παραπάνω απεικονίζονται στην Εικόνα 6.1.



Εικόνα 6.1: Έναρξη της μετάφρασης σε αλληλουχία mRNA.

6.3 Σχετικές Εργασίες

Η ακριβής πρόβλεψη του σημείου έναρξης της μετάφρασης σε αλληλουχίες mRNA ή cDNA έχει μελετηθεί αρκετά από τη δεκαετία του 1980. Η ιδιαίτερη σημασία του έγκειται στο γεγονός ότι οι in vitro μέθοδοι εντοπισμού είχαν μεγάλο κόστος και απαιτούσαν πολύ χρόνο για να ολοκληρωθούν. Έτσι, ήταν έντονη η ανάγκη να κατευθυνθούν ή και να αντικατασταθούν από υπολογιστικές μεθόδους που θα μπορούσαν να προσφέρουν την επιθυμητή πληροφορία με ελάχιστο κόστος και με υψηλή ακρίβεια. Το

γεγονός όμως ότι δεν είναι γνωστές όλες οι λεπτομέρειες του μηχανισμού της έναρξης της μετάφρασης, έχει οδηγήσει την επιστημονική κοινότητα στη δημιουργία ενός αριθμού υπολογιστικών εργαλείων, με αρκετά καλή ακρίβεια. Ωστόσο, υπάρχουν ακόμη περιθώρια βελτίωσης και το πρόβλημα οπωσδήποτε δε θεωρείται τετριμμένο. Ένα ακόμη στοιχείο που καθιστά το πρόβλημα αυτό ακόμη πιο ενδιαφέρον είναι η διαρκής αλληλούχιση ενός μεγάλου αριθμού οργανισμών των οποίων το γονιδίωμα ακόμη δεν έχει υπομνηματιστεί και μελετηθεί.

Από το 1982, το ζήτημα της πρόβλεψης του σημείου έναρξης της μετάφρασης προσεγγίστηκε με μεθόδους βιολογικές, στατιστικές αλλά και με τεχνικές μηχανικής μάθησης και ανακάλυψης γνώσης σε βάσεις δεδομένων. Οι Stormo et al. [1982] χρησιμοποίησαν τον αλγόριθμο perceptron για το διαχωρισμό των σημείων έναρξης της μετάφρασης. Η Kozak [1987] ανέπτυξε τον πρώτο πίνακα βαρών για τον εντοπισμό τους σε μια αλληλουχία cDNA. Το πρότυπο συναίνεσης (consensus pattern) που προέκυψε από τον πίνακα αυτόν είναι το εξής: GCC[**AG**]CCatg**G** (με έντονα γράμματα παρουσιάζονται οι πιο συντηρημένες θέσεις). Στο μεταξύ, οι Kozak & Shatkin [1978] είχαν προτείνει ένα μοντέλο σάρωσης για την έναρξη της μετάφρασης, το οποίο αργότερα επεκτάθηκε από την Kozak [1989]. Σύμφωνα με αυτό, η μετάφραση ξεκινά στο πρώτο κωδικόνιο έναρξης που πληροί κατάλληλες προϋποθέσεις. Το μοντέλο αυτό ονομάζεται μοντέλο σάρωσης ριβοσώματος (ribosome scanning model) [Cigan et al., 1988; Agarwal & Bafna, 1998].

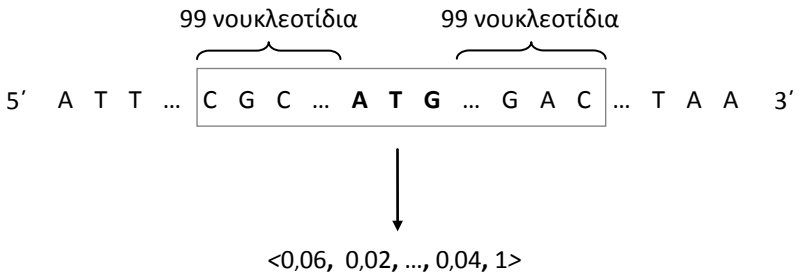
Οι δημοφιλέστερες τεχνικές ανακάλυψης γνώσης που έχουν εφαρμοστεί για την πρόβλεψη του σημείου έναρξης της μετάφρασης περιλαμβάνουν τεχνητά νευρωνικά δίκτυα [Hatzigeorgiou, 2002; Pedersen & Nielsen, 1997], μηχανές διανυσμάτων υποστήριξης [Zien et al., 2000], γκαουσιανά μικτά μοντέλα [Li et al., 2005], μεθόδους γραμμικού διαχωρισμού [Salamov et al., 1998] και τεχνικές που βασίζονται σε στατιστική και ομοιότητα [Nishikawa et al., 2001]. Οι Liu & Wong [2003] εφάρμοσαν τεχνικές δημιουργίας χαρακτηριστικών (πρότυπα νουκλεοτιδίων) και επιλογής χαρακτηριστικών και χρησιμοποίησαν διάφορους αλγόριθμους ταξινόμησης. Αργότερα οι Liu et al. [2004] ακολούθησαν την ίδια προσέγγιση τριών βημάτων, χρησιμοποιώντας όμως πρότυπα αμινοξέων κατά τη δημιουργία χαρακτηριστικών. Οι Nadershahi et al. [2004] παρουσίασαν μια ενδιαφέρουσα συγκριτική μελέτη πέντε μεθόδων πρόβλεψης του σημείου έναρξης της μετάφρασης σε ετικέτες εκφρασμένων αλληλουχιών (Expressed Sequence Tags - ESTs), οι οποίες αποτελούν μικρά τμήματα DNA (συνήθως 200-500 νουκλεοτιδίων) που δημιουργούνται από την αλληλούχιση του ενός ή και των δύο άκρων ενός εκφρασμένου γονιδίου.

6.4 Η Μέθοδος Πλειοψηφίας

Στην ενότητα αυτή προτείνεται μια νέα προσέγγιση για την πρόβλεψη του σημείου έναρξης της μετάφρασης [Tzanis et al., 2006b]. Η προσέγγιση αυτή αξιοποιεί την πληροφορία που προσφέρουν κάποια χαρακτηριστικά, μερικά από τα οποία προτείνονται στα πλαίσια της διατριβής και παράλληλα εκμεταλλεύεται τα πλεονεκτήματα των

ομάδων ταξινομητών. Συγκεκριμένα, η προτεινόμενη μέθοδος συνδυάζει έναν αριθμό από ταξινομητές (Εδάφιο 6.4.2), οι προβλέψεις των οποίων συνδυάζονται με τις μεθόδους πλειοψηφίας (majority voting) ή σταθμισμένης πλειοψηφίας (weighted majority voting).

Η αποστολή των ταξινομητών είναι να προβλέψουν αν ένα κωδικόνιο ATG[13] είναι σημείο έναρξης της μετάφρασης ή όχι. Για την εκπαίδευση των ταξινομητών χρησιμοποιείται ένας αριθμός από θετικά και αρνητικά παραδείγματα τα οποία εξάγονται από τις αλληλουχίες cDNA ενός συνόλου εκπαίδευσης. Κάθε αλληλουχία cDNA περιλαμβάνει ένα θετικό παράδειγμα (κωδικόνιο ATG που είναι το σημείο έναρξης της μετάφρασης) και αρκετα αρνητικά παραδείγματα (τα υπόλοιπα κωδικόνια ATG). Κάθε παράδειγμα αναπαρίσταται από ένα διάνυσμα, οι συνιστώσες του οποίου αποτελούν τις τιμές κάποιων χαρακτηριστικών (Εδάφιο 6.4.1) που υπολογίζονται σε μια περιοχή 99 νουκλεοτιδίων ανοδικά του κωδικονίου ATG και 99 νουκλεοτιδίων καθοδικά αυτού. Η τελευταία συνιστώσα του διανύσματος υποδηλώνει την τάξη του συγκεκριμένου παραδείγματος (0: αρνητικό ή 1: θετικό). Η διαδικασία παρουσιάζεται στην Εικόνα 6.2. Με την ίδια διαδικασία εξάγονται και τα διανύσματα των παραδειγμάτων αξιολόγησης, καθώς και των νέων στιγμιότυπων.



Εικόνα 6.2: Εξαγωγή διανύσματος χαρακτηριστικών από αλληλουχία cDNA.

6.4.1 Χαρακτηριστικά

Η επιλογή των κατάλληλων χαρακτηριστικών για την περιγραφή των δεδομένων σε ένα πεδίο εφαρμογής είναι μια από τις σημαντικότερες διαδικασίες και ενδεχομένως η πλέον καθοριστική. Αν τα χαρακτηριστικά που επιλεγούν δεν είναι κατάλληλα, τότε είναι πολύ πιθανό η διαδικασία ανακάλυψης γνώσης να αποτύχει. Στην προτεινόμενη προσέγγιση έχει δοθεί ιδιαίτερη βαρύτητα στην επιλογή των κατάλληλων χαρακτηριστικών και μάλιστα προτείνεται ένα σύνολο από νέα χαρακτηριστικά που δεν έχουν αξιοποιηθεί σε προγενέστερες μελέτες του προβλήματος της πρόβλεψης του σημείου έναρξης της μετάφρασης. Ο Πίνακας 6.1 παρουσιάζει τα χαρακτηριστικά που χρησιμοποιούνται στην προτεινόμενη προσέγγιση.

Κάποια από αυτά τα χαρακτηριστικά (1, 2 και 12-15) έχουν χρησιμοποιηθεί σε προηγούμενες μελέτες [Zeng et al., 2002; Liu & Wong, 2003; Liu et al., 2004]. Ωστόσο,

[13] Χρησιμοποιείται ο συμβολισμός ATG αντί AUG, καθώς γίνεται αναφορά σε αλληλουχίες cDNA.

κάποια σύνολα χαρακτηριστικών (3 [Tzanis et al., 2005a; Tzanis & Vlahavas, 2006; Tzanis et al., 2006b] και 4-11 [Tzanis et al., 2006b]) προτείνονται και μελετώνται στα πλαίσια της παρούσας διατριβής. Πρέπει να σημειωθεί ότι στο σύνολο χαρακτηριστικών 1 τα *x* και *y* μπορούν να είναι είτε αμινοξέα, είτε κωδικόνια λήξης της μετάφρασης. Το σύνολο χαρακτηριστικών 2 περιλαμβάνει χαρακτηριστικά που καταμετρούν τη διαφορά του αριθμού εμφανίσεων άλλων χαρακτηριστικών συχνότητας εμφάνισης (1 και 3-10) μεταξύ των περιοχών ανοδικά και καθοδικά του ATG. Επιπλέον, τα χαρακτηριστικά 12 και 13 βασίζονται στο πρότυπο συναίνεσης της Marilyn Kozak (GCC[AG]CCatgG) [1987].

Πίνακας 6.1: Σύνολο χαρακτηριστικών.

	Χαρακτηριστικά	**Περιγραφή**
		Χαρακτηριστικά Συχνότητας Εμφάνισης
1	↑*x*, ↑*xy* ↓*x*, ↓*xy*	Αριθμός εμφανίσεων των αμινοξέων *x*, *y* ανοδικά και καθοδικά, αντίστοιχα.
2	↕*x*	Διαφορά αριθμού εμφανίσεων του χαρακτηριστικού συχνότητας εμφάνισης *x* ανοδικά - καθοδικά.
3	↑ *x*: θέση-κωδ. *k* ↓ *x*: θέση-κωδ. *k*	Αριθμός εμφανίσεων του νουκλεοτιδίου *x* στη θέση *k* ($k \in \{1, 2, 3\}$) των κωδικονίων ανοδικά και καθοδικά, αντίστοιχα (στο ίδιο πλαίσιο ανάγνωσης με το ATG).
4	↑υδρόφοβα ↓υδρόφοβα	Αριθμός εμφάνισης υδρόφοβων αμινοξέων ανοδικά και καθοδικά, αντίστοιχα.
5	↑υδρόφιλα ↓υδρόφιλα	Αριθμός εμφάνισης υδρόφιλων αμινοξέων ανοδικά και καθοδικά, αντίστοιχα.
6	↑όξινα ↓όξινα	Αριθμός εμφάνισης όξινων αμινοξέων ανοδικά και καθοδικά, αντίστοιχα.
7	↑βασικά ↓βασικά	Αριθμός εμφάνισης βασικών αμινοξέων ανοδικά και καθοδικά, αντίστοιχα.
8	↑αρωματικά ↓αρωματικά	Αριθμός εμφάνισης αρωματικών αμινοξέων ανοδικά και καθοδικά, αντίστοιχα.
9	↑αλειφατικά ↓αλειφατικά	Αριθμός εμφάνισης αλειφατικών αμινοξέων ανοδικά και καθοδικά, αντίστοιχα.
10	↑μη αρωμ./αλειφ. ↓μη αρωμ./αλειφ.	Αριθμός εμφάνισης μη-αρωματικών, μη-αλειφατικών αμινοξέων ανοδικά και καθοδικά, αντίστοιχα.
		Χαρακτηριστικά Θέσης
11	θέση -3*k* θέση 3(*k*+1)	Αμινοξύ που κωδικοποιείται από το κωδικόνιο στη θέση (-3*k* ή 3(*k*+1) ανοδικά και καθοδικά, αντίστοιχα (στο ίδιο πλαίσιο ανάγνωσης με το ATG με $k \geq 1$).
		Boolean Χαρακτηριστικά
12	↑[AG]: θέση -3	Παρουσία ή μη των νουκλεοτιδίων A ή G στη θέση -3.
13	↓G: θέση +4	Παρουσία ή μη του νουκλεοτιδίου G στη θέση +4.
14	↑ATG	Παρουσία ή μη άλλου κωδικονίου ATG ανοδικά.
15	↓λήξη	Παρουσία ή μη κωδικονίου λήξης καθοδικά.

Το σύνολο χαρακτηριστικών 3 περιγράφει την περιοδική εμφάνιση συγκεκριμένων νουκλεοτιδίων σε κάποια από τις τρεις θέσεις των κωδικονίων που βρίσκονται στο ίδιο πλαίσιο ανάγνωσης με το κωδικόνιο ATG που αποτελεί το σημείο αναφοράς (Εικόνα 6.3). Το σύνολο χαρακτηριστικών 11 καταγράφει τα αμινοξέα που κωδικοποιούνται από τα κωδικόνια που βρίσκονται στο ίδιο πλαίσιο ανάγνωσης με το κωδικόνιο ATG από τις θέσεις -99 έως +102. Η αρίθμηση των νουκλεοτιδικών θέσεων μιας αλληλουχίας ως προς το κωδικόνιο αναφοράς παρουσιάζεται στην Εικόνα 6.4. Σε προηγούμενες μελέτες [Zeng et al., 2002; Liu & Wong, 2003] έχουν χρησιμοποιηθεί αντίστοιχα χαρακτηριστικά θέσης για νουκλεοτίδια και όχι για αμινοξέα.

Τέλος, τα σύνολα χαρακτηριστικών 4-10 βασίζονται στις χημικές ιδιότητες των αμινοξέων που συνθέτουν μια πολυπεπτιδική αλυσίδα. Τέτοιου είδους χαρακτηριστικά δεν έχουν χρησιμοποιηθεί σε προηγούμενες μελέτες του προβλήματος. Οι χημικές ιδιότητες των αμινοξέων μιας πολυπεπτιδικής αλυσίδας καθορίζουν τις αλληλεπιδράσεις που συμβαίνουν μεταξύ τους (π.χ. ένα θετικά φορτισμένο αμινοξύ έλκεται από ένα αρνητικά φορτισμένο). Συνεπώς, οι αλληλεπιδράσεις μεταξύ των αμινοξέων καθορίζουν τη δομή του πρωτεϊνικού μορίου, η οποία προσδιορίζει τη λειτουργία της πρωτεΐνης. Επομένως, η λειτουργικότητα μιας πρωτεΐνης κρύβεται στην αλληλουχία των αμινοξέων της και τις χημικές ιδιότητες αυτών. Ωστόσο, η αλληλουχία των αμινοξέων μιας πρωτεΐνης δεν είναι παρά η αντίστοιχη μεταφρασμένη αλληλουχία mRNA (ή cDNA). Βάσει των παραπάνω δεδομένων κρίθηκε απαραίτητη η μελέτη των χαρακτηριστικών που βασίζονται στις χημικές ιδιότητες των πρωτεϊνών.

Θέση:	**1**	**2**	**3**	**1**	**2**	**3**				**1**	**2**	**3**	**1**	**2**	**3**	
5′	T	G	A	A	T	A	**A**	**T**	**G**	G	G	C	T	A	C	3′

Εικόνα 6.3: Αρίθμηση των θέσεων των νουκλεοτιδίων μέσα στα κωδικόνια που βρίσκονται στο ίδιο πλαίσιο ανάγνωσης με το κωδικόνιο ATG που αποτελεί το σημείο αναφοράς.

Θέση:	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**	**+1**	**+2**	**+3**	**+4**	**+5**	**+6**	
5′	T	G	A	A	T	A	**A**	**T**	**G**	G	G	C	3′

Εικόνα 6.4: Αρίθμηση των νουκλεοτιδίων μιας αλληλουχίας ως προς το κωδικόνιο αναφοράς.

Επιλογή Χαρακτηριστικών

Για την επιλογή των καταλληλότερων χαρακτηριστικών μελετήθηκαν δύο δημοφιλή μέτρα αξιολόγησης της πληροφορίας. Τα μέτρα αυτά παρέχουν μια κατάταξη των χαρακτηριστικών βάσει της πληροφορίας που τα χαρακτηριστικά παρέχουν για το διαχωρισμό των παραδειγμάτων των δύο τάξεων. Τα μέτρα αυτά είναι τα εξής:

- Μέτρο X^2. Αξιολογεί ένα χαρακτηριστικό υπολογίζοντας την τιμή του στατιστικού X^2 σε σχέση με την τάξη των παραδειγμάτων.

- Μέτρο λόγου κέρδους. Αξιολογεί ένα χαρακτηριστικό υπολογίζοντας το λόγο κέρδους (gain ratio) σε σχέση με την τάξη των παραδειγμάτων.

6.4.2 Ταξινομητές

Η υλοποίηση της προτεινόμενης μεθόδου χρησιμοποιεί 7 διαφορετικούς ταξινομητές. Όπως έχει ήδη αναφερθεί, οι προβλέψεις των ταξινομητών αυτών συνδυάζονται με τις μεθόδους πλειοψηφίας και σταθμισμένης πλειοψηφίας. Για τη δημιουργία των ταξινομητών χρησιμοποιήθηκαν οι προκαθορισμένες παράμετροι της βιβλιοθήκης Weka, έκτος από τις περιπτώσεις όπου αναφέρεται κάτι διαφορετικό. Οι ταξινομητές που δημιουργήθηκαν είναι οι εξής:

- Αφελής ταξινομητής Bayes (Naïve Bayes) [John & Langley, 1995].
- Δένδρο ταξινόμησης (αλγόριθμος C4.5) [Quinlan, 1993].
- Ταξινομητής *k*-Κοντινότερων γειτόνων (*k*-Nearest Neighbours – kNN) [Aha et al., 1991]. Ο αριθμός των κοντινότερων γειτόνων ορίστηκε *k*=1.
- Μηχανή διανυσμάτων υποστήριξης (SVM). Για τη δημιουργία αυτού του ταξινομητή χρησιμοποιήθηκε ο αλγόριθμος που παρέχει η βιβλιοθήκη Weka (Sequential Minimal Optimization – SMO [Platt, 1998; Keerthi et al., 2001]).
- Νευρωνικό δίκτυο χωρίς κρυφά επίπεδα και με ανάστροφη μετάδοση σφάλματος.
- Πίνακας απόφασης (decision table) [Kohavi, 1995b].
- Ταξινομητής προτασιακών κανόνων ταξινόμησης που δημιουργήθηκε με τον αλγόριθμο RIPPER (Repeated Incremental Pruning to Produce Error Reduction) [Cohen, 1995].

6.4.3 Πειραματική Αξιολόγηση

Σε αυτό το εδάφιο παρουσιάζονται το σύνολο δεδομένων που χρησιμοποιήθηκε για την εκπαίδευση και την αξιολόγηση της προτεινόμενης μεθόδου και τα αποτελέσματα από τα πειράματα που διεξήχθησαν.

Σύνολο Δεδομένων

Για την πειραματική αξιολόγηση της προτεινόμενης μεθόδου χρησιμοποιήθηκε ένα σύνολο 3.312 γονιδιακών αλληλουχιών (σύνολο δεδομένων «Verebrates»), οι οποίες ελήφθησαν από σπονδυλωτά ζώα (vertebrates) [Pedersen & Nielsen, 1997]. Οι αλληλουχίες του συγκεκριμένου συνόλου δεδομένων έχουν εξαχθεί από τη βάση δεδομένων GenBank (έκδοση 95) [Benson et al., 1997]. Οι αλληλουχίες που επιλέχθηκαν προέρχονται από γονίδια πυρηνικού DNA με καταγεγραμμένο σημείο έναρξης της μετάφρασης

και υπέστησαν επεξεργασία, με σκοπό την αφαίρεση των ιντρονίων[14]. Μετά την επεξεργασία, ελήφθησαν υπόψη μόνο εκείνες οι αλληλουχίες που περιλάμβαναν τουλάχιστον 10 νουκλεοτίδια ανοδικά του σημείου έναρξης της μετάφρασης, και τουλάχιστον 150 νουκλεοτίδια καθοδικά αυτού. Επιπλέον, το σύνολο δεδομένων έχει υποστεί προσεκτική επεξεργασία για την κατά το δυνατό ελαχιστοποίηση της πλεονάζουσας πληροφορίας.

Αξιολόγηση Προτεινόμενων Χαρακτηριστικών

Ο Πίνακας 6.2 παρουσιάζει τα 10 πρώτα χαρακτηριστικά των κατατάξεων των χαρακτηριστικών βάσει των δύο μέτρων, X^2 και λόγου κέρδους. Ανάμεσα στα 10 πρώτα χαρακτηριστικά της κατάταξης του μέτρου X^2 περιλαμβάνονται 8 από τα χαρακτηριστικά που προτείνονται στα πλαίσια της παρούσας διατριβής. Αντίστοιχα, μεταξύ των 10 πρώτων χαρακτηριστικών της κατάταξης του μέτρου λόγου κέρδους περιλαμβάνονται 6 από τα χαρακτηριστικά που προτείνονται στα πλαίσια της διατριβής. Πέντε από τα προτεινόμενα χαρακτηριστικά (↓G: θέση-κωδ. 1, ↓υδρόφοβα, ↕υδρόφοβα, ↓μη αρωμ./αλειφ., ↕μη αρωμ./αλειφ.) εμφανίζονται ταυτόχρονα μεταξύ των 10 πρώτων χαρακτηριστικών των κατατάξεων και των δύο μέτρων.

Πίνακας 6.2: Τα 10 πρώτα χαρακτηριστικά της κατάταξης βάσει των δύο μέτρων.

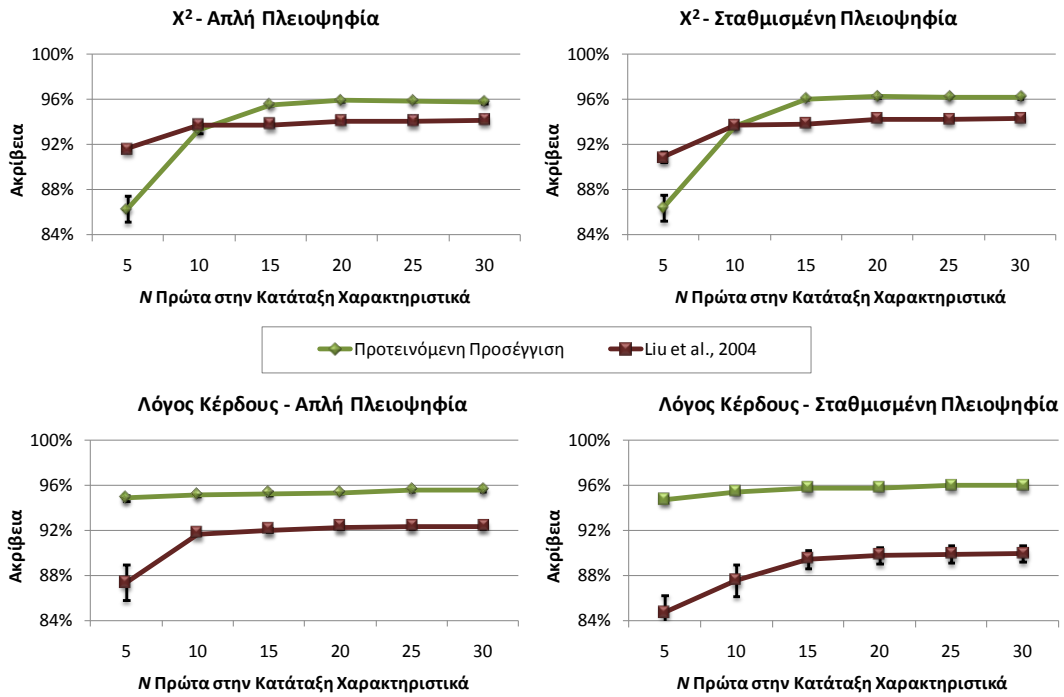
	Μέτρο X^2	**Μέτρο Λόγου Κέρδους**
1	↑ATG	↑ATG
2	↓G: θέση-κωδ. 1	↓λήξη
3	↓υδρόφοβα	↑M
4	↓μη αρωμ./αλειφ.	↑[AG]: θέση -3
5	↓C: θέση-κωδ. 3	↓μη αρωμ./αλειφ.
6	↓λήξη	↓G: θέση-κωδ. 1
7	↓αλειφατικά	↕μη αρωμ./αλειφ.
8	↕μη αρωμ./αλειφ.	↓υδρόφοβα
9	↕υδρόφοβα	↕υδρόφοβα
10	↓T: θέση-κωδ. 2	↓C: θέση-κωδ. 2

Η Εικόνα 6.5 παρουσιάζει μια σύγκριση της απόδοσης των ομάδων ταξινομητών που δημιουργήθηκαν χρησιμοποιώντας τα χαρακτηριστικά της προτεινόμενης μεθόδου έναντι των ομάδων ταξινομητών που δημιουργήθηκαν χρησιμοποιώντας τα χαρακτηριστικά προηγούμενης μελέτης [Liu et al., 2004]. Η μελέτη αυτή πρότεινε μια ιδιαίτερα αποτελεσματική προσέγγιση για την πρόβλεψη του σημείου έναρξης της μετάφρασης σε σχέση με τις προγενέστερες της προσεγγίσεις. Οι ράβδοι σφάλματος των γραφημάτων αναπαριστούν την τυπική απόκλιση. Πρέπει να σημειωθεί ότι σε κάποιες περιπτώσεις η τυπική απόκλιση είναι πολύ μικρή με αποτέλεσμα να μην είναι εμφανείς οι

[14] Ουσιαστικά αυτή η επεξεργασία μετατρέπει μια γονιδιακή αλληλουχία σε αλληλουχία παρόμοια με την αντίστοιχή της αλληλουχία cDNA.

ράβδοι σφάλματος. Ως διαδικασία επικύρωσης χρησιμοποιήθηκε η 10-πλή σταυρωτή επικύρωση.

Όπως παρατηρείται στα γραφήματα, η χρήση των χαρακτηριστικών που προτείνονται στην παρούσα διατριβή οδηγεί στη δημιουργία αποτελεσματικότερων ταξινομητών[15] και ομάδων ταξινομητών. Το παραπάνω συμπέρασμα δεν ισχύει στις περιπτώσεις των 5 και 10 χαρακτηριστικών με την κατάταξη του μέτρου X^2. Ο λόγος για τον οποίο συμβαίνει αυτό είναι ότι δεν συμπεριλήφθηκαν στα πρώτα 5 τα χαρακτηριστικά ↓λήξη και ↑[AG]: θέση -3 (πρότυπο συναίνεσης της Kozak) και στα πρώτα 10 το χαρακτηριστικό ↑[AG]: θέση -3. Η σημασία των συγκεκριμένων χαρακτηριστικών καταδεικνύεται από τη ραγδαία βελτίωση της απόδοσης που παρουσιάζεται όταν αυτά συμπεριληφθούν στην κατάταξη (μεταβάσεις από 5 σε 10 και από 10 σε 15 χαρακτηριστικά στα γραφήματα για το μέτρο X^2).



Εικόνα 6.5: Σύγκριση μεταξύ ομάδων ταξινομητών που δημιουργήθηκαν χρησιμοποιώντας τα χαρακτηριστικά της προτεινόμενης μεθόδου ή τα χαρακτηριστικά της μελέτης [Liu et al., 2004] (ράβδοι σφάλματος: ±τυπική απόκλιση).

Αξιολόγηση Προτεινόμενης Προσέγγισης

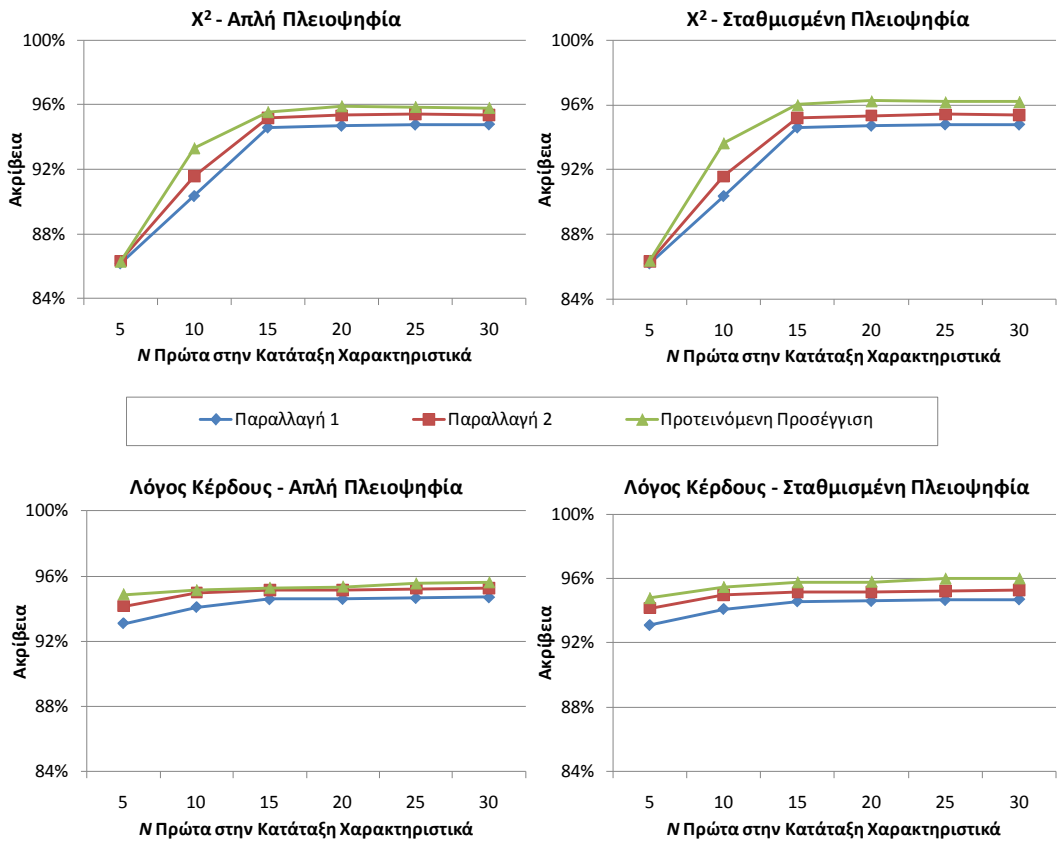
Στην παράγραφο αυτή παρουσιάζονται τα αποτελέσματα πειραμάτων που εκτελέστηκαν για τη σύγκριση της προτεινόμενης προσέγγισης με δύο παραλλαγές αυτής. Η προτεινόμενη μέθοδος και οι δύο παραλλαγές της χρησιμοποιούν μέρος από τις κατατάξεις των χαρακτηριστικων που περιλαμβάνει ο Πίνακας 6.1. Σε όλες τις προσεγγίσεις

[15] Σε όλες σχεδόν τις περιπτώσεις και για καθέναν από τους 7 ταξινομητές που δημιουργήθηκαν, η προσέγγιση που χρησιμοποίησε τα προτεινόμενα χαρακτηριστικά υπερτερεί έναντι της προγενέστερης προσέγγισης των Liu et al. [2004].

χρησιμοποιούνται ομάδες ταξινομητών (μέθοδος πλειοψηφίας) για την ταξινόμηση των κωδικονίων ATG ως σημείων έναρξης της μετάφρασης ή μη. Οι διαφορές κάθε προσέγγισης παρουσιάζονται παρακάτω:

- *Παραλλαγή 1*. Στην προσέγγιση αυτή όλα τα κωδικόνια ATG που λαμβάνουν την πλειοψηφία ταξινομούνται ως σημεία έναρξης της μετάφρασης.
- *Παραλλαγή 2*. Στην προσέγγιση αυτή μόνο το πρώτο κωδικόνιο ATG μιας αλληλουχίας cDNA το οποίο θα λάβει την πλειοψηφία θα ταξινομηθεί ως σημείο έναρξης της μετάφρασης. Τα υπόλοιπα κωδικόνια ATG της ίδιας αλληλουχίας cDNA δε θεωρούνται σημεία έναρξης της μετάφρασης ακόμη και αν έχουν λάβει την πλειοψηφία. Η προσέγγιση αυτή ουσιαστικά αποτελεί υλοποίηση του μοντέλου σάρωσης του ριβοσώματος.
- *Προτεινόμενη προσέγγιση*. Στην προτεινόμενη προσέγγιση μόνο το κωδικόνιο ATG μιας αλληλουχίας cDNA το οποίο θα λάβει τη μεγαλύτερη πλειοψηφία θα ταξινομηθεί ως σημείο έναρξης της μετάφρασης. Τα υπόλοιπα κωδικόνια ATG της ίδιας αλληλουχίας cDNA δε θεωρούνται σημεία έναρξης της μετάφρασης ακόμη και αν έχουν λάβει την πλειοψηφία.

Τα αποτελέσματα των πειραμάτων παρουσιάζονται στην Εικόνα 6.6. Ως διαδικασία επικύρωσης χρησιμοποιήθηκε η 10-πλή σταυρωτή επικύρωση.



Εικόνα 6.6: Σύγκριση της προτεινόμενης προσέγγισης με τις δύο παραλλαγές της.

Όπως παρατηρείται στα γραφήματα της εικόνας, σε κάθε περίπτωση η προτεινόμενη προσέγγιση υπερέχει των παραλλαγών της. Επιπλέον, η παραλλαγή 2 υπερέχει της παραλλαγής 1. Το συμπερασμα που προκύπτει από τα παραπάνω είναι ότι οι προσέγγισεις που επιβάλλεται να προβλέπουν μόνο ένα κωδικόνιο ATG ανά αλληλουχία cDNA ως σημείο έναρξης της μετάφρασης είναι αποτελεσματικότερες. Περαιτέρω αύξηση της απόδοσης παρέχεται αν ληφθεί ως σημείο έναρξης της μετάφρασης το κωδικόνιο ATG που θα λάβει τη μεγαλύτερη πλειοψηφία, παρά το πρώτο στην αλληλουχία κωδικόνιο ATG που θα λάβει απλά την πλειοψηφία.

Ο Πίνακας 6.3 παρουσιάζει τα αποτελέσματα διάφορων δημοφιλών προσεγγίσεων που χρησιμοποίησαν το σύνολο δεδομένων «Vertebrates». Η αποτελεσματικότητα των προσεγγίσεων εκτιμάται βάσει των μετρικών της ακρίβειας και της προσαρμοσμένης ακρίβειας. Τα στοιχεία προγενέστερων μελετών έχουν ληφθεί από τα αποτελέσματα που παρουσίασαν οι συγγραφείς στις αντίστοιχες εργασίες τους. Όπως παρουσιάζεται στον πίνακα, οι προτεινόμενες στα πλαίσια της διατριβής προσεγγίσεις, δηλαδή η μέθοδος πλειοψηφίας [Tzanis et al., 2006b] και η μέθοδος MANTIS [Tzanis et al., 2007] εμφανίζουν αρκετά υψηλή αποτελεσματικότητα στο συγκεκριμένο σύνολο δεδομένων, ξεπερνώντας όλες τις προγενέστερες προσεγγίσεις.

Πίνακας 6.3: Αποτελέσματα μελετών για το σύνολο δεδομένων «Verebrates».

Προσέγγιση	Ακρίβεια	Προσαρμοσμένη Ακρίβεια
Pedersen & Nielsen, 1997	85,00%	82,50%
Zien et al., 2000	88,10%	82,00%
Rajapakse and Ho, 2005	96,10%	95,35%
Liu et al., 2004	92,45%	88,34%
Tzanis et al., 2006b	96,25%	-
MANTIS[16]	97,26%	96,30%

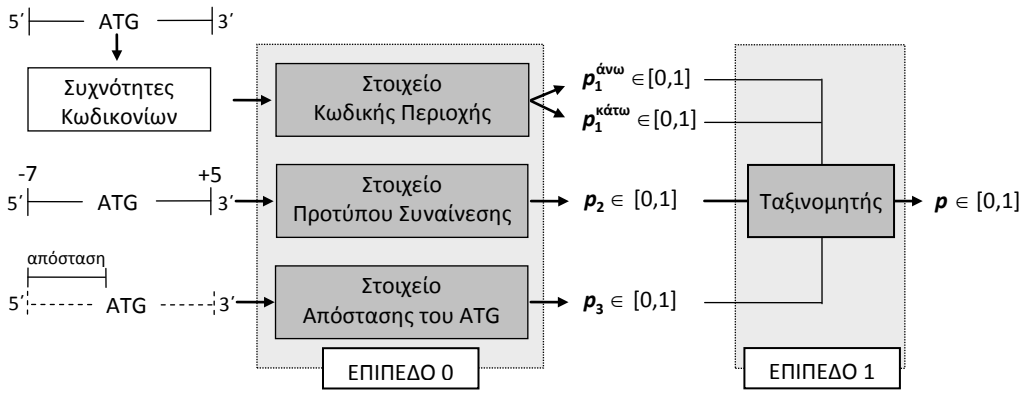
6.5 Η Μέθοδος StackTIS

Στην ενότητα αυτή παρουσιάζεται μια αποτελεσματικότερη και πιο ολοκληρωμένη προσέγγιση που αναπτύχθηκε στα πλαίσια της διατριβής με σκοπό την ακριβή πρόβλεψη του σημείου έναρξης της μετάφρασης. Η μέθοδος αυτή ονομάζεται StackTIS[17] και αποτελείται από τρία βασικά στοιχεία πρόβλεψης, τα οποία συνεισφέρουν στην τελική πρόβλεψη. Το καθένα από αυτά είναι προσανατολισμένο σε ένα διαφορετικό μέρος του προβλήματος και μπορεί να περιλαμβάνει έναν ή περισσότερους ταξινομητές. Οι ταξινομητές που ανήκουν στο ίδιο στοιχείο επιδιώκουν να προβλέψουν κατά πόσο ένα υποψήφιο σημείο έναρξης της μετάφρασης (κωδικόνιο ATG) είναι πραγματικά το

[16] Η συγκεκριμένη μέθοδος [Tzanis et al., 2007] θα παρουσιαστεί στο Εδάφιο 6.5.6. Το όνομα προέρχεται από τους όρους ***MA**chi**N**e Learning* και *Translation Initiation Site* (***TIS***) και παρομοιάζει την ικανότητα πρόβλεψης της μεθόδου με αυτή των μάντεων.

[17] Το όνομα προέρχεται από τους όρους ***Stack**ing* και *Translation Initiation Site* (***TIS***).

σημείο έναρξης της μετάφρασης βασιζόμενοι στα ιδιαίτερα χαρακτηριστικά του υποπροβλήματος που αντιπροσωπεύει το στοιχείο αυτό. Συγκεκριμένα, το πρώτο στοιχείο (*στοιχείο κωδικής περιοχής*) έχει ως σκοπό την αποτύπωση της διαφοράς των ιδιοτήτων κωδικής περιοχής εκατέρωθεν του κωδικονίου ATG, το δεύτερο (*στοιχείο προτύπου συναίνεσης*), επιδιώκει την ανακάλυψη προτύπων συναίνεσης (consensus patterns) σε μια πολύ μικρή περιοχή εκατέρωθεν του κωδικονίου ATG και το τρίτο είναι ένα μοντέλο που λαμβάνει υπόψη τη θέση του ATG στην αλληλουχία. Στην Εικόνα 6.7 φαίνεται αναλυτικότερα ο τρόπος λειτουργίας της μεθόδου StackTIS συμπεριλαμβάνοντας την είσοδο που λαμβάνει και την έξοδο που παράγει.



Εικόνα 6.7: Λειτουργία της μεθόδου StackTIS.

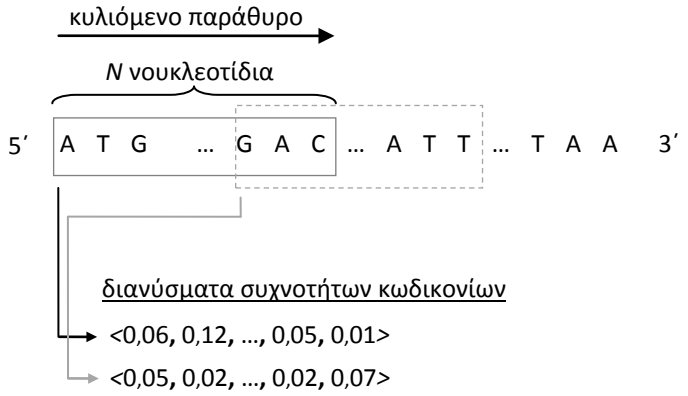
6.5.1 Το Στοιχείο Κωδικής Περιοχής

Το στοιχείο κωδικής περιοχής αποτελείται από έναν ταξινομητή που έχει ως σκοπό την αναγνώριση της διαφοράς των ιδιοτήτων κωδικής περιοχής εκατέρωθεν ενός υποψήφιου σημείου έναρξης της μετάφρασης σε μια αλληλουχία cDNA. Συγκεκριμένα, η βασική αποστολή του στοιχείου αυτού είναι να ανακαλύπτει ικανοποιητικού μεγέθους (συνήθως > 90 νουκλεοτιδίων) τμήματα της αλληλουχίας cDNA τα οποία έχουν ως αφετηρία τους ένα κωδικόνιο ATG και παρουσιάζουν χαρακτηριστικά κωδικής περιοχής. Αντίθετα, τα τμήματα της αλληλουχίας cDNA που βρίσκονται ανοδικά του ίδιου κωδικονίου ATG, πρέπει να παρουσιάζουν ελάχιστα χαρακτηριστικά κωδικής περιοχής. Στις περιπτώσεις που ισχύει κάτι τέτοιο, τα υπο εξέταση κωδικόνια ATG έχουν αρκετά μεγάλη πιθανότητα να είναι σημεία έναρξης της μετάφρασης.

Για την εκπαίδευση του ταξινομητή χρησιμοποιείται ένας αριθμός από θετικά και αρνητικά παραδείγματα τα οποία εξάγονται από τις αλληλουχίες cDNA ενός συνόλου εκπαίδευσης. Η εξαγωγή των παραδειγμάτων γίνεται με τη χρήση ενός κυλιόμενου παραθύρου σταθερού μεγέθους. Το παράθυρο αυτό μετατοπίζεται κατά μήκος της αλληλουχίας με καθορισμένο βήμα. Για κάθε θέση στην οποία τοποθετείται το παράθυρο εξάγεται ένα παράδειγμα εκπαίδευσης, το οποίο αναπαρίσταται από ένα διάνυσμα 64 συνιστωσών. Κάθε συνιστώσα αντιστοιχεί στη συχνότητα ενός συγκεκριμένου κωδικονίου. Η συχνότητα κάθε κωδικονίου υπολογίζεται από την καταμέτρηση των εμφανίσεών του στο τμήμα της αλληλουχίας που καλύπτεται από το αντίστοιχο παρά-

θυρο. Καθώς το παράθυρο μετατοπίζεται κατά μήκος της αλληλουχίας cDNA, ένα νέο διάνυσμα υπολογίζεται και με αυτόν τον τρόπο παράγεται ένα παράδειγμα εκπαίδευσης. Η διαδικασία παρουσιάζεται στην Εικόνα 6.8. Με την ίδια διαδικασία εξάγονται και τα διανύσματα των παραδειγμάτων ελέγχου, καθώς και των νέων στιγμιότυπων.

Τα θετικά παραδείγματα εκπαίδευσης εξάγονται από ένα κυλιόμενο παράθυρο σταθερού μήκους *N*, η αρχή του οποίου τοποθετείται στο πρώτο νουκλεοτίδιο που ακολουθεί το σημείο έναρξης της μετάφρασης κάθε αλληλουχίας cDNA του συνόλου εκπαίδευσης. Στα πλαίσια της διατριβής έχουν εκτελεστεί πειράματα για διάφορες τιμές του μήκους *N* του παραθύρου (Εικόνα 6.9). Τα αρνητικά παραδείγματα εκπαίδευσης εξάγονται από κυλιόμενο παράθυρο που έχει ίδιο μήκος με το παράθυρο που εξάγει και τα θετικά παραδείγματα. Ωστόσο, σε αυτήν την περίπτωση τα παραδείγματα εξάγονται από τις αμετάφραστες περιοχές 5′ UTR και 3′ UTR, καθώς και από τα τμήματα των δύο μη-κωδικών πλαισίων ανάγνωσης που καθορίζονται από τα όρια της κωδικής περιοχής. Αξίζει να σημειωθεί ότι τα αρνητικά παραδείγματα που εξάγονται από τα δύο μη-κωδικά πλαίσια ανάγνωσης επιλέγονται από τμήματα που δεν περιέχουν κωδικόνια λήξης της μετάφρασης. Αυτό γίνεται με σκοπό να διατηρηθεί η ευαισθησία του ταξινομητή στον εντοπισμό αλλαγών των συχνοτήτων των υπόλοιπων κωδικονίων παρά στην παρουσία ή απουσία των κωδικονίων λήξης. Η παρουσία κωδικονίου λήξης σε ένα τμήμα αλληλουχίας σημαίνει πως το συγκεκριμένο τμήμα δεν ανήκει σε κωδική περιοχή, καθώς οι κωδικές περιοχές δεν περιλαμβάνουν κωδικόνια λήξης της μετάφρασης. Συνεπώς, στις περιπτώσεις που εμφανίζεται ένα κωδικόνιο λήξης σε κάποιο τμήμα αλληλουχίας cDNA, η απόφαση του ταξινομητή είναι περιττή. Τελικά, είναι επιθυμητό ο ταξινομητής να εστιάζει στη διάκριση τμημάτων κωδικής περιοχής από τμήματα μη-κωδικής περιοχής, τα οποία δεν περιλαμβάνουν κωδικόνια λήξης.



Εικόνα 6.8: Εξαγωγή διανυσμάτων συχνοτήτων κωδικονίων για το στοιχείο κωδικής περιοχής.

Ο αριθμός των θετικών παραδειγμάτων εκπαίδευσης που εξάγονται από μια αλληλουχία cDNA είναι ίσος με τον αριθμό των αρνητικών παραδειγμάτων εκπαίδευσης που εξάγονται από την ίδια αλληλουχία. Με αυτόν τον τρόπο προκύπτει ένα συμμετρικό ως προς την τάξη των παραδειγμάτων σύνολο εκπαίδευσης. Ο αριθμός των παραδειγμάτων που εξάγονται από κάθε αλληλουχία περιορίζεται από δύο παραμέτρους. Η πρώτη

παράμετρος είναι το ελάχιστο επιτρεπτό μήκος μετατόπισης του κυλιόμενου παραθύρου (*minOffset*). Η παράμετρος αυτή χρησιμοποιείται με σκοπό την αποτροπή δημιουργίας παραδειγμάτων εκπαίδευσης που παρουσιάζουν μεγάλη ομοιότητα μεταξύ τους. Για παράδειγμα, αν ένα παράθυρο που περιλαμβάνει 50 κωδικόνια μετατοπίζεται κατά ένα κωδικόνιο κάθε φορά, τότε τα διαδοχικά παραδείγματα που θα προκύπτουν θα διαφέρουν μόλις κατά 1/50 ή 2%, με αποτέλεσμα τη δημιουργία ενός συνόλου εκπαίδευσης που περιλαμβάνει πολλά και ουσιαστικά, επαναλαμβανόμενα παραδείγματα. Στα πειράματα που παρουσιάζονται παρακάτω (Εδάφιο 6.5.6) η τιμή της παραμέτρου αυτής ορίστηκε σταθερή και ίση με 30 νουκλεοτίδια. Η δεύτερη παράμετρος είναι ο μέγιστος επιτρεπτός αριθμός παραδειγμάτων εκπαίδευσης για μια τάξη ανά αλληλουχία (*maxExamplesPerClass*). Χρησιμοποιείται ώστε να αποτραπεί η εξαγωγή μεγάλου αριθμού παραδειγμάτων από μια αλληλουχία. Είναι προφανές πως αν δε λαμβανόταν υπόψη αυτή η παράμετρος, τότε οι μεγαλύτερες αλληλουχίες cDNA θα προσέφεραν περισσότερα παραδείγματα εκπαίδευσης. Ωστόσο, είναι επιθυμητό κάθε αλληλουχία cDNA να συνεισφέρει κατά το δυνατό τον ίδιο αριθμό παραδειγμάτων με τις υπόλοιπες αλληλουχίες. Η τιμή της παραμέτρου αυτής ορίστηκε, επίσης, σταθερή κατά τη διεξαγωγή των πειραμάτων και ίση με 4 παραδείγματα ανα τάξη για κάθε αλληλουχία, δηλαδή συνολικά 8 παραδείγματα εκπαίδευσης ανα αλληλουχία cDNA.

Το πλήθος των παραδειγμάτων που τελικά εξάγονται και αντιπροσωπεύουν κάθε τάξη, υπολογίζεται βάσει των περιορισμών που θέτουν οι παραπάνω παράμετροι, τα μήκη των κωδικών και των αμετάφραστων περιοχών, καθώς και η αναλογία θετικών και αρνητικών παραδειγμάτων, που πρέπει να είναι 1:1. Τέλος, αξίζει να σημειωθεί ότι η αναλογία των αρνητικών παραδειγμάτων που εξάγονται από τις 5΄ UTR και 3΄ UTR επιδιώκεται να είναι όσο το δυνατόν πιο κοντά στο 1:1, ώστε να αντιπροσωπεύονται και οι δύο περιοχές στον ίδιο βαθμό.

Η έξοδος του στοιχείου κωδικής περιοχής περιλαμβάνει μια πρόβλεψη για την περιοχή ανοδικά του κωδικονίου ATG ($p_1^{\text{άνω}}$) και μια πρόβλεψη για την περιοχή καθοδικά του κωδικονίου ATG ($p_1^{\text{κάτω}}$). Θεωρητικά, οι τιμές αυτών των προβλέψεων για ένα σημείο έναρξης της μετάφρασης θα πρέπει να είναι $p_1^{\text{άνω}} \cong 0$ και $p_1^{\text{κάτω}} \cong 1$.

6.5.2 Το Στοιχείο Προτύπου Συναίνεσης

Σε προηγούμενες μελέτες [Kozak, 1987; Hatzigeorgiou, 2002; Rajapakse & Ho, 2005;] έχει καταδειχθεί ότι για τον εντοπισμό του σημείου έναρξης της μετάφρασης είναι ιδιαίτερα σημαντική η εξέταση της αλληλουχίας σε μια μικρή περιοχή γύρω από αυτό. Αυτός είναι και ο σκοπός του στοιχείου προτύπου συναίνεσης, το οποίο χρησιμοποιεί *αλυσίδες Markov* (*Markov chains*) προκειμένου να αποτυπώσει το πρότυπο συναίνεσης, το οποίο αρχίζει στη θέση -7 και τελειώνει στη θέση +5 (Εικόνα 6.4), με σημείο αναφοράς το υποψήφιο σημείο έναρξης της μετάφρασης. Η χρήση αλυσίδων Markov επιτρέπει την αποτύπωση όχι μόνο της πιθανότητας εμφάνισης ενός νουκλεοτιδίου σε μια συγκεκριμένη θέση, όπως κάνουν οι κλασικοί αλγόριθμοι εύρεσης προτύπων

συναίνεσης, αλλά και την αποτύπωση των πιθανών εξαρτήσεων της εμφάνισης ενός νουκλεοτιδίου από τις εμφάνισεις προηγούμενων στην αλληλουχία νουκλεοτιδίων.

Μια *αλυσίδα Markov* (*Markov chain*) είναι μια αλληλουχία καταστάσεων ενός συστήματος που έχει τη *ιδιότητα Markov*, η οποία περιγράφεται από την Εξίσωση 6.1.

$$P(s_i \mid s_{i-1}s_{i-2}...s_0) = P(s_i \mid s_{i-1}) \qquad 6.1$$

Η Εξίσωση 6.1 περιγράφει την αλυσίδα Markov 1ης τάξης, όπου κάθε κατάσταση εξαρτάται από την αμέσως προηγούμενή της. Υπάρχουν, ωστόσο, αλυσίδες Markov ανώτερης τάξης. Η τάξη μιας αλυσίδας Markov αντιστοιχεί στον αριθμό των αμέσως προηγούμενων καταστάσεων από τις οποίες εξαρτάται κάθε κατάσταση. Μια αλυσίδα Markov τάξης *k* (k^{th} *order Markov chain*) περιγράφεται από την Εξίσωση 6.2.

$$P(s_i \mid s_{i-1}s_{i-2}...s_0) = P(s_i \mid s_{i-1}s_{i-2}...s_{i-k}) \qquad 6.2$$

Μια αλυσίδα Markov χαρακτηρίζεται *ομοιογενής* (*homogeneous*) όταν οι πιθανότητες μετάβασης από τη μία κατάσταση στην άλλη δεν εξαρτώνται από το χρόνο (δηλ. τη θέση των καταστάσεων μέσα στην αλληλουχία), όπως περιγράφεται από την Εξίσωση 6.3.

$$P(s_{i+1} = x \mid s_i = y) = P(s_i = x \mid s_{i-1} = y) \qquad 6.3$$

Αντίστοιχα, μια αλυσίδα Markov χαρακτηρίζεται *μη-ομοιογενής* (*non-homogeneous*) όταν οι πιθανότητες μετάβασης από τη μία κατάσταση στην άλλη εξαρτώνται από τη θέση των καταστάσεων μέσα στην αλληλουχία. Αυτό γενικά περιγράφεται από την Εξίσωση 6.4. Μια μη-ομοιογενής αλυσίδα Markov μπορεί να καταγράψει την κατανομή των πιθανοτήτων μετάβασης ξεχωριστά για κάθε θέση μέσα στην αλληλουχία.

$$P(s_{i+1} = x \mid s_i = y) \neq P(s_i = x \mid s_{i-1} = y) \qquad 6.4$$

Η μοντελοποίηση ενός υποψήφιου σημείου έναρξης της μετάφρασης με αλυσίδα Markov γίνεται ως εξής: κάθε νουκλεοτίδιο που βρίσκεται στις θέσεις -7 έως +5, σε σχέση με το υποψήφιο σημείο έναρξης της μετάφρασης, αντιστοιχεί σε μια κατάσταση της αλυσίδας. Η εκ των υστέρων (a posteriori) πιθανότητα μιας αλληλουχίας με δεδομένη μια αλυσίδα Markov τάξης *k* (*M*) παρουσιάζεται από την Εξίσωση 6.5. Ουσιαστικά, με την Εξίσωση 6.5 μπορεί να παραχθεί μια πιθανότητα που δηλώνει κατά πόσο ένα νέο στιγμιότυπο (αλληλουχία που μοντελοποιεί ένα υποψήφιο σημείο έναρξης της μετάφρασης) έχει παραχθεί από την αλυσίδα *M* (δηλ. ανήκει στην τάξη/κλάση που μοντελοποιεί η *M*).

$$P(s_0 s_1 \ldots s_L \mid M) = \prod_{i=k+1}^{L} P(s_i \mid s_{i-1} s_{i-2} \ldots s_{i-k}) \qquad 6.5$$

Για τη δημιουργία από τα παραδείγματα του συνόλου εκπαίδευσης μιας αλυσίδας Markov που περιγράφει μια τάξη/κλάση χρησιμοποιήθηκε ο τύπος που περιγράφεται στην Εξίσωση 6.6. Ο τύπος αυτός εκτιμά τις πιθανότητες μετάβασης μεταξύ των καταστάσεων της αλυσίδας Markov.

$$\hat{P}(s_i \mid s_{i-1} s_{i-2} \ldots s_{i-k}) = \frac{\# s_i s_{i-1} s_{i-2} \ldots s_{i-k}}{\# s_{i-1} s_{i-2} \ldots s_{i-k}} \qquad 6.6$$

Για τη δημιουργία ενός ταξινομητή με αλυσίδες Markov εκπαιδεύονται δύο αλυσίδες, μια χρησιμοποιώντας μόνο παραδείγματα της θετικής τάξης και μια άλλη χρησιμοποιώντας μόνο παραδείγματα της αρνητικής τάξης. Όταν παρουσιάζεται ένα νέο στιγμιότυπο για ταξινόμηση, κάθε μια από τις δύο αλυσίδες παράγει μια εκτίμηση πιθανότητας για το στιγμιότυπο αυτό (S_+ και S_-). Οι δύο εκτιμήσεις κλιμακώνονται, ώστε το άθροισμά τους να είναι ίσο με τη μονάδα, ως εξής:

$$S'_+ = \frac{S_+}{S_+ + S_-} \text{ και } S'_- = \frac{S_-}{S_+ + S_-} \qquad 6.7$$

Η έξοδος του στοιχείου προτύπου συναίνεσης (p_2) είναι η S'_+.

6.5.3 Το Στοιχείο Απόστασης του ATG

Το στοιχείο απόστασης του ATG εκτιμά ένα κωδικόνιο ATG ως σημείο έναρξης της μετάφρασης ή μη, βάσει της απόστασής του από το 5΄ άκρο της αλληλουχίας. Πρακτικά, αντιπροσωπεύει την πληροφορία που δίνεται από το μοντέλο σάρωσης του ριβοσώματος. Σύμφωνα με το μοντέλο αυτό, το ριβόσωμα σαρώνει την αλληλουχία έως ότου συναντήσει το πρώτο κωδικόνιο ATG που βρίσκεται σε ένα κατάλληλο περιβάλλον νουκλεοτιδίων. Αυτό, ουσιαστικά, σημαίνει ότι το σημείο έναρξης της μετάφρασης είναι το κωδικόνιο ATG που έχει ταξινομηθεί ως σημείο έναρξης της μετάφρασης από έναν ταξινομητή και βρίσκεται πλησιέστερα στο 5΄ άκρο. Στην περίπτωση αυτή, όλα τα υπόλοιπα κωδικόνια ATG ταξινομούνται ως αρνητικά, ακόμα και εκείνα που συγκεντρώνουν μεγαλύτερη εκτίμηση πιθανότητας από τον ταξινομητή. Ωστόσο, το μοντέλο αυτό έχει ορισμένα μειονεκτήματα. Για παράδειγμα, σε κάποιες περιπτώσεις, συμβαίνει να υπάρχει ακόμα ένα κωδικόνιο ATG, όχι πολύ μετά το επιλεγμένο ATG και με μεγαλύτερη τιμή πρόβλεψης. Αν δεν είχε ληφθεί υπόψη το μοντέλο σάρωσης ριβοσώματος αλλά μόνο ο ταξινομητής, θα είχε επιλεγεί το δεύτερο κωδικόνιο ATG ως σημείο έναρξης της μετάφρασης. Οποιοδήποτε από τα δύο κωδικόνια είναι πολύ πιθανό να είναι το σημείο έναρξης της μετάφρασης. Αυτό που τελικά θα επιλεγεί εξαρτάται από την πολιτική επιλογής που ακολουθεί κάθε προσέγγιση.

Το συγκεκριμένο πρόβλημα μελετήθηκε παραπάνω (Εικόνα 6.6), όπου η παραλλαγή 2 βασιζόταν στο μοντέλο σάρωσης του ριβοσώματος και θα επέλεγε το πρώτο κωδικόνιο ATG του παραπάνω παραδείγματος. Αντίθετα, η προτεινόμενη προσέγγιση που βασίζεται μόνο στην πρόβλεψη του ταξινομητή θα επέλεγε το δεύτερο κωδικόνιο ATG του παραπάνω παραδείγματος. Όπως παρουσιάστηκε στην Εικόνα 6.6, η προτεινόμενη προσέγγιση υπερέχει της παραλλαγής 2 σε αποτελεσματικότητα, όμως σε κάποιες περιπτώσεις η σωστή επιλογή είναι αυτή που κάνει η παραλλαγή 2. Για το λόγο αυτό είναι απαραίτητος ο συνδυασμός του μοντέλου σάρωσης του ριβοσώματος και της πρόβλεψης του ταξινομητή με τη χρήση ενός μοντέλου που μαθαίνει πώς να τα συνδυάζει.

Ο συνδυασμός του μοντέλου σάρωσης του ριβοσώματος με τις πρόβλεψεις ταξινομητών πραγματοποιείται στη μέθοδο StackTIS μέσω της σύντηξης των προβλέψεων των επιμέρους στοιχείων πρόβλεψης (Εδάφιο 6.5.4). Το στοιχείο κωδικής περιοχής και το στοιχείο προτύπου συναίνεσης αποτελούν τους ταξινομητές που παρέχουν τις προβλέψεις, ενώ το στοιχείο απόστασης του ATG αντιπροσωπεύει το μοντέλο σάρωσης του ριβοσώματος.

Το στοιχείο απόστασης του ATG περιλαμβάνει δύο μοντέλα, ένα που εκπαιδεύεται χρησιμοποιώντας μόνο τα παραδείγματα της θετικής τάξης και άλλο ένα που εκπαιδεύεται χρησιμοποιώντας μόνο τα παραδείγματα της αρνητικής τάξης. Οι εκτιμήσεις πιθανότητας των δύο μοντέλων κλιμακώνονται στο διάστημα [0,1], όπως στις εξισώσεις 6.7. Η έξοδος του στοιχείου απόστασης του ATG (p_3) είναι η κλιμακωμένη στο διάστημα [0,1] εκτίμηση πιθανότητας της θετικής τάξης (S'_+).

6.5.4 Σύντηξη των Επιμέρους Προβλέψεων

Το τελικό στάδιο στη μέθοδο StackTIS είναι η σύντηξη των επιμέρους προβλέψεων των στοιχείων πρόβλεψης που περιγράφηκαν παραπάνω. Η συγκεκριμένη διαδικασία σύντηξης που υλοποιείται στη μέθοδο StackTIS αποτελεί μια δημοφιλή μέθοδο σύντηξης ταξινομητών, που ονομάζεται *στοιβαγμένη γενίκευση* (*stacked generalization*) ή *στοίβαγμα* (*stacking*). Η στοιβαγμένη γενίκευση είναι ένα σχήμα ελαχιστοποίησης του σφάλματος γενίκευσης ενός ή περισσοτέρων μοντέλων. Σύμφωνα με την τεχνική αυτή, ένας αριθμός ταξινομητών (ταξινομητές επιπέδου 0) οι οποίοι εκπαιδεύονται με τα αρχικά δεδομένα (δεδομένα επιπέδου 0), παράγουν τα δεδομένα εισόδου (δεδομένα επιπέδου 1) ενός ταξινομητή επιπέδου 1. Στη μέθοδο StackTIS τα δεδομένα επιπέδου 1 παράγονται από τους ταξινομητές των τριών στοιχείων πρόβλεψης που περιγράφηκαν παραπάνω.

Ο Πίνακας 6.4 παρουσιάζει τον αλγόριθμο βάσει του οποίου υλοποιείται η μέθοδος StackTIS, συνοψίζοντας όλα όσα περιγράφηκαν στις προηγούμενες παραγράφους. Πρέπει να σημειωθεί ότι χρησιμοποιείται ένα σύνολο των δεδομένων για την εκπαίδευση των στοιχείων πρόβλεψης και ένα διαφορετικό σύνολο των δεδομένων για την εκπαίδευση του ταξινομητή επιπέδου 1.

Πίνακας 6.4: Ο αλγόριθμος που υλοποιεί τη μέθοδο StackTIS.

Είσοδος: Ένα σύνολο *S* αλληλουχιών cDNA με γνωστό σημείο έναρξης της μετάφρασης (*TIS_Position*) για την εκπαίδευση των στοιχείων πρόβλεψης επιπέδου 0, ένα σύνολο *S'* αλληλουχιών cDNA με γνωστό σημείο έναρξης της μετάφρασης (*TIS_Position*) για την εκπαίδευση του ταξινομητή επιπέδου 1, το μήκος του παραθύρου για την εξαγωγή παραδειγμάτων του στοιχείου κωδικής περιοχής (*winLength*), το ελάχιστο επιτρεπτό μήκος μετατόπισης του κυλιόμενου παραθύρου (*minOffset*), ο μέγιστος επιτρεπτός αριθμός παραδειγμάτων εκπαίδευσης για μια τάξη ανά αλληλουχία (*maxExamplesPerClass*), μια μέθοδος ταξινόμησης (*ClassificationMethod0*) για τη δημιουργία του ταξινομητή κωδικής περιοχής και μια μέθοδος ταξινόμησης (*ClassificationMethod1*) για τη δημιουργία του ταξινομητή επιπέδου 1.

Έξοδος: Οι ταξινομητές που αποτελούν το σύστημα πρόβλεψης.

```
for each (<s, TIS_Position> ∈ S)
  <N, offset> ← numCodingExamples(
                  s, winLength, minOffset, maxExamplesPerClass)
  Coding+ ← Coding+ ∪ extractPositiveCodonFrequencyVectors(
                        s, winLength, N, offset)
  Coding- ← Coding- ∪ extractNegativeCodonFrequencyVectors(
                        s, winLength, N, offset)
  Concensus+ ← Concensus+ ∪ extractPositiveConsensusExample(
                            s, TIS_Position)
  Concensus- ← Concensus- ∪ extractRandomNagativeConsensusExample(
                            s,  TIS_Position)
  ATG_Dist+ ← ATG_Dist+ ∪ TIS_Position
  for each (ATG ∈ s | position(ATG) ≠ TIS_Position)
    ATG_Dist- ← ATG_Dist- ∪ calculateATG_Distance(ATG, s)
CodCl ← buildClassifier(Coding+ ∪ Coding-, ClassificationMethod0)
ConCl+ ← buildMarkovChain(Concensus+)
ConCl- ← buildMarkovChain(Concensus-)
ATGCl+ ← buildClassifier(ATG_Dist+)
ATGCl- ← buildClassifier(ATG_Dist-)
for each (<s, TIS_Position> ∈ S´)
  <p1^άνω, p1^κάτω, p2, p3>←getPredictions(s,CodCl,ConCl+,ConCl-,ATGCl+,ATGCl-)
  examples1 ← examples1 ∪ <p1^άνω, p1^κάτω, p2, p3>
Level1_Cl ← buildClassifier(examples1, ClassificationMethod1)
return <CodCl, ConCl+, ConCl-, ATGCl+, ATGCl-, Level1_Cl>
```

6.5.5 Τελική Πρόβλεψη του Σημείου Έναρξης της Μετάφρασης

Στο εδάφιο αυτό παρουσιάζεται η συνολική διαδικασία που ακολουθείται από τη μέθοδο StackTIS για την πρόβλεψη του σημείου έναρξης της μετάφρασης σε μια αλληλουχία cDNA. Η διαδικασία αυτή περιλαμβάνει τα εξής βήματα:

1. Εύρεση της μακρύτερης περιοχής καθενός εκ των τριών πλαισίων ανάγνωσης (στην κατεύθυνση 5'→3'), η οποία δεν περιλαμβάνει κωδικόνια λήξης της μετάφρασης.

2. Εύρεση όλων των κωδικονίων ATG που περιλαμβάνονται στην περιοχές που βρέθηκαν στο βήμα 1. Αυτά τα κωδικόνια και μόνον αυτά, θεωρούνται υποψήφια σημεία έναρξης της μετάφρασης.

3. Για καθένα από τα κωδικόνια ATG που εντοπίστηκαν στο βήμα 2, εφαρμόζονται τα ακόλουθα:

   i. Χρησιμοποιείται το στοιχείο κωδικής περιοχής για τον υπολογισμό μιας πρόβλεψης για την περιοχή ανοδικά του κωδικονίου ATG ($p_1^{\text{άνω}}$) και μιας πρόβλεψης για την περιοχή καθοδικά του κωδικονίου ATG ($p_1^{\text{κάτω}}$). Οι δύο προβλέψεις υπολογίζονται από δύο στιγμιότυπα που εξάγονται από την αλληλουχία cDNA χρησιμοποιώντας ένα παράθυρο ίδιου μεγέθους με το παράθυρο που χρησιμοποιήθηκε για την εξαγωγή των παραδειγμάτων εκπαίδευσης. Το στιγμιότυπο που χρησιμοποιείται για τον υπολογισμό της $p_1^{\text{άνω}}$ εξάγεται από παράθυρο του οποίου το τέλος συμπίπτει με το τελευταίο νουκλεοτίδιο πριν από το κωδικόνιο ATG. Αντίθετα, το στιγμιότυπο που χρησιμοποιείται για τον υπολογισμό της $p_1^{\text{κάτω}}$ εξάγεται από παράθυρο του οποίου η αρχή συμπίπτει με το πρώτο νουκλεοτίδιο που ακολουθεί το κωδικόνιο ATG.

   ii. Χρησιμοποιείται το στοιχείο προτύπου συναίνεσης για τον υπολογισμό της πρόβλεψης p_2 ($S_+^{'}$) για την περιοχή που αρχίζει στη θέση -7 και τελειώνει στη θέση +5 εκατέρωθεν του κωδικονίου ATG.

   iii. Χρησιμοποιείται το στοιχείο απόστασης του ATG για τον υπολογισμό της πρόβλεψης p_3 ($S_+^{'}$) βάσει της απόστασης του κωδικονίου ATG από το 5′ άκρο της αλληλουχίας cDNA.

   iv. Οι προβλέψεις που υπολογίστηκαν παραπάνω χρησιμοποιούνται ως είσοδος του ταξινομητή επιπέδου 1, ο οποίος θα υπολογίσει την τελική πρόβλεψη για το κωδικόνιο ATG.

4. Επιστρέφεται μια φθίνουσα κατάταξη των κωδικονίων ATG σύμφωνα με την τελική πρόβλεψη, με το πρώτο στη σειρά (υψηλότερη τιμή πρόβλεψης) να λογίζεται ως το σημείο έναρξης της μετάφρασης.

6.5.6 Πειραματική Αξιολόγηση

Στο εδάφιο αυτό παρουσιάζονται τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση και την αξιολόγηση της μεθόδου StackTIS, καθώς και τα αποτελέσματα από τα πειράματα που διεξήχθησαν.

Σύνολα Δεδομένων

Για την πειραματική αξιολόγηση της μεθόδου StackTIS χρησιμοποιήθηκαν τρία σύνολα αλληλουχιών cDNA. Το ένα από αυτά (H. sapiens 1) έχει χρησιμοποιηθεί και σε προη-

γούμενες μελέτες [Hatzigeorgiou, 2002; Li et al., 2005; Liu et al., 2005], ενώ τα άλλα δύο (H. sapiens 1 και O. sativa) δημιουργήθηκαν στα πλαίσια της παρούσας διατριβής και με αφορμή την αξιολόγηση της μεθόδου StackTIS. Όπως γίνεται κατανοητό από τα ονόματα των συνόλων δεδομένων τα δύο εξ' αυτών περιλαμβάνουν cDNA αλληλουχίες του ανθρώπου (Homo sapiens), ενώ το ένα περιλαμβάνει cDNA αλληλουχίες του ρυζιού (Oryza sativa). Το συγκεκριμένο είδος ρυζιού έχει σχετικά μικρό γονιδίωμα (περίπου $430x10^6$ βάσεις), φημίζεται για την ευκολία με την οποία μπορεί να τροποποιηθεί γενετικά και αποτελεί οργανισμό μοντέλο για τα δημητριακά.

Αλληλουχίες H. sapiens 1

Το σύνολο δεδομένων «H. sapiens 1» περιλαμβάνει 480 αλληλουχίες cDNA του οργανισμού Homo sapiens, δηλαδή του ανθρώπου [Hatzigeorgiou, 2002]. Εξήχθη από τη βάση δεδομένων πρωτεϊνών Swissprot[18]. Για το σκοπό αυτό συλλέχθηκαν όλες οι ανθρώπινες πρωτεΐνες των οποίων τα αμινοτελικά άκρα έχουν αλληλουχισθεί σε επίπεδο αμινοξέων και ελέγχθηκαν προσεκτικά. Στη συνέχεια, ανακτήθηκαν τα πλήρη mRNAs που αντιστοιχούν στις πρωτεΐνες με σημείο έναρξης της μετάφρασης του οποίου η ύπαρξη είχε επικυρωθεί πειραματικά. Τέλος, επιλέχθηκαν οι cDNA αλληλουχίες που αντιστοιχούν στα ανακτημένα mRNAs.

Αλληλουχίες H. sapiens 2

Το σύνολο δεδομένων «H. sapiens 2» περιλαμβάνει 2.351 αλληλουχίες cDNA του ανθρώπου [Tzanis et al., 2011]. Για τη δημιουργία του ακολουθήθηκε η διαδικασία που περιγράφεται παρακάτω. Αρχικά, εξήχθησαν από τη βάση δεδομένων Swiss-Prot όλες οι αναγνωρισμένες πλήρεις αλληλουχίες πρωτεϊνων του ανθρώπου (18.962 καταχωρήσεις). Ταυτόχρονα, εξήχθησαν από τη συλλογή αλληλουχιών αναφοράς (RefSeq[19]) της βάσης δεδομένων GenBank όλες οι αλληλουχίες cDNA του ανθρώπου (52.491 καταχωρήσεις). Στη συνέχεια, οι πρωτεϊνικές αλληλουχίες υποβλήθηκαν σε ολική, κατά ζεύγη στοίχιση αλληλουχιών μεταξύ τους, με σκοπό να επιλεγεί ένα μόνο μέρος αυτών, ώστε να μειωθεί η ομοιότητά τους. Συγκεκριμένα, επιλέχθηκε ένα σύνολο πρωτεϊνικών αλληλουχιών, τέτοιο ώστε οποιεσδήποτε δύο αλληλουχίες του συνόλου να έχουν μεταξύ τους ποσοστιαία ομοιότητα μικρότερη του 30%. Η διαδικασία αυτή πραγματοποιήθηκε με σκοπό την ελαχιστοποίηση της πόλωσης του συνόλου δεδομένων λόγω ομοιότητας μεταξύ των αλληλουχιών που περιλαμβάνει. Έπειτα, οι αλληλουχίες cDNA που είχαν εξαχθεί μεταφράστηκαν και συγκρίθηκαν με το επεξεργασμένο σύνολο των πρωτεϊνικών αλληλουχιών. Προέκυψαν 8.113 αλληλουχίες cDNA, οι οποίες ταυτίστηκαν με κάποια από τις πρωτεϊνικές αλληλουχίες. Τελικά, από τις 8.113 αλληλουχίες cDNA επιλέχθηκαν οι 2.351 που ικανοποιούσαν τους περιορισμούς μήκους της κωδικής

[18] http://www.expasy.org/sprot (προσπελάστηκε: 17/05/2011).

[19] Η συλλογή αλληλουχιών αναφοράς RefSeq [Pruitt, 2007] έχει ως σκοπό την παροχή ενός περιεκτικού, ολοκληρωμένου, μη-πλεονάζοντος, καλώς υπομνηματισμένου συνόλου αλληλουχιών γονιδιακού DNA, μεταγράφων και πρωτεϊνών.

περιοχής και των αμετάφραστων περιοχών (5′ UTR και 3′ UTR). Οι τελευταίοι περιορισμοί τέθηκαν με σκοπό τη διατήρηση αλληλουχιών που θα προσέφεραν ικανό αριθμό θετικών και αρνητικών παραδειγμάτων, ώστε να είναι δυνατή η αποτελεσματικότερη εκπαίδευση και αξιολόγηση των ταξινομητών.

Αλληλουχίες O. sativa

Το σύνολο δεδομένων «O. sativa» περιλαμβάνει 652 αλληλουχίες cDNA του ρυζιού [Tzanis et al., 2011]. Για τη δημιουργία του ακολουθήθηκε η ίδια διαδικασία με αυτή που περιγράφηκε παραπάνω για τη δημιουργία του συνόλου δεδομένων «H. sapiens 2». Από τη βάση δεδομένων Swiss-Prot εξήχθησαν 2.262 πρωτεϊνικές αλληλουχίες και από τη συλλογή αλληλουχιών αναφοράς (RefSeq) της βάσης δεδομένων GenBank 23.309 αλληλουχίες cDNA του ρυζιού. Έπειτα από την ολική στοιχιση και τη σύγκριση των μεταφρασμένων αλληλουχιών cDNA με τις εναπομείνασες πρωτεϊνικές αλληλουχίες, προέκυψαν 824 αλληλουχίες cDNA. Τελικά, από τις 824 αλληλουχίες cDNA επιλέχθηκαν οι 652 που ικανοποιούσαν τους περιορισμούς μήκους της κωδικής περιοχής και των αμετάφραστων περιοχών.

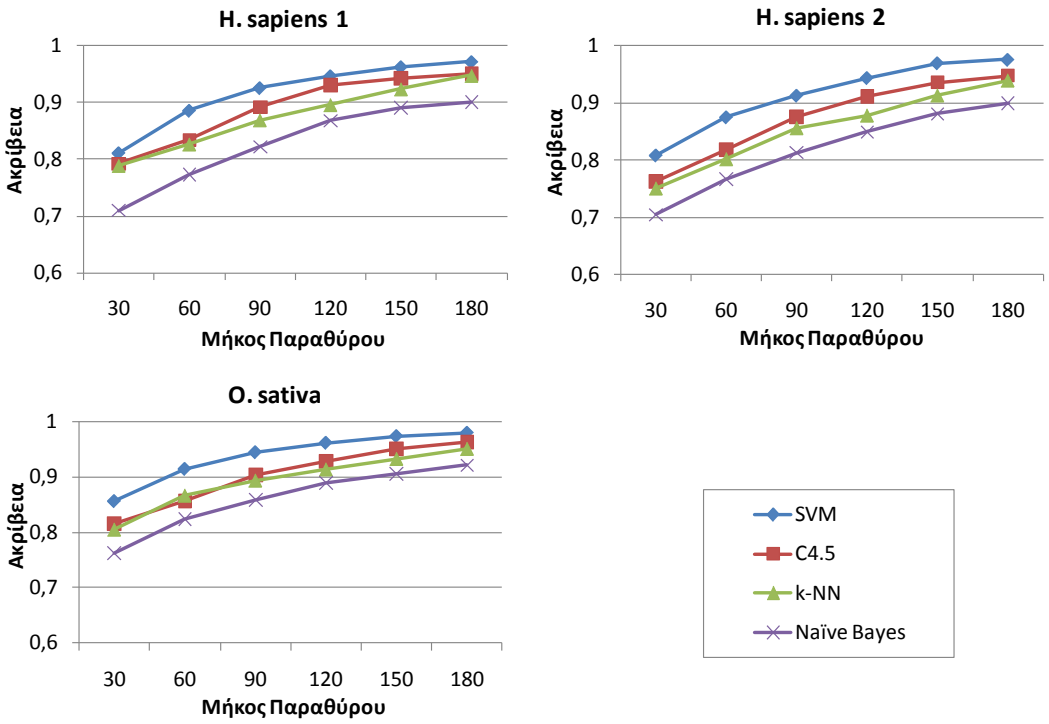
Αξιολόγηση του Στοιχείου Κωδικής Περιοχής

Για την αξιολόγηση των ταξινομητών του στοιχείου κωδικής περιοχής εκτελέστηκαν πειράματα με αλγορίθμους που αντιπροσωπεύουν τέσσερεις διαφορετικές κατηγορίες αλγορίθμων ταξινόμησης. Για τη δημιουργία των ταξινομητών χρησιμοποιήθηκαν οι προκαθορισμένες παράμετροι της βιβλιοθήκης Weka, εκτός αν αναφέρεται διαφορετικά. Οι ταξινομητές που δημιουργήθηκαν είναι οι εξής:

- Αφελής ταξινομητής Bayes (Naïve Bayes).
- Δένδρο ταξινόμησης (αλγόριθμος C4.5).
- Ταξινομητής *k*-Κοντινότερων γειτόνων (*k*-Nearest Neighbours – *k*-NN). Για τη δημιουργία αυτού του ταξινομητή πραγματοποιήθηκε κανονικοποίηση των τιμών των χαρακτηριστικών και αυτόματη επιλογή του βέλτιστου αριθμού κοντινότερων γειτόνων (μεταξύ 1 και 15) με τη διαδικασία σταυρωτής επικύρωσης με μία παράλειψη (leave-one-out cross-validation).
- Μηχανή διανυσμάτων υποστήριξης (SVM). Για τη δημιουργία αυτού του ταξινομητή έγινε χρήση της βιβλιοθήκης WLSVM [El-Manzalawy & Honavar, 2005]. Χρησιμοποιήθηκε πυρήνας RBF και οι παράμετροι κόστους (cost) και γ (gamma) επιλέχθηκαν μέσω διαδικασίας αναζήτησης πλέγματος (grid search[20]) με χρήση μιας ενσωματωμένης διαδικασίας σταυρωτής επικύρωσης. Η έξοδος του ταξινομητή είναι εκτιμήσεις πιθανότητας.

[20] Περισσότερες πληροφορίες για τη συγκεκριμένη διαδικασία αναζήτησης παρέχονται στο υλικό τεκμηρίωσης των πακέτων και κλάσεων της βιβλιοθήκης αλγορίθμων μηχανικής μάθησης Weka. URL: http://weka.sourceforge.net/doc.dev/weka/classifiers/meta/GridSearch.html (προσπελάστηκε: 17/05/2011).

Η Εικόνα 6.9 παρουσιάζει τα αποτελέσματα των πειραμάτων που διεξήχθησαν με σκοπό την αξιολόγηση των ταξινομητών του στοιχείου κωδικής περιοχής, οι οποίοι περιγράφηκαν παραπάνω. Ως διαδικασία επικύρωσης χρησιμοποιήθηκε η 10-πλή σταυρωτή επικύρωση. Επιπλέον, η παραπάνω διαδικασία επαναλήφθηκε έξι φορές για κάθε ταξινομητή, χρησιμοποιώντας κάθε φορά διαφορετικό μήκος κυλιόμενου παραθύρου για τη δημιουργία των συνόλων εκπαίδευσης και αξιολόγησης.



Εικόνα 6.9: Αξιολόγηση των ταξινομητών του στοιχείου κωδικής περιοχής.

Όπως παρουσιάζεται στα γραφήματα της εικόνας, ο ταξινομητής μηχανής διανυσμάτων υποστήριξης (SVM) υπερτερεί σε αποτελεσματικότητα, που μετράται με την ακρίβεια ταξινόμησης, έναντι των υπόλοιπων ταξινομητών. Επιπλέον, παρατηρείται ότι η ακρίβεια ταξινόμησης αυξάνεται καθώς αυξάνεται και το μήκος του κυλιόμενου παραθύρου που χρησιμοποιήθηκε για τη δημιουργία των συνόλων εκπαίδευσης και αξιολόγησης. Η βελτίωση της ακρίβειας ταξινόμησης με τη χρήση κυλιόμενων παραθύρων μεγαλύτερου μήκους είναι αναμενόμενη, επειδή όταν χρησιμοποιείται μεγαλύτερος αριθμός νουκλεοτιδίων για την εξαγωγή των παραδειγμάτων εκπαίδευσης, τότε παρέχεται περισσότερη πληροφορία στον ταξινομητή, ο οποίος μπορεί να την αξιοποιήσει, να εκπαιδευτεί καλύτερα και να επιτύχει μεγαλύτερη ακρίβεια πρόβλεψης. Μια ακόμη ενδιαφέρουσα παρατήρηση είναι ότι ο ρυθμός αύξησης της ακρίβειας ταξινόμησης σε σχέση με την αύξηση του μήκους του παραθύρου είναι φθίνων. Συνεπώς, υπάρχει μια ανώτατη τιμή, που στην προκειμένη περίπτωση δεν απέχει πολύ από το μέγιστο μήκος παραθύρου (180 νουκλεοτίδια) που χρησιμοποιήθηκε, πάνω από την

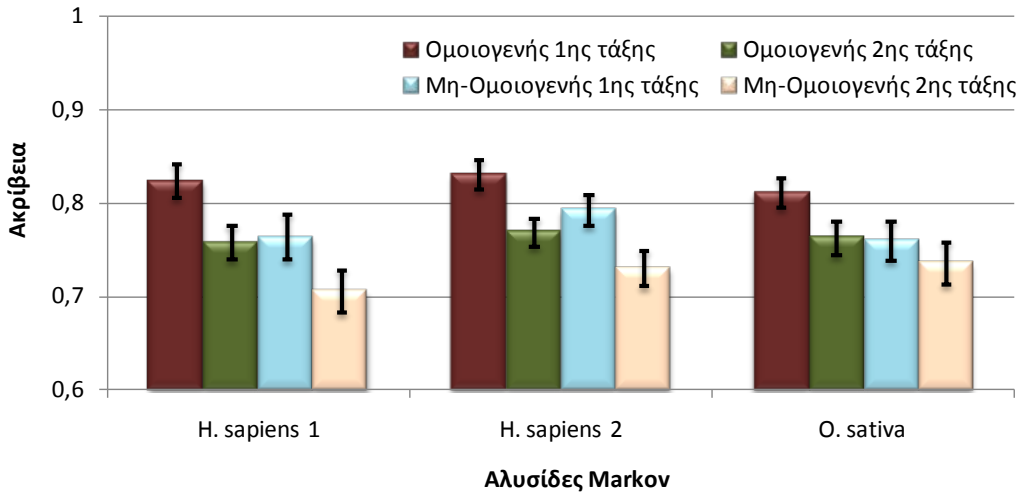
οποία η βελτίωση της ακρίβειας ταξινόμησης είναι ασήμαντη. Επιπλέον, το μήκος του κυλιόμενου παραθύρου πρέπει να έχει μια περιορισμένη τιμή, λόγω του ότι οι αλληλουχίες cDNA και οι κωδικές τους περιοχές έχουν και αυτές περιορισμένο μήκος. Για το λόγο αυτό στα πειράματα που θα περιγραφούν στη συνέχεια έχουν χρησιμοποιηθεί κυλιόμενα παράθυρα μήκους 120 νουκλεοτιδίων, που θεωρείται μια ικανοποιητική τιμή για την αντιμετώπιση τόσο των μεγάλων όσο και των μικρών αλληλουχιών cDNA.

Αξιολόγηση του Στοιχείου Προτύπου Συναίνεσης

Για την αξιολόγηση των ταξινομητών του στοιχείου προτύπου συναίνεσης εκτελέστηκαν πειράματα με τις τέσσερεις παρακάτω αλυσίδες Markov:

- Ομοιογενής αλυσίδα Markov 1ης τάξης.
- Ομοιογενής αλυσίδα Markov 2ης τάξης.
- Μη-ομοιογενής αλυσίδα Markov 1ης τάξης.
- Μη-ομοιογενής αλυσίδα Markov 2ης τάξης.

Η Εικόνα 6.10 παρουσιάζει τα αποτελέσματα των πειραμάτων που διεξήχθησαν με σκοπό την αξιολόγηση των ταξινομητών του στοιχείου προτύπου συναίνεσης. Ως διαδικασία επικύρωσης χρησιμοποιήθηκε και σε αυτήν την περίπτωση η 10-πλή σταυρωτή επικύρωση.



Εικόνα 6.10: Αξιολόγηση των ταξινομητών του στοιχείου προτύπου συναίνεσης (ράβδοι σφάλματος: ±τυπική απόκλιση).

Όπως παρουσιάζεται στο γράφημα της εικόνας, η ομοιογενής αλυσίδα Markov 1ης τάξης επιτυγχάνει την καλύτερη αποτελεσματικότητα, που μετράται με την ακρίβεια ταξινόμησης, έναντι των υπόλοιπων αλυσίδων Markov. Μάλιστα η υπεροχή της είναι σημαντική και υποδηλώνει ότι μια χαμηλότερου βαθμού αλυσίδα Markov μοντελοποιεί καλύτερα την πληροφορία που περικλύεται στην εγγύς περιοχή γύρω από ένα υποψήφιο σημείο έναρξης της μετάφρασης (κωδικόνιο ATG).

Μελέτη της Βαθμονόμησης των Ταξινομητών Επιπέδου 0

Ένας ταξινομητής *C*, που για κάθε παράδειγμα *x*, παράγει μια πρόβλεψη $0 \leq s(x) \leq 1$, καλείται *καλώς βαθμονομημένος* (*well-callibrated*), αν η εμπειρική πιθανότητα συμμετοχής σε μια τάξη *c* $P(c|s(x) = s)$ συγκλίνει στην τιμή πρόβλεψης $s(x) = s$, καθώς ο αριθμός των παραδειγμάτων τείνει στο άπειρο [Murphy & Winkler, 1977; Zadrozny & Elkan, 2002]. Πρακτικά, αυτό σημαίνει πως το 100*s*% των παραδειγμάτων για τα οποία ο ταξινομητής *C* αποδίδει μια τιμή πρόβλεψης *s*, θα πρέπει να είναι μέλη της τάξης *c*.

Στην παρούσα διατριβή μελετήθηκε η επίδραση της *βαθμονόμησης* (*callibration*) των πιθανοτικών προβλέψεων που παράγουν ως έξοδο οι ταξινομητές επιπέδου 0, πριν χρησιμοποιηθούν ως είσοδος στον ταξινομητή επιπέδου 1. Στα πλαίσια αξιολόγησης της μεθόδου StackTIS πραγματοποιήθηκαν πειράματα με προβλέψεις ταξινομητών που υποβλήθηκαν σε διαδικασία βαθμονόμησης αλλά και με προβλέψεις ταξινομητών που δεν υποβλήθηκαν σε αυτήν τη διαδικασία. Ο αλγόριθμος που χρησιμοποιήθηκε για τη βαθμονόμηση των ταξινομητών είναι ο *Pool Adjacent Violators* (*PAV*) [Ayer et al., 1955], ο οποίος πραγματοποιεί *ισοτονική παλινδρόμηση* (*isotonic regression*).

Τα αποτελέσματα αυτής της μελέτης οδήγησαν στο συμπέρασμα ότι η διαδικασία της βαθμονόμησης των ταξινομητών επιπέδου 0 δεν οδηγεί σε στατιστικά σημαντική βελτίωση της αποτελεσματικότητας του ταξινομητή επιπέδου 1. Αξίζει να σημειωθεί ότι τα αποτελέσματα αυτά συμβαδίζουν με τα συμπεράσματα προγενέστερης μελέτης [Caruana et al., 2006]. Οι συγγραφείς της παραπάνω εργασίας διερεύνησαν την επίδραση της βαθμονόμησης στη διαδικασία της επιλογής ομάδας ταξινομητών (ensemble selection) και ανακάλυψαν ότι τα βαθμονομημένα μοντέλα παρουσίαζαν βελτίωση μόνο του τετραγωνικού σφάλματος (squared error) και της διασταυρωμένης εντροπίας (cross-entropy). Μάλιστα, σύμφωνα με τη γνώμη των συγγραφέων, η βελτίωση παρουσιάστηκε εξαιτίας της επιλογής καλύτερων μοντέλων και όχι λόγω της βαθμονόμησης των μοντέλων.

Λαμβάνοντας υπόψη τα παραπάνω και για να αποφευχθεί η άσκοπη επιβάρυνση της μεθόδου StackTIS, η διαδικασία βαθμονόμησης των ταξινομητών δεν συμπεριλήφθηκε στη μέθοδο.

Αξιολόγηση της Μεθόδου StackTIS

Για την αξιολόγηση της μεθόδου StackTIS χρησιμοποιήθηκε μια μη τετριμμένη τυπική προσέγγιση (θα καλείται από εδώ και στο εξής «τυπική προσέγγιση») ως μέθοδος σύγκρισης. Αυτή η προσέγγιση αποτελείται από το συνδυασμό ενός στοιχείου κωδικής περιοχής και ενός στοιχείου προτύπου συναίνεσης ακολουθούμενα από την εφαρμογή του μοντέλου σάρωσης του ριβοσώματος. Στην προσέγγιση αυτή τα δύο στοιχεία συνδυάστηκαν με τη διαδικασία της στοιβαγμένης γενίκευσης, όπως και στη μέθοδο StackTIS, ενώ το μοντέλο σάρωσης του ριβοσώματος εφαρμόστηκε στην τελική πρόβλεψη που παρέχει το περιγραφόμενο σύστημα, έτσι ώστε να είναι δυνατή η σύγκρισή

του με τη μέθοδο StackTIS. Επιπλέον, η μεθόδος StackTIS συγκρίθηκε με την προγενέστερη μέθοδο MANTIS [Tzanis et al., 2007] της οποίας αποτελεί επέκταση.

Η μέθοδος MANTIS είναι μια αποτελεσματική προσέγγιση του προβλήματος της πρόβλεψης του σημείου έναρξης της μετάφρασης, η οποία έχει αξιολογηθεί με εκτενή πειράματα σε σύνολα αλληλουχιών cDNA διάφορων οργανισμών. Τα πειράματα έδειξαν πως η συγκεκριμένη μέθοδος επιτυχγάνει βελτιωμένη απόδοση σε σύγκριση με τις δημοφιλέστερες προγενέστερές της προσεγγίσεις. Η μέθοδος MANTIS αποτελείται και αυτή από τρία στοιχεία πρόβλεψης αντίστοιχα με εκείνα της μεθόδου StackTIS. Οι βασικές διαφορές της μεθόδου StackTIS από τη μέθοδο MANTIS παρουσιάζονται παρακάτω:

- Ο αριθμός των χαρακτηριστικών (Πίνακας 6.1 εκτός από τα χαρακτηριστικά 12, 13) που χρησιμοποιούνται για τη δημιουργία των ταξινομητών του στοιχείου κωδικής περιοχής της μεθόδου MANTIS είναι αρκετά μεγάλος (> 150). Επιπλέον, πολλά από τα χαρακτηριστικά αυτά έχουν εκ των πραγμάτων μεγάλη συσχέτιση μεταξύ τους. Για το λόγο αυτό χρησιμοποιούνται μέθοδοι επιλογής και μετασχηματισμού χαρακτηριστικών, όπως η ανάλυση κυρίων συνιστωσών (Principal Components Analysis – PCA), που μετασχηματίζει το χώρο των χαρακτηριστικών σε ένα νέο χώρο από πολύ λιγότερα και ορθογώνια μεταξύ τους χαρακτηριστικά. Αντίθετα, ο αριθμός των χαρακτηριστικών που χρησιμοποιούνται για τη δημιουργία των ταξινομητών του στοιχείου κωδικής περιοχής της μεθόδου StackTIS είναι 64, δηλαδή οι συχνότητες των 64 κωδικονίων και έτσι η χρήση της μεθόδου ανάλυσης κυρίων συνιστωσών δεν είναι απαραίτητη και αποφεύγεται.

- Όπως αναφέρθηκε στο Εδάφιο 6.5.1 σχετικά με την εκπαίδευση του ταξινομητή του στοιχείου κωδικής περιοχής της μεθόδου StackTIS, τα αρνητικά παραδείγματα που εξάγονται από τα δύο μη-κωδικά πλαίσια ανάγνωσης επιλέγονται από τμήματα που δεν περιέχουν κωδικόνια λήξης της μετάφρασης. Όπως εξηγήθηκε, ο σκοπός είναι να διατηρηθεί η ευαισθησία του ταξινομητή στον εντοπισμό αλλαγών των συχνοτήτων των υπόλοιπων κωδικονίων παρά στην παρουσία ή απουσία των κωδικονίων λήξης. Η ικανοποίηση του συγκεκριμένου περιορισμού δεν ελέγχεται στη μέθοδο MANTIS.

- Η μέθοδος StackTIS επιλέγει και χρησιμοποιεί έναν ταξινομητή (αυτόν που παρουσιάζει την καλύτερη απόδοση) για τη δημιουργία καθενός από τα τρία στοιχεία πρόβλεψης. Αντίθετα, η μέθοδος MANTIS δεν περιλαμβάνει τη συγκεκριμένη δυνατότητα της επιλογής, γεγονός που συχνά οδηγεί στη μείωση της αποτελεσματικότητας της μεθόδου λόγω της παρουσίας των πιο αδύναμων ταξινομητών.

- Διαφορές παρουσιάζονται και στη διαδικασία επιλογής των υποψήφιων σημείων έναρξης της μετάφρασης. Στη μέθοδο MANTIS όλα τα κωδικόνια ATG θεω-

ρούνται υποψήφια. Ωστόσο, στη μέθοδο StackTIS πολλά κωδικόνια ATG απορρίπτονται, αφού θεωρούνται υποψήφια μόνο εκείνα που περιλαμβάνονται στο επιμηκέστερο ανοικτό πλαίσιο ανάγνωσης καθενός εκ των τριών πλαισίων ανάγνωσης.

- Για την εξαγωγή παραδειγμάτων εκπαίδευσης και αξιολόγησης των ταξινομητών του στοιχείου της κωδικής περιοχής της μεθόδου MANTIS, χρησιμοποιείται ένα παράθυρο για κάθε κωδικόνιο ATG, τα οποίο μάλιστα στο κέντρο του έχει το αντίστοιχο κωδικόνιο ATG. Αντίθετα, τα παράθυρα που χρησιμοποιούνται για την εξαγωγή παραδειγμάτων εκπαίδευσης στη μέθοδο StackTIS δεν είναι κεντραρισμένα σε κάποιο κωδικόνιο ATG, ενώ για την πρόβλεψη ενός στιγμιότυπου χρησιμοποιούνται δύο παράθυρα, ένα για την περιοχή καθοδικά του κωδικονίου ATG και ένα για την περιοχή ανοδικά του κωδικονίου ATG.

Στη συγκεκριμένη μελέτη χρησιμοποιήθηκαν και αξιολογήθηκαν δύο αλγόριθμοι για τη δημιουργία ταξινομητών επιπέδου 1, ο αλγόριθμος γραμμικής παλινδρόμησης πολλαπλών μεταβλητών απόκρισης (Multi-response Linear Regression – MLR) και ο αλγόριθμος M5΄. Ο MLR θεωρείται κατάλληλος για τέτοιου είδους προσεγγίσεις, καθώς κατά τη στοιβαγμένη γενίκευση, που χρησιμοποιείται στη μέθοδο StackTIS, είναι απαραίτητη η χρήση πιθανοτικών προβλέψεων με τιμή που μπορεί να κυμαίνεται από 0 έως 1 παρά δυαδικών προβλέψεων με τιμή 0 ή 1 [Ting & Witten, 1999]. Ο αλγόριθμος MLR δημιουργεί ταξινομητές υπολογίζοντας πολλαπλές εξαρτημένες μεταβλητές χρησιμοποιώντας ένα κοινό σύνολο μεταβλητών εισόδου, ενώ ο αλγόριθμος M5΄ δημιουργεί ταξινομητές με αναπαράσταση δενδροειδούς μορφής για τάξεις συνεχούς συνόλου τιμών. Η στοιβαγμένη γενίκευση με ταξινομητή M5΄ έχει προταθεί ως επέκταση της αντίστοιχης μεθόδου με χρήση ταξινομητή MLR παρουσιάζοντας βελτιωμένη απόδοση [Dzeroski & Zenko, 2004].

Πραγματοποιήθηκαν εκτενή πειράματα με τα τρία σύνολα δεδομένων που περιγράφηκαν παραπάνω. Η διαδικασία επικύρωσης που χρησιμοποιήθηκε είναι η 10-πλή σταυρωτή επικύρωση. Ως ταξινομητής του στοιχείου κωδικής περιοχής χρησιμοποιήθηκε μια μηχανή διανυσμάτων υποστήριξης, η οποία όπως παρουσιάστηκε παραπάνω (Εικόνα 6.9), επιτυγχάνει την καλύτερη απόδοση. Ως ταξινομητής του στοιχείου προτύπου συναίνεσης χρησιμοποιήθηκε μια ομοιογενής αλυσίδα Markov 1ης τάξης, η οποία όπως επίσης παρουσιάστηκε παραπάνω (Εικόνα 6.10), επιτυγχάνει την καλύτερη απόδοση. Πρέπει να τονιστεί ότι όλα τα στάδια της διαδικασίας ανακάλυψης γνώσης (εξαγωγή χαρακτηριστικών, εκπαίδευση και αξιολόγηση ταξινομητών επιπέδου-0 και επιπέδου-1) ενσωματώθηκαν στη διαδικασία επικύρωσης, καθώς επίσης και ότι όλες οι επαναλήψεις πραγματοποιήθηκαν με τις ίδιες παρμέτρους και τα ίδια δεδομένα για κάθε συγκρινόμενη μέθοδο. Όλα τα παραπάνω εξασφαλίζουν μια όσο το δυνατόν πιο δίκαιη αξιολόγηση για όλες τις συγκρινόμενες μεθόδους.

Ο Πίνακας 6.5 παρουσιάζει τα αποτελέσματα της σύγκρισης των μεθόδων StackTIS, MANTIS και της τυπικής προσέγγισης με ταξινομητή επιπέδου 1 τον ταξινομητή MLR. Ο Πίνακας 6.6 παρουσιάζει τα αποτελέσματα της αντίστοιχης σύγκρισης με ταξινομητή επιπέδου 1 τον ταξινομητή M5΄. Όπως παρουσιάζεται στους δύο πίνακες, η μέθοδος StackTIS υπερτερεί έναντι των άλλων δύο προσεγγίσεων με κάθε σύνολο δεδομένων. Επιπλέον, ο ταξινομητής M5΄ παρουσιάζει ελαφρώς καλύτερη απόδοση από τον ταξινομητή MLR.

Πίνακας 6.5: Σύγκριση των μεθόδων StackTIS, MANTIS και της τυπικής προσέγγισης με ταξινομητή επιπέδου-1 τον ταξινομητή MLR.

	Ακρίβεια			**Προσαρμοσμένη Ακρίβεια**		
Σύνολο Δεδομένων	StackTIS	MANTIS	Τυπική Προσέγγιση	StackTIS	MANTIS	Τυπική Προσέγγιση
H. sapiens 1	99,32%	99,08%	98,70%	94,28%	92,99%	86,05%
H. sapiens 2	97,06%	96,14%	93,02%	96,42%	95,68%	89,14%
O. sativa	97,26%	97,12%	92,32%	96,87%	96,09%	88,04%

Πίνακας 6.6: Σύγκριση των μεθόδων StackTIS, MANTIS και της τυπικής προσέγγισης με ταξινομητή επιπέδου-1 τον ταξινομητή M5΄.

	Ακρίβεια			**Προσαρμοσμένη Ακρίβεια**		
Σύνολο Δεδομένων	StackTIS	MANTIS	Τυπική Προσέγγιση	StackTIS	MANTIS	Τυπική Προσέγγιση
H. sapiens 1	99,38%	99,14%	98,90%	94,52%	93,42%	87,62%
H. sapiens 2	97,36%	96,18%	93,16%	96,97%	95,79%	90,45%
O. sativa	97,82%	97,24%	92,89%	97,11%	96,82%	89,93%

Για τη στατιστική σύγκριση της μεθόδου StackTIS με τη μέθοδο MANTIS και την τυπική προσέγγιση εφαρμόστηκε η μη παραμετρική δοκιμασία προσημασμένης κατάταξης Wilcoxon (Wilcoxon signed-rank test). Πραγματοποιήθηκε σύγκριση των μεθόδων για κάθε σύνολο δεδομένων ανά επανάληψη της διαδικασίας της 10-πλής σταυρωτής επικύρωσης. Βάσει της δοκιμασίας, η υπεροχή της μεθόδου StackTIS έναντι της τυπικής προσέγγισης για καθένα από τα τρία σύνολα δεδομένων αποδεικνύεται στατιστικά σημαντική με συντελεστή βεβαιότητας 99% και για τα δύο μέτρα αξιολόγησης που χρησιμοποιήθηκαν (ακρίβεια και προσαρμοσμένη ακρίβεια). Η υπεροχή είναι στατιστικά σημαντική και για τους δύο ταξινομητές επιπέδου 1 (MLR και M5΄). Ο Πίνακας 6.7 περιλαμβάνει τα αποτελέσματα της στατιστικής σύγκρισης της μεθόδου StackTIS έναντι της μεθόδου MANTIS. Το «+(α)» υποδηλώνει τη στατιστικά σημαντική υπεροχή της μεθόδου StackTIS με συντελεστή βεβαιότητας 1-α. Πρέπει να σημειωθεί ότι σε όλες τις περιπτώσεις η υπεροχή της μεθόδου StackTIS έναντι της μεθόδου MANTIS είναι στατιστικά σημαντική με συντελεστή βεβαιότητας τουλάχιστον 90% και για τα δύο μέτρα αποτελεσματικότητας.

Πίνακας 6.7: Στατιστική σύγκριση της μεθόδου StackTIS με τη μέθοδο MANTIS.

	Ακρίβεια		Προσαρμοσμένη Ακρίβεια	
Σύνολο Δεδομένων	MLR	M5′	MLR	M5′
H. sapiens 1	+ (0,05)	+ (0,05)	+ (0,01)	+ (0,01)
H. sapiens 2	+ (0,01)	+ (0,01)	+ (0,01)	+ (0,01)
O. sativa	+ (0,10)	+ (0,05)	+ (0,05)	+ (0,05)

Ο Πίνακας 6.8 παρουσιάζει τα αποτελέσματα διάφορων δημοφιλών προσεγγίσεων που χρησιμοποίησαν το σύνολο δεδομένων «H. sapiens 1». Η αποτελεσματικότητα των προσεγγίσεων εκτιμάται βάσει των μετρικών της ακρίβειας και της προσαρμοσμένης ακρίβειας (Εδάφιο 2.4.3). Τα στοιχεία προγενέστερων μελετών έχουν ληφθεί από τα αποτελέσματα που παρουσίασαν οι συγγραφείς στις αντίστοιχες εργασίες τους. Όπως παρουσιάζεται στον πίνακα, οι προτεινόμενες προσεγγίσεις (MANTIS και StackTIS) εμφανίζουν αρκετά υψηλή αποτελεσματικότητα στο συγκεκριμένο σύνολο δεδομένων.

Πίνακας 6.8: Αποτελέσματα μελετών για το σύνολο δεδομένων «H. sapiens 1».

Προσέγγιση	Ακρίβεια	Προσαρμοσμένη Ακρίβεια
Hatzigeorgiou, 2002	-	94,00%
Li et al., 2005	-	95,24%
Liu et al., 2005	98,46%	85,34%
MANTIS	99,14%	93,42%
StackTIS	99,38%	94,52%

6.6 Συμπεράσματα

Σε αυτό το κεφάλαιο παρουσιάστηκαν διάφορες προσεγγίσεις για την πρόβλεψη του σημείου έναρξης της μετάφρασης. Προτάθηκαν και μελετήθηκαν νέα χαρακτηριστικά, τα οποία παρουσίασαν σημαντική ικανότητα χαρακτηρισμού των τάξεων στα πειράματα που διεξήχθησαν. Επίσης, αξιολογήθηκε ο τρόπος επιλογής του κωδικονίου που λογίζεται ως το σημείο έναρξης της μετάφρασης βάσει των αποφάσεων ταξινομητών, καθώς και βάσει του μοντέλου σάρωσης του ριβοσώματος.

Η πιο ολοκληρωμένη και αποτελεσματική μέθοδος προβλεψης του σημείου έναρξης της μετάφρασης που προτάθηκε στα πλαίσια της παρούσας διατριβής, είναι η μέθοδος StackTIS. Η μέθοδος αυτή αποτελείται από τρία βασικά στοιχεία πρόβλεψης, τα οποία συνεισφέρουν στην τελική πρόβλεψη. Το καθένα από αυτά είναι προσανατολισμένο σε ένα διαφορετικό μέρος του προβλήματος και μπορεί να περιλαμβάνει έναν ή περισσότερους ταξινομητές. Αξίζει να σημειωθεί πως η χρήση του στοιχείου απόστασης του ATG παρέχει στη μέθοδο τα πλεονεκτήματα του μοντέλου σάρωσης του ριβοσώματος, ενώ ταυτόχρονα ο συνδυασμός του με τα άλλα στοιχεία περιορίζει την επίδραση των μειονεκτημάτων του συγκεκριμένου μοντέλου. Τα τρία επιμέρους στοιχεία συνδυάζονται με μια ιδιαίτερα αποτελεσματική μέθοδο συνδυασμού ταξινομητών,

τη στοιβαγμένη γενίκευση. Η έξοδος που παρέχει η μέθοδος StackTIS είναι μια εκτίμηση της πιθανότητας κάποιου υποψήφιου στιγμιότυπου να είναι το σημείο έναρξης της μετάφρασης. Μάλιστα, στην περίπτωση τυπικών αλληλουχιών cDNA που περιλαμβάνουν περισσότερα από ένα υποψήφια σημεία έναρξης της μετάφρασης επιστρέφεται μια φθίνουσα κατάταξη των προβλέψεων βάσει της εκτιμηθείσας πιθανότητας.

Το πλεονέκτημα της χρήσης εκτιμήσεων πιθανότητας (τιμές από 0 έως 1) έναντι δυαδικών προβλέψεων (τιμές 0 ή 1) είναι ότι για κάθε πρόβλεψη ουσιαστικά παρέχεται και η εκτίμηση της βεβαιότητας με την οποία η μέθοδος περέχει την πρόβλεψη. Αυτό είναι ιδιαίτερα σημαντικό όταν περισσότερα από ένα κωδικόνια ATG έχουν προβλεφθεί ως σημεία έναρξης της μετάφρασης ή όταν είναι επιθυμητό να ληφθούν υπόψη και εναλλακτικά κωδικόνια ATG, που θα μπορούσαν να είναι το σημείο έναρξης της μετάφρασης της υπο εξέταση αλληλουχίας.

Εκτενή πειράματα με τρία σύνολα δεδομένων έδειξαν ότι η μέθοδος StackTIS παρουσιάζει ιδιαίτερα υψηλή αποτελεσματικότητα που ξεπερνά τις άλλες δημοφιλείς προσεγγίσεις που έχουν προταθεί για την αντιμετώπιση του συγκεκριμένου προβλήματος. Μάλιστα, η υπεροχή της μεθόδου StackTIS βάσει των μέτρων της ακρίβειας και της προσαρμοσμένης ακρίβειας αποδεικνύεται πως είναι στατιστικά σημαντική.

Πρόβλεψη του Σημείου Αποκοπής και Πολυαδενυλίωσης

7

7.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζεται η μέθοδος PolyA-iEP, η οποία αναπτύχθηκε στα πλαίσια της διατριβής για την πρόβλεψη του σημείου αποκοπής και πολυαδενυλίωσης (cleavage-polyadenylation site). Η αποκοπή και πολυαδενυλίωση είναι μια μετα-μεταγραφική διαδικασία που λαμβάνει χώρα μόνο στους ευκαρυωτικούς οργανισμούς και περιλαμβάνει την αποκοπή ενός τμήματος από το 3΄ άκρο ενός μορίου mRNA και την αντικατάσταση του αποκοπέντος τμήματος από μια αλληλουχία καταλοίπων αδενίνης. Στόχος της διαδικασίας αυτής είναι η προστασία του μορίου mRNA από αποικοδόμηση, ώστε αυτό να φτάσει άθικτο στην περιοχή της πρωτεϊνοσύνθεσης, μεταφέροντας έτσι αυτούσια την πληροφορία του DNA.

Η προτεινόμενη μέθοδος υλοποιήθηκε στη γλώσσα προγραμματισμού JAVA, ενώ χρησιμοποιήθηκε και η βιβλιοθήκη αλγορίθμων μηχανικής μάθησης Weka [Witten & Frank, 2005].

7.2 Βασικές Έννοιες

Στην ενότητα αυτή περιγράφονται οι βασικές έννοιες που απαιτούνται για την κατανόηση της λειτουργίας της μεθόδου PolyA-iEP. Παρουσιάζονται έννοιες που αφορούν τις βιολογικές διαδικασίες της αποκοπής και της πολυαδενυλίωσης καθώς και κάποιες τεχνικές και αλγοριθμικές λεπτομέρειες σχετικά με την εξόρυξη αναδυόμενων προτύπων και την ταξινόμηση βάσει αυτών.

7.2.1 Αποκοπή και Πολυαδενυλίωση

Η διαδικασία της αποκοπής και πολυαδενυλίωσης λαμβάνει χώρα αμέσως μετά τη μεταγραφή. Το ένζυμο που καταλύει τη διαδικασία της μεταγραφής, η RNA πολυμερά-

ση, μεταγράφει πέρα από την θέση που αντιστοιχεί στο 3΄ άκρο του mRNA. Μερικές αλληλουχίες στο mRNA αναγνωρίζονται ως στόχοι για ενδονουκλεοτιδική πέψη, δηλαδή αποκοπή της αλληλουχίας. Μετά την αποκοπή ακολουθεί η πολυαδενυλίωση, δηλαδή η προσθήκη μιας σειράς νουκλεοτιδίων αδενυλικού οξέος, που καλείται πολυαδενυλικό οξύ ή πολύ(A). Κατά αντιστοιχία το σημείο αποκοπής και πολυαδενυλίωσης αναφέρεται και ως σημείο πολύ(A)[21], ενώ η 3΄ τερματική αλληλουχία από κατάλοιπα αδενίνης συχνά περιγράφεται ως ουρά πολυ(A), ενώ το μόριο mRNA που έχει αυτό το χαρακτηριστικό δηλώνεται ως πολυ(A)$^+$.

Ένα κοινό χαρακτηριστικό των μορίων mRNA στους περισσότερους ανώτερους ευκαρυώτες, είναι η παρουσία της αλληλουχίας AAUAAA σε μια περιοχή 11-30 νουκλεοτιδίων ανοδικά της θέσης που θα προστεθεί η ουρά πολυ(A). Η διαγραφή ή μετάλλαξη του εξαμερούς AAUAAA εμποδίζει την παραγωγή του πολυαδενυλιωμένου 3΄ άκρου [Sheets et al., 1990]. Το σήμα χρειάζεται τόσο για την αποκοπή όσο και για την πολυαδενυλίωση.

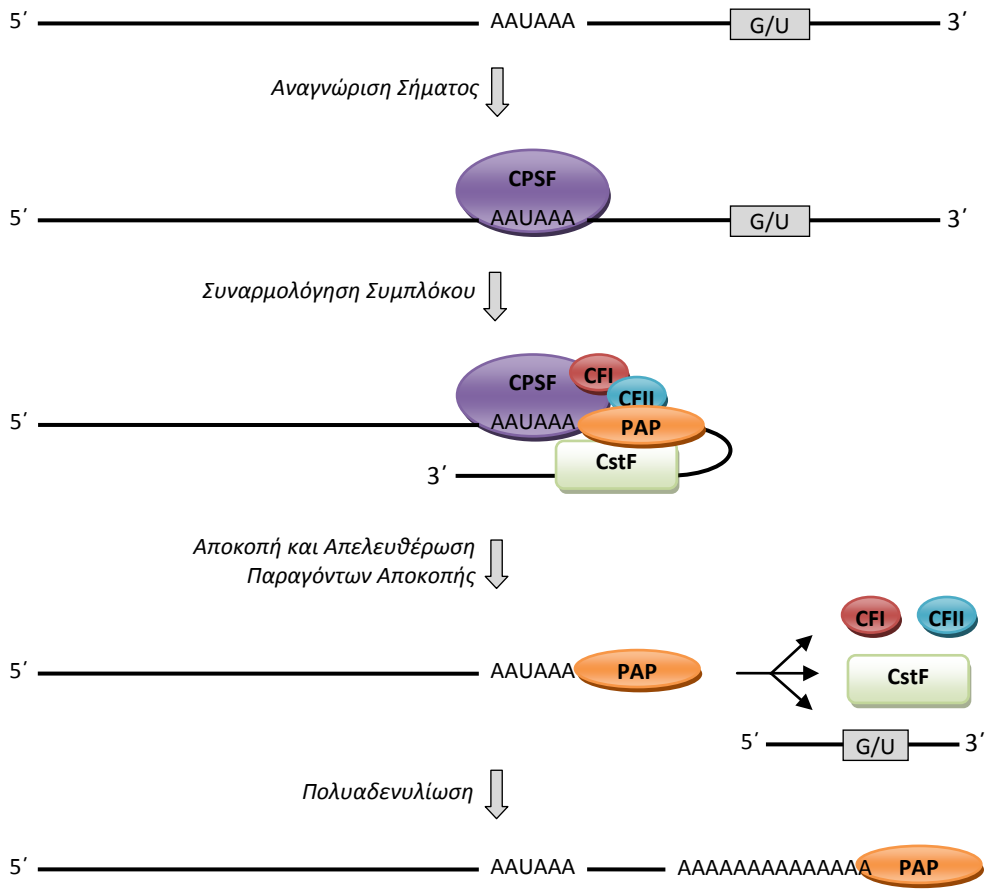
Ένα σύμπλοκο πρωτεϊνών αναλαμβάνει να διεκπεραιώσει τόσο την αποκοπή όσο και την πολυαδενυλίωση. Το σύμπλοκο που πραγματοποιεί την αποκοπή και την πολυαδενυλίωση αποτελείται από μία σειρά εξειδικευμένων παραγόντων οι βασικότεροι από τους οποίους είναι οι εξής (Εικόνα 7.1):

- Ένας ειδικός παράγοντας CPSF (Cleavage and Polyadenylation Specificity Factor), ο οποίος αρχικά αναγνωρίζει την αλληλουχία AAUAAA και έπειτα κατευθύνει τις ενέργειες που απαιτούνται για τη διεξαγωγή της αποκοπής και πολυαδενυλίωσης.
- Μια ενδονουκλεάση που αποτελείται από τους παράγοντες αποκοπής I και II (CFI και CFII) και πραγματοποιεί την αποκοπή του mRNA.
- Ένας παράγοντας ενίσχυσης CstF (Cleavage stimulatory Factor), που προσδένεται σε μια αλληλουχία πλούσια σε G/U καθοδικά της περιοχής αποκοπής.
- Μια πολυ(A) πολυμεράση (PAP), για τη σύνθεση της ουράς πολύ(A).

Η πολυαδενυλίωση σταθεροποιεί το mRNA, προστατεύοντας το από αποικοδόμηση από το 3΄ άκρο, καθώς το 5΄ άκρο είναι ήδη προστατευμένο με την καλύπτρα (Εδάφιο 3.5.3). Η RNA πολυμεράση συνεχίζει τη μεταγραφή πέραν του σημείου αποκοπής, αλλά το επιπλέον τμήμα του mRNA αποικοδομείται γρήγορα, καθώς το 5΄ άκρο που δημιουργείται από την αποκοπή είναι απροστάτευτο. Μόνο μετά την ολοκλήρωση όλων των τροποποιήσεων και της επεξεργασίας, μπορεί το mRNA να εξαχθεί από τον πυρήνα στο κυτταρόπλασμα. Κατά μέσο όρο, το mRNA καθυστερεί περίπου 20 λεπτά για να εξαχθεί από τον πυρήνα. Μόλις το mRNA εισέλθει στο κυτταρόπλασμα, αναγνωρίζεται από τα ριβοσώματα και μεταφράζεται.

---

[21] Οι όροι *σημείο αποκοπής και πολυαδενυλίωσης*, *σημείο αποκοπής, σημείο πολυαδενυλίωσης,* και *σημείο πολύ(A)* χρησιμοποιούνται εναλλακτικά προσδιορίζοντας το ίδιο ακριβώς σημείο.



Εικόνα 7.1: Η διαδικασία της αποκοπής και πολυαδενυλίωσης του mRNA.

Τα δύο κυριότερα προβλήματα με τα οποία ασχολείται η Βιοπληροφορική, σε σχέση με την αποκοπή και πολυαδενυλίωση, είναι ο προσδιορισμός του σημείου στο οποίο θα γίνει η αποκοπή και εν συνεχεία η πολυαδενυλίωση και ο προσδιορισμός των *ρυθμιστικών στοιχείων* (*cis-regulatory elements*), δηλαδή σημάτων ή προτύπων, που βρίσκονται ανοδικά (upstream) και καθοδικά (downstream) του σημείου αυτού.

Πρόσφατες έρευνες έχουν δείξει ότι περισσότερα από τα μισά ανθρώπινα γονίδια έχουν πολλαπλά σημεία πολυαδενυλίωσης [Tian et al., 2005; Yan & Marr, 2005]. Όπως είναι προφανές η *εναλλακτική πολυαδενυλίωση (alternative polyadenylation)* συμβάλλει στην πολυπλοκότητα που παρουσιάζουν τα ανθρώπινα κύτταρα σε σχέση με τους άλλους ευκαρυώτες παράγοντας μόρια mRNA με διαφορετικές 3′ αμετάφραστες περιοχές (3′ UTRs) και κωδικοποιώντας έτσι διάφορων ειδών πρωτεΐνες. Η εναλλακτική πολυαδενυλίωση μπορεί να έχει πολλές επιδράσεις στον μεταβολισμό του mRNA, καθώς τα 3′ UTRs περιέχουν διάφορα ρυθμιστικά στοιχεία όπως στοιχεία πλούσια σε αδενίνη και γουανίνη (AU-rich elements), που είναι υπεύθυνα για τη σταθερότητα του mRNA ή άλλα στοιχεία απαραίτητα για την μετάφρασή του. Η επίδραση της εναλλακτικής πολυαδενυλίωσης στην διαδικασία της μετάφρασης είναι συνήθως συνδεδεμένη με το *εναλλακτικό μάτισμα* (*alternative splicing*) [Edwalds-Gilbert et al., 1997], πράγμα που έχει αποδειχθεί για διάφορα γονίδια.

7.2.2 Ταξινόμηση με Αναδυόμενα Πρότυπα

Όπως έχει ήδη περιγραφεί τα αναδυόμενα πρότυπα είναι σύνολα αντικειμένων των οποίων η υποστήριξη διαφέρει σημαντικά μεταξύ δύο συνόλων δεδομένων και μπορούν να χρησιμοποιηθούν για ταξινόμηση (Εδάφιο 2.3.2). Για το σκοπό αυτό υπολογίζονται ζεύγη βαθμολογιών σύμφωνα με τις Εξισώσεις 2.5 και 2.6. Η μία βαθμολογία καταδεικνύει αν η αλληλουχία είναι θετική και η άλλη αν είναι αρνητική. Η τελική απόφαση μπορεί να ληφθεί με τη σύγκριση των δυο τιμών και την ταξινόμηση της αλληλουχίας στην τάξη με τη μεγαλύτερη βαθμολογία. Ωστόσο, επειδή το μέγεθος των δύο συνόλων αναδυόμενων προτύπων μπορεί να διαφέρει σημαντικά οι βαθμολογίες πρέπει να προσαρμοστούν κατάλληλα. Στην παρούσα διατριβή έχουν μελετηθεί τρεις μέθοδοι για την προσαρμογή των βαθμολογιών [Tzanis et al., 2008], οι οποίες περιγράφονται παρακάτω:

1. Η πρώτη μέθοδος προτάθηκε από τους [Dong et al., 1999]. Υπολογίζει δυο μέσες βαθμολογίες, $base_+$ και $base_-$ για την θετική και αρνητική τάξη αντίστοιχα. Η βαθμολογία $base_+$ ορίζεται υπολογίζοντας τη θετική βαθμολογία για κάθε θετική αλληλουχία και θέτοντας ως $base_+$ το μέσο όρο των βαθμολογιών. Αντίστοιχα, υπολογίζεται και η $base_-$.

2. Η δεύτερη μέθοδος χρησιμοποιεί την εντροπία με σκοπό να επιλεγεί ένα κατώφλι για το εξής κλάσμα:

$$\frac{score(T,+)}{score(T,-)}.$$

   Αυτό το κλάσμα υπολογίζεται για κάθε αλληλουχία (αρνητική ή θετική) στο σύνολο δεδομένων και βρίσκεται ένα σημείο *entropy_threshold*, όπου μεγιστοποιείται το κέρδος πληροφορίας. Μια νέα αλληλουχία ταξινομείται ως θετική, αν το κλάσμα της υπερβαίνει το *entropy_threshold*, αλλιώς ταξινομείται ως αρνητική.

3. Η τρίτη μέθοδος περιλαμβάνει το συνδυασμό των δύο παραπάνω. Προτείνεται ένα κατώφλι για το κλάσμα της μεθόδου 2. Το κατώφλι υπολογίζεται ως εξής:

$$entropy_base = \frac{entropy_threshold + \frac{base_+}{base_-}}{2} \qquad 7.1$$

Η μέθοδος 1 ευνοεί την τάξη με το μικρότερο πλήθος παραδειγμάτων στα δεδομένα εκπαίδευσης. Αντίθετα, η περίπτωση στο 2 ευνοεί την τάξη με τα περισσότερα παραδείγματα στα δεδομένα εκπαίδευσης. Για αυτόν το λόγο προτάθηκε και χρησιμοποιείται η μέθοδος 3, η οποία μετριάζει τη μεροληψία των άλλων δύο μεθόδων. Πρέπει να διευκρινιστεί ότι η ταξινόμηση με βάση τις παραπάνω περιγραφείσες μεθόδους δε λαμβάνει χώρα στη μέθοδο PolyA-iEP, καθώς οι βαθμολογίες που υπολογίζονται τροφοδοτούνται ως είσοδος σε ταξινομητή, ο οποίος τις συνδυάζει και λαμβάνει την τελική

απόφαση. Ωστόσο, χρησιμοποιούνται στα πειράματα που παρουσιάζονται παρακάτω με σκοπό την αξιολόγηση της μεθόδου PolyA-iEP (Εδάφιο 7.5.4).

Ενδιαφέροντα Αναδυόμενα Πρότυπα

Ένα μειονέκτημα των αναδυόμενων προτύπων (Εδάφιο 2.3.2) είναι ότι το πλήθος τους μπορεί να είναι πολύ μεγάλο, ιδίως όταν το κατώφλι υποστήριξης (support) και το κατώφλι του ρυθμού ανάπτυξης (growth rate) έχουν τεθεί πολύ χαμηλά. Η αύξηση των ορίων δεν είναι απαραίτητα σωστή λύση καθώς υπάρχει κίνδυνος κάποια πολύτιμα αναδυόμενα πρότυπα να μην ανακαλυφθούν. Για παράδειγμα, αν το κατώφλι της υποστήριξης τεθεί αρκετά υψηλό, τότε τα αναδυόμενα πρότυπα με χαμηλή υποστήριξη αλλά υψηλό ρυθμό ανάπτυξης θα χαθούν. Από την άλλη, αν το κατώφλι του ρυθμού ανάπτυξης τεθεί υψηλό, τότε τα αναδυόμενα πρότυπα με χαμηλό ρυθμό ανάπτυξης αλλά υψηλή υποστήριξη θα χαθούν. Για την επίλυση αυτού του προβλήματος έχουν προταθεί κάποια μέτρα ενδιαφέροντος με στόχο τη μείωση του αριθμού των προτύπων που ανακαλύπτονται χωρίς να θυσιάζονται πολύτιμα πρότυπα ή τουλάχιστον να θυσιάζονται όσο το δυνατόν πιο λίγα. Ένα τέτοιο μέτρο ενδιαφέροντος είναι τα *Χι Αναδυόμενα Πρότυπα* (*Chi Emerging Patterns*) [Fan, 2004], που ορίζονται παρακάτω.

Δοθέντος ενός συνόλου δεδομένων-υπόβαθρο D_1 και ενός συνόλου δεδομένων-στόχος D_2, ένα σύνολο αντικειμένων X λέγεται χι αναδυόμενο πρότυπο, αν πληρούνται όλες οι παρακάτω προϋποθέσεις:

1. $supp_2(X) \geq \sigma$, όπου σ είναι το κατώφλι της ελάχιστης υποστήριξης.
2. $gr_{1\to2}(X) \geq \rho$, όπου ρ είναι το κατώφλι του ελάχιστου ρυθμού ανάπτυξης.
3. $\forall Y \subset X, gr_{1\to2}(Y) < gr_{1\to2}(X)$
4. $|X| = 1 \vee |X| > 1 \wedge (\forall Y \subset X \wedge |Y| = |X| - 1 \wedge chi(X,Y) \geq \eta)$, όπου η = 3,84 είναι ένα κατώφλι ελάχιστης τιμής της συνάρτησης $chi(X, Y)$, η οποία υπολογίζεται βάσει της δοκιμασίας X^2.

Η πρώτη συνθήκη διασφαλίζει ότι τα πρότυπα που έχουν ανακαλυφθεί έχουν την ελάχιστη επιθυμητή υποστήριξη στο σύνολο δεδομένων εκπαίδευσης, έτσι ώστε να μπορούν να γενικευθούν και σε νέα δεδομένα. Η δεύτερη συνθήκη διασφαλίζει ότι τα πρότυπα που ανακαλύφθηκαν έχουν μια επαρκή ικανότητα διάκρισης μεταξύ των δύο συνόλων δεδομένων (D_1 και D_2). Η τρίτη συνθήκη χρησιμοποιείται για να απομακρύνει εκείνα τα αναδυόμενα πρότυπα που έχουν ένα υποσύνολο με ίση ή μεγαλύτερη υποστήριξη και ίσο ή μεγαλύτερο ρυθμό ανάπτυξης (οποιοδήποτε σύνολο έχει υποστήριξη μεγαλύτερη ή ίση με τα υπερσύνολά του, συνεπώς το μόνο που ελέγχεται στη συνθήκη είναι ο ρυθμός ανάπτυξης). Από τη στιγμή που ισχύει η συνθήκη 3 και λόγω του ότι το υποσύνολο Y έχει μικρότερο αριθμό αντικειμένων από το X και άρα είναι περισσότερο γενικό από το X, δεν υπάρχει λόγος να διατηρηθεί το X. Τέλος, η τέταρτη συνθήκη διασφαλίζει ότι ένα αναδυόμενο πρότυπο έχει στατιστικά σημαντική (95%) διαφορά

κατανομής υποστήριξης μεταξύ συνόλου δεδομένων-στόχος και συνόλου δεδομένων-υπόβαθρο σε σχέση με τη διαφορά κατανομής υποστήριξης των μικρότερων κατά ένα αντικείμενο υποσυνόλων του.

Για τους λόγους που περιγράφηκαν παραπάνω επιλέχθηκε η χρήση των χι αναδυόμενων προτύπων στη μέθοδο PolyA-iEP.

7.3 Σχετικές Εργασίες

Μια από τις πρώτες προσεγγίσεις στο πρόβλημα της πολυαδενυλίωσης ήταν η εργασία των Salamov & Solovyev [1997] που ανέπτυξαν ένα πρόγραμμα με το όνομα POLYAH για την αναγνώριση των σημείων πολυαδενυλίωσης των 3΄ άκρων ανθρώπινων πρόδρομων αλληλουχιών mRNA. Ο αλγόριθμος βασιζόταν σε μια γραμμική συνάρτηση διάκρισης (Linear Discriminant Function LDF), η οποία εκπαιδεύτηκε για να ξεχωρίζει τα σημεία πολυαδενυλίωσης από άλλες περιοχές του ανθρώπινου γονιδιώματος που περιείχαν την ακολουθία AAUAAA. Η ακρίβεια της μεθόδου εκτιμήθηκε σε ένα σύνολο από 131 περιοχές πολυαδενυλίωσης και 1.466 περιοχές ανθρώπινου γονιδιώματος που περιείχαν την ακολουθία AAUAAA. Η ευαισθησία της μεθόδου είναι 86% και η ειδικότητα 51%.

Το 1999 οι Tabaska & Zhang ανέπτυξαν το polyadq, ένα πρόγραμμα για πρόβλεψη σημάτων πολυαδενυλίωσης σε ανθρώπινο γονιδίωμα. Το πρόγραμμα ανακαλύπτει τα σήματα χρησιμοποιώντας δύο γραμμικές συναρτήσεις διάκρισης. Μια συνάρτηση εξειδικευμένη για τα σήματα που σχετίζονται με την ακολουθία AAUAAA και μια άλλη για τα σήματα που σχετίζονται με την ακολουθία AUUAAA. Το polyadq προβλέπει σήματα πολυαδενυλίωσης με συντελεστή συσχέτισης (correlation coefficient) 0,413 σε ολόκληρα γονίδια και 0,512 στα δύο τελευταία εξόνια των γονιδίων.

Οι Van Helden et al. [2000], προσέγγισαν το πρόβλημα της εύρεσης του σημείου πολυαδενυλίωσης με στατιστικές μεθόδους. Άλλες ενδιαφέρουσες προσεγγίσεις παρουσιάστηκαν από τους Graber et al. [2002] και Hajarnavis et al. [2004] που χρησιμοποίησαν Κρυφά Μοντέλα Markov (Hidden Markov Models).

Οι Liu et al. [2003] πρότειναν μια μέθοδο μηχανικής μάθησης για την πρόβλεψη σημάτων πολυαδενυλίωσης σε αλληλουχίες ανθρώπινου RNA, αναλύοντας χαρακτηριστικά γύρω από αυτά. Η μέθοδος αποτελείτο από 3 στάδια: (1) Παραγωγή υποψήφιων χαρακτηριστικών από τις αρχικές αλληλουχίες. (2) Επιλογή των σχετικών χαρακτηριστικών χρησιμοποιώντας αλγόριθμο βασισμένο στην εντροπία. (3) Χρησιμοποίηση των χαρακτηριστικών που προέκυψαν για την κατασκευή μιας μηχανής διανυσμάτων υποστήριξης (SVM) για την πρόβλεψη των σημείων πολυαδενυλίωσης.

Οι Hu et al. [2005], ανέπτυξαν ένα πρόγραμμα με την ονομασία PROBE (Polyadenylation-Related Oligonucleotide Bidimensional Enrichment) με σκοπό να εντοπίσουν ρυθμιστικά στοιχεία που βρίσκονταν στην ίδια αλληλουχία με το σημείο πολυαδενυλίωσης (cis-elements). Αποτέλεσμα της εργασίας τους ήταν η ανακάλυψη πολλών νέων ρυθμιστικών στοιχείων κυρίως πλούσιων σε γουανίνη αλλά και η διαπί-

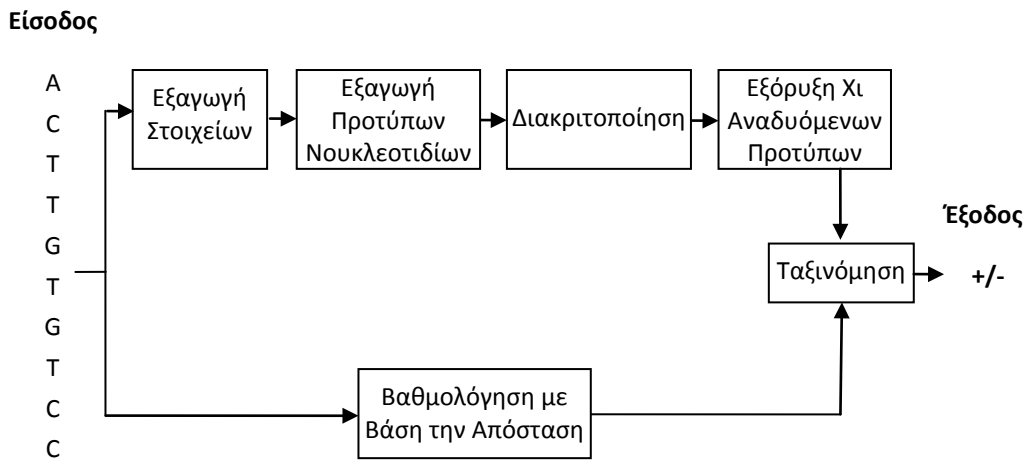
στωση ότι πολλά ρυθμιστικά στοιχεία σε φυτικά κύτταρα υπήρχαν ως ρυθμιστικά στοιχεία και σε ασθενή σημεία πολυαδενυλίωσης. Ασθενές σημείο πολυαδενυλίωσης στη διαδικασία της εναλλακτικής πολυαδενυλίωσης είναι το σημείο στο οποίο δεν γίνεται συχνά αποκοπή σε σχέση με το ισχυρό. Έτσι, γίνεται η υπόθεση ότι κάποια ρυθμιστικά στοιχεία υπάρχουν σε όλα τα ευκαρυωτικά κύτταρα και διατηρήθηκαν κατά την εξελικτική διαδικασία.

Ένα χρόνο αργότερα οι Cheng et al. [2006], μέλη του ίδιου εργαστηρίου, ερεύνησαν αν αυτά τα ρυθμιστικά στοιχεία μπορούν να χρησιμοποιηθούν για να προβλεφθούν με επιτυχία σημεία πολυαδενυλίωσης. Ανέπτυξαν μια μηχανή διανυσμάτων υποστήριξης, χρησιμοποιώντας τα ρυθμιστικά στοιχεία της προηγούμενης μελέτης. Το πρόγραμμα που αναπτύχθηκε ονομάστηκε polya_svm και πέτυχε βελτιωμένη απόδοση σε σχέση με το polyadq.

Πιο πρόσφατα, οι Ji et al. [2007], οι οποίοι χρησιμοποίησαν συμπεράσματα της προγενέστερης εργασίας των Loke et al. [2005], ανέπτυξαν ένα πρόγραμμα βασισμένο σε Γενικευμένα Κρυφά Μοντέλα Markov (Generalized Hidden Markov Models) για την πρόβλεψη σημείων πολυαδενυλίωσης σε βιολογικές αλληλουχίες του φυτού Arabidopsis thaliana.

7.4 Η Μέθοδος PolyA-iEP

Η μέθοδος PolyA-iEP αναπτύχθηκε στα πλαίσια της διατριβής με σκοπό την αποτελεσματική πρόβλεψη σημείων πολυαδενυλίωσης στα 3΄ άκρα αλληλουχιών mRNA. Μολονότι η παρούσα μελέτη επικεντρώνεται σε δεδομένα του οργανισμού Arabidopsis thaliana, η PolyA-iEP είναι μια πλήρως παραμετροποιήσιμη μέθοδος και μπορεί να εφαρμοστεί και σε δεδομένα άλλων οργανισμών, όπως για παράδειγμα του ανθρώπου. Η υλοποίηση του έγινε στην γλώσσα προγραμματισμού JAVA και αποτελείται από έναν αριθμό σταδίων, τα οποία παρουσιάζονται στην Εικόνα 7.2 και αναλύονται εκτενώς στις παραγράφους που ακολουθούν.



Εικόνα 7.2: Η αρχιτεκτονική της μεθόδου PolyA-iEP.

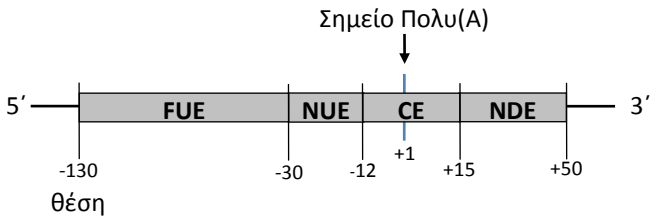
7.4.1 Εξαγωγή Στοιχείων

Η περιοχή γύρω από ένα σημείο πολυαδενυλίωσης χωρίζεται σε έναν αριθμό από στοιχεία (elements). Αυτά τα στοιχεία αποτελούν αλληλουχίες νουκλεοτιδίων και διαφέρουν μεταξύ τους ως προς το μήκος τους και τη συχνότητα εμφάνισης των νουκλεοτιδίων που περιέχουν. Ως αποτέλεσμα των παραπάνω διαφορών, κάθε στοιχείο περιλαμβάνει διαφορετικά πρότυπα αλληλουχιών. Αυτό καταδεικνύει την ανάγκη αναζήτησης προτύπων σε καθένα από τα στοιχεία ξεχωριστά. Επιπλέον, η αναζήτηση συσχετίσεων ανάμεσα στα πρότυπα των διαφορετικών στοιχείων φαίνεται να παρουσιάζει ενδιαφέρον. Αυτή η αναζήτηση οδηγεί σε νέα εκτεταμένα πρότυπα που αποτελούν συνδυασμούς των προτύπων διαφόρων στοιχείων. Αυτή η νέα μορφή προτύπων, είναι πιθανό να περιέχει περισσότερη πληροφορία και επομένως τα πρότυπα αυτά χαρακτηρίζονται από μεγαλύτερη ικανότητα διάκρισης σε σχέση με τα απλά πρότυπα κάθε στοιχείου. Η ανακάλυψη των εκτεταμένων προτύπων και η χρήση τους για αποτελεσματικότερη ταξινόμηση είναι ένα καινοτόμο χαρακτηριστικό στοιχείο της μεθόδου PolyA-iEP.

Τα βασικά στοιχεία που έχουν προταθεί στη βιβλιογραφία [Loke et al., 2005] είναι τρία και παίρνουν το όνομά τους από τη θέση τους σε σχέση με το σημείο αποκοπής και πολυαδενυλίωσης. Στη μέθοδο PolyA-iEP περιλαμβάνεται και ένα τέταρτο στοιχείο. Τα τέσσερα στοιχεία είναι τα εξής:

- *Άπω Ανοδικό Στοιχείο* (*Far Upstream Element – FUE*).
- *Εγγύς Ανοδικό Στοιχείο* (*Near Upstream Element – NUE*).
- *Στοιχείο Αποκοπής* (*Cleavage Element – CE*).
- *Εγγύς Καθοδικό Στοιχείο* (*Near Downstream Element – NDE*).

Το εγγύς καθοδικό στοιχείο δεν είναι βαρύνουσας σημασίας, αλλά για λόγους πληρότητας έχει συμπεριληφθεί στη μέθοδο PolyA-iEP. Η Εικόνα 7.3 συνοψίζει σχηματικά τα τέσσερα στοιχεία γύρω από το σημείο πολυαδενυλίωσης.



Εικόνα 7.3: Τα στοιχεία γύρω από το σημείο πολυαδενυλίωσης.

7.4.2 Εξαγωγή των Προτύπων Νουκλεοτιδίων

Καθένα από τα στοιχεία τα οποία περιγράφηκαν παραπάνω αναπαρίστανται από ένα διάνυσμα που περιέχει τις συχνότητες όλων των νουκλεοτιδικών προτύπων. Τα πρότυπα περιλαμβάνουν όλα τα δυνατά πρότυπα νουκλεοτιδίων με μέγεθος 1 έως *k*. Το k ορίζεται από τον χρήστη. Επιπλέον, η μέθοδος PolyA-iEP περιλαμβάνει την παραγωγή

ενός αριθμού προτύπων που περιέχουν χαρακτήρες διττής σημασίας ως επέκταση των αρχικών προτύπων. Τα νέα αυτά πρότυπα είναι πιο γενικά καθώς περιλαμβάνουν την πληροφορία δύο απλών προτύπων. Το μήκος τους επίσης μπορεί να κυμαίνεται από 1 έως *k*. Το αλφάβητο για την παραγωγή αυτών των προτύπων παρουσιάζει ο Πίνακας 7.1. Στις τελευταίες 6 γραμμές του πίνακα παρουσιάζονται όλοι οι χαρακτήρες διττής σημασίας που χρησιμοποιούνται στη μέθοδο. Για παράδειγμα, το πρότυπο «AWT» είναι ένα έγκυρο πρότυπο και υποδηλώνει είτε την ύπαρξη του «AAT» είτε του «ATT». Συνοψίζοντας τα παραπάνω, κάθε αλληλουχία αναπαρίσταται στη μέθοδο PolyA-iEP από έναν αριθμό διανυσμάτων. Κάθε στοιχείο ενός διανύσματος αντιστοιχεί στην συχνότητα ενός έγκυρου προτύπου. Ο αριθμός των διανυσμάτων για κάθε αλληλουχία είναι ίσος με τον αριθμό των στοιχείων (Εδάφιο 7.4.1) γύρω από το σημείο πολυαδενυλίωσης, που στην προκειμένη περίπτωση είναι τέσσερα.

Πίνακας 7.1: Το αλφάβητο των προτύπων.

Γράμμα Αλφαβήτου	Νουκλεοτίδια	
A	A	**A**denine
C	C	**C**ytocine
G	G	**G**uanine
T	T	**T**hymine
R	A ή G	pu**R**ine
Y	C ή T	p**Y**rimidine
M	A ή C	a**M**ino
K	G ή T	**K**eto
S	C ή G	**S**trong (3 δεσμοί H)
W	A ή T	**W**eak (2 δεσμοί H)

7.4.3 Διακριτοποίηση

Στη μέθοδο PolyA-iEP πραγματοποιείται δυαδική διακριτοποίηση (παράγονται δύο διακριτές τιμές), με σκοπό την παραγωγή δεδομένων σε μορφή συναλλαγών. Η μέθοδος διακριτοποίησης που χρησιμοποιείται βασίζεται στην εντροπία και το κέρδος πληροφορίας και υλοποιείται βάσει των Εξισώσεων 2.7 και 2.8 (Εδάφιο 2.4.2). Τα διανύσματα που δημιουργήθηκαν στο προηγούμενο στάδιο και είχαν ως στοιχεία τους τις συχνότητες των προτύπων, σε αυτό το στάδιο αντικαθίστανται από συναλλαγές. Οι συναλλαγές αυτές περιλαμβάνουν τα αναγνωριστικά (ID) των προτύπων που εμφανίζονται με αρκετά μεγάλη συχνότητα στις αντίστοιχες αλληλουχίες, ώστε μετά τη διακριτοποίηση λαμβάνουν την τιμή 1 και όχι την τιμή 0.

7.4.4 Εξόρυξη Χι Αναδυόμενων Προτύπων

Τα συναλλακτικά δεδομένα που παράγονται στο προηγούμενο στάδιο χρησιμοποιούνται για την ανακάλυψη χι αναδυόμενων προτύπων. Για το σκοπό αυτό επεκτάθηκε και προσαρμόστηκε ο FP-Growth [Han et al., 2000], ένας αλγόριθμος που χρησιμοποιείται για την εξόρυξη συχνών συνόλων αντικειμένων (Εδάφιο 2.3.2). Ο τροποποιημένος αλγόριθμος λαμβάνει ως είσοδο δύο σύνολα δεδομένων, το σύνολο δεδομένων-υπόβαθρο και το σύνολο δεδομένων-στόχο και ανακαλύπτει όλα τα χι αναδυόμενα πρότυπα βάσει των παραμέτρων που δίνονται από το χρήστη (ελάχιστα όρια υποστήριξης και ρυθμού ανάπτυξης). Αξίζει να αναφερθεί ότι τα πρότυπα που προκύπτουν από το PolyA-iEP είναι εκτεταμένα, καθώς προκύπτουν από συνδυασμούς προτύπων διαφορετικών στοιχείων και όχι από ένα στοιχείο μόνο. Σε αυτό το στάδιο παράγονται δύο σύνολα χι αναδυόμενων προτύπων, E_+ και E_-, για τα σύνολα θετικών και αρνητικών παραδειγμάτων αντίστοιχα.

Στα πλαίσια της παρούσας μελέτης το σύνολο των αρνητικών παραδειγμάτων αποτελείται από τρία διαφορετικά είδη αλληλουχιών που δεν περιέχουν σημεία πολυαδενυλίωσης (αλληλουχίες 5΄ UTR, κωδικές αλληλουχίες και ιντρόνια). Οι αλληλουχίες αυτές παρουσιάζουν τελείως διαφορετική κατανομή νουκλεοτιδίων και αν και οι τρεις κατηγορίες αρνητικών αλληλουχιών αντιμετωπιστούν ως μία, τότε υπάρχει απώλεια πληροφορίας και συνεπώς μειωμένη αποτελεσματικότητα ταξινόμησης. Για το λόγο αυτό, στη συγκεκριμένη μελέτη έγινε εξόρυξη τεσσάρων ζευγών από σύνολα χι αναδυόμενων προτύπων (E_+/E_-). Το ένα ζεύγος σχετίζεται με το διαχωρισμό των θετικών αλληλουχιών από όλες τις αρνητικές και τα υπόλοιπα τρία σχετίζονται με στο διαχωρισμό των θετικών αλληλουχιών από κάθε κατηγορία αρνητικών αλληλουχιών ξεχωριστά.

Ένα παράδειγμα εκτεταμένου χι αναδυόμενου προτύπου που μπορεί να εξορυχθεί με τη μέθοδο PolyA-iEP είναι το ακόλουθο:

{FUE_TGGA, NUE_CT, NDE_CYG} : 0,32

Το παραπάνω εκτεταμένο πρότυπο σχετίζει την παρουσία του προτύπου TGGA στο στοιχείο FUE του προτύπου CT στο στοιχείο NUE και του προτύπου CYG στο στοιχείο NDE. Το πρότυπο CYG περιέχει τον διττό χαρακτήρα Y που αντιπροσωπεύει την παρουσία πυριμιδίνης (C ή T) και συνεπώς αντιπροσωπεύει την παρουσία του προτύπου CCT ή του προτύπου CTG. Η ισχύς (strength) του εκτεταμένου αυτού προτύπου είναι 0,32.

7.4.5 Βαθμολόγηση με Βάση τα Χι Αναδυόμενα Πρότυπα

Όπως έχει ήδη περιγραφεί τα αναδυόμενα πρότυπα μπορούν να χρησιμοποιηθούν για ταξινόμηση (Εδάφιο 2.3.2). Σε αυτό το στάδιο της μεθόδου PolyA-iEP, τα ζεύγη E_+/E_- των αναδυόμενων προτύπων που παρήχθησαν προηγουμένως χρησιμοποιούνται για να βαθμολογήσουν ένα στιγμιότυπο του συνόλου δεδομένων, δηλαδή μια αλληλουχία, προσδιορίζοντας κατά πόσο είναι θετικό ή αρνητικό. Για το σκοπό αυτό υπολογίζονται ζεύγη βαθμολογιών σύμφωνα με τις Εξισώσεις 2.5 και 2.6 που έχουν περιγραφεί στο

Εδάφιο 2.3.2. Η πρώτη βαθμολογία δείχνει κατά πόσο η αλληλουχία είναι θετική και η δεύτερη κατά πόσο είναι αρνητική.

Ο συνολικός αριθμός των βαθμολογιών που παράγονται σε αυτό το στάδιο είναι οκτώ. Δύο βαθμολογίες, μια για την θετική και μια για την αρνητική τάξη, παράγονται για κάθε μια από τις παρακάτω περιπτώσεις ταξινόμησης αλληλουχιών.

- Θετικές αλληλουχίες έναντι όλων των αρνητικών αλληλουχιών.
- Θετικές αλληλουχίες έναντι αλληλουχιών 5΄ UTR.
- Θετικές αλληλουχίες έναντι κωδικών αλληλουχιών.
- Θετικές αλληλουχίες έναντι ιντρονίων.

7.4.6 Βαθμολόγηση με Βάση την Απόσταση

Η βαθμολόγηση με βάση την απόσταση είναι ένα στάδιο που αποτελεί καινοτομία της μεθόδου PolyA-iEP σε σύγκριση με τις μεθόδους που παρουσιάζονται στη βιβλιογραφία και είναι ανεξάρτητη από τα στάδια που παρουσιάστηκαν προηγουμένως. Σε αυτό το στάδιο υπολογίζονται οι συχνότητες του κάθε νουκλεοτιδίου σε κάθε θέση της αλληλουχίας και δημιουργούνται δύο πίνακες για κάθε τάξη. Για παράδειγμα, το νουκλεοτίδιο A έχει συχνότητα 0,15 στην θέση 1 των αλληλουχιών (Πίνακας 7.2) και η συχνότητα αυτή είναι τέταρτη σε σειρά σε σχέση με τις συχνότητες των άλλων νουκλεοτιδίων στην θέση αυτή (Πίνακας 7.3). Αντίστοιχα, το νουκλεοτίδιο G έχει συχνότητα 0,40 και είναι πρώτο σε σχέση με τα νουκλεοτίδια στη θέση αυτή. Οι πίνακες που παράγονται περιέχουν αυτά τα στοιχεία. Ο πρώτος, τις συχνότητες του κάθε νουκλεοτιδίου για κάθε θέση της αλληλουχίας και ο δεύτερος την κατάταξη βάσει της συχνότητας που έχει το νουκλεοτίδιο αυτό σε σχέση με τα άλλα νουκλεοτίδια για κάθε θέση της αλληλουχίας.

Για να υπολογιστεί η απόσταση μιας αλληλουχίας από μια τάξη ή υποτάξη (αλληλουχίες 5΄ UTR, κωδικές αλληλουχίες ή ιντρόνια) αρχικά η αλληλουχία μετατρέπεται σε ένα διάνυσμα που περιέχει τις θέσεις των νουκλεοτιδίων σύμφωνα με το δεύτερο πίνακα για μια τάξη ή υποτάξη. Έπειτα, υπολογίζεται η απόσταση από το μοναδιαίο διάνυσμα, το οποίο ουσιαστικά αντιπροσωπεύει την αλληλουχία. Στο τέλος, η υπολογισμένη απόσταση διαιρείται με το μήκος της αλληλουχίας. Για παράδειγμα, δεδομένου του δεύτερου πίνακα (Πίνακας 7.3) το διάνυσμα που υπολογίζεται για την αλληλουχία ATGGC είναι <4, 1, 2,5, 1, 2>. Η απόσταση Μανχάταν του διανύσματος αυτού από το μοναδιαίο διάνυσμα <1, 1, 1, 1, 1> είναι 5,5. Διαιρώντας την απόσταση αυτή με το μήκος του διανύσματος, προκύπτει η βαθμολογία 1,1. Αυτή είναι η μέση απόσταση των νουκλεοτιδίων της αλληλουχίας από την τάξη ή υποτάξη στην οποία ανήκει ο πίνακας από τον οποίο ελήφθησαν οι τιμές. Όσο μικρότερη είναι η υπολογιζόμενη απόσταση, τόσο μεγαλύτερη είναι η πιθανότητα να ανήκει η συγκεκριμένη αλληλουχία στην τάξη που αντιπροσωπεύεται από τον πίνακα βάσει του οποίου έγιναν οι υπολογισμοί.

Πίνακας 7.2: Παράδειγμα πίνακα με τις συχνότητες των νουκλεοτιδίων.

Νουκλεοτίδιο	Θέση στην Αλληλουχία				
	1	2	3	4	5
A	0,15	0,08	0,28	0,10	0,10
C	0,20	0,22	0,22	0,30	0,30
G	0,40	0,33	0,25	0,40	0,30
T	0,25	0,37	0,25	0,20	0,30

Πίνακας 7.3: Παράδειγμα πίνακα με τις κατατάξεις των νουκλεοτιδίων.

Νουκλεοτίδιο	Θέση στην Αλληλουχία				
	1	2	3	4	5
A	4	4	1	4	4
C	3	3	4	2	2
G	1	2	2,5	1	2
T	2	1	2,5	3	2

7.4.7 Ταξινόμηση

Οι βαθμολογίες που προκύπτουν στα προηγούμενα στάδια χρησιμοποιούνται ως είσοδος σε έναν ταξινομητή ο οποίος λαμβάνει την τελική απόφαση για την ταξινόμηση μιας αλληλουχίας ως θετικής ή αρνητικής, δηλαδή περιέχουσας ή μη περιέχουσας σημείο αποκοπής και πολυαδενυλίωσης, αντίστοιχα. Πιο συγκεκριμένα, ένα σύνολο από 13 βαθμολογίες δίνονται ως είσοδος στον ταξινομητή. Οι οκτώ από τις βαθμολογίες αυτές προέρχονται από τα βήματα της βαθμολόγησης με βάση τα χι αναδυόμενα πρότυπα και οι 5 από αυτές προέρχονται από το βήμα της βαθμολόγησης με βάση την απόσταση. Οποιοσδήποτε αλγόριθμος ταξινόμησης μπορεί να χειριστεί πραγματικές τιμές αριθμητικών δεδομένων μπορεί να χρησιμοποιηθεί για τη δημιουργία του ταξινομητή.

Ο Πίνακας 7.4 παρουσιάζει τον αλγόριθμο βάσει του οποίου υλοποιείται η μέθοδος PolyA-iEP, συνοψίζοντας όλα τα στάδια που περιλαμβάνει η μέθοδος, τα οποία περιγράφηκαν στις προηγούμενες παραγράφους. Πρέπει να σημειωθεί ότι χρησιμοποιείται ένα σύνολο των δεδομένων για την εξαγωγή των χι αναδυόμενων προτύπων και τη δημιουργία των πινάκων συχνοτήτων και κατάταξης συχνοτήτων και ένα διαφορετικό σύνολο των δεδομένων για την εκπαίδευση του ταξινομητή που συνδυάζει τις βαθμολογίες που προέρχονται από τις δύο διαφορετικές μεθόδους βαθμολόγησης των αλληλουχιών.

Πίνακας 7.4: Ο αλγόριθμος που υλοποιεί τη μέθοδο PolyA-iEP.

Είσοδος: Ένα σύνολο *S* αλληλουχιών cDNA με γνωστό σημείο πολυαδενυλίωσης (*PAS_Position*), ένα σύνολο *S΄* αλληλουχιών cDNA με γνωστό σημείο σημείο πολυαδενυλίωσης (*PAS_Position*) για την εκπαίδευση του ταξινομητή, ένας πίνακας με τις θέσεις έναρξης των διαφορετικών στοιχείων (*ElementsStartPositions*), ένας πίνακας με τις θέσεις τερματισμού των διαφορετικών στοιχείων (*ElementsEndPositions*), το ελάχιστο (*minLen*) και το μέγιστο (*maxLen*) μήκος, καθώς και το αλφάβητο (*alphabet*) των νουκλεοτιδικών προτύπων και μια μέθοδος ταξινόμησης (*ClassificationMethod1*) για τη δημιουργία του ταξινομητή.

Έξοδος: Οι παράμετροι και οι ταξινομητές που αποτελούν το σύστημα πρόβλεψης.

```
for each (<s, PAS_Position> ∈ S)
  Elements ← extractElements(s, PAS_Position,
              ElementsStartPositions, ElementsEndPositions)
  for each (e ∈ Elements)
    NuclPatFreq[e] ← NuclPatFreq[e] ∪ countNucleotidePetterns(
                                        minLen,maxLen,alphabet)
discretize(NuclPatFreq)
Transactions ← convertToTransactions(NuclPatFreq)
<AllEP+,AllEP-,5UTREP+,5UTREP-,CodingEP+,CodingEP-,IntronEP+,IntronEP->
  ← mineChiEmergingPatterns(min_sup, min_gr)
<Positives, AllNegatives, 5UTRs, Coding, Introns>
  ← calculateFrequencyRankingMatrix(S)
for each (<s, PAS_Position> ∈ S´)
  EP_Scores ← getEPScores(s, AllEP+, AllEP-, 5UTREP+, 5UTREP-,
                          CodingEP+, CodingEP-,IntronEP+,IntronEP-)
  Dist_Scores ← getDistScores(s, Positives, AllNegatives, 5UTRs,
                              Coding, Introns)
  examples1 ← examples1 ∪ <EP_Scores, Dist_Scores>
C ← buildClassifier(examples1, ClassificationMethod)
return <AllEP+, AllEP-, 5UTREP+, 5UTREP-, CodingEP+, CodingEP-,
        IntronEP+, IntronEP-, Positives, AllNegatives, 5UTRs, Coding,
        Introns, C>
```

7.5 Πειραματική Αξιολόγηση

Στην ενότητα αυτή παρουσιάζονται τα πειράματα που διεξήχθησαν με σκοπό την αξιολόγηση της μεθόδου PolyA-iEP. Στα πλαίσια της διεξαγωγής των πειραμάτων χρησιμοποιήθηκαν κάποια σύνολα δεδομένων τα οποία, επίσης, παρουσιάζονται στην ενότητα αυτή.

7.5.1 Δεδομένα

Χρησιμοποιήθηκαν τέσσερα σύνολα αλληλουχιών mRNA του οργανισμού Arabidopsis thaliana. Το ένα σύνολο περιέχει 6.209 θετικά παραδείγματα, τα οποία είναι αλληλουχίες mRNA που περιέχουν σημείο πολυαδενυλίωσης, ενώ τα άλλα τρία σύνολα περιέχουν αρνητικά παραδείγματα (864 αλληλουχίες 5΄ UTR, 1.501 κωδικές αλληλουχίες και 1.581 ιντρόνια, αντίστοιχα). Τα δεδομένα αυτά έχουν χρησιμοποιηθεί και σε προηγού-

μενες εργασίες [Ji et al., 2007; Koh & Wong, 2007]. Όλες οι αλληλουχίες έχουν μήκος 400 νουκλεοτιδίων. Οι θετικές αλληλουχίες έχουν ένα σημείο πολυαδενυλίωσης στη θέση 301 και έχουν υποστεί διαδικασία στοίχισης μεταξύ τους με σκοπό την απόρριψη εκείνων που έχουν μεγάλη ομοιότητα με τις υπόλοιπες [Koh & Wong., 2007]. Πιο συγκεκριμένα, στο επιλεγμένο σύνολο των 6.209 θετικών αλληλουχιών δυο οποιεσδήποτε αλληλουχίες έχουν το πολύ 70% ομοιότητα μεταξύ τους.

Ο Πίνακας 7.5 περιγράφει τα δεδομένα που χρησιμοποιήθηκαν. Για λόγους που εξυπηρετούν την αξιοπιστία και την αποτελεσματικότητα της πειραματικής διαδικασίας τα αρχικά σύνολα δεδομένων διαιρέθηκαν σε υποσύνολα βάσει δύο κριτηρίων. Το πρώτο κριτήριο που έχει αναφερθεί και προηγουμένως σχετίζεται με τα βιολογικά χαρακτηριστικά της κάθε αρνητικής αλληλουχίας που μπορεί να είναι αλληλουχία 5΄ UTR, κωδική αλληλουχία ή ιντρόνιο. Ο διαχωρισμός των αλληλουχιών βάσει αυτού του κριτηρίου προσδιορίζεται από τις γραμμές που περιέχει ο Πίνακας 7.5.

Το δεύτερο κριτήριο σχετίζεται με τη διαδικασία που πρέπει να ακολουθηθεί για τη δημιουργία και αξιολόγηση ενός μοντέλου. Για το λόγο αυτό οι αλληλουχίες διαχωρίστηκαν με τυχαίο τρόπο σε τρία υποσύνολα. Το πρώτο χρησιμοποιείται για την εξόρυξη χι αναδυόμενων προτύπων, το δεύτερο για την εκπαίδευση του ταξινομητή και το τρίτο για την αξιολόγηση του ταξινομητή. Ο αριθμός των αλληλουχιών που περιέχονται σε καθένα από τα τρία υποσύνολα υπολογίστηκε λαμβάνοντας υπόψη ότι πρέπει να υπάρχουν επαρκή δεδομένα για την εξαγωγή των αναδυόμενων προτύπων και την εκπαίδευση του ταξινομητή. Ο διαχωρισμός των αλληλουχιών βάσει του δεύτερου κριτηρίου προσδιορίζεται από τις στήλες που περιέχει ο Πίνακας 7.5.

Πίνακας 7.5: Σύνολα Αλληλουχιών.

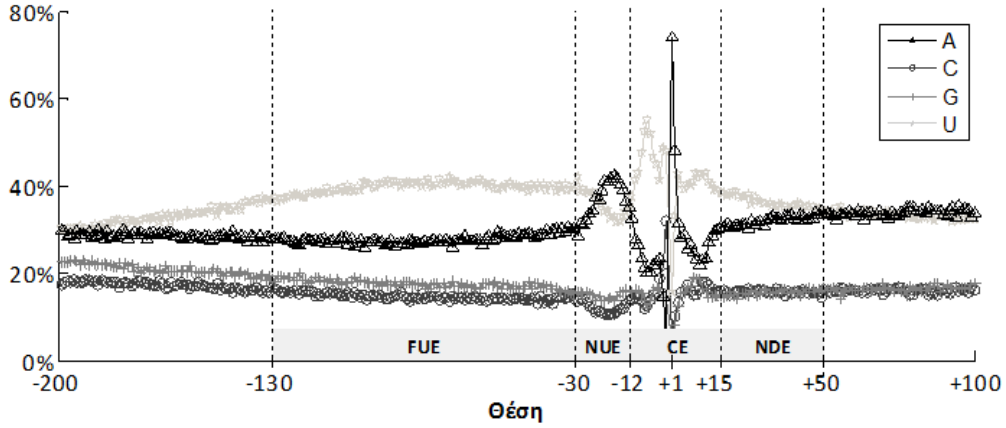
	Εξόρυξη	**Εκπαίδευση**	**Αξιολόγηση**	**Σύνολο**
Θετικές Αλληλουχίες	2.794	2.173	1.242	6.209
Αρνητικές Αλληλουχίες	1.777	1.380	789	3.946
5΄ UTR	389	302	173	864
Κωδικές	676	525	300	1.501
Ιντρόνια	712	553	316	1.581
Σύνολο (Θετικές + Αρνητικές)	4.571	3.553	2.031	10.155

7.5.2 Εξερευνητική Ανάλυση

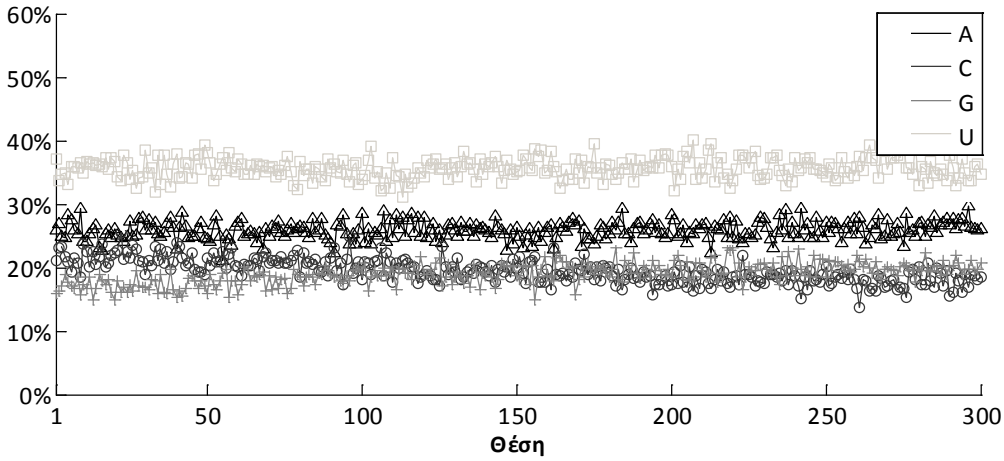
Η αποτελεσματικότητα ταξινόμησης των θετικών αλληλουχιών έναντι των αλληλουχιών κάθε αρνητικής υποτάξης παρουσιάζει σημαντική μεταβλητότητα. Για τη μελέτη της μεταβλητότητας αυτής δημιουργήθηκαν γραφικές παραστάσεις της κατανομής των νουκλεοτιδίων των αλληλουχιών της θετικής τάξης (3΄ άκρο mRNA) και των αλληλουχιών καθεμιάς από τις αρνητικές υποτάξεις (αλληλουχίες 5΄ UTR, κωδικές αλληλουχίες και ιντρόνια). Η Εικόνα 7.4 παρουσιάζει την κατανομή των νουκλεοτιδίων στις αλλη-

λουχίες της θετικής τάξης από τις θέσεις -200 έως +100 σε σχέση με το σημείο πολυαδενυλίωσης, το οποίο τοποθετείται στη θέση +1. Όπως παρατηρείται στην εικόνα, οι διαφορές της κατανομής των νουκλεοτιδίων μεταξύ των διαφορετικών στοιχείων (FUE, NUE, CE NDE) είναι εμφανείς. Η κατανομή των νουκλεοτιδίων στις αλληλουχίες 5΄ UTR (Εικόνα 7.5), στις κωδικές αλληλουχίες (Εικόνα 7.6) και στα ιντρόνια (Εικόνα 7.7) παρουσιάζεται διαφορετική. Όπως αποκαλύπτουν οι γραφικές παραστάσεις η κατανομή νουκλεοτιδίων των θετικών αλληλουχιών παρουσιάζει τη μεγαλύτερη ομοιότητα με την κατανομή νουκλεοτιδίων των ιντρονίων και τη μικρότερη ομοιότητα με την κατανομή νουκλεοτιδίων των κωδικών αλληλουχιών. Αυτές οι παρατηρήσεις είναι σύμφωνες με τη δυσκολία ταξινόμησης των αντίστοιχων αλληλουχιών καθώς και με τα πειραματικά αποτελέσματα που παρουσιάζονται στις επόμενες παραγράφους.

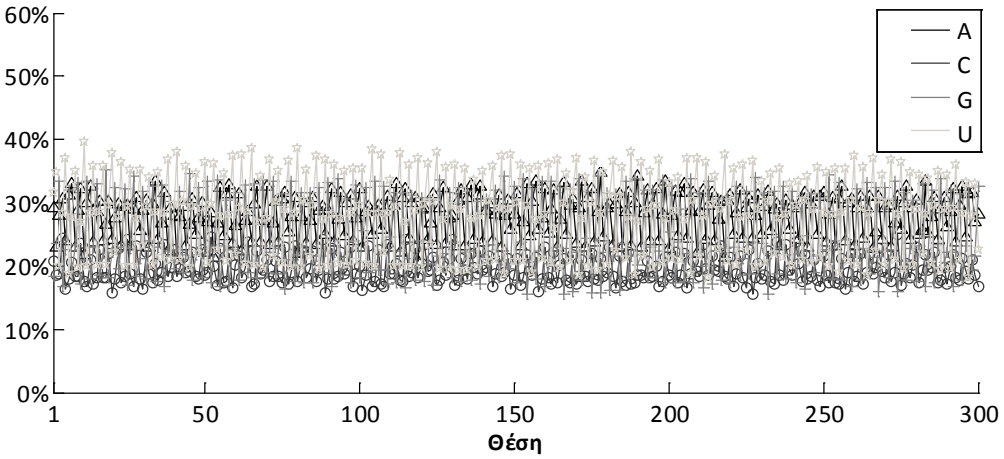
Συγκρίνοντας την κατανομή νουκλεοτιδίων των θετικών αλληλουχιών με την κατανομή νουκλεοτιδίων των ιντρονίων, γίνεται αντιληπτό ότι στα ιντρόνια το συχνότερα εμφανιζόμενο νουκλεοτίδιο είναι η ουρακίλη, ακολουθούμενη από την αδενίνη, τη γουανίνη και την κυτοσίνη με σειρά φθίνουσας συχνότητας εμφάνισης. Η ίδια κατανομή παρουσιάζεται και στην περιοχή ανοδικά του στοιχείου NUE των θετικών αλληλουχιών και μάλιστα είναι περισσότερο όμοια στο στοιχείο FUE. Αυτό το τμήμα της θετικής αλληλουχίας είναι αρκετά σημαντικό για τη σωστή ταξινόμησή της και για το λόγο αυτόν τα ιντρόνια διαχωρίζονται δύσκολα από τις θετικές αλληλουχίες. Αντίθετα, η κατανομή νουκλεοτιδίων των κωδικών αλληλουχιών παρουσιάζει σημαντική διαφορά σε σχέση με την αντίστοιχη των θετικών αλληλουχιών. Τέλος, οι αλληλουχίες 5΄ UTR έχουν παρόμοια κατανομή με αυτή των ιντρονίων, αλλά διαφέρει από την κατανομή των θετικών αλληλουχιών περισσότερο από ότι διαφέρει η κατανομή που παρουσιάζουν τα ιντρόνια. Οπότε, οι αλληλουχίες 5΄ UTR μπορούν να διαχωριστούν περισσότερο αποτελεσματικά από τις θετικές αλληλουχίες, από όσο τα ιντρόνια από τις θετικές αλληλουχίες.



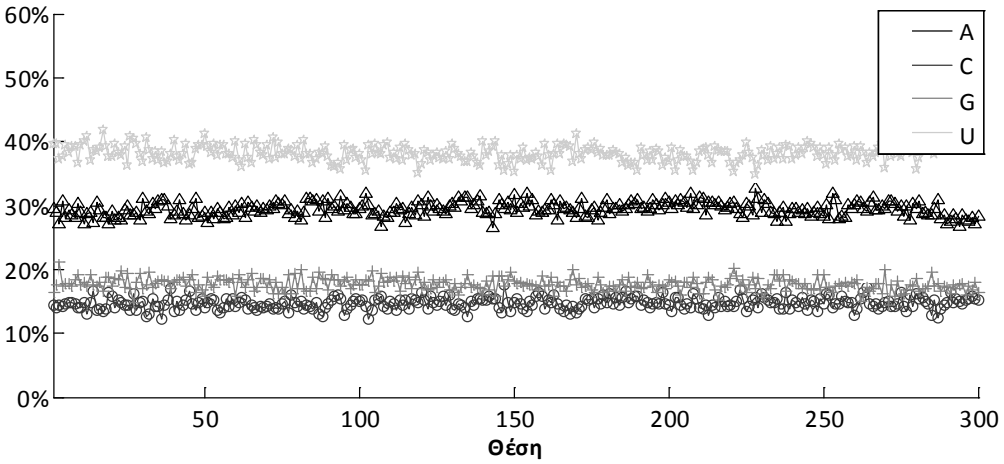
Εικόνα 7.4: Κατανομή νουκλεοτιδίων στις θέσεις -200 έως +100 γύρω από το σημείο πολύ(Α) στο 3΄ άκρο των αλληλουχιών mRNA.



Εικόνα 7.5: Κατανομή νουκλεοτιδίων στις αλληλουχίες 5΄ UTR.



Εικόνα 7.6: Κατανομή νουκλεοτιδίων στις κωδικές αλληλουχίες.



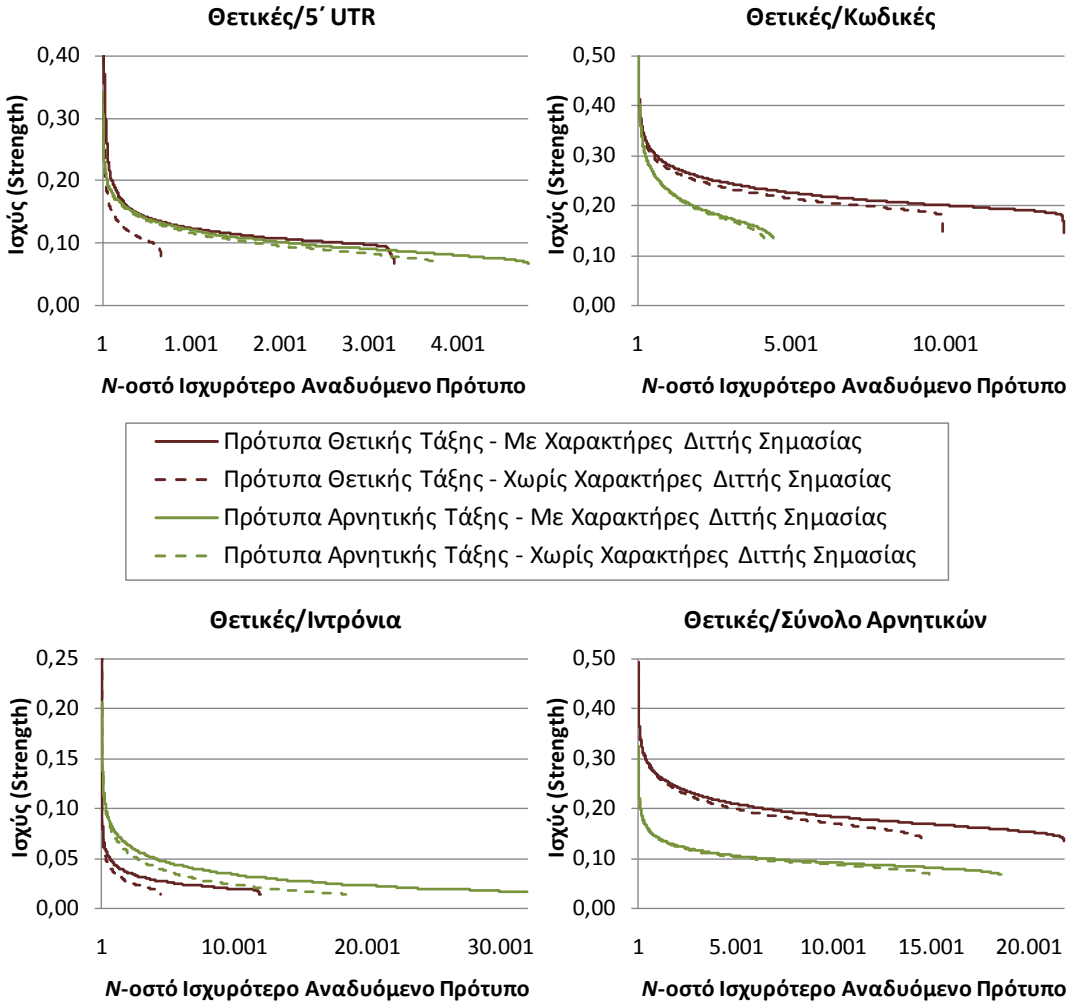
Εικόνα 7.7: Κατανομή νουκλεοτιδίων στα ιντρόνια.

7.5.3 Συνεισφορά των Χαρακτήρων Διττής Σημασίας

Για την αποτίμηση της συνεισφοράς των προτύπων που περιέχουν χαρακτήρες διττής σημασίας, συγκρίθηκαν οι αποδόσεις της μεθόδου PolyA-iEP με τη χρήση και χωρίς τη

χρήση τέτοιων χαρακτήρων. Τα πιο ισχυρά χι αναδυόμενα πρότυπα που εξήχθησαν τόσο από τα θετικά όσο και από τα αρνητικά παραδείγματα και των δύο προσεγγίσεων παρουσιάζονται στην Εικόνα 7.8. Τα πρότυπα είναι ταξινομημένα βάσει της ισχύος τους (strength). Τα γραφήματα παρουσιάζουν τα πρότυπα που εξήχθησαν από τους εξής συνδυασμούς δεδομένων: θετικές αλληλουχίες/αλληλουχίες 5΄ UTR, θετικές αλληλουχίες/κωδικές αλληλουχίες, θετικές αλληλουχίες/ιντρόνια, θετικές αλληλουχίες/όλες οι αρνητικές αλληλουχίες.

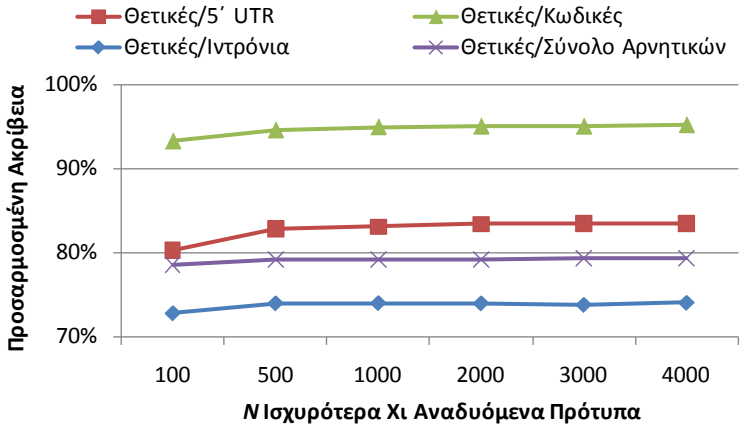
Όπως φαίνεται σε όλα τα γραφήματα, η προσέγγιση που περιλαμβάνει πρότυπα με χαρακτήρες διττής σημασίας, παρέχει ισχυρότερα χι αναδυόμενα πρότυπα, όταν ληφθούν υπόψη τα πρώτα *N* πιο ισχυρά πρότυπα. Συνεπώς, η συνεισφορά των χαρακτήρων διττής σημασίας στην ποιότητα των αναδυόμενων προτύπων είναι σημαντική και η συμπερίληψή τους οδηγεί στην εξόρυξη καλύτερων προτύπων. Μια σημαντική παρατήρηση είναι η χαμηλή ισχύς των προτύπων της ταξινόμησης θετικών αλληλουχιών/ιντρονίων εν συγκρίσει με τις υπόλοιπες ταξινομήσεις. Αυτό καταδεικνύει την εγγενή δυσκολία διαχωρισμού των δύο τάξεων.



Εικόνα 7.8: Συνεισφορά των χαρακτήρων διττής σημασίας στην ισχύ των χι αναδυόμενων προτύπων.

7.5.4 Αξιολόγηση της Βαθμολόγησης με Βάση τα Χι Αναδυόμενα Πρότυπα

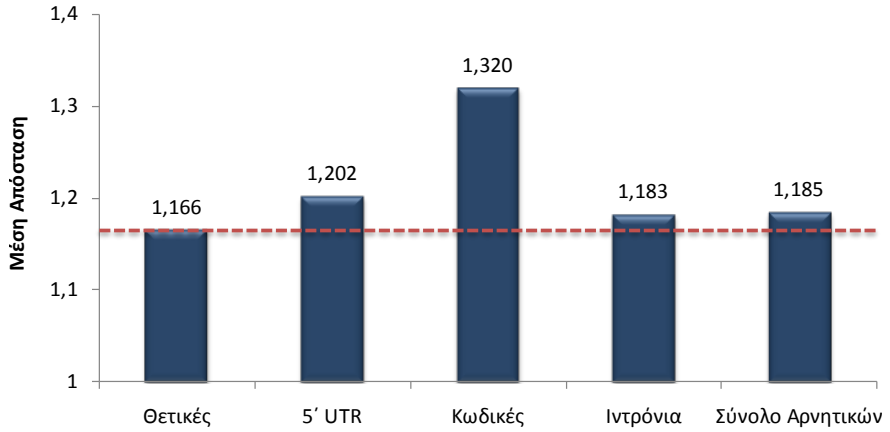
Για να εκτιμηθεί η δυνατότητα των χι αναδυόμενων προτύπων για ταξινόμηση έγιναν μια σειρά από πειράματα που περιλάμβαναν διαφορετικό αριθμό από τα πιο ισχυρά χι αναδυόμενα πρότυπα. Η ταξινόμηση πραγματοποιήθηκε σύμφωνα με τις εξισώσεις 2.5 και 2.6 και την Εξίσωση 7.1 για την προσαρμογή των βαθμολογιών. Τα αποτελέσματα παρουσιάζονται στην Εικόνα 7.9. Όπως γίνεται φανερό, η ακρίβεια αυξάνεται καθώς αυξάνεται ο αριθμός των χι αναδυόμενων προτύπων που συμμετέχουν στην ταξινόμηση. Ωστόσο, η αύξηση της ακρίβειας μετά από ένα συγκεκριμένο αριθμό προτύπων (περίπου 500) είναι μικρή.



Εικόνα 7.9: Απόδοση ταξινόμησης με χρήση των ισχυρότερων χι αναδυόμενων προτύπων.

7.5.5 Αξιολόγηση της Βαθμολόγησης με Βάση την Απόσταση

Η Εικόνα 7.10 παρουσιάζει τη μέση απόσταση των αλληλουχιών της θετικής τάξης από κάθε τάξη και υποτάξη. Συγκεκριμένα, δημιουργήθηκαν πέντε πίνακες κατάταξης νουκλεοτιδίων (Εδάφιο 7.4.6) χρησιμοποιώντας πέντε διαφορετικά σύνολα δεδομένων εκπαίδευσης (θετικές αλληλουχίες, αλληλουχίες 5΄ UTR, κωδικές αλληλουχίες, ιντρόνια, σύνολο των αρνητικών αλληλουχιών). Έπειτα, χρησιμοποιώντας ένα σύνολο αξιολόγησης με θετικές αλληλουχίες υπολογίστηκε η μέση απόσταση από κάθε τάξη και υποτάξη.



Εικόνα 7.10: Μέση απόσταση αλληλουχιών της θετικής τάξης από κάθε τάξη και υποτάξη.

Όπως ήταν αναμενόμενο κοντινότερη τάξη στις θετικές αλληλουχίες είναι η θετική. Η επόμενη πιο κοντινή υποτάξη είναι τα ιντρόνια και η πιο απομακρυσμένη τάξη είναι οι κωδικές αλληλουχίες. Μια σημαντική παρατήρηση είναι ότι η μέση απόσταση των θετικών αλληλουχιών από την αρνητική τάξη (Σύνολο Αρνητικών) είναι σχεδόν ίση με τη μέση απόστασή τους από την υποτάξη των ιντρονίων. Αυτό δείχνει ότι ο διαχωρισμός των θετικών αλληλουχιών από τις αρνητικές παρουσιάζει σχεδόν την ίδια δυσκολία με το διαχωρισμό των θετικών αλληλουχιών από τα ιντρόνια.

7.5.6 Αξιολόγηση του Ταξινομητή

Στο παρόν εδάφιο παρουσιάζεται η αξιολόγηση της μεθόδου PolyA-iEP συνολικά. Ο ταξινομητής που δημιουργείται χρησιμοποιεί τις βαθμολογίες που παράγονται στα προηγούμενα στάδια και βάσει αυτών λαμβάνει την τελική απόφαση. Διεξήχθησαν πειράματα με 7 διαφορετικούς ταξινομητές (κάποιοι από αυτούς δημιουργούνται με διαφορετική παραμετροποίηση των ίδιων αλγορίθμων). Οι αλγόριθμοι που χρησιμοποιήθηκαν περιλαμβάνονται στη βιβλιοθήκη αλγορίθμων μηχανικής μάθησης Weka [Witten & Frank, 2005]. Οι επτά ταξινομητές που δημιουργήθηκαν είναι οι εξής:

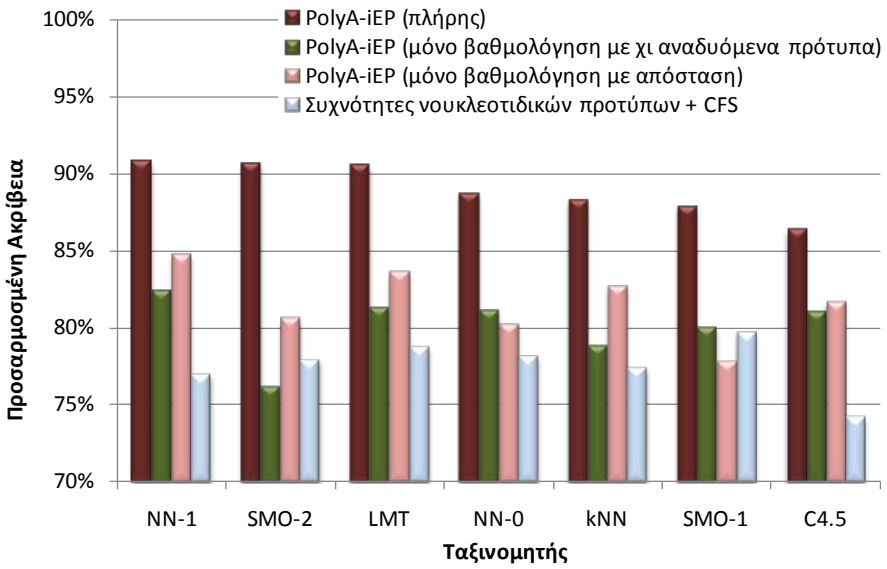
- Νευρωνικό δίκτυο χωρίς κρυφά επίπεδα και με ανάστροφη μετάδοση σφάλματος (NN–0).
- Νευρωνικό δίκτυο με ένα κρυφό επίπεδο και με ανάστροφη μετάδοση σφάλματος (NN–1).
- Μηχανή διανυσμάτων υποστήριξης (SMO–1) με γραμμικό πυρήνα. Χρησιμοποιήθηκε ο αλγόριθμος ελάχιστης σειριακής βελτιστοποίησης (sequential minimal optimization) [Platt, 1998] που δημιουργεί γενικευμένους γραμμικούς ταξινομητές.
- Μηχανή διανυσμάτων υποστήριξης (SMO–2) με τετραγωνικό πυρήνα. Χρησιμοποιήθηκε ο ίδιος αλγόριθμος που χρησιμοποιήθηκε για τη δημιουργία του SMO-1.
- Δένδρο λογιστικών μοντέλων (logistic model tree – LMT). Αποτελεί δένδρο ταξινόμησης με συναρτήσεις λογιστικής παλινδρόμησης στα φύλλα. [Landwehr et al., 2005; Sumner et al., 2005].
- Δένδρο ταξινόμησης (C4.5). Χρησιμοποιήθηκε ο αλγόριθμος C4.5 [Quinlan, 1993]. Για το κλάδεμα του δένδρου χρησιμοποιήθηκε η τεχνική reduced-error pruning.
- *k*–Κοντινότεροι γείτονες (*k*–NN) [Aha et al., 1991]. Ο αριθμός των γειτόνων *k* επιλέχθηκε με διαδικασία σταυρωτής επικύρωσης.

Με σκοπό την αξιολόγηση της αναγκαιότητας συνδυασμού των δύο μεθόδων βαθμολόγησης της PolyA-iEP στα πειράματα που έγιναν συμπεριλήφθηκαν δύο επιπλέων

εκδοχές της PolyA-iEP, μια που δεν περιλαμβάνει τη βαθμολόγηση με βάση την απόσταση και μια που δεν περιλαμβάνει τη βαθμολόγηση με βάση τα χι αναδυόμενα πρότυπα. Επίσης, για να εκτιμηθεί η συνολική απόδοση της μεθόδου PolyA-iEP, συγκρίθηκε με μια βασική προσέγγιση η οποία περιλαμβάνει τη χρήση ενός μεγάλου πλήθους χαρακτηριστικών που αναπαριστούν τη συχνότητα προτύπων νουκλεοτιδίων σε ένα στιγμιότυπο (αλληλουχία). Πιο συγκεκριμένα, κάθε στιγμιότυπο αναπαρίσταται από ένα διάνυσμα που περιέχει τις συχνότητες των 5.460 προτύπων νουκλεοτιδίων. Το μήκος των προτύπων κυμαίνεται από 1 έως 6. Η προσέγγιση αυτή παρουσιάζει αρκετά μειωμένη απόδοση. Συγκεκριμένα, κανένας ταξινομητής δεν επιτυγχάνει προσαρμοσμένη ακρίβεια πάνω από 75%. Για το λόγο αυτό, προστέθηκε ένα βήμα επιλογής χαρακτηριστικών με σκοπό την απομάκρυνση των λιγότερο σχετικών χαρακτηριστικών και τη βελτίωση της απόδοσης. Η μέθοδος που χρησιμοποιήθηκε για την επιλογή χαρακτηριστικών βασίζεται στη συσχέτιση (correlation–based feature selection – CFS) [Hall, 1999]. Με τη μέθοδο αυτή επιλέχθηκαν 121 χαρακτηριστικά και βελτιώθηκε σημαντικά η απόδοση της προσέγγισης.

Στην Εικόνα 7.11 παρουσιάζονται τα αποτελέσματα της σύγκρισης όλων των προσεγγίσεων με χρήση όλων των ταξινομητών. Η απόδοση μετράται με το μέτρο της προσαρμοσμένης ακρίβειας. Τα συγκεκριμένα πειράματα αφορούν το γενικό πρόβλημα ταξινόμησης θετικές αλληλουχίες/σύνολο αρνητικών αλληλουχιών. Όπως φαίνεται στο γράφημα, η μέθοδος PolyA-iEP είναι εμφανώς καλύτερη από οποιαδήποτε άλλη μέθοδο, με το νευρωνικό δίκτυο με ένα κρυφό επίπεδο να έχει την μεγαλύτερη απόδοση, πετυχαίνοντας προσαρμοσμένη ακρίβεια 91%.

Ο Πίνακας 7.6 παρουσιάζει την ευαισθησία (sensitivity), την ειδικότητα (specificity) και την προσαρμοσμένη ακρίβεια (adjusted accuracy) της μεθόδου PolyA-iEP για κάθε έναν από τους ταξινομητές που χρησιμοποιήθηκαν.



Εικόνα 7.11: Προσαρμοσμένη ακρίβεια των ταξινομητών.

Μια άλλη προσέγγιση που μπορεί να χρησιμοποιηθεί, ώστε να συγκριθεί με την PolyA-iEP είναι ο ταξινομητής των χι αναδυόμενων προτύπων που δημιουργείται σύμφωνα με τις εξισώσεις 2.5 και 2.6. Αυτή η προσέγγιση είναι διαφορετική γιατί οι βαθμολογίες που προκύπτουν από την εξόρυξη των αναδυόμενων προτύπων δεν χρησιμοποιούνται για την τροφοδότηση ενός ταξινομητή αλλά γίνεται απ' ευθείας ταξινόμηση βάσει των προαναφερθεισών εξισώσεων. Η προσαρμοσμένη ακρίβεια του συγκεκριμένου ταξινομητή δεν ξεπερνά το 80% και επομένως σε κάθε περίπτωση είναι χειρότερη από τη μέθοδο PolyA-iEP.

Πίνακας 7.6: Απόδοση των ταξινομητών της μεθόδου PolyA-iEP.

Ταξινομητής	Ευαισθησία	Ειδικότητα	Προσαρμοσμένη Ακρίβεια
NN-1	93,7%	88,2%	91,0%
SMO-2	93,1%	88,3%	90,7%
LMT	93,1%	88,1%	90,6%
NN-0	91,7%	85,7%	88,7%
k-NN	91,9%	84,7%	88,3%
SMO-1	91,7%	84,0%	87,9%
C4.5	90,5%	82,3%	86,4%

Ο Πίνακας 7.7 παρουσιάζει τη στατιστική σύγκριση με επίπεδο εμπιστοσύνης 95% των 7 ταξινομητών της μεθόδου PolyA-iEP για το διαχωρισμό των θετικών αλληλουχιών από το σύνολο των αρνητικών αλληλουχιών. Η κατά ζεύγη σύγκριση των ταξινομητών πραγματοποιήθηκε χρησιμοποιώντας το T-Test και τη διαδικασία 10-πλής σταυρωτής επικύρωσης. Το σύμβολο «+» σε ένα κελί του πίνακα δηλώνει ότι ο ταξινομητής της αντίστοιχης γραμμής πέτυχε στατιστικά σημαντικά καλύτερη απόδοση από τον ταξινομητή της αντίστοιχης στήλης. Παρόμοια, το σύμβολο «-» σε ένα κελί του πίνακα δηλώνει ότι ο ταξινομητής της αντίστοιχης γραμμής πέτυχε στατιστικά σημαντικά χειρότερη απόδοση από τον ταξινομητή της αντίστοιχης στήλης. Η απουσία στατιστικά σημαντικής διαφοράς μεταξύ δύο ταξινομητών δηλώνεται με το σύμβολο «0». [22]

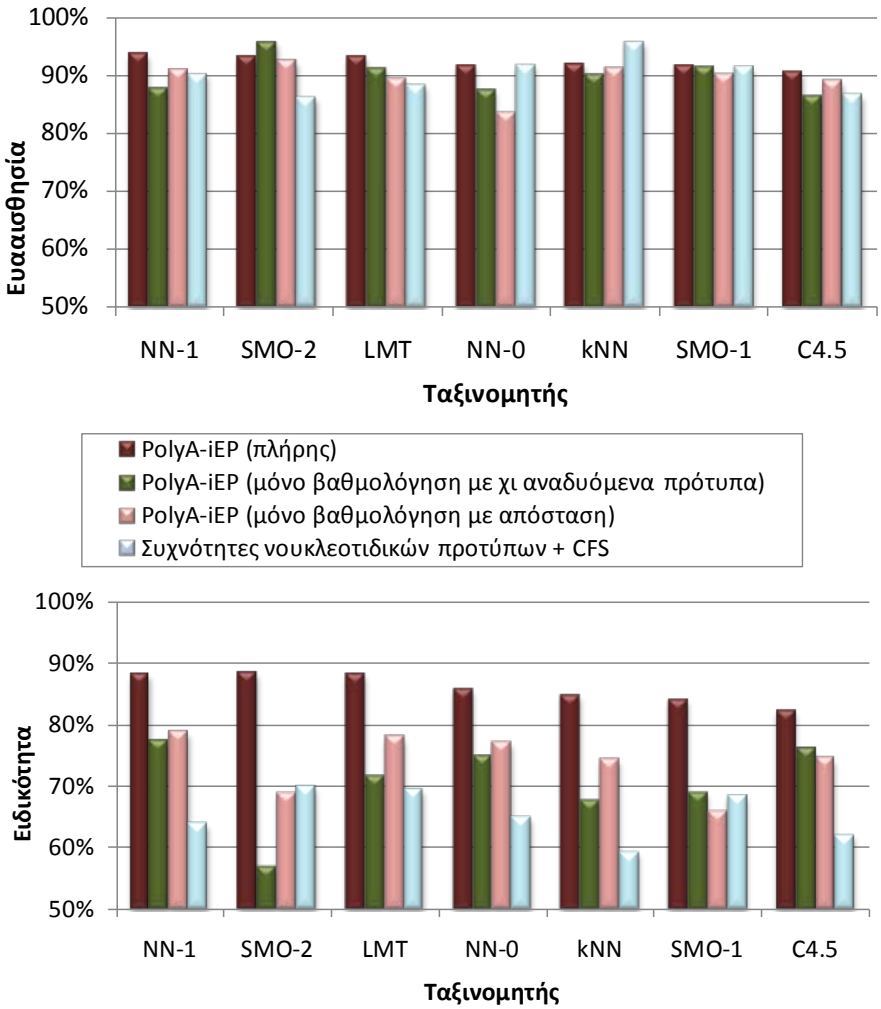
Όπως γίνεται φανερό από τον πίνακα, τρεις ταξινομητές (NN-1, SMO-2 και LMT) είναι όλοι τους στατιστικά σημαντικά καλύτεροι σε σύγκριση με τους υπόλοιπους (NN-0, k-NN,SMO-1 και C4.5). Ωστόσο, κανένας από τους ταξινομητές αυτής της τριάδας δεν είναι στατιστικά σημαντικά καλύτερος από τους άλλους δύο. Συνεπώς, οι συγκεκριμένοι τρεις ταξινομητές έχουν την ίδια απόδοση και υπερτερούν έναντι των άλλων.

[22] Ο Πίνακας 7.7 είναι ουσιαστικά τριγωνικός, καθώς η υπεροχή (+) και η υστέρηση (-) μεταξύ δύο ταξινομητών είναι έννοιες αντίθετες. Οπότε, όταν εμφανίζεται το σύμβολο «+» σε ένα κελί του πίνακα, τότε στο συμμετρικό ως προς την κύρια διαγώνιο κελί εμφανίζεται το σύμβολο «-» και αντίστροφα. Επίσης, το σύμβολο «0» εμφανίζεται και στα δύο συμμετρικά κελιά.

Πίνακας 7.7: Στατιστική σύγκριση των ταξινομητών για την ταξινόμηση θετικές/σύνολο αρνητικών (+/-/0: ο ταξινομητής της γραμμής υπερέχει/υστερεί/δε διαφέρει).

	NN-1	SMO-2	LMT	NN-0	*k*-NN	SMO-1	C4.5
NN-1		0	0	+	+	+	+
SMO-2	0		0	+	+	+	+
LMT	0	0		+	+	+	+
NN-0	-	-	-		0	0	+
***k*-NN**	-	-	-	0		0	+
SMO-1	-	-	-	0	0		0
C4.5	-	-	-	-	-	0	

Η Εικόνα 7.12 παρουσιάζει γραφικά την ευαισθησία και την ειδικότητα των ταξινομητών για όλες τις συγκρινόμενες προσεγγίσεις.



Εικόνα 7.12: Ευαισθησία και ειδικότητα των ταξινομητών.

Όπως γίνεται φανερό από την Εικόνα 7.12, αυτό που πραγματικά πετυχαίνει η μέθοδος PolyA-iEP, είναι η σημαντική βελτίωση της εξαιρετικά χαμηλής ειδικότητας που παρουσιάζουν οι άλλες προσεγγίσεις, καθώς επίσης και η μη-αμελητέα (στις περισσότερες περιπτώσεις) βελτίωση της ευαισθησίας. Ο βασικός λόγος για τον οποίο επιτυγχάνεται αυτή η βελτίωση της απόδοσης είναι η συμμετοχή των δύο ξεχωριστών μεθόδων βαθμολόγησης, της βαθμολόγησης με βάση τα χι αναδυόμενα πρότυπα και της βαθμολόγησης με βάση την απόσταση. Αν εξεταστεί προσεκτικά η Εικόνα 7.11, γίνεται αντιληπτό ότι η βαθμολόγηση με βάση τα χι αναδυόμενα πρότυπα είναι αποτελεσματικότερη από τη βαθμολόγηση με βάση την απόσταση, όταν χρησιμοποιούνται γραμμικοί ταξινομητές (NN-0 και SMO-1). Ακριβώς το αντίθετο παρατηρείται, όταν χρησιμοποιούνται μη-γραμμικοί ταξινομητές (NN-1, SMO-2, LMT, *k*-NN και C4.5). Αυτή η παρατήρηση αποτελεί μια σαφή ένδειξη, ότι η προτεινόμενη μέθοδος βαθμολόγησης με βάση την απόσταση κατορθώνει να αναπαραστήσει επιτυχώς τη μη-γραμμική συνιστώσα του προβλήματος που αντιμετωπίζεται στο παρόν κεφάλαιο. Η μη-γραμμικότητα του προβλήματος δεν αναπαρίσταται ικανοποιητικά σε καμία από τις υπόλοιπες προσεγγίσεις και γι' αυτό η αποτελεσματικότητά τους είναι χαμηλή.

7.6 Συμπεράσματα

Η πρόβλεψη του σημείου πολυαδενυλίωσης ή σημείου πολύ(Α) σε μια αλληλουχία mRNA είναι ένα σημαντικό πεδίο έρευνας, που προσελκύει την προσοχή ερευνητών της βιοπληροφορικής, της βιολογίας και της ιατρικής. Το ενδιαφέρον εστιάζεται στην αποτελεσματική πρόβλεψη του σημείου πολύ(Α) και στην ανακάλυψη νέων προτύπων και σημάτων γύρω από αυτό. Η μέθοδος που παρουσιάστηκε στο παρόν κεφάλαιο κινείται και προς τις δύο κατευθύνσεις. Οι δυσκολίες στην πρόβλεψη του σημείου πολύ(Α) προέρχονται κυρίως από την απουσία σημαντικά συντηρημένων σημάτων γύρω από αυτό. Τον Αύγουστο του 2009 οι Mayr & Bartel δημοσίευσαν μια μελέτη τους σε φυσιολογικά και καρκινικά κύτταρα. Τα αποτελέσματα της μελέτης έδειξαν μια ισχυρή συσχέτιση μεταξύ του μήκους της περιοχής 3΄ UTR των αλληλουχιών mRNA ογκογονιδίων και του επιπέδου έκφρασης αυτών των ογκογονιδίων. Το ενδιαφέρον σε αυτήν την περίπτωση είναι ότι το μήκος της περιοχής 3΄ UTR των αλληλουχιών mRNA καθορίζεται από τη θέση του σημείου πολύ(Α) μέσα στην αλληλουχία mRNA. Είναι προφανές ότι ερευνητική μελέτη της διαδικασίας της πολυαδενυλίωσης μπορεί να συνεισφέρει σημαντικά στην κατανόηση της λειτουργίας των οργανισμών και των ασθενειών και αναμένεται στο μέλλον να αποτελέσει ένα από τα πιο ενδιαφέροντα πεδία της βιοπληροφορικής.

Η μέθοδος PolyA-iEP που παρουσιάστηκε σε αυτό το κεφάλαιο της διατριβής αντιμετωπίζει αποτελεσματικά το πρόβλημα της πρόβλεψης του σημείου πολύ(Α) και εστιάζει κυρίως στο διαχωρισμό των αλληλουχιών που περιέχουν ένα σημείο πολύ(Α) από τις αλληλουχίες που δεν περιέχουν τέτοιο σημείο. Η συγκεκριμένη πτυχή του προβλήματος δεν έχει μελετηθεί επαρκώς σε άλλες εργασίες. Η μέθοδος αποτελείται

από δύο βασικά τμήματα. Το πρώτο βασίζεται στα αναδυόμενα πρότυπα και το δεύτερο σε μια διαδικασία βαθμολόγησης των αλληλουχιών βάσει της απόστασής τους από διάφορες τάξεις και υποτάξεις αλληλουχιών. Το δεύτερο αυτό τμήμα αποτελεί καινοτομία της μεθόδου. Καινοτομία αποτελούν, επίσης, τα εκτεταμένα πρότυπα, που μπορούν να εκτείνονται σε περισσότερα από ένα στοιχεία και ανακαλύπτονται από το πρώτο τμήμα της μεθόδου. Η αποτελεσματικότητα της μεθόδου PolyA-iEP είναι αρκετά μεγάλη και στα πειράματα που διεξήχθησαν έφτασε το 93,7% σε ευαισθησία και το 88,2% σε ειδικότητα. Ένα άλλο σημαντικό στοιχείο της μεθόδου είναι ότι είναι απόλυτα παραμετροποιήσιμη και τα μοντέλα της μπορούν εύκολα να επανεκπαιδευτούν και να χρησιμοποιηθούν με αλληλουχίες από διάφορους οργανισμούς.

Επίλογος

8

Σε αυτό το κεφάλαιο συγκεντρώνονται τα σημαντικότερα συμπεράσματα που προέκυψαν κατά την εκπόνηση της διατριβής και παρουσιάζονται κάποιες ιδέες για μελλοντικές επεκτάσεις.

8.1 Συμπεράσματα

Αντικείμενο έρευνας της παρούσας διατριβής αποτέλεσε η ανακάλυψη γνώσης από βιολογικά δεδομένα. Η διατριβή κινήθηκε στα πλαίσια των ερευνητικών περιοχών της ανακάλυψης γνώσης από βάσεις δεδομένων και της βιοπληροφορικής. Αυτά τα δύο ερευνητικά πεδία προέκυψαν ως αποτέλεσμα της προόδου των επιστημών της πληροφορικής και της βιολογίας, έχοντας ως βασικό τους σκοπό να καλύψουν την ανάγκη της αποδοτικής και αποτελεσματικής διαχείρισης και ανάλυσης μεγάλων όγκων δεδομένων.

Στο δεύτερο κεφάλαιο της διατριβής παρουσιάστηκε το ερευνητικό πεδίο της ανακάλυψης γνώσης από βάσεις δεδομένων και στο τρίτο κεφάλαιο παρουσιάστηκαν τα ερευνητικά πεδία της μοριακής βιολογίας και της βιοπληροφορικής. Αυτά τα δύο κεφάλαια παρέχουν το απαιτούμενο υπόβαθρο για την κατανόηση της ερευνητικής εργασίας που πραγματοποιήθηκε και παρουσιάζουν τις βασικές κατευθύνσεις της έρευνας αυτών των ερευνητικών πεδίων, καθώς και τα οφέλη που προκύπτουν από τις εφαρμογές τους. Μάλιστα, το εύρος και η βαρύνουσα σημασία των εφαρμογών τους καθιστούν σαφές γιατί αυτά τα πεδία προσελκύουν το ενδιαφέρον των ερευνητών και αναμένεται να το προσελκύσουν ακόμη περισσότερο στο μέλλον.

Στο τέταρτο κεφάλαιο της διατριβής προτάθηκε ένα νέο πρόβλημα, που αποσκοπεί στην ανακάλυψη αμοιβαίως αποκλειόμενων αντικειμένων. Το συγκεκριμένο πρόβλημα μπορεί να ενταχθεί στο ευρύτερο πεδίο ανακάλυψης συσχετίσεων και ως εκ τούτου μπορεί να θεωρηθεί επέκταση του παραδοσιακού προβλήματος της ανακάλυψης

κανόνων συσχέτισης. Μάλιστα, η γνώση ότι δύο αντικείμενα είναι αμοιβαίως αποκλειόμενα, μπορεί να αποτελέσει σημαντική πληροφορία κατά την αναζήτηση άγνωστων ταξονομικών σχέσεων μεταξύ τους. Η συμβολή της παρούσας διατριβής στο αντικείμενο αυτό είναι ο ορισμός του προβλήματος, ο ορισμός μέτρων αμοιβαίου αποκλεισμού και η υλοποίηση ένας αλγόριθμου ανακάλυψης αμοιβαίως αποκλειόμενων αντικειμένων.

Στο πέμπτο κεφάλαιο της διατριβής παρουσιάστηκε μια προσέγγιση για την αποτελεσματική και αποδοτική ταξινόμηση δεδομένων γονιδιακής έκφρασης αξιοποιώντας την πληροφορία που παρέχουν οι συσχετίσεις μεταξύ γονιδιακών εκφράσεων. Για την αξιοποίηση των συσχετίσεων γίνεται χρήση ενός αλγόριθμου εξόρυξης συχνών συνόλων αντικειμένων. Τα πειράματα που διεξήχθησαν έδειξαν ότι ένας μικρός αριθμός χαρακτηριστικών, τα οποία είναι συχνά σύνολα γονιδιακών εκφράσεων, είναι ικανός να προσφέρει πολύ αποτελεσματικούς ταξινομητές με ακρίβεια που κυμαίνεται στο 95%. Ο αριθμός των απαιτούμενων χαρακτηριστικών μπορεί να είναι μικρότερος από 100. Το όφελος σε αυτήν την περίπτωση είναι διπλό, καθώς βελτιώνεται η αποτελεσματικότητα των ταξινομητών, αλλά μπορούν και να μειωθούν σημαντικά οι διαστάσεις των δεδομένων και κατ' επέκταση το υπολογιστικό κόστος.

Στο έκτο κεφάλαιο της διατριβής παρουσιάστηκαν προσεγγίσεις που υλοποιήθηκαν στα πλαίσια της παρούσας διατριβής για την πρόβλεψη του σημείου έναρξης της μετάφρασης. Προτάθηκαν και μελετήθηκαν νέα χαρακτηριστικά, τα οποία παρουσίασαν σημαντική ικανότητα χαρακτηρισμού των τάξεων στα πειράματα που διεξήχθησαν. Επίσης, αξιολογήθηκε ο τρόπος επιλογής του σημείου έναρξης της μετάφρασης βάσει των αποφάσεων ταξινομητών, καθώς και βάσει του μοντέλου σάρωσης του ριβοσώματος. Η πιο ολοκληρωμένη και αποτελεσματική μέθοδος πρόβλεψης του σημείου έναρξης της μετάφρασης που προτάθηκε στα πλαίσια της διατριβής, είναι η μέθοδος StackTIS, που αποτελείται από τρία βασικά στοιχεία πρόβλεψης, καθένα από τα οποία είναι προσανατολισμένο σε ένα διαφορετικό μέρος του προβλήματος. Αξίζει να σημειωθεί, πως η χρήση του στοιχείου απόστασης του ATG παρέχει στη μέθοδο τα πλεονεκτήματα του μοντέλου σάρωσης του ριβοσώματος, ενώ ταυτόχρονα ο συνδυασμός του με τα άλλα στοιχεία περιορίζει την επίδραση των μειονεκτημάτων του συγκεκριμένου μοντέλου. Τα τρία επιμέρους στοιχεία συνδυάζονται με μια ιδιαίτερα αποτελεσματική μέθοδο συνδυασμού ταξινομητών, τη στοιβαγμένη γενίκευση. Η έξοδος της μεθόδου StackTIS είναι μια φθίνουσα κατάταξη των πιθανοτικών προβλέψεων. Το πλεονέκτημα των πιθανοτικών έναντι των δυαδικών αποφάσεων είναι ότι παρέχουν εκτός από την ταξινόμηση ενός στιγμιότυπου και μια εκτίμηση της βεβαιότητας με την οποία αυτό ταξινομείται. Το χαρακτηριστικό αυτό είναι ιδιαίτερα σημαντικό όταν περισσότερα από ένα κωδικόνια ATG έχουν προβλεφθεί ως σημεία έναρξης της μετάφρασης ή όταν είναι επιθυμητό να ληφθούν υπόψη και εναλλακτικά σημεία έναρξης της μετάφρασης. Εκτενή πειράματα έδειξαν ότι η μέθοδος StackTIS παρουσιάζει ιδιαίτερα υψηλή αποτελεσματικότητα η οποία βάσει των μέτρων της ακρίβειας και της προσαρμοσμένης

ακρίβειας αποδεικνύεται πως είναι στατιστικά σημαντική σε σχέση με άλλες δημοφιλείς προσεγγίσεις.

Στο έβδομο κεφάλαιο της διατριβής παρουσιάστηκε η μέθοδος PolyA-iEP, η οποία έχει ως σκοπό την αποτελεσματική πρόβλεψη του σημείου αποκοπής και πολυαδενυλίωσης. Η μέθοδος εστιάζει στο διαχωρισμό των αλληλουχιών που περιέχουν ένα σημείο πολυαδενυλίωσης από τις αλληλουχίες που δεν περιέχουν τέτοιο σημείο, ένα θέμα το οποίο δεν έχει μελετηθεί επαρκώς σε άλλες μελέτες. Η PolyA-iEP αποτελείται από δύο βασικά μέρη. Το πρώτο βασίζεται στα αναδυόμενα πρότυπα και το δεύτερο, που αποτελεί καινοτομία της διατριβής, σε μια διαδικασία βαθμολόγησης των αλληλουχιών βάσει της απόστασής τους από διάφορες τάξεις και υποτάξεις αλληλουχιών. Καινοτομία αποτελούν, επίσης, τα εκτεταμένα πρότυπα, που μπορούν να εκτείνονται σε περισσότερα από ένα στοιχεία και ανακαλύπτονται από το πρώτο τμήμα της μεθόδου. Η αποτελεσματικότητα της μεθόδου PolyA-iEP είναι πολύ μεγάλη αγγίζοντας το 93,7% σε ευαισθησία (sensitivity) και το 88,2% σε ειδικότητα (specificity).

8.2 Μελλοντικές Επεκτάσεις

Στις σκέψεις για μελλοντικές επεκτάσεις που αφορούν τον αλγόριθμο ανακάλυψης αμοιβαίως αποκλειόμενων αντικειμένων, που παρουσιάστηκε στο τέταρτο κεφάλαιο, περιλαμβάνονται τοσο η βελτίωση της απόδοσης του, όσο και η χρήση του σε περισσότερα πεδία εφαρμογών. Μάλιστα, στην τελευταία περίπτωση θα ήταν ιδιαίτερα θεμιτό να υπάρχει διαθέσιμη ταξονομική πληροφορία για τα δεδομένα, ώστε να αξιολογηθεί ακριβέστερα η αποτελεσματικότητα του αλγορίθμου.

Μια από τις σκέψεις για μελλοντική εργασία σχετικά με την προσέγγιση της ταξινόμησης δεδομένων γονιδιακής έκφρασης που ελήφθησαν με την τεχνική SAGE (κεφάλαιο 5) είναι η προσαρμογή της μεθόδου για εφαρμογή σε δεδομένα μικροσυστοιχιών (microarrays). Παράλληλα, ιδιαίτερο ενδιαφέρον θα είχε η μελέτη της επίδρασης των σφαλμάτων αλληλούχισης καθώς και άλλων πιθανών πηγών θορύβου στην αποτελεσματικότητα της ταξινόμησης.

Σχετικά με τις προσεγγίσεις που σχετίζονται με την πρόβλεψη του σημείου έναρξης της μετάφρασης, όπως είναι η μέθοδος StackTIS, που παρουσιάστηκε στο έκτο κεφάλαιο, οι ιδέες για μελλοντική επέκταση επικεντρώνονται στην πρόβλεψη του σημείου έναρξης της μετάφρασης σε ατελείς αλληλουχίες. Τέτοιες αλληλουχίες είναι οι ετικέτες εκφρασμένων αλληλουχιών (ESTs), οι οποίες δεν είναι πλήρεις και μάλιστα είναι συχνή η παρουσία σφαλμάτων κύλισης του πλαισίου ανάγνωσης. Για τους λόγους αυτούς απαιτείται ειδική διαχείριση των αλληλουχιών αυτών, η οποία μπορεί να πραγματοποιηθεί μέσω κατάλληλων τροποποιήσεων της μεθόδου StackTIS. Επίσης, στις σκέψεις για μελλοντική εργασία περιλαμβάνεται η χρήση της μεθόδου StackTIS για το συμπληρωματικό πρόβλημα της αναγνώρισης μη κωδικών αλληλουχιών RNA, όπως είναι τα microRNAs.

Τα μελλοντικά σχέδια επέκτασης της μεθόδου PolyA-iEP, που παρουσιάστηκε στο έβδομο κεφάλαιο, περιλαμβάνουν τη μελέτη για χρήση πιο πολύπλοκων μεθόδων ταξινόμησης, όπως είναι οι ομάδες ταξινομητών, αλλά και την εφαρμογή της μεθόδου για πρόβλεψη του σημείου πολυαδενυλίωσης σε αλληλουχίες του ανθρώπινου γονιδιώματος. Επίσης, ιδιαίτερο ενδιαφέρον παρουσιάζει η μελέτη του προβλήματος της εναλλακτικής πολυαδενυλίωσης και θα μπορούσε να αποτελέσει μια από τις μελλοντικές εργασίες. Μάλιστα, η πρόσφατα δημοσιευμένη μελέτη [Mayr & Bartel, 2009] που συσχετίζει την πολυαδενυλίωση με τον καρκίνο, μπορεί να αποτελέσει κίνητρο για την έρευνα προς αυτήν την κατεύθυνση.

Βιβλιογραφία

Agarwal, P.K. and Bafna, V. (1998). The Ribosome Scanning Model for Translation Initiation: Implications for Gene Prediction and Full-Length cDNA Detection. In *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology (ISMB '98)*, AAAI Press, 2-7.

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining Association Rules Between Sets of Items in Large Databases. *Proceedings of the ACM SIGMOD Conference on Management of Data*, 207-216.

Agrawal, R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Databases*, 478-499.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A.I. (1996). Fast discovery of association rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Editors), In *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, California, USA, 307-328.

Aha, D.W., Kibler, D., and Albert, M.K. (1991). Instance-based learning algorithms. *Machine Learning*, 6, 37-66.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3), 403-410.

Alves, A., Zagoruiko, N., Okun, O., Kutnenko, O., and Borisova, I. (2005). Predictive Analysis of Gene Expression Data from Human SAGE Libraries. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, pp. 60–71.

Ayer, M. Brunk, H., Ewing, G., and Reid, W. (1955). An empirical distribution function for sampling with incomplete information, *The Annals of Mathematical Statistics*, vol. 26, no. 4, pp. 641-647.

Becquet, C. Blachon, S., Jeudy, B., Boulicaut, J.-F., and Gandrillon, O. (2002). Strong-Association-Rule Mining for Large-Scale Gene-Expression Data Analysis: A Case Study on Human SAGE Data. *Genome Biology*, 3(12): research0067.

Benson, D., Boguski, M., Lipman, D., and Ostell, J. (1997). Genbank. *Nucleic Acids Research*, 25, 1–6.

Berberidis, C., Tzanis, G., and Vlahavas, I. (2005). Mining for Contiguous Frequent Itemsets in Transaction Databases, In *Proceedings of the IEEE 3rd International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, Sofia, Bulgaria, 679-685.

Bezdek, J.C., Ehrlich, R., Full, W. (1984). FCM: The Fuzzy C-Means Clustering Algorithm, *Computers and Geosciences*, 10, 191-203.

Caruana, R. Munson, A., and Niculescu-Mizil, A. (2006). Getting the Most Out of Ensemble Selection. In *Proccedings of the 6th IEEE International Conference on Data Mining (ICDM'06)*, Hong Kong, pp.828-833.

Chang, C.-C. and Lin, C.-J. (2005). *LIBSVM - A Library for Support Vector Machines*. URL: http://www.csie.ntu.edu.tw/~cjlin/libsvm (προσπελάστηκε: 17/05/2011)

Chen, X. and Petrounias, I. (2000). Discovering Temporal Association Rules: Algorithms, Language and System. In *Proceedings of the 16th International Conference on Data Engineering*, pp. 306.

Cheng, Y. Miura, R.M., and Tian, B. (2006). Prediction of mRNA polyadenylation sites by support vector machine. *Bioinformatics*, 22(19), 2320-2325.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson, J.D. (2003). Multiple Sequence Alignment with the Clustal Series of Programs. *Nucleic Acids Research*, 31 (13): 3497–3500.

Cigan, A., Feng, L., and Donahue, T. (1988). tRNA functions in directing the scanning ribosome to the start site of translation. *Science*, 242, 93-97.

Cohen, W. (1995). Fast Effective Rule Induction. In *Proceedings of the 12th International Conference on Machine Learning*. Morgan Kaufmann, Lake Tahoe, USA, 80-89.

Connor, M. and Ferguson Smith, M. (1997). *Essential Medical Genetics*, (5th Edition). Wiley-Blackwell, MA, USA.

Cortes, C. and Vapnik, V., Support-Vector Networks, *Machine Learning*, 20, 1995.

Dietterich, T. (1999). Machine Learning. In Wilson, R.A. and Keil F.C. (Eds.). *The MIT Encyclopedia of the Cognitive Sciences*, MIT Press.

Dong, G., and Li, J. (1999). Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 43–52.

Dunham, M.H. (2002). *Data Mining: Introductory and Advanced Topics*, Prentice Hall, Upper Saddle River, New Jersey, USA.

Dzeroski, S. and Zenko, B., (2004). Is Combining Classifiers with Stacking Better than Selecting the Best One?, Machine Learning, vol. 54, 255-273.

Edelstein, H. (1999). Introduction to Data Mining and Knowledge Discovery, (3rd Edition), Two Crows Corporation. URL: http://www.twocrows.com (προσπελάστηκε: 17/05/2011)

Edwalds-Gilbert, G., Veraldi, K.L., Milcarek, C. (1997). Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Research*, 25,13, 2547-2561.

El-Manzalawy, Y. and Honavar, V. (2005). *WLSVM: Integrating LibSVM into Weka Environment*. URL: http://www.cs.iastate.edu/~yasser/wlsvm (προσπελάστηκε: 17/05/2011).

Esseghir, M.A., Ben Yahia, S., and Abdelhak, S. (2005). Localizing Compact Set of Genes Involved in Cancer Diseases Using an Evolutionary Connectionist Approach. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, pp. 78-83.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, USA, 226-231.

Fan, H. (2004). *Efficient Mining of Interesting Emerging Patterns and Their Effective Use in Classification*, PhD Thesis, University of Melbourne, Australia.

Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. From Data Mining to Knowledge Discovery: An Overview. In Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds.). (1996). *Advances in Knowledge Discovery and Data Mining*, AAAI Press/The MIT Press, Menlo Park, California, USA, pp. 1-34.

Gamberoni, G. and Storari, S. (2004). Supervised and Unsupervised Learning Techniques for Profiling SAGE Results. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Pisa, Italy, pp. 121–126.

Gandrillon, O. (2004). Guide to the Gene Expression Data. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Pisa, Italy, 116–120.

Gasmi, G., Hamrouni, T., Abdelhak, S., Ben Yahia, S., and Mephu Nguifo, E. (2005). Extracting Generic Basis of Association Rules from SAGE Data. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, pp. 84–89.

Ghosh, D. and Chinnaiyan, A.M. (2002). Mixture Modeling of Gene Expression Data from Microarray Experiments, *Bioinformatics*, 18: 275-286.

Glasgow, J., Jurisica, I., and Rost, B. (2004). AI and Bioinformatics, *AI Magazine*, 25(1): 7-8.

Graber, J.H., McAllister, G.D., and Smith, T.F. (2002). Probabilistic prediction of Saccharomyces cerevisiae mRNA 3'-processing sites. *Nucleic Acids Research*, 30(8), 1851-1858.

Hall, M.A. (1999). *Correlation-based Feature Subset Selection for Machine Learning*. PhD Thesis, University of Waikato, Hamilton, New Zealand.

Han, J., & Fu, Y. (1994). Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Databases. In *AAAI Workshop on Knowledge Discovery in Databases*.157-168.

Han, J. and Fu, Y. (1995). Discovery of Multiple-Level Association Rules from Large Databases. In *Proceedings of the 21st International Conference on Very Large Databases*. 420-431.

Han, J. and Kamber, M. (2001). *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, USA.

Han, J., Pei, J., Yin, Y. (2000). Mining Frequent Patterns Without Candidate Generation. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Dallas, Texas, USA, 1-12.

Hajarnavis, A., Korf, I., and Durbin, R. (2004). A probabilistic model of 3′-end formation in Caenorhabditis elegans. *Nucleic Acids Research*, 32(11), 3392–3399.

Hatzigeorgiou, A. (2002). Translation Initiation Start Prediction in Human cDNAs with High Accuracy, *Bioinformatics*, vol. 18, no. 2, 343-350.

Hatzikos, E., Tsoumakas, G., Tzanis, G., Bassiliades, N., and Vlahavas, I. (2008). An Empirical Study of Sea Water Quality Prediction, *Knowledge-Based Systems*, Elsevier, 21(6): 471-478.

Hebert, C. Blachon, S., and Cremilleux, B. (2005). Mining δ-strong Characterization Rules in Large SAGE Data. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, pp. 90-101.

Hirsh, H. and Noordewier, M. (1994). Using Background Knowledge to Improve Inductive Learning of DNA Sequences, *Proceedings of the 10th IEEE Conference on Artificial Intelligence for Applications*, 351-357.

Holsheimer, M. and Siebes, A. (1994). *Data Mining: The Search for Knowledge in Databases*, Technical Report CS-R9406, CWI, Amsterdam, The Netherlands.

Houle, J.L., Cadigan, W., Henry, S., Pinnamaneni, A., and Lundahl, S. (2000) *Database Mining in the Human Genome Initiative*, Whitepaper, Bio-databases.com, Amita Corporation. URL: http://www.biodatabases.com/whitepaper01.html (προσπελάστηκε: 17/05/2011).

Hu, J., Lutz, C. S., Wilusz, J., and Tian, B. (2005). Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA*, 11(10), 1485-1493.

Hunter, L (1993). Molecular Biology for Computer Scientists, *Artificial Intelligence and Molecular Biology*, L. Hunter (Ed.), AAAI Press.

Hunter, L. (2004) Life and Its Molecules: A Brief Introduction, *AI Magazine,* 25(1), 9-22.

Ji, G., Zheng, J., Shen, Y., Wu, X., Jiang, R., Lin, Y., Loke, J., Davis, K, Reese, G., and Li, Q. (2007). Predictive modeling of plant messenger RNA polyadenylation sites. *BMC Bioinformatics*, 8:43.

John, G.H. and Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Mateo, USA, 338-345.

Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., and Murthy, K.R.K. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*. 13(3):637-649.

Kerr, K.M. and Churchill, G.A. (2001). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments, *Proceedings of the National Academy of Sciences*, 98, 8961-8965.

Kira, K., Rendell, L.A. (1992). A Practical Approach to Feature Selection. In Proceedings of the 9th International Workshop on Machine Learning, pp. 249-256.

Koh, C.H. and Wong, L. (2007). Recognition of polyadenylation sites from Arabidopsis genomic sequenses. In *Proceedings of 18th International Conference on Genome Informatics*, 73-82.

Kohavi, R. (1995a). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, In *Proceedings of 14th International Joint Conference on Artificial Intelligence*, 1137-1143.

Kohavi, R. (1995b). The Power of Decision Tables. In *Proceedings of the 8th European Conference on Machine Learning (ECML'95)*, Springer Verlag, LNAI 914, 174-189.

Kononenko, I. (1994). Estimating Attributes: Analysis and Extensions of RELIEF. In Proceedings of the European Conference on Machine Learning, pp. 171-182.

Koperski, K. and Han, J. (1995). Discovery of Spatial Association Rules in Geographic Information Databases. In *Proceedings of the 4th International Symposium on Large Spatial Databases*. 47-66.

Kozak, M. (1987). An Analysis of 5'-Noncoding Sequences from 699 Vertebrate Messenger RNAs. *Nucleic Acids Research*, 15(20), 8125-8148.

Kozak, M. (1996). Interpreting cDNA sequences: some insights from studies on translation. *Mammalian Genome*, 7, 563–574.

Kozak, M. (1989). The Scanning Model for Translation: An Update. *Journal of Cell Biology*, 108(2), 229-241.

Kozak, M. and Shatkin, A. J. (1978). Migration of 40 S Ribosomal Subunits on Messenger RNA in the Presence of Edeine. *Journal of Biological Chemistry*, 253(18), 6568-6577.

Krallinger, M. and Valencia, A. (2005). Text-Mining and Information-Retrieval Services for Molecular Biology. *Genome Biology*, 6(224).

Landwehr, N., Hall, M., and Frank, E.(2005). Logistic Model Trees. *Machine Learning*. 95(1-2), 161-205.

Laudon, K.C. and Laudon, J.P. (2001). *Essentials of Management Information Systems (4th Edition),* Prentice Hall, Upper Saddle River, New Jersey, USA.

Li, G., Leong, T.-Y., and Zhang, L. (2005). Translation Initiation Sites Prediction with Mixture Gaussian Models in Human cDNA Sequences, *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 17, 1152-1160.

Lin, H.-T. and Li, L. (2005). Analysis of SAGE Results with Combined Learning Techniques. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, pp. 102–113.

Lipman, D. and Pearson, W. (1985). Rapid and Sensitive Protein Similarity Searches. *Science*, 227: 1435-1441.

Lipman, D. and Pearson, W. (1988). Improved Tools for Biological Sequence Comparison. In *Proceedings of the National Academy of Sciences*, 85: 2444-2448.

Liu, H., Han, H., Li, J., and Wong, L. (2004). Using Amino Acid Patterns to Accurately Predict Translation Initiation Sites, *In Silico Biology*, vol. 4, no. 3, 255-269.

Liu, H. and Wong, L. (2003). Data Mining Tools for Biological Sequences. Journal of Bioinformatics and Computational Biology. 1(1), 139-168.

Liu, H., Han, H., Li, J., and Wong, L. (2003). An in-silico method for prediction of polyadenylation signals in human sequences. In *Proceedings of the 14th International Conference on Genome Informatics*, 84-93.

Loke, J., Stahlberg, E.A., Strenski, D.G., Haas, B.J. Wood, P.C., & Li, Q.Q. (2005). Compilation of mRNA polyadenylation signals in Arabidopsis revealed a new signal element and potential secondary structures. *Plant Physiology*, 138, 1457-1468.

Ma, Q. and Wang, J.T.L. (1999). Biological Data mining Using Bayesian Neural Networks: A Case Study, *International Journal on Artificial Intelligence Tools*, Special Issue on Biocomputing, 8(4), 433-451.

Mannila, H., Toivonen, H., and Verkamo, A.I. (1994). Efficient Algorithms for Discovering Association Rules. In *Proceedings of AAAI Workshop on Knowledge Discovery in Databases*, 181-192.

Martinez, R., Christen, R., Pasquier, C., and Pasquier, N. (2005). Exploratory Analysis of Cancer SAGE Data. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Porto, Portugal, pp. 72–77.

Mayr, C. and Bartel, D.P. (2009). Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138, 673–684.

Meselson, M., and Stahl, F.W. (1958). The Replication of DNA in Escherichia coli. *Proceedings of the National Academy of Sciences*, 44, 671-82.

Mitchell, T. (1997). Machine Learning, McGraw Hill, New York, USA.

Murphy, A. and Winkler, R. (1977). Reliability of subjective probability forecasts of precipitation and temperature. Applied Statistics, 26(1), 41-47.

Nadershahi, A., Fahrenkrug, S. C., & Ellis, L. B. (2004). Comparison of computational methods for identifying translation initiation sites in EST data. *Bioinformatics*, 5(14).

Needleman, S.B., and Wunch, C.D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, *Journal of Molecular Biology*, 48: 443-453.

Ng, R.T., Sander, J., and Sleumer, M.C. (2001). Hierarchical Cluster Analysis of SAGE Data for Cancer Profiling. *Proceedings of BIOKDD Conference*, 65-72.

Nishikawa, T., Ota, T., and Isogai, T. (2001). Prediction whether a Human cDNA Sequence Contains Initiation Codon by Combining Statistical Information and Similarity with Protein Sequences. *Bioinformatics*, 16(11), 960-967.

Pedersen, A.G., and Nielsen, H. (1997). Neural Network Prediction of Translation Initiation Sites in Eukaryotes: Perspectives for EST and Genome analysis. In *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, USA, 226-233.

Platt, J. (1998). Fast Training of Support Vector Machines using Sequential Minimal Optimization. In Schoelkopf, B., Burges, C., and Smola, A. (editors), *Advances in Kernel Methods - Support Vector Learning*, MIT Press.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, 35(Database issue), D61-D65.

Quinlan, J.R. *The Effect of Noise on Concept Learning*. In Michalski, R.S., Carbonell, J.G., and Mitchell, T. (Eds.). (1986a). Machine Learning, an Artificial Intelligenece Approach, vol. 2, Morgan Kaufmann, San Mateo, California, USA, pp. 149-166.

Quinlan, J.R. Induction of Decision Trees. (1986b). *Machine Learning*, 1: 81-106.

Quinlan, J. R. (1990). Learning Logical Definitions from Relations. *Machine Learning* 5(3), 239-266.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, USA (1993).

Rajapakse, J.C. and Ho, L.S. (2005). Markov Encoding for Detecting Signals in Genomic Sequences. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 2, 131-142.

Ramanathan, C.S. and Davison, D.B. (2002). Pharmaceutical Bioinformatics and Drug Discovery. In C.W. Sensen (Ed.), *Biotechnology: Genomics and Bioinformatics*, (Vol. 5b): 123-144, Wiley-VCH.

Rice, P. Longden, I., and Bleasby, A (2000). EMBOSS: The European Molecular Biology Open Software Suite, *Trends in Genetics*, 16(6): 276-277.

Rioult, F. (2004). Mining Strong Emerging Patterns in Wide SAGE Data. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Pisa, Italy, pp. 484–487.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Cornell Aeronautical Laboratory, *Psychological Review*, 65(6), pp. 386–408.

Salamov, A.A., and Solovyev, V. (1997). Recognition of 3′-processing sites of human mRNA precursors. *Computer Applications in the Biosciences*, 13, 23-28.

Salamov, A.A., Nishikawa, T., and Swindells, M.B. (1998). Assessing Protein Coding Region Integrity in cDNA Sequencing Projects. *Bioinformatics*, 14(5), 384-390.

Savasere, A., Omiecinski, E, and Navathe, S.B. (1998). Mining for Strong Negative Associations in a Large Database of Customer Transactions. In *Proceedings of the 14th International Conference on Data Engineering*. 494-502.

Schena, M, Shalon, D, Davis, R.W., and Brown, P.O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* 270, 467-470.

Schuler, G.D. (1998). Sequence Alignment and Database Searching, In A.D. Baxevanis and B.F.F. Ouellette (Eds.), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 145-171, Wiley-Liss, Inc.

Sheets, M.D., Ogg, S.C., and Wickens, M.P. (1990). Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Research*. 18(19): 5799-805.

Shi, L. (2001). DNA Microarray (Genome Chip), 2002. URL: http://www.gene-chips.com (προσπελάστηκε: 17/05/2011).

Smith, T.F. and Waterman, M.S. (1981). Identification of Common Molecular Subsequences, *Journal of Molecular Biology*, 147(1), 195-197.

Smolkin, M. and Ghosh, D. (2003). Cluster Stability Scores for Microarray Data in Cancer Studies, *BMC Bioinformatics*, 4(36).

Srikant, R. and Agrawal, R. (1995). Mining Generalized Association Rules. In *Proceedings of the 21st Conference on Very Large Databases*. 407-419.

Stormo, G. D., Schneider, T. D., Gold, L., and Ehre, A. (1982). Use of the 'Perceptron' Algorithm to Distinguish Translational Initiation Sites in E. coli. *Nucleic Acids Research*, 10(9), 2997-3011.

Sumner, M., Frank, E., and Hall, M. (2005). Speeding up Logistic Model Tree Induction. In *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 675-683.

Teng, C.M. (2002). Learning form Dissociations. In *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery*. 11-20.

The Gene Ontology Consortium. (2000). Gene ontology: tool for the unification of biology. Nature Genetics, 25(1): 25-29.

Thomas, S. and Sarawagi, S. (1998). Mining Generalized Association Rules and Sequential Patterns Using SQL Queries. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*. 344-348.

Tian, B., Hu, J., Zhang, H., and Lutz, C.S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*, 33: 201-212.

Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botststein, D., and Brown, P. (1999). Clustering Methods for the Analysis of DNA Microarray Data. Technical Report, Stanford University.

Ting, M.T. and Witten, I.H. (1999). Issues in Stacked Generalization, Journal of Artificial Intelligence Research, vol. 10, 271-289.

Tumer, K. and Ghosh, J. (1996). Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8(34): 385–403.

Tung, A. K. H., Lu, H., Han, J., and Feng, L. (2003). Efficient Mining of Intertransaction Association Rules. *IEEE Transactions On Knowledge And Data Engineering*. 15(1), 43-56.

Tzanis, G. and Berberidis, C. (2007). Mining for Mutually Exclusive Items in Transaction Databases. *International Journal of Data Warehousing and Mining*, Idea Group Publishing, 3(3), 45-59.

Tzanis, G., Berberidis, C., Alexandridou, A., and Vlahavas, I. (2005a). Improving the Accuracy of Classifiers for the Prediction of Translation Initiation Sites in Genomic Sequences. In *Proceedings of the 10th Panhellenic Conference on Informatics (PCI'2005)*, Volos, Greece, 426–436.

Tzanis, G., Berberidis, C., and Vlahavas, I. (2005b). Biological Data Mining, *Encyclopedia of Database Technologies and Applications*, L.C. Rivero, J.H. Doorn, and V.E. Ferraggine (Eds.), IDEA Group Publishing.

Tzanis, G., Berberidis, C., and Vlahavas, I. (2006a). On the Discovery of Mutually Exclusive Items in a Market Basket Database, In *Proceedings of the 2nd ADBIS Workshop on Data Mining and Knowledge Discovery*, Thessaloniki, Greece, pp. 1-12.

Tzanis, G., Berberidis, C., and Vlahavas, I. (2006b). A Novel Data Mining Approach for the Accurate Prediction of Translation Initiation Sites, In Proceedings of the 7th International Symposium on Biological and Medical Data Analysis, Springer-Verlag, LNCS 4345, 92-103, Thessaloniki, Greece.

Tzanis, G. Berberidis, C., and Vlahavas, I. (2007). MANTIS: A Data Mining Methodology for Effective Translation Initiation Site Prediction. In *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, Lyon, France, 6343-6347.

Tzanis, G., Berberidis, C., and Vlahavas, I. (2009). Machine Learning and Data Mining in Bioinformatics. *Handbook of Research on Innovations in Database Technologies and Applications: Current and Future Trends*, Ferraggine, V.E., Doorn, J.H., Rivero, L.C. (Eds.), IGI Global.

Tzanis, G., Katakis, I., Partalas, I., and Vlahavas, I. (2006c). Modern Applications of Machine Learning, In *Proceedings of the 1st Annual SEERC Doctoral Student Conference*, Thessaloniki, Greece.

Tzanis, G., Kavakiotis, I., Vlahavas, I. (2008). Polyadenylation Site Prediction Using Interesting Emerging Patterns, In *Proceedings of the 8th IEEE International Conference on Bioinformatics and Bioengineering*, IEEE, Athens, Greece, 1-7.

Tzanis, G., Kavakiotis, I., Vlahavas, I. (2011). PolyA-iEP: A Data Mining Method for the Effective Prediction of Polyadenylation Sites, *Expert Systems with Applications*, Elsevier, 38(10): 12398-12408.

Tzanis, G. and Vlahavas, I. (2006). Prediction of Translation Initiation Sites Using Classifier Selection. In *Proceedings of the 4th Hellenic Conference on Artificial Intelligence (SETN'06)*, G. Antoniou, G. Potamias, D. Plexousakis, C. Spyropoulos (Eds.), Springer-Verlag, LNAI 3955, Heraklion, Greece, 367-377.

Tzanis, G. and Vlahavas, I. (2007a). Mining High Quality Clusters of SAGE Data, In *Proceedings of the 2nd VLDB Workshop on Data Mining in Bioinformatics*, Vienna, Austria, September 23, 2007.

Tzanis, G. and Vlahavas, I. (2007b). Accurate Classification of SAGE Data Based on Frequent Patterns of Gene Expression. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, Patras, Greece, 96-100.

Tzanis, G. and Vlahavas, I. (2010). Mining for Mutually Exclusive Gene Expressions, In *Proceedings of the 6th Hellenic Conference on Artificial Intelligence*, Springer-Verlag, Athens, Greece, 255-264.

Van Helden, J., Del Olmo, M., and Perez-Ortin, J.E. (2000). Statistical analysis of yeast genomic downsream sequences reveals putative polyadenylation signals. *Nucleic Acids Research*, 28(4), 1000-1010.

Velculescu, V.E., Zhang, L. Vogelstein, B., and Kinzler, K.W. (1995). Serial Analysis of Gene Expression, *Science*, 270 (5235), 484-487.

Whishart, D.S. (2002). Tools for Protein Technologies. In C.W. Sensen (ed.), *Biotechnology: Genomics and Bioinformatics*, (Vol. 5b), 325-344, Wiley-VCH.

Whittaker, R.H. (1969). New Concepts of Kingdoms of Organisms. *Science*, 163, 150–160.

Witten, I.H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87, 4576-4579.

Wolpert, D.H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241–259.

Wu, X., Zhang, C., and Zhang, S. (2004). Efficient Mining of both Positive and Negative Association Rules. *ACM Transactions on Information Systems*. 22(3), 381-405.

Xing, E.P., Jordan, M.I., and Karp, R.M. (2001). Feature Selection for High-Dimensional Genomic Microarray Data. In *Proceedings of the 18th International Conference on Machine Learning*, 601-608.

Yan, J. and Marr, T.G. (2005). Computational analysis of 3′-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Research*, 15: 369-375.

Yang, Q. and Wu, X. (2006). 10 Challenging Problems in Data Mining Research, *International Journal of Information Technology & Decision Making*, 5(4), 497-604.

Yeung, K.Y., Medvedovic, M., and Bumgarner, R.E. (2003). Clustering Gene-Expression Data with Repeated Measurements, *Genome Biology*, 4(5), R34.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proccedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002)*, Edmonton, Alberta, Canada, pp. 694-699.

Zagoruiko, N., Borisova, I., Kiselev, A., Kutnenko, O. (2005). Algorithm GRAD for selection of informative genetic characteristics. In *Proceedings of the 2nd Moscow Conference on Computational Molecular Biology*, Moscow, Russia.

Zeng F., Yap H., and Wong, L. (2002). Using Feature Generation and Feature Selection for Accurate Prediction of Translation Initiation Sites. In Proceedings of the 13th International Conference on Genome Informatics, Tokyo, Japan, 192-200.

Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T., and Muller, R.-K. (2000). Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites, *Bioinformatics*, 16(9), 799-807.

Αλεπόρου-Μαρίνου, Β., Αργυροκαστρίτης, Α., Κομητοπούλου, Α., Πιαλόγλου, Π., και Σγουρίτσα, Β. (2003). *Βιολογία Θετικής Κατεύθυνσης Γ΄ Τάξης Ε.Λ.*, Ο.Ε.Δ.Β.

Βαζιριγιάννης, Μ. και Χαλκίδη, Μ. (2003). *Εξόρυξη Γνώσης από Βάσεις Δεδομένων*, Εκδόσεις Τυπωθήτω – Γ. Δαρδάνος, Αθήνα.

Βλαχάβας, Ι., Κεφαλάς, Π., Βασιλειάδης, Ν., Κόκκορας, Φ., και Σακελλαρίου, Η. (2006). *Τεχνητή Νοημοσύνη* (3η Έκδοση), Εκδόσεις Β. Γκιούρδας Εκδοτική, Θεσσαλονίκη.

Καψάλης, Α. Μπουρμπουχάκης, Ι., Περάκης, Β., και Σαλαμαστράκης, Σ. (2003). *Βιολογία Γενικής Παιδείας Β΄ Τάξης Ε.Λ.*, Ο.Ε.Δ.Β.

Μπίκης, Κ., Μακρής, Χ., και Κοσσίδα, Σ. *Ηλεκτρονική Διαχείριση Βιολογικών Δεδομένων*. Στο Κοσσίδα, Σ. (2008). Βιοπληροφορική: Δυνατότητες και Προοπτικές, ΙΙΒΕΑΑ, Ακαδημία Αθηνών, σελίδες 33-52.

Λίστα Δημοσιεύσεων

Άρθρα σε Περιοδικά

1. Tzanis, G. and Berberidis, C. (2007). Mining for Mutually Exclusive Items in Transaction Databases. *International Journal of Data Warehousing and Mining*, Idea Group Publishing, 3(3): 45-59.

2. Hatzikos, E., Tsoumakas, G., Tzanis, G., Bassiliades, N., and Vlahavas, I. (2008). An Empirical Study of Sea Water Quality Prediction, *Knowledge-Based Systems*, Elsevier, 21(6): 471-478.

3. Tzanis, G., Kavakiotis, I., Vlahavas, I. (2011). PolyA-iEP: A Data Mining Method for the Effective Prediction of Polyadenylation Sites, *Expert Systems with Applications*, Elsevier, 38(10): 12398-12408.

4. Tzanis, G. Berberidis, C., and Vlahavas, I. StackTIS: A Stacked Generalization Approach for Effective Prediction of Translation Initiation Sites, *Computers in Biology and Medicine*, Elsevier (υποβληθέν).

Άρθρα σε Πρακτικά Συνεδρίων

1. Berberidis, C., Tzanis, G., and Vlahavas, I. (2005). Mining for Contiguous Frequent Itemsets in Transaction Databases, In *Proceedings of the IEEE 3rd International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, Sofia, Bulgaria, 679-685.

2. Tzanis, G., Berberidis, C., Alexandridou, A., and Vlahavas, I. (2005). Improving the Accuracy of Classifiers for the Prediction of Translation Initiation Sites in Genomic Sequences. In *Proceedings of the 10th Panhellenic Conference on Informatics (PCI'2005)*, Volos, Greece, 426–436.

3. Tzanis, G. and Vlahavas, I. (2006). Prediction of Translation Initiation Sites Using Classifier Selection. In *Proceedings of the 4th Hellenic Conference on Artificial Intelligence (SETN'06)*, G. Antoniou, G. Potamias, D. Plexousakis, C. Spyropoulos (Eds.), Springer-Verlag, LNAI 3955, Heraklion, Greece, 367-377.

4. Tzanis, G., Katakis, I., Partalas, I., and Vlahavas, I. (2006). Modern Applications of Machine Learning, In *Proceedings of the 1st Annual SEERC Doctoral Student Conference*, Thessaloniki, Greece.

5. Tzanis, G., Berberidis, C., and Vlahavas, I. (2006). On the Discovery of Mutually Exclusive Items in a Market Basket Database, In *Proceedings of the 2nd ADBIS Workshop on Data Mining and Knowledge Discovery*, Thessaloniki, Greece, 1-12.

6. Tzanis, G., Berberidis, C., and Vlahavas, I. (2006). A Novel Data Mining Approach for the Accurate Prediction of Translation Initiation Sites, In Proceedings of the

7th International Symposium on Biological and Medical Data Analysis, Thessaloniki, Greece, 92-103.

7. Tzanis, G. Berberidis, C., and Vlahavas, I. (2007). MANTIS: A Data Mining Methodology for Effective Translation Initiation Site Prediction. In *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, Lyon, France, 6343-6347.

8. Tzanis, G. and Vlahavas, I. (2007). Mining High Quality Clusters of SAGE Data, In *Proceedings of the 2nd VLDB Workshop on Data Mining in Bioinformatics*, Vienna, Austria.

9. Tzanis, G. and Vlahavas, I. (2007). Accurate Classification of SAGE Data Based on Frequent Patterns of Gene Expression. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, Patras, Greece, 96-100.

10. Tzanis, G., Kavakiotis, I., Vlahavas, I. (2008). Polyadenylation Site Prediction Using Interesting Emerging Patterns, In *Proceedings of the 8th IEEE International Conference on Bioinformatics and Bioengineering*, IEEE, Athens, Greece, 1-7.

11. Tzanis, G. and Vlahavas, I. (2010). Mining for Mutually Exclusive Gene Expressions, In *Proceedings of the 6th Hellenic Conference on Artificial Intelligence*, Springer-Verlag, Athens, Greece, 255-264.

Κεφάλαια σε Συλλογικούς Τόμους

1. Tzanis, G., Berberidis, C., and Vlahavas, I. (2005). Biological Data Mining, *Encyclopedia of Database Technologies and Applications*, L.C. Rivero, J.H. Doorn, and V.E. Ferraggine (Eds.), IDEA Group Publishing.

2. Tzanis, G., Berberidis, C., and Vlahavas, I. (2009). Machine Learning and Data Mining in Bioinformatics. *Handbook of Research on Innovations in Database Technologies and Applications: Current and Future Trends*, Ferraggine, V.E., Doorn, J.H., Rivero, L.C. (Eds.), IGI Global.

Λίστα Ετεροαναφορών

Σύνολο Ετεροαναφορών: 52

Tzanis, G., Berberidis, C., and Vlahavas, I. (2005). Biological Data Mining, *Encyclopedia of Database Technologies and Applications*, L.C. Rivero, J.H. Doorn, and V.E. Ferraggine (Eds.), IDEA Group Publishing.

1. Khorsi, A. (2006). Knowledge Representation, A.I. applications and Data-mining. In *Proceedings of Conférence Internationale sur l'Informatique et ses Applications*, Algeria, 2006.
2. Kianmehr, K., Alshalalfa, M., and Alhajj, R. (2009). Fuzzy clustering-based discretization for gene expression classification, *Knowledge and Information Systems*, 24: 441-465.
3. Leon, F., Aignatoaiei, B.I., and Zaharia, M.H. (2009). Performance Analysis of Algorithms for Protein Structure Classification. In *Proceedings of the 20th International Workshop on Database and Expert Systems Application*, 203-207.
4. Kaur, A., Kaur, M., and Kaur, R.K. (2010). Using Cluster Analysis for Protein Secondary Structure Prediction. *International Journal of Computer Applications* 4(12): 20–22.
5. Salle, P. (2010). *Les motifs séquentiels pour les données issues des puces ADN*, PhD Thesis, Université Montpellier II, 2010.
6. Singh, R., Deol, S.K., and Sandhu, P.S. (2010). Chou-Fasman Method for Protein Structure Prediction using Cluster Analysis. In *Proceedings of World Academy of Science, Engineering and Technology* 72: 982-987.
7. Zheng, G., Jiang, M., He, X. Zhao, J., Guo, H., Chen, G., Zha, Q., and Lu, A. (2011). Discrete Derivative: A Data Slicing Algorithm for Exploration of Sharing Biological Networks between Rheumatoid Arthritis and Coronary Heart Disease. *BioData Mining*, 4:18.

Berberidis, C., Tzanis, G., and Vlahavas, I. (2005). Mining for Contiguous Frequent Itemsets in Transaction Databases, In *Proceedings of the IEEE 3rd International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, Sofia, Bulgaria, 679-685.

1. Aljandal, W.A. (2008). *Itemset size-sensitive interestingness measures for association rule mining and link prediction*. PhD Thesis, Department of Computing and Information Sciences, College of Engineering, Kansas State University, 2008.

2. Duemong, F. (2009). *AMFIST: An Algorithm for Mining Frequent Itemsets based on the Similarities of Transactions*. Master Thesis, Department of Computer Science, Prince of Songkla University.

Tzanis, G., Berberidis, C., Alexandridou, A., and Vlahavas, I. (2005). Improving the Accuracy of Classifiers for the Prediction of Translation Initiation Sites in Genomic Sequences. In *Proceedings of the 10th Panhellenic Conference on Informatics (PCI'2005)*, Volos, Greece, 426–436.

1. Lee, J., Blain, S., Casas, M., Kenny, D., Berall, G., and Chau, T. (2006). A Radial Basis Classifier for the Automatic Detection of Aspiration in Children with Dysphagia. *Journal of NeuroEngineering and Rehabilitation*, 3(14).
2. Ma, C., Zhou, D., and Zhou, Y. (2006). Feature Mining and Integration for Improving the Prediction Accuracy of Translation Initiation Sites in Eukaryotic mRNAs. In *Proceedings of the 5th International Conference on Grid and Cooperative Computing Workshops*, IEEE Computer Society, 349-356.
3. Nobre, C.N. (2007). *Desenvolvimento de uma Metodologia para Previsao de Sitios de Inicio de Traducao*. Master Thesis, Universidade Federal de Minas Gerais.
4. Nobre, C.N., Ortega, J.M., de Padua Braga, A. (2007). High Efficiency on Prediction of Translation Initiation Site (TIS) of RefSeq Sequences. In *Proceedings of the 2nd Brazilian Symposium on Bioinformatics (BSB 2007)*, 138-148, 2007.
5. Yuntao, Z. and Baomiao, D. (2007). Application of RS-GA-SVM method in the recognition of translation initiation sites of eukaryotic DNAs. *Computers and Applied Chemistry*, 24(6): 794-798.
6. Gao, T., Tian, Y., Shao, X., and Deng, N. (2008). Accurate Prediction of Translation Initiation Sites by Universum SVM. In *Proceedings of the 2nd International Symposium on Optimization and Systems Biology*, 279–286, China.
7. Kongmanee, T. (2008). *Genome Sequences Analysis Using Neural Networks*. Master Thesis. Prince of Songkla University.
8. Lee, J. (2009). *Investigation of Accelerometry, Mechanomyography, and Nasal Airflow Signals for Abnormal Swallow Detection*, PhD Thesis, Department of Electrical and Computer Engineering University of Toronto.
9. Lee, J. Steele, C.M., and Chau, T. (2009). Swallow segmentation with artificial neural networks and multi-sensor fusion, *Medical Engineering & Physics*, 2009.
10. Wettayaprasit, W., Kongmanee, T., and Vanichayobon, S. (2009). The Efficiency of n-gram and TF-IDF Techniques for Translation Initiation Site Prediction. In *Proceedings of the Conference on Knowledge and Smart Technologies*, 2009.

11. Nhan, B.R. and Chau, T. (2010). Classifying Affective States Using Thermal Infrared Imaging of the Human Face. *IEEE Transactions on Biomedical Engineering,* 57(4): 679-987.

12. Wang, J. and Sun, Y.F. (2010). Recognition of Alternative Splice Sites Based on Discrete Cosine Transform Learned by SVM With the Mixed Pseudo-image Coding Method, In *Proceedings of the the 4th IEEE International Conference on Bioinformatics and Biomedical Engineering*, 2010.

13. Lee, J., Steele, C.M., and Cha, T. (2011). Classification of healthy and abnormal swallows based on accelerometry and nasal airflow signals. *Artificial Intelligence in Medicine*, Elsevier.

Tzanis, G. and Vlahavas, I. (2006). Prediction of Translation Initiation Sites Using Classifier Selection. In *Proceedings of the 4th Hellenic Conference on Artificial Intelligence (SETN'06)*, G. Antoniou, G. Potamias, D. Plexousakis, C. Spyropoulos (Eds.), Springer-Verlag, LNAI 3955, Heraklion, Greece, 367-377.

1. Selpi, Bryant, C.H., Kemp, G.J.L., Cvijovic, M. (2006). A First Step towards Learning which uORFs Regulate Gene Expression. *Journal of Integrative Bioinformatics*, 3(2), 31-45.

2. Nobre, C.N. (2007). *Desenvolvimento de uma Metodologia para Previsao de Sitios de Inicio de Traducao*. Master Thesis, Universidade Federal de Minas Gerais.

3. Nobre, C.N., Ortega, J.M., de Padua Braga, A. (2007). High Efficiency on Prediction of Translation Initiation Site (TIS) of RefSeq Sequences. In *Proceedings of the 2nd Brazilian Symposium on Bioinformatics (BSB 2007)*, 138-148, 2007.

4. Kongmanee, T. (2008). *Genome Sequences Analysis Using Neural Networks*. Master Thesis. Prince of Songkla University.

5. Tikole, S. and Sankararamakrishnan, R. (2008). Prediction of Translation Initiation Sites in Human mRNA Sequences with AUG Start Codon in Weak Kozak Context: A Neural Network Approach. *Biochemical and Biophysical Research Communications*. 369(4), 1166-1168, 2008.

6. Zeng, J. and Alhajj, R. (2008). Predicting translation initiation sites using a multi-agent architecture empowered with reinforcement learning. In *Proceedings of the IEEE Computational Intelligence in Bioinformatics and Computational Biology*, 241-248.

7. Zeng, J. and Alhajj, R., and Demetrick, D. (2008). Effectiveness of Applying Codon Usage Bias for Translational Initiation Sites Prediction. In *Proceedings of the IEEE International Conference of Bioinformatics and Biomedicine*, 121-126.

8. Zeng, J. and Alhajj, R. (2009). Integrating Swarm Intelligent Algorithms for Translation Initiation Sites Prediction. In *Innovations in Swarm Intelligence*, Studies in Cmputational Intelligence, Springer, 2009.

Tzanis, G., Berberidis, C., and Vlahavas, I. (2006). On the Discovery of Mutually Exclusive Items in a Market Basket Database, In *Proceedings of the 2nd ADBIS Workshop on Data Mining and Knowledge Discovery*, Thessaloniki, Greece, 1-12.

1. Aljandal, W.A. (2008). *Itemset size-sensitive interestingness measures for association rule mining and link prediction*. PhD Thesis, Department of Computing and Information Sciences, College of Engineering, Kansas State University.
2. Σιντόρης, Κ. (2008). *Διερεύνηση Πρακτικών Τεχνολογίας Λογισμικού στην Ελληνική Βιομηχανία Λογισμικού με Στατιστικές Μεθόδους και Τεχνικές Εξόρυξης Δεδομένων*. Master Thesis. Department of Informatics, University of Macedonia.

Tzanis, G., Berberidis, C., and Vlahavas, I. (2006). A Novel Data Mining Approach for the Accurate Prediction of Translation Initiation Sites, In Proceedings of the 7th International Symposium on Biological and Medical Data Analysis, Thessaloniki, Greece, 92-103.

1. Wang, J., and Sun, Y.F. (2010). Recognition of Alternative Splice Sites Based on Discrete Cosine Transform Learned by SVM With the Mixed Pseudo-image Coding Method. In *Proceedings of the the 4th IEEE International Conference on Bioinformatics and Biomedical Engineering*, 1-4.
2. Nobre, C.N. (2007). *Desenvolvimento de uma Metodologia para Previsao de Sitios de Inicio de Traducao*. Master Thesis, Universidade Federal de Minas Gerais.
3. Nobre, C.N., Ortega, J.M., de Padua Braga, A. (2007). High Efficiency on Prediction of Translation Initiation Site (TIS) of RefSeq Sequences. In *Proceedings of the 2nd Brazilian Symposium on Bioinformatics (BSB 2007)*, 138-148, 2007.
4. Kongmanee, T. (2008). *Genome Sequences Analysis Using Neural Networks*. Master Thesis. Prince of Songkla University.

Tzanis, G. and Berberidis, C. (2007). Mining for Mutually Exclusive Items in Transaction Databases. *International Journal of Data Warehousing and Mining*, Idea Group Publishing, 3(3): 45-59.

1. Gorla, N., and Betty, P.W.Y. (2008). Vertical Fragmentation in Databases Using Data-Mining Technique. *International Journal of Data Warehousing and Mining*, Idea Group Publishing 4(3): 35-53.
2. Hamrouni, T., Yahia, S.B., and Nguifo, E.M. (2008). GARM: Generalized Association Rule Mining. In *Proceedings of the 6th Conference on Concept Lattices and Their Applications*, Olomouc, Czech Republic, 145-156.

3. Hamrouni, T., Yahia, S.B., and Nguifo, E.M. (2010). Generalization of association rules through disjunction. *Annals of Mathematics and Artificial Intelligence*, Springer, 59(2): 201-222.

4. Hamrouni, T., Yahia, S.B., and Nguifo, E.M. (2009). Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets. *Data and Knowledge Engineering*, 68(10): 1091-1111.

Tzanis, G. Berberidis, C., and Vlahavas, I. (2007). MANTIS: A Data Mining Methodology for Effective Translation Initiation Site Prediction. In *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, Lyon, France, 6343-6347.

1. Kongmanee, T. (2008). *Genome Sequences Analysis Using Neural Networks*. Master Thesis. Prince of Songkla University.

2. Kongmanee, T., Vanichayobon, S., and Wettayaprasit, W. (2009). The TF-IDF and Neural Networks Approach for Translation Initiation Site Prediction. In *Proceedings of the 2nd IEEE International Conference on Computer Science and Information Technology*, IEEE, 318-322.

3. Wettayaprasit, W., Kongmanee, T., and Vanichayobon, S. (2009). The Efficiency of n-gram and TF-IDF Techniques for Translation Initiation Site Prediction. In *Proceedings of the Conference on Knowledge and Smart Technologies*, 2009.

Tzanis, G. and Vlahavas, I. (2007). Mining High Quality Clusters of SAGE Data, In *Proceedings of the 2nd VLDB Workshop on Data Mining in Bioinformatics*, Vienna, Austria.

1. Tang, D.M., Zhua, Q.X., and Yanga, F. (2010). A Poisson-based adaptive affinity propagation clustering for SAGE data. *Computational Biology and Chemistry*. 34(1): 63-70.

Tzanis, G. and Vlahavas, I. (2007). Accurate Classification of SAGE Data Based on Frequent Patterns of Gene Expression. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, Patras, Greece, 96-100.

1. Yang, C.-H., Shih, T.-M., and Chuang, L.-Y. (2009). Reducing SAGE Data Using Genetic Algorithms, In *Proceedings of World Academy of Science, Engineering and Technology*, 41.

Hatzikos, E., Tsoumakas, G., Tzanis, G., Bassiliades, N., and Vlahavas, I. (2008). An Empirical Study of Sea Water Quality Prediction, *Knowledge-Based Systems*, Elsevier, 21(6): 471-478.

1. Kotis, K., Papasalouros, A., and Nikitakos, N. (2009). Supporting Decision Making in Maritime Environmental Protection with a Knowledge-based Education and

Awareness Approach, In *Proceedings of AIAI'09 Workshop on Artificial Intelligence Applications in Environmental Protection*, Thessaloniki, Greece, 35-45.

2. El-Mahrouk, M.E., El-Nady, M.F., and Hegazi, M.A. (2010). Effect of Diluted Seawater Irrigation and Exogenous Proline Treatments on Growth, Chemical Composition and Anatomical Characteristics of Conocarpus Erectus L. *Journal of Agricultural Research Kafr El Sheikh University*, 36(4).
3. Li, F., Li, D., Wei, Y., Daokun, M., and Ding, Q. (2010). Dissolved oxygen prediction in apostichopus japonicus aquaculture ponds by BP neural network and AR model. *Sensor Letters*, 8(1): 95-101.
4. Papasalouros, A., Kotis, K., and Nikitakos, N. (2010). Towards an Intelligent Tutoring System for Environmental Decision Makers, *Environmental Engineering and Management Journal*, 9(2): 197-204.
5. Xuan, W., Jiake, L., Deti, X. (2010). A hybrid approach of support vector machine with particle swarm optimization for water quality prediction. In *Proceedings of the 5th International Conference on Computer Science and Education (ICCSE)*, 1158-1163.
6. Martinez-Alvarez, F., Troncoso, A. Morales-Esteban, A., and Riquelme, J.C. (2011). Computational Intelligence Techniques for Predicting Earthquakes. In *Proceedings of the International Conference on Hybrid Artificial Intelligent Systems (HAIS'11)*, LNAI 6679, 287-294.

Tzanis, G., Kavakiotis, I., Vlahavas, I. (2008). Polyadenylation Site Prediction Using Interesting Emerging Patterns, In *Proceedings of the 8th IEEE International Conference on Bioinformatics and Bioengineering*, IEEE, Athens, Greece, 1-7.

1. Liu, S.-K. (2010). *Prediction of mRNA polyadenylation sites in human genes by nucleotide composition*. Master Thesis, Department of Computer Science and Information Engineering, National Central University, Taiwan.