ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Αναγνώριση Διπλώματος Πρωτεϊνών με χρήση Δικτύων Ομοιότητας και Γενετικών Αλγορίθμων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΜΥΡΣΙΝΗΣ Α. ΚΑΦΩΡΟΥ

Επιβλέπουσα : Κωνσταντίνα Νικήτα
Καθηγήτρια Ε.Μ.Π.

Αθήνα, Ιούλιος 2009

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Αναγνώριση Διπλώματος Πρωτεϊνών με χρήση Δικτύων Ακολουθιών και Γενετικών Αλγορίθμων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΜΥΡΣΙΝΗΣ Α. ΚΑΦΩΡΟΥ

Επιβλέπουσα : Κωνσταντίνα Νικήτα
Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 20η Ιουλίου 2009.

.....	.....	.....
Κωνσταντίνα Νικήτα	Νικόλαος Ουζούνογλου	Δημήτρης Κουτσούρης
Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2009

.....

ΜΥΡΣΙΝΗ Α. ΚΑΦΩΡΟΥ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

© 2009 – All rights reserved

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την Καθηγήτρια κ. Κωνσταντίνα Νικήτα για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα και για την επίβλεψη της παρούσας διπλωματικής εργασίας. Επίσης, θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα Γιάννη Βαλαβάνη για την πολύτιμη βοήθειά του στην εκπόνηση της εργασίας και τις καίριες συμβουλές του. Τέλος, θα ήθελα να πω ένα μεγάλο ευχαριστώ στην οικογένεια μου και σε όλους εκείνους που που ήταν δίπλα μου όλα αυτά τα χρόνια.

Περίληψη

Η παρούσα διπλωματική εργασία προσεγγίζει το πρόβλημα της αναγνώρισης διπλώματος πρωτεϊνών (fold recognition), δηλαδή της κατηγοριοποίησης πρωτεϊνικών ακολουθιών σε κατάλληλες δομικές κατηγορίες με βάση χαρακτηριστικά που εξάγονται από την ακολουθία. Το πρόβλημα της αναγνώρισης διπλώματος πρωτεϊνών αντιμετωπίζεται εδώ ως πρόβλημα ταξινόμησης για τη λύση του οποίου εφαρμόζονται εξελιγμένες μέθοδοι υπολογιστικής νοημοσύνης. Γίνεται χρήση ενός συνόλου 311 πρωτεϊνικών ακολουθιών που κατανέμονται σε 27 γνωστά διπλώματα. Κάθε πρωτεϊνική ακολουθία του διαθέσιμου συνόλου δεδομένων χαρακτηρίζεται από 125 χαρακτηριστικά ακολουθίας που σχετίζονται με την αμινοξική σύσταση της καθώς και την υδροφοβικότητα, την πόλωση, την πολωσιμότητα, τον όγκο Van der Waals, και την προβλεπόμενη δευτεροταγή δομή των καταλοίπων της. Η μεθοδολογία που αναπτύσσεται συνίσταται i) στην ανάπτυξη κατάλληλου ταξινομητή βασισμένου σε δίκτυα ομοιότητας πρωτεϊνών που υπολογίζει τη συγγένεια μιας πρωτεϊνικής ακολουθίας με ακολουθίες γνωστού διπλώματος με βάση τα χρησιμοποιούμενα χαρακτηριστικά ii) εφαρμογή γενετικού αλγορίθμου για την επιλογή των πιο εύρωστων χαρακτηριστικών από το αρχικό σύνολο των 125 χαρακτηριστικών ακολουθίας. Η χρήση των 125 χαρακτηριστικών ακολουθίας από τον ταξινομητή οδήγησε σε μέση ορθότητα ίση περίπου με 32% σε σύνολα ελέγχου που προέκυψαν κατά τη διασταυρωμένη επικύρωση από 10 μέρη. Η απόδοση του αλγορίθμου είναι συγκρίσιμη με την απόδοση ταξινομητών μεγαλύτερης πολυπλοκότητας που άλλοι ερευνητές χρησιμοποίησαν στο ίδιο πρόβλημα. Η εφαρμογή του γενετικού αλγορίθμου μείωσε τη διάσταση του χρησιμοποιούμενου συνόλου χαρακτηριστικών και αύξησε τη μέση ορθότητα του ταξινομητή κατά 5%.

Λέξεις κλειδιά: Βιοπληροφορική, δομική βιοπληροφορική, πρωτεϊνομική, δίκτυα ομοιότητας πρωτεϊνών, αναγνώριση διπλώματος, γενετικός αλγόριθμος, ταξινομητής, χαρακτηριστικά ακολουθίας

Abstract

The present thesis addresses the task of protein fold recognition, that is the categorization of protein sequences into appropriate structural classes based on sequence-derived features. We regard the problem as a classification one and we employ methods from the field of computational intelligence. We used a dataset of 311 protein sequences that belong to 27 folds. Every protein sequence of the dataset is characterized by 125 sequence-derived features that have been used widely in the literature, and are related with amino-acid composition, polarity, polarizability, normalized Van der Waals volume, hydrophobicity and the predicted structure of its residues. Our methodology consists of two tasks. First we developed a convenient classifier based on protein similarity networks that measures the affinities of a query protein with all other sequences of known fold. The query protein is assigned the fold of the sequence with the greatest affinity to it, with respect to the sequence-derived features. Then we used a genetic algorithm for the selection of the most robust features among the 125 of the initial dataset.

The initial classifier led to an average total accuracy of approximately 32%. In order to evaluate the accuracy of the method we used the resampling technique of 10-fold cross validation. The accuracy of our algorithm is comparable with that of more complex classifiers, developed by researchers to confront the same problem. The application of the genetic algorithm decreased the dimension of the set of the sequence-derived features and improved the accuracy of the classifier up to 5%.

Keywords: Bioinformatics, structural Bioinformatics, proteomics, fold recognition, protein similarity networks, classifier, genetic algorithm, sequence-derived features

Περιεχόμενα

Αναγνώριση Διπλώματος Πρωτεϊνών με χρήση Δικτύων Ομοιότητας και Γενετικών Αλγορίθμων **1**

Αναγνώριση Διπλώματος Πρωτεϊνών με χρήση Δικτύων Ακολουθιών και Γενετικών Αλγορίθμων **3**

Περιεχόμενα ..... **i**

Κατάλογος Σχημάτων ..... **iii**

Κατάλογος Πινάκων ..... **iv**

1 Εισαγωγή **1**

1.1 Βιοπληροφορική 1

1.2 Αντικείμενο Διπλωματικής 4

1.2.1 Συνεισφορά *5*

1.3 Οργάνωση κειμένου 6

2 Θεωρητικό υπόβαθρο **7**

2.1 Ο κώδικας της ζωής 7

2.2 Πρωτεΐνες 12

2.2.1 Σύσταση *12*

2.2.2 Ρόλος *15*

2.2.3 Πρωτεϊνική δομή *17*

2.3 Έννοιες από την επιστήμη των υπολογιστών 23

2.3.1 Ταξινομητές *23*

2.3.2 Γράφοι - Βιολογικά Δίκτυα *25*

2.3.3 Γενετικοί Αλγόριθμοι *27*

3 Αναγνώριση Διπλώματος Πρωτεϊνών **31**

3.1 Πρόβλεψη Δομής Πρωτεϊνών 31

3.2 Κατηγοριοποίηση Υπερ-δευτεροταγών δομών 33
3.2.1 SCOP (Structural Classification Of Proteins database) *33*
3.2.2 CATH *34*
3.3 Μέθοδοι Πρόγνωσης Διπλώματος Πρωτεϊνών 36
3.3.1 Μέθοδοι προτυποποίησης με ομολογία (homology modeling) *36*
3.3.2 Μέθοδοι ύφανσης πρωτεϊνών (protein threading) *37*
3.3.3 Μέθοδοι απαρχής (ab initio) *38*
3.3.4 Συμπεράσματα *38*
4 Υλοποίηση Ταξινομητή 41
4.1 Εισαγωγή 41
4.2 Υλοποίηση συστήματος 44
4.2.1 Κατασκευή Δικτύου Ομοιότητας Ακολουθιών *45*
4.2.2 Επαναληπτικός Αλγόριθμος *45*
4.3 Πειραματική Αξιολόγηση 47
4.3.1 Διαφορετικά περιθώρια λάθους *48*
4.3.2 Διαφορετικές ομάδες χαρακτηριστικών *50*
4.3.3 Σύγκριση επιτυχίας στα διαφορετικά διπλώματα *52*
5 Βελτιστοποίηση με χρήση Γενετικού Αλγορίθμου 57
5.1 Βασικά στοιχεία Γενετικού Αλγόριθμου 57
5.2 Κατασκευή δίκτυου ομοιότητας δομών 59
5.3 Συνάρτηση Καταλληλότητας 59
5.4 Αποτελέσματα και αξιολόγηση χρήσης ΓΑ 60
6 Επίλογος 63
6.1 Σύνοψη 63
6.2 Συμπεράσματα 64
7 Βιβλιογραφία 65
Παράρτημα 67

Κατάλογος Σχημάτων

Εικόνα 1.1 Βιοπληροφορική : σύμπραξη δύο επιστημονικών πεδίων 3

Εικόνα 2.1 Οι βάσεις του νουκλεϊκών οξέων και οι δεσμοί που αναπτύσσονται μεταξύ τους..9

Εικόνα 2.2 Διαδικασία πρωτεϊνόσύνθεσης σε ένα ριβόσωμα. 11

Εικόνα 2.3 Κεντρικό βιολογικό δόγμα.....12

Εικόνα 2.4 Χημικός τύπος αμινοξέων.....14

Εικόνα 2.5 Οργάνωση των δομών των πρωτεϊνών.....18

Εικόνα 2.6 Γραφική αναπαράσταση α-έλικας.....20

Εικόνα 2.7 Γραφική αναπαράσταση αντιπαράλληλης και παράλληλης β-πτυχωτής επιφάνειας 21

Εικόνα 2.8 Τοπολογικό διάγραμμα αντιπαράλληλης β-πτυχωτής επιφάνειας.....22

Εικόνα 4.1 Ποσοστά επιτυχίας % του αλγορίθμου δεδομένου ότι το σωστό δίπλωμα που αναζητούμε εμφανίζεται στα πρώτα x αποτελέσματα (TA_n, n=1,2,3). Για n=1 πρόκειται για την αυτόματη λειτουργία.....50

Εικόνα 4.2 Αριθμός σωστών ταξινομήσεων σε σχέση με τις υπάρχουσες περιπτώσεις για κάθε δίπλωμα (TA_1, για 125 χαρακτηριστικά και α=0).....55

Εικόνα 4.3 Ποσοστό επιτυχούς ανεύρεσης του διπλώματος μίας πρωτεϊνικής ακολουθίας σε σχέση με το δίπλωμα στο οποίο ανήκει. 55

Εικόνα 5.1 Γραφική αναπαράσταση αποτελέσματα προ και μετά την εφαρμογή του γενετικού αλγορίθμου για τη βελτιστοποίηση του ταξινομητή. 62

Κατάλογος Πινάκων

Πίνακας 2.1 Αμινοξέα που απαντώνται στους οργανισμούς 13

Πίνακας 4.1: Ομάδες χαρακτηριστικών εξαγμένων από την ακολουθία 42

Πίνακας 4.2 Ο αριθμός των ακολουθιών και των δομών που αντιστοιχούν σε κάθε ένα απο τα 27 διπλώματα.....43

Πίνακας 4.4 Ποσοστό % επιτυχίας εύρεσης του σωστού διπλώματος για τις πρωτεϊνικές ακολουθίες για διάφορες τιμές του *α* 49

Πίνακας 4.5 Ποσοστά επιτυχίας του ταξινομητή για τις τρεις κατηγορίες θεώρησης σωστού αποτελέσματος,έξι ομάδες χαρακτηριστικών και τρεις τιμές του α.....51

Πίνακας 4.6 Ποσοστά επιτυχίας του ταξινομητή για κάθε δίπλωμα για τις τρεις κατηγορίες θεώρησης σωστού αποτελέσματος.....53

Πίνακας 5.1 Συγκριτικά αποτελέσματα προ και μετά εφαρμογής γενετικού αλγορίθμου για τη βελτιστοποίηση του ταξινομητή.....61

1

Εισαγωγή

Στο κεφάλαιο αυτό γίνεται αρχικά μια εισαγωγή στο γενικότερο πεδίο της Βιοπληροφορικής στο οποίο εντάσσεται η παρούσα διπλωματική εργασία (Ενότητα 1.1). Στην Ενότητα 1.2 παρουσιάζεται το αντικείμενο της εργασίας και η συνεισφορά της. Τέλος, στην ενότητα 1.3 γίνεται συνοπτική αναφορά στο περιεχόμενο του κάθε κεφαλαίου της εργασίας.

1.1 *Βιοπληροφορική*

Το διεπιστημονικό πεδίο της Βιοπληροφορικής (Bioinformatics) είναι το αποτέλεσμα της σύμπραξης των επιστημών της βιολογίας και της πληροφορικής. Η ραγδαία εξελισσόμενη επιστήμη των υπολογιστών συνεργάζεται με τη σύγχρονη βιολογία με τελικό σκοπό την αποθήκευση, τη διαχείριση, την επεξεργασία και την ερμηνεία της βιολογικής πληροφορίας (1).

Η Βιοπληροφορική αντλεί γνώση από την υπολογιστική ανάλυση των βιολογικών δεδομένων. Τα βιολογικά δεδομένα μπορεί να αφορούν σε πληροφορία προερχόμενη από το γενετικό κώδικα αλλά και πειραματικά αποτελέσματα από διάφορες πηγές, στατιστικά στοιχεία ασθενών, αλλά και επιστημονική βιβλιογραφία. Η βιοπληροφορική, εκτός από την πληροφορική και τη βιολογία, χρησιμοποιεί γνώσεις

από πολλές άλλες επιστήμες όπως από τη στατιστική, τη χημεία, τα θεωρητικά μαθηματικά, τη βιοχημεία, τη φυσική, αλλά και τη γλωσσολογία. Είναι ένας ταχέως αναπτυσσόμενος κλάδος με πολλαπλές εφαρμογές σε διάφορα πεδία της ιατρικής, της φαρμακολογίας αλλά και της ίδιας της βιολογίας (2).

Η χρήση της πληροφορικής στη βιολογία σαν ιδέα ξεκίνησε ήδη από τις αρχές του 19ου αιώνα όταν οι επιστήμονες προσπαθούσαν να προσδιορίσουν βιολογικούς κανόνες με χρήση επαγωγικών συλλογισμών (για παράδειγμα: A.J. Lotka, Elements of Physical Biology, 1925). Ο όρος Βιοπληροφορική επινοήθηκε από την Paulien Hogeweg, θεωρητική βιολόγο του Πανεπιστημίου της Ουτρέχτης, το 1978 για τη μελέτη υπολογιστικών διεργασιών σε βιολογικά συστήματα. Η σχέση μεταξύ της επιστήμης υπολογιστών και της βιολογίας ισχυροποιήθηκε πολύ ομαλά. Η ανάπτυξη γρήγορων επεξεργαστών και η αντίληψη ότι το πεδίο της βιολογίας μπορεί να τους χρησιμοποιήσει για να ξεπεράσει πολλά εμπόδια, καθιέρωσε την βιοπληροφορική ως επιστημονικό πεδίο. Σε συνδυασμό με το γεγονός ότι ο ρυθμός παραγωγής δεδομένων κατά τη βιολογική έρευνα αυξανόταν ολοένα και περισσότερο, λόγω της εισαγωγής σύγχρονων τεχνολογιών στα βιολογικά εργαστήρια, που απαιτούσαν επεξεργασία και ερμηνεία, κατέστησε τη συμβολή της απαραίτητη. Επιπλέον, σύντομα βρέθηκε ότι υπάρχει ισχυρή αναλογία μεταξύ του γενετικού κώδικα και ενός προγράμματος υπολογιστή, και αποδείχθηκε ότι ο γενετικός κώδικας αναπαριστά μία μηχανή Turing (2).Ενώ αρχικά σκοπός της Βιοπληροφορικής ήταν η ανάλυση ακολουθιών νουκλεϊκών οξέων και πρωτεϊνών, σήμερα περιλαμβάνει τη χρήση μεθόδων πληροφορικής στη γονιδιωματική και την πρωτεϊνομική, στη δομική και συστημική βιολογία αλλά και στην ιατρική (1). Πλέον, εφαρμογές βιοπληροφορικής είναι διαθέσιμες για χρήση από κάθε ενδιαφερόμενο, στο διαδίκτυο.

Καθώς είναι ένα πεδίο ραγδαία εξελισσόμενο γίνονται προσπάθειες να καθιερωθεί ένας ευρέως αποδεκτός ορισμός του. Σύμφωνα με το αγγλικό λεξικό της Οξφόρδης (Oxford English Dictionary) (3) ορίζεται ότι *η Βιοπληροφορική αντιλαμβάνεται τη Βιολογία σε επίπεδο μορίων (όμοια με τη φυσικοχημεία) και εφαρμόζει τεχνικές πληροφορικής (εφαρμοσμένα μαθηματικά, επιστήμη των υπολογιστών, στατιστική) με στόχο την κατανόηση και οργάνωση της πληροφορίας που σχετίζεται με τα μόρια, σε μακρο-κλίμακα*. Άλλοι επιστήμονες και πανεπιστημιακά ιδρύματα υποστηρίζουν ένα πιο στενό ορισμό του πεδίου: *η χρήση των ηλεκτρονικών υπολογιστών για την επίλυση προβλημάτων της μοριακής βιολογίας*. Ενώ ο ορισμός που υιοθετεί το πανεπιστήμιο του Stanford συμπεριλαμβάνει και τις περιοχές της υπολογιστικής βιολογίας

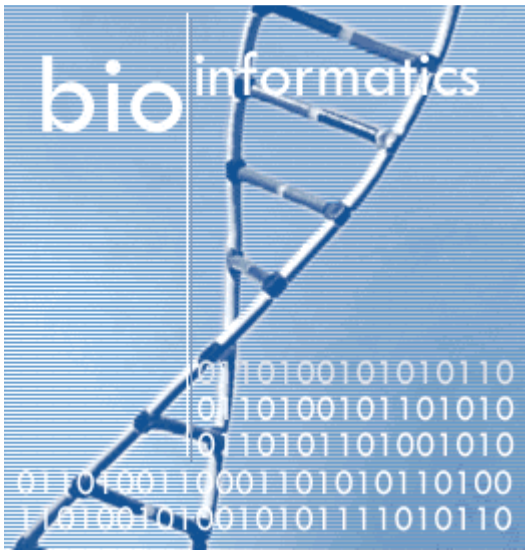
(computational biology) καθώς και την μοντελοποίηση πληθυσμών (population modeling).

Ο σκοπός της βιοπληροφορικής μπορεί να περιγραφεί ως τριπλός (1):

- Οργάνωση των δεδομένων ώστε να εξασφαλίζεται εύκολη πρόσβαση των ερευνητών σε αυτά και να επιτρέπεται η καταχώρηση νέων δεδομένων καθώς παράγονται.
- Ανάπτυξη εργαλείων και πόρων για την ανάλυση των δεδομένων.
- Χρήση των εργαλείων για την ανάλυση και ερμηνεία των δεδομένων με τρόπο χρήσιμο για τη Βιολογία.

Στη συνέχεια αναφέρονται παραδοσιακά αντικείμενα του πεδίου της Βιοπληροφορικής που πραγματεύονται την επεξεργασία βιολογικών δεδομένων σε επίπεδο ακολουθιών και δομών μακρομορίων:

- Στοίχιση και σύγκριση ακολουθιών DNA, RNA και πρωτεϊνών.
- Εύρεση γονιδίων σε ακολουθίες DNA και χαρτογράφησή τους στα χρωμοσώματα.
- Πρόβλεψη δομής DNA και RNA
- Κατηγοριοποίηση και πρόβλεψη δομής πρωτεϊνών
- Πρόβλεψη λειτουργίας πρωτεϊνών
- Σχεδιασμός μορίων και πρόσδεση



Εικόνα 1.1 Βιοπληροφορική : σύμπραξη δύο επιστημονικών πεδίων

Τελευταία, όμως, έχουν αναδυθεί νέες περιοχές που αφορούν σε υψηλότερα επίπεδα οργάνωσης της βιολογικής πληροφορίας. Οι προσδοκίες της Ιατρικής από τους κλάδους αυτούς, που επεξεργάζονται τη βιολογική πληροφορία στα ανώτερα επίπεδα οργάνωσης του κυττάρου, είναι τεράστιες και σχετίζονται με την ανακάλυψη νέων βιολογικών δεικτών (biomarkers) των νοσημάτων, την αναγνώριση νέων θεραπευτικών στόχων, την ανάπτυξη μεθόδων πρώιμης διάγνωσης και την εύρεση εξατομικευμένων θεραπειών. Μέχρι τώρα ο κλάδος της Βιοπληροφορικής έχει παίξει καταλυτικό ρόλο σε μεγάλα επιστημονικά εγχειρήματα, όπως το Πρόγραμμα Ανθρώπινου Γονιδιώματος (Human Research Project), για την αποκωδικοποίηση του ανθρώπινου γονιδιώματος και θέτει παρακαταθήκη για το ρόλο που θα παίξει στο μέλλον.

1.2 *Αντικείμενο Διπλωματικής*

Ένα από τα βασικά πεδία της βιοπληροφορικής είναι η δομική βιοπληροφορική (structural bioinformatics) που μελετά την τρισδιάστατη δομή των μακρομορίων (πρωτεΐνες, DNA, RNA). Στοχεύει στην αποθήκευση, ανάλυση και οπτικοποίηση της δομικής πληροφορίας των μακρομορίων σε διάφορα επίπεδα και την εξαγωγή χρήσιμης γνώσης για τη λειτουργία τους. Ένα από τα σημαντικότερα σκέλη της δομικής βιοπληροφορικής είναι η υπολογιστική πρόγνωση της δομής των πρωτεϊνών. Η σημασία της έγκειται στο γεγονός ότι ο πειραματικός προσδιορισμός των πρωτεϊνών είναι δύσκολος, με αποτέλεσμα να έχει συσσωρευτεί μεγάλος αριθμός ακολουθιών πρωτεϊνών με άγνωστη δομή.

Η παρούσα διπλωματική εργασία καταπιάνεται με το πρόβλημα της αναγνώρισης διπλώματος πρωτεϊνών (fold recognition). Δηλαδή γίνεται κατηγοριοποίηση πρωτεϊνικών ακολουθιών σε δομικές κατηγορίες λαμβάνοντας υπόψη χαρακτηριστικά που προέρχονται από την πρωτεϊνική ακολουθία (sequence-derived features). Υλοποιείται ένας ταξινομητής (classifier) για να βρίσκει τη σωστή δομική κατηγορία για μία ακολουθία αναζήτησης γνωρίζοντας: ορισμένα χαρακτηριστικά της ακολουθία αυτής, τις δομικές κατηγορίες στις οποίες ανήκει ένα σύνολο από άλλες ακολουθίες, καθώς και τα αντίστοιχα χαρακτηριστικά των πρωτεϊνών του συνόλου αυτού.

Στην προκειμένη περίπτωση χρησιμοποιείται ένα σύνολο δεδομένων 311 πρωτεϊνικών ακολουθιών που κατανέμονται σε 27 γνωστά διπλώματα. Κάθε πρωτεϊνική ακολουθία του συνόλου δεδομένων χαρακτηρίζεται από 125 χαρακτηριστικά ακολουθίας που σχετίζονται με συγκεκριμένες βιοχημικές ιδιότητές της. Οι ιδιότητες αυτές είναι: η αμινοξική σύσταση της, η υδροφοβικότητα, η πόλωση, η πολωσιμότητα, ο κανονικοποιημένος όγκος Van der Waals, και η προβλεπόμενη δευτεροταγής δομή των καταλοίπων της. Ο ταξινομητής βασίζεται σε έναν αλγόριθμο που χρησιμοποιεί δίκτυα ομοιότητας πρωτεϊνών (Protein Sequence Similarity Networks), κατασκευασμένα με βάση τα 125 χαρακτηριστικά της ακολουθίας. Ο ταξινομητής υλοποιήθηκε σε περιβάλλον Matlab και ακολούθησαν πολλά πειράματα με μεταβολή διαφορετικών παραμέτρων κάθε φορά, ώστε να αξιολογηθεί κατάλληλα. Τα πειράματα έγιναν με τη μέθοδο της διασταυρωμένης επικύρωσης από 10 μέρη, και η μέση ορθότητα του ταξινομητή βρέθηκε 36%.

Στη συνέχεια εφαρμόστηκε ένας γενετικός αλγόριθμος (genetic algorithm), με σκοπό την επιλογή των χαρακτηριστικών εκείνων (από το αρχικό σύνολο των 125), που περιέχουν σημαντικότερη πληροφορία για τη δομή των ακολουθιών. Για το σκοπό αυτό υλοποιήθηκε και το δίκτυο ομοιότητας των δομών των πρωτεϊνών (Protein Structure Similarity Network). Ο γενετικός αλγόριθμος πρότεινε τη χρήση 48 χαρακτηριστικών από το αρχικό σύνολο, για τη βελτιστοποίηση του αποτελέσματος. Έπειτα, επαναλήφθηκαν τα πειράματα ώστε να γίνει επαναξιολόγηση της ορθότητας του ταξινομητή. Μετά την εφαρμογή του γενετικού αλγορίθμου και την αξιοποίηση των αποτελεσμάτων του, η μέση ορθότητα του ταξινομητή αυξήθηκε κατά 5%. Αξίζει να σημειωθεί ότι παρόλο που η πολυπλοκότητα του αλγορίθμου διατηρείται σε χαμηλά επίπεδα, η απόδοσή του είναι συγκρίσιμη με την απόδοση ταξινομητών μεγαλύτερης πολυπλοκότητας, που έχουν προταθεί για το ίδιο πρόβλημα.

1.2.1 Συνεισφορά

Η βασική συνεισφορά της παρούσας εργασίας συνοψίζεται στα ακόλουθα σημεία.

- Προτείνεται λύση για το πρόβλημα της αναγνώρισης του διπλώματος πρωτεϊνικών ακολουθιών χρησιμοποιώντας εξελιγμένες μεθόδους υπολογιστικής νοημοσύνης.

- Γίνεται μία λεπτομερής ανάλυση του προβλήματος καθώς και αναφορά στις προηγούμενες σχετικές εργασίες.
- Υλοποιείται ένας ταξινομητής που μπορεί να χρησιμοποιηθεί για κάθε πρωτεϊνική ακολουθία δεδομένου ότι μπορούμε να εξάγουμε χαρακτηριστικά από αυτή.
- Χρησιμοποιούνται δίκτυα ομοιότητας πρωτεϊνικών ακολουθιών κατασκευασμένα με βάση πληροφορίες από τις ίδιες τις ακολουθίες ή από τις δομές τους, που προέκυψαν από εργαλεία διαθέσιμα στο διαδίκτυο.
- Χρησιμοποιείται ένας γενετικός αλγόριθμος για τη βελτιστοποίηση του ταξινομητή, που αυξάνει την απόδοσή του.
- Γίνεται παρουσίαση μιας σειράς πειραμάτων καθώς και των αποτελεσμάτων και των συμπερασμάτων που προέκυψαν από αυτά.

1.3 *Οργάνωση κειμένου*

Το περιεχόμενο της διπλωματικής εργασίας οργανώνεται στα ακόλουθα κεφάλαια.

Στο κεφάλαιο 2 παρουσιάζεται το θεωρητικό υπόβαθρο πάνω στο οποίο βασίστηκε η παρούσα εργασία και κάποιες βασικές έννοιες που χρησιμοποιήθηκαν. Επειδή η εργασία έχει άμεση σχέση με το αντικείμενο της βιολογίας, θεωρήθηκε σκόπιμη μια αναφορά στις βασικές έννοιες του πεδίου. Επίσης γίνεται μια εισαγωγή στη θεωρία των γράφων και των γενετικών αλγορίθμων.

Στο κεφάλαιο 3 γίνεται εκτενής αναφορά στο πρόβλημα της αναγνώρισης διπλώματος πρωτεϊνών, στη σημασία του και στις μεθόδους που χρησιμοποιούνται για τη λύση του.

Το κεφάλαιο 4 αφορά στην ανάλυση του αλγορίθμου και στην υλοποίηση του συστήματος ταξινόμησης. Γίνεται πειραματική αξιολόγηση και σχολιασμός των αποτελεσμάτων. Το κεφάλαιο 5 αναφέρεται στη χρήση γενετικού αλγορίθμου για τη βελτιστοποίηση του μοντέλου ταξινομητή που προτείνεται. Παρουσιάζονται τα χαρακτηριστικά του νέου μοντέλου, καθώς και τα αποτελέσματα των πειραμάτων που έγιναν με τα νέα δεδομένα.

Στο κεφάλαιο 6 γίνεται μια σύνοψη της εργασίας, παρουσιάζονται τα συμπεράσματα που προέκυψαν και προτείνονται κάποιες μελλοντικές επεκτάσεις.

Στο κεφάλαιο 7 δίνεται η βιβλιογραφία που χρησιμοποιήθηκε.

Στο Παράρτημα παρατίθεται ο κώδικας που χρησιμοποιήθηκε για την υλοποίηση του αλγορίθμου που προτείνεται στο περιβάλλον Matlab. Παρουσιάζονται οι σημαντικότερες συναρτήσεις με σχόλια.

2

Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό παρουσιάζονται κάποιες βασικές έννοιες που χρησιμοποιήθηκαν στην παρούσα εργασία και είναι απαραίτητες για την κατανόηση του προβλήματος. Στην ενότητα 2.1 γίνεται μια αναφορά στις βασικές βιολογικές έννοιες με έμφαση στις πρωτεΐνες, στη 2.2 γίνεται μια παρουσίαση των εννοιών που σχετίζονται με την πρόβλεψη των δομών των πρωτεϊνών, ενώ στη 2.3 παρουσιάζεται το υπόβαθρο από την επιστήμη των υπολογιστών, πάνω στο οποίο βασίστηκε η εργασία.

2.1 *Ο κώδικας της ζωής*

Ένας από τους σημαντικότερους κλάδους της Βιολογίας είναι η Μοριακή Βιολογία. Η Μοριακή Βιολογία μελετά τη δομή και την έκφραση της γενετικής πληροφορίας σε μοριακό επίπεδο, δηλαδή τη λειτουργία των βιολογικών μακρομορίων των κυττάρων και την έκφρασή τους. Το πεδίο έρευνάς της έχει κοινά χαρακτηριστικά με άλλα πεδία της βιολογίας, ιδιαίτερα με τη γενετική και τη βιοχημεία.

Αναλυτικότερα, η μικρότερη δομική και λειτουργική μονάδα ζωής είναι το κύτταρο. Το κύτταρο είναι μια συστηματικά οργανωμένη ομάδα μορίων που βρίσκονται σε συνεχή αλληλεπίδραση μεταξύ τους. Τα χημικά στοιχεία που

απαντώνται σε ένα κύτταρο είναι περιορισμένα. Ο Άνθρακας (C), το Υδρογόνο (H), το Οξυγόνο (O), το Άζωτο (N), ο Φωσφόρος (P) και το Θείο (S), αποτελούν και το 99% περίπου του βάρους του. Τα χημικά συστατικά του είναι δυνατόν να ταξινομηθούν σε ανόργανα: νερό (H_2O) και μεταλλικά ιόντα, και οργανικά: πρωτεΐνες, υδατάνθρακες, λίπη και νουκλεϊκά οξέα (4).

Τα κύτταρα διαχωρίζονται από το περιβάλλον τους αλλά και επικοινωνούν με αυτό, μέσω της κυτταρικής μεμβράνης τους. Στο εσωτερικό τους βρίσκεται το κυτταρόπλασμα στο οποίο περιέχονται τα οργανίδια του κυττάρου, δηλαδή διαχωρισμένες δομικές μονάδες που εκτελούν τις διάφορες λειτουργίες. Τα οργανίδια ποικίλλουν από κύτταρο σε κύτταρο, ανάλογα με τη λειτουργία του. Τα διάφορα οργανίδια είναι: Κυτταρική μεμβράνη, Ενδοπλασματικό δίκτυο, το οποίο διακρίνεται σε αδρό και λείο, Πυρήνας, Πυρηνίσκος, Σύμπλεγμα Golgi, Μικροϊνίδια και μικροσωληνίσκοι, τα οποία αποτελούν τον Κυτταρικό σκελετό, Ριβόσωμα, Κυστίδιο, Μιτοχόνδρια, Κενοτόπιο, Λυσόσωμα, Κεντροσωμάτιο (μόνο σε ζωικά κύτταρα), Χλωροπλάστες, Χυμοτόπιο (μόνο σε φυτικά κύτταρα).

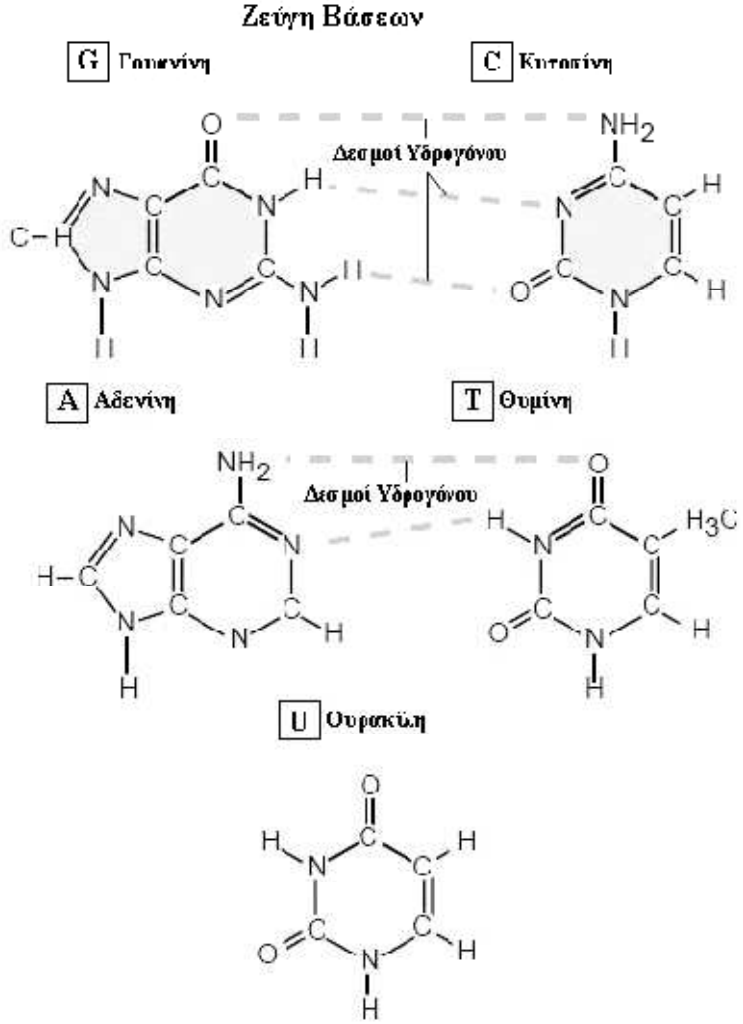
Κάθε κύτταρο και κατ'επέκταση κάθε οργανισμός κατασκευάζει τις δομές του και πραγματοποιεί τις λειτουργίες του σύμφωνα με μια σειρά πληροφοριών που έχει κληρονομήσει από τους προγόνους του. Οι πληροφορίες αυτές είναι καταγραμμένες στο DNA του, δηλαδή στο μόριο που αποτελεί το γενετικό υλικό των κυττάρων. Συνοπτικά οι λειτουργίες του γενετικού υλικού είναι οι εξής:

- Η αποθήκευση της γενετική πληροφορίας. Στο DNA (ή στο RNA των RNA-ιών) περιέχονται οι πληροφορίες που καθορίζουν όλα τα χαρακτηριστικά ενός οργανισμού και οργανώνονται σε λειτουργικές μονάδες, τα γονίδια.
- Η διατήρηση και η μεταβίβαση της γενετικής πληροφορίας από κύτταρο σε κύτταρο και από οργανισμό σε οργανισμό, που εξασφαλίζονται με τον αυτοδιπλασιασμό του DNA.
- Η έκφραση των γενετικών πληροφοριών, που επιτυγχάνεται με τον έλεγχο της σύνθεσης των πρωτεϊνών.

Τα είδη των νουκλεϊκών οξέων στα κύτταρα είναι δύο: το δεσοξυριβονουκλεϊκό οξύ (DNA) και το ριβονουκλεϊκό οξύ (RNA). Τα νουκλεϊκά οξέα συντίθενται από δομικές μονάδες που ονομάζονται νουκλεοτίδια. Τα νουκλεοτίδια είναι σύνθετα και αποτελούνται από τρία μέρη:

- ένα μόριο σακχάρου, πεντόζη, δεσοξυριβόζη για το DNA, ριβόζη για το RNA,
- ένα μόριο φωσφορικού οξέος, που συνδέεται στο σάκχαρο,
- μια αζωτούχο βάση, που επίσης συνδέεται στο σάκχαρο.

Ενα δεσοξυριβο-νουκλεοτίδιο μπορεί να έχει μια από τις ακόλουθες αζωτούχες βάσεις: Αδενίνη (A), Γουανίνη (G) Κυτοσίνη (C) ή Θυμίνη (T). Ένα ριβονουκλεοτίδιο αντί για θυμίνη (T) έχει Ουρακίλη (U). Οι δύο βάσεις A και G είναι μεγαλύτερες και ανήκουν στην ομάδα των πουρινών, ενώ οι άλλες τρείς, C, T και U είναι μικρότερες και ανήκουν στις πυριμιδίνες.



Εικόνα 2.1 Οι βάσεις του νουκλεϊκών οξέων και οι δεσμοί που αναπτύσσονται μεταξύ τους.

Το DNA αποτελείται από δύο πολυνουκλεοτιδικές αλυσίδες, τους κλώνους, που συνδέονται μεταξύ τους με δεσμούς υδρογόνου σχηματίζοντας μια δεξιόστροφη έλικα. Οι δεσμοί υδρογόνου δημιουργούνται ανάμεσα στις αζωτούχες βάσεις, που

είναι κάθετες στον άξονα του μορίου και προεξέχουν προς τα μέσα. Δεσμοί υδρογόνου σχηματίζονται μεταξύ των συμπληρωματικών βάσεων, δύο μεταξύ της αδενίνης και της θυμίνης και τρεις μεταξύ της γουανίνης και της κυτοσίνης. Το μόριο του RNA σε αντίθεση με αυτό του DNA είναι μονόκλωνο και περιέχει ουρακίλη. Δεδομένου ότι τα σάκχαρα και το φωσφορικό οξύ είναι ίδια σε όλα τα νουκλεοτίδια, μπορούμε να προσδιορίσουμε μια νουκλεοτιδική αλυσίδα μόνο από την ακολουθία βάσεών της. Αυτό το συμπέρασμα είναι πολύ χρήσιμο στη Βιοπληροφορική.

Ακολουθεί μια περιγραφή των διαδικασιών στις οποίες μετέχουν τα νουκλεϊκά οξέα, με ιδιαίτερη έμφαση στη μετάφραση - τη διαδικασία σύνθεσης των πρωτεϊνών (5).

- Αντιγραφή

  Πριν από τη διαίρεση των κυττάρων και το σχηματισμό των θυγατρικών τους, το γενετικό υλικό τους διπλασιάζεται. Κατά τη διαδικασία αυτή το πατρικό DNA χρησιμοποιείται ως καλούπι για την δημιουργία θυγατρικών αλυσίδων. Τα κύτταρα διαθέτουν ένα σημαντικό αριθμό εξειδικευμένων ενζύμων και άλλων πρωτεϊνών που λειτουργούν ταυτόχρονα και καταλύουν τις χημικές αντιδράσεις με μεγάλη ταχύτητα και εκπληκτική ακρίβεια. Έτσι, καταλύονται οι δεσμοί υδρογόνου που συγκρατούν τους δυο κλώνους μεταξύ τους, και κάθε παλιά αλυσίδα χρησιμεύει ως καλούπι για την κατασκευή μιας νέας. Κάθε θυγατρικό μόριο αποτελείται από μια πατρική και μια νέα αλυσίδα (ημισυντηρητικός τρόπος διπλασιασμού).

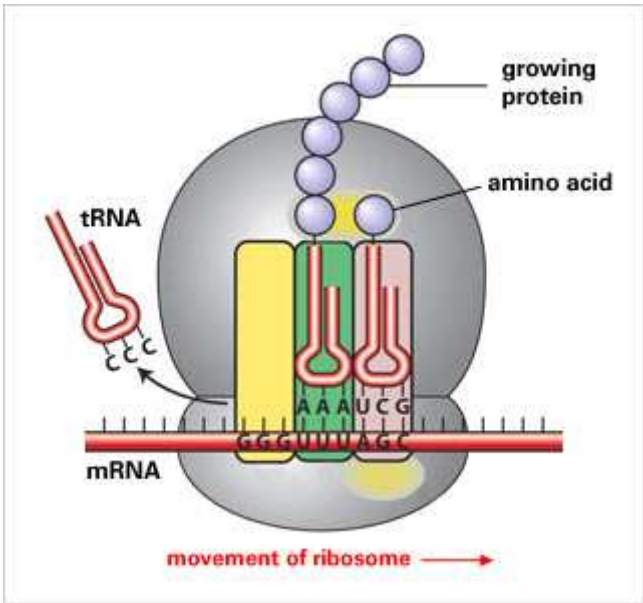
- Μεταγραφή

  Το πρώτο βήμα για την έκφραση της γενετικής πληροφορίας είναι η μεταφορά της από το DNA στο RNA μέσω της διαδικασίας της μεταγραφής. Για τη κατασκευή του RNA, ξετυλίγεται η αλυσίδα του DNA σε μια συγκεκριμένη περιοχή και τοποθετούνται συμπληρωματικές βάσεις απέναντι από τη μία από τις δυο έλικες. Αφού τελειώσει η διαδικασία οι δύο κλώνοι επανασυνδέονται. Υπάρχουν πολλοί τύποι RNA που συντίθενται με αυτό τον τρόπο, με σημαντικότερο το αγγελιαφόρο RNA (mRNA), που μεταφέρει την πληροφορία της σύνθεσης των πρωτεϊνών από τον πυρήνα στα ριβοσώματα. Υπάρχουν 3 ακόμη είδη RNA τα οποία μετέχουν στη διαδικασία της πρωτεϊνοσύνθεσης με έμμεσο τρόπο, το μεταφορικό RNA (tRNA), που μεταφέρει τα αμινοξέα, το ριβοσωμικό (rRNA), που είναι δομικό συστατικό

των ριβοσωμάτων και το μικροπυρηνικό (snRNA), που παίζει καθοριστικό ρόλο στη σύνθεση του ίδιου του RNA.

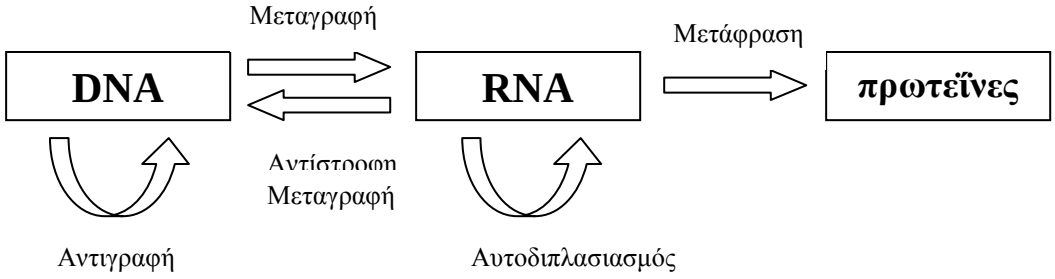
- Μετάφραση

  Η μετάφραση είναι η αποκωδικοποίηση των πληροφοριών που μεταφέρει το mRNA με σκοπό τη σύνθεση κάποιας συγκεκριμένης πρωτεΐνης. Πραγματοποιείται στα ριβοσώματα και απαιτεί τη συμμετοχή αρκετών πρωτεϊνών και ενέργειας. Κάθε τριάδα βάσεων (κωδικόνιο) στο mRNA αντιστοιχεί και σε ένα συγκεκριμένο αμινοξύ. Πολλά αμινοξέα δημιουργούν μια πολυπεπτιδική αλυσίδα, την πρωταρχική δομή της πρωτεΐνης. Το tRNA μεταφέρει βάσει του κανόνα της συμπληρωματικότητας, (βλ.Εικόνα 2.2) το κατάλληλο αμινοξύ στο ριβόσωμα, το οποίο ενώνεται με το προηγούμενό του με πεπτιδικό δεσμό. Έτσι σχηματίζεται η πολυπεπτιδική αλυσίδα. Η δημιουργία του πεπτιδικού δεσμού καταλύεται από το ένζυμο «πεπτιδυλο-τρανσφεράση». Αμέσως μετά ή και κατά τη διάρκεια της πρωτεϊνοσύνθεσης, τα αμινοξικά κατάλοιπα υφίστανται μετατροπές, τις μετα-μεταγραφικές τροποποιήσεις. Οι τροποποιήσεις αυτές καθορίζουν τις φυσικές και χημικές ιδιότητες, την αναδίπλωση, τη δραστικότητα καθώς και τη λειτουργικότητα των πρωτεϊνών.



Εικόνα 2.2 Διαδικασία πρωτεϊνόσύνθεσης σε ένα ριβόσωμα.

Οι παραπάνω ιδιότητες των νουκλεϊκών οξέων συνοψίζονται στο κεντρικό δόγμα της Μοριακής Βιολογίας. Το DNA αυτοδιπλασιάζεται και μεταγράφεται παράγοντας RNA. Το RNA μεταφράζεται και παράγει πρωτεΐνες. Επίσης σε μερικές περιπτώσεις ιών το RNA έχει τη δυνατότητα να αυτοδιπλασιάζεται αλλά και να μεταγράφεται αντίστροφα, παράγοντας DNA. (βλ. Εικόνα 2.3)



Εικόνα 2.3 Κεντρικό δόγμα Μοριακής Βιολογίας

2.2 *Πρωτεΐνες*

2.2.1 Σύσταση

Οι πρωτεΐνες (proteins) ή πολυπεπτίδια (polypeptides) είναι μεγάλες οργανικές ενώσεις που αποτελούνται από αμινοξέα διατεταγμένα σε γραμμικές αλυσίδες. Το μοριακό τους βάρος κυμαίνεται από 10.000 μέχρι 1.000.000. Όλες οι πρωτεΐνες περιέχουν Άνθρακα (C), Οξυγόνο (O), Άζωτο (N) και οι περισσότερες από αυτές Θείο (S). Οι δομικές τους μονάδες, τα αμινοξέα, συνδέονται μεταξύ τους με πεπτιδικό δεσμό, με ταυτόχρονη αποβολή ενός μορίου νερού. Στις πρωτεΐνες των ζωντανών οργανισμών απαντώνται 20 διαφορετικά αμινοξέα. Καθένας από τους δυνατούς συνδυασμούς σε μια τριπλέτα (triplet) βάσεων, αντιστοιχίζεται σε ένα από τα 20 αμινοξέα, χρησιμοποιώντας τον κανόνα του ενός ή των τριών γραμμάτων (βλ. Πίνακα 2.1). Άρα, κάθε αμινοξύ κωδικοποιείται από μια τουλάχιστον τριάδα γενετικού κώδικα DNA.

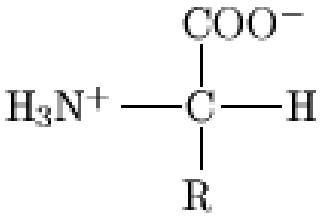
Αμινοξύ	Κωδικοποιήση 3 γραμμάτων	Κωδικοποίηση 1 γράμματος
Αλανίνη (Alanine)	Ala	A
Αργινίνη (Arginine)	Arg	R
Ασπαραγίνη (Asparagine)	Asn	N
Ασπαρτικό Οξύ (Aspartic acid)	Asp	D
Κυστεΐνη (Cys)	Cys	C
Γλουταμινικό Οξύ (Glutamic Acid)	Glu	E
Γλουταμίνη (Gloutamine)	Gln	Q
Γλυκίνη (Glycine)	Gly	G
Ιστιδίνη (His)	His	H
Ισολευκίνη (Isoleucine)*	Ile	I
Λευκίνη (Leucine)*	Leu	L
Λυσίνη (Lysine)*	Lys	K
Μεθιονίνη (Methionine)*	Met	M
Φαινυλαλανίνη (Phenylalanine)*	Phe	F
Προλίνη (Proline)	Pro	P
Σηρίνη (Serine)	Ser	S
Θρεονίνη (Threonine)*	Thr	T
Τρυπτοφάνη (Tryptophan)*	Trp	W
Τυροσίνη (Tyrosine)	Tyr	Y
Βαλίνη (Valine)*	Val	V

Πίνακας 2.1 Αμινοξέα που απαντώνται στους οργανισμούς (τα απαραίτητα συμβολίζονται με *).

Γενικά, στον ανθρώπινο οργανισμό, 8 από τα 20 αμινοξέα που χρησιμοποιούνται στη σύνθεση πρωτεϊνών δεν μπορούν να συντεθούν από τον ίδιο τον οργανισμό και πρέπει να λαμβάνονται από την τροφή. Για το λόγο αυτό καλούνται απαραίτητα ή βασικά αμινοξέα. Επίσης, 4 από τα 20 είναι ημιαπαραίτητα, αφού δεν μπορούν να συντεθούν στα παιδιά. Τα υπόλοιπα 8 συντίθενται μέσω των μεταβολικών μονοπατιών.

Όλα τα αμινοξέα έχουν μια γενική βιοχημική σύσταση. Αποτελούνται από μια ή περισσότερες αμινομάδες ($-NH_2$), από μια ή περισσότερες καρβοξυλομάδες (-COOH), ένα άτομο υδρογόνου (H) και μια πλευρική αλυσίδα R, που συνδέονται σε ένα άτομο άνθρακα. Διακρίνονται σε α-αμινοξέα, β-αμινοξέα, γ-αμινοξέα κ.ο.κ. ανάλογα με το σε ποιο άτομο άνθρακα -μετά το καρβοξύλιο- βρίσκεται η αμινομάδα. Τα **α-αμινοξέα** (δηλαδή αυτά που έχουν την αμινομάδα και το καρβοξύλιο στο ίδιο άτομο άνθρακα, που συμβολίζεται ως Cα) αποτελούν τα δομικά συστατικά των

πρωτεϊνών (βλ. Εικόνα 2.4). Η δομή είναι ίδια και για τα 20 αμινοξέα εκτός από τις πλευρικές ομάδες R.



Εικόνα 2.4 Χημικός τύπος αμινοξέων

Κάθε πρωτεΐνη μπορεί να αποτελείται από μία ή περισσότερες αλυσίδες αμινοξέων ή αλλιώς, πεπτιδικές αλυσίδες. Η μονοδιάστατη αμινοξική ακολουθία της πρωτεΐνης ονομάζεται πρωτοταγής δομή (primary structure).

Τα αμινοξέα μπορούν να ομαδοποιηθούν ανάλογα με τις ιδιότητες των πλευρικών αλυσίδων τους (R):

- Αμινοξέα με αλειφατικές, μη πολικές πλευρικές αλυσίδες:

  Αλανίνη, Βαλίνη, Γλυκίνη, Ισολευκίνη, Λευκίνη, Μεθειονίνη, Προλίνη.

- Αμινοξέα με αλειφατικές, πολικές πλευρικές αλυσίδες:

  Θρεονίνη, Σερίνη, Κυστεΐνη.

- Αμινοξέα με αρωματικές πλευρικές αλυσίδες:

  Φαινυλαλανίνη, Τυροσίνη, Τρυπτοφάνη.

- Αμινοξέα με βασικές πλευρικές αλυσίδες:

  Λυσίνη, Αργινίνη, Ιστιδίνη.

- Αμινοξέα με όξινες πλευρικές αλυσίδες:

  Ασπαραγινικό, Γλουταμινικό.

- Αμινοξέα με καρβοξυλαμίδια στις πλευρικές αλυσίδες τους:

  Ασπαραγίνη, Γλουταμίνη.

2.2.2 Ρόλος

Οι πρωτεΐνες παίζουν πολύ σημαντικό ρόλο στις βιολογικές διεργασίες των κυττάρων και κατ'επέκταση των οργανισμών. Ο ρόλος αυτός καθορίζεται από την αλληλουχία των αμινοξέων τους και από τις μετα-μεταγραφικές μετατροπές. Οι πρωτεΐνες, με κριτήριο τη λειτουργία τους διακρίνονται σε 2 ευρύτερες κατηγορίες: τις δομικές, που αποτελούν δομικά συστατικά των κυττάρων και κατ'επέκταση των οργανισμών, και τις λειτουργικές, που συμβάλλουν στις διάφορες λειτουργίες. Συγκεκριμένα η δράση των πρωτεϊνών συνδέεται με τα παρακάτω (5) :

- **Δομή**

Οι πρωτεΐνες παίζουν σημαντικό ρόλο στη δομή σε επίπεδο κυττάρου, καθώς οι κυτταρικές μεμβράνες αποτελούνται σε μεγάλο ποσοστό από πρωτεΐνες, αλλά και σε επίπεδο οργανισμού. Για παράδειγμα η αντοχή δέρματος και οστών οφείλονται στο κολλαγόνο και την ελαστίνη, δυο ινώδεις πρωτεΐνες.

- **Λειτουργίες**

Καταλυτική δράση

Ο πιο γνωστός ρόλος των πρωτεϊνών στα κύτταρα είναι ο ρόλος τους ως ένζυμα, τα οποία καταλύουν χημικές αντιδράσεις. Ο ρόλος των ενζύμων είναι πολύ στοχευμένος, με αποτέλεσμα να καταλύουν ελάχιστες, ακόμα και μόνο μια, χημικές αντιδράσεις. Τα ένζυμα έχουν πολύ ουσιαστικό ρόλο στη βασικές διεργασίες του κυττάρου: στην αντιγραφή, τη μεταγραφή, τη μετάφραση και τις μετα-μεταγραφικές τροποποιήσεις. Η δράση των ενζύμων επιταχύνει τις χημικές αντιδράσεις από 10^6 μέχρι και 10^{17} φορές. Από δομική άποψη, αν και τα ένζυμα μπορεί να αποτελούνται από εκατοντάδες αμινοξέα, μόνο ένα μικρό ποσοστό των αμινοξέων συμμετέχει στην κατάλυση.

Μηχανισμός άμυνας

Τα αντισώματα είναι πρωτεϊνικά μόρια, που παράγονται από τα β-λεμφοκύτταρα του ανοσοποιητικού μας συστήματος, όταν ένα αντιγόνο (παθογόνος μικροοργανισμός, ιός ή ξένο υλικό), προσβάλλει τον οργανισμό. Τα αντισώματα αντιδρούν με το αντιγόνο και το εξουδετερώνουν. Στην πραγματικότητα, ένα αντίσωμα αναγνωρίζει μόνο μια περιοχή του αντιγόνου, που ονομάζεται αντιγονικός καθοριστής. Κάθε είδος αντισώματος που αναγνωρίζει ένα αντιγονικό καθοριστή παράγεται από μια ομάδα όμοιων β-λεμφοκυττάρων, που αποτελούν ένα κλώνο. Τα αντισώματα που παράγονται από ένα κλώνο β-λεμφοκυττάρων ονομάζονται μονοκλωνικά. Τα μονοκλωνικά αντισώματα είναι πολύ σημαντικά στην ιατρική και χρησιμοποιούνται ως διαγνωστικά για την ανίχνευση ασθενειών ή ως εξειδικευμένα φάρμακα εναντίον παθογόνων μικροοργανισμών ή ακόμα και καρκινικών κυττάρων.

Μεταφορά και αποθήκευση ουσιών

Πολλές πρωτεΐνες συνδέονται με συγκεκριμένα μικρά βιομόρια με σκοπό τη μεταφορά τους σε άλλες περιοχές ενός πολυκύτταρου οργανισμού ή την αποθήκευσή τους. Για παράδειγμα η αιμοσφαιρίνη είναι υπεύθυνη για τη μεταφορά οξυγόνου και διοξειδίου του άνθρακα στο αίμα σπονδυλωτών. Η μυοσφαιρίνη μεταφέρει οξυγόνο και το αποθηκεύει προσωρινά στους μύες των σπονδυλωτών. Ενώ η τρανσφεράση μεταφέρει σίδηρο στο αίμα, ο οποίος αποθηκεύεται στο συκώτι, ως σύμπλοκο με τη φερριτίνη, μια ακόμα πρωτεΐνη.

Λειτουργία Κίνησης

Οι πρωτεΐνες είναι τα κύρια συστατικά των μυών, των οποίων η συστολή επιτυγχάνεται με την ολισθητική κίνηση δύο ειδών πρωτεϊνικών νηματίων, της ακτίνης και της μυοσίνης. Επίσης η κίνηση των χρωμοσωμάτων στη μίτωση και η μέσω μαστιγίων προώθηση του σπέρματος, επιτυγχάνεται με συσταλτά πρωτεϊνικά συγκροτήματα.

Διακυτταρική επικοινωνία

Οι πρωτεΐνες παίζουν πολύ σημαντικό ρόλο στη διακυτταρική επικοινωνία. Για παράδειγμα οι ιντερφερόνες είναι αντιιικές πρωτεΐνες, που παράγονται από κύτταρα που έχουν μολυνθεί από ιούς. Οι πρωτεΐνες αυτές, επάγουν την παραγωγή άλλων πρωτεϊνών από τα γειτονικά υγιή κύτταρα, οι οποίες εμποδίζουν των πολλαπλασιασμό των ιών σε αυτά. Οι ιντερφερόνες είναι οικογένεια συγγενών πρωτεϊνών που ταξινομούνται ανάλογα με τη χημική και βιολογική ενεργότητά τους σε τρεις ομάδες: τις ιντερφερόνες α, β και γ. Παρόμοιο ρόλο έχει και η ινσουλίνη, η οποία μεταφέρει ένα σήμα από το κύτταρο στο οποίο συντίθεται σε άλλα κύτταρα απομακρυσμένων ιστών. Επίσης πολύ σημαντικός είναι ο ρόλος των πρωτεϊνών της κυτταρικής μεμβράνης ως υποδοχείς εξωτερικών σημάτων με σκοπό την πρόκληση ανάλογων χημικών αντιδράσεων στο εσωτερικό του κυττάρου.

Έλεγχος της ανάπτυξης και της διαφοροποίησης

Οι πρωτεΐνες χρησιμεύουν στην ελεγχόμενη έκφραση των γενετικών πληροφοριών, ουσιώδη διαδικασία για την κανονική ανάπτυξη και διαφοροποίηση των κυττάρων.

Ρύθμιση λειτουργιών

Πολλές λειτουργίες ρυθμίζονται από ειδικές πρωτεΐνες, όπως για παράδειγμα η όραση που ρυθμίζεται από τη ροδοψίνη η οποία είναι μια πρωτεΐνη-φωτοϋποδοχέας στα ραβδία του αμφιβληστροειδή, αλλά και η μετάδοση νευρικών παλμών που ρυθμίζεται από την ακετυλχολιχίνη.

Υπάρχουν ακόμα πολλοί ειδικοί ρόλοι, που έχουν οι πρωτεΐνες στην ζωή των κυττάρων και των οργανισμών. Επίσης οι πρωτεΐνες παίζουν πολύ σημαντικό ρόλο στην έναρξη αλλά και πορεία πολλών ασθενειών στον ανθρώπινο οργανισμό.

2.2.3 Πρωτεϊνική δομή

2.2.3.1 Γενικά

Πριν οι πρωτεΐνες μπορέσουν να πραγματοποιήσουν αυτές τις σημαντικές λειτουργίες, αυτοδιαμορφώνουν την τρισδιάστατη δομή τους, ή «αναδιπλώνονται». Η διαδικασία της αναδίπλωσης των πρωτεϊνών, παρότι είναι ζωτική και θεμελιώδης ουσιαστικά για όλη τη βιολογία, παραμένει από πολλές πλευρές ένα μυστήριο.

Όταν οι πρωτεΐνες δεν αναδιπλώνονται σωστά («λανθασμένη αναδίπλωση»), μπορεί να υπάρξουν σοβαρές επιπτώσεις. Σε λανθασμένη αναδίπλωση οφείλονται πολλές ασθένειες, όπως η Νόσος του Αλτσχάιμερ, η νόσος Κρόιτζφελντ-Γιάκομπ, η πλαγία μυοατροφική σκλήρυνση, η νόσος του Χάντιγκτον, η νόσος του Πάρκινσον, πολλοί καρκίνοι και σχετιζόμενα με τον καρκίνο σύνδρομα.

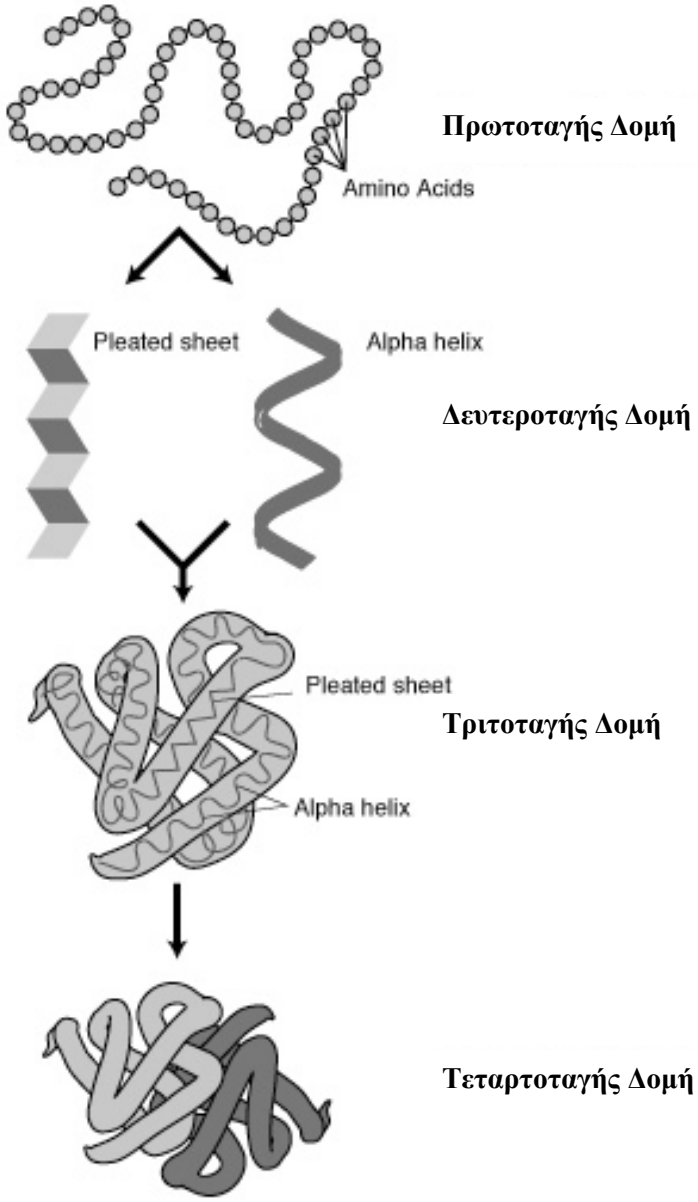
Εκτός από την λανθασμένη αναδίπλωση, μπορεί να επέλθει και καταστροφή ενός σωστού διπλώματος υπό την επίδραση περιβαλλοντικών παραγόντων, για παράδειγμα έκθεση σε ακραίες τιμές θερμοκρασίας ή PH. Τότε οι πρωτεΐνες υφίστανται μετουσίωση, δηλαδή σπάνε οι δεσμοί που έχουν αναπτυχθεί μεταξύ των πλευρικών ομάδων, καταστρέφεται η τρισδιάστατη δομή τους και η πρωτεΐνες χάνουν τη λειτουργικότητά τους.

Κάθε είδος πρωτεΐνης αναδιπλώνεται σε ένα πολύ συγκεκριμένο σχήμα κάθε φορά με βοήθεια άλλων ενζύμων ή χωρίς. Αυτό το σχήμα είναι το πιο σταθερό που μπορεί να αποκτήσει η πρωτεΐνη.

2.2.3.2 Οργάνωση των δομών των πρωτεϊνών

- **Πρωτοταγής Δομή:** είναι η μονοδιάστατη αμινοξική ακολουθία
- **Δευτεροταγής Δομή:** προκύπτει από την αναδίπλωσή της αλυσίδας των αμινοξέων και την κανονική στερεοδιάταξη τμημάτων της. Συνίσταται στη δημιουργία υπομονάδων χωρικής οργάνωσης, όπως α-έλικα (alpha helix), β-πτυχωτή επιφάνεια (β-pleated sheet), στροφή (turn).
- **Τριτοταγής Δομή:** δημιουργείται υπό την επίδραση αλληλεπιδράσεων μεταξύ των αμινοξέων (ηλεκτροστατικές δυνάμεις. Δυνάμεις Van der Waals, υδρόφοβες αλληλεπιδράσεις, υδρογονικοί δεσμοί) και αντιστοιχεί στην τρισδιάστατη δομή ή στερεοδιάταξη των ατόμων.

- **Τεταρτοταγής Δομή:** είναι ο συνδυασμός των επιμέρους πολυπεπτιδικών αλυσίδων σε ένα ενιαίο πρωτεϊνικό μόριο. Κατά τη δημιουργία της, αμινοξέα απομακρυσμένα στην ακολουθία πλησιάζουν στις τρεις διαστάσεις, για να φτιάξουν μια λειτουργική περιοχή, ένα ενεργό κέντρο.



Εικόνα 2.5 Οργάνωση των δομών των πρωτεϊνών

2.2.3.3 Στοιχεία δευτεροταγούς δομής των πρωτεϊνών.

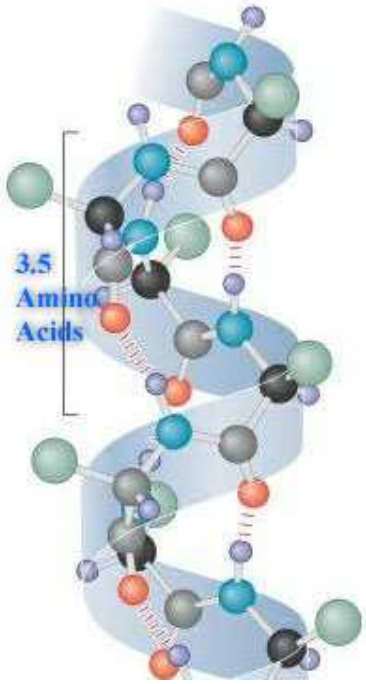
Η αμινοξική αλληλουχία μιας πρωτεϊνικής πολυπεπτιδικής αλυσίδας ονομάζεται πρωτοταγής δομή. Διαφορετικές περιοχές της αλληλουχίας δημιουργούν τοπικές κανονικές δευτεροταγείς δομές, όπως α-έλικες ή β-κλώνους. Η τριτοταγής δομή δημιουργείται με το πακετάρισμα τέτοιων δομικών στοιχείων σε μία και αποτελείται από μία ή περισσότερες συμπαγείς μονάδες που ονομάζονται επικράτειες (domains). Η τελική πρωτεΐνη μπορεί να περιλαμβάνει διάφορες πολυπεπτιδικές αλυσίδες διευθετημένες σε αλυσίδες, την τεταρτοταγή δομή. Κατά τη δημιουργία μιας τέτοιας τριτοταγούς και τεταρτοταγούς δομής, οι πολυπεπτιδικές αλυσίδες μπορούν να αναδιπλώνονται σε κανονικές επαναλαμβανόμενες δομές οι οποίες ορίζουν την ***δευτεροταγή δομή*** των πρωτεϊνών. Οι κυριότερες από αυτές είναι:

(i) α-Έλικα

(ii) β-Πτυχωτή Επιφάνεια

2.2.3.3.1 α-έλικα

Η α-έλικα είναι ένα από τα κύρια στοιχεία δευτεροταγούς δομής των πρωτεϊνών. Είναι δεξιόστροφη και προκύπτει από τη σύνδεση της αμινομάδας του i-οστού καταλοίπου και του οξυγόνου του 4i-οστού καταλοίπου, με δεσμούς υδρογόνου. Μία τυπική δομή α-έλικας έχει τουλάχιστον 5 κατάλοιπα (συνήθως 10) που κάνουν περίπου 3 στροφές. Έχει 3,6 αμινοξικά κατάλοιπα ανά στροφή, αριθμός που αντιστοιχεί σε βήμα 5.4 Å. Αυτό σημαίνει ότι κάθε κατάλοιπο απέχει από το πλησιέστερό του, 1.5 Å. Η ακτίνα r της έλικας είναι 2,3 Å. Η εντροπία σχηματισμού της είναι μεγάλη και αξίζει να σημειωθεί ότι δεν συμμετέχουν όλα τα αμινοξέα στο σχηματισμό της α-έλικας με την ίδια επιτυχία (6).



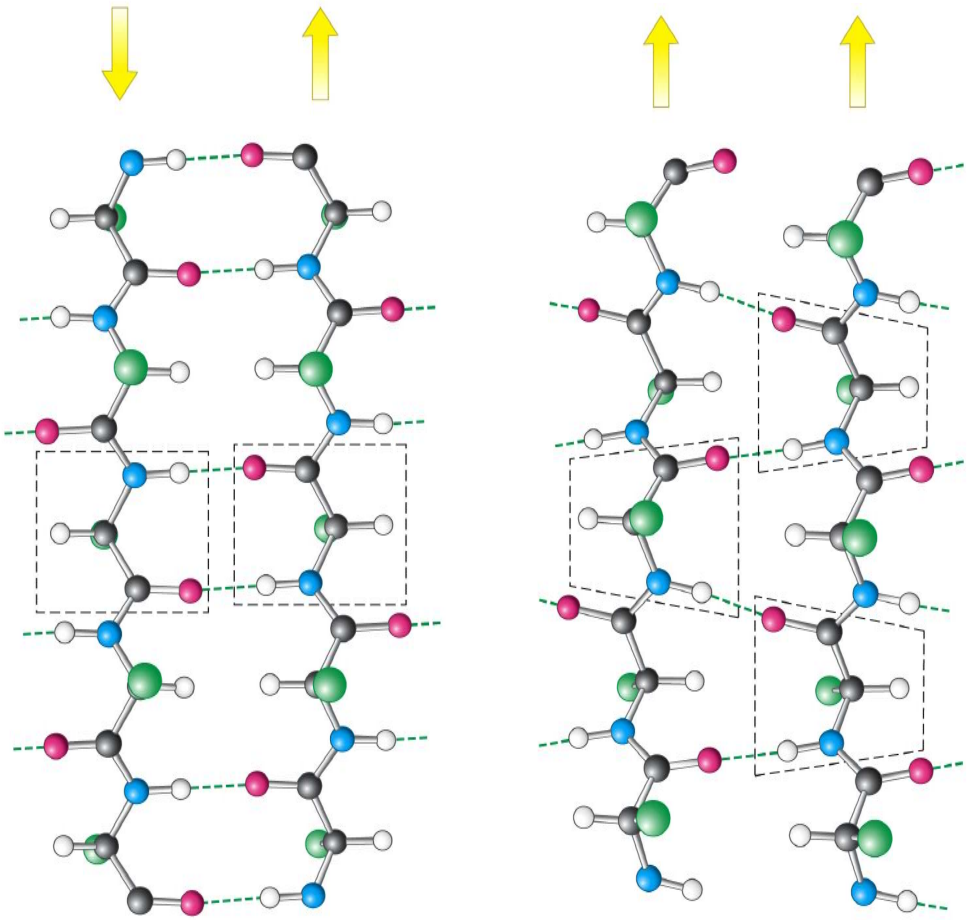
Εικόνα 2.6 Γραφική αναπαράσταση α-έλικας

Πολύ συχνά οι πλευρικές αλυσίδες των καταλοίπων που περιέχονται στις α-έλικες είναι υδρόφιλες ή υδρόφοβες με συγκεκριμένη περιοδικότητα. Επειδή η ιδιότητα αλλάζει συνήθως ανά 3 ή 4 κατάλοιπα και όπως αναφέρθηκε προηγουμένως, υπάρχουν 3,6 κατάλοιπα ανά στροφή, οι α-έλικες καταλήγουν να έχουν μία υδρόφοβη και μία υδρόφιλη επιφάνεια. Υπάρχουν και αριστερόστροφες α-έλικες αλλά δεν απαντώνται σε πρωτεΐνες επειδή είναι πολύ ασταθείς. Οι α-έλικες απαντώνται πολύ συχνά σε πρωτεΐνες μεμβρανών αλλά και σε δομικές και υδατοδιαλυτές πρωτεΐνες.

2.2.3.3.2 β-πτυχωτές επιφάνειες

Οι β-πτυχωτές επιφάνειες αποτελούν το δεύτερο κύριο στοιχείο δευτεροταγούς δομής στις πρωτεΐνες. Για να σχηματιστούν οι β-πτυχωτές επιφάνειες προηγείται ο σχηματισμός υδρογονικών δεσμών. Αυτοί οι δεσμοί μπορεί να σχηματιστούν είτε μεταξύ των αμινομάδων και καρβοξυλομάδων διαφορετικών αλυσίδων αμινοξέων, είτε μπορεί να σχηματιστούν μεταξύ τμημάτων της ίδιας αμινοξικής ακολουθίας. Αυτά τα τμήματα ονομάζονται β-κλώνοι (β-strands). Οι β-κλώνοι δεν είναι εκτεταμένοι ώστε να ευνοείται ο σχηματισμός ισχυρών δεσμών υδρογόνου, με αποτέλεσμα να σχηματίζεται στην επιφάνειά τους πτύχωση. Από το γεγονός αυτό προέρχεται και η ονομασία β-πτυχωτές επιφάνειες.

Οι β-κλώνοι μπορούν να είναι όλοι αντιπαράλληλοι ή όλοι παράλληλοι αλλά και σε κάποιες περιπτώσεις μεικτοί. Όταν οι β-κλώνοι είναι αντιπαράλληλοι, σχηματίζεται η αντιπαράλληλη πτυχωτή επιφάνεια, μίας και οι γειτονικές β-πτυχώσεις έχουν αντίθετες κατευθύνσεις. Οι δεσμοί υδρογόνου μεταξύ C=O και H-N συνδέουν ένα και μόνο αμινοξύ στην γειτονική β-πτύχωση. (βλ.Εικόνα 2.7) Οι πλευρικές αλυσίδες (στην εικόνα με πράσινο χρώμα) είναι κατ'εναλλαγή άνω και κάτω από το επίπεδο της β-πτύχωσης. Όταν οι κλώνοι είναι παράλληλοι σχηματίζεται η παράλληλη επιφάνεια στην οποία οι γειτονικές β-πτυχώσεις έχουν την ίδια κατεύθυνση. Οι δεσμοί υδρογόνου μεταξύ C=O και H-N συνδέουν κάθε αμινοξύ της μίας πτύχωσης με δύο διαφορετικά αμινοξέα στην γειτονική β-πτύχωση (βλ.Εικόνα 2.7). Κοινό αποτέλεσμα της δημιουργίας και των δυο κατηγοριών των επιφανειών, είναι η σταθεροποίηση της δομής.



Εικόνα 2.7 Γραφική αναπαράσταση αντιπαράλληλης και παράλληλης β-πτυχωτής επιφάνειας

Στα τοπολογικά διαγράμματα, οι β-πτυχωτές επιφάνειες απεικονίζονται συνήθως με βέλη που δείχνουν τόσο την κατεύθυνση κάθε β-κλώνου όσο και τον τρόπο με τον οποίο συνδέονται οι κλώνοι μεταξύ τους κατά μήκος της πολυπεπτιδικής αλυσίδας.

Γειτονικοί αντιπαράλληλοι β-κλώνοι συνδέονται με βρόχους φουρκέτας. Τέτοιοι βρόχοι είναι συχνά μικροί και δεν έχουν κάποια κανονική δευτεροταγή δομή.

Εικόνα 2.8 Τοπολογικό διάγραμμα αντιπαράλληλης β-πτυχωτής επιφάνειας

2.2.3.3.3 Υπερδευτεροταγείς δομές

Εκτός από τις δευτεροταγείς δομές υπάρχουν και οι υπερδευτεροταγείς δομές οι οποίες είναι κανονικοί συνδυασμοί δευτεροταγών δομών, δηλαδή α-ελίκων και β-πτυχωτών επιφανειών. Οι υπερδευτεροταγείς δομείς χωρίζονται σε 4 κατηγορίες.

- **αα**: δυο α-έλικες με αντιπαράλληλο προσανατολισμό.
- **β-μαίανδρος**: αντιπαράλληλη β-πτυχωτή επιφάνεια, στο εσωτερικό της οποίας υπάρχουν β-στροφές.
- **βαβ**: παράλληλη β-πτυχωτή επιφάνεια, στο εσωτερικό της οποίας βρίσκεται μία α-έλικα.
- **ελληνικό κλειδί**: αντιπαράλληλη β-πτυχωτή επιφάνεια που αναδιπλώνεται με τον εαυτό της.

Οι α-έλικες βρίσκονται συνήθως στην εξωτερική πλευρά των πρωτεϊνών σε αντίθεση με τις β-πτυχωτές επιφάνειες που βρίσκονται βαθιά στο εσωτερικό τους.
Και οι δυο δομές είναι το ίδιο σημαντικές και συνηθισμένες στις πρωτεΐνες. Για παράδειγμα στις σφαιρικές υδατοδιαλυτές πρωτεΐνες συναντάμε τις α-έλικες με συχνότητα περίπου 30% ενώ τις β-πτυχωτές επιφάνειες με συχνότητα περίπου 25 %. Οι β-πτυχωτές επιφάνειες είναι πιο ισχυρές και σταθερές δομές από τις α-έλικες. Αξίζει να σημειωθεί ότι πρωτεΐνες με παρόμοιες διαμορφώσεις συνήθως σχετίζονται με μία συγκεκριμένη λειτουργία.

2.3 *Έννοιες από την επιστήμη των υπολογιστών*

2.3.1 Ταξινομητές

Το πρόβλημα αναγνώρισης διπλώματος πρωτεϊνών αντιμετωπίστηκε ως πρόβλημα ταξινόμησης προτύπων. Για το λόγο αυτό θεωρήθηκε σκόπιμη μία αναφορά στους ταξινομητές, οι οποίοι γενικά αποτελούν χρήσιμο αντικείμενο στα χέρια των επιστημόνων πολλών πεδίων έρευνας. Οι ταξινομητές βασίζονται στην εξέταση των χαρακτηριστικών ενός αντικειμένου, το όποιο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Η εργασία της κατηγοριοποίησης απαιτεί καλό ορισμό των κατηγοριών (κλάσεων) και το σύνολο, που χρησιμοποιείται για την εκπαίδευση του μοντέλου, να αποτελείται από προκατηγοριοποιημένα παραδείγματα. Ο αριθμός των κλάσεων συνήθως είναι περιορισμένος (7). Σκοπός είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να κατηγοριοποιήσει δεδομένα που δεν έχουν ακόμα κατηγοριοποιηθεί. Οι πιο γνωστές τεχνικές που χρησιμοποιούνται στην ταξινόμηση είναι:

- Δέντρα Αποφάσεων (Decision Trees)
- Νευρωνικά Δίκτυα (Neural Networks)
- Κ-πλησιέστερων γειτόνων (k-Nearest Neighbors, k-NN)
- k-means
- Μηχανές Υποστήριξης Διανυσμάτων (Support Vector Machines, SVM)
- Bayesian μέθοδοι

Όλες οι παραπάνω τεχνικές στηρίζονται στην ιδέα της «εκπαίδευσης» με τη βοήθεια ενός υποσυνόλου δεδομένων (σύνολο εκπαίδευσης). Το μοντέλο κατασκευάζεται περιγράφοντας ένα προκαθορισμένο σύνολο από κατηγορίες δεδομένων. Στη συνέχεια τα δεδομένα εκπαίδευσης αναλύονται από έναν αλγόριθμο κατηγοριοποίησης για να κατασκευάσουν το μοντέλο. Τα δεδομένα εκπαίδευσης επιλέγονται τυχαία από ένα μεγαλύτερο πληθυσμό και ανήκουν σε μία από τις προκαθορισμένες κατηγορίες. Η κατηγορία στην οποία ανήκουν τα δεδομένα εκπαίδευσης θεωρείται γνωστή και το βήμα αυτό λέγεται «εποπτευόμενη μάθηση». Στη συνέχεια χρησιμοποιούνται τα δοκιμαστικά δεδομένα για να υπολογίσουν την ακρίβεια του μοντέλου. Αν η ακρίβεια είναι αποδεκτή το μοντέλο χρησιμοποιείται για

κατηγοριοποίηση μελλοντικών δεδομένων των οποίων η κατηγοριοποίηση είναι άγνωστη.

Δύο από τις σημαντικότερες μεθόδους ταξινόμησης προτύπων βασίζονται σε στατιστικές μεθόδους και στα τεχνητά νευρωνικά δίκτυα (8).

2.3.1.1 Στατιστικές Μέθοδοι:

Στη στατιστική προσέγγιση κάθε πρότυπο περιγράφεται από ένα διάνυσμα *n*-χαρακτηριστικών και θεωρείται ότι αποτελεί σημείο του *n*-διάστατου χώρου. Με βάση έναν αριθμό προτύπων, που είναι γνωστό σε ποια κατηγορία ανήκουν, χωρίζεται ο *n*-διάστατος χώρος σε αντίστοιχες περιοχές. Το κάθε πρότυπο προς ταξινόμηση αντιστοιχίζεται σε μία από αυτές τις κατηγορίες με βάση τη στατιστική ανάλυση των χαρακτηριστικών ως προς τα χαρακτηριστικά των *a-priori* δοσμένων προτύπων. Παραδείγματα στατιστικών μεθόδων αποτελούν οι γραμμικοί ταξινομητές και οι ταξινομητές κοντινής γειτονίας.

Επειδή στο πρόβλημά μας χρησιμοποιούμε ταξινομητή κοντινής γειτονίας, γίνεται μία εκτενέστερη αναφορά σε αυτό: Η ταξινόμηση κοντινής γειτονίας είναι μία στατιστική μέθοδος επιβλεπόμενης ταξινόμησης, δηλαδή είναι γνωστές οι κατηγορίες του προβλήματος και ένας αριθμός προτύπων που ανήκουν σε αυτές. Ο αλγόριθμος της μεθόδου ταξινομεί κάθε πρότυπο σε εκείνη τη κατηγορία από την οποία απέχει λιγότερο, με βάση μία απόσταση, όπως για παράδειγμα, Ευκλείδεια, Τετραγωνική και Hamming απόσταση. Η μέθοδος αυτή μπορεί να έχει τη μορφή της ταξινόμησης του κοντινότερου γείτονα ή των k-κοντινότερων γειτόνων.

2.3.1.2 Τεχνητά Νευρωνικά Δίκτυα:

Τα Τεχνητά Νευρωνικά Δίκτυα ή απλά Νευρωνικά Δίκτυα (ΝΔ) μπορούν να θεωρηθούν συστήματα παράλληλης σύνδεσης μεγάλου αριθμού απλών επεξεργαστών συνδεμένων μεταξύ τους. Κάθε επεξεργαστής αντιστοιχεί σε ένα τεχνητό νευρώνα, η λειτουργία του οποίου βασίζεται στις αρχές του βιολογικού νευρώνα του νευρικού συστήματος. Τα περισσότερα ΝΔ έχουν την ιδιότητα να εκπαιδεύονται, μαθαίνοντας από την εμπειρία, και να ταξινομούν τα εφαρμοζόμενα στην είσοδο τους δεδομένα. Τα σημαντικότερα ΝΔ που χρησιμοποιούνται για ταξινόμηση είναι τα πολυεπίπεδα ΝΔ πρόσθιας τροφοδότησης (Multilayer Feed-Forward Neural Networks-MLF NN) και τα ΝΔ συναρτήσεων ακτινικής βάσης (Radial Basis Function-RBF), τα οποία

αναλύονται εκτενώς στη συνέχεια, αλλά και τα δίκτυα ανταγωνιστικής μάθησης (Competitive learning NN), ή τα δίκτυα αυτοοργανούμενου χάρτη (Kohonen-Network ή Self-Organizing Map-SOM).

2.3.2 Γράφοι - Βιολογικά Δίκτυα

2.3.2.1 Θεωρία γράφων

Με τον όρο γράφος εννοείται μία συλλογή σημείων και γραμμών που συνδέουν ένα μέρος αυτών. Τα σημεία ονομάζονται κόμβοι ή διανύσματα (nodes/vertices) και οι γραμμές ακμές (edges). Ο γράφος συνήθως σημειώνεται σαν G ή σαν G (V,E), όπου V είναι το σύνολο των κόμβων και $E \subseteq V \times V$ και E το σύνολο των ακμών. Συχνά χρησιμοποιείται το n για την αντιπροσώπευση του |V| και m για την αντιπροσώπευση του |E|. Μια άλλη παραλλαγή είναι η χρησιμοποίηση του V (G) για την αντιπροσώπευση των κόμβων και E (G) για τις ακμές. Ο γράφος είναι μη κατευθυνόμενος, αν οι ακμές είναι μη κατευθυνόμενες, στην αντίθετη περίπτωση είναι κατευθυνόμενος. Κόμβοι που συνδέονται από μία ακμή λέγονται γειτονικοί κόμβοι. Το σύνολο των γειτόνων του κόμβου u (η γειτονιά του) σημειώνεται σαν N(u) και η κοντινότεροι γείτονες του αποτελούν το σύνολο N[u], το οποίο ορίζεται σαν N[u]= N(u){u}. Ο βαθμός (degree) ενός κόμβου είναι ο αριθμός των ακμών που παρουσιάζει ο κόμβος. Στους κατευθυνόμενους γράφους, εσωτερικός βαθμός (in-degree) είναι ο αριθμός των ακμών που καταλήγουν σε ένα κόμβο, ενώ ο εξωτερικός βαθμός (out-degree) είναι ο αριθμός των ακμών που ξεκινούν από αυτόν τον κόμβο.

Ο γράφος θεωρείται συμπληρωμένος όταν υπάρχει ακμή μεταξύ όλων των ζευγαριών κόμβων και ονομάζεται κλίκα (clique). Μονοπάτι (P_k) σε έναν γράφο είναι η αλληλουχία κόμβων και ακμών, έτσι ώστε κάθε κόμβος να ανήκει στις ακμές πριν και μετά από αυτόν και οι κόμβοι δεν επαναλαμβάνονται μέσα στο γράφο. Μήκος μονοπατιού είναι ο αριθμός των ακμών μέσα σε αυτό και το κοντινότερο μονοπάτι μεταξύ δυο κόμβων u και v σημειώνεται σαν d(u,v). Υπογράφος ενός γράφου G ορίζεται το σύνολο των κόμβων και των ακμών που ανήκουν στο G. Το ελάχιστο όριο ακμών ενός γράφου G είναι το σύνολο S, με το S ∪- 36 - να έχει το ελάχιστο μέγεθος από όλα τα σύνολα ακμών που μπορούν να αποσυναρμολογήσουν το γράφο αν απομακρυνθούν. Συνδετικότητα ακμών (edge connectivity) είναι ο ελάχιστος αριθμός ακμών, που αν απομακρυνθεί θα αποσυναρμολογήσει το γράφο. Τέλος, ένας

γράφος λέγεται ζυγισμένος αν υπάρχει μία συνάρτηση βάρους μεταξύ ακμών ή κόμβων και ανάλογα μπορεί να οριστεί το ελάχιστο ζυγισμένο όριο ακμών.

Υπάρχουν πολλοί τρόποι για να αποθηκευθεί ένας γράφος σε ένα υπολογιστικό σύστημα. Αυτό εξαρτάται από τη δομή των δεδομένων του γράφου αλλά και του αλγόριθμου που χειρίζεται τα δεδομένα αυτά. Εδώ γίνεται αναφορά στους δύο πιο διαδεδομένους τρόπους, τις λίστες και τους πίνακες. Οι λίστες χρησιμοποιούνται κυρίως σε αραιούς γράφους μιας και έχουν μικρές απαιτήσεις σε επίπεδο μνήμης, ενώ οι πίνακες ενδείκνυνται για εφαρμογές με μεγάλους γράφους και απαιτείται μεγάλη χωρητικότητα μνήμης.

- **Λίστες:**

Λίστα εμφάνισης (Incidence list): οι ακμές αντιπροσωπεύονται από μία συστοιχία που περιλαμβάνει ζευγάρια κόμβων, πιθανά βάρη και άλλα δεδομένα.

Λίστα γειτονιάς (Adjacency list): κάθε κόμβος έχει μία λίστα όλων των κόμβων, με τους οποίους είναι γείτονας.

- **Πίνακες:**

Πίνακας εμφάνισης (Incidence matrix): ο γράφος αντιπροσωπεύεται από ένα πίνακα με E (ακμές) επί V κόμβους (vertices), όπου [ακμή, κόμβος] περιλαμβάνει τα δεδομένα της ακμής (απλούστερη περίπτωση: 1 – συνδεδεμένο, 0 – όχι συνδεδεμένο).

Laplacian ή Kirchhoff ή Admittance Πίνακας: ορίζεται σαν τη διαφορά μεταξύ πίνακα βαθμών και πίνακα γειτονιάς και γι'αυτό παρέχει δύο ειδών πληροφορίες για τους κόμβους.

Πίνακας γειτονίας (Adjacency matrix): ορίζεται σαν ένας N×N πίνακας, όπου N ο αριθμός των κόμβων στο γράφο. Αν υπάρχει ακμή μεταξύ του κόμβου χ και του κόμβου y, τότε το στοιχείο $M_{x,y}$ είναι 1, αλλιώς 0

Συχνά συνδυάονται οι μέθοδοι απεικόνισης λίστα και πίνακα. Μία δομή που βασίζεται στη θεωρία των γράφων, απλουστεύοντάς την, είναι τα **δίκτυα**. Τα δίκτυα παρέχουν ένα διαισθητικό και οπτικό τρόπο για την οργάνωση μεγάλων ποσοτήτων πληροφορίας. Σε ένα βιολογικό δίκτυο μπορεί να αποκαλυφθεί τόσο η γενική εικόνα του δικτύου όσο και τοπικές λεπτομέρειες (9).

Με τον όρο δίκτυο εννοούμε έναν μεγάλο αριθμό κόμβων (που μπορεί να αναπαριστούν γονίδια, μεταβολίτες ή πρωτεΐνες) και ακμές που δείχνουν τη σχέση

μεταξύ των κόμβων. Υπάρχουν δυο τύποι περιγραφής των δικτύων. Στον πρώτο τύπο η ακμή αναπαριστά φυσική αλληλεπίδραση μεταξύ των κόμβων όπως συμβαίνει στα δίκτυα αλληλεπίδρασης πρωτεϊνών. Στα μεταβολικά δίκτυα οι μεταβολίτες συνδέονται επειδή συμμετέχουν στην ίδια ενζυματική αντίδραση. Ο δεύτερος τρόπος περιγραφής είναι ο θεωρητικός, όπου για παράδειγμα η ακμή σε ένα δίκτυο μπορεί να υποδηλώνει το βαθμό ομοιότητας μεταξύ των κορυφών

2.3.3 Γενετικοί Αλγόριθμοι

Όταν έχουμε προβλήματα ταξινόμησης κάποιων δεδομένων σε κατηγορίες -όπως εδώ πρωτεΐνες σε ομάδες ίδιας δομής- για την ταξινόμηση των οποίων χρησιμοποιούμε σύνολα χαρακτηριστικών των ίδιων των δεδομένων, προσπαθούμε να μειώσουμε τη διάσταση του συνόλου των χαρακτηριστικών. Χρησιμοποιώντας μικρά σύνολα χαρακτηριστικών πετυχαίνουμε καλύτερα αποτελέσματα κατηγοριοποίησης αλλά και ελαχιστοποιούμε το κόστος των ταξινομητών. Ταυτόχρονα μπορούμε να αναγνωρίσουμε τα σχετικά βιολογικά χαρακτηριστικά για το δοθέν πρόβλημα. Αναζητούνται, λοιπόν, τα πιο εύρωστα (robust) χαρακτηριστικά των προς ταξινόμηση, δεδομένων, αυτά, που κάνουν τη διάκριση (separability) των κλάσεων πιο σαφή.

Οι Γενετικοί Αλγόριθμοι (Genetic Algorithms) είναι ένα πολύ χρήσιμο εργαλείο για το σκοπό αυτό. Οι ΓΑ είναι εμπνευσμένοι από τη Δαρβινική θεωρία της εξέλιξης της φύσης. Γενικά, εκτελούν μία αναζήτηση στο χώρο των υποψήφιων λύσεων με σκοπό την εύρεση αποδεκτών, σύμφωνα με κάποιο κριτήριο, λύσεων.

Αναλυτικότερα η φιλοσοφία των ΓΑ είναι η εξής: Αρχικά δημιουργείται με τυχαίο τρόπο ένα σύνολο από Π υποψήφιες λύσεις του προβλήματος. Αξίζει να σημειωθεί ότι το αρχικό αυτό σύνολο απαρτίζεται από λύσεις οι οποίες είναι επί το πλείστον μη αποδεκτές (είτε επειδή δεν είναι βέλτιστες, είτε επειδή δεν είναι έγκυρες). Έστω N το πλήθος των στοιχείων του συνόλου Π. Οι λύσεις αυτές βαθμολογούνται από μία συνάρτηση καταλληλότητας (fitness function). Η βαθμολόγησή τους συνίσταται στην αντιστοίχηση κάθε υποψήφιας λύσης σε έναν αριθμό, ο οποίος δηλώνει την εγγύτητα της υποψήφιας μη αποδεκτής λύσης, ως προς κάποια αποδεκτή (10).

Στη συνέχεια από τον αρχικό πληθυσμό, προκύπτει ένας νέος πληθυσμός Π΄ με χρήση κατάλληλων λειτουργιών που επιδρούν στα άτομα του αρχικού πληθυσμού. Ο νέος πληθυσμός Π΄, αποτελείται από το σύνολο των απογόνων και συνήθως αποτελεί

βελτίωση του προηγούμενου πληθυσμού. Η διαδικασία επαναλαμβάνεται για το νέο πληθυσμό Π΄, ενώ οι πιο συνηθισμένες συνθήκες τερματισμού είναι η εύρεση μιας τέλειας λύσης με βάση τη συνάρτηση καταλληλότητας ή η σύγκλιση όλων των λύσεων σε μία. Ο γενετικός αλγόριθμος εκτελεί μία αναζήτηση στο χώρο των υποψηφίων λύσεων, με στόχο την εύρεση κάποιας λύσης που μεγιστοποιεί τη συνάρτηση καταλληλότητας. Η αναζήτηση αυτή είναι παράλληλη, καθώς σε κάθε υποψήφια λύση, μπορεί να εκτελεστεί ξεχωριστή αναζήτηση. Γίνονται μικρές αλλαγές στις υποψήφιες λύσεις του πληθυσμού και επιλέγονται πάντα οι καλύτερες, βάσει της συνάρτησης καταλληλότητας. Η αναζήτηση επικεντρώνεται στις περισσότερο κατάλληλες λύσεις, χωρίς όμως να αγνοούνται οι υπόλοιπες, καθώς υπάρχει πάντα ο κίνδυνος να παγιδευτεί η διαδικασία σε τοπικό μέγιστο (local maximum).

Λαμβάνοντας υπόψη τη μεταφορά εννοιών από τη Βιολογία στο χώρο της υπολογιστικής νοημοσύνης, το σύνολο των πιθανών λύσεων αναφέρεται ως πληθυσμός. Η κάθε υποψήφια λύση αναφέρεται ως χρωμόσωμα, που στην κλασσική προσέγγιση των ΓΑ αναπαρίσταται με μία συμβολοσειρά (string) ενός πεπερασμένου αλφαβήτου. Τα επιμέρους τμήματα που κωδικοποιούν κάποιο χαρακτηριστικό ονομάζονται γονίδια. Η διασταύρωση (crossover) αναφέρεται στην ανταλλαγή τμήματος της αλυσίδας μεταξύ των υποψήφιων λύσεων και η μετάλλαξη (mutation) συνίσταται στην αλλαγή του δυαδικού ψηφίου μιας τυχαίας επιλεγείσας θέσης με ένα επίσης τυχαία επιλεγμένο δυαδικό ψηφίο. Το σύνολο των παραμέτρων που αναπαρίστανται από ένα συγκεκριμένο γονίδιο που μας ενδιαφέρει αναφέρεται σαν γονότυπος (genotype). Επίσης, όπως στην Βιολογία, η προσαρμογή ή ποιότητα ενός οργανισμού (fitness) ορίζεται τυπικά ως η πιθανότητα που έχει ο οργανισμός να αναπαραχθεί, στους ΓΑ η προσαρμογή ή ποιότητα μιας λύσης είναι η πιθανότητα της να είναι η νικητήρια σε ένα ανταγωνιστικό περιβάλλον. Από τα άτομα του πληθυσμού, τα οποία διασταυρώνονται μεταξύ τους αλλά και μεταλλάσσονται, προκύπτουν οι απόγονοι (offsprings), οι οποίοι συμμετέχουν σε νέες διασταυρώσεις, υπόκεινται σε νέες μεταλλάξεις και δίνουν νέους απογόνους. Τελικά μετά την ολοκλήρωση ενός αριθμού γενεών προκύπτει ο απόγονος που αντιστοιχεί στη βέλτιστη λύση του προβλήματος.

Η απλούστερη μορφή ΓΑ περιλαμβάνει τρεις τελεστές: επιλογή (selection) , διασταύρωση (crossover) και μετάλλαξη (mutation). Οι τελεστές αυτοί περιγράφονται παρακάτω (11) :

2.3.3.1 Επιλογή

Ο τελεστής αυτός επιλέγει χρωμοσώματα στον πληθυσμό προς αναπαραγωγή. Όσο μεγαλύτερη είναι η ποιότητα του χρωμοσώματος (fitter), τόσο αυξάνεται η πιθανότητα του να επιλεγεί περισσότερες φορές προς αναπαραγωγή.

2.3.3.2 Διασταύρωση

Ο τελεστής αυτός επιλέγει τυχαία μία θέση και ανταλλάσσει τις υποακολουθίες πριν και μετά από αυτή τη θέση μεταξύ δύο χρωμοσωμάτων για να παραγάγει δύο απογόνους. Για παράδειγμα, οι ακολουθίες 10000100 και 11111111 μπορούν να διασταυρωθούν μετά την τρίτη θέση κάθε μίας για να δώσουν τους δύο απογόνους: 10011111 και 11100100. Ο τελεστής διασταύρωσης μιμείται χονδρικά τη βιολογική διασταύρωση μεταξύ δύο απλοειδών (μονού χρωμοσώματος) οργανισμών.

2.3.3.3 Μετάλλαξη

Ο τελεστής αυτός, που εφαρμόζεται μετά τη διασταύρωση, αλλάζει τυχαία μερικά από τα ψηφία ενός χρωμοσώματος. Για παράδειγμα, η ακολουθία 0000100 θα μπορούσε να μεταλλαχθεί στη δεύτερη θέση, οπότε το αποτέλεσμα θα ήταν 0100100. Η μετάλλαξη μπορεί να λάβει χώρα σε κάθε θέση του χρωμοσώματος με κάποια πιθανότητα, που είναι συνήθως πολύ μικρή (π.χ. 0.001).

Οι γενετικοί αλγόριθμοι έχουν εφαρμοστεί σε πλήθος εφαρμογών. Μερικές από αυτές είναι οι ακόλουθες:

Βελτιστοποίηση: Αριθμητική βελτιστοποίηση και συνδυαστική βελτιστοποίηση σε εφαρμογές, όπως σχεδίαση κυκλωμάτων και χρονοπρογραμματισμός εργασιών.

Αυτόματος προγραμματισμός: Προβλήματα εξέλιξης υπολογιστικών προγραμμάτων για ειδικές εργασίες και σχεδίαση άλλων υπολογιστικών δομών, όπως κυτταρικά αυτόματα και δίκτυα κατάταξης (sorting).

Μάθηση μηχανής: Προβλήματα μηχανικής μάθησης για εργασίες όπως ταξινόμηση, σύντηξη σημάτων, εκτίμηση και πρόβλεψη. Επίσης προβλήματα βέλτιστης επιλογής παραμέτρων βιομηχανικών ελεγκτών, ρομποτικών συστημάτων, νευρωνικών δικτύων, ασαφών συστημάτων κλπ.

Οικονομικά Συστήματα: Προβλήματα μοντελοποίησης οικονομικών διεργασιών, διεργασιών καινοτομίας κλπ.

Βιολογικά συστήματα: Σπουδή προβλημάτων γενετικής των πληθυσμών, εξέλιξης των ειδών και μάθησης των ατόμων.

Οικολογικά συστήματα: Προβλήματα μοντελοποίησης οικολογικών φαινομένων, όπως ροή φυσικών πηγών, συμβίωση, ανάπτυξη παρασίτων κλπ..

Κοινωνικά συστήματα: Προβλήματα κοινωνικής συμπεριφοράς, συνεργασίας και επικοινωνίας .

Τεχνολογικά συστήματα: Συστήματα οργάνωσης, επίβλεψης και ελέγχου βιομηχανικών συστημάτων, ρομποτικά συστήματα κατασκευής και υπηρεσιών, συστήματα παραγωγής και διανομής ενέργειας, συστήματα μεταφορών και επικοινωνιών κ.λπ. (11).

3

Αναγνώριση Διπλώματος Πρωτεϊνών

Στο κεφάλαιο αυτό γίνεται μία εκτενής περιγραφή του ζητήματος της πρόβλεψης των πρωτεϊνικών δομών (Ενότητα 3.1). Επίσης παρουσιάζεται η κατηγοριοποίηση των πρωτεϊνικών υπερ-δευτεροταγών δομών (Ενότητα 3.2) και παρουσιάζονται διάφορες μέθοδοι που χρησιμοποιούνται ευρέως για την αντιμετώπιση του προβλήματος (Ενότητα 3.3).

3.1 *Πρόβλεψη Δομής Πρωτεϊνών*

Ενώ οι πρωτεϊνικές ακολουθίες εξάγονται με ταχείς ρυθμούς, η ταυτοποίηση της δομή τους στο χώρο εργαστηριακά είναι μία αργή και δαπανηρή διαδικασία. Οι μέθοδοι που χρησιμοποιούνται είναι: η κρυσταλλογραφία με ακτίνες Χ ή περίθλαση ηλεκτρονίων, ο πυρηνικός μαγνητικός συντονισμός και η φασματομετρία μάζας. Οι συγκεκριμένες μέθοδοι λειτουργούν διαφορετικά σε διαφορετικές κατηγορίες πρωτεϊνικών ομάδων. Το 1971 ιδρύθηκε στα Brookhaven National Laboratories (BNL) των ΗΠΑ η Protein Data Bank, η οποία περιελάμβανε 7 δομές μακρομορίων, όπως αυτές προέκυψαν από κρυσταλλογραφικές μελέτες. Ο ρυθμός αύξησης των εγγραφών στη δεκαετία του '70 ήταν πολύ μικρός. Από το 1980 και μετά λόγω της τεχνολογικής εξέλιξης σε κάθε στάδιο του προσδιορισμού δομών ο ρυθμός προσθήκης δεδομένων στην Protein Data Bank αυξάνεται δραματικά αλλά πάντα ο αριθμός των πρωτεϊνών των οποίων είναι γνωστή η δομή, παραμένει πολύ μικρός.

Αναλυτικότερα, η βάση δεδομένων των γνωστών πρωτεϊνών Protein Data Bank περιέχει 6.010.749 εγγραφές (12) ενώ η δομή είναι γνωστή μόνο για τις 47.873 από αυτές (13). Κάθε μήνα προστίθενται στη βάση πρωτεϊνών 200 περίπου πρωτεΐνες.

Την ίδια στιγμή γνωρίζουμε ότι ορισμένοι συνδυασμοί στοιχείων δευτεροταγούς δομής, με συγκεκριμένη γεωμετρική διευθέτηση, - αποτελούν την υπερ-δευτεροταγή δομή - συσχετίζονται άμεσα με συγκεκριμένες λειτουργίες. Για παράδειγμα, ο συνδυασμός helix-loop-helix αποτελεί ένα μοτίβο για πρόσδεση σε DNA. Έτσι, επιτυχής πρόγνωση των στοιχείων δευτεροταγούς δομής, συγκεκριμένα, μπορεί να βοηθήσει στην αποκάλυψη τέτοιων σχηματισμών ανώτερης τάξης που παραπέμπει, συχνά, σε καθορισμένους βιολογικούς ρόλους.

Λαμβάνοντας πάντα υπόψη και τη σημασία της δομής για τη λειτουργία της πρωτεΐνης (βλ. Ενότητα 2.2), γίνεται εύκολα αντιληπτό πόσο σημαντικό είναι να αναπτυχθούν μέθοδοι πρόγνωσης της δομής των πρωτεϊνών. Βέβαια οφείλει κανείς να είναι ιδιαίτερα προσεκτικός όταν προσπαθεί να επεκτείνει τα συμπεράσματα του και στη λειτουργία των πρωτεϊνών. Αν και έχει αποδειχθεί ότι πρωτεΐνες με πολύ μικρή ομολογία (μικρότερη από 30%) σε επίπεδο αμινοξικής ακολουθίας, φαίνεται να έχουν το ίδιο δίπλωμα στο χώρο και την ίδια λειτουργία, στην πράξη, η μεγάλη πλειοψηφία ζευγών πρωτεϊνών με παρόμοιες δομές παρουσιάζουν σχετικά χαμηλή ομολογία στην πρωτοταγή τους δομή. Παρόλα αυτά, είναι δυνατόν πρωτεΐνες όμοιες σε μεγάλο βαθμό να έχουν εντελώς διαφορετικούς βιολογικούς ρόλους. Χαρακτηριστικό παράδειγμα είναι η κρυσταλλίνη στο φακό του οφθαλμού (με καθαρά δομικό ρόλο) η οποία είναι εμφανές προϊόν πρόσφατης εξέλιξης από μεταβολικά ένζυμα. Επίσης, σε αρκετές περιπτώσεις, δομική ή ακόμη και λειτουργική ομοιότητα εμφανίζεται για περιορισμένα τμήματα πρωτεϊνών. Για παράδειγμα η ύπαρξη προτύπου "δακτύλου ψευδαργύρου" (Zink-finger motif), είναι μία καλή ένδειξη πως η συγκεκριμένη περιοχή είναι δυνατό να προσδένεται σε DNA ή RNA, δε μπορούμε, όμως, να αποκλείσουμε απόλυτα την πιθανότητα ώστε, συνολικά, η πρωτεΐνη να συμμετέχει σε κάποια άλλη βιολογική λειτουργία.

3.2 *Κατηγοριοποίηση Υπερ-δευτεροταγών δομών*

Από το 1959 που προσδιορίσθηκε η πρώτη τρισδιάστατη δομή, η δομή της μυοσφαιρίνης μέχρι τώρα, η πολυπλοκότητα και η ποικιλομορφία των πρωτεϊνικών δομών αυξάνεται παράλληλα με τον αριθμό των νέων προσδιοριζόμενων μακρομορίων. Έτσι, από τότε μέχρι σήμερα, έχει αναγνωρισθεί ότι οι πρωτεΐνες αποτελούνται από αυτοτελή ημιανεξάρτητα δομικά στοιχεία που μπορούν να είναι συνεχή ή μη συνεχή στην πρωτεϊνική ακολουθία . Έτσι, προφανώς, το πρώτο βήμα για την κατάταξη των πρωτεϊνών σε τάξεις, κλάσεις, οικογένειες κλπ είναι η ταυτοποίηση των δομικών στοιχείων τους.

Η σημασία της κατάταξης των πρωτεϊνών σε κατηγορίες δεν έγκειται μόνο στο να ερμηνεύσουμε καλύτερα τις σχέσεις μεταξύ αλληλουχίας, δομής και λειτουργίας αλλά και τις βαθύτερες εξελικτικές διαδικασίες, που έχουν οδηγήσει στην δημιουργία αυτών των ποικιλόμορφων δομικών σχηματισμών.

Οι δομικές ομοιότητες των πρωτεϊνών πολύ συχνά μαρτυρούν την εξελικτική τους προέλευση. Στην πορεία της εξέλιξης έλαβαν χώρα διάφορες προσθήκες, αφαιρέσεις και αντικαταστάσεις αμινοξέων. Αναφορικά με τις πρωτεΐνες, που η εξελικτική τους σχέση ξεκινά στο μακρινό παρελθόν, οι αλλαγές αυτές μπορεί να είναι τόσες πολλές που να μην έχουν παρόμοια δομή αλλά να διατηρούν την παρόμοια αρχική τους λειτουργία. Αυτό μπορεί να συμβεί επειδή τα αμινοξέα στα ενεργά κέντρα παρέμειναν αναλλοίωτα. Άρα είναι ξεκάθαρο πως πρέπει να είμαστε ιδιαίτερα προσεκτικοί στην κατηγοριοποίηση των πρωτεϊνών, για να αποφευχθεί η εξαγωγή λάθος συμπερασμάτων.

Στη συνέχεια περιγράφονται δυο από τα πιο διαδεδομένα συστήματα κατηγοριοποίησης πρωτεϊνικών δομών (6).

3.2.1 SCOP (Structural Classification Of Proteins database)

Η Βάση Πρωτεϊνικών Δομών (SCOP - Structural Classification of Proteins database for the investigation of sequences and structures.) περιγράφει τις δομικές και εξελικτικές σχέσεις μεταξύ των πρωτεϊνών γνωστής δομής κατατεθειμένων στην Protein Data Bank (PDB). Η βάση διατηρείται από τα μέσα της δεκαετίας του 90 στο κέντρο ερευνών Μοριακής Βιολογίας και Πρωτεϊνικής Μηχανικής στο Cambridge

της Αγγλίας (Laboratory of Molecular Biology and Centre for Protein Engineering - MRC) (14).

Οι πρωτεΐνες ταξινομούνται με βάση τις σχέσεις αυτές με μη αυτοματοποιημένο τρόπο αλλά με τη συμβολή του ανθρώπινου παράγοντα. Τα βασικά επίπεδα ταξινόμησης είναι η οικογένεια (Family), η υπερ-οικογένεια (Superfamily), το δίπλωμα (Fold) και η τάξη (Class).

Οικογένεια (**Family**): Η ομοιότητα σε επίπεδο ακολουθίας ανάμεσα στα μέλη μιας οικογένειας είναι της τάξης του 30% και άνω. Υπάρχουν όμως περιπτώσεις όπου οι δομές και η λειτουργία είναι παρόμοιες υποδηλώνοντας κοινό πρόγονο ενώ η ομοιότητα σε επίπεδο ακολουθίας να είναι μικρότερη του 30% (σφαιρίνες, 15%). Η οικογένεια υποδηλώνει ξεκάθαρη εξελικτική σχέση μεταξύ των μελών της.

Υπερ-οικογένεια (Superfamily): Στο επίπεδο της υπερ-οικογένειας κατατάσσονται πρωτεΐνες που εμφανίζουν πολύ μικρή ομοιότητα στο επίπεδο της ακολουθίας, αλλά τα δομικά τους χαρακτηριστικά και η λειτουργία τους υποδηλώνουν πιθανή κοινή προέλευση.

Δίπλωμα (Fold): Οι πρωτεΐνες που εμφανίζουν το ίδιο δίπλωμα έχουν τα ίδια σε μεγάλο βαθμό χαρακτηριστικά δευτεροταγούς δομής και κοινή τοπολογία. Πρωτεΐνες που έχουν το ίδιο δίπλωμα αλλά δεν είναι όμοιες από άποψη αμινοξικής ακολουθίας, έχουν ορισμένα περιφερειακά στοιχεία της δευτεροταγούς τους δομής ανόμοια και όσον αφορά στο μέγεθος και στη διαμόρφωση. Πρωτεΐνες που εμφανίζουν κοινό δίπλωμα δεν είναι απαραίτητο να έχουν κοινή εξελικτική προέλευση.

Τάξη (Class): Τέσσερις κύριες δομικές κατηγορίες πρωτεϊνών έχουν ταυτοποιηθεί με βάση το δίπλωμα των στοιχείων δευτεροταγούς δομής τους, τις **all-α** (η δομή σχηματίζεται από α-έλικες), **all-β** (η δομή αποτελείται από β-πτυχωτές επιφάνειες), **α/β** (α-έλικες και β-πτυχωτές επιφάνειες εναλλάσσονται στην δομή της πρωτεΐνης) και **α+β** (α-έλικες και β-πτυχωτές επιφάνειες βρίσκονται σε διακριτές περιοχές της δομής).

3.2.2 CATH

Η βάση δεδομένων CATH (το όνομα της προέρχεται από τα αρχικά γράμματα των επιπέδων ταξινόμησης που ορίζει) ταξινομεί ιεραρχικά τις πρωτεϊνικές δομές που βρίσκονται στην PDB με βασικό κριτήριο τις αυτοτελείς δομικές περιοχές που τις απαρτίζουν. Επιπλέον, οι πρωτεΐνες αυτές πρέπει να έχουν προσδιοριστεί με

διακριτικότητα καλύτερη από 3 Å . Σε αντίθεση με τη SCOP, χρησιμοποιεί περισσότερο αυτοματοποιημένες μεθόδους ταξινόμησης, αλλά σε συγκεκριμένες περιπτώσεις συμβάλει και ο ανθρώπινος παράγοντας (15).

Για την ταξινόμηση χρησιμοποιούνται πέντε κύρια επίπεδα : Τάξη (Class), Αρχιτεκτονική (Architecture), Τοπολογία (Topology), Ομόλογη Υπέρ-οικογένεια (Homologous superfamily) και Αλληλουχία (Sequence).

Τάξη (Class): Η τάξη προσδιορίζεται τα στοιχεία δευτεροταγούς δομής μιας αυτοτελούς δομικής περιοχής μιας πρωτεΐνης (domain) αλλά και το πακετάρισμα τους. Τέσσερις κύριες ομάδες πρωτεϊνικών δομικών στοιχείων έχουν ταυτοποιηθεί : (i) κυρίως α (mainly-α), (ii) κυρίως β (mainly-β), (iii) α-β, που περιλαμβάνει εναλλασσόμενες α/β και α+β δομές, και (iv) δομές με χαμηλό ποσοστό δευτεροταγούς δομής.

Αρχιτεκτονική (Architecture): Η ταξινόμηση γίνεται με βάση την οργάνωση των στοιχείων δευτεροταγούς δομής, μη λαμβάνοντας όμως υπόψη τον τρόπο διασύνδεσής τους.

Τοπολογία (Topology): Η τοπολογία προσδιορίζεται από τον προσανατολισμό και τον τρόπο σύνδεσης των στοιχείων δευτεροταγούς δομής. Με βάση μία σειρά από εμπειρικές παραμέτρους και χρησιμοποιώντας αλγόριθμους που επιτυγχάνουν την ολική σύγκριση των δομικών στοιχείων, οι πρωτεΐνες ομαδοποιούνται σε διαφορετικές κατηγορίες τοπολογίας. Δομές που κατά το 60% μοιάζουν με την μικρότερη ταξινομούνται στο ίδιο επίπεδο τοπολογίας.

Ομόλογη Υπερ-οικογένεια (Homology superfamily): Σε αυτό το επίπεδο ομαδοποιούνται όλα εκείνα τα δομικά στοιχεία που έχουν 35% ομοιότητα μεταξύ τους στο επίπεδο της αλληλουχίας τους, με αποτέλεσμα να θεωρείται ότι προέρχονται από ένα κοινό πρόγονο.

Αλληλουχία (Sequence): Στο επίπεδο αυτό, τα μέλη χαρακτηρίζονται από από ομοιότητα των αλληλουχιών τους σε ποσοστό μεγαλύτερο από 30%, άρα παρόμοιες δομές και λειτουργίες.

3.3 *Μέθοδοι Πρόγνωσης Διπλώματος Πρωτεϊνών*

Το να γνωρίζουμε τη δομή μιας πρωτεΐνης είναι καίριο για να καταλάβουμε πώς λειτουργεί αλλά και να για την στοχεύσουμε με φάρμακα. Μία μικρή πρωτεΐνη μπορεί να αποτελείται από 100 αμινοξέα, αλλά οι ανθρώπινες πρωτεΐνες είναι συνήθως πολύ μεγαλύτερες και αποτελούνται μέχρι και από 1000 αμινοξέα. Ακόμα και οι μικρότερες πρωτεΐνες μπορούν να αναδιπλωθούν στο χώρο με πάρα πολλούς τρόπους γιατί έχουν πολλούς βαθμούς ελευθερίας, το οποίο είναι ευρέως γνωστό ως παράδοξο του Levinthal. Δηλαδή ότι ο αριθμός των δυνατών διαμορφώσεων στο χώρο μιας πολυπεπτιδικής αλυσίδας είναι τόσο μεγάλος, που η αξιολόγηση όλων των περιπτώσεων είναι αδύνατη (1).

Οι υπολογιστικές μέθοδοι που έχουν αναπτυχθεί μπορούν να κατηγοριοποιηθούν σε μεθόδους προτυποποίησης με ομολογία (homology modeling), ύφανσης πρωτεϊνών (protein threading) και απαρχής μεθόδους (ab initio). Οι δύο πρώτες μέθοδοι είναι ευριστικές, δηλαδή χρησιμοποιούν, ήδη προσδιορισμένες δομές για να προσδιορίσουν τη δομή μιας ακολουθίας αγνώστου δομής. Είναι προσεγγίσεις βασισμένες στη γνώση (knowledge based) που αποκτούν ολοένα και περισσότερο ενδιαφέρον. Αντίθετα οι απαρχής μέθοδοι δεν βασίζονται σε προηγούμενη γνώση, αλλά χρησιμοποιούν φυσικές αρχές (physics based methods). Όλες οι μέθοδοι πρόγνωσης της δομής των πρωτεϊνών αξιολογούνται κάθε δυο χρόνια με απαρχή το 1994, στα πλαίσια του διαγωνισμού CASP (Critical Assessment of Techniques for Protein Structure Prediction).

3.3.1 Μέθοδοι προτυποποίησης με ομολογία (homology modeling)

Η μέθοδος προτυποποίησης με ομολογία εφαρμόζεται σε μια άγνωστη δομή, αν βρεθεί ομόλογή της με γνωστή δομή. Με την προσέγγιση αυτή, προσπαθούμε να προσδιορίσουμε θεωρητικά την άγνωστη δομή μιας πρωτεΐνης. Γενικά, πρωτεΐνες με ομοιότητα στην ακολουθία άνω του 30%, η οποία εκτείνεται σε μήκος μεγαλύτερο των 80 καταλοίπων, θεωρείται πως διπλώνουν στο χώρο με παρόμοιο τρόπο. Τουλάχιστον αυτό ισχύει για τις περιοχές μη τυχαίας δευτεροταγούς δομής. Η γνωστή δομή λέγεται Template (οδηγός), ενώ η άγνωστη Model (μοντέλο).

Με τον τρόπο αυτό έγινε δυνατόν να μεγαλώσει το πλήθος των 'γνωστών' 3D-δομών από 7.000 σε περίπου 50.000. Η δυσκολία βρίσκεται στη σωστή προσαρμογή των πλευρικών αλυσίδων, πρόβλημα του οποίου η επίλυση επιβάλει τη χρήση ενεργειακών υπολογισμών, καθώς και δεδομένων για όλες τις διευθετήσεις πλευρικών αλυσίδων σε λυμένες πρωτεϊνικές δομές.

Αναλυτικότερα η διαδικασία της προτυποποίησης με ομολογία ακολουθεί τα εξής βήματα:

- Εντοπισμός της ακολουθίας του μοντέλου στις βάσεις δεδομένων
- Εύρεση μιας κατάλληλης πρωτεΐνης που να ικανοποιεί τις απαιτήσεις ομολογίας με την ακολουθία του μοντέλου
- Εντοπισμός της πειραματικά προσδιορισμένης δομής-οδηγού στις βάσεις δεδομένων και έλεγχος της αξιοπιστίας της
- Στοίχιση των δύο ακολουθιών
- Δημιουργία του μοντέλου, με ειδικό λογισμικό το οποίο δέχεται ως δεδομένα τη δομή-οδηγό και τη στοίχιση των ακολουθιών
- Οπτική παρατήρηση του μοντέλου, και έλεγχος για πιθανές ανεπίτρεπτες αποστάσεις μεταξύ ατόμων
- Ενδεχόμενη βελτίωση της στοίχισης των ακολουθιών, με σκοπό τη βελτίωση του μοντέλου
- Δημιουργία του νέου μοντέλου
- Ενεργειακή βελτιστοποίηση του μοντέλου, με χρήση κατάλληλου λογισμικού
- Έλεγχος του βελτιστοποιημένου μοντέλου
- Αξιολόγηση του μοντέλου, ταιριάζοντάς το στο χώρο με κάποια πειραματικά προσδιορισμένη δομή (τη δομή-οδηγό ή άλλη)

3.3.2 Μέθοδοι ύφανσης πρωτεϊνών (protein threading)

Οι μέθοδοι ύφανσης πρωτεϊνών χρησιμοποιούνται σε περιπτώσεις όπου δεν εμφανίζεται υψηλή ομολογία, τεχνικές 'Threading' (ή Remote Homology Modelling, όπως αλλιώς αναφέρονται) είναι δυνατόν να αποκαλύψουν 'απομακρυσμένες' ομολογίες, οι οποίες δεν ανιχνεύονται με τη χρήση συμβατικών μεθόδων στοίχισης ακολουθιών. Αυτό μπορεί να συμβεί επειδή ακολουθίες με μη εμφανή ομοιότητα μπορεί να έχουν παρόμοια αναδίπλωση, χωρίς να έχουν εξελικτική σχέση. Η εύρεση

της δομής μπορεί να αναχθεί στην αναγνώριση της αναδίπλωσης ανάμεσα σε γνωστές δομές.

Για την ύφανση μιας πρωτεΐνης αγνώστου δομής είναι απαραίτητα: (i) ένα αντιπροσωπευτικό σύνολο T που περιλαμβάνει τις δομές πρότυπα (folds template), (ii) μία συνάρτηση ενέργειας, κατάλληλη για την αξιολόγηση στοίχισης της ακολουθίας s με μία δομή t του συνόλου T, (iii) ένας αλγόριθμος ύφανσης για την εύρεση της βέλτιστης στοίχισης για κάθε ζεύγος ακολουθίας s-t, (iv) ένα κριτήριο αξιολόγησης της δομής που προβλέπεται. Η προβλεπόμενη δομή δεν περιέχει λεπτομέρειες για τις θέσεις των ατόμων των καταλοίπων, άρα μπορεί να αποτελέσει στη συνέχεια είσοδο σε εργαλείο προτυποποίησης πρωτεϊνών με ομολογία.

3.3.3 Μέθοδοι απαρχής (ab initio)

Οι απαρχής μέθοδοι δεν χρησιμοποιούν πρότυπα δομών, αλλά επιχειρούν να προβλέψουν τη δομή μιας πρωτεΐνης από την αμινοξική της ακολουθία. Στις μεθόδους αυτές κοινά απαραίτητα στοιχεία είναι: (i) κατάλληλη αναπαράσταση της πρωτεΐνης στο χώρο, (ii)κατάλληλη ενεργειακή συνάρτηση συμβατή με την αναπαράσταση της πρωτεΐνης, (iii)αποδοτικός και αξιόπιστος αλγόριθμος για την αναζήτηση του χώρου λύσεων και την ελαχιστοποίηση της ενέργεια, ώστε να προσεγγιστεί η δομή της πρωτεΐνης. Το ιδανικό θα ήταν να μπορούν να συμπεριληφθούν όλες οι δυνάμεις-αλληλεπιδράσεις που ασκούνται μεταξύ όλων των ατόμων. Όμως, μία τέτοια εξαντλητική αναζήτηση θα καθιστούσε το χώρο λύσεων τεράστιο, με αποτέλεσμα να χρησιμοποιούνται τελικά απλοποιημένα πεδία δυνάμεων και αλγόριθμοι που δεν αξιολογούν όλες τις δυνατές λύσεις (MONTE CARLO ή γενετικοί αλγόριθμοι).

3.3.4 Συμπεράσματα

Αν και οι απαρχής μέθοδοι φαινόταν να αντικαθίστανται από τις μεθόδους προτυποποίησης με ομολογία ή ύφανσης, τελευταία έχουν εμφανιστεί νέες απαρχής μέθοδοι που κάνουν χρήση δομικής πληροφορίας για στοιχειώδη τμήματα ακολουθιών.

Είναι γεγονός ότι μερικές φορές οι μέθοδοι προτυποποίησης με ομολογία ή ύφανσης δίνουν φορές εντυπωσιακά αποτελέσματα, δεν είναι όμως δυνατό να

θεωρήσουμε πως έχει λυθεί το πρόβλημα του διπλώματος των πρωτεϊνών. Αν και οι μέθοδοι αυτές αφήνουν μεγάλες υποσχέσεις για το μέλλον, είναι επικίνδυνο να υπερεκτιμηθούν τα αποτελέσματά τους. Για το λόγο αυτό, πολύ συχνά καταφεύγουμε σε απλουστευμένες προγνώσεις, που μπορεί να μη δίνουν πλήρως τις τρισδιάστατες δομές αλλά αποκαλύπτουν σημαντικά μυστικά για τη δομή των πρωτεϊνών.

Στη συγκεκριμένη εργασία προτείνεται για τη λύση του προβλήματος μια μέθοδος, που εντάσσεται στις απαρχής μεθόδους, η οποία χρησιμοποιεί χαρακτηριστικά εξαγμένα από την ακολουθία με σκοπό την πρόβλεψη της δομής.

4

Υλοποίηση Ταξινομητή

Το κεφάλαιο αυτό αφορά στην ανάλυση του αλγορίθμου που χρησιμοποιείται για την υλοποίηση του ταξινομητή. Στην ενότητα 4.1 γίνεται μια εισαγωγή, στην ενότητα 4.2 γίνεται αναλυτική επεξήγηση της υλοποίησης και στην ενότητα 4.3 παρουσιάζονται τα πειράματα που έγιναν για την αξιολόγηση του ταξινομητή.

4.1 *Εισαγωγή*

Η γενική ιδέα αντιμετώπισης του προβλήματος αναγνώρισης διπλώματος πρωτεϊνών, είναι η ταξινόμηση πρωτεϊνικών ακολουθιών σε διπλώματα -σύμφωνα με τη SCOP ταξινόμηση (βλ. Ενότητα 3.2.1)- βάσει χαρακτηριστικών που εξάγονται από την ίδια την ακολουθία. Για να υλοποιηθεί ο ταξινομητής χρησιμοποιούμε ένα σύνολο δεδομένων που έχει χρησιμοποιηθεί και σε άλλες εργασίες. Το σύνολο δεδομένων αποτελείται από 311 πρωτεΐνες που είναι ταξινομημένες σε 27 κύρια διπλώματα. Δημιουργήσαμε δίκτυα ομοιότητας των πρωτεϊνικών αυτών ακολουθιών, με σκοπό να αποδώσουμε σε ορισμένες πρωτεϊνικές ακολουθίες του ίδιου συνόλου, που θεωρούμε ότι ανήκουν σε άγνωστο δίπλωμα, το σωστό. Τέλος ελέγξαμε την ακρίβεια του ταξινομητή.

Στο δίκτυο ομοιότητας πρωτεϊνικών ακολουθιών (Protein Sequence Similarity Network - PSeSN) η ομοιότητα μεταξύ των πρωτεϊνικών ακολουθιών-κορυφών μετράται χρησιμοποιώντας ένα σύνολο 125 χαρακτηριστικών, τα οποία προέρχονται από την ίδια την ακολουθία. Τα χαρακτηριστικά αυτά αφορούν στην αμινοξική σύσταση (20 χαρακτηριστικά), στην προβλεπόμενη δευτεροταγή δομή των καταλοίπων της (21 χαρακτηριστικά), στον κανονικοποιημένο όγκο van der Waals (21 χαρακτηριστικά), την υδροφοβικότητα (21 χαρακτηριστικά), στην πολικότητα (21 χαρακτηριστικά) και στην πολωσιμότητα (21 χαρακτηριστικά). (βλ. Πίνακα 4.1)

Αναλυτικότερα, η ομάδα των χαρακτηριστικών που αφορούν στη σύσταση των αμινοξέων, είναι ένα διάνυσμα που περιέχει το ποσοστό εμφάνισης του κάθε αμινοξέος στην πρωτεϊνική ακολουθία. Τα χαρακτηριστικά των υπόλοιπων κατηγοριών προέκυψαν από την κατηγοριοποίηση όλων των αμινοξέων σε τρεις κλάσεις για κάθε κατηγορία. Για την υδροφοβικότητα, για παράδειγμα, τα αμινοξέα χωρίστηκαν στις κατηγορίες: πολικό, ουδέτερο και υδροφοβικό (13).

Αμινοξική Σύσταση			Προβλεπόμενη 2ο-ταγής δομή			Υδροφοβικότητα			Van der Waals			Πολικότητα			Πολωσιμότητα		
1	…	20	21	…	41	42	…	62	63	…	83	84	…	104	105	…	125

Πίνακας 4.1: Ομάδες χαρακτηριστικών εξαγμένων από την ακολουθία

Στην επόμενη σελίδα παραθέτουμε τον πίνακα που παρουσιάζει το σύνολο των πρωτεϊνικών ακολουθιών που χρησιμοποιήθηκαν ταξινομημένων κατά δίπλωμα και κλάση. Στις δύο τελευταίες στήλες βρίσκεται ο αριθμός των ακολουθιών που ανήκουν στο κάθε δίπλωμα και ο αριθμός των δομών οι οποίες είναι διαθέσιμες.

Πρωτεϊνικό Δίπλωμα		***Αριθμός Ακολουθιών***	***Αριθμός Δομών***
	κλάση α		
1	Globin-like	13	13
2	Cytocrome *c*	7	6
3	DNA-binding 3-helical bundle	12	12
4	4-helical up-and-down bundle	7	7
5	4-helical cytokines	9	7
6	Alpha; EF-hand	6	6
	κλάση β		
7	Immunoglobulin-like *β*-sandwich	30	27
8	Cupredoxins	9	9
9	Viral coat and capsid proteins	16	16
10	ConA-likelectins/glucanases	7	6
11	SH3-like barrel	8	8
12	OB-fold	13	13
13	Trefoil	8	7
14	Trypsin-like serine proteases	9	8
15	Lipocalins	9	9
	κλάση α/β		
16	(TIM)-barrel	29	28
17	FAD (also NAD)-binding motif	11	11
18	Flavodoxin-like	11	11
19	NAD(P)-binding Rossmann-fold	13	13
20	P-loop containing nucleotide	10	9
21	Thioredoxin-like	9	9
22	Ribonuclease H-like motif	10	9
23	Hydrolases	11	11
24	Periplasmic binding protein-like	11	11
	κλάση α+β		
25	*β*-grasp	7	7
26	Ferredoxin-like	13	11
27	Small inhibitors, toxins, lectins	13	12
		311	**296**

Πίνακας 4.2 Ο αριθμός των ακολουθιών και των δομών που αντιστοιχούν σε κάθε ένα από τα 27 διπλώματα.

4.2 *Υλοποίηση συστήματος*

Επείδή είναι πολύ σύνηθες ένας αλγόριθμος μηχανικής μάθησης -ιδιαιτέρως αν είναι ταξινομητής- να λειτουργεί καλά για ένα συγκεκριμένο σύνολο αλλά να μην έχει καλή ικανότητα γενίκευσης, δηλαδή να μην παρέχει ικανοποιητικά αποτελέσματα σε ένα σύνολο αξιολόγησης, χρησιμοποιήσαμε τη μέθοδο διασταυρωμένης επικύρωσης ή διεπικύρωσης (cross-validation). Η τεχνική αυτή χρησιμοποιείται κυρίως σε μοντέλα των οποίων στόχος είναι η πρόβλεψη αποτελεσμάτων, για να εκτιμηθεί η ακρίβεια του μοντέλου πρόβλεψης στην πράξη. Εκτιμά πόσο καλά το μοντέλο θα λειτουργήσει σε άγνωστα δεδομένα, δηλαδή πόσο καλή είναι η ικανότητα γενίκευσής του. Σύμφωνα με τη μέθοδο αυτή το αρχικό σύνολο εκπαίδευσης διαιρείται σε n συμπληρωματικά υποσύνολα. Σε κάθε επανάληψη χρησιμοποιούνται τα n-1 υποσύνολα για εκπαίδευση (training set) και το υποσύνολο που απομένει χρησιμοποιείται για αξιολόγηση (testing or validating set). Η διαδικασία επαναλαμβάνεται n φορές, ώστε κάθε υποσύνολο να χρησιμοποιηθεί για μια φορά ως σύνολο αξιολόγησης. Στο τέλος των επαναλήψεων εκτιμάται η συνολική ικανότητα ταξινόμησης του συστήματος.

Στο συγκεκριμένο πρόβλημα χρησιμοποιήσαμε διασταυρωμένη επικύρωση 10 επαναλήψεων (10-fold cross validation). Δηλαδή οι 311 διαθέσιμες ακολουθίες κατανέμονται έτσι ώστε να έχουμε 10 διαφορετικά σύνολα εκπαίδευσης (που αποτελούνται από το 90% των διαθέσιμων πρωτεϊνικών ακολουθιών) και 10 διαφορετικά σύνολα αξιολόγησης (που αποτελούνται από το υπόλοιπο 10% των διαθέσιμων πρωτεϊνικών ακολουθιών που απομένει κάθε φορά). Σε κάθε επανάληψη της μεθόδου διασταυρωμένης επικύρωσης, το σύνολο εκπαίδευσης χρησιμοποιείται για να κατασκευαστεί το δίκτυο ομοιότητας πρωτεϊνικών ακολουθιών (296 πρωτεϊνικές ακολουθίες) και το σύνολο αξιολόγησης (32 πρωτεϊνικές ακολουθίες) για τον έλεγχο της μεθόδου. Το συγκεκριμένο σύνολο ακολουθιών έχει χρησιμοποιηθεί στην προσπάθεια επίλυσης του ίδιου προβλήματος και από άλλους ερευνητές (13) (14) (15) (16).

4.2.1 Κατασκευή Δικτύου Ομοιότητας Ακολουθιών

Το δίκτυο πρωτεϊνικών ακολουθιών που κατασκευάζεται με βάση τα 125 χαρακτηριστικά που εξάγονται από την ακολουθία, έχει ως κορυφές τις 296 ακολουθίες πρωτεϊνών, οι οποίες ανήκουν στο σύνολο εκπαίδευσης και ακμές με βάρη (14). Κάθε πρωτεϊνική ακολουθία-κορυφή χαρακτηρίζεται από ένα διάνυσμα 125 χαρακτηριστικών που εξάγονται από αυτή. Το βάρος των ακμών του δικτύου αντιστοιχεί σε μία μετρική σχέση ομοιότητας των κορυφών που ενώνει. Ισούται με το αντίστροφο της ευκλείδιας απόστασης, ανάμεσα στα διανύσματα, που αντιστοιχούν στις εκάστοτε δύο κορυφές του δικτύου που ενώνει η εν λόγω ακμή. Σκοπός της χρήσης της ευκλείδειας απόστασης είναι να μετρήσει πόσο απέχουν οι δύο παρατηρήσεις, στην προκειμένη περίπτωση τα δύο διανύσματα χαρακτηριστικών των πρωτεϊνών.

Γενικεύοντας τον τύπο της ευκλείδειας απόστασης για p=125 χαρακτηριστικά και θεωρώντας ότι τα δυο διανύσματα χαρακτηριστικών είναι $\boldsymbol{x} = (x_1, x_2, \dots, x_p)$ και $\mathbf{y} = (y_1, y_2, \dots, y_p)$ προκύπτει η παρακάτω εξίσωση:

$$d(x, y) = \sqrt{\sum_{i=1}^{p} (\mathrm{x}_i - y_i)^2} \qquad (1)$$

Με βάση την παραπάνω εξίσωση και αντιστρέφοντας το αποτέλεσμα που προέκυψε για κάθε ακμή $\left({}^{1}/_{d}\right)$, υπολογίζεται ο πίνακας των βαρών των ακμών του δικτύου ακολουθιών. Με τον τρόπο αυτό μια ακμή που ενώνει δυο πρωτεΐνες-κορυφές έχει μεγαλύτερο βάρος όταν οι πρωτεΐνες αυτές έχουν παρόμοια χαρακτηριστικά, δηλαδή μεγάλο βαθμό ομοιότητας μεταξύ τους. Ενώ αντίθετα, μία ακμή έχει μικρότερο βάρος όσο μεγαλύτερη είναι η διαφορά των δυο πρωτεϊνών-κορυφών που συνδέει.

Για τις ανάγκες της υλοποίησης σε Matlab, το δίκτυο αναπαραστάθηκε ως πίνακας γειτνίασης με βάρη. (βλ. Ενότητα 2.3)

4.2.2 Επαναληπτικός Αλγόριθμος

Το δίκτυο ομοιότητας πρωτεϊνικών ακολουθιών υλοποιήθηκε με σκοπό να υπολογιστεί η συγγένεια μιας πρωτεϊνικής ακολουθίας, της οποίας το δίπλωμα θεωρείται άγνωστο και στο εξής θα αναφέρεται ως ακολουθία-αναζήτησης, (query

protein), με τις υπόλοιπες πρωτεϊνικές ακολουθίες-κορυφές του δικτύου, των οποίων το δίπλωμα θεωρείται γνωστό.

Για το σκοπό αυτό χρησιμοποιήθηκε η τεχνική των τυχαίων διαδρομών (random walks). Αφού υπολογιστούν οι βαθμοί συσχέτισης μεταξύ της ακολουθίας αναζήτησης και των υπόλοιπων πρωτεϊνικών ακολουθιών και τελειώσει η αναζήτηση, αποδίδεται στην ακολουθία αναζήτησης το δίπλωμα της πρωτεϊνικής ακολουθίας με την οποία παρουσιάζει τη μεγαλύτερη συγγένεια, βασιζόμενη στα χαρακτηριστικά που έχουν εξαχθεί από τις ακολουθίες.

Έστω $x_1, x_2, \ldots, x_m$ το σύνολο των διανυσμάτων (καθένα από αυτά περιλαμβάνει τα 125 χαρακτηριστικά) που αναπαριστούν τις πρωτεϊνικές ακολουθίες στο δίκτυο ομοιότητας πρωτεϊνών. Θεωρούμε ότι το διάνυσμα x_1 αποτελεί την ακολουθία αναζήτησης. Υπολογίζουμε τις τιμές του βαθμού συσχέτισης για όλα τα δυνατά ζεύγη πρωτεϊνικών ακολουθιών, συμπεριλαμβανομένης και της ακολουθίας αναζήτησης. Οι τιμές αυτές αποθηκεύονται σε ένα πίνακα K, διαστάσεων mxm. Συγκεκριμένα, στη θέση K_{ij} του πίνακα, αποθηκεύεται ο βαθμός συσχέτισης μεταξύ των πρωτεϊνικών ακολουθιών x_i και x_j. Οι τιμές του πίνακα κανονικοποιούνται έτσι ώστε για όλα τα i να ισχύει:

$$\sum_{j=2}^{m} K_{ij} = 1 \qquad (2)$$

Δηλαδή για παράδειγμα το στοιχείο K_{12} του πίνακα K αντιστοιχεί στο αντίστροφο της ευκλείδειας απόστασης των χαρακτηριστικών της πρωτεϊνικής ακολουθίας 1 (εν προκειμένω της ακολουθίας αναζήτησης) με την πρωτεϊνική ακολουθία 2. Επίσης ορίζεται $y_i(t),\ i = 2, \ldots m$ ο βαθμός συσχέτισης της ακολουθίας αναζήτησης με τις υπόλοιπες πρωτεϊνικές ακολουθίες του δικτύου. Δηλαδή για παράδειγμα: $y_3(4)$ είναι ο βαθμός συσχέτισης της πρωτεΐνης-ερώτημα με την τρίτη πρωτεϊνική ακολουθία, όταν t=4.

Αρχικοποιούμε το σύνολο τιμών $y_i(0) = 0$ για κάθε $i = 2, \ldots m$ για τη χρονική στιγμή (t=0), και για κάθε επόμενο βήμα η τιμή των $y_i(t+1),\ i = 2, \ldots m$ δίνεται από τη σχέση:

$$y_i(t+1) = K_{1i} + a \sum_{j=2}^{m} K_{ji} y_i(t) \qquad (3)$$

Ο πολλαπλασιαστικός παράγοντας α παίρνει τιμές στο διάστημα [0,1] και προσδιορίζεται πριν την εκτέλεση του αλγορίθμου. Όταν α=0 σημαίνει ότι ο αλγόριθμος ταξινομεί τις πρωτεϊνικές ακολουθίες-κορυφές του δικτύου βασιζόμενος μόνο στην ευκλείδεια απόσταση μεταξύ αυτών και της ακολουθίας αναζήτησης. Ενώ όταν α=1 σημαίνει ότι ο αλγόριθμος εκτελεί αναζήτηση σε όλο το δίκτυο. Δηλαδή όταν α=0 το αποτέλεσμα θα βασίζεται μόνο στον αρχικό βαθμό συσχέτισης μεταξύ των χαρακτηριστικών της ακολουθίας αναζήτησης με τις υπόλοιπες πρωτεϊνικές ακολουθίες, όπως αυτός αποτυπώνεται στην πρώτη γραμμή του πίνακα K_{1i}, χωρίς να λαμβάνει υπόψη του το βαθμό συσχέτισης των υπόλοιπων πρωτεϊνικών ακολουθιών μεταξύ τους, όπως αυτός αποτυπώνεται στα υπόλοιπα στοιχεία του πίνακα K.

Ο αλγόριθμος σταματάει τις επαναλήψεις όταν δυο τιμές $y_i(t)$ και $y_i(t+1)$ συγκλίνουν, δηλαδή απέχουν μεταξύ τους λιγότερο από κάποια τιμή κατωφλίου (threshold). Στη συγκεκριμένη προσέγγιση οι τιμές του βαθμού συσχέτισης της ακολουθίας αναζήτησης για κάθε μία από τις υπόλοιπες πρωτεϊνικές ακολουθίες συμπεριλήφθηκαν σε διάνυσμα **y** με *(i-1)* στοιχεία και θεωρήθηκε ότι ο αλγόριθμος συγκλίνει σε συγκεκριμένες τιμές όταν η ευκλείδεια απόσταση των διανυσμάτων δυο διαδοχικών επαναλήψεων $[\boldsymbol{y_i}(t), \boldsymbol{y_i}(t+1)]$ ήταν μικρότερη της τιμής 0.000001 (14).

Μετά το τέλος των επαναλήψεων οι βαθμοί συσχέτισης της ακολουθίας αναζήτησης με τις υπόλοιπες πρωτεϊνικές ακολουθίες ταξινομούνται και αποδίδεται στην ακολουθία αναζήτησης το δίπλωμα της πρωτεΐνης εκείνης με την οποία υπολογίστηκε ότι έχει το μεγαλύτερο βαθμό συσχέτισης.

4.3 *Πειραματική Αξιολόγηση*

Για τη μέτρηση της επιτυχίας της μεθόδου ορίστηκε η μεταβλητή TA (total accuracy-ολική ακρίβεια), η οποία δηλώνει την ποσοστιαία επιτυχία απόδοσης του σωστού διπλώματος στις ακολουθίες αναζήτησης, του συνόλου αξιολόγησης. Η τιμή της TA υπολογίστηκε για κάθε ένα από τα 10 ζεύγη συνόλων εκπαίδευσης και αξιολόγησης και στη συνέχεια ελήφθη προς αξιολόγηση η μέση τιμή της. Για να υπάρχει ποικιλία συνθηκών μεταβάλλονται διάφορες παράμετροι μεταξύ των πειραμάτων.

4.3.1 Διαφορετικά περιθώρια λάθους

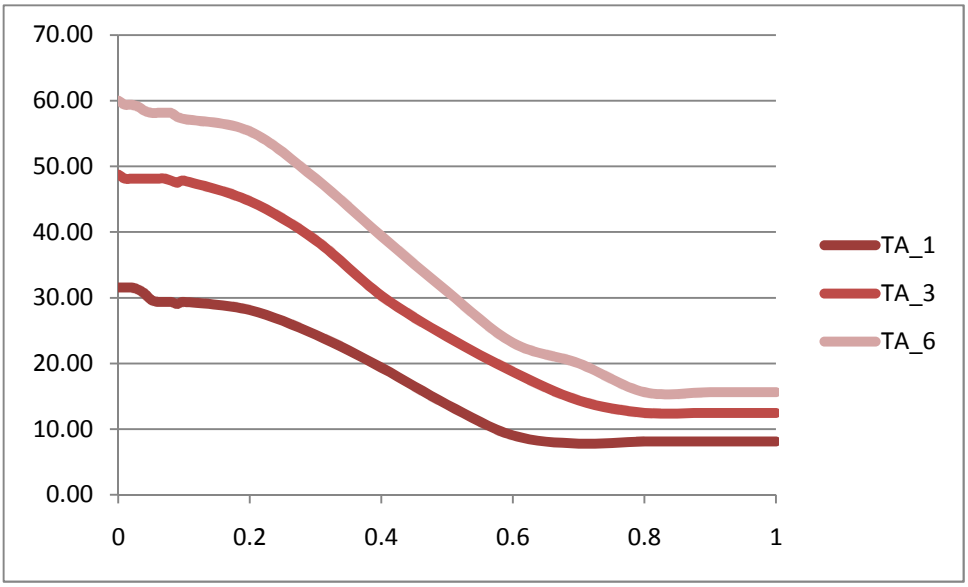
Αρχικά ο αλγόριθμος χρησιμοποιήθηκε σε 10 σύνολα διασταυρωμένης επικύρωσης και βρέθηκε το ποσοστό απόδοσης του σωστού διπλώματος για τιμές του ***α*** της εξίσωσης (4.3), δεδομένου ότι δεχόμαστε ως έξοδο του αλγορίθμου το δίπλωμα της πρωτεϊνικής ακολουθίας που παρουσιάζει τη μεγαλύτερη συγγένεια με την εκάστοτε πρωτεΐνη-ερώτημα (TA_1) (βλ. Πίνακας 4.3: $1^{η}$ στήλη). Αυτό το ποσοστό επιτυχίας επιτυγχάνεται όταν χρησιμοποιείται ο ταξινομητής πλήρως αυτόματα. Όμως, αν θεωρήσουμε ότι ο αλγόριθμος μπορεί να προτείνει κάποια διπλώματα ως πιθανότερα και ένας ερευνητής να επιλέξει από ένα σύνολο ποιο είναι το σωστό με περαιτέρω επεξεργασία, μπορούμε να θεωρήσουμε σωστό το αποτέλεσμα του αλγορίθμου αν εμφανίζεται το σωστό δίπλωμα της ακολουθίας αναζήτησης στα n πρώτα αποτελέσματα, ταξινομημένα κατά βαθμό συσχέτισης (affinity).

Δηλαδή μπορούμε να δεχτούμε ως σωστό το αποτέλεσμα του αλγορίθμου αν το δίπλωμα της ακολουθίας αναζήτησης είναι το ίδιο με το δίπλωμα μίας εκ των τριών πρωτεϊνικών ακολουθιών που παρουσιάζουν τη μεγαλύτερη συγγένεια με την ακολουθία αναζήτησης (TA_3). Επιπλέον μπορούμε να δεχτούμε ως σωστό το αποτέλεσμα όταν το δίπλωμα της πρωτεΐνης-ερώτημα είναι το δίπλωμα μίας εκ των έξι πρωτεϊνικών ακολουθιών με τις οποίες υπάρχει ο μεγαλύτερος βαθμός συγγένειας (TA_6).

Έτσι από τα πειράματα ταξινόμησης για τις 3 κατηγορίες ποσοστών επιτυχίας TA_1 TA_3 TA_6 και για τιμές του ***α*** από 0 έως 0.1 με βήμα 0.01 και από 0.1 έως 1 με βήμα 0.1 προκύπτει ο πίνακας που φαίνεται παρακάτω. Στον πίνακα έχει επισημανθεί το μεγαλύτερο ποσοστό επιτυχίας για κάθε κατηγορία. Έτσι παρατηρούμε ότι στην αυτόματη λειτουργία του αλγορίθμου το μεγαλύτερο ποσοστό επιτυχίας επιτυγχάνεται για α=0,...,0.03 . Ενώ στις άλλες δυο κατηγορίες καλύτερα αποτελέσματα έχουμε για α=0. Αξίζει να σημειωθεί ότι όσο μεγαλύτερες τιμές παίρνει το ***α*** τόσο αυξάνει και η πολυπλοκότητα του αλγορίθμου άρα και ο χρόνος εκτέλεσης. Παρόλα αυτά παρατηρούμε τα μεγαλύτερα ποσοστά επιτυχίας για πολύ μικρά ***α***, δηλαδή όταν δεν λαμβάνουμε σχεδόν καθόλου υπόψη τον δεύτερο όρο της σχέσης (3).

α	*TA_1*	*TA_3*	*TA_6*
0	31.56	48.75	60.00
0.01	31.56	48.13	59.37
0.02	31.56	48.13	59.37
0.03	31.25	48.13	59.06
0.04	30.63	48.13	58.44
0.05	29.69	48.13	58.13
0.06	29.38	48.13	58.13
0.07	29.38	48.13	58.13
0.08	29.38	47.81	58.13
0.09	29.06	47.50	57.50
0.1	29.38	47.81	57.19
0.2	28.13	44.69	55.31
0.3	24.38	38.75	48.13
0.4	19.38	30.31	39.38
0.5	13.75	24.06	30.94
0.6	9.06	18.75	23.13
0.7	7.81	14.38	20.00
0.8	8.13	12.50	15.63
0.9	8.13	12.50	15.63
1	8.13	12.50	15.63

Πίνακας 4.4 Ποσοστό % επιτυχίας εύρεσης του σωστού διπλώματος για τις πρωτεϊνικές ακολουθίες για διάφορες τιμές του *α* (μέση τιμή για τα 10 ζεύγη της μεθόδου διεπικύρωσης).



Εικόνα 4.1 Ποσοστά επιτυχίας % του αλγορίθμου δεδομένου ότι το σωστό δίπλωμα που αναζητούμε εμφανίζεται στα πρώτα x αποτελέσματα (TA_n, n=1,2,3). Για n=1 πρόκειται για την αυτόματη λειτουργία.

Αξίζει να σημειωθεί ότι ένας τυχαίος ταξινομητής θα είχε ποσοστό επιτυχίας: (1/27x100)%=3.7% ενώ ο συγκεκριμένος ταξινομητής έχει ποσοστό επιτυχίας σε αυτόματη λειτουργία εύρεσης διπλώματος 31.56%. Δηλαδή, επειδή οι κατηγορίες διπλωμάτων που υπάρχουν στο σύνολο δεδομένων είναι 27, ένας τυχαίος ταξινομητής θα είχε 1 στις 27 πιθανότητες σωστής κατηγοριοποίησης.

4.3.2 Διαφορετικές ομάδες χαρακτηριστικών

Μια ακόμα ενδιαφέρουσα προσέγγιση ήταν να χρησιμοποιήσουμε κάποιες ομάδες από τα 125 χαρακτηριστικά των πρωτεϊνικών ακολουθιών - κορυφών και να δούμε πώς επηρεάζονται τα αποτελέσματα. Έτσι αρχικά χρησιμοποιήσαμε τα 20 χαρακτηριστικά από την αμινοξική σύσταση (20 χαρακτηριστικά). Στη συνέχεια προσθέσαμε μία-μία τις υπόλοιπες ομάδες των χαρακτηριστικών - την προβλεπόμενη δευτεροταγή δομή των καταλοίπων, τον κανονικοποιημένο όγκο van der Waals, την υδροφοβοκότητα, την πολικότητα και την πολωσιμότητα - ελέγχοντας τα αποτελέσματα κάθε φορά. Δηλαδή τρέξαμε τον αλγόριθμο κρατώντας 20, 41, 62, 83, 104 και 125 χαρακτηριστικά. Το πείραμα έγινε για α=0, για α=0.03 και α=0.3 και για τις τρεις κατηγορίες θεώρησης σωστού του αποτελέσματος. (TA_1, TA_3, TA_6). Τα

αποτελέσματα παρατίθενται στους παρακάτω πίνακες και επισημαίνεται το μεγαλύτερο ποσοστό επιτυχίας για κάθε ***α.***

	α	Αριθμός χαρακτηριστικών					
		20	***41***	***62***	***83***	***104***	***125***
TA_1	***0***	26.25	26.88	32.19	33.43	33.44	31.56
	0.03	25.93	21.87	31.88	32.81	33.44	31.25
	0.3	19.69	19.69	28.75	28.43	26.87	24.38

	α	Αριθμός χαρακτηριστικών					
		20	***41***	***62***	***83***	***104***	***125***
TA_3	***0***	42.5	39.38	53.125	52.19	50.94	48.75
	0.03	41.85	39.38	52.5	52.19	50	48.13
	0.3	36.88	30.94	43.75	42.5	39.06	38.75

	α	Αριθμός χαρακτηριστικών					
		20	***41***	***62***	***83***	***104***	***125***
TA_6	***0***	55.93	50.31	60.44	61.56	62.81	60
	0.03	54.06	50	62.81	61.25	62.19	59.06
	0.3	46.88	42.50	53.13	52.5	51.25	48.13

Πίνακας 4.5 Ποσοστά επιτυχίας του ταξινομητή για τις τρεις κατηγορίες θεώρησης σωστού αποτελέσματος, έξι ομάδες χαρακτηριστικών και τρεις τιμές του α.

Τα ποσοστά επιτυχίας του ταξινομητή παρουσιάζουν μία ιδιαιτερότητα. Ενώ θα ήταν αναμενόμενο όσα πιο πολλά γνωστά χαρακτηριστικά έχουμε για μία πρωτεϊνική ακολουθία, τόσο πιο εύκολο να είναι να βρούμε το δίπλωμα της, κάτι τέτοιο δεν συμβαίνει. Άρα καταλήγουμε στο συμπέρασμα ότι κάποια χαρακτηριστικά από τα 125 της ακολουθίας περιέχουν σημαντικότερη πληροφορία για τη δομή της, σε σχέση με τα υπόλοιπα. Για το λόγο αυτό στο επόμενο κεφάλαιο υλοποιείται ένας γενετικός αλγόριθμος, με σκοπό την εύρεση των πιο εύρωστων χαρακτηριστικών και την χρήση τους στο τελικό μοντέλο ταξινομητή. (βλ.Κεφάλαιο 5)

4.3.3 Σύγκριση επιτυχίας στα διαφορετικά διπλώματα

Θεωρήσαμε επίσης ενδιαφέρον να δούμε αν ο ταξινομητής λειτουργεί διαφορετικά για πρωτεϊνικές ακολουθίες που ανήκουν σε διαφορετικές ομάδες διπλωμάτων. Αν και ο αριθμός των διαθέσιμων πρωτεϊνικών ακολουθιών αναζήτησης που ανήκουν σε κάθε ένα από τα 27 διπλώματα δεν είναι ο ίδιος (βλ. Πίνακας 1), υπολογίσαμε το ποσοστό επιτυχίας για κάθε τύπο διπλώματος. Δηλαδή, για τις 32 ακολουθίες αναζήτησης του καθενός από τα 10 σύνολα αξιολόγησης υπολογίσαμε το ποσοστό επιτυχίας του ταξινομητή ανάλογα με το δίπλωμα. Τα πειράματα έγιναν για τρεις τιμές του $\boldsymbol{\alpha}$, α=0, α=0.03 και α=0.3, για τις τρεις κατηγορίες ακρίβειας του αποτελέσματος (TA_1, TA_3 και TA_6) και τμηματικά για τις ομάδες χαρακτηριστικών που προέρχονται από την πρωτεϊνική ακολουθία, όπως και στο προηγούμενο σύνολο πειραμάτων (20, 41, 62, 83, 104 και 125 χαρακτηριστικά).

Για λόγους συντομίας στον πίνακα 4.6 παρουσιάζονται τα ποσοστά επιτυχίας του ταξινομητή για κάθε δίπλωμα και για τις τρεις κατηγορίες ακρίβειας, για $\boldsymbol{\alpha}$=0 και για 125 χαρακτηριστικά. Για $\boldsymbol{\alpha}$=0 και οι τρεις κατηγορίες παρουσιάζουν την καλύτερη τους συμπεριφορά. Στη δεύτερη στήλη καταγράφεται ο αριθμός των πρωτεϊνικών ακολουθιών-ερωτημάτων που ανήκουν στο συγκεκριμένο δίπλωμα, για καλύτερη αξιολόγηση των αποτελεσμάτων.

Δίπλωμα	***Αριθμός Ακολουθιών*** *(με αυτό το δίπλωμα)*	***Ποσοστά Επιτυχίας***		
		TA_1	***TA_3***	***TA_6***
1	16	87,50	93,75	93,75
2	8	37,50	37,50	37,50
3	5	80,00	80,00	100,00
4	3	0,00	0,00	0,00
5	7	14,29	71,43	71,43
6	6	0,00	33,33	33,33
7	31	19,35	51,61	67,74
8	8	0,00	25,00	25,00
9	13	53,85	92,31	92,31
10	9	33,33	33,33	33,33
11	9	0,00	0,00	22,22
12	17	17,65	17,65	17,65
13	14	14,29	21,43	42,86
14	8	37,50	37,50	37,50
15	6	50,00	100,00	100,00
16	30	66,67	83,33	100,00
17	15	33,33	60,00	86,67
18	14	0,00	0,00	14,29
19	18	16,67	44,44	66,67
20	11	54,55	54,55	54,55
21	6	16,67	16,67	33,33
22	13	0,00	46,15	53,85
23	9	11,11	22,22	88,89
24	14	71,43	85,71	100,00
25	7	0,00	42,86	42,86
26	11	0,00	0,00	0,00
27	12	50,00	58,33	58,33

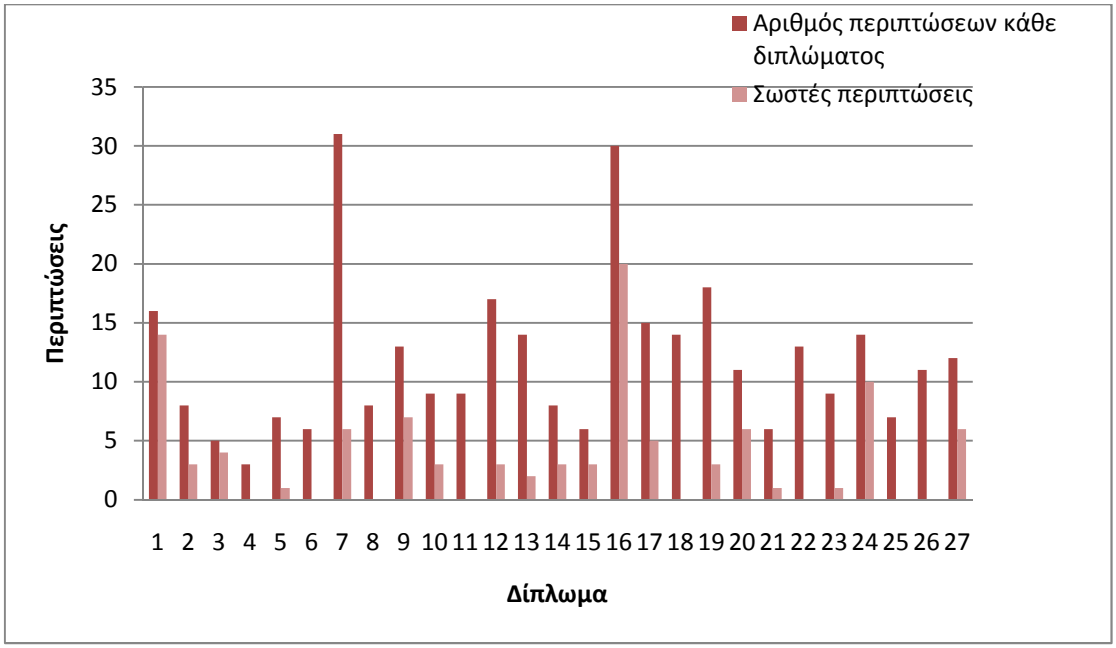
Πίνακας 4.6 Ποσοστά επιτυχίας του ταξινομητή για κάθε δίπλωμα για τις τρεις κατηγορίες θεώρησης σωστού αποτελέσματος.

Αρχικά παρατηρήσαμε ότι τα αποτελέσματα για κάθε δίπλωμα δεν παρουσιάζουν αξιοσημείωτες μεταβολές καθώς μεταβάλλεται το ***α***. Επίσης δεν παρουσιάζουν σημαντικές μεταβολές ούτε για διαφορετικές ομάδες χαρακτηριστικών, δηλαδή τα επί μέρους αποτελέσματα συμβαδίζουν με τη γενικότερη εικόνα μεταβολής του μέσου όρου της ακρίβειας.

Για κάποια διπλώματα ο ταξινομητής έχει μεγαλύτερη επιτυχία ενώ για άλλα όχι. Για παράδειγμα όσον αφορά τις πρωτεϊνικές ακολουθίες που ανήκουν στο δίπλωμα 1 (globin-like) σε όλες τις περιπτώσεις παρουσιάζει πάντα επιτυχία μεγαλύτερη του 80%, καθώς και όσον αφορά εκείνες που ανήκουν στο δίπλωμα 24 (periplasmic binding protein-like), ποσοστά επιτυχίας πάνω από 70%. Αντίθετα για εκείνες που ανήκουν στο δίπλωμα 18 ή εκείνες που ανήκουν στο δίπλωμα 26 τα ποσοστά επιτυχίες παραμένουν σε όλες τις περιπτώσεις πολύ χαμηλά.

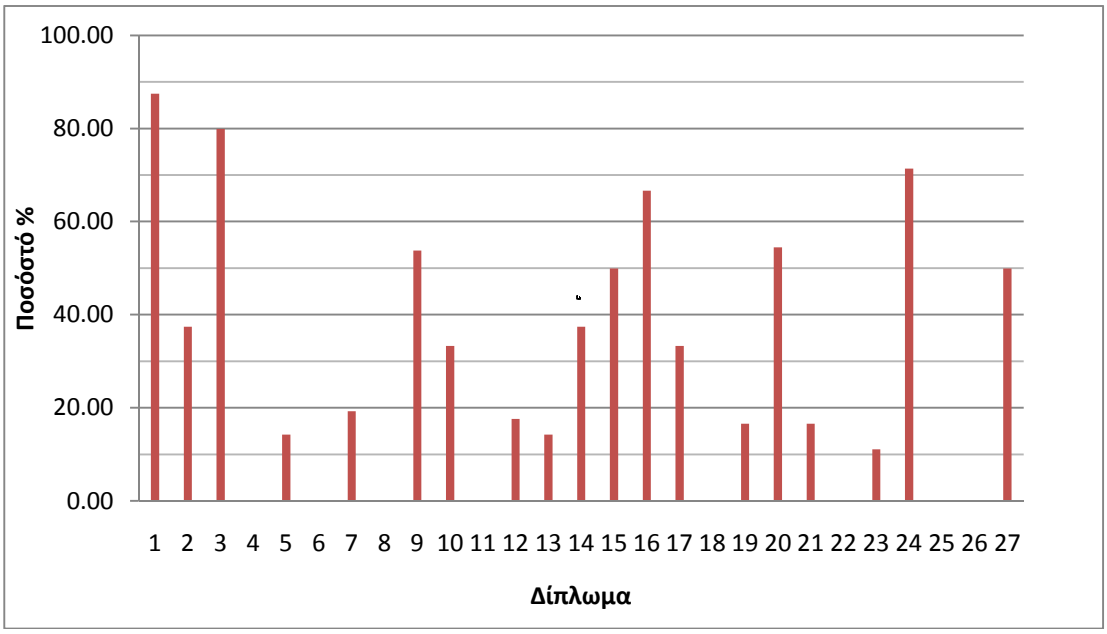
Επιπλέον, θεωρήσαμε σημαντικό να παρατηρήσουμε τι συμβαίνει όταν ο ταξινομητής προτείνει λάθος δίπλωμα για την ακολουθία αναζήτησης. Ακόμα και τότε το λάθος δίπλωμα που προτείνει ανήκει -στην συντριπτική πλειονότητα των περιπτώσεων- στην ίδια κλάση με το σωστό. Για παράδειγμα για το δίπλωμα 18 (Flavodoxin-like), που τα ποσοστά επιτυχίας είναι πολύ μικρά, προτείνονται ως σωστά τα διπλώματα 16 ((TIM)-barrel), 19 (NAD(P)-binding Rossmann-fold), 20 (P-loop containing nucleotide) και 24 (Periplasmic binding protein-like) τα οποία ανήκουν στην ίδια κλάση με το δίπλωμα 18 (α/β κλάση). Αυτό συμβαίνει γιατί τα χαρακτηριστικά που προέρχονται από την ακολουθία παρουσιάζουν μεγάλη ομοιότητα, ανάμεσα στις πρωτεϊνικές ακολουθίες που ανήκουν στην ίδια κλάση. Παρόμοιες παρατηρήσεις μπορούν να γίνουν και για τα υπόλοιπα διπλώματα με μικρά ποσοστά επιτυχίας.

Τέλος παρατηρήσαμε ότι όταν τα ποσοστά επιτυχίας είναι μικρά για ακρίβεια TA_1, σε κάποιες περιπτώσεις βελτιώνονται δραματικά αν επιτρέψουμε το περιθώριο των 6 αποτελεσμάτων (ακρίβεια TA_6) ενώ σε άλλες όχι, χωρίς να μπορεί να εξαχθεί κανόνας για αυτό. Παραθέτουμε ένα γράφημα που παρουσιάζει την επιτυχία σε κάθε δίπλωμα, δεδομένου ότι ο ταξινομητής λειτουργεί αυτόματα (TA_1), για 125 χαρακτηριστικά και α=0.



Εικόνα 4.2 Αριθμός σωστών ταξινομήσεων σε σχέση με τις υπάρχουσες περιπτώσεις για κάθε δίπλωμα (TA_1, για 125 χαρακτηριστικά και α=0)

Επίσης παραθέτουμε γραφικά τα αποτελέσματα για την ακρίβεια του ταξινομητή σε κάθε δίπλωμα σε ποσοστιαία αναλογία.



Εικόνα 4.3 Ποσοστό επιτυχούς ανεύρεσης του διπλώματος μια πρωτεϊνικής ακολουθίας σε σχέση με το δίπλωμα στο οποίο ανήκει.

5

Βελτιστοποίηση με χρήση Γενετικού Αλγορίθμου

Στο συγκεκριμένο κεφάλαιο επανεξετάζουμε το πρόβλημα της αναγνώρισης διπλώματος με χρήση του ίδιου αλγορίθμου για την ταξινόμηση των πρωτεϊνικών ακολουθιών, αλλά με προεπιλογή των πιο εύρωστων χαρακτηριστικών ακολουθίας. Στην ενότητα 5.1 γίνεται μία παρουσίαση του Γενετικού Αλγορίθμου ενώ στην ενότητα 5.2 παραθέτουμε τα αποτελέσματα που προέκυψαν από την αξιοποίηση των αποτελεσμάτων που έδωσε ο Γενετικός Αλγόριθμος, στην υλοποίηση του ταξινομητή.

5.1 *Βασικά στοιχεία Γενετικού Αλγόριθμου*

Χρησιμοποιώντας μικρότερο σύνολο χαρακτηριστικών σκοπός μας είναι να υλοποιήσουμε ένα ταχύτερο και αποδοτικότερο ταξινομητή. Για την επιλογή των χαρακτηριστικών αυτών χρησιμοποιούμε τις δυνατότητες των ΓΑ σε συνδυασμό με την προσπάθεια επίτευξης μέγιστης δυνατής ομοιότητας του δικτύου ομοιότητας ακολουθιών και του δικτύου ομοιότητας δομών των ακολουθιών (Protein Structural Similarity Network - PStSN). Δηλαδή, θέλουμε να χρησιμοποιήσουμε στην κατασκευή του δικτύου των ακολουθιών τα χαρακτηριστικά εκείνα που τον κάνουν να μοιάζει όσο το δυνατό περισσότερο με το δίκτυο των δομών.

Όπως περιγράφηκε και στην ενότητα 2.3 αρχικά επιλέγεται από το Γενετικό Αλγόριθμο ένας τυχαίος αρχικός πληθυσμός λύσεων. Για την επεξήγηση της λειτουργίας του αλγορίθμου θεωρούμε ότι το χρωμόσωμα του κάθε ατόμου του αρχικού πληθυσμού είναι μία δυαδική μάσκα, μία σειρά δηλαδή από δυαδικά ψηφία (0 ή 1) με μήκος ίσο με την αρχική διάσταση του συνόλου των χαρακτηριστικών, δηλαδή 125. Σε κάθε ένα από τα 125 χαρακτηριστικά του αρχικού συνόλου αντιστοιχεί και ένα δυαδικό ψηφίο. Αν το ψηφίο αυτό είναι 1, τότε το αντίστοιχο χαρακτηριστικό συμπεριλαμβάνεται στη λύση ενώ αν είναι 0 δε συμπεριλαμβάνεται. Μια υποψήφια λύση του προβλήματος για αρχικό σύνολο 125 χαρακτηριστικών φαίνεται στο Σχήμα 5.2. Στην υποψήφια αυτή λύση το $1^{ο}$ χαρακτηριστικό συμπεριλαμβάνεται ενώ το $3^{ο}$ δε συμπεριλαμβάνεται κ.ο.κ. .

1	1	0	- - - - -	1	0

125 ψηφία

Σχήμα 5.2 Το χρωμόσωμα μιας υποψήφιας λύσης για το πρόβλημα.

Στη δικιά μας υλοποίηση οι τιμές που αποδίδει στα χαρακτηριστικά του τυχαίου χρωμοσώματος ο ΓΑ, δεν είναι μόνο 0 και 1. Άρα και στο τελικό χρωμόσωμα σε κάθε χαρακτηριστικό αντιστοιχεί μία αριθμητική τιμή. Εμείς από αυτή την τελική τιμή κρατάμε μόνο εκείνα τα χαρακτηριστικά για τα οποία η τιμή ξεπερνά ένα κατώφλι (συγκεκριμένα το 0.5).

Τα άτομα του πληθυσμού διασταυρώνονται μεταξύ τους με πιθανότητα που εξαρτάται από τη τιμή της συνάρτησης ποιότητας τους. Σαν συνάρτηση ποιότητας χρησιμοποιείται μία συνάρτηση που κατασκευάζει κάθε φορά, για κάθε χρωμόσωμα υποψήφιας λύσης, το δίκτυο ομοιότητας ακολουθιών και το συγκρίνει με το δίκτυο ομοιότητας δομών, υπολογίζοντας ένα κόστος μη-ομοιότητας (dissimilarity cost). Όταν η συνάρτηση κόστους παίρνει την ελάχιστη τιμή, δηλαδή τα δίκτυα παρουσιάζουν τα μέγιστη ομοιότητα, κρατάμε μόνο τα χαρακτηριστικά εκείνα, από τα αρχικά 125, για τα οποία επετεύχθη αυτό.

5.2 *Κατασκευή δίκτυου ομοιότητας δομών*

Για να γίνει κατασκευή του δικτύου ομοιότητας δομών, προηγήθηκε αναζήτηση των δομών των πρωτεϊνικών ακολουθιών στην Τράπεζα Πρωτεϊνικών Δεδομένων (Protein Data Bank - PDB). Η δομή ήταν διαθέσιμη για τις 296 από αυτές καθώς για τις υπόλοιπες υπήρχαν ελλιπείς πληροφορίες σχετικά με τις ατομικές συντεταγμένες.

Για το λόγο αυτό κρατήσαμε και από το αρχικό σύνολο των 311 πρωτεϊνικών ακολουθιών τις 296 για τις οποίες είχαμε διαθέσιμες δομικές πληροφορίες. Το δίκτυο ομοιότητας δομών είναι ένα δίκτυο με κορυφές τις 296 πρωτεϊνικές ακολουθίες (πλέον κρατάμε μόνο 296 και στο δίκτυο ομοιότητας ακολουθιών) και ακμές με βάρη. Η ομοιότητα ή όχι των πρωτεϊνικών ακολουθιών-κορυφών δηλώνεται με το βάρος των ακμών ανάμεσά τους. Τα βάρη στις ακμές τοποθετήθηκαν, σύμφωνα με το περιβάλλον DaliLite. Σύμφωνα με το κριτήριο αυτό, προκύπτει μία τιμή κατά τη στοίχιση των δομών των ακολουθιών (z-score), που υποδηλώνει ομοιότητα. Έτσι, όταν ισχύει το κριτήριο: $Z \geq 2$, σημαίνει ότι οι δομές που αντιστοιχούν στις εκάστοτε δύο πρωτεϊνικές ακολουθίες παρουσιάζουν μεγάλη ομοιότητα, κριτήριο που έχει χρησιμοποιηθεί και σε προηγούμενες μελέτες.

Αναπαριστούμε το γράφων ομοιότητας δομών σαν ένα πίνακα γειτνίασης 296x296, όπου περιέχονται κανονικοποιημένες οι τιμές του Z, αν ίσχυε το κριτήριο ομοιότητας $Z \geq 2$, αλλιώς τοποθετούμε την τιμή 0.001.

5.3 *Συνάρτηση Καταλληλότητας*

Ο γενετικός αλγόριθμος προσπαθεί να κατασκευάσει ένα γράφο ομοιότητας ακολουθιών ο οποίος να μοιάζει όσο περισσότερο γίνεται στο δίκτυο ομοιότητας δομών, χρησιμοποιώντας κάθε φορά διαφορετικά υποσύνολα χαρακτηριστικών. Αυτό προσπαθεί να το πετύχει χρησιμοποιώντας μία συνάρτηση καταλληλότητας (fitness function), που υπολογίζει το μέγεθος της διαφοράς ανάμεσα στους δυο γράφους και το επιστρέφει σαν κόστος. Το δίκτυο ομοιότητας πρωτεϊνικών ακολουθιών κατασκευάζεται, για κάθε διαφορετικό υποσύνολο χαρακτηριστικών και υπολογίζεται κάθε φορά η διαφορετική συνάρτηση κόστους. Θεωρώντας ότι και το δίκτυο ομοιότητας πρωτεϊνικών ακολουθιών αναπαρίσταται ως ένας πίνακας γειτνίασης (A)

η διαφορά του από τον πίνακα γειτνίασης του δικτύου των δομών (B), υπολογίζεται την εξής συνάρτηση κόστους :

$$DC = \sum_{i}^{296} \sum_{j}^{296} \left| A_{ij} - B_{ij} \right| \qquad (4.1)$$

Αρχικά ο γενετικός αλγόριθμος επιλέγει τυχαία κάποια από τα 125 χαρακτηριστικά που λειτουργούν ως ανεξάρτητες μεταβλητές, ώστε να φτιάξει ένα δίκτυο ομοιότητας ακολουθιών. Ο πίνακας γειτνίασης του δικτύου αυτού κανονικοποιείται από 0.001 έως 1 όπως έγινε και με τον πίνακα γειτνίασης των δομών. Στη συνέχεια υπολογίζει τη συνάρτηση καταλληλότητας μέσω του υπολογισμού της συνάρτησης κόστους, βάσει της σχέσης (4.1). Σε κάθε διαφορετικό χρωμόσωμα, αντιστοιχεί ένα διαφορετικό δίκτυο ομοιότητας πρωτεϊνικών ακολουθιών και μια διαφορετική τιμή για την συνάρτηση καταλληλότητας.

5.4 *Αποτελέσματα και αξιολόγηση χρήσης ΓΑ*

Ο ΓΑ υλοποιήθηκε σε Matlab και ως μέγεθος πληθυσμού επιλέχθηκε 100 (100 χρωμοσώματα). Δεν ορίστηκε χρονικός περιορισμός στη προσπάθεια βελτιστοποίησης αλλά ο τερματισμός του αλγορίθμου ορίστηκε να συμβεί μετά από 49 γενεές. Αρχικά η τιμή της συνάρτησης καταλληλότητας, που προέκυπτε από την απόσταση των δύο δικτύων ήταν 18221, βάσει του πρώτου τυχαία επιλεγμένου χρωμοσώματος. Τελικά ο γενετικός αλγόριθμος βρήκε τιμή στη συνάρτηση καταλληλότητας στο τελικό σημείο ίση με 3272 και ως τελικό χρωμόσωμα, ένα σύνολο τιμών για τα 125 γονίδια-χαρακτηριστικά (ελάχιστη τιμή: -1.3986 και μέγιστη τιμή: 2.2511). Έγινε επιλογή και κρατήσαμε μόνο τα χαρακτηριστικά εκείνα για τα οποία η τιμή ήταν μεγαλύτερη από 0,5.

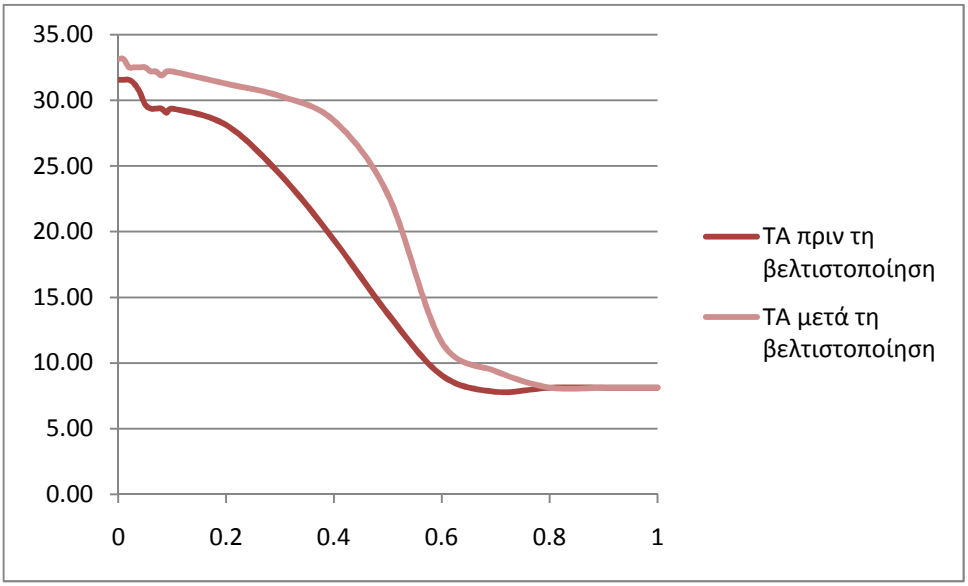
Τα χαρακτηριστικά που υπέδειξε ο γενετικός αλγόριθμος και κατ’επέκταση θεωρήθηκαν ότι περιέχουν σημαντική βιολογική πληροφορία για τη δομή των πρωτεϊνικών ακολουθιών, ανήκουν στις παρακάτω κατηγορίες: αμινοξική σύσταση (3 χαρακτηριστικά), προβλεπόμενη δευτεροταγής δομή των καταλοίπων της ακολουθίας (5 χαρακτηριστικά), κανονικοποιημένος όγκος van der Waals (4 χαρακτηριστικά), υδροφοβοκότητα (19 χαρακτηριστικά), πολικότητα (10

χαρακτηριστικά) και την πολωσιμότητα (7 χαρακτηριστικά), σύνολο 48 χαρακτηριστικά. Αξιολογώντας τη βιολογική τους σημασία, βλέπουμε ότι η υδροφοβικότητα καθορίζει την πρωτεϊνική δομή σε μεγαλύτερο βαθμό από από ότι η αμινοξική σύσταση.

Αφού κρατήσαμε 48 από τα αρχικά 125 χαρακτηριστικά επαναλάβαμε τα πειράματα αξιολόγησης του ταξινομητή με τη μέθοδο της διασταυρωμένης επικύρωσης, χρησιμοποιώντας το ίδιο σύνολο πρωτεϊνικών ακολουθιών. Στον Πίνακα 5.1 παρουσιάζονται τα αποτελέσματα ακρίβειας της αυτόματης λειτουργίας του ταξινομητή για διάφορες τιμές του ***α***, και γίνεται σύγκριση τους με τα αντίστοιχα αποτελέσματα πριν την επιλογή των πιο εύρωστων χαρακτηριστικών με τη χρήση του ΓΑ.

α	***TA_1 μετά τη βελτιστοποίηση***	***TA_1 πριν τη βελτιστοποίηση***	***Απόλυτη Μεταβολή***	***Μεταβολή %***
0	33.13	31.56	+1.56	+4.95%
0.01	33.13	31.56	+1.56	+4.95%
0.02	32.50	31.56	+0.94	+2.97%
0.03	32.50	31.25	+1.25	+4.00%
0.04	32.50	30.63	+1.88	+6.12%
0.05	32.50	29.69	+2.81	+9.47%
0.06	32.19	29.38	+2.81	+9.57%
0.07	32.19	29.38	+2.81	+9.57%
0.08	31.88	29.38	+2.50	+8.51%
0.09	32.19	29.06	+3.13	+10.75%
0.1	32.19	29.38	+2.81	+9.57%
0.2	31.25	28.13	+3.13	+11.11%
0.3	30.31	24.38	+5.94	+24.36%
0.4	28.44	19.38	+9.06	+46.77%
0.5	22.81	13.75	+9.06	+65.91%
0.6	11.56	9.06	+2.50	+27.59%
0.7	9.38	7.81	+1.56	+20.00%
0.8	8.13	8.13	+0.00	+0.00%
0.9	8.13	8.13	+0.00	+0.00%
1	8.13	8.13	+0.00	+0.00%

Πίνακας 5.1 Συγκριτικά αποτελέσματα προ και μετά εφαρμογής γενετικού αλγορίθμου για τη βελτιστοποίηση του ταξινομητή.



Εικόνα 5.1 Γραφική αναπαράσταση αποτελεσμάτων προ και μετά την εφαρμογή του γενετικού αλγορίθμου για τη βελτιστοποίηση του ταξινομητή.

Παρατηρούμε ότι ο ταξινομητής εμφανίζει βελτίωση της τάξης του 5% για τα καλύτερα προηγούμενα αποτελέσματα, και δίνει τελικό ποσοστό ακρίβειας αυτόματης πρόγνωσης του σωστού διπλώματος ίσο με **33.13%** Για τιμές του α που δίνουν χειρότερα αποτελέσματα ακρίβειας, η βελτίωση φτάνει ακόμα και το 65%.

6

Επίλογος

Με το κεφάλαιο αυτό ολοκληρώνεται η παρούσα διπλωματική εργασία. Στην Ενότητα 8.1 γίνεται μία σύνοψη της εργασίας και παρουσιάζονται κάποια βασικά συμπεράσματα. Στην ενότητα 8.2 περιγράφονται ορισμένες ενέργειες που θα μπορούσαν μελλοντικά να επεκτείνουν την εργασία αυτή.

6.1 *Σύνοψη*

Το πρόβλημα της αναγνώρισης διπλώματος πρωτεϊνών απασχολεί πολύ την κοινότητα της Βιοπληροφορικής. Αυτό οφείλεται στο γεγονός ότι η δομή των πρωτεϊνών είναι καίριας σημασίας για τη ζωή των έμβιων όντων και την καθορίζει με πολλούς τρόπους. Επειδή όμως δεν μπορούν όλες οι δομές να ταυτοποιηθούν εργαστηριακά, γίνεται προσπάθεια πρόγνωσής τους με χρήση εξελιγμένων μεθόδων μηχανικής μάθησης.

Στην εργασία αυτή υλοποιήθηκε ένας ταξινομητής με σκοπό την κατηγοριοποίηση πρωτεϊνών αγνώστου διπλώματος σε 27 από κύριες κατηγορίες της ταξινόμησης κατά SCOP. Σκοπός μας ήταν να δημιουργηθεί ένα μοντέλο το οποίο μπορεί να εφαρμοστεί για να κατηγοριοποιήσει δεδομένα (πρωτεΐνες) που δεν έχουν

ακόμα κατηγοριοποιηθεί. Στη δικιά μας προσέγγιση του προβλήματος χρησιμοποιήσαμε δίκτυα ομοιότητας πρωτεϊνικών ακολουθιών και δομών.

Η διαδικασία βασίστηκε στην χρησιμοποίηση χαρακτηριστικών εξαγμένων από την ακολουθία ενός συνόλου δεδομένων που χρησιμοποιείται συχνά για την αξιολόγηση, παρόμοιων προτάσεων για την αντιμετώπιση του προβλήματος αναγνώρισης διπλώματος. Αλλά ο μεγάλος αριθμός των χαρακτηριστικών σε προβλήματα ταξινόμησης συνδέεται γενικά στη βιβλιογραφία με θόρυβο στα δεδομένα και λιγότερο αποδοτικούς ταξινομητές. Για το λόγο αυτό προσπαθήσαμε να χρησιμοποιήσουμε τα χαρακτηριστικά εκείνα για τα οποία το δίκτυο ομοιότητας ακολουθιών μοιάζει όσο το δυνατόν περισσότερο με το δίκτυο ομοιότητας δομών. Αυτό έγινε μέσω της συνάρτησης καταλληλότητας ενός γενετικού αλγορίθμου και επιλέχτηκαν τα πιο εύρωστα χαρακτηριστικά, με αποτέλεσμα να βελτιωθεί η απόδοση του ταξινομητή κατά 5% (στην αυτόματη λειτουργία).

6.2 *Συμπεράσματα*

Υπάρχουν πολλοί τρόποι αντιμετώπισης του προβλήματος αναγνώρισης διπλώματος πρωτεϊνών που έχουν αναπτυχθεί από διάφορους ερευνητές. Σκοπός μας ήταν να δημιουργήσουμε ένα ταξινομητή μικρής πολυπλοκότητας, που να είναι γενικευμένος και να μπορεί να εφαρμοστεί σε όλα τα σύνολα δεδομένων πρωτεϊνικών ακολουθιών ανεξάρτητα από το μέγεθός τους και τον αριθμό των χαρακτηριστικών που εξάγονται από αυτές. Η απόδοση του αλγορίθμου είναι συγκρίσιμη με την απόδοση ταξινομητών μεγαλύτερης πολυπλοκότητας που άλλοι ερευνητές χρησιμοποίησαν στο ίδιο πρόβλημα.

Τα ενθαρρυντικά αποτελέσματα που προέκυψαν δίνουν έναυσμα για μελλοντικές επεκτάσεις της εργασίας. Για παράδειγμα θα μπορούσαν να χρησιμοποιηθούν και άλλες μετρικές για την κατασκευή του δικτύου ομοιότητας πρωτεϊνικών ακολουθιών. Επίσης θα μπορούσαν να γίνουν πειράματα με μεγαλύτερα σύνολα δεδομένων είτε ακολουθιών είτε χαρακτηριστικών. Τέλος θα μπορούσε να δημιουργηθεί και ένα γραφικό περιβάλλον στο οποίο θα μπορούσαν να εισάγονται οι πρωτεϊνικές ακολουθίες με τα χαρακτηριστικά τους από το χρήστη του ταξινομητή και μετά την επεξεργασία να εμφανίζονται τα αποτελέσματα.

7

Βιβλιογραφία

1. **Κ. Νικήτα, Καθηγήτρια ΕΜΠ.** *Προσομοίωση Φυσιολογικών Συστημάτων.* Αθήνα : ΕΜΠ, 2007.

2. **Linge, Michael Nilges and Jens P.** *Unité de Bio–Informatique* . Paris : Institut Pasteur, 2003.

3. **Dictionary, Oxford English.** *http://www.oed.com.*

4. **Εμμ.Φραγκούλης.** *Βιοχημεία.,* Κ.Ε. Σέκερης, 1984.

5. **Αλεπόρου-Μαρίνου Β., Αργυροκαστρίτης Α., Κομητοπούλου Α., Πιαλόγλου Π., Σγουρίτσα Β.** *Βιολογία Θετικής Κατεύθυνσης Γ'τάξης Ενιαίου Λυκείου.* Αθήνα : ΟΕΔΒ, 2005.

6. **Μ. Καρμιράντζου, Β. Προμπονάς, Σ. Χαμόδρακας.** *Μαθήματα Βιοφυσικής http://biophysics.biol.uoa.gr/courses.* Αθήνα : Τμήμα Βιολογίας, 2008.

7. **Τσιράκης Ν., Σκούρα Α.** *Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης.* , Πανεπιστήμιο Πατρών.

8. **Ι., Βαλαβάνης.** *Ανάπτυξη υπολογιστικύ συστήματος υποστήριξης διάγνωσης εστιακών βλαβών ήπατος από εικόνες αξονικής τομογραφίας με χρήση χαρακτηριστικών υφής και τεχνικών πολλαπλής ταξινόμησης.* Αθήνα : ΕΜΠ, 2003.

9. **Κ., Δημητρακοπούλου.** *Αναγνώριση λειτουργικών υπο-δομών στο πρωτεϊνικό δίκτυο του SACCHAROMYCES CEREVISIAE συνδυάζοντας δεδομένα έκφρασης γονιδίων και αλληλεπίδρασης πρωτεϊνών,* Πανεπιστήμιο Πατρών, 2007.

10. **Βλαχάβας, Κεφαλάς, Βασιλειάδης, Κόκκορας, Σακελαρίου.** *Τεχνητή Νοημοσύνη*, Γκιούρδας, 2006.

11. **Τζαφέστας Σ., Καθηγητής ΕΜΠ.** *Υπολογιστική Νοημοσύνη.* Αθήνα : s.n., 2002.

12. **R.Apweiler et. al,** *UniProt: the Universal Protein knowledge base.* D115-D119,Nucl. Ac. Res., 2004, Vol. 32.

13. **H.M.Berman et. al,** *The Protein Data Bank.* pp. 235-242, Nucl. Ac. Res., 2000, Vol. 28.

14. **Murzin et. al,** *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* 1995.

15. **University College of London,** *http://www.biochem.ucl.ac.uk/bsm/cath/*

16. *Multi-class protein fold recognition using support vector machines and neural networks.* **C.H.Q. Ding, I.Dubchak.** Bioinformatics, vol.7, pp.349-358, 2001.

17. **A.C.Tan, D.Gilbert, and Y.Deville.,** *Multi -class protein fold classification using a new ensemble machine learning approach.* pp.206-217, Genome Informatics, 2003, Vol. 14.

18. **S.Y.M. Shi, P.N. Suganthan, and K.Deb.,** *Multi -class protein fold recognition using multi-objective evolutionary algorithms.* IEEE, Proc IEEE CIBCB, 2004.

19. **A. Chinnasamy, W.K. Sung and A.Mittal.,** *Protein Structure and Fold Prediction using tree-augmented naive bayesian classifier.* Stanford CA : World Scientific Press, 2004.

20. **Valavanis I., Spyrou G., Nikita K.** *Analysis and Comparison of Protein Sequence- and Structure - based Similarity Networks,* submitted to Biomedical Informatics, 2008.

21. **Weston J., Andre E., Zhou D., Leslie C., Noble W.S.** *Protein Ranking: From local to global structure in the protein similarity Network,* PNAS, 2004.

22. **I.Dubchak et. al,** *Recognition of a protein fold in the context of the Stuctural Classification of Proteins (SCOP) classification,* Proteins, vol.35, pp.401-407, 1999.

23. **I. K. Valavanis, G. M. Spyrou, K. S. Nikita.** , *Protein Similarity Networks and Genetic Algorithm Driven Feature Selection for Fold Recognition.* Athens : 8 th IEEE International Conference on Bioinformatics and Bioengineering, October 2008.

24. **J.Weston, A.Eliseef,D.Zhou,C.S.Leslie,W.S.Noble.,** *Protein Ranking: From local to global structure in the protein similarity network,* Proc. Natl. Acad. Sci.USA, vol.101, pp. 6559-6563, 2004.

Τα παρακάτω χρήσιμα links χρησιμοποιήθηκαν σε αρκετά σημεία της εργασίας:

http://en.wikipedia.org

http://www.ebi.ac.uk/

http://el.science.wikia.com/wiki

Παράρτημα

Στο παράρτημα παρατίθενται οι κυριότερες συναρτήσεις που υλοποιήθηκαν σε Matlab για την δημιουργία του ταξινομητή, του γενετικού αλγορίθμου και την εκτέλεση των πειραμάτων.

main (βασική συνάρτηση μέσω της οποίας έτρεχε ο αλγόριθμος, με διαφορετική υλοποίηση για κάθε περιθώριο λάθους)

```
clear;
table = zeros (10,1);
global set;
global comparison;
global percentage;

%success_each_protein = zeros (320,2); %TA_1
%success_each_protein = zeros (320,4); %TA_3
%success_each_protein = zeros (320,7); %TA_6
%success_each_protein = zeros (320,11); %TA_10

for set = 1:10
set
global CV_train_sets;
global CV_test_sets;

load -mat 10_CV_tests_all_125_feats_01_and_indexed_targets.mat

percentage = test_cases_TA1(set,set);

table(set,1) = percentage;

if set==1
    success_each_protein = comparison;
else
    success_each_protein = vertcat(success_each_protein, comparison);
end

end

fold_success_matrix = zeros(27,4);
```

```
input_array = success_each_protein;

%TA_1
for j=1:27
counter = 0;
counter_correct = 0;

    for i=1:320

        if input_array(i,1)==j
            counter=counter+1;

            if input_array(i,1)==input_array(i,2)
               counter_correct= counter_correct+1;

            end

        end

    end
    fold_success_matrix(j,1) = j;
    fold_success_matrix(j,2) = counter;
    fold_success_matrix(j,3) = counter_correct;
    fold_success_matrix(j,4) = counter_correct*(100/counter);

end


%TA_3
% for j=1:27
% counter = 0;
% counter_correct = 0;
%
%     for i=1:320
%
%         if input_array(i,1)==j
%             counter=counter+1;
%
%             if ((input_array(i,1)==input_array(i,2)) ||
(input_array(i,1)==input_array(i,3))||(input_array(i,1)==input_array(
i,4)))
%                counter_correct= counter_correct+1;
%
%             end
%
%         end
%
%     end
%     fold_success_matrix(j,1) = j;
%     fold_success_matrix(j,2) = counter;
%     fold_success_matrix(j,3) = counter_correct;
%     fold_success_matrix(j,4) = counter_correct*(100/counter);
%
% end

%TA_6
% for j=1:27
% counter = 0;
% counter_correct = 0;
```

```
%
%     for i=1:320
%
%         if input_array(i,1)==j
%             counter=counter+1;
%
%             if ((input_array(i,1)==input_array(i,2)) ||
(input_array(i,1)==input_array(i,3))||(input_array(i,1)==input_array(
i,4))||(input_array(i,1)==input_array(i,5))||(input_array(i,1)==input
_array(i,6))||(input_array(i,1)==input_array(i,7)))
%                counter_correct= counter_correct+1;
%
%             end
%
%         end
%
%     end
%     fold_success_matrix(j,1) = j;
%     fold_success_matrix(j,2) = counter;
%     fold_success_matrix(j,3) = counter_correct;
%     fold_success_matrix(j,4) = counter_correct*(100/counter);
%
% end
```

Test_cases (με διαφορετική υλοποίηση για κάθε περιθώριο λάθους, καλείται μέσω της main και υπολογίζει την επιτυχία σε κάθε δίπλωμα)

```
function percentage = test_cases(i_test,i_train)


global CV_train_sets;
global CV_test_sets;
global comparison;

train_set_f = CV_train_sets{1,i_train};   % o pinakas train_set einai
enas pinakas m-1xp pou periexei ta  xaraktiristika kai to fold twn
prwteinwn
test_set_f = CV_test_sets{1,i_test};      % o pinakas test_set einai
enas pinakas 32xp pou periexei ta 125 xaraktiristika kai to fold

[n,p] = size(test_set_f);


%apothikeysi tis stilis twn fold tou test_set gia sygkrisi twn
apotelesmatwn
right_folds = zeros (n,1);      %arxikopoiisi
right_folds = test_set_f(:,p);    %apothikeusi tis stilis twn fold


my_results = zeros (n,1);  %arxikopoiisi enos pinaka stili gia tin
apothikeusi twn apotelesmatwn

for all_queries=1:n     % opou n to megethos tou test set
```

```
    query_protein = test_set_f(all_queries,:);          % oles oi
prwteines tou test_set
    query_output =
graph_based_classification_v3_TA1(query_protein,train_set_f);
    my_results (all_queries,1)= query_output;

end

correct = 0;

for i_n=1:n
    if (((my_results(i_n,1)) == (right_folds(i_n,1))))
       correct = correct+1;
    end


end
percentage = correct*(100/n);
comparison = horzcat(right_folds, my_results);
```

graph_based_classification (η βασική συνάρτηση που καλείται μέσω της test_cases, κατασκευάζει το δίκτυο ομοιότητας πρωτεϊνών, υπολογίζει τους βαθμούς συσχέτισης, δέχεται ως είσοδο την ακολουθία αναζήτησης και τις υπόλοιπες πρωτεϊνικές ακολουθίες γνωστού διπλώματος, και δίνει ως έξοδο το προτεινόμενο ως σωστό δίπλωμα. Και αυτή έχει διαφορετική υλοποίηση για κάθε περιθώριο λάθος απόδοσης διπλώματος και αριθμό χαρακτηριστικών)

```
function query_output
=graph_based_classification_v3_TA1(query_protein,train_set)

%% Train_set (input1)
                                                  % apodosi tis timis
tou b ston c wste na diagrafei i 126i stili,i stili twn fold
m = size(train_set, 1) + 1;                        % m o arithmos twn
grammwn tou pinaka twn prwteinwn


%% Query_Protein (input2)

c = query_protein;                                % apodosi tis timis
tou query_features ston c wste na diagrafei i stili twn fold
p = size(query_protein, 2);
num_of_features = p - 1;                  %sto test set exw kai tin
stili tou fold



con_array = [query_protein;train_set];      % topothetisi stin 1i
grammi tou pinaka twn xaraktiristikwn tis query protein

    con_array(:,p) = [];          % diagrafi tis stilis twn fold
```

```
    e = zeros(m,m);     % arxikopoisi enos mxm pinaka e wste na
apothikeutoun ekei oi times tou e

    for h = 1:1:m             %

        for l = 1:1:m           % sygkrisi tis h me tis upoloipes m-1
prwteines
                                    % thelw ta stoixeia b[h,:] kai
b[l,:]
            d = con_array(h,:) - con_array(l,:);        % o pinakas
con_array einai 1xnum_of_features kai periexei ta pi-qi twn
xaraktiristikwn tis h kai h+1 prwteinis
            d = d.^2;                   % (pi-qi)^2
            d = sum(d);                 % athroisma olwn twn
tetragwnwn twn diaforwn twn  xaraktiristikwn
            d = sqrt(d);                % euresi tis timis tis
apostasis

            d = 1/d;                    % euresi antistrofou
            e(h,l) = d;                 % apothikeusi tis timis tis
apostasis sti thesi h,h+1..m-1 enos pinaka e
            e(l,h) = d;

        end
    end


for i=1:m

   values_i=e(i,:);
   values_i(i)=[];
   e(i,i)=max(values_i);
end

%% Normalization of similarity matrix k

    k = zeros (m,m);            %arxikopoiisi tou pinaka k
    k(1:m,1)=e(:,1);

    for i = 1:m               %gia kathe grammi
        sum_xi = 0;

        for j = 2:m
               sum_xi = sum_xi + e(i,j);
        end


        for l = 2:m
            k(i,l) = e(i,l)/sum_xi;
        end

    end


%% Epanaliptiki sxesi
% k - normalized similarity matrix
```

```
% y - the affinity of the query protein to the sequences-vertices
x2.....xm

    a = 0;
    y_previous = zeros(1, m-1);
    y_current = ones(1, m-1);     % arxikopoiisi tetoia gia na
mpainei mesa sto while
    t = 0;                       % metabliti pou deixnei ton arithmo
twn epanalipsewn
    y = 0;                       % arxikopoiisi tou affinity gia kathe
i-j


    while (((sqrt(sum(((y_current) - (y_previous)).^2))))> 0.000001)
% threshold σύγκλισης
        t = t + 1;
        y_previous = y_current;

        for i_affinity = 2:m         % upologismos tou affinity tis
query me kathe mia apo tis m-1 prwteines
            sum_k = 0;

            for g = 2:m;            % upologismos tou sum_k
                sum_k = sum_k + k(g,i_affinity);
            end

            y_current(1,(i_affinity-1)) = k(1,i_affinity) + (a *
sum_k * y_previous(1,i_affinity-1));

         end

    end

%% Anazitisi tou swstoterou fold

[f, r] = sort(y_current);


%% Apodosi tou fold stin ekdodo (output)
query_output = train_set(r(1, m-1), p);
```

create_struct (μέσω της συνάρτησης αυτής κατασκευάζεται ο γράφος των δομών που χρησιμοποιείται μέσα στι συνάρτηση καταλληλότητας του γενετικού αλγορίθμου, ο οποίος και αποθηκεύεται στο αρχείο adj_struct_matrix.mat)

```
%Σκοπός αυτού του αρχείου είναι η δημιουργία του πίνακα
%adj_struct_matrix.mat που θα έχει μέγεθος (i,j) και θα περιέχει τις
%συχετίσεις των δομών.

load results_corrected_v2.mat
%αρχικά αφαιρώ από τον πίνακα z_scores τις τελευταίες γραμμές και
στήλες
my_matrix = z_scores;
```

```
my_matrix(298,:) = [];
my_matrix(297,:) = [];
my_matrix(:,298) = [];
my_matrix(:,297) = [];


%% Εύρεση min max του z_scores για την μετέπειτα κανονικοποίηση του
πίνακα my_matrix
m = 1;

for i = 1:296                %για όλες τις γραμμές και όλες τις στήλες
    for j = 1:296
        if (z_scores(i,j) ~= (-1)) && (i~=j);
            trial_array(1,m) = z_scores(i,j);
            m = m + 1;
        end
    end
end

z_scores_min = min(trial_array);
z_scores_max = max(trial_array);


%% Δημιουργία του πίνακα my_matrix

for i = 1:296
    for j = 1:296

        if z_scores(i,j) == -1
           my_matrix(i,j) = 0.001;
        end

        if i==j
           my_matrix(i,j) = 1;
        end

        if (z_scores(i,j) ~= (-1)) && (i~=j)
            my_matrix(i,j) = ((z_scores(i,j)-z_scores_min)*(0.99-
0.01)/(z_scores_max-z_scores_min))+0.01;     %υπολογισμός
κανονικοποιημένης τιμής για κάθε κελί του πίνακα
        end
    end
end

a = min(my_matrix,[],1);
b = max(my_matrix,[],1);

adj_struct_matrix = my_matrix;

savefile = 'adj_struct_matrix.mat';
save(savefile, 'adj_struct_matrix');
```

my_graph_similarity_fitnessfun (συνάρτηση καταλληλότητας για την υλοποίηση του γενετικού αλγόριθμου, δέχεται ως είσοδο το εκάστοτε

χρωμόσωμα και δίνει ως έξοδο μία τιμή μη-ομοιότητας των δυο δικτύων, για το χρωμόσωμα αυτό)

```
function[fitness]=my_graph_similarity_fitnessfun(which_feats_to_consi
der)

load features_min_max_normed_values.mat
load results_corrected_v2.mat
load adj_struct_matrix.mat
load sequence_matrix_296_features_125.mat
global time
global stili
global all_fitness

%% Με βάση to which_feats_to_cosider κρατάω κάποια από τα 125
χαρακτηριστικά
features_added_in_new_matrix=0;
for i=1:125
    if  which_feats_to_consider(1,i) > 0.5
        features_added_in_new_matrix =
features_added_in_new_matrix+1;
        sequence_matrix_n_features(1:296,
features_added_in_new_matrix) =
sequence_matrix_296_features_125(1:296,i);
    end
end
time=time+1;
time


sequence_matrix_296x296 = zeros(296,296);     % arxikopoisi enos mxm
pinaka e wste na apothikeutoun ekei oi times tou e

    for h = 1:1:296             % ews m-1
        for l = 1:1:296    % sygkrinw tin h me tis upoloipes 296
proteines thelw ta stoixeia b[h,:] kai b[l,:]

            d = sequence_matrix_n_features(h,:) -
sequence_matrix_n_features(l,:);       % periexei ta pi-qi twn
xaraktiristikwn tis h kai h+1 prwteinis
            d = d.^2;                   % (pi-qi)^2
            d = sum(d);                 % athroisma olwn twn
tetragwnwn twn diaforwn twn  xaraktiristikwn
            d = sqrt(d);                % euresi tis timis tis
apostasis


            d = 1/d;                    % euresi antistrofou
            sequence_matrix_296x296(h,l) = d;                 %
apothikeusi tis timis tis apostasis sti thesi h,h+1..m-1 enos pinaka
e
            sequence_matrix_296x296(l,h) = d;

        end
    end
```

```
%% Κανονικοποιήση


 m = 1;
    for i = 1:296                %για όλες τις γραμμές και όλες τις
στήλες
        for j = 1:296
            if  (i~=j);
                array(1,m) = sequence_matrix_296x296 (i,j);
                m = m + 1;
            end
        end
    end

sequence_matrix_296x296_min = min(array);
sequence_matrix_296x296_max = max(array);



    for i = 1:296
        for j = 1:296

            if  i==j
                adj_sequence_matrix(i,j) = 1;
            end

            if  i~=j
                adj_sequence_matrix(i,j) =
((sequence_matrix_296x296(i,j)-sequence_matrix_296x296_min)*(0.99-
0.001)/(sequence_matrix_296x296_max-
sequence_matrix_296x296_min))+0.001;
            end

        end
    end


    fitness = sum(sum(abs(adj_struct_matrix-adj_sequence_matrix)));
    all_fitness(1,stili)= fitness;
    stili = stili + 1;
end
```

my_optimize_sequence_features (κλήση του γενετικού αλγορίθμου και ορισμός παραμέτρων, το τελικό χρωμόσωμα είναι ο πίνακας which_feats_to _consider)

```
global time
global stili
global all_fitness
time=1;
stili=1;
```

```
options =
gaoptimset('PopulationSize',100,'Generations',1,'StallTimeLimit',Inf)
;
[which_feats_to_consider fval reason output population scores] =
ga(@my_graph_similarity_fitnessfun, 125, options);
```