«Πρόβλημα αναγνώρισης της αναδίπλωσης μιας πρωτεΐνης: Μια πρόταση επίλυσης σε Λογικό Προγραμματισμό με Διαχείριση Περιορισμών»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΔΙΑΜΑΝΤΟΠΟΥΛΟΥ ΝΙΚΟΛΑΟΥ

Α.Ε.Μ.: 405

Επιβλέπων: Λέκτορας Σωτήριος Κωτσιαντής

Πάτρα, Σεπτέμβριος 2014

Πανεπιστήμιο Πατρών
Σχολή Θετικών Επιστημών
Τμήμα Μαθηματικών
ΜΠΣ « Μαθηματικά και Σύγχρονες Εφαρμογές»

Τριμελής Επιτροπή

1. **Κωτσιαντής Σωτήριος:** Λέκτορας Τμήματος Μαθηματικών, Πανεπιστημίου Πατρών

2. **Ράγγος Όμηρος:** Επίκουρος Καθηγητής Τμήματος Μαθηματικών, Πανεπιστημίου Πατρών

3. **Καββαδίας Δημήτριος:** Επίκουρος Καθηγητής Τμήματος Μαθηματικών, Πανεπιστημίου Πατρών

Ευχαριστίες

Η εκπόνηση της παρούσας διπλωματικής εργασίας πραγματοποιήθηκε υπό την επίβλεψη του Λέκτορα κ. Κωτσιαντή Σωτήριου (Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών) και θα ήθελα να τον ευχαριστήσω θερμά που μου έδωσε την ευκαιρία να ασχοληθώ με ένα τόσο σημαντικό τομέα, όπως αυτός της Βιοπληροφορικής. Η καθοδήγηση και η βοήθειά του ήταν καθοριστικής σημασίας για την εργασία.

Θα ήθελα επίσης να ευχαριστήσω τους Επίκουρους Καθηγητές, κ. Ράγγο Όμηρο και κ. Καββαδία Δημήτριο, για την επίσης πολύτιμη βοήθεια και συμβολή που είχαν για την εκπόνηση της εργασίας.

Τέλος θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την στήριξη που μου προσέφεραν και για την υπομονή τους.

Περίληψη

Η κατανόηση των μοριακών μηχανισμών της ζωής απαιτεί την αποκωδικοποίηση των λειτουργιών που εκτελούν οι πρωτεΐνες σε έναν οργανισμό. Δεκάδες χιλιάδες πρωτεΐνες έχουν μελετηθεί τα τελευταία χρόνια σε στόχο την εύρεση της τρισδιάστατης δομής τους που στην ουσία καθορίζει και την λειτουργία τους. Παρ' όλα αυτά οι πειραματικές μέθοδοι που χρησιμοποιούνται παρότι είναι ακριβείς μέθοδοι είναι ακόμα ιδιαιτέρως δύσκολες και απαιτητικές σε χρόνο και οικονομικό κόστος. Για το λόγο αυτό επιχειρείται, με υπολογιστικές μεθόδους, να μειωθεί το κόστος της πρόβλεψης της 3D δομής μιας πρωτεΐνης όταν η γραμμική ακολουθία των αμινοξέων που την αποτελούν είναι γνωστή.

Η παρούσα διπλωματική εργασία προσεγγίζει το πρόβλημα αυτό ως πρόβλημα βελτιστοποίησης για τη λύση του οποίου εφαρμόζεται μία δηλωτική προσέγγιση σε Λογικό Προγραμματισμό με Διαχείριση Περιορισμών. Η πρόταση βασίζεται σε ένα προσωποκεντρικό κυβικό πλέγμα ως τοπολογικό μοντέλο για την χωροθέτηση της πρωτεΐνης. Χρησιμοποιούνται επίσης πληροφορίες που αφορούν τυχόν δευτερεύουσες δομές που υπάρχουν στην πρωτεΐνη καθώς και άλλα heuristics για να περιοριστεί σημαντικά ο χώρος αναζήτησης. Τα πρώτα αποτελέσματα σε πραγματικές πρωτεΐνες είναι ενθαρρυντικά τόσο όσο αφορά την ακρίβεια και την χρονική επίδοση αλλά και όσον αφορά την κλιμάκωση του προβλήματος.

Λέξεις κλειδιά: Λογικός Προγραμματισμός με Διαχείριση Περιορισμών - CLP(FD), Βιοπληροφορική , Αναδίπλωση πρωτεΐνης.

Περιεχόμενα

1 ΕΙΣΑΓΩΓΗ 10

1.1 ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ 10
1.2 ΑΝΤΙΚΕΙΜΕΝΟ ΔΙΠΛΩΜΑΤΙΚΗΣ 13
1.2.1 ΣΥΝΕΙΣΦΟΡΑ 13
1.3 ΟΡΓΑΝΩΣΗ ΚΕΙΜΕΝΟΥ 14

2 ΠΡΩΤΕΪΝΕΣ 15

2.1 ΡΟΗ ΤΗΣ ΓΕΝΕΤΙΚΗΣ ΠΛΗΡΟΦΟΡΙΑΣ 15
2.1.1 ΒΙΟΛΟΓΙΚΑ ΜΑΚΡΟΜΟΡΙΑ 15
2.1.2 ΚΕΝΤΡΙΚΟ ΔΟΓΜΑ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ 18
2.1.2.1 Αντιγραφή 18
2.1.2.2 Μεταγραφή 19
2.1.2.3 Μετάφραση 19
2.2 ΔΟΜΗ ΠΡΩΤΕΪΝΩΝ 21
2.2.1 ΑΜΙΝΟΞΕΑ: ΤΑ ΔΟΜΙΚΑ ΣΤΟΙΧΕΙΑ ΤΩΝ ΠΡΩΤΕΪΝΩΝ 22
2.2.2 ΕΠΙΠΕΔΑ ΔΟΜΗΣ ΤΩΝ ΠΡΩΤΕΪΝΩΝ 29
2.3 ΣΤΟΙΧΕΙΑ ΔΕΥΤΕΡΟΤΑΓΟΥΣ ΔΟΜΗΣ ΠΡΩΤΕΪΝΩΝ 33
2.3.1 Α-ΕΛΙΚΑ 34
2.3.2 Β-ΠΤΥΧΩΤΕΣ ΕΠΙΦΑΝΕΙΕΣ 36
2.3.3 ΜΗ ΚΑΝΟΝΙΚΗ ΔΟΜΗ (COIL) 40
2.4 ΣΤΕΡΕΟΧΗΜΙΚΗ ΔΙΑΜΟΡΦΩΣΗ ΠΡΩΤΕΪΝΩΝ 41
2.5 ΛΕΙΤΟΥΡΓΙΕΣ ΤΩΝ ΠΡΩΤΕΪΝΩΝ 43
2.5.1 ΕΝΖΥΜΙΚΗ ΚΑΤΑΛΥΤΙΚΗ ΔΡΑΣΗ 43
2.5.2 ΜΗΧΑΝΙΣΜΟΣ ΑΜΥΝΑΣ (ΑΝΟΣΟΠΡΟΣΤΑΣΙΑ) 44
2.5.3 ΜΕΤΑΦΟΡΑ ΚΑΙ ΑΠΟΘΗΚΕΥΣΗ ΟΥΣΙΩΝ 44
2.5.4 ΛΕΙΤΟΥΡΓΙΑ ΚΙΝΗΣΗΣ 45
2.5.5 ΔΙΑΚΥΤΤΑΡΙΚΗ ΕΠΙΚΟΙΝΩΝΙΑ 45
2.5.6 ΜΗΧΑΝΙΚΗ ΥΠΟΣΤΗΡΙΞΗ 46
2.5.7 ΈΛΕΓΧΟΣ ΤΗΣ ΑΝΑΠΤΥΞΗΣ ΚΑΙ ΤΗΣ ΔΙΑΦΟΡΟΠΟΙΗΣΗΣ 46
2.5.8 ΔΗΜΙΟΥΡΓΙΑ ΚΑΙ ΜΕΤΑΔΟΣΗ ΝΕΥΡΙΚΩΝ ΠΑΛΜΩΝ 46
2.6 ΑΛΛΗΛΕΠΙΔΡΑΣΗ ΠΡΩΤΕΪΝΩΝ 47

3 ΠΡΟΒΛΗΜΑ ΑΝΑΓΝΩΡΙΣΗΣ ΑΝΑΔΙΠΛΩΣΗΣ ΠΡΩΤΕΪΝΩΝ 49

3.1 ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ 49
3.2 ΜΕΘΟΔΟΙ ΠΡΟΓΝΩΣΗΣ ΑΝΑΔΙΠΛΩΣΗΣ ΠΡΩΤΕΪΝΩΝ 50
3.2.1 ΜΕΘΟΔΟΙ ΠΡΟΤΥΠΟΠΟΙΗΣΗΣ ΜΕ ΟΜΟΛΟΓΙΑ (HOMOLOGY MODELING) 53

3.2.2 ΠΡΟΒΛΕΨΗ ΔΟΜΗΣ ΜΕ ΤΟ ΛΟΓΙΣΜΙΚΟ MODELLER 57
3.2.3 ΜΕΘΟΔΟΙ ΥΦΑΝΣΗΣ ΠΡΩΤΕΪΝΩΝ (PROTEIN THREADING) 59
3.2.4 ΜΕΘΟΔΟΙ ΑΠΑΡΧΗΣ (AB INITIO) 60
3.3 ΣΥΜΠΕΡΑΣΜΑΤΑ 60

4 ΛΟΓΙΚΟΣ ΠΡΟΓ/ΣΜΟΣ ΜΕ ΔΙΑΧΕΙΡΙΣΗ ΠΕΡΙΟΡΙΣΜΩΝ 62

4.1 ΛΟΓΙΚΟΣ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΣ 62
4.1.1 Η ΓΛΩΣΣΑ PROLOG 64
4.2 ΠΡΟΒΛΗΜΑΤΑ ΙΚΑΝΟΠΟΙΗΣΗΣ ΠΕΡΙΟΡΙΣΜΩΝ 65
4.2.1 ΑΔΥΝΑΜΙΕΣ ΤΟΥ ΛΟΓΙΚΟΥ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ 66
4.3 Η ΔΙΑΦΟΡΟΠΟΙΗΣΗ ΤΟΥ ΛΟΓΙΚΟΥ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ ΜΕ ΔΙΑΧΕΙΡΙΣΗ ΠΕΡΙΟΡΙΣΜΩΝ 67
4.4 ΣΥΓΚΡΙΝΟΝΤΑΣ ΛΟΓΙΚΟ ΠΡΟΓ/ΣΜΟ ΜΕ ΤΟΝ ΛΟΓΙΚΟ ΠΡΟΓ/ΣΜΟ ΜΕ ΠΕΡΙΟΡΙΣΜΟΥΣ 70
4.5 ΑΛΓΟΡΙΘΜΟΙ ΕΛΕΓΧΟΥ ΣΥΝΕΠΕΙΑΣ 73
4.5.1 ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ ΑΛΓΟΡΙΘΜΟΥ ΣΥΝΕΠΕΙΑΣ 74
4.6 ΔΗΜΟΦΙΛΗ CLP ΣΥΣΤΗΜΑΤΑ 77

5 ΥΛΟΠΟΙΗΣΗ ΣΕ CLP(FD) 80

5.1 ΕΙΣΑΓΩΓΗ 80
5.2 ΣΧΕΤΙΚΕΣ ΕΡΓΑΣΙΕΣ 83
5.3 ΠΡΟΤΕΙΝΟΜΕΝΗ ΧΩΡΟΘΕΤΗΣΗ - ΜΟΝΤΕΛΑ ΚΡΥΣΤΑΛΛΙΚΟΥ ΠΛΕΓΜΑΤΟΣ 85
5.4 ΠΕΡΙΟΡΙΣΜΟΙ 88
5.4.1 ΧΩΡΙΚΟΙ ΠΕΡΙΟΡΙΣΜΟΙ ΠΑΝΩ ΣΤΟ ΠΚΚ ΠΛΕΓΜΑ 88
5.4.2 ΠΕΡΙΟΡΙΣΜΟΣ ΣΥΜΠΥΚΝΩΣΗΣ 89
5.4.3 ΠΕΡΙΟΡΙΣΜΟΙ ΠΟΥ ΑΦΟΡΟΥΝ ΤΙΣ ΔΕΥΤΕΡΕΥΟΥΣΕΣ ΔΟΜΕΣ ΜΙΑΣ ΠΡΩΤΕΪΝΗΣ 90
5.5 ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΗΣ ΕΝΕΡΓΕΙΑΣ ΤΗΣ ΠΡΩΤΕΪΝΗΣ 91
5.6 ΥΛΟΠΟΙΗΣΗ ΤΟΥ ΠΑΠ ΣΕ CLP(FD) 93
5.6.1 ΤΟ ΚΑΤΗΓΟΡΗΜΑ 'PERIORISMOI' 94
5.7 ΑΝΑΖΗΤΗΣΗ ΛΥΣΕΩΝ 98
5.7.1 ΚΑΤΑΧΩΡΙΣΗ ΤΙΜΩΝ ΣΤΙΣ ΜΕΤΑΒΛΗΤΕΣ ΣΤΟ ΠΚΚ ΠΛΕΓΜΑ 98
5.7.2 ΔΙΑΔΟΣΗ ΠΕΡΙΟΡΙΣΜΩΝ 98
5.7.3 ΠΡΩΤΗ ΔΟΚΙΜΗ ΤΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ 99
5.8 ΧΡΗΣΗ ΤΗΣ ΔΕΥΤΕΡΕΥΟΥΣΑΣ ΔΟΜΗΣ 100
5.9 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ 104
5.10 ΔΙΕΡΕΥΝΩΜΕΝΕΣ ΣΤΡΑΤΗΓΙΚΕΣ ΑΝΑΖΗΤΗΣΗΣ 107
5.11 ΔΙΕΡΕΥΝΩΜΕΝΗ ΣΤΡΑΤΗΓΙΚΗ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΗΣ 109
5.12 ΝΈΑ ΤΕΧΝΙΚΉ ΑΝΑΖΉΤΗΣΗΣ (BOUNDED BLOCK FAILS HEURISTIC) 109
5.12.1 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΜΕ ΚΑΙ ΧΩΡΙΣ ΤΟ BBF HEURISTIC 112
5.13 ΣΥΓΚΡΙΣΗ ΜΕ ΤΟ ΣΥΣΤΗΜΑ HMMSTR/ROSETTA 114

6 ΕΠΙΛΟΓΟΣ 115

6.1 ΣΥΝΟΨΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ 115
6.2 ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ 116

7 ΒΙΒΛΙΟΓΡΑΦΙΑ 118

8 ΠΑΡΑΡΤΗΜΑ 1 – ΚΥΡΙΑ ΚΑΤΗΓΟΡΗΜΑΤΑ 122

9 ΠΑΡΑΡΤΗΜΑ 2 - ΠΡΩΤΕΪΝΕΣ 136

Κατάλογος Εικόνων

Εικόνα 1.1 Σύμπραξη Βιολογίας - Πληροφορικής 12
Εικόνα 2.1 Οι βάσεις του DNA-RNA και οι δεσμοί που αναπτύσσονται μεταξύ τους 18
Εικόνα 2.2 Διαδικασία σύνθεσης πρωτείνης σε ένα ριβόσωμα 20
Εικόνα 2.3 Κεντρικό δόγμα Μοριακής Βιολογίας 20
Εικόνα 2.4 Χημικός τύπος αμινοξέως 23
Εικόνα 2.5 Το αμινοξύ Σερίνη 23
Εικόνα 2.6 Το αμινοξύ Αλανίνη (στα αριστερά) και Φαινυλαλανίνη (στα δεξιά) 24
Εικόνα 2.7 Η κατάταξη των αμινοξέων σε 4 κατηγορίες 27
Εικόνα 2.8 Σχηματισμός πεπτιδικού δεσμού σε μια πρωτεΐνη 28
Εικόνα 2.9 Τα 4 δομικά επίπεδα οργάνωσης μιας πρωτεΐνης 31
Εικόνα 2.10 Τα 4 δομικά επίπεδα μιας πρωτεΐνης 31
Εικόνα 2.11 Γραφική απεικόνιση α-έλικας 35
Εικόνα 2.12 Αντιπαράλληλη και Παράλληλη β-πτυχωτή επιφάνεια 37
Εικόνα 2.13 Τοπολογικό διάγραμμα αντιπαράλληλης β-πτύχωσης 39
Εικόνα 2.14 Η πρωτείνη ‘2IGD’ με χρωματισμένες τις δευτερεύουσες δομές της 40
Εικόνα 2.15 Τα 4 βασικά είδη υπερδευτεροταγών δομών 32
Εικόνα 2.16 Δομικά στοιχεία (domains) 33
Εικόνα 2.17 Τριτοταγής δομή της πρωτεΐνης ‘2IGD’ 43
Εικόνα 2.18 Λειτουργίες πρωτεϊνών 46
Εικόνα 3.1 Συσχέτιση της μεθόδου πρόβλεψης της δομής μιας πρωτείνης με την ακρίβεια των μοντέλων που προκύπτουν και τις εφαρμογές για κάθε μοντέλο 52
Εικόνα 4.1 Γράφος περιορισμών 75
Εικόνα 4.2 Πορεία υπολογισμού του προβλήματος διάταξης εργασιών με προτεραιότητες 77
Εικόνα 5.1 Δείγματα 3D κρυσταλλικών πλεγμάτων 87
Εικόνα 5.2 Ένας κύβος του ΠΚΚ πλέγματος. Οι έντονες γραμμές είναι οι ακμές. Οι διακεκομμένες γραμμές είναι οι ακμές μέσα στον κύβο. Στα δεξιά, φαίνεται το πλήες σύνολο των ακμών. 88
Εικόνα 5.3 α-έλικα 102
Εικόνα 5.4 Β-πτυχωτή επιφάνεια 104
Εικόνα 5.5 Η BBF μέθοδος αναζήτησης 110

Κατάλογος Πινάκων

Πίνακας 2.1 Είδη οργανικών μακρομορίων 16
Πίνακας 2.2 Αμινοξέα των οργανισμών (τα απαραίτητα συμβολίζονται με *) 25
Πίνακας 4.1 Prolog κώδικας για το 'SEND MORE MONEY' πρόβλημα 72
Πίνακας 4.2 CLP κώδικας για το 'SEND MORE MONEY' πρόβλημα 72
Πίνακας 4.3 Πίνακας περιορισμών για το πρόβλημα καθορισμού της σειράς εργασιών 74
Πίνακας 5.1 Ενέργειες αμινοξέων 92
Πίνακας 5.2 Πειραματικά αποτελέσματα 106
Πίνακας 5.3 Σύγκριση απόδοσης με το σύστημα στο [24] 107
Πίνακας 5.4 Πειραματικά αποτελέσματα με και χωρίς το BBF 112
Πίνακας 5.5 Συγκρίσεις με τις προβλέψεις του συστήματος Rosetta 114

1 Εισαγωγή

Στο κεφάλαιο αυτό γίνεται μια εισαγωγή στο 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 (**Εικόνα 1.1**). Η ραγδαία εξελισσόμενη επιστήμη των υπολογιστών συνεργάζεται με τη βιολογία με τελικό σκοπό την αποθήκευση, την επεξεργασία και την ερμηνεία των βιολογικών δεδομένων [1].

Τα βιολογικά δεδομένα μπορεί να αφορούν δεδομένα του γενετικού κώδικα, πειραματικά αποτελέσματα επιστημονικής βιβλιογραφίας, στατιστικά στοιχεία ασθενών, κ.α. Η Βιοπληροφορική, εκτός από την πληροφορική και τη βιολογία, χρησιμοποιεί γνώσεις και από πολλές άλλες επιστήμες όπως τη στατιστική, τη χημεία, τα θεωρητικά μαθηματικά, τη βιοχημεία και τη φυσική. Είναι ένας ταχέως αναπτυσσόμενος κλάδος με πολλαπλές εφαρμογές σε διάφορα πεδία της βιολογίας, της ιατρικής και της φαρμακολογίας [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 (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 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, αφού 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.

Συνοψίζοντας, ο σκοπός της Βιοπληροφορικής μπορεί να περιγραφεί ως τριπλός [1]:

1. Οργάνωση των δεδομένων ώστε να εξασφαλίζεται εύκολη πρόσβαση των ερευνητών σε αυτά και να επιτρέπεται η καταχώρηση νέων δεδομένων καθώς παράγονται.
2. Ανάπτυξη εργαλείων και πόρων για την ανάλυση των δεδομένων.

3. Χρήση των εργαλείων για την ανάλυση και ερμηνεία των δεδομένων με τρόπο χρήσιμο για τη Βιολογία.

Στη συνέχεια αναφέρονται τα κυριότερα 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 Σύμπραξη Βιολογίας – Πληροφορικής [45]

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 αναδίπλωσης πρωτεϊνών (fold recognition).

1.2.1 Συνεισφορά

Η βασική συνεισφορά της παρούσας εργασίας συνοψίζεται στα ακόλουθα σημεία:

- Γίνεται μία λεπτομερής ανάλυση του προβλήματος αναγνώρισης της αναδίπλωσης πρωτεϊνών καθώς και αναφορά στις υπολογιστικές προσεγγίσεις που το επιλύουν.
- Προτείνεται λύση για το πρόβλημα χρησιμοποιώντας μια εξελιγμένη τεχνολογία τεχνητής νοημοσύνης, την τεχνολογία Constraint Logic Programming σε Finite Domains - CLP(FD), δηλ. το λογικό προγραμματισμό με περιορισμούς σε διάκριτα πεδία ορισμού.
- Υλοποιείται ένα CLP(FD) σύστημα που μπορεί να χρησιμοποιηθεί αποδοτικά για την πρόβλεψη της δομής για σχετικά μεγάλες πρωτεϊνικές ακολουθίες.
- 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. Επειδή η εργασία έχει άμεση σχέση με το αντικείμενο της βιολογίας, στο Κεφάλαιο 2 θεωρήθηκε σκόπιμη μια αναφορά στις βασικές έννοιες του πεδίου. Στο κεφάλαιο 3 παρουσιάζεται το θεωρητικό υπόβαθρο πάνω στο οποίο βασίστηκε η παρούσα εργασία με μια εισαγωγή στη θεωρία της CLP(FD) τεχνολογίας που χρησιμοποιήθηκε για την επίλυση του προβλήματος αναδίπλωσης πρωτεϊνών. Στο κεφάλαιο 4 γίνεται εκτενής αναφορά στο πρόβλημα της αναγνώρισης αναδίπλωσης πρωτεϊνών, στη σημασία του και παρατίθενται οι σημαντικότερες μέθοδοι που χρησιμοποιούνται για τη λύση του.

Το κεφάλαιο 5 αναφέρεται στην ανάλυση του αλγορίθμου και στην υλοποίηση του συστήματος που προτείνεται για την επίλυση του προβλήματος αναδίπλωσης πρωτεϊνών, κάνοντας χρήση της CLP(FD) τεχνολογίας. Παρατίθενται τα πειραματικά αποτελέσματα συνοδευόμενα από τα εξαχθέντα συμπεράσματα και γίνεται σχολιασμός των αποτελεσμάτων. Στο κεφάλαιο 6 γίνεται μια σύνοψη της εργασίας, παρουσιάζονται τα συμπεράσματα που προέκυψαν και προτείνονται κάποιες μελλοντικές επεκτάσεις. Στο κεφάλαιο 7 δίνεται η βιβλιογραφία από την οποία αντλήθηκε το απαραίτητο γνωστικό υπόβαθρο για την εκτέλεση της παρούσας εργασίας.

Τέλος, στο Παράρτημα 1 παρατίθεται ο κώδικας που χρησιμοποιήθηκε για την υλοποίηση του αλγορίθμου σε περιβάλλον Sicstus Prolog, με σχολιασμό των σημαντικότερων κατηγορημάτων. Στο Παράρτημα 2 δίνονται τα κατηγορήματα που περιγράφουν τις πρωτεΐνες πάνω στις οποίες εξήχθησαν τα πειραματικά αποτελέσματα της εργασίας.

2 Πρωτεΐνες

Στο κεφάλαιο αυτό παρουσιάζονται κάποιες βασικές βιολογικές έννοιες, με έμφαση στις πρωτεΐνες, που χρησιμοποιήθηκαν στην παρούσα εργασία και είναι απαραίτητες για την κατανόηση του προβλήματος πρόβλεψης των δομών των πρωτεϊνών.

2.1 Ροή της Γενετικής Πληροφορίας

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 και την έκφρασή τους. Το πεδίο έρευνάς της έχει κοινά χαρακτηριστικά με άλλα πεδία της βιολογίας, ιδιαίτερα με τη γενετική και τη βιοχημεία.

2.1.1 Βιολογικά Μακρομόρια

Σύμφωνα με τα ευρήματα της Μοριακής Βιολογίας, η μικρότερη δομική και λειτουργική μονάδα ζωής είναι το κύτταρο. Το κύτταρο είναι μια 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 το 99% περίπου του βάρους του. Τα χημικά συστατικά του ταξινομούνται σε 1) ανόργανα: νερό (H_2O) και μεταλλικά ιόντα και 2) οργανικά: πρωτεΐνες, υδατάνθρακες, λίπη και νουκλεϊκά οξέα [4]. Στον **Πίνακας 2.1** φαίνονται τα τέσσερα κύρια είδη των βιολογικών μακρομορίων, τα δομικά τους στοιχεία και οι βασικές λειτουργίες τους.

Macromolecule	Subunit	Function	Example
PROTEINS			
Globular	Amino acids	Modified glucose	Hemoglobin
Structural	Amino acids	Catalysis; transport Support	Hair; silk
NUCLEIC ACIDS			
DNA	Nucleotides	Encodes genes	Chromosomes
RNA	Nucleotides	Needed for gene expression	Messenger RNA
LIPIDS			
Fats	Glycerol and three fatty acids	Energy storage	Butter; corn oil; soap
Phospholipids	Glycerol, two fatty acids, phosphate, and polar R groups	Cell membranes	Lecithin
Prostaglandins	Five-carbon rings with two nonpolar tails	Chemical messengers	Prostaglandin E (PGE)
Steroids	Four fused carbon rings	Membranes; hormones	Cholesterol; estrogen
Terpenes	Long carbon chains	Pigments; structural	Carotene; rubber
CARBOHYDRATES			
Starch, glycogen	Glucose	Energy storage	Potatoes
Cellulose	Glucose	Cell walls	Paper; strings of celery
Chitin		Structural support	Crab shells

Πίνακας 2.1 Είδη βιολογικών μακρομορίων [47]

Κάθε είδος μακρομορίου έχει ξεχωριστές χημικές ιδιότητες που του επιτρέπουν να φέρει σε πέρας κάποιες συγκεκριμένες λειτουργίες του οργανισμού. Η χημεία αυτών των μακρομορίων είναι η ίδια σε όλους τους οργανισμούς και αυτό έχει ως πλεονέκτημα οι οργανισμοί να λαμβάνουν τις απαραίτητες βιοχημικές ουσίες τρώγοντας άλλους οργανισμούς. Για παράδειγμα, όταν τρώμε ένα μήλο λαμβάνουμε τους πολυσακχαρίτες και τις πρωτεΐνες που αυτό περιέχει και τα οποία χρησιμοποιούνται από τον ανθρώπινο οργανισμό.

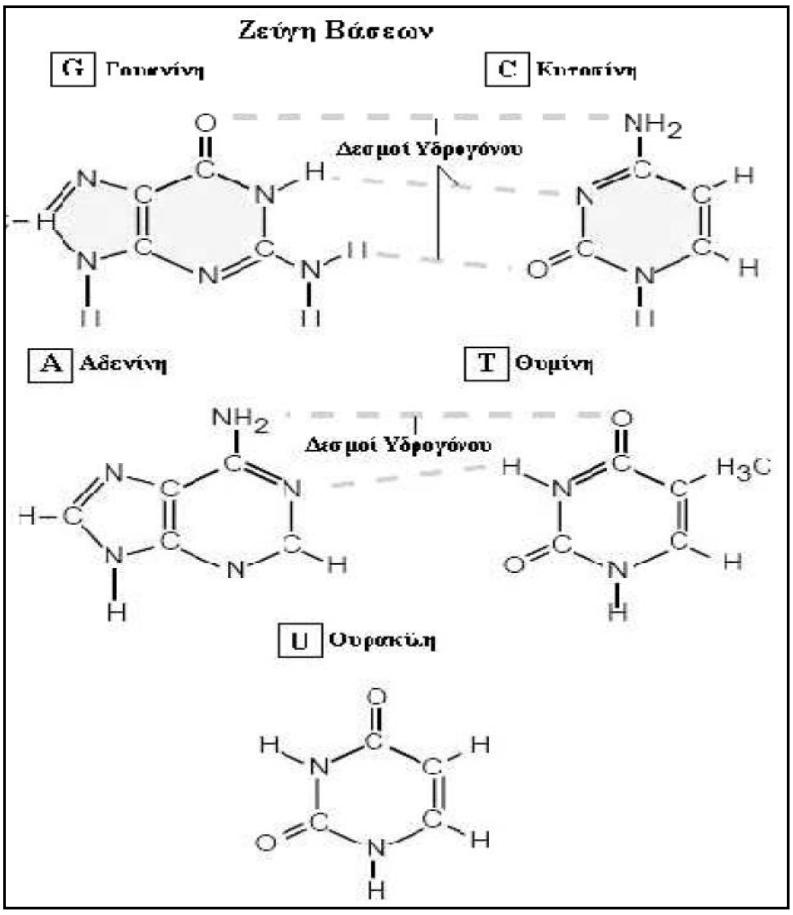
Τα κύτταρα διαχωρίζονται από το περιβάλλον τους αλλά και επικοινωνούν με αυτό, μέσω της κυτταρικής μεμβράνης τους. Στο εσωτερικό τους βρίσκεται το κυτταρόπλασμα στο οποίο περιέχονται τα οργανίδια του κυττάρου, δηλαδή διαχωρισμένες δομικές μονάδες που εκτελούν τις διάφορες λειτουργίες. Κάθε κύτταρο και 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 γενετικής πληροφορίας. Στο DNA 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 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. Ενα νουκλεοτίδιο του DNA μπορεί να έχει μια από τις ακόλουθες αζωτούχες βάσεις: Αδενίνη (A), Γουανίνη (G) Κυτοσίνη (C) ή Θυμίνη (T). Ένα νουκλεοτίδιο του RNA αντί για θυμίνη (T) έχει Ουρακίλη (U).

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 (βλ. **Εικόνα 2.1**). Το μόριο του RNA σε αντίθεση με αυτό του DNA είναι μονόκλωνο.



Εικόνα 2.1 Οι βάσεις του DNA-RNA και οι δεσμοί που αναπτύσσονται μεταξύ τους [45]

2.1.2 Κεντρικό δόγμα Μοριακής Βιολογίας

Ακολουθεί μια περιγραφή των διαδικασιών στις οποίες μετέχουν τα νουκλεϊκά οξέα, με ιδιαίτερη έμφαση στη μετάφραση, δηλ. την διαδικασία σύνθεσης των πρωτεϊνών [5].

2.1.2.1 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.

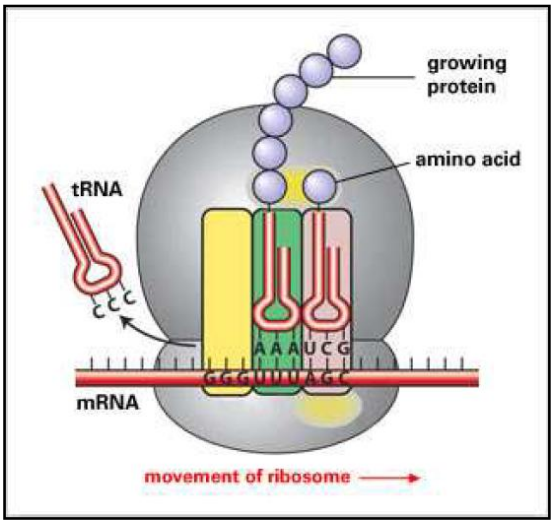
2.1.2.2 Μεταγραφή

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.

2.1.2.3 Μετάφραση

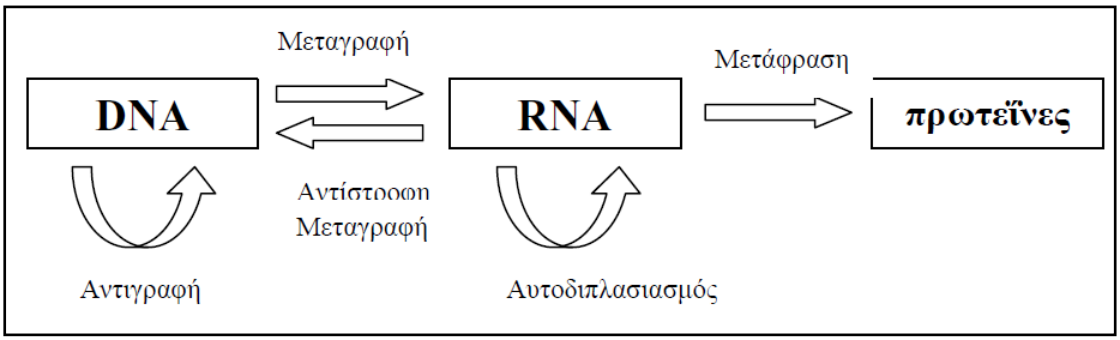
Η μετάφραση είναι η αποκωδικοποίηση των πληροφοριών που μεταφέρει το mRNA με σκοπό τη σύνθεση κάποιας συγκεκριμένης πρωτεΐνης. Πραγματοποιείται στα ριβοσώματα και απαιτεί τη συμμετοχή αρκετών πρωτεϊνών και ενέργειας. Κάθε τριάδα βάσεων (κωδικόνιο) στο mRNA αντιστοιχεί και σε ένα συγκεκριμένο αμινοξύ. 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-μεταφραστικές τροποποιήσεις. Οι τροποποιήσεις αυτές καθορίζουν τις φυσικές και χημικές ιδιότητες, την αναδίπλωση, τη δραστικότητα καθώς και τη λειτουργικότητα των πρωτεϊνών.



Εικόνα 2.2 Διαδικασία σύνθεσης πρωτεΐνης σε ένα ριβόσωμα [45]

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 [45]

2.2 Δομή πρωτεϊνών

Το βασικότερο συστατικό της λειτουργίας των οργανισμών είναι οι πρωτεΐνες καθώς είναι υπεύθυνες για την πλειοψηφία των ουσιαστικότερων λειτουργιών των κυττάρων ενός οργανισμού. Το όνομα «πρωτεΐνη» διατυπώθηκε από τον Berzelius και χρησιμοποιήθηκε για πρώτη φορά στο εγχειρίδιο του Mulder (1840). Η λέξη προέρχεται από την ελληνική λέξη «πρωτεύω». Οι πρωτεΐνες φέρουν το όνομα τους δίκαια, μιας και πολυάριθμες λειτουργίες του οργανισμού οφείλονται σε ειδικές πρωτεΐνες. Είναι άγνωστη και αδιανόητη η ύπαρξη ζωής χωρίς πρωτεΐνες.

Οι πρωτεΐνες είναι συστατικά κάθε κυττάρου, σχηματίζουν τα συσταλτά στοιχεία καθώς και τα ένζυμα που ελευθερώνουν την απαραίτητη ενέργεια για τη διατήρηση της ζωής, ενώ στο αίμα παίζουν το ρόλο μεταφορικών μορίων. Οι πρωτεΐνες είναι ένα από τα τέσσερα είδη βιολογικών μακρομορίων και αποτελούνται από μικρότερα δομικά στοιχεία, τα οποία ονομάζονται αμινοξέα [1]. Η ακολουθία των αμινοξέων της κάθε πρωτεΐνης καθορίζει τη δομή και τις λειτουργίες της και είναι αποτέλεσμα της κωδικοποίησης της ακολουθίας νουκλεοτιδίων (γονίδια) του κυττάρου. Συνεπώς, όλη η πληροφορία που είναι απαραίτητη για τη δημιουργία των πρωτεϊνών ενός κυττάρου βρίσκεται στα γονίδιά του.

Οι πρωτεΐνες παρουσιάζουν μεγάλη ποικιλία ιδιοτήτων και σχήματος, π.χ., οι πρωτεΐνες των μαλλιών, του δέρματος, των τενόντων και των οστών είναι λεπτές και μακριές. Έτσι, ένα μόριο κολλαγόνου (πρωτεΐνης των τενόντων και των οστών) είναι περίπου κυλινδρικό, 2990 Å σε μήκος με διάμετρο μόνο 14 Å. Περίπου 300.000 τέτοια μόρια ενώνονται στις άκρες τους για να κατασκευάσουν ένα τένοντα 10 cm μακρύ και περίπου 1012 τέτοιες στήλες είναι αναγκαίες για τη δημιουργία ενός τένοντα διαμέτρου 5 mm. Παρόμοιες είναι όλες οι ινώδεις πρωτεΐνες, τα κύρια δομικά υλικά των οργανισμών.

Αντίθετα, τα πρωτεϊνικά μόρια που παίζουν ρόλο στις χημικές μεταβολές που πραγματοποιούνται στους ζωντανούς οργανισμούς είναι πιο συμπαγή σε σχήμα, παρόλο που έχουν μεγάλη ποικιλία μεγεθών. Παραδείγματα τέτοιων

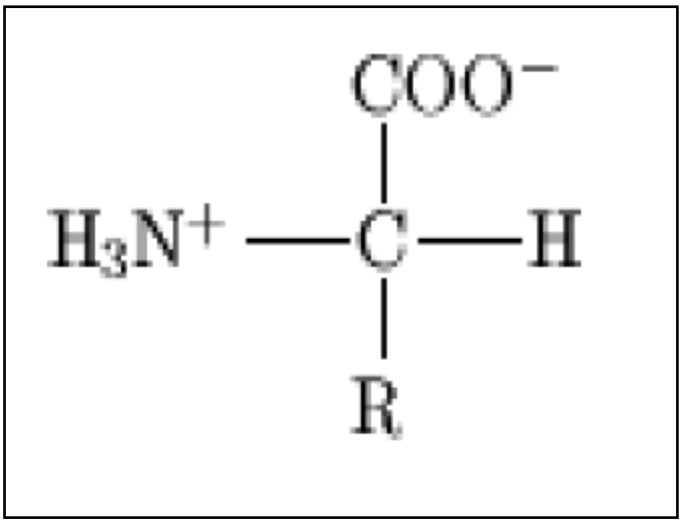
σφαιρικών πρωτεϊνών είναι η αιμοσφαιρίνη, ο μεταφορέας του οξυγόνου στο αίμα, τα αντισώματα, οι πρωτεΐνες που ενώνονται με ξένες ουσίες και βοηθούν στην αποβολή τους και, κυρίως, τα ένζυμα, οι οργανικοί καταλύτες που ελέγχουν τις χημικές αντιδράσεις που πραγματοποιούνται στους ζωντανούς οργανισμούς.

Πολλές σφαιρικές πρωτεΐνες μπορούν να απομονωθούν από κύτταρα, να καθαριστούν και να κρυσταλλωθούν. Οι δομές τους είναι δυνατόν να προσδιοριστούν με μεθόδους κρυσταλλογραφίας ακτίνων-Χ, όπως γίνεται και για μικρότερα βιολογικά μόρια. Η μοναδική διαφορά είναι η μεγαλύτερη πολυπλοκότητά τους που οφείλεται στο γεγονός ότι ένα πρωτεϊνικό μόριο αποτελείται από έναν πολύ μεγάλο αριθμό ατόμων.

Στην ενότητα αυτή γίνεται αρχικά μια περιγραφή των δομικών στοιχείων των πρωτεϊνών που είναι τα αμινοξέα. Στη συνέχεια, παρουσιάζονται οι δομές των πρωτεϊνών με έμφαση στη δευτερεύουσα δομή μιας πρωτεΐνης (α-έλικα και β-πτυχωτή επιφάνεια) και στην τριτογενή (τρισδιάστατη) δομή μιας πρωτεΐνης. Τέλος αναφέρονται οι βιολογικές λειτουργίες των πρωτεϊνών σε έναν οργανισμό και αναλύεται η αλληλεπίδραση των πρωτεϊνών με άλλα μόρια ενός οργανισμού.

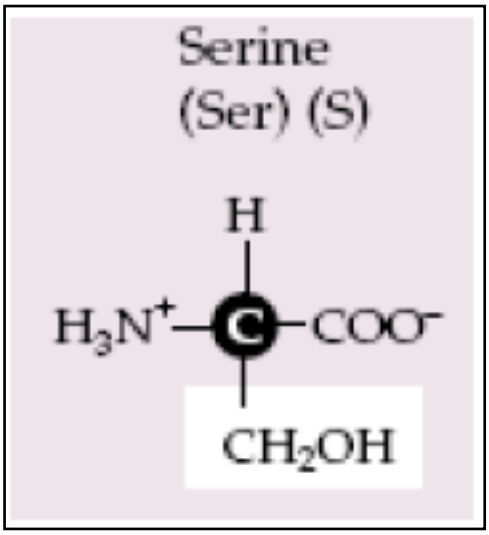
2.2.1 Αμινοξέα: τα δομικά στοιχεία των πρωτεϊνών

Οι πρωτεΐνες αποτελούνται από μικρότερα δομικά στοιχεία, τα οποία ονομάζονται αμινοξέα Τα αμινοξέα παρουσιάζουν όλα την ίδια γενική δομή. Ένα κεντρικό άτομο άνθρακα (που ονομάζεται κατά συνθήκη α-άτομο άνθρακα, Cα) συνδέεται με 4 διαφορετικά άτομα ή ομάδες ατόμων: ένα άτομο υδρογόνου (-H), μία αμινομάδα (-NH2), μία καρβοξυλική ομάδα (-COOH) και μία από 20 δυνατές διαφορετικές πλευρικές αλυσίδες διαφόρων χημικών τύπων (-R) [2], όπως φαίνεται στην **Εικόνα 2.4**. Τα 20 διαφορετικά αμινοξέα διακρίνονται ανάλογα με τις R-ομάδες ή πλευρικές αλυσίδες τους. Οι πλευρικές αλυσίδες των αμινοξέων είναι σημαντικές καθώς καθορίζουν την τρισδιάστατη δομή και τις λειτουργίες μιας πρωτεΐνης.



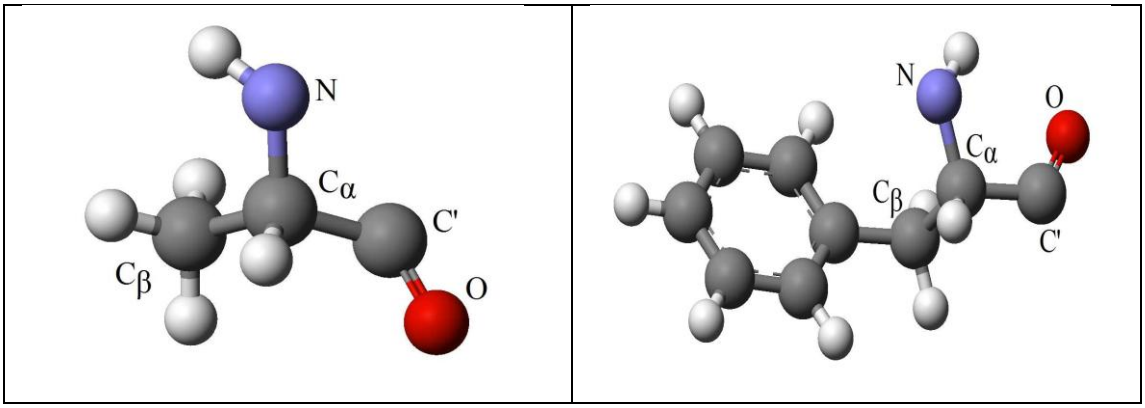
Εικόνα 2.4 Χημικός τύπος αμινοξέως [45]

Στην **Εικόνα 2.5** φαίνεται το αμινοξύ ‘σερίνη’ και σε άσπρο πλέγμα ξεχωρίζει η πλευρική αλυσίδα αυτού του αμινοξέως.



Εικόνα 2.5 Το αμινοξύ Σερίνη [47]

Στην **Εικόνα 2.6** φαίνεται η τρισδιάσταση δομή των αμινοξέων Αλανίνη και Φαινυλαλανίνη, αντίστοιχα.



Εικόνα 2.6 Το αμινοξύ Αλανίνη (στα αριστερά) και Φαινυλαλανίνη (στα δεξιά) [49]

Στις πρωτεΐνες των ζωντανών οργανισμών απαντώνται 20 διαφορετικά αμινοξέα. Καθένας από τους δυνατούς συνδυασμούς σε μια τριάδα βάσεων του DNA αντιστοιχίζεται σε ένα από τα 20 αμινοξέα. Άρα, κάθε αμινοξύ κωδικοποιείται από μια τουλάχιστον τριάδα γενετικού κώδικα DNA. Στον **Πίνακας 2.2** φαίνονται τα 20 αμινοξέα, όπως κωδικοποιούνται στην βιβλιογραφία, χρησιμοποιώντας τον κανόνα του ενός ή των τριών γραμμάτων.

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.

Η σύνδεση των μορίων δύο αμινοξέων, με τον χαρακτηριστικό τρόπο που απαντάται στις πρωτεΐνες, περιλαμβάνει μία απλή αντίδραση συμπύκνωσης, στην οποία αποβάλλεται ένα μόριο νερού και σχηματίζεται ένας πεπτιδικός δεσμός. Το μήκος των πρωτεϊνών κυμαίνεται από μικρές αλυσίδες (ribonuclease A) που αποτελούνται από 51 αμινοξέα έως και μεγάλες αλυσίδες (apolipoprotein B) που αποτελούνται από 4.636 αμινοξέα [2]. Αν και αποτελούνται μόνο από είκοσι αμινοξέα, ο θεωρητικός αριθμός των διαφορετικών πρωτεϊνών είναι τεράστιος αν αναλογιστούμε ότι για μία

ακολουθία 100 αμινοξέων (σχετικά μικρή ακολουθία) υπάρχουν 20^{100} διαφορετικοί συνδυασμοί που οδηγούν σε διαφορετικές πρωτεΐνες.

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 2.2 Αμινοξέα των οργανισμών (τα απαραίτητα συμβολίζονται με *) [46]

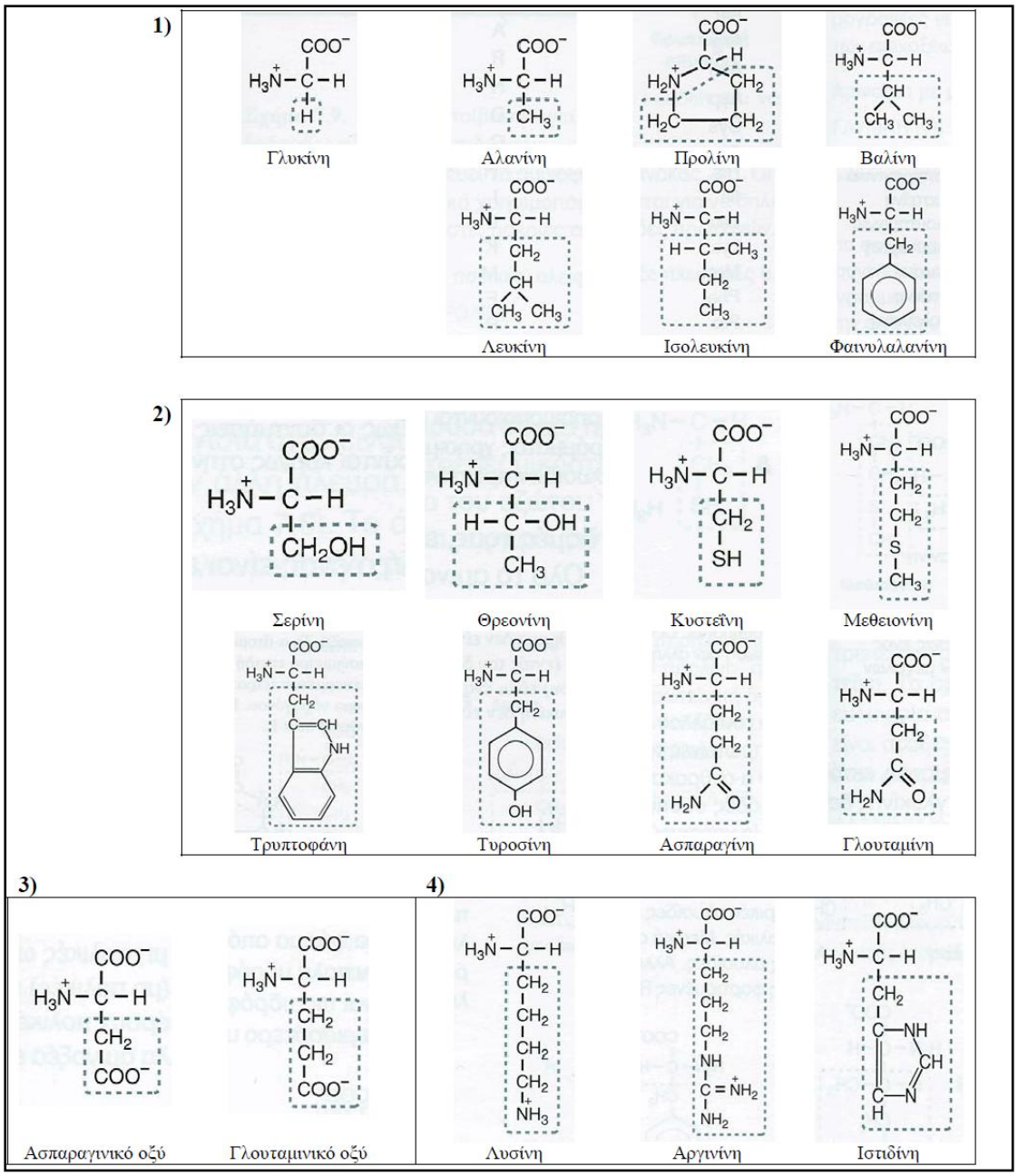
Τα αμινοξέα χωρίζονται σε τέσσερις ομάδες, που παρουσιάζονται στην **Εικόνα 2.7**, σύμφωνα με τις ιδιότητες των πλευρικών αλυσίδων [2]. Αυτές οι ιδιότητες καθορίζουν σε μεγάλο βαθμό την τρισδιάστατη δομή της πρωτεΐνης καθώς οι δεσμοί που είναι υπεύθυνοι για τη δημιουργία και σταθεροποίησή της είναι οι δεσμοί μεταξύ των πλευρικών αλυσίδων των αμινοξέων.

Ορισμένες πλευρικές αλυσίδες είναι ηλεκτρικά πολωμένες (+1,-1), άλλες είναι πολικές (υδροφιλικές) και άλλες μη πολικές (υδροφοβικές).

- *1η ομάδα αμινοξέων:* Τα πέντε αμινοξέα (Arg, His, Lys, Asp, Glu) έχουν ηλεκτρικά πολωμένες πλευρικές αλυσίδες που προσελκύουν το νερό (υδροφιλικά) και αντίθετα πολωμένα ιόντα όλων των ειδών.
- *2η ομάδα αμινοξέων:* Τα πέντε αμινοξέα (Ser, Thr, Asn, Gln, Tyr) έχουν πολικές πλευρικές αλυσίδες που δημιουργούν ασθενής δεσμούς με το νερό και άλλα πολικά συστατικά. Και αυτά τα αμινοξέα είναι υδροφιλικά.
- *3η ομάδα αμινοξέων:* Τα εφτά αμινοξέα (Ala, Ile, Leu, Met, Phe, Trp, Val) έχουν μη πολικές πλευρικές αλυσίδες και στο υγρό περιβάλλον του εσωτερικού του κυττάρου έχουν την τάση να συγκροτούνται στο εσωτερικό της πρωτεΐνης. Αυτά τα αμινοξέα είναι υδροφοβικά.
- *4η ομάδα αμινοξέων:* Τέλος, τα τρία αμινοξέα (Cys, Gly, Pro) είναι ειδικές περιπτώσεις, αν και οι πλευρικές τους αλυσίδες είναι κυρίως υδροφοβικές.

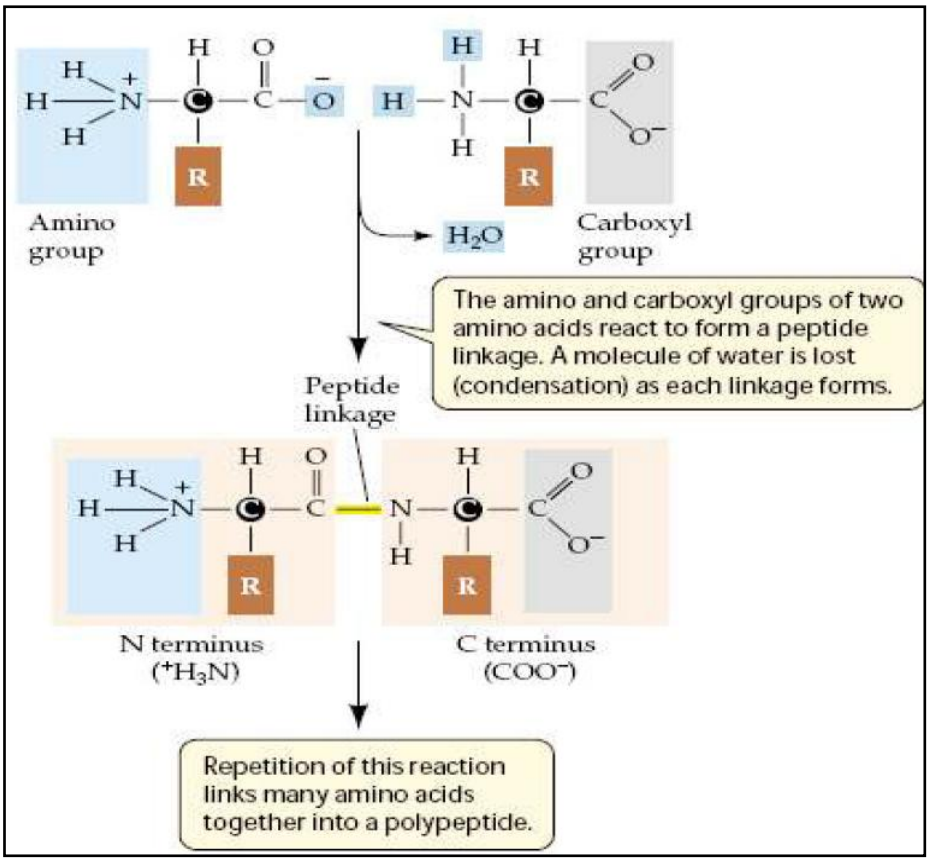
Μια άλλη κατηγοριοποίηση των αμινοξέων ανάλογα με τις ιδιότητες των πλευρικών αλυσίδων τους είναι η παρακάτω:

- 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.



Εικόνα 2.7 Η κατάταξη των αμινοξέων σε 4 κατηγορίες [46]

Τα αμινοξέα ενώνονται μεταξύ τους με πεπτιδικό δεσμό και για αυτό το λόγο η ακολουθία αμινοξέων ονομάζεται και πολυπεπτιδική αλυσίδα [2]. Σε αυτόν το δεσμό αντιδρούν μεταξύ τους η καρβοξυλική ομάδα του ενός αμινοξέος με την αμινομάδα του άλλου και ταυτόχρονα ελευθερώνεται και ένα μόριο νερού. Η επαναληπτική αλληλουχία ατόμων κατά μήκος της αλυσίδας αποκαλείται πολυπεπτιδικός σκελετός. Στην **Εικόνα 2.8** φαίνεται μία απλοποιημένη περιγραφή αυτής της αντίδρασης.



Εικόνα 2.8 Σχηματισμός πεπτιδικού δεσμού σε μια πρωτεΐνη [47]

Κάθε πρωτεΐνη είναι μια ακολουθία αμινοξέων η οποία έχει μία γραμμική διάταξη, δηλαδή έχει μια αρχή και ένα τέλος. Το σημείο που ορίζει την αρχή της ακολουθίας είναι η ελεύθερη αμινομάδα του πρώτου αμινοξέος και ονομάζεται 'N terminus' ενώ το σημείο που ορίζει το τέλος της ακολουθίας είναι η ελεύθερη καρβοξυλική ομάδα του τελευταίου αμινοξέος και ονομάζεται 'C terminus'. Όλες οι άλλες αμινομάδες και καρβοξυλικές ομάδες (εκτός αυτών που υπάρχουν στις πλευρικές αλυσίδες) εμπλέκονται στους πεπτιδικούς δεσμούς της αλυσίδας και δεν θεωρούνται 'ελεύθερες', ανέπαφες ομάδες. Αυτή η διάταξη της ακολουθίας των αμινοξέων αναφέρεται και ως 'N-C' προσανατολισμός των πρωτεϊνών [6].

Μία πολυπεπτιδική αλυσίδα αποτελείται από ένα κανονικά επαναλαμβανόμενο μέρος, την κύρια αλυσίδα (main chain) και ένα μεταβλητό μέρος, τις χαρακτηριστικές πλευρικές αλυσίδες ή ομάδες (side chains) των αμινοξικών καταλοίπων. Η κύρια αλυσίδα ονομάζεται πολλές φορές σπονδυλική στήλη (backbone).

Πολλές πρωτεΐνες, όπως η μυοσφαιρίνη, αποτελούνται από μία πολυπεπτιδική αλυσίδα. Άλλες αποτελούνται από δύο ή περισσότερες αλυσίδες, οι οποίες μπορεί να είναι ίδιες ή διαφορετικές. Για παράδειγμα, η αιμοσφαιρίνη έχει δύο αλυσίδες ενός είδους και δύο ενός άλλου. Οι αλυσίδες συγκρατούνται μεταξύ τους με ομοιοπολικούς δεσμούς. Εναλλακτικά, οι πολυπεπτιδικές αλυσίδες μερικών πρωτεϊνών που αποτελούνται από πολλές αλυσίδες συνδέονται με δισουλφιδικούς δεσμούς.

2.2.2 Επίπεδα δομής των πρωτεϊνών

Το 1959, οι Linderstrom-Lang και Schelman πρότειναν από τις μελέτες τους ότι στις πρωτεΐνες μπορούν να αναγνωριστούν τέσσερα επίπεδα οργάνωσης: η πρωτοταγής, η δευτεροταγής, η τριτοταγής και η τεταρτοταγής δομή [1, 4]. Παρακάτω δίνεται μία περιγραφή αυτών των τεσσάρων δομών καθώς παίζουν καθοριστικό ρόλο στις λειτουργίες των πρωτεϊνών.

1) **Πρωτοταγής δομή**: Η ακριβής ακολουθία των αμινοξέων μιας πρωτεΐνης αποτελεί την πρωτοταγή δομή της. Οι συντομογραφίες (ένα γράμμα) των αμινοξέων χρησιμοποιούνται για την καταγραφή των ακολουθιών τους. Στα υψηλότερα επίπεδα των πρωτεϊνικών δομών, οι τοπικές αναδιπλώσεις δίνουν στο μακρομόριο το τελικό λειτουργικό του σχήμα, αλλά όλα αυτά τα επίπεδα προέρχονται από την πρωτοταγή δομή των πρωτεϊνών. Δηλαδή, οι ιδιότητες της ακριβής ακολουθίας των αμινοξέων καθορίζουν τον τρόπο που ένα μακρομόριο τυλίγεται και αναδιπλώνεται, ώστε να σταθεροποιήσει τη δομή του.

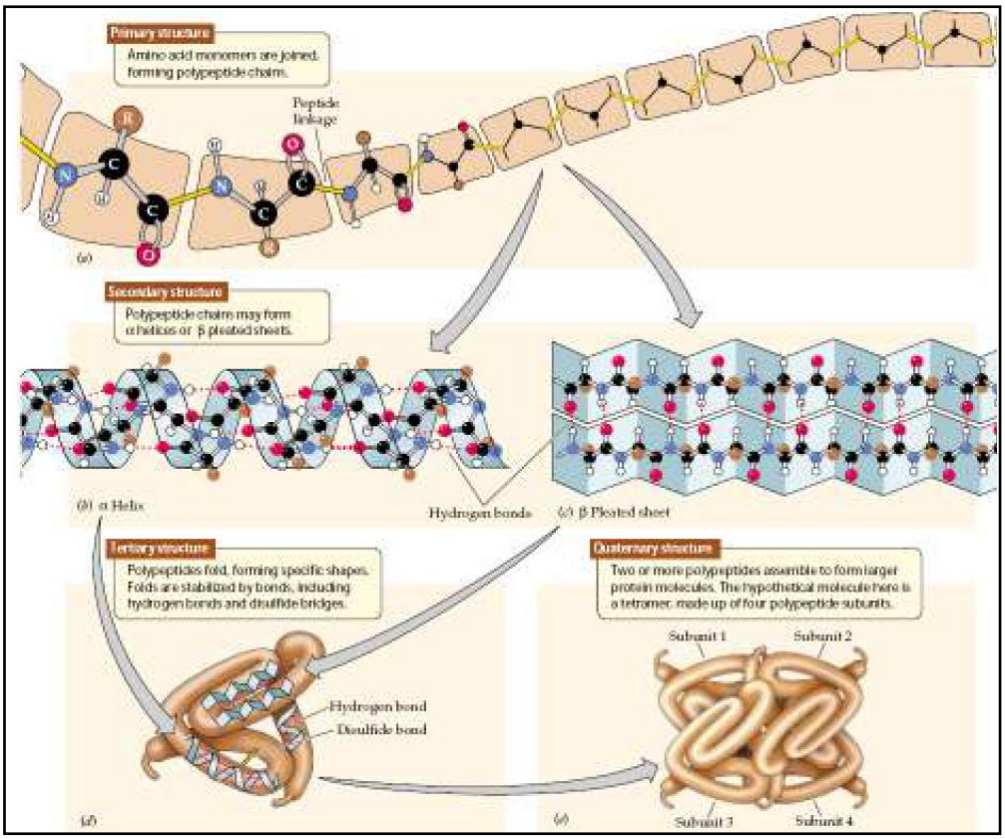
2) **Δευτεροταγής δομή**: Η δευτεροταγής δομή μιας πρωτεΐνης αποτελείται από συμμετρικά, επαναλαμβανόμενα πρότυπα σε διαφορετικές περιοχές της αλυσίδας της. Υπάρχουν δύο βασικοί τύποι δευτεροταγούς δομής οι οποίοι καθορίζονται από τους δεσμούς υδρογόνου μεταξύ των αμινοξέων και οι οποίοι είναι η α-έλικα και η β-πτυχωτή επιφάνεια.

3) **Τριτοταγής δομή**: Η πολυπεπτιδική αλυσίδα κάμπτεται και αναδιπλώνεται στο χώρο έχοντας σαν αποτέλεσμα την διαμόρφωση της τριτοταγούς δομής της. Μια πλήρης περιγραφή της τριτοταγούς δομής της πρωτεΐνης διευκρινίζει τη θέση κάθε ατόμου του μακρομορίου σε σχέση με τα άλλα στις τρεις διαστάσεις του χώρου. Οι α-έλικες και β-πτυχωτές έλικες συμβάλλουν στη διαμόρφωση της τριτοταγούς δομής και εντοπίζονται σε ορισμένες περιοχές του μακρομορίου. Παράλληλα, οι περισσότερες περιοχές των μακρομορίων αποτελούνται από μορφές δευτεροταγούς δομής που είναι μοναδικές για κάθε πρωτεΐνη. Οι δεσμοί που είναι υπεύθυνοι για την δημιουργία και σταθεροποίηση της τριτοταγούς δομής είναι οι δεσμοί μεταξύ των πλευρικών αλυσίδων των αμινοξέων [2].

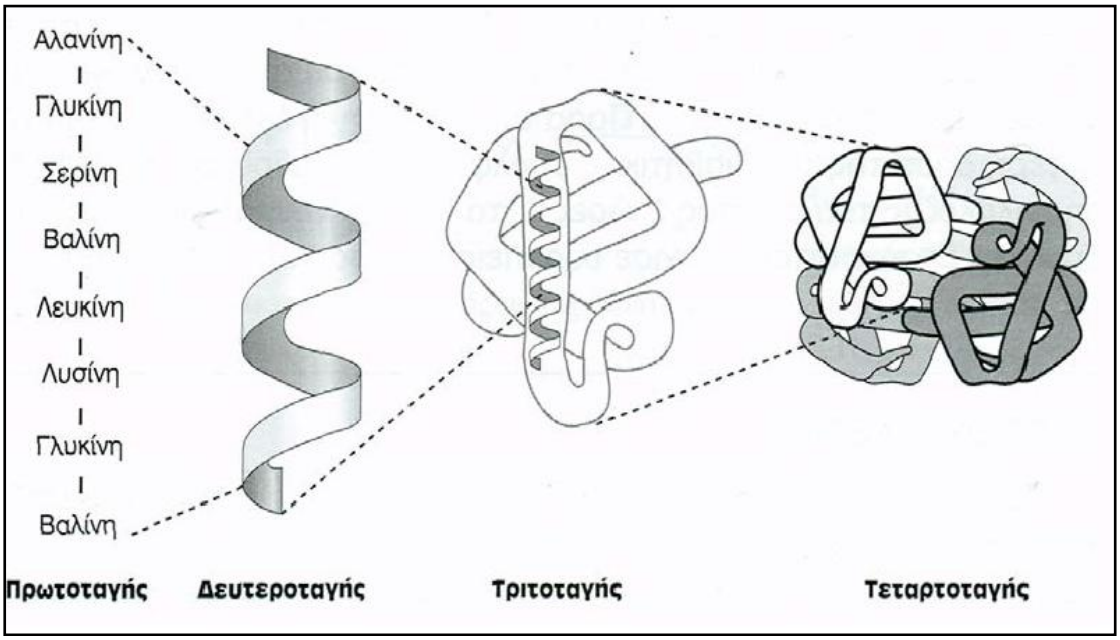
4) **Τεταρτοταγής δομή**: Πολλές λειτουργικές πρωτεΐνες αποτελούνται από δύο ή περισσότερες πολυπεπτιδικές αλυσίδες, οι οποίες ονομάζονται υπομονάδες. Η τεταρτοταγής δομή της πρωτεΐνης είναι αποτέλεσμα του τρόπου σύνδεσης και αλληλεπίδρασης των υπομονάδων της [2].

Αφού, όπως αποδείχθηκε αρκετά πειστικά με πειράματα αναδιάταξης, η αμινοξική ακολουθία περιέχει κωδικοποιημένη όλη την αναγκαία πληροφορία ώστε μία πρωτεΐνη να διπλωθεί στη φυσική τριδιάστατη δομή της, έπεται ότι υπάρχει μία σχέση αλληλεξάρτησης μεταξύ των 4 επιπέδων οργάνωσης, τέτοια ώστε στοιχεία ενός κατώτερου επιπέδου οργάνωσης να καθορίζουν στοιχεία των ανώτερων επιπέδων.

Στην **Εικόνα 2.9** και στην **Εικόνα 2.10** φαίνονται τα τέσσερα δομικά επίπεδα οργάνωσης των πρωτεϊνών.



Εικόνα 2.9 Τα 4 δομικά επίπεδα οργάνωσης μιας πρωτεΐνης [47]



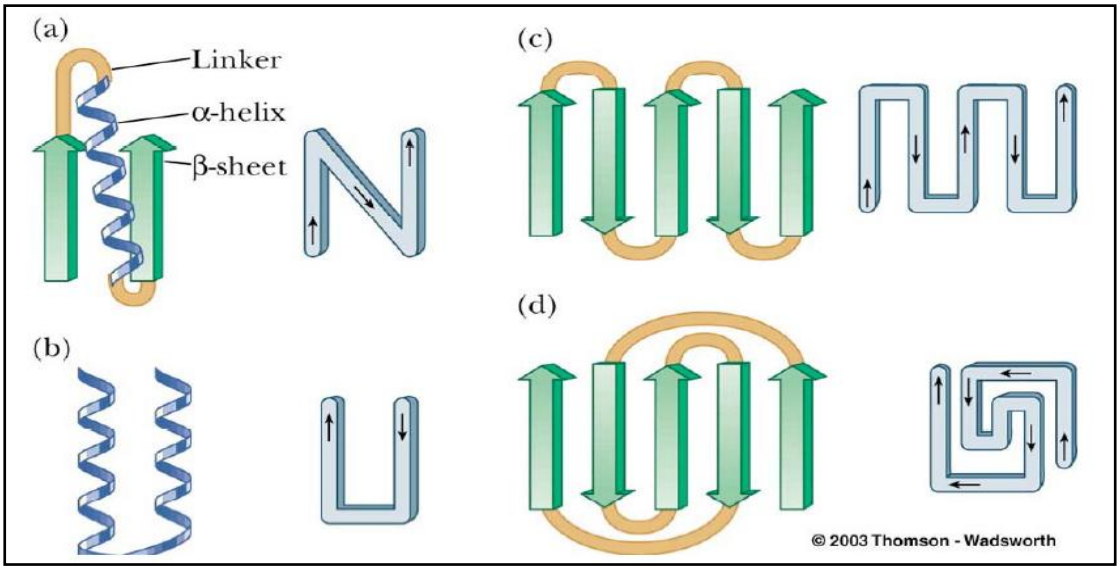
Εικόνα 2.10 Τα 4 δομικά επίπεδα μιας πρωτεΐνης [46]

Πρόσφατες πληροφορίες που πάρθηκαν μετά από μελέτες της δομής πολλών πρωτεϊνών με κρυσταλλογραφία ακτίνων Χ επιτρέπουν την προσθήκη στο σχήμα των 4 επιπέδων δύο επιπλέον επιπέδων οργάνωσης, της

υπερδευτεροταγούς δομής (supersecondary structure) και του **δομικού στοιχείου ή μονάδας ή αυτοτελούς περιοχής** (domain).

Οι υπερδευτεροταγείς δομές αποτελούν κανονικούς συνδυασμούς δευτεροταγών δομών [5] και η αρκετά συχνή εμφάνισή τους δείχνει ότι ευνοούνται για κινητικούς λόγους κατά τη διαδικασία της αναδίπλωσης ή/και για ενεργειακούς λόγους. Έχουν αναφερθεί τέσσερα βασικά είδη υπερδευτεροταγών δομών:

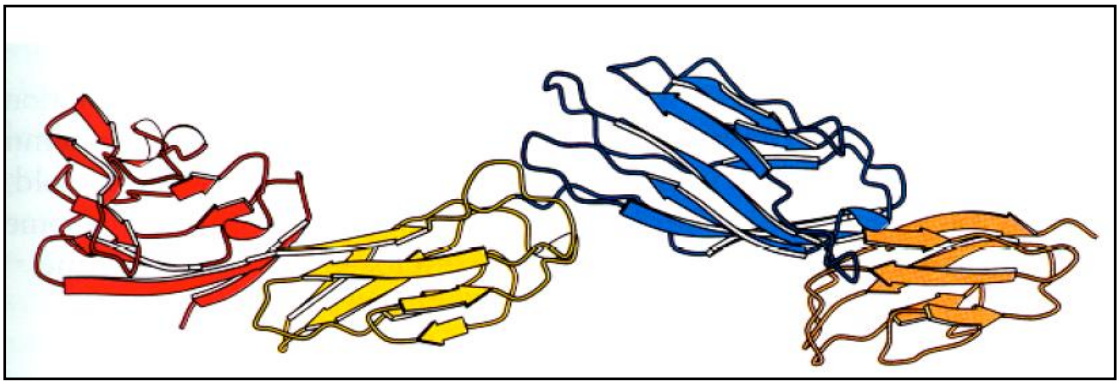
1. βαβ - δύο παράλληλοι κλώνοι β-πτυχωτών επιφανειών, ανάμεσα στους οποίους παρεμβάλλεται μία α-έλικα (δες **Εικόνα 2.11., (a))**,
2. αα - δύο αντιπαράλληλες α-έλικες (δες **Εικόνα 2.11., (b))**,
3. β μαίανδρος - μία αντιπαράλληλη β-πτυχωτή επιφάνεια, η οποία σχηματίζεται από αντιπαράλληλους β-κλώνους, ανάμεσα στους οποίους παρεμβάλλονται β-στροφές (δες **Εικόνα 2.11., (c))**,
4. Ελληνικό Κλειδί - μία επαναληπτική υπερδευτεροταγής δομή, η οποία σχηματίζεται όταν μία αντιπαράλληλη πτυχωτή επιφάνεια αναδιπλώνεται με τον εαυτό της (δες **Εικόνα 2.11., (d))**.



Εικόνα 2.11 Τα 4 βασικά είδη υπερδευτεροταγών δομών [46]

Τα δομικά στοιχεία ή μονάδες ή αυτοτελείς περιοχές (**Εικόνα 2.**) σχηματίζονται όταν πολύ μεγάλες πολυπεπτιδικές αλυσίδες τείνουν να διπλώνονται σε δύο ή περισσότερα διακριτά σφαιρικά τμήματα, τα οποία

συνδέονται σχετικά χαλαρά μεταξύ τους. Μικρά τμήματα συνδέουν συνήθως τα γειτονικά δομικά στοιχεία, τα οποία μπορεί να είναι ανεξάρτητα μεταξύ τους από δομικής και λειτουργικής απόψεως. Έχει επίσης αναφερθεί ότι δομικά στοιχεία με παρόμοιες διαμορφώσεις σχετίζονται συνήθως με μία συγκεκριμένη λειτουργία.



Εικόνα 2.12 Δομικά στοιχεία (domains) [46]

2.3 Στοιχεία δευτεροταγούς δομής πρωτεϊνών

Κατά τη δημιουργία μιας τριτοταγούς δομής μιας πρωτεΐνης, 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**

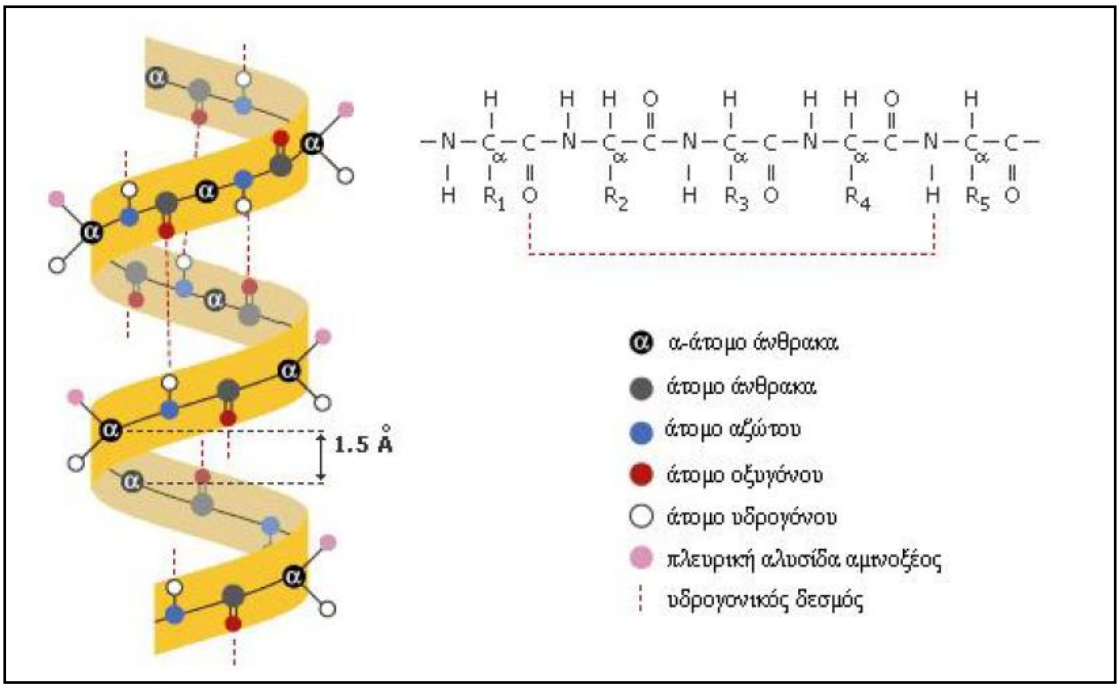
Οι δύο αυτές δομές είναι το ίδιο σημαντικές και συνηθισμένες στις πρωτεΐνες. Για παράδειγμα στις σφαιρικές υδροδιαλυτές πρωτεΐνες συναντάμε τις α-έλικες με συχνότητα περίπου 30% ενώ τις β-πτυχωτές επιφάνειες με συχνότητα περίπου 25%. Οι β-πτυχωτές επιφάνειες είναι πιο ισχυρές και σταθερές δομές από τις α-έλικες.

2.3.1 α-έλικα

Η α-έλικα είναι μία δεξιόστροφη έλικα η οποία είναι αποτέλεσμα των δεσμών υδρογόνου που δημιουργούνται μεταξύ του υδρογόνου της ένωσης N-H ενός αμινοξέος και του οξυγόνου της ένωσης C=O ενός άλλου αμινοξέος. Οι πλευρικές αλυσίδες των αμινοξέων εκτείνονται εξωτερικά της σπονδυλικής στήλης της έλικας. Ένας παράγοντας που καθορίζει την ύπαρξη α-έλικας σε μια πρωτεΐνη είναι η παρουσία μεγάλων πλευρικών αλυσίδων, οι οποίες εμποδίζουν τη δημιουργία δεσμών υδρογόνου μεταξύ των αμινοξέων και επομένως και τη διαμόρφωση α-έλικας [2].

Οι α-έλικες βρίσκονται συνήθως στην εξωτερική πλευρά των πρωτεϊνών σε αντίθεση με τις β-πτυχωτές επιφάνειες που βρίσκονται βαθιά στο εσωτερικό τους. Οι α-έλικες είναι μία ιδιαίτερα απαντώμενη δευτεροταγής δομή (ποσοστό εμφάνισης σε σφαιρικές υδατοδιαλυτές πρωτεΐνες ~ 30%). Σχηματίζονται όταν σε μία πολυπεπτιδική αλυσίδα οι γωνίες γύρω από τους δεσμούς N-Ca και Ca-C' (φ και ψ) παίρνουν τιμές αντίστοιχα -57° και -47° περίπου, για μία σειρά διαδοχικών, τουλάχιστον 5, καταλοίπων.

Η α-έλικα είναι δεξιόστροφη. Σταθεροποιείται από την ύπαρξη υδρογονικών δεσμών, σχεδόν παράλληλων προς τον άξονά της. Οι δεσμοί αυτοί σχηματίζονται ανάμεσα στο υδρογόνο της αμινομάδας (-NH) του αμινοξέως i και στο οξυγόνο της καρβοξυλομάδας του αμινοξέως i-4. Σαν έλικα, έχει ένα χαρακτηριστικό βήμα που είναι 5,4 Å περίπου και έχει 3,6 αμινοξέα ανά στροφή. Επομένως, η απόσταση δύο διαδοχικών αμινοξέων παράλληλα με τον άξονα της έλικας είναι περίπου 1,5 Å (**Εικόνα 2.12**). Η ακτίνα της α-έλικας είναι 2,3 Å.



Εικόνα 2.12 Γραφική απεικόνιση α-έλικας [46]

Επειδή, πολλές φορές, οι α-έλικες βρίσκονται στην επιφάνεια των πρωτεϊνών, υπάρχει η τάση οι πλευρικές αλυσίδες των καταλοίπων που συμμετέχουν στον σχηματισμό τους να μεταβάλλονται από υδρόφοβες σε υδρόφιλες, με περιοδικότητα 3 ή 4 κατάλοιπα (πράγμα που φαίνεται στην αμινοξική ακολουθία). Σε αυτές τις περιπτώσεις, οι α-έλικες παρουσιάζουν μία υδρόφοβη και μία υδρόφιλη επιφάνεια.

Δεν συμμετέχουν όλα τα αμινοξικά κατάλοιπα με την ίδια επιτυχία στο σχηματισμό α-έλικας. Έτσι, η Αλανίνη, η Γλουταμίνη, η Λευκίνη και η Μεθειονίνη ευνοούν τη δημιουργία ελίκων ενώ η Προλίνη, η Γλυκίνη και η Σερίνη όχι. Οι α-έλικες, εκτός από τη μεγάλη συχνότητα εμφάνισής τους σε σφαιρικές υδατοδιαλυτές πρωτεΐνες, απαντώνται και σε δομικές (ινώδεις) πρωτεΐνες, όπως στις α-κερατίνες, στην ακτίνη, τη μυοσίνη, την τροπομυοσίνη κλπ. καθώς και σε πρωτεΐνες μεμβρανών όπως π.χ. στη ροδοψίνη και βακτηριοροδοψίνη, σε ριβοσωμικές πρωτεΐνες, πρωτεΐνες ιών κλπ.

Η μεγάλη συχνότητα εμφάνισης της δεξιόστροφης α-έλικας οφείλεται στην πολύ σταθερή ενεργειακή κατάσταση που διαθέτει (κατάσταση ελάχιστης

ενέργειας). Υπάρχουν και αριστερόστροφες α-έλικες αλλά δεν απαντώνται σε πρωτεΐνες επειδή είναι πολύ ασταθείς.

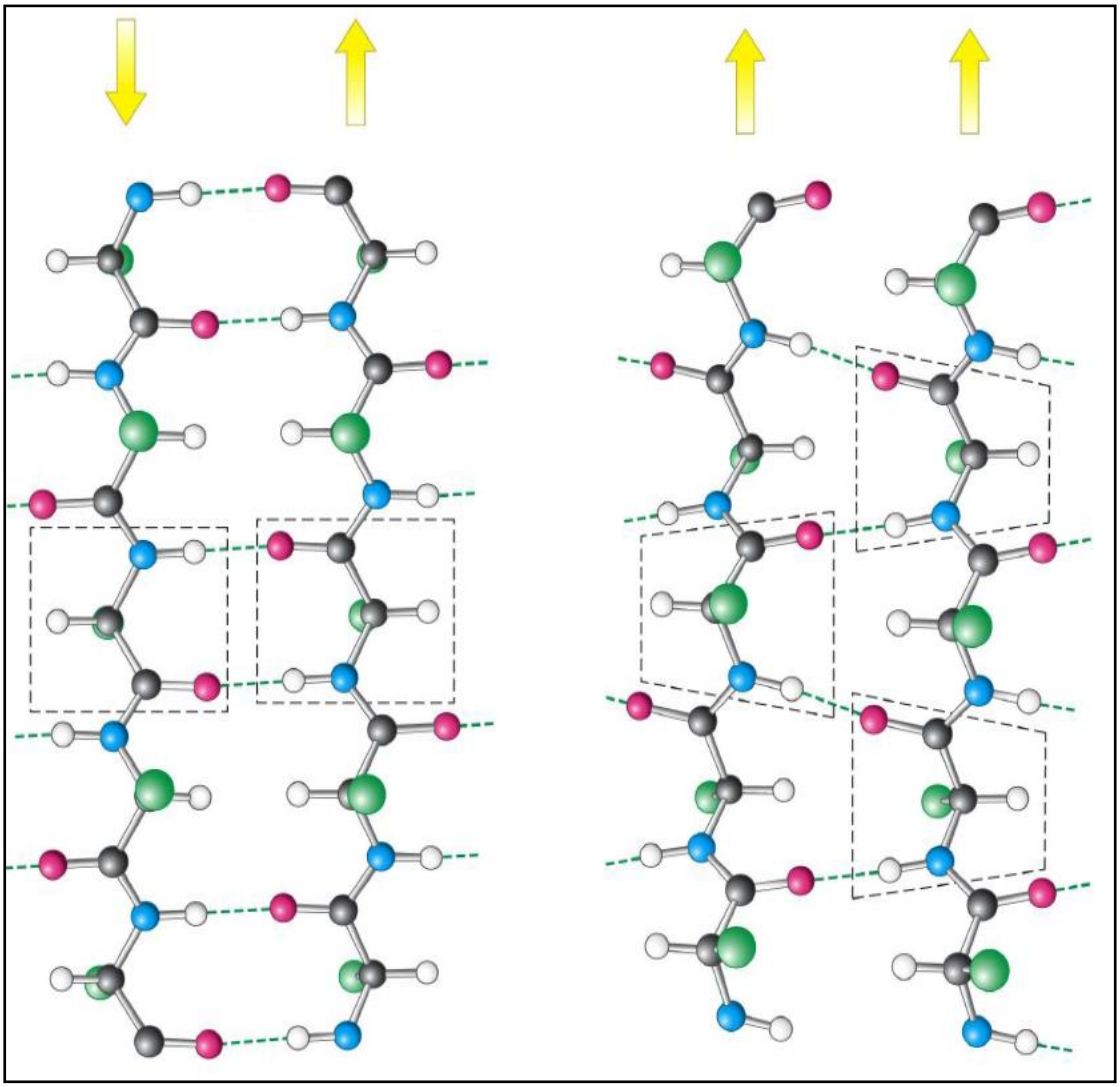
2.3.2 Β-πτυχωτές επιφάνειες

Οι β-πτυχωτές επιφάνειες (ή β-πτυχωτά φύλλα) είναι η δεύτερη κύρια μορφή δευτεροταγούς δομής μιας πρωτεΐνης που προτάθηκε θεωρητικά από τους Pauling και Corey [7, 8], όπως και η α-έλικα. Για να σχηματιστούν οι β-πτυχωτές επιφάνειες προηγείται ο σχηματισμός υδρογονικών δεσμών, είτε μεταξύ των αμινομάδων και καρβοξυλομάδων διαφορετικών αλυσίδων αμινοξέων, είτε μεταξύ διαφορετικών τμημάτων της ίδιας αμινοξικής ακολουθίας [2]. Αυτά τα τμήματα ονομάζονται β-κλώνοι (β-strands).

Οι αμινο- και καρβοξυλο- ομάδες προσανατολίζονται περίπου κάθετα στον άξονα της αλυσίδας, η οποία είναι σχεδόν εντελώς εκτεταμένη, ώστε να είναι δυνατή η δημιουργία υδρογονικών δεσμών με αντίστοιχες ομάδες διπλανών αλυσίδων (β-κλώνων). Η ανάγκη σχηματισμού ισχυρών υδρογονικών δεσμών (περίπου ευθύγραμμων) έχει σαν αποτέλεσμα οι β-κλώνοι να μην είναι τελείως εκτεταμένοι, με συνέπεια η επιφάνεια που σχηματίζεται από αυτούς να παρουσιάζει μία ελαφρά κάμψη ή πτύχωση, απ' όπου και «πτυχωτή» η ονομασία της.

Είναι δυνατόν, οι β-κλώνοι που αποτελούν μία β-πτυχωτή επιφάνεια να έχουν την ίδια διεύθυνση, οπότε η δομή ονομάζεται παράλληλη β-πτυχωτή επιφάνεια ή αντίθετες διευθύνσεις, οπότε δημιουργείται η αντιπαράλληλη β-πτυχωτή επιφάνεια (βλ. **Εικόνα 2.13**). Σπανιότερα, μία β-πτυχωτή επιφάνεια είναι μείγμα παράλληλων και αντιπαράλληλων β-κλώνων. Οι δεσμοί υδρογόνου μεταξύ C=O και H-N συνδέουν ένα και μόνο αμινοξύ στην γειτονική β-πτύχωση. Οι πλευρικές αλυσίδες (στην εικόνα με πράσινο χρώμα) τοποθετούνται κάθετα στον άξονα των β-κλώνων και 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 2.13**). Κοινό αποτέλεσμα της δημιουργίας και των δυο κατηγοριών των επιφανειών είναι η σταθεροποίηση της δομής.



Εικόνα 2.13 Αντιπαράλληλη και Παράλληλη β-πτυχωτή επιφάνεια [45]

Στις β-πτυχωτές επιφάνειες παρατηρούνται 3 σημαντικές περιοδικότητες: Η επαναλαμβανόμενη περίοδος κατά μήκος των πολυπεπτιδικών αλυσίδων (β-κλώνων), η απόσταση μεταξύ δύο β-κλώνων (που συνδέονται με υδρογονικούς δεσμούς) και η απόσταση μεταξύ δύο διαδοχικών πτυχωτών επιφανειών (στην περίπτωση που περισσότερες από μία πτυχωτές επιφάνειες «πακετάρονται» η μία πάνω στην άλλη).

Η πρώτη περιοδικότητα είναι περίπου 6,5 Å στις παράλληλες β-πτυχωτές επιφάνειες και 7,0 Å στις αντιπαράλληλες (ανά δύο αμινιξέα κατά μήκος των β-κλώνων). Η απόσταση μεταξύ δύο β-κλώνων είναι περίπου 4,7 Å. Η απόσταση μεταξύ δύο διαδοχικών β-πτυχωτών επιφανειών εξαρτάται από το μέγεθος των πλευρικών αλυσίδων των αμινοξέων και ποικίλλει από 3,5 έως 12 Å περίπου.

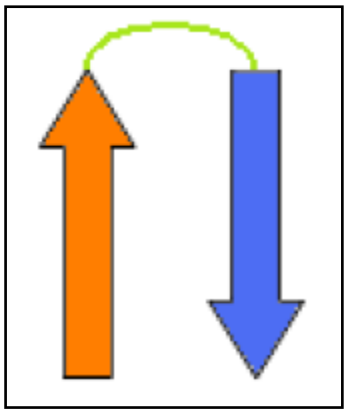
Οι περισσότερες β-πτυχωτές επιφάνειες περιέχουν λιγότερους από 6 β-κλώνους, χωρίς να έχει παρατηρηθεί «προτίμηση» για συγκεκριμένο αριθμό. Εξάλλου, υπάρχει μικρή προτίμηση οι β-κλώνοι να έχουν περιττό αριθμό αμινοξέων. Δεν παρατηρήθηκε επίσης προτίμηση για παράλληλες ή αντιπαράλληλες επιφάνειες, αλλά παράλληλες επιφάνειες με λιγότερους από 4 β-κλώνους είναι σπάνιες. Μεικτές πτυχωτές επιφάνειες, δηλαδή με παράλληλους και αντιπαράλληλους κλώνους εμφανίζονται μεν, αλλά είναι λιγότερο συχνές από ότι θα αναμενόταν με τυχαία ανάμειξη των διευθύνσεων. Οι κλώνοι δηλαδή των β-πτυχωτών επιφανειών επιδεικνύουν κάποια «συνεργατικότητα»: τείνουν να είναι όλοι παράλληλοι ή αντιπαράλληλοι.

Αμινοξέα που ευνοούν τη δημιουργία β-πτυχωτών επιφανειών, είναι κυρίως αμινοξέα με ογκώδεις υδρόφοβες αλυσίδες (Βαλίνη, Ισολευκίνη, Λευκίνη, Τρυπτοφάνη, Κυστεΐνη, Τυροσίνη), οι οποίες διευθετούνται σχετικά άνετα στη στερεοδιάταξη της β-πτυχωτής επιφάνειας (το «πακετάρισμα» των πλευρικών αλυσίδων τους ευνοεί σημαντικά τις αλληλεπιδράσεις van der Waals).

Οι παράλληλες β-πτυχωτές επιφάνειες είναι συνήθως «θαμμένες» στο εσωτερικό των πρωτεϊνών και οι δύο επιφάνειές τους συνήθως «προστατεύονται» από άλλα τμήματα της κύριας αλυσίδας (συχνά α-έλικες), όντας ταυτόχρονα υδρόφοβες. Οι αντιπαράλληλες πτυχωτές επιφάνειες πολλές φορές έχουν τη μία πλευρά τους εκτεθειμένη στο διαλύτη και την άλλη «θαμμένη», πράγμα που απεικονίζεται στην αμινοξική τους ακολουθία, η οποία εμφανίζει εναλλακτικά υδρόφιλα και υδρόφοβα κατάλοιπα. Οι β-πτυχωτές επιφάνειες παρουσιάζουν γενικά μεγαλύτερη υδροφοβικότητα στο κέντρο τους παρά στους ακραίους β-κλώνους.

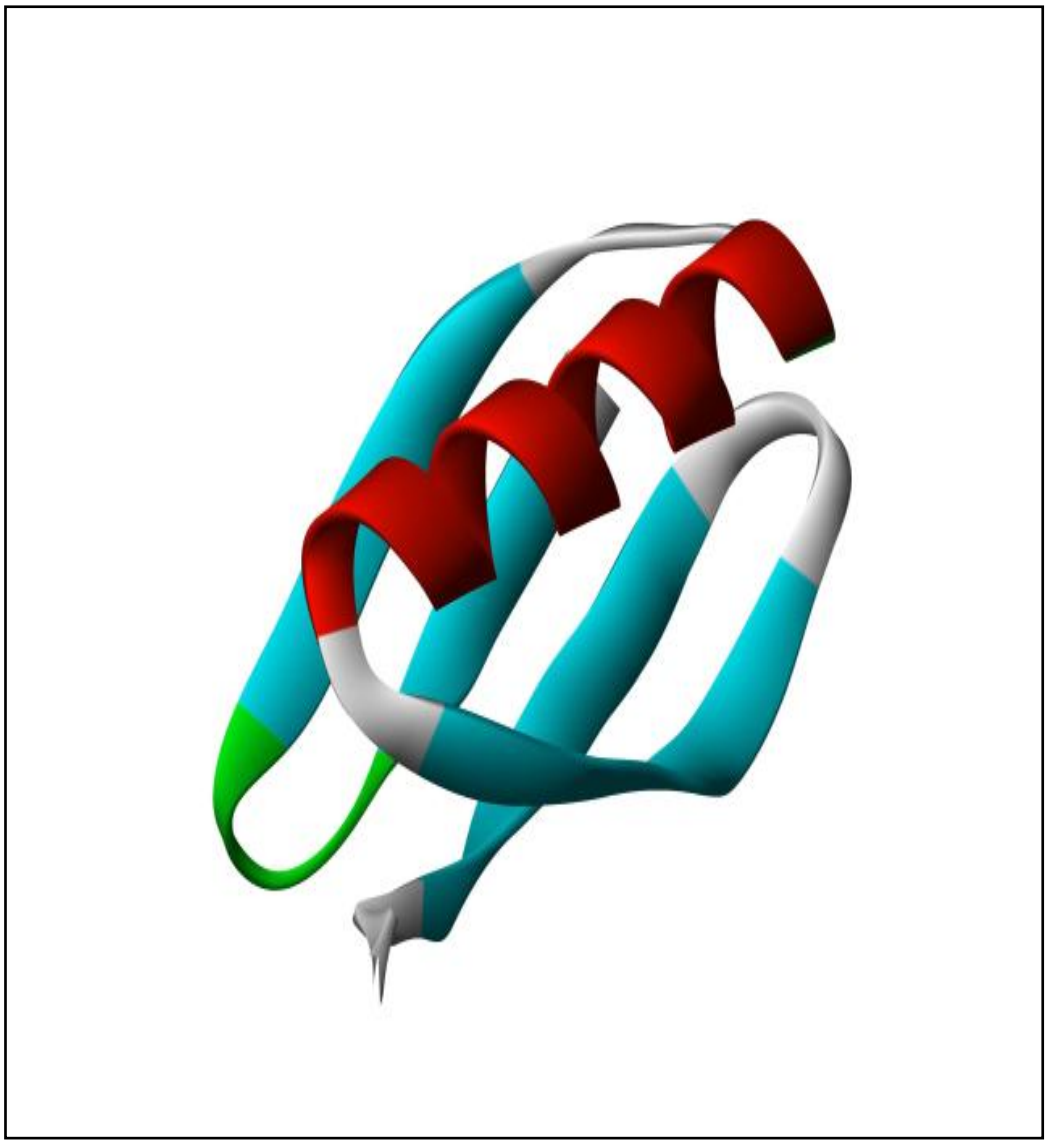
Η συχνότητα εμφάνισης των β-πτυχωτών επιφανειών σε σφαιρικές υδατοδιαλυτές πρωτεΐνες είναι μεγάλη. Είναι της τάξης του 20-25% (και διαρκώς αυξάνεται όσο προστίθενται καινούριες κρυσταλλογραφικά επιλυθείσες δομές). Βρίσκονται σε πολλές τάξεις πρωτεϊνών: τις πρωτεάσες, τις αφυδρογονάσες κλπ. Επίσης εμφανίζονται σε ινώδεις (δομικές) πρωτεΐνες του χορίου των μεταξοσκωλήκων, τις κερατίνες των φτερών και λεπιών, στην ουρά φάγων κλπ. Είναι εξίσου διαδεδομένη και σημαντική δευτεροταγής δομή όπως και η α-έλικα.

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 (βλ. **Εικόνα 2.14**). Γειτονικοί αντιπαράλληλοι β-κλώνοι συνδέονται με βρόχους φουρκέτας. Τέτοιοι βρόχοι είναι συχνά μικροί και δεν έχουν κάποια κανονική δευτεροταγή δομή.



Εικόνα 2.14 Τοπολογικό διάγραμμα αντιπαράλληλης β-πτύχωσης [45]

Ενδεικτικά, στην παρακάτω **Εικόνα 2.15** φαίνεται η τρισδιάσταση δομή της πρωτεΐνης ‘ 2IGD’, όπου με κόκκινο χρώμα παρουσιάζονται οι έλικες, με μπλε χρώμα οι β-πτυχωτές επιφάνειες.



Εικόνα 2.15 Η πρωτείνη '2IGD' με χρωματισμένες τις δευτερεύουσες δομές της [49]

2.3.3 Μη Κανονική Δομή (coil)

Με τον όρο μη κανονική δομή (coil ή random coil) χαρακτηρίζονται τα τμήματα εκείνα μίας πολυπεπτιδικής αλυσίδας που δεν είναι α-έλικα ή β-πτυχωτή επιφάνεια. Οι περισσότερες πρωτεΐνες αποτελούνται από συνδυασμούς στοιχείων δευτεροταγούς δομής, α-ελίκων και β-πτυχωτών επιφανειών, τα οποία συνδέονται από τμήματα μη κανονικής δομής με διάφορα μήκη και ακανόνιστο σχήμα. Αυτά τα τμήματα μη κανονικής δομής βρίσκονται στην επιφάνεια του μορίου. Οι καρβοξυλομάδες και αμινομάδες της κύριας

αλυσίδας αυτών των τμημάτων, οι οποίες γενικά δεν σχηματίζουν υδρογονικούς δεσμούς, είναι εκτεθειμένες στο διαλύτη και μπορούν να σχηματίσουν υδρογονικούς δεσμούς με μόρια νερού. Πληροφορίες σχετικά με την ύπαρξη πιθανών κανονικοτήτων αυτών των τμημάτων δεν έχουν ακόμη τεκμηριωθεί.

2.4 Στερεοχημική διαμόρφωση πρωτεϊνών

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

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. Η μακριά πρωτεϊνική αλυσίδα αναδιπλώνεται στο χώρο λαμβάνοντας μία συγκεκριμένη τρισδιάστατη μορφή, η οποία σταθεροποιείται χάρη σε ασθενείς μη ομοιοπολικούς δεσμούς που δημιουργούνται τόσο από άτομα του πολυπεπτιδικού σκελετού, όσο και από άτομα των πλευρικών αλυσίδων των αμινοξέων. Οι ασθενείς αυτοί δεσμοί είναι δεσμοί υδρογόνου, ιοντικοί δεσμοί και δεσμοί van der Waals. Οι δεσμοί αυτοί μόνοι τους είναι αρκετά ασθενείς σε σχέση με τους ομοιοπολικούς δεσμούς του μορίου έτσι, είναι απαραίτητοι πολλοί μη ομοιοπολικοί δεσμοί, ώστε να συγκρατηθούν δύο περιοχές της διπλωμένης πολυπεπτιδικής αλυσίδας. Επομένως, η σταθερότητα κάθε διπλωμένης δομής επηρεάζεται από τη συνολική ισχύ πολλών τέτοιων δεσμών.

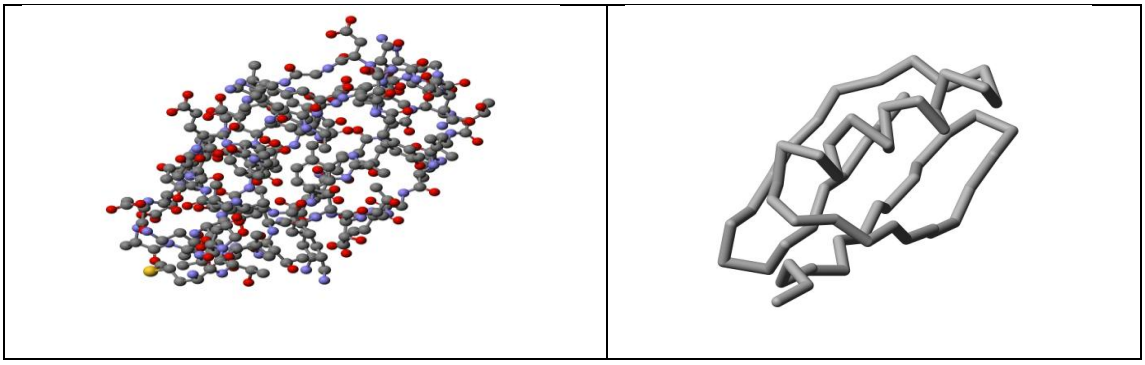
Ένα τέταρτο είδος ασθενούς δύναμης παίζει σημαντικό ρόλο στον καθορισμό του σχήματος μιας πρωτεΐνης. Για την κατανόηση της συγκεκριμένης αλληλεπίδρασης πρέπει να αναφέρουμε ότι οι πλευρικές αλυσίδες των αμινοξέων κατατάσσονται σε δύο μεγάλες ομάδες τις Υδρόφοβες (μη-

πολικές), που δεν διαθέτουν στο σύνολο τους κάποιο ηλεκτρικό φορτίο ή διαθέτουν πολύ μικρό και χάνουν τη δομή τους όταν βρεθούν σε επαφή με μόρια νερού, καθώς επίσης και τις Υδρόφιλες (πολικές) τα μέλη των οποίων διαθέτουν κάποιο σημαντικό θετικό ή αρνητικό φορτίο και δεν επηρεάζονται από την παρουσία μορίων νερού στο περιβάλλον τους.

Σύμφωνα με τα παραπάνω χαρακτηριστικά των πλευρικών ομάδων έχουμε την εξής διαδικασία: τα υδρόφοβα μόρια μεταξύ τους και οι μη πολικές πλευρικές αλυσίδες συγκεκριμένων αμινοξέων, όταν βρεθούν σε υδατικό περιβάλλον, συνωθούνται μαζί έτσι ώστε να ελαχιστοποιηθεί η διασπαστική δράση τους στο δίκτυο των υδρογονοδεσμών του νερού. Επομένως, ένας σημαντικός παράγοντας που κατευθύνει την αναδίπλωση ή πτύχωση (folding) κάθε πρωτεΐνης είναι η κατανομή των πολικών και μη πολικών αμινοξέων της.

Ο μηχανισμός, πιο συγκεκριμένα, είναι ο εξής: Οι μη πολικές (υδρόφοβες) πλευρικές αλυσίδες μιας πρωτεΐνης τείνουν να συναθροίζονται στο εσωτερικό του μορίου αποφεύγοντας έτσι την επαφή με το νερό, αντίθετα οι πολικές (υδρόφιλες) πλευρικές αλυσίδες τείνουν να διατάσσονται κοντά στην εξωτερική επιφάνεια του μορίου, όπου μπορεί να σχηματίσουν δεσμούς υδρογόνου με το νερό ή με άλλα πολικά μόρια. Όταν πολικά αμινοξέα βρίσκονται κρυμμένα στο εσωτερικό του μορίου συνήθως συνδέονται με δεσμούς υδρογόνου με άλλα πολικά αμινοξέα ή με τον πολυπεπτιδικό σκελετό.

Κατά συνέπεια, ένα πρωτεϊνικό μόριο διατάσσεται στο χώρο ακολουθώντας του νόμους των παραπάνω δεσμών δημιουργώντας μια σχετικά χαοτική δομή. Μέσα στη δομή αυτή όμως παρατηρούμε, όπως αναφέρθηκε, και κάποια επαναλαμβανόμενα μοτίβα, όπως η α-έλικα και το β-πτυχωτό φύλλο. Οι δομές αυτές δημιουργούνται λόγω αλληλεπιδράσεων των ατόμων κυρίως του πεπτιδικού σκελετού και μας δίνουν ιδιαίτερα σταθερές δομές που συνάδουν και με το φαινόμενο της υδροφοβίας και υδροφιλίας. Στην **Εικόνα 2.16** φαίνεται η τριτοταγής δομή της πρωτεΐνης ‘2IGD’, τόσο στην πλήρη δομή των ατόμων της όσο και στην μορφή μόνο των Ca ατόμων.



Εικόνα 2.16 Τριτοταγής δομή της πρωτεΐνης '2IGD'
(δεξιά φαίνονται όλα τα άτομα – αριστερά μόνο τα Ca άτομα) [49]

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 μιας ορθής αναδίπλωσης 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.

2.5 Λειτουργίες των πρωτεϊνών

Ο ρόλος των πρωτεϊνών είναι σημαντικός για όλες σχεδόν τις λειτουργίες των οργανισμών. Αυτό διαφαίνεται από τις πολλές λειτουργίες τις οποίες συντελούν και οι οποίες μπορούν να διαχωριστούν σε οκτώ γενικές κατηγορίες [2], οι οποίες είναι οι εξής:

2.5.1 Ενζυμική καταλυτική δράση

Ο πιο γνωστός ρόλος των πρωτεϊνών στα κύτταρα είναι ο ρόλος τους ως ένζυμα, τα οποία καταλύουν χημικές αντιδράσεις. Σχεδόν όλες οι χημικές

αντιδράσεις στα βιολογικά συστήματα καταλύονται από ειδικά μακρομόρια που ονομάζονται ένζυμα. Μερικές από τις αντιδράσεις αυτές είναι αρκετά απλές, όπως π.χ. η ενυδάτωση του CO2. Άλλες, όπως η αντιγραφή ενός ολοκλήρου χρωμοσώματος, είναι αρκετά πολύπλοκες. Τα ένζυμα έχουν πολύ ουσιαστικό ρόλο στις βασικές διεργασίες του κυττάρου: στην αντιγραφή, τη μεταγραφή, τη μετάφραση και τις μετα-μεταγραφικές τροποποιήσεις. Η δράση των ενζύμων επιταχύνει τις χημικές αντιδράσεις από 106 μέχρι και 1017 φορές. Από δομική άποψη, αν και τα ένζυμα μπορεί να αποτελούνται από εκατοντάδες αμινοξέα, μόνο ένα μικρό ποσοστό των αμινοξέων συμμετέχει στην κατάλυση. Περισσότερα από 100 ένζυμα έχουν ταυτοποιηθεί και πολύ περισσότερα έχουν κρυσταλλωθεί. Το αξιοσημείωτο είναι ότι όλα τα γνωστά ένζυμα είναι πρωτεΐνες.

2.5.2 Μηχανισμός άμυνας (ανοσοπροστασία)

Ειδικές πρωτεΐνες χρησιμοποιούν το σχήμα τους για να αναγνωρίσουν ξένους εισβολείς και καρκινικά κύτταρα. Αυτές οι πρωτεΐνες ονομάζονται αντισώματα και αποτελούν τον κύριο κορμό του ανοσοποιητικού συστήματος. Τα αντισώματα 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.

2.5.3 Μεταφορά και αποθήκευση ουσιών

Πολλές πρωτεΐνες συνδέονται με συγκεκριμένα μικρά μόρια και ιόντα 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 στα ερυθροκύτταρα. Η μυοσφαιρίνη μεταφέρει οξυγόνο και το αποθηκεύει προσωρινά στους μύες των σπονδυλωτών. Η τρανσφεράση μεταφέρει σίδηρο στο πλάσμα του αίματος, ο οποίος αποθηκεύεται στο συκώτι ως σύμπλοκο με τη φερριτίνη, μια διαφορετική πρωτεΐνη.

2.5.4 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 προώθηση του σπέρματος μέσω μαστιγίων επιτυγχάνεται με συσταλτά πρωτεϊνικά συγκροτήματα.

2.5.5 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.

2.5.6 Μηχανική υποστήριξη

Ινώδεις και νηματοειδείς πρωτεΐνες έχουν δομικό ρόλο στους οργανισμούς. Για παράδειγμα η αντοχή δέρματος και οστών οφείλονται στο κολλαγόνο και την ελαστίνη, δυο πρωτεΐνες που σχηματίζουν ίνες.

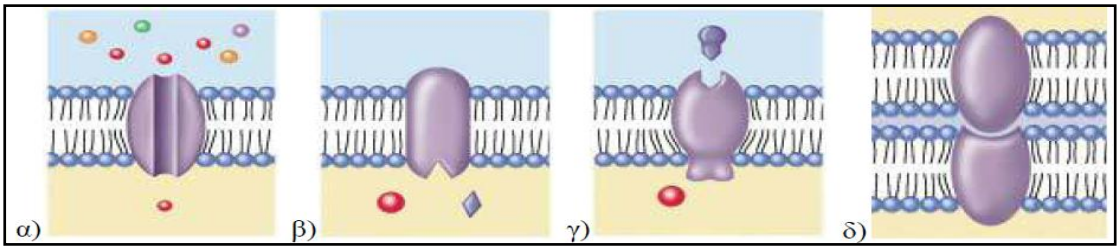
2.5.7 Έλεγχος της ανάπτυξης και της διαφοροποίησης

Οι πρωτεΐνες χρησιμεύουν στην ελεγχόμενη έκφραση των γενετικών πληροφοριών, ουσιώδη διαδικασία για την κανονική ανάπτυξη και διαφοροποίηση των κυττάρων. Μόνο ένα μικρό κλάσμα του γονιδιώματος ενός κυττάρου εκφράζεται κάθε φορά. Στα βακτήρια, πρωτεΐνες-καταστολείς, είναι σπουδαίοι παράγοντες ελέγχου που «σιγούν» ειδικά τμήματα του DNA ενός κυττάρου.

2.5.8 Δημιουργία και μετάδοση νευρικών παλμών

Η απόκριση των νευρικών κυττάρων σε ειδικά ερεθίσματα υποβοηθείται από πρωτεΐνες-υποδοχείς. Για παράδειγμα, η ροδοψίνη είναι ο φωτοϋποδοχέας-πρωτεΐνη στα ραβδία του αμφιβληστροειδή. Μόρια υποδοχείς που μπορούν να «διεγερθούν» με ειδικά μόρια, όπως η ακετυλοχολίνη, είναι υπεύθυνα για τη μετάδοση των νευρικών παλμών στις συνάψεις, δηλαδή τις επαφές μεταξύ νευρικών κυττάρων. Υπάρχουν ακόμα πολλοί ειδικοί ρόλοι, που έχουν οι πρωτεΐνες στην ζωή των κυττάρων και των οργανισμών.

Στην **Εικόνα 2.17** φαίνονται οι κυριότερες λειτουργίες των πρωτεϊνών.



Εικόνα 2.17 Λειτουργίες πρωτεϊνών (μεταφορά μορίων – ένζυμο – υποδοχέας χημικών ουσιών – σύνδεση κυττάρων) [47]

2.6 Αλληλεπίδραση πρωτεϊνών

Οι βιολογικές ιδιότητες μιας πρωτεΐνης εξαρτώνται από την ικανότητά τους να αλληλεπιδρούν με άλλα μόρια. Για παράδειγμα, τα αντισώματα συνδέονται με τους ιούς ή τα βακτήρια και με τον τρόπο αυτό ενεργοποιούν τους αμυντικούς μηχανισμούς του οργανισμού. Όλες οι πρωτεΐνες συνδέονται με άλλα μόρια. Σε κάθε περίπτωση η σύνδεση αυτή παρουσιάζει μεγάλη εξειδίκευση. Αυτό σημαίνει ότι κάθε πρωτεΐνη μπορεί να αλληλεπιδράσει μόνο με ένα ή έστω λίγα από τα χιλιάδες μόρια που έρχεται σε επαφή. Η ουσία που συνδέεται με την πρωτεΐνη μπορεί να είναι είτε ένα ιόν, είτε ένα μικρό μόριο, είτε κάποιο μακρομόριο (πρωτεΐνη) και αποκαλείται συνδέτης ή προσδέτης (ligand) για τη συγκεκριμένη πρωτεΐνη.

Η ικανότητα μιας πρωτεΐνης να συνδέεται εκλεκτικά και με μεγάλη συγγένεια με ένα συνδέτη οφείλεται στο σχηματισμό ενός συνόλου ασθενών, μη ομοιοπολικών δεσμών (δεσμοί υδρογόνου, ιοντικοί δεσμοί, έλξεις van der Waals) και ευνοϊκών υδρόφοβων αλληλεπιδράσεων. Από μόνος του κάθε ένας από τους παραπάνω δεσμούς είναι ασθενής. Έτσι, προϋπόθεση για να αναπτυχθεί μία αποτελεσματική αλληλεπίδραση είναι να σχηματιστούν ταυτόχρονα πολλοί ασθενείς δεσμοί. Αυτό μπορεί να επιτευχθεί μόνο αν το επιφανειακό περίγραμμα του μορίου του συνδέτη ταιριάζει πολύ καλά με την πρωτεΐνη, όπως το «χέρι μέσα σε ένα γάντι».

Η περιοχή της πρωτεΐνης που αλληλεπιδρά με το συνδέτη αποκαλείται θέση σύνδεσης. Η θέση σύνδεσης συνήθως αποτελείται από μία κοιλότητα στην επιφάνεια της πρωτεΐνης που σχηματίζεται από μία συγκεκριμένη διάταξη αμινοξέων. Τα αμινοξέα αυτά ανήκουν σε περιοχές της πολυπεπτιδικής αλυσίδας που απέχουν αρκετά μεταξύ τους. Εν τούτοις, έχουν πλησιάσει πολύ κοντά χάρη στην πτύχωση της πρωτεΐνης.

Επιπλέον, μία μεγάλη πρωτεΐνη μπορεί να αποτελείται από δύο ή περισσότερες πρωτεΐνες. Αυτό συμβαίνει όταν μια θέση σύνδεσης αναγνωρίζει την επιφάνεια μιας δεύτερης πρωτεΐνης και τότε η ισχυρή σύνδεση των δύο πτυχωμένων πολυπεπτιδικών αλυσίδων στη θέση αυτή,

δημιουργεί μια μεγαλύτερη πρωτεΐνη με μοναδική γεωμετρία. Στην απλούστερη περίπτωση, δύο (ή παραπάνω) ταυτόσημες αλυσίδες συνδέονται μεταξύ τους σχηματίζοντας ένα συμμετρικό σύμπλεγμα γνωστό ως διμερές (πολυμερές). Η ακεραιότητα του συμπλέγματος εξασφαλίζεται από τις αλληλεπιδράσεις ανάμεσα σε δύο ταυτόσημες θέσεις σύνδεσης.

Μελετώντας όλα τα παραπάνω στοιχεία αντιλαμβανόμαστε ότι η πλήρης αναγνώριση της λειτουργίας μιας πρωτεΐνης θα μπορούσε να γίνει μόνο με την αναγνώριση της τρισδιάστατης δομής της πρωτεΐνης, πρόβλημα που αποτελεί το κυριότερο σήμερα στόχο της Βιοπληροφορικής και στόχο της παρούσας εργασίας.

3 Πρόβλημα Αναγνώρισης Αναδίπλωσης Πρωτεϊνών

Στο κεφάλαιο αυτό γίνεται μία εκτενής περιγραφή του προβλήματος της πρόβλεψης των πρωτεϊνικών δομών (Ενότητα 3.1). Στην Ενότητα 3.2 παρουσιάζονται οι μέθοδοι που χρησιμοποιούνται ευρέως για την αντιμετώπιση του προβλήματος. Στην Ενότητα 3.3 παρουσιάζεται συνοπτικά ένα από τα πιο διαδεδομένα λογισμικά πρόβλεψης της δομής μιας πρωτεΐνης, το λογισμικό Modeller. Το κεφάλαιο κλείνει στην Ενότητα 3.4 με τα συμπεράσματα που έχουν προκύψει στην βιβλιογραφία από την σύγκριση των μεθόδων που αναφέρθηκαν στην προηγούμενη ενότητα.

3.1 Περιγραφή του προβλήματος

Ενώ οι πρωτεϊνικές ακολουθίες εξάγονται με ταχείς ρυθμούς, η εργαστηριακή ταυτοποίηση της δομής τους στο χώρο είναι μία αργή και δαπανηρή διαδικασία. Οι εργαστηριακές μέθοδοι που χρησιμοποιούνται είναι:

- η κρυσταλλογραφία με ακτίνες Χ,
- ο πυρηνικός μαγνητικός συντονισμός και
- η φασματομετρία μάζας.

Το 1971 ιδρύθηκε στα Brookhaven National Laboratories (BNL) των ΗΠΑ η Τράπεζα Δεδομένων Πρωτεϊνών - Protein Data Bank (PDB), η οποία περιελάμβανε 7 3D 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, ο ρυθμός προσθήκης δεδομένων στην τράπεζα δεδομένων πρωτεϊνών αυξάνεται δραματικά, αλλά πάντα ο αριθμός των πρωτεϊνών των

οποίων είναι γνωστή η δομή, παραμένει πολύ μικρός. Σύμφωνα με τους [9, 10], η τράπεζα PDB περιέχει 6.010.749 εγγραφές, ενώ η δομή είναι γνωστή μόνο για τις 47.873 από αυτές. Κάθε μήνα προστίθενται στη τράπεζα πρωτεϊνών 200 περίπου πρωτεΐνες. Το να γνωρίζουμε τη δομή μιας πρωτεΐνης είναι καίριο για να καταλάβουμε πώς λειτουργεί αλλά και να για την στοχεύσουμε με φάρμακα σε περίπτωση δυσλειτουργίας. Γίνεται λοιπόν εύκολα αντιληπτό πόσο σημαντικό είναι να αναπτυχθούν άλλες πιο γρήγορες και οικονομικές μέθοδοι πρόγνωσης της δομής των πρωτεϊνών σε σχέση με τις υπάρχουσες χρονοβόρες και ακριβές εργαστηριακές μεθόδους πρόγνωσης.

Η εύρεση μιας ταχείας υπολογιστικής μεθόδου για την πρόγνωση της δομής μιας πρωτεΐνης με ακρίβεια γίνεται άμεσα αντιληπτό πόσο σημαντική θα ήταν για την πρόοδο της Βιολογίας. Έτσι σχεδόν ολόκληρος ο κλάδος της βιοπληροφορικής στράφηκε σε αυτήν την κατεύθυνση. Τα οφέλη από την εύρεση μιας τέτοιας μεθόδου είναι ιδιαίτερα σημαντικά, καθώς θα επιταχύνει σημαντικά τη μελέτη του ανθρώπινου γονιδιώματος, κύριοι εκφραστές του οποίου είναι οι πρωτεΐνες, ανοίγοντας νέους ορίζοντες στην Ιατρική, τη Βιολογία αλλά και την κατανόηση της ζωής. Επιπλέον, η εύρεση μιας τέτοιας μεθόδου θα βοηθούσε στην εξακρίβωση της αλληλεπίδρασης μεταξύ δύο πρωτεϊνικών μορίων, βρίσκοντας άμεση εφαρμογή στη φαρμακολογία και συγκεκριμένα στη σχεδίαση νέων αποτελεσματικότερων και ασφαλέστερων φαρμάκων, βελτιώνοντας σημαντικά την καθημερινότητα και βιωσιμότητα πολλών ανθρώπων (ενδεικτικά τονίζεται ότι τα νέας γενιάς αντικαρκινικά φάρμακα βασίζονται στη σχέση πρωτεΐνης-συνδέτη για να στοχεύσουν σε καρκινικά κύτταρα χωρίς να επηρεάσουν σημαντικά τα υγιή κύτταρα του ασθενούς).

3.2 Μέθοδοι Πρόγνωσης Αναδίπλωσης Πρωτεϊνών

Μία μικρή πρωτεΐνη μπορεί να αποτελείται από λίγα αμινοξέα, αλλά οι ανθρώπινες πρωτεΐνες είναι συνήθως πολύ μεγαλύτερες και αποτελούνται μέχρι και από 1000 αμινοξέα. Ακόμα και οι μικρότερες πρωτεΐνες μπορούν να

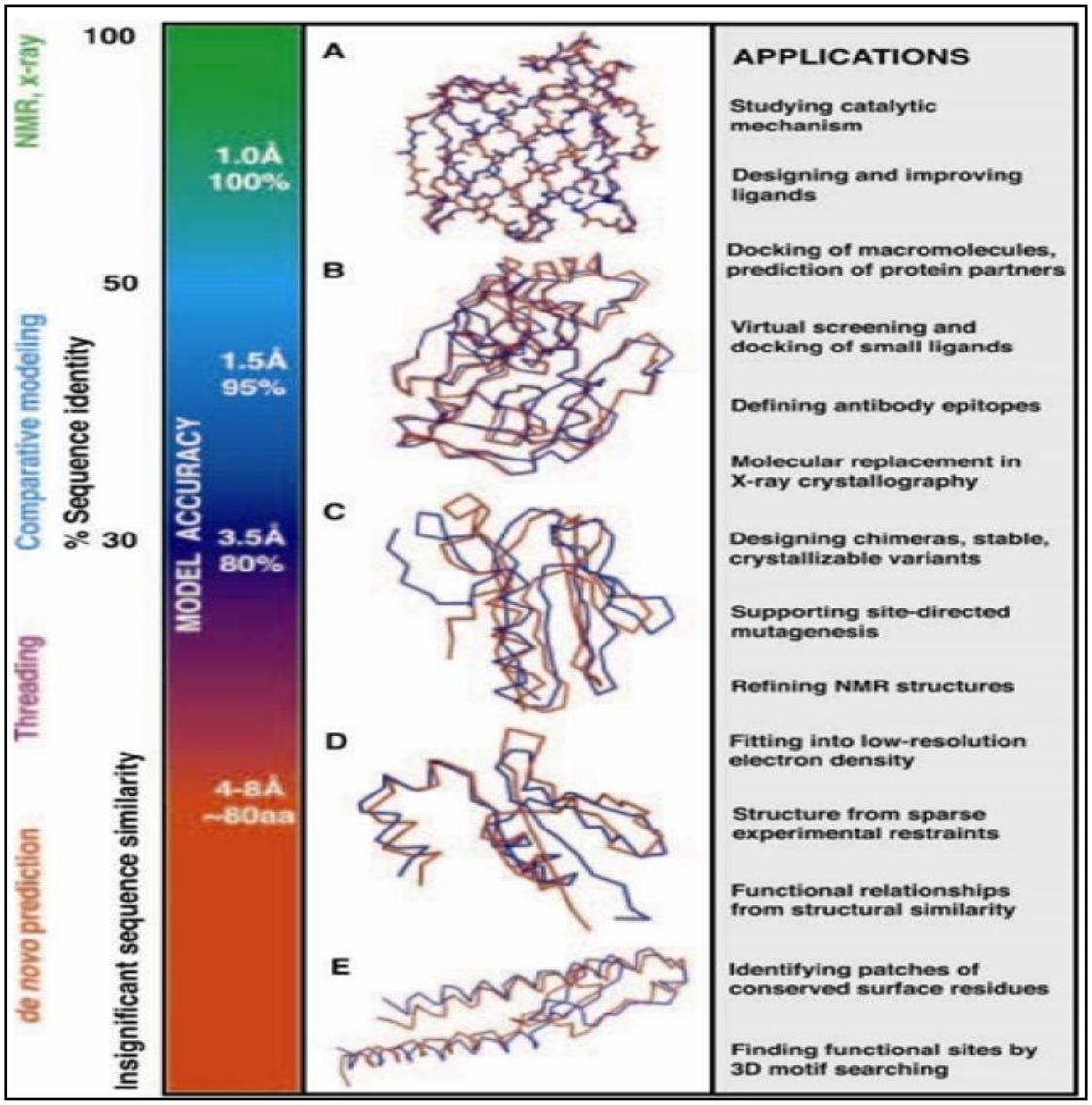
αναδιπλωθούν στο χώρο με πάρα πολλούς τρόπους γιατί έχουν πολλούς βαθμούς ελευθερίας, το οποίο είναι ευρέως γνωστό ως παράδοξο του Levinthal. Δηλαδή ότι ο αριθμός των δυνατών διαμορφώσεων στο χώρο μιας πολυπεπτιδικής αλυσίδας είναι τόσο μεγάλος, που η αξιολόγηση όλων των περιπτώσεων είναι αδύνατη.

Από την στιγμή που βρέθηκε εργαστηριακά η πρώτη δομή πρωτεΐνης, οι ερευνητές, χημικοί/βιοχημικοί και βιολόγοι οι οποίοι χρησιμοποιούσαν υπολογιστικές προσεγγίσεις επιχείρησαν να αναπτύξουν προσεγγίσεις ή πρωτόκολλα, με στόχο την πρόβλεψη και την προσομοίωση μιας τρισδιάστατης δομής με μοναδικό στοιχείο και αφετηρία την ακολουθία αμινοξέων της πρωτεΐνης. Αξίζει να αναφερθεί ότι η πλειοψηφία των πρώτων λογισμικών Βιοπληροφορικής επιχειρούσαν να επιλύσουν το «πρόβλημα της αναδίπλωσης των πρωτεϊνών». Μετά από 40 χρόνια εξέλιξης στο πεδίο αυτό, η πρόβλεψη και η προσομοίωση δομών πρωτεϊνών, συνεχίζει να αποτελεί ένα από τα πιο ενεργά ερευνητικά πεδία στη Δομική Βιοπληροφορική, στη Βιομοριακή Προσομοίωση, στη Δομική Βιοχημεία/Βιολογία και στο σχεδιασμό ενώσεων με βιολογικό και φαρμακευτικό ενδιαφέρον. Η ερευνητική δραστηριότητα στο συγκεκριμένο πεδίο, εκτός από το μεγάλο αριθμό επιστημονικών δημοσιεύσεων που αποδίδει ανά έτος, κατάφερε να οδηγήσει στην αποτελεσματική εφαρμογή νέων υπολογιστικών προσεγγίσεων. Οι προσεγγίσεις αυτές μπορούν να τοποθετηθούν σε τρεις άξονες:

1) προσομοίωση μέσω ομολογίας - συγκριτική προσομοίωση,
2) προσομοίωση μέσω threading, και
3) προσομοίωση ab initio, ή de novo προσομοίωση.

Υπάρχει αρκετή βιβλιογραφία για τις προσεγγίσεις αυτές. Λεπτομερείς πληροφορίες μπορεί ο αναγνώστης να βρει στο [11]. Και οι τρεις μέθοδοι είναι ουσιαστικά προσεγγιστικές προβλέψεις, με την έννοια ότι οι δομές οι οποίες προκύπτουν είναι δομικά μοντέλα και όχι δομές οι οποίες βασίζονται σε καθαρά πειραματικά δεδομένα, όπως αυτά προκύπτουν από την κρυσταλλογραφία ακτίνων Χ και από τον πυρηνικό μαγνητικό συντονισμό NMR. Οι δύο πρώτες μέθοδοι είναι ευριστικές, δηλαδή χρησιμοποιούν ήδη

προσδιορισμένες δομές για να προσδιορίσουν τη δομή μιας πρωτεΐνης αγνώστου δομής. Είναι προσεγγίσεις που αξιοποιούν την υπάρχουσα γνώση και προσπαθούν να εξάγουν συμπεράσματα από τα υφιστάμενα πειραματικά δεδομένα σχετικά με τη δομή μιας πρωτεΐνης, και αποκτούν ολοένα και περισσότερο ενδιαφέρον. Αντίθετα, οι απαρχής μέθοδοι δεν βασίζονται σε προηγούμενη γνώση, αλλά χρησιμοποιούν φυσικές και χημικές αρχές. Όλες οι μέθοδοι πρόγνωσης της δομής των πρωτεϊνών αξιολογούνται κάθε δυο χρόνια από το 1994, στα πλαίσια του διαγωνισμού CASP (Critical Assessment of Techniques for Protein Structure Prediction).



Εικόνα 3.1 Συσχέτιση της μεθόδου πρόβλεψης της δομής μιας πρωτείνης με την ακρίβεια των μοντέλων που προκύπτουν και τις εφαρμογές για κάθε μοντέλο [48]

Στην **Εικόνα 3.1** παρουσιάζεται μια σχηματική απεικόνιση της ποιότητας και ακρίβειας των δομικών μοντέλων ανάλογα με την μέθοδο μοριακής προσομοίωσης. Τα παραδείγματα της προσομοίωσης μέσω ομολογίας βασίζονται σε περίπου 60% (A), 40% (B) και 30% (C) ταύτιση της αμινοξικής ακολουθίας. Τα παραδείγματα της ab initio (de novo) προσομοίωσης παρουσιάζονται στο (D) και (E). Οι δομές που προβλέφθηκαν παρουσιάζονται με κόκκινο και οι πραγματικές δομές με μπλε.

Η ακρίβεια των δομών μειώνεται από το (A) στο (E) αλλά όπως γίνεται αντιληπτό η συνολική δομή των μοντέλων παραμένει σχετικά κοντά με την πραγματική δομή. Στη δεξιά στήλη του σχήματος συσχετίζεται η ακρίβεια ενός μοντέλου με την εφαρμογή που μπορεί να έχει ένα τέτοιο μοντέλο στην κατανόηση βιολογικών φαινομένων ή στην δομική μελέτη άλλων μορίων.

3.2.1 Μέθοδοι προτυποποίησης με ομολογία (homology modeling)

Ανάμεσα στις τρεις διαθέσιμες προσεγγίσεις, η πιο αποτελεσματική και ακριβής είναι η προσομοίωση μέσω ομολογίας (homology modeling). Η προσομοίωση μέσω ομολογίας αποδίδει σχετικά υψηλής ακρίβειας τρισδιάστατα δομικά μοντέλα πρωτεϊνών, βασισμένα σε ομοιότητες στην αμινοξική ακολουθία με κάποια άλλη διαθέσιμη τρισδιάστατη δομή στη βάση δεδομένων πρωτεϊνών (PDB). Στις δομές οι οποίες προκύπτουν από την προσομοίωση μέσω ομολογίας, η ποιότητα των μοντέλων εξαρτάται σημαντικά από το βαθμό της ομοιότητας ανάμεσα στην ακολουθία της αγνώστου δομής πρωτεΐνης και της ακολουθίας της πρωτεΐνης με την οποία επιχειρείται σύγκριση της ομολογίας τους. Οι πρωτεΐνες οι οποίες εμφανίζουν το μεγαλύτερο βαθμό ταύτισης είναι και αυτές για τις οποίες επιτυγχάνεται η προσομοίωσή τους με μεγαλύτερη ακρίβεια και αξιοπιστία. Ως γενικός κανόνας, η επιτυχία στην προσομοίωση μπορεί να οριστεί ως η μέση συμφωνία των συντεταγμένων μεταξύ της πρωτεΐνης της οποίας η δομή προσομοιάζεται (πρωτεΐνη – στόχος) και της δομής της πρωτεΐνης που χρησιμοποιείται ως εκμαγείο (η γνωστή και πειραματικά υπολογισμένη δομή,

και είναι η σύγκλιση κατά 0,3 Å για κάθε 10% αύξηση της ταύτισης/ομολογίας ανάμεσα στις δύο ακολουθίες.

Η προσομοίωση μέσω ομολογίας δεν μπορεί να εφαρμοστεί με αποτελεσματικότητα και αξιοπιστία για την προσομοίωση και την πρόβλεψη δομών πρωτεϊνών οι οποίες εμφανίζουν ποσοστό ταύτισης αμινοξικής ακολουθίας μικρότερο του 25-30% με την πρωτεΐνη αναφοράς. Παρ' όλα αυτά σε συγκεκριμένες σπάνιες περιπτώσεις, η προσομοίωση μέσω ομολογίας μπορεί να χρησιμοποιηθεί ακόμα και με ποσοστό ταύτισης μικρότερο του 25%. Στην περίπτωση αυτή συνυπολογίζονται και άλλα δομικά χαρακτηριστικά, όπως η τοπολογία, η ύπαρξη συμπαραγόντων όπως προσθετικές ομάδες (π.χ. αίμη), μεταλλικά ιόντα με παρόμοια αμινοξέα, θέσεις πρόσδεσης, πεπτιδικά τμήματα σύνδεσης μεταξύ τομέων της πρωτεΐνης, κ.λ.π.

Η προσομοίωση μέσω ομολογίας είναι μια πολυδιάστατη διαδικασία, η οποία περιλαμβάνει την αντιστοίχιση των ακολουθιών, τη δομική ταυτοποίηση, την αναζήτηση σε βάσεις δεδομένων, τη διερεύνηση της καταλληλότερης διαμόρφωσης του πεπτιδικού σκελετού και των πλευρικών αλυσίδων, την διευθέτηση των τμημάτων τα οποία δεν ταυτίζονται ανάμεσα στις δύο ακολουθίες (τα οποία συνήθως αποτελούν τις λεγόμενες θηλιές), την ελαχιστοποίηση της ενέργειας του δομικού μοντέλου που προκύπτει από την διαδικασία και την αποτίμηση της δομής που προκύπτει για να καταλήξει τελικά στη δημιουργία ενός αξιόπιστου δομικού μοντέλου.

Πιο συγκεκριμένα, η προσομοίωση μέσω ομολογίας μπορεί να αναλυθεί σε πέντε βασικά στάδια:

1) Αντιστοίχιση της αμινοξικής ακολουθίας της πρωτεΐνης με την άγνωστη δομή στην ακολουθία της πρωτεΐνης με τη γνωστή δομή, και επιλογή μιας διαθέσιμης δομής αυτής της πρωτεΐνης από την PDB.
2) Χρήση της αντιστοίχισης των ακολουθιών για την επιλογή και την αντικατάσταση τμημάτων του σκελετού (συνήθως θηλιές, οι δομές των οποίων είναι συγκεντρωμένες και διαθέσιμες σε ειδικές βιβλιοθήκες

ακολουθιών γνωστές για το σχηματισμό θηλιών), τα οποία χρειάζεται να υποστούν διαμορφωτικές τροποποιήσεις κατά την προσομοίωση. Κάτι τέτοιο είναι απαραίτητο όταν μεταξύ των υπό σύγκριση ακολουθιών υπάρχουν τμήματα τα οποία εμφανίζονται επιπλέον ή απουσιάζουν σε κάποια από τις δύο ακολουθίες.

3) Αντικατάσταση των πλευρικών αλυσίδων των αμινοξέων που διαφέρουν ανάμεσα στις δύο ακολουθίες, είτε λόγω της εισαγωγής είτε λόγω της έλλειψης πεπτιδικού τμήματος αμινοξέων στην ακολουθία των αμινοξέων της πρωτεΐνης-στόχου και επιτυχή προσομοίωση/μοντελοποίηση της διαμόρφωσης πεπτιδικών τμημάτων τα οποία σχηματίζουν θηλιές (loops).
4) Βελτιστοποίηση του μοντέλου που προκύπτει από την προσομοίωση, με χρήση πρωτοκόλλων Μοριακής Μηχανικής/Δυναμικής για την ελαχιστοποίηση της ενέργειας του συστήματος (μήκη/γωνίες δεσμών, στερεοχημικές αλληλεπιδράσεις, δεσμοί-Η, κ.λ.π.).
5) Έλεγχος της ποιότητας, σε επίπεδο ανάλυσης και ορθότητας, των μοριακών δομών με χρήση λογισμικού μοριακής απεικόνισης και αποτίμησης μοριακών δομών (π.χ. PROCHECK).

Το πλέον καθοριστικό στάδιο στη διαδικασία προσομοίωσης μέσω ομολογίας είναι το πρώτο βήμα, δηλαδή η αντιστοίχιση των δύο ακολουθιών. Εσφαλμένη αντιστοίχιση ανάμεσα στις δύο ακολουθίες των πρωτεϊνών θα δημιουργήσει ένα ντόμινο λαθών ανάμεσα στα διάφορα στάδια της διαδικασίας προσομοίωσης, οδηγώντας τελικά σε ένα παραπλανητικό τρισδιάστατο δομικό μοντέλο της πρωτεΐνης.

Τα πιθανά σφάλματα στα δομικά μοντέλα τα οποία προκύπτουν από την διαδικασία της συγκριτικής προσομοίωσης μπορούν να εστιαστούν στα εξής σημεία:

I. Σφάλματα στη διάταξη/πακετάρισμα των πλευρικών αλυσίδων των αμινοξέων (παρατηρούνται μεγαλύτερες διαφορές ανάμεσα στις πλευρικές αλυσίδες των αμινοξέων της πρωτεΐνης-στόχου και της πρωτεΐνης αναφοράς σε σχέση με τον πεπτιδικό σκελετό, τα άτομα του οποίου παραμένουν ίδια).

II. Σφάλματα και αποκλίσεις από την πρωτεΐνη αναφοράς ακόμα και σε σωστά αντιστοιχισμένες περιοχές της ακολουθίας ως επακόλουθο της διαφοράς στην συγκεκριμένη περιοχή ανάμεσα στην πρωτεΐνη αναφοράς και την πρωτεΐνη – στόχο (αυτό μπορεί να οφείλεται είτε στο μέσο, διάλυμα ή στερεά κατάσταση και στον τρόπο υπολογισμού της δομής, είτε στα φυσικοχημικά χαρακτηριστικά των πλευρικών αλυσίδων των αμινοξέων όταν αυτά διαφέρουν ανάμεσα σε πρωτεΐνη – στόχο και αναφοράς).
III. Σφάλματα στην περιοχή όπου δεν υπάρχει διαθέσιμη δομή αναφοράς (εκμαγείο), και οφείλεται σε απουσία πεπτιδικών τμημάτων στη δομή
IV. Σφάλματα λόγω μη ακριβούς ή παραπλανητικής αντιστοίχισης των ακολουθιών της πρωτεΐνης-στόχου και της πρωτεΐνης αναφοράς.
V. Σφάλματα τα οποία οφείλονται σε λάθος επιλογή της πρωτεΐνης αναφοράς, ένα πρόβλημα το οποίο είναι συχνό όταν συγκρίνονται μακρινές μεταξύ τους πρωτεΐνες (<25% ταύτιση αμινοξικών ακολουθιών).

Αρχικά η διαδικασία της προσομοίωσης μέσω ομολογίας αποτελούσε μια πολύπλοκη διαδικασία, προσιτή μόνο σε όσους διέθεταν σημαντικές γνώσεις στη χρήση του λειτουργικού UNIX και γενικά υπολογιστικών προγραμμάτων. Ένας άλλος περιοριστικός παράγοντας ήταν η μικρή διαθεσιμότητα εξειδικευμένων προγραμμάτων μοριακής απεικόνισης καθώς και Η/Υ υψηλής υπολογιστικής ισχύος. Σήμερα, η εξέλιξη στην τεχνολογία των υπολογιστών και του λογισμικού περιόρισε σημαντικά τη χρονική διάρκεια των επί μέρους σταδίων. Έτσι η προσομοίωση μέσω ομολογίας μπορεί πλέον να εκτελείται καθημερινά και να εφαρμόζεται με ευχέρεια ακόμα και από μη εξειδικευμένους με υπολογιστικά συστήματα ερευνητές. Επιπλέον εκτός των εμπορικά διαθέσιμων λογισμικών υπάρχουν αρκετά αξιόπιστα λογισμικά ευρέως διαδεδομένα και ελεύθερης χρήσης, όπως το λογισμικό MODELLER [12]. Το λογισμικό αυτό είναι διαθέσιμο τόσο για λειτουργικά συστήματα UNIX και LINUX όσο και για την πλατφόρμα των WINDOWS.

Πρόσφατα η διαδικασία της μοριακής προσομοίωσης έχει γίνει διαθέσιμη μέσω διακομιστών στο διαδίκτυο. Πλέον υπάρχει ένας σημαντικός αριθμός

διακομιστών, όπως ο SWISS-MODEL όπου μπορεί να πραγματοποιηθεί αυτόματα, με μικρή συμμετοχή του χειριστή, η διαδικασία της προσομοίωσης μέσω ομολογίας από οποιονδήποτε ερευνητή.

Βέβαια οφείλει κανείς να είναι ιδιαίτερα προσεκτικός όταν προσπαθεί να επεκτείνει τα συμπεράσματα του για την δομή των πρωτεϊνών και στη λειτουργία των πρωτεϊνών. Πρωτεΐνες όμοιες σε μεγάλο βαθμό είναι δυνατόν 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.

3.2.2 Πρόβλεψη δομής με το λογισμικό MODELLER

Το MODELLER είναι ένα από τα πιο ευρέως διαδεδομένα λογισμικά για την πρόβλεψη/προσομοίωση δομής μέσω ομολογίας. Στην πιο απλή του μορφή, τα εισαγόμενα δεδομένα είναι η αμινοξική ακολουθία της πρωτεΐνης-στόχου, μία ή περισσότερες πρωτεΐνες αναφοράς, οι ατομικές συντεταγμένες (από την PDB) της δομής των πρωτεϊνών αναφοράς και ένα μικρό αρχείο εντολών (script). Με βάση αυτά τα δεδομένα, το MODELLER υπολογίζει αυτόματα ένα δομικό μοντέλο της πρωτεΐνης-στόχου μέσα σε λίγα λεπτά σε έναν κοινό Η/Υ, χωρίς να χρειάζεται να παρέμβει ο χρήστης.

Το MODELLER κατασκευάζει δομικά μοντέλα μέσω της ικανοποίησης κάποιων διαμορφωτικών περιορισμών (spatial restraints). Οι περιορισμοί αυτοί περιλαμβάνουν:

A. περιορισμούς εξ ομολογίας, που προέρχονται από την αντιστοίχιση ακολουθίας με τις πρωτεΐνες αναφοράς και εφαρμόζονται σε ατομικές αποστάσεις και δίεδρες γωνίες της ακολουθίας-στόχου,
B. στερεοχημικούς περιορισμούς όπως προτιμώμενα μήκη και γωνίες δεσμών, που προέρχονται από το πεδίο δυνάμεων (force field) μοριακής μηχανικής,
C. στατιστικές προτιμήσεις για δίεδρες γωνίες μη-δεσμικές ατομικές αποστάσεις, προερχόμενες από μία βάση γνωστών πρωτεϊνικών δομών και
D. προαιρετικά, μη-αυτόματους περιορισμούς βασισμένους σε πληροφορίες από φασματοσκοπία NMR, πειράματα αλληλεπίδρασης πρωτεϊνών με χημικές ενώσεις διασταυρούμενης διασύνδεσης (cross-linking), φθορισμομετρία, ηλεκτρονική μικροσκοπία, κατευθυνόμενη μεταλλαξιγένεση κ.ά.

Οι παραπάνω περιορισμοί εκφράζονται ως πιθανολογικές συναρτήσεις και συνδυάζονται σε μια συνάρτηση (objective function). Η συνάρτηση αυτή βελτιστοποιείται με τρόπο ώστε να παραβιάζεται όσο το δυνατόν μικρότερος αριθμός περιορισμών. Η διαδικασία της βελτιστοποίησης ακολουθεί τη μέθοδο της μεταβλητής συνάρτησης (variable target function), δηλαδή γίνεται σε κύκλους, με την αρχική συνάρτηση να λαμβάνει υπ' όψιν μόνο τους τοπικούς περιορισμούς (που αφορούν τα κοντινά αμινοξέα) και στη συνέχεια αυτή να ενσωματώνει όλο και πιο μακρινούς, μέχρι το σύνολό τους να συμπεριληφθεί στην τελική συνάρτηση. Οι μέθοδοι που χρησιμοποιούνται κατά τη βελτιστοποίηση είναι η μέθοδος συζευγμένου ανύσματος (conjugate gradients method) και η προσομοίωση ανόπτησης (simulated annealing). Η όλη διαδικασία είναι παρόμοια με τον υπολογισμό τριτοταγούς διαμόρφωσης στη φασματοσκοπία NMR.

Εκτός από τη δημιουργία δομικών μοντέλων με τη μέθοδο των διαμορφωτικών περιορισμών, το MODELLER μπορεί επίσης να πραγματοποιήσει de novo προσομοίωση της δομής των θηλιών (loops). Οι θηλιές στη δομή μιας πρωτεΐνης μπορεί να διαφοροποιούνται σε μεγάλο βαθμό, ακόμη και μεταξύ πρωτεϊνών με υψηλή ομολογία. Συχνά όμως

παίζουν σημαντικό ρόλο στη λειτουργία της πρωτεΐνης, συμμετέχοντας στο σχηματισμό των θέσεων δέσμευσης υποκαταστατών ή στο ενεργό κέντρο. Η προσομοίωση της δομής των θηλιών πραγματοποιείται με μια διαδικασία διαμορφωτικής έρευνας και βελτιστοποίησης. Οι χρησιμοποιούμενοι αλγόριθμοι είναι η μέθοδος συζευγμένου ανύσματος και η προσομοίωση ανόπτησης.

Το MODELLER προσφέρει και τη δυνατότητα εκτέλεσης διαφόρων βοηθητικών εργασιών, όπως αντιστοίχιση ακολουθιών ή δομών για δύο ή περισσότερες πρωτεΐνες, δημιουργία φυλογενετικών δέντρων κ.ά. Το πρόγραμμα στη μορφή για τοπική εγκατάσταση σε Η/Υ δεν περιλαμβάνει γραφικό περιβάλλον, αλλά λειτουργεί με αρχεία γραπτών εντολών (scripts).

3.2.3 Μέθοδοι ύφανσης πρωτεϊνών (protein threading)

Οι μέθοδοι ύφανσης πρωτεϊνών χρησιμοποιούνται σε περιπτώσεις όπου δεν εμφανίζεται υψηλή ομολογία. Οι 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 Τ 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 Τ, (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 αμινοξέων, άρα μπορεί να αποτελέσει στη συνέχεια είσοδο σε εργαλείο προτυποποίησης πρωτεϊνών με ομολογία.

3.2.4 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 ενέργειας, 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 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 εντυπωσιακά αποτελέσματα, δεν είναι όμως δυνατό να θεωρήσουμε πως έχει λυθεί το πρόβλημα της αναδίπλωσης 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 μια πολλά υποσχόμενη τεχνολογία που προέρχεται από την κοινότητα της Τεχνητής Νοημοσύνης, την τεχνολογία Constraint Logic Programming in Finite Domains (CLP(FD)), που αποτελεί το αντικείμενο του επόμενου κεφαλαίου.

4 Λογικός προγ/σμός με διαχείριση περιορισμών

Σε αυτό το κεφάλαιο παρουσιάζονται περιληπτικά οι κύριες ιδέες της τεχνολογίας του λογικού προγραμματισμού με διαχείριση περιορισμών (Constraint Logic Programming – CLP), η οποία αποτελεί το υπόβαθρο από την επιστήμη των υπολογιστών, πάνω στο οποίο βασίστηκε η εργασία. Οι περισσότερες από αυτές τις ιδέες χρησιμοποιούνται στην εργασία. Αυτή η μικρή εισαγωγή καλύπτει αυτά που πρέπει να γνωρίζει ο αναγνώστης. Για περισσότερες λεπτομέρειες μπορεί να γίνει αναφορά στο κλασικό βιβλίο των Marriot και Stuckey [13] και σε ένα πιο πρόσφατο βιβλίο του Apt [14].

4.1 Λογικός προγραμματισμός

Η ανάπτυξη του λογικού προγραμματισμού συνδέεται με την ανάγκη επίλυσης προβλημάτων χαρακτηριζόμενων από μεταβλητότητα. Π.χ., προβλήματα με πολλές παραλλαγές, όπως είναι τα συνδυαστικά προβλήματα ή τα προβλήματα βελτιστοποίησης που δεν μπορούν να προκαθοριστούν πλήρως, είναι συχνά δύσκολο να λυθούν αλγοριθμικά. Ακόμα και στις περιπτώσεις που αυτό είναι εφικτό, μία μικρή αλλαγή στις προδιαγραφές απαιτεί τη μορφοποίηση ή και τον ανασχεδιασμό του αλγορίθμου.

Η αρχή του λογικού προγραμματισμού μπορεί να αποδοθεί στους Kowalski και Colmerauer. Ο Kowalski μορφοποίησε την διαδικαστική ερμηνεία της προτασιακής λογικής του Horn. Έδειξε πως ένας κανόνας «A, αν B_1 και B_2 και... και B_n» μπορεί να διαβαστεί και να υλοποιηθεί σαν μια διαδικασία

διαδικαστικής γλώσσας προγραμματισμού, όπου A είναι το κεφάλι της διαδικασίας και B_i το σώμα της.

Η βασική διαφοροποίηση του λογικού προγραμματισμού, και του συγγενικού του, του συναρτησιακού προγραμματισμού, από τις άλλες προγραμματιστικές γλώσσες, έγκειται στην απουσία σαφών οδηγιών για τη λύση ενός προβλήματος. Σε ένα πρόγραμμα λογικού προγραμματισμού, τα δεδομένα, οι υποθέσεις ενός προβλήματος (εκφρασμένες με τη μορφή κανόνων) και ο στόχος, είναι αρκετά για την επίλυσή του.

Πολλές εφαρμογές της τεχνητής νοημοσύνης κατάφεραν να επωφεληθούν από την δύναμη του λογικού προγραμματισμού. Μια ιδιαίτερη περιοχή στην οποία ο λογικός προγραμματισμός προσφέρει πολλά πλεονεκτήματα είναι σαν μια τυπική και αφηρημένη γλώσσα καθορισμού απαιτήσεων για την ανάπτυξη λογισμικού. Μια λογική γλώσσα εφοδιασμένη με πρωτογενείς όρους της θεωρίας συνόλων είναι εξαιρετική για την φάση καθορισμού απαιτήσεων στο σχεδιασμό λογισμικού. Η σημασιολογία της προτασιακής λογικής είναι σαφής επιτρέποντας την έκφραση λεπτομερειών με μικρές εκφράσεις. Η βασική ιδέα είναι η χρήση μιας γλώσσας απαιτήσεων βασισμένη στην λογική πρώτης τάξης και ενός μεταφραστή/διερμηνέα ώστε οι απαιτήσεις να μπορούν να εκτελεστούν. Μια τέτοια δυνατότητα έχει το πλεονέκτημα εντοπισμού αρκετών σφαλμάτων ενός έργου λογισμικού στο στάδιο προσδιορισμού απαιτήσεων.

Όταν χρησιμοποιείται λογικός προγραμματισμός, η διαφορά ανάμεσα στις φάσεις της απαιτήσεων και της υλοποίησης του λογισμικού τείνει να μην υπάρχει. Προς το παρόν, δεν είναι δυνατόν να παραχθεί αυτόματα ένα αποδοτικό λογισμικό, πάντως είναι δυνατόν να παραχθεί ένα πρωτότυπο σε μικρό χρόνο που θα δρα όπως το τελικό λογισμικό με το μειονέκτημα της μικρότερης απόδοσης. Με άλλα λόγια ο χρήστης μπορεί να ελέγξει το λογισμικό και να τροποποιήσει τις απαιτήσεις ωσότου όλες οι δυνατότητες μοντελοποιηθούν σωστά.

4.1.1 Η γλώσσα Prolog

Η γλώσσα Prolog υλοποιεί τις ιδέες του λογικού προγ/σμού. Η γλώσσα αυτή αναπτύχθηκε την δεκαετία του 70 από τον B. Kowalski [15], ο οποίος ανακάλυψε ότι ο κατηγορηματικός λογισμός θα μπορούσε να χρησιμοποιηθεί σαν γλώσσα προγ/σμού, από δύο ομάδες ερευνητών καθοδηγούμενες από τους A. Colmerauer και D. Warren που υλοποίησαν αυτές τις ιδέες και από την εργασία πάνω στην μέθοδο της ενοποίησης που έκανε ο J. Robinson το 1965.

Από άποψη υλοποίησης, οι πρώτες εκδόσεις της Prolog ήταν μη αποδοτικές. Πολύ καλύτερη απόδοση επιτεύχθηκε με την υλοποίηση της αφηρημένης μηχανής του Warren (Warren Abstract Machine, WAM για συντομία). Ιδιαίτερη έμφαση δόθηκε στην υλοποίηση του αλγόριθμου ενοποίησης, χρησιμοποιώντας την διαδικασία του Paterson-Wegman [16] και τον αλγόριθμο Martelli-Montanari που μπορούσαν να λύσουν ένα πρόβλημα σε γραμμικό χρόνο. Πάντως, σε τυπικές υλοποιήσεις ο αλγόριθμος ενοποίησης τρέχει σε τετραγωνικό χρόνο στην χειρότερη περίπτωση, αλλά είναι ευκολότερο να κωδικοποιηθεί.

Η αφηρημένη μηχανή του WAM έχει δύο μειονεκτήματα:

1) βασικές δομές δεδομένων που προσφέρονται από διαδικαστικές γλώσσες δεν είναι άμεσα διαθέσιμες, αν και κάποιες υλοποιήσεις παρέχουν βιβλιοθήκες για τον σκοπό αυτό και
2) παρέχονται διασυνδέσεις σε άλλες γλώσσες για την κωδικοποίηση ρουτινών που είναι δύσκολο να κωδικοποιηθούν στην Prolog.

Η Prolog βασίζεται στην λογική πρώτης τάξης και τα γηγενή αντικείμενα είναι σύμβολα χωρίς κάποια ιδιαίτερη σημασία. Η εκτέλεση ενός Prolog προγ/τος γίνεται με μια αναζήτηση σε μια συλλογή δυνατών λύσεων (χώρος αναζήτησης) που ορίζονται από τις προδιαγραφές του χρήστη. Η διαδικασία βασίζεται στην ενοποίηση συντακτικών οντοτήτων. Η αναζήτηση όλων των λύσεων μπορεί να οδηγήσει σε κάποια μονοπάτια αδιεξόδου. Αυτά τα

μονοπάτια εγκαταλείπονται μέσα από μια διαδικασία πισωγυρίσματος που κατευθύνει διαφορετικά την αναζήτηση μέσω διαφορετικών επιλογών. Αυτή η προσέγγιση γίνεται γρήγορα μη αποδοτική, εξαιτίας απαιτήσεων χρόνου και χώρου.

4.2 Προβλήματα Ικανοποίησης περιορισμών

Μια πολύ μεγάλη κατηγορία προβλημάτων της καθημερινότητας είναι τα λεγόμενα συνδυαστικά προβλήματα ή αλλιώς προβλήματα ικανοποίησης περιορισμών, όπως η σύνταξη του ωρολογίου προγράμματος για την κατανομή των ωρών εργασίας (timetabling), ο σχεδιασμός ενεργειών για την οργάνωση γραμμών παραγωγής (planning), κλπ. Ειδικότερα, ένα πρόβλημα ικανοποίησης περιορισμών ορίζεται ως ακολούθως:

Ένα πεπερασμένο I σύνολο μεταβλητών: $\{X_1, X_2, X_3,...X_K\}$
Αντίστοιχα πεπερασμένα πεδία ορισμού: $\{D_1, D_2, D_3,...D_K\}$
Ένα σύνολο περιορισμών: $\{C_1, C_2, C_3,...C_m\}$

Ένας περιορισμός C_i $\{C_{i1}, C_{i2}, C_{i3}, ... C_{iK}\}$ μεταξύ n μεταβλητών από το σύνολο I είναι το υποσύνολο του καρτεσιανού γινομένου D_{i1} x D_{i2} x...D_{in} το οποίο καθορίζει τις τιμές των μεταβλητών που είναι συμβατές μεταξύ τους. Ανάλογα με το πόσες μεταβλητές περιλαμβάνει ένας περιορισμός χαρακτηρίζεται ως:

- μοναδιαίος (unary) όταν περιλαμβάνει μια μεταβλητή, π.χ. $V_1>0$
- δυαδικός (binary) όταν περιλαμβάνει δύο μεταβλητές, π.χ. $V_1>V_2$
- ανώτερης τάξης (higher order) όταν περιλαμβάνει περισσότερες, π.χ. $V_1+V_2+V_3>0$.

Η επίλυση ενός προβλήματος περιορισμών συνίσταται στην ανάθεση τιμών στις μεταβλητές, έτσι ώστε να ικανοποιούνται οι περιορισμοί. Στην περίπτωση ύπαρξης συνάρτησης βελτιστοποίησης, επιδιώκεται ο προσδιορισμός της λύσης εντός του συνόλου των δυνατών συνδυασμών, η οποία να ικανοποιεί τους περιορισμούς και να βελτιστοποιεί την τιμή της συνάρτησης.

4.2.1 Αδυναμίες του Λογικού προγραμματισμού

Ο λογικός προγραμματισμός προσφέρει έναν εύκολο τρόπο περιγραφής διακριτών συνδυαστικών προβλημάτων, με χρήση του μη ντετερμινισμού αλλά και της σχεσιακής σχέσης του προγραμματιστικού περιβάλλοντος. Παρόλα αυτά, το μοναδικό είδος αποδεκτών περιορισμών, είναι οι ισότητες μεταξύ όρων συντακτικά ίδιων.

Έστω για παράδειγμα ο παρακάτω κανόνας:

- triangle(X,Y,Z) :- X>=0, Y>=0, Z>=0, X+Y>=Z, X+Z>=Y, Y+Z>=X.

Εάν υποβάλλουμε την ερώτηση:

- ?- triangle(3,4,5).

στην κλασσική Prolog θα πάρουμε την απάντηση:

- yes.

Αν όμως υποβάλλουμε την ερώτηση:

- ?- triangle(3,4,Z).

δεν θα πάρουμε καμία απάντηση (εκτός ίσως από κάποιο μήνυμα λάθους). Ο λόγος είναι ότι στην κλασσική Prolog οι αριθμητικές σχέσεις (ισότητες, ανισότητες) μπορούν να ελεγχθούν μόνο εφόσον οι μεταβλητές έχουν όλες πάρει τιμή.

Έτσι, η τεχνική που αναγκαστικά θα χρησιμοποιηθεί για την αναζήτηση λύσεων στο πρόβλημα θα είναι η τεχνική παραγωγή & έλεγχος (generate & test). Συγκεκριμένα, η βασική δομή ενός Prolog προγράμματος θα είναι η εξής:

find (X) ←
generate (X) ,
test (X) .

Το κατηγόρημα generate(X), τυπικά μη ντετερμινιστικό, είναι υπεύθυνο να επιστρέψει μια απαρίθμηση των λύσεων που είναι συχνά εκθετικές στον αριθμό. Κάθε φορά που το κατηγόρημα generate επιστρέφει μια λύση στο X, το κατηγόρημα test ντετερμινιστικά και αποδοτικά ελέγχει αν το X είναι

αποδεκτό με δεδομένους τους περιορισμούς του προβλήματος. Η στρατηγική είναι ορθή αλλά μη αποδοτική από τη στιγμή που κάθε καταχώρηση του χώρου αναζήτησης παράγεται πριν τον επακόλουθο έλεγχο.

Σαν αποτέλεσμα η αποδοτικότητα του προγράμματος είναι πολύ χαμηλή, ιδίως για τα μεγάλου μεγέθους συνδυαστικά προβλήματα, στα οποία υπάρχει συνδυαστική έκρηξη του χώρου αναζήτησης. Η πραγματική δυσκολία, προκειμένου για την αποδοτικότητα του προγράμματος, έγκειται στον περιορισμό του δέντρου αναζήτησης, ένα σημαντικό εργαλείο που δεν προσφέρει ο κλασσικός λογικός προγραμματισμός.

Η εμπορική επιτυχία της Prolog επιβραδύνθηκε από την μη αποδοτικότητά της σε σύγκριση με τις διαδικαστικές γλώσσες όπως η C και έτσι η χρήση της περιορίστηκε με τα χρόνια. Η τάση αυτή αντιστράφηκε με την εισαγωγή του χειρισμού των περιορισμών στις λογικές γλώσσες. Τα νέα αυτά χαρακτηριστικά επιτρέπουν την συγγραφή αποδοτικών προγραμμάτων με την χρήση του κλασσικού δηλωτικού στυλ.

4.3 Η διαφοροποίηση του Λογικού Προγραμματισμού με Διαχείριση Περιορισμών

Ο Λογικός Προγραμματισμός με Διαχείριση Περιορισμών αντιπροσωπεύει μία νέα κλάση περιγραφικών (declarative) προγραμματιστικών γλωσσών προγραμματισμού, η οποία συνδυάζει το μη-ντετερμινισμό με την επίλυση περιορισμών. Αιτία ανάπτυξής του είναι η αδυναμία του αμιγώς λογικού προγραμματισμού αναφορικά με την αποτελεσματική επίλυση μεγάλης κλίμακας συνδυαστικών προβλημάτων ή προβλημάτων βελτιστοποίησης.

Ο χειρισμός των περιορισμών και η διάδοση της πληροφορίας μέσω αυτών αποτέλεσαν αντικείμενο έρευνας στην κοινότητα της Τεχνητής Νοημοσύνης για πολλά συναπτά έτη. Προκειμένου για τη δηλωτική περιγραφή και την αποτελεσματική επίλυση των πολύπλοκων συνδυαστικών προβλημάτων αναπτύχθηκε μία νέα τεχνολογία λογισμικού, γνωστή ως λογικός

προγραμματισμός με διαχείριση περιορισμών (Constraint Logic Programming, CLP).

Η CLP τεχνολογία διατηρεί όλα τα πλεονεκτήματα του λογικού προγραμματισμού, όπως η δηλωτικότητα, ο μικρός χρόνος ανάπτυξης και η εύκολη συντήρηση, ενώ παρέχει επίσης μία λύση στο πρόβλημα αποδοτικότητας του κλασσικού λογικού προγραμματισμού. Ως αποτέλεσμα αυτού, οι εφαρμογές που αναπτύσσονται είναι συγκρίσιμες με τους αλγορίθμους που χρησιμοποιούνται στις επιτακτικές γλώσσες, όπως η C και η Pascal. Έτσι, οι χρήστες μπορούν να αναζητούν τη φυσική έκφραση του προβλήματος με την χρήση περιορισμών, χωρίς να αναλωθούν στην αναζήτηση του κατάλληλου αλγόριθμου επίλυσής του.

Η βασική διαφοροποίηση του CLP από τον κλασσικό Λογικό Προγραμματισμό, είναι η αντικατάσταση της διαδικασίας δοκιμής και σφάλματος (generate & test) και της διαδικασίας της ενοποίησης, από μεθόδους χειρισμού και επίλυσης περιορισμών, όπως είναι οι τεχνικές συνέπειας. Η μεθοδολογία των τεχνικών συνέπειας προσανατολίζεται στην αφαίρεση των «ασυμβίβαστων» τιμών από τα πεδία τιμών των μεταβλητών, προκειμένου να μειωθεί το διάστημα αναζήτησης. Οι μεταβλητές παριστάνονται ως κόμβοι γράφων, ενώ οι περιορισμοί ως τόξα μεταξύ των κόμβων, που διαμορφώνουν το καλούμενο «δίκτυο περιορισμού». Οι αλγόριθμοι συνέπειας ταξινομούνται σε τρεις βασικές κατηγορίες: τη συνέπεια κόμβων, τη συνέπεια τόξων και τη συνέπεια μονοπατιών, σύμφωνα με το βαθμό ελέγχων συνέπειας που εκτελούν. Ένα CLP πρόγραμμα έχει την ίδια δηλωτικά δομή με ένα Prolog πρόγραμμα και έτσι ένα έτοιμο Prolog πρόγραμμα μπορεί να μετατραπεί σε CLP πρόγραμμα με πολύ μικρή προσπάθεια.

Η πρώτη εξέλιξη του CLP σε σύγκριση με μια καθαρά αφηρημένη λογική δομή είναι η προσθήκη αντικειμένων με ειδική σημασιολογία στις διάφορες περιοχές εφαρμογής, δηλ. ακέραιους και πραγματικούς αριθμούς και τις αλγεβρικές πράξεις που ορίζονται ανάμεσά τους. Γι’ αυτό υπάρχει μια οικογένεια γλωσσών, μια για κάθε περιοχή εφαρμογής. Ένα CLP πρόγ/μα περιέχει

λογικές εκφράσεις (περιορισμούς) , π.χ. X > 0, Y + Z < 15, οι οποίοι θα πρέπει να ικανοποιούνται για τον σωστό τερματισμό του προγ/τος. Σε αυτό το πλαίσιο, ο αλγόριθμος της ενοποίησης (που είναι η καρδιά της Prolog) δεν μπορεί να γίνει εκμεταλλεύσιμος και συχνά επεκτείνεται ή/και αντικαθίσταται από έναν επιλυτή περιορισμών στη συγκεκριμένη περιοχή εφαρμογής. Ο επιλυτής είναι υπεύθυνος να ελέγχει αν οι περιορισμοί είναι ικανοποιήσιμοι. Για λόγους απόδοσης, οι επιλυτές είναι αυξητικοί με την έννοια ότι με δεδομένο ένα επιλυμένο σύνολο περιορισμών είναι δυνατόν να προσθέσουμε έναν ακόμα περιορισμό χωρίς να απαιτείται ο επανυπολογισμός όλου του συνόλου περιορισμών. Πολλές τεχνικές επίλυσης περιορισμών είναι δανεισμένοι από άλλες περιοχές της επιστήμης Η/Υ, π.χ. επιχειρησιακή έρευνα, μέθοδος simplex για γραμμικούς περιορισμούς σε πραγματικούς αριθμούς, κλπ.

Οι νεώτερες εκδόσεις των πιο γνωστών υλοποιήσεων της Prolog έχουν επεκτείνει τη γλώσσα με τη δυνατότητα να χειρίζεται μεταβλητές με πεδίο ορισμού και περιορισμούς μεταξύ των μεταβλητών (CLP βιβλιοθήκες). Τα διάφορα συστήματα διαφέρουν συνήθως στο είδος των περιορισμών που υποστηρίζουν (π.χ. γραμμικές ανισότητες, ανισότητα μεταξύ 2 ή πολλών μεταβλητών κλπ).

Όταν μετακινούμαστε από μια καθαρά λογική γλώσσα σε μια CLP γλώσσα, υπάρχει μια νοητική μετάθεση της κύριας μεθοδολογίας για λύση του προβλήματος: από παραγωγή και έλεγχο (generate and test) σε περιορισμοί και έλεγχος (constrain and generate). Η ιδέα είναι να περιορίζουμε αρχικά τις μεταβλητές με την καταχώρηση του πεδίου ορισμού τους και τον καθορισμό των περιορισμών ανάμεσά τους. Σε μια δεύτερη φάση, μόνο οι αποδεκτές λύσεις παράγονται με την βασική ιδέα καθυστέρησης όσο είναι δυνατόν των καταχωρήσεων στις μεταβλητές.

Η γενική μορφή ενός CLP προγράμματος για την επίλυση προβλημάτων ικανοποίησης περιορισμών είναι η εξής:

```
solve(Variables):-
        read_data(Data),
        setup_constraints(Data, Variables),
        labeling(Variables).
```

όπου, το κατηγόρημα read_data/1 ορίζεται από εμάς και διαβάζει κάποια δεδομένα σχετικά με το πρόβλημα.
Το κατηγόρημα setup_constraints/2 ορίζεται από εμάς και ορίζει τους περιορισμούς επί των μεταβλητών, βάσει των δεδομένων που διαβάσαμε. Το κατηγόρημα labeling/1 είναι ενσωματωμένο στο CLP σύστημα και λύνει το πρόβλημα, δηλ. εξερευνά το χώρο αναζήτησης με τις τεχνικές συνέπειας των περιορισμών.

4.4 Συγκρίνοντας Λογικό Προγ/σμό με τον Λογικό Προγ/σμό με περιορισμούς

Ο Herbert A. Simon, ένας επιστήμονας που μελέτησε επιλύσεις προβλημάτων με την χρήση της τεχνολογίας CLP, χρησιμοποίησε τα δημοφιλή προβλήματα κρυπταριθμητικής για να δείξει τα πλεονεκτήματα διαχείρισης των περιορισμών στον λογικό προγραμματισμό. Ως ενδεικτικό παράδειγμα πρότεινε το πρόβλημα εύρεσης μιας ένα-προς-ένα αντιστοίχισης μεταξύ των γραμμάτων S,E,N,D,M,O,R,Y και των αριθμών 0,1,2,...,9, έτσι ώστε να ισχύει η παραπάτω ισότητα:

SEND +
MORE =
MONEY

Οι μεταβλητές του προβλήματος είναι τα γράμματα S,E,N,D,M,O,R,Y. Τα πεδία τιμών τους είναι το σύνολο ακεραίων 0,1,...,9. Οι περιορισμοί είναι ότι οι μεταβλητές πρέπει να πάρουν διαφορετικές τιμές μεταξύ τους και ότι πρέπει να ισχύει η παραπάνω εξίσωση. Το πρόβλημα έχει 108 δυνατές

καταχωρήσεις (η συνδυαστική έκρηξη των 8 μεταβλητών με 10 καταχωρήσεις η καθεμία).

Έγινε ένας πειραματικός έλεγχος για να συγκριθούν οι αποδόσεις του προβλήματος όταν κωδικοποιηθεί στην βασική Prolog και σε μια CLP γλώσσα. Στον **Πίνακας 4.1** παρουσιάζεται ο κώδικας Prolog για την υλοποίηση του 'SEND MORE MONEY' προβλήματος, ενώ στον **Πίνακας 4.2** παρουσιάζεται ο ισοδύναμος κώδικας στο CLP σύστημα CHIP.

Ο κώδικας Prolog χρησιμοποιεί την generate and test προσέγγιση για την επίλυση του προβλήματος. Ειδικότερα, το κατηγόρημα generate είναι μη ντετερμινιστικό και χρησιμοποιεί το ενσωματωμένο κατηγόρημα member για να απαριθμήσει τις δυνατές καταχωρήσεις από μια λίστα (δηλ., από 0 έως 9) σε κάθε μεταβλητή. Το κατηγόρημα test υλοποιεί τον έλεγχο των περιορισμών πάνω στις τιμές των μεταβλητών. Το κατηγόρημα different υλοποιεί το σύνολο των τετραγωνικών ελέγχων για να εξασφαλίσει ότι κάθε ζευγάρι μεταβλητών έχει διακριτή τιμή από τις άλλες μεταβλητές.

Ο CLP κώδικας, γραμμένος με την constrain and generate προσέγγιση περιέχει τα ίδια δομικά στοιχεία. Η φάση των περιορισμών τις ιδιότητες που πρέπει να ικανοποιούν οι λύσεις (χρήση του τελεστή #= και χρήση του ενσωματωμένου κατηγορήματος all_different). Το κατηγόρημα labeling ξεκινά την εξερεύνηση του δεντρου αναζήτησης με τις τεχνικές συνέπειας περιορισμών.

Οι αποδόσεις αυτών των δύο προσεγγίσεων είναι ως εξής. Η υλοποίηση της Prolog πήρε 143 δευτερόλεπτα για να καταλήξει την αναζήτηση (εκτελέστηκε μια εξαντλητική αναζήτηση), ενώ ο CLP κώδικας πήρε λιγότερο από 0.01 δευτερόλεπτα. Αυτό είναι ένα ξεκάθαρο παράδειγμα της δύναμης της διάδοσης περιορισμών που είναι ενσωματωμένη στην CLP τεχνολογία.

```
lp([S,E,N,D,M,O,R,Y]) :-
  generate([S,E,N,D,M,O,R,Y]),
  test([S,E,N,D,M,O,R,Y]).

generate([]).
generate([A|R]):-
  member(A,[0,1,2,3,4,5,6,7,8,9]),
  generate(R).

test([S,E,N,D,M,O,R,Y]) :-
  S>0, M>0,
  different([S,E,N,D,M,O,R,Y]),
  1000*S + 100*E + 10*N + D + 1000*M + 100*O + 10*R + E
  =:= 10000*M + 1000*O + 100*N + 10*E + Y.

different([_]).
different([A,B|R]):-
  different1(A,[B|R]),
  different([B|R]).

different1(_,[]).
different1(A,[B|R]):-
  A =\= B,
  different1(A,R).
```

Πίνακας 4.1 Prolog κώδικας για το 'SEND MORE MONEY' πρόβλημα [50]

```
clp([S,E,N,D,M,O,R,Y]) :-
  constrain([S,E,N,D,M,O,R,Y]),
  labeling([],[S,E,N,D,M,O,R,Y]).

constrain([S,E,N,D,M,O,R,Y]):-
  domain([S,E,N,D,M,O,R,Y], 0, 9),
  S#>0, M#>0,
  all_different([S,E,N,D,M,O,R,Y]),
  1000*S + 100*E + 10*N + D + 1000*M + 100*O + 10*R + E
  #= 10000*M + 1000*O + 100*N + 10*E + Y.
```

Πίνακας 4.2 CLP κώδικας για το 'SEND MORE MONEY' πρόβλημα [50]

4.5 Αλγόριθμοι ελέγχου συνέπειας

Όπως έχει ήδη αναφερθεί, η μέθοδος της δοκιμής και σφάλματος (generate and test) και η μέθοδος της κλασσικής οπισθοδρόμησης δεν είναι αποδοτικές, καθώς οι περιορισμοί στην περίπτωση αυτή χρησιμοποιούνται για να μειώσουν εκ των υστέρων το δέντρο αναζήτησης. Η φράση «εκ των υστέρων» υποδηλώνει ότι έχει προηγηθεί η αποτυχία, και επομένως έχει ήδη σπαταληθεί αρκετός προγραμματιστικός χρόνος.

Η βασική ιδέα των αλγορίθμων ελέγχου συνέπειας είναι η όσο το δυνατόν μεγαλύτερη μείωση του χώρου αναζήτησης πριν την ανάθεση τιμών.
Η γενική μορφή αυτού του αλγορίθμου είναι η εξής:

- Για κάθε περιορισμό αφαίρεσε από τα πεδία τιμών των μεταβλητών τις τιμές εκείνες που δεν μπορούν να συμμετέχουν στην τελική λύση.
- Στο μειωμένο χώρο αναζήτησης που προκύπτει από το προηγούμενο βήμα εφάρμοσε έναν κλασσικό αλγόριθμο αναζήτησης για να βρεθεί η λύση (π.χ. πρώτα κατά βάθος).
- Σε κάθε βήμα (ανάθεση τιμής) αυτής της αναζήτησης εφάρμοσε ξανά τον αλγόριθμο ελέγχου συνέπειας έτσι ώστε να αφαιρεθούν τυχόν τιμές από τα πεδία των μεταβλητών οι οποίες δεν μπορούν να συμμετέχουν στην λύση.

Παραδείγματα τεχνικών συνέπειας είναι οι τεχνικές «forward checking» και «looking ahead», οι οποίες εξασφαλίζουν εκ των προτέρων για την επόμενη μεταβλητή στην οποία δεν έχει αποδοθεί ακόμα τιμή, ότι υπάρχει μία τουλάχιστον τιμή για την οποία ικανοποιούνται, ως εκείνο το σημείο, οι περιορισμοί. Έτσι, απορρίπτονται συνδυασμοί που θα κόστιζαν σε προγραμματιστικό χρόνο, και δεν θα είχαν το επιθυμητό αποτέλεσμα. Για να γίνει αυτό, αφιερώνουν αρκετό χρόνο σε κάθε κόμβο προκειμένου να μειώσουν το συνολικό τελικό αριθμό των κόμβων που επεξεργάζονται. Εξάλλου, λόγω της διαδικασίας της διάδοσης (propagation), οι συνέπειες των

επιλογών που λαμβάνουν χώρα διαδίδονται, γεγονός που πιθανόν οδηγεί στην έγκαιρη πρόληψη άλλων αποτυχιών.

4.5.1 Παράδειγμα χρήσης αλγορίθμου συνέπειας

Για να περιγραφεί η χρήση ενός αλγορίθμου συνέπειας περιορισμών, ας πάρουμε ως παράδειγμα συνδυαστικού προβλήματος ένα πρόβλημα εκτέλεσης δραστηριοτήτων με προτεραιότητες. Συγκεκριμένα, πρέπει να ορισθεί η σειρά με την οποία θα εκτελεστούν τέσσερις εργασίες Α, Β, Γ, Δ. Λόγω όμως της φύσης του προβλήματος, η εργασία Α πρέπει να εκτελεστεί μετά από την Δ, η Γ πριν από την Β, και η Β πριν από την Α. Μια μοντελοποίηση του προβλήματος είναι η εξής:

- Μεταβλητές: Α, Β, Γ, Δ
- Πεδία τιμών: DA=DB=DΓ=DΔ = {1,2,3,4}
- Περιορισμοί (όπως φαίνονται στον **Πίνακας 4.3**)

Α ≠ Β (C1)	**Β ≠ Γ (C4)**	**Α > Δ (C7)**
Α ≠ Γ (C2)	**Β ≠ Δ (C5)**	**Γ < Β (C8)**
Α ≠ Δ (C3)	**Γ ≠ Δ (C6)**	**Β < Α (C9)**

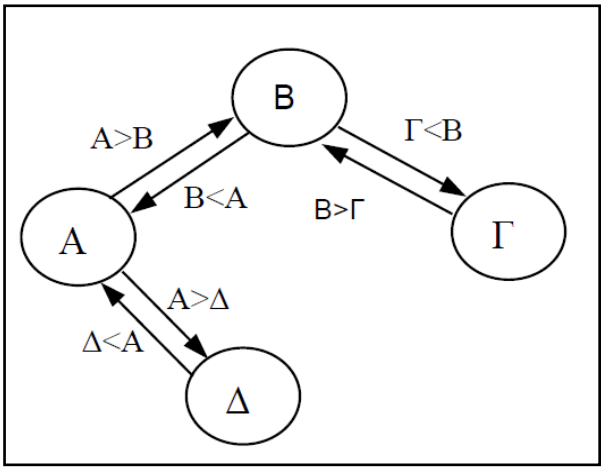
Πίνακας 4.3 Πίνακας περιορισμών για το πρόβλημα καθορισμού της σειράς εργασιών [51]

Το πρόβλημα έχει τρεις λύσεις, τις ακόλουθες:
Α=4, Β=2, Γ=1, Δ=3, δηλαδή η σειρά είναι: ΓΒΔΑ
Α=4, Β=3, Γ=1, Δ=2, δηλαδή η σειρά είναι:ΓΔΒΑ
Α=4, Β=3, Γ=2, Δ=1, δηλαδή η σειρά είναι: ΔΓΒΑ

Το ερώτημα που τίθεται είναι πώς φθάνει κανείς στις παραπάνω λύσεις, με την χρήση των αλγορίθμων ελέγχου συνέπειας. Γενικά, οι περιορισμοί

αναπαρίστανται με έναν γράφο περιορισμών. Ενδεικτικά, στην **Εικόνα 4.1** φαίνεται ο γράφος περιορισμών για τους περιορισμούς

- A > Δ
- Γ < B
- B < A



Εικόνα 4.1 Γράφος περιορισμών [51]

Τα αρχικά πεδία τιμών των μεταβλητών είναι:

- A ∈ {1,2,3,4}
- B ∈ {1,2,3,4}
- Γ ∈ {1,2,3,4}
- Δ ∈ {1,2,3,4}

Λόγω του ότι B < A (C9), η μεταβλητή B δε μπορεί σε καμιά περίπτωση να πάρει την τιμή 4, αλλά ούτε και η A να πάρει την τιμή 1, άρα:

- A ∈ {2,3,4}
- B ∈ {1,2,3}
- Γ ∈ {1,2,3,4}
- Δ ∈ {1,2,3,4}

Λόγω του ότι Γ < B (C8), η Γ δεν μπορεί να πάρει την τιμή 3 ούτε και την τιμή 4, ενώ η B δε μπορεί να πάρει την τιμή 1, άρα:

- A ∈ {2,3,4}
- B ∈ {2,3}
- Γ ∈ {1,2}

- Δ ∈ {1,2,3,4}

Λόγω του ότι A > Δ (C7) η Δ δεν μπορεί να πάρει την τιμή 4, άρα:

- A ∈ {2,3,4}
- B ∈ {2,3}
- Γ ∈ {1,2}
- Δ ∈ {1,2,3}

Το πεδίο της B έχει μεταβληθεί, οπότε ο περιορισμός C9 πρέπει να επανεξεταστεί. Λόγω του ότι B < A (C9) δεν μπορεί να υπάρχει η τιμή 2 στο πεδίο της A, άρα:

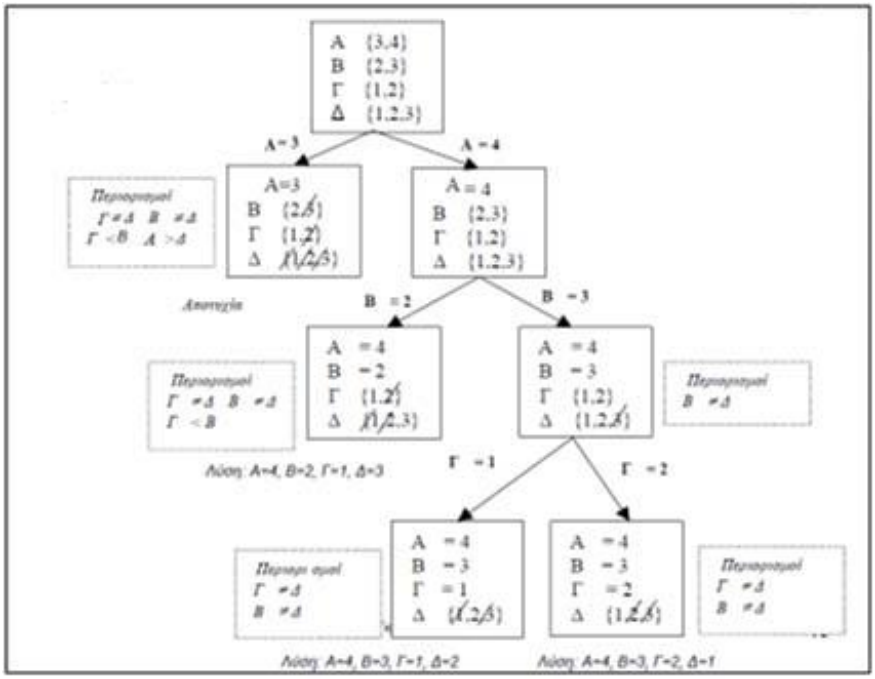
- A ∈ {3,4}
- B ∈ {2,3}
- Γ ∈ {1,2}
- Δ ∈ {1,2,3}

Έτσι, μετά από την παραπάνω διαδικασία οι πιθανοί συνδυασμοί γίνονται $2\times2\times2\times3 = 24$, σε σχέση με τους 256 που υπήρχαν αρχικά. Στο σημείο αυτό τα πεδία τιμών των μεταβλητών δεν μπορούν να περιοριστούν παραπέρα. ‘Ετσι, επιλέγεται στην τύχη μια μεταβλητή και της αποδίδεται αυθαίρετα μία από τις τιμές της. Αυτό προκαλεί επιπλέον περιορισμούς στα πεδία των υπολοίπων μεταβλητών.

Η παραπάνω διαδικασία επαναλαμβάνεται μέχρι:

- Όλες οι μεταβλητές να πάρουν μία τιμή.
- Κάποια μεταβλητή να μείνει χωρίς τιμές
- Το σημείο αυτό αποτελεί σημείο υπαναχώρησης

Όλη η πορεία του υπολογισμού του προβλήματος δείχνεται στην παρακάτω **Εικόνα 4.2.**



Εικόνα 4.2 Πορεία υπολογισμού του προβλήματος διάταξης εργασιών με προτεραιότητες [51]

Ο γράφος αναζήτησης διασχίζεται συνήθως πρώτα-κατά-βάθος. Υπάρχουν βέβαια ευρετικές τεχνικές για αύξηση της απόδοσης, όπως π.χ.:

- Να επιλέξουμε να δώσουμε τιμή πρώτα στις μεταβλητές με τις λιγότερες τιμές.
- Εναλλακτικά, να επιλέξουμε να δώσουμε τιμή πρώτα στις μεταβλητές εκείνες που συμμετέχουν στους περισσότερους περιορισμούς.
- Αφού επιλέξουμε μια μεταβλητή, στη συνέχεια επιλέγουμε να της αναθέσουμε πρώτα την τιμή εκείνη που προκαλεί τη μικρότερη απόρριψη τιμών από τα πεδία των υπολοίπων μεταβλητών.

4.6 Δημοφιλή CLP συστήματα

Από τα πιο δημοφιλή CLP συστήματα είναι το σύστημα CHIP (Constraint Handling In Prolog). Αναπτύχθηκε στο Ευρωπαϊκό Κέντρο Ερευνών (European Computer-Industry Research Centre, ECRC) και έχει

χρησιμοποιηθεί από πολλούς ερευνητές. Η CHIP είναι μια γλώσσα λογικού προγραμματισμού έντονα επηρεασμένη από τη θεωρία των τεχνικών συνέπειας. Η εφαρμογή τους στον λογικό προγραμματισμό προτάθηκε από τον Galliard με σκοπό τη βελτίωση της συμπεριφοράς κατά τη διαδικασία αναζήτησης. Πιο συγκεκριμένα, ο ερευνητής πρότεινε τη χρήση των περιορισμών σε ένα ενεργό ρόλο, με στόχο την εκ των προτέρων μείωση του δέντρου αναζήτησης. Αξίζει να σημειωθεί, ότι η ιδέα αυτή αντιβαίνει στην κλασσική οπισθοδρόμηση και τη μεθοδολογία δοκιμής και σφάλματος που χρησιμοποιούνται από τον κλασσικό λογικό προγραμματισμό, και έχουν ως αρχή τον εκ των υστέρων έλεγχο των περιορισμών. Βασική ιδέα είναι πλέον η αποτελεσματική ολοκλήρωση και επικοινωνία της ντετερμινιστικής διαδικασίας εκτίμησης και εφαρμογής περιορισμών και της μη-ντετερμινιστικής διαδικασίας αναζήτησης.

Επιπρόσθετα, καθώς τα περισσότερα συνδυαστικά προβλήματα περιλαμβάνουν μία συνάρτηση βελτιστοποίησης, η λύση της συνάρτησης αυτής γίνεται στην CHIP με τη βοήθεια κατηγορημάτων. Για παράδειγμα, και αναφορικά με μια συνάρτηση ελαχιστοποίησης, όταν βρεθεί μια λύση με τιμή α, ο περιορισμός (τιμή συνάρτησης < α) προστίθεται δυναμικά στο πρόγραμμα. Καθώς συνεχίζεται η αναζήτηση μίας καινούργιας λύσης, όλα τα τμήματα του διαστήματος έρευνας, για τα οποία η τιμή της συνάρτησης είναι μεγαλύτερη του α, απορρίπτονται. Η βέλτιστη λύση θεωρείται ότι έχει επιτευχθεί όταν δεν προκύπτει μία καινούργια, καλύτερη από την πιο πρόσφατη.

Κάθε CLP σύστημα περιλαμβάνει διαφορετικούς επιλυτές περιορισμών για διαφορετικά υπολογιστικά πεδία. Δηλ. παρέχει βιβλιοθήκες με επιλυτές περιορισμών, ούτως ώστε να μπορεί να επιλεγεί ο κατάλληλος, ανάλογα με τη φύση του εκάστοτε προβλήματος.

Ένα άλλο επίσης δημοφιλές και ελεύθερο CLP σύστημα είναι το ECLIPSE [41], το οποίο υποστηρίζει πολλές διαφορετικές περιπτώσεις περιορισμών, όπως:

- την βιβλιοθήκη ic (interval constraints) που υποστηρίζει περιορισμούς επί ακεραίων και πραγματικών αριθμών, των οποίων τα πεδία ορισμού είναι διαστήματα, περιορισμούς με περισσότερες από δύο μεταβλητές,
- ειδικούς περιορισμούς με πολλές μεταβλητές για προβλήματα χρονοπρογραμματισμού,
- βιβλιοθήκη για περιορισμούς συνόλων ic_sets,
- περιορισμούς σχετικούς με πεπερασμένα σύνολα ακεραίων, κλπ.

Η εύρεση λύσεων στο σύστημα ECLIPSE γίνεται με δύο κατηγορήματα που ξεκινούν τη διαδικασία εύρεσης ανάθεσης τιμών στις μεταβλητές:

- indomain(X): Η μεταβλητή X είναι μια μεταβλητή για την οποία έχουν ορισθεί κάποιοι περιορισμοί. Η κλήση indomain(X) αναθέτει στην X μια από τις τιμές που αυτή μπορεί να πάρει, ενώ ταυτόχρονα ενημερώνει όλες τις μεταβλητές που συμμετέχουν σε περιορισμούς με την X.
- labeling(List): Η λίστα List περιλαμβάνει ένα σύνολο μεταβλητών για τις οποίες έχουν ορισθεί κάποιοι περιορισμοί. Η κλήση στην labeling επιχειρεί να βρει μια ανάθεση τιμών στις μεταβλητές της List, η οποία να είναι συμβατή με όλους τους περιορισμούς όπου συμμετέχουν οι μεταβλητές.

Συνήθως καλούμε την labeling με όλες τις μεταβλητές του προβλήματος.

5 Υλοποίηση σε CLP(FD)

5.1 Εισαγωγή

Οι πρωτεΐνες είναι υπεύθυνες για κάθε σχεδόν λειτουργία ενός οργανισμού. Η ακολουθία των στοιχείων (αμινοξέων) που αποτελούν μια πρωτεΐνη ονομάζεται πρωτοταγής (1D) δομή της πρωτεΐνης. Μια λειτουργική πρωτεΐνη είναι μια διπλωμένη ακολουθία των αμινοξέων στον τρισδιάστατο (3D) χώρο. Αρκετά πειράματα έχουν αποδείξει ότι η συνέπεια μιας προσπάθειας αναδιάταξης μιας πρωτεΐνης στον 3D χώρο είναι η πρωτεΐνη αυτή να επιστρέφει άμεσα στην φυσική της διαμόρφωση [1]. Η 3D δομή της πρωτεΐνης χαρακτηρίζει την λειτουργία της. Η γνώση της 3D δομής μιας πρωτεΐνης είναι ζωτικής σημασίας σε πολλές βιοιατρικές εφαρμογές, π.χ. για σχεδιασμό κατάλληλων φαρμάκων για την βελτίωση της λειτουργικότητάς της.

Το Πρόβλημα Αναδίπλωσης Πρωτεΐνης (ΠΑΠ) είναι το πρόβλημα πρόβλεψης της τρισδιάστατης δομής μιας πρωτεΐνης όταν η γραμμική ακολουθία των αμινοξέων που την αποτελούν (δηλαδή η χημική της δομή) είναι γνωστή. Τις τελευταίες δεκαετίες οι περισσότεροι επιστήμονες έχουν καταλήξει ότι η φυσική χωροθέτηση μιας πρωτεΐνης έχει την ελάχιστη ελεύθερη ενέργεια. Ως εκ τούτου, μια εξαρχής (de novo) στρατηγική επίλυσης του ΠΑΠ είναι να αναχθεί σε ένα πρόβλημα εύρεσης μιας ακολουθίας σημείων στον 3D χώρο που ελαχιστοποιεί την τιμή της συνολικής ενέργειας της πρωτεΐνης, η οποία εξαρτάται από το 3D σχήμα της πρωτεΐνης. Πρόκειται δηλ. για ένα πρόβλημα βελτιστοποίησης με αρκετούς περιορισμούς που πηγάζουν από γνωστές φυσικές και χημικές ιδιότητες των αμινοξέων. Συνήθως, η τιμή της ενέργειας ορίζεται μέσω μιας συνάρτησης της οποίας οι τιμές εξαρτώνται από την απόσταση και τον τύπο ανάμεσα σε οποιοδήποτε ζεύγος αμινοξέων [17].

Η έκδοση του προβλήματος σαν πρόβλημα απόφασης (δηλ., αν υπάρχει μία αναδίπλωση της ακολουθίας με ενέργεια μικρότερη από μια συγκεκριμένη τιμή) έχει αποδειχθεί ότι είναι NP-πλήρης [18, 19]. Ακόμα κι αν η συνάρτηση της ενέργειας μπορεί να υπολογιστεί σε πολυωνυμικό χρόνο, η έκδοση απόφασης του ΠΑΠ παραμένει NP-πλήρης. Παρόλο που απαιτεί πολύ χρόνο για την επίλυσή του, το ΠΑΠ είναι τόσο κρίσιμης σημασίας που αξίζει να επιλυθεί έστω και προσεγγιστικά. Είναι ένα πρόβλημα κρίσιμης σπουδαιότητας για την Βιολογία και την Βιοτεχνολογία και το γεγονός ότι το τυπικό μήκος μιας πρωτεΐνης δεν είναι τεράστιο ενθαρρύνει τις προσπάθειες επίλυσής του.

Το πρόβλημα είναι εξαιρετικά πολύπλοκο και γι' αυτό πρέπει να απλοποιηθεί χωρίς βέβαια να χαθεί η βιολογική του διάσταση. Μια κοινή απλοποίηση είναι η χρήση ενός διάκριτου χωρικού μοντέλου πλέγματος για τις θέσεις των αμινοξέων στο χώρο [20, 21, 22]. Στην εργασία χρησιμοποιείται ένα τοπολογικό μοντέλο που βασίζεται σε πλέγμα σχήματος κύβου μήκους 2 που συμπεριλαμβάνει στα σημεία του και το κέντρο κάθε επιφάνειας του κύβου.

Οποιαδήποτε γνώση σχετικά με τον μηχανισμό αναδίπλωσης μιας πρωτεΐνης σαφώς βελτιώνει κάθε προηγούμενη προσπάθεια επίλυσης του προβλήματος. Ένα από τα συστατικά αυτού του μηχανισμού που πλέον γνωρίζουμε είναι το γεγονός ότι μερικές ειδικές διαμορφώσεις, όπως α-έλικες και β-πτυχωτές επιφάνειες είναι αναγνωρίσιμες στις περισσότερες πρωτεΐνες πριν την προσπάθεια εύρεσης της 3D δομής τους. Αυτά τα δομικά στοιχεία καλούνται Δευτεροταγής Δομή της πρωτεΐνης. Η μεγάλη ακρίβεια που έχει επιτευχθεί στις προβλέψεις δευτερεύουσας δομής [23] επιτρέπει την συμπερίληψη αυτής της πρόβλεψης στο κύριο πρόβλημα πρόβλεψης της τριτοταγούς δομής της πρωτεΐνης. Στην εργασία διερευνάται πόσο αποτελεσματική είναι σε όρους επίδοσης και ακρίβειας, η εισαγωγή περιορισμών που αφορούν την δευτεροταγή δομή της πρωτεΐνης.

Χρησιμοποιούμε την τεχνολογία Constraint Logic Programming (CLP), ειδικότερα την τεχνολογία Constraint Logic Programming σε Finite Domains

(CLP(FD)), για να μοντελοποιήσουμε και να λύσουμε αποδοτικά το ΠΑΠ. Η επιλογή του CLP(FD) είναι φυσική για τους εξής λόγους, σύμφωνα με το [14]:

- Το CLP(FD) είναι ένα παράδειγμα προγραμματισμού κατάλληλο για προβλήματα βελτιστοποίησης.
- Παρέχει δηλωτική και υψηλού επιπέδου μοντελοποίηση, σε συνδυασμό με αποδοτικές στρατηγικές επίλυσης και αναζήτησης.
- Μας επιτρέπει να προσθέσουμε εύκολα νέους περιορισμούς που προκύπτουν ως νέα γνώση καθώς και διαφορετικές στρατηγικές αναζήτησης.

Τα βασικά ερευνητικά ερωτήματα που πραγματεύεται η εργασία είναι τα εξής δύο:

1. Πόσο αποτελεσματική σε όρους απόδοσης και ακρίβειας είναι η χρήση της CLP(FD) για το ΠΑΠ;
2. Μπορεί ένας CLP(FD) αλγόριθμος να εφαρμοστεί για μεγάλα ΠΑΠ, δηλ. να κλιμακωθεί για μεγάλες πρωτεΐνες;

Μια πρώτη προσέγγιση στην χρήση του CLP(FD) για το ΠΑΠ έχει παρουσιαστεί στο [24]. Σε αυτήν την εργασία επεξεργαζόμαστε αυτή την πρόταση με σκοπό να βελτιώσουμε την απόδοση και την κλιμάκωση. Ο τελικός σκοπός της εργασίας είναι να αποδείξουμε ότι:

- οι δυνατότητες μοντελοποίησης και βελτιστοποίησης του CLP(FD) είναι πολύ καλές για τον χειρισμό του ΠΑΠ; είναι δυνατό να μοντελοποιήσουμε εύκολα τις δευτερεύουσες δομές μιας πρωτεΐνης, γεγονός που δεν είναι εύκολο να γίνει με άλλες υλοποιήσεις.
- οι δυνατότητες μοντελοποίησης υψηλού επιπέδου μας επιτρέπουν να προσθέσουμε ή να τροποποιήσουμε εύκολα περιορισμούς όταν αυτοί γίνουν διαθέσιμοι, καθώς επίσης να εξερευνήσουμε την χρήση διαφορετικών στρατηγικών αναζήτησης;
- αποδοτικότητα και κλιμάκωση μπορούν να επιτευχθούν για πραγματικά προβλήματα.

Το πρώτο βήμα σε αυτήν την διαδικασία είναι η επαναμοντελοποίηση του προβλήματος όπως προτείνεται στο [24] με στόχο την καλύτερη επίδοση. Εισάγουμε μια νέα δομή περιορισμών, ιδιαίτερα για την περιγραφή της δευτερεύουσας δομής της πρωτεΐνης. Έπειτα ελέγχεται η αποτελεσματικότητα χρήσης της πληροφορίας αυτής στον αλγόριθμο. Τέλος, αυτοί οι επιπρόσθετοι περιορισμοί συνδυάζονται με κάποιες ευρετικές μεθόδους για καθοδήγηση της εξερεύνησης του χώρου αναζήτησης, οδηγώντας σε μια πιο αποδοτική περικοπή.

Χρησιμοποιώντας την νέα δομή περιορισμών και τα νέα heuristics, πετυχαίνουμε μια επιτάχυνση της τάξης 50% σε σχέση με την επίδοση του κώδικα που παρουσιάζεται στο [24] και βελτιωμένη κλιμάκωση. Μελετάται η αποτελεσματικότητα της μεθόδου σε κάποιες μικρομεσαίες πρωτεΐνες των οποίων η δομή είναι γνωστή, καταλήγοντας σε επιτυχή εφαρμογή του αλγορίθμου σε πρωτεΐνες μήκους ως και 100 αμινοξέα.

5.2 Σχετικές εργασίες

Η βιβλιογραφία επάνω στο ΠΑΠ είναι εκτεταμένη [25, 26]. Το ΠΑΠ έχει αναγνωριστεί σαν μια μεγάλη πρόκληση [27], και έχει αντιμετωπιστεί με μια ποικιλία προσεγγίσεων (π.χ., συγκριτική μοντελοποίηση μέσω ομολογίας, νημάτωση, εξαρχής προσέγγιση).

Όσον αφορά την κοινότητα των περιορισμών, απλοποιημένες εκδόσεις του ΠΑΠ, βασισμένες στο HP μοντέλο, έχουν μελετηθεί και επιτυχώς λυθεί για πρωτεΐνες μεγέθους 30-40 στο [28, 29, 30]. Το μοντέλο HP διαχωρίζει τα αμινοξέα σε δύο κατηγορίες (H-υδροφοβικά, και P-υδροφιλικά) και το πρόβλημα ανάγεται σε ένα πρόβλημα εύρεσης εκείνης της διαμόρφωσης που μεγιστοποιεί τις επαφές των H αμινοξέων. Το πρόβλημα έχει μελετηθεί σε 2D τετραγωνικά πλέγματα [19, 31], 2D τριγωνικά πλέγματα [20], 3D τετραγωνικά μοντέλα [32] και προσωποκεντρικά κυβικά πλέγματα [33]. Στο [34] το σχετικό

πρόβλημα εύρεσης της 3D δομής για την παράπλευρη αλυσίδα μελετάται επίσης μέσα στο CLP πλαίσιο.

Οι Backofen και Will έχουν εκτεταμένα μελετήσει την HP μοντελοποίηση του προβλήματος σε προσωποκεντρικά κυβικά πλέγματα ([28, 29, 30]. Η προσέγγιση αυτή είναι κατάλληλη για καθολικές πρωτεΐνες, επειδή η κύρια δύναμη που καθοδηγεί την διαδικασία αναδίπλωσης είναι η ηλεκτρική ενέργεια που παράγεται από τα H και P αμινοξέα και τα προσωποκεντρικά κυβικά πλέγματα είναι μια από τις καλύτερες και απλούστερες προσεγγίσεις του 3D χώρου. Η προσέγγιση των Backofen και Will είναι πολύ ενδιαφέρουσα, αλλά ο μεγάλος βαθμός αφαιρετικότητας του μοντέλου τους δεν δίνει πάντα ακριβή αποτελέσματα. Ειδικότερα, οι τοπικές διαμορφώσεις της μορφής έλικας και πτυχώσεων συχνά δεν ανευρίσκονται. Σε σύγκριση με την εργασία των Backofen και Will, η προσέγγιση της παρούσας εργασίας τροποποιεί το μοντέλο της ενεργειακής συνεισφοράς, επεκτείνοντας τις αλληλεπιδράσεις ανάμεσα στις κατηγορίες H και P αμινοξέων σε αλληλεπιδράσεις ανάμεσα σε κάθε ζεύγος αμινοξέων σε κοντινή απόσταση.

Γενικότερα, η χρήση της τεχνολογίας περιορισμών στο πλαίσιο του ΠΑΠ είναι αρκετά περιορισμένη. Στην αναφορά [35], η Prolog χρησιμοποιείται για την υλοποίηση ευρετικών μεθόδων για την περικοπή του χώρου αναζήτησης για την πρόβλεψη α-ελίκων και β-πτυχωτών τοπολογιών. Στην αναφορά [36], κατανεμημένη αναζήτηση και συνεχής βελτιστοποίηση χρησιμοποιούνται για την εξαρχής πρόβλεψη της δομής μιας πρωτεΐνης.

5.3 Προτεινόμενη χωροθέτηση - Μοντέλα κρυσταλλικού πλέγματος

Αρκετά μοντέλα έχουν προταθεί για τις 3D ιδιότητες των πρωτεϊνών, δηλ. για τις αποδεκτές θέσεις στο χώρο για κάθε αμινοξύ. Η θέση ενός σημείου αναπαριστά την θέση ενός αμινοξέως και είναι μια τριάδα τιμών {x, y, z} (3D μοντέλα). Οι μεταβλητές x, y και z μπορεί να είναι πραγματικοί αριθμοί, οπότε οι πρωτεΐνες μπορεί να έχουν οποιαδήποτε θέση στο χώρο, ή ακέραιοι αριθμοί οπότε τα αμινοξέα μπορούν να πάρουν ένα πεπερασμένο αριθμό θέσεων. Είναι προφανές ότι στην πρώτη περίπτωση προκύπτουν δύσκολα διαχειρίσιμα υπολογιστικά μοντέλα, για το λόγο αυτό προτιμούμε την δεύτερη επιλογή, δηλ. διάκριτα τοπολογικά μοντέλα.

Τα μοντέλα κρυσταλλικού πλέγματος χρησιμοποιούνται από παλιά για την πρόβλεψη της δομής μιας πρωτεΐνης. Οι δομές πλέγματος έχουν υιοθετηθεί σε διάφορα επιστημονικά πεδία, παρέχοντας μια διαχειρίσιμη διακριτοποίηση του 3D χώρου. Στο χώρο της βιολογίας διευκολύνουν ιδιαίτερα την διερεύνηση της φυσικοχημικής οργάνωσης των μοριακών δομών. Ειδικότερα, τα τελευταία χρόνια οι δομές πλέγματος έχουν ιδιαίτερο ενδιαφέρον για την μελέτη του ΠΑΠ [21]. Ένα δυνατό μοντέλο είναι ένα πλέγμα του οποίου τα σημεία είναι ένα υποσύνολο του N^3. Όμως, σ αυτήν την περίπτωση οι μόνες επιτρεπτές γωνίες ανάμεσα σε τρία συνεχόμενα αμινοξέα είναι 90◦ και 180◦. Αυτό το πλέγμα είναι πολύ περιοριστικό για μια ρεαλιστική μοντελοποίηση του προβλήματος και γι’ αυτό έχει εγκαταλειφθεί.

Στο [33] αναφέρεται ότι το ΠροσωποΚεντρικό Κυβικό (ΠΚΚ) πλέγμα είναι ένα κατάλληλο, ρεαλιστικό μοντέλο για την τρισδιάστατη διαμόρφωση μιας πρωτεΐνης. Αυτό το μοντέλο βασίζεται σε κύβους με πλευρές μήκους 2 , στο οποίο τα αποδεκτά σημεία είναι οι κορυφές του κύβου μαζί με το κεντρικό σημείο κάθε επιφάνειας του κύβου. Τα σημεία που απέχουν απόσταση √2 συνδέονται μεταξύ τους και η απόσταση αυτή ονομάζεται τμήμα πλέγματος. Κάθε σημείο γειτνιάζει με 12 σημεία (βλ. **Εικόνα 5.2**).

Κάθε αμινοξύ μπορεί να τοποθετηθεί σε ένα σημείο του προσωποκεντρικού κυβικού πλέγματος και να συνδεθεί με ένα γειτονικό αμινοξύ με μια από τις 12 δυνατές συνδέσεις. Η γωνία ανάμεσα σε 3 αμινοξέα μπορεί να είναι 60◦, 90◦, 120◦ και 180◦. Κάποιοι περιορισμοί όμως στις πρωτεΐνες κάνουν τις τιμές 60◦ και 180◦ αδύνατες. Γι' αυτό μόνο συνδέσεις 90◦ και 120◦ είναι αποδεκτές. Τέτοιος περιορισμός δεν υπάρχει στις γωνίες ανάμεσα σε 4 αμινοξέα. Μία επαφή ορίζεται ανάμεσα σε δύο μη γειτονικά αμινοξέα όταν η απόστασή τους είναι 2 τμήματα πλέγματος. Ένα τμήμα πλέγματος ανταποκρίνεται φυσικά σε απόσταση 3,8 Å, που είναι περίπου η απόσταση Van der Waals ανάμεσα στα 2 κεντρικά άτομα άνθρακα δύο συνεχόμενων αμινοξέων. Συνεπώς, πρέπει να επιβληθεί ο περιορισμός ότι 2 μη συνεχόμενα αμινοξέα πρέπει να απέχουν περισσότερο δύο τμήματα πλέγματος.

Ορισμός 1 (Πλέγμα). Ένα κρυσταλλικό πλέγμα είναι ένας γράφος (P,E), όπου P είναι ένα σύνολο από 3D σημεία $(x, y, z) \in Z^3$, που συνδέονται από μη κατευθυνόμενες ακμές (E). Τα πλέγματα περιέχουν ισχυρές συμμετρίες και παρουσιάζουν τακτικά πρότυπα που επαναλαμβάνονται στον χώρο. Αν όλοι κόμβοι έχουν τον ίδιο βαθμό δ (δηλ., ο αριθμός των ακμών που εξέρχονται από έναν κόμβο είναι δ), τότε λέμε ότι το πλέγμα είναι δ-συνδεδεμένο.

Με δεδομένα δύο σημεία $A,B \in P$, και $\{Ax, Ay, Az\}$, $\{Bx, By, Bz\}$ τις συντεταγμένες τους στον 3D χώρο, ορίζονται οι ακόλουθες δύο έννοιες της απόστασης ανάμεσα στα σημεία του πλέγματος:

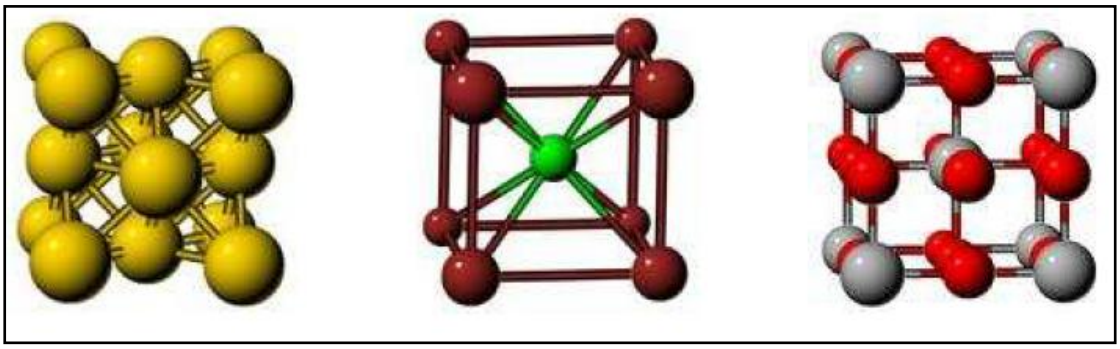
- $\text{ΤΕΤΡΑΓΩΝΗ_ΕΥΚΛΕΙΔΙΑ_ΑΠΟΣΤΑΣΗ}\ (A,B) = (Bx - Ax)^2 + (By - Ay)^2 + (Bz - Az)^2$

- $\text{ΜΕΓΙΣΤΗ_ΑΠΟΣΤΑΣΗ_ΚΑΤΕΥΘΥΝΣΗΣ}(A,B) = \max\{|Bx - Ax|, |By - Ay|, |Bz - Az|\}$

Ισχύει πάντα ότι

$\text{ΜΕΓΙΣΤΗ_ΑΠΟΣΤΑΣΗ_ΚΑΤΕΥΘΥΝΣΗΣ}(A,B)^2 \leq \text{ΤΕΤΡΑΓΩΝΗ_ΕΥΚΛΕΙΔΙΑ_ΑΠΟΣΤΑΣΗ}\ (A,B).$

Δείγματα 3D κρυσταλλικών πλεγμάτων φαίνονται παρακάτω, στην **Εικόνα 5.1**.



Εικόνα 5.1 Δείγματα 3D κρυσταλλικών πλεγμάτων [50]

Σε αυτήν την εργασία χρησιμοποιείται ένας συγκεκριμένος τύπος πλέγματος, το λεγόμενο ΠροσωποΚεντρικό Κυβικό (ΠΚΚ) πλέγμα.

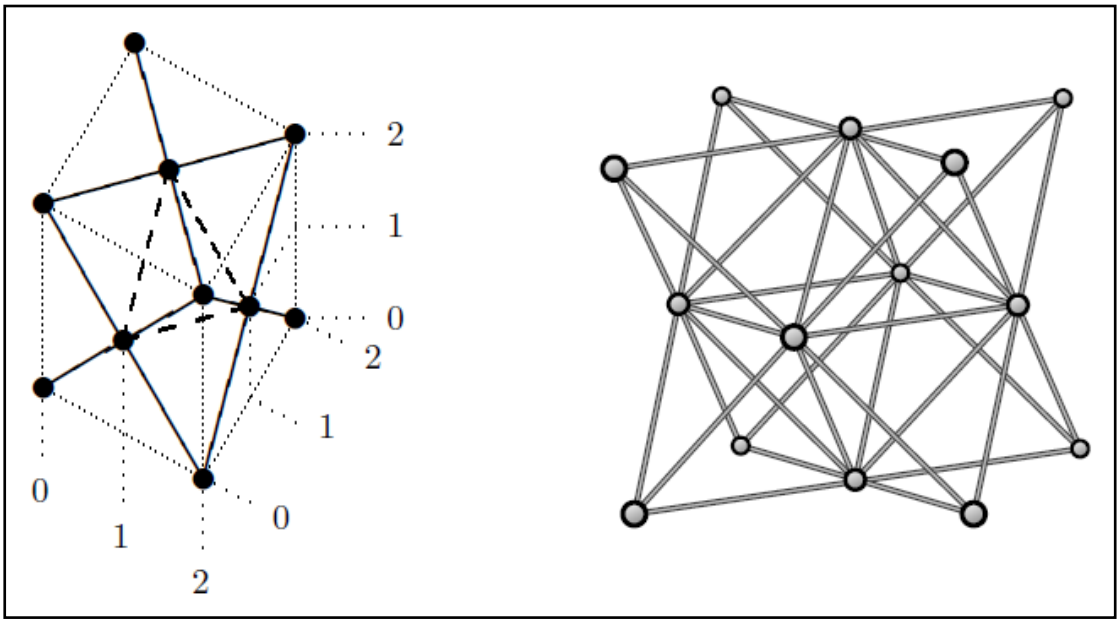
Ορισμός 2 (ΠροσωποΚεντρικό Κυβικό πλέγμα – ΠΚΚ πλέγμα). Ένα ΠροσωποΚεντρικό Κυβικό (ΠΚΚ) πλέγμα (P,E) ορίζεται από τις ακόλουθες ιδιότητες:

- $P = \{(x, y, z) \mid x, y, z \in Z \wedge (x + y + z)$ είναι άρτιο$\}$;
- $E = \{(A,B) \mid A,B \in P, \text{ΤΕΤΡΑΓΩΝΗ_ΕΥΚΛΕΙΔΙΑ_ΑΠΟΣΤΑΣΗ } (A,B) = 2\}$.

Δηλαδή, ένα ΠΚΚ μοντέλο βασίζεται σε κύβους με πλευρές μήκους 2, όπου το κεντρικό σημείο κάθε επιφάνειας του κύβου είναι επίσης σημείο του πλέγματος. Ο πρακτικός κανόνας για τον υπολογισμό των σημείων που ανήκουν στο πλέγμα είναι ο έλεγχος αν το άθροισμα των συντεταγμένων των σημείων (x,y,z) είναι άρτιο (**Εικόνα 5.2**). Ζευγάρια σημείων (x_i, y_i, z_i), (x_j, y_j, z_j) με τετράγωνη Ευκλείδεια απόσταση 2 συνδέονται και σχηματίζουν τις ακμές του πλέγματος; η απόστασή τους ονομάζεται μονάδα πλέγματος, για την οποία ισχύει ότι

- $|x_i - x_j| + |y_i - y_j| + |z_i - z_j| = 2$.

Το ΠΚΚ πλέγμα είναι 12-συνδεδεμένο. Σύμφωνα με το [33], το ΠΚΚ πλέγμα θεωρείται ένα κατάλληλο, ρεαλιστικό μοντέλο για τις τρισδιάστατες διαμορφώσεις των πρωτεϊνών.



Εικόνα 5.2 Ένας κύβος του ΠΚΚ πλέγματος. Οι έντονες γραμμές είναι οι ακμές. Οι διακεκομμένες γραμμές είναι οι ακμές μέσα στον κύβο. Στα δεξιά, φαίνεται το πλήρες σύνολο των ακμών. [50]

5.4 Περιορισμοί

5.4.1 Χωρικοί Περιορισμοί πάνω στο ΠΚΚ πλέγμα

Η χωροθέτηση μιας πρωτεΐνης στο ΠΚΚ πλέγμα επιβάλλει μια σειρά από χωρικούς περιορισμούς. Μια αναδίπλωση της πρωτεΐνης $S = s_1 \ldots s_n$, όπου s_i είναι το i-οστό αμινοξύ είναι μια καταχώριση τιμών στις μεταβλητές θέσης $V_1, V_2, \ldots, V_n$ κάθε αμινοξέως που ικανοποιεί τους παρακάτω περιορισμούς.

1) Για κάθε $i \in \{1,\ldots, n-1\}$,

   ΤΕΤΡΑΓΩΝΗ_ΕΥΚΛΕΙΔΙΑ_ΑΠΟΣΤΑΣΗ $(V_i, V_{i+1}) = 2$.

Γνωρίζοντας από την βιβλιογραφία ότι δύο συνεχόμενα αμινοξέα διαχωρίζονται από μια σταθερή απόσταση (τυπικά, 3.8 Å), πρέπει όλα τα γειτονικά αμινοξέα στην πρωτογενή ακολουθία να απεικονίζονται σε γειτονικά σημεία του πλέγματος. Το γεγονός αυτό μεταφράζεται σε τετράγωνη ευκλείδεια απόσταση ίση με 2.

2) Όλες οι μεταβλητές θέσης $\{V_1, V_2, \ldots, V_n\}$ πρέπει να έχουν διαφορετικές τιμές μεταξύ τους. Δηλ. όλα τα αμινοξέα πρέπει να καταλαμβάνουν διαφορετικές θέσεις στο πλέγμα, αφού δεν πρέπει να υπάρχουν βρόγχοι.

3) Για κάθε $i \in \{2, \ldots, n-1\}$,
$\text{ΤΕΤΡΑΓΩΝΗ_ΕΥΚΛΕΙΔΙΑ_ΑΠΟΣΤΑΣΗ}\ (V_{i-1}, V_{i+1}) \leq 7$,

Στο ΠΚΚ πλέγμα, κάθε σημείο έχει 12 γειτονικά σημεία και η γωνία ανάμεσα σε τρία συνεχόμενα σημεία μπορεί να πάρει τιμές 60◦, 90◦, 120◦, και 180◦. Ογκομετρικοί περιορισμοί στις πρωτεΐνες αποκλείουν τις τιμές 60◦ και 180◦. Για το λόγο αυτό, το μοντέλο περιορίζει να υπάρχουν μόνο γωνίες 90◦ και 120◦.

4) Για κάθε $i, j \in \{1, \ldots, n\}, |i-j| >= 2$,
$\text{ΤΕΤΡΑΓΩΝΗ_ΕΥΚΛΕΙΔΙΑ_ΑΠΟΣΤΑΣΗ}(V_i, V_j) > 4$.

Δηλ. κάθε δυάδα μη συνεχόμενων αμινοξέων πρέπει να διαχωρίζεται από περισσότερες από μια μονάδες πλέγματος, αφού σύμφωνα με τις υπάρχουσες παρατηρήσεις δύο μη συνεχόμενα αμινοξέα δεν μπορούν να βρίσκονται σε απόσταση μιας μονάδας πλέγματος, επειδή τότε οι όγκοι τους θα επικαλύπτονταν.

5.4.2 Περιορισμός συμπύκνωσης

Βασισμένοι σε μια μελέτη επάνω στις δομές των πρωτεϊνών, οι Huang Και Powers στο [13] κατέληξαν στο συμπέρασμα ότι το μέγιστο μήκος μιας

πρωτεΐνης δίνεται από τον τύπο $2.2n^{0.38}$ Å, όπου n είναι το πλήθος των αμινοξέων της πρωτεΐνης. Έτσι, για κάθε ζεύγος αμινοξέων είναι λογικό να πρέπει να επιβληθεί ένα άνω όριο στην απόστασή τους. Αυτό γίνεται με την χρήση ενός παράγοντα σύμπτυξης. Έτσι, εισάγεται ο ακόλουθος περιορισμός:

- Για κάθε $i, j \in \{1, \ldots, n\}$,

  $\text{ΜΕΓΙΣΤΗ_ΑΠΟΣΤΑΣΗ_ΚΑΤΕΥΘΥΝΣΗΣ}(Vi, Vj) \leq cf * n,$

όπου cf είναι ο παράγοντας συμπύκνωσης, που εκφράζεται σαν ένας αριθμός ανάμεσα σε 0 και 1 και n είναι το μήκος της πρωτεΐνης. Ο παράγοντας συμπύκνωσης καθορίζει μια προσεγγιστική μέγιστη απόσταση ανάμεσα στα αμινοξέα. Η τιμή του cf θα μπορεί να δίνεται από στον χρήστη, αλλά υπάρχει και μια εξ' ορισμού τιμή που έχει υπολογισθεί με βάση τον τύπο $2.2n^{0.38}$ Å, ώστε να απεικονίζει την μονάδα πλέγματος στην πραγματική απόσταση ανάμεσα στα αμινοξέα.

5.4.3 Περιορισμοί που αφορούν τις δευτερεύουσες δομές μιας πρωτεΐνης

Μαζί με την πρωτογενή ακολουθία αμινοξέων, το μοντέλο της εργασίας συμπεριλαμβάνει και την γνώση σχετικά με την παρουσία ιδιαίτερων δευτερευουσών δομών στην πρωτεΐνη (π.χ. έλικας ή πτυχωτής επιφάνειας). Όπως αναφέρθηκε στο Κεφάλαιο 2, η τελική διαμόρφωση μιας πρωτεΐνης μπορεί να εμπεριέχει αρκετά δευτερογενή δομικά στοιχεία (δηλ. τις έλικες και τις πτυχώσεις) που τακτοποιούνται σε συγκεκριμένα μοτίβα. Οι έλικες αποτελούνται από 5 ως 40 συνεχόμενα αμινοξέα σε δεξιόστροφη έλικα, με 3.6 αμινοξέα ανά στροφή. Οι πτυχωτές επιφάνειες αποτελούνται από εκτεταμένα φύλλα των 5 ως 10 αμινοξέων. Κάθε φύλλο αποτελείται από συνεχόμενα αμινοξέα αλλά τα φύλλα που συμμετέχουν σε μια πτύχωση δεν είναι απαραίτητα συνεχόμενα στην πρωτεϊνική ακολουθία. Ένα άλλο σημαντικό δομικό χαρακτηριστικό των πρωτεϊνών είναι η ικανότητα των κυστεινών να σχηματίζουν ομοιοπολικούς δεσμούς μέσω των ατόμων θείου που διαθέτουν, σχηματίζοντας έτσι γέφυρες που επιβάλουν σημαντικούς περιορισμούς

επαφής. Αυτό το είδος πληροφορίας είναι συχνά διαθέσιμο, είτε από πειράματα είτε από προβλέψεις.

Υπάρχουν αλγόριθμοι βασισμένοι σε νευρωνικά δίκτυα [25], οι οποίοι μπορούν να προβλέψουν με μεγάλη ακρίβεια της τάξης του 75% την δευτερογενή δομή μιας πρωτεΐνης (π.χ., τα συστήματα PHD ή PSI-pred). Όλη αυτή η πληροφορία χρησιμοποιείται για να επιβάλλει αρκετούς τοπικούς περιορισμούς, αναγκάζοντας μια ακολουθία σημείων να έχουν μια αυστηρή χωρική μορφή (δες Ενότητα 5.6).

5.5 Μοντελοποίηση της ενέργειας της πρωτεΐνης

Η ενέργεια της πρωτεΐνης ορίζεται από το άθροισμα των ενεργειών που συνεισφέρουν όλα τα ζευγάρια των μη συνεχόμενων αμινοξέων. Αυτές οι ενέργειες εξαρτώνται από την απόσταση και των τύπο των αμινοξέων. Η ενεργειακή συνεισφορά είναι μηδενική αν τα δύο αμινοξέα είναι σε απόσταση μεγαλύτερη από 2 μονάδες πλέγματος. Για το ΠΚΚ πλέγμα θεωρούμε ότι δύο αμινοξέα Α, Β είναι σε επαφή ώστε να αλληλεπιδρούν και άρα συνεισφέρουν ενεργειακά όταν η τετράγωνη Ευκλείδεια απόσταση μεταξύ τους έχει τιμή 4, δηλ. όταν απέχουν δύο μονάδες πλέγματος.

Κάθε φορά λοιπόν που εντοπίζεται μια επαφή ανάμεσα σε ένα ζεύγος αμινοξέων, υπολογίζεται μια ιδιαίτερη ενεργειακή συνεισφορά, ανάλογα με το ζεύγος των αμινοξέων, σύμφωνα με τον **Πίνακας 5.1**. Οι τιμές του πίνακα αυτού έχουν υπολογιστεί με στατιστική ανάλυση των φυσικών χωροθετήσεων των εργαστηριακά ταυτοποιημένων πρωτεινών [17, 37].

Για κάποια ζεύγη αμινοξέων η συνεισφορά είναι θετική, ενώ για άλλα είναι αρνητική. Αφού αναζητείται η διαμόρφωση με την ελάχιστη ενέργεια, προτιμώνται οι αρνητικές συνεισφορές. Έτσι, οι τιμές ενέργειας των προβλεπόμενων δομών καταλήγουν να είναι αρνητικές.

	CYS	MET	PHE	ILE	LEU	VAL	TRP	TYR	ALA	GLY
CYS	-3.477	-2.240	-2.424	-2.410	-2.343	-2.258	-2.080	-1.892	-1.700	-1.101
MET	-2.240	-1.901	-2.304	-2.286	-2.208	-2.079	-2.090	-1.834	-1.517	-0.897
PHE	-2.424	-2.304	-2.467	-2.530	-2.491	-2.391	-2.286	-1.963	-1.750	-1.034
ILE	-2.410	-2.286	-2.530	-2.691	-2.647	-2.568	-2.303	-1.998	-1.872	-0.885
LEU	-2.343	-2.208	-2.491	-2.647	-2.501	-2.447	-2.222	-1.919	-1.728	-0.767
VAL	-2.258	-2.079	-2.391	-2.568	-2.447	-2.385	-2.097	-1.790	-1.731	-0.756
TRP	-2.080	-2.090	-2.286	-2.303	-2.222	-2.097	-1.867	-1.834	-1.565	-1.142
TYR	-1.892	-1.834	-1.963	-1.998	-1.919	-1.790	-1.834	-1.335	-1.318	-0.818
ALA	-1.700	-1.517	-1.750	-1.872	-1.728	-1.731	-1.565	-1.318	-1.119	-0.290
GLY	-1.101	-0.897	-1.034	-0.885	-0.767	-0.756	-1.142	-0.818	-0.290	0.219
THR	-1.243	-0.999	-1.237	-1.360	-1.202	-1.240	-1.077	-0.892	-0.717	-0.311
SER	-1.306	-0.893	-1.178	-1.037	-0.959	-0.933	-1.145	-0.859	-0.607	-0.261
ASN	-0.788	-0.658	-0.790	-0.669	-0.524	-0.673	-0.884	-0.670	-0.371	-0.230
GLN	-0.835	-0.720	-0.807	-0.778	-0.729	-0.642	-0.997	-0.687	-0.323	0.033
GLU	-0.616	-0.409	-0.482	-0.402	-0.291	-0.298	-0.613	-0.631	-0.235	-0.097
ASP	-0.179	-0.209	-0.419	-0.439	-0.366	-0.335	-0.624	-0.453	-0.039	0.443
HIS	-1.499	-1.252	-1.330	-1.234	-1.176	-1.118	-1.383	-1.222	-0.646	-0.325
ARG	-0.771	-0.611	-0.805	-0.854	-0.758	-0.664	-0.912	-0.745	-0.327	-0.050
LYS	-0.112	-0.146	-0.270	-0.253	-0.222	-0.200	-0.391	-0.349	0.196	0.589
PRO	-1.196	-0.788	-1.076	-0.991	-0.771	-0.886	-1.278	-1.067	-0.374	-0.042

	THR	SER	ASN	GLN	ASP	GLU	HIS	ARG	LYS	PRO
CYS	-1.243	-1.306	-0.788	-0.835	-0.616	-0.179	-1.499	-0.771	-0.112	-1.196
MET	-0.999	-0.893	-0.658	-0.720	-0.409	-0.209	-1.252	-0.611	-0.146	-0.788
PHE	-1.237	-1.178	-0.790	-0.807	-0.482	-0.419	-1.330	-0.805	-0.270	-1.076
ILE	-1.360	-1.037	-0.669	-0.778	-0.402	-0.439	-1.234	-0.854	-0.253	-0.991
LEU	-1.202	-0.959	-0.524	-0.729	-0.291	-0.366	-1.176	-0.758	-0.222	-0.771
VAL	-1.240	-0.933	-0.673	-0.642	-0.298	-0.335	-1.118	-0.664	-0.200	-0.886
TRP	-1.077	-1.145	-0.884	-0.997	-0.613	-0.624	-1.383	-0.912	-0.391	-1.278
TYR	-0.892	-0.859	-0.670	-0.687	-0.631	-0.453	-1.222	-0.745	-0.349	-1.067
ALA	-0.717	-0.607	-0.371	-0.323	-0.235	-0.039	-0.646	-0.327	0.196	-0.374
GLY	-0.311	-0.261	-0.230	0.033	-0.097	0.443	-0.325	-0.050	0.589	-0.042
THR	-0.617	-0.548	-0.463	-0.342	-0.382	-0.192	-0.720	-0.247	0.155	-0.222
SER	-0.548	-0.519	-0.423	-0.260	-0.521	-0.161	-0.639	-0.264	0.223	-0.199
ASN	-0.463	-0.423	-0.367	-0.253	-0.344	0.160	-0.455	-0.114	0.271	-0.018
GLN	-0.342	-0.260	-0.253	0.054	0.022	0.179	-0.290	-0.042	0.334	-0.035
GLU	-0.382	-0.521	-0.344	0.022	0.179	0.634	-0.664	-0.584	-0.176	0.189
ASP	-0.192	-0.161	0.160	0.179	0.634	0.933	-0.324	-0.374	-0.057	0.257
HIS	-0.720	-0.639	-0.455	-0.290	-0.664	-0.324	-1.078	-0.307	0.388	-0.346
ARG	-0.247	-0.264	-0.114	-0.042	-0.584	-0.374	-0.307	0.200	0.815	-0.023
LYS	0.155	0.223	0.271	0.334	-0.176	-0.057	0.388	0.815	1.339	0.661
PRO	-0.222	-0.199	-0.018	-0.035	0.189	0.257	-0.346	-0.023	0.661	0.129

Πίνακας 5.1 Ενέργειες αμινοξέων [17]

5.6 Υλοποίηση του ΠΑΠ σε CLP(FD)

Σε αυτήν την ενότητα περιγράφεται η προτεινόμενη CLP(FD) μοντελοποίηση του ΠΑΠ, με έμφαση στα κύρια κατηγορήματα υλοποίησης. Χρησιμοποιήθηκε η βιβλιοθήκη 'clpfd' της SICStus PROLOG 3.9.1 [38]. Ο πλήρης κώδικας δίνεται στο Παράρτημα 1 της εργασίας.

Ο κύριος όρος του προγράμματος είναι ένας κλασσικός όρος «Περιορισμοί & Αναζήτηση» της μορφής:

```
anadiplosi ( Kwdikos, Xronos, Sympyknwsh ):-
      arxikopoihsh,
      prwteinh(Kwdikos, Prwtogenhs, Deuterogenhs),
      Arxikos_xronos,
      periorismoi(Prwtogenhs, Deuterogenhs, Tritogenhs,
            Energeia, Sympyknwsh, Taks_tritogenhs,
            Lista_me_Energeies),
      emfanise_xrono('Χρόνος Περιορισμών: '),nl,!,
      anazhthsh_lyshs(Kwdikos, Prwtogenhs, Deuterogenhs,
            Sympyknwsh, Xronos, Tritogenhs, Energeia,
            Lista_me_Energeies),
      emfanise_xrono('Συνολικός χρόνος: '),
      emfanise_apotelesma(Prwtogenhs,Tritogenhs),
      write('Ενέργεια = '), write(Energeia).
```

Η μεταβλητή εισόδου `Kwdikos` είναι ο κωδικός της πρωτεΐνης για την οποία αναζητείται η 3D δομή, η μεταβλητή εισόδου `Xronos` είναι ένα χρονικό όριο που μπορεί να τεθεί για την αναζήτηση σε secs και η μεταβλητή εισόδου `Sympyknwsh` είναι ο παράγοντας συμπύκνωσης της λύσης.

Το κατηγόρημα 'prwteinh', ορίζεται εξωτερικά στο βοηθητικό αρχείο 'prwteines.pl' και μας επιτρέπει να προσπελάσουμε την πρωτογενή και τις δευτερογενείς δομές μιας πρωτεΐνης με βάση τον κωδικό της, στις μεταβλητές εισόδου `Prwtogenhs` και `Deuterogenhs`, αντίστοιχα.
Για παράδειγμα,

```
prwteinh(Kwdikos, Prwtogenhs, Deuterogenhs):-
      Kwdikos = '1LE0',
      Prwtogenhs = [s,w,t,w,e,g,n,k,w,t,w,k],
      Deuterogenhs = [ptyxwsh (2,4), ptyxwsh (9,11)].
```

Το κατηγόρημα ‘periorismoi’ δημιουργεί τις μεταβλητές που αναπαριστούν τις θέσεις των αμινοξέων και προσθέτει τους κατάλληλους περιορισμούς για τις μεταβλητές αυτές. Η μεταβλητή εξόδου Tritogenhs είναι η λίστα με τριάδες μεταβλητών που αναπαριστούν τις θέσεις των αμινοξέων, δηλ. η τριτοταγής διαμόρφωση της πρωτεΐνης. Το κατηγόρημα ‘anazhthsh_lyshs’ αναζητά την λύση στο χώρο αναζήτησης. Το κατηγόρημα ‘emfanise_apotelesma’ εκτυπώνει την έξοδο σε κατάλληλη μορφή, επιτρέποντάς την οπτικοποίηση της ευρισκόμενης 3D δομής της πρωτεΐνης με βιολογικό λογισμικό απεικόνισης. Η μεταβλητή εξόδου Energeia είναι η τιμή εξόδου με την ενέργεια της πρωτεΐνης (οι τιμές του πίνακα ενεργειών πολλαπλασιάζονται με το 1000 για να προκύψουν ακέραιες τιμές ενέργειας).

Ένα παράδειγμα χρήσης του προγράμματος είναι ο ακόλουθος υπολογισμός:

```
?- anadiplosi ('1LE0', 120, 0.6).
          1 s 12 12 12
          2 w 13 13 12
          3 t 14 13 13
          4 w 15 12 13
          5 e 16 13 13
          6 g 17 12 13
          7 n 17 11 12
          8 k 18 10 12
          9 w 17 9 12
          10 t 16 9 13
          11 w 15 10 13
          12 k 14 10 12
          Ενέργεια = -4085
```

5.6.1 Το κατηγόρημα ‘periorismoi’

Το κατηγόρημα periorismoi ορίζεται ως εξής:

```
periorismoi(Prwtogenhs, Deuterogenhs, Tritogenhs, Energeia,
Sympyknwsh, Taks_tritogenhs,  Lista_me_Energeies):-
      elegxos(Prwtogenhs),
      length(Prwtogenhs,N),
```

```
dhmiourgia_tritogenhs (N, Tritogenhs),
pedio_orismou(Tritogenhs, N),
apofygh_brogxwn(Tritogenhs, N),
apofygh_symmetriwn(N, Tritogenhs),
periorismoi_epomenou(Tritogenhs),
periorismoi_apostashs(Tritogenhs),
periorismoi_sympyknwshs(Tritogenhs,N, Sympyknwsh),
periorismoi_gwnies(Tritogenhs),
deutereuousa_domh(Deuterogenhs, Tritogenhs),
periorismoi_energeias(Prwtogenhs, Tritogenhs,
     Energeia, Lista_me_Energeies).
```

Το κατηγόρημα ‘elegxos’ ελέγχει την πρωτογενή λίστα αμινοξέων για τυχόν λανθασμένο κωδικό κάποιου αμινοξέως. Το κατηγόρημα ‘dhmiourgia_tritogenhs’ δημιουργεί μια λίστα με τριπλέτες N μεταβλητών που χρησιμοποιούνται σαν συντεταγμένες των N αμινοξέων, όπου N είναι το μήκος της πρωτογενούς λίστας Prwtogenhs.

Το κατηγόρημα ‘pedio_orismou’ περιορίζει στο διάστημα 0 .. 2*N όλες τις μεταβλητές θέσης των αμινοξέων X_i, Y_i, Z_i, δηλ. καθορίζει το μέγιστο μήκος μιας πρωτεΐνης. Επιπλέον, επιβάλλει το άθροισμα των συντεταγμένων κάθε σημείου $X_i + Y_i + Z_i$ να είναι άρτιο, λόγω της ιδιομορφίας του ΠΚΚ πλέγματος.

Το κατηγόρημα ‘apofygh_brogxwn’ επιβάλλει όλες οι τριάδες θέσεων να είναι διαφορετικές. Χρησιμοποιείται το ενσωματωμένο κατηγόρημα ‘all_different’ σε μια νέα λίστα μεταβλητών [I_1, …, I_n] όπου $Ii \#= (X_i*P*P) + (Y_i*P) + Z_i$, για μία κατάλληλη τιμή P.

Το κατηγόρημα ‘apofygh_symmetriwn’ εισάγει περιορισμούς με στόχο την απομάκρυνση περιττών αποδεκτών διαμορφώσεων. Ειδικότερα, τίθεται ότι το επόμενο σημείο για το [N, N, N] θα είναι το [N + 1, N + 1, N], εάν ο N είναι άρτιος (εάν N είναι περιττός, προστίθεται 1 σε όλες τις μεταβλητές για να ικανοποιηθεί ο περιορισμός αρτιότητας του αθροίσματος των μεταβλητών).

Το κατηγόρημα ‘periorismoi_epomenou’ επιβάλλει τον περιορισμό για τις θέσεις όλων των συνεχόμενων αμινοξέων $[X_i, Y_i, Z_i]$ και $[X_{i+1}, Y_{i+1},$

Z_{i+1}] να είναι γειτονικά σημεία στο πλέγμα. Χρησιμοποιούνται οι μεταβλητές Dx,Dy και Dz, περιορίζονται στο διάστημα 0..1 και επιβάλλονται οι ακόλουθοι περιορισμοί:

```
domain([Dx,Dy,Dz],0,1),
Dx #= abs(Xi-Xi+1),
Dy #= abs(Yi-Yi+1),
Dz #= abs(Zi-Zi+1),
sum([Dx,Dy,Dz], #= ,  2).
```

Το κατηγόρημα 'periorismoi_apostashs' προσθέτει περιορισμούς στις μεταβλητές για να εξασφαλίσει ότι κάθε δύο μη συνεχόμενα σημεία απέχουν Ευκλείδεια απόσταση τουλάχιστον 2. Για κάθε αμινοξύ j>i+1 επιβάλλεται:

```
Dx #= (Xi - Xj) * (Xi - Xj),
Dy #= (Yi - Yj) * (Yi - Yj),
Dz #= (Zi - Zj) * (Zi - Zj),
sum([Dx,Dy,Dz], #>, 2).
```

Το κατηγόρημα 'periorismoi_sympyknwshs΄ προσθέτει περιορισμούς στις μεταβλητές για να εξασφαλίσει ότι η μέγιστη διάμετρος της πρωτεΐνης είναι N*Sympyknwsh, όπου N είναι το πλήθος των αμινοξέων της πρωτεΐνης και Sympyknwsh είναι ο παράγοντας συμπύκνωσης που δίνεται ως είσοδος από τον χρήστη του προγράμματος.

Το κατηγόρημα 'periorismoi_gwnies' περιορίζει κάθε 3 συνεχόμενα σημεία να μπορούν να σχηματίσουν μόνο γωνίες 90◦ ή 120◦. Με δεδομένα τρία συνεχόμενα σημεία με συντεταγμένες (X_1, Y_1, Z_1), (X_2, Y_2, Z_2), (X_3, Y_3, Z_3), περιορίζονται τα δύο αντίστοιχα διανύσματα των σχετικών αποστάσεων να είναι:

1) διαφορετικά ώστε να μην σχηματίζεται γωνία κάμψης 180◦ και

2) το βαθμωτό τους γινόμενο να είναι >= 0, ώστε να μην επιτρέπεται γωνία 60◦, η οποία δίνει τιμή -1.

Συγκεκριμένα, αρκεί να τεθούν οι ακόλουθοι περιορισμοί:

```
Ax #= X2-X1,
Ay #= Y2-Y1,
Az #= Z2-Z1,
Bx #= X3-X2,
By #= Y3-Y2,
Bz #= Z3-Z2,
%%% γωνία κάμψης <> 180 μοίρες
10 * Ax + 5 * Ay + Az #\= 10 *  Bx + 5 * By + Bz,
%%% γωνία κάμψης >= 90°
bathmwto_ginomeno ([Ax,Ay,Az],[Bx,By,Bz],gwnia),
gwnia #>= 0.
```

όπου το κατηγόρημα 'bathmwto_ginomeno' εκτελεί το άθροισμα των γινομένων δύο διανυσμάτων.

Το κατηγόρημα 'periorismoi_energeias' υπολογίζει την ενέργεια της διαμόρφωσης στην μεταβλητή Energeia και την λίστα των επιμέρους ενεργειών στην μεταβλητή Lista_me_Energeies. Συγκεκριμένα, υπολογίζει το άθροισμα των μεταβλητών En που ορίζονται από κάθε ζεύγος αμινοξέων A_i, A_j της πρωτογενούς ακολουθίας Prwtogenhs, με j>i+1, ως εξής:

```
pinakas(Ai, Aj, Energeia),
En in {0, Energeia },
DX #= abs(X1 - X2),
DY #= abs(Y1 - Y2),
DZ #= abs(Z1 - Z2),
2 #= DX + DY + DZ  #<=> En #= Energeia, %%% Συνεισφορά επαφής
```

όπου το κατηγόρημα 'pinakas' επιστρέφει την τιμή ενέργειας για τα αμινοξέα A_i, A_j από τον πίνακα των ενεργειών. Ο αριθμός 2 στην τελευταία σειρά επιβάλλει η απαιτούμενη απόσταση επαφής να είναι 2 μονάδες. (Σημ. οι μεταβλητές DX, DY και DZ δεν περιορίζονται στο διάστημα 0..1, όπως συμβαίνει στον περιορισμό για τα συνεχόμενα αμινοξέα στο κατηγόρημα 'periorismoi_epomenou'). Η μεταβλητή Lista_me_Energeies είναι η λίστα

όλων των μεταβλητών `En` που ορίζονται παραπάνω και στην ουσία αναπαριστούν την ενεργειακή συνεισφορά κάθε ζεύγους αμινοξέων.

5.7 Αναζήτηση λύσεων

5.7.1 Καταχώριση τιμών στις μεταβλητές στο ΠΚΚ πλέγμα

Από την στιγμή που καθοριστούν οι περιορισμοί, ξεκινά η διαδικασία της αναζήτησης. Η διαδικασία της αναζήτησης είναι η υλοποίηση μιας συγκεκριμένης διαδικασίας απόδοσης τιμών στις μεταβλητές, διάδοσης των περιορισμών και πιθανών πισωγυρισμάτων [15]. Η εξέλιξη του υπολογισμού μπορεί να απεικονιστεί σαν την κατασκευή ενός δέντρου αναζήτησης, όπου οι εσωτερικοί κόμβοι ανταποκρίνονται στην επιλογή της τιμής μιας μεταβλητής (labeling) ενώ οι ακμές ανταποκρίνονται στην διάδοση της καταχώρισης αυτής σε άλλες μεταβλητές (διαμέσου των περιορισμών).

Συνήθως, χρησιμοποιούνται δύο κύριες στρατηγικές επιλογής μεταβλητών:

1) η αριστερόστροφη στρατηγική (leftmost), η οποία επιλέγει την πιο αριστερή μεταβλητή που δεν έχει πάρει τιμή σαν επόμενη καταχώριση και
2) η στρατηγική πρώτης αποτυχίας (first-fail), η οποία επιλέγει την μεταβλητή με το μικρότερο πεδίο ορισμού.

5.7.2 Διάδοση περιορισμών

Η διάδοση των περιορισμών ενεργοποιείται όταν το πεδίο ορισμού μιας μεταβλητής τροποποιηθεί. Π.χ. αν υποτεθεί μια κατάσταση όπου οι μεταβλητές $G = \{V_1, \ldots, V_{k-1}\}$ έχουν πάρει συγκεκριμένες τιμές, V_k είναι η μεταβλητή που πρόκειται να λάβει τιμή και $NG = \{V_{k+1}, \ldots, V_n\}$ είναι οι υπόλοιπες μεταβλητές. Το πρώτο βήμα μετά την καταχώρηση τιμής στην V_k,

είναι να ελεγχεί η συνέπεια των περιορισμών της μορφής $C(V_k, V_i)$, όπου $V_i \in G$ (πρόκειται για τον έλεγχο συνέπειας κόμβου). Η επόμενη φάση διάδοσης διαιρείται σε δύο βήματα. Καταρχήν, όλοι οι περιορισμοί της μορφής $C(V_k, V_j)$ επεξεργάζονται, όπου $V_j \in NG$. Αυτό το βήμα διαδίδει τα νέα όρια της V_k στις μεταβλητές που δεν έχουν ακόμα πάρει τιμή. Αμέσως μετά εφαρμόζεται ο έλεγχος ορίων στους περιορισμούς της μορφής $C(V_i, V_j)$, όπου $V_i \in NG$.

Ειδικότερα, στο ΠΑΠ αν η μεταβλητή V_{i-1} έχει πάρει τιμή και η μεταβλητή V_i πρόκειται να λάβει τιμή, υπάρχουν μόνο 12 δυνατές τιμές που επιτρέπει το πλέγμα - δεδομένου ότι το πλέγμα είναι 12-συνδεδεμένο και τα συνεχόμενα αμινοξέα απέχουν μια μονάδα πλέγματος. Έτσι επεκτείνεται το δέντρο αναζήτησης με μόνο αυτές τις 12 τιμές που είναι συμβατές με το τρέχον πεδίο ορισμού της V_i. Μια ιδιαίτερη περίπτωση, αλλά πολύ κοινή, συμβαίνει όταν δύο συνεχόμενες μεταβλητές θέσης, π.χ. V_{i-1} και V_{i-2} έχουν συγκεκριμένες τιμές. Σε αυτήν την περίπτωση, η αλληλεπίδραση των περιορισμών περιορίζει τις τιμές της V_i σε 6 το πολύ δυνατές τιμές (λόγω των περιορισμών για τις δυνατές γωνίες κάμψης 90° ή 120°), και μόνο αυτές χρησιμοποιούνται για την επέκταση του δέντρου αναζήτησης. Η χρήση αυτής της τεχνικής διάδοσης περιορισμών επιταχύνει την διαδικασία σημαντικά σε σχέση με μια διαδικασία που εξερευνά όλα τα δυνατά σημεία.

5.7.3 Πρώτη δοκιμή του προγράμματος

Σε πρώτη φάση δοκιμάστηκε η απόδοση και η κλιμάκωση του προγράμματος χωρίς την χρήση της πληροφορίας για την δευτερεύουσα δομή των πρωτεϊνών, δηλ. χωρίς την ενσωμάτωση του κατηγορήματος ‘deutereuousa_domh‘ που επιβάλλει τους περιορισμούς δευτερεύουσας δομής. Για την αναζήτηση λύσεων χρησιμοποιείται το ενσωματωμένο κατηγόρημα `labeling` ως εξής:

```
labeling([ff,minimize(Energeia)], Tritogenhs).
```

Γίνεται δηλ. εξερεύνηση του δέντρου αναζήτησης με την χρήση της first-fail (πρώτη αποτυχία) στρατηγικής επιλογής μεταβλητής, αφού ως 1η παράμετρος του κατηγορήματος ‘`labeling`’ τίθεται η επιλογή `ff`. Επίσης, χρησιμοποιείται η μέθοδος branch-and-bound για βελτιστοποίηση. Χρησιμοποιήθηκε ο ενσωματωμένος branch-and-bound βελτιστοποιητής, με την χρήση της επιλογής `minimize`.

Σύμφωνα με τις δοκιμές που πραγματοποιήθηκαν, το πρόγραμμα βρίσκει λύσεις σε αποδεκτό χρόνο (λίγα λεπτά) μόνο για μικρές πρωτεΐνες μήκους N ≤ 20, γεγονός που προτρέπει για την ενσωμάτωση της δευτερεύουσας δομής με στόχο την εφαρμογή του προγράμματος σε μεγαλύτερες πρωτεΐνες.

5.8 Χρήση της δευτερεύουσας δομής

Σ’ αυτήν την ενότητα παρουσιάζεται ο τρόπος ενσωμάτωσης της δευτερεύουσας δομής μιας πρωτεΐνης, για να βελτιωθεί το δηλωτικό πρωτότυπο που παρουσιάστηκε στην προηγούμενη ενότητα. Όπως αναφέρθηκε στο Κεφάλαιο 2, η δευτερεύουσα δομή μιας πρωτεΐνης μπορεί να προβλεφτεί με αρκετά μεγάλη ακρίβεια (π.χ., από τα προγράμματα `PhD' ή PSI-pred).

Για τα πειράματα που έγιναν η πληροφορία της δευτερεύουσας δομής ανακτήθηκε από τη βάση πρωτεϊνών PDB [10] και καταγράφηκε χειροκίνητα στο αρχείο των πρωτεϊνών ‘`prwteines.pl`’. Μελλοντικά προγραμματίζεται να τροποποιηθεί ο παρόν κώδικας ώστε να χρησιμοποιεί αυτόματα την πρόβλεψη που δίνει το πρόγραμμα PhD ή άλλα σχετικά προγράμματα. Η πληροφορία της δευτερεύουσας δομής μιας πρωτεΐνης έχει την εξής μορφή:

- `elika(i, j)`: όπου ορίζει ότι τα στοιχεία `i, i+1, …, j` της ακολουθίας εισόδου σχηματίζουν μία έλικα,

- `ptyxwsh(i, j)`: όπου ορίζει ότι τα στοιχεία `i, i+1, …, j` της ακολουθίας εισόδου σχηματίζουν μία πτυχωτή επιφάνεια.

- `gefyra(i, j)`: μαρτυρά την παρουσία γέφυρας κυστεινών ανάμεσα στα στοιχεία `i, i+1, …, j` (δηλ. ότι τα στοιχεία αυτά είναι πολύ κοντά το ένα στο άλλο).

Όλη η παραπάνω πληροφορία ανακτάται από το κυρίως πρόγραμμα μέσα από το αρχείο των πρωτεϊνών '`prwteines.pl`' και αποθηκεύεται στη λίστα `Deuterogenhs`.

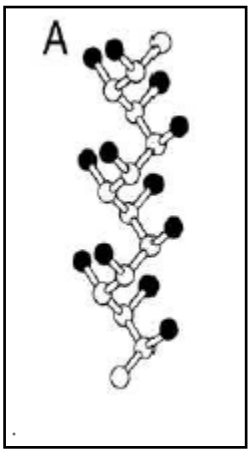
Το μοντέλο περιορισμών για την περιγραφή της δευτερεύουσας δομής που προτείνεται στο [24] υιοθετεί μία επιπρόσθετη αναπαράσταση με δείκτες κατευθύνσεων για να περιγράψει τις θέσεις των αμινοξέων. Συγκεκριμένα, αφού στο ΠΚΚ πλέγμα για δύο δεδομένα σημεία μόνο 6 δυνατές κατευθύνσεις επιτρέπονται για το επόμενο σημείο, μία λίστα από (N – 2) αριθμούς που κυμαίνονται από 1..6 μπορεί να περιγράψει επακριβώς την θέση του κάθε αμινοξέως σε μια πρωτεΐνη μεγέθους N. Αυτή η αναπαράσταση κρίνεται κατάλληλη για την επιβολή περιορισμών που αφορούν έλικες και πτυχώσεις, διότι οι δευτερεύουσες δομές εξαιτίας του επαναλαμβανόμενου σχήματός τους, περιγράφονται απλά με μια ακολουθία κατευθύνσεων (γωνιών). Ειδικότερα, οι έλικες περιγράφονται ως ακολουθίες της μορφής 1-3-4-6-1-3-4-6-… (μπορούν επίσης να ξεκινούν από 3, 4, ή 6) στις μεταβλητές των δεικτών κατεύθυνσης. Ενώ οι β-πτυχώσεις περιγράφονται ως ακολουθίες της μορφής: 2-3-2-3-… (ή 3-2-3-2-…).

Όμως οι δείκτες κατεύθυνσης έπρεπε να συνδεθούν με τις Καρτεσιανές συντεταγμένες και αυτό γινόταν με ένα επιπρόσθετο σύνολο περιορισμών. Παρατηρήθηκε όμως ότι αν και το παραγόμενο μοντέλο ήταν ξεκάθαρο και δηλωτικό, οι παραγόμενοι περιορισμοί ήταν μη γραμμικοί αφού απαιτούσαν διανυσματικά γινόμενα και επομένως είχαν φτωχή διάδοση κατά την διάρκεια της αναζήτησης. Αυτή ήταν και η κύρια αιτία για την μικρή σχετικά απόδοση και την έλλειψη κλιμάκωσης για μεγαλύτερες πρωτεΐνες.

Στην παρούσα εργασία επιλέχθηκε να μοντελοποιηθεί η περιγραφή των δευτερευουσών δομών αντί για τους δείκτες κατεύθυνσης με τις Καρτεσιανές συντεταγμένες. Δηλ. έγινε προσπάθεια να περιοριστούν όσο τον δυνατόν οι μη γραμμικοί περιορισμοί, ευελπιστώντας σε καλύτερη απόδοση και κλιμάκωση. Ειδικότερα, η περιγραφή των δευτερευουσών δομών με Καρτεσιανές συντεταγμένες έγινε από το κατηγόρημα '`deutereuousa_domh`' ως εξής:

α-έλικες: η μοντελοποίηση των α-ελικών (**Εικόνα 5.3**) βασίζεται στην παρατήρηση ότι αρκεί να περιοριστούν τα 5 πρώτα αμινοξέα της έλικας για να υπάρξει εγγύηση για το σχήμα της έλικας - το σχήμα αυτό μπορεί να διαδοθεί στο υπόλοιπο της έλικας μέσα από απλές ισότητες διανυσμάτων. Ειδικότερα, με δεδομένα τα 5 πρώτα αμινοξέα μιας έλικας με συντεταγμένες $<X_1,Y_1,Z_1>$, . . . , $<X_5,Y_5,Z_5>$, οι ακόλουθοι όροι μπορούν να περιγράψουν την οργάνωση της έλικας με βάση τις σχετικές αποστάσεις ανάμεσα στα σημεία αυτά στο ΠΚΚ πλέγμα:

Το 1^o με το 5^o σημείο πρέπει να έχουν μια ίδια συντεταγμένη , το 1^o με το 3^o καθώς και το 3^o με το 5^o σημείο πρέπει να έχουν δύο ίδιες συντεταγμένες, το 1^o με το 4^o καθώς και το 2^o με το 5^o σημείο πρέπει να απέχουν σε όλες τις συνταγμένες τους απόσταση μικρότερη από 2.



Εικόνα 5.3 α-έλικα [50]

Οι παραπάνω όροι μπορούν να περιγραφούν με CLP(FD) περιορισμούς ως εξής:

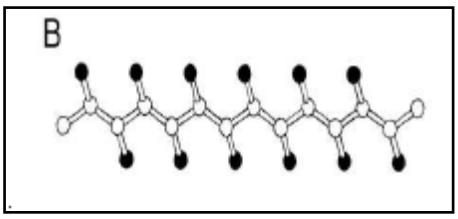
```
Dx #=X5-X1, Dy #=Y5-Y1, Dz #=Z5-Z1,
Dx1#=X3-X1, Dy1#=Y3-Y1, Dz1#=Z3-Z1,
Dx2#=X5-X3, Dy2#=Y5-Y3, Dz2#=Z5-Z3,
Dx3#=abs(X4-X1), Dy3#=abs(Y4-Y1), Dz3#=abs(Z4-Z1),
Dx4#=abs(X5-X2), Dy4#=abs(Y5-Y2), Dz4#=abs(Z5-Z2),
count(0,[Dx,Dy,Dz],#=,1),
count(0,[Dx1,Dy1,Dz1],#=,2),
count(0,[Dx2,Dy2,Dz2],#=,2),
Dx3 #=<2, Dy3 #=<2, Dz3 #=<2,
Dx4 #=<2, Dy4 #=<2, Dz4 #=<2.
```

Εξαιτίας της αλληλεπίδρασης των περιορισμών, η λίστα `[Dx,Dy,Dz]` καταλήγει να είναι το διάνυσμα `(±2,±2,0)` για οποιαδήποτε συνδυασμό των συντεταγμένων των 5 σημείων. Για κάθε αμινοξύ s_i στην θέση $\{X_i,Y_i,Z_i\}$, η θέση του αμινοξέως s_{i+5} θα είναι $\{X_i+Dx,Y_i+Dy,Z_i+Dz\}$. Έτσι, περιορισμοί της μορφής $Dx\#=X_{i+5}-X_i$, $Dy\#=Y_{i+5}-Y_i$, $Dz\#=Z_{i+5}-Z_i$ προστίθενται για κάθε επόμενο ζεύγος αμινοξέων s_i, s_{i+5} στην έλικα για να εξασφαλίσουν την διάδοση του σχήματος της έλικας.

β-πτυχώσεις: η μοντελοποίηση μιας β-πτύχωσης (**Εικόνα 5.4**) είναι πολύ απλούστερη. Με δεδομένα 3 συνεχόμενα αμινοξέα με συντεταγμένες $\{X_1,Y_1,Z_1, \ldots ,X_3,Y_3,Z_3\}$, για να σχηματιστεί μια γωνία ±90◦ που περιγράφει την πτύχωση, το πρώτο και το τελευταίο σημείο πρέπει να έχουν δύο ίδιες συντεταγμένες και μια συντεταγμένη σε απόσταση 2, γεγονός που περιγράφεται από τους παρακάτω όρους:

```
Dx #= X3-X1, Dy #= Y3-Y1, Dz #= Z3-Z1,
count(0,[Dx,Dy,Dz],#=,2),
abs(Dx + Dy + Dz) #=2,
```

Οι παραπάνω όροι επιβάλλονται και στις υπόλοιπες τριάδες αμινοξέων της πτυχωτής επιφάνειας με ισότητες διανυσμάτων, χρησιμοποιώντας ένα βήμα 3 αμινοξέων.



Εικόνα 5.4 Β-πτυχωτή επιφάνεια [50]

Γέφυρα κυστεινών: η υλοποίηση αυτών των δομών στο [24] έγινε επιβάλλοντας μια μέγιστη Ευκλείδεια απόσταση 6 ανάμεσα στα δύο αμινοξέα που συνδέονταν με γέφυρα με τον παρακάτω περιορισμό:

$abs(X_1-X_2) + abs(Y_1-Y_2) + abs(Z_1-Z_2) \#=< 6$

Επιλέχθηκε να αντικατασταθεί ο περιορισμός της Ευκλείδειας απόστασης 6, με την προσέγγιση τα δύο σημεία να απέχουν σε όλες τις συντεταγμένες τους απόσταση έως 4, δηλ:

$Dx\#=abs(X_1-X_2),\ Dy\#=abs(Y_1-Y_2),\ Dz\#=abs(Z_1-Z_2),$
$Dx \#=< 4,\ Dy \#=< 4,\ Dz \#=< 4.$

5.9 Πειραματικά αποτελέσματα

Το αναθεωρημένο πρόγραμμα ελέγχθηκε καταρχήν με μικρομεσαίες πρωτεΐνες, των οποίων η ταυτότητα προέρχεται από την Τράπεζα Δεδομένων Πρωτεϊνών PDB [10], απ' όπου μπορεί να αποκτηθεί τόσο η πρωτοταγής όσο και η δευτεροταγής δομή μιας πρωτεΐνης. Τα υπολογιστικά αποτελέσματα αναφέρονται στον **Πίνακας 5.2** (χρησιμοποιήθηκε SICStus PROLOG 3.9.1 [38] και ένα PC Intel Core 3 3.2GHz, 4GB RAM). Τα σύμβολα που χρησιμοποιούνται για την δευτερογενή δομή είναι, 'gefyra' για την γέφυρα, 'pryxwsh' για την πτύχωση και 'elika' για την έλικα.

Για να ελεγχεί η ποιότητα της πρόβλεψης χρησιμοποιήθηκε ως μέτρο η μέση τετραγωνική απόκλιση (Root Mean Square Deviation - RMSD) των ατόμων α-άνθρακα του κορμού της πρωτεΐνης. Το RMSD εκφράζει την δομική διαφορά δύο πρωτεϊνών μετρώντας την μέση απόσταση ανάμεσα στα άτομα α-άνθρακα δύο πρωτεϊνών. Ως μοντέλο αναφοράς χρησιμοποιήθηκε αυτό της Τράπεζας Δεδομένων Πρωτεϊνών [10] (όπου υπάρχει ένα πλήρες μοντέλο ατόμων με πραγματικές συντεταγμένες), εξάγοντας τον κορμό της πρωτεΐνης για να επιτευχθεί μια ισοδύναμη αλυσίδα ατόμων α-άνθρακα. Όταν συγκρίνονται τα αποτελέσματα λαμβάνεται υπόψη η απόσταση ανάμεσα στα ζευγάρια των αντίστοιχων ατόμων α-άνθρακα, δηλ. ανάμεσα στην πρόβλεψη του προγράμματος και σε μια βέλτιστα υπερτιθέμενη δομή που εξάγεται από την Τράπεζα Δεδομένων Πρωτεϊνών.

Μελετώντας τα αποτελέσματα παρατηρείται ενδεικτικά ότι για την (1EDP) (17 αμινοξέα) το RMSD σφάλμα είναι 4,2 Å, ενώ για την (1ZDD) (34 αμινοξέα) είναι 3,9 Å. Η πρωτεΐνη (1ZDD) συντίθεται κυρίως από α-έλικες που μπορούν να προσεγγισθούν καλά στο πλέγμα, γι' αυτό το σφάλμα είναι μικρότερο αν και πρόκειται για μεγαλύτερη πρωτεΐνη. Είναι αξιοσημείωτο ότι για τις περισσότερες μικρές πρωτεΐνες με μήκος από 12 ως 40 αμινοξέα, η απόκλιση είναι μάλλον μικρή γύρω στα 5.0 Å, δηλ. της τάξης απόστασης ανάμεσα σε δύο συνεχόμενα αμινοξέα. Για την πρωτεΐνη (1ED0) (46 αμινοξέα) υπάρχει μια απόκλιση 7.3 Å, ενώ για την μεγαλύτερη πρωτεΐνη (1ENH) (54 αμινοξέα), μια πρωτεΐνη που προσκολλάται στο DNA με πιο περίπλοκη δομή, υπάρχει μια απόκλιση 10 Å.

Είναι προφανές ότι τα αποτελέσματα επηρεάζονται από τα λάθη που ενυπάρχουν στην χρήση του ΠΚΚ πλέγματος και είναι ανεξάρτητα από την ποιότητα των στρατηγικών λύσης που χρησιμοποιούνται. Τα λάθη αυτά είναι δυνατόν να περιοριστούν εκτελώντας κάποια βήματα μοριακής δυναμικής στο αρχικό αποτέλεσμα στο ΠΚΚ πλέγμα – όπως την γενικευμένη μέθοδο Born [39] που υλοποιείται στο πρόγραμμα CHARMM [40]. Μετά από τέτοια διαδικασία, το μόριο με μια πλήρη περιγραφή ατόμων είναι κοντύτερα στο

τοπικό ελάχιστο ενέργειας και γενικά περιορίζει τα σφάλματα που προέρχονται από την χρήση του ΠΚΚ πλέγματος.

Πρωτεΐνη	Μέγεθος	Δευτερεύουσα Δομή	Χρόνος	Ενέργεια	RMSD
1LE0	12	[ptyxwsh(2,4),ptyxwsh(9,11)]	0.54s	-6654	3.9
1KVG	12	[gefyra(2,11),ptyxwsh(2,4),ptyxwsh(9,11)]	2.34s	-14234	3.3
1LE3	16	[ptyxwsh(2,6),ptyxwsh(11,15)]	4.35s	-10984	5.1
1EDP	17	[gefyra(1,15),gefyra(3,11),elika(9,15)]	0.45s	-16112	4.2
1PG1	18	[gefyra(6,15),gefyra(8,13), ptyxwsh(4,9),ptyxwsh(12,17)]	0.52s	-26546	3.8
1ZDD	34	[gefyra(5,34),elika(3,13),elika(20,33)]	3m 13s	-9542	3.9
1VII	36	[elika(4,8),elika(15,18),elika(23,32)]	3h 23m	-24124	9.2
2GP8	40	[elika(6,21),elika(26,38)]	1m 5s	-17126	4.8
1ED0	46	[gefyra(3,40),gefyra(4,32),gefyra(16,26), elika(7,18),elika(23,30), ptyxwsh(2,4),ptyxwsh(33,34)]	12h.14m.	-27232	7.1
1ENH	54	[elika(8,20),elika(26,36),elika(40,52), ptyxwsh(22,23)]	2h 13m	-26764	9.5
2ERA	61	[ptyxwsh(2,5), ptyxwsh(13,17), ptyxwsh(34,41), ptyxwsh(23,31), ptyxwsh(50,55)]	17h 14m	-37010	10.6

Πίνακας 5.2 Πειραματικά αποτελέσματα

Μελετώντας τώρα την επίδραση που έχει η προτεινόμενη αναπαράσταση της δευτερεύουσας δομής σε σχέση με την μοντελοποίηση που προτείνεται στο [24] σε κοινές πρωτεΐνες και σε ίδιο Η/Υ (Intel Core 3 3GHz, 4GB RAM) προκύπτει ο συγκριτικός **Πίνακας 5.3**. Ο πίνακας αυτός δείχνει τις διαφορές στις στήλες ‘Το σύστημα στο [24] και ‘Το προτεινόμενο σύστημα’. Τα αποτελέσματα δείχνουν μια σημαντική επιτάχυνση στην απόδοση της τάξης από 4 ως 50; Μόνο σε μια περίπτωση παρατηρήθηκε επιβράδυνση. Πάντως, αν και τα αποτελέσματα είναι καλά για σχετικά μικρές πρωτεΐνες, πάλι δεν δίνουν ιδιαίτερα σημαντική επιτάχυνση για μεγαλύτερες πρωτεΐνες. Για το λόγο αυτό, στην επόμενη ενότητα δοκιμάζονται νέα heuritstic αναζήτησης που πιθανολογείται η ικανότητά τους να διερευνούν πιο αποδοτικά το δέντρο αναζήτησης που παράγουν οι μεγαλύτερες πρωτεΐνες.

Πρωτεΐνη	Μέγεθος	Το προτεινόμενο σύστημα	Το σύστημα στο [24]	Επιτάχυνση
1LE0	12	0.54s	2.8s	5.2
1KVG	12	2.34s	9.8s	4.2
1LE3	16	4.35s	3.9s	0.9
1EDP	17	0.45s	22.23s	49.4
1PG1	18	0.52s	13.88s	26.7
1ZDD	34	3m 13s	26m 40s	8.6

Πίνακας 5.3 Σύγκριση απόδοσης με το σύστημα στο [24]

5.10 Διερευνώμενες στρατηγικές αναζήτησης

Ακόμα και μετά την περικοπή του χώρου αναζήτησης που επιτυγχάνουν οι περιορισμοί της δευτερεύουσας δομής, ο χώρος αναζήτησης παραμένει πολύ μεγάλος όταν χρησιμοποιηθούν μεγάλες πρωτεΐνες. Οι πρωτεΐνες που μπορεί να βρεθεί με ικανοποιητική ακρίβεια και σε λογικό χρόνο η 3D δομή τους έχουν μέγεθος περίπου 40 αμινοξέα, αλλά αυτό δεν επαρκεί. Έτσι, διερευνάται η εισαγωγή κάποιων heuristics για να περιοριστεί περαιτέρω ο χώρος αναζήτησης.

Στην αρχή διερευνήθηκαν κάποιες πολλά υποσχόμενες ιδέες αναζήτησης που αναφέρονται στην βιβλιογραφία και είναι ήδη ενσωματωμένες στο CLP(FD) σύστημα ECLIPSE [41]. Για τον λόγο αυτό μεταφέρθηκε καταρχήν το πρόγραμμα από την παρούσα υλοποίηση της SICStus Prolog στο σύστημα ECLIPSE με την πραγματοποίηση μικρών αλλαγών. Παρά το γεγονός ότι οι ιδέες αυτές έμοιαζαν θεωρητικά πολλά υποσχόμενες, όταν χρησιμοποιήθηκαν απέτυχαν να προσφέρουν κάποιο πλεονέκτημα σε σχέση με την μέχρι τότε υλοποίηση.

Ειδικότερα, ελέγχησαν οι ακόλουθες στρατηγικές αναζήτησης:

I. Η Limited Discrepancy Search (LDS).

II. Η Bounded Backtracking Search (BBS).
III. Η Depth Bounded Search (DBS) σε συνδυασμό με την LDS.

Η LDS στρατηγική ψάχνει τους γείτονες μια αποδεκτής λύσης που διαφέρουν μόνο για ένα περιορισμένο αριθμό διαφορετικών επιλογών [42]. Η BBS στρατηγική περιορίζει τον αριθμό των βημάτων πισωγυρίσματος που γίνονται στο δέντρο αναζήτησης. Η DBS στρατηγική επεκτείνει πλήρως ένα καθορισμένο αριθμό επιπέδων από την ρίζα και έπειτα εξερευνεί το υπόλοιπο δέντρο με μια καθορισμένη τεχνική (LDS στα πειράματα που πραγματοποιήθηκαν).

Το ότι τελικά αυτά τα heuristics δεν ταιριάζουν στο πρόβλημα μπορεί να γίνει αντιληπτό και νοητικά ως εξής:
Ιδεατά, κάποιος θα ήθελε να εξερευνήσει το χώρο αναζήτησης λαμβάνοντας υπόψη ότι:

- μικρές αλλαγές στην αναδίπλωση μιας πρωτεΐνης δεν επηρεάζουν ιδιαίτερα την καθολική ενέργεια - που είναι το αντίθετο με αυτό που κάνει η LDS στρατηγική;
- από την στιγμή που βρεθεί μια λύση, είναι πιθανόν ότι μια άλλη καλύτερη λύση βρίσκεται σε ένα πλήρως διαφορετικό μέρος του δέντρου αναζήτησης - που είναι σε αντίθεση με αυτό που κάνει η BBS στρατηγική, η οποία εκτελεί ένα σταθερό αριθμό από πισωγυρίσματα και έτσι κρατά την αναζήτηση κοντά στην λύση που έχει ήδη βρεθεί;
- η πλήρης εξερεύνηση των πρώτων επιπέδων του δέντρου αναζήτησης (όπως γίνεται στην DBS) είναι χρονοβόρα και έτσι ένα πιο επιλεκτικό heuristic πρέπει να χρησιμοποιηθεί.

Πάντως, τα πειράματα που έγιναν έδειξαν ότι μερικές στρατηγικές αναζήτησης (π.χ., η BBS στρατηγική) μπορεί να παράγει λύσεις που είναι υποβέλτιστες σε όρους ενέργειας, αλλά σε πολύ μικρό χρόνο. Αυτό μπορεί να είναι χρήσιμο σε καταστάσεις όπου μια βέλτιστη λύση απαιτεί πάρα πολύ χρόνο και έτσι μια προσεγγιστική λύση στην βέλτιστη θα ήταν ικανοποιητική.

5.11 Διερευνώμενη στρατηγική επιλογής μεταβλητής

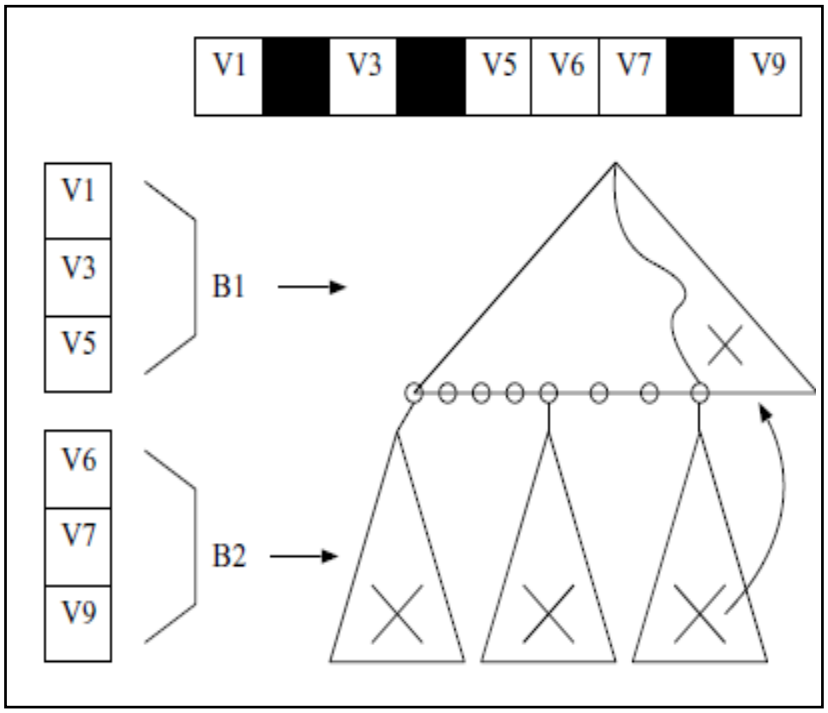
Σαν εναλλακτική της στρατηγικής επιλογής μεταβλητής ‘πρώτη αποτυχία’ (first-fail) δοκιμάστηκε η απόδοση τιμών πρώτα στις X συντεταγμένες κάθε σημείου και μετά στις άλλες δύο συντεταγμένες με την ελπίδα ότι κάποια διάδοση θα περικόψει μέρη του δέντρου αναζήτησης. Η διαίσθηση ήταν ότι η δομή πλέγματος δημιουργεί ισχυρές εξαρτήσεις ανάμεσα στις συντεταγμένες; Έτσι δίνοντας τιμή σε μια συντεταγμένη πιθανολογούταν να επηρεάζονται γρήγορα και οι υπόλοιπες συντεταγμένες. Δυστυχώς και σε αυτή την περίπτωση η απόδοση ήταν φτωχή, εξαιτίας της έλλειψης διάδοσης (δίνοντας τιμή σε μια μεταβλητή δεν είναι αρκετά επαρκές για την διάδοση στις γειτονικές μεταβλητές).

5.12 Νέα τεχνική αναζήτησης (Bounded Block Fails heuristic)

Τελικώς, διερευνήθηκε άλλο ένα νέο heuristic για την αναζήτηση του δέντρου, που ονομάζεται Bounded Block Fails (BBF). Αυτή η τεχνική αν και είναι γενική και μπορεί να εφαρμοστεί σε κάθε τύπο αναζήτησης, αποδείχθηκε εξαιρετικά αποτελεσματική όταν εφαρμόστηκε στο ΠΑΠ. Η BBF ευρετική μέθοδος για αναζήτηση λύσης πρωτοπροτάθηκε στο [99] και υλοποιείται όπως περιγράφεται παρακάτω.

Η μέθοδος εμπλέκει την έννοια του μπλοκ. Με δεδομένη μια λίστα αδέσμευτων μεταβλητών και μεταβλητών που έχουν πάρει τιμή $V = [V_1, \ldots, V_n]$, οι μεταβλητές στο σύνολο V χωρίζονται σε μπλοκ σταθερού μεγέθους k, έτσι ώστε η συνένωση όλων των μπλοκ $B_1, B_2, \ldots, B_\ell$ δίνει την ταξινομημένη λίστα των μη δεσμευμένων μεταβλητών V , όπου ℓ είναι ο αριθμός των μπλοκ και ισχύει $\ell \leq n/k$.
Δηλ. ένα μπλοκ B_i είναι μια υπολίστα του V μεγέθους k που αποτελείται από μη δεσμευμένες μεταβλητές. Η συνένωση όλων των μπλοκ $B_1, B_2, \ldots, B_\ell$ μας

δίνει την ταξινομημένη λίστα με αδέσμευτες μεταβλητές που υπάρχουν στην V ,όπου $\ell \leq n/k$. Τα μπλοκ επιλέγονται δυναμικά και εξαιρούν μερικές από τις αρχικές μεταβλητές που έχουν ήδη πάρει τιμή λόγω της διάδοσης περιορισμών. Έτσι ο αριθμός των μπλοκ μπορεί να είναι μικρότερος από n/k και μπορεί να μην είναι σταθερός κατά την διάρκεια αναζήτησης. Η **Εικόνα 5.5** δείχνει ένα παράδειγμα για $k = 3$, με μια λίστα από 9 μεταβλητές. Οι μαυρισμένες περιοχές αναπαριστούν δεσμευμένες μεταβλητές.



Εικόνα 5.5 Η BBF μέθοδος αναζήτησης [50]

Η ευρετική μέθοδος χωρίζει την αναζήτηση ανάμεσα στα ℓ μπλοκ. Εσωτερικά, κάθε μπλοκ B_i παίρνει τιμές σύμφωνα με την επιθυμητή στρατηγική τιμών. Όταν ένα μπλοκ B_i πάρει πλήρως τιμές, η αναζήτηση μετακινείται στο επόμενο μπλοκ B_{i+1}, αν υπάρχει. Αν η καταχώριση τιμών στο μπλοκ B_{i+1} αποτύχει, η αναζήτηση πισωγυρίζει στο μπλοκ B_i. Στο σημείο αυτό, υπάρχουν δύο επιλογές: αν οι φορές κατά τις οποίες το B_{i+1} έχει πλήρως αποτύχει είναι λιγότερες από ένα δεδομένο κατώφλι t_i, τότε η διαδικασία συνεχίζει παράγοντας μια ακόμα λύση για το B_i και επανεισαγωγή στο B_{i+1}. Διαφορετικά, αν έχουν συμβεί πολλές αποτυχίες, δηλ. έχει ξεπεραστεί το

κατώφλι t_i, τότε η μέθοδος παράγει μια αποτυχία για το B_i και πισωγυρίζει σε ένα προηγούμενο μπλοκ. Να σημειωθεί ότι ο αριθμός των αποτυχιών περιλαμβάνει και κανονικές αποτυχίες αναζήτησης καθώς και αυτές που προκαλούνται από την παρούσα στρατηγική. Η λίστα $t_1, \ldots, t_\ell$ με τα κατώφλια καθορίζει την συμπεριφορά της ευρετικής μεθόδου. Η **Εικόνα 5.5** έχει ως υπόθεση $t_1=3$ και δείχνει ότι μετά την τρίτη αποτυχία του B_2, η αναζήτηση στο B_1 αποτυγχάνει επίσης.

Μιλώντας διαισθητικά, το BBF heuristic προτείνει να αναθεωρήσουμε προηγούμενες επιλογές (δηλ., μια πιο δραστική αναθεώρηση της δομής που έχει παραχθεί μέχρι τότε) αντί της εξερεύνησης όλου του χώρου δυνατοτήτων κάτω από το μπλοκ που γεννά αποτυχίες (αντί δηλ. μια πιο τοπική αναθεώρηση της δομής). Σε αυτές τις περιπτώσεις, μπορούμε να διαπεράσουμε λύσεις που παράγουν αποτυχίες μπλοκ, επειδή άλλες θα ανακαλυφθούν ακολουθώντας άλλες επιλογές σε προηγούμενα μπλοκ. Στο πλαίσιο αναζήτησης λύσεων στο 3D πλέγμα, πολλές αποτυχίες στον παρόν κλάδο του δέντρου σημαίνουν ότι η μερική χωρική δομή που έχει φτιαχτεί μέχρι τότε δεν μας επιτρέπει να προχωρήσουμε χωρίς να παραβιαστούν κάποιοι περιορισμοί.

Αν υπάρχει μεγάλη πυκνότητα αποδεκτών λύσεων η BBF μέθοδος επιτρέπει να εξαιρεθούν κάποιες λύσεις, ανάλογα με τις τιμές του κατωφλιού και να έτσι να περικοπεί ο χώρος αναζήτησης χωρίς το ρίσκο να χαθεί η βέλτιστη λύση. Πάντως, η μέθοδος BBF είναι γενικά μη πλήρης, δηλ. μπορεί και να χάσει την βέλτιστη λύση.

Το συμπέρασμα που προέκυψε από την εφαρμογή της μεθόδου είναι ότι η BBF στρατηγική αποδίδει καλύτερα τόσο χρονικά όσο και ποιοτικά από την μέχρι τώρα υλοποίηση χωρίς το BBF και αυτό φαίνεται στα επόμενα πειραματικά αποτελέσματα.

5.12.1 Πειραματικά αποτελέσματα με και χωρίς το BBF heuristic

Στον **Πίνακας 5.4**, υπολογίσθηκαν οι χρόνοι εκτέλεσης και οι τιμές ενέργειας για ένα σύνολο πρωτεϊνών, με και χωρίς την εφαρμογή του BBF. Για την BBF στρατηγική, χρησιμοποιήθηκαν οι τιμές $k = 3$, $t_1 = 5$ και $t_i = 3$ για $i>=1$, οι οποίες μετά από αρκετά πειράματα αποδεδείχθηκε ότι δίνουν καλή απόδοση. Οι είσοδοι που σημειώνονται με (*) σημαίνουν εκπνοή χρόνου (δηλ. η εκτέλεση σταμάτησε μετά το πέρασμα του καθορισμένου χρόνου).

Όνομα Πρωτεΐνης	Μέγεθος Πρωτεΐνης	Χωρίς BBF		Με BBF		Επιτάχυνση
		Χρόνος	Ενέργεια	Χρόνος	Ενέργεια	
1LE0	12	0.54s	-6654	0.52s	-6112	1.03
1KVG	12	2.34s	-14234	2.14s	-13745	1.09
1LE3	16	4.35s	-10984	3.52s	-10674	1.23
1EDP	17	0.45s	-16112	0.39s	-16112	1.15
1PG1	18	0.52s	-26546	0.47s	-26546	1.10
1ZDD	34	3m 13s	-9542	3m	-9352	1.07
1VII	36	3h 23m	-24124	1h 12m	-24124	3.2
2GP8	40	1m 5s	-17126	1m 2s	-17126	1.04
1ED0	46	(*) 10h		1h 46m	-31580	
1ENH	54	2h 13m	-26764	1h 10m	-25549	2
2ERA	61	(*) 10h		4h 35m	-38206	
1SN1	63	(*) 10h		6h 42m	-42116	
1YPA	63	(*) 10h		6h 24m	-45678	
1FVS	72	(*) 10h		1h 35m	-52346	
1B4R	80	(*) 10h		2h 19m	-74378	

Πίνακας 5.4 Πειραματικά αποτελέσματα με και χωρίς το BBF

Θα πρέπει να τονιστούν δύο σημαντικές παρατηρήσεις. Καταρχήν, η υλοποίηση του BBF πραγματοποιήθηκε στην SICStus Prolog σε υψηλό

επίπεδο; έτσι όταν υπήρχε αποτυχία, καλούνταν πολλά κατηγορήματα ‘fail’, γεγονός που είχε ένα σχετικά υψηλό κόστος στο χειρισμό του δέντρου αναζήτησης. Αυτό έκανε την υλοποίηση μη αποδοτική και αυτό φαίνεται ξεκάθαρα στις μικρότερες πρωτεΐνες. Παρόλα αυτά, υπάρχουν προφανή πλεονεκτήματα για τους χρόνους εκτέλεσης και τις βέλτιστες λύσεις για τις μεγαλύτερες πρωτεΐνες. Είναι προφανές ότι μια χαμηλού επιπέδου υλοποίηση του BBF θα οδηγούσε σε πολύ καλύτερη απόδοση.

Η δεύτερη παρατήρηση είναι ότι το BBF είναι σχεδιασμένο να χειρίζεται μεγάλου μεγέθους εισόδους. Για τις μικρότερες πρωτεΐνες, ο αριθμός των μπλοκ τείνει να είναι μικρός και το BBF κάνει λίγα πισωγυρίσματα μπλοκ. Στα πειράματα χρησιμοποιήθηκε μπλοκ μεγέθους 3 (που ανταποκρίνονται στις τρεις συντεταγμένες ενός αμινοξέως). Προσπάθειες για μείωση αυτής της τιμής (δηλ., αύξηση του αριθμού των μπλοκ) δημιουργούσε πτώση της απόδοσης, εξαιτίας του πολύ μικρού μεγέθους των μπλοκ (που υπερισχύει της αρχικής ιδέας του BBF). Για τον χειρισμό μεγαλύτερων πρωτεϊνών είναι δυνατόν να χρησιμοποιηθούν μεγαλύτερα μπλοκ (π.χ. $k=9$), που περιλαμβάνουν περισσότερους κόμβους μέσα στο αντίστοιχο υποδέντρο. Έτσι, μια αποτυχία από την BBF στρατηγική περικόπτει ένα μεγαλύτερο κομμάτι του δέντρου και επιταχύνει την αναζήτηση. Παντως για να ισσοροπήσει το ποσό επιτάχυνσης με την ακρίβεια, πρέπει να συντονιστούν κατάλληλα οι παραμέτροι t_i (π.χ. με μεγάλες τιμές t_i έχουμε πιο αργά αλλά πιο ακριβή αποτελέσματα).

Πρωτεΐνες με μέγεθος μεγαλύτερο ακόμα και από 100 αμινοξέα μπορούν να επιλυθούν από το σύστημα. Το αποτέλεσμα αυτό είναι σημαντικά καλύτερο από αυτό που αναφέρεται στο [24], όπου πρωτεΐνες με μέγεθος μεγαλύτερο από 70 αμινοξέα χρειάζονται ώρες για να επιλυθούν. Αυτή η βελτίωση είναι σημαντική εξαιτίας της NP-πληρότητας του ΠΑΠ, όπου μια αύξηση στο μήκος της πρωτεΐνης μεταφράζεται σε μια εκθετική αύξηση του χρόνου εκτέλεσης. Το BBF heuristic και η νέα δομή περιορισμών παρέχει πιο αποτελεσματική περικοπή του δέντρου αναζήτησης, δίνοντας καλύτερες ποιοτικά και γρήγορες λύσεις. Η ισορροπία ανάμεσα σε ποιότητα και ταχύτητα ελέγχεται από το

κατώφλι του BBF: μεγαλύτερες τιμές παρέχουν μια πιο λεπτομερειακή αναζήτηση και άρα πιο ποιοτικές αλλά πιο αργές λύσεις.

5.13 Σύγκριση με το σύστημα HMMSTR/Rosetta

Επίσης, έγινε μια σύγκριση των παραπάνω αποτελεσμάτων με αυτά του συστήματος πρόβλεψης HMMSTR/Rosetta [44]. Το σύστημα αυτό δεν χρησιμοποιεί πλέγμα σαν χωρικό μοντέλο: τα αμινοξέα μπορούν να πάρουν οποιαδήποτε θέση στο χώρο R^3. Το σύστημα HMMSTR/Rosetta χρησιμοποιήθηκε σαν ένας εξαρχής (de novo) προγνώστης (απενεργοποιώντας τις επιλογές για ομολογία και psi-blast). Η σύγκριση βέβαια δεν είναι δίκαια επειδή στο μοντέλο της εργασίας η δευτερεύουσα δομή είναι γνωστή και δεν προβλέπεται. Οι χρόνοι του συστήματος HMMSTR/Rosetta παίρνονται από τα αρχεία αποτελεσμάτων αλλά δεν είναι ξεκάθαρο σε ποιο μηχάνημα αναφέρονται. Τα αποτελέσματα υπάρχουν στον **Πίνακας 5.5**. Η HMMSTR πρόβλεψη είναι σαφώς ταχύτερη, αλλά οι προβλέψεις της παρούσας εργασίας (που περιέχουν γνωστές δευτερεύουσες δομές) βελτιώνουν το RMSD για μεγαλύτερα προβλήματα.

Όνομα Πρωτεΐνης	Μέγεθος Πρωτεΐνης	Το σύστημά μας		Το σύστημα Rosetta	
		Χρόνος	RMSD	Χρόνος	RMSD
1ZDD	34	3m 43s	5.2	5m.35s	3.5
1VII	36	1h 12m	5.6	5m.35s.	4.2
2GP8	40	1m 2s	3.6	6m.35s.	6.4
1ED0	46	1h 46m	6.2	7m.23s.	8.9

Πίνακας 5.5 Συγκρίσεις με τις προβλέψεις του συστήματος Rosetta

6 Επίλογος

Με το κεφάλαιο αυτό ολοκληρώνεται η παρούσα διπλωματική εργασία. Στην Ενότητα 6.1 γίνεται μία σύνοψη της εργασίας και παρουσιάζονται κάποια βασικά συμπεράσματα. Στην ενότητα 6.2 περιγράφονται ορισμένες ενέργειες που θα μπορούσαν μελλοντικά να επεκτείνουν την εργασία αυτή.

6.1 Σύνοψη και Συμπεράσματα

Το πρόβλημα της αναγνώρισης αναδίπλωσης 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 νέες πρωτεΐνες που προκύπτουν καθημερινά να ταυτοποιηθούν εργαστηριακά, γίνεται προσπάθεια πρόγνωσης της 3D δομής τους με χρήση εξελιγμένων υπολογιστικών μεθόδων.

Σε αυτή την εργασία μοντελοποιήθηκε το Πρόβλημα Αναδίπλωσης Πρωτεϊνών (ΠΑΠ) σαν έναν πρόβλημα βελτιστοποίησης που αφορά την τοποθέτηση των αμινοξέων της πρωτεΐνης σε ένα ΠροσωποΚεντρικό Κυβικό (ΠΚΠ) πλέγμα - ένα βιολογικά συμβατό χωρικό μοντέλο - με ταυτόχρονη ελαχιστοποίηση της συνάρτησης ενέργειας της πρωτεΐνης, χρησιμοποιώντας έναν έτοιμο πίνακα ενεργειών. Για την υλοποίηση χρησιμοποιήθηκε το CLP(FD) παράδειγμα προγραμματισμού. Παρουσιάστηκε ένας διαφορετικός τρόπους οργάνωσης των περιορισμών σε σχέση με την μέχρι τώρα CLP(FD) προσέγγιση της βιβλιογραφίας και δοκιμάστηκαν και αναλύθηκαν διαφορετικές στρατηγικές αναζήτησης για την επίλυση. Έμφαση δόθηκε σε ένα νέο heuristic αναζήτησης, το Bounded Block Fails (BBF), που ταιριάζει στο πρόβλημα δίνοντας πολύ καλά αποτελέσματα απόδοσης, ακρίβειας και κλιμάκωσης.

Το βασικό συμπέρασμα που προκύπτει είναι ότι οι τεχνικές διαχείρισης περιορισμών που υλοποιούνται στην CLP(FD) τεχνολογία μας επιτρέπουν να πετύχουμε μια φυσική και αρκετά δηλωτική κωδικοποίηση του προβλήματος, επιτρέποντας την εύκολη τροποποίηση και επεκτασιμότητα, όπως π.χ. την απλή εισαγωγή νέων περιορισμών που αναπαριστούν επιπρόσθετη γνώση για μια πρωτεΐνη (π.χ., γνώση για πρόσθετα στοιχεία δευτερεύουσας δομής).

Η εργασία εισάγει νέες ιδέες για βελτίωση της απόδοσης, της ακρίβειας και της κλιμάκωσης σε σχέση με την CLP(FD) υλοποίηση που έχει παρουσιαστεί στο [24]. Οι ιδέες αυτές αφορούν: (1) την κατάλληλη επαναδιατύπωση των περιορισμών, κυρίως αυτών που περιγράφουν τις δευτερεύουσες δομές μιας πρωτεΐνης και (2) την διερεύνηση και υλοποίηση heuristics αναζήτησης που ταιριάζουν στο ΠΑΠ.

Η CLP(FD) υλοποίηση που προτείνεται βελτιώνει την απόδοση που έχει παρουσιαστεί στο [24] με μια επιτάχυνση της τάξης από 4 ως 50 σε μικρές και μεσαίου μεγέθους πρωτεΐνες, βελτιώνοντας και την ακρίβεια της λύσης. Όπως φαίνεται από τα πειραματικά αποτελέσματα με το προτεινόμενο μοντέλο μπορούν να επιλυθούν πρωτεΐνες μήκους περίπου 100 αμινοξέων, μέγεθος το οποίο δεν είναι εφικτό με τα υπάρχοντα CLP(FD) μοντέλα.

6.2 Μελλοντικές επεκτάσεις

Τα ενθαρρυντικά αποτελέσματα που προέκυψαν δίνουν έναυσμα για μελλοντικές επεκτάσεις της εργασίας. Η παρούσα εργασία μπορεί να αποτελέσει το σημείο έναρξης για περαιτέρω ανάλυση του προβλήματος, με στόχο την επίλυση πολύ μεγαλύτερων πρωτεϊνών με μήκος ακόμα και 500 αμινοξέα. Η δηλωτικότητα του κώδικα επιτρέπει την εύκολη υλοποίηση νέας γνώσης με την εισαγωγή νέων περιορισμών (π.χ. περιορισμούς για την πλευρική αλυσίδα, τοπολογικούς περιορισμούς, περιορισμούς για υπερδευτεροταγείς δομές, κ.α.) καθώς και την εισαγωγή νέων heuristics αναζήτησης με σκοπό την περαιτέρω περικοπή του χώρου αναζήτησης.

Μια ιδέα για διεύρυνση στην κλιμάκωση της παρούσας υλοποίησης για το χειρισμό μεγαλύτερων πρωτεϊνών θα μπορούσε να είναι η εξής: Όταν χειριζόμαστε μεγαλύτερες πρωτεΐνες συνήθως συναντάμε καταστάσεις όπου οι διαμορφώσεις διαφόρων υποακολουθιών είναι ήδη γνωστές (π.χ. από ομολογία). Τότε, το ΠΑΠ γίνεται ισοδύναμο με την πρόβλεψη τοποθέτησης λίγων διαμορφωμένων υποακολουθιών που συνδέονται από Μη Κανονικές Δομές (coils). Χρησιμοποιώντας τις Ευκλείδειες αποστάσεις ανάμεσα σε κάθε ζευγάρι αμινοξέων στις γνωστές υποακολουθίες σαν επιπρόσθετους περιορισμούς στο υπόλοιπο μοντέλο περιορισμών, η παρούσα υλοποίηση θα μπορούσε να δώσει μια ικανοποιητική λύση σε μεγαλύτερες πρωτεΐνες.

Η δηλωτική φύση της CLP(FD) κωδικοποίησης δίνει μια πλήρη ελευθερία στην εξερεύνηση εναλλακτικών στρατηγικών αναζήτησης, μια εκ των οποίων θα μπορούσε να ήταν η χρήση παράλληλης αναζήτησης. Ειδικότερα, η εξερεύνηση του χώρου αναζήτησης λύσεων είναι μια καθαρά μη-ντετερμινιστική διαδικασία, με περιορισμένη αλληλεπίδραση ανάμεσα στους διακριτούς κλάδους του δέντρου αναζήτησης, αφού κάθε κλάδος οδηγεί σε μια πιθανά διαφορετική λύση του προβλήματος. Αυτό είναι ένα πολύ βολικό πλαίσιο για την παράλληλη διερεύνηση διαφορετικών κλάδων του δέντρου αναζήτησης, που θα εκτελούνται από διαφορετικά στιγμιότυπα της μηχανής επίλυσης περιορισμών.

Ιδιαίτερα ενδιαφέρουσα θα ήταν επίσης η υλοποίηση ενός ειδικού συστήματος περιορισμών που διαθέτει γηγενείς μεταβλητές με διακριτές τιμές 3D σημεία του πλέγματος και πρωτογενείς περιορισμούς βασικές χωρικές σχέσεις μέσα στην δομή του πλέγματος. Σχεδιάζοντας ειδικές τεχνικές επίλυσης και διάδοσης περιορισμών και υλοποιώντας κάποια από τα προτεινόμενα heuristics, όπως το BBF, σε χαμηλό επίπεδο μπορεί να επιτευχθεί πολύ καλύτερη απόδοση και κλιμάκωση λύσεων.

7 Βιβλιογραφία

[1] Κ. Ε. Κ. Νικήτα, Προσομοίωση Φυσιολογικών Συστημάτων, Αθήνα : ΕΜΠ, 2007.

[2] M. N. a. J. P. Linge, Unité de Bio–Informatique, Paris : Institut Pasteur.

[3] Dictionary, Oxford English. http://www.oed.com, Oxford University Press.

[4] Εμμ.Φραγκούλης, Βιοχημεία, Κ.Ε. Σέκερης, 1984.

[5] Α. Α. Κ. Α. Π. Π. Αλεπόρου-Μαρίνου Β., Βιολογία Θετικής Κατεύθυνσης Γ' Τάξης Ενιαίου Λυκείου, Αθήνα: ΟΕΔΒ, 2005.

[6] Μ. Καρμιράντζου, «Μαθήματα Βιοφυσικής,» http://biophysics.biol.uoa.gr/courses-Πανεπιστημιο Αθηνων, Τμήμα Βιολογίας, Αθήνα, 2008.

[7] Pauling L. & Corey R.B., “The Pleated Sheet , A New Layer Configuration of Polypeptide Chains”, 1951.

[8] Pauling, L . , Corey, R. B., & Branson, H. R. , “The Structure of Proteins: Two Hydrogen-Bonded Hel ical Configurations of the Polypeptide Chain”, 1951.

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

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

[11] Σπυρούλιας, Γ.; Μοριακή Προσομοίωση. 2008, Τμήμα Φαρμακευτικής, Πανεπιστήμιο Πατρών.

[12] Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M. S.; Eramian, D.; Shen, M. Y.; Pieper, U.; Sali, A.; Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics 2006, Chapter 5, Unit 5 6.

[13] K. Marriott and P. J. Stuckey. Programming with constraints. MIT Press, 1998.

[14] K. R. Apt. Principles of constraint programming. Cambridge University press, 2003.

[15] R. Kowalski. Predicate logic as logic programming language. Proc. of IFIP, congress 74, 569–574, 1974.

[16] M. Paterson and W. Wegman. Linear unification. J. of Computer and System Sciences, 16(2):158–167, 1978.

[17] M. Berrera, H. Molinari, and F. Fogolari. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. BMC Bioinformatics, 4(8), Epub, 2003.

[18] B. Berger and T. Leighton, «Protein folding in the hydrophobic-hydrophilic (HP) model is NP-comolete,» σε *Proc. of Second Annual International Conferences on Computational Molecular Biology (RECOMB'98) pp. 30-39*, New York, 1998.

[19] P. Crescenzi, D. Goldman, C. Papadimitrou, A. Piccolboni, and M. Yannakakis, «On the complexity of protein folding,» σε *Proc. of Symposium on Theory of Computing (STOC)*, ACM Press, 1998, pp. pp. 597-603.

[20] R. Agarwala, S. Batzoglou, V. Danˇc´ık, S. Decatur, S. Hannenhalli, M. Farach, S. Muthukrishnan, and S. Skiena, «Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model.,» *Journal of Computational Biology,* pp. 4:275-296, 1997.

[21] J. Skolnick and A.Kolinski, «Reduced models of proteins and their applications,» *Polymer,* pp. 45:511-524, 2004.

[22] L. Toma and S. Toma, «Folding simulation of protein models on the structure-based cubo-octahedral lattice with contact interactions algorithm,» *Protein Science,* pp. 8:196-202, 1999.

[23] B. Rost and C. Sander, «Prediction of protein secondary structure at better than 70% accuracy,» *Journal of Molecular Biology,* pp. 232:584-599, 1993.

[24] A. Dal Pal`u, A. Dovier, and F. Fogolari, Constraint logic programming approach to protein structure prediction, BMC Bioinformatics, 2004.

[25] R. Bonneau and D. Baker, «Ab initio protein structure prediction: progress and prospects,» *Annual Review of Biophysics and Biomolecular Structure,* pp. 30:173-189, 2001.

[26] J. Skolnick and A. Kolinski, « Computational studies of protein folding,» *Computing in Science and Engineering,* pp. 3(5):40-50, 2001.

[27] Committee on Mathematical Challenges from Computational Chemistry, «Mathematical challenges from theoretical-computational chemistry,» National Research Council, 1995.

[28] R. Backofen, «The protein structure prediction problem: a constraint optimization approach using a new lower bound,» *Constraints,* pp. 6(2-3):223-255, 2001.

[29] R. Backofen and S. Will, «A constraint-based approach to structure prediction for simplified protein models that outperform other existing methods.,» σε *International Conference on Logic Programming*, Springer Verlag, 2003, pp. pp. 49-71.

[30] R. Backofen and S. Will., « Fast, constraint-based threading of HP sequences to hydrophobic cores.,» σε *Principles and practice of Constraint Programming*, Springer Verlag, 2001, pp. pp. 495-508.

[31] A. Newman., «A new algorithm for protein folding in the HP model,» σε *Symposium on Discrete Algorithms*, ACM Press, 2002, p. pp. 876–885.

[32] W. Hart and S. Israil. Fast protein folding in the hydrophobic-hytrophilic model within three-eighths of optimal. J. of Computational Biology, pages 53–96, 1996.

[33] G. Raghunathan and R. L. Jernigan. Ideal architecture of residue packing and its observation in protein structures. Protein Science, 6:2072–2083, 1997.

[34] M. T. Swain and G. J. L. Kemp. A CLP approach to the protein side-chain placement problem. In T.Walsh, editor, CP'01, LNCS vol. 2239, pages 479-493,2001.

[35] D. Clark, J. Shirazi, and C. Rawlings, «Protein topology prediction through constraint-based search and the evaluation of topological folding rules,» *Protein Engineering,* pp. 4:752-760, 1991.

[36] S. Forman, Torsion Angle Selection and Emergent Non-local secondary structure in protein structure prediction. PhD thesis, University of Iowa, 2001.

[37] S. Miyazawa and R. L. Jernigan, «Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading.,» *Journal of Molecular Biology,* pp. 256(3):623-644, 1996.

[38] Swedish Institute of Computer Science, , «SICStus Prolog User's Manual, Release 3.12.,» Swedish Institute of Computer Science, , 2005.

[39] D. Qiu, P. Shenkin, F. Hollinger and W. Still. The gb/sa continuum model for solvation. A fast analytical method for the calculation of approximate born radii. J. Phys. Chem., 101:3005–3014, 1997.

[40] B. R. Brooks et al.Charmm: A program for macromolecular energy minimization and dynamics calculations.J. Comput. Chem., 4:187–217, 1983.

[41] A. M. Cheadle et al. ECLiPSe: An Introduction. Technical Report IC-Parc 03–1, IC-Parc, 2003.

[42] M.L.G. William and D. Harvey. Limited discrepancy search. IJCAI, pages 607–615, Morgan Kaufmann, 1995.

[43] A. Dal Pal`u, A. Dovier, and E. Pontelli, «Heuristics, optimizations, and parallelism for protein structure prediction in CLP(FD),» σε *Principles and Practice of Declarative Programming*, ACM Press, 2005, pp. 230-241.

[44] K. Simons et al. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999, 3:171–176.

[45] Μ. Καφώρου, «Αναγνώριδη Διπλώματος Πρωτεινών με χρήση Δικτύων Ομοιότητας και Γενετικών Αλγορίθμων», Διπλωματική εργασία, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ, ΕΜΠ:2009.

[46] Ε. Παππάς, «Ανάπτυξη Συστήματος Πρόβλεψης Δευτεροταγούς δομής πρωτεινών με χρήση τεχνικών κοντινότερων γειτόνων και πιθανοτικών νευρωνικών δικτύων», Διπλωματική εργασία, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ, ΕΜΠ:2004

[47] Κατσιάνης Ν., «Συγκριτική ανάλυση αλγορίθμων ανίχνευσης διαμεμβρανικών τμημάτων σε πρωτεΐνες», Διπλωματική εργασία, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ, ΑΠΘ:2007

[48] Δημητρόπουλος Ν., «Υπολογιστική μελέτη Δομής και Δυναμικής βιομοριακών συμπλόκων της α1 υπομονάδας του νικοτινικού υποδοχέα της Ακετυλοχολίνης (nAChR) με άλφα-νευροτοξίνες», Διπλωματική εργασία, Τμήμα Φαρμακευτικής, Πανεπιστήμιο Πατρών:2010

[49] http://www.chemie.fu-berlin.de/chemistry/bio/amino-acids_en.html, Ινστιτούτο Χημείας, Τμήμα Βιολογίας-Χημείας_Φαρμακευτικής, Βερολίνο

[50] Alessandro Dal Palu, «Constraint Programming Approaches to the Protein Structure Prediction Problem», Phd. thesis, 2006

[51] Ι. Βλαχάβας, Π.Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας, Η. Σακελλαρίου, «Τεχνητή Νοημοσύνη – Β' τόμος», Κεφ. 6 – Ικανοποποίηση Περιορισμών.

8 Παράρτημα 1 – Κύρια Κατηγορήματα

Στο παράρτημα αυτό παρατίθεται ο CLP(FD) κώδικας υλοποίησης του συστήματος, με σχολιασμό των βασικότερων κατηγορημάτων.

```
% SICStus Prolog Κώδικας για το Πρόβλημα Αναδίπλωσης Πρωτεινών
σε CLP(FD)
% Τελευταία Τροποποίηση: 20 Αυγούστου 2014


:-use_module(library(clpfd)).
:-use_module(library(lists)).
:-use_module(library(terms)).
:-use_module(library(timeout)).
:-use_module(library(system),[]).
:-consult('prwteines.pl').

% ενημερώνει το κατάλογο εργασίας στον κατάλογο του καλούντος
προγ/τος που καλεί
:-system:working_directory(A,A), write(Katalogos ergasias:
'),write(A),nl.


% ΚΥΡΙΟ ΚΑΤΗΓΟΡΗΜΑ: anadiplosi/1-2-3
% ΕΙΣΟΔΟΣ: ο κωδικός της πρωτεΐνης από το αρχείο prwteines.pl
%        Ένα χρονικό όριο για την αναζήτηση σε secs
%        Ο παράγοντας συμπύκνωσης της λύσης
% ΕΞΟΔΟΣ: Η ελάχιστη ενέργεια που δίδει η διαμόρφωση που
% βρέθηκε
% ΠΑΡΑΔΕΙΓΜΑ:  anadiplosi ('1ed0').
%         Εξ' ορισμού χρονικό όριο : 24 ώρες,
%          Εξ' ορισμού παράγοντας συμπύκνωσης:  4*N/5
%           anadiplosi ('1ED0',120).
%         Χρονικό όριο: 120secs,
%          Εξ' ορισμού παράγοντας συμπύκνωσης:  4*N/5
%           anadiplosi ('1ED0',120,0.6).
%         Χρονικό όριο: 120secs, παράγοντας συμπύκνωσης: 0.6

anadiplosi ( Kwdikos ):-
      prwteinh(Kwdikos, Prwtogenhs, _),
      length(Prwtogenhs, N),
      CF is (4.4*N**0.38)/N,
      anadiplosi (Kwdikos, 36000, CF).

anadiplosi ( Kwdikos, Xronos ):-
      prwteinh(Kwdikos, Prwtogenhs, _),
      length(Prwtogenhs, N),
```

```
      CF is (4.4*N**0.38)/N,
      anadiplosi (Kwdikos, Xronos, CF).

anadiplosi ( Kwdikos, Xronos, Sympyknwsh ):-
      arxikopoihsh,
      prwteinh(Kwdikos, Prwtogenhs, Deuterogenhs),
      Arxikos_xronos,
      periorismoi(Prwtogenhs, Deuterogenhs, Tritogenhs,
            Energeia, Sympyknwsh, Taks_tritogenhs,
            Lista_me_Energeies),
      emfanise_xrono('Χρόνος Περιορισμών: '),nl,!,
      anazhthsh_lyshs(Kwdikos, Prwtogenhs, Deuterogenhs,
            Sympyknwsh, Xronos, Tritogenhs, Energeia,
            Lista_me_Energeies),
      emfanise_xrono('Συνολικός χρόνος: '),
      emfanise_apotelesma(Prwtogenhs,Tritogenhs),
      write('Ενέργεια = '), write(Energeia).



%$$$$$$$$$$$$$$$$$$$$
% ΟΡΙΣΜΟΣ ΠΕΡΙΟΡΙΣΜΩΝ
%$$$$$$$$$$$$$$$$$$$$

%  ΚΑΤΗΓΟΡΗΜΑ: periorismoi
%  ΕΙΣΟΔΟΣ: Prwtogenhs, Deuterogenhs
%  ΕΞΟΔΟΣ: Οι λίστες Tritogenhs,
%          η περιορισμένη τιμή Energeia,
%          η λίστα Lista_me_Energeies με την συνεισφορά των
επαφών στην ενέργεια, και
%          μια σειρά περιορισμών στις μεταβλητές

periorismoi(Prwtogenhs, Deuterogenhs, Tritogenhs, Energeia,
Sympyknwsh, Taks_tritogenhs,  Lista_me_Energeies):-
      elegxos(Prwtogenhs),
      length(Prwtogenhs,N),
      dhmiourgia_tritogenhs (N, Tritogenhs),
      pedio_orismou(Tritogenhs, N),
      apofygh_brogxwn(Tritogenhs, N),
      periorismoi_epomenou(Tritogenhs),
      periorismoi_apostashs(Tritogenhs),
      periorismoi_sympyknwshs(Tritogenhs,N, Sympyknwsh),
      periorismoi_gwnies(Tritogenhs),
      deutereuousa_domh(Deuterogenhs, Tritogenhs),
      apofygh_symmetriwn(N, Tritogenhs),
      periorismoi_energeias(Prwtogenhs, Tritogenhs,
            Energeia, Lista_me_Energeies).


% ΚΑΤΗΓΟΡΗΜΑ: dhmiourgia_tritogenhs/2
% ΕΙΣΟΔΟΣ: N, αριθμός >= 0, το μήκος της πρωτογενούς αλυσίδας
αμινοξέων
% ΕΞΟΔΟΣ: μια λίστα με τριπλέτες N μεταβλητών που
χρησιμοποιούνται σαν συντεταγμένες

dhmiourgia_ Tritogenhs (N, Tritogenhs):-
```

```
    Ntripl is N*3,
    length(Tritogenhs,Ntripl).


% PREDICATE: pedio_orismou/2
% ΕΙΣΟΔΟΣ: N, αριθμός >= 0,
%        Tritogenhs, η λίστα με τις τριπλέτες των μεταβλητών
% ΑΠΟΤΕΛΕΣΜΑ: προσθέτει περιορισμούς στις μεταβλητές ώστε οι
τιμές τους να είναι στο διάστημα 0 .. 2*N

pedio_orismou([X,Y,Z |Ypoloipes], N) :-
     MHKOS_KYBOU is N * 2,
     domain([X,Y,Z],0,MHKOS_KYBOU),
     sum([X,Y,Z], #=, Athroisma),
     artio(Athroisma),
     pedio_orismou(Ypoloipes, N).
pedio_orismou([], _).



% ΚΑΤΗΓΟΡΗΜΑ: apofygh_brogxwn/2 (Βοηθητικό:
theseis_se_akeraious/3)
% ΕΙΣΟΔΟΣ:  Tritogenhs, η λίστα με τις τριπλέτες μεταβλητών
και το N μέγεθος της πρωτείνης
% ΑΠΟΤΕΛΕΣΜΑ: προσθέτει περιορισμούς στις μεταβλητές
%        για την αποφυγή ίσων τριπλετών

apofygh_brogxwn(Tritogenhs, N):-
     ARITHMOS is 2*N+1,
     theseis_se_akeraious(Tritogenhs, Eswteriki_lista,
          ARITHMOS),
     all_different(Eswteriki_lista).

theseis_se_akeraious([[X,Y,Z]|R],[I|S], ARITHMOS):-
    I #= X* ARITHMOS * ARITHMOS +Y* ARITHMOS +Z,
    theseis_se_akeraious( R , S, ARITHMOS).
    theseis_se_akeraious([],[], _).



% ΚΑΤΗΓΟΡΗΜΑ: periorismoi_epomenou/1 (Βοηθητικό: epomeno/6)
% ΕΙΣΟΔΟΣ: Tritogenhs, η λίστα με τις τριπλέτες μεταβλητών
% ΑΠΟΤΕΛΕΣΜΑ: προσθέτει περιορισμούς στις μεταβλητές
%        για να εξασφαλίσει ότι οι τριπλέτες είναι συνεχόμενες

periorismoi_epomenou([]).
periorismoi_epomenou([_]).
periorismoi_epomenou([A,B|C]) :-
    epomeno(A,B),
    periorismoi_epomenou([B|C]).
epomeno([X_i,Y_i,Z_i],[X_si,Y_si,Z_si]):-
      domain([Dx,Dy,Dz],0,1),
      DX #= abs(X_i - X_si),
      DY #= abs(Y_i - Y_si),
      DZ #= abs(Z_i - Z_si),
```

```
       sum([DX,DY,DZ], #= ,  2).


% ΚΑΤΗΓΟΡΗΜΑ: periorismoi_apostashs/1 (Βοηθητικό: apostash/4,
oxi_epomeno/6)
% ΕΙΣΟΔΟΣ: Tritogenhs, η λίστα με τις τριπλέτες μεταβλητών
% ΑΠΟΤΕΛΕΣΜΑ: προσθέτει περιορισμούς στις μεταβλητές
%        για να εξασφαλίσει ότι οι μη συνεχόμενες τριπλέτες
%        έχουν απόσταση στο πλέγμα > 1.
% ΣΗΜΕΙΩΣΗ: Εισάγει μη γραμμικούς περιορισμούς

periorismoi_apostashs([_]).
periorismoi_apostashs([P1,P2|R]):-
    apostash(P1,R),
    periorismoi_apostashs([P2|R]).

apostash(_,[]).
apostash(P1,[P2|R]):-
    oxi_epomeno(P1,P2),
    apostash(P1,R).

oxi_epomeno([X1,Y1,Z1],[X2,Y2,Z2]):-
      Dx #= (X1-X2) * (X1-X2),
      Dy #= (Y1-Y2) * (Y1-Y2),
      Dz #= (Z1-Z2) * (Z1-Z2),
      sum([Dx,Dy,Dz], #>, 2).


%$$$$$$$$$$$$$$$$$$$$
% ΚΑΤΗΓΟΡΗΜΑ: periorismoi_sympyknwshs /3
% ΕΙΣΟΔΟΙ: Tritogenhs, η λίστα με τις τριπλέτες των
μεταβλητών, το μέγεθος της πρωτεΐνης N και CF ο παράγοντας
συμπύκνωσης
% ΑΠΟΤΕΛΕΣΜΑ: προσθέτει FD περιορισμούς σε εκείνες τις
μεταβλητές για να εξασφαλίσει ότι η γραμμική απόσταση ανάμεσα
στα αμινοξέα είναι μικρή
%$$$$$$$$$$$$$$$$$$$$


periorismoi_sympyknwshs (Tritogenhs, N, Sympyknwsh):-
     Apostash is integer(N*Sympyknwsh),
     apost_periorismoi(Tritogenhs, Apostash).


apost_periorismoi([],_).
apost_periorismoi([_,_,_],_).
apost_periorismoi([X1,Y1,Z1,X2,Y2,Z2|R], Apostash) :-
    apost_bohth(X1,[X2,Y2,Z2|R], Apostash),
    apost_bohth(Y1,[Y2,Z2|R], Apostash),
    apost_bohth(Z1,[Z2|R], Apostash),
    apost_periorismoi([X2,Y2,Z2|R], Apostash).
apost_bohth(_,[],_).
apost_bohth(Z1,[Z2], Apostash):-
    orio(Z1,Z2, Apostash).
apost_bohth(Y1,[Y2,_], Apostash):- %% Y
```

```
    orio(Y1,Y2, Apostash).
apost_bohth(X1,[X2,_,_|R], Apostash):- %% X ή Y ή Z
    orio(X1,X2, Apostash),
    apost_bohth(X1,R, Apostash).
orio(A,B,Dist):-
  A - B #< Apostash,
  B - A #< Apostash.


%$$$$$$$$$$$$$$$$$$$$
% ΚΑΤΗΓΟΡΗΜΑ: periorismoi_gwnies /1
% ΕΙΣΟΔΟΣ: Η τριτοταγής δομή.
% ΑΠΟΤΕΛΕΣΜΑ: Περιορίζουμε τα δύο διανύσματα των σχετικών
αποστάσεων να είναι:
% 1) διαφορετικά ώστε να μην σχηματίζεται γωνία κάμψης 180
μοίρες
% 2) το βαθμωτό τους γινόμενο να είναι >= 0, ώστε να μην
επιτρέπεται γωνία 60 μοιρών η οποία δίνει τιμή -1
%$$$$$$$$$$$$$$$$$$$$

periorismoi_gwnies([X1,Y1,Z1,X2,Y2,Z2,X3,Y3,Z3|R]):- !,
    Ax #= X2-X1,
    Ay #= Y2-Y1,
    Az #= Z2-Z1,
    Bx #= X3-X2,
    By #= Y3-Y2,
    Bz #= Z3-Z2,
   %%% γωνία κάμψης <> 180 μοίρες
    10 * Ax + 5 * Ay + Az #\= 10 *  Bx + 5 * By + Bz,
   %%% γωνία κάμψης >= 90 μοίρες
    bathmwto_ginomeno ([Ax,Ay,Az],[Bx,By,Bz],gwnia),
    gwnia #>= 0,
    periorismoi_gwnies([X2,Y2,Z2,X3,Y3,Z3|R]).
periorismoi_gwnies(_).


%$$$$$$$$$$$$$$$$$$$$
% ΚΑΤΗΓΟΡΗΜΑ: deutereuousa_domh /2
% ΕΙΣΟΔΟΣ: Η πληροφορία της δευτερογενούς δομής και η
τριτοταγής δομή.
% ΑΠΟΤΕΛΕΣΜΑ: Επιβάλλει τους περιορισμούς που εξασφαλίζουν την
δημιουργία της δευτερεύουσας δομής
%$$$$$$$$$$$$$$$$$$$$

deutereuousa_domh ([], _).
deutereuousa_domh ([gefyra(A1,A2)|Ypol], Tritogenhs):-
    !,
    Periorismos_gefyras(A1,A2, Tritogenhs),
    deutereuousa_domh(Ypol, Tritogenhs).

%%% ELIKA
deutereuousa_domh([elika(Arxh, Telos)|Ypol], Tritogenhs):-
     !,
     perioxh_deyter(Arxh, Telos, Tritogenhs, Kommati_Trit),
     periorismoi_elikas(Kommati_Trit),
     deutereuousa_domh(SecRest, Tritogenhs).
```

```
%%% PTYXWSH
deutereuousa_domh([ptyxwsh(Arxh, Telos)| Ypol], Tritogenhs):-
     !,
     perioxh_deyter(Arxh, Telos, Tritogenhs,Kommati_Trit),
     periorismoi_ptyxwshs(Kommati_Trit),
     deutereuousa_domh(Ypol, Tritogenhs).



perioxh_deyter(1,End,_Tertiary,[]):-
    End<1,!.
perioxh_deyter(1,End,Tertiary,Sub):-
     E3 is End*3,
     length(Sub,E3),
     prefix(Sub, Tertiary).
perioxh_deyter(S,E,[_X,_Y,_Z|Tertiary],Sub):-
    S>1,!,
    S1 is S-1,
    E1 is E-1,
    perioxh_deyter(S1,E1,Tertiary,Sub).


periorismoi_ptyxwshs([]).
periorismoi_ptyxwshs([_,_,_]):-!.
periorismoi_ptyxwshs([_,_,_,_,_,_]):-!.
periorismoi_ptyxwshs([X1,Y1,Z1,X2,Y2,Z2,X3,Y3,Z3|R]):-
   Dx #= X3-X1,
   Dy #= Y3-Y1,
   Dz #= Z3-Z1,
   count(0,[Dx,Dy,Dz],#=,2),  %% δύο ίδιες συντεταγμένες
   abs(Dx+Dy+Dz) #=2,      %% μια συντεταγμένη σε απόσταση 2

periorismoi_ptyxwshs_diadosh([X1,Y1,Z1,X2,Y2,Z2,X3,Y3,Z3|R],[D
x,Dy,Dz]).


%%% το σχήμα διαδίδεται στα υπόλοιπα αμινοξέα της πτύχωσης με
ισότητες διανυσμάτων με την χρήση ενός βήματος 3 αμινοξέων

periorismoi_ptyxwshs_diadosh ([_,_,_,_,_,_],_):-!.
periorismoi_ptyxwshs_diadosh ([_,_,_],_):-!.
periorismoi_ptyxwshs_diadosh ([],_).
periorismoi_ptyxwshs_diadosh
([X1,Y1,Z1,X2,Y2,Z2,X3,Y3,Z3|R],[Dx,Dy,Dz]):-
   Dx #= X3-X1,
   Dy #= Y3-Y1,
   Dz #= Z3-Z1,
   periorismoi_ptyxwshs_diadosh
      ([X2,Y2,Z2,X3,Y3,Z3|R],[Dx,Dy,Dz]).



periorismoi_elikas([]).
periorismoi_elikas(L) :-
        length(L,NN),
```

```
        NN >= 15, !,
        L=[X1,Y1,Z1,
           X2,Y2,Z2,
           X3,Y3,Z3,
           X4,Y4,Z4,
           X5,Y5,Z5
           |_] ,
   Dx #= X5-X1,
   Dy #= Y5-Y1,
   Dz #= Z5-Z1,
   count(0,[Dx,Dy,Dz],#=,1),  %% μια ίδια συντεταγμένη
   Dx1 #= X3-X1,  %% μισή έλικα
   Dy1 #= Y3-Y1,
   Dz1 #= Z3-Z1,
   count(0,[Dx1,Dy1,Dz1],#=,2),  %% δύο ίδιες συντεταγμένες
   Dx2 #= X5-X3,  %% δεύτερη μισή έλικα
   Dy2 #= Y5-Y3,
   Dz2 #= Z5-Z3,
   count(0,[Dx2,Dy2,Dz2],#=,2),  %% δύο ίδιες συντεταγμένες

   Dx3 #= abs(X4-X1),  %% δεύτερη μισή έλικα
   Dy3 #= abs(Y4-Y1),
   Dz3 #= abs(Z4-Z1),
   Dx3 #=<2,
   Dy3 #=<2,
   Dz3 #=<2,      % δεν μπορεί να είναι το διάνυσμα 3 1 0

   Dx4 #= abs(X5-X2),  %% δεύτερη μισή έλικα
   Dy4 #= abs(Y5-Y2),
   Dz4 #= abs(Z5-Z2),
   Dx4 #=<2,
   Dy4 #=<2,
   Dz4 #=<2,      % δεν μπορεί να είναι το διάνυσμα 3 1 0

   periorismoi_elikas_l(L,[Dx,Dy,Dz]).
periorismoi_elikas(_):-
   write('Προσοχή: πολύ μικρή έλικα!.'),nl.


periorismoi_elikas_ diadosh ([],_).
periorismoi_elikas_ diadosh ([_,_,_],_):-!.
periorismoi_elikas_ diadosh ([_,_,_,_,_,_],_):-!.
periorismoi_elikas_ diadosh ([_,_,_,_,_,_,_,_,_],_):-!.
periorismoi_elikas_ diadosh ([_,_,_,_,_,_,_,_,_,_,_,_],_):-!.
periorismoi_elikas_ diadosh (L,[Dx,Dy,Dz]):-
        L=[X1,Y1,Z1,
           X2,Y2,Z2,
           X3,Y3,Z3,
           X4,Y4,Z4,
           X5,Y5,Z5
           |R],
   Dx #= X5-X1,
   Dy #= Y5-Y1,
   Dz #= Z5-Z1,
   periorismoi_elikas_ diadosh
      ([X2,Y2,Z2,X3,Y3,Z3,X4,Y4,Z4,X5,Y5,Z5|R],[Dx,Dy,Dz]).
```

```
%$$$$$$$$$$$$$$$$$$$$
% ΚΑΤΗΓΟΡΗΜΑ: Periorismos_gefyras/3 (βοηθητικό κατηγόρημα
syntetagmenes/3)
% ΕΙΣΟΔΟΣ: Οι αριθμοί των 2 αμινοξέων και η τριτοταγής δομή.
% ΑΠΟΤΕΛΕΣΜΑ: Περιορίζουμε το άθροισμα των τετραγώνων των
αποστάσεων να είναι ως 12
%$$$$$$$$$$$$$$$$$$$$


Periorismos_gefyras (A1, A2, Tritogenhs):-
    syntetagmenes (Tritogenhs,A1,P1),
    syntetagmenes (Tritogenhs,A2,P2),
    gefyra(P1,P2).


gefyra ([X1,Y1,Z1],[X2,Y2,Z2]):-
    Dx #= abs(X1-X2),
    Dy #= abs(Y1-Y2),
    Dz #= abs(Z1-Z2),
    Dx #=< 2,
    Dy #=< 2,
    Dz #=< 2.



% ΚΑΤΗΓΟΡΗΜΑ: apofygh_symmetriwn
% ΕΙΣΟΔΟΣ: Το μήκος της πρωτεΐνης και η τριτοταγής δομή
% ΑΠΟΤΕΛΕΣΜΑ: Επιβάλει μια σχέση συμμετριών σε κάποια
γειτονικά σημεία

apofygh_symmetriwn(0,[]).
apofygh_symmetriwn(1,_).
apofygh_symmetriwn(Mhkos,[[Mhkos,Mhkos,Mhkos],[M,M,Mhkos]|_]):
-
      artio(Mhkos),
       M #= Mhkos+1.
apofygh_symmetriwn(Mhkos,[[N,N,N],[M,M,N]|_]):-
       peritto(Mhkos),
       N #= Mhkos + 1,
       M #= N+1.


%$$$$$$$$$$$$$$$$$$$$
% ΚΑΤΗΓΟΡΗΜΑ: periorismoi_energeias/4
% ΕΙΣΟΔΟΣ: Η πρωτοταγής λίστα και η τριτοταγής λίστα.
% ΕΞΟΔΟΣ:  Υπολογίζει την ενέργεια της διαμόρφωσης στην
μεταβλητή Energeia (σαν περιορισμό)και την λίστα των επιμέρους
%$$$$$$$$$$$$$$$$$$$$


periorismoi_energeias(Prwtogenhs, Tritogenhs, Energeia,
Lista_me_Energeies):-
```

```
      lista_epimerous_energeies(Prwtogenhs, Tritogenhs,
                                  Lista_me_Energeies),
      sum(Lista_me_Energeies,#=,Energeia).

% ΚΑΤΗΓΟΡΗΜΑΤΑ: lista_epimerous_energeies (βοηθητικά
κατηγορήματα list_pot_amino, pinakas)
% ΕΙΣΟΔΟΣ: The Prwtogenhs list of aminoacids,
%        The Tritogenhs list of triples,
% ΕΞΟΔΟΣ: Η λίστα Lista_me_Energeies με τις μεταβλητές που
σχετίζονται με κάθε δυνατή ενέργεια λόγω επαφής που θα
υπολογισθούν στην μεταβλητή Energeia (σαν περιορισμός)
%        Και η τιμή της ενέργειας της διαμόρφωσης

lista_epimerous_energeies([_],[_],[]).
lista_epimerous_energeies([Ami1,Ami2|ListaAmin],[Pos1,Pos2|Lis
taPos],Eksodos_lista_energeiwn):-
          lista_energ_amino(Ami1,Pos1,ListaAmin,ListaPos,L),
          lista_epimerous_energeies([Ami2|ListaAmin],
[Pos2|ListaPos], Lista_me_Energeies),
      append(L,Lista_me_Energeies,Eksodos_lista_energeiwn).

lista_energ_amino(_Aminoksi,_Thesi,[],[],[]).
lista_energ_amino(Amino,[X1,Y1,Z1],[Amino2|ListaAmin],[[X2,Y2,
Z2]|ListaPos],[En|L]):-
          pinakas(Amino, Amino2, Energeia),
          En in {0, Energeia },
          DX #= abs(X1 - X2),
          DY #= abs(Y1 - Y2),
          DZ #= abs(Z1 - Z2),
          2 #= DX + DY + DZ  #<=> En #= Energeia, %%%
Συνεισφορά επαφής

lista_energ_amino(Amino,[X1,Y1,Z1],ListaAmin,ListaPos,L).


%$$$$$$$$$$$$$$$$$$$
% ΦΑΣΗ ΑΝΑΖΗΤΗΣΗΣ
% ΣΗΜΕΙΩΣΗ: time_out/3 είναι ένα ενσωματωμένο μετα-κατηγόρημα
% :- time_out( Goal, Time, L)
% 1. Επιστρέφει L = success αν μια λύση για τον στόχο Goal
βρεθεί σε <= Time ms.
% 2. Επιστρέφει L = time_out αν μια λύση για τον στόχο Goal
δεν βρεθεί σε χρόνο <= Time ms.
% 3. Επιστρέφει fails αν σε χρόνο <= Time ms αποδειχθεί ότι
δεν υπάρχουν λύσεις για τον στόχο Goal.
%$$$$$$$$$$$$$$$$$$$

anazhthsh_lyshs(Xronos, Tritogenhs, Energeia) :-
      Xronos_se_millisecs is integer(Xronos * 1000),
         time_out(
            (write('ξεκινά η αναζήτηση …'),nl,

            labeling([ff,minimize(Energeia)], Tritogenhs)),
            Xronos_se_millisecs, Apotel),
         (Apotel == time_out -> write('Προσοχή: Δεν είναι
αρκετός ο χρόνος για την εύρεση λύσης !!!'),nl; true),
```

```
          print_results(Kwdikos, Xronos, Prwtogenhs,
Deuterogenhs, Tritogenhs ,Sympyknwsh,b).




%$$$$$$$$$$$$$$$$$$$$
% Τα 20 αμινοξέα
%$$$$$$$$$$$$$$$$$$$$

Aminoksi(a). Aminoksi(c). Aminoksi(d).
Aminoksi(e). Aminoksi(f). Aminoksi(g).
Aminoksi(h). Aminoksi(i). Aminoksi(k).
Aminoksi(l). Aminoksi(m). Aminoksi(n).
Aminoksi(p). Aminoksi(q). Aminoksi(r).
Aminoksi(s). Aminoksi(t). Aminoksi(v).
Aminoksi(w). Aminoksi(y).

%$$$$$$$$$$$$$$$$$$$$
% ΠΙΝΑΚΑΣ ΕΝΕΡΓΕΙΩΝ
%$$$$$$$$$$$$$$$$$$$$


% c
pinakas(c,c,-3477).   pinakas(c,m,-2240).   pinakas(c,f,-2424).
pinakas(c, i, -2410). pinakas(c, l, -2343). pinakas(c, v, -2258).
pinakas(c, w, -2080). pinakas(c, y, -1892). pinakas(c, a, -1700).
pinakas(c, g, -1101). pinakas(c, t, -1243). pinakas(c, s, -1306).
pinakas(c, n, -788).  pinakas(c, q, -835).  pinakas(c, d, -616).
pinakas(c, e, -179).  pinakas(c, h, -1499). pinakas(c, r, -771).
pinakas(c, k, -112).  pinakas(c, p, -1196).
% m
pinakas(m, m, -1901). pinakas(m, f, -2304). pinakas(m, i, -2286).
pinakas(m, l, -2208). pinakas(m, v, -2079). pinakas(m, w, -2090).
pinakas(m, y, -1834). pinakas(m, a, -1517). pinakas(m, g, -897).
pinakas(m, t, -999).  pinakas(m, s, -893).  pinakas(m, n, -658).
pinakas(m, q, -720).  pinakas(m, d, -409).  pinakas(m, e, -209).
pinakas(m, h, -1252). pinakas(m, r, -611).  pinakas(m, k, -146).
pinakas(m, p, -788).
% f
pinakas(f, f, -2467). pinakas(f, i, -2530). pinakas(f, l, -2491).
pinakas(f, v, -2391). pinakas(f, w, -2286). pinakas(f, y, -1963).
pinakas(f, a, -1750). pinakas(f, g, -1034). pinakas(f, t, -1237).
pinakas(f, s, -1178). pinakas(f, n, -790).  pinakas(f, q, -807).
pinakas(f, d, -482).  pinakas(f, e, -419).  pinakas(f, h, -1330 ).
pinakas(f, r, -805).  pinakas(f, k, -270).  pinakas(f, p, -1076).
% i
pinakas(i, i, -2691). pinakas(i, l, -2647). pinakas(i, v, -2568).
pinakas(i, w, -2303). pinakas(i, y, -1998). pinakas(i, a, -1872).
pinakas(i, g, -885).  pinakas(i, t, -1360). pinakas(i, s, -1037).
pinakas(i, n, -669).  pinakas(i, q, -778).  pinakas(i, d, -402).
pinakas(i, e, -439).  pinakas(i, h, -1234). pinakas(i, r, -854).
pinakas(i, k, -253).  pinakas(i, p, -991).
% l
pinakas(l, l, -2501). pinakas(l, v, -2447). pinakas(l, w, -2222).
pinakas(l, y, -1919). pinakas(l, a, -1728). pinakas(l, g, -767).
pinakas(l, t, -1202). pinakas(l, s, -959).  pinakas(l, n, -524).
pinakas(l, q, -729).  pinakas(l, d, -291).  pinakas(l, e, -366).
```

```
pinakas(l, h, -1176). pinakas(l, r, -758).  pinakas(l, k, -222).
pinakas(l, p, -771).
% v
pinakas(v, v, -2385). pinakas(v, w, -2097). pinakas(v, y, -1790).
pinakas(v, a, -1731). pinakas(v, g, -756).  pinakas(v, t, -1240).
pinakas(v, s, -933).  pinakas(v, n, -673).  pinakas(v, q, -642).
pinakas(v, d, -298).  pinakas(v, e, -335).  pinakas(v, h, -1118).
pinakas(v, r, -664).  pinakas(v, k, -200).  pinakas(v, p, -886).
% w
pinakas(w, w, -1867).  pinakas(w, y, -1834). pinakas(w, a, -1565).
pinakas(w, g, -1142).  pinakas(w, t, -1077). pinakas(w, s, -1145).
pinakas(w, n, -884).   pinakas(w, q, -997).  pinakas(w, d, -613).
pinakas(w, e, -624).   pinakas(w, h, -1383). pinakas(w, r, -912).
pinakas(w, k, -391).   pinakas(w, p, -1278).
% y
pinakas(y, y, -1335).  pinakas(y, a, -1318). pinakas(y, g, -818).
pinakas(y, t, -892).   pinakas(y, s, -859).  pinakas(y, n, -670).
pinakas(y, q, -687).   pinakas(y, d, -631).  pinakas(y, e, -453).
pinakas(y, h, -1222).  pinakas(y, r, -745).  pinakas(y, k, -349).
pinakas(y, p, -1067).
% a
pinakas(a, a, -1119).  pinakas(a, g, -290).  pinakas(a, t, -717).
pinakas(a, s, -607).   pinakas(a, n, -371).  pinakas(a, q, -323).
pinakas(a, d, -235).   pinakas(a, e, -039).  pinakas(a, h, -646).
pinakas(a, r, -327).   pinakas(a, k, 196).   pinakas(a, p, -374).
% g
pinakas(g, g, 219).    pinakas(g, t, -311).  pinakas(g, s, -261).
pinakas(g, n, -230).   pinakas(g, q, 033).   pinakas(g, d, -097).
pinakas(g, e, 443).    pinakas(g, h, -325).  pinakas(g, r, -050).
pinakas(g, k, 589).    pinakas(g, p, -042).
% t
pinakas(t, t, -617).   pinakas(t, s, -548).  pinakas(t, n, -463).
pinakas(t, q, -342).   pinakas(t, d, -382).  pinakas(t, e, -192).
pinakas(t, h, -720).   pinakas(t, r, -247).  pinakas(t, k, 155).
pinakas(t, p, -222).
% s
pinakas(s, s, -519).   pinakas(s, n, -423).  pinakas(s, q, -260).
pinakas(s, d, -521).   pinakas(s, e, -161).  pinakas(s, h, -639).
pinakas(s, r, -264).   pinakas(s, k, 223).   pinakas(s, p, -199).
% n
pinakas(n, n, -367).   pinakas(n, q, -253).  pinakas(n, d, -344).
pinakas(n, e, 160).    pinakas(n, h, -455).  pinakas(n, r, -114).
pinakas(n, k, 271).    pinakas(n, p, -018).
% q
pinakas(q, q, 054).    pinakas(q, d, 022).   pinakas(q, e, 179).
pinakas(q, h, -290).   pinakas(q, r, -042).  pinakas(q, k, 334).
pinakas(q, p, -035).
% d
pinakas(d, d, 179).    pinakas(d, e, 634).   pinakas(d, h, -664).
pinakas(d, r, -584).   pinakas(d, k, -176).  pinakas(d, p, 189).
% e
pinakas(e, e, 933).    pinakas(e, h, -324).  pinakas(e, r, -374).
pinakas(e, k, -057).   pinakas(e, p, 257).
% h
pinakas(h, h, -1078).  pinakas(h, r, -307).  pinakas(h, k, 388).
pinakas(h, p, -346).
% r
pinakas(r, r, 200).    pinakas(r, k, 815).   pinakas(r, p, -023).
% k
pinakas(k, k, 1339).   pinakas(k, p, 661).
% p
```

```
pinakas(p, p, 129).

% Συμμετρικές τιμές
pinakas(P1,P2,C):-
          pinakas(P2,P1,C),!.

%%% Κατηγορήματα εμφάνισης

emfanise_apotelesma( Prwtogenhs, Tritogenhs ) :-
   emfanise_apotelesma( 1 , Prwtogenhs, Tritogenhs ).
emfanise_apotelesma( I , [Amino|R], [X,Y,Z|N]) :-
      write(I), write('\t'),
      write(Amino), write('\t'),
      write(X), write('\t'),
      write(Y), write('\t'),
      write(Z), write('\t'),
   nl,
      I1 is I + 1,
      emfanise_apotelesma( I1, R, N).
emfanise_apotelesma(_ , [], []).

%%% Κατηγορήματα χρόνου

Arxikos_xronos :-
                 statistics(runtime,[T|_]),
                 assert(Arxh(T)).
emfanise_xrono(Mhnyma) :-
  Arxh(Arxikos_xronos),
  statistics(runtime,[Trexon_xronos|_]),
  Diafora_xronou is Trexon_xronos - Arxikos_xronos,
  write(Mhnyma),
  emfanise_apotelesma_se_wres_lepta_deyter(Diafora_xronou),
  write('\t').

emfanise_apotelesma_se_wres_lepta_deyter ( K ) :-
  K < 60000,!,
  S is integer(K/1000),
     M is K mod 1000,
   write(S),
      write(','),
      write(M),
      write('s').
emfanise_apotelesma_se_wres_lepta_deyter ( K ) :-
  K < 3600000,!,
  M is integer(K / 60000),
  S is K mod 60000,
  write(M),
  write('min.'),
  emfanise_apotelesma_se_wres_lepta_deyter (S).
emfanise_apotelesma_se_wres_lepta_deyter ( K ) :-
  H is integer(K / 3600000),
  M is K mod 3600000,
  write(H),
  write('hours.'),
  emfanise_apotelesma_se_wres_lepta_deyter (M).
```

```
%$$$$$$$$$$$$$$$$$$$
% Απομάκρυνση βεβαιωμένων γεγονότων,
% που παρέμειναν στην μνήμη λόγω απόρριψης της εκτέλεσης

katharise :-
  retract(start(_)),
  katharise.
katharise.


%$$$$$$$$$$$$$$$$$$$
% ΚΑΤΗΓΟΡΗΜΑ: elegxos/1
% ΕΙΣΟΔΟΣ: Μια λίστα
% ΕΞΟΔΟΣ: Αληθής αν η λίστα περιέχει αμινοξέα
%$$$$$$$$$$$$$$$$$$$


elegxos([T|R]):-
    ground(T),
    aminoksi(T),!,
    elegxos(R).
elegxos([_T|_R]):-
    write('Λανθασμένη ακολουθία αμινοξέων'),nl,
    fail.
elegxos([]).


%$$$$$$$$$$$$$$$$$$$
% Κατηγόρημα για άρτιο/περιττό.
%$$$$$$$$$$$$$$$$$$$


artio (X) :-
    0 #= X mod 2.

peritto(X):-
     1 #= X mod 2.


%$$$$$$$$$$$$$$$$$$$
% ΚΑΤΗΓΟΡΗΜΑ: bathmwto_ginomeno
% ΕΙΣΟΔΟΣ: V1, V2
% ΕΞΟΔΟΣ: V1 X V2.
% ΣΗΜΕΙΩΣΗ: Πρόκειται για μη γραμμικούς περιορισμούς
%$$$$$$$$$$$$$$$$$$$


bathmwto_ginomeno ([Ax,Ay,Az],[Bx,By,Bz],S):-
          Px #= Ax*Bx,
          Py #= Ay*By,
          Pz #= Az*Bz,
        sum([Px,Py,Pz],'#=',S).


%$$$$$$$$$$$$$$$$$$$
```

```
% ΚΑΤΗΓΟΡΗΜΑ: syntetagmenes /3
% ΕΙΣΟΔΟΣ: Η τριτοταγής δομή και ο αριθμός ενός αμινοξέος
% ΕΞΟΔΟΣ: Οι συντεταγμένες [X,Y,Z] του αμινοξέος
%$$$$$$$$$$$$$$$$$$$


syntetagmenes(Tritogenhs, A ,[X,Y,Z]):-
    T1 is (A-1)*3+1,
    T2 is T1 + 1,
    T3 is T2 + 1,
    nth(T1, Tritogenhs,X),
    nth(T2, Tritogenhs,Y),
    nth(T3, Tritogenhs,Z).
```

9 Παράρτημα 2 - Πρωτεΐνες

Στο παράρτημα 2 παρατίθενται οι πρωτεΐνες που χρησιμοποιήθηκαν ως σύνολο ελέγχου του CLP(FD) προγράμματος που υλοποιήθηκε στην παρούσα εργασία για την πρόβλεψη της 3D διαμόρφωσης μιας πρωτεΐνης. Όλες οι πρωτεΐνες καθώς και οι δευτερεύουσες δομές τους προέρχονται από την γνωστή Βάση Δεδομένων Πρωτεϊνών PDB [10].

```
prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% Μήκος: 12
     KWDIKOS = '1LE0',
     Prwtogenhs = [s,w,t,w,e,g,n,k,w,t,w,k] ,
     Deuterogenhs = [ ptyxwsh(2,4), ptyxwsh(9,11) ].

prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% Μήκος: 12
        KWDIKOS = '1KVG',
        Prwtogenhs = [s,c,h,f,g,p,l,g,w,v,c,k],
        Deuterogenhs = [ gefyra(2,11),
                         ptyxwsh(3,4), ptyxwsh(9,11)].

prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% Μήκος: 16
         KWDIKOS = '1LE3',
       Prwtogenhs = [g,e,w,t,w,d,d,a,t,k,
                     t,w,t,w,t,e],
       Deuterogenhs = [ptyxwsh(2,6), ptyxwsh(11,15)].


prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%%% Μήκος: 17
        KWDIKOS = '1EDP',
         Prwtogenhs = [c,s,c,s,s,l,m,d,k,e,
                       c,v,y,f,c,h,l],
         Deuterogenhs = [gefyra(1,15), gefyra(3,11),
                         elika(9,15)
                        ].
prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% Μήκος: 18
        KWDIKOS = '1PG1',
       Prwtogenhs = [r,g,g,r,l,c,y,c,r,r,
                     r,f,c,v,c,v,g,r],
      Deuterogenhs = [gefyra(6,15), gefyra(8,13),
                      ptyxwsh(4,9), ptyxwsh(12,17) ].
```

```
prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%%% Μήκος: 34
        KWDIKOS = '1ZDD',
       Prwtogenhs = [f,n,m,q,c,q,r,r,f,y,
                  e,a,l,h,d,p,n,l,n,e,
                  e,q,r,n,a,k,i,k,s,i,
                  r,d,d,c],
        Deuterogenhs = [
                 gefyra(5, 34),
                      elika(3,13),elika(20,33)].



prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%%% Μήκος: 36
       KWDIKOS = '1VII',
        Prwtogenhs = [m,l,s,d,e,d,f,k,a,v,
                   f,g,m,t,r,s,a,f,a,n,
                      l,p,l,w,k,q,q,n,l,k,
                      k,e,k,g,l,f],
       Deuterogenhs = [elika(4,8),elika(15,18),elika(23,32)].


prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%%% Μήκος: 40
        KWDIKOS = '2GP8',
       Prwtogenhs = [i,t,g,d,v,s,a,a,n,k,
                    d,a,i,r,k,q,m,d,a,a,
                        a,s,k,g,d,v,e,t,y,r,
                        k,l,k,a,k,l,k,g,i,r],
        Deuterogenhs = [elika(6,21),elika(26,38)].


prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% Μήκος: 46
     KWDIKOS = '1ED0',
     Prwtogenhs = [k,s,c,c,p,n,t,t,g,r,
                 n,i,y,n,a,c,r,l,t,g,
                     a,p,r,p,t,c,a,k,l,s,
                     g,c,k,i,i,s,g,s,t,c,
                     p,s,d,y,p,k],
     Deuterogenhs = [gefyra(3,40),gefyra(4,32),gefyra(16,26),
                   elika(7,18),elika(23,30),
                   ptyxwsh(3,4),ptyxwsh(33,34)].


prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% Μήκος: 54
     KWDIKOS = '1ENH',
     Prwtogenhs = [r,p,r,t,a,f,s,s,e,q,
                 l,a,r,l,k,r,e,f,n,e,
                     n,r,y,l,t,e,r,r,r,q,
                     q,l,s,s,e,l,g,l,n,e,
                     a,q,i,k,i,w,f,q,n,k,
                     r,a,k,i],
```

```
     Deuterogenhs = [elika(8,20),elika(26,36),elika(40,52),
                    ptyxwsh(22,23)].


prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%%% Μήκος = 61
  KWDIKOS = '2ERA',
  Prwtogenhs = [r,i,c,f,n,h,q,g,s,q,
              p,q,t,t,k,t,c,s,p,g,
              e,s,s,c,y,n,k,q,w,s,
              d,f,r,g,t,i,i,e,r,g,
              c,g,c,p,t,v,k,p,g,i,
              k,l,s,c,c,e,s,e,v,c,
              n],
  Deuterogenhs = [ ptyxwsh(2,5), ptyxwsh(13,17),
ptyxwsh(34,41),
                   ptyxwsh(23,31), ptyxwsh(50,55)].
prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%%% Μήκος = 63
  KWDIKOS = '1SN1',
  Prwtogenhs = [v,r,d,a,y,i,a,k,p,h,
              n,c,v,y,e,c,a,r,n,e,
              y,c,n,d,l,c,t,k,n,g,
              a,k,s,g,y,c,q,w,v,g,
              k,y,g,n,g,c,w,c,i,e,
              l,p,d,n,v,p,i,r,v,p,
              g,k,c],
  Deuterogenhs = [ gefyra(12,63), gefyra(16,36),
gefyra(22,46), gefyra(26,48),
                  elika(19,28), ptyxwsh(2,5), ptyxwsh(33,39),
ptyxwsh(43,51)].


prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%%% Μήκος = 63
  KWDIKOS = '1YPA',
  Prwtogenhs = [m,k,t,e,w,p,e,l,v,g,
              k,a,v,a,a,a,k,k,v,i,
              l,q,d,k,p,e,a,q,i,i,
              v,l,p,v,g,t,i,v,t,m,
              e,y,r,i,d,r,v,r,l,f,
              v,d,k,l,d,n,i,a,q,v,
              p,r,v],
  Deuterogenhs = [ elika(31,43), ptyxwsh(24,29)].


prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% prwteinh Μήκος: 71
  KWDIKOS = '1FVS',
  Prwtogenhs = [a,r,e,v,i,l,a,v,h,g,
              m,t,c,s,a,c,t,n,t,i,
              n,t,q,l,r,a,l,k,g,v,
              t,k,c,d,i,s,l,v,t,n,
              e,c,q,v,t,y,d,n,e,v,
              t,a,d,s,i,k,e,i,i,e,
```

```
            d,c,g,f,d,c,e,i,l,r,
            d],
  Deuterogenhs = [ elika(13,26), elika(51,62), ptyxwsh(30,36),
ptyxwsh(41,46), ptyxwsh(2,9), ptyxwsh(65,70)].


prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% prwteinh Μήκος: 79
  KWDIKOS = '1B4R',
  Prwtogenhs = [a,t,l,v,g,p,h,g,p,l,
            a,s,g,q,l,a,a,f,h,i,
            a,a,p,l,p,v,t,a,t,r,
            w,d,f,g,d,g,s,a,e,v,
            d,a,a,g,p,a,a,s,h,r,
            y,v,l,p,g,r,y,h,v,t,
            a,v,l,a,l,g,a,g,s,a,
            l,l,g,t,d,v,q,v,e],
  Deuterogenhs = [ ptyxwsh(2,4), ptyxwsh(14,21),
ptyxwsh(46,51), ptyxwsh(39,43), ptyxwsh(28,32),
ptyxwsh(55,64), ptyxwsh(69,78)].


prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% prwteinh Μήκος: 86
  KWDIKOS = '1NFJ',
  Prwtogenhs = [e,h,v,v,y,v,g,n,k,p,
            v,m,n,y,v,l,a,t,l,t,
            q,l,n,e,g,a,d,e,v,v,
            i,k,a,r,g,r,a,i,s,r,
            a,v,d,v,a,e,i,v,r,n,
            r,f,m,p,g,v,k,v,k,e,
            i,k,i,d,t,e,e,l,e,s,
            e,q,g,r,r,s,n,v,s,t,
            i,e,i,v,l,a],
  Deuterogenhs = [ elika(10,24), elika(36,52), ptyxwsh(2,5),
ptyxwsh(28,34), ptyxwsh(77,86), ptyxwsh(57,67)].


prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% prwteinh Μήκος: 91
  KWDIKOS = '1JNS',
  Prwtogenhs = [a,k,t,a,a,a,l,h,i,l,
            v,k,e,e,k,l,a,l,d,l,
            l,e,q,i,k,n,g,a,d,f,
            g,k,l,a,k,k,h,s,i,c,
            p,s,g,k,r,g,g,d,l,g,
            e,f,r,q,g,q,m,v,p,a,
            f,d,k,v,v,f,s,c,p,v,
            l,e,p,t,g,p,l,h,t,q,
            f,g,y,h,i,i,k,v,l,y,
            r],
  Deuterogenhs = [ elika(14,27), elika(30,36), elika(43,47),
elika(60,67), ptyxwsh(50,53), ptyxwsh(3,11), ptyxwsh(82,89),
ptyxwsh(74,79)].
```

```
prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% prwteinh Μήκος: 94
  KWDIKOS = '1CI5',
  Prwtogenhs = [s,s,q,q,i,y,g,v,k,y,
              g,n,v,t,f,h,v,p,s,n,
              q,p,l,k,e,v,l,w,k,k,
              q,k,d,k,v,a,e,l,e,n,
              s,e,f,r,a,f,s,s,f,k,
              n,r,v,y,l,d,t,k,s,g,
              s,l,t,i,y,n,l,t,s,s,
              d,e,d,e,y,e,m,e,s,p,
              n,i,t,d,s,m,k,f,f,l,
              y,v,g,e],
  Deuterogenhs = [ elika(49,52), ptyxwsh(3,8), ptyxwsh(85,93),
ptyxwsh(74,79), ptyxwsh(26,30), ptyxwsh(33,39),
ptyxwsh(42,45)].
prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% prwteinh Μήκος: 96
  KWDIKOS = '1HS7',
%% commented the short elika
  Prwtogenhs = [t,n,q,k,t,k,e,l,s,n,
              l,i,e,t,f,a,e,q,s,r,
              v,l,e,k,e,c,t,k,i,g,
              s,k,r,d,s,k,e,l,r,y,
              k,i,e,t,e,l,i,p,n,c,
              t,s,v,r,d,k,i,e,s,n,
              i,l,i,h,q,n,g,k,l,s,
              a,d,f,k,n,l,k,t,k,y,
              q,s,l,q,q,s,y,n,q,r,
              k,s,l,f,p,l],
  Deuterogenhs = [ elika(2,29), elika(35,45), elika(45,60),
%elika(61,66),
                elika(66,94)].



prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% prwteinh Μήκος: 97
  KWDIKOS = '1J27',
  Prwtogenhs = [m,k,a,y,l,g,l,y,t,a,
              r,l,e,t,p,a,r,s,l,k,
              e,k,r,a,l,i,k,p,a,l,
              e,r,l,k,a,r,f,p,v,s,
              a,a,r,l,y,g,l,d,a,w,
              g,y,e,v,v,g,f,t,l,l,
              g,n,d,p,a,w,v,e,e,t,
              m,r,a,a,a,r,f,l,a,e,
              a,g,g,f,q,v,a,l,e,e,
              f,r,l,e,a,f,e],
  Deuterogenhs = [ elika(18,37), elika(63,81), ptyxwsh(40,45),
ptyxwsh(52,61), ptyxwsh(2,13), ptyxwsh(85,97)].



prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% prwteinh Μήκος: 98
```

```
  KWDIKOS = '2NCM',
  Prwtogenhs = [r,v,l,q,v,d,i,v,p,s,
             q,g,e,i,s,v,g,e,s,k,
             f,f,l,c,q,v,a,g,d,a,
             k,d,k,d,i,s,w,f,s,p,
             n,g,e,k,l,s,p,n,q,q,
             r,i,s,v,v,w,n,d,d,d,
             s,s,t,l,t,i,y,n,a,n,
             i,d,d,a,g,i,y,k,c,v,
             v,t,a,e,d,g,t,q,s,e,
             a,t,v,n,v,k,i,f],
  Deuterogenhs = [ gefyra(24,79), ptyxwsh(3,8),
ptyxwsh(19,28), ptyxwsh(61,67), ptyxwsh(52,57),
ptyxwsh(11,15), ptyxwsh(87,98), ptyxwsh(75,83),
ptyxwsh(34,39), ptyxwsh(43,45)].


prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% prwteinh Μήκος: 98
  KWDIKOS = '1BM8',
  Prwtogenhs = [q,i,y,s,a,r,y,s,g,v,
             d,v,y,e,f,i,h,s,t,g,
             s,i,m,k,r,k,k,d,d,w,
             v,n,a,t,h,i,l,k,a,a,
             n,f,a,k,a,k,r,t,r,i,
             l,e,k,e,v,l,k,e,t,h,
             e,k,v,q,g,g,f,g,k,y,
             q,g,t,w,v,p,l,n,i,a,
             k,q,l,a,e,k,f,s,v,y,
             d,q,l,k,p,l,f,d],
  Deuterogenhs = [ elika(33,39), elika(44,54), elika(77,86),
elika(90,97), ptyxwsh(2,7), ptyxwsh(10,16), ptyxwsh(21,25),
ptyxwsh(61,63), ptyxwsh(72,74)].




prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% prwteinh Μήκος: 100
  KWDIKOS = '1JHG',
  Prwtogenhs = [s,a,a,m,a,e,q,r,h,q,
             e,w,l,r,f,v,d,l,l,k,
             n,a,y,q,n,d,l,h,l,p,
             l,l,n,l,m,l,t,p,d,e,
             r,e,a,l,g,t,r,v,r,i,
             i,e,e,l,l,r,g,e,m,s,
             q,r,e,l,k,n,e,l,g,a,
             g,i,a,t,i,t,r,g,s,n,
             s,l,k,a,a,p,v,e,l,r,
             q,w,l,e,e,v,l,l,k,s],
  Deuterogenhs = [ elika(2,24), elika(28,35), elika(38,56),
elika(61,68), elika(72,84), elika(87,97)].



prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
```

```
%% prwteinh Μήκος: 101
  KWDIKOS = '1UKU',
  Prwtogenhs = [m,i,i,v,y,t,t,f,p,d,
              w,e,s,a,e,k,v,v,k,t,
              l,l,k,e,r,l,i,a,c,a,
              n,l,r,e,h,r,a,f,y,w,
              w,e,g,k,i,e,e,d,k,e,
              v,g,a,i,l,k,t,r,e,d,
              l,w,e,e,l,k,e,r,i,k,
              e,l,h,p,y,d,v,p,a,i,
              i,r,i,d,v,d,d,v,n,e,
              d,y,l,k,w,l,i,e,e,t,
              k],
  Deuterogenhs = [ elika(10,24), elika(58,60), elika(61,73),
elika(89,100), ptyxwsh(29,41), ptyxwsh(44,57), ptyxwsh(2,8),
ptyxwsh(81,84)].


prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% prwteinh Μήκος: 101
  KWDIKOS = '1P1L',
  Prwtogenhs = [m,h,n,f,i,y,i,t,a,p,
              s,l,e,e,a,e,r,i,a,k,
              r,l,l,e,k,k,l,a,a,c,
              v,n,i,f,p,i,k,s,f,f,
              w,w,e,g,k,i,e,a,a,t,
              e,f,a,m,i,v,k,t,r,s,
              e,k,f,a,e,v,r,d,e,v,
              k,a,m,h,s,y,t,t,p,c,
              i,c,a,i,p,i,e,r,g,l,
              k,e,f,l,d,w,i,d,e,t,
              v],
  Deuterogenhs = [ elika(11,25), elika(62,74), elika(90,102),
ptyxwsh(30,42), ptyxwsh(45,59), ptyxwsh(2,9), ptyxwsh(81,85)].



prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
%% prwteinh Μήκος: 102
  KWDIKOS = '',
  Prwtogenhs = [m,e,e,v,v,l,i,t,v,p,
              s,e,e,v,a,r,t,i,a,k,
              a,l,v,e,e,r,l,a,a,c,
              v,n,i,v,p,g,l,t,s,i,
              y,r,w,q,g,e,v,v,e,d,
              q,e,l,l,l,l,v,k,t,t,
              t,h,a,f,p,k,l,k,e,r,
              v,k,a,l,h,p,y,t,v,p,
              e,i,v,a,l,p,i,a,e,g,
              n,r,e,y,l,d,w,l,r,e,
              n,t],
  Deuterogenhs = [ elika(11,25), elika(63,75), elika(91,101),
ptyxwsh(30,42), ptyxwsh(47,60), ptyxwsh(2,9), ptyxwsh(82,86)].


prwteinh(KWDIKOS, Prwtogenhs, Deuterogenhs):-
```

```
%% prwteinh Μήκος: 104
  KWDIKOS = '1TQG',
  Prwtogenhs = [g,s,h,m,e,y,l,g,v,f,
             v,d,e,t,k,e,y,l,q,n,
             l,n,d,t,l,l,e,l,e,k,
             n,p,e,d,m,e,l,i,n,e,
             a,f,r,a,l,h,t,l,k,g,
             m,a,g,t,m,g,f,s,s,m,
             a,k,l,c,h,t,l,e,n,i,
             l,d,k,a,r,n,s,e,i,k,
             i,t,s,d,l,l,d,k,i,f,
             a,g,v,d,m,i,t,r,m,v,
             d,k,i,v],
  Deuterogenhs = [ elika(2,30), elika(33,54), elika(56,75),
elika(81,103)].
```