**ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ**
**ΙΑΤΡΙΚΗ ΣΧΟΛΗ**
**ΜΟΡΦΟΛΟΓΙΚΟΣ–ΚΛΙΝΙΚΟΕΡΓΑΣΤΗΡΙΑΚΟΣ ΤΟΜΕΑΣ**
**ΕΡΓΑΣΤΗΡΙΟ ΙΑΤΡΙΚΗΣΦΥΣΙΚΗΣ**
**ΔΙΕΥΘΥΝΤΗΣ: Καθηγητής ΜΑΡΓΑΡΙΤΑ ΤΖΑΦΛΙΔΟΥ**

# «Επεξεργασία και ανάλυση βιολογικών δεδομένων με τη χρήση ευφυών υπολογιστικών μεθόδων»

**ΚΩΝΣΤΑΝΤΙΝΟΣ Π. ΕΞΑΡΧΟΣ**

**ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ**

**ΙΩΑΝΝΙΝΑ 2011**

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ**
**ΙΑΤΡΙΚΗ ΣΧΟΛΗ**
**ΜΟΡΦΟΛΟΓΙΚΟΣ–ΚΛΙΝΙΚΟΕΡΓΑΣΤΗΡΙΑΚΟΣ ΤΟΜΕΑΣ**
**ΕΡΓΑΣΤΗΡΙΟ ΙΑΤΡΙΚΗΣΦΥΣΙΚΗΣ**
**ΔΙΕΥΘΥΝΤΗΣ: Καθηγητής ΜΑΡΓΑΡΙΤΑ ΤΖΑΦΛΙΔΟΥ**

# «Επεξεργασία και ανάλυση βιολογικών δεδομένων με τη χρήση ευφυών υπολογιστικών μεθόδων»

**ΚΩΝΣΤΑΝΤΙΝΟΣ Π. ΕΞΑΡΧΟΣ**

**ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ**

**ΙΩΑΝΝΙΝΑ 2011**

*«Η έγκριση της διδακτορικής διατριβής από την Ιατρική Σχολή του Πανεπιστημίου Ιωαννίνων δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα»*

Ν. 5343/32, άρθρο 202, παράγραφος 2

**Ημερομηνία αίτησης του κ. Έξαρχου Κωνσταντίνου**: 17-5-2007

**Ημερομηνία ορισμού Τριμελούς Συμβουλευτικής Επιτροπής**: 607ª/29-5-2007

**Μέλη Τριμελούς Συμβουλευτικής Επιτροπής**:

Επιβλέπων

Ρήγας Κωνσταντίνος Επίκουρος Καθηγητής Ιατρικής Φυσικής Ιατρικής Σχολής Πανεπιστημίου Ιωαννίνων

Μέλη

Πολίτου Αναστασία Επίκουρη Καθηγήτρια Βιολογικής Χημείας Ιατρικής Σχολής Πανεπιστημίου Ιωαννίνων

Φωτιάδης Δημήτριος Αναπληρωτής Καθηγητής Ιατρικής Πληροφορικής Τμήματος Πληροφορικής Πανεπιστημίου Ιωαννίνων

**Ημερομηνία ορισμού θέματος**: 14-6-2007

«Επεξεργασία και ανάλυση βιολογικών δεδομένων με τη χρήση ευφυών υπολογιστικών μεθόδων»

**ΔΙΟΡΙΣΜΟΣ ΕΠΤΑΜΕΛΟΥΣ ΕΞΕΤΑΣΤΙΚΗΣ ΕΠΙΤΡΟΠΗΣ** : 714ª/7-6-2011

| | |
|---|---|
| **Τζαφλίδου Μαργαρίτα** | Καθηγήτρια Ιατρικής Φυσικής Ιατρικής Σχολής Πανεπιστημίου Ιωαννίνων |
| **Φωτιάδης Δημήτριος** | Καθηγητής Τμήματος Μηχανικών Επιστήμης Υλικών Πανεπιστημίου Ιωαννίνων |
| **Φώτσης Θεόδωρος** | Καθηγητής Βιολογικής Χημείας Ιατρικής Σχολής Πανεπιστημίου Ιωαννίνων |
| **Λύκας Αριστείδης** | Αναπληρωτής Καθηγητής Τμήματος Πληροφορικής Πανεπιστημίου Ιωαννίνων |
| **Πολίτου Αναστασία** | Επίκουρη Καθηγήτρια Βιολογικής Χημείας Ιατρικής Σχολής Πανεπιστημίου Ιωαννίνων |
| **Ρήγας Κωνσταντίνος** | Επίκουρος Καθηγητής Ιατρικής Φυσικής Ιατρικής Σχολής Πανεπιστημίου Ιωαννίνων |
| **Τρογκάνης Αναστάσιος** | Επίκουρος Καθηγητής Τμήματος Βιολογικών Εφαρμογών και Τεχνολογιών Ιατρικής Σχολής Πανεπιστημίου Ιωαννίνων |

Έγκριση Διδακτορικής Διατριβής με βαθμό **«ΑΡΙΣΤΑ»** στις **8-7-2011**

**ΠΡΟΕΔΡΟΣ ΙΑΤΡΙΚΗΣ ΣΧΟΛΗΣ**

**Ιωάννης Γουδέβενος**

Καθηγητής Παθολογίας-Καρδιολογίας

**Η Γραμματέας της Σχολής**

**ΠΑΡΑΣΚΕΥΗ ΣΒΕΝΤΖΟΥΡΗ -ΖΩΗ**

*Στον παππού μου*

# Πρόλογος

Η παρούσα διδακτορική διατριβή εκπονήθηκε στο Εργαστήριο Ιατρικής Φυσικής της Ιατρικής Σχολής του Πανεπιστημίου Ιωαννίνων σε συνεργασία με τη Μονάδα Ιατρικής Τεχνολογίας και Ευφυών Πληροφοριακών Συστημάτων του Τμήματος Μηχανικών Επιστήμης Υλικών του Πανεπιστημίου Ιωαννίνων.

Η εργασία στα πλαίσια της διδακτορικής διατριβής πραγματοποιήθηκε υπό την επίβλεψη του κ. Κωνσταντίνου Ρήγα, Επίκ. Καθηγητή της Ιατρικής Σχολής. Μέλη της τριμελούς συμβουλευτικής επιτροπής ήταν ο κ. Δημήτριος Φωτιάδης, Καθηγητής στο Τμήμα Επιστήμης και Τεχνολογίας Υλικών και η κα. Αναστασία Πολίτου, Επίκ. Καθηγήτρια της Ιατρικής Σχολής. Θα ήθελα να ευχαριστήσω θερμά την τριμελή συμβουλευτική επιτροπή που καθοδήγησε και οριοθέτησε κάθε φάση της διατριβής μου.

Θα ήθελα να ευχαριστήσω ειλικρινά τον κ. Δημήτριο Φωτιάδη, για τις ευκαιρίες που μου παρείχε, την ουσιαστική καθοδήγηση σε όλη τη διάρκεια της διατριβής και κυρίως για την υποστήριξη που μου πρόσφερε τόσο σε επιστημονικό και υλικοτεχνικό επίπεδο, όσο ηθικά και πνευματικά. Επίσης θα ήθελα να ευχαριστήσω τους: κ. Αναστάσιο Τρογκάνη, Επικ. Καθηγητή του Τμήματος Βιολογικών Εφαρμογών και Τεχνολογιών, κ. Κωνσταντίνο Παπαλουκά, Επικ. Καθηγητή του Τμήματος Βιολογικών Εφαρμογών και Τεχνολογιών και κ. Γεώργιο Γκωλέτση, Λέκτορα του Τμήματος Οικονομικών.

Οφείλω ένα πολύ μεγάλο ευχαριστώ στον αδερφό μου, την Πάττυ και τους γονείς μου που με στηρίζουν στις προσπάθειές μου και υποστηρίζουν τα όνειρά μου. Ένα μεγάλο ευχαριστώ στο Γεώργιο Ρήγα (ψηλό) και στο Μάνο Γεωργιλή που με διαφορετικό τρόπο ο καθένας με βοήθησαν σημαντικά να ολοκληρώσω την διατριβή μου. Τέλος, ένα μεγάλο ευχαριστώ οφείλω στους φίλους και συνεργάτες μου από την Μονάδα Ιατρικής Τεχνολογίας και Ευφυών Πληροφοριακών Συστημάτων.

Η διδακτορική διατριβή χρηματοδοτήθηκε από τα έργα INTREPID (IST-2002-507464), POCEMON (ICT-2007-216088) και NeoMark (ICT-2007-224483).

Κωνσταντίνος Έξαρχος
Υποψήφιος Διδάκτορας Ιατρικής Σχολής
Πανεπιστημίου Ιωαννίνων

## Περιεχόμενα

## Λίστα εικόνων

## Λίστα πινάκων

# Συντομογραφίες

| Συντομογραφία | Αγγλικός όρος | Ελληνικός όρος |
|---|---|---|
| NMR | Nuclear Magnetic Resonance | Πυρηνικός Μαγνητικός Συντονισμός |
| PPIases | Peptidyl Prolyl Isomerases | Πεπτιδικές Προλινικές Ισομεράσες |
| SVM | Support Vector Machine | Μηχανή Διανυσμάτων Υποστήριξης |
| IDP | Intrinsically Disordered Proteins | Ρεομορφικές πρωτεΐνες |
| PTM | Post-translational Modifications | Μετα-μεταφραστικές τροποποιήσεις |
| HPRD | Human Protein Reference Database | - |
| DIP | Database of Interacting Proteins | - |
| HPV | Human Pappiloma Virus | Ιός ανθρώπινων κονδυλωμάτων |
| ANN | Artificial Neural Networks | Τεχνητά Νευρωνικά Δίκτυα |
| PSSM | Position Specific Scoring Matrix | - |
| PDB | Protein Data Bank | - |
| PISCES | Protein Sequence Culling Server | - |
| VADAR | Volume Area Dihedral Angle Reporter | - |
| Se | Sensitivity | Ευαισθησία |
| Sp | Specificity | Ειδίκευση |
| PPV | Positive Predictive Value | Θετική προγνωστική αξία |
| Acc | Accuracy | Ακρίβεια |
| ROC | Receiver Operating Characteristic | - |
| AUC | Area Under Curve | Εμβαδόν κάτω από την καμπύλη ROC |
| RBF | Radial Basis Function | Πυρήνας Ακτινικής Βάσης |
| PCA | Principal Component Analysis | Ανάλυση Κύριων Συνιστωσών |
| FDR | False Discovery Rate | Ρυθμός σφάλματος |
| ELM | Eukaryotic Linear Motif | - |
| TRG | Targeting/localization | Στόχευση/εντόπιση |
| LIG | Ligand/binding | Πρόσδεση/υποκατάσταση |
| CLV | Cleavage site | Θέση αποκοπής |
| MOD | Post-translational Modifications | Μετα-μεταφραστικές τροποποιήσεις |
| MiMI | Michigan Molecular Interactions | - |
| LRR | Leucine Rich Repeats | Πρωτεϊνικές περιοχές πλούσιες σε λευκίνη |
| SMOTE | Synthetic Minority Oversampling Technique | - |
| CFS | Correlation-based Feature Selection | - |
| BN | Bayes Network | Μπεϋζιανό δίκτυο |
| DT | Decision Tree | Δέντρο απόφασης |
| RF | Random Forests | Τυχαία δάση |
| DBN | Dynamic Bayesian Networks | Δυναμικά Μπεϋζιανά Δίκτυα |

# Εισαγωγή

Η παρούσα διδακτορική διατριβή εστιάζεται στην ανάπτυξη και εφαρμογή ευφυών υπολογιστικών μεθόδων για την οργάνωση, επεξεργασία, ανάλυση και κατανόηση μεγάλου όγκου βιολογικών δεδομένων. Απώτερος σκοπός είναι η σταδιακή αποκρυπτογράφηση του γενετικού υποβάθρου ασθενειών και κατ' επέκταση η αποδοτικότερη και στοχευμένη αντιμετώπισή τους. Η μελέτη ανάγεται σε ένα επαγωγικό πλαίσιο έρευνας που δομείται σταδιακά από το ειδικό προς το γενικό. Η έρευνα ξεκινάει από το κατώτερο επίπεδο οργάνωσης των πρωτεϊνών - την πρωτοταγή δομή - όπου μελετάμε τους δεσμούς που αναπτύσσονται μεταξύ των αμινοξέων. Στη συνέχεια, εξετάζουμε τις εξαρτήσεις και αλληλεπιδράσεις μεταξύ των πρωτεϊνικών μορίων, εστιάζοντας συγκεκριμένα σε πρωτεΐνες που είτε μερικώς είτε πλήρως είναι ρεομορφικές, δηλαδή δεν φέρουν σαφή τριτοταγή δομή. Οι ρεομορφικές πρωτεΐνες λόγω της εγγενούς ευμεταβλητότητάς τους, έχει βρεθεί ότι συμμετέχουν σε πλήθος κυτταρικών λειτουργιών και συνακόλουθα έχουν συσχετιστεί με την επαγωγή σοβαρών ασθενειών. Στη συνέχεια της διδακτορικής διατριβής επεκτείνουμε και συνδυάζουμε την ανάλυση από το μικροσκοπικό επίπεδο με μακροσκοπικές παρατηρήσεις και δεδομένα. Συγκεκριμένα, συλλέγουμε βιολογικά δεδομένα που αφορούν την κυτταρική και συστημική λειτουργία του οργανισμού, καθώς και κλινικά δεδομένα (ιατρικό ιστορικό και απεικονιστικά δεδομένα) που αφορούν σε φαινοτυπικό πλέον επίπεδο ανατομικές οντότητες και τον οργανισμό ως σύνολο. Εφαρμόζουμε αυτή την πολύπλευρη και πολυπαραγοντική ανάλυση σε μια πολύπλοκη νόσο όπως ο καρκίνος – και πιο συγκεκριμένα στον στοματικό καρκίνο – που φέρει εκφάνσεις σε όλα τα επίπεδα της φυσιολογίας του οργανισμού.

Στο 1ο κεφάλαιο παραθέτουμε εισαγωγικές έννοιες για όλα τα επιμέρους ερευνητικά πεδία που εξετάζονται, είτε άπτονται της παρούσας διδακτορικής διατριβής. Συγκεκριμένα, οι τομείς όπου αναλύονται οι βασικές έννοιες και παρουσιάζεται η σχετική βιβλιογραφία είναι: τα επίπεδα οργάνωσης των πρωτεϊνών, ο ισομερισμός του πεπτιδικού δεσμού, τα δίκτυα πρωτεϊνικών αλληλεπιδράσεων, οι ρεομορφικές πρωτεΐνες και ο λειτουργικός τους ρόλος, η μικροσκοπική και μακροσκοπική θεώρηση του καρκίνου και ιδιαίτερα ο μηχανισμός εξέλιξης του στοματικού καρκίνου.

Στο 2ο κεφάλαιο παρουσιάζουμε μια μεθοδολογία για την πρόβλεψη της διαμόρφωσης του πεπτιδικού δεσμού, μεταξύ των αμινοξέων μιας πρωτεΐνης. Εξάγοντας ένα πλήθος χαρακτηριστικών με βιολογική σημασία αμιγώς από την πρωτοταγή ακολουθία των αμινοξέων, προβλέπουμε τη διαμόρφωση του πεπτιδικού δεσμού. Συνεχίζοντας στο ίδιο ερευνητικό πεδίο, στο 3ο κεφάλαιο εστιάζουμε το ενδιαφέρον μας στην εξαγωγή ακολουθιακών προτύπων, που χαρακτηρίζουν και περιγράφουν με έναν εύληπτο τρόπο τις ισομερείς διαμορφώσεις του πεπτιδικού δεσμού. Στη συνέχεια χρησιμοποιούμε τα εξαχθέντα πρότυπα ώστε να αναδείξουμε τις λειτουργικές συσχετίσεις των cis πεπτιδικών δεσμών.

Στο 4ο κεφάλαιο μελετάμε τα δίκτυα αλληλεπίδρασης των ρεομορφικών πρωτεϊνών και μέσα από αυτά αναδεικνύουμε τον τρόπο με τον οποίο επάγουν τον λειτουργικό τους ρόλο καθώς και την συνακόλουθη συμμετοχή τους σε πλήθος ασθενειών στον άνθρωπο. Εντοπίζουμε κατά τρόπο συστηματικό ακολουθιακά πρότυπα μέσω των οποίων οι ρεομορφικές πρωτεΐνες αλληλεπιδρούν με ένα πλήθος πρωτεϊνικών μορίων και επιτελούν τον ετερόκλητο λειτουργικό τους ρόλο.

Στο 5ο κεφάλαιο αναπτύσσουμε μια ολιστική και συνδυαστική προσέγγιση με σκοπό την έγκαιρη πρόβλεψη της επανεμφάνισης του στοματικού καρκίνου. Συγκεκριμένα i) αναλύουμε γενετικά δεδομένα για να εντοπίσουμε σε μικροσκοπικό επίπεδο τον βιολογικό μηχανισμό που καθοδηγεί την εξέλιξη της νόσου και στη συνέχεια ii) αναλύουμε πληροφορίες σχετικά με τις μακροσκοπικές εκφάνσεις του στοματικού καρκίνου σε απεικονιστικά δεδομένα καθώς και το κλινικό προφίλ του ασθενούς. Η προτεινόμενη πολυπαραγοντική και πολυεπίπεδη ανάλυση μας βοηθά να εντοπίσουμε τους παράγοντες που επιδρούν καθοριστικά στην εξέλιξη της νόσου και κατ' επέκταση να ανιχνεύσουμε έγκαιρα και με ακρίβεια μια ενδεχόμενη υποτροπή.

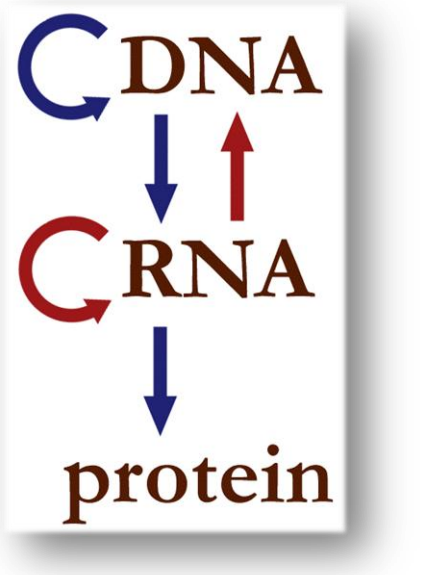
Η συνεισφορά της παρούσας διδακτορικής διατριβής εντοπίζεται στα ακόλουθα σημεία: (i) στην ανάδειξη του βιολογικού μηχανισμού που καθορίζει την διαμόρφωση του πεπτιδικού δεσμού, (ii) στον συστηματικό εντοπισμό λειτουργικών συσχετίσεων των πρωτεϊνικών περιοχών που φέρουν cis πεπτιδικούς δεσμούς, (iii) στην εξαγωγή ακολουθιακών προτύπων που σηματοδοτούν και επάγουν τις αλληλεπιδράσεις και λειτουργίες των ρεομορφικών πρωτεϊνών, (iv) στην συστηματική καταγραφή του τρόπου με τον οποίο οι ρεομορφικές πρωτεΐνες συμμετέχουν σε γενετικές ασθένειες, (v) στη μελέτη του γενετικού υποβάθρου πολυπαραγοντικών ασθενειών (στοματικός καρκίνος) και την ανάδειξη γενετικών παραγόντων που συμβάλλουν στην εξέλιξή τους, (vi) στη

συνδυαστική ανάλυση κλινικών και βιολογικών δεδομένων για την πολύπλευρη πλαισίωση γενετικών ασθενειών, καταγράφοντας μεταβολές στα κύτταρα, τα συστήματα, τους ιστούς, μέχρι και ολόκληρο τον οργανισμό.

# 1ο ΚΕΦΑΛΑΙΟ: Εισαγωγικές έννοιες - βιβλιογραφική ανασκόπηση

## *1.1 Βιοπληροφορική και εισαγωγικές έννοιες βιολογίας*

Η διατύπωση του κεντρικού δόγματος της μοριακής βιολογίας [1, 2] και η συνακόλουθη ανακάλυψη της διπλής έλικας του DNA αποτελούν ουσιαστικά και την απαρχή της Βιοπληροφορικής, δηλαδή του επιστημονικού κλάδου που χρησιμοποιεί αρχές Στατιστικής και Πληροφορικής για να επιλύσει ζητήματα Μοριακής Βιολογίας. Τα βασικά σημεία του κεντρικού δόγματος, που αφορά στην ροή της γενετικής πληροφορίας σε βιολογικά συστήματα, απεικονίζονται στην Εικόνα 1 και αναλύονται αμέσως πιο κάτω.




**Εικόνα 1:** Το κεντρικό δόγμα της μοριακής βιολογίας.

Σε αδρές γραμμές η γενετική πληροφορία "ρέει" από το DNA στο RNA και εν συνεχεία στις πρωτεΐνες, περιγράφοντας με αυτόν τον τρόπο την ζωή και όλες τις διεργασίες της ως μια γραμμική ροή πληροφορίας. Το DNA χρησιμοποιώντας ένα αλφάβητο 4 γραμμάτων (A: αδενίνη, T: θυμίνη, C: κυτοσίνη, G: γουανίνη) αποτελεί το βασικό σχέδιο για έναν οργανισμό. Για να χρησιμοποιηθεί αυτός ο τεράστιος όγκος

πληροφορίας που βρίσκεται καταγεγραμμένος στο DNA ώστε να παραχθούν πρωτεΐνες, πρέπει το DNA να μετατραπεί σε αγγελιοφόρο RNA (mRNA) μέσω της διαδικασίας που καλείται μεταγραφή. Ακολούθως, το mRNA μεταφράζεται σε πρωτεΐνες περνώντας κατ' αυτόν τον τρόπο στο αλφάβητο των 20 γραμμάτων (αμινοξέα). Η ακολουθία των αμινοξέων καθορίζει την τρισδιάστατη δομή του πρωτεϊνικού μορίου [3], την εντόπισή του στο κύτταρο, την αλληλεπίδρασή του με άλλες πρωτεΐνες, και συνεπώς την εν γένει λειτουργία του. Αυτή η αρθρωτή/ακολουθιακή προσέγγιση των βιολογικών συστημάτων, όπως περιγράφεται στο κεντρικό δόγμα, αλλά και κυρίως η αναπαράσταση των βιολογικών δεδομένων (DNA, RNA, πρωτεϊνών) με τη μορφή ψηφιακής πληροφορίας, άνοιξαν ένα παράθυρο στην Βιολογία για την εισβολή της Πληροφορικής.

Θα μπορούσαμε να αντιπαραβάλλουμε την ροή της πληροφορίας στους οργανισμούς με την διαδικασία εκτέλεσης ενός προγράμματος στον ηλεκτρονικό υπολογιστή. Συγκεκριμένα, όταν εκτελούμε ένα πρόγραμμα, τότε αυτό αντιγράφεται από την μνήμη ανάγνωσης (ROM) στην μνήμη προσπέλασης (RAM). Από εκεί και έπειτα και για όσο παραμένει ανοιχτός ο υπολογιστής, το πρόγραμμα εκτελείται βάσει των εντολών που έχουν περαστεί στην RAM. Αυτός ο σχεδιασμός επιτρέπει την εκτέλεση του προγράμματος, διασφαλίζοντας το πρωτότυπο που βρίσκεται στην μνήμη ROM από οποιαδήποτε πιθανή βλάβη. Το DNA (μνήμη ROM) αποτελεί λοιπόν το πρωτότυπο λειτουργίας του κυττάρου που μέσω του χρηστικού του αντιγράφου σε μορφή RNA (μνήμη RAM) δίδει τις εντολές για όλες τις κυτταρικές λειτουργίες.

## *DNA/RNA και γενετικός κώδικας*

Στο πιο βασικό του επίπεδο το DNA δομείται από μια σειρά μικρότερων μορίων που ονομάζονται νουκλεοτίδια. Ακολούθως, το κάθε νουκλεοτίδιο αποτελείται από 3 μονάδες: ένα σάκχαρο, μια φωσφορική ομάδα και μια αζωτούχο βάση, η οποία και διακρίνει τα νουκλεοτίδια μεταξύ τους (Εικόνα 2).


**Εικόνα 2:** Απλοποιημένη απεικόνιση ενός νουκλεοτιδίου.

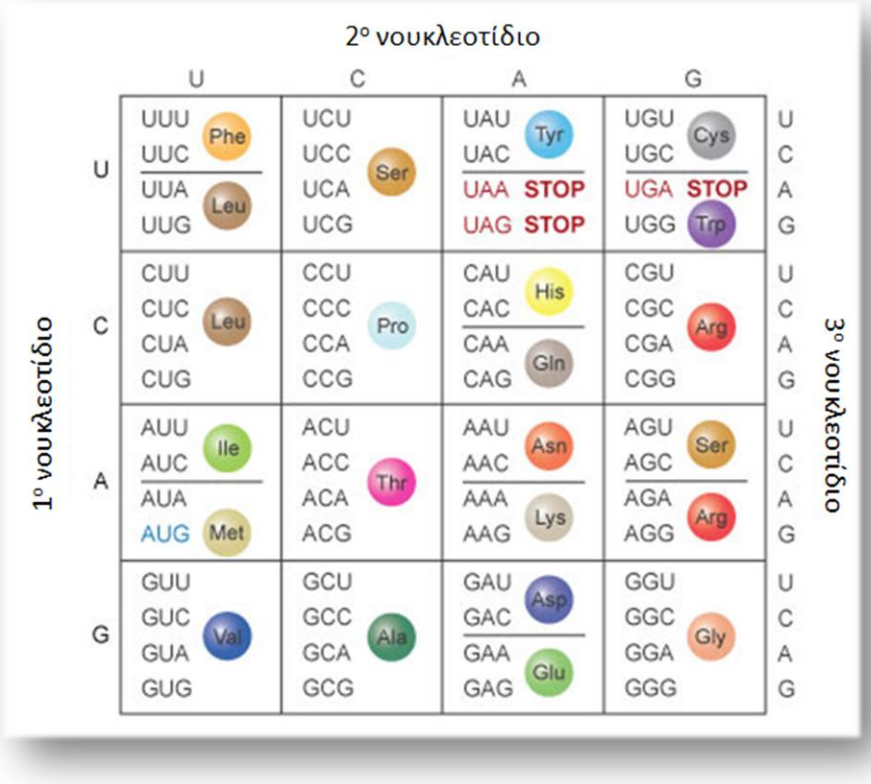
Τα περισσότερα μόρια του DNA αποτελούνται από δυο αλυσίδες διαπλεκόμενες μεταξύ τους κατά τέτοιο τρόπο ώστε στην τρισδιάστατη δομή τους να αποτελούν μια διπλή έλικα, όπου ο βασικός κορμός σακχάρου-φωσφορικού βρίσκεται στο εξωτερικό και οι βάσεις στο εσωτερικό (Εικόνα 3). Οι δύο κλώνοι του DNA συγκρατούνται με δεσμούς υδρογόνου μεταξύ των αζωτούχων βάσεων στο κέντρο της έλικας. Συγκεκριμένα, η αδενίνη σχηματίζει δυο δεσμούς υδρογόνου με τη θυμίνη (A-T) και η κυτοσίνη τρεις υδρογονοδεσμούς με τη γουανίνη (C-G), και καλούνται ζεύγη βάσεων.

**Εικόνα 3:** Η τρισδιάστατη δομή του DNA προσομοιάζει μια σπειροειδή σκάλα.

Η αλληλουχία των βάσεων κατά μήκος της αλυσίδας του DNA αποτελεί τις γενετικές πληροφορίες – τις οδηγίες δηλαδή για την διατήρηση της ζωής του, καθιστώντας το DNA ένα αποθηκευτικό μόριο. Αυτές οι οδηγίες στη συνέχεια καθορίζουν την σύνθεση

πλειάδας βιομορίων, των πρωτεϊνών, οι οποίες συγκροτούν και ενορχηστρώνουν έναν οργανισμό.

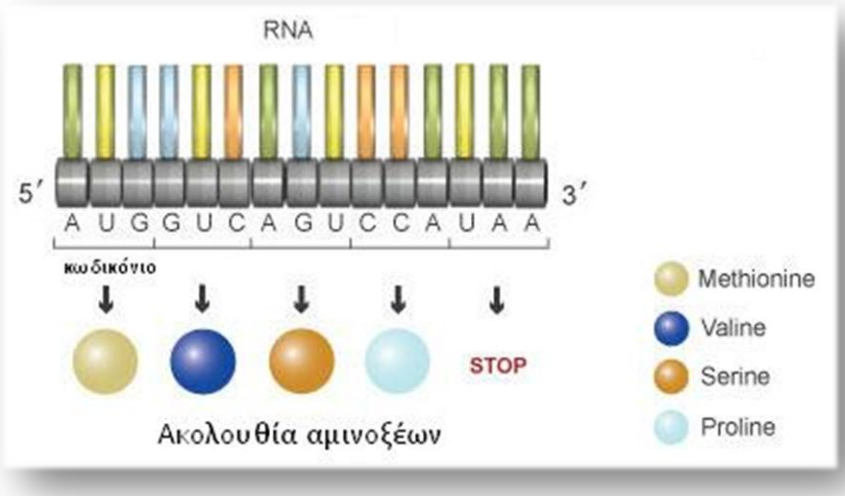
Όπως αναφέρθηκε και πιο πάνω, το DNA αρχικά μεταγράφεται σε RNA, το οποίο αποτελείται από τις βάσεις αδενίνη, ουρακίλη (U), κυτοσίνη και γουανίνη, τηρώντας τα ζεύγη A-U και C-G. Το RNA είναι το ενδιάμεσο μόριο για τη ροή των γενετικών πληροφοριών, και μεταφράζεται σε πρωτεΐνες χρησιμοποιώντας ως λεξικό το γενετικό κώδικα (Εικόνα 4).




**Εικόνα 4:** Ο γενετικός κώδικας.

Ο γενετικός κώδικας εκφράζει την σχέση μεταξύ της αλληλουχίας των βάσεων στο DNA (ή στο αντίγραφο του RNA) και της αλληλουχίας των αμινοξέων στις πρωτεΐνες. Τα αμινοξέα, που είναι 20 στον αριθμό κωδικοποιούνται από τριπλέτες βάσεων, που ονομάζονται κωδικόνια, αρχίζοντας από ένα συγκεκριμένο σημείο (Εικόνα 5). Από τα 64 κωδικόνια, τα 61 καθορίζουν συγκεκριμένα αμινοξέα ενώ τα άλλα τρία (UAA, UAG και UGA) είναι κωδικόνια τερματισμού της αλυσίδας. Επομένως, για τα περισσότερα

αμινοξέα υπάρχουν περισσότερες από μία κωδικεύουσες τριπλέτες, δηλαδή, ο γενετικός κώδικας είναι εκφυλισμένος.




**Εικόνα 5:** Το κάθε αμινοξύ κωδικοποιείται από μια ομάδα τριών βάσεων (κωδικόνιο).

Ο μεγάλος εκφυλισμός του γενετικού κώδικα είναι τεράστιας βιολογικής σημασίας αφού ελαχιστοποιεί και "καταπνίγει" τα βλαβερά αποτελέσματα των μεταλλάξεων. Σε αντίθετη περίπτωση, αν δηλαδή ο γενετικός κώδικας δεν ήταν εκφυλισμένος, τότε 20 κωδικόνια θα κωδικοποιούσαν 20 αμινοξέα και άρα θα υπήρχαν 44 κωδικόνια τερματισμού. Βάσει πιθανοτήτων λοιπόν, μια ενδεχόμενη αλλαγή βάσης, λόγω μετάλλαξης, πιθανότατα θα οδηγούσε σε τερματισμό της αλυσίδας. Επίσης, αξίζει να σημειωθεί ότι δεν υπάρχει επικάλυψη μεταξύ των κωδικονίων του γενετικού κώδικα, δηλαδή, μια τριπλέτα βάσεων αποτελεί μέρος ενός και μόνο κωδικονίου, οπότε και μια πιθανή σημειακή μετάλλαξη θα επηρεάσει μονάχα ένα κωδικόνιο και κατ' επέκταση αμινοξύ.

Με τη μετάφραση λοιπόν του RNA (που έχει προκύψει έπειτα από την μεταγραφή του DNA) σε πρωτεΐνες, ολοκληρώνεται η αλληλουχία βημάτων που περιγράφει το κεντρικό δόγμα της μοριακής βιολογίας. Όπως αναφέρθηκε και προηγουμένως, η βάση του κεντρικού δόγματος είναι η ιεραρχική ροή της πληροφορίας, δηλαδή, ενώ το DNA και το RNA περιέχουν όλη την πληροφορία που τελικά μεταφράζεται σε πρωτεΐνες, το αντίστροφο δεν ισχύει. Το DNA λοιπόν περιλαμβάνει επιπρόσθετη πληροφορία που δεν θα μπορούσε να εξαχθεί από την πρωτεϊνική ακολουθία, όπως οδηγίες για τον έλεγχο της

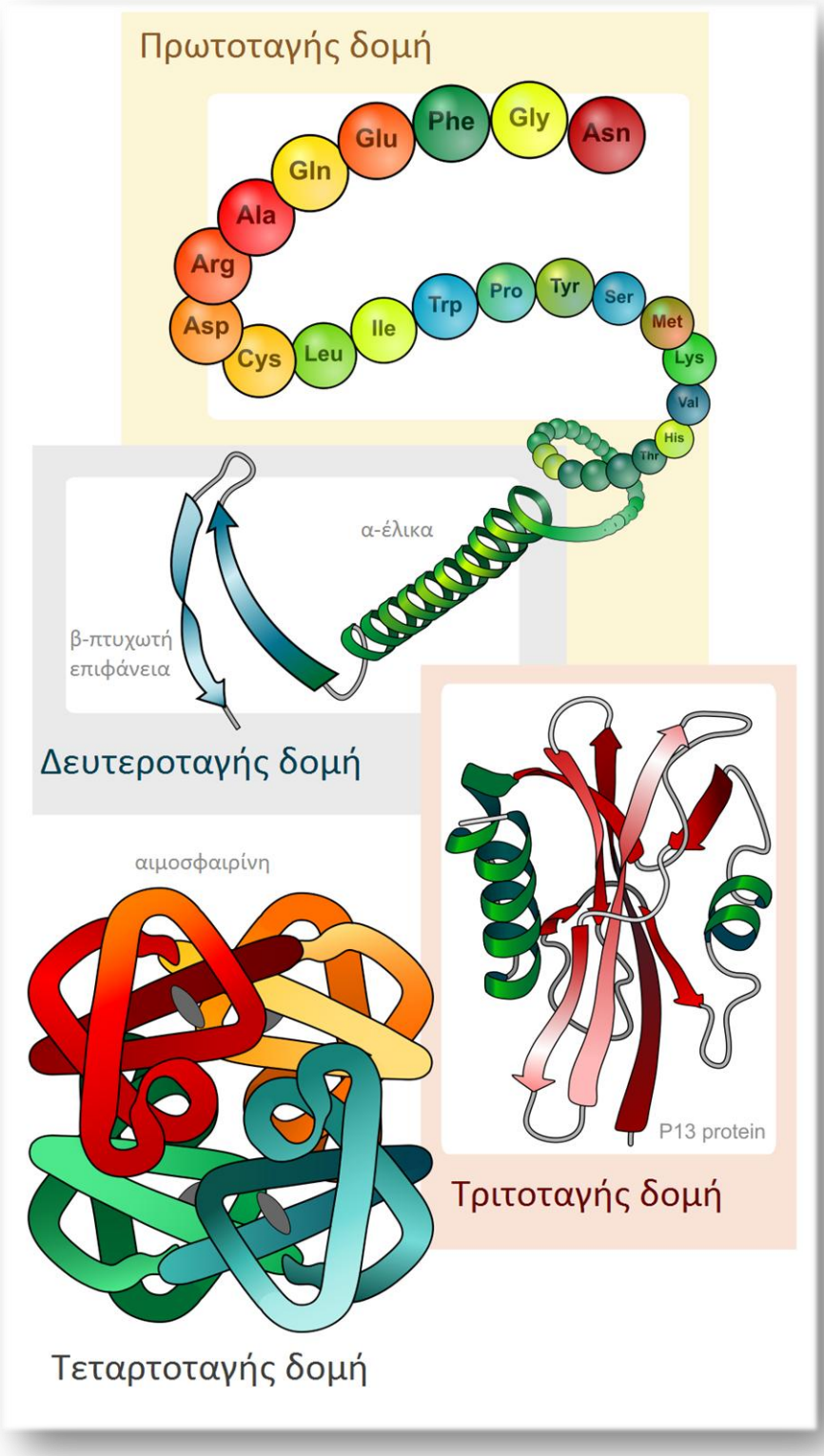
γονιδιακής έκφρασης. Συνεπώς τίθεται ευλόγως το ερώτημα γιατί να μελετήσουμε τις πρωτεΐνες όταν το DNA μπορεί να προσφέρει την ίδια και κατά κανόνα περισσότερη πληροφορία; Η απάντηση είναι ότι εάν ήμασταν σε θέση να κατανοήσουμε και να συστηματοποιήσουμε όλο το εύρος την πληροφορίας που βρίσκεται "κρυμμένη" στο DNA δεν θα υπήρχε η ανάγκη ανάπτυξης μεθόδων για την ανάλυση των πρωτεϊνικών ακολουθιών. Δυστυχώς, όμως για την ώρα αυτό δεν ισχύει και διαφαίνεται από το γεγονός ότι συνεχώς αναπτύσσονται μέθοδοι για τον εντοπισμό των περιοχών του DNA που τελικά μεταφράζονται σε πρωτεΐνες και ποιά άλλα τμήματά του επιτελούν άλλες λειτουργίες. Εστιάζοντας λοιπόν στις πρωτεΐνες παρακάμπτουμε ενδεχόμενες ασάφειες που ενυπάρχουν στις νουκλεοτιδικές ακολουθίες [4]. Συνεπώς, η παράλληλη εξερεύνηση και των δυο πεδίων, που κατά τα άλλα είναι άρρηκτα δεμένα μεταξύ, τους αποτελεί την προσφορότερη οδό για την πληρέστερη περιγραφή των πολύπλοκων βιολογικών συστημάτων.

## *Πρωτεΐνες*

Οι πρωτεΐνες αποτελούν πιθανότατα την πιο πολύπλοκη χημική οντότητα στη φύση, και αποτελούν ακρογωνιαίο λίθο για όλες τις βιολογικές διαδικασίες χάρη στην πλειάδα των κρίσιμων λειτουργιών που επιτελούν. Συγκεκριμένα δρουν ως καταλύτες, μεταφέρουν και αποθηκεύουν ποικιλία μορίων (π.χ. οξυγόνο), παρέχουν μηχανική στήριξη και ανοσολογική προστασία, μεταβιβάζουν νευρικά ερεθίσματα, ελέγχουν την ανάπτυξη και την διαφοροποίηση καθώς και πλήθος άλλων λειτουργιών.

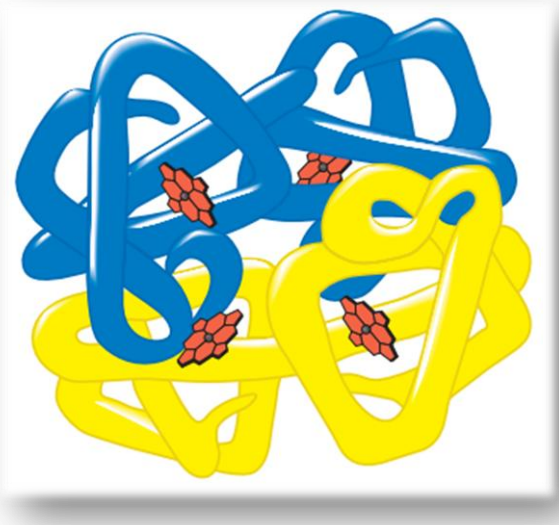
Για την επιτέλεση όλων αυτών των λειτουργιών, οι πρωτεΐνες πρέπει να λάβουν συγκεκριμένη αναδίπλωση στο χώρο ή και πολλές φορές να συνεργαστούν με άλλα πρωτεϊνικά μόρια. Ξεκινώντας, λοιπόν, από μια μακροσκοπική και εκ-του-αποτελέσματος θεώρηση των πρωτεϊνών, και προχωρώντας βαθύτερα στις πιο πρωτογενείς οντότητες που τις δομούν, διατρέχουμε αντίστροφα την διαδεδομένη ιεραρχική προσέγγιση για την διάκριση των επιπέδων οργάνωσης των πρωτεϊνών [5]. Συγκεκριμένα, θα περιγράψουμε πρώτα την τεταρτοταγή δομή και θα προχωρήσουμε ακολούθως προς την τριτοταγή και δευτεροταγή καταλήγοντας στην πρωτοταγή δομή, αποδομώντας κατ' αυτόν τον τρόπο κάθε επίπεδο οργάνωσης σε επιμέρους οντότητες (Εικόνα 6). Ο σκοπός μας είναι να καταλήξουμε στην πρωτοταγή δομή, η οποία παρά την σχετική της απλότητα είναι και

αυτή που ενέχει και κωδικοποιεί σε μεγάλο βαθμό την πολυπλοκότητα των υπόλοιπων επιπέδων οργάνωσης των πρωτεϊνών.




**Εικόνα 6:** Αντίστροφη ιεραρχική δομή οργάνωσης των πρωτεϊνών.

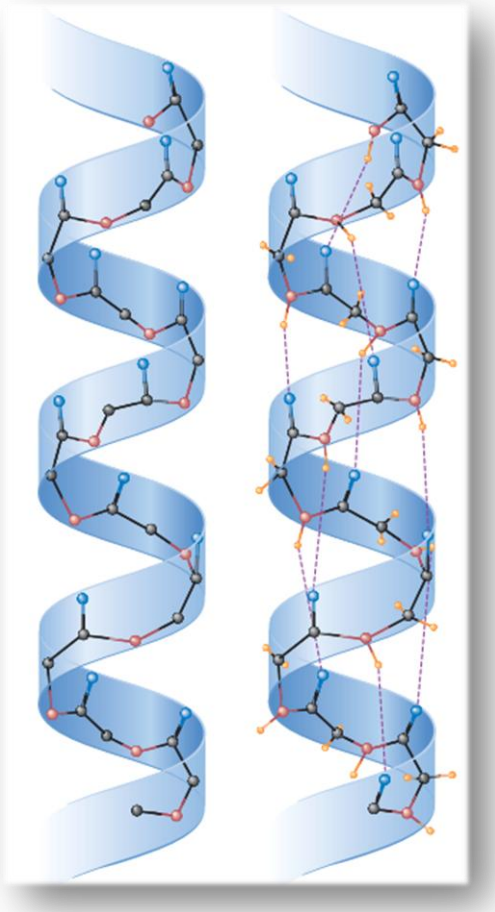
Η τεταρτοταγής δομή των πρωτεϊνών αφορά την κατά περιπτώσεις συνεργασία δύο ή και περισσοτέρων πολυπεπτιδίων με σκοπό την επιτέλεση κάποιας λειτουργίας. Οι περισσότερες πρωτεΐνες με μοριακό βάρος μεγαλύτερο από 50.000 αποτελούνται από δύο ή περισσότερες μη-ομοιοπολικά συνδεδεμένες πολυπεπτιδικές αλυσίδες. Ο σχηματισμός του ολιγομερούς γίνεται αυθόρμητα και οφείλεται σε υδρόφοβες αλληλεπιδράσεις των ομάδων που βρίσκονται στην επιφάνεια των επιμέρους υπομονάδων. Χαρακτηριστικό παράδειγμα τεταρτοταγούς δομής αποτελεί η αιμοσφαιρίνη του αίματος που μεταφέρει το οξυγόνο στους ιστούς, και αποτελείται από δύο ζεύγη διαφορετικών πρωτεϊνικών αλυσίδων, δύο α-αλυσίδες και δύο β-αλυσίδες (Εικόνα 7). Μόνο όταν τα τέσσερα πεπτίδια ενωθούν μεταξύ τους καθίσταται λειτουργική η προκύπτουσα πρωτεΐνη.



**Εικόνα 7:** Δομή της αιμοσφαιρίνης.

Προχωρώντας ένα βήμα πιο χαμηλά σε επίπεδο δομικής οργάνωσης των πρωτεϊνών, συναντούμε την τριτοταγή δομή, η οποία περιγράφει πολυπεπτιδικές αλυσίδες που έχουν αναδιπλωθεί κατάλληλα ώστε να καθίστανται λειτουργικές είτε μεμονωμένες είτε ως δομικά στοιχεία πιο σύνθετων πολυπεπτιδικών σχηματισμών. Η τριτοταγής δομή λοιπόν αφορά την διαμόρφωση μιας πολυπεπτιδικής αλυσίδας στο χώρο με σκοπό την επιτέλεση κάποιας λειτουργίας. Στο παράδειγμα της αιμοσφαιρίνης που αναφέρθηκε παραπάνω, κάθε μια από τις τέσσερις αλυσίδες πριν ενωθεί με τις υπόλοιπες για να σχηματίσουν την αιμοσφαιρίνη, έχει περάσει ένα στάδιο αναδίπλωσης ώστε να λάβει συγκεκριμένη τρισδιάστατη δομή.

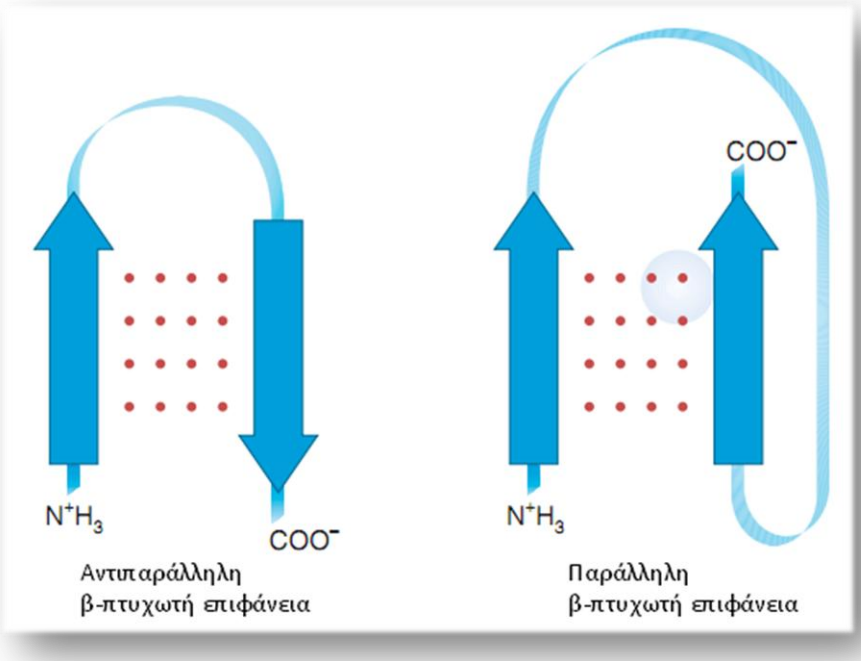
Μια προσεκτική επισκόπηση της τριτοταγούς δομής θα αποκαλύψει επαναλήψεις και συνδυασμούς απλούστερων δομών που επίσης φέρουν συγκεκριμένη αναδίπλωση στο χώρο. Οι πιο βασικές δομές αυτού του τύπου, που συνθέτουν ως επί το πλείστον την δευτεροταγή δομή οργάνωσης των πρωτεϊνών είναι η α-έλικα (α-helix) και η β-πτυχωτή επιφάνεια (β-pleated sheet) [6]. Η α-έλικα αποτελεί το συνηθέστερα απαντώμενο στοιχείο δευτεροταγούς δομή και έχει σχήμα που μοιάζει με ραβδόμορφο σπείραμα. Η χαρακτηριστική δομή της (Εικόνα 8) σταθεροποιείται με ένα επαναλαμβανόμενο μοτίβο δεσμών υδρογόνου ανάμεσα σε κοντινά άτομα του πολυπεπτιδικού σκελετού.



**Εικόνα 8:** Χαρακτηριστική δομή α-έλικας.

Η διαμόρφωση της β-πτυχωτής επιφάνειας ή β-επιφάνειας ήταν η δεύτερη διαμόρφωση που προτάθηκε ως ενεργειακά σταθερή δομή στις πρωτεΐνες [6]. Στην δομή αυτή οι πολυπεπτιδικές αλυσίδες είναι τοποθετημένες παράλληλα και με τέτοιο τρόπο ώστε να σχηματίζεται ο μέγιστος αριθμός δεσμών υδρογόνου, μεταξύ ατόμων του

πολυπεπτιδικού σκελετού. Κάθε τμήμα της πολυπεπτιδικής αλυσίδας που συμμετέχει σε μια β-πτυχωτή επιφάνεια ονομάζεται β-κλώνος. Ένα ζεύγος παρακείμενων κλώνων που αλληλεπιδρούν μέσω δεσμών υδρογόνου μπορεί να έχουν την ίδια ή αντίθετη κατεύθυνση, χαρακτηρίζοντας κατ' επέκταση και ολόκληρη την πτυχωτή επιφάνεια ως παράλληλη ή αντιπαράλληλη, αντιστοίχως (Εικόνα 9).



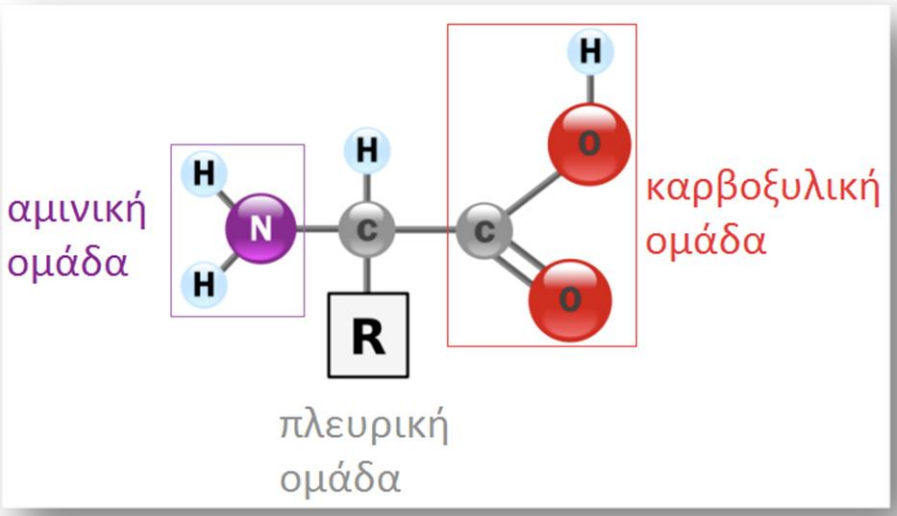

**Εικόνα 9:** Χαρακτηριστική δομή παράλληλης και αντιπαράλληλης β-πτυχωτής επιφάνειας.

Στο αμέσως χαμηλότερο και τελευταίο επίπεδο οργάνωσης των πρωτεϊνών βρίσκεται η πρωτοταγής δομή που περιγράφει ουσιαστικά γραμμικές ακολουθίες από τους θεμέλιους λίθους των πρωτεϊνών, τα αμινοξέα. Οι πρωτεΐνες του οργανισμού προκύπτουν από διαφορετικούς συνδυασμούς μεταξύ των 20 αμινοξέων (Πίνακας 1). Το ίδιο αλφάβητο των 20 χαρακτήρων είναι αυτό που δομεί όλες τις πρωτεΐνες, σε όλα τα γνωστά είδη από τα βακτήρια μέχρι τον άνθρωπο.

**Πίνακας 1:** Τα 20 αμινοξέα που συνθέτουν τις πρωτεΐνες των ζωντανών οργανισμών.

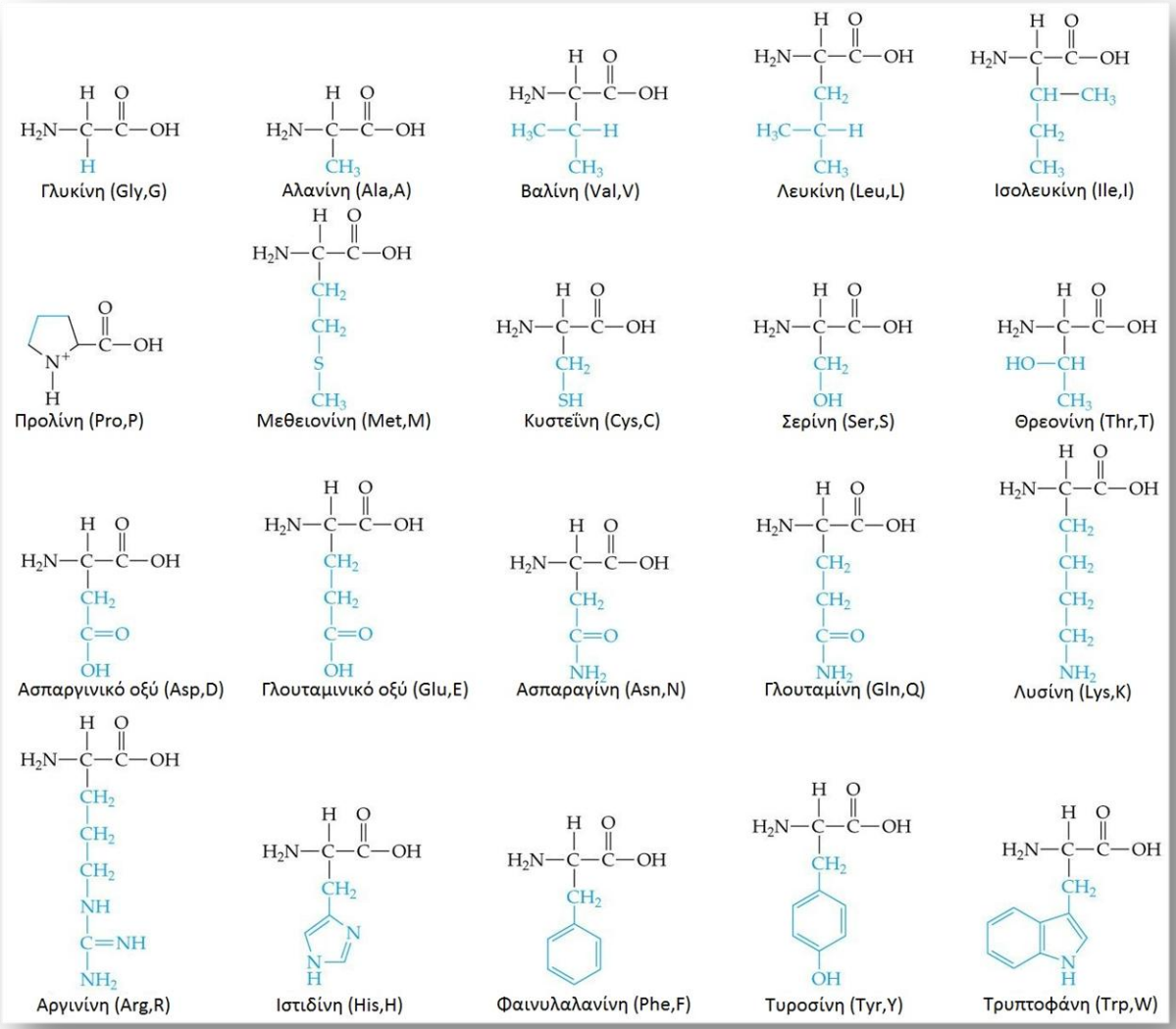
| Αμινοξύ | Αγγλική ονομασία | Συντομογραφία 3 γραμμάτων | Συντομογραφία 1 γράμματος |
|---|---|---|---|
| Αλανίνη | Alanine | Ala | A |
| Αργινίνη | Arginine | Arg | R |
| Ασπαραγίνη | Asparagine | Asn | N |
| Ασπαραγινικό οξύ | Aspartic acid | Asp | D |
| Βαλίνη | Valine | Val | V |
| Γλουταμίνη | Glutamine | Gln | Q |
| Γλουταμινικό οξύ | Glutamic acid | Glu | E |
| Γλυκίνη | Glycine | Gly | G |
| Θρεονίνη | Threonine | Thr | T |
| Ισολευκίνη | Isoleucine | Ile | I |
| Ιστιδίνη | Histidine | His | H |
| Κυστεΐνη | Cysteine | Cys | C |
| Λευκίνη | Leucine | Leu | L |
| Λυσίνη | Lysine | Lys | K |
| Μεθειονίνη | Methionine | Met | M |
| Προλίνη | Proline | Pro | P |
| Σερίνη | Serine | Ser | S |
| Τρυπτοφάνη | Tryptophan | Trp | W |
| Τυροσίνη | Tyrosine | Tyr | Y |
| Φαινυλαλανίνη | Phenylalanine | Phe | F |

Η γενική δομή των αμινοξέων απεικονίζεται στην Εικόνα 10. Το κεντρικό άτομο άνθρακα (α-άνθρακας) συνδέεται με μια αμινική ομάδα, μια καρβοξυλική ομάδα, ένα άτομο υδρογόνου και μια ομάδα R ή πλευρική αλυσίδα.




**Εικόνα 10:** Γενική δομή των αμινοξέων.

Η χαρακτηριστική ομάδα R είναι και αυτή ουσιαστικά που διακρίνει και διαφοροποιεί τα αμινοξέα μεταξύ τους. Υπάρχουν 20 είδη πλευρικών αλυσίδων στις πρωτεΐνες, που διαφέρουν μεταξύ τους ως προς το μέγεθος, το σχήμα, το φορτίο, τη δυνατότητα σχηματισμού δεσμών υδρογόνου, την υδροφοβικότητα και την χημική αντιδραστικότητα. Χάρη στην ποικιλία και τις ιδιότητες των πλευρικών αλυσίδων, το κάθε αμινοξύ παρουσιάζει ιδιαίτερα χαρακτηριστικά, καθώς επίσης και ομοιότητες είτε σε μικρό είτε σε μεγαλύτερο βαθμό με άλλα αμινοξέα (Εικόνα 11).



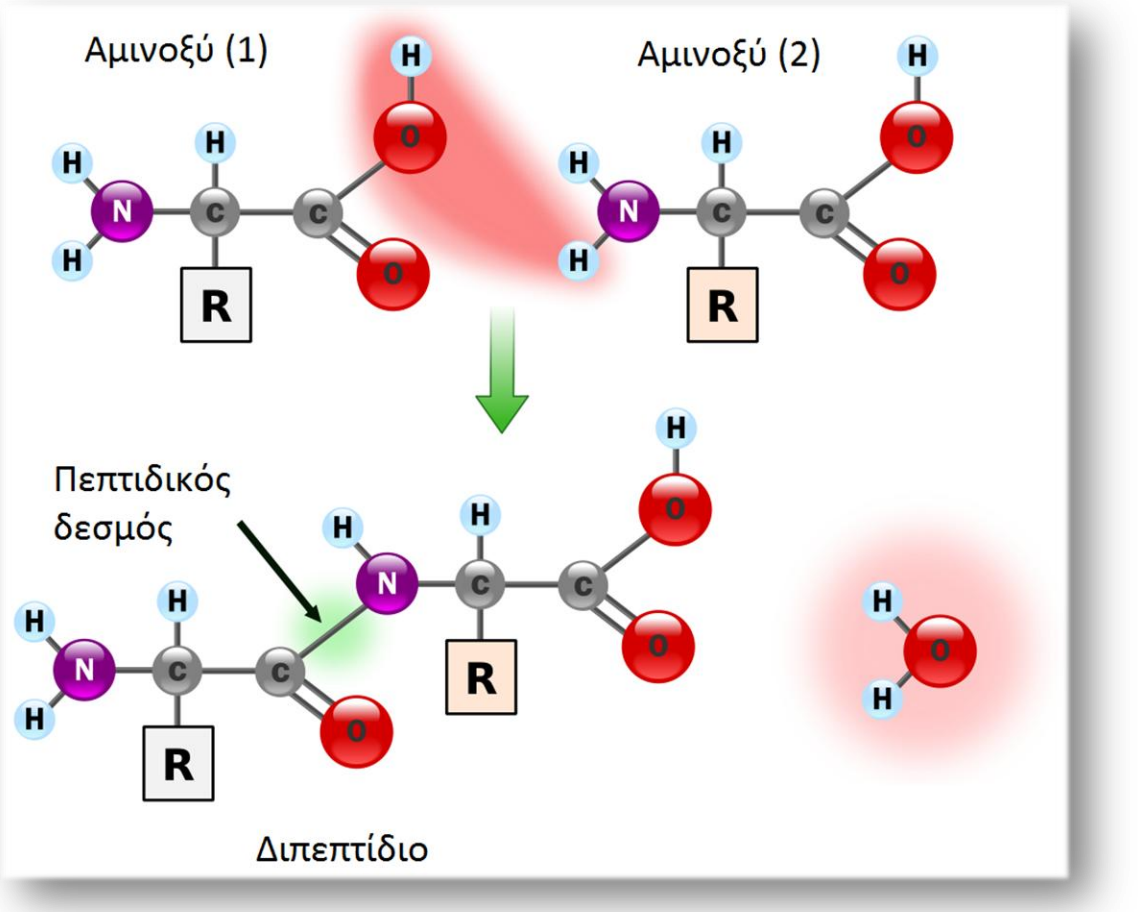

**Εικόνα 11:** Τα 20 βασικά αμινοξέα.

Όπως παρατηρούμε στην Εικόνα 11, η γλυκίνη και η αλανίνη είναι τα πιο απλά αμινοξέα, μιας και η πλευρική τους αλυσίδα αποτελείται από ένα άτομο υδρογόνου και μια μεθυλική ομάδα (-$CH_3$), αντίστοιχα. Μεγαλύτερες υδατανθρακικές αλυσίδες διαθέτουν η βαλίνη, η λευκίνη και η ισολευκίνη. Ακολούθως, η μεθειονίνη περιλαμβάνει στην πλευρική της αλυσίδα έναν θειοαιθέρα (-S-), παρ' όλα αυτά λογίζεται και ως αλειφατικό αμινοξύ. Οι μεγαλύτερες αλειφατικές αλυσίδες είναι κατά κανόνα υδρόφοβες, δηλαδή, τείνουν να δημιουργούν συσσωματώματα ώστε να αποφεύγουν την επαφή με το νερό. Η προλίνη διαθέτει επίσης αλειφατική πλευρική αλυσίδα, διαφέρει όμως από τα άλλα αμινοξέα διότι η πλευρική της ομάδα συνδέεται και με το άτομο του αζώτου και με τον α-άνθρακα. Η χαρακτηριστική δομή της προλίνης μπορεί να επηρεάσει σημαντικά την

τρισδιάστατη αναδίπλωση του πρωτεϊνικού μορίου, μιας και ο δακτύλιος την κάνει πιο άκαμπτη από τα υπόλοιπα αμινοξέα [7]. Η φαινυλαλανίνη, η τυροσίνη και η τρυπτοφάνη διαθέτουν στο μόριό τους αρωματικό δακτύλιο, συγκεκριμένα η φαινυλαλανίνη όπως υποδηλώνει και το όνομά της περιέχει φαινολικό δακτύλιο, η τρυπτοφάνη ινδολικό συνδεδεμένο με μια μεθυλενική ομάδα (-$CH_2$-) και η τυροσίνη έχει στο δακτύλιό της υδροξύλιο. Η τυροσίνη και η τρυπτοφάνη είναι ελαφρώς υδρόφοβες, ενώ η φαινυλαλανίνη παρουσιάζει καθαρά υδροφοβικό χαρακτήρα. Τα αμινοξέα σερίνη και θρεονίνη, περιέχουν αλειφατικές υδροξυλικές ομάδες, γεγονός που τα καθιστά ιδιαιτέρως υδρόφιλα. Η κυστεΐνη μοιάζει αρκετά με τη σερίνη, με τη διαφορά ότι το υδροξύλιο (-OH) της σερίνης έχει αντικατασταθεί από μια σουλφυδρυλική (-SH) ομάδα. Η λυσίνη και η αργινίνη χαρακτηρίζονται από τις σχετικά μακριές πλευρικές αλυσίδες. Τα δύο αυτά αμινοξέα μαζί και με την ιστιδίνη είναι θετικά φορτισμένα και κατά συνέπεια υδρόφιλα. Τέλος, το ασπαραγινικό οξύ και το γλουταμινικό οξύ περιέχουν όξινες πλευρικές αλυσίδες.

## *1.2 Πεπτιδικοί δεσμοί*

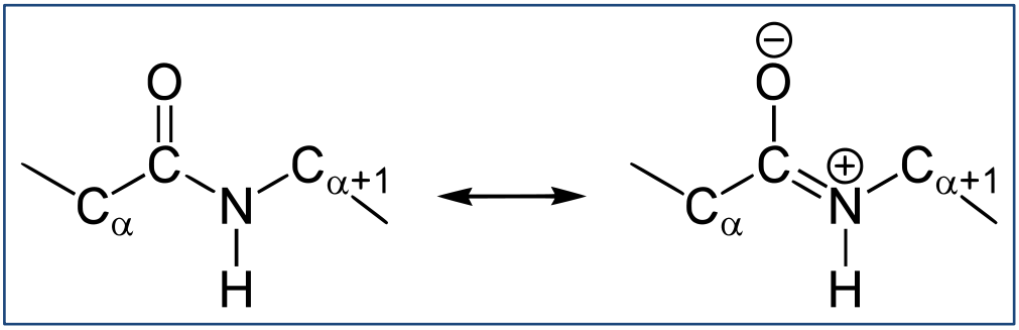
Όπως αναφέρθηκε και προηγουμένως, στη βάση της πρωτεϊνικής οργάνωσης βρίσκεται η πρωτοταγής δομή, η οποία κωδικοποιεί τα ανώτερα επίπεδα οργάνωσης των πρωτεϊνών και παίζει κατ' αυτόν τον τρόπο καθοριστικό ρόλο στις βιολογικές διεργασίες του κυττάρου και του οργανισμού. Η πρωτοταγής δομή αφορά την γραμμική αλληλουχία αμινοξέων διαδοχικά συνδεδεμένων μεταξύ τους με πεπτιδικό δεσμό. Συγκεκριμένα, η καρβοξυλική ομάδα ενός αμινοξέος αντιδρά με την αμινική ομάδα του επόμενου οπότε και σχηματίζεται ένα διπεπτίδιο συνοδευόμενο από την απώλεια ενός μορίου ύδατος (Εικόνα 12). Σε μικρά πεπτίδια, ο αριθμός των αμινοξέων καταδεικνύεται από τα προθέματα δι- (2 αμινοξέα), τρι- (3 αμινοξέα), τετρα- (4 αμινοξέα), κ.ο.κ.




**Εικόνα 12:** Η δημιουργία του πεπτιδικού δεσμού.

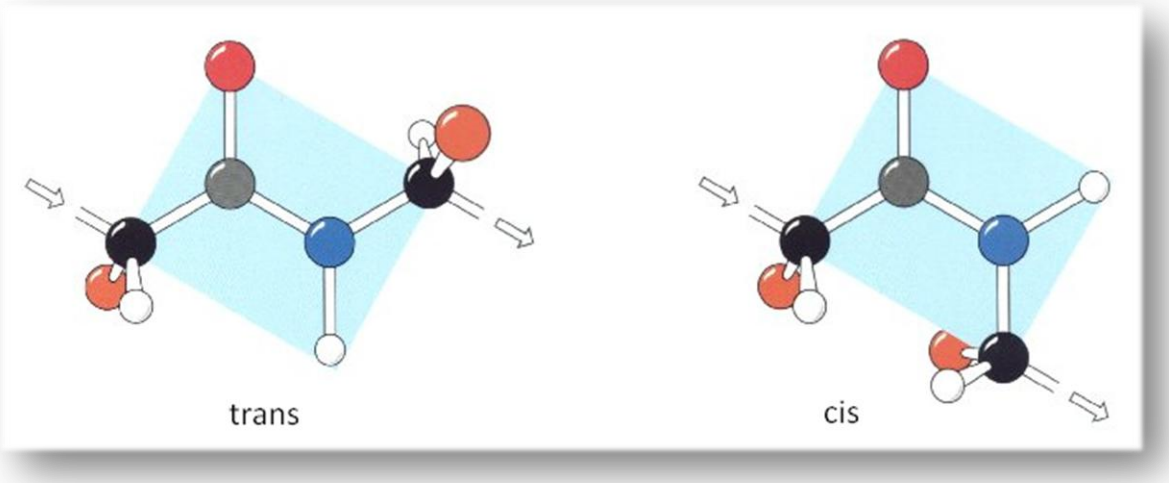
Μια σειρά αμινοξέων που ενώνονται με πεπτιδικούς δεσμούς σχηματίζουν μια πολυπεπτιδική αλυσίδα, όπου κάθε μονάδα αμινοξέος στο πολυπεπτίδιο καλείται κατάλοιπο. Μια πολυπεπτιδική αλυσίδα διαθέτει πολικότητα καθώς στο ένα άκρο έχει την αμινική ομάδα και στο άλλο άκρο την καρβοξυλική ομάδα. Συμβατικά θεωρούμε ότι το αμινοτελικό άκρο σηματοδοτεί την αρχή της πολυπεπτιδικής αλυσίδας ενώ το καρβοξυλικό βρίσκεται στο τέλος της.

Από γεωμετρικής απόψεως, ο πεπτιδικός δεσμός είναι βασικά επίπεδος, δηλαδή, σε κάθε ζεύγος αμινοξέων οι δύο διαδοχικοί α-άνθρακες, η ομάδα -CO του πρώτου αμινοξέος και η ομάδα -NH του δεύτερου αμινοξέος βρίσκονται στο ίδιο επίπεδο. Αυτό οφείλεται στο γεγονός ότι ο πεπτιδικός δεσμός έχει εν μέρει χαρακτήρα διπλού δεσμού (Εικόνα 13), ο οποίος αποτρέπει την περιστροφή γύρω από τον εαυτό του.




**Εικόνα 13:** Δομές συντονισμού του πεπτιδικού δεσμού.

Το γεγονός ότι ο δεσμός δεν περιστρέφεται περιορίζει τις δυνατές στερεοδιατάξεις του πεπτιδικού δεσμού και εξηγεί την επίπεδη φύση του. Στην *trans* διαμόρφωση οι δύο α-άνθρακες βρίσκονται σε διαφορετικές μεριές σε σχέση με τον πεπτιδικό δεσμό, ενώ στην *cis* διαμόρφωση βρίσκονται στην ίδια μεριά, όπως ακριβώς υποδηλώνεται και από την κυριολεκτική μετάφραση των λατινικών όρων *cis*, *trans* (Εικόνα 14).

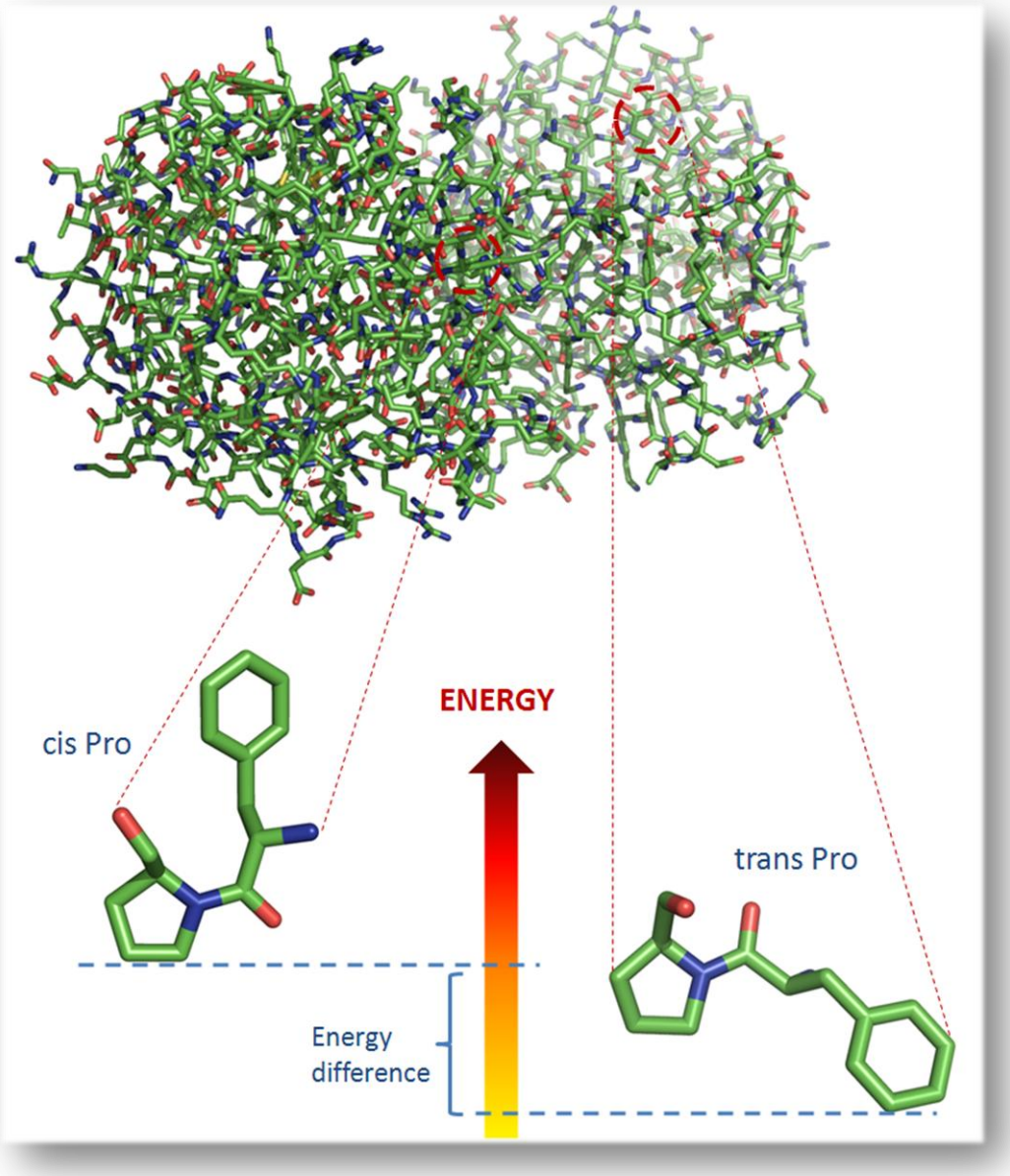



**Εικόνα 14:** Γεωμετρικά ισομερή του πεπτιδικού δεσμού.

Για τον χαρακτηρισμό των δεσμών σε *cis* και *trans* προσδιορίζεται η τιμή της δίεδρης γωνίας ω [$C_a$(i)-C(i)-N(i+1)-$C_a$(i+1)], που λαμβάνει τιμές ω=0° και ω=180°, αντίστοιχα για την κάθε διαμόρφωση. Αξίζει να σημειωθεί ότι η απόσταση $C_a$(i)-$C_a$(i+1) στην *cis* διαμόρφωση, είναι κατά 1Å βραχύτερη σε σχέση με την *trans* διαμόρφωση και γι' αυτό υπάρχει μεγάλη εξάρτηση ανάμεσα στην ανάλυση που έχει προσδιοριστεί η τρισδιάστατη δομή της πρωτεΐνης και στο πλήθος των *cis* πεπτιδικών δεσμών που εντοπίζονται [8]. Το γεγονός αυτό είχε ως αποτέλεσμα πολλοί *cis* δεσμοί είτε να περάσουν απαρατήρητοι, είτε

να σημανθούν εσφαλμένα ως *trans*. Γι' αυτό και σε πολλές περιπτώσεις έχει επαναληφθεί ο προσδιορισμός της τρισδιάστατης δομής ορισμένων πρωτεϊνών, με μεγαλύτερη πλέον διακριτική ικανότητα, εντοπίζοντας κατά συνέπεια και μεγαλύτερο πλήθος *cis* πεπτιδικών δεσμών. Η έλλειψη πρωτεϊνικών δομών υψηλής ευκρίνειας και η συνακόλουθη ασάφεια που αφορούσε το χαρακτηρισμό και τη διάκριση των πεπτιδικών δεσμών, αποτελούσε τροχοπέδη για την εξερεύνηση ιδιαίτερα των *cis* διαμορφώσεων. Παρ' όλα αυτά, τα τελευταία χρόνια υπάρχει ένας μεγάλος ρυθμός καταχώρησης πρωτεϊνικών δομών υψηλής ανάλυσης σε βάσεις δεδομένων στο διαδίκτυο.

Σε κάθε περίπτωση, οι *cis* πεπτιδικοί δεσμοί εντοπίζονται σε πολυπεπτίδια με σαφώς μικρότερη συχνότητα σε σχέση με τους *trans* [9], λόγω υψηλότερου ενεργειακού φορτίου (Εικόνα 15). Ωστόσο, όταν ο πεπτιδικός δεσμός έχει ως δεύτερο κατάλοιπο την προλίνη, οπότε και θα σηματοδοτείται για χάριν συντομίας ως X-Pro (όπου X μπορεί να είναι οποιοδήποτε από τα 20 αμινοξέα), η διαφορά ενέργειας μεταξύ των *cis* και *trans* διαμορφώσεων είναι μικρότερη σε σχέση με έναν X-nonPro (όπου nonPro είναι οποιοδήποτε αμινοξύ εκτός από την προλίνη) πεπτιδικό δεσμό. Συγκεκριμένα, μόνο το 0.03% των X-nonPro πεπτιδικών δεσμών βρίσκονται σε *cis* διαμόρφωση, ενώ όσον αφορά τους X-Pro πεπτιδικούς δεσμούς, το αντίστοιχο ποσοστό είναι 5.2% [8]. Από εδώ και στο εξής οι X-Pro και X-nonPro πεπτιδικοί δεσμοί σε *cis* διαμόρφωση θα καλούνται *cis*-Pro και *cis*-nonPro, αντίστοιχα, ενώ στην περίπτωση της *trans* διαμόρφωσης θα καλούνται *trans*-Pro και *trans*-nonPro. Επίσης, σε μελέτες που πραγματοποιήθηκαν χρησιμοποιώντας μικρού μήκους πολυπεπτιδικές αλυσίδες με τη χρήση Πυρηνικού Μαγνητικού Συντονισμού (NMR: Nuclear Magnetic Resonance), καταδείχτηκε ότι η πρωτοταγής ακολουθία των αμινοξέων καθορίζει την διαμόρφωση του πεπτιδικού δεσμού [10]. Δηλαδή, η τελική διαμόρφωση ενός πεπτιδικού δεσμού μεταξύ δύο αμινοξέων, καθορίζεται σε ένα μεγάλο βαθμό από τα κατάλοιπα που βρίσκονται εκατέρωθεν και περιβάλλουν τον εν λόγω πεπτιδικό δεσμό. Προς την ίδια κατεύθυνση έχει παρατηρηθεί ότι οι δύο διαμορφώσεις του πεπτιδικού δεσμού εντοπίζονται με διαφορετική συχνότητα στα βασικά στοιχεία της δευτεροταγούς δομής των πρωτεϊνών [11, 12]. Έχει ακόμη καταδειχτεί ότι οι *cis* δεσμοί εμφανίζουν προτίμηση σε περιοχές του πρωτεϊνικού μορίου που βρίσκονται περισσότερο εκτεθειμένες στο διαλύτη [12].




**Εικόνα 15:** Ισομερείς διαμορφώσεις ενός Phe-Pro πεπτιδικού δεσμού.

Οι *cis* πεπτιδικοί δεσμοί είναι πολύ σημαντικοί για ένα πλήθος βιολογικών διεργασιών. Συγκεκριμένα, όσον αφορά τους X-Pro πεπτιδικούς δεσμούς, ο ισομερισμός τους καταλύεται από τα ένζυμα Peptidyl Prolyl Isomerases (PPIases), που συμμετέχουν ενεργά στην επαγωγή και εξέλιξη πολλών ασθενειών όπως το AIDS, διάφορες μορφές καρκίνου, Alzheimer καθώς και άλλες νευροεκφυλιστικές παθήσεις [13]. Πρόσφατες μελέτες έχουν δείξει ότι ο ισομερισμός ενός X-Pro πεπτιδικού δεσμού δρα ως μοριακός διακόπτης, ο οποίος μπορεί να ελέγχει το εύρος, το χρονισμό και τη διάρκεια μιας κυτταρικής διεργασίας. Κατά συνέπεια, ο έλεγχος του θα μπορούσε να αποτελέσει πρόσφορο στόχο

για τη δράση φαρμάκων καθώς και θεραπευτικές παρεμβάσεις [14]. Ένα ακόμη γεγονός που καταδεικνύει την μεγάλη σημασία των *cis*-Pro πεπτιδικών δεσμών είναι ότι τα αμινοξέα (προλίνες) που φέρουν αυτούς τους δεσμούς, διατηρούνται εξελικτικά περισσότερο από τα γειτονικά τους αμινοξέα, τα οποία διατηρούνται σε ίδιο βαθμό με την υπόλοιπη πολυπεπτιδική αλυσίδα [15]. Επίσης, σημαντικό εύρημα αποτελεί και ο συχνός εντοπισμός *cis* πεπτιδικών δεσμών, και ιδίως της κατηγορίας X-nonPro, μέσα ή κοντά σε περιοχές υψηλής λειτουργικής σημασίας, όπως ενεργά κέντρα πρωτεϊνών, ή εν γένει σε περιοχές που σχετίζονται με την λειτουργία του πρωτεϊνικού μορίου [11, 16, 17]. Έχει προταθεί ακόμη πως οι *cis* πεπτιδικοί δεσμοί, λόγω της αυξημένης ενέργειας που ενέχουν, αποτελούν αποθήκη ενέργειας για την πρωτεΐνη [16]. Τέλος, ο ισομερισμός του πεπτιδικού δεσμού, παίζει καθοριστικό ρόλο στην αναδίπλωση της πολυπεπτιδικής αλυσίδας στο χώρο, τον τεμαχισμό της αλυσίδας, την μεταφορά σημάτων μέσα στο κύτταρο αλλά και την μεταβίβαση μηνυμάτων και μορίων διαμέσου της κυτταρικής μεμβράνης [13, 18].

Λόγω της μεγάλης σημασίας που έχουν οι *cis* πεπτιδικοί δεσμοί στην διαμόρφωση της τρισδιάστατης δομής των πρωτεϊνών και ακολούθως στην λειτουργία τους, έχουν προταθεί στη βιβλιογραφία διάφορες μεθοδολογίες για την διάκριση μεταξύ των *cis* και *trans* διαμορφώσεων. Η πρώτη εργασία προς αυτή την κατεύθυνση [19], εστίασε αποκλειστικά σε X-Pro πεπτιδικούς δεσμούς και χρησιμοποίησε αποκλειστικά στοιχεία πρωτοταγούς δομής για να προβλέψει την διαμόρφωση του πεπτιδικού δεσμού. Συγκεκριμένα, χρησιμοποίησε ένα σύνολο 242 X-Pro δεσμών για να εξάγει κανόνες που διαχωρίζουν τους *cis* από τους *trans* δεσμούς. Οι κανόνες αυτοί αφορούσαν είτε αποκλειστικά τα αμινοξικά κατάλοιπα, είτε ομαδοποιήσεις αυτών με βάση κάποιες φυσικοχημικές ιδιότητες αυτών (υδροφοβικότητα, πολικότητα, φορτίο, μέγεθος, αλειφατικός/αρωματικός χαρακτήρας). Μεταγενέστερα, οι Wang *et al.* [20] κωδικοποιώντας το κάθε αμινοξύ ως 20-διάστατο διάνυσμα εισήγαγαν την αλληλουχία των αμινοξέων ως είσοδο σε μια Μηχανή Διανυσμάτων Υποστήριξης (SVM: Support Vector Machine) [21] με πολυωνυμικό πυρήνα ώστε να διακρίνουν τις *cis*/*trans* διαμορφώσεις X-Pro πεπτιδικών δεσμών. Ο αλγόριθμος COPS [22] είναι η πρώτη προσπάθεια που δεν στοχεύει αποκλειστικά σε X-Pro δεσμούς αλλά διακρίνει και X-Pro και X-nonPro πεπτιδικούς δεσμούς. Ο αλγόριθμος λαμβάνει υπόψη μόνο την δευτεροταγή δομή από τριπλέτες αμινοξέων και χρησιμοποιεί μια επέκταση των παραμέτρων Chou-Fasman [23] για να προβλέψει την διαμόρφωση του πεπτιδικού δεσμού. Η πιο πρόσφατη μελέτη πραγματοποιήθηκε από τους Song *et al.* [24] με τη δημιουργία του CISPEPpred, ενός

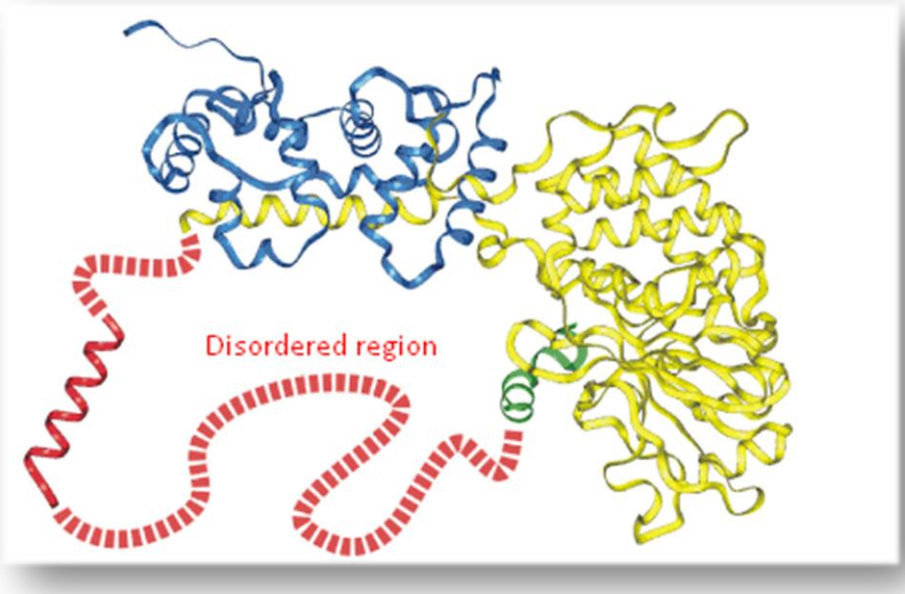
αλγορίθμου για την πρόβλεψη της διαμόρφωσης μόνο X-Pro πεπτιδικών δεσμών. Ο αλγόριθμος βασίζεται σε μια Μηχανή Υποστήριξης Απόφασης που χρησιμοποιεί ως είσοδο στοιχεία πολλαπλής στοίχισης ακολουθιών [25] με την μορφή Position Specific Scoring Matrices (PSSMs) καθώς και την προβλεπόμενη δευτεροταγή δομή των αμινοξέων [26].

Όλες οι παραπάνω μεθοδολογίες, εκτός ίσως από την πρώτη [19], έχουν ως κοινό παρονομαστή την κατασκευή ενός διανύσματος εισόδου το οποίο εν συνεχεία χρησιμοποιείται για την εκπαίδευση ενός ταξινομητή για την διάκριση μεταξύ *cis* και *trans* διαμορφώσεων. Κατ' αυτόν το τρόπο, παρόλο που μπορούμε να διακρίνουμε τους *cis* από τους *trans* δεσμούς, δεν λαμβάνουμε πληροφορία για τους λόγους και τα χαρακτηριστικά που οδήγησαν στη μια ή στην άλλη διαμόρφωση. Συνεπώς, λαμβάνουμε αδιαφανώς ως έξοδο ένα δυαδικό αποτέλεσμα χωρίς να είμαστε σε θέση να κατανοήσουμε περαιτέρω το μοριακό και χημικό υπόβαθρο του *cis*/*trans* ισομερισμού. Προς αυτή την κατεύθυνση, έχουν παρουσιαστεί στη βιβλιογραφία μεθοδολογίες για να επιλύσουν άλλα, παρόμοια προβλήματα που αφορούν στη δομή των πρωτεϊνών. Συγκεκριμένα, οι μεθοδολογίες αυτές, δεν εστιάζουν στην πρόβλεψη-ταξινόμηση, αλλά περιλαμβάνουν ως βασικό βήμα την εξαγωγή ακολουθιακών προτύπων που περιγράφουν την υπό μελέτη πρωτεϊνική περιοχή. Τα ακολουθιακά αυτά πρότυπα ως επί το πλείστον αφορούν την αλληλουχία των αμινοξέων στην πολυπεπτιδική αλυσίδα αλλά κατ' επέκταση μπορούν να εμπλουτιστούν και με πληροφορία σχετικά με τις φυσικοχημικές ιδιότητες των αμινοξέων. Τέτοιου τύπου μελέτες, είχαν αρχικά εφαρμοστεί για την διάκριση μεταξύ των βασικών δομικών στοιχείων της δευτεροταγούς οργάνωσης των πρωτεϊνών [27, 28] καθώς και πιο πρόσφατα για να περιγράψουν τους λόγους που οδηγούν συγκεκριμένες πεπτιδικές περιοχές ή ακόμη και ολόκληρες πρωτεΐνες να είναι ρεομορφικές, όσον αφορά την τρισδιάστατη δομή τους στο χώρο [29].

## *1.3 Ρεομορφικές πρωτεϊνικές περιοχές*

Όπως είδαμε και παραπάνω, οι περιοχές που φέρουν *cis* πεπτιδικούς δεσμούς, είτε πρόκειται για X-Pro είτε για X-nonPro, έχουν μεγάλη βιολογική και λειτουργική σημασία για το κύτταρο. Ένα επίσης ιδιαίτερα ενδιαφέρον ερευνητικό πεδίο που έχει ανακύψει τα τελευταία χρόνια είναι η ύπαρξη πρωτεϊνών που είτε τμηματικά είτε στην ολότητά τους

δεν έχουν σαφή τρισδιάστατη δομή (Εικόνα 16) αλλά χαρακτηρίζονται από δομική ευμεταβλητότητα [30]. Οι περιοχές αυτές ή οι πρωτεΐνες που φέρουν τέτοιες περιοχές χαρακτηρίζονται ρεομορφικές (IDP: Intrinsically Disordered Proteins) και έχει αποδειχτεί ότι εμπλέκονται σε ένα μεγάλο πλήθος κυτταρικών λειτουργιών αλλά και στην επαγωγή πολλών ασθενειών [31].




**Εικόνα 16:** Πρωτεΐνη με ρεομορφική περιοχή στην ακολουθία της.

Η ιδέα των ρεομορφικών πρωτεϊνών, παρόλο που δεν είναι καινούρια έχει προκαλέσει έντονο ερευνητικό ενδιαφέρον τα τελευταία χρόνια. Ουσιαστικά, αποτελεί ένα πλήγμα στην επί σειρά ετών κυριαρχούσα θεωρία "κλειδί και κλειδαριά", όπου για να είναι λειτουργική μια πρωτεΐνη πρέπει να έχει σαφή, σταθερή και μοναδική τρισδιάστατη δομή. Αυτή τη στατική θεώρηση για την λειτουργικότητα των πρωτεϊνών ενίσχυσαν πλήθος ερευνών τα μετέπειτα χρόνια αλλά και οι χιλιάδες τρισδιάστατες δομές πρωτεϊνών που προσδιορίστηκαν ως επί το πλείστον με κρυσταλλογραφία ακτινών Χ. Στην πραγματικότητα όμως, υπάρχουν πάρα πολλές πρωτεΐνες που παρουσιάζουν περιοχές ρεομορφικής τρισδιάστατης δομής, οι οποίες μάλιστα είναι και απαραίτητες για την λειτουργία τους.

Παρά την υψηλή ευμεταβλητότητα στην τρισδιάστατη δομή ή μάλλον χάρη σε αυτήν, οι ρεομορφικές πρωτεΐνες εμπλέκονται σε ένα μεγάλο εύρος σημαντικών λειτουργιών [32-34] με βασική αυτήν της σηματοδότησης. Ο καθοριστικός ρόλος των ρεομορφικών

πρωτεϊνών στην σηματοδότηση ενισχύεται και από το γεγονός ότι τα ευκαρυωτικά κύτταρα διαθέτουν έναν πολύ μεγάλο αριθμό ρεομορφικών πρωτεϊνών, σε σχέση με τα βακτήρια και τα αρχαία/αρχαιοβακτήρια [35, 36]. Επίσης, αξίζει να σημειωθεί ότι περισσότερες από το 70% των σηματοδοτικών πρωτεϊνών αλλά και η συντριπτική πλειοψηφία των καρκινικών πρωτεϊνών διαθέτουν μεγάλες ρεομορφικές περιοχές στο μόριο τους [37].

Ο ιδιαίτερος δομικός τους χαρακτήρας τους επιτρέπει να αντιδρούν με πολλές άλλες πρωτεΐνες και να επιτυγχάνουν μεγάλη εξειδίκευση και ταυτόχρονα μικρή χημική συγγένεια, ιδιαιτέρως σε λειτουργίες σχετικές με κυτταρική ρύθμιση και έλεγχο βιολογικών μονοπατιών [38-40]. Έχει παρατηρηθεί επίσης μεγάλη συσχέτιση ρεομορφικών πρωτεϊνικών περιοχών με μετά-μεταφραστικές τροποποιήσεις (PTM: Post-translational Modifications) [31, 34], όπως για παράδειγμα υδροξυλίωση, μεθυλίωση, φωσφορυλίωση κ.ά. Το ευρύ φάσμα λειτουργιών που έχουν αποδοθεί στις ρεομορφικές πρωτεΐνες καταδεικνύει τον συμπληρωματικό τους χαρακτήρα ως προς τις λειτουργίες των αυστηρά δομημένων πρωτεϊνών και κατ' επέκταση την μεγάλη σημασία τους για την συνολική ρύθμιση και επιβίωση του κυττάρου [31]. Ένα ακόμη σημαντικό πεδίο με το οποίο έχουν συσχετιστεί οι ρεομορφικές πρωτεΐνες είναι και οι μεταγραφικοί παράγοντες. Συγκεκριμένα, πολλοί μεταγραφικοί παράγοντες διαθέτουν στο πρωτέωμά τους μεγάλες ρεομορφικές περιοχές που συμμετέχουν στην λειτουργία τους και με αυτόν τον τρόπο επηρεάζουν σημαντικά την γονιδιακή έκφραση, την ανάπτυξη του κυττάρου και την διαφοροποίηση [31].

Η συμμετοχή και συσχέτιση των ρεομορφικών πρωτεϊνών με έναν τόσο μεγάλο αριθμό λειτουργιών, που καλύπτουν μάλιστα ένα ευρύ φάσμα ετερόκλητων βιολογικών ακολουθιών, τις έχει ενοχοποιήσει και για την επαγωγή και εξέλιξη διαφόρων ασθενειών. Λόγω της δυνατότητάς τους να αλληλεπιδρούν με πολλαπλούς στόχους, έχουν βρεθεί ως συστατικά πολλών βιολογικών μονοπατιών που σχετίζονται με ασθένειες. Ο τρόπος που επάγουν μια ασθένεια σχετίζεται με σφάλματα που λαμβάνουν χώρα κατά την αναγνώριση πρωτεϊνών, ρύθμιση διεργασιών, την περαιτέρω σηματοδότηση καθώς και την αναδίπλωση στο χώρο. Ιδιαίτερα αν αναλογιστούμε το πλήθος των πρωτεϊνών με τις οποίες αλληλεπιδρούν άμεσα ή σχετίζονται και επηρεάζουν έμμεσα οι ρεομορφικές πρωτεΐνες, οδηγούμαστε σε ένα ντόμινο αλυσιδωτών αντιδράσεων και εσφαλμένων λειτουργιών που έχουν ως επακόλουθο μια πληθώρα ασθενειών [31, 41]. Ο ρόλος των ρεομορφικών πρωτεϊνών στον καρκίνο έχει μελετηθεί εκτενώς, όπου έχουν εντοπιστεί σε

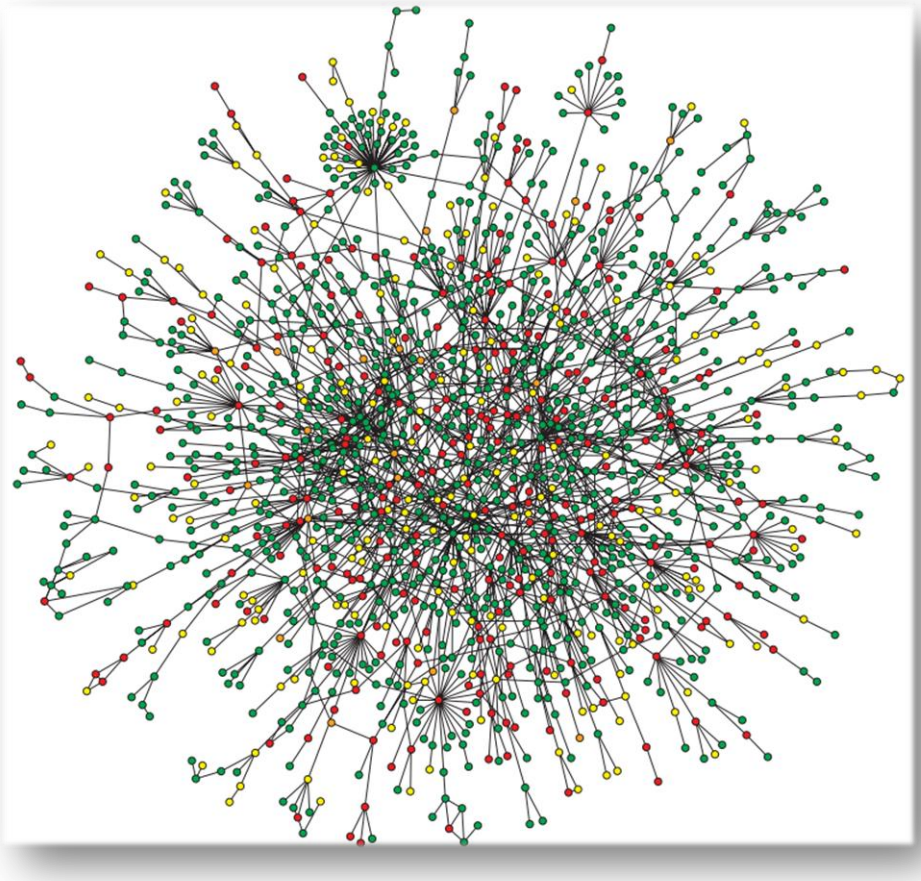
πολλές σημαντικές πρωτεΐνες που σχετίζονται με τον καρκίνο, όπως P53, AFP και BRCA1, μεγάλες ρεομορφικές περιοχές και ιδίως σε λειτουργικές περιοχές τους [37]. Άλλες σημαντικές οικογένειες ασθένειών με τις οποίες έχουν συσχετιστεί ρεομορφικές πρωτεΐνες είναι τα καρδιαγγειακά νοσήματα, νευροεκφυλιστικές ασθένειες (όπως Alzheimer, Parkinson κ.ά.), καθώς ακόμη και ενδοκρινολογικές παθήσεις όπως ο διαβήτης [31].

Ο καταλυτικός ρόλος που διαδραματίζουν οι ρεομορφικές πρωτεΐνες σε ένα μεγάλο εύρος ετερόκλητων λειτουργιών και βιολογικών μονοπατιών, όπως είναι αναμενόμενο τις καθιστά καθοριστικά μέλη σε δίκτυα αλληλεπίδρασης πρωτεϊνών. Πράγματι, η εγγενής τους ευμεταβλητότητα, που κωδικοποιείται από την πρωτοταγή ακολουθία, έχει ως αποτέλεσμα να αλληλεπιδρούν είτε άμεσα είτε έμμεσα με έναν μεγάλο αριθμό πρωτεϊνών. Οι περιοχές αυτές μέσω των οποίων λαμβάνει χώρα η αλληλεπίδραση μεταξύ δυο πρωτεϊνών έχουν μήκος 3-8 αμινοξέα και συνήθως χαρακτηρίζονται από δομική ευμεταβλητότητα. Τέτοιου τύπου πεπτιδικές περιοχές, που μπορούν να περιγραφούν εύληπτα με ακολουθιακά πρότυπα έχουν βρεθεί να συμμετέχουν καθοριστικά σε πολυάριθμες κυτταρικές λειτουργίες. Συνεπώς, η μελέτη της των ρεομορφικών πρωτεϊνών σε ένα δίκτυο αλληλεπίδρασης, μπορεί να φανερώσει ιδιαίτερα σημαντικές πληροφορίες για τον τρόπο με τον οποίο οι ρεομορφικές πρωτεΐνες επάγουν μια συγκεκριμένη λειτουργία μέσω ακολουθιακών προτύπων αλλά και κατ' επέκταση μια δυσλειτουργία ή και ασθένεια [31].

## *1.4 Δίκτυα αλληλεπίδρασης πρωτεϊνών και ασθένειες*

Στην προηγούμενη ενότητα είδαμε πως οι ρεομορφικές πρωτεΐνες σχετίζονται με ένα ευρύ φάσμα λειτουργιών γι' αυτό και κατέχουν καθοριστικό ρόλο σε πρωτεϊνικά δίκτυα αλληλεπιδράσεων, όπου σχετίζονται και αντιδρούν με πολλά και ετερόκλητα πρωτεϊνικά μόρια [42]. Ειδικά τα τελευταία χρόνια, η ανάλυση των γονιδιωμάτων από οργανισμούς-μοντέλα αλλά και στον άνθρωπο έχουν ανοίξει τον δρόμο για την μελέτη της μοριακής εξέλιξης, της ανακάλυψη νέων πρωτεϊνικών λειτουργιών, αλλά και σημαντικότερα την διαλεύκανση της επαγωγής και εξέλιξης ασθενειών σε μοριακό επίπεδο. Η έρευνα σε αυτό το πεδίο εστιάζεται σε τέσσερις βασικές κατευθύνσεις: τον εντοπισμό νέων γονιδίων μιας ασθένειας, την μελέτη του δικτύου αλληλεπιδράσεων που σχετίζονται με μια νόσο, την

αναγνώριση υποδικτύων που εμπλέκονται με νόσους και την ακριβέστερη και πληρέστερη σταδιοποίηση μιας ασθένειας [43]. Υπό αυτό το πρίσμα θεώρησης των ασθενειών, προκύπτουν πολλαπλές εφαρμογές που πιθανά θα προάγουν την εξατομικευμένη ιατρική, την φαρμακολογία, αλλά και την πρόληψη.



**Εικόνα 17:** Δίκτυο αλληλεπίδρασης πρωτεϊνών.

Προς αυτή την κατεύθυνση, έχουν γίνει πολλές προσπάθειες για τη συγκέντρωση και συστηματική καταγραφή δυαδικών αλληλεπιδράσεων μεταξύ πρωτεϊνών. Μερικές από τις κυριότερες και πληρέστερες βάσεις δεδομένων αυτού του τύπου που αξίζει να αναφερθούν είναι οι Human Protein Reference Database (HPRD) [44], Database of Interacting Proteins (DIP) [45], IntAct [46] καθώς και πολλές άλλες που συνοψίζονται στην βάση Pathguide [47]. Πιο κάτω παρατίθενται ορισμένες σημαντικές μελέτες που αφορούν την ανάλυση και συστηματική μελέτη ασθενειών με βάση το μοριακό τους υπόβαθρο όπως καταγράφεται σε δίκτυα αλληλεπίδρασης πρωτεϊνών.

- Οι Jonsson και Bates [48] μελέτησαν την συνδεσιμότητα ενός δικτύου αλληλεπίδρασης αποτελούμενου από 346 πρωτεΐνες που εμπλέκονται με

διάφορους τύπους καρκίνου στον άνθρωπο. Κατέληξαν στο συμπέρασμα ότι οι καρκινικές σε σχέση με τις μη-καρκινικές πρωτεΐνες τείνουν να μετέχουν σε μεγαλύτερα και πιο συμπαγή συμπλέγματα πρωτεϊνών. Σε μια παρόμοια μελέτη, που εστίαζε αποκλειστικά όμως στον καρκίνου του πνεύμονα, παρατηρήθηκε ότι τα γονίδια που εκφράζονται σε αυτόν τον τύπο καρκίνου εμφανίζουν περισσότερες αλληλεπιδράσεις σε σχέση με τα γονίδια που αποσιωπούνται [49].

- Τα παραπάνω συμπεράσματα αντέκρουσε μια σχετική έρευνα που στόχευε στην οικοδόμηση ενός δικτύου γονιδίων που σχετίζονται με την επαγωγή ασθενειών στον άνθρωπο [50]. Συγκεκριμένα, δεν διαπιστώθηκε στατιστικά σημαντικότερη συνδεσιμότητα των γονιδίων που σχετίζονται με ασθένειες σε σχέση με τα υπόλοιπα γονίδια του δικτύου. Μια πιθανή εξήγηση για την υψηλή συνδεσιμότητα που αποδίδεται ορισμένες φορές σε γονίδια σχετικά με ασθένειες, είναι το γεγονός ότι λόγω των δυσμενών προεκτάσεων και συνεπειών τους έχουν μελετηθεί πιο διεξοδικά σε σχέση με άλλα, πιο "απενοχοποιημένα", γονίδια.
- Σημαντικές προσπάθειες έχουν επίσης καταβληθεί για τον εντοπισμό νέων γονιδίων που εμπλέκονται με μια ασθένεια, αλλά και για τον σαφή προσδιορισμό του τρόπου με τον οποίο ένα γονίδιο συμμετέχει και επηρεάζει την εξέλιξη της ασθένειας. Για το σκοπό αυτό έχουν υλοποιηθεί διάφοροι αλγόριθμοι [51-53] που χρησιμοποιώντας ως είσοδο ένα μεγάλο πλήθος από γονιδιακές αλληλεπιδράσεις, επιδιώκουν να τις συσχετίσουν με αντίστοιχους φαινοτύπους. Σε εννοιολογικά παρόμοιες μελέτες, έχουν επεκταθεί οι λίστες με γονίδια που ενοχοποιούνται για μια ασθένεια, μελετώντας αποκλειστικά το υπάρχον δίκτυο αλληλεπίδρασης γονιδίων μιας ασθένειας και εντοπίζοντας κόμβους με καταλυτικό ρόλο στο δίκτυο [54, 55]. Σε ορισμένες περιπτώσεις, οι περαιτέρω πειραματικές μελέτες έχουν επαληθεύσει και εξακριβώσει την εκ-των-δεδομένων αποκτηθείσα γνώση.
- Μια άλλη σημαντική συμβολή των δικτύων αλληλεπίδρασης γονιδίων, αφορά στην εκμετάλλευσή τους για την σταδιοποίηση ασθενειών, καθώς και τον διαχωρισμό ή την κατάταξη ασθενών με βάση έναν δείκτη κινδύνου [56]. Η περαιτέρω μελέτη των υποδικτύων που σχετίζονται με τις επιμέρους ομάδες

ασθενών μπορεί να αποκαλύψει σημαντικές συσχετίσεις και πληροφορίες σχετικά με την πορεία της νόσου.

- Αυτή η ολοένα αυξανόμενη μελέτη των δικτύων αλληλεπίδρασης και η "εισβολή" τους στην ιατρική πρακτική, έχει ανοίξει ένα σημαντικό παράθυρο και στον τομέα της φαρμακολογίας. Συγκεκριμένα, γνωρίζοντας το μόριο-στόχο, και τα μονοπάτια αλληλεπιδράσεων στο οποίο συμμετέχει, μπορούμε συστηματικά να μειώσουμε δυνητικούς αλλά ανεπιθύμητους ενδιάμεσους στόχους, επιτυγχάνοντας κατ' αυτόν τον τρόπο βελτιωμένη και αυξημένη επιλεκτικότητα αλλά και εξάλειψη πιθανών δυσμενών παρενεργειών [57].

Όλες αυτές, καθώς και πολλές άλλες εφαρμογές των δικτύων αλληλεπίδρασης, θα συνεχίσουν να αντιμετωπίζουν τεχνολογικές, βιολογικές και αλγοριθμικές προκλήσεις. Παρά την συνεχή και ολοένα αυξανόμενη ροή βιολογικών δεδομένων, σε ό,τι αφορά τον άνθρωπο και τα δίκτυα ασθενειών του, τα υπάρχοντα δεδομένα παραμένουν ελάχιστα και σε πολλές περιπτώσεις μη στοχευμένα, λόγω της εγγενούς πολυπλοκότητας, του όγκου και της ποικιλομορφίας που τα διακρίνει. Παρ' όλα αυτά η αποσαφήνιση των μηχανισμών που διέπουν τις ανθρώπινες ασθένειες σε μοριακό επίπεδο και η συνακόλουθη εκμετάλλευσή τους στον τομέα της διάγνωσης και της θεραπείας, αποτελεί θεμελιώδη στόχο και επιδίωξη.

## *1.5 Κλινικο-γενετικά δεδομένα στον καρκίνο*

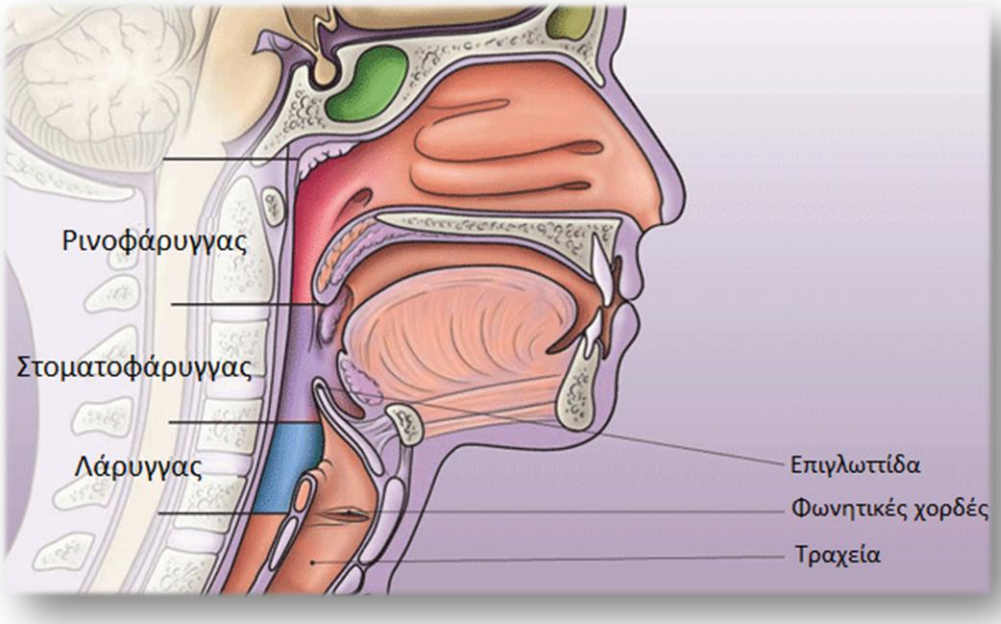
Παρόλο που το γενετικό υπόβαθρο των ασθενειών έχει αρχίσει τα τελευταία χρόνια να μελετάται πιο συστηματικά, πρέπει σε κάθε περίπτωση να εντάσσεται μέσα σε ένα ολιστικό πλαίσιο που λαμβάνει υπόψη περιβαλλοντικούς παράγοντες, το κλινικό ιστορικό του ασθενούς, διατροφικές ή και άλλες καθημερινές συνήθειες (π.χ. κάπνισμα), προγενέστερες νόσους, πιθανές μολύνσεις, φαρμακευτικές αγωγές, καθώς και πολλές άλλες παραμέτρους.

Βάσει επιδημιολογικών μελετών, ο καρκίνος αποτελεί την κύρια αιτία θανάτου παγκοσμίως [58], γι' αυτό και βρίσκεται στο επίκεντρο ερευνητικών μελετών για την αποκάλυψη των παραγόντων που οδηγούν στην εμφάνιση και εξέλιξή του, αλλά και στην έγκαιρη πρόγνωση και θεραπεία του. Ο καρκίνος, λόγω της ποικιλομορφίας που τον

χαρακτηρίζει αλλά και των ετερόκλητων στοιχείων που έχει βρεθεί ότι σχετίζονται με την επαγωγή του, απαιτεί πολυσυστηματική και πολυπαραγοντική προσέγγιση. Ιδιαίτερο βάρος έχει δοθεί τα τελευταία χρόνια στην εκμετάλλευση γενετικών παραγόντων καθώς και στην συσχέτισή τους με αντίστοιχα κλινικά ευρήματα. Συγκεκριμένα, η εμβάθυνση στο γενετικό υπόβαθρο του καρκίνου, έχει συμβάλει στην κατανόηση της βιολογίας του καρκίνου, στην έγκαιρη αναγνώριση ατόμων που βρίσκονται σε υψηλό κίνδυνο ανάπτυξης κάποιας κακοήθειας, στον χαρακτηρισμό και την σταδιοποίηση της νόσου, την ανακάλυψη νέων θεραπειών, εξατομικευμένων και προσαρμοσμένων στο γενετικό αποτύπωμα του καρκίνου. Κατά συνέπεια, η επέκταση της γνώσης μας σε μοριακό πλέον επίπεδο μπορεί να συνεισφέρει σε όλες τις πτυχές της διαχείρισης του καρκίνου, συμπεριλαμβανομένης της πρόληψης, της διάγνωσης καθώς και της θεραπείας [59].

Σημείο αναφοράς για την συμβολή του μοριακού υποβάθρου μιας ασθένειας ως προς την καταπολέμησή της, αποτελεί η ανάπτυξη του σκευάσματος Gleevec που συχνά αναφέρεται και ως το φάρμακο-θαύμα στην θεραπεία του καρκίνου [60]. Το Gleevec, αφορά την χρόνια μυελογενή λευχαιμία, μια σπάνια μορφή καρκίνου, όπου η συστηματική μελέτη και αποκάλυψη της βιολογίας της νόσου, οδήγησε σταδιακά στην θεαματική ίασή της, ανοίγοντας ουσιαστικά ένα παράθυρο που μετέθετε την βιολογία μιας νόσου από το αμιγώς ερευνητικό-θεωρητικό επίπεδο, στην ιατρική πρακτική. Ένα ακόμη από τα πιο πρόσφατα παραδείγματα, που καταδεικνύει την συνεισφορά της μελέτης ενός καρκίνου σε μοριακό επίπεδο, είναι και η ανακάλυψη του μηχανισμού που ο καρκίνος του μαστού δημιουργεί μεταστατικές εστίες στον εγκέφαλο και μάλιστα εξηγεί πως τα καρκινικά κύτταρα διαπερνούν τον αιματοεγκεφαλικό φραγμό [61].

Ένας τύπος καρκίνου, που παρόλο που δεν βρίσκεται στο προσκήνιο, αφορά ένα σημαντικό μέρος του πληθυσμού και λόγω συχνότητας αλλά και λόγω θνησιμότητας είναι ο καρκίνος του στόματος. Συγκεκριμένα, ο καρκίνος του στόματος περιλαμβάνει νεοπλασίες στην στοματική κοιλότητα, τον φάρυγγα και τον λάρυγγα (Εικόνα 18). Βρίσκεται στην $8^{η}$ θέση παγκοσμίως όσον αφορά την συχνότητα εντόπισής του, με περισσότερα από 500.000 άτομα να διαγιγνώσκονται με στοματικό καρκίνο ετησίως [62].




**Εικόνα 18:** Ανατομικά στοιχεία που αφορούν τον στοματικό καρκίνο.

Γενικά, ο καρκίνος του στόματος αποτελεί μια πολύπλοκη και πολυπαραγοντική ασθένεια, που χαρακτηρίζεται από κλινική, φαινοτυπική, παθολογική και βιολογική ετερογένεια. Η εξέλιξη και η πρόοδος του καρκίνου αυτού πιστεύεται ότι προκύπτει από πολλαπλές βαθμιαίες και αλυσιδωτές αλλοιώσεις των κυτταρικών και μοριακών δομών του τοπικού επιθηλίου. Πρωταρχικοί παράγοντες που έχουν ενοχοποιηθεί για την εμφάνιση καρκίνου στο στόμα, αποτελούν το κάπνισμα και η κατανάλωση αλκοόλ και ιδίως ο συνδυασμός τους [63]. Επίσης, έχει παρατηρηθεί σημαντική συσχέτιση με το φύλο, αφού οι άντρες έχουν διπλάσιες πιθανότητες να διαγνωστούν με στοματικό καρκίνο σε σχέση με τις γυναίκες. Άλλοι παράγοντες κινδύνου που έχουν σχετιστεί με συγκεκριμένες εντοπίσεις νεοπλασιών στην στοματοφαρρυγγική οδό είναι η έκθεση στον ήλιο (καρκίνος στο χείλος) [63] καθώς και η μόλυνση από τον ιό HPV (καρκίνος στο στοματοφάρυγγα και την βάση της γλώσσας) [64]. Μια πολύ σημαντική παράμετρος που πρέπει να τονιστεί για τον στοματικό καρκίνο είναι οι ιδιαίτερα δυσμενείς επιπτώσεις που έχει στην ποιότητα ζωής και την καθημερινότητα του ασθενούς, μιας και επηρεάζει συνήθως καταλυτικά την εξωτερική του εμφάνιση και δυσχεραίνει σημαντικά την ομιλία καθώς και την πρόσληψη και κατάποση της τροφής.

Τα καρκινικά κύτταρα, πέρα από την αρχική τους εντόπιση, έχουν την τάση να εξαπλώνονται σε γειτονικές περιοχές του λαιμού, τους πνεύμονες ή σε άλλα μέρη του

σώματος. Πολύ συχνά εμφανίζονται μεταστάσεις σε παρακείμενους λεμφαδένες, μέσω του λεμφικού συστήματος το οποίο διευκολύνει την εξάπλωση των καρκινικών κυττάρων. Η συνεχής μελέτη της νόσου, και η συνακόλουθη βελτίωση των πρωτοκόλλων θεραπείας, έχουν οδηγήσει σε υψηλά ποσοστά επιτυχούς εξαφάνισης της νόσου [65]. Παρ' όλα αυτά υπάρχει ένα κρίσιμο στάδιο κατά την εξέλιξη της νόσου, που καλείται ύφεση, κατά την οποία δεν υπάρχει, ή τουλάχιστον δεν ανιχνεύεται, κλινική, εργαστηριακή ή απεικονιστική απόδειξη την νεοπλασματικής μάζας και ο ασθενής θεωρείται ότι έχει απαλλαγεί από τον καρκίνο. Ωστόσο, ακόμη και σε αυτό το σημείο κάποια μη-ανιχνεύσιμα υπολείμματα της νόσου εξακολουθούν να είναι παρόντα και μπορεί να οδηγήσουν σε μια επανεμφάνιση ή μετάσταση της νόσου. Συγκεκριμένα, όσον αφορά τον καρκίνο του στόματος, τα ποσοστά τοπικής επανεμφάνισης της ασθένειας, μετά την επιτυχή θεραπεία και το στάδιο της ύφεσης, είναι της τάξης του 25-48% [66]. Μάλιστα λόγω της ιδιαίτερα επιθετικής και διεισδυτικής φύσης του συγκεκριμένου καρκίνου, τα ποσοστά αυτά μετρώνται σε διάστημα διετίας από την στιγμή της θεραπείας (χειρουργική αφαίρεση, χημειοθεραπεία, ακτινοθεραπεία) του ασθενούς [66].

Εξίσου απογοητευτική με τα παραπάνω ποσοστά, είναι και η δυνατότητα έγκαιρης πρόβλεψης μια ενδεχόμενης επανεμφάνισης, γεγονός που μπορεί να αποδοθεί εν μέρει και στα δυσμενή ιστολογικά χαρακτηριστικά που παρουσιάζει η εικόνα του συγκεκριμένου καρκίνου [66]. Η καθυστερημένη ανίχνευση μιας επικείμενης επανεμφάνισης, μπορεί να οδηγήσει σε σοβαρότατες επιπλοκές, με ιδιαίτερα δυσμενείς επιπτώσεις, ενώ αντίθετα έχει παρατηρηθεί ότι η αντιμετώπιση της νόσου στα αρχικά στάδια εξέλιξης έχει σημαντικά βελτιωμένη επιβίωση, σε σχέση με πιο προχωρημένα στάδια [67]. Οι μέθοδοι που έχουν αναπτυχθεί μέχρι σήμερα για να προβλέψουν την έκβαση της νόσου μετά την ύφεση, αντιμετωπίζουν τμηματικά συνήθως την ασθένεια και όχι στην ολότητά της, επιτυγχάνοντας κατά συνέπεια ανεπαρκή αποτελέσματα. Ειδικά στη μοριακή βάση της νόσου, οι διαθέσιμοι βιολογικοί δείκτες είναι περιορισμένοι σε αριθμό αλλά και αποτελεσματικότητα [68, 69]. Ο αποτελεσματικός συνδυασμός των ήδη γνωστών βιολογικών δεικτών, αλλά και η ανακάλυψη νέων, σε συνάρτηση με κλινικούς προδιαθεσικούς παράγοντες θα ωφελήσει σημαντικά στην διαστρωμάτωση των ασθενών με βάση την πιθανότητα επανεμφάνισης και θα διευκολύνει σημαντικά την παρακολούθηση εξέλιξης της νόσου κατά τη διάρκεια της ύφεσης.

Στο γενικό πλαίσιο που αφορά την πρόγνωση επανεμφάνισης του καρκίνου αλλά και την εν γένει μοντελοποίηση εξέλιξης της νόσου, χωρίς να εστιάζουμε αποκλειστικά στην

περίπτωση του στοματικού καρκίνου, έχουν προταθεί διάφορες προσεγγίσεις στη βιβλιογραφία. Ως επί το πλείστον τα μοντέλα αυτά υπολογίζουν ένα σκορ κινδύνου για την επανεμφάνιση, που προκύπτει από τον γραμμικό συνδυασμό μιας σειράς μεταβλητών που επηρεάζουν την εξέλιξη της νόσου. Με βάση αυτό το σκορ χρησιμοποιούνται κάποιοι κανόνες για τα ταξινομήσουν τους ασθενείς σε διάφορες κατηγορίες κινδύνου [70, 71]. Πιο πρόσφατες προσεγγίσεις χρησιμοποιούν έναν αλγόριθμο μηχανικής μάθησης, όπως τα τεχνητά νευρωνικά δίκτυα (ANN: Artificial Neural Networks) ή SVM όπου με βάση ένα διάνυσμα εισόδου προβλέπει την κλάση που ανήκει ο ασθενής [72, 73]. Ωστόσο, οι περισσότερες από αυτές τις προσεγγίσεις ενέχουν την λογική "μαύρου κουτιού" εξάγοντας αδιαφανώς μια απόφαση, χωρίς να συνοδεύεται συνήθως από επαρκή αιτιολόγηση. Μια ακόμη σημαντική παράμετρος που υπεισέρχεται στο πρόβλημα πρόβλεψης ενδεχόμενης επανεμφάνισης είναι η δυνατότητα των αλγορίθμων να συμπεριλάβουν και να αντικατοπτρίσουν την εξέλιξη της νόσου στον χρόνο. Δηλαδή, να συσχετίσουν την πιθανότητα μιας ενδεχόμενης επανεμφάνισης μέσα σε ένα αδρό χρονικό πλαίσιο. Συνεπώς, κατά την πρόβλεψη επανεμφάνισης του καρκίνου, οι κρίσιμες παράμετροι που πρέπει να λαμβάνονται υπόψη είναι: η ακρίβεια, η αιτιολόγηση και ο χρόνος, ή με απλά λόγια, να δίνονται απαντήσεις στα ερωτήματα: *αν*, *πότε* και *γιατί*.

## *1.6 Συνεισφορά διδακτορικής διατριβής*

Η παρούσα διδακτορική διατριβή εστιάζεται στην ανάπτυξη και εφαρμογή ευφυών υπολογιστικών μεθόδων για την οργάνωση, επεξεργασία, ανάλυση και κατανόηση μεγάλου όγκου βιολογικών δεδομένων. Απώτερος σκοπός είναι η σταδιακή αποκρυπτογράφηση του γενετικού υποβάθρου ασθενειών και κατ' επέκταση η αποδοτικότερη και στοχευμένη αντιμετώπισή τους.

Η μελέτη ανάγεται σε ένα επαγωγικό πλαίσιο έρευνας που δομείται σταδιακά από το ειδικό προς το γενικό. Συγκεκριμένα, ξεκινάμε από το κατώτερο επίπεδο οργάνωσης των πρωτεϊνών, δηλαδή την πρωτοταγή δομή, μελετώντας την διαμόρφωση του πεπτιδικού δεσμού που συνδέει τα αμινοξέα της πολυπεπτιδικής αλυσίδας. Οι πεπτιδικοί δεσμοί βρίσκονται ως επί το πλείστον στην *trans* διαμόρφωση και σπάνια στην *cis*. Ο ακριβής εντοπισμός της τελευταίας μπορεί να συμβάλλει σε ένα πλήθος εφαρμογών με μεγάλη βιολογική σημασία – από τον λεπτομερέστερο καθορισμό της δομής των πρωτεϊνικών

μορίων και την εξακρίβωση της λειτουργίας τους, μέχρι τον έλεγχο του κυτταρικού κύκλου και την αποκρυπτογράφηση γενετικών ασθενειών. Επίσης, εκτός από την αναγνώριση και διάκριση μεταξύ των ισομερών διαμορφώσεων του πεπτιδικού δεσμού στοχεύουμε στην ανάδειξη των παραγόντων που καθορίζουν την διαμόρφωση του πεπτιδικού δεσμού.

Στο επόμενο βήμα, μελετάμε τις εξαρτήσεις μεταξύ των πρωτεϊνικών μορίων και την συμμετοχή τους στα πολύπλοκα δίκτυα αλληλεπίδρασης πρωτεϊνών. Εστιάζουμε σε πρωτεΐνες που είναι είτε μερικώς είτε στην ολότητά τους ρεομορφικές. Οι ρεομορφικές περιοχές, όμοια με τις περιοχές που φέρουν cis πεπτιδικούς δεσμούς ενέχουν ιδιαίτερη λειτουργική σημασία για το κύτταρο. Πρόκειται για περιοχές όπου η πολυπεπτιδική αλυσίδα δεν λαμβάνει σαφή δομή στο χώρο και χάρη σε αυτή την εγγενή ευμεταβλητότητα οι ρεομορφικές πρωτεΐνες έχουν συσχετιστεί με πλήθος βιολογικών λειτουργιών καθώς και με την επαγωγή σοβαρότατων ασθενειών. Συγκεκριμένα, σε αυτό το πεδίο μελέτης εξάγουμε και αναλύουμε τα ακολουθιακά πρότυπα που καθορίζουν τις αλληλεπιδράσεις, την λειτουργία και τις περαιτέρω συσχετίσεις των ρεομορφικών πρωτεϊνών με γενετικές ασθένειες στον άνθρωπο. Δομούμε λοιπόν κατά τρόπο συστηματικό ένα σύνολο από εύληπτες και περιγραφικές απεικονίσεις που καταγράφουν τον τρόπο και τον μηχανισμό δράσης των ρεομορφικών πρωτεϊνών προς την επιτέλεση των λειτουργιών τους.

Στη συνέχεια της διδακτορικής διατριβής επεκτείνουμε και συνδυάζουμε την ανάλυση από το μικροσκοπικό επίπεδο με μακροσκοπικές παρατηρήσεις και δεδομένα. Συγκεκριμένα, συλλέγουμε βιολογικά δεδομένα που αφορούν την κυτταρική και συστημική λειτουργία του οργανισμού, καθώς και κλινικά δεδομένα (ιατρικό ιστορικό και απεικονιστικά δεδομένα) που αφορούν σε φαινοτυπικό πλέον επίπεδο ανατομικές οντότητες και τον οργανισμό ως σύνολο. Παρά την έντονη "εισβολή" των βιολογικών δεδομένων στην αποκρυπτογράφηση γενετικών ασθενειών, η κλινική εξέταση δεν παύει να αποτελεί ακρογωνιαίο λίθο στην ιατρική διάγνωση οπότε και ο συνδυασμός των δύο κατά τρόπο συμπληρωματικό συνιστά την προσφορότερη οδό. Εφαρμόζουμε αυτή την πολύπλευρη και πολυπαραγοντική ανάλυση σε μια πολύπλοκη νόσο όπως ο καρκίνος – και πιο συγκεκριμένα στον στοματικό καρκίνο – που φέρει εκφάνσεις σε όλα τα επίπεδα της φυσιολογίας του οργανισμού. Ο σκοπός της ανάλυσης είναι η αναγνώριση των επιμέρους σταδίων που λαμβάνουν χώρα κατά την εξέλιξη της νόσου καθώς και των

σημαντικότερων κλινικών παραγόντων που καταδεικνύουν την πρόοδο της θεραπείας του ασθενούς.

Κατά την επαγωγική ανάλυση ένα πρόβλημα κατακερματίζεται σε απλούστερα προβλήματα που μπορούν να αντιμετωπιστούν πιο συστηματικά και αποτελεσματικά και ενδέχεται να άπτονται πολλών επιμέρους εφαρμογών. Αξίζει να σημειωθεί όμως ότι ιδίως σε ό,τι αφορά τα βιολογικά δεδομένα, ο όγκος της παραγόμενης πληροφορίας είναι συνήθως δυσθεώρητος, οπότε και η ανάπτυξη κατάλληλων ευφυών αλγορίθμων για την ανάλυση, επεξεργασία και κατανόησή τους είναι απαραίτητη. Στην παρούσα διδακτορική διατριβή συνδυάζουμε τα επιστημονικά πεδία της Μηχανικής Μάθησης και της Υπολογιστικής Νοημοσύνης με σκοπό την αντιμετώπιση προβλημάτων που ανακύπτουν από την Βιολογία και την Ιατρική.

Η συνεισφορά της παρούσας διδακτορικής διατριβής εντοπίζεται στα ακόλουθα σημεία: (i) στην ανάδειξη του βιολογικού μηχανισμού που καθορίζει την διαμόρφωση του πεπτιδικού δεσμού, (ii) στον συστηματικό εντοπισμό λειτουργικών συσχετίσεων των πρωτεϊνικών περιοχών που φέρουν cis πεπτιδικούς δεσμούς, (iii) στην εξαγωγή ακολουθιακών προτύπων που σηματοδοτούν και επάγουν τις αλληλεπιδράσεις και λειτουργίες των ρεομορφικών πρωτεϊνών, (iv) στην συστηματική καταγραφή του τρόπου με τον οποίο οι ρεομορφικές πρωτεΐνες συμμετέχουν σε γενετικές ασθένειες, (v) στην μελέτη του γενετικού υποβάθρου πολυπαραγοντικών ασθενειών (στοματικός καρκίνος) και την ανάδειξη γενετικών παραγόντων που συμβάλλουν στην εξέλιξή τους, (vi) στη συνδυαστική ανάλυση κλινικών και βιολογικών δεδομένων για την πολύπλευρη πλαισίωση γενετικών ασθενειών, καταγράφοντας μεταβολές στα κύτταρα, τα συστήματα, τους ιστούς, μέχρι και ολόκληρο τον οργανισμό.

# 2ο ΚΕΦΑΛΑΙΟ: Πρόβλεψη διαμόρφωσης πεπτιδικού δεσμού

## *2.1 Σκοπός*

Στο παρόν κεφάλαιο παρουσιάζουμε μια μεθοδολογία για την πρόβλεψη της διαμόρφωσης του πεπτιδικού δεσμού μεταξύ των αμινοξέων μιας πολυπεπτιδικής αλυσίδας. Σε αντίθεση με τις περισσότερες εργασίες της βιβλιογραφίας που επικεντρώνονται μόνο σε δεσμούς X-Pro, η παρούσα προσέγγιση προβλέπει τη διαμόρφωση του πεπτιδικού δεσμού μεταξύ δυο οποιονδήποτε αμινοξέων, λαμβάνοντας υπόψη και τους σπάνιους αλλά πολύ σημαντικούς *cis*-nonPro δεσμούς. Συγκεκριμένα, ομαδοποιούμε τους πεπτιδικούς δεσμούς σε τέσσερις κατηγορίες, *cis*-Pro. *cis*-nonPro, *trans*-Pro και *trans*-nonPro, στις οποίες προσπαθούμε να ταξινομήσουμε τους πεπτιδικούς δεσμούς μιας πρωτεΐνης. Για να το πετύχουμε αυτό, εξάγουμε από την απλή ακολουθιακή πληροφορία της πρωτοταγούς δομής, μια σειρά από χαρακτηριστικά που έχει βρεθεί ότι συσχετίζονται με την διαμόρφωση του πεπτιδικού δεσμού. Το πρώτο χαρακτηριστικό αφορά στοιχεία πολλαπλής στοίχισης ακολουθιών με τη μορφή PSSMs, που έχει αποδειχτεί ότι αποτελεί ένα ιδιαίτερα χρήσιμο χαρακτηριστικό σε πολλά προβλήματα από τον τομέα της Βιοπληροφορικής [26]. Τα υπόλοιπα χαρακτηριστικά που εξάγουμε από κάθε αμινοξύ της πολυπεπτιδικής αλυσίδας και χρησιμοποιούμε ως είσοδο είναι: η δευτεροταγής δομή, η έκθεση στο διαλύτη καθώς και 6 σημαντικές φυσικοχημικές ιδιότητές. Επίσης, κατασκευάζουμε μια συνεπτυγμένη μορφή των PSSM που ενέχουν πληροφορία σχετική με τις φυσικοχημικές ιδιότητες του κάθε αμινοξέος (PSSMX), ένα χαρακτηριστικό που έχει προταθεί και χρησιμοποιηθεί επιτυχώς για την πρόβλεψη ρεομορφικών περιοχών στις πρωτεΐνες [74]. Επιπλέον, εφαρμόζουμε έναν αλγόριθμο επιλογής χαρακτηριστικών ώστε να απαλείψουμε χαρακτηριστικά που είτε είναι άσχετα είτε πλεονάζοντα, δυσχεραίνοντας σε κάθε περίπτωση την διαδικασία της ταξινόμησης των πεπτιδικών δεσμών. Η έξοδος του αλγορίθμου επιλογής χαρακτηριστικών μελετάται διεξοδικά ώστε να προσδιοριστεί η συνεισφορά του κάθε χαρακτηριστικού, αλλά και του κάθε γειτονικού αμινοξέος στην τελική διαμόρφωση του πεπτιδικού δεσμού. Μεγάλη

προσοχή δίνεται ώστε να ξεπεραστεί το μείζον θέμα ανισοκατανομής των δεδομένων στις κλάσεις, καθώς και στην κατά το δυνατόν πιο αξιόπιστη εκτίμηση του σφάλματος ταξινόμησης.

## *2.2 Δεδομένα*

Το σύνολο δεδομένων που χρησιμοποιήσαμε κατά την ανάπτυξη της παρούσας μεθοδολογίας περιλαμβάνει 3050 πρωτεϊνικές ακολουθίες από την βάση Protein Data Bank (PDB) [75] που επιλέχθηκαν με το εργαλείο Protein Sequence Culling Server (PISCES) [76]. Οι παραπάνω δομές έχουν προσδιοριστεί με την χρήση κρυσταλλογραφίας ακτινών Χ και ανάλυση μεγαλύτερη από 2.0 Å, R-factor μικρότερο από 0.25, ενώ η μεταξύ τους ομοιότητα δεν ξεπερνάει το 25%. Από τις τρισδιάστατες συντεταγμένες που είναι καταχωρημένες στην PDB, για να εξάγουμε τις απαραίτητες δίεδρες γωνίες μεταξύ των αμινοξέων της πολυπεπτιδικής αλυσίδας χρησιμοποιήσαμε το εργαλείο Volume Area Dihedral Angle Reporter (VADAR) [77]. Οι δεσμοί των οποίων η δίεδρη γωνία βρίσκεται στο διάστημα [-30°, +30°] χαρακτηρίζονται ως *cis* ενώ όλοι οι υπόλοιποι ως *trans*. Οι δεσμοί που προκύπτουν κατανέμονται στις τέσσερις κλάσεις ως εξής: 318 *cis*-nonPro, 1416 *cis*-Pro, 30657 *trans*-Pro και 657968 *trans*-nonPro.

## *2.3 Μεθοδολογία*

Η προτεινόμενη μεθοδολογία περιλαμβάνει 3 στάδια όπως φαίνεται στην Εικόνα 19: α) εξαγωγή χαρακτηριστικών, β) επιλογή χαρακτηριστικών και γ) κατηγοριοποίηση πεπτιδικών δεσμών. Στο πρώτο στάδιο εξάγουμε από την αμινοξική ακολουθία τα PSSM, τη δευτεροταγή δομή, την έκθεση στο διαλύτη και τις φυσικοχημικές ιδιότητες. Το διάνυσμα εισόδου που αποτελείται από αυτά τα χαρακτηριστικά ονομάζεται FV-PSSM. Ακολούθως, ο συνδυασμός των PSSM με τις φυσικοχημικές ιδιότητες δίδει τα συνεπτυγμένα PSSM (PSSMX) που μαζί με την δευτεροταγή δομή και την έκθεση στο διαλύτη συνιστούν το δεύτερο διάνυσμα εισόδου, που καλείται FV-PSSMX. Στη συνέχεια ο αλγόριθμος επιλογής χαρακτηριστικών εντοπίζει τα χαρακτηριστικά με την μέγιστη

διακριτική ικανότητα τα οποία εν συνεχεία εισάγονται σε έναν ταξινομητή για να διακρίνουμε σε ποια κατηγορία ανήκει ο κάθε πεπτιδικός δεσμός.


**Εικόνα 19:** Τα στάδια της προτεινόμενης μεθοδολογίας.

*Εξαγωγή χαρακτηριστικών*

Ένα μεγάλο πλήθος χαρακτηριστικών εξάγονται από την αμινοξική ακολουθία και θα αποτελέσουν το διάνυσμα εισόδου που θα χρησιμοποιηθεί στα επόμενα στάδια της μεθοδολογίας. Αρχικά εξάγουμε τα PSSM με τη χρήση του προγράμματος PSI-BLAST [25], όπου συγκεκριμένα εκτελούμε τρείς επαναλήψεις απέναντι στην βάση δεδομένων NCBI nr [78] με κατώφλι $10^{-3}$ για την παράμετρο E-value. Οι εξαχθείσες τιμές κανονικοποιούνται στη συνέχεια στο διάστημα [0,1]. Για την πρόβλεψη της δευτεροταγούς δομής χρησιμοποιούμε το πρόγραμμα PSIPRED [79] το οποίο παράγει δείκτες αξιοπιστίας για τρείς πιθανές καταστάσεις της δευτεροταγούς δομής, συγκεκριμένα H (helix), E (strand) και Coil (L). Ακολούθως, για τον υπολογισμό της έκθεσης στο διαλύτη, χρησιμοποιούμε το πρόγραμμα RVP-net [80], το οποίο μάλιστα υπολογίζει πραγματικές τιμές έκθεσης για κάθε κατάλοιπο στην πρωτεϊνική ακολουθία,

εισάγοντας κατ' αυτόν τον τρόπο περισσότερη πληροφορία από ότι η χρήση δυαδικών ή εν γένει διακριτών προβλέψεων. Επίσης, για κάθε αμινοξύ εξάγουμε και 6 φυσικοχημικές ιδιότητες, συγκεκριμένα: όγκο, υδροφοβικότητα, πολικότητα, φορτίο, αρωματικός/αλειφατικός χαρακτήρας [81, 82], οι τιμές των οποίων παρατίθενται στον Πίνακας 2.

**Πίνακας 2:** Φυσικοχημικές ιδιότητες των 20 αμινοξέων.

| Ιδιότητα / Αμινοξύ | Όγκος | Υδροφοβικότητα | Πολικότητα | Φορτίο | Αρωματικός χαρακτήρας | Αλειφατικός χαρακτήρας |
|---|---|---|---|---|---|---|
| **A** | 31 | 0.61 | 8.1 | 0 | 0 | 0 |
| **R** | 124 | 0.6 | 10.5 | 1 | 0 | 0 |
| **N** | 56 | 0.06 | 11.6 | 0 | 0 | 0 |
| **D** | 54 | 0.46 | 13 | 1 | 0 | 0 |
| **C** | 55 | 1.07 | 5.5 | 0 | 0 | 0 |
| **Q** | 85 | 0 | 10.5 | 0 | 0 | 0 |
| **E** | 83 | 0.47 | 12.3 | 1 | 0 | 0 |
| **G** | 3 | 0.07 | 9 | 0 | 0 | 0 |
| **H** | 96 | 0.61 | 10.4 | 1 | 1 | 0 |
| **I** | 111 | 2.22 | 5.2 | 0 | 0 | 1 |
| **L** | 111 | 1.53 | 4.9 | 0 | 0 | 1 |
| **K** | 119 | 1.15 | 11.3 | 1 | 0 | 0 |
| **M** | 105 | 1.18 | 5.7 | 0 | 0 | 0 |
| **F** | 132 | 2.02 | 5.2 | 0 | 1 | 0 |
| **P** | 32.5 | 1.95 | 8 | 0 | 0 | 0 |
| **S** | 32 | 0.05 | 9.2 | 0 | 0 | 0 |
| **T** | 61 | 0.05 | 8.6 | 0 | 0 | 0 |
| **W** | 170 | 2.65 | 5.4 | 0 | 1 | 0 |
| **Y** | 136 | 1.88 | 6.2 | 0 | 1 | 0 |
| **V** | 84 | 1.32 | 5.9 | 0 | 0 | 1 |

Στην συνέχεια, κατασκευάζουμε τα PSSMX, που όπως αναφέρθηκε αποτελούν συνεπτυγμένη μορφή των PSSM λαμβάνοντας υπόψη τις φυσικοχημικές ιδιότητες των

αμινοξέων. Το συγκεκριμένο χαρακτηριστικό έχει προταθεί και χρησιμοποιηθεί επιτυχώς για την πρόβλεψη ρεομορφικών περιοχών σε πρωτεϊνικές ακολουθίες [74, 83]. Στην προτεινόμενη μεθοδολογία, κάθε καταχώρηση $p_{ik}$ που αφορά την θέση $i$ και την ιδιότητα $k$ στον πίνακα PSSMX ορίζεται ως εξής:

$$p_{ik} = \sum_{j=1}^{20} AP_{kj} \times x_{ij} \,, \qquad (1)$$

όπου $AP_{kj}$ είναι η τιμή της κάθε φυσικοχημικής ιδιότητας και $x_{ij}$ είναι η τιμή του $j$ αμινοξέος στην θέση $i$ του PSSM. Έτσι προκύπτει ένα διάνυσμα εισόδου με μικρότερη διάσταση από τα PSSM που ενέχει και πληροφορία σχετική με τις φυσικοχημικές ιδιότητες του κάθε αμινοξέος.

Τα εξαχθέντα χαρακτηριστικά χρησιμοποιούνται για να σχηματίσουν τα δύο διανύσματα εισόδου, δηλαδή το FV-PSSM που αποτελείται από PSSM, δευτεροταγή δομή, έκθεση στο διαλύτη και φυσικοχημικές ιδιότητες, και το FV-PSSMX που αποτελείται από τα PSSMX, τη δευτεροταγή δομή και την έκθεση στο διαλύτη. Όλα τα παραπάνω χαρακτηριστικά εξάγονται κάθε φορά από ένα κυλιόμενο παράθυρο μήκους 11 αμινοξέων, δηλαδή λαμβάνοντας υπόψη τα ±5 γειτονικά αμινοξέα που περιβάλλουν στην πρωτοταγή αμινοξική ακολουθία έναν πεπτιδικό δεσμό. Συνεπώς, τα διανύσματα εισόδου που προκύπτουν FV-PSSM και FV-PSSMX, περιέχουν 331 και 111 χαρακτηριστικά, αντιστοίχως.

## *Επιλογή χαρακτηριστικών*

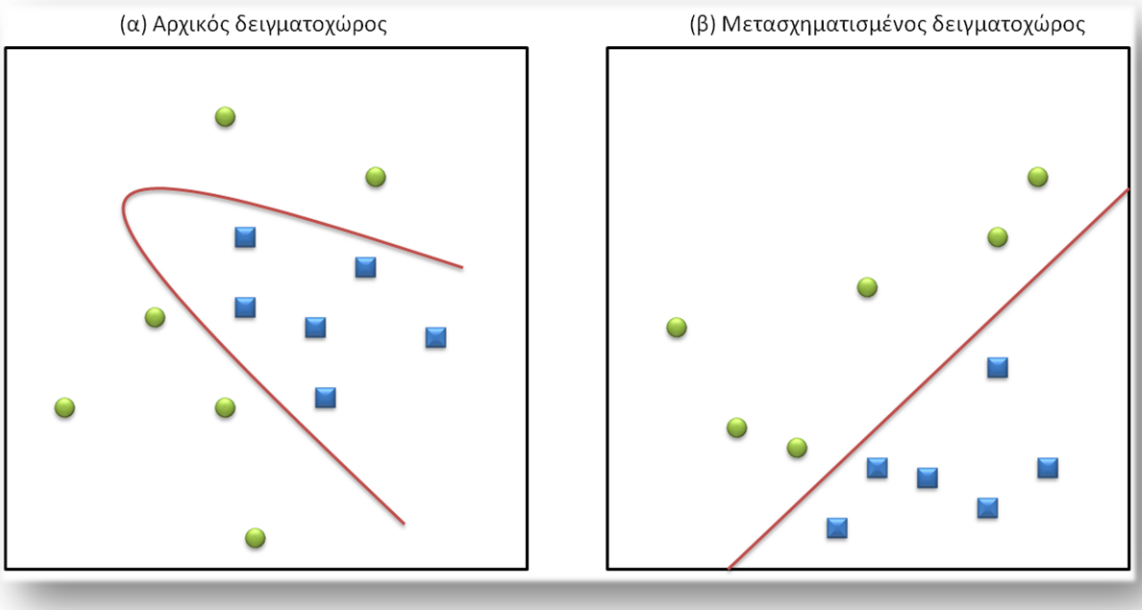
Μερικά από τα χαρακτηριστικά που περιγράφονται παραπάνω, ενδέχεται να είναι περιττά, άχρηστα ή συσχετισμένα μεταξύ τους, υποβιβάζοντας την απόδοση του ταξινομητή. Από θεωρητική άποψη, η βέλτιστη επιλογή χαρακτηριστικών περιλαμβάνει έναν εξαντλητικό έλεγχο όλων των πιθανών $2^{n-1}$ συνδυασμών, όπου n είναι το πλήθος των χαρακτηριστικών στο αρχικό διάνυσμα εισόδου. Στην βιβλιογραφία παρατίθενται δύο γενικές κατηγορίες αλγορίθμων επιλογής χαρακτηριστικών, οι filter και οι wrapper [21]. Οι filter αλγόριθμοι είναι ανεξάρτητοι από τον τελικό αλγόριθμο ταξινόμησης και βασίζονται στον υπολογισμό κάποιων στατιστικών μέτρων για να αξιολογήσουν την συνεισφορά είτε μεμονωμένων χαρακτηριστικών είτε υποσυνόλων χαρακτηριστικών. Οι αλγόριθμοι wrapper χρησιμοποιούν το τελικό αλγόριθμο ταξινόμησης ως "μαύρο κουτί"

για να εκτιμήσουν την διακριτική ικανότητα υποσυνόλων χαρακτηριστικών υπολογίζοντας ουσιαστικά την συνολική ακρίβεια της πρόβλεψης. Οι αλγόριθμοι wrapper συνήθως έχουν καλύτερα αποτελέσματα από τους filter αλγόριθμους μιας και είναι συντονισμένοι με τον αλγόριθμο ταξινόμησης [84]. Στην παρούσα εργασία χρησιμοποιούμε έναν wrapper αλγόριθμο που πραγματοποιεί την αναζήτηση των υποσυνόλων χαρακτηριστικών με τον BestFirtst αλγόριθμο [21] και έχει ως τελικό ταξινομητή ένα SVM.

### *Κατηγοριοποίηση πεπτιδικών δεσμών*

Στο επόμενο στάδιο της προτεινόμενης μεθοδολογίας, χρησιμοποιούμε ένα SVM για να προβλέψουμε την πιθανή διαμόρφωση του πεπτιδικού δεσμού, σε μια από τις τέσσερις κλάσεις, δηλαδή *cis*-Pro, *cis*-nonPro, *trans*-Pro και *trans*-nonPro. Τα SVM έχουν χρησιμοποιηθεί σε ένα μεγάλο εύρος επιστημονικών πεδίων, συμπεριλαμβανομένης και την Βιοπληροφορικής, δίδοντας ιδιαίτερα καλά αποτελέσματα [85].

Έστω ότι κάθε δείγμα του συνόλου εκπαίδευσης αποτελείται από ζεύγη της μορφής $(x_i,y_i)$, $i=1,2,...,l$ όπου $x_i$ είναι το διάνυσμα εισόδου και $y_i$ η κλάση στην οποία ανήκει το δείγμα εισόδου. Αρχικά ο ταξινομητής SVM μετασχηματίζει το διάνυσμα εισόδου σε έναν άλλο δειγματοχώρο μεγαλύτερης διάστασης με τη χρήση μιας συνάρτησης πυρήνα, όπου τα δείγματα εισόδου είναι γραμμικώς ή εν γένει πιο σαφώς διαχωρισμένα (Εικόνα 20).




**Εικόνα 20:** (α) Αρχικός και (β) μετασχηματισμένος δειγματοχώρος.

Στον μετασχηματισμένο πλέον δειγματοχώρο το SVM αναζητά το υπερεπίπεδο που διαχωρίζει τα δείγματα εισόδου με το μικρότερο δυνατό σφάλμα. Η συνάρτηση απόφασης με βάση την οποία ταξινομούνται νέα, άγνωστα δείγματα ορίζεται ως εξής:

$$f(x) = sign(\sum_{i=1}^{l} a_i y_i K(x_i, y_i) + b), \quad (2)$$

όπου $a_i$ είναι βάρη, $b$ είναι μια σταθερά (bias term) και $K(x_i,y_i)$ είναι η συνάρτηση πυρήνα. Οι προαναφερθείσες παράμετροι καθορίζονται μεγιστοποιώντας την συνάρτηση 3:

$$\sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j K(x_i, y_j), \quad (3)$$

όπου:

$$\sum_{i=1}^{l} a_i y_i = 0 \quad \text{και} \quad 0 \le a_i \le C\,. \quad (4)$$

Η μεταβλητή $C$ αποτελεί παράγοντα που ουσιαστικά ελέγχει τη σχέση του σφάλματος ταξινόμησης και του περιθωρίου ανάμεσα από το υπερεπίπεδο απόφασης και τα πιο ακραία δείγματα από την κάθε κλάση. Στην προτεινόμενη μεθοδολογία χρησιμοποιούμε την ακόλουθη πολυωνυμική συνάρτηση πυρήνα:

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^d, \quad (5)$$

όπου $d$ είναι ο βαθμός του πολυωνύμου.

Τα SVM παρουσιάζουν καλές επιδόσεις σε πολυδιάστατα δεδομένα εισόδου. Επίσης, παρόλο που έχουν σχεδιαστεί για να αντιμετωπίζουν δυαδικά προβλήματα, δηλαδή προβλήματα δύο κλάσεων, μπορούν εύκολα να επεκταθούν και για προβλήματα πολλαπλών κλάσεων. Συγκεκριμένα, αποσυνθέτουν το πρόβλημα διάκρισης μεταξύ πολλαπλών κλάσεων σε δυαδικά προβλήματα της μορφής ένας-εναντίον-όλων. Έστω $Y = \{y_1, y_2, ..., y_k\}$ οι κλάσεις στις οποίες κατανέμονται τα δεδομένα εισόδου, για κάθε κλάση $y_i \in Y$, δημιουργείται ένα ξεχωριστό δυαδικό πρόβλημα ταξινόμησης όπου όλα τα δείγματα εισόδου που ανήκουν στην κλάση $y_i$ θεωρούνται θετικά, και όλα τα υπόλοιπα αρνητικά. Κατασκευάζουμε, κατ' αυτόν τον τρόπο έναν δυαδικό ταξινομητή για την κλάση $y_i$ που διακρίνει τα δείγματα της κλάσης $y_i$ από όλα τα υπόλοιπα δείγματα. Όταν λοιπόν, θέλουμε να ταξινομήσουμε ένα νέο δείγμα, όλοι οι δυαδικοί ταξινομητές αναθέτουν την κλάση που προέβλεψαν και πλειοψηφικά ταξινομείται το άγνωστο δείγμα [21].

## *2.4 Αποτελέσματα*

Για να αξιολογήσουμε την προτεινόμενη μεθοδολογία, χρησιμοποιήσαμε τα εξής στατιστικά μέτρα: ευαισθησία (*Se*: sensitivity), θετική προγνωστική αξία (*PPV*: positive predictive value) και ακρίβεια (*Acc*: accuracy). Η ευαισθησία εκφράζει τα θετικά δείγματα που ταξινομήθηκαν σωστά, η θετική προγνωστική αξία εκφράζει τα αρνητικά δείγματα που ταξινομήθηκαν σωστά και η ακρίβεια δείχνει το ποσοστό των δειγμάτων που ταξινομήθηκαν συνολικά σωστά από τον ταξινομητή. Οι σχέσεις για τον υπολογισμό των παραπάνω μέτρων δίνονται παρακάτω:

$$Se = TP/(TP+FN), \quad (6)$$

$$PPV = TP/(TP+FP), \quad (7)$$

$$Acc = (TP+TN)/(TP+TN+FP+FN), \quad (8)$$

όπου οι όροι *TP*, *TN*, *FP*, *FN* εξηγούνται αμέσως παρακάτω στον Πίνακας 3.

**Πίνακας 3:** Βοηθητικός πίνακας για την εξαγωγή των στατιστικών μέτρων αξιολόγησης.

| | | **Πραγματική κλάση** | |
|---|---|---|---|
| | | Θετικό | Αρνητικό |
| **Προβλεπόμενη κλάση** | Θετικό | True Positive (*TP*) | False Positive (*FP*) |
| | Αρνητικό | False Negative (*FN*) | True Negative (*TN*) |

Επίσης, κατά την αξιολόγηση του ταξινομητή χρησιμοποιούμε και την καμπύλη ROC (Receiver Operating Characteristic) όπου κάθε σημείο επί της καμπύλης δίνει την ευαισθησία και την θετική προγνωστική αξία του ταξινομητή. Το εμβαδόν της περιοχής κάτω από την καμπύλη ROC (AUC: Area Under Curve) θεωρείται σημαντικός δείκτης για την αξιολόγηση της ακρίβειας και επίδοσης μια μεθόδου.

Ο προτεινόμενος αλγόριθμος ταξινόμησης συγκρίθηκε σε ένα αντιπροσωπευτικό δείγμα του συνόλου δεδομένων με τρείς ταξινομητές: ένα SVM με πυρήνα ακτινικής βάσης (RBF: Radial Basis Function) με παραμέτρους $c$=1.0 και $\gamma$=0.01, ένα δέντρο

απόφασης και ένα τεχνητό νευρωνικό δίκτυο. Στους δύο πρώτους ταξινομητές για επιλογή χαρακτηριστικών χρησιμοποιήσαμε έναν wrapper αλγόριθμο ενώ στα τεχνητά νευρωνικά δίκτυα η μείωση της διάστασης του διανύσματος εισόδου πραγματοποιήθηκε με την μέθοδο ανάλυσης κύριων συνιστωσών (PCA: Principal Component Analysis) όπου εξαιρέθηκαν χαρακτηριστικά τα οποία συμβάλλουν σε ποσοστό λιγότερο από 5% της συνολικής διακύμανσης [21]. Συγκριτικά αποτελέσματα μεταξύ των τεσσάρων ταξινομητών παρατίθενται στον Πίνακας 4.

**Πίνακας 4:** Συγκριτικά αποτελέσματα 4 αλγορίθμων ταξινόμησης.

| | **Wrapper + SVM (πολυωνυμικός πυρήνας)** | | **Wrapper + SVM (RBF πυρήνας)** | | **PCA+ANN** | | **Wrapper + Δέντρο απόφασης** | |
|---|---|---|---|---|---|---|---|---|
| | ***Se(%)*** | ***PPV(%)*** | ***Se(%)*** | ***PPV(%)*** | ***Se(%)*** | ***PPV(%)*** | ***Se(%)*** | ***PPV(%)*** |
| ***cis*-Pro** | 77 | 70 | 62 | 59 | 60 | 60 | 56 | 55 |
| ***cis*-nonPro** | 75 | 71 | 77 | 59 | 57 | 57 | 61 | 56 |
| ***cis*** | 76 | 71 | 70 | 59 | 59 | 59 | 59 | 56 |
| ***trans*-Pro** | 66 | 74 | 57 | 60 | 61 | 60 | 55 | 56 |
| ***trans*-nonPro** | 69 | 74 | 46 | 67 | 57 | 57 | 53 | 57 |
| ***trans*** | 68 | 74 | 52 | 64 | 59 | 59 | 54 | 57 |
| **Accuracy** | **72** | | **60** | | **58** | | **56** | |

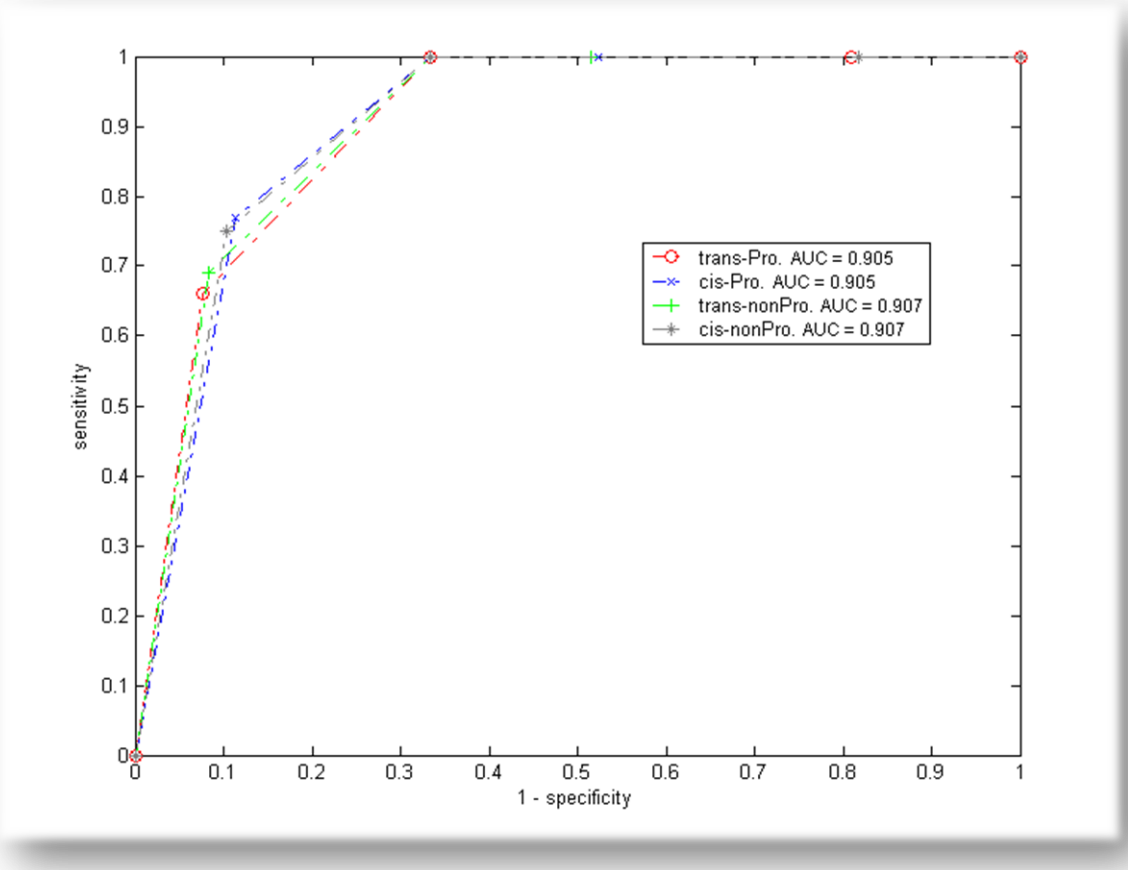
Επιπρόσθετα, μελετούμε την επίδραση δύο διανυσμάτων εισόδου (FV-PSSM και FV-PSSMX) ως προς την διάκριση των διαμορφώσεων του πεπτιδικού δεσμού. Συγκεκριμένα, το διάνυσμα FV-PSSM αποτελείται από τα PSSM, τη δευτεροταγή δομή, την έκθεση στο διαλύτη και τις φυσικοχημικές ιδιότητες, ενώ το FV-PSSMX περιλαμβάνει τα PSSMX, τη δευτεροταγή δομή και την έκθεση στο διαλύτη. Πιο κάτω περιγράφεται η διαδικασία που αποτελείται από δύο προσεγγίσεις και ακολουθήθηκε για την όσο το δυνατόν εγκυρότερη και πληρέστερη αξιολόγηση των διανυσμάτων εισόδου αλλά και της προτεινόμενης μεθοδολογίας γενικά. Να υπενθυμίσουμε σε αυτό το σημείο ότι τα δεδομένα εισόδου δεν είναι ισοκατανεμημένα στις τέσσερεις κλάσεις αλλά υπάρχουν 318 *cis*-nonPro, 1416 *cis*-Pro, 30657 *trans*-Pro και 657968 *trans*-nonPro δεσμοί. Αρχικά, λαμβάνουμε τυχαία 200 δεσμούς χωρίς επανατοποθέτηση από κάθε κλάση (δηλαδή ~70% της μικρότερης κλάσης

*cis*-nonPro) και κατασκευάζουμε ένα σύνολο 800 δεσμών όπου υπάρχει ίσος αριθμός δειγμάτων από την κάθε κλάση. Επαναλαμβάνουμε αυτή τη διαδικασία τυχαία 10 φορές ώστε να προκύψουν 10 σύνολα που το καθένα έχει 800 δεσμούς εξίσου κατανεμημένους στις τέσσερεις κλάσεις. Κατά την πρώτη προσέγγιση αξιολόγησης εφαρμόζουμε σε κάθε ένα από τα 10 σύνολα την μέθοδο διασταυρούμενης επικύρωσης n ομάδων (n-fold cross-validation) με n=10. Κατά τη μέθοδο αυτή ένα σύνολο δεδομένων διαιρείται σε 10 ομάδες που καθεμία περιέχει 1/10 των αρχικών δεδομένων. Εν συνεχεία τα 9/10 των δεδομένων χρησιμοποιούνται για την εκπαίδευση του ταξινομητή και το υπολειπόμενο 1/10 για έλεγχο. Ακολούθως, οι ομάδες κυλίονται κάθε φορά κατά μία θέση ώστε τελικά όλες να έχουν χρησιμοποιηθεί ακριβώς μία φορά για έλεγχο. Εμείς πραγματοποιούμε 10-fold cross validation σε κάθε ένα από τα 10 ισοκατανεμημένα σύνολα και υπολογίζουμε τον μέσο όρο. Κατά την δεύτερη προσέγγιση, χρησιμοποιούμε κάθε ένα από 10 σύνολα που περιγράφηκαν παραπάνω για την εκπαίδευση του ταξινομητή και όλους τους υπόλοιπους δεσμούς κάθε φορά (δηλαδή 118 *cis*-nonPro, 1216 *cis*-Pro, 30457 *trans*-Pro και 657768 *trans*-nonPro) για έλεγχο. Έπειτα, υπολογίζουμε τον μέσο όρο των αποτελεσμάτων από κάθε σύνολο ελέγχου [21, 86]. Τα στατιστικά μέτρα που προκύπτουν από τις δύο προσεγγίσεις αξιολόγησης και για κάθε διάνυσμα εισόδου παρατίθενται στον Πίνακας 5. Μέσα στις παρενθέσεις και δίπλα από κάθε τιμή στατιστικού μέτρου υπάρχει η τυπική απόκλιση της τιμής αυτής.

**Πίνακας 5:** Συνοπτικά αποτελέσματα της προτεινόμενης μεθοδολογίας.

| | | FV-PSSM | | FV-PSSMX | |
|---|---|---|---|---|---|
| | | ***Se(%)*** | ***PPV(%)*** | ***Se(%)*** | ***PPV(%)*** |
| **Σύνολα ελέγχου** | ***cis*-Pro** | 64.63(6.54) | 60.90(1.73) | 68.92(4.92) | 59.80(1.48) |
| | ***cis*-nonPro** | 70.26(7.01) | 62.80(1.55) | 72.72(4.90) | 64.30(0.95) |
| | ***cis*** | 67.45(6.78) | 61.85(1.64) | 70.82(4.91) | 62.05(1.22) |
| | ***trans*-Pro** | 58.28(7.11) | 62.50(1.65) | 53.60(4.87) | 63.50(1.96) |
| | ***trans*-nonPro** | 58.32(4.44) | 66.60(4.14) | 59.55(2.46) | 68.90(3.31) |
| | ***trans*** | 58.30(5.76) | 64.55(2.90) | 56.58(3.67) | 66.20(2.64) |
| | **Accuracy** | **62.87(4.37)** | | **63.70(2.35)** | |
| | | | | | |
| **10-fold cross validation** | ***cis*-Pro** | 71.55(7.01) | 69.46(4.81) | 73.35(3.42) | 64.41(2.76) |
| | ***cis*-nonPro** | 77.40(4.34) | 68.08(2.64) | 76.35(2.67) | 66.07(2.36) |
| | ***cis*** | 74.45(5.68) | 68.77(3.73) | 74.85(3.05) | 65.24(2.56) |
| | ***trans*-Pro** | 67.75(8.99) | 70.71(3.42) | 59.20(5.98) | 68.97(1.96) |
| | ***trans*-nonPro** | 64.65(3.99) | 73.92(3.83) | 60.70(3.65) | 71.97(2.69) |
| | ***trans*** | 66.20(6.49) | 72.32(3.63) | 59.95(4.82) | 70.47(2.33) |
| | **Accuracy** | **70.23(2.11)** | | **67.4(1.69)** | |

Για το διάνυσμα εισόδου FV-PSSM κατασκευάζουμε την καμπύλη ROC (Εικόνα 21) για κάθε μία από τις τέσσερεις κλάσεις, όπου η κάθε καμπύλη υποδηλώνει την απόδοση της μεθόδου ως προς την πρόβλεψη μιας κατηγορίας έναντι όλων των υπολοίπων [87]. Επίσης, υπολογίζουμε το εμβαδόν AUC κάτω από κάθε καμπύλη, όπου σε κάθε περίπτωση είναι πάνω από 0.905.




**Εικόνα 21:** Καμπύλη ROC για τις τέσσερεις κλάσεις.

Τέλος, στον Πίνακας 6 παρατίθενται συγκριτικά αποτελέσματα ανάμεσα στην προτεινόμενη μεθοδολογία και τις μεθόδους που περιγράφονται για το αντίστοιχο πρόβλημα στην βιβλιογραφία (υποκεφάλαιο 1.2: Πεπτιδικός δεσμός). Εκτός από στατιστικά αποτελέσματα οι μέθοδοι συγκρίνονται με βάση το μέγεθος του αρχικού συνόλου δεδομένων, τον σκοπό της πρόβλεψης, των χαρακτηριστικών του διανύσματος εισόδου και της τεχνικής αξιολόγησης. Περαιτέρω τεχνικά χαρακτηριστικά αλλά και ποιοτική σύγκριση των μεθόδων παρατίθενται στην επόμενη ενότητα (υποκεφάλαιο 2.5: Συζήτηση-συμπεράσματα).

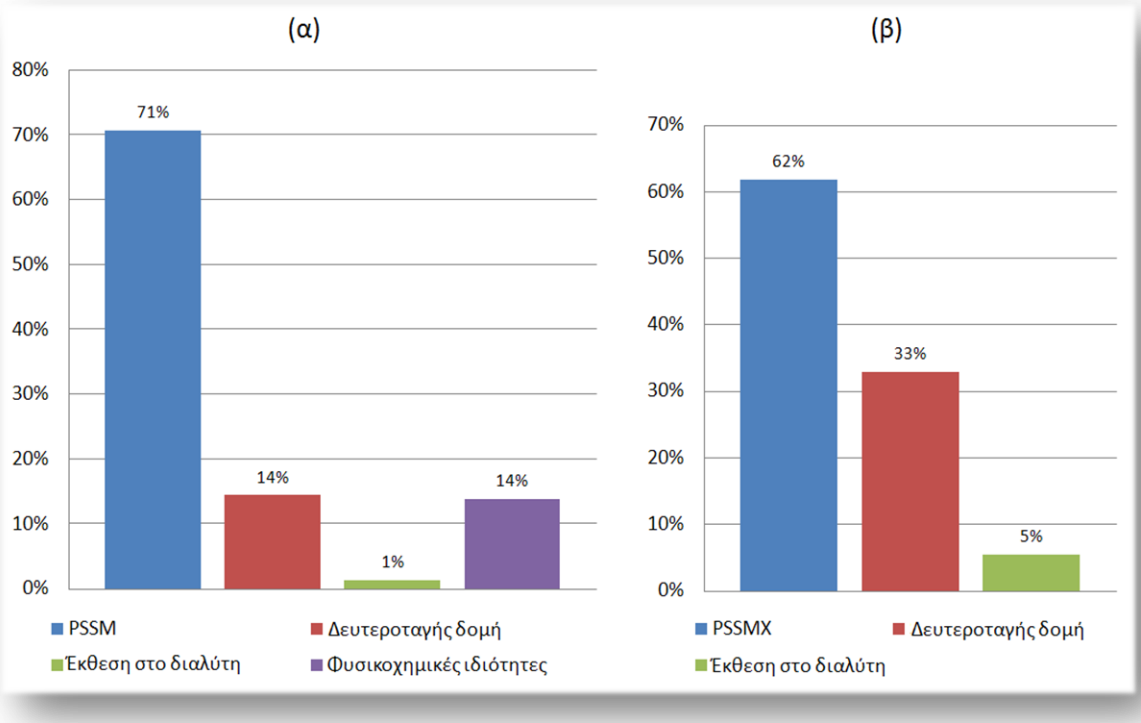
**Πίνακας 6:** Σύγκριση μεταξύ της προτεινόμενης μεθοδολογίας και αυτών στην βιβλιογραφία.

| Μεθοδολογία | Σύνολο δεδομένων | Σκοπός | Χαρακτηριστικά | Τεχνική αξιολόγησης | Se (%) | Acc (%) |
|---|---|---|---|---|---|---|
| Ακολουθιακά πρότυπα [19] | 242 | X-Pro | Φυσικοχημικές ιδιότητες | Ανεξάρτητο σύνολο ελέγχου | 73 | 86 |
| SVM [20] | 2193 | X-Pro | Ακολουθία αμινοξέων | cross validation | 77 | 77 |
| Παράμετροι Chou-Fasman [22] | 8584 | X-Pro & X-nonPro | Δευτεροταγής δομή | cross validation (10-fold) | 35 | 66 |
| SVM [24] | 2424 | X-Pro | PSSM, δευτεροταγής δομή | cross validation (5-fold) | 71 | 71 |
| **Προτεινόμενη μεθοδολογία** | 3050 | X-Pro & X-nonPro | PSSM, δευτεροταγής δομή, έκθεση στο διαλύτη, φυσικοχημικές ιδιότητες | cross validation (10-fold) | 75 | 70 |
| | | X-Pro | | | 72 | |
| | | **X-nonPro** | | | **77** | |

## *2.5 Συζήτηση-συμπεράσματα*

Η παρούσα μεθοδολογία αποτελείται από τρία στάδια και έχει αναπτυχθεί με σκοπό την πρόβλεψη της διαμόρφωσης του πεπτιδικού δεσμού μεταξύ δύο οποιονδήποτε αμινοξέων. Στο πρώτο στάδιο ένας μεγάλος αριθμός χαρακτηριστικών εξάγεται από την πρωτοταγή ακολουθία των αμινοξέων, τα οποία στο επόμενο βήμα αξιολογούνται από έναν αλγόριθμο επιλογής χαρακτηριστικών ώστε να προκύψει το υποσύνολο χαρακτηριστικών με τη μεγαλύτερη διακριτική ικανότητα. Το υποσύνολο αυτό στο επόμενο βήμα εισάγεται σε έναν ταξινομητή SVM με σκοπό να προβλέψει την διαμόρφωση του πεπτιδικού δεσμού για X-Pro και X-nonPro ζεύγη αμινοξέων. Συγκεκριμένα χρησιμοποιήθηκαν δύο διανύσματα εισόδου, FV-PSSM και FV-PSSMX, όπου έδωσαν συγκρίσιμα αποτελέσματα.

Το τελικό διάνυσμα εισόδου και στις δύο περιπτώσεις ήταν ιδιαιτέρως εκτενές και περιείχε έναν σημαντικό αριθμό χαρακτηριστικών είτε συσχετισμένων μεταξύ τους, είτε εν γένει μειωμένης διακριτικής ικανότητας ως προς τον τελικό σκοπό της ταξινόμησης, δυσχεραίνοντας κατ' αυτόν τον τρόπο την εκπαίδευση του ταξινομητή. Όπως έχει αναφερθεί, για την απαλοιφή των περιττών χαρακτηριστικών και την διατήρηση των πιο σημαντικών, χρησιμοποιήθηκε ο wrapper αλγόριθμος, ο οποίος συχνά υπερτερεί σε σχέση με άλλους αλγόριθμους επιλογής χαρακτηριστικών. Το πλήθος των χαρακτηριστικών που διατηρήθηκαν από τον wrapper αλγόριθμο και αποτέλεσαν το τελικό διάνυσμα εισόδου στον αλγόριθμο ταξινόμησης ποικίλλει από 8 έως 27, που αποτελούν και αυτά με την μεγαλύτερη διακριτική ικανότητα για το συγκεκριμένο πρόβλημα ταξινόμησης. Στη συνέχεια για κάθε διάνυσμα εισόδου, υπολογίζουμε την συνεισφορά κάθε χαρακτηριστικού στο τελικό διάνυσμα εισόδου, όπως διαμορφώθηκε από τον αλγόριθμο επιλογής χαρακτηριστικών. Συγκεκριμένα, για το διάνυσμα FV-PSSM υπολογίζουμε το ποσοστό του τελικού διανύσματος εισόδου ανήκουν σε PSSM, δευτεροταγή δομή, έκθεση στο διαλύτη ή φυσικοχημικές ιδιότητες. Τα αποτελέσματα προβάλλονται στην Εικόνα 22-α. Ομοίως υπολογίζουμε τα αντίστοιχα ποσοστά για το FV-PSSMX όπου η συνεισφορά των PSSMX, της δευτεροταγούς δομής και της έκθεσης στο διαλύτη φαίνονται στην Εικόνα 22-β.



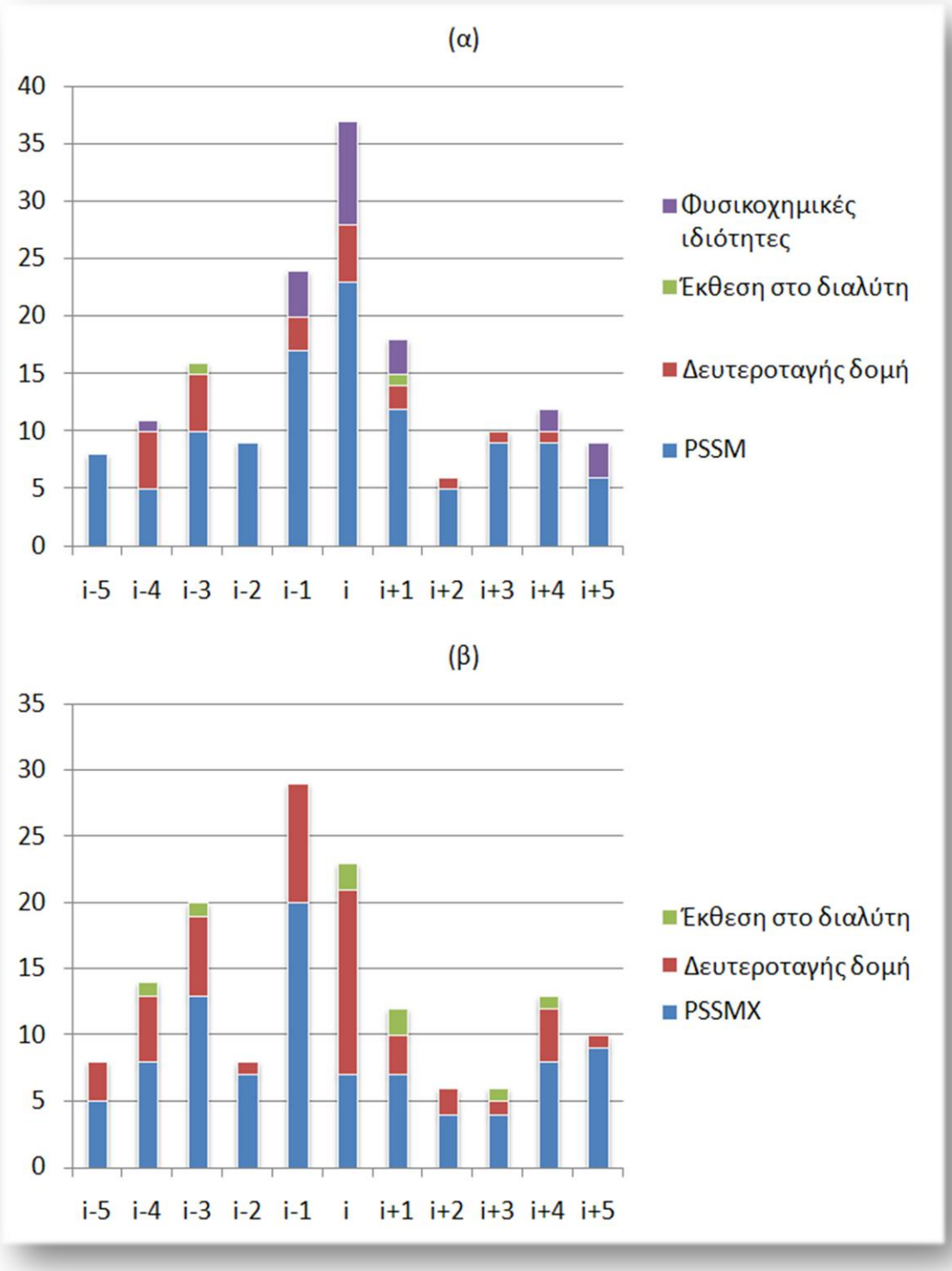

**Εικόνα 22:** Συνεισφορά κάθε χαρακτηριστικού στο τελικό διάνυσμα εισόδου (α) FV-PSSM, (β) FV-PSSMX.

Η ποσοστιαία συνεισφορά του κάθε χαρακτηριστικού στο τελικό διάνυσμα εισόδου, είναι ενδεικτική σε κάθε περίπτωση για την συμβολή και την επιρροή του χαρακτηριστικού στην διαμόρφωση του πεπτιδικού δεσμού. Από την Εικόνα 22-α είναι σαφές ότι τα PSSM παίζουν πρωταρχικό ρόλο στον ισομερισμό του πεπτιδικού δεσμού (X-Pro και X-nonPro), επιβεβαιώνοντας και ενισχύοντας έτσι το γεγονός ότι η γειτονική ακολουθία ασκεί επιρροή στην δομή του πεπτιδικού δεσμού. Μάλιστα, αξίζει να σημειωθεί ότι τα PSSM δεν αντικατοπτρίζουν την απλή ακολουθία των αμινοξέων αλλά εμπεριέχουν πληροφορία σχετική με την συχνότητα των αμινοξέων και την διατήρησή τους εξελικτικά στις πρωτεΐνες. Όσον αφορά την δευτεροταγή δομή αλλά και τις φυσικοχημικές ιδιότητες των αμινοξέων, βλέπουμε να επηρεάζουν εξίσου την διαμόρφωση του πεπτιδικού δεσμού, ως ένα βαθμό όχι καθοριστικό, αλλά σίγουρα όχι αμελητέο. Από την άλλη, η έκθεση στο διαλύτη φαίνεται να ασκεί ιδιαιτέρως περιορισμένη επιρροή στον ισομερισμό του πεπτιδικού δεσμού. Τα συμπεράσματα αυτά είναι σε συμφωνία με τη βιβλιογραφία [12] όπου έχει δειχτεί η σημαντική συσχέτιση του

*cis/trans* ισομερισμού με τη δευτεροταγή δομή, αλλά και η σαφώς μικρότερη επίδραση της έκθεσης στο διαλύτη.

Ελαφρώς διαφοροποιημένη είναι η κατάσταση όσον αφορά το διάνυσμα εισόδου FV-PSSMX (Εικόνα 22-β), όπου η δευτεροταγής δομής επηρεάζει σε μεγαλύτερο βαθμό την διαμόρφωση του πεπτιδικού δεσμού από ότι στο διάνυσμα FV-PSSM. Όπως και πριν, η έκθεση στο διαλύτη φαίνεται να ασκεί σχεδόν αμελητέα επίδραση στον πεπτιδικό δεσμό, ενώ τα PSSMX επηρεάζουν σημαντικά τον ισομερισμό του πεπτιδικού δεσμού. Παρ' όλα αυτά, τα PSSMX λόγω της πληροφορίας που ενέχουν σχετικά με τις φυσικοχημικές ιδιότητες των αμινοξέων, επιπρόσθετα από τα PSSM, θα αναμέναμε να συμμετέχουν πιο καθοριστικά στο τελικό διάνυσμα εισόδου.

Επίσης, μελετήσαμε και την επίδραση του κάθε ενός από τα ±5 γειτονικά αμινοξέα στην διαμόρφωση του πεπτιδικού δεσμού. Ομοίως με πριν, υπολογίσαμε το ποσοστό με το οποίο εκπροσωπείται το κάθε κατάλοιπο στο τελικό διάνυσμα εισόδου (FV-PSSM και FV-PSSMX) μετά την εφαρμογή του αλγόριθμου επιλογής χαρακτηριστικών. Τα αποτελέσματα παρατίθενται στην Εικόνα 23.




**Εικόνα 23:** Συνεισφορά γειτονικών αμινοξέων στην διαμόρφωση του πεπτιδικού δεσμού για τα διανύσματα (α) FV-PSSM και (β) FV-PSSMX.

Παρατηρούμε συγκεκριμένα για το FV-PSSM (Εικόνα 23-α), ότι εκτός από το αμινοξύ i που αποτελεί το δεύτερο συστατικό του πεπτιδικού δεσμού, τα αμέσως γειτνιάζοντα

αμινοξέα i±1, διατηρούνται σε μεγάλο βαθμό στο τελικό διάνυσμα εισόδου και κατ' επέκταση ασκούν σημαντική επίδραση στην τελική διαμόρφωση του πεπτιδικού δεσμού. Αξίζει να σημειωθεί πως η συνεισφορά των αμινοξέων i±2 είναι κατάτι μικρότερη από την συνεισφορά των αμινοξέων i±3 και i±4 που βρίσκονται πιο μακριά στην πολυπεπτιδική αλυσίδα. Εκτός από αυτή την ελαφρώς παράδοξη και ενδιαφέρουσα παρατήρηση, η συνεισφορά των γειτονικών αμινοξέων εκατέρωθεν του πεπτιδικού δεσμού φθίνει καθώς απομακρυνόμαστε προς τις δύο κατευθύνσεις της πολυπεπτιδικής αλυσίδας. Όσον αφορά το διάνυσμα εισόδου FV-PSSMX, η συνεισφορά του κάθε αμινοξέος αντικατοπτρίζεται στην Εικόνα 23-β. Παρατηρούμε και πάλι την καθοριστική επιρροή των αμινοξέων i, i±1 στην διαμόρφωση του πεπτιδικού δεσμού, με επαυξημένο τον ρόλο του αμινοξέος i-1, σε σχέση με το διάνυσμα FV-PSSM. Όπως και πριν, παρατηρούμε πάλι ότι παραδόξως η επίδραση των αμινοξέων i±2 ως προς τον ισομερισμό του πεπτιδικού δεσμού είναι μικρότερη από τα πιο απομακρυσμένα i-3 και i±4.

Το γεγονός ότι η προτεινόμενη μεθοδολογία δεν έχει αναπτυχθεί και αξιολογηθεί στο ίδια σύνολα δεδομένων με τις υπόλοιπες μεθοδολογίες που παρουσιάζονται στην βιβλιογραφία, δεν επιτρέπει την ασφαλή εξαγωγή συμπερασμάτων κατά την μεταξύ τους σύγκριση. Παρ' όλα αυτά, στον Πίνακας 6 παραθέτουμε μια ποσοτική αλλά και ποιοτική σύγκριση των μεθόδων αυτών, σε σχέση με την προτεινόμενη. Η πρώτη μεθοδολογία που παρουσιάστηκε [19], χρησιμοποιεί ένα σύνολο 242 X-Pro πεπτιδικών δεσμών για να εξάγει ακολουθιακά πρότυπα που αφορούν είτε μεμονωμένα αμινοξέα είτε φυσικοχημικές ιδιότητες αυτών, και ακολούθως χρησιμοποιεί τα εξαχθέντα πρότυπα για να διακρίνει τις *cis* από τις *trans* διαμορφώσεις, αλλά μόνον για τους X-Pro πεπτιδικούς δεσμούς. Τα πρότυπα αυτά λαμβάνονται με βάση μια περιοχή ±6 γειτονικών αμινοξέων, αφού πρώτα δοκιμάστηκε ανεπιτυχώς η εξαγωγή τους από ένα ιδιαίτερα περιορισμένο παράθυρο τριών καταλοίπων (i, i±1). Παρά τα ενθαρρυντικά αποτελέσματα, το σχετικά μικρό σύνολο δεδομένων (242 X-Pro πεπτιδικοί δεσμοί) αποδυναμώνει την αξιοπιστία των εξαχθέντων προτύπων, τα οποία όταν δοκιμάστηκαν αργότερα σε εκτενέστερα σύνολα δεδομένων δεν έδειξαν ικανοποιητικά αποτελέσματα. Σε μια άλλη προσέγγιση [20], τα αμινοξέα της πρωτεϊνικής ακολουθίας κωδικοποιήθηκαν με δυαδικά 20-διάστατα διανύσματα τα οποία αποτέλεσαν την είσοδο σε έναν SVM ταξινομητή, ο οποίος διέκρινε μεταξύ *cis*-Pro και *trans*-Pro διαμορφώσεων. Και αυτή η μεθοδολογία εστίασε μόνο στους X-Pro πεπτιδικούς δεσμούς δίνοντας ακρίβεια πρόβλεψης, 70% και 77%, όταν αξιολογήθηκε σε ανεξάρτητο σύνολο ελέγχου και με την μέθοδο jackknife-test, αντιστοίχως. Η μέθοδος jackknife-test

αποτελεί ειδική περίπτωση της μεθόδου αξιολόγησης n-fold cross validation, όπου το n είναι ίσο με τον συνολικό αριθμό δειγμάτων στο σύνολο δεδομένων. Κατ' αυτόν τον τρόπο, αν ένα σύνολο δεδομένων αποτελείται από k δείγματα, χρησιμοποιούμε τα k-1 για εκπαίδευση και το εναπομένον δείγμα για έλεγχο, επαναλαμβάνοντας την ίδια διαδικασία k φορές μέχρις ότου όλα τα δείγματα έχουν χρησιμοποιηθεί ακριβώς μια φορά για έλεγχο. Η συγκεκριμένη μέθοδος αξιολόγησης, παρόλο που χρησιμοποιεί στο έπακρο τα δείγματα του συνόλου δεδομένων, ενδέχεται να οδηγήσει μερικές φορές σε υπερεκτίμηση της ακρίβειας της μεθόδου [86]. Η επόμενη μεθοδολογία που παρουσιάζεται στη βιβλιογραφία αφορά στον αλγόριθμο COPS [22] όπου λαμβάνοντας υπόψη μόνο την δευτεροταγή δομή από τριπλέτες αμινοξέων προβλέπει την διαμόρφωση του πεπτιδικού δεσμού για κάθε ζεύγος αμινοξέων. Παρόλο που αυτή η μεθοδολογία είναι και η μόνη που προβλέπει την διαμόρφωση του δεσμού μεταξύ δυο οποιονδήποτε αμινοξέων, το περιορισμένο πλήθος χαρακτηριστικών που χρησιμοποίησε, αλλά και το μικρό μέγεθος παραθύρου που έλαβε υπόψη οδήγησε σε χαμηλή συνολική ακρίβεια. Η πιο πρόσφατη μεθοδολογία συνιστά τον αλγόριθμο CISPEPpred [24], που επίσης εστιάζει αποκλειστικά σε X-Pro πεπτιδικούς δεσμούς, και χρησιμοποιεί ως είσοδο PSSM και δευτεροταγή δομή από ένα παράθυρο ±5 γειτονικών αμινοξέων. Τα χαρακτηριστικά αυτά εισάγονται σε έναν ταξινομητή SVM με πυρήνα RBF και πετυχαίνει ακρίβεια πρόβλεψης για τις δύο κλάσεις (*cis*-Pro και *trans*-nonPro) 71%, κατά την αξιολόγηση με 5-fold cross validation.

Επιπρόσθετα, η προτεινόμενη μεθοδολογία συγκρίθηκε και με τρεις ακόμη αλγορίθμους ταξινόμησης συνδυασμένους με έναν αλγόριθμο επιλογής χαρακτηριστικών. Συγκεκριμένα, αξιολογήθηκε η απόδοση ενός wrapper αλγορίθμου και ενός SVM με RBF πυρήνα, της μεθόδου PCA με ένα ANN, καθώς και ενός wrapper με δέντρο απόφασης. Λεπτομερή αποτελέσματα παρατίθενται στον Πίνακας 4. Το SVM με τον πυρήνα RBF πέτυχε συνολική ακρίβεια πρόβλεψης μειωμένη κατά 12% σε σχέση με την προτεινόμενη μεθοδολογία, ενώ εξίσου υποδεέστερη είναι και η απόδοσή του όσον αφορά τα υπόλοιπα στατιστικά μέτρα που υπολογίστηκαν. Στην περίπτωση του ANN και του δέντρου απόφασης, τα στατιστικά μέτρα αξιολόγησης είναι ακόμη πιο μειωμένα, αναδεικνύοντας την υπεροχή της προτεινόμενης μεθοδολογίας ως προς την πρόβλεψη της διαμόρφωσης του πεπτιδικού δεσμού.

Συμπερασματικά, παρουσιάσαμε μια μεθοδολογία που αφορά στην πρόβλεψη της διαμόρφωσης του πεπτιδικού δεσμού, μεταξύ οποιονδήποτε αμινοξέων. Συγκεκριμένα, ομαδοποιήσαμε τους πεπτιδικούς σε τέσσερις κλάσεις (*cis*-Pro, *cis*-nonPro, *trans*-Pro και

*trans*-nonPro) ώστε να ανιχνεύσουμε και τους ιδιαίτερα σημαντικούς από βιολογικής απόψεως *cis*-nonPro δεσμούς. Κατά την προσπάθεια διάκρισης μεταξύ των κατηγοριών του πεπτιδικού δεσμού, δοκιμάσαμε και αξιολογήσαμε δύο διανύσματα εισόδου, τα οποία λαμβάνουν υπόψη ένα μεγάλο αριθμό χαρακτηριστικών που μπορούν να εξαχθούν από την πρωτοταγή ακολουθία των αμινοξέων. Επιπλέον, η προτεινόμενη μεθοδολογία χρησιμοποιεί έναν αποδοτικό και αποτελεσματικό αλγόριθμο επιλογής χαρακτηριστικών με σκοπό τον εντοπισμό των χαρακτηριστικών με την μεγαλύτερη διακριτική ικανότητα και άρα των πιο καθοριστικών ως προς την διαμόρφωση του πεπτιδικού δεσμού. Ακόμη, μελετήθηκε η προβλεπτική ικανότητα του κάθε γειτονικού αμινοξέος ως προς την επίδρασή του στον ισομερισμό του πεπτιδικού δεσμού. Μια πιθανή μελλοντική μελέτη θα μπορούσε να επικεντρωθεί στην πρόβλεψη της διαμόρφωσης του πεπτιδικού δεσμού λαμβάνοντας υπόψη την επίδραση που ασκούν αμινοξέα που βρίσκονται κοντά στο χώρο παρά στην πρωτοταγή πρωτεϊνική ακολουθία.

# 3ο ΚΕΦΑΛΑΙΟ: Εξαγωγή προτύπων στη γειτονιά *cis* πεπτιδικών δεσμών

## *3.1 Σκοπός*

Στο κεφάλαιο αυτό θα παρουσιάσουμε μια μεθοδολογία για την συστηματική ανάλυση πρωτεϊνικών περιοχών που περιέχουν *cis* πεπτιδικούς δεσμούς. Όπως αναφέρθηκε και στην εισαγωγή της διατριβής, παρόλο που έχουν προταθεί διάφορες μεθοδολογίες για την πρόβλεψη της διαμόρφωσης του πεπτιδικού δεσμού, οι μεθοδολογίες αυτές ως επί το πλείστον εκπαιδεύουν έναν αλγόριθμο μηχανικής μάθησης ο οποίος με βάση ένα σύνολο χαρακτηριστικών εισόδου, διακρίνει *cis* και *trans* ισομερή. Κατ' αυτόν τον τρόπο λαμβάνεται η προβλεπόμενη έξοδος, η οποία έχει συνήθως αδιαφανώς προκύψει, οπότε και δεν εμπλουτίζεται η υπάρχουσα βιολογική γνώση γύρω από τον *cis/trans* ισομερισμό. Επίσης όπως έχει αναφερθεί και παραπάνω, η πλειοψηφία των μεθόδων της βιβλιογραφίας, επικεντρώνονται αποκλειστικά στους X-Pro πεπτιδικούς δεσμούς και αγνοούν τον ιδιαιτέρως σημαντικό ισομερισμό των X-nonPro πεπτιδικών δεσμών.

Αντίθετα, στην παρούσα μεθοδολογία, προτείνουμε ένα σύνολο βημάτων για την εξαγωγή ακολουθιακών προτύπων με φυσική και βιολογική σημασία από περιοχές με *cis*-Pro και *cis*-nonPro πεπτιδικούς δεσμούς. Ο σκοπός μας είναι ο εντοπισμός μη-τυχαίων προτύπων στη γειτονιά ενός πεπτιδικού δεσμού, που να καταδεικνύουν την διαμόρφωσή του. Τα εξαχθέντα πρότυπα, αφορούν εκτός από μεμονωμένα αμινοξέα και διάφορες ομαδοποιήσεις αμινοξέων με βάση τις φυσικοχημικές και δομικές τους ιδιότητες. Η συστηματική μελέτη και ανάλυση αυτών των προτύπων μπορεί να αποκαλύψει τους παράγοντες που ευνοούν (και αντίστοιχα αποτρέπουν) την μία ή την άλλη διαμόρφωση. Ακολούθως, τα εξαχθέντα πρότυπα αντιπαραβάλλονται συστηματικά έναντι βιολογικών βάσεων που περιλαμβάνουν ζεύγη προτύπων και των αντίστοιχων λειτουργιών που αυτά επιτελούν, όπως έχει διαπιστωθεί πειραματικά στην βιβλιογραφία. Επομένως, αναδεικνύονται οι πιθανές λειτουργικές προτιμήσεις των *cis* προτύπων και κατ' επέκταση *cis* πεπτιδικών δεσμών, γεγονός ιδιαιτέρως σημαντικό για τους *cis*-nonPro πεπτιδικούς

δεσμούς που η σπανιότητά τους δυσχεραίνει τη συστηματική επισκόπηση και ανάλυσή τους.

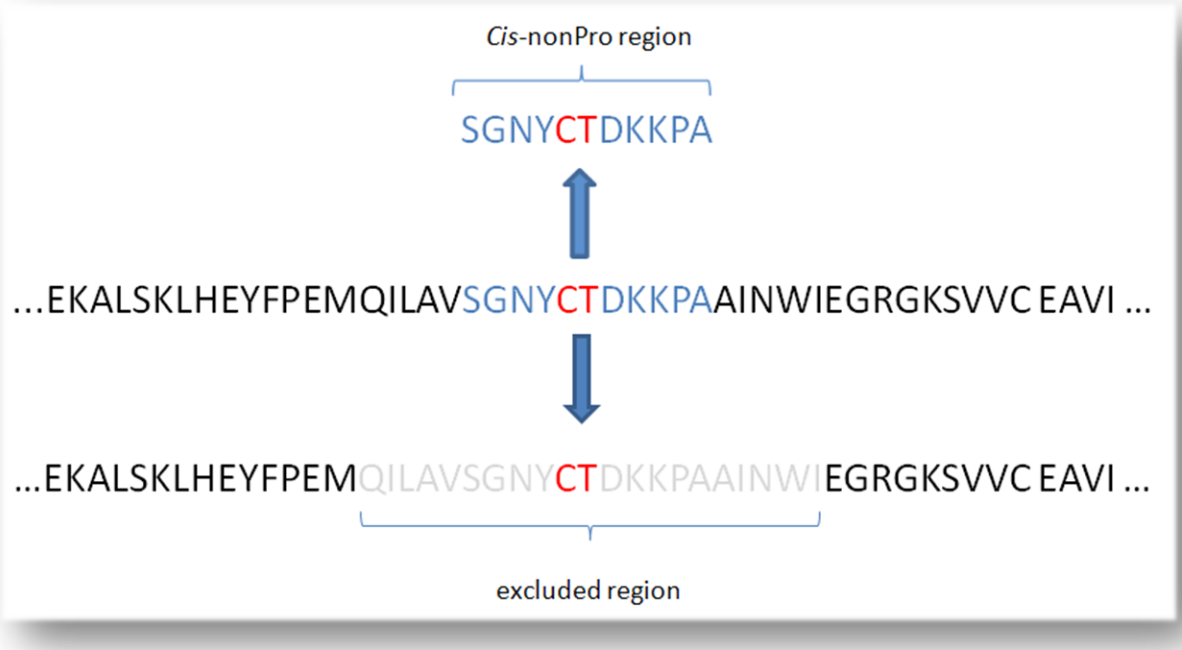
Συνολικά, η προτεινόμενη μεθοδολογία συνιστά μια μάλλον γενική και ευρεία αλληλουχία βημάτων με διπλή συνεισφορά. Πρώτον, αποτελεί μια ενορχηστρωμένη προσέγγιση για την εξαγωγή ακολουθιακών προτύπων από περιοχές ενδιαφέροντος σε πρωτεϊνικές ακολουθίες. Τα πρότυπα που προκύπτουν προσφέρουν μια εύληπτη και ντετερμινιστική αναπαράσταση των υπό μελέτη περιοχών, που λόγω της βιολογικής σημασίας που ενέχουν μπορούν να αποκαλύψουν σημαντικές πληροφορίες με φυσική σημασία. Δεύτερον, ιδίως όσον αφορά φαινόμενα που απαντώνται σπάνια στις πρωτεΐνες, τα ακολουθιακά πρότυπα βοηθούν στην γενίκευση και συστηματική καταγραφή της υπάρχουσας γνώσης, Μάλιστα, η λειτουργική ανάλυση των ακολουθιακών προτύπων μπορεί να αποκαλύψει πιθανές λειτουργικές συσχετίσεις στην περίπτωση άγνωστων ή εν γένει ανεξερεύνητων πρωτεϊνικών ακολουθιών.

## *3.2 Δεδομένα*

Το σύνολο δεδομένων που χρησιμοποιήθηκε κατά την ανάπτυξη της παρούσας μεθοδολογίας, έχει προέλθει από την PDB [75]. Συγκεκριμένα, 3050 πρωτεϊνικές δομές έχουν επιλεχθεί σύμφωνα με τα ακόλουθα κριτήρια: ανάλυση δομής με χρήση κρυσταλλογραφίας ακτινών Χ σε ανάλυση μεγαλύτερη των 2Å, η μέγιστη ομοιότητα μεταξύ των ακολουθιών δεν ξεπερνά το 25% και η παράμετρος ακρίβειας R-factor είναι μικρότερη από 0.25. Στη συνέχεια χρησιμοποιήσαμε το πρόγραμμα VADAR [77] ώστε να υπολογίσουμε την δίεδρη γωνία ω μεταξύ των διαδοχικών αμινοξέων, βάσει των τρισδιάστατων συντεταγμένων που περιέχει η PDB για κάθε άτομο της πρωτεΐνης. Οι πεπτιδικοί δεσμοί με δίεδρες γωνίες εντός του εύρους ±30º καταχωρήθηκαν ως *cis*, ενώ δεσμοί με γωνίες στο διάστημα [150º, 210º] καταχωρήθηκαν ως *trans*. Να σημειωθεί ότι το ίδιο ισχύει και για τους X-Pro και για τους X-nonPro πεπτιδικούς δεσμούς.

Η κατανομή των πεπτιδικών δεσμών στις παραπάνω πρωτεϊνικές ακολουθίες είναι ως εξής: 1417 *cis*-Pro, 30668 *trans*-Pro, 318 *cis*-nonPro και 685716 *trans*-nonPro. Για κάθε πεπτιδικό δεσμό κατασκευάζουμε ένα παράθυρο μεγέθους 11 αμινοξέων που λαμβάνει υπόψη τα ±5 γειτονικά αμινοξέα που επηρεάζουν την διαμόρφωση του δεσμού [24, 88]. Ειδικά για την περίπτωση των *trans* πεπτιδικών δεσμών που βρίσκονται εκατέρωθεν *cis*-

nonPro πεπτιδικών δεσμών, δεν κατασκευάζονται παράθυρα και εξαιρούνται από την μελέτη, γιατί εμφανίζουν μεγάλη επικάλυψη με τα αντίστοιχα παράθυρα των *cis*-nonPro πεπτιδικών δεσμών και θα οδηγούσαν πιθανά σε διαστρέβλωση των εξαχθέντων προτύπων (Εικόνα 24). Προκύπτουν, λοιπόν, τέσσερα σύνολα, $D_{cis\text{-Pro}}$, $D_{trans\text{-Pro}}$, $D_{cis\text{-nonPro}}$ και $D_{trans\text{-nonPro}}$, που το κάθε ένα περιέχει παράθυρα μήκους 11 αμινοξέων που περιβάλλουν τον κάθε τύπο πεπτιδικού δεσμού. Τα πέντε πρώτα και πέντε τελευταία αμινοξέα της κάθε πολυπεπτιδικής αλυσίδας εξαιρούνται από την μελέτη γιατί δεν διαθέτουν αρκετό αριθμό γειτονικών καταλοίπων.

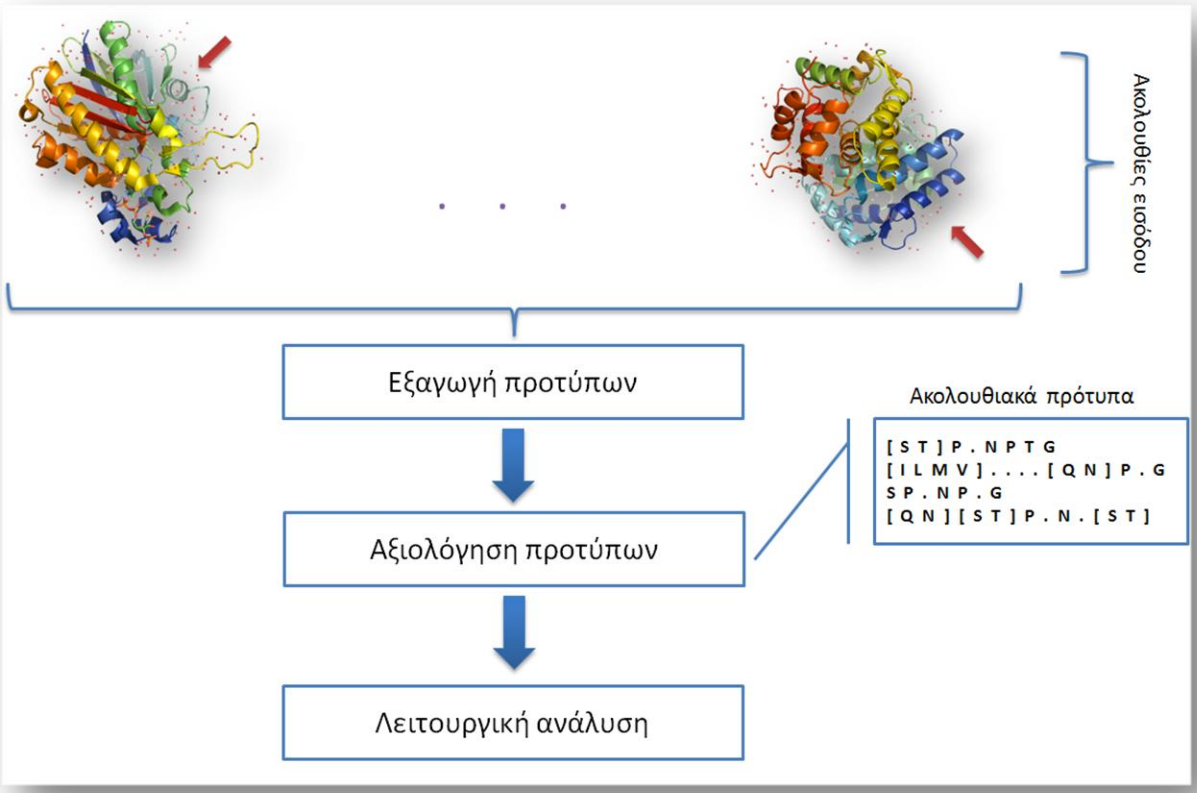



**Εικόνα 24:** Κατασκευή των συνόλων δεδομένων $D_{cis\text{-nonPro}}$ και $D_{trans\text{-nonPro}}$.

## *3.3 Μεθοδολογία*

Η προτεινόμενη μεθοδολογία αποτελείται από τρία στάδια που απεικονίζονται συνοπτικά στην Εικόνα 25. Συγκεκριμένα, σε κάθε πρωτεϊνική ακολουθία εντοπίζουμε τους *cis*-Pro και *cis*-nonPro και με τη χρήση ενός αλγορίθμου εξαγωγής ακολουθιακών προτύπων, εντοπίζουμε για κάθε κατηγορία δεσμού ξεχωριστά, αντιπροσωπευτικά πρότυπα που περιγράφουν τις περιοχές που περιβάλλουν τον κάθε τύπο δεσμού.

Ακολούθως, τα εξαχθέντα πρότυπα που αφορούν τους *cis*-Pro και *cis*-nonPro δεσμούς, αντιπαραβάλλονται με τα αντίστοιχα *trans* πρότυπα (*trans*-Pro και *trans*-nonPro) ώστε να διατηρηθούν μόνο τα πιο επιλεκτικά. Με βάση την τελική λίστα ακολουθιακών προτύπων για τους *cis*-Pro και *cis*-nonPro πεπτιδικούς δεσμούς ανατρέχουμε σε βάσεις δεδομένων που περιέχουν ακολουθιακά πρότυπα των οποίων η λειτουργία έχει επιβεβαιωθεί πειραματικά, ώστε να αναδείξουμε τις πιθανές λειτουργικές τάσεις των *cis* προτύπων.




**Εικόνα 25:** Εξαγωγή ακολουθιακών προτύπων και λειτουργική ανάλυση αυτών.

## *Εξαγωγή προτύπων*

Οι λίστες $D_{cis\text{-Pro}}$ και $D_{cis\text{-nonPro}}$ που περιέχουν παράθυρα αμινοξέων γύρω από *cis*-Pro και *cis*-nonPro δεσμούς αντίστοιχα, εισάγονται εκ περιτροπής στον αλγόριθμο TEIRESIAS [89] ώστε να εντοπίσουμε ακολουθιακά πρότυπα που εμφανίζονται με μεγάλη συχνότητα στις παραπάνω περιοχές. Ο TEIRESIAS είναι ένας αλγόριθμος

εντοπισμού ακολουθιακών προτύπων που περιλαμβάνει κάποια χαρακτηριστικά και δυνατότητες ειδικά στοχευμένες για βιολογικές ακολουθίες. Η ροή εκτέλεσής του αποτελείται από δύο φάσεις: την αναζήτηση και την συγχώνευση. Κατά την φάση της αναζήτησης εντοπίζονται στο σύνολο των ακολουθιών εισόδου, όλα τα στοιχειώδη πρότυπα που ικανοποιούν ένα σύνολο περιορισμών που έχουμε θέσει κατά την εκτέλεση του αλγορίθμου. Όσα πρότυπα ικανοποιούν αυτούς τους περιορισμούς καταγράφονται και ακολούθως κατά την φάση της συγχώνευσης, συνενώνονται προοδευτικά σε μεγαλύτερα και πιο συγκεκριμένα ακολουθιακά πρότυπα, μέχρις ότου καταγραφούν όλα τα πρότυπα που υπάρχουν στο σύνολο εισόδου. Τα κριτήρια που τίθενται κατά την εκτέλεση του αλγορίθμου αφορούν τα χαρακτηριστικά των εξαγόμενων προτύπων. Συγκεκριμένα το L αντιστοιχεί στον ελάχιστο αριθμό σταθερών χαρακτήρων, το W αφορά το μέγιστο μήκος του προτύπου και τέλος το K αποτελεί τον βαθμό στήριξης (support), δηλαδή το πλήθος εμφανίσεων του προτύπου στο σύνολο των ακολουθιών εισόδου. Στην παρούσα μεθοδολογία επιλέχτηκαν οι ακόλουθες τιμές για τις παραπάνω παραμέτρους: L=3, W=11 και K=2. Η τιμή L=3, αποτελεί την ελάχιστη τιμή για την οποία έχει αποδειχτεί ότι συγκλίνει η φάση συγχώνευσης [89] του TEIRESIA, το μέγιστο μήκος W=11 είναι όσο και το μήκος όλων των ακολουθιακών παραθύρων εισόδου, ενώ για τον βαθμό στήριξης επιλέχτηκε αρχικά μια χαμηλή τιμή, δεδομένου ότι κατά το στάδιο της αξιολόγησης θα πραγματοποιηθεί σχολαστικός έλεγχος των εξαχθέντων προτύπων.

Ακολούθως, για κάθε σύνολο ακολουθιών εισόδου ($D_{cis\text{-Pro}}$ και $D_{cis\text{-nonPro}}$) εκτελούνται τρία είδη εξαγωγής προτύπων: i) ακριβής εξαγωγή προτύπων, ii) εξαγωγή προτύπων λαμβάνοντας υπόψη χημικές ομαδοποιήσεις των αμινοξέων ([AG], [DE], [FY], [KR], [ILMV], [QN], [ST]) και iii) εξαγωγή προτύπων λαμβάνοντας υπόψη δομικές ομαδοποιήσεις των αμινοξέων ([CS], [DLN], [EQ], [FHWY], [ITV], [KMR]). Τα αμινοξέα μέσα στις αγκύλες θεωρούνται ισοδύναμα κατά την εκτέλεση του αλγορίθμου εξαγωγής προτύπων. Η χρήση χημικών και δομικών ομαδοποιήσεων των αμινοξέων είναι πιθανό να αποκαλύψει κρυμμένα πρότυπα στη γειτονιά των *cis*-Pro και *cis*-nonPro πεπτιδικών δεσμών. Για παράδειγμα, αν μια θέση στο σύνολο των ακολουθιών εισόδου καταλαμβάνεται με σχετικά μεγάλη συχνότητα από τα αμινοξέα φαινυλαλανίνη, ιστιδίνη, τρυπτοφάνη ή τυροσίνη, θα αντιστοιχισθεί στην ομάδα [FHWY] υποδηλώνοντας ροπή συνολικά προς αρωματικά αμινοξέα.

*Αξιολόγηση προτύπων*

Μετά την εξαγωγή των ακολουθιακών προτύπων από τις *cis*-Pro και *cis*-nonPro περιοχές, κάθε εξαχθέν πρότυπο συγκρίνεται με μεγάλες πρωτεϊνικές βάσεις δεδομένων καθώς και με ακολουθίες ελέγχου (*trans*), ώστε να αξιολογηθεί η σημαντικότητά του. Αρχικά, για κάθε πρότυπο υπολογίζεται η πιθανότητα το συγκεκριμένο πρότυπο να προκύψει κατά τύχη, δηλαδή να εντοπιστεί σε ένα εκτενές σύνολο πρωτεϊνικών ακολουθιών [90, 91]. Επίσης, είναι πολύ σημαντικό να εξαιρέσουμε πρότυπα που εντοπίζονται με μεγάλη συχνότητα και στα σύνολα δεδομένων των *trans* πεπτιδικών δεσμών. Τα σύνολα ελέγχου λοιπόν που χρησιμοποιούνται για την αξιολόγηση των *cis*-Pro και *cis*-nonPro προτύπων, είναι τα $D_{trans\text{-Pro}}$ και $D_{trans\text{-nonPro}}$, αντίστοιχα. Υπολογίζουμε λοιπόν για κάθε ακολουθιακό πρότυπο ένα μέτρο αξιολόγησης (*Score*, εξίσωση 9) που ουσιαστικά συγκρίνει αναλογικά την εκπροσώπηση του προτύπου στα *cis* και *trans* σύνολα ακολουθιών. Έστω ότι *P* είναι το υπό μελέτη πρότυπο και *M(P)* το σύνολο των ακολουθιακών περιοχών με τις οποίες ταιριάζει το πρότυπο *P*, οπότε ορίζεται η ακόλουθη εξίσωση:

$$Score = \frac{\left|D_{cis} \cap M(P)\right|}{\left|D_{cis} \cap M(P)\right| + balance \times \left|D_{trans} \cap M(P)\right|} \qquad (9)$$

Ο παράγοντας *balance* είναι μια σταθερά εξισορρόπησης που ουσιαστικά αντισταθμίζει την ανισοκατανομή των *cis* και *trans* δεσμών. Συγκεκριμένα για την περίπτωση των *cis*-Pro δεσμών $balance = |D_{cis\text{-}Pro}|/|D_{trans\text{-}Pro}|$ και για την περίπτωση των *cis*-nonPro δεσμών $balance = |D_{cis\text{-}nonPro}|/|D_{trans\text{-}nonPro}|$. Στην εξίσωση 9, $|D_{cis} \cap M(P)|$ είναι ο αριθμός των *cis* ακολουθιών που ταιριάζουν με το πρότυπο *P* και $|D_{trans} \cap M(P)|$ είναι αντίστοιχα ο αριθμός των *trans* ακολουθιών που ταιριάζουν με το πρότυπο P. Επομένως, η εξίσωση 9, τροποποιείται όπως φαίνεται στην εξισώσεις 10 και 11 αμέσως πιο κάτω, για την περίπτωση των *cis*-Pro και *cis*-nonPro πεπτιδικών δεσμών αντίστοιχα.

$$Score = \frac{\left|D_{cis-\mathrm{Pro}} \cap M(P)\right|}{\left|D_{cis-\mathrm{Pro}} \cap M(P)\right| + balance \times \left|D_{trans-\mathrm{Pro}} \cap M(P)\right|} \qquad (10)$$

$$Score = \frac{\left|D_{cis-non\,Pro} \cap M(P)\right|}{\left|D_{cis-non\,Pro} \cap M(P)\right| + balance \times \left|D_{trans-non\,Pro} \cap M(P)\right|} \quad (11)$$

Η εισαγωγή του παράγοντα *balance* αποτελεί ένα μέτρο για την αντιμετώπιση της ανισοκατανομής των κλάσεων στα *cis* και *trans* σύνολα δεδομένων, χωρίς να υπεισέρχονται ζητήματα δειγματοληψίας, που αφορούν την απαλοιφή πιθανώς σημαντικών δειγμάτων από την πλειοψηφούσα κλάση (*trans*). Όσον αφορά το *Score*, αυτό λαμβάνει τιμές μέσα στο διάστημα [0,1], όπου τιμές κοντά στο 0 δείχνουν έντονη συσχέτιση με το σύνολο των *trans* δεσμών ενώ αντίστοιχα τιμές κοντά στο 1 καταδεικνύουν έντονη συσχέτιση προς τους *cis* δεσμούς. Κατά συνέπεια, τιμές που λαμβάνουμε πάνω από το 0.5 σηματοδοτούν τάση προς συσχέτιση με τους *cis* δεσμούς. Το όριο που έχουμε θέσει στην συγκεκριμένη μελέτη είναι 0.90, ώστε να λάβουμε πρότυπα με σαφή συσχέτιση προς τους *cis* πεπτιδικούς δεσμούς.

Ακολούθως, στη διαδικασία αξιολόγησης, εφαρμόζουμε ένα επιπλέον κριτήριο, που αφορά το πλήθος εμφάνισης κάθε προτύπου στο σύνολο των *trans* (X-Pro και X-nonPro) πεπτιδικών δεσμών (δηλαδή $D_{trans\text{-Pro}}$ και $D_{trans\text{-nonPro}}$) όπως φαίνεται στην εξίσωση 12:

$$\begin{aligned} &if\ \left|D_{trans} \cap M(P)\right| = 0\ \ then\ \left|D_{cis} \cap M(P)\right| \geq 3 \\ &else\ \left|D_{cis} \cap M(P)\right| \geq 4. \end{aligned} \quad (12)$$

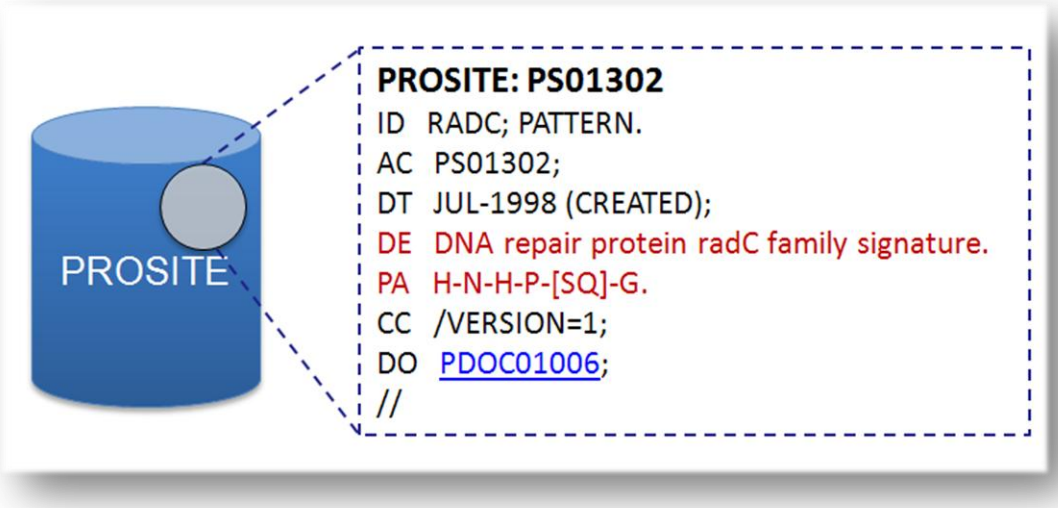
Με το παραπάνω κριτήριο διασφαλίζουμε ότι ακόμη και στην περίπτωση που ένα πρότυπο δεν ταιριάζει με καμία ακολουθία από το σύνολο ελέγχου ($D_{trans\text{-Pro}}$ ή $D_{trans\text{-nonPro}}$) τότε θα πρέπει και πάλι να ελέγχεται το πλήθος των ακολουθιών με τις οποίες ταιριάζει στο σύνολο $D_{cis\text{-Pro}}$ ή $D_{cis\text{-nonPro}}$, αντίστοιχα. Δηλαδή, να μην καταχωρείται ένα πρότυπο ότι παρουσιάζει μεγάλη συσχέτιση με τους *cis* δεσμούς, αν δεν ταιριάζει τουλάχιστον με τρείς *cis* ακολουθίες, ακόμη και αν ταιριάζει με 0 *trans* ακολουθίες οπότε και θα δίνει *Score*=1. Ελαφρώς πιο αυστηρό είναι το κατώφλι που αφορά το πλήθος των *cis* ακολουθιών που πρέπει να ταιριάζει κάθε πρότυπο, όταν αυτό ταιριάζει με μια τουλάχιστον *trans* ακολουθία.

Επιπρόσθετα, όσον αφορά τους X-nonPro πεπτιδικούς δεσμούς, επιχειρούμε να απαλείψουμε πλεονάζοντα ή και περιττά πρότυπα, δηλαδή πρότυπα που παρουσιάζουν

ομοιότητα μεταξύ τους και κατά συνέπεια τα σύνολα ακολουθιών που ανασύρουν εμφανίζουν σχετική αλληλοεπικάλυψη. Ο τελικός σκοπός είναι να καταλήξουμε σε ένα σύνολο ακολουθιακών προτύπων που παρουσιάζουν σχέση 1-N σε σχέση με τις *cis* ακολουθίες στο σύνολο $D_{cis\text{-nonPro}}$ διατηρώντας ταυτόχρονα σε χαμηλά επίπεδα την αντιπροσώπευσή τους στο σύνολο $D_{trans\text{-nonPro}}$. Ο αλγόριθμος που απαλείφει τα πλεονάζοντα πρότυπα τερματίζει είτε όταν επιτύχει 100% κάλυψη (Coverage) των *cis*-nonPro ακολουθιών, είτε όταν έχουν αξιολογηθεί όλα τα ακολουθιακά πρότυπα εισόδου. Ο αριθμός των *trans*-nonPro ακολουθιών που ανασύρει το κάθε πρότυπο αποτελεί το σφάλμα που υπεισέρχεται στο τελικό σύνολο ακολουθιακών προτύπων (FDR: False Discovery Rate).

## *Λειτουργική ανάλυση*

Στο επόμενο στάδιο της παρούσας μεθοδολογίας, για κάθε ακολουθιακό πρότυπο που προέκυψε, επιχειρούμε να ανακαλύψουμε τις πιθανές λειτουργικές του συσχετίσεις. Γι' αυτό το λόγο αναζητούμε ομοιότητες του κάθε εξαχθέντος προτύπου με πρότυπα που έχουν καταχωρηθεί σε βάσεις δεδομένων και των οποίων η λειτουργία έχει επιβεβαιωθεί πειραματικά. Συγκεκριμένα, τα πρότυπα που αφορούν τους *cis*-Pro πεπτιδικούς δεσμούς αντιπαραβάλλονται με τα πρότυπα της βάσης PROSITE [92], με τη χρήση ενός αλγορίθμου που αναζητά εάν υπάρχει επικάλυψη μεταξύ δυο ακολουθιακών προτύπων [93]. Η βάση PROSITE, ουσιαστικά αποτελείται από ένα σύνολο ακολουθιακών πρωτεϊνικών προτύπων, καταχωρημένων με τη μορφή κανονικών εκφράσεων, όπου το κάθε πρότυπο συνοδεύεται από την λειτουργία που έχει βρεθεί πειραματικά ότι επιτελεί, καθώς και από πλήθος άλλων πληροφοριών. Μία τυπική καταχώρηση της βάσης PROSITE φαίνεται στην Εικόνα 26.




**Εικόνα 26:** Τυπική εγγραφή της βάσης PROSITE.

Από την κάθε καταχώρηση αυτό που μας ενδιαφέρει περισσότερο για τον σκοπό της παρούσας μελέτης συνοψίζεται στις δύο γραμμές που έχουν τονιστεί με έντονα (κόκκινα) γράμματα στην Εικόνα 26, δηλαδή η κανονική έκφραση του προτύπου και η αντίστοιχη λειτουργία που αυτό επιτελεί. Ο αλγόριθμος σύγκρισης ακολουθιακών προτύπων που χρησιμοποιείται αναζητά επικάλυψη μεταξύ του *cis*-Pro προτύπου και της κανονικής έκφρασης που υπάρχει στην PROSITE. Κατά την σύγκριση αναζητείται αρχικά αν επικαλύπτεται μερικώς είτε και πλήρως το *cis*-Pro πρότυπο από την κανονική έκφραση στην PROSITE και στη συνέχεια το αντίστροφο, δηλαδή, εάν το *cis*-Pro πρότυπο υπερτίθεται του PROSITE προτύπου, οδηγώντας έτσι σε εξαντλητική αναζήτηση πιθανών επικαλύψεων μεταξύ των δύο προτύπων.

Στην περίπτωση των προτύπων στη γειτονιά των *cis*-nonPro πεπτιδικών δεσμών, η βάση αναφοράς που χρησιμοποιούμε για να εντοπίσουμε πιθανές λειτουργικές συσχετίσεις των εξαχθέντων προτύπων είναι η δικτυακή βάση Eukaryotic Linear Motif (ELM) [94]. Κατ' αντιστοιχία με την βάση PROSITE, η βάση ELM περιέχει επίσης ζεύγη ακολουθιακών προτύπων με την επιβεβαιωμένη λειτουργία που έχει βρεθεί ότι αυτά επιτελούν. Οι λειτουργίες που είναι καταχωρημένες στην ELM διακρίνονται σε τέσσερις ευρείες κατηγορίες: i) στόχευση/εντόπιση (targeting/localization: TRG), ii) μετα-μεταγραφικές τροποποιήσεις (post-translational modifications: MOD), iii) πρόσδεση/υποκατάσταση (ligand/binding: LIG) και iv) θέσεις αποκοπής (cleavage: CLV). Για την σύγκριση μεταξύ των *cis*-nonPro προτύπων και των κανονικών εκφράσεων που

περιγράφουν τις λειτουργίες που είναι καταχωρημένες στην βάση ELM, χρησιμοποιήθηκε το πρόγραμμα Comparimotif [95]. Σε κάθε σύγκριση ανατίθεται ένα σκορ (CM_score), και ποσοτικοποιεί την ομοιότητα μεταξύ δύο πρωτεϊνικών προτύπων λαμβάνοντας υπόψη την μεταξύ τους επικάλυψη, το μήκος τους και τον εκφυλισμό τους.

## *3.4 Αποτελέσματα - συζήτηση*

Από την παραπάνω μεθοδολογική ανάλυση προέκυψαν λίστες με ακολουθιακά πρότυπα, που περιγράφουν με εύληπτο τρόπο την γειτονιά των *cis*-Pro και *cis*-nonPro πεπτιδικών δεσμών. Οι λίστες αυτές που περιγράφουν τα ακολουθιακά πρότυπα με τη μορφή κανονικών εκφράσεων, μπορούν ακολούθως να συμβάλλουν στην γενίκευση των υπαρχουσών γνώσεων σχετικά με τις *cis* διαμορφώσεις, καθώς και να χρησιμοποιηθούν περαιτέρω για να εξαχθούν πληροφορίες που αφορούν τις λειτουργικές τους συσχετίσεις.

### *Cis-Pro πεπτιδικοί δεσμοί*

Ο Πίνακας 7 περιέχει τα 20 καλύτερα ακολουθιακά πρότυπα στη γειτονιά των *cis*-Pro πεπτιδικών δεσμών, όπως προέκυψαν από την προτεινόμενη μεθοδολογία.

**Πίνακας 7:** Τα 20 καλύτερα ακολουθιακά πρότυπα των *cis*-Pro περιοχών.

| Ακριβής εξαγωγή προτύπων | | | Χημικές ομαδοποιήσεις αμινοξέων | | | Δομικές ομαδοποιήσεις αμινοξέων | | |
|---|---|---|---|---|---|---|---|---|
| **Πρότυπο** | **Σκορ** | **Σημαντι-κότητα** | **Πρότυπο** | **Σκορ** | **Σημαντι-κότητα** | **Πρότυπο** | **Σκορ** | **Σημαντι-κότητα** |
| FE.P...F | 1 | -15.5 | [ST]P.NPTG | 1 | -24.4 | [DLN].[DLN]...[KMR].[ITV][CS] | 1 | -15.2 |
| SP.NP.G | 0.99 | -19.8 | [ST]PNNP.G | 1 | -25.0 | [CS]..NPTG | 1 | -19.7 |
| V...EP...H | 0.99 | -15.4 | [ILMV]..P.NP.G | 1 | -18.4 | SP.NP.G | 0.99 | -19.8 |
| GPY.G | 0.99 | -14.9 | [ST]PN..T[AG] | 1 | -18.5 | [CS]...[FHWY].[FHWY].N | 0.99 | -12.6 |
| G...GPY | 0.98 | -15.0 | [ST]P.NP.G | 0.99 | -18.8 | [EQ].P[FHWY][ITV].V | 0.99 | -17.7 |
| P.NPTG | 0.98 | -19.6 | [ILMV][KR]EP[FY] | 0.99 | -17.5 | [FHWY]P.E[FHWY]I | 0.99 | -19.1 |
| PNNP.G | 0.98 | -20.0 | SP.NP.G | 0.99 | -19.8 | [ITV]..P.NP.G | 0.99 | -18.8 |
| PY..SG | 0.98 | -14.7 | [ILMV][KR]EPF | 0.99 | -17.9 | GPY.G | 0.99 | -14.9 |
| NNPT | 0.97 | -14.6 | [ILMV]....[QN]P.G | 0.99 | -12.6 | [CS].PNNP | 0.99 | -19.6 |
| V.....N..T | 0.97 | -8.1 | [QN]..FV[FY] | 0.99 | -13.9 | V...EP...H | 0.99 | -15.4 |

| P..YP.K | 0.97 | -15.5 | E.GP[FY] | 0.99 | -14.8 | [ITV].SP.[DLN]P | 0.99 | -17.6 |
|---|---|---|---|---|---|---|---|---|
| N...P.PE | 0.97 | -14.8 | GPY.G | 0.99 | -14.9 | PY..[CS]G | 0.98 | -14.5 |
| GPY...G | 0.97 | -15.2 | PNNP[ST] | 0.99 | -19.4 | P.NPTG | 0.98 | -19.6 |
| SP.N..G | 0.97 | -14.4 | [ILMV].[ST].N...G | 0.99 | -12.2 | G...GPY | 0.98 | -15.0 |
| S..NP.G | 0.96 | -14.2 | GY..[ILMV].K | 0.99 | -13.2 | PNNP.G | 0.98 | -20.0 |
| T...NP.G | 0.96 | -14.3 | C[ST]..NP | 0.99 | -14.3 | [ITV][FHWY]E...F | 0.98 | -13.0 |
| PYG.S | 0.96 | -15.0 | V...EP...H | 0.99 | -15.4 | PY..SG | 0.98 | -14.7 |
| QL.....Y | 0.96 | -9.1 | [QN][ST]P.N.[ST] | 0.99 | -17.9 | S..P.D[KMR][DLN] | 0.98 | -18.1 |
| R...Y.P | 0.96 | -9.3 | [ILMV]...[ST]P.N.T | 0.99 | -17.3 | K.P[FHWY]T..[ITV] | 0.98 | -18.5 |
| N.K..F | 0.96 | -9.0 | P[FY]PE.[AG] | 0.99 | -19.5 | [KMR]..GP[FHWY]..[ITV] | 0.98 | -18.2 |

Τα πρότυπα που περιέχονται στον Πίνακας 7 ακολουθούν κάποιες τυπικές συμβάσεις των κανονικών εκφράσεων. Συγκεκριμένα, η τελεία "." σημαίνει ότι αυτή η θέση μπορεί να καταληφθεί από οποιοδήποτε αμινοξύ, δηλαδή, κατά την αναζήτηση ακολουθιακών προτύπων δεν παρατηρήθηκε κάποιο συγκεκριμένο αμινοξύ με σημαντικά μεγαλύτερη συχνότητα από τα υπόλοιπα. Επίσης, οι αγκύλες χρησιμοποιούνται για να δείξουν ότι τα αμινοξέα που περικλείουν είναι ισοδύναμα, και παρατηρούνται με την ίδια ή παρόμοια συχνότητα στη συγκεκριμένη θέση του υπό μελέτη προτύπου.

Παρατηρώντας τον Πίνακας 7 βλέπουμε ότι τα πρότυπα που έχουν εξαχθεί λαμβάνοντας υπόψη είτε χημικές είτε δομικές ομαδοποιήσεις των αμινοξέων παρουσιάζουν σχετικά μικρότερες τιμές όσον αφορά τη σημαντικότητα. Όπως αναφέρθηκε και προηγούμενα η σημαντικότητα ουσιαστικά εκφράζει αντίστροφα το πόσο πιθανό είναι να εντοπιστεί ένα πρότυπο κατά τύχη σε μια μεγάλη πρωτεϊνική βάση. Αυτό καταδεικνύει ότι οι χημικές και δομικές ιδιότητες των αμινοξέων συμβάλλουν στην διάκριση μεταξύ των *cis*-Pro και *trans*-Pro ισομερών διαμορφώσεων. Μάλιστα αυτό το συμπέρασμα ενισχύεται αν αναλογιστούμε ότι τα πρότυπα που έχουν εξαχθεί κατά την ακριβή εξαγωγή προτύπων δεν περιλαμβάνουν την ασάφεια που υπεισέρχεται με τις ομαδοποιήσεις και είναι πιο συγκεκριμένα, οπότε και θα ανέμενε κανείς να παρουσιάζουν ακόμη χαμηλότερες τιμές ως προς την σημαντικότητα.

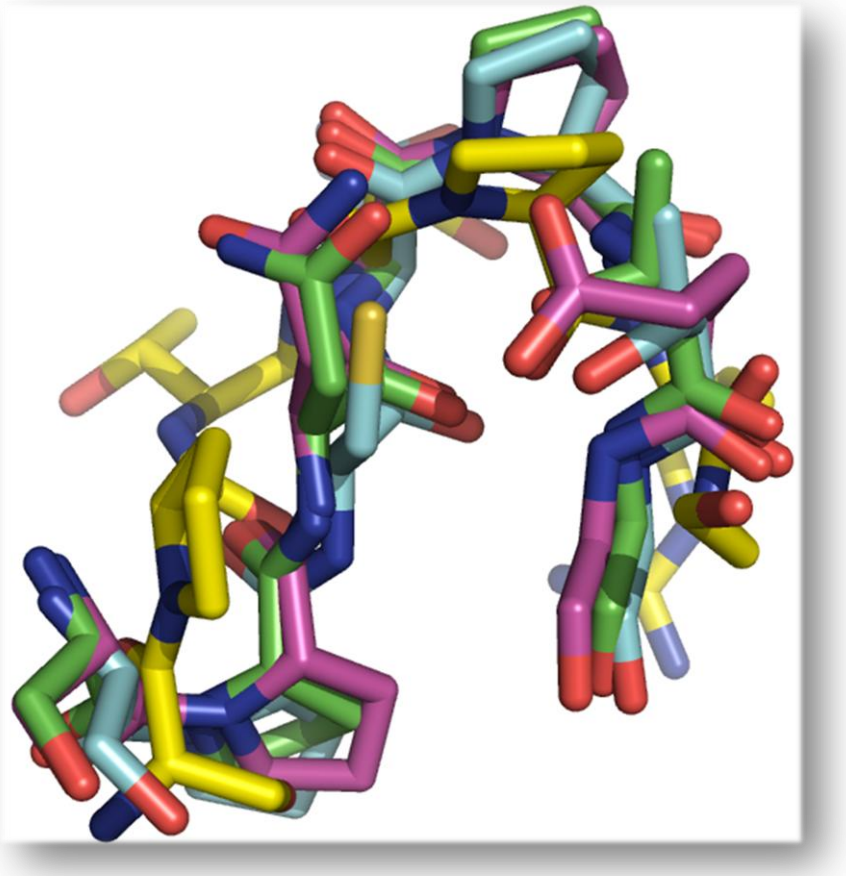
Επίσης, πολλά πρότυπα είναι κοινά, είτε εξ ολοκλήρου είτε με μικρές τροποποιήσεις, και στους τρεις τρόπους εξαγωγής προτύπων. Αυτές οι παραλλαγές σε κάθε περίπτωση φαίνεται απλώς να αποτελούν παρεκκλίσεις γύρω από ένα βασικό πρότυπο. Χαρακτηριστικά παραδείγματα τέτοιων προτύπων είναι τα "SP.NP.G", GPY.G" και "V…EP…H", που έχουν παραχθεί σε κάθε τύπο εξαγωγής προτύπων και όλα τους έχουν

αποφέρει πολύ υψηλές βαθμολογίες. Στον Πίνακας 8 παρουσιάζονται για εποπτικούς λόγους οι παραλλαγές του βασικού προτύπου "SP.NP.G" που ενυπάρχουν μόνο στον Πίνακας 7 με τα 20 καλύτερα πρότυπα. Τα πρότυπα έχουν στοιχιστεί ώστε να είναι εμφανείς οι μεταξύ τους ομοιότητες.

**Πίνακας 8:** Παραλλαγές του βασικού προτύπου "SP.NP.G".

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Ακριβής εξαγωγή προτύπων** | | | | | S | P | . | N | P | . | G |
| | | | | | | P | . | N | P | T | G |
| | | | | | | P | N | N | P | . | G |
| | | | | | | | N | N | P | T | |
| | | | | | S | P | . | N | . | . | G |
| | | | | | S | . | . | N | P | . | G |
| | | | | T | . | . | . | N | P | . | G |
| **Χημική ομαδοποίηση αμινοξέων** | | | | | [ST] | P | . | N | P | T | G |
| | | | | | [ST] | P | N | N | P | . | G |
| | | | [ILMV] | . | . | P | . | N | P | . | G |
| | | | | | [ST] | P | N | . | . | T | [AG] |
| | | | | | [ST] | P | . | N | P | . | G |
| | | | | | | P | N | N | P | [ST] | |
| | | | [ILMV] | . | [ST] | . | N | . | . | . | G |
| | | | | [QN] | [ST] | P | . | N | . | [ST] | |
| | [ILMV] | . | . | . | [ST] | P | . | N | . | T | |
| **Δομική ομαδοποίηση αμινοξέων** | | | | | [CS] | . | . | N | P | T | G |
| | | | [ITV] | . | . | P | . | N | P | . | G |
| | | | | [CS] | . | P | N | N | P | | |
| | | | [ITV] | . | S | P | . | [DLN] | P | | |
| | | | | | | P | . | N | P | T | G |
| | | | | | | P | N | N | P | . | G |

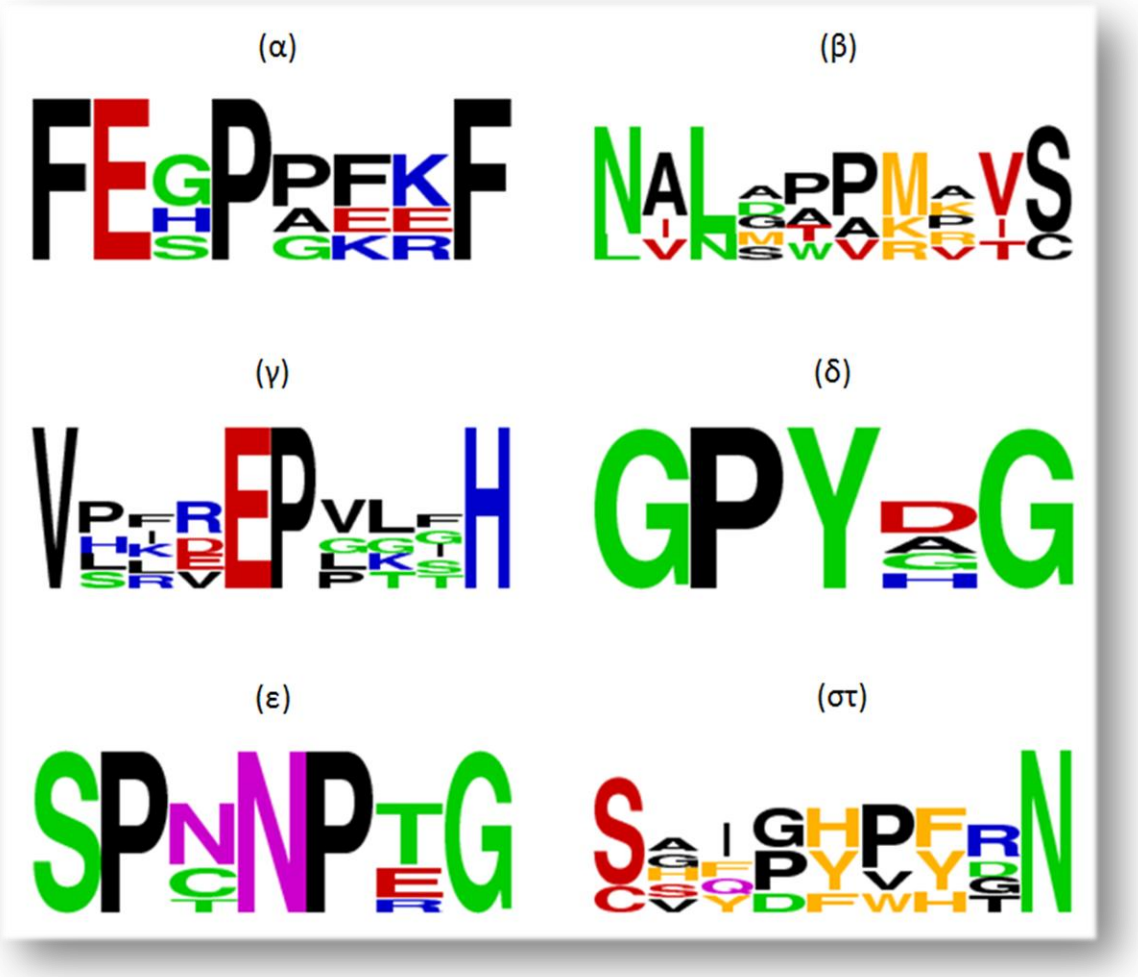
Επιπρόσθετα, για το παραπάνω πρότυπο ("SP.NP.G") φαίνεται στην Εικόνα 27 μια δομική στοίχιση μεταξύ των ακολουθιών από τις οποίες έχει προκύψει, ώστε να οπτικοποιήσουμε σε αδρές γραμμές την διάταξη των αμινοξέων που αποτελούν το πρότυπο στο χώρο.



**Εικόνα 27:** Απεικόνιση του προτύπου "SP.NP.G".

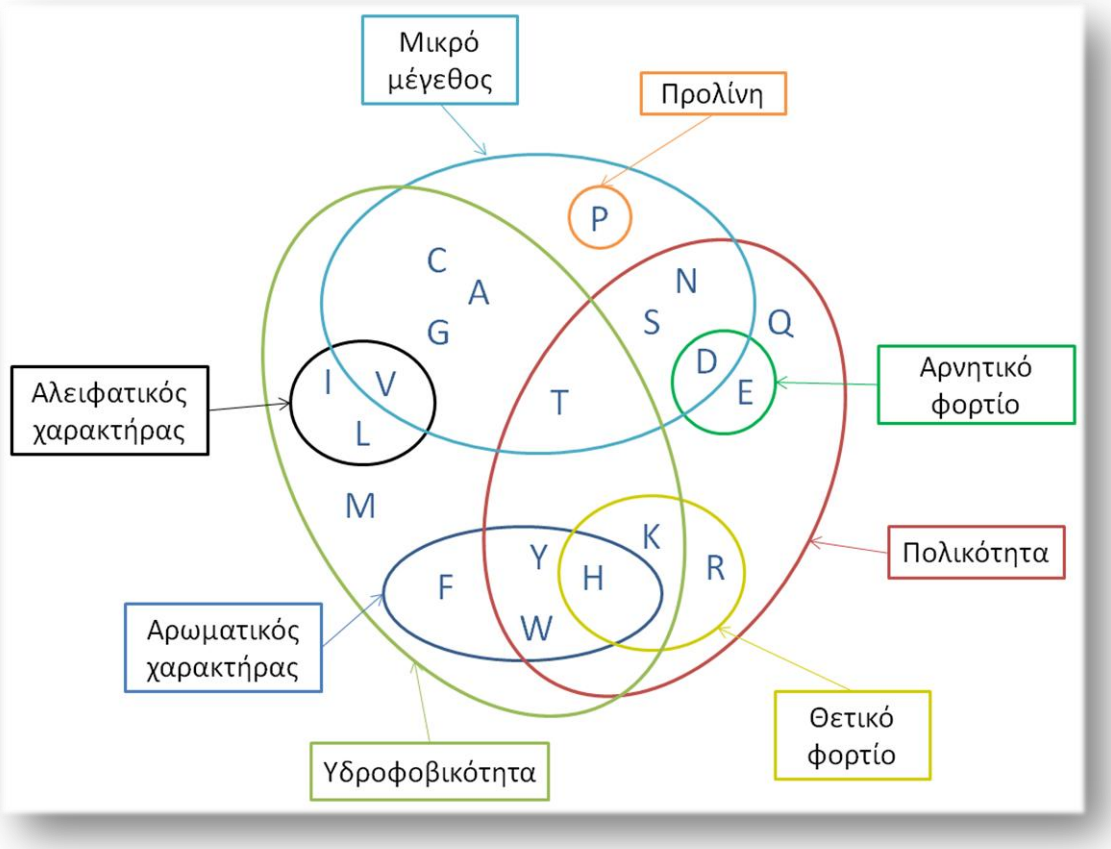
Αξιολογώντας τα 20 καλύτερα πρότυπα που προέκυψαν, αρκετές ενδιαφέρουσες παρατηρήσεις και συμπεράσματα μπορούν να εξαχθούν. Μικρά, πολικά αμινοξέα, ιδίως η σερίνη και η ασπαραγίνη, βρίσκονται συνήθως ως συστατικά του *cis*-Pro πεπτιδικού δεσμού. Αυτό παρατηρείται κυρίως στο μοτίβο "SP.NP.G" και τις παραλλαγές του, όπου ο *cis*-Pro πεπτιδικός δεσμός παρατηρείται συχνά μεταξύ των αμινοξέων S-P και N-P. Η παρατήρηση αυτή συνάδει επίσης και με την σχετική βιβλιογραφία που αφορά τις προτιμήσεις των αμινοξέων ως προς τον σχηματισμό *cis*-Pro πεπτιδικών δεσμών [11, 19]. Επιπλέον, η γλυκίνη βρίσκεται με υψηλή συχνότητα, είτε ως αμινοξύ που συμμετέχει στη δημιουργία του πεπτιδικού δεσμού είτε στην γειτονιά του. Συγκεκριμένα, μόνο στον Πίνακας 7 η γλυκίνη εντοπίζεται σε 33 από τα 60 πρότυπα. Μερικά αντιπροσωπευτικά πρότυπα είναι το "SP.NP.G" και οι διάφορες παραλλαγές του, καθώς επίσης και τα: "GPY.G", "G…GPY", "[KMR]..GP[FWHY]..[ITV]" και "PY..SG". Το μικρό σε μέγεθος κατάλοιπο της γλυκίνης αυξάνει την πιθανότητα για σχηματισμό *cis* πεπτιδικού δεσμού, πιθανώς λόγω έλλειψης στερεοχημικής παρεμπόδισης. Σε πλήθος ερευνητικών εργασιών

έχει τονιστεί η εμφάνιση της γλυκίνης με μεγάλη συχνότητα στην άμεση γειτονιά *cis*-Pro πεπτιδικών δεσμών [9-12, 15, 19]. Έχει επίσης αποδειχτεί ότι η γλυκίνη, σε αντίθεση με θετικά φορτισμένα κατάλοιπα, δρα ως σταθεροποιητικός παράγοντας όταν εντοπίζεται σε θέσεις μετά τον πεπτιδικό δεσμό. Ακόμη, έχει προταθεί στη βιβλιογραφία ότι αρωματικά αμινοξέα (φαινυλαλανίνη, τρυπτοφάνη, ιστιδίνη και τυροσίνη) καταλαμβάνουν συχνά θέσεις που βρίσκονται σε άμεση γειτνίαση με *cis*-Pro πεπτιδικούς δεσμούς. Αυτή η τάση προς αρωματικά αμινοξέα επιβεβαιώνεται και από τα εξαχθέντα πρότυπα, αν και στην παρούσα μελέτη παρατηρείται σαφής προτίμηση προς τα αμινοξέα τυροσίνη και φαινυλαλανίνη. Χαρακτηριστικά παραδείγματα αποτελούν τα πρότυπα "FE.P…F", "GPY.G", "PY..SG", "[ILMV][KR]EP[FY]" και "[CS]…[FWHY].[FWHY].N" που περιέχουν πλήθος αρωματικών καταλοίπων. Παρόμοιες παρατηρήσεις έχουν καταγραφεί και στην βιβλιογραφία [11, 15] και μάλιστα είχε τονιστεί και η μεγάλη συχνότητα εμφάνισης των αμινοξέων τυροσίνη και φαινυλαλανίνη να έπονται του πεπτιδικού δεσμού, ενισχύοντας κατ' αυτόν τον τρόπο περαιτέρω την αξιοπιστία των εξαχθέντων προτύπων. Επιπλέον, παρατηρούμε μια αρκετά υψηλή συχνότητα προς β-διακλαδισμένα (b-branched) αμινοξέα (βαλίνη, ισολευκίνη και θρεονίνη), όπως φαίνεται στα πρότυπα "V…EP…H", "[DLN].[DLN]…[KMR].[ITV][CS]", "[EQ].P[FHWY][ITV].V" και "[ITV]..P.NP.G". Σύμφωνα με τη σχετική βιβλιογραφία [11], τα β-διακλαδισμένα αμινοξέα προσφέρουν την απαραίτητη δομική παρεμπόδιση για τον ισομερισμό ενός δεσμού από *trans*-Pro σε *cis*-Pro. Αξίζει να σημειωθεί ότι η γειτονιά των *cis*-Pro πεπτιδικών δεσμών κατά βάση παρουσιάζει έλλειψη σε αλανίνη και ασπαρτικό οξύ, μάλιστα η αλανίνη φαίνεται να εμφανίζεται ως επί το πλείστον ως μέρος του ομαδοποίησης [AG]. Στην περίπτωση του ασπαρτικού οξέως μπορούμε να παρατηρήσουμε ότι παρότι εντοπίζεται σε τρία από τα πρότυπα που σημείωσαν ιδιαιτέρως υψηλά σκορ ("[DLN].[DLN]...[KMR].[ITV][CS]", "[ITV].SP.[DLN]P", “S..P.D[KMR][DLN]"), αυτό συμβαίνει μόνον όταν λαμβάνονται υπόψη οι δομικές ομαδοποιήσεις των αμινοξέων. Επομένως η εντόπισή του στα παραπάνω πρότυπα οφείλεται κατά κύριο λόγο στην ομαδοποίηση [DLN] όπου συμμετέχει. Μερικά αντιπροσωπευτικά πρότυπα από αυτά που περιέχονται στον Πίνακας 7 απεικονίζονται με τη χρήση ακολουθιακών λογότυπων (sequence logo) στην Εικόνα 28 προσφέροντας μια γραφική και πιο διαισθητική απεικόνιση [96].




**Εικόνα 28:** Ακολουθιακά λογότυπα για τα πιο αντιπροσωπευτικά πρότυπα των *cis*-Pro πεπτιδικών δεσμών. (α) "FE.P...F", (β) "[DLN]. [DLN]... [KMR]. [ITV] [CS]", (γ) "V...EP...H", (δ) "GPY.G", (ε) "SP.NP.G" και (στ) "[CS]...[FHWY]. [FHWY].N".

Επιπλέον, για τους *cis*-Pro πεπτιδικούς δεσμούς παραθέτουμε πρότυπα που προκύπτουν έπειτα από διαχωρισμό των αμινοξέων με βάση ένα σύνολο από φυσικοχημικές ιδιότητες. Οι ιδιότητες που λαμβάνονται υπόψη, καθώς και οι η κατανομή των 20 αμινοξέων σε αυτές παρατίθενται με τη μορφή διαγράμματος Venn στην Εικόνα 29. Σχηματίζονται κατ' αυτόν τον τρόπο 8 δυαδικά χαρακτηριστικά, όπου το κάθε αμινοξύ είτε φέρει είτε δεν φέρει την καθεμιά από τις 8 φυσικοχημικές ιδιότητες.




**Εικόνα 29:** Κατανομή των αμινοξέων σε 8 σημαντικές φυσικοχημικές ιδιότητες.

Τα πρότυπα που προκύπτουν με βάση τις 8 φυσικοχημικές ιδιότητες παρατίθενται στον Πίνακας 9. Για κάθε φυσικοχημική ιδιότητα παρουσιάζονται τα 10 καλύτερα πρότυπα, χρησιμοποιώντας δυαδική αναπαράσταση. Κατά σύμβαση, με "0" σηματοδοτείται μια θέση που καταλαμβάνεται από αμινοξύ που δεν φέρει την υπό μελέτη φυσικοχημική ιδιότητα, ενώ το "1" αντιπροσωπεύει αμινοξύ που φέρει την συγκεκριμένη ιδιότητα. Όπως και προηγούμενα, η τελεία "." λογίζεται ως οποιοδήποτε από τα 20 αμινοξέα.

**Πίνακας 9:** Τα 10 καλύτερα πρότυπα με βάση 8 φυσικοχημικές ιδιότητες.

| | Πρότυπο | Σκορ | Σημαντι-κότητα | | Πρότυπο | Σκορ | Σημαντι-κότητα |
|---|---|---|---|---|---|---|---|
| **Υδροφοβικότητα** | 100010.0.00 | 0.88 | -25.62 | **Αλειφατικός χαρακτήρας** | 111.0000010 | 0.90 | -28.43 |
| | 00011010.00 | 0.86 | -29.12 | | 11100000010 | 0.89 | -31.66 |
| | 00011000111 | 0.86 | -32.00 | | 111000.0010 | 0.89 | -28.38 |
| | 100010.000 | 0.85 | -25.68 | | 111.00.0010 | 0.89 | -25.15 |
| | 10001000..0 | 0.85 | -25.67 | | 101000101.0 | 0.88 | -28.45 |
| | 11101011000 | 0.85 | -31.53 | | 1110001.1 | 0.87 | -23.33 |
| | 1000.00000 | 0.84 | -26.22 | | 010.0010101 | 0.86 | -28.45 |
| | 1101011000 | 0.84 | -28.09 | | 01..0010101 | 0.85 | -25.17 |
| | 0011010000 | 0.83 | -29.15 | | 111.00100.0 | 0.85 | -25.18 |
| | 00.11010000 | 0.83 | -29.14 | | 0100.001110 | 0.85 | -28.34 |
| **Πολικότητα** | 00011001000 | 0.87 | -31.46 | **Αρωματικός χαρακτήρας** | 11010..100 | 0.93 | -23.78 |
| | 00.11001000 | 0.86 | -27.80 | | 1101…10 | 0.92 | -17.53 |
| | 00111001000 | 0.85 | -31.44 | | 1101..010 | 0.91 | -20.68 |
| | 11000011000 | 0.84 | -31.39 | | 1101.0.10 | 0.90 | -20.68 |
| | 00000000011 | 0.84 | -31.47 | | 01.010.0010 | 0.87 | -25.02 |
| | 00101000001 | 0.83 | -31.51 | | 010010100.0 | 0.88 | -28.22 |
| | 0001100.000 | 0.83 | -27.69 | | 10010100.0 | 0.88 | -25.11 |
| | 11001000111 | 0.82 | -31.39 | | 0.101000010 | 0.88 | -28.19 |
| | 11010000011 | 0.82 | -31.39 | | 0.0.0010110 | 0.88 | -24.91 |
| | 110010.0111 | 0.81 | -27.73 | | 01.010.001 | 0.87 | -21.91 |
| **Μικρό μέγεθος** | 01011100101 | 0.84 | -31.44 | **Θετικό φορτίο** | 110.10010 | 0.87 | -23.35 |
| | 11111101001 | 0.83 | -31.32 | | 0.00011100 | 0.81 | -24.56 |
| | 10111100011 | 0.83 | -31.40 | | 0000011100 | 0.80 | -27.67 |
| | 1011100101 | 0.82 | -27.79 | | --- | --- | --- |
| | 01101101001 | 0.82 | -31.53 | | --- | --- | --- |
| | 10111110100 | 0.81 | -31.36 | | --- | --- | --- |
| | 01101101.01 | 0.80 | -27.81 | | --- | --- | --- |
| | 0101110.101 | 0.80 | -27.73 | | --- | --- | --- |
| | 01101101101 | 0.80 | -31.57 | | --- | --- | --- |
| | --- | --- | --- | | --- | --- | --- |
| **Προλίνη** | 000001110.0 | 0.81 | -29.86 | **Αρνητικό φορτίο** | 01.0.000011 | 0.80 | -24.94 |
| | 0000.1101 | 0.80 | -23.84 | | --- | --- | --- |
| | 000001101 | 0.80 | -26.89 | | --- | --- | --- |
| | --- | --- | --- | | --- | --- | --- |
| | --- | --- | --- | | --- | --- | --- |
| | --- | --- | --- | | --- | --- | --- |
| | --- | --- | --- | | --- | --- | --- |
| | --- | --- | --- | | --- | --- | --- |

Παρατηρώντας τον Πίνακας 9, μια ενδιαφέρουσα παρατήρηση που προκύπτει είναι ότι για ορισμένες φυσικοχημικές ιδιότητες (π.χ. θετικό και αρνητικό φορτίο των αμινοξέων) διατηρούνται ελάχιστα πρότυπα. Συμπεραίνουμε, λοιπόν, ότι οι ιδιότητες αυτές διαδραματίζουν μάλλον ήσσονος σημασίας ρόλο ως προς την διάκριση των *cis*-Pro και *trans*-Pro ισομερών διαμορφώσεων του πεπτιδικού δεσμού, συνεπώς, οι ιδιότητες αυτές κατανέμονται ομοιόμορφα στην γειτονία των *cis*-Pro και *trans*-Pro δεσμών. Πράγματι,

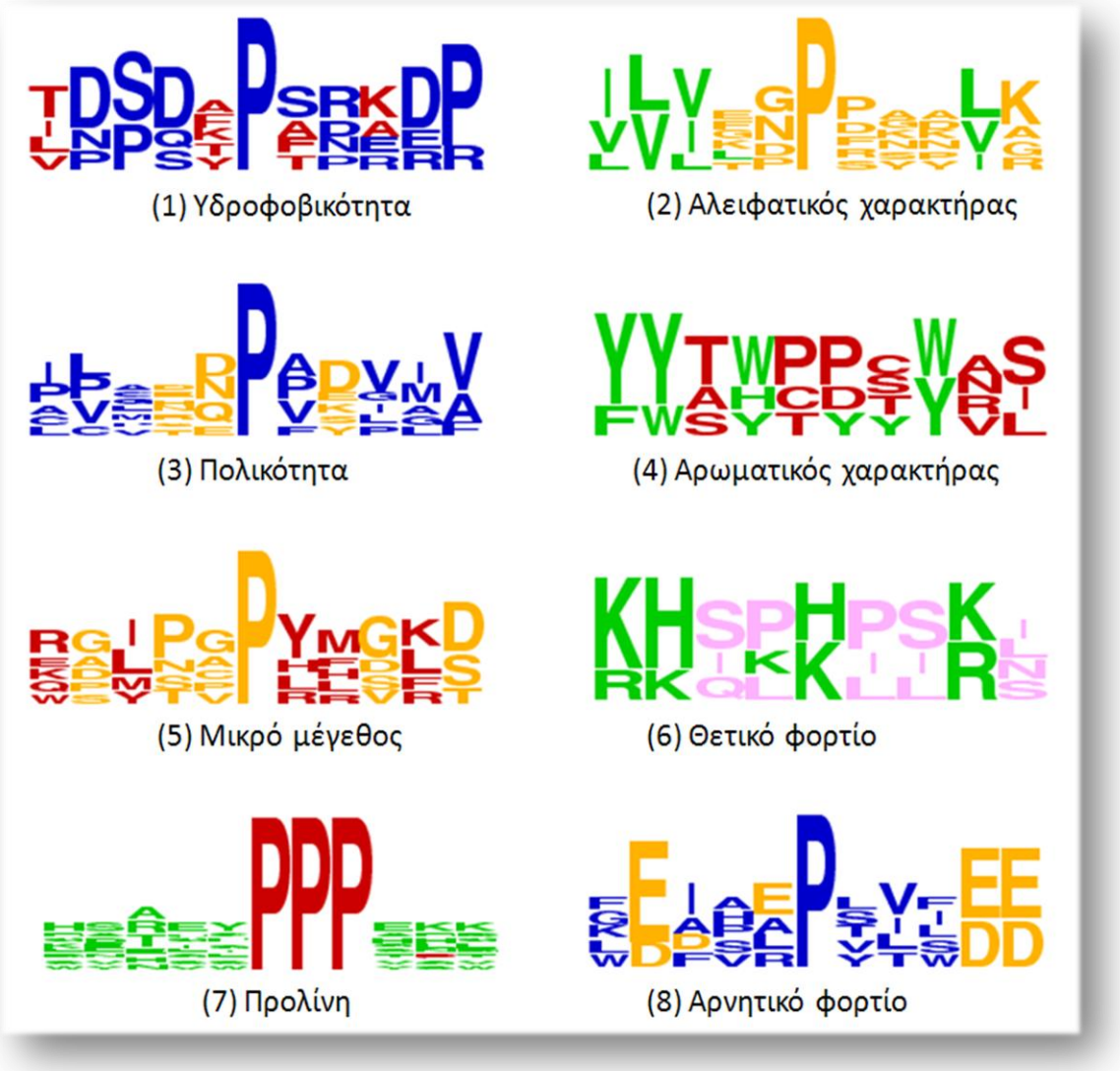
ούτε στην βιβλιογραφία έχει αναφερθεί συσχετισμός της διαμόρφωσης ενός πεπτιδικού δεσμού και του φορτίου των γειτνιαζόντων αμινοξέων. Όσον αφορά τα πρότυπα που προέκυψαν παρατηρώντας τις εμφανίσεις των προλινών στην γειτονιά *cis*-Pro πεπτιδικών δεσμών, λίγα πρότυπα έχουν διατηρηθεί ως σημαντικά, δίδοντας παράλληλα σχετικά χαμηλά σκορ, γεγονός που δεν μας επιτρέπει να εξάγουμε αξιόπιστες συσχετίσεις βάσει αυτής της ιδιότητας.

Όπως φαίνεται στον Πίνακας 9 τα υψηλότερα σκορ παρατηρούνται όταν λογίζεται ο αρωματικός χαρακτήρας των γειτονικών αμινοξέων. Σχεδόν όλα τα πρότυπα που έχουν διατηρηθεί ως σημαντικά, διαθέτουν αρωματικά αμινοξέα (π.χ. "11010..100", "1101…10", "1101..010", "1101.0.10"). Παρόμοιες παρατηρήσεις έχουν αναφερθεί και προηγουμένως στα πρότυπα που αφορούσαν μεμονωμένα αμινοξέα είτε και χημικές ή δομικές ομαδοποιήσεις αυτών. Η θετική επίδραση των αρωματικών αμινοξέων στην σταθεροποίηση *cis*-Pro πεπτιδικών δεσμών θεωρείται ότι οφείλεται στην δομική παρεμπόδιση του αρωματικού δακτυλίου που ωθεί τον σχηματισμό *cis* διαμόρφωσης [11], ωστόσο η συγκεκριμένη θεωρία δεν έχει πλήρως επιβεβαιωθεί.

Εξίσου υψηλά σκορ με τον αρωματικό χαρακτήρα των αμινοξέων παρατηρούνται και στην περίπτωση που λαμβάνεται υπόψη ο υδροφοβικός, πολικός και αλειφατικός χαρακτήρας των αμινοξέων. Ειδικά στην περίπτωση των προτύπων με βάση τον αλειφατικότητα των αμινοξέων, συχνά παρατηρούμε μια σειρά από τρία διαδοχικά αλειφατικά κατάλοιπα. Ακολούθως, όσον αφορά την πολικότητα, σε πολλά από τα εξαχθέντα πρότυπα παρατηρούμε ζεύγη πολικών αμινοξέων, όπως φαίνεται στα πρότυπα: "00011001000", "11000011000", "00000000011", "11010000011" καθώς και τις παραλλαγές αυτών. Παρόμοια συμπεράσματα μπορούν να εξαχθούν και σε σχέση με την υδροφοβικότητα των αμινοξέων όπου ζεύγη υδροφοβικών καταλοίπων εμφανίζονται με μεγάλη συχνότητα στα πρότυπα του Πίνακας 9. Αντιπροσωπευτικά πρότυπα είναι τα: "00011010.00", "00011000111", "11101011000", "1101011000" και "0011010000".

Τέλος, όταν λογίζεται το μέγεθος των αμινοξέων και ιδίως όταν επικεντρωνόμαστε σε μικρά κατάλοιπα, παρατηρούμε μεγάλες συστάδες από μικρά αμινοξέα σε σειρά, όπως φαίνεται συγκεκριμένα στα πρότυπα: "11111101001", "10111100011" και "10111110100". Γενικά μικρά αμινοξέα έχει αναφερθεί στη βιβλιογραφία ότι παρατηρούνται στην γειτονιά *cis*-Pro πεπτιδικών δεσμών αφού το μικρό τους μέγεθος δεν θέτει εμπόδιο σε πιθανή στρέψη του πεπτιδικού δεσμού [11]. Τα καλύτερα πρότυπα για

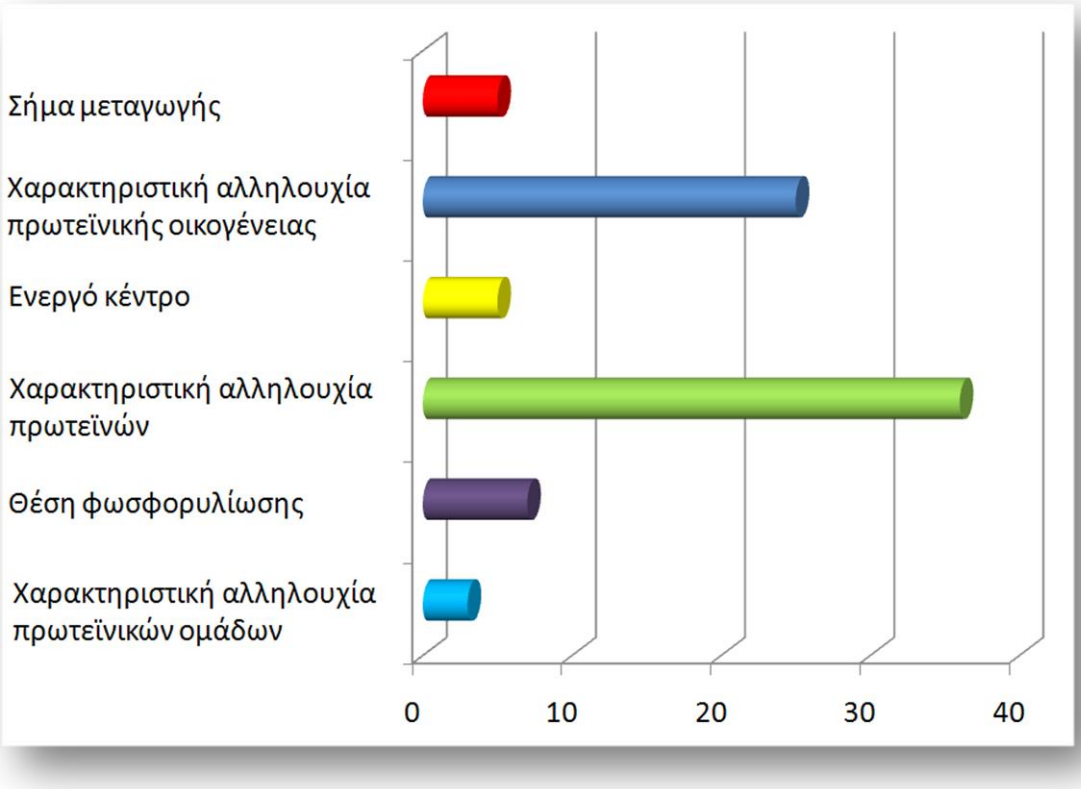
κάθε φυσικοχημική ιδιότητα απεικονίζονται στην Εικόνα 30 με τη χρήση ακολουθιακών λογοτύπων.




**Εικόνα 30:** Ακολουθιακά λογότυπα για τα πιο αντιπροσωπευτικά πρότυπα λαμβάνοντας υπόψη τις φυσικοχημικές ιδιότητες των αμινοξέων.

Στη συνέχεια αφού εξάγουμε τα ακολουθιακά πρότυπα που περιγράφουν την γειτονιά των *cis*-Pro πεπτιδικών δεσμών, επιχειρούμε να εντοπίσουμε τις λειτουργικές κατηγορίες στις οποίες ανήκουν. Λόγω της μεγάλης σπανιότητας των *cis* πεπτιδικών δεσμών, η χρήση ακολουθιακών προτύπων για αυτό τον σκοπό είναι ιδιαιτέρως πρόσφορη αφού τα πρότυπα γενικεύουν την υπάρχουσα γνώση. Στην Εικόνα 31 απεικονίζονται οι λειτουργικές ομάδες με τις οποίες σχετίζονται συχνά οι *cis*-Pro πεπτιδικοί δεσμοί. Ο οριζόντιος άξονας

αναφέρεται στο συνολικό πλήθος των *cis*-Pro προτύπων που ανέσυραν την συγκεκριμένη λειτουργική ομάδα της PROSITE [92]. Παρατηρούμε σημαντική ροπή προς χαρακτηριστικές αλληλουχίες μεμονωμένων πρωτεϊνών (protein signature) καθώς και πρωτεϊνικών οικογενειών (family signature), ενώ σχετικά χαμηλότερη είναι συχνότητα με την οποία ανασύρονται λειτουργικές περιοχές που αφορούν σήματα μεταγωγής (targeting signals), ενεργά κέντρα (active sites), θέσεις φωσφορυλίωσης (phosphorylation sites) και χαρακτηριστικές αλληλουχίες πρωτεϊνικών ομάδων (domain signatures). Αξίζει να σημειωθεί ότι η προτεινόμενη μεθοδολογία κατάδειξε κάποιες λειτουργίες των *cis*-Pro περιοχών που είχαν ήδη επιβεβαιωθεί πειραματικά στην βιβλιογραφία, αλλά ανακάλυψε και πλήθος νέων που χρήζουν περαιτέρω διερεύνησης.




**Εικόνα 31:** Λειτουργικές ομάδες με τις οποίες σχετίζονται τα *cis*-Pro ακολουθιακά πρότυπα.

Εκ των πραγμάτων βέβαια αναμέναμε τα πρότυπα που περιγράφουν και χαρακτηρίζουν *cis*-Pro πρωτεϊνικές περιοχές να σχετίζονται με σημαντικές κυτταρικές λειτουργίες, όπως έχει διαπιστωθεί στην βιβλιογραφία πειραματικά [8, 9, 11, 15], και

αποδείχτηκε συστηματικά πλέον από την παρούσα μεθοδολογική ανάλυση. Η λειτουργική συσχέτιση των *cis*-Pro πεπτιδικών δεσμών μπορεί να συναχθεί και από το γεγονός ότι οι πρωτεϊνικές αλληλουχίες που φέρουν *cis*-Pro πεπτιδικούς δεσμούς έχει αποδειχτεί ότι διατηρούνται εξελικτικά περισσότερο από τις γειτονικές περιοχές τους που φέρουν αμιγώς *trans* πεπτιδικούς δεσμούς, οι οποίες διατηρούνται εξελικτικά εξίσου με ολόκληρη την υπόλοιπη πρωτεϊνική ακολουθία [15].

## *Cis-nonPro πεπτιδικοί δεσμοί*

Όπως αναφέρθηκε και προηγουμένως, λόγω της μεγάλης σπανιότητας των *cis*-nonPro πεπτιδικών δεσμών, εφαρμόστηκαν κάποια επιπλέον βήματα ώστε να εξαχθεί όσο το δυνατόν πιο περιορισμένος αριθμός από σαφή και συγκεκριμένα ακολουθιακά πρότυπα που περιγράφουν την γειτονιά των *cis*-nonPro πεπτιδικών δεσμών. Η έξοδος από τον αλγόριθμο TEIRESIAS [89] περιλαμβάνει 4815 ακολουθιακά πρότυπα όταν εκτελούμε ακριβή εξαγωγή προτύπων, ενώ στην περίπτωση που λαμβάνονται υπόψη χημικές και δομικές ομαδοποιήσεις των αμινοξέων η έξοδος αποτελείται από 38904 και 32812 πρότυπα, αντίστοιχα. Στο επόμενο βήμα της μεθοδολογίας, τα εξαχθέντα πρότυπα αξιολογούνται με βάση το σκορ, όπως αυτό ορίστηκε στην εξίσωση 11, για την περίπτωση των X-nonPro πεπτιδικών δεσμών. Το σκορ λαμβάνει τιμές στο διάστημα [0,1], όπου τιμές κοντά στο 0.50 αντιστοιχούν σε πρότυπα που δεν εμφανίζουν σαφή προτίμηση προς κάποιο τύπο πεπτιδικού δεσμού, ενώ τιμές μεγαλύτερες από 0.50 υποδηλώνουν συσχέτιση με τους *cis*-nonPro πεπτιδικούς δεσμούς. Το κατώφλι που θέσαμε για να λάβουμε όσο το δυνατόν πιο αντιπροσωπευτικά πρότυπα των *cis*-nonPro δεσμών είναι 0.90, απαλείφοντας κατ' αυτόν τον τρόπο πρότυπα που εμφανίζουν συχνά στην γειτονιά *trans*-nonPro πεπτιδικών δεσμών. Η εφαρμογή της τιμής κατωφλίου οδηγεί σε σημαντική μείωση των αρχικά εξαχθέντων ακολουθιακών προτύπων, και συγκεκριμένα για κάθε τύπο εξαγωγής προτύπων, δηλαδή ακριβή εξαγωγή, εξαγωγή με χρήση χημικών ομαδοποιήσεων των αμινοξέων και εξαγωγή με χρήση δομικών ομαδοποιήσεων των αμινοξέων, καταλήγουμε σε 1622, 8251 και 7347 πρότυπα αντίστοιχα. Τα πρότυπα που προκύπτουν και από αυτό το βήμα παρουσιάζουν σημαντική ομοιότητα μεταξύ τους, και άρα πολλά πρότυπα ενδέχεται να ταιριάζουν με την ίδια περιοχή στην πρωτεϊνική ακολουθία. Για τον σκοπό αυτό εφαρμόζουμε έναν αλγόριθμο ο οποίος για κάθε ομάδα παρόμοιων προτύπων που

ταιριάζουν με την ίδια περιοχή, διατηρεί μόνο τον καλύτερο αντιπρόσωπο. Συγκεκριμένα για να το πετύχει αυτό ταξινομεί τα πρότυπα με βάση το σκορ σε φθίνουσα σειρά και εκ περιτροπής διατρέχει με κάθε πρότυπο τα σύνολα με τις πρωτεϊνικές περιοχές των *cis*-nonPro ($D_{cis\text{-nonPro}}$) και *trans*-nonPro ($D_{trans\text{-nonPro}}$) πεπτιδικών δεσμών, ούτως ώστε να επιτύχει την πλήρη κάλυψη των *cis*-nonPro δεσμών ανασύροντας όσο το δυνατόν μικρότερο αριθμό *trans*-nonPro δεσμών (FDR). Στον Πίνακας 10 παραθέτουμε τα σύνολα των προτύπων μετά από κάθε βήμα καθώς και την κάλυψη και πιθανά σφάλματα που επισύρει το κάθε σύνολο ακολουθιακών προτύπων.

**Πίνακας 10:** Επισκόπηση των προτύπων σε κάθε βήμα της προτεινόμενης μεθοδολογίας.

| **Ακριβής εξαγωγή προτύπων** | | | |
|---|---|---|---|
| | *TEIRESIAS* | *Score > 0.90* | *Τελικό σύνολο* |
| ***Αριθμός προτύπων*** | 4815 | 1622 | 231 |
| ***Coverage (%)*** | 100% | - | 100 |
| ***FDR (%)*** | 3.58 | - | 0.25 |
| **Χημικές ομαδοποιήσεις αμινοξέων** | | | |
| | *TEIRESIAS* | *Score > 0.90* | *Τελικό σύνολο* |
| ***Αριθμός προτύπων*** | 38904 | 8251 | 235 |
| ***Coverage (%)*** | 100 | - | 100 |
| ***FDR (%)*** | 6.79 | - | 0.03 |
| **Δομικές ομαδοποιήσεις αμινοξέων** | | | |
| | *TEIRESIAS* | *Score > 0.90* | *Τελικό σύνολο* |
| ***Αριθμός προτύπων*** | 32812 | 7347 | 225 |
| ***Coverage (%)*** | 100 | - | 100 |
| ***FDR (%)*** | 6.69 | - | 0.02 |

Παρατηρούμε ότι στις αρχικές λίστες προτύπων όπως προέκυψαν από την εφαρμογή του αλγορίθμου TEIRESIAS, παρόλο που εξασφαλίζεται 100% κάλυψη, ανασύρεται ιδιαίτερα υψηλός αριθμός *trans*-nonPro δεσμών. Ιδίως αν αναλογιστούμε το τεράστιο πλήθος των *trans*-nonPro πεπτιδικών δεσμών στο αρχικό σύνολο δεδομένων (685716) τότε εύκολα συμπεραίνουμε ότι η διακριτική αξία των εξαχθέντων προτύπων είναι σχεδόν αμελητέα. Όμως, η κατάσταση αλλάζει δραματικά μετά την εφαρμογή των επόμενων βημάτων της προτεινόμενης μεθοδολογίας. Συγκεκριμένα, το τελικό σύνολο προτύπων

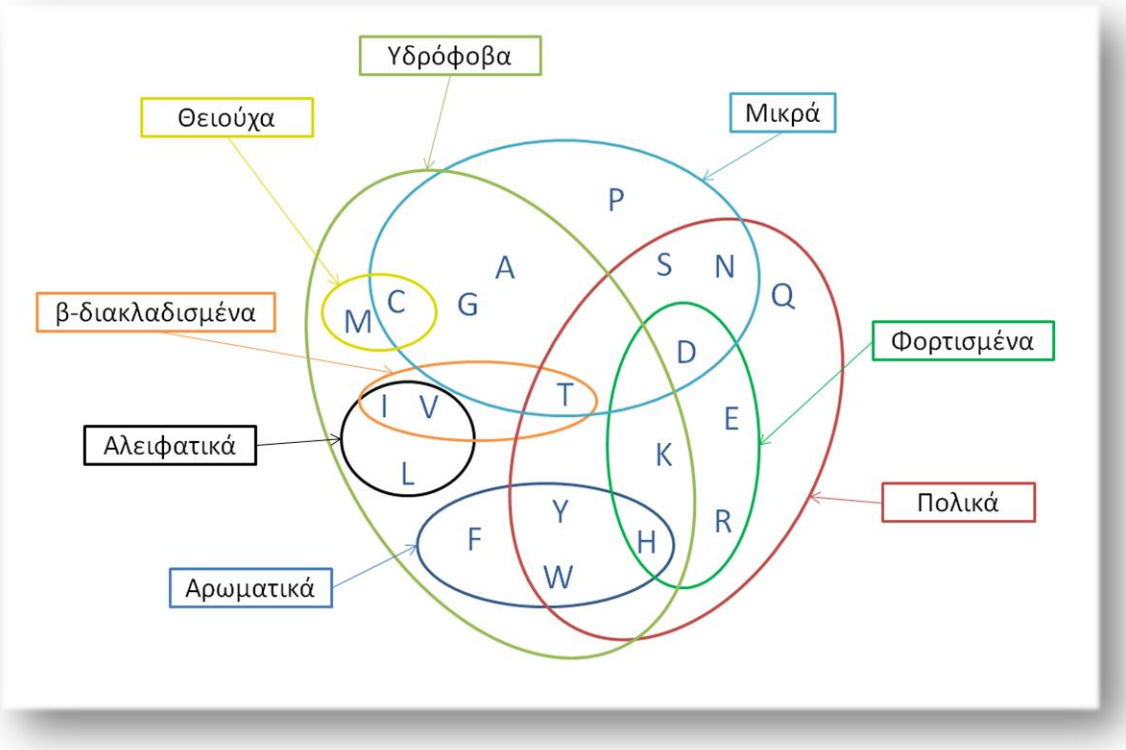
περιέχει μεν πολύ λιγότερα πρότυπα, η κάλυψη όμως είναι επίσης 100% και ο αριθμός σφαλμάτων είναι πλέον αμελητέος. Βέβαια και πάλι όσον αφορά την ακριβή εξαγωγή προτύπων, στο τελικό σύνολο που προκύπτει το FDR παρόλο που αρχικά δείχνει σχετικά μικρό (FDR=0.25%), σε απόλυτους αριθμούς αντιστοιχεί σε 1714 *trans*-nonPro πεπτιδικούς δεσμούς. Όμως στην περίπτωση των χημικών και δομικών ομαδοποιήσεων, τα σύνολα των ακολουθιακών προτύπων που προκύπτουν είναι σαφώς πιο συγκεκριμένα και επιλεκτικά αφού επιτυγχάνουν κάλυψη 100% διατηρώντας παράλληλα το FDR σε πολύ χαμηλά επίπεδα, και ποσοστιαία αλλά και σε απόλυτους αριθμούς.

Ο Πίνακας 11 περιέχει ένα αντιπροσωπευτικό υποσύνολο των ακολουθιακών προτύπων που περιγράφουν την γειτονιά των *cis*-nonPro πεπτιδικών δεσμών. Συγκεκριμένα περιέχονται τα 20 πρότυπα από κάθε τύπο εξαγωγής προτύπων που σημείωσαν τα υψηλότερα σκορ βάσει της εξίσωσης 11. Επιπρόσθετα για κάθε πρότυπο παρατίθεται ένα μέτρο σημαντικότητας που ουσιαστικά δείχνει την πιθανότητα το συγκεκριμένο πρότυπο να έχει προκύψει τυχαία. Τα υπογραμμισμένα αμινοξέα είναι αυτά που φέρουν τον *cis*-nonPro πεπτιδικό δεσμό στις αρχικές πρωτεϊνικές ακολουθίες από τις οποίες προέκυψαν.

**Πίνακας 11:** Τα 20 πιο αντιπροσωπευτικά πρότυπα των *cis*-nonPro πεπτιδικών δεσμών.

| **Ακριβής εξαγωγή προτύπων** | | | **Χημικές ομαδοποιήσεις αμινοξέων** | | | **Δομικές ομαδοποιήσεις αμινοξέων** | | |
|---|---|---|---|---|---|---|---|---|
| **Πρότυπο** | **Σκορ** | **Σημαντι-κοτητα** | **Πρότυπο** | **Σκορ** | **Σημαντι-κοτητα** | **Πρότυπο** | **Σκορ** | **Σημαντι-κοτητα** |
| KPGKGRRK | 1 | -37.67 | KPGKGRRK | 1 | -37.67 | KPGKGRRK | 1 | -37.67 |
| EDGTKEPLL | 1 | -42.25 | G.[AG][DE].K..SL | 1 | -23.91 | S.S[ITV]H..N | 1 | -19.73 |
| HAESGEYG L | 1 | -44.51 | [ILMV][ILMV].[AG].D.AT | 1 | -21.70 | E.V[DLN].[KMR]P | 1 | -18.28 |
| LGTVINQL | 1 | -36.61 | G[AG].[DE][ILMV]K.[ILM V]S[ILMV] | 1 | -31.33 | MLQ...[ITV].[KMR] | 1 | -19.03 |
| ADEAT | 1 | -20.01 | G..[FY]W[QN]..D[ST] | 1 | -25.89 | [EQ].GYT.R | 1 | -20.32 |
| ALNALKLV T | 1 | -41.86 | LGTVINQL | 1 | -36.61 | [KMR]..QGY..R | 1 | -20.23 |
| YFT...I | 1 | -14.57 | EDGTKEPLL | 1 | -42.25 | EDGTKEPLL | 1 | -42.25 |
| CLA..VN | 1 | -20.96 | GA.D[DE]A[ST] | 1 | -24.88 | LGTVINQL | 1 | -36.61 |
| R..DP,...VV | 1 | -20.01 | ADEAT | 1 | -20.01 | HAESGEYGL | 1 | -44.51 |
| H.YSQ | 1 | -15.76 | [ST]..A[DE]G.A | 1 | -18.38 | ADEAT | 1 | -20.01 |
| VYL..L...Y | 1 | -20.29 | [AG][ILMV]..L[KR]L..D | 1 | -20.11 | ALNALKLVT | 1 | -41.86 |
| NAW..D | 1 | -15.89 | T.R.E..A.[ILMV] | 1 | -18.48 | GG...[KMR]M..L | 1 | -19.48 |
| A...KHF.G.G | 1 | -26.56 | [ST].LN.LK[ILMV] | 1 | -23.04 | [DLN]L.EL..E[EQ] | 1 | -20.75 |
| L,.SRGF | 1 | -19.77 | [AG].HF[ILMV]GD | 1 | -24.32 | [EQ]..P.,[FHWY]P.E | 1 | -19.33 |
| REPDP | 1 | -21.25 | MLQ[QN]...[ILMV][KR] | 1 | -23.90 | QL...N.L.[KMR][DLN] | 1 | -23.32 |
| G.MFW | 1 | -16.96 | [AG]KHF.G.G | 1 | -25.89 | A[FHWY].[FHWY]E...EN | 1 | -24.56 |
| L.G..VVP..S | 1 | -24.86 | HAESGEYGL | 1 | -44.51 | M[FHWY].[EQ][FHWY]. D[ITV] | 1 | -23.90 |
| MDHSNY | 1 | -28.88 | [KR][ILMV].P.[ILMV][ST]. .[FY] | 1 | -21.32 | [ITV]..G[ITV].T.[ITV].V | 1 | -22.12 |
| VL.G..TNI | 1 | -25.03 | [AG][ST].D..GP | 1 | -18.86 | [KMR]Y...N.V[CS] | 1 | -19.41 |
| L..A..V.SS | 1 | -18.77 | G...M.C..I | 1 | -16.47 | [FHWY]..KG.[ITV].R[IT V] | 1 | -23.29 |

Παρατηρούμε από τον Πίνακας 11 ότι τα εξαχθέντα πρότυπα σημειώνουν ιδιαιτέρως υψηλά σκορ, διατηρώντας παράλληλα την τιμή της σημαντικότητας σε χαμηλά επίπεδα καταδεικνύοντας ότι η τελική λίστα με τα εξαχθέντα πρότυπα είναι και χαρακτηριστική των *cis*-nonPro πεπτιδικών δεσμών αλλά απαρτίζεται από πρότυπα που η πιθανότητα να βρεθούν τυχαία σε ένα μεγάλο σύνολο πρωτεϊνικών ακολουθιών είναι πολύ μικρή. Ειδικά στην περίπτωση που λαμβάνονται υπόψη οι χημικές και δομικές ομαδοποιήσεις των αμινοξέων τα σκορ που επιτυγχάνονται είναι σε κάθε περίπτωση μεγαλύτερα από 0.99 υπογραμμίζοντας την μεγάλη συσχέτιση των εξαχθέντων προτύπων με *cis*-nonPro δεσμούς. Για την διευκόλυνση της συζήτησης γύρω από τα εξαχθέντα πρότυπα, παραθέτουμε στην Εικόνα 32, την κατανομή των 20 αμινοξέων σε 8 σημαντικές φυσικοχημικές ιδιότητες με βιολογική σημασία.




**Εικόνα 32:** Ομαδοποιήσεις των αμινοξέων με βάση 8 σημαντικές φυσικοχημικές ιδιότητες.

Από την παρατήρηση και αξιολόγηση των εξαχθέντων προτύπων, πολλά σημαντικά νέα συμπεράσματα μπορούν να συναχθούν, αλλά παράλληλα επιβεβαιώνεται η υπάρχουσα γνώση, ενισχύοντας έτσι την αξιοπιστία της προτεινόμενης μεθοδολογικής ανάλυσης στο σύνολό της. Στον Πίνακας 12, παραθέτουμε την ποσοστιαία σύσταση των προτύπων από κάθε αμινοξύ καθώς και κάθε ομάδα αμινοξέων, στην περίπτωση που χρησιμοποιήθηκαν οι χημικές και δομικές ομαδοποιήσεις αυτών κατά την εξαγωγή των προτύπων. Για να υπολογίσουμε την συχνότητα εμφάνισης του κάθε αμινοξέος στο σύνολο των προτύπων διαιρέσαμε το πλήθος των εμφανίσεων του, δια το σύνολο των σταθερών χαρακτήρων (όλοι οι χαρακτήρες εκτός από την τελεία ".") που περιέχονται σε όλα τα πρότυπα για κάθε τύπο εξαγωγής προτύπων.

**Πίνακας 12:** Συχνότητα εμφάνισης των αμινοξέων στα εξαχθέντα πρότυπα.

| | **Συχνότητα αμινοξέων (%)** | | |
|---|---|---|---|
| | Ακριβής εξαγωγή προτύπων | Χημικές ομαδοποιήσεις αμινοξέων | Δομικές ομαδοποιήσεις αμινοξέων |
| **A** | 8 | 5 | 6 |
| **R** | 3 | 2 | 2 |
| **N** | 4 | 3 | 3 |
| **D** | 6 | 4 | 4 |
| **C** | 1 | 1 | 1 |
| **E** | 5 | 4 | 3 |
| **Q** | 2 | 1 | 2 |
| **G** | 13 | 9 | 10 |
| **H** | 3 | 2 | 2 |
| **I** | 5 | 3 | 3 |
| **L** | 9 | 6 | 6 |
| **K** | 4 | 3 | 3 |
| **M** | 2 | 1 | 1 |
| **F** | 4 | 3 | 2 |
| **P** | 6 | 4 | 4 |
| **S** | 7 | 5 | 5 |
| **T** | 6 | 3 | 3 |
| **W** | 2 | 2 | 1 |
| **Y** | 4 | 3 | 2 |
| **V** | 7 | 4 | 4 |
| **AG/ITV** | - | 6 | 10 |
| **DE/DLN** | - | 3 | 7 |
| **FY/KMR** | - | 3 | 5 |
| **QN/EQ** | - | 2 | 2 |
| **ILMV/FHWY** | - | 12 | 7 |
| **ST/CS** | - | 5 | 2 |
| **KR/-** | - | 3 | - |

Όσον αφορά την ακριβή εξαγωγή προτύπων, παρατηρούμε ότι τα ακολουθιακά πρότυπα που προκύπτουν, περιέχουν σε μεγάλο βαθμό γλυκίνη, αλλά επίσης δείχνουν σημαντική προτίμηση προς τα αμινοξέα λευκίνη, αλανίνη, σερίνη και βαλίνη. Το ίδιο

ισχύει και για τους άλλους δύο τύπους εξαγωγής προτύπων, όπου όμως παρατηρείται παρόμοια συχνότητα εμφάνισης με την βαλίνη για τα αμινοξέα προλίνη, ασπαρτικό οξύ και γλουταμινικό οξύ. Τα εν λόγω κατάλοιπα, έχει διαπιστωθεί και στη βιβλιογραφία ότι εντοπίζονται συχνά στην γειτονιά *cis*-nonPro πεπτιδικών δεσμών, ιδίως η γλυκίνη και η αλανίνη [11]. Αξίζει να σημειωθεί ότι εκτός από τα αμινοξέα με μεγάλη συχνότητα εμφάνισης στο σύνολο των εξαχθέντων προτύπων, και τα αμινοξέα που εντοπίζονται σπάνια στα πρότυπα είναι επίσης κοινά και στους τρείς τύπους εξαγωγής προτύπων. Συγκεκριμένα, η κυστεΐνη παρουσιάζει το μικρότερο ποσοστό αντιπροσώπευσης στα *cis*-nonPro πρότυπα (1% σε όλους τους τύπους εξαγωγής προτύπων), και ακολουθούν η μεθειονίνη, η γλουταμίνη και η τρυπτοφάνη. Ενδιαφέρον παρουσιάζει το γεγονός, ότι και τα τέσσερα κατάλοιπα που σημειώνουν μεγάλη συχνότητα εμφάνισης, είναι μικρά σε όγκο, εξυπηρετώντας τον σχηματισμό *cis* διαμόρφωσης λόγω της μικρής στερεοχημικής παρεμπόδισης. Ακόμη, παρατηρούμε ότι η γλυκίνη, η αλανίνη και η βαλίνη είναι υδροφοβικά αμινοξέα, καταδεικνύοντας μια σχετική προτίμηση των *cis*-nonPro διαμορφώσεων προς τα μικρά και υδροφοβικά κατάλοιπα. Εξαίρεση σε αυτόν τον κανόνα αποτελεί η κυστεΐνη, που είναι μικρό και υδρόφοβο αμινοξύ όμως εντούτοις συναντάται πολύ σπάνια στα εξαχθέντα πρότυπα. Μια πιθανή εξήγηση για αυτήν την παρατήρηση είναι το γεγονός ότι η κυστεΐνη περιέχει την έντονα δραστική θειική ομάδα, που την διακρίνει και από τα υπόλοιπα τρία μικρά και υδρόφοβα αμινοξέα (γλυκίνη, αλανίνη και βαλίνη). Εκτός από την κυστεΐνη, και η μεθειονίνη ανήκει στα θειούχα αμινοξέα και όπως αναφέρθηκε προηγουμένως επίσης εντοπίζεται με πολύ χαμηλή συχνότητα στα *cis*-nonPro ακολουθιακά πρότυπα. Συνεπώς, παρατηρούμε ότι η ιδιαιτέρως δραστική ομάδα του θείου, ασκεί αρνητική επίδραση στον σχηματισμό και σταθεροποίηση της *cis*-nonPro διαμόρφωσης. Επίσης, παρατηρούμε μικρά ποσοστά εμφάνισης των αμινοξέων γλουταμίνη και τρυπτοφάνη, γεγονός που δείχνει την αρνητική συσχέτιση μεταξύ των *cis*-nonPro πεπτιδικών δεσμών και ογκωδών, πολικών καταλοίπων.

Στην περίπτωση που λαμβάνονται υπόψη οι χημικές ομαδοποιήσεις των αμινοξέων, η ομάδα [ILMV] παρουσιάζει την μεγαλύτερη συχνότητα εμφάνισης, με την [AG] να ακολουθεί. αντιθέτως, η ομάδα των αμινοξέων [QN] παρουσιάζει μια από τις μικρότερες συχνότητες αντιπροσώπευσης. Κρίνοντας από την επικρατούσα ομάδα αμινοξέων [ILMV], και ιδίως από τα αμινοξέα ισολευκίνη, λευκίνη και βαλίνη, παρατηρούμε μια σημαντική συσχέτιση των *cis*-nonPro προτύπων με αλειφατικά αμινοξέα. Η υψηλή συχνότητα που παρουσιάζει η ομάδα [AG] συνάδει με τις προηγούμενες παρατηρήσεις

που αφορούν μεμονωμένα τα αμινοξέα αλανίνη και γλυκίνη, και ενισχύεται περαιτέρω λόγω της συχνής εμφάνισής του συνδυασμού τους. Σε ό,τι αφορά την δομική ομαδοποίηση των αμινοξέων, η ομάδα καταλοίπων [ITV] εμφανίζει , μαζί με την γλυκίνη, την υψηλότερη συχνότητα εμφάνισης, ανάμεσα στα πρότυπα των *cis*-nonPro πεπτιδικών δεσμών. Οι ομάδες αμινοξέων [DLN] και [FHWY] επίσης παρατηρούνται σε ένα σημαντικό βαθμό στα εξαχθέντα πρότυπα. Επίσης, η συχνή εμφάνιση της ομάδας [ITV] καταδεικνύει την σημαντική συσχέτιση μεταξύ των β-διακλαδισμένων αμινοξέων και των *cis*-nonPro πεπτιδικών δεσμών, η οποία έχει επίσης διαπιστωθεί και στην βιβλιογραφία [11], ιδίως σε θέσεις που προηγούνται του πεπτιδικού δεσμού, σταθεροποιώντας κατ' αυτόν τον τρόπο την *cis* διαμόρφωση. Στην ίδια εργασία διαπιστώθηκε επίσης υψηλή συχνότητα μικρών και πολικών αμινοξέων (σερίνη, ασπαραγίνη και ασπαρτικό οξύ), παρατήρηση που μόνο μερικώς επιβεβαιώνεται στο σύνολο των ακολουθιακών προτύπων που εξήχθηκαν στην παρούσα μελέτη. Συγκεκριμένα, όταν παρατηρούμε μεμονωμένα τα αμινοξέα η σερίνη είναι το μόνο κατάλοιπο που παρουσιάζει υψηλή συχνότητα στα *cis*-nonPro πρότυπα, σε αντίθεση με το ασπαρτικό οξύ και την ασπαραγίνη που καταλαμβάνουν την $8^{η}$ και $13^{η}$ θέση αντίστοιχα, κατά την ακριβή εξαγωγή προτύπων, ενώ σε παρόμοια επίπεδα βρίσκονται οι αντίστοιχες συχνότητες όταν λογίζονται η χημική και δομική ομαδοποίηση των αμινοξέων. Η ασυμφωνία βέβαια αυτή γεφυρώνεται από την εκπροσώπηση την ασπαραγίνης και του ασπαρτικού οξέως μέσα από την υψηλή συχνότητα της ομάδας [DLN]. Μια ακόμη ομάδα αμινοξέων που παρουσιάζει ιδιαιτέρως υψηλή συχνότητα είναι η [FHWY] που περιλαμβάνει αρωματικά κατάλοιπα. Αξίζει να σημειωθεί ότι τα αμινοξέα που απαρτίζουν την παραπάνω ομάδα όταν λογίζονται μεμονωμένα δεν παρουσιάζουν σημαντική εκπροσώπηση στα εξαχθέντα πρότυπα. Η έντονη αυτή συσχέτιση των *cis*-nonPro πεπτιδικών δεσμών με τα αρωματικά αμινοξέα, έχει επίσης παρατηρηθεί στη βιβλιογραφία [11, 97] όπου έχει διατυπωθεί ο σταθεροποιητικός ρόλος που ασκούν στην *cis* διαμόρφωση.

Επίσης, εκτιμούμε την συχνότητα του κάθε αμινοξέος ως συστατικό του πεπτιδικού δεσμού, είτε ως το πρώτο είτε ως το δεύτερο αμινοξύ που απαρτίζει τον υπό μελέτη πεπτιδικό δεσμό. Οι συγκεκριμένες συχνότητες παρατίθενται στον Πίνακας 13.

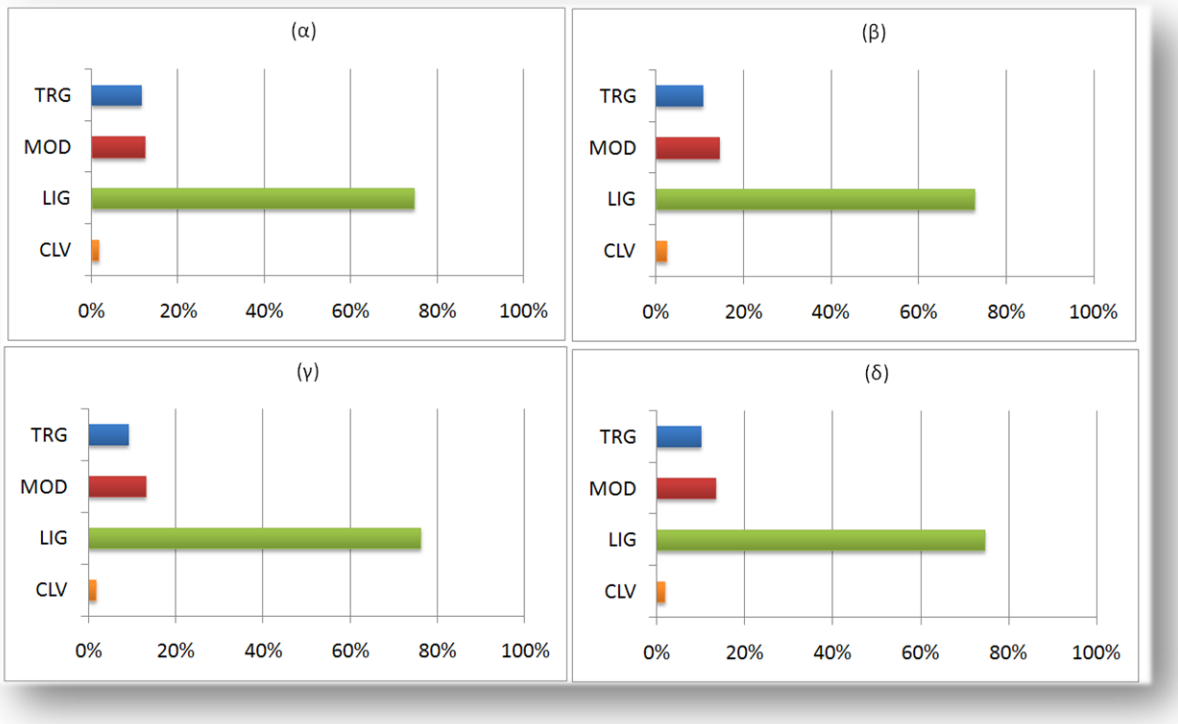
**Πίνακας 13:** Κατανομή των αμινοξέων στην πρώτη και δεύτερη θέση του πεπτιδικού δεσμού.

| | **Πρώτο αμινοξύ του πεπτιδικού δεσμού** | **Δεύτερο αμινοξύ του πεπτιδικού δεσμού** | **Συχνότητα εμφάνισης αμινοξέων στην πρώτη θέση του πεπτιδικού δεσμού (%)** | **Συχνότητα εμφάνισης αμινοξέων στη δεύτερη θέση του πεπτιδικού δεσμού (%)** |
|---|---|---|---|---|
| **A** | 22 | 23 | 7 | 7 |
| **R** | **7** | **15** | **2** | **5** |
| **N** | 14 | 15 | 4 | 5 |
| **D** | 23 | 26 | 7 | 8 |
| **C** | 4 | 5 | 1 | 2 |
| **E** | 20 | 27 | 6 | 8 |
| **Q** | 14 | 9 | 4 | 3 |
| **G** | 62 | 52 | 19 | 16 |
| **H** | 7 | 7 | 2 | 2 |
| **I** | 6 | 7 | 2 | 2 |
| **L** | 11 | 9 | 3 | 3 |
| **K** | 16 | 17 | 5 | 5 |
| **M** | 5 | 4 | 2 | 1 |
| **F** | 12 | 15 | 4 | 5 |
| **P** | 24 | 0 | 8 | 0 |
| **S** | **13** | **28** | **4** | **9** |
| **T** | **12** | **22** | **4** | **7** |
| **W** | **16** | **2** | **5** | **1** |
| **Y** | 14 | 16 | 4 | 5 |
| **V** | 16 | 19 | 5 | 6 |

Αρχικά, παρατηρούμε το δεύτερο συστατικό του πεπτιδικού δεσμού να καταλαμβάνεται με μεγάλη συχνότητα από συγκεκριμένα κατάλοιπα, και συγκεκριμένα από τα αμινοξέα: γλυκίνη [9],σερίνη, ασπαρτικό οξύ και γλουταμινικό οξύ. Η υψηλή συχνότητα που σημειώνει το ασπαρτικό οξύ και το γλουταμινικό οξύ στη δεύτερη θέση του πεπτιδικού δεσμού καταδεικνύει μια υψηλή τάση η συγκεκριμένη θέση να καταλαμβάνεται από αρνητικά φορτισμένα κατάλοιπα. Ελαφρώς μικρότερη συχνότητα όσον αφορά την ίδια θέση παρατηρείται για τα αμινοξέα αλανίνη και θρεονίνη. Επίσης, υπάρχουν κάποια αμινοξέα που εντοπίζονται ως συστατικά των *cis*-nonPro πεπτιδικών

δεσμών με ιδιαιτέρως μικρή συχνότητα, όπως η μεθειονίνη, η τρυπτοφάνη, η κυστεΐνη, η ιστιδίνη και η ισολευκίνη. Αξίζει να παρατηρήσουμε ότι κάποια αμινοξέα παρουσιάζονται με έντονα άνιση συχνότητα ως πρώτο και δεύτερο συστατικό του πεπτιδικού δεσμού. Σε αυτήν την κατηγορία ανήκουν τα αμινοξέα αργινίνη, γλουταμίνη, σερίνη, θρεονίνη και κυρίως η τρυπτοφάνη, και παρουσιάζονται με έντονα γράμματα στον Πίνακας 13.

Στη συνέχεια, το σύνολο των εξαχθέντων προτύπων που αφορούν τους *cis*-nonPro πεπτιδικούς δεσμούς αντιπαραβάλλεται με τα ακολουθιακά πρότυπα που είναι καταχωρημένα στην βάση ELM [94]. Με αυτόν τον τρόπο αναζητούμε ομοιότητες μεταξύ των *cis*-nonPro προτύπων και προτύπων των οποίων η λειτουργία έχει εξακριβωθεί πειραματικά. Η αντιπαραβολή αυτή σκοπό έχει να αναδείξει κάποιες ήδη γνωστές λειτουργικές συσχετίσεις των *cis*-nonPro πεπτιδικών δεσμών, αλλά πρώτιστα να ανακαλύψει πλήθος νέων. Όπως περιγράφτηκε και παραπάνω, τα πρότυπα που προέκυψαν και με τους τρείς τρόπους εξαγωγής προτύπων (δηλαδή ακριβής εξαγωγή προτύπων, χημικές ομαδοποιήσεις των αμινοξέων και δομικές ομαδοποιήσεις αμινοξέων), συγκρίνονται με τα πρότυπα της βάσης ELM χρησιμοποιώντας τον αλγόριθμο Comparomotif [95], ο οποίος ποσοτικοποιεί τον επικάλυψη μεταξύ δύο ακολουθιακών προτύπων. Τα αποτελέσματα που λαμβάνονται προβάλλονται στην Εικόνα 33, αμέσως πιο κάτω. Επιπρόσθετα, για να αναδείξουμε τις λειτουργικές ομάδες σχετίζονται στατιστικά σημαντικά με τα *cis*-nonPro ακολουθιακά πρότυπα, εκτελούμε $x^2$-test ανάμεσα στην παρατηρούμενη συχνότητα (όπως αυτή υπολογίζεται από το Comparimotif) και την συχνότητα που ενέχει η κάθε λειτουργική ομάδα στο σύνολο της βάσης ELM. Ο κάθετος άξονας περιέχει τις τέσσερεις λειτουργικές ομάδες που περιλαμβάνονται στην ELM, δηλαδή, i) στόχευση/εντόπιση (targeting/localization: TRG), ii) μετα-μεταγραφικές τροποποιήσεις (post-translational modifications: MOD), iii) πρόσδεση/υποκατάσταση (ligand/binding: LIG) και iv) θέσεις αποκοπής (cleavage: CLV), και ο οριζόντιος άξονας δείχνει την συχνότητα εκπροσώπησης της κάθε λειτουργικής ομάδας στο σύνολο των εξαχθέντων *cis*-nonPro προτύπων.




**Εικόνα 33:** Λειτουργικές συσχετίσεις των *cis*-nonPro πεπτιδικών δεσμών, χρησιμοποιώντας (α) ακριβή εξαγωγή προτύπων, (β) χημικές ομαδοποιήσεις αμινοξέων, (γ) δομικές ομαδοποιήσεις των αμινοξέων και (δ) μέσος όρος.

Όπως παρατηρούμε σε όλα τα γραφήματα της Εικόνα 33, η κατανομή των ακολουθιακών προτύπων είναι αρκετά παρόμοια σε για κάθε τύπο εξαγωγής προτύπων. Συγκεκριμένα, η κατηγορία LIG σχετίζεται με το 80% περίπου των εξαχθέντων ακολουθιακών προτύπων, και ακολουθούν οι ομάδες TRG και MOD που αντιπροσωπεύουν η καθεμία περίπου το 15% των λειτουργικών συσχετίσεων των *cis*-nonPro προτύπων. Πολύ μικρό είναι το ποσοστό που καταλαμβάνει η λειτουργική ομάδα CLV, σε κάθε τύπο εξαγωγής προτύπων. Οι τιμές αυτές συγκρίνονται μέσω $x^2$-test με τις συχνότητες αντιπροσώπευσης της κάθε λειτουργικής ομάδας μέσα στη βάση ELM, όπου είναι TRG: 13%, MOD: 25%, LIG: 54% και CLV: 7%. Χρησιμοποιώντας μια αυστηρή τιμή κατωφλίου ($10^{-2}$) για την στατιστική σημαντικότητα του $x^2$-test, αναδεικνύεται στατιστικά σημαντική συσχέτιση των εξαχθέντων προτύπων με την λειτουργική ομάδα LIG, και μάλιστα με κάθε τύπο εξαγωγής προτύπων. Η σημαντική αυτή συσχέτιση ενισχύεται περαιτέρω από την σχετική βιβλιογραφία [98] όπου έχει αναφερθεί η στενή σχέση των *cis*-nonPro πεπτιδικών δεσμών με την πρόσδεση μορίων και την εντόπιση των καταλοίπων στο ενεργό κέντρο ενζύμων. Επίσης, έχει διαπιστωθεί ότι *cis*-nonPro

πεπτιδικοί δεσμοί εντοπίζονται με μεγάλη συχνότητα είτε μέσα, είτε σε γειτνιάζουσες θέσεις με τα ενεργά κέντρα ενζύμων [11, 18, 97], που αποτελούν κατεξοχήν θέσεις πρόσδεσης μορίων.

Ελαφρώς μικρότερη συχνότητα αποδίδεται στην λειτουργική ομάδα που αφορά τις μετα-μεταφραστικές τροποποιήσεις (MOD), όπου πράγματι έχει επίσης διαπιστωθεί πειραματικά την βιβλιογραφία συσχέτιση μεταξύ της επιτέλεσης μετα-μεταφραστικών τροποποιήσεων και της εντόπισης των *cis*-nonPro πεπτιδικών δεσμών [97]. Παρ' όλα αυτά στατιστικά σημαντική συσχέτιση παρατηρείται μόνο κατά την ακριβή εξαγωγή προτύπων, καθιστώντας την συγκεκριμένη συσχέτιση λιγότερο αξιόπιστη. Τέλος, σε ότι αφορά την λειτουργική ομάδα CLV, το ποσοστό των ακολουθιακών προτύπων που ανέσυραν μέλη αυτής της ομάδας από την βάση ELM ήταν ιδιαιτέρως μικρό, γεγονός που δεν συνιστά κάποια έντονη συσχέτιση των *cis*-nonPro δεσμών με λειτουργίες που αφορούν θέσεις αποκοπής (CLV). Πράγματι, ούτε στην βιβλιογραφία έχει διαπιστωθεί η συγκεκριμένη συσχέτιση.

## *3.5 Συμπεράσματα*

Στο παρόν κεφάλαιο παρουσιάσαμε μια μεθοδολογική ανάλυση για την εξαγωγή ακολουθιακών προτύπων σε πρωτεϊνικά μόρια. Συγκεκριμένα, εστιάσαμε στις περιοχές που περιβάλλουν τους *cis*-Pro και *cis*-nonPro πεπτιδικούς δεσμούς, οι οποίοι ενέχουν μεγάλη σημασία για ένα πλήθος σημαντικών λειτουργιών καθώς και για την εν γένει κυτταρική ρύθμιση και λειτουργία. Για κάθε τύπο *cis* (*cis*-Pro και *cis*-nonPro) πεπτιδικού δεσμού εντοπίσαμε ακολουθιακά πρότυπα που παρουσιάζουν μεγάλη συσχέτιση με τις εν λόγω *cis* περιοχές και ταυτόχρονα μικρή συσχέτιση με τις αντίστοιχες *trans* (*trans*-Pro και *trans*-nonPro) περιοχές. Σε ότι αφορά τους *cis*-Pro πεπτιδικούς δεσμούς, ανάμεσα στα εξαχθέντα πρότυπα παρατηρήθηκε σημαντική τάση προς το αμινοξύ γλυκίνη, και εν γένει προς μικρά σε μέγεθος κατάλοιπα. Εξίσου σημαντική τάση σημειώθηκε και για αρωματικά καθώς και για β-διακλαδισμένα αμινοξέα. Όσον για τα πρότυπα που εξήχθησαν από την γειτονιά των *cis*-nonPro πεπτιδικών δεσμών, παρατηρήθηκε μια αρνητική συσχέτιση με θειούχα αμινοξέα, καταδεικνύοντας τον αρνητικό ρόλο της έντονα δραστικής θειικής ομάδας ως προς την σταθεροποίηση των *cis*-nonPro διαμορφώσεων. Όμοια με τους *cis*-Pro πεπτιδικούς δεσμούς, και στην περίπτωση των *cis*-nonPro πεπτιδικών δεσμών, θετική

επίδραση ασκούν τα μικρά σε μέγεθος αμινοξέα, όπως η γλυκίνη, αλανίνη και λευκίνη, ενώ επίσης με μεγάλη συχνότητα παρατηρούνται αρωματικά κατάλοιπα. Επίσης για κάθε τύπο *cis* πεπτιδικού δεσμού χρησιμοποιώντας τα εξαχθέντα πρότυπα εντοπίσαμε συστηματικά πιθανές λειτουργικές συσχετίσεις. Οι *cis*-Pro πεπτιδικοί δεσμοί ανευρίσκονται συχνά σε τμήματα της αλληλουχίας που είναι χαρακτηριστικά μεμονωμένων πρωτεϊνών και οικογενειών πρωτεϊνών καθώς επίσης και σε σημαντικές λειτουργικά περιοχές όπως είναι τα ενεργά κέντρα ενζύμων και μετα-μεταγραφικές τροποποιήσεις (π.χ. θέσεις φωσφορυλίωσης). Οι *cis*-nonPro πεπτιδικοί δεσμοί συχνά εμπλέκονται σε λειτουργίες σχετικές με πρόσδεση/υποκατάσταση, στόχευση καθώς και με πλήθος μετα-μεταγραφικών τροποποιήσεων.

# 4ο ΚΕΦΑΛΑΙΟ: Ρεομορφικές πρωτεΐνες και δίκτυα αλληλεπιδράσεων

## *4.1 Σκοπός*

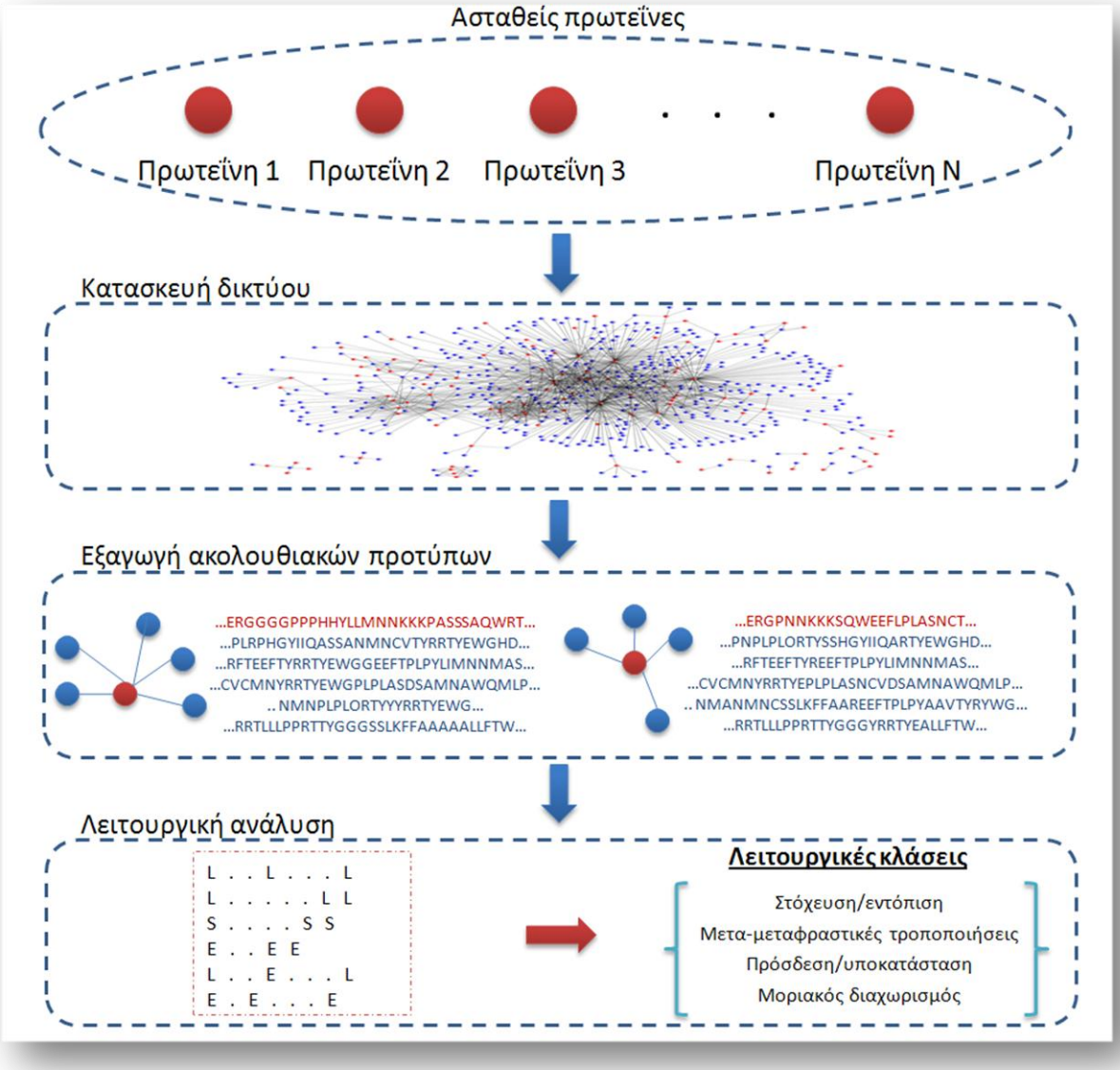
Στο παρόν κεφάλαιο θα μας απασχολήσουν οι ρεομορφικές πρωτεΐνες και ιδίως η συμμετοχή τους σε σημαντικές κυτταρικές λειτουργίες [32-34] αλλά και ο τρόπος με τον οποίον εμπλέκονται και επάγουν ένα πλήθος ασθενειών στον άνθρωπο, όπως ο καρκίνος, καρδιαγγειακές παθήσεις, διαβήτης κ.ά. [31]. Οι ρεομορφικές πρωτεΐνες χαρακτηρίζονται από εγγενή ευμεταβλητότητα της τρισδιάστατης δομής τους [30], γεγονός που τους προσφέρει μεγάλη ευελιξία ως προς τις αλληλεπιδράσεις τους με άλλα πρωτεϊνικά μόρια. Χάρη σε αυτήν την πληθώρα και ποικιλία των αλληλεπιδράσεων και κατά συνέπεια την λειτουργική τους αναγκαιότητα, οι ρεομορφικές πρωτεΐνες έχουν προσελκύσει έντονο ερευνητικό ενδιαφέρον τα τελευταία χρόνια. Ένα μεγάλο μέρος αυτών των αλληλεπιδράσεων που είναι απαραίτητες για την λειτουργικότητα του κυττάρου και εν γένει του οργανισμού, οφείλονται και επάγονται από μικρά τμήματα στην πρωτεϊνική ακολουθία που μπορούν να περιγραφούν με ακολουθιακά πρότυπα (κανονικές εκφράσεις). Η ανακάλυψη αυτών των ακολουθιακών τμημάτων είναι πολύ σημαντική για την κατανόηση, σε επίπεδο πρωτοταγούς δομής, του τρόπου επαγωγής και του μοριακού μηχανισμού που καθοδηγεί μια κυτταρική λειτουργία. Ωστόσο, ο εντοπισμός και η ανακάλυψη ακολουθιακών προτύπων που επάγουν μια λειτουργία, απαιτεί χρονοβόρες πειραματικές διαδικασίες, από τις οποίες συνήθως προκύπτει ένας μικρός αριθμός προτύπων που αφορούν μεμονωμένες λειτουργίες.

Η μεθοδολογία που θα παρουσιάσουμε αμέσως πιο κάτω εκμεταλλεύεται πληροφορίες που αφορούν δίκτυα αλληλεπίδρασης πρωτεϊνών, με σκοπό τον συστηματικό πλέον εντοπισμό ακολουθιακών προτύπων που συμμετέχουν στις αλληλεπιδράσεις των ρεομορφικών πρωτεϊνών με πλήθος άλλων πρωτεϊνικών μορίων. Συγκεκριμένα, δομούμε ένα πρωτεϊνικό δίκτυο γύρω από ρεομορφικές πρωτεΐνες, ανασύροντας αλληλεπιδράσεις που έχουν αποθηκευτεί σε σχετικές βάσεις δεδομένων, και έχουν επιβεβαιωθεί πειραματικά, και στη συνέχεια επιχειρούμε να εντοπίσουμε εκείνα τα ακολουθιακά

πρότυπα που πιθανώς επάγουν την σηματοδότηση είτε και την καθεαυτού επιτέλεση μιας λειτουργίας. Εστιάζουμε αποκλειστικά σε ρεομορφικές πρωτεΐνες, καταρχήν λόγω της μεγάλης λειτουργικής τους σημασίας, αλλά και επειδή έχει αποδειχτεί ότι ρεομορφικές περιοχές φέρουν πληθώρα ακολουθιακών προτύπων με έντονη λειτουργική δράση [99]. Πρωταρχικός μας σκοπός είναι η ανακάλυψη αυτών των προτύπων με έναν συστηματικό τρόπο, και ακολούθως η μελέτη της σύστασης και των ιδιαίτερων χαρακτηριστικών που αυτά φέρουν.

## *4.2 Μεθοδολογία*

Τα βήματα της προτεινόμενης μεθοδολογίας απεικονίζονται στην Εικόνα 34. Αρχικά, συγκεντρώνουμε ένα σύνολο πρωτεϊνών που είτε είναι πλήρως ρεομορφικές είτε φέρουν ρεομορφικές περιοχές στην ακολουθία τους, και μάλιστα όλες εκφράζονται στον άνθρωπο. Στη συνέχεια, εκμεταλλευόμαστε πληροφορίες σχετικές με τις αλληλεπιδράσεις των εν λόγω πρωτεϊνών με άλλα πρωτεϊνικά μόρια, τις οποίες εξάγουμε από βάσεις δεδομένων που συγκεντρώνουν συστηματικά πρωτεϊνικές αλληλεπιδράσεις, όπως αυτές προκύπτουν από πειραματικές μελέτες. Στο δίκτυο που προκύπτει εντοπίζουμε τις ρεομορφικές πρωτεΐνες που συνιστούν τους κόμβους του δικτύου (δηλαδή, τις πρωτεΐνες που εμφανίζουν υψηλό βαθμό συνδεσιμότητας με άλλα πρωτεϊνικά μόρια) και ανακτούμε τις ακολουθίες των πρωτεϊνών που αλληλεπιδρούν. Η κεντρική μας υπόθεση είναι ότι για να αλληλεπιδρά μια συγκεκριμένη πρωτεΐνη (ρεομορφική) με έναν μεγάλο αριθμό ετερόκλητων πρωτεϊνών, τότε αυτές θα διαθέτουν στην ακολουθία τους ένα τμήμα (ακολουθιακό πρότυπο) το οποίο ευνοεί και επάγει την επακόλουθη λειτουργία. Ανάμεσα σε αυτές τις ακολουθίες αναζητούμε ακολουθιακά πρότυπα, με την μορφή κανονικών εκφράσεων, που εκπροσωπούνται σημαντικά στις υπό μελέτη ακολουθίες. Ακολούθως, τα εξαχθέντα πρότυπα αντιπαραβάλλονται με ακολουθίες ελέγχου ώστε να αποφευχθούν πρότυπα που εμφανίζονται με μεγάλη συχνότητα στο γενικό σύνολο των πρωτεϊνών και να παραμείνουν ακολουθιακά πρότυπα ιδιαιτέρως συγκεκριμένα και επιλεκτικά. Στη συνέχεια το τελικό σύνολο προτύπων αναλύεται διεξοδικά ως προς την αμινοξική του σύσταση καθώς και τις ενδεχόμενες λειτουργικές συσχετίσεις.




**Εικόνα 34:** Διαγραμματική απεικόνιση της προτεινόμενης μεθοδολογικής ανάλυσης.

## *Ρεομορφικές πρωτεΐνες*

Το αρχικό σύνολο ρεομορφικών πρωτεϊνών εξάγεται από την βάση DISPROT [100], η οποία φιλοξενεί αμιγώς ρεομορφικές πρωτεΐνες, δηλαδή πρωτεΐνες που είναι ρεομορφικές είτε τμηματικά είτε στην ολότητά τους. Από τις 523 πρωτεΐνες που περιέχονται στην βάση DISPROT εξαιρούμε αυτές που δεν εκφράζονται στον άνθρωπο, καταλήγοντας σε ένα σύνολο 193 πρωτεϊνών για τις οποίες ανακτούμε επίσης τα αντίστοιχα γονίδια που τις

κωδικοποιούν. Το σύνολο των πρωτεϊνών/γονιδίων που προκύπτει αποτελεί την βάση πάνω στην οποία δομείται η ακόλουθη μεθοδολογική ανάλυση.

### *Κατασκευή δικτύου*

Στο προηγούμενο βήμα συλλέχθηκαν στο σύνολό τους οι μέχρι σήμερα χαρακτηρισμένες ρεομορφικές πρωτεΐνες που κωδικοποιούνται στον άνθρωπο, ώστε να κατασκευάσουμε για αυτές τις πρωτεΐνες ένα δίκτυο αλληλεπιδράσεων με άλλα πρωτεϊνικά μόρια. Συγκεκριμένα, για τις εν λόγω πρωτεΐνες αναζητούμε στην βάση Michigan Molecular Interactions (MiMI) [101] πιθανές πρωτεϊνικές αλληλεπιδράσεις που έχουν εντοπιστεί πειραματικά. Η βάση MiMI ουσιαστικά συγκεντρώνει και ενοποιεί δυαδικές αλληλεπιδράσεις όπως αυτές έχουν καταχωρηθεί σε βάσεις όπως η Human Protein Reference Database (HPRD) [44], η IntAct [46], σύνολα δεδομένων από το Center for Cancer Systems Biology of Harvard [102], καθώς και πλήθος άλλων. Για κάθε ρεομορφική πρωτεΐνη εισόδου, ανασύρουμε όλα τα αλληλεπιδρώντα με αυτήν μόρια, καταλήγοντας σε ένα δίκτυο πρωτεϊνικών αλληλεπιδράσεων που αποτελείται από 3995 κόμβους και 52678 ακμές. Επιπρόσθετα, για το προκύπτον δίκτυο υπολογίζουμε ένα πλήθος ποσοτικών δεικτών που περιγράφουν τα τοπολογικά χαρακτηριστικά του. Συγκεκριμένα, εξάγονται τα ακόλουθα χαρακτηριστικά: i) συντελεστής ομαδοποίησης (clustering coefficient): ένα μέτρο που δείχνει τον βαθμό στον οποίο οι κόμβοι ενός δικτύου τείνουν να είναι συνδεδεμένοι μεταξύ τους, ii) συνδεσιμότητα (degree distribution): ο αριθμός των ακμών που συνδέονται με έναν κόμβο, iii) μέση συνδεσιμότητα (average degree): ο μέσος όρος συνδεσιμότητας για όλο το δίκτυο, iv) διάμετρος δικτύου (network diameter): ο ελάχιστος αριθμός ακμών που ενώνουν τους δύο πιο απομακρυσμένους κόμβους του δικτύου, v) βαθμός κεντρικότητας του δικτύου (network centralization): περιγράφει την ομοιόμορφη κατανομή των ακμών ανάμεσα στους κόμβους του δικτύου, vi) ετερογένεια του δικτύου (network heterogeneity): η τάση του δικτύου να περιλαμβάνει πυκνά συνδεδεμένους κόμβους, vii) κεντρικότητα διαμεσολάβησης (betweenness centrality): ο βαθμός του ελέγχου που ασκεί ένας κόμβος στους κόμβους με τους οποίους αλληλεπιδρά και viii) κεντρικότητα εγγύτητας (closeness centrality): ένα μέτρο που δείχνει την ταχύτητα με την οποία η πληροφορία από έναν κόμβο "ρέει" στο σύνολο των κόμβων που επηρεάζει/αλληλεπιδρά.

*Εξαγωγή ακολουθιακών προτύπων*

Έπειτα από την κατασκευή του δικτύου αλληλεπίδρασης πρωτεϊνών, υπολογίζουμε για κάθε ρεομορφική πρωτεΐνη του δικτύου τον βαθμό συνδεσιμότητας (degree of connectivity) της με τις υπόλοιπες πρωτεΐνες του δικτύου. Επιλέγουμε στη συνέχεια μόνο το κορυφαίο 20% των κόμβων με βάση τον βαθμό συνδεσιμότητας ξεδιαλέγοντας κατ' αυτόν τον τρόπο τους κόμβους υψηλού βαθμού (που συχνά αποκαλούνται *κέντρα* - hubs), οι οποίοι είναι 43 στο υπό μελέτη δίκτυο. Παρόλο που το κατώφλι για την επιλογή των κέντρων του δικτύου (20%) δεν αποτελεί αυστηρό κριτήριο, εντούτοις έχει χρησιμοποιηθεί στη βιβλιογραφία για τον εντοπισμό κέντρων [103, 104]. Στον Πίνακας 14 παρατίθενται οι 43 πρωτεΐνες-κέντρα που προκύπτουν στο τελικό σύνολο καθώς και ο βαθμός συνδεσιμότητας της καθεμιάς. Παρατηρούμε ότι σε κάθε περίπτωση ο αριθμός των πρωτεϊνών με τις οποίες υπάρχει αλληλεπίδραση υπερβαίνει τις 57 και σε ορισμένες περιπτώσεις περιλαμβάνει πολλές εκατοντάδες.

**Πίνακας 14:** Γονίδια με τον μεγαλύτερο βαθμό συνδεσιμότητας στο δίκτυο.

| Γονίδιο | Βαθμός | Γονίδιο | Βαθμός | Γονίδιο | Βαθμός | Γονίδιο | Βαθμός |
|---|---|---|---|---|---|---|---|
| MYC | 986 | NCBP1 | 169 | RPA1 | 102 | HMGA1 | 70 |
| MAX | 950 | NCBP2 | 164 | NFKBIA | 96 | BCL2L1 | 68 |
| TP53 | 305 | AR | 162 | CD4 | 94 | VHL | 66 |
| GRB2 | 231 | CCNH | 147 | GSK3B | 92 | MDM2 | 64 |
| ESR1 | 214 | HNRNPA1 | 141 | HRAS | 90 | POU2F1 | 63 |
| POLR2H | 208 | SHC1 | 135 | PLK1 | 87 | FOS | 60 |
| RAC1 | 206 | RAF1 | 116 | BCL2 | 87 | NCOA3 | 60 |
| EGFR | 205 | CDKN1A | 116 | RXRA | 82 | CTDP1 | 60 |
| RELA | 198 | SP1 | 104 | CCNB1 | 82 | SRF | 58 |
| SMAD4 | 184 | ETF1 | 103 | ARHGAP1 | 79 | CDKN2A | 57 |
| BRCA1 | 178 | NR3C1 | 103 | CDKN1B | 78 | | |

Προκύπτουν, λοιπόν, κατ' αυτόν τον τρόπο 43 σύνολα (δηλαδή ένα για κάθε πρωτεΐνη-κέντρο), όπου το καθένα περιέχει τις ακολουθίες των πρωτεϊνών που έχει βρεθεί ότι αλληλεπιδρούν με την εν λόγω πρωτεΐνη-κέντρο. Στις ακολουθίες καθενός από αυτά τα σύνολα εφαρμόζουμε ένα σύνολο βημάτων προεπεξεργασίας: αρχικά, διατηρούμε μόνο

πρωτεΐνες που παρουσιάζουν μεταξύ τους το πολύ 25% ομοιότητα, δηλαδή, από ομάδες πρωτεϊνών που εμφανίζουν τουλάχιστον 25% ομοιότητα, διατηρείται μόνο μια πρωτεΐνη-εκπρόσωπος [105]. Με αυτόν τον τρόπο εξαλείφουμε το ενδεχόμενο ένα ακολουθιακό πρότυπο να παρουσιάσει παραπλανητικά μεγάλο βαθμό εκπροσώπησης κατά την εξαγωγή προτύπων. Στη συνέχεια, αφαιρούμε από τις ακολουθίες που παραμένουν πρωτεϊνικές περιοχές που συνήθως δεν φέρουν ακολουθιακά πρότυπα με λειτουργική σημασία, τέτοιες περιοχές είναι διαμεμβρανικά τμήματα (transmembrane segments), περιελιγμένα σπειράματα (coiled-coils), περιοχές που περιέχουν κολλαγόνο κ.α. [106]. Με τα βήματα που περιγράφηκαν παραπάνω αφαιρούμε από κάθε σύνολο πρωτεϊνών, πλεονάζουσες πρωτεϊνικές περιοχές και εν γένει περιοχές στην ακολουθία του πρωτεϊνικού μορίου που ενδεχόμενα θα δυσχέραιναν την εξαγωγή προτύπων ή ακόμη χειρότερα θα οδηγούσαν σε εσφαλμένα πρότυπα και συμπεράσματα. Ακολούθως, και σύμφωνα με την αρχική και θεμέλια υπόθεσή μας, εφόσον πλήθος πρωτεϊνών αλληλεπιδρούν με μια συγκεκριμένη, τότε στην πλειοψηφία τους αυτές οι πρωτεΐνες θα περιέχουν ένα τμήμα στην ακολουθία τους που θα σηματοδοτεί και θα επάγει την αλληλεπίδραση. Τα σύνολα που προκύπτουν έπειτα από την προαναφερθείσα προεπεξεργασία, αποτελούν την είσοδο του αλγορίθμου TEIRESIA [89] ώστε να εντοπιστούν ακολουθιακά πρότυπα, με τη μορφή κανονικών εκφράσεων, που ενυπάρχουν με μεγάλη συχνότητα στις ακολουθίες εισόδου. Κατά την εκτέλεση του αλγορίθμου, στοιχειώδη πρότυπα που υπερβαίνουν έναν ελάχιστο βαθμό στήριξης (support) ανάμεσα στις ακολουθίες εισόδου διατηρούνται και στη συνέχεια συνενώνονται ώστε να προκύψουν πιο συγκεκριμένα και επιλεκτικά πρότυπα. Στην παρούσα ανάλυση ο βαθμός στήριξης που επιλέχτηκε είναι K=3, δηλαδή κάθε πρότυπο πρέπει να υπάρχει σε τουλάχιστον τρείς διαφορετικές ακολουθίες. Τα εξαχθέντα πρότυπα απαιτούμε να έχουν μήκος τριών έως οκτώ χαρακτήρων [107, 108], εκ των οποίων τουλάχιστον τρεις καταλαμβάνονται από σταθερούς χαρακτήρες (literals), ώστε να συγκλίνει ο αλγόριθμος εξαγωγής προτύπων [91].

Στο επόμενο βήμα τα πρότυπα που εξάγονται από τον TEIRESIA, υπόκεινται σε μια διαδικασία αξιολόγησης που περιλαμβάνει δύο στάδια. Στο πρώτο στάδιο, υπολογίζεται για κάθε εξαχθέν πρότυπο η αντιπροσώπευσή (representation) του στις ακολουθίες εισόδου, ώστε κάθε πρότυπο να αφορά ένα σημαντικό ποσοστό των δεδομένων εισόδου. Συγκεκριμένα, κάθε πρότυπο πρέπει να παρατηρείται τουλάχιστον στο 40% των ακολουθιών εισόδου, όπου δεν υπολογίζονται περισσότερες από μια παραλλαγές για κάθε ακολουθία. Το κατώφλι αυτό ορίζεται δυναμικά για καθένα από τα 43 σύνολα, μιας και το

πλήθος των ακολουθιών που απαρτίζουν κάθε σύνολο ποικίλει σημαντικά. Κατά το δεύτερο στάδιο αξιολόγησης, υπολογίζεται η στατιστική σημαντικότητα του κάθε προτύπου. Συγκεκριμένα, εκτελούμε $x^2$-test (εξίσωση 13) ώστε να εντοπίσουμε εκείνα τα πρότυπα που εμφανίζονται στο υπό μελέτη σύνολο ακολουθιών εισόδου, στατιστικά σημαντικά περισσότερο από ότι θα ανέμενε κάποιος με βάση ένα αντιπροσωπευτικό σύνολο ελέγχου.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad (13)$$

όπου $O_i$ και $E_i$ συνιστούν τις παρατηρούμενες και προσδοκώμενες τιμές αντίστοιχα, και n είναι ο αριθμός των κελιών στο πίνακα τιμών που δομείται κατά την εκτέλεση του $x^2$-test. Για να υπολογίσουμε την προσδοκώμενη συχνότητα, σχηματίσαμε ένα σύνολο ακολουθιών ελέγχου, που περιλαμβάνει 7000 πρωτεΐνες τυχαία επιλεγμένες από την Protein Data Bank [75], όπου επίσης έχουν υποστεί ακριβώς τα ίδια στάδια προεπεξεργασίας. Η p-value που προκύπτει για κάθε ακολουθιακό πρότυπο από την εκτέλεση του $x^2$-test, αποτελεί μέτρο για να εκτιμήσουμε κατά πόσον ένα πρότυπο εμφανίζεται στατιστικά σημαντικά περισσότερες φορές στο σύνολο εισόδου σε σχέση με το σύνολο ελέγχου. Το κατώφλι που επιλέχτηκε για την p-value, είναι $10^{-2}$, όπου πρότυπα με υψηλότερες τιμές απορρίπτονται [99]. Συνολικά, ο συνδυασμός των μέτρων της αντιπροσώπευσης και της στατιστικής σημαντικότητας που υπολογίζονται για κάθε ακολουθιακό πρότυπο, θα αναδείξει πρότυπα που περιγράφουν πλήρως και με συγκεκριμένο και σαφή τρόπο τις πρωτεϊνικές περιοχές που επάγουν σημαντικές λειτουργίες των ρεομορφικών πρωτεϊνών.

### *Λειτουργική ανάλυση*

Στο επόμενο βήμα της μεθοδολογικής μας ανάλυσης, κάθε πρότυπο που έχει προκύψει με κάθε είδος εξαγωγής ακολουθιακών προτύπων (ακριβής εξαγωγή προτύπων, χημικές ομαδοποιήσεις αμινοξέων, δομικές ομαδοποιήσεις αμινοξέων) αντιπαραβάλλεται και συγκρίνεται εκ περιτροπής με κάθε εγγραφή της βάσης ELM [94]. Πιο συγκεκριμένα, κάθε ένα από τα πρότυπα που έχουν διατηρηθεί ως περιγραφικά, συγκρίνεται με τη χρήση

του αλγορίθμου Comparimotif [95] με κάθε πρότυπο στην ELM, ώστε να ανακαλύψουμε πιθανές λειτουργικές συσχετίσεις των εξαχθέντων προτύπων. Η ELM αποτελεί μια σαφώς δομημένη βάση δεδομένων όπου οι λειτουργίες ομαδοποιούνται σε τέσσερεις ευρείες λειτουργικές ομάδες: i) στόχευση/εντόπιση (targeting/localization: TRG), ii) μετα-μεταγραφικές τροποποιήσεις (post-translational modifications: MOD), iii) πρόσδεση/υποκατάσταση (ligand/binding: LIG) και iv) θέσεις αποκοπής (cleavage: CLV). Αξίζει να σημειωθεί ότι ο αλγόριθμος Comparimotif δεν εντοπίζει μόνο ακριβείς επικαλύψεις προτύπων, αλλά χρησιμοποιεί την εξελικτική σχέση μεταξύ των αμινοξέων ώστε να σκοράρει κάθε ζεύγος χαρακτήρων μεταξύ των υπό σύγκριση προτύπων.

## *4.3 Αποτελέσματα - συζήτηση*

Από την μεθοδολογική ανάλυση που περιγράφηκε παραπάνω, προέκυψε μια λίστα ακολουθιακών προτύπων που σηματοδοτούν και συμμετέχουν στην επαγωγή διαφόρων λειτουργιών των ρεομορφικών πρωτεϊνών. Η προτεινόμενη ανάλυση, εντοπίζει με συστηματικό τρόπο, βάσει υπερ-αντιπροσώπευσης, ακολουθιακά πρότυπα με λειτουργικό ρόλο, αποφεύγοντας έτσι την δυσχερή και χρονοβόρα πειραματική διαδικασία εντοπισμού.

### *Δίκτυο αλληλεπιδράσεων*

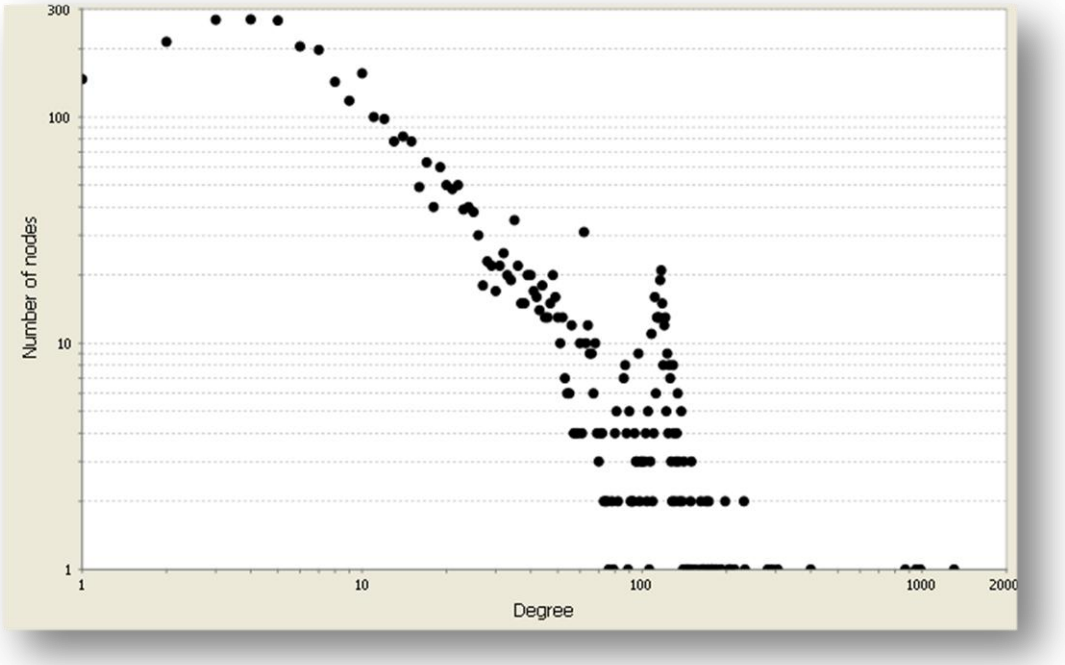
Ο Πίνακας 15 περιέχει τα τοπολογικά χαρακτηριστικά του πρωτεϊνικού δικτύου αλληλεπιδράσεων, και συγκεκριμένα τα εξής: συντελεστής ομαδοποίησης (clustering coefficient), διάμετρος δικτύου (network diameter), μέση συνδεσιμότητα (average degree), βαθμός κεντρικότητας του δικτύου (network centralization) και ετερογένεια του δικτύου (network heterogeneity).

**Πίνακας 15:** Τοπολογικά χαρακτηριστικά του δικτύου αλληλεπίδρασης.

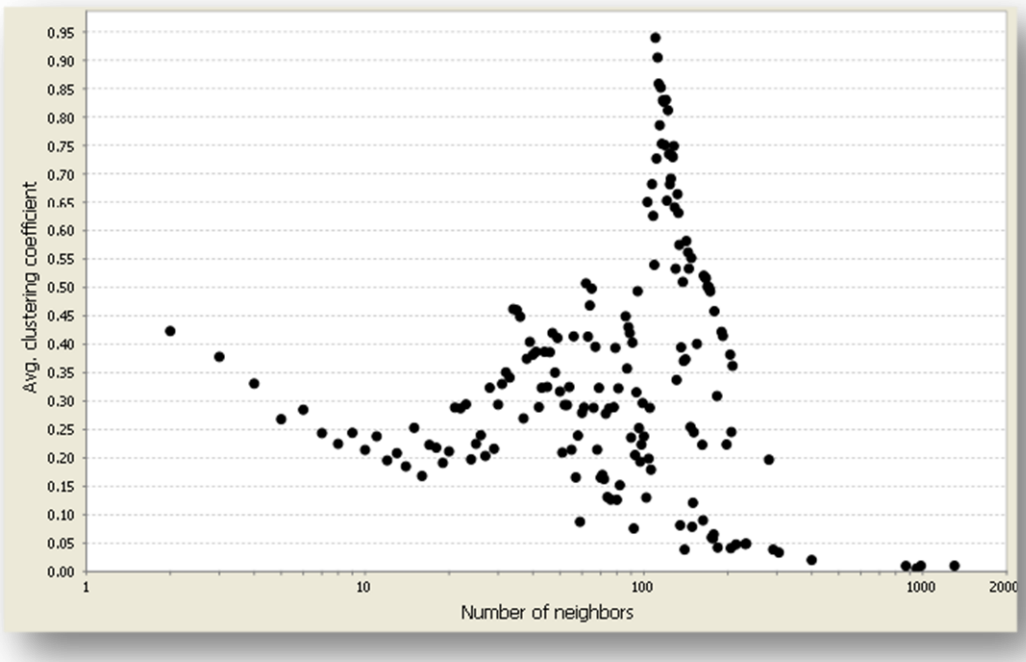
| Χαρακτηριστικό | Τιμή |
|---|---|
| **Συντελεστής ομαδοποίησης** | 0.306 |
| **Διάμετρος δικτύου** | 7 |
| **Μέση συνδεσιμότητα** | 26.639 |
| **Βαθμός κεντρικότητας δικτύου** | 0.322 |
| **Ετερογένεια δικτύου** | 1.824 |

Επίσης, στην Εικόνα 35, απεικονίζονται οι γραφικές παραστάσεις που αφορούν την κατανομή της συνδεσιμότητας του δικτύου (degree distribution) με βάση τους κόμβους του, την κατανομή του συντελεστή ομαδοποίησης (clustering coefficient), την κεντρικότητα διαμεσολάβησης (betweenness centrality) και την κεντρικότητα εγγύτητας (closeness centrality).

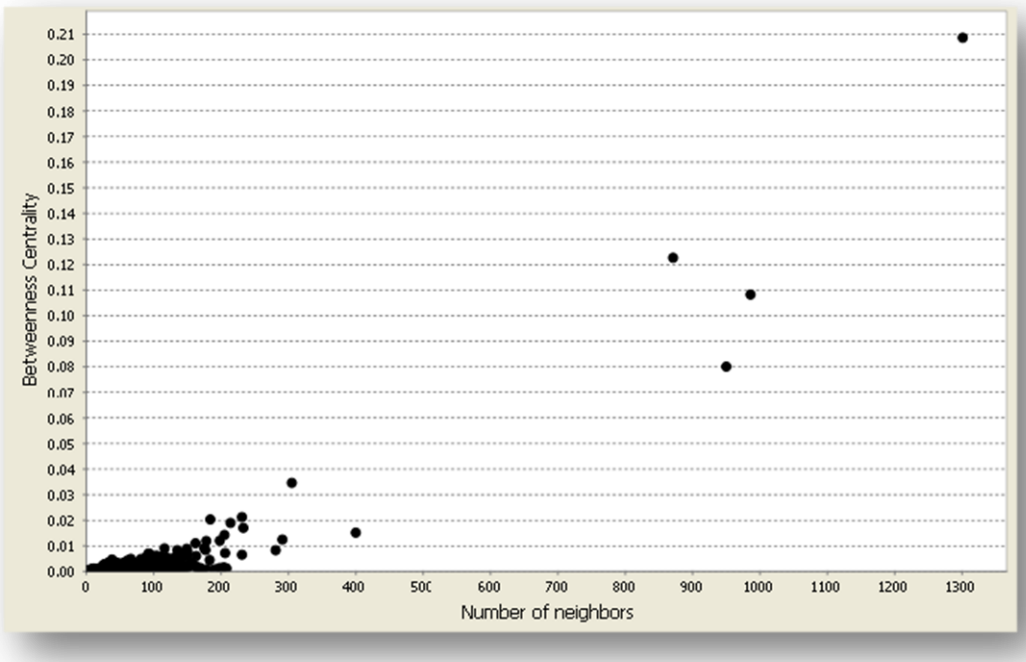
(α)



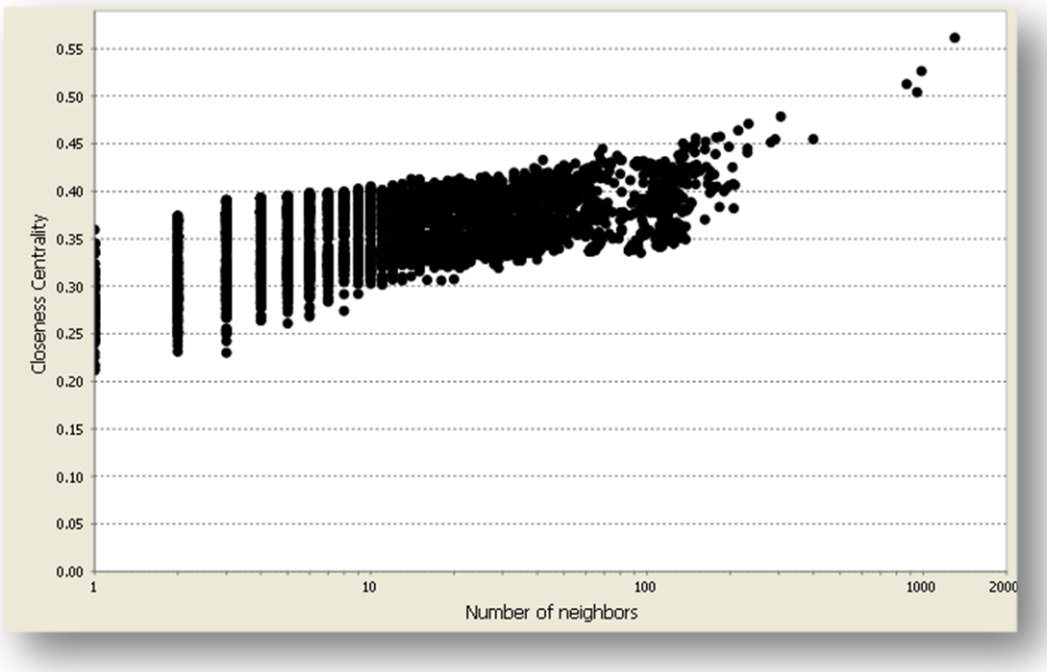
(β)


Avg. clustering coefficient
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1
10
100
1000
2000
Number of neighbors

(γ)


Betweenness Centrality
0.21
0.20
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
Number of neighbors

(δ)




**Εικόνα 35:** Γραφικές παραστάσεις τοπολογικών χαρακτηριστικών του δικτύου. (α) κατανομή της συνδεσιμότητας του δικτύου, (β) κατανομή του συντελεστή ομαδοποίησης, (γ) κεντρικότητα διαμεσολάβησης και (δ) κεντρικότητα εγγύτητας.

Όπως παρατηρούμε στην Εικόνα 35-α, καθώς αυξάνεται ο βαθμός συνδεσιμότητας, μειώνεται ο αριθμός των κόμβων που φέρουν τον συγκεκριμένο βαθμό. Παρ' όλα αυτά από το γράφημα δεν παρατηρείται κάποιο σημείο όπου το πλήθος τον κόμβων να φθίνει σαφώς έντονα, οπότε και θα οριοθετούσε τους κόμβους-κέντρα από τους υπόλοιπους κόμβους του δικτύου. Όσον αφορά την κατανομή του συντελεστή ομαδοποίησης (Εικόνα 35-β) πολύ λίγοι κόμβοι φέρουν τιμές κοντά στο μηδέν, και ιδίως αυτοί που εμφανίζουν υψηλό βαθμό συνδεσιμότητας, ενώ η πλειοψηφία των κόμβων παρουσιάζει τιμές κοντά στο 0.35. Επίσης, μια έντονα αυξητική τάση παρατηρείται για κόμβους που έχουν περίπου 100 γείτονες (δηλαδή, πρωτεΐνες με τις οποίες αλληλεπιδρούν). Στην Εικόνα 35-γ παρατηρούμε ότι η κεντρικότητα διαμεσολάβησης εμφανίζει ιδιαιτέρως χαμηλές τιμές, γεγονός που δείχνει ότι οι συντριπτική πλειοψηφία των κόμβων του δικτύου τείνουν να αποτελούν μέλη από ομαδοποιήσεις κόμβων, παρά να διαμεσολαβούν και να συνενώνουν

διαφορετικές ομάδες κόμβων. Όσο για το γράφημα της κεντρικότητας εγγύτητας (Εικόνα 35-δ) παρατηρούμε μια έντονη τάση των κόμβων προς τιμές γύρω από το 0.35.

### *Ακολουθιακά πρότυπα*

Στον Πίνακας 16 παρατίθενται τα πιο αντιπροσωπευτικά πρότυπα, όπως προέκυψαν από την μεθοδολογική μας ανάλυση, για κάθε ένα από τα 43 κέντρα του δικτύου μας. Όπως αναφέρθηκε και παραπάνω, από το σύνολο των ακολουθιακών προτύπων που προέκυψαν για κάθε κέντρο απαλείψαμε αυτά που εκπροσωπούνταν σε λιγότερο από το 40% των ακολουθιών εισόδου, καθώς επίσης και αυτά για τα οποία η p-value από το $x^2$-test ήταν μεγαλύτερη από $10^{-2}$. Παρ' όλα αυτά για μερικές πρωτεΐνες-κέντρα διατηρήθηκε ένας σχετικά μεγάλος αριθμός προτύπων. Σε αυτήν την περίπτωση ο Πίνακας 16 περιλαμβάνει μόνο τα κορυφαία πέντε ακολουθιακά πρότυπα για κάθε κέντρο. Επίσης, σε μερικές περιπτώσεις η εφαρμογή των παραπάνω κατωφλίων διατήρησε πολύ λίγα πρότυπα, ενώ σε μερικές περιπτώσεις κανένα (ETF1).

**Πίνακας 16:** Λίστα με τα πέντε κορυφαία πρότυπα για κάθε κόμβο-κέντρο του δικτύου.

| **Κέντρο** | **Πρότυπο** | **p-value** | **Αντ.* (%)** | **Κέντρο** | **Πρότυπο** | **p-value** | **Αντ.* (%)** | **Κέντρο** | **Πρότυπο** | **p-value** | **Αντ.* (%)** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MYC | LL.....L | 1.69E-80 | 45 | HNRNPA1 | EE.E | 4.18E-28 | 42 | RXRA | L..LL | 4.94E-15 | 57 |
| | L..L...L | 1.97E-47 | 43 | | L...L..L | 2.53E-08 | 41 | | S.....LL | 3.43E-25 | 57 |
| | L...L..L | 4.50E-48 | 42 | | E..EE | 1.01E-24 | 41 | | L..L..L | 5.33E-16 | 57 |
| | LL..L | 6.11E-52 | 41 | | L..E.L | 1.14E-11 | 40 | | S.L....L | 1.87E-27 | 57 |
| | L.....LL | 3.17E-66 | 40 | SHC1 | LL.....L | 2.56E-21 | 62 | | SS.S | 2.58E-35 | 54 |
| MAX | LL.....L | 5.58E-63 | 42 | | L..SL | 5.46E-28 | 60 | CCNB1 | S.LL | 1.97E-23 | 52 |
| | L..L...L | 2.19E-39 | 41 | | L..LL | 7.69E-17 | 58 | | LSS | 3.40E-30 | 52 |
| TP53 | L..L...L | 4.60E-12 | 40 | | L..L...L | 3.85E-13 | 57 | | S..S.S | 1.70E-42 | 51 |

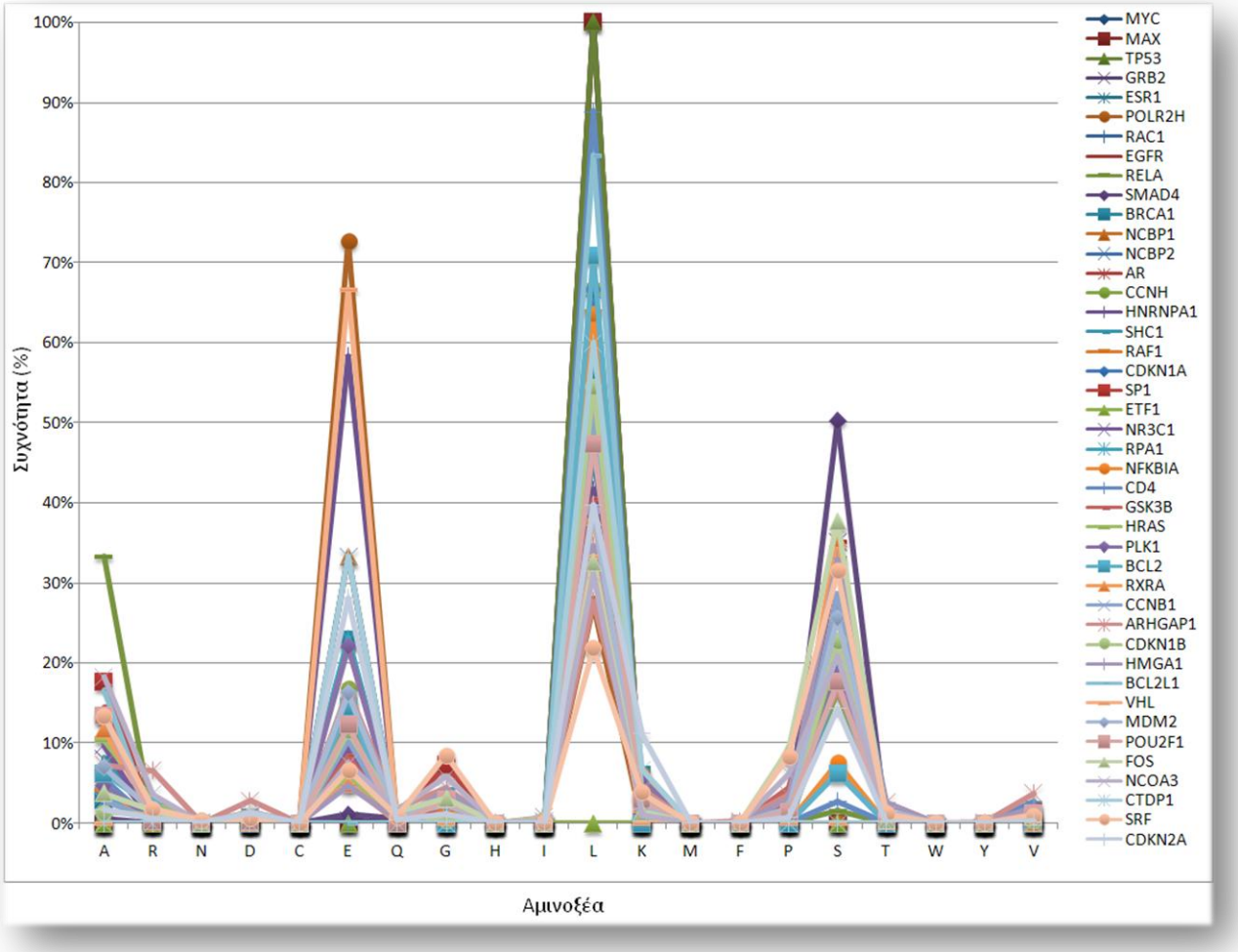
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | L..L..L | 1.02E-16 | 40 | | L...LL | 3.39E-18 | 57 | | L..L.L | 2.47E-14 | 49 |
| GRB2 | L...L..L | 2.54E-15 | 50 | RAF1 | L..L...L | 6.96E-09 | 51 | | SS.S | 1.28E-31 | 49 |
| | LLL | 2.61E-30 | 50 | | L..L.L | 7.64E-14 | 51 | ARHGAP1 | ELL | 1.36E-24 | 79 |
| | S....SS | 4.88E-65 | 50 | | L.....SL | 1.29E-17 | 51 | | L.E.L | 9.23E-19 | 79 |
| | L.L.L | 9.41E-28 | 49 | | L...L..L | 1.28E-08 | 49 | | E.E...E | 4.48E-56 | 79 |
| | LL.....L | 3.19E-18 | 49 | | L..E...L | 2.51E-12 | 49 | | L.L.E | 1.07E-25 | 76 |
| ESR1 | L..LL | 3.15E-28 | 55 | CDKN1A | L....LL | 1.68E-18 | 50 | | L.EL | 1.22E-18 | 76 |
| | L...L..L | 1.66E-20 | 51 | | L...LL | 8.51E-15 | 50 | CDKN1B | S.L.S | 1.60E-31 | 61 |
| | L.E.L | 1.89E-20 | 51 | | L.LL | 4.29E-16 | 49 | | L....E.L | 1.10E-12 | 59 |
| | L..L...L | 1.42E-16 | 48 | | L.E.L | 1.97E-09 | 47 | | L...LL | 1.06E-12 | 56 |
| | LL...L | 1.50E-25 | 48 | | LL.....L | 6.40E-12 | 47 | | L...E..L | 1.16E-16 | 56 |
| POLR2H | EE.E | 2.10E-55 | 48 | SP1 | L....LS | 1.51E-22 | 54 | | L..L...L | 3.02E-08 | 54 |
| | EE.....E | 1.05E-39 | 44 | | L.SS | 9.62E-36 | 54 | HMGA1 | L...LL | 1.27E-16 | 56 |
| | L...L..L | 1.98E-14 | 44 | | SSS | 2.33E-38 | 53 | | S...SS | 4.21E-44 | 55 |
| | E.EE | 1.69E-49 | 43 | | S..SS | 6.01E-37 | 53 | | LL.A | 8.06E-16 | 55 |
| | EE..E | 2.23E-34 | 42 | | SLS | 3.74E-27 | 53 | | SS...S | 4.30E-43 | 53 |
| RAC1 | L.L....L | 1.17E-34 | 64 | ETF1 | - | - | - | | L..LL | 4.92E-12 | 53 |
| | L.....LL | 1.09E-33 | 64 | NR3C1 | L..LL | 3.03E-14 | 52 | BCL2L1 | L..L...L | 1.93E-06 | 45 |
| | L...L..L | 2.97E-22 | 61 | | E.L...L | 1.82E-11 | 49 | | L.A.A | 3.02E-09 | 43 |
| | L..LL | 1.10E-24 | 61 | | L..S.L | 2.93E-24 | 49 | | L..LL | 3.30E-06 | 42 |
| | LL..L | 4.64E- | 61 | | L...L..L | 5.15E- | 48 | | LL....L | 1.79E- | 42 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 25 | | | | 10 | | | | 09 | |
| | LL.....L | 3.77E-22 | 53 | | L..L..L | 4.19E-12 | 48 | VHL | EE....L | 1.75E-08 | 41 |
| | L..LL | 4.33E-15 | 47 | | L...L..L | 1.61E-15 | 55 | | S.S...S | 2.75E-33 | 53 |
| EGFR | LL...L | 5.18E-19 | 47 | | L..LL | 3.05E-16 | 53 | | SS...S | 1.16E-33 | 51 |
| | LL..L | 2.31E-15 | 47 | RPA1 | LL..L | 1.65E-16 | 53 | MDM2 | EE...L | 1.55E-13 | 51 |
| | L...LL | 6.06E-18 | 47 | | L....EE | 1.78E-25 | 53 | | L..LL | 2.27E-09 | 51 |
| | L...L..L | 8.77E-13 | 44 | | L.E...L | 1.28E-21 | 53 | | L.E.L | 2.26E-08 | 51 |
| | A.A.A | 1.64E-16 | 43 | | L..L.E | 4.20E-13 | 56 | | LL.....L | 2.13E-14 | 57 |
| RELA | L.E....L | 1.53E-13 | 43 | | L...L.E | 1.72E-08 | 46 | | EE....L | 4.44E-12 | 51 |
| | E..EE | 1.26E-35 | 43 | NFKBIA | L.L....L | 5.25E-10 | 44 | POU2F1 | L.E.L | 1.17E-08 | 51 |
| | L..LL | 4.14E-13 | 42 | | K..L..L | 3.30E-12 | 44 | | S.L..S | 2.83E-27 | 51 |
| | LL..L | 1.39E-19 | 52 | | L.E.L | 6.06E-06 | 43 | | A..L...L | 1.32E-10 | 51 |
| | L..L...L | 6.16E-16 | 52 | | LL..L | 9.82E-14 | 49 | | S.S...S | 3.22E-40 | 67 |
| SMAD4 | L.....LL | 5.51E-22 | 50 | | L.L..L | 1.48E-18 | 49 | | S....SS | 1.03E-36 | 61 |
| | LL...L | 1.87E-20 | 49 | CD4 | L...LL | 7.70E-13 | 45 | FOS | L.....EL | 3.13E-12 | 61 |
| | SL...L | 7.55E-26 | 49 | | L..L.L | 2.36E-14 | 45 | | S..SS | 2.61E-27 | 58 |
| | L...L..L | 1.19E-22 | 55 | | LL....L | 3.21E-15 | 44 | | S.SS | 3.38E-28 | 58 |
| | L..LL | 4.16E-22 | 53 | | L..LL | 1.32E-17 | 59 | | L..L.E | 7.52E-10 | 63 |
| BRCA1 | L.E....L | 1.13E-20 | 52 | GSK3B | L...L..L | 9.94E-15 | 57 | NCOA3 | L.LL | 1.34E-11 | 59 |
| | L..L...L | 1.49E-16 | 51 | | L..L.S | 7.44E-19 | 51 | | LL....L | 1.13E-10 | 56 |
| | S.S...S | 6.00E-67 | 51 | | L..L...L | 6.36E-10 | 51 | | L..S.L | 7.52E-14 | 56 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NCBP1 | LL..L | 4.61E-14 | 45 | | E..E.L | 4.22E-15 | 51 | | S...S.L | 8.64E-18 | 56 |
| | L...LL | 1.48E-16 | 45 | HRAS | L..L...L | 4.05E-16 | 65 | CTDP1 | E....LL | 1.03E-09 | 48 |
| | EE.E | 6.78E-32 | 42 | | L...L..L | 1.28E-14 | 61 | | L....K.L | 4.83E-11 | 46 |
| | L...L..L | 2.35E-09 | 41 | | L.L..L | 2.11E-21 | 61 | | L...LL | 2.76E-07 | 42 |
| | S....S.S | 6.96E-49 | 41 | | LL...L | 3.72E-18 | 59 | | EE.E | 4.77E-16 | 42 |
| NCBP2 | LL..L | 7.34E-12 | 42 | | L....LL | 1.51E-19 | 57 | | L..LE | 4.92E-07 | 40 |
| | EE.E | 2.45E-30 | 42 | PLK1 | E..L..L | 1.89E-13 | 52 | SRF | S..SS | 2.54E-26 | 61 |
| | L...L..L | 7.62E-09 | 41 | | E..K.L | 1.21E-14 | 52 | | GS.S | 1.48E-24 | 61 |
| AR | L..LL | 1.77E-22 | 53 | | SL....L | 4.85E-23 | 52 | | S.SS | 6.50E-24 | 57 |
| | S.S...S | 2.35E-70 | 52 | | E.L..L | 3.40E-09 | 50 | | S.S..S | 7.18E-29 | 57 |
| | L..L...L | 7.35E-17 | 51 | | E...LL | 1.40E-14 | 50 | | S...SS | 2.22E-26 | 57 |
| | SSS | 2.03E-62 | 50 | BCL2 | L..L...L | 9.12E-12 | 54 | CDKN2A | L...L..L | 7.20E-10 | 56 |
| | S....SS | 9.53E-73 | 50 | | A.L..L | 2.08E-09 | 46 | | EE..E | 3.06E-20 | 54 |
| CCNH | L...LL | 7.90E-16 | 43 | | LL..L | 2.09E-08 | 44 | | EEE | 9.65E-23 | 51 |
| | SS...S | 6.75E-52 | 42 | | LL.L | 5.39E-10 | 43 | | E..EK | 1.44E-16 | 51 |
| | LL....L | 2.21E-18 | 41 | | LL....L | 2.17E-11 | 43 | | E...E.E | 1.61E-25 | 51 |
| | E....LL | 2.83E-13 | 41 | | | | | | | | |
| | L....EE | 6.87E-20 | 41 | | | | | | | | |

*Αντ.: αντιπροσώπευση

Στην συζήτηση που ακολουθεί, τα συμπεράσματα που προκύπτουν έχουν εξαχθεί βάσει του όλων των ακολουθιακών προτύπων που αντιστοιχούν σε κάθε πρωτεΐνη-κέντρο και όχι από το υποσύνολο αυτών όπως φαίνονται στον Πίνακας 16. Είναι αξιοσημείωτο, ότι παρότι ασχολούμαστε με ετερόκλητα σύνολα πρωτεϊνών, τα πρότυπα που έχουν προκύψει παρουσιάζουν σημαντικές ομοιότητες μεταξύ τους, πιθανώς λόγω της βάσης που τα συνδέει, ότι δηλαδή αφορούν και εξυπηρετούν τις λειτουργίες ρεομορφικών πρωτεϊνών. Συνολικά, παρατηρούμε ότι τους δομικούς λίθους των εξαχθέντων ακολουθιακών προτύπων συνιστούν ως επί το πλείστον τα αμινοξέα λευκίνη (L), σερίνη (S), γλουταμινικό οξύ (E) και η αλανίνη (A). Η κατανομή των 20 αμινοξέων στα πρότυπα που προέκυψαν παρουσιάζεται στην Εικόνα 36. Συγκεκριμένα για κάθε αμινοξύ μετρήθηκε ο μέσος όρος των εμφανίσεών του ανάμεσα στα πρότυπα που εξήχθησαν για κάθε κέντρο του δικτύου. Σαφείς κορυφές παρατηρούνται για τα αμινοξέα λευκίνη, γλουταμινικό οξύ, σερίνη καθώς και ελαφρώς μικρότερες για την αλανίνη και τη γλυκίνη. Επίσης, παρατηρούμε ότι τα υπόλοιπα αμινοξέα σχεδόν στο σύνολό τους εκπροσωπούνται σε πολύ μικρό βαθμό ανάμεσα στα εξαχθέντα πρότυπα.




**Εικόνα 36:** Κατανομή των αμινοξέων στα εξαχθέντα πρότυπα.

Παρατηρούμε ότι η συντριπτική πλειοψηφία των προτύπων περιέχει είτε αμιγώς είτε τουλάχιστον σε μεγάλο βαθμό το αμινοξύ λευκίνη. Θυμίζει, λοιπόν, εν μέρει περιοχές των πρωτεϊνικών ακολουθιών με διαδοχικά αμινοξέα λευκίνης (Leucine Rich Repeats: LRR) [109, 110]. Αυτή η ειδική κατηγορία πρωτεϊνικών ακολουθιών έχει βρεθεί ότι παίζει σημαντικό ρόλο σε αλληλεπιδράσεις μεταξύ πρωτεϊνικών μορίων, καθώς επίσης και κατά την αναγνώριση και πρόσδεση πεπτιδίων [110]. Επομένως, ο έντονα αλληλεπιδραστικός χαρακτήρας των ρεομορφικών πρωτεϊνών φαίνεται να επάγεται σε μεγάλο βαθμό από LRRs. Επιπρόσθετα, ανάμεσα στις πολλαπλές λειτουργίες όπου συμμετέχουν οι LRRs, έχει αποκαλυφθεί και μια έντονη συσχέτισή τους με ασθένειες στον άνθρωπο [111]. Επίσης, μεγάλη προτίμηση παρατηρείται ανάμεσα στα εξαχθέντα πρότυπα και προς το γλουταμινικό οξύ. Στην βιβλιογραφία έχει αναφερθεί συσχέτιση μεταξύ ρεομορφικών

πρωτεϊνών και πεπτιδίων πλούσιων σε γλουταμινικό οξύ, και συγκεκριμένα στην περίπτωση των ραβδίων του αμφιβληστροειδούς [112].

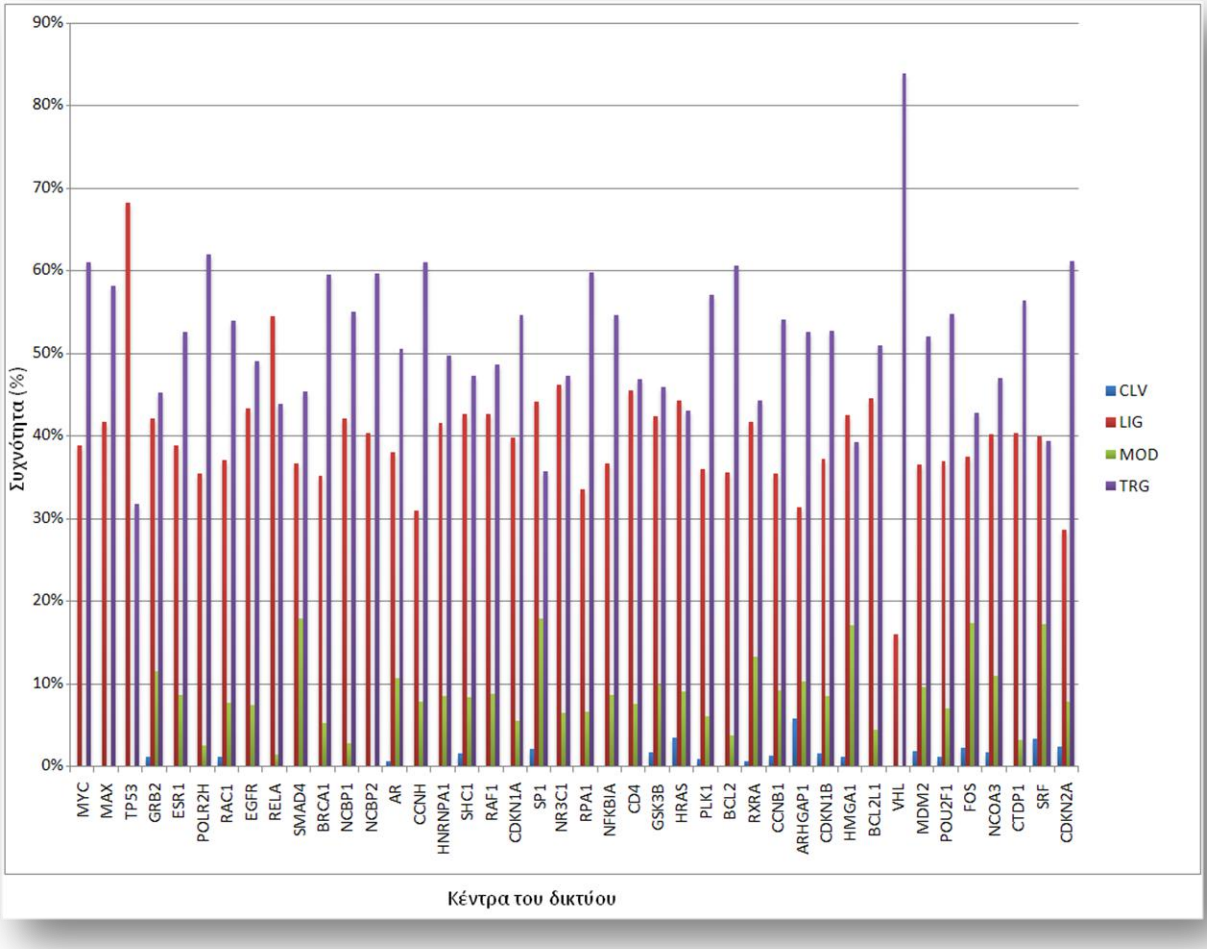
Στη βιβλιογραφία [113, 114] έχει επίσης μελετηθεί, συνολικά στις πρωτεΐνες, η σύσταση ακολουθιακών προτύπων που εμπλέκονται με διάφορες κυτταρικές λειτουργίες. Συγκεκριμένα, παρατηρήθηκαν διαδοχικές επαναλήψεις μεμονωμένων καταλοίπων που συμμετέχουν στην επαγωγή πρωτεϊνικών αλληλεπιδράσεων, καθώς και σε πλήθος λειτουργιών. Τα αμινοξέα που παρατηρήθηκε ότι αποτελούν αυτού του τύπου τα πρότυπα είναι η λευκίνη, η αλανίνη, η σερίνη, η γλουταμίνη, η γλυκίνη, το γλουταμινικό οξύ και η προλίνη [113]. Σε μια παρόμοια ανάλυση μελετήθηκαν ακολουθιακά πρότυπα προερχόμενα από διάφορους οργανισμούς, ώστε να εντοπιστούν προτιμήσεις προς συγκεκριμένα κατάλοιπα. Επικεντρώνοντας μόνο στον άνθρωπο, τα αμινοξέα που παρατηρήθηκαν με μεγαλύτερη συχνότητα είναι το γλουταμινικό οξύ, η σερίνη, η λευκίνη, η αλανίνη, η γλυκίνη και η προλίνη. Κατά συνέπεια, υπάρχει ένας βασικός πυρήνας αμινοξέων που είναι κοινά ανάμεσα στην βιβλιογραφία και αυτά που προέκυψαν από την δική μας ανάλυση. Υπάρχουν, ωστόσο κάποια αμινοξέα τα οποία στην παρούσα ανάλυση δεν παρατηρήθηκαν με εξίσου μεγάλη συχνότητα, το οποίο οφείλεται στο γεγονός ότι επικεντρωθήκαμε σε πρότυπα που επάγουν λειτουργίες αποκλειστικά ρεομορφικών πρωτεϊνών. Ακόμη, σε μια σχετική μελέτη στην βιβλιογραφία [114] παρατηρήθηκε έντονη συσχέτιση ανάμεσα σε πρότυπα που φέρουν σειρές από διαδοχικά αμινοξέα λευκίνης, γλουταμινικού οξέος και σερίνης με διαύλους ασβεστίου και πλήθος ασθενειών όπως Huntington, διάφορους τύπους λευχαιμίας, καρκίνο. Η παρατήρηση αυτή συνάδει με τα αποτελέσματα από την παρούσα ανάλυση όπου μελετήθηκε η προτίμηση των εξαχθέντων προτύπων ως προς τα βιολογικά μονοπάτια που συμμετέχουν καθώς και με πολυάριθμες εργασίες που αφορούν στον λειτουργικό ρόλο των ρεομορφικών πρωτεϊνών [31-35].

## *Λειτουργικές συσχετίσεις*

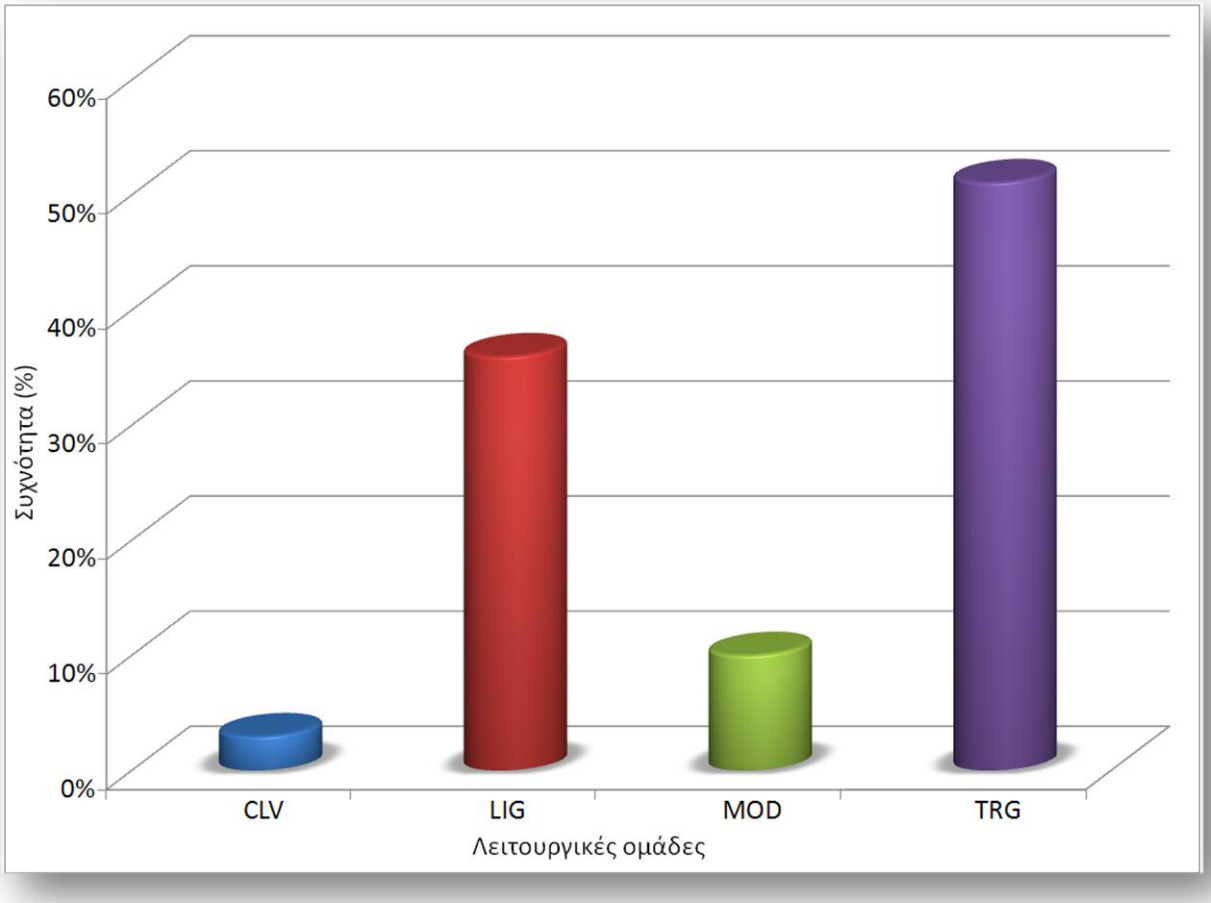
Στην συνέχεια επιχειρούμε να εντοπίσουμε πιθανές λειτουργικές συσχετίσεις των εξαχθέντων προτύπων. Όπως περιγράφηκε και παραπάνω, τα πρότυπα από κάθε μία από τις 43 ρεομορφικές πρωτεΐνες-κέντρα συγκρίνονται με τα πρότυπα της βάσης ELM [94] των οποίων η λειτουργία είναι πειραματικά επιβεβαιωμένη. Όταν η ομοιότητα μεταξύ των δύο υπό σύγκριση προτύπων είναι μεγάλη, τότε στο πρότυπο που εξήχθη με την

προτεινόμενη ανάλυση ανατίθεται η αντίστοιχη λειτουργία από το πρότυπο της ELM. Να υπενθυμίσουμε σε αυτό το σημείο, ότι τα πρότυπα της ELM είναι ομαδοποιημένα σε τέσσερεις ευρείες λειτουργικές ομάδες: i) στόχευση/εντόπιση (targeting/localization: TRG), ii) μετα-μεταγραφικές τροποποιήσεις (post-translational modifications: MOD), iii) πρόσδεση/υποκατάσταση (ligand/binding: LIG) και iv) θέσεις αποκοπής (cleavage: CLV). Στην Εικόνα 37 βλέπουμε την συχνότητα εκπροσώπησης της κάθε λειτουργικής ομάδας στα εξαχθέντα πρότυπα. Συγκεκριμένα, η Εικόνα 37-α δείχνει την αντίστοιχη ποσοστιαία εκπροσώπηση για κάθε κόμβο-κέντρο του δικτύου, ενώ η Εικόνα 37-β συνοψίζει την μέση κατανομή όλων των προτύπων από όλα τα κέντρα του δικτύου.

(α)



(β)




**Εικόνα 37:** Κατανομή των εξαχθέντων προτύπων στις λειτουργικές ομάδες της ELM. (α) Για κάθε κέντρο του δικτύου, (β) συνολικά για όλα τα πρότυπα.

Στην Εικόνα 37-α παρατηρούμε ότι η εκπροσώπηση των τεσσάρων λειτουργικών ομάδων ανά τους 43 κόμβους-κέντρα του δικτύου αλληλεπίδρασης, καταδεικνύουν μια έντονη τάση των ρεομορφικών πρωτεϊνών προς συγκεκριμένες κυτταρικές λειτουργίες. Συγκεκριμένα, οι σαφώς πλειοψηφούσες λειτουργικές κλάσεις είναι οι TRG και LIG. Η παρατήρηση αυτή πράγματι επιβεβαιώνεται από τη σχετική βιβλιογραφία όπου η συσχέτιση των ρεομορφικών πρωτεϊνών με λειτουργίες που αφορούν πρόσδεση και στόχευση έχει παρατηρηθεί σε πλήθος μελετών [31-34]. Η δική μας μεθοδολογική ανάλυση, επιπλέον, εντοπίζει συστηματικά και καταδεικνύει τα ακολουθιακά πρότυπα που επάγουν τις συγκεκριμένες λειτουργίες για κάθε ρεομορφική πρωτεΐνη που μελετάμε. Όσον αφορά την λειτουργική ομάδα MOD, παρατηρείται ελαφρώς μειωμένη συχνότητα εκπροσώπησής της στα εξαχθέντα πρότυπα, και στις δύο εικόνες. Στην βιβλιογραφία έχει

επίσης παρατηρηθεί έντονη συσχέτιση ανάμεσα σε ρεομορφικές πρωτεΐνες και περιοχές όπου λαμβάνουν χώρα μετα-μεταγραφικές τροποποιήσεις [34]. Στην περίπτωση της λειτουργικής ομάδας CLV, δεν παρατηρείται κάποια σημαντική εκπροσώπησή της στο σύνολο των ακολουθιακών προτύπων, και μάλιστα σε αρκετές πρωτεΐνες-κέντρα η συχνότητα εμφάνισής της είναι μηδενική. Πράγματι, ούτε στην βιβλιογραφία έχει αναφερθεί συσχετισμός, είτε στο σύνολο των ρεομορφικών πρωτεϊνών είτε που να αφορά μεμονωμένες ρεομορφικές πρωτεΐνες, καταδεικνύοντας κάποια προτίμηση των ρεομορφικών πρωτεϊνών προς λειτουργίες σχετικές με θέσεις αποκοπής.

## *4.4 Συμπεράσματα*

Στο παρόν κεφάλαιο επικεντρωθήκαμε σε μια ειδική κατηγορία πρωτεϊνών που έχουν προσελκύσει έντονο ερευνητικό ενδιαφέρον, τις ρεομορφικές πρωτεΐνες, οι οποίες ουσιαστικά αποτελούν πολυπεπτίδια που είτε τμηματικά, είτε στην ολότητά τους παρουσιάζουν δομική ευμεταβλητότητα και ευελιξία ως προς την τρισδιάστατη δομή τους. Ως επακόλουθο, εμπλέκονται σε πλήθος αλληλεπιδράσεων με άλλα πρωτεϊνικά μόρια και έχει βρεθεί ότι συμμετέχουν σε έναν μεγάλο αριθμό σημαντικών κυτταρικών λειτουργιών καθώς και στην επαγωγή σοβαρών ασθενειών (π.χ. καρκίνος, διαβήτης, καρδιαγγειακά νοσήματα κ.ά.). Στην πλειοψηφία τους, αυτές οι λειτουργίες σηματοδοτούνται και επιτελούνται από μικρού μήκους ακολουθιακά πρότυπα, όπου η ανίχνευσή τους απαιτεί κανονικά χρονοβόρες πειραματικές διαδικασίες. Στην προτεινόμενη μεθοδολογία, εκμεταλλευόμαστε πληροφορίες που αφορούν πρωτεϊνικές αλληλεπιδράσεις ώστε να δομήσουμε ένα πρωτεϊνικό δίκτυο όπου οι ρεομορφικές πρωτεΐνες αποτελούν κόμβους-κέντρα με υψηλό βαθμό συνδεσιμότητας. Εν συνεχεία, αναζητούμε στο σύνολο των μορίων που αλληλεπιδρούν με μια συγκεκριμένη ρεομορφική πρωτεΐνη, ακολουθιακά πρότυπα που ενυπάρχουν στις ακολουθίες αυτές. Τα εξαχθέντα πρότυπα αφού αξιολογηθούν σχολαστικά έναντι ακολουθιών ελέγχου ώστε να διατηρηθούν τα πιο συγκεκριμένα και περιγραφικά πρότυπα αναλύονται ως προς την αμινοξική τους σύσταση. Συγκεκριμένα, παρατηρούμε ότι η τελική λίστα προτύπων που προκύπτει δομείται σχεδόν στο σύνολό της από τα αμινοξέα λευκίνη, γλουταμινικό οξύ, σερίνη, αλανίνη και γλυκίνη, είτε μεμονωμένα είτε από συνδυασμούς αυτών. Επίσης, κάποια αμινοξέα εκλείπουν σχεδόν εξ ολοκλήρου από τα εξαχθέντα πρότυπα (π.χ. ιστιδίνη, μεθειονίνη,

φαινυλαλανίνη, τρυπτοφάνη, τυροσίνη). Το σύνολο των προτύπων που εντοπίσαμε αποτελούν μια συστηματική και εύληπτη περιγραφή για τον τρόπο που επάγουν οι ρεομορφικές πρωτεΐνες τις λειτουργίες τους. Ακολούθως, η γνώση και κατανόηση του μοριακού υποβάθρου μιας κυτταρικής λειτουργίας μπορεί να αποτελέσει σημαντικό βήμα ως προς τον προαγωγή, παρεμπόδιση και εν γένει ρύθμισή της.

# 5ο ΚΕΦΑΛΑΙΟ: Ανάλυση κλινικών και γενετικών δεδομένων - στοματικός καρκίνος

## *5.1 Σκοπός*

Στο παρόν κεφάλαιο θα παρουσιάσουμε μια ολοκληρωμένη μελέτη για την πολύπλευρη παρακολούθηση ασθενών με στοματικό καρκίνο, με σκοπό την μοντελοποίηση της εξέλιξης της νόσου, ώστε να είμαστε σε θέση να προβλέψουμε κατά το δυνατόν ακριβέστερα και νωρίτερα μια πιθανή επανεμφάνιση της νόσου. Για τον σκοπό αυτό συγκεντρώνουμε μια σειρά ετερογενών κλινικών, απεικονιστικών και γενετικών δεδομένων, όπου αναζητούμε τους παράγοντες εκείνους που επάγουν μια επικείμενη υποτροπή της νόσου και την πιθανή συνακόλουθη επανεμφάνιση του καρκίνου. Η προτεινόμενη μεθοδολογική ανάλυση δομείται σε δύο επίπεδα, α) την πρόβλεψη πιθανής επανεμφάνισης χρησιμοποιώντας δεδομένα από την αρχική περίοδο της νόσου, δηλαδή κατά την διάγνωση-θεραπεία, και β) την παρακολούθηση της εξέλιξης της νόσου στο διάστημα που έπεται της θεραπείας για διάστημα 18 μηνών. Στην πρώτη περίπτωση καταγράφουμε χαρακτηριστικά που συνιστούν το κλινικό προφίλ του ασθενούς, απεικονιστικά δεδομένα από την περιοχή εντόπισης του όγκου καθώς και γενετικά δεδομένα που αφορούν την έκφραση μιας σειράς γονιδίων στον καρκινικό ιστό. Στη συνέχεια επεξεργαζόμαστε και αναλύουμε τα παραπάνω δεδομένα στο σύνολό τους ώστε να ανιχνεύσουμε μια πιθανή επανεμφάνιση της νόσου. Κατά την δεύτερη περίπτωση αποσκοπούμε στην συστηματική μελέτη της εξέλιξης της νόσου στο χρονικό διάστημα που ακολουθεί μετά την θεραπεία. Συγκεκριμένα, κατακερματίζουμε το διάστημα παρακολούθησης του ασθενούς, που συνολικά είναι 18 μήνες, σε διαδοχικά διαστήματα 3 μηνών όπου συλλέγουμε και αναλύουμε μια σειρά δεδομένων, δομώντας κατ' αυτόν τον τρόπο ένα μοντέλο εξέλιξης της νόσου όπου λαμβάνουμε υπόψη πλέον και την διάσταση του χρόνου. Τα δεδομένα που χρησιμοποιούνται σε αυτήν την ανάλυση είναι η έκφραση ενός μεγάλου αριθμού γονιδίων που λαμβάνουμε από το αίμα.

Συνολικά ο απώτερος σκοπός της ανάλυσής μας είναι διττός: i) ο εντοπισμός ενός μικρού αριθμού χαρακτηριστικών (κλινικών, απεικονιστικών και γενετικών) που παίζουν κυρίαρχο ρόλο στην εξέλιξη της νόσου. Η γνώση αυτών των παραγόντων δύναται να αποκαλύψει σημαντικά χαρακτηριστικά για τους λόγους που κατευθύνουν την εξέλιξη της νόσου, συμβάλλοντας στην κατανόηση και συνακόλουθη αποκωδικοποίησή της. ii) Η χρησιμοποίηση και συστηματική ανάλυση αυτών των χαρακτηριστικών για να υπολογίσουμε την πιθανότητα του κάθε ασθενούς προς υποτροπή και επανεμφάνιση του καρκίνου. Κατά τη δεύτερη περίπτωση μάλιστα είμαστε σε θέση να υπολογίσουμε όχι μόνο το εάν αλλά και το πότε είναι πιθανότερο να παρουσιαστεί η επανεμφάνιση. Γνωρίζοντας εκ των προτέρων εάν και πότε είναι πιθανόν να παρουσιαστεί μια ενδεχόμενη υποτροπή είμαστε σε θέση να τροποποιήσουμε κατάλληλα την θεραπευτική αγωγή του ασθενούς, δηλαδή, είτε εντείνοντας την αγωγή στους ασθενείς με υψηλή πιθανότητα επανεμφάνισης, είτε απαλλάσσοντας τους ασθενείς με καλή πρόγνωση από τις επιβλαβείς συνέπειες μιας εντατικής και επιθετικής αγωγής (π.χ. χημειοθεραπεία).

## *5.2 Εισαγωγή*

Όπως αναφέρθηκε και στο εισαγωγικό κεφάλαιο της διδακτορικής διατριβής, ο στοματικός καρκίνος αφορά νεοπλασίες που εμφανίζονται στην στοματική κοιλότητα, τον φάρυγγα και τον λάρυγγα. Καταλαμβάνει την 8$^{η}$ θέση παγκοσμίως στην κλίμακα κατάταξης καρκίνων με βάση την συχνότητα εμφάνισής τους, αφού περισσότερα από 500.000 άτομα διαγιγνώσκονται με καρκίνο του στόματος ετησίως [62]. Αποτελεί μια σύνθετη και πολυπαραγοντική νόσο που οι εκφάνσεις της εντοπίζονται και καταγράφονται στο κλινικό προφίλ του ασθενούς, σε απεικονιστικές εξετάσεις (π.χ. αξονική και μαγνητική τομογραφία) και φυσικά σε μοριακό επίπεδο στις διακυμάνσεις της έκφρασης των γονιδίων στην περιοχή του καρκίνου. Η διαρκής βελτίωση των πρωτοκόλλων θεραπείας (π.χ. χειρουργική αφαίρεση, χημειοθεραπεία, ακτινοθεραπεία) έχει επιτύχει ιδιαιτέρως υψηλά ποσοστά επιτυχίας ως προς την απαλλαγή από τον καρκίνο [65]. Το στάδιο που ακολουθεί την επιτυχή απαλοιφή των καρκινικών κυττάρων καθώς και των γειτονικών προσβεβλημένων λεμφαδένων, οδηγεί σε μια κατάσταση που καλείται *ύφεση,* όπου ψήγματα της ασθένειας δεν είναι πλέον δυνατόν να ανιχνευθούν. Είναι σημαντικό δε να τονιστεί ότι λόγω της επιθετικότητας και διεισδυτικότητας του συγκεκριμένου

καρκίνου, ακόμη και μετά την επιτυχή και αποτελεσματική αντιμετώπισή του (δηλαδή στο στάδιο της ύφεσης), το ποσοστό επανεμφάνισής του κυμαίνεται στο 25-48% [66].

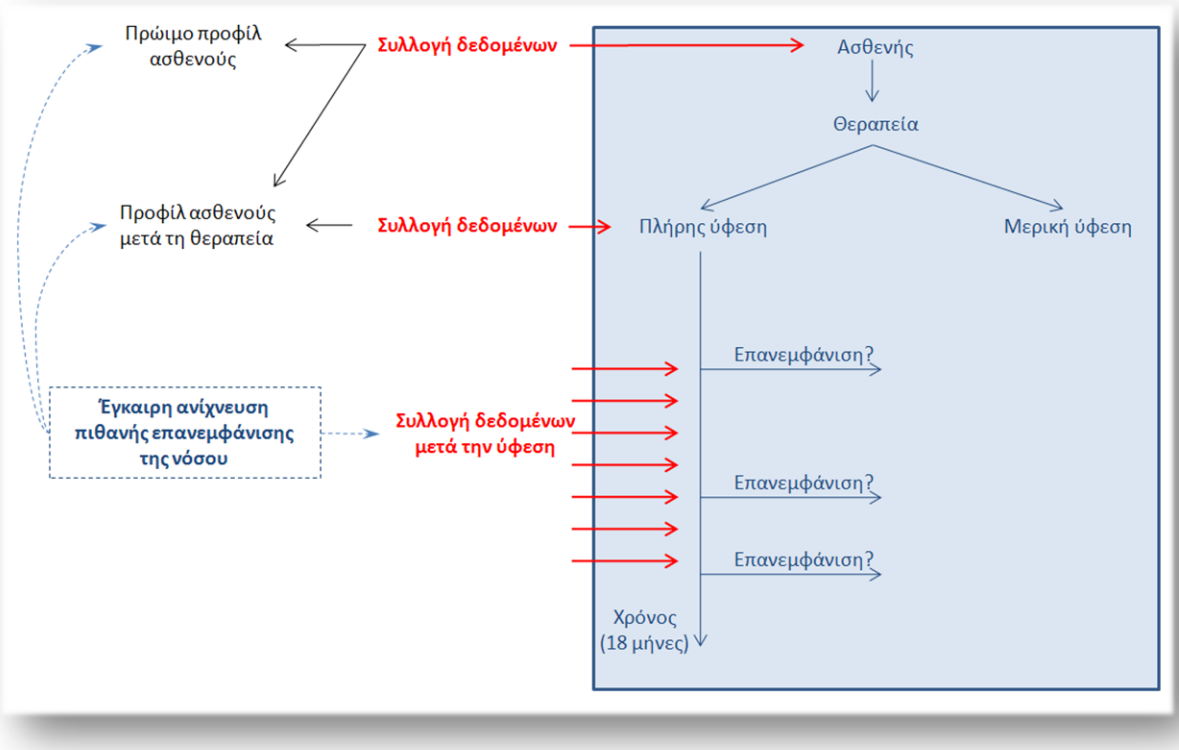
Μέχρι σήμερα, πλήθος παραγόντων έχουν ενοχοποιηθεί για την εμφάνιση και εξέλιξη της νόσου. Το κάπνισμα και η κατανάλωση αλκοόλ, συνιστούν τους βασικούς παράγοντες κινδύνου που έχουν συσχετισθεί με τον στοματικό καρκίνο, από κλινικής απόψεως [63]. Επίσης, έχει παρατηρηθεί ότι οι άνδρες έχουν διπλάσια πιθανότητα εμφάνισης της νόσου, καθιστώντας το φύλο έναν ακόμη σημαντικό προδιαθεσικό παράγοντα [63]. Για ειδικές εντοπίσεις νεοπλασιών στην στοματοφαρυγγική οδό, όπως είναι καρκίνοι στο χείλος ή στην βάση της γλώσσας, έχουν ενοχοποιηθεί η έκθεση στον ήλιο [63] και η μόλυνση από τον HPV [64], αντίστοιχα. Επίσης, όσον αφορά την μοριακή βάση της νόσου, οι γονιδιακοί παράγοντες που έχουν συσχετισθεί είναι σχετικά περιορισμένοι σε πλήθος καθώς και σε αποτελεσματικότητα [68, 69]. Προς αυτή την κατεύθυνση έχουν προταθεί διάφορες προσεγγίσεις στη βιβλιογραφία με σκοπό τον εντοπισμό επανεμφάνισης της νόσου, είτε σε γειτονικά κύτταρα, είτε απομακρυσμένες μεταστάσεις. Στις εργασίες [115, 116] εντοπίζεται ένα γενετικό προφίλ, που περιλαμβάνει έναν περιορισμένο αριθμό γονιδίων, τα οποία διακρίνουν ασθενείς με και χωρίς μετάσταση πρωτογενούς στοματικού καρκίνου στους γειτονικούς λεμφαδένες. Κατά παρόμοιο τρόπο, στην [117] επιχειρείται η έγκαιρη πρόβλεψη μιας πιθανής μετάστασης στοματικού καρκίνου σε απομακρυσμένο ιστό. Στις εργασίες [118-120] μελετώνται συγκεκριμένα νεοπλασίες στην γλώσσα, και γίνεται προσπάθεια να προσδιοριστεί η εξέλιξη της νόσου, ούτως ώστε να είμαστε σε θέση να προβλέψουμε μια ενδεχόμενη υποτροπή ή επανεμφάνισή της. Συνολικά, ο συνδυασμός των υπαρχόντων παραγόντων κινδύνου που έχουν εντοπιστεί από κλινικο-γενετική σκοπιά, καθώς και η ανακάλυψη νέων θα συμβάλλει σημαντικά στην κατανόηση της εξέλιξης του στοματικού καρκίνου και κατά συνέπεια στην έγκαιρη πρόβλεψη πιθανής υποτροπής.

## *5.3 Μεθοδολογία-αποτελέσματα*

### *Κλινικό σενάριο*

Για να γίνει κατανοητή η προτεινόμενη ανάλυση, είναι σημαντικό να διασαφηνιστεί το κλινικό σενάριο (Εικόνα 38) βάσει του οποίου έγινε η συλλογή των ετερογενών κλινικών,

απεικονιστικών και γενετικών δεδομένων. Αρχικά, ένας ασθενής διαγιγνώσκεται με στοματικό καρκίνο, οπότε και συλλέγουμε σε αυτό το στάδιο κλινικά δεδομένα που αφορούν το ιστορικό του ασθενούς, απεικονιστικά δεδομένα (αξονική και μαγνητική τομογραφία) από την πρωτογενή εστία του καρκίνου και τους γειτονικούς λεμφαδένες καθώς και γενετικά δεδομένα που ουσιαστικά περιλαμβάνουν την έκφραση μιας σειράς γονιδίων από τον ίδιο τον καρκινικό ιστό. Αυτά τα δεδομένα που προέρχονται από τα πρώτα στάδια της διάγνωσης, δομούν το "πρώιμο προφίλ" του ασθενούς. Εν συνεχεία, ο ασθενής υπόκειται σε κατάλληλη θεραπευτική αγωγή, που περιλαμβάνει είτε χειρουργική αφαίρεση του όγκου, είτε χημειο/ακτινο-θεραπεία, είτε και συνδυασμό αυτών. Έπειτα, από τους ασθενείς που οδηγήθηκαν σε πλήρη ύφεση της ασθένειας, συλλέγουμε σε τακτά διαστήματα δεδομένα που μεταβάλλονται με την πάροδο του χρόνου. Σε αυτά τα δεδομένα συγκαταλέγονται η έκφραση των γονιδίων από το αίμα που κυκλοφορεί στην περιοχή γύρω από τον αρχικό όγκο. Αυτό συμβαίνει γιατί μετά την αρχική εξάλειψη του όγκου ακολουθεί μια σειρά από συνεδρίες χημειοθεραπείας ή/και ακτινοθεραπείας που θεωρείται ότι δύναται να μεταβάλλει την έκφραση των γονιδίων στο αίμα. Συγκεκριμένα, τα παραπάνω δεδομένα λαμβάνονται από τον ασθενή κάθε τρεις μήνες μετά τη θεραπευτική αντιμετώπιση και για συνολικό διάστημα 18 μηνών (δηλαδή στους 0, 3, 6, 9, 12, 15 και 18 μήνες μετά την αρχική θεραπεία). Όλο το φάσμα των δεδομένων που συλλέγονται αναλύονται συστηματικά ώστε να μοντελοποιήσουμε την πορεία εξέλιξης της νόσου και εντοπίσουμε κατά το δυνατόν πιο έγκαιρα μια πιθανή επανεμφάνιση της νόσου.




**Εικόνα 38:** Κλινικό σενάριο.

## *Συλλογή δεδομένων*

Στην παρούσα ανάλυση μελετήθηκε η εξέλιξη του στοματικού καρκίνου σε ένα σύνολο 100 ασθενών που προήλθαν από δύο μεγάλες ογκολογικές κλινικές στην Ιταλία και Ισπανία. Σύμφωνα με τη σχετική βιβλιογραφία, σε ένα ποσοστό 70-80% εξ' αυτών αναμένεται να επιτευχθεί πλήρης ίαση οπότε και θα εισέλθουν στο στάδιο της ύφεσης. Από αυτούς τους ασθενείς περίπου 25-48% θα παρουσιάσουν υποτροπή και επανεμφάνιση της νόσου, μέσα στο χρονικό διάστημα των 18 μηνών που θα τους παρακολουθούμε και θα συλλέγουμε δεδομένα. Αυτός είναι και ο αντικειμενικός σκοπός της έρευνάς μας, δηλαδή, να αναλύσουμε τα δεδομένα που προέρχονται από αυτούς τους ασθενείς ώστε να καταφέρουμε να διακρίνουμε τους μεν από τους δε και να εντοπίσουμε τους παράγοντες με την υψηλότερη διακριτική ικανότητα.

Κατά την κατασκευή των συνόλων δεδομένων, τίθεται ένα ζήτημα που έχει να κάνει με την κατηγορία στην οποία ανήκει ο κάθε ασθενής. Παρόλο που για τους ασθενείς με

επανεμφάνιση η επιλογή είναι ξεκάθαρη, όσον αφορά την άλλη κατηγορία, που αποτελεί κατά κάποιον τρόπο και το σύνολο ελέγχου, η επιλογή ενέχει κάποια ασάφεια και αμφιβολία. Αυτό έγκειται στο γεγονός ότι κάθε ασθενής δυνητικά ανά πάσα στιγμή ενδέχεται να παρουσιάσει υποτροπή της νόσου που θα οδηγήσει σε πιθανή επανεμφάνιση. Επομένως, πρέπει να θέσουμε ένα χρονικό κατώφλι. Ασθενείς των οποίων η διάρκεια παρακολούθησης έχει ξεπεράσει το κατώφλι αυτό χωρίς υποψία υποτροπής θεωρούνται υγιείς και δομούν το σύνολο ελέγχου. Στην παρούσα ανάλυση, το κατώφλι αυτό επιλέχτηκε να είναι 12 μήνες μετά την αρχική θεραπεία του ασθενούς. Οι ασθενείς που είτε δεν έχουν παρουσιάσει υποτροπή είτε δεν έχουν παρακολουθηθεί για διάστημα μεγαλύτερο των 12 μηνών εξαιρούνται από την ανάλυση. Το σύνολο των ασθενών που λαμβάνονται υπόψη συνοψίζεται στον Πίνακας 17. Με έντονα (κόκκινα) φαίνονται οι ασθενείς με επανεμφάνιση ενώ με πράσινο οι ασθενείς που μέχρι στιγμής βρίσκονται ακόμη σε ύφεση. Κάθε ασθενής σηματοδοτείται με ένα νούμερο που αποσκοπεί στην ανωνυμοποίηση και διάκριση μεταξύ των ασθενών. Ο Πίνακας 17, όπως παρατηρούμε έχει διαιρεθεί σε τρία τμήματα, διαχωρίζοντας έτσι τα υποσύνολα των ασθενών με βάση τον τύπο δεδομένων που υπάρχουν διαθέσιμα και λαμβάνονται υπόψη στην παρούσα μελέτη.

**Πίνακας 17:** Σύνοψη των ασθενών που χρησιμοποιήθηκαν στην παρούσα μελέτη.

| **Κλινικά δεδομένα (σύνολο: 41)** | | | | | |
|---|---|---|---|---|---|
| 90 | 92 | **91** | **93** | 98 | 99 |
| 100 | 101 | 102 | 95 | **146** | 104 |
| 150 | **151** | **103** | 106 | **156** | 149 |
| 141 | 148 | 134 | 145 | 96 | 94 |
| **140** | 142 | 157 | **163** | **169** | 183 |
| 178 | 182 | 181 | 179 | 180 | 186 |
| **187** | **188** | **174** | 223 | **222** | |
| | | | | | |
| **Απεικονιστικά δεδομένα (σύνολο: 21)** | | | | | |
| 90 | 92 | **93** | 94 | 95 | 96 |
| 99 | 100 | 145 | **146** | 148 | 149 |

| | | | | | |
|---|---|---|---|---|---|
| **151** | **156** | **169** | **174** | 178 | 179 |
| 181 | 182 | 183 | | | |
| | | | | | |
| **Γενετικά δεδομένα (σύνολο: 25)** | | | | | |
| 90 | 92 | 94 | 95 | 96 | 98 |
| 99 | 100 | 101 | 102 | 104 | 106 |
| 134 | 145 | 148 | 149 | 150 | 157 |
| **151** | **156** | **146** | **103** | **93** | **163** |
| **174** | | | | | |

Κλινικά δεδομένα

Όσον αφορά τα κλινικά δεδομένα που συλλέγονται από κάθε ασθενή, αυτά συνιστούν πέρα από το κλασικό ιατρικό ιστορικό που λαμβάνεται, ένα σύνολο παραγόντων κινδύνου που συγκεκριμένα αφορούν τον στοματικό καρκίνο, την ταξινόμηση/σταδιοποίηση του καρκίνου σύμφωνα με το σύστημα TNM [121] καθώς και ένα πλήθος εξειδικευμένων εξετάσεων και βιολογικών δεικτών. Το σύνολο των κλινικών δεδομένων περιέχονται στον Πίνακας 18.

**Πίνακας 18:** Συνοπτική παράθεση των κλινικών χαρακτηριστικών.

| **Κλινικά χαρακτηριστικά** | |
|---|---|
| Ecog Status | Eating Habits |
| Weight | BMI |
| Height | Substance Exposition |
| Diabetes | Precancerous Lesions |
| Allergies | Duration |
| Cholesterol | Immunosuppressor Treatments Presence |
| Hypertension | Immuno Duration |
| Familiar History Of Malignance | Immuno Type |
| Smoker | Tumor Maximum Diameter |
| Smoking Habits | Tumor Thickness |
| Quantity Per Day | Depth Of Invasion |

| | |
|---|---|
| Smoking For | Basaloid Features |
| Ex Smoker | Lympho Plasmacytic Rection |
| Quitted Smoking | Lympho Plasmacytic Invasion |
| Alcohol | Perineural Invasion |
| Drinking Habits | Degree Of Cells Keratinisation |
| Mechanical Trauma | Nuclear Pleomorphism |
| Mobile Prosthesis | Number Of Mitoses Per 10HPF |
| Dental Cusps | Grade Of Differentation |
| Galvanic Current | Surgical Margins |
| Oral Hygiene | Martinez-Gimeno Score |
| Infection | Anneroths Mod Score |
| Type Of Infection | D2_40Stain |
| Physical Agents | P53_STAIN |
| Type Of Physical Agent | P16Ink4aStain |
| Diet Deficit | EGFR Stain |
| Fe Haematic Concentration | CyclinD1Stain |
| Plummer Vinson | Ki67Stain |
| Hb Haematic Concentration | HPV_DNA |
| B12 Vitamins Haematic Concentration | T Staging |
| A Vitamins Haematic Concentration | N Staging |
| E Vitamins Haematic Concentration | M Staging |
| Folati | |

Στον αμέσως επόμενο πίνακα (Πίνακας 19) φαίνονται αναλυτικά οι ασθενείς που λαμβάνονται υπόψη καθώς και οι μήνες παρακολούθησης του καθενός.

**Πίνακας 19:** Κατανομή των ασθενών με βάση τους μήνες παρακολούθησης.

| Ασθενής | Μήνες παρακολούθησης | Ασθενής | Μήνες παρακολούθησης |
|---|---|---|---|
| 90 | 18 | 145 | 18 |
| 92 | 12 | 96 | 18 |
| **91** | **3** | 94 | 15 |
| **93** | **9** | **140** | **12** |
| 98 | 18 | 142 | 15 |
| 99 | 15 | 157 | 18 |
| 100 | 15 | **163** | **9** |
| 101 | 12 | **169** | **6** |
| 102 | 12 | 183 | 12 |
| 95 | 12 | 178 | 18 |
| **146** | **15** | 182 | 15 |
| 104 | 12 | 181 | 21 |
| 150 | 15 | 179 | 21 |
| **151** | **9** | 180 | 18 |
| **103** | **6** | 186 | 24 |
| 106 | 18 | **187** | **15** |
| **156** | **6** | **188** | **3** |
| 149 | 12 | **174** | **6** |
| 141 | 15 | 223 | 24 |
| 148 | 12 | **222** | **6** |
| 134 | 18 | | |

Απεικονιστικά δεδομένα

Σε κάθε ασθενή επίσης πραγματοποιούνται αξονική και μαγνητική τομογραφία ώστε να διαπιστωθούν επακριβώς τα όρια του όγκου και των γειτονικών λεμφαδένων, καθώς και για να εξαχθούν κάποια επιπλέον χαρακτηριστικά που αφορούν την σύσταση, υφή και σχήμα αυτών. Τα χαρακτηριστικά που εξάγονται από κάθε εικόνα και χρησιμοποιούνται περαιτέρω για την διάκριση μεταξύ των ασθενών παρουσιάζονται στον Πίνακας 20.

**Πίνακας 20:** Σύνοψη των απεικονιστικών χαρακτηριστικών που χρησιμοποιούνται.

| Απεικονιστικά δεδομένα | |
|---|---|
| Contrast Take-Up Rate | Carotid Infiltration |
| Minor Axis Bigger 10mm | Cuteneous Invasion |
| Extra Nodal Spreading | Site |
| Shape Deviation | Side |
| Texture | Side Relative To Tumor |
| Water Content | Cluster |
| Necrosis | Number Of Lymph Nodes |
| Central Necrosis | Number Lymph Nodes Bigger 3 |
| Bone Infiltration | |

Στον Πίνακας 21 που ακολουθεί, παρουσιάζεται η κατανομή των ασθενών με βάση τους μήνες που βρίσκονται υπό παρακολούθηση χωρίς να σημειωθεί επανεμφάνιση ή για αυτούς που έχουν υποστεί υποτροπή ο μήνας της επανεμφάνισης.

**Πίνακας 21:** Κατανομή των ασθενών με βάση τον μήνα παρακολούθησης.

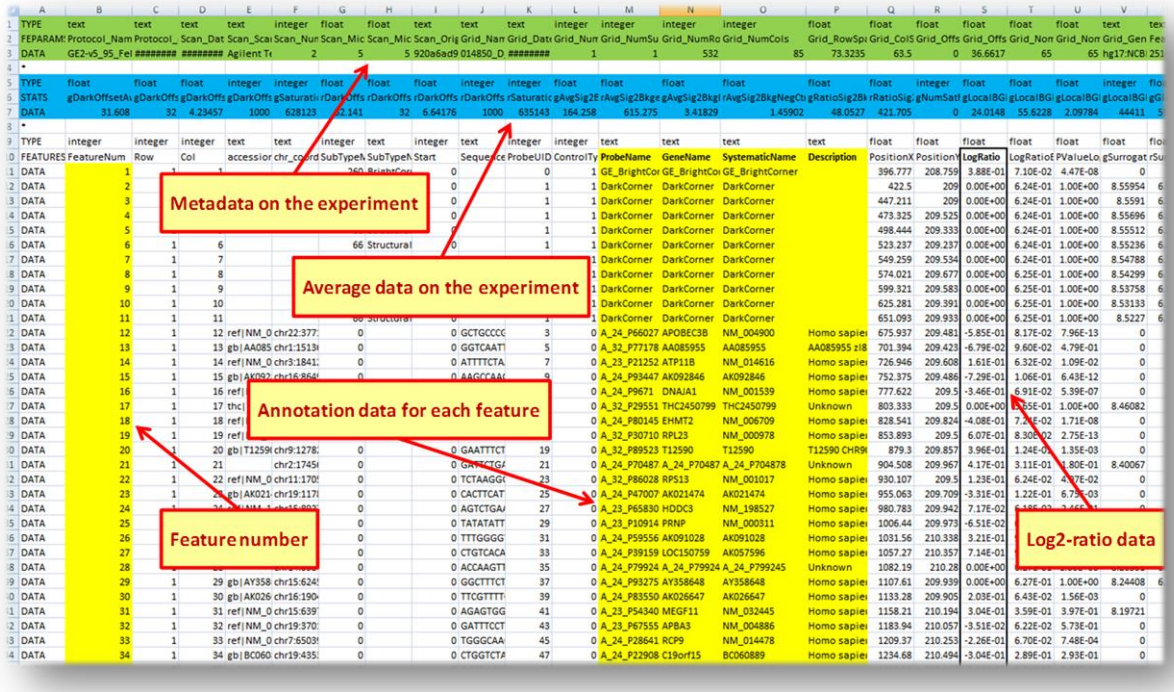
| Ασθενής | Μήνες παρακολούθησης |
|---|---|
| 90 | 18 |
| 92 | 9 |
| **93** | **9** |
| 94 | 15 |
| 95 | 12 |
| 96 | 18 |
| 99 | 15 |
| 100 | 15 |
| 145 | 18 |
| **146** | **15** |
| 148 | 12 |

| | |
|---|---|
| 149 | 12 |
| **151** | **9** |
| **156** | **6** |
| **169** | **6** |
| **174** | **6** |
| 178 | 18 |
| 179 | 21 |
| 181 | 21 |
| 182 | 15 |
| 183 | 12 |

Γενετικά δεδομένα

Η τρίτη κατηγορία δεδομένων που συλλέγουμε συνίσταται στην καταγραφή της γονιδιακής έκφρασης α) από δείγμα του καρκινικού ιστού που λαμβάνεται πριν ή κατά τη διάρκεια της θεραπείας και β) από το αίμα που κυκλοφορεί στην περιοχή του καρκίνου, το οποίο και λαμβάνεται σε διαδοχικά χρονικά διαστήματα μετά την θεραπεία και κατά τη διάρκεια της ύφεσης.

Και στις δύο περιπτώσεις λαμβάνεται η γονιδιακή έκφραση από 45015 γονίδια και αποθηκεύεται σε αρχεία FE (Feature extraction). Κάθε αρχείο FE περιλαμβάνει το όνομα του γονιδίου, τον βαθμό έκφρασής του, την κανονικοποιημένη τιμή έκφρασης και πλήθος άλλων μετα-δεδομένων που δεν μας απασχολούν στην παρούσα ανάλυση. Ένα τυπικό αρχείο FE παρουσιάζεται στην Εικόνα 39. Να σημειωθεί ότι σε όλες τις μετρήσεις, όλων των ασθενών, χρησιμοποιήθηκε η ίδια συσκευή εξαγωγής μικροσυστοιχιών με τις ίδιες ακριβώς ρυθμίσεις ώστε να αποφευχθεί οποιαδήποτε άλλη πηγή μεταβλητότητας πέρα από τις διακυμάνσεις στην έκφραση των γονιδίων.




**Εικόνα 39:** Τυπική μορφή ενός αρχείου FE.

Στην περίπτωση που καταγράφουμε την έκφραση των γονιδίων από το αίμα, όπως είναι προφανές ένα ξεχωριστό FE αρχείο προκύπτει για κάθε επίσκεψη του ασθενούς σύμφωνα με το προκαθορισμένο πρωτόκολλο, δηλαδή στους 0, 3, 6, 9, 12, 15 και 18 μήνες μετά την θεραπεία.

Στους πίνακες που ακολουθούν παρουσιάζεται η κατανομή των ασθενών με βάση τον μήνα παρακολούθησης που βρίσκονται. Όσον αφορά τα δεδομένα που λαμβάνονται κατά τα αρχικά στάδια της θεραπείας από τον καρκινικό ιστό η κατανομή των ασθενών υπάρχει στον Πίνακας 22, ενώ ο Πίνακας 23 δείχνει αντίστοιχα ην κατανομή των ασθενών όταν η γονιδιακή έκφραση λαμβάνεται από το αίμα.

**Πίνακας 22:** Κατανομή των ασθενών βάσει γενετικών δεδομένων από τον καρκινικό ιστό.

| Ασθενής | Μήνες παρακολούθησης |
|---|---|
| 90 | 18 |
| 92 | 12 |
| **93** | **9** |
| 94 | 15 |
| 95 | 12 |
| 96 | 18 |
| 98 | 18 |
| 99 | 15 |
| 100 | 15 |
| 101 | 12 |
| 102 | 12 |
| **103** | **6** |
| 104 | 12 |
| 106 | 18 |
| 134 | 18 |
| 145 | 18 |
| **146** | **15** |
| 148 | 12 |
| 149 | 12 |
| 150 | 15 |
| **151** | **9** |
| **156** | **6** |
| 157 | 18 |
| **163** | **9** |
| **174** | **6** |

**Πίνακας 23:** Κατανομή των ασθενών βάσει γενετικών δεδομένων από το αίμα.

| Ασθενής | Μήνες παρακολούθησης |
|---|---|
| 92 | 3 |
| **93** | **6** |
| 97 | 3 |
| 146 | 3 |
| 147 | 3 |
| **156** | **6** |

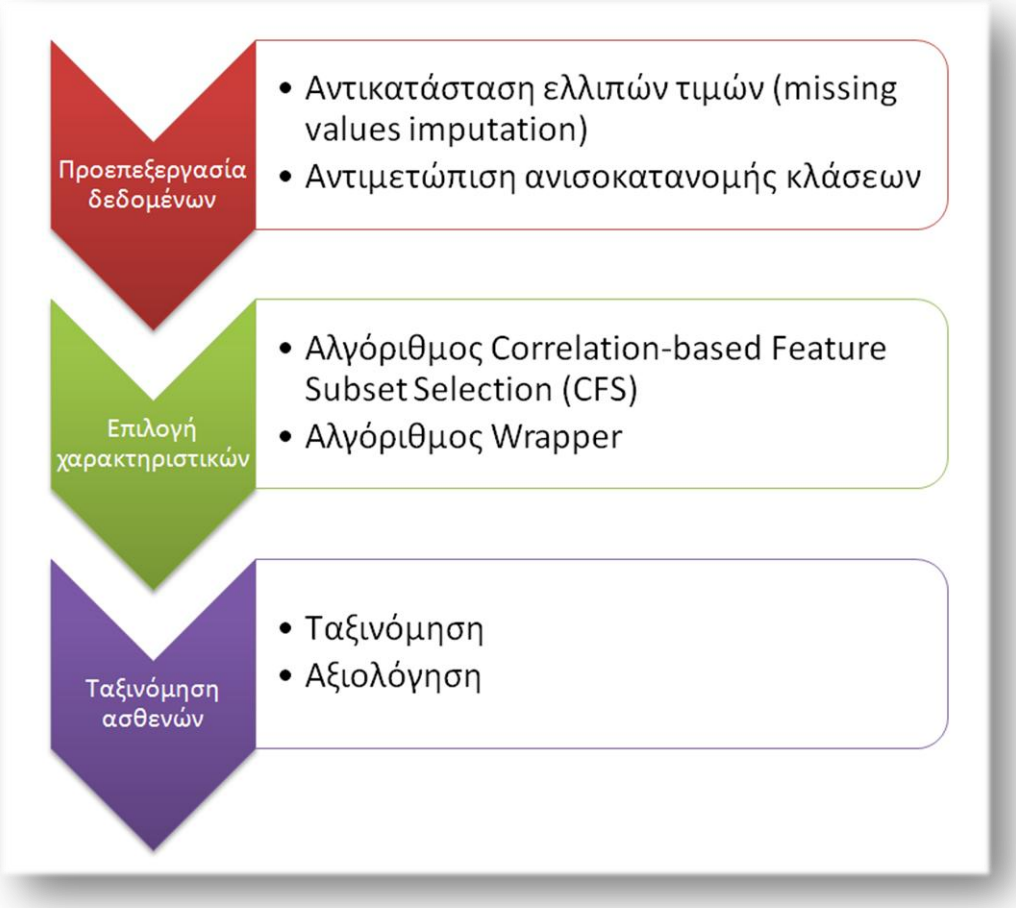
## *Πρόβλεψη επανεμφάνισης της νόσου*

Όπως αναφέραμε και προηγουμένως, η παρούσα ανάλυση δομείται σε δύο επίπεδα, α) την πρόβλεψη πιθανής επανεμφάνισης χρησιμοποιώντας δεδομένα από την αρχική περίοδο της νόσου, δηλαδή κατά την διάγνωση-θεραπεία, και β) την παρακολούθηση και μοντελοποίηση της εξέλιξης της νόσου στο διάστημα που έπεται της θεραπείας και για ένα διάστημα 18 μηνών. Στην ενότητα αυτή θα εστιάσουμε στο πρώτο επίπεδο της ανάλυσης, δηλαδή, αξιοποιώντας δεδομένα που προέρχονται αμιγώς από τα πρώτα στάδια της νόσου και συνιστούν το "πρώιμο προφίλ" του ασθενούς, υπολογίζουμε την πιθανότητα επανεμφάνισης του καρκίνου. Συγκεκριμένα, για τον σκοπό αυτό χρησιμοποιούμε κλινικά, απεικονιστικά και γενετικά δεδομένα (γονιδιακή έκφραση από τον καρκινικό ιστό), με σκοπό να εντοπίσουμε από κάθε τύπο δεδομένων τα πιο σημαντικά χαρακτηριστικά που καταδεικνύουν μια πιθανή υποτροπή και στη συνέχεια να υπολογίσουμε την πιθανότητα της υποτροπής. Χρησιμοποιώντας αρχικά κάθε τύπο δεδομένων εκ περιτροπής, δομούμε κάθε φορά έναν ξεχωριστό ταξινομητή, και στη συνέχεια υλοποιούμε ένα μετα-ταξινομητή που συνδυάζει τις επιμέρους προβλέψεις βάσει πλειοψηφίας.

### Ανάλυση κλινικών δεδομένων

Σε αυτήν την ενότητα, χρησιμοποιώντας αποκλειστικά τα κλινικά χαρακτηριστικά που λαμβάνονται από κάθε ασθενή κατά τα πρώτα στάδια της διάγνωσης-θεραπείας

υλοποιούμε έναν ταξινομητή που υπολογίζει βάσει αυτών των δεδομένων την πιθανότητα ενός ασθενούς να παρουσιάσει επανεμφάνιση του καρκίνου. Στην Εικόνα 40 βλέπουμε τα βήματα που ακολουθούνται κατά την ανάλυση των κλινικών δεδομένων.




**Εικόνα 40:** Μεθοδολογία ανάλυσης κλινικών δεδομένων.

Αρχικά, τα δεδομένα εισόδου υφίστανται προεπεξεργασία που αφορά στην αντικατάσταση των αγνώστων/ελλιπών τιμών. Συγκεκριμένα τιμές που λείπουν από κατηγορικά και αριθμητικά χαρακτηριστικά αντικαθίστανται από την επικρατούσα τιμή και τον μέσο όρο αντίστοιχα. Να σημειωθεί ότι τα χαρακτηριστικά που παρουσίαζαν ελλιπείς τιμές σε ποσοστό μεγαλύτερο από το 90% των ασθενών, εξαιρέθηκαν από την μελέτη. Τα χαρακτηριστικά αυτά φαίνονται στον Πίνακας 24.

**Πίνακας 24:** Κλινικά χαρακτηριστικά στα οποία >90% ασθενών είχαν ελλιπείς τιμές.

| Duration | Fe Haematic Concentration |
|---|---|
| Immuno Duration | Hb Haematic Concentration |
| Immuno Type | B12 Vitamins Haematic Concentration |
| Quitted Smoking | A Vitamins Haematic Concentration |
| Type Of Infection | E Vitamins Haematic Concentration |
| Type Of Physical Agent | Folati |

Στη συνέχεια ανακύπτει το πρόβλημα της ανισοκατανομής των κλάσεων. Όπως είδαμε στον Πίνακας 19, οι 41 διαθέσιμοι ασθενείς χωρίζονται σε δύο κλάσεις, αυτούς που έχουν επανεμφάνιση και αυτούς που δεν έχουν, όπου η κάθε κλάση περιέχει 13 και 28 ασθενείς, αντίστοιχα. Από μια δειγματοληπτική ματιά, τρεις προσεγγίσεις μπορούν να εφαρμοστούν, υπερδειγματοληψία της μειοψηφούσας κλάσης, υποδειγματοληψία της πλειοψηφούσας κλάσης, και συνδυασμός αυτών. Η πρώτη προσέγγιση ενέχει το μειονέκτημα ότι πιθανώς θορυβώδη δείγματα πολλαπλασιάζονται υποβαθμίζοντας την ποιότητα του συνόλου εισόδου, κατά την δεύτερη προσέγγιση απαλείφουμε πιθανώς πολύτιμα δεδομένα από ασθενείς που δεν έχουν επανεμφάνιση, ενώ η τρίτη προσέγγιση ενδεχόμενα συνδυάζει τα μειονεκτήματα και των δύο τεχνικών [21]. Εφαρμόζουμε λοιπόν τον αλγόριθμο SMOTE (Synthetic Minority Oversampling Technique) [122], ο οποίος ουσιαστικά κατασκευάζει δείγματα της μειοψηφούσας κλάσης, χρησιμοποιώντας βασιζόμενος σε μια τεχνική k-NN (k-κοντινότερων γειτόνων). Κατ' αυτόν τον τρόπο λαμβάνουμε υπόψη όλα τα διαθέσιμα δείγματα της πλειοψηφούσας κλάσης, χωρίς να πολλαπλασιάζουμε μεμονωμένα πιθανώς λανθασμένα δείγματα, αφού τα νέα δείγματα που κατασκευάζονται προκύπτουν από γραμμικό συνδυασμό όλων των διαθέσιμων δειγμάτων της μειοψηφούσας κλάσης. Το σύνολο δεδομένων που προκύπτει αποτελείται πλέον από 56 ασθενείς, ισοκατανεμημένους στις δύο κλάσεις.

Όπως είναι λογικό, ανάμεσα στα κλινικά χαρακτηριστικά που παρουσιάζονται στον Πίνακας 18, κάποια από αυτά εμφανίζουν μεγάλη συσχέτιση μεταξύ τους, ενώ κάποια άλλα ενδεχομένως να μην είναι ιδιαιτέρως συσχετισμένα με την κλάση που επιθυμούμε να προβλέψουμε. Σκοπός μας λοιπόν είναι η αφαίρεση αυτών των περιττών και πλεοναζόντων χαρακτηριστικών που δυσχεραίνουν την ταξινόμηση, καθώς και η ανάδειξη

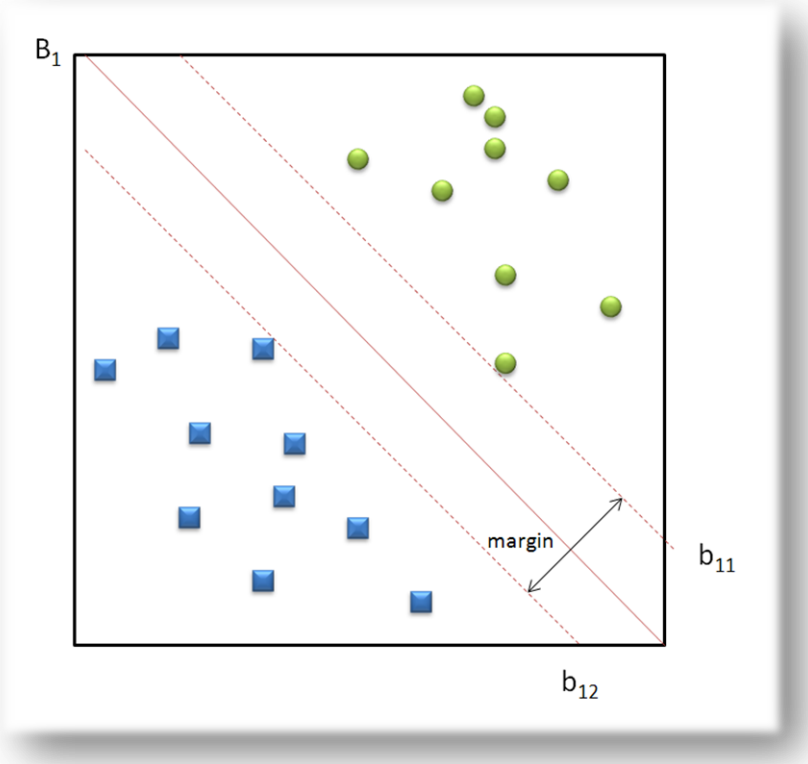
των χαρακτηριστικών που εμφανίζουν την μεγαλύτερη διακριτική ικανότητα και συνεπώς αποτελούν καλούς δείκτες κατά την προσπάθεια πρόβλεψης πιθανών επανεμφανίσεων μεταξύ των ασθενών. Για τον σκοπό αυτό χρησιμοποιούμε δύο αλγορίθμους επιλογής χαρακτηριστικών, τον αλγόριθμο CFS (Correlation-Based Feature Selection) [123, 124] και τον αλγόριθμο wrapper [84]. Ο αλγόριθμος CFS ουσιαστικά υπολογίζει την συσχέτιση μεταξύ των χαρακτηριστικών εισόδου καθώς και την συσχέτιση του κάθε χαρακτηριστικού με την κλάση. Προφανώς, η έξοδος του τελικά συνίσταται από αυτά τα χαρακτηριστικά που φέρουν μικρή συσχέτιση μεταξύ τους και μεγάλη συσχέτιση με την κλάση. Ο αλγόριθμος wrapper αποτελεί μια κατηγορία αλγορίθμων ταξινόμησης. Παρόλο που η βασική ιδέα πάνω στην οποία δομείται η ροή λειτουργίας του είναι αρκετά απλή, το τελικό σύνολο χαρακτηριστικών που δίνει ως έξοδο παρουσιάζει πολύ καλά αποτελέσματα και υπερτερεί των υπολοίπων αλγορίθμων επιλογής χαρακτηριστικών. Συγκεκριμένα, ο wrapper αλγόριθμος κατασκευάζει όλους τους πιθανούς συνδυασμούς χαρακτηριστικών εισόδου, και τους αξιολογεί χρησιμοποιώντας εσωτερικά τον τελικό ταξινομητή. Το τελικό σύνολο χαρακτηριστικών που αποδίδει στην έξοδο είναι αυτό που πέτυχε το υψηλότερο ποσοστό ακρίβειας κατά την ταξινόμηση.

Στο επόμενο βήμα υλοποιούμε μια σειρά από αλγόριθμους ταξινόμησης με σκοπό να διαχωρίσουμε τους ασθενείς με βάση την πιθανότητα επανεμφάνισης καρκίνου, σε δύο σύνολα, αυτούς με υψηλή και αυτούς με χαμηλή πιθανότητα επανεμφάνισης. Συγκεκριμένα, οι αλγόριθμοι που αναπτύχθηκαν είναι οι εξής: δίκτυα Bayes (Bayes Network: BN), Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks: ANN), Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines: SVM), δέντρα απόφασης (Decision Trees: DT) και τυχαία δάση (Random Forests: RF).

Τα δίκτυα Bayes έχουν χρησιμοποιηθεί ευρέως για την αναπαράσταση γνώσης, όπως επίσης και για ταξινόμηση, σε ένα ευρύ πεδίο ετερόκλητων εφαρμογών, μεταξύ των οποίων και η ιατρική [125]. Η δομή ενός δικτύου Bayes ουσιαστικά είναι ένας κατευθυνόμενος ακυκλικός γράφος (Directed Acyclic Graph: DAG), όπου οι κόμβοι του αντιστοιχούν σε μεταβλητές και οι κατευθυνόμενες ακμές σηματοδοτούν σχέση εξάρτησης μεταξύ των μεταβλητών που ενώνουν. Η γραφική αναπαράσταση ενός δικτύου Bayes επιτρέπει την εύληπτη απεικόνιση των συσχετίσεων και εξαρτήσεων που ενυπάρχουν μεταξύ των μεταβλητών. Επίσης, πολλά ευρέως γνωστά μοντέλα, όπως Kalman φίλτρα και μοντέλα αυτοσυσχέτισης (autoregressive models) μπορούν να αναπαρασταθούν και να υλοποιηθούν ως δίκτυα Bayes.

Τα τεχνητά νευρωνικά δίκτυα (ANN) αποτελούν έναν από τους πιο δημοφιλείς ταξινομητές, που έχουν χρησιμοποιηθεί σε πλήθος εφαρμογών [126]. Το νευρωνικό δίκτυο είναι ένα δίκτυο από απλούς υπολογιστικούς κόμβους (νευρώνες) διασυνδεδεμένους μεταξύ τους. Το μοντέλο είναι εμπνευσμένο από τα βιοηλεκτρικά δίκτυα που δημιουργούνται στον εγκέφαλο ανάμεσα στους νευρώνες (νευρικά κύτταρα) και στις συνάψεις (σημεία επαφής των νευρικών απολήξεων). Στο μαθηματικό μοντέλο των νευρωνικών δικτύων υπάρχουν κομβικά σημεία (nodes) στα οποία καταλήγουν συνδέσεις από άλλους κόμβους του δικτύου, στις οποίες συνήθως αποδίδεται κάποιο βάρος. Οι νευρώνες είναι το δομικό στοιχείο του δικτύου. Υπάρχουν δύο είδη νευρώνων, οι νευρώνες εισόδου και οι υπολογιστικοί νευρώνες. Οι νευρώνες εισόδου δεν υπολογίζουν τίποτα, μεσολαβούν ανάμεσα στις εισόδους του δικτύου και τους υπολογιστικούς νευρώνες. Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν τις εισόδους τους με τη χρήση βαρών και υπολογίζουν το άθροισμα του γινομένου. Το άθροισμα που προκύπτει είναι το όρισμα της συνάρτησης μεταφοράς. Πρακτικά, ένα νευρωνικό δίκτυο βελτιστοποιεί μία συνάρτηση, σύμφωνα με κάποιους περιορισμούς, δίδοντας ως έξοδο μια σειρά πιθανοτήτων, βάσει των οποίων επιτελεί ταξινόμηση.

Τα SVM, που έχουν επίσης εξηγηθεί και σε προηγούμενο κεφάλαιο, αποτελούν έναν από τους πιο αποτελεσματικούς αλγορίθμους που εφαρμόζονται για ταξινόμηση και έχουν εφαρμοστεί σε πληθώρα προβλημάτων [127]. Έχουν τις ρίζες τους στη Στατιστική, και ουσιαστικά αναζητούν το υπερεπίπεδο που διαχωρίζει τα δείγματα δύο ή περισσοτέρων κατηγοριών με το μικρότερο σφάλμα. Το τελικό υπερεπίπεδο επιλέγεται ώστε να απέχει κατά το δυνατόν περισσότερο από τα πιο δυσδιάκριτα δείγματα κάθε κατηγορίας (Εικόνα 41). Μια επίσης πολύ ενδιαφέρουσα ιδιότητα των SVM, είναι ο μετασχηματισμός που μπορούν να επιτελούν στα δείγματα εισόδου ώστε να τα απεικονίσουν σε έναν δειγματοχώρο μεγαλύτερης διάστασης όπου τα δείγματα είναι γραμμικά διαχωρίσιμα.




**Εικόνα 41:** Το SVM επιλέγει το υπερεπίπεδο που ελαχιστοποιεί το σφάλμα ταξινόμησης.

Τα δέντρα απόφασης αποτελούν μια διαφανή και εύληπτη διαδικασία ταξινόμησης όπου βασίζεται σε ένα σύνολο ελέγχων. Συγκεκριμένα, κάθε κόμβος του δέντρου αντιπροσωπεύει έναν έλεγχο του οποίου οι εκβάσεις έχουν καθοριστεί κατά την διαδικασία εκπαίδευσης. Η διαδικασία επαγωγής με τη σειρά της αφορά στον διαδοχικό έλεγχο των τιμών του διανύσματος εισόδου από όλους τους κόμβους του δέντρου, καταλήγοντας σε έναν τερματικό κόμβο-φύλλο όπου περιέχει και την προβλεπόμενη κλάση στην οποία κατηγοριοποιείται το δείγμα εισόδου. Το μεγάλο πλεονέκτημα των δέντρων απόφασης είναι η διαφανής αρχιτεκτονική τους που προσφέρει επεξήγηση των ληφθέντων αποφάσεων όμως αντισταθμίζεται από το γεγονός ότι έχουν τη δυνατότητα να οριοθετούν μονάχα γραμμικά διαχωριστικά υπερεπίπεδα.

Τα τυχαία δάση [128] είναι ταξινομητές που δομούνται από μια συλλογή δέντρων απόφασης. Έχουν ανακύψει σχετικά πρόσφατα στην βιβλιογραφία όμως τα υψηλά αποτελέσματα που δίνουν, έχουν ωθήσει στην ευρεία εφαρμογή τους σε πολλά και ετερόκλητα επιστημονικά πεδία [129]. Κάθε τυχαίο δάσος, όπως θα περίμενε κανείς και διαισθητικά, αποτελείται από πολλά δέντρα απόφασης. Κατά την διαδικασία ταξινόμησης, κατασκευάζονται πολλοί μεμονωμένοι ταξινομητές (δέντρα απόφασης) βάσει υποσυνόλων

του διανύσματος εισόδου οι οποίοι και εκχωρούν για κάθε δεδομένο διάνυσμα εισόδου μία ψήφο για την απόφαση που υπολογίζουν ως πιο πιθανή στην έξοδο. Εν συνεχεία ο αλγόριθμος των τυχαίων δασών συγκεντρώνει τις μεμονωμένες ψήφους και δίνει ως έξοδο την κλάση με τις περισσότερες ψήφους. Η δομή των τυχαίων δασών που υλοποιούν ταξινομητές με υποσύνολα των χαρακτηριστικών του διανύσματος εισόδου, συμβάλλει στον εντοπισμό αλληλεπιδράσεων μεταξύ των μεταβλητών που ενδεχομένως να παρέμεναν απαρατήρητες στο συνολικό διάνυσμα εισόδου.

Για την αξιολόγηση και ποσοτική σύγκριση των μεθόδων που εφαρμόσαμε στο σύνολο εισόδου, χρησιμοποιήθηκαν τα μέτρα: ακρίβεια, ευαισθησία και ειδίκευση, τα οποία ορίζονται αμέσως πιο κάτω στις εξισώσεις 14-16:

$$\textit{Ευαισθησία} = TP/(TP+FN), \tag{14}$$

$$\textit{Ειδίκευση} = TN/(FP+TN), \tag{15}$$

$$\textit{Ακρίβεια} = (TP+TN)/(TP+TN+FP+FN), \tag{16}$$

όπου οι όροι *TP*, *TN*, *FP*, *FN* ορίζονται βάσει του Πίνακας 25:

**Πίνακας 25:** Βοηθητικός πίνακας για την εξαγωγή των στατιστικών μέτρων αξιολόγησης.

| | | **Πραγματική κλάση** | |
|---|---|---|---|
| | | Θετικό | Αρνητικό |
| **Προβλεπόμενη κλάση** | Θετικό | True Positive (*TP*) | False Positive (*FP*) |
| | Αρνητικό | False Negative (*FN*) | True Negative (*TN*) |

Η μέθοδος που χρησιμοποιήθηκε για αξιολόγηση είναι η 10-fold cross validation, η οποία χρησιμοποιείται κυρίως σε σύνολα δεδομένων με περιορισμένο αριθμό δειγμάτων, όπως συμβαίνει και στην περίπτωσή μας. Η συγκεκριμένη μέθοδος έχει περιγραφεί και σε προηγούμενη ενότητα, υπενθυμίζουμε λοιπόν ότι η βασική ροή εφαρμογής της κατακερματίζει το σύνολο δειγμάτων σε 10 τυχαία ισοκατανεμημένα υποσύνολα δεδομένων, όπου χρησιμοποιεί τα 9/10 για εκπαίδευση και στη συνέχεια εφαρμόζει το εκπαιδευμένο μοντέλο στο υπολειπόμενο 1/10 που αποτελεί και το σύνολο ελέγχου. Η ίδια

διαδικασία επαναλαμβάνεται 10 φορές, κυλίοντας τα 10 υποσύνολα που έχουν προκύψει, ώστε το κάθε ένα να έχει χρησιμοποιηθεί εκ περιτροπής ακριβώς μια φορά ως σύνολο ελέγχου.

Συνολικά, λοιπόν, οι συνδυασμοί που δοκιμάστηκαν και αξιολογήθηκαν χρησιμοποιώντας τα κλινικά χαρακτηριστικά ως είσοδο, αφορούν την αντικατάσταση των ελλιπών τιμών, την εφαρμογή των χαρακτηριστικών εισόδου είτε χωρίς την εφαρμογή αλγορίθμου επιλογής, είτε χρησιμοποιώντας τους αλγορίθμους CFS και wrapper, και τέλος κατά την ταξινόμηση χρησιμοποιήθηκαν οι αλγόριθμοι: δίκτυα Bayes (BN), Τεχνητά Νευρωνικά Δίκτυα (ANN), Μηχανές Διανυσμάτων Απόφασης (SVM), δέντρα απόφασης (DT) και τυχαία δάση (RF). Τα αποτελέσματα που προέκυψαν σε κάθε περίπτωση, παρουσιάζονται στους πίνακες που ακολουθούν.

Ο Πίνακας 26 περιέχει τα αποτελέσματα που προέκυψαν χωρίς επιλογή χαρακτηριστικών και χωρίς να πραγματοποιήσουμε αντικατάσταση ελλιπών τιμών.

**Πίνακας 26:** Αποτελέσματα χωρίς αντικατάσταση ελλιπών τιμών και χωρίς την εφαρμογή αλγορίθμου για επιλογή χαρακτηριστικών.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 78.6 | 82.1 | 75 |
| **ANN** | 64.3 | 64.3 | 64.3 |
| **SVM** | 66.1 | 75 | 57.1 |
| **DT** | 67.9 | 78.6 | 57.1 |
| **RF** | 66.1 | 75 | 57.1 |

Ο Πίνακας 27 αμέσως πιο κάτω περιέχει αποτελέσματα χωρίς αντικατάσταση ελλιπών και αγνώστων τιμών, και μετά την εφαρμογή του αλγορίθμου CFS για επιλογή χαρακτηριστικών. Τα χαρακτηριστικά που διατηρήθηκαν από τον αλγόριθμο επιλογής χαρακτηριστικών ως πιο σημαντικά είναι τα εξής: hypertension, familiar history of malignance, infection, basaloid features, lymphoplasmacytic reaction, lymphovascular invasion, surgical margins and T staging.

**Πίνακας 27:** Αποτελέσματα χωρίς αντικατάσταση αγνώστων τιμών και μετά την εφαρμογή του αλγορίθμου CFS για επιλογή χαρακτηριστικών.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 80 | 78.6 | 82.1 |
| **ANN** | 73.2 | 78.6 | 67.9 |
| **SVM** | 76.8 | 82.1 | 71.4 |
| **DT** | 67.9 | 75 | 60.7 |
| **RF** | 78.6 | 78.6 | 78.6 |

Ο Πίνακας 28 περιέχει τα αποτελέσματα που έδωσε ο κάθε αλγόριθμος, χωρίς να πραγματοποιήσουμε αντικατάσταση ελλιπών τιμών και μετά την εφαρμογή του αλγορίθμου wrapper για επιλογή χαρακτηριστικών. Συγκεκριμένα τα χαρακτηριστικά που διατηρήθηκαν ως πιο σημαντικά και με τη μεγαλύτερη διακριτική ικανότητα για κάθε ταξινομητή είναι, για το BN: oral hygiene, basaloid features, lymphoplasmacytic reaction και lymphovascular invasion; για τα ANN: galvanic current, precancerous lesions, basaloid features, number of mitoses per 10HPF και KI67 stain; για τα SVM: familiar history of malignance, mobile prosthesis, oral hygiene, infection and lymphoplasmacytic reaction; για τα DT: familiar history of malignance, mobile prosthesis, number of mitoses per 10HPF και P16Ink4a Stain; και για τα RF: cholsesterol, smoking habits, ex-smoker, mobile prosthesis, basaloid features, lymphoplasmacytic reaction, degree of cell keratinisation, Martinez-Gimeno score και D240 stain.

**Πίνακας 28:** Αποτελέσματα που προέκυψαν χωρίς αντικατάσταση ελλιπών τιμών και επιλογή χαρακτηριστικών με τον αλγόριθμο wrapper.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 80 | 78.6 | 82.1 |
| **ANN** | 75 | 78.6 | 71.4 |
| **SVM** | 87.5 | 89.3 | 85.7 |
| **DT** | 75 | 82.1 | 67.9 |
| **RF** | 82.1 | 82.1 | 82.1 |

Στην συνέχεια αντικαθιστούμε ελλιπείς και άγνωστες τιμές, με τις επικρατούσες τιμές και τους μέσους όρους για τα κατηγορικά και αριθμητικά χαρακτηριστικά, αντίστοιχα. Στο διάνυσμα εισόδου που προκύπτει εφαρμόζουμε και πάλι τα βήματα που εφαρμόστηκαν και παραπάνω, και τα αποτελέσματα παρατίθενται στους επόμενους πίνακες.

Συγκεκριμένα, στον Πίνακας 29, βλέπουμε τα αποτελέσματα που λάβαμε με κάθε αλγόριθμο ταξινόμησης, χωρίς να εκτελέσουμε προηγουμένως επιλογή χαρακτηριστικών.

**Πίνακας 29:** Αποτελέσματα που προέκυψαν έπειτα από αντικατάσταση αγνώστων τιμών, χωρίς χρήση αλγορίθμου επιλογής χαρακτηριστικών.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 78.6 | 78.6 | 78.6 |
| **ANN** | 73.2 | 75 | 71.4 |
| **SVM** | 71.4 | 71.4 | 71.4 |
| **DT** | 62.5 | 64.3 | 60.7 |
| **RF** | 67.9 | 71.4 | 64.3 |

Ακολούθως, στον Πίνακας 30, βλέπουμε τα αποτελέσματα που λάβαμε, αφού εφαρμόσαμε τον αλγόριθμο CFS για επιλογή χαρακτηριστικών. Τα χαρακτηριστικά που καταδείχτηκαν ως πιο σημαντικά είναι: hypertension, familiar history of malignance, mobile prosthesis, oral hygiene, infection, eating habits, basaloid features, grade differentiation και T staging.

**Πίνακας 30:** Αποτελέσματα που προέκυψαν έπειτα από την εφαρμογή του αλγορίθμου CFS για επιλογή χαρακτηριστικών.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 75 | 85.7 | 64.3 |
| **ANN** | 71.4 | 82.1 | 60.7 |
| **SVM** | 75 | 78.6 | 71.4 |
| **DT** | 62.5 | 57.1 | 67.9 |
| **RF** | 66.1 | 78.6 | 53.6 |

Τέλος, στον Πίνακας 31 παρατίθενται τα αποτελέσματα που προέκυψαν μετά την εφαρμογή του αλγορίθμου wrapper για επιλογή χαρακτηριστικών. Τα συγκεκριμένα χαρακτηριστικά που διατηρήθηκαν ως πιο σημαντικά και με τη μεγαλύτερη διακριτική ικανότητα για κάθε έναν από τους ταξινομητές είναι τα εξής: BN: ecog status, familiar history of malignance, smoking habits, mobile prosthesis, galvanic current, infection, basaloid features, lymphoplasmacytic reaction, grade differentiation and EGFR stain; για τα ANN: familiar history of malignance, smoking duration, infection, eating habits, substance exposition, basaloid features, perineural invasion, nuclear pleomorphism and T staging; για τα SVM: lymphoplasmacytic reaction, degree of keratinisation and N staging; για τα DT: ecog status, familiar history of malignance, smoking duration, mobile prosthesis, oral hygiene, lymphoplasmacytic reaction and perineural invasion; και τέλος για τα RF: ex-smoker, oral hygiene, grade differentiation, Martinez Gimeno and EGFR stain.

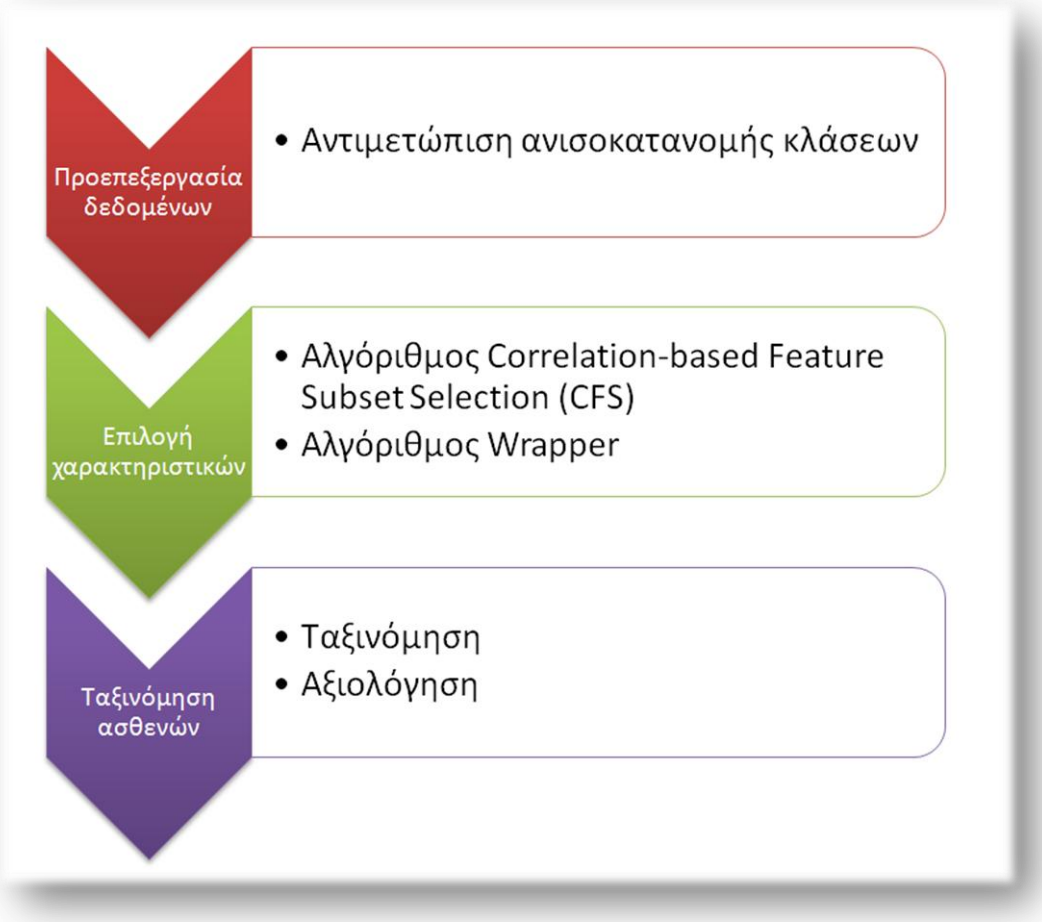
**Πίνακας 31:** Αποτελέσματα που προέκυψαν μετά την εφαρμογή του αλγορίθμου wrapper για επιλογή χαρακτηριστικών.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 82.1 | 85.7 | 78.6 |
| **ANN** | 83.9 | 82.1 | 85.7 |
| **SVM** | 82.1 | 85.7 | 78.6 |
| **DT** | 80.4 | 89.3 | 71.4 |
| **RF** | 78.6 | 71.4 | 85.7 |

Συγκρίνοντας τα αποτελέσματα που προέκυψαν πριν και μετά την αντικατάσταση των αγνώστων τιμών παρατηρούμε ότι δεν υπάρχουν σημαντικές διακυμάνσεις μεταξύ των δύο περιπτώσεων. Σημαντική δε παρατήρηση αποτελεί το γεγονός ότι η επιλογή ευνοεί τα αποτελέσματα της ταξινόμησης, ιδίως στην περίπτωση που χρησιμοποιείται ο αλγόριθμος wrapper. Παρόμοια είναι η κατάσταση είτε επιτελούμε αντικατάσταση αγνώστων τιμών είτε όχι.

Ανάλυση απεικονιστικών δεδομένων

Στη συνέχεια υλοποιούμε μια μεθοδολογία για να υπολογίσομε την πιθανότητα επανεμφάνισης καρκίνου, χρησιμοποιώντας αμιγώς απεικονιστικά δεδομένα, δηλαδή χαρακτηριστικά που εξάγονται από εικόνες αξονικής και μαγνητικής τομογραφίας. Τα στάδια της προτεινόμενης μεθοδολογίας παρατίθεται στην Εικόνα 42.




**Εικόνα 42:** Μεθοδολογία ανάλυσης των απεικονιστικών δεδομένων.

Κατά παρόμοιο τρόπο με τα κλινικά δεδομένα, και τα απεικονιστικά υφίστανται αρχικά κάποια στάδια προεπεξεργασίας ώστε να βελτιωθεί η ποιότητα του διανύσματος εισόδου. Εφαρμόζουμε λοιπόν τον αλγόριθμο SMOTE που με βάσει τα υπάρχοντα δείγματα της κάθε κλάσης, κατασκευάζει νέα μέχρις ότου οι δύο κλάσεις να εκπροσωπούνται εξίσου στο τελικό σύνολο δεδομένων. Ακολούθως, είτε χρησιμοποιούμε το διάνυσμα εισόδου ως έχει για να προβλέψουμε την πιθανότητα κάθε ασθενούς να παρουσιάσει επανεμφάνιση της ασθένειας, είτε εφαρμόζουμε αλγορίθμους επιλογής

χαρακτηριστικών για να εντοπίσουμε εκείνα τα χαρακτηριστικά που συνιστούν καθοριστικούς παράγοντες ως προς την επανεμφάνιση του καρκίνου. Συγκεκριμένα οι αλγόριθμοι επιλογής χαρακτηριστικών που εφαρμόζονται είναι όπως και προηγούμενα, ο CFS και ο αλγόριθμος wrapper. Σε κάθε περίπτωση, δηλαδή, είτε χωρίς επιλογή χαρακτηριστικών, είτε έπειτα από την εφαρμογή των αλγορίθμων CFS ή wrapper, το διάνυσμα εισόδου που προκύπτει εισάγεται σε μια σειρά ταξινομητών ώστε να προβλεφθεί εάν και με ποια πιθανότητα ένας ασθενής ενδέχεται να παρουσιάσει επανεμφάνιση καρκίνου. Οι ταξινομητές που δοκιμάστηκαν είναι οι εξής: δίκτυα Bayes, Τεχνητά Νευρωνικά Δίκτυα, Μηχανές Διανυσμάτων Υποστήριξης, δέντρα απόφασης και τυχαία δάση.

Στον Πίνακας 32, παρατίθενται τα αποτελέσματα που προέκυψαν, χωρίς την εφαρμογή αλγορίθμου για την επιλογή χαρακτηριστικών.

**Πίνακας 32:** Αποτελέσματα που προέκυψαν χωρίς επιλογή χαρακτηριστικών.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 91.4 | 89.7 | 93.1 |
| **ANN** | 70.7 | 69 | 72.4 |
| **SVM** | 65.5 | 72.4 | 58.6 |
| **DT** | 77.6 | 75.9 | 79.3 |
| **RF** | 79.3 | 86.2 | 72.4 |

Στην συνέχεια, χρησιμοποιούμε τον αλγόριθμο CFS για να πραγματοποιήσουμε επιλογή χαρακτηριστικών. Συγκεκριμένα, τα χαρακτηριστικά που διατηρήθηκαν από τον αλγόριθμο CFS είναι: Water Content, Necrosis, bone infiltration, carotid infiltration και side of tumor, και τα αποτελέσματα που έδωσαν με κάθε ταξινομητή φαίνονται στον Πίνακας 33.

**Πίνακας 33:** Αποτελέσματα που προέκυψαν μετά την εφαρμογή του αλγορίθμου CFS.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 82.8 | 79.3 | 86.2 |
| **ANN** | 79.3 | 82.8 | 75.9 |
| **SVM** | 67.2 | 86.2 | 48.3 |
| **DT** | 74.1 | 72.4 | 75.9 |
| **RF** | 74.1 | 82.8 | 65.5 |

Τέλος, χρησιμοποιούμε τον αλγόριθμο wrapper για να πραγματοποιήσουμε επιλογή χαρακτηριστικών, οπότε και για κάθε ταξινομητή προκύπτει μια λίστα χαρακτηριστικών όπως φαίνεται πιο κάτω. Για το BN: contrast take-up rate, minor axis bigger than 10mm, water content, necrosis, central necrosis, site και side; για το ANN: necrosis και carotid infiltration; για τα SVM: extra nodal spreading, shape deviation, necrosis και carotid infiltration; για το DT: shape deviation, water content και necrosis; και για RF: extra nodal spreading, water content, necrosis και bone infiltration. Τα επιμέρους αποτελέσματα παρατίθενται στον Πίνακας 34.

**Πίνακας 34:** Αποτελέσματα που προέκυψαν μετά την εφαρμογή του αλγορίθμου wrapper.

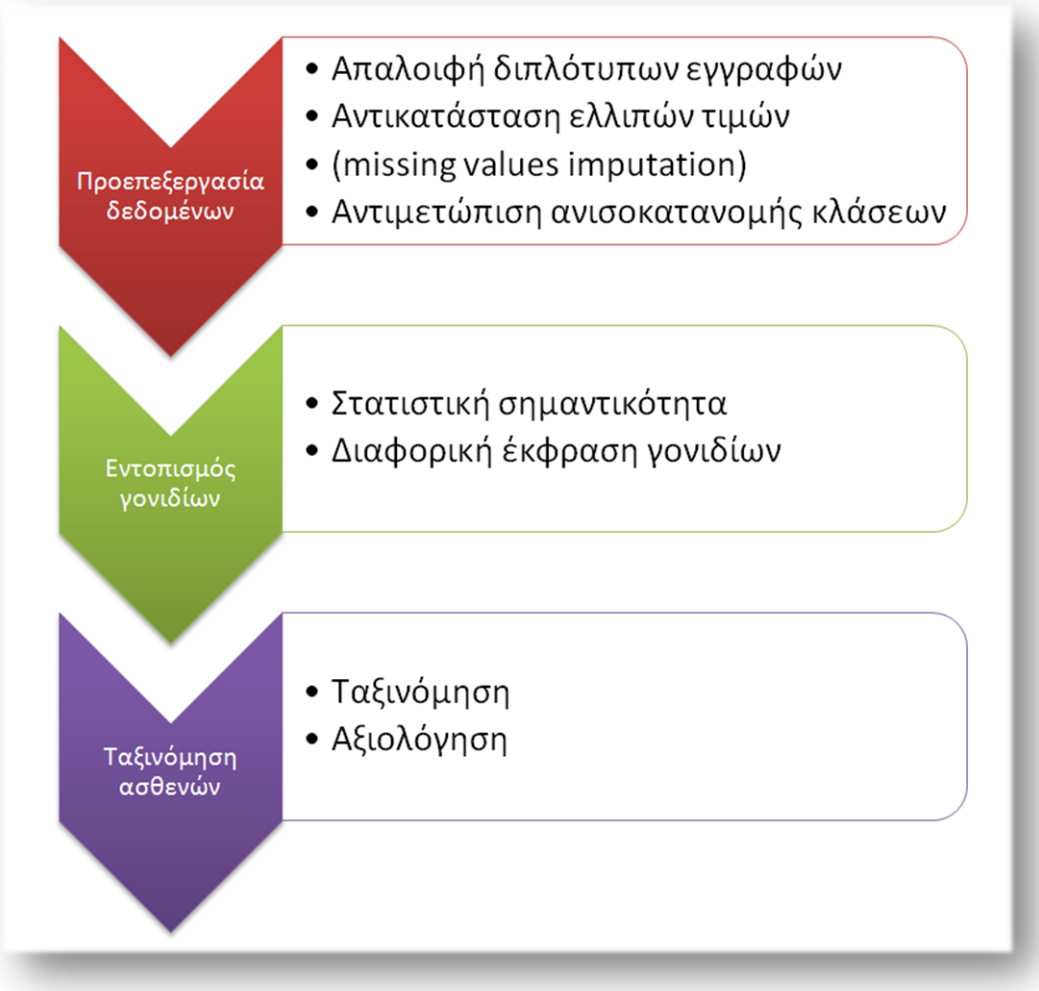
| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 89.7 | 89.7 | 89.7 |
| **ANN** | 81 | 82.8 | 79.3 |
| **SVM** | 79.3 | 86.2 | 72.4 |
| **DT** | 81 | 75.9 | 86.2 |
| **RF** | 82.8 | 79.3 | 86.2 |

Όπως και στην περίπτωση των κλινικών δεδομένων, και κατά την ανάλυση των απεικονιστικών δεδομένων, συνολικά υψηλότερα αποτελέσματα προέκυψαν έπειτα από την εφαρμογή του αλγορίθμου wrapper για επιλογή χαρακτηριστικών. Αξίζει να

σημειώσουμε βέβαια, ότι την καλύτερη απόδοση παρουσίασε ο αλγόριθμος Bayes, και μάλιστα όταν χρησιμοποιήθηκε όλο το διάνυσμα εισόδου, χωρίς να πραγματοποιήσουμε επιλογή χαρακτηριστικών.

Ανάλυση γενετικών δεδομένων

Κατά την ανάλυση γενετικών δεδομένων, χρησιμοποιούμε την έκφραση των γονιδίων, όπως λήφθηκε από τον καρκινικό ιστό κατά τα πρώτα στάδια της θεραπείας, με σκοπό να εντοπίσουμε εκείνα τα γονίδια που εκφράζονται με διαφορετικό τρόπο ανάμεσα σε ασθενείς με και χωρίς επανεμφάνιση της ασθένειας. Στη συνέχεια, χρησιμοποιούμε την έκφραση αυτών των γονιδίων, ως είσοδο σε έναν ταξινομητή που σκοπό έχει να διαχωρίσει ασθενείς με και χωρίς επανεμφάνιση και μάλιστα να υπολογίσει και την πιθανότητα που έχει κάποιος ασθενής να παρουσιάσει μια ενδεχόμενη επανεμφάνιση. Η μεθοδολογία που ακολουθήσαμε για την ανάλυση των γενετικών δεδομένων παρουσιάζεται στην Εικόνα 43. Αρχικά, εκτελούμε κάποια βασικά βήματα προεπεξεργασίας ώστε να βελτιωθεί η ποιότητα του διανύσματος εισόδου. Έπειτα, εντοπίζουμε τα γονίδια που εκφράζονται κατά τρόπο διαφορετικό ανάμεσα σε ασθενείς με και χωρίς επανεμφάνιση, τα οποία εν συνεχεία εισάγονται σε έναν ταξινομητή με σκοπό τον διαχωρισμό των ασθενών βάσει της πιθανότητας ανάπτυξης επανεμφάνισης.




**Εικόνα 43:** Μεθοδολογία ανάλυσης γενετικών δεδομένων.

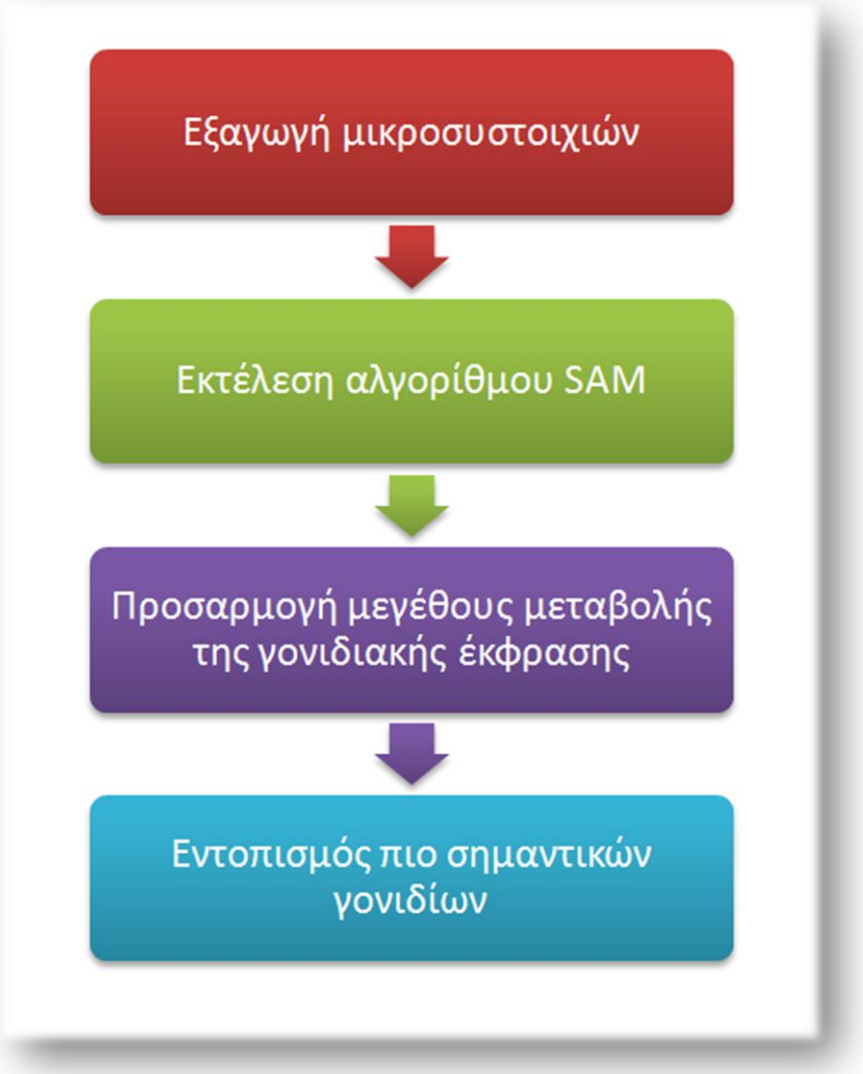
Όπως αναφέρθηκε και προηγουμένως, από κάθε ασθενή εξάγουμε την γονιδιακή έκφραση σε τμήμα του καρκινικού ιστού. Συγκεκριμένα, λαμβάνουμε την έκφραση 45015 γονιδίων σε λογαριθμικές τιμές με βάση το δύο (log2-ratio values) οι οποίες και αρχικά υφίστανται τα παρακάτω στάδια προεπεξεργασίας:

- απαλοιφή γονιδίων ελέγχου
- απαλοιφή διπλότυπων γονιδίων
- απαλοιφή γονιδίων χαμηλής ποιότητας
- αντικατάσταση ή απαλοιφή ελλιπών τιμών

Από την παραπάνω προεπεξεργασία προκύπτουν 33491 γονίδια που χρησιμοποιούνται στα επόμενα στάδια της ανάλυσης. Οι ασθενείς για τους οποίους υπάρχουν διαθέσιμα

δεδομένα φαίνονται στον Πίνακας 22. Κατά την οικοδόμηση του συνόλου δεδομένων, οι κατηγορίες που λαμβάνονται υπόψη για τον σκοπό της μελέτης μας, είναι είτε ασθενείς με επανεμφάνιση, είτε ασθενείς που παραμένουν υγιείς (χωρίς εμφανή τουλάχιστον σημάδια επανεμφάνισης της νόσου), για ένα επαρκές χρονικό διάστημα, όπου στην παρούσα ανάλυση είναι 12 μήνες.

Αναζητούμε λοιπόν, ανάμεσα στο ευρύ σύνολο των γονιδίων που εξάγουμε, το υποσύνολο γονιδίων των οποίων η έκφραση μεταξύ ασθενών με και χωρίς επανεμφάνιση διαφέρει σημαντικά. Τα γονίδια αυτά συνθέτουν τους παράγοντες που από βιολογική σκοπιά παίζουν καθοριστικό ρόλο στην εξέλιξη της επανεμφάνισης και διαθέτουν την μεγαλύτερη διακριτική ικανότητα ως προς τον διαχωρισμό των δύο κλάσεων ασθενών που μελετάμε. Για τον σκοπό αυτό, αρχικά εφαρμόζουμε τον αλγόριθμο SAM (Significance Analysis of Microarrays) [130] όπου μετρά την διαφορική έκφραση του κάθε γονιδίου μεταξύ των ασθενών των δύο κλάσεων. Η χρήση των κλασικών στατιστικών τεστ (π.χ. t-test) μπορεί εύκολα να οδηγήσει σε εσφαλμένα και παραπλανητικά συμπεράσματα. Θεωρώντας ένα σύνολο 30000 χαρακτηριστικών/γονιδίων και θέτοντας ως κατώφλι για την p-value το 0.01, που είναι μια τυπική τιμή για την εξασφάλιση στατιστικής σημαντικότητας, θα ανιχνεύαμε περίπου 300 γονίδια στην τύχη. Για αυτό και σε σύνολα δεδομένων με πολλά χαρακτηριστικά και λίγα δείγματα, μεμονωμένα στατιστικά τεστ δεν αποτελούν σωστή επιλογή. Από την άλλη, το SAM κατασκευάζει πολλά υποσύνολα δεδομένων, πραγματοποιώντας αντιμεταθέσεις, και εκτελεί σε καθένα από αυτά t-test. Από κάθε τέτοιο σύνολο και για κάθε γονίδιο, εξάγεται ένα σκορ που αφορά στην σημαντικότητα του γονιδίου, που ουσιαστικά εκφράζει την διαφορική του έκφραση μεταξύ των δύο κλάσεων των ασθενών, λαμβάνοντας υπόψη και την τυπική απόκλιση στην έκφραση του γονιδίου. Επίσης, πραγματοποιήσαμε πλήθος στατιστικών τεστ σε σχέση με το μέγεθος μεταβολής (R) στην έκφραση ενός γονιδίου στους ασθενείς των δύο κλάσεων, ώστε να θεωρηθεί σημαντική. Η συγκεκριμένη σειρά βημάτων που ακολουθούμε για τον εντοπισμό των γονιδίων με την μεγαλύτερη διακριτική ικανότητα και κατά συνέπεια των πιο σημαντικών φαίνεται στην Εικόνα 44.




**Εικόνα 44:** Στάδια εντοπισμού των πιο σημαντικών γονιδίων.

Ο Πίνακας 35, περιέχει τα αποτελέσματα για τα διάφορα πειράματα που εκτελέστηκαν, δηλαδή το πλήθος των γονιδίων που αναγνωρίστηκαν ως πιο σημαντικά, για διάφορες τιμές που αφορούν το μέγεθος μεταβολής της γονιδιακής έκφρασης και τον ρυθμό εσφαλμένων εντοπισμών (FDR: False Discovery Rate).

**Πίνακας 35:** Πλήθος γονιδίων που καταδείχτηκαν ως πιο σημαντικά για τις διάφορες τιμές του μεγέθους μεταβολής της γονιδιακής έκφρασης.

| Μέγεθος μεταβολής (R) | # σημαντικών γονιδίων | FDR (%) | # εσφαλμένα εντοπισμένων γονιδίων |
|---|---|---|---|
| **1.0** | 2 | 0.55 | 1.1 |
| **1.2** | 1 | 0 | 0 |
| **1.5** | 40 | 13 | 5.5 |
| **1.8** | 6 | 0 | 0 |
| **2.0** | 0 | 0 | 0 |
| **2.5** | 0 | 0 | 0 |

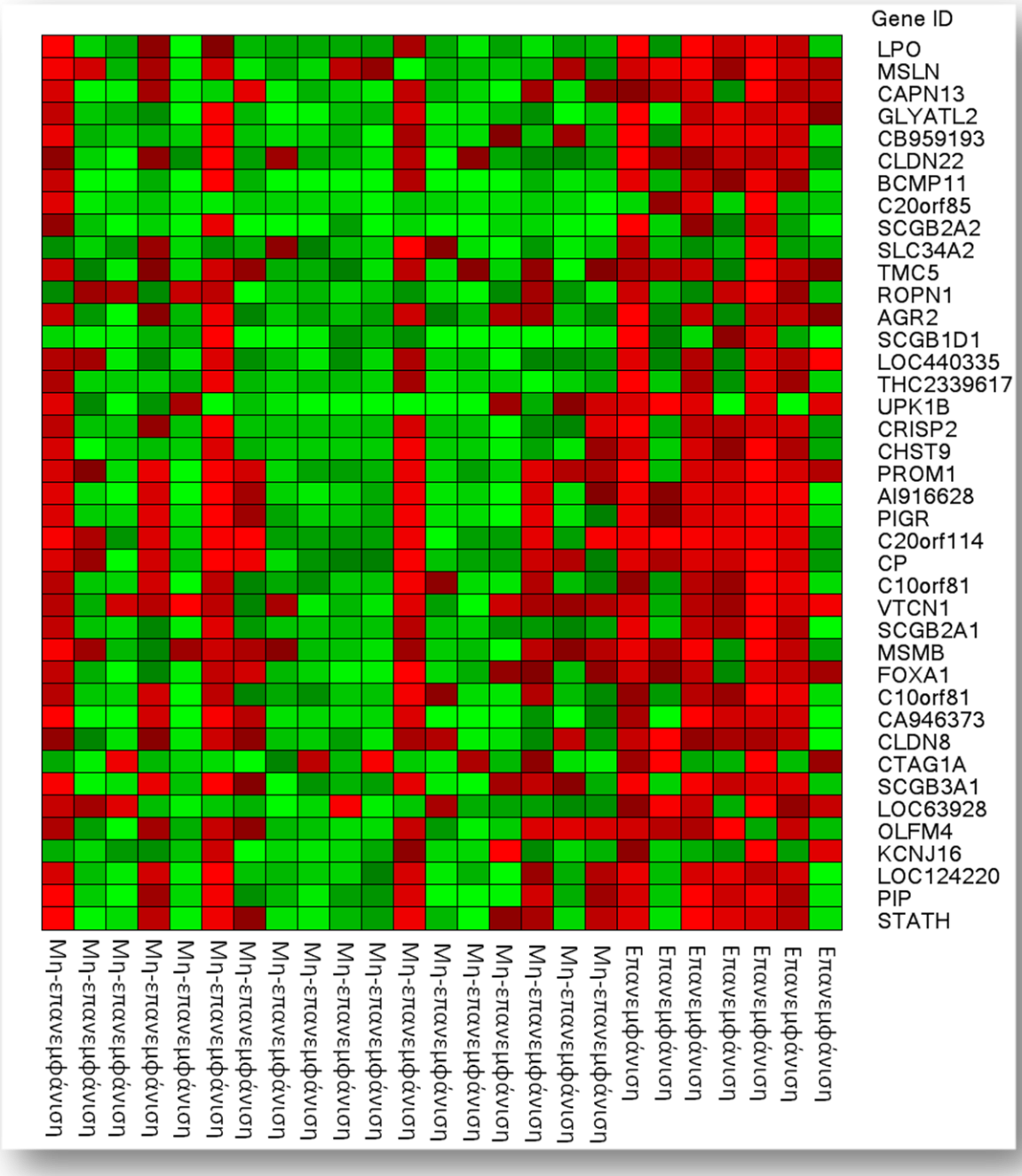
Ακολούθως, ο Πίνακας 36 περιλαμβάνει τα συγκεκριμένα γονίδια που καταδείχτηκαν ως πιο σημαντικά, λαμβάνοντας υπόψη μέγεθος μεταβολής τουλάχιστον 1.5 μεταξύ των δύο κλάσεων των ασθενών.

**Πίνακας 36:** Γονίδια που καταδείχτηκαν ως πιο σημαντικά.

| Όνομα γονιδίου | Μέγεθος μεταβολής |
|---|---|
| **LPO** | 1.6 |
| **MSLN** | 1.5 |
| **CAPN13** | 1.6 |
| **GLYATL2** | 1.9 |
| **CB959193** | 1.7 |
| **CLDN22** | 1.7 |
| **BCMP11** | 1.9 |
| **C20orf85** | 1.5 |
| **SCGB2A2** | 1.8 |
| **SLC34A2** | 1.8 |
| **TMC5** | 1.6 |
| **ROPN1** | 1.6 |
| **AGR2** | 1.8 |
| **SCGB1D1** | 1.5 |

| | |
|---|---|
| **LOC440335** | 1.6 |
| **THC2339617** | 1.7 |
| **UPK1B** | 1.6 |
| **CRISP2** | 1.8 |
| **CHST9** | 1.7 |
| **PROM1** | 1.6 |
| **AI916628** | 1.9 |
| **PIGR** | 1.6 |
| **C20orf114** | 1.5 |
| **CP** | 1.5 |
| **C10orf81** | 1.6 |
| **VTCN1** | 1.6 |
| **SCGB2A1** | 1.7 |
| **MSMB** | 1.7 |
| **FOXA1** | 1.6 |
| **C10orf81** | 1.5 |
| **CA946373** | 1.5 |
| **CLDN8** | 1.6 |
| **CTAG1A** | 1.6 |
| **SCGB3A1** | 1.6 |
| **LOC63928** | 1.5 |
| **OLFM4** | 1.6 |
| **KCNJ16** | 1.5 |
| **LOC124220** | 1.6 |
| **PIP** | 1.9 |
| **STATH** | 1.7 |

Η Εικόνα 45 δείχνει εποπτικά τον βαθμό έκφρασης των 40 γονιδίων που εντοπίστηκαν ως πιο σημαντικά στο πλήθος των ασθενών που λήφθηκαν υπόψη και από τις δύο κλάσεις.




**Εικόνα 45:** Γραφική απεικόνιση της γονιδιακής έκφρασης μεταξύ ασθενών με και χωρίς επανεμφάνιση της ασθένειας.

Για την περαιτέρω επιβεβαίωση των γονιδίων που παίζουν καθοριστικό ρόλο στην επαγωγή επανεμφάνισης της ασθένειας, στο σύνολο των 33491 γονιδίων που προκύπτει μετά την αρχική προεπεξεργασία, εφαρμόζουμε μια σειρά αλγορίθμων που αποσκοπούν στον εντοπισμό γονιδίων που εκφράζονται κατά τρόπο διαφορετικό ανάμεσα σε δύο

σύνολα δειγμάτων. Συγκεκριμένα, εφαρμόσαμε τους αλγορίθμους eBayes [131, 132], PLS-CV [124], RF-MDA [128] καθώς και έναν αλγόριθμο που συνδυάζει τα επιμέρους σύνολα που προκύπτουν από το κάθε αλγόριθμο ξεχωριστά [133]. Θέτοντας τα 40 γονίδια ως όριο κατά την έξοδο του κάθε αλγορίθμου, τα γονίδια που καταδείχτηκαν από τον κάθε αλγόριθμο ως πιο σημαντικά παρουσιάζονται στον Πίνακας 37.

**Πίνακας 37:** Γονίδια που καταδείχτηκαν ως πιο σημαντικά έπειτα από την εφαρμογή μιας σειράς αλγορίθμων.

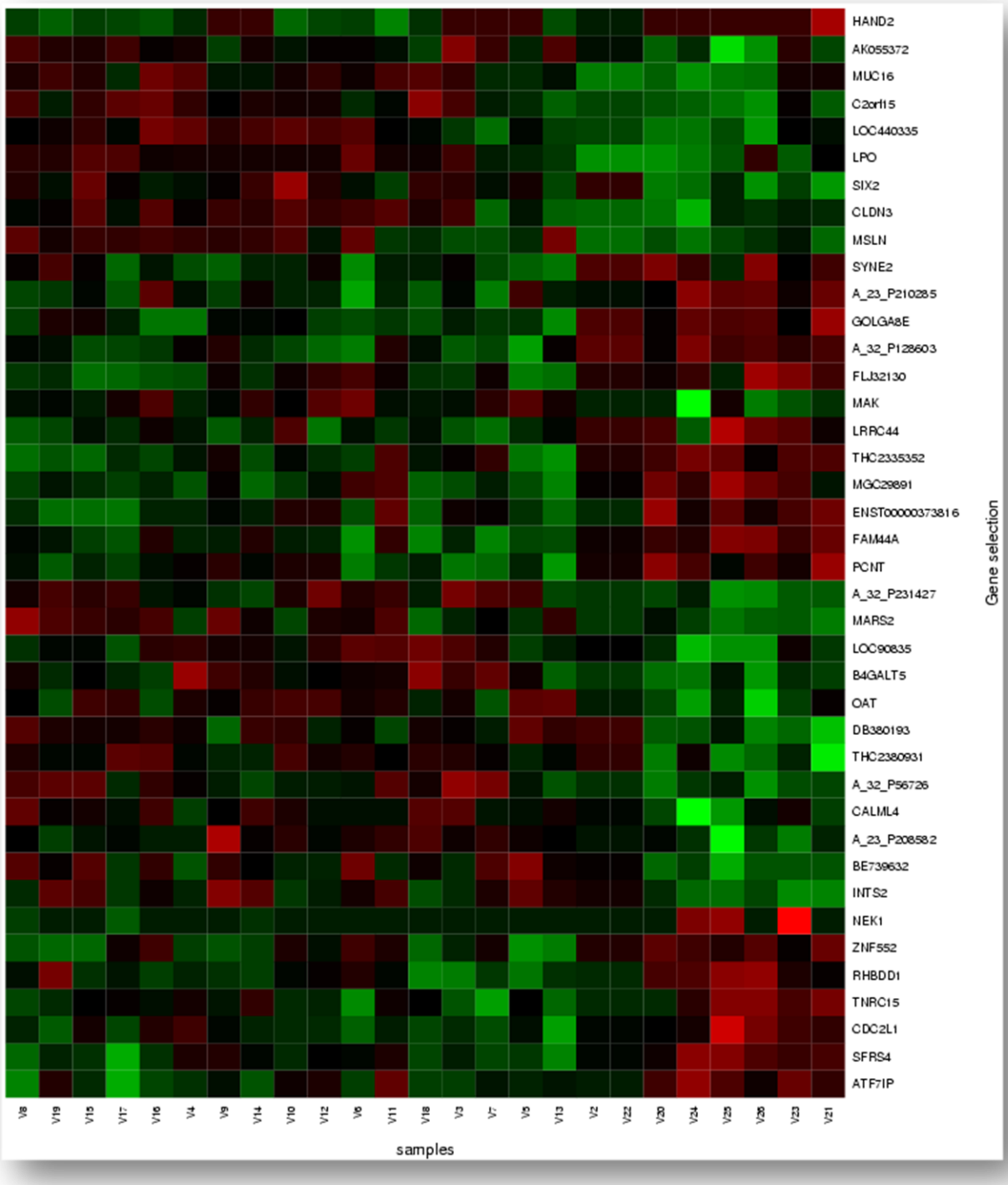
| eBayes | PLS-CV | RF-MDA | Συνδυασμός μεθόδων |
|---|---|---|---|
| FAM44A | GLYATL2 | MYO18B | GOLGA8E |
| GOLGA8E | SCGB2A2 | AA991174 | SFRS4 |
| A_32_P231427 | BCMP11 | BQ365891 | MGC29891 |
| A_32_P128603 | THC2339617 | SLC11A2 | A_32_P128603 |
| THC2335352 | AI916628 | RAB1A | THC2335352 |
| SFRS4 | CTAG1A | NOTCH2NL | AI475779 |
| MGC29891 | CB959193 | THC2441546 | FLJ32130 |
| PCNT | LOC63928 | AW138098 | A_32_P231427 |
| C2orf15 | CRISP2 | TLE6 | LPO |
| SIX2 | CHST9 | HOXD9 | C20orf85 |
| CDC2L1 | PIP | ENST00000373816 | THC2339617 |
| BE739632 | NEFL | A_24_P58899 | NR2E1 |
| MAK | CLDN22 | CASP8AP2 | CLDN3 |
| THC2380931 | FOXA1 | CASKIN2 | MGAM |
| ATF7IP | TMPRSS11D | AA765447 | FOXJ1 |
| MUC16 | NEF3 | PTPLAD2 | AKAP14 |
| HAND2 | SCGB1D1 | THC2270160 | HSPB3 |
| DB380193 | LOC440335 | MGC2752 | THC2410387 |
| A_23_P210285 | SCGB2A1 | ZDHHC5 | LOC613212 |
| FLJ32130 | MSMB | ENST00000292543 | GSPT2 |
| ENST00000373816 | TMC5 | CLDN22 | ANKRD12 |
| ZNF552 | AGR2 | THC2345565 | ENST00000373816 |
| AK055372 | UPK1B | BG216680 | A_23_P208582 |
| INTS2 | SPINK5 | CGB2 | A_32_P93584 |

| | | | |
|---|---|---|---|
| MARS2 | SLC34A2 | CDC2L2 | ANKRD42 |
| MSLN | CAPN13 | LOC388335 | BE739632 |
| LOC90835 | FLJ32130 | A_24_P512530 | THC2430017 |
| LOC440335 | STATH | NUCB2 | FOXG1B |
| CLDN3 | CRNN | PSG11 | SYT8 |
| B4GALT5 | HORMAD1 | VIM | ENST00000358162 |
| LRRC44 | FLJ46385 | U69195 | RAB12 |
| TNRC15 | ROPN1 | C17orf48 | THC2305868 |
| SYNE2 | PIGR | A_24_P50390 | CALML4 |
| NEK1 | CLDN8 | THC2307332 | B4GALT5 |
| A_32_P56726 | MSLN | KIAA0644 | THC2440787 |
| LPO | PAGE5 | ATXN10 | C19orf15 |
| CALML4 | SLC44A4 | KIF1B | CA420688 |
| RHBDD1 | AK124097 | F13A1 | AW972815 |
| A_23_P208582 | C20orf85 | CUGBP2 | YSK4 |
| OAT | HAND2 | PCP4 | SEMA3E |

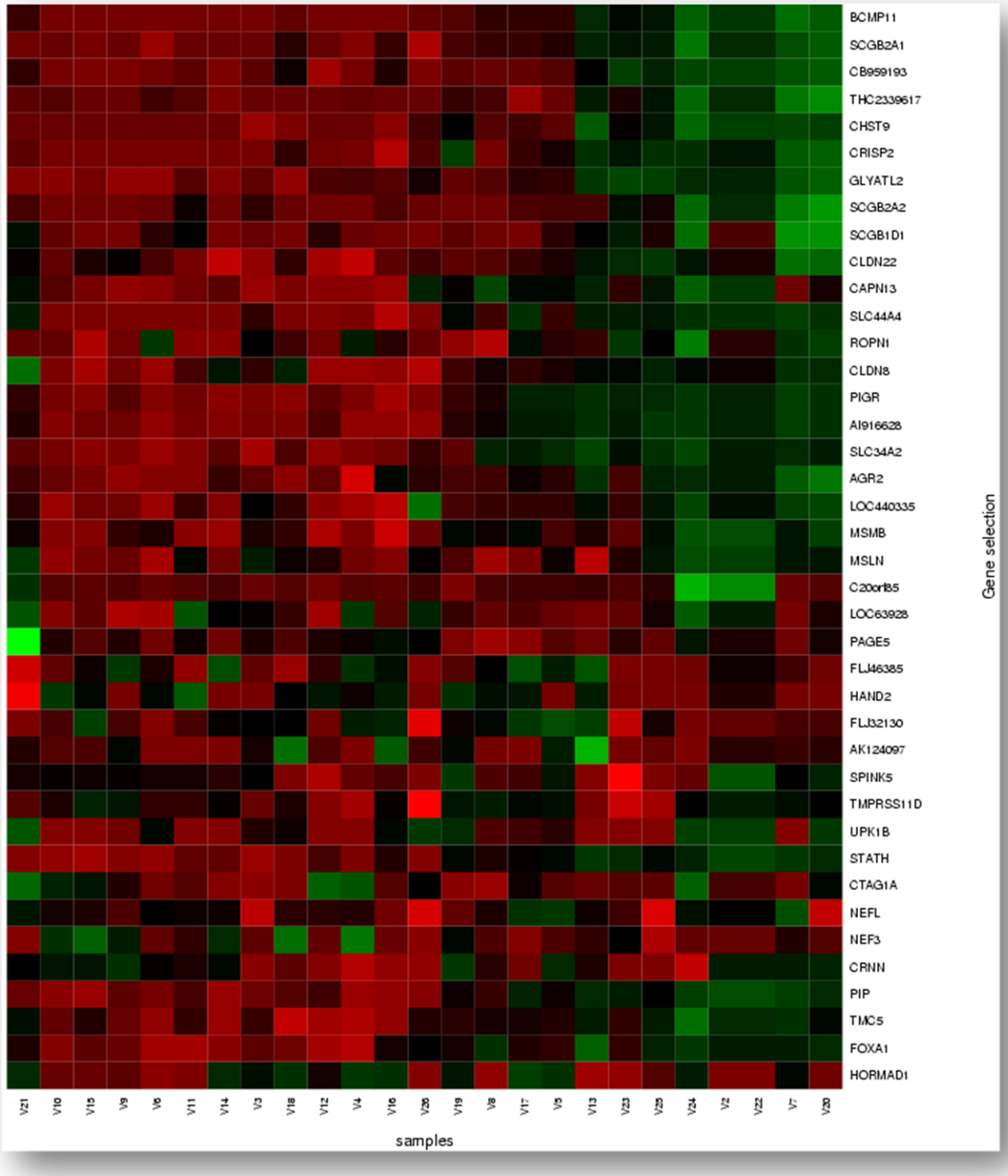
Αξίζει να σημειωθεί, ότι με τη μεθοδολογία που χρησιμοποιήσαμε παραπάνω για τον εντοπισμό των πιο σημαντικών γονιδίων προέκυψαν πολλά γονίδια κοινά με αυτά που παρατίθενται στον Πίνακας 37. Συγκεκριμένα, 29 από τα 40 γονίδια που εντοπίσαμε, είναι κοινά με τουλάχιστον μια από τις μεθοδολογίες των οποίων τα αποτελέσματα παρουσιάζονται στον Πίνακας 37, ενισχύοντας έτσι την αξιοπιστία και την εγκυρότητα των εξαχθέντων αποτελεσμάτων.

Στις Εικόνα 46 που ακολουθεί παρουσιάζονται εποπτικά οι εκφράσεις των γονιδίων που καταδείχτηκαν από κάθε αλγόριθμο ως πιο σημαντικά.

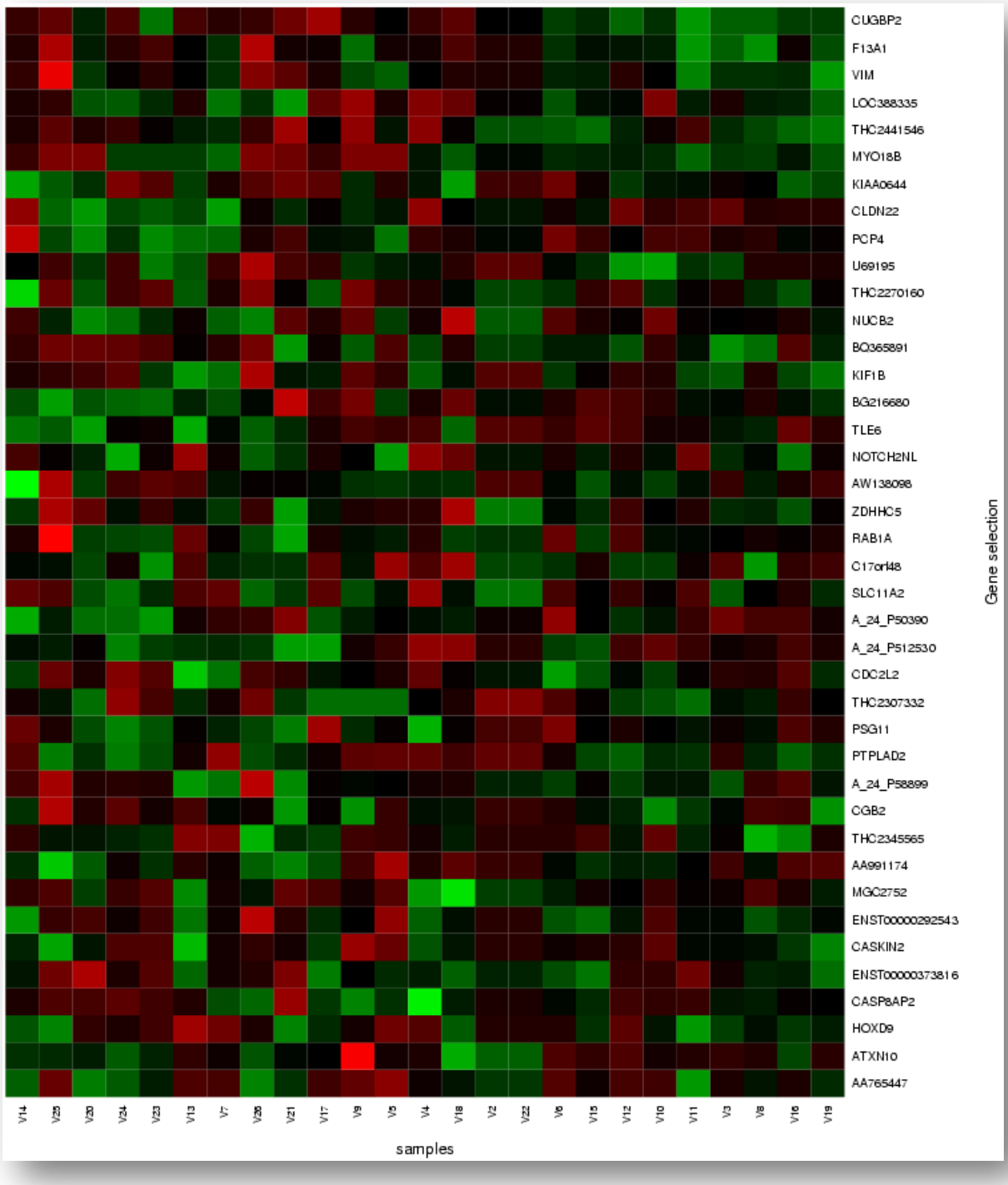
(α)


HAND2
AK055372
MUC16
C2orf15
LOC440335
LPO
SIX2
CLDN3
MSLN
SYNE2
A_23_P210285
GOLGA8E
A_32_P128603
FLJ32130
MAK
LRRC44
THC2335352
MGC29891
ENST00000373816
FAM44A
PCNT
A_32_P231427
MARS2
LOC90835
B4GALT5
OAT
DB380193
THC2380931
A_32_P56726
CALML4
A_23_P208582
BE739632
INTS2
NEK1
ZNF552
RHBDD1
TNRC15
CDC2L1
SFRS4
ATF7IP
Gene selection
V8 V19 V15 V17 V16 V4 V9 V14 V10 V12 V6 V11 V18 V3 V7 V5 V13 V2 V22 V20 V24 V25 V26 V23 V21
samples

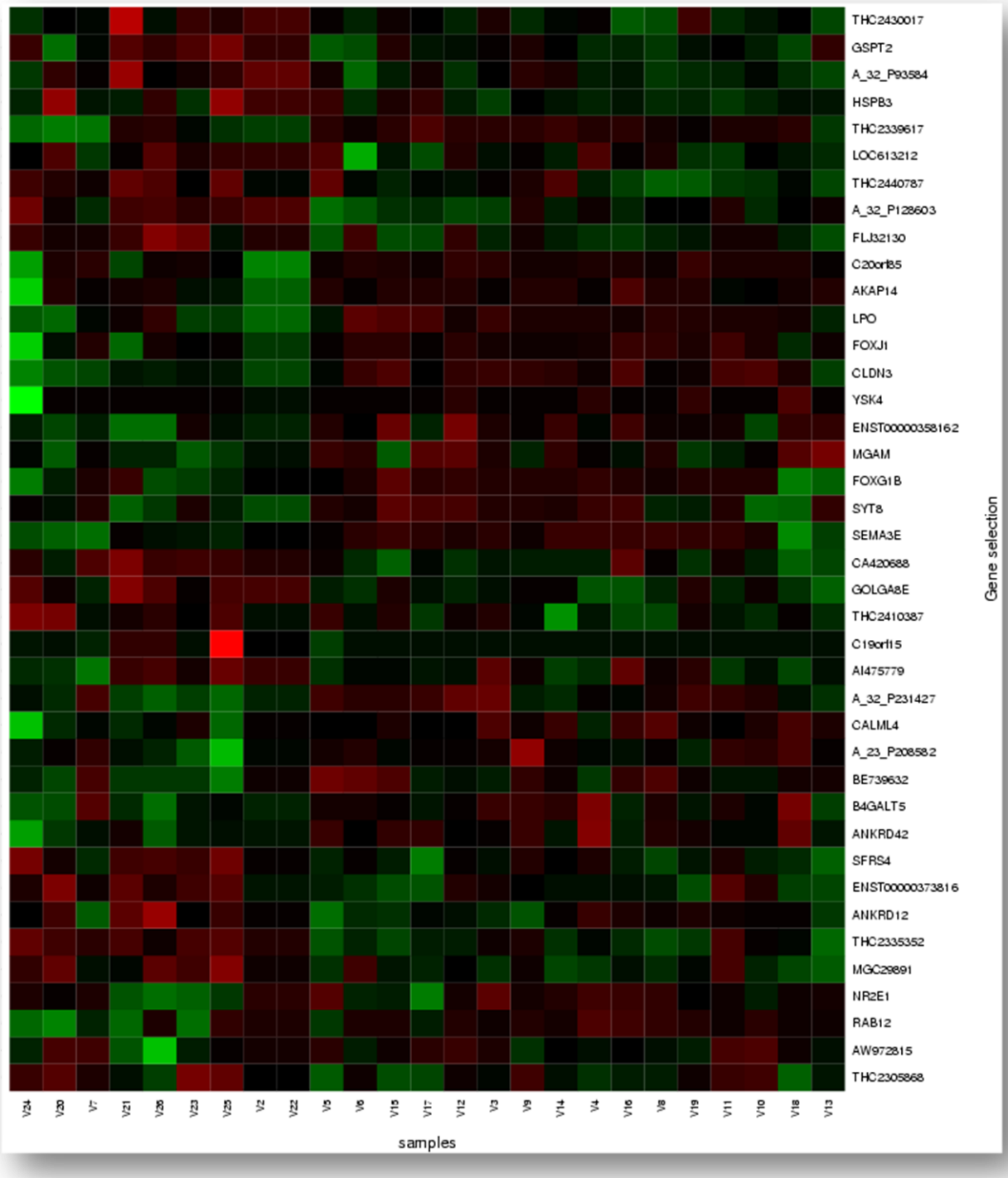
(β)


BCMP11
SCGB2A1
CB959193
THC2339617
CHST9
CRISP2
GLYATL2
SCGB2A2
SCGB1D1
CLDN22
CAPN13
SLC44A4
ROPN1
CLDN8
PIGR
AI916628
SLC34A2
AGR2
LOC440335
MSMB
MSLN
C20orf85
LOC63928
PAGE5
FLJ46385
HAND2
FLJ32130
AK124097
SPINK5
TMPRSS11D
UPK1B
STATH
CTAG1A
NEFL
NEF3
CRNN
PIP
TMC5
FOXA1
HOXMAD1
Gene selection
V21 V10 V15 V9 V6 V11 V14 V3 V18 V12 V4 V16 V26 V19 V8 V17 V5 V13 V23 V25 V24 V2 V22 V7 V20
samples

(γ)


CUGBP2
F13A1
VIM
LOC388335
THC2441546
MYO18B
KIAA0644
CLDN22
PCP4
U69195
THC2270160
NUCB2
BQ365891
KIF1B
BG216680
TLE6
NOTCH2NL
AW138098
ZDHHC5
RAB1A
C17orf48
SLC11A2
A_24_P50390
A_24_P512530
CDC2L2
THC2307332
PSG11
PTPLAD2
A_24_P58899
CGB2
THC2345565
AA991174
MGC2752
ENST00000292543
CASKIN2
ENST00000373816
CASP8AP2
HOXD9
ATXN10
AA765447
Gene selection
V14
V25
V20
V24
V23
V13
V7
V26
V21
V17
V9
V5
V4
V18
V2
V22
V6
V15
V12
V10
V11
V3
V8
V16
V19
samples

(δ)




**Εικόνα 46:** Παρουσίαση της γονιδιακής έκφρασης έπειτα από την εφαρμογή κάθε αλγορίθμου: (α) eBayes, (β) PLS-CV, (γ) RF-MDA και (δ) συνδυασμός όλων των παραπάνω.

Τα 40 γονίδια που παρατίθενται στον Πίνακας 36, θα αποτελέσουν την είσοδο για τα επόμενα βήματα της μεθοδολογίας μας, που αφορά στην χρήση και αξιοποίηση αμιγώς γενετικών δεδομένων, και δη της γονιδιακής έκφρασης, με σκοπό την διάκριση μεταξύ ασθενών με και χωρίς επανεμφάνιση της νόσου, καθώς και τον υπολογισμό της πιθανότητας που ανατίθεται σε κάθε πρόβλεψη.

Όπως και κατά την ανάλυση των κλινικών και απεικονιστικών δεδομένων, όμοια και στα γενετικά δεδομένα, αρχικά εφαρμόζουμε τον αλγόριθμο SMOTE [122] ώστε να αντιμετωπίσουμε την ανισοκατανομή των κλάσεων στο σύνολο δεδομένων μας. Ακολούθως, είτε χρησιμοποιούμε όλα τα γονίδια που προέκυψαν προηγουμένως, είτε επιτελούμε περαιτέρω επιλογή χαρακτηριστικών και σε κάθε περίπτωση εισάγουμε τα επιμέρους χαρακτηριστικά/γονίδια σε μια σειρά ταξινομητών ώστε να διακρίνουμε ασθενείς με και χωρίς επανεμφάνιση της ασθένειας. Οι ταξινομητές που υλοποιούμε είναι οι: δίκτυο Bayes, Τεχνητά Νευρωνικά Δίκτυα, Μηχανές Διανυσμάτων Υποστήριξης, δέντρα απόφασης, τυχαία δάση.

Ο Πίνακας 38 περιέχει τα αποτελέσματα που προέκυψαν από κάθε ταξινομητή όταν χρησιμοποιήσαμε τις εκφράσεις των 40 γονιδίων ως είσοδο στους ταξινομητές, χωρίς περαιτέρω επιλογή χαρακτηριστικών.

**Πίνακας 38:** Αποτελέσματα που προέκυψαν χωρίς επιλογή χαρακτηριστικών.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 83.3 | 88.9 | 77.8 |
| **ANN** | 83.3 | 88.9 | 77.8 |
| **SVM** | 88.9 | 94.4 | 83.3 |
| **DT** | 77.8 | 66.7 | 88.9 |
| **RF** | 77.8 | 77.8 | 77.8 |

Στη συνέχεια πραγματοποιούμε επιλογή χαρακτηριστικών με τον αλγόριθμο CFS, οπότε και διατηρούνται τα εξής γονίδια: MSLN, CAPN13, GLYATL2, CLDN22, CTAG1A και LOC63928. Χρησιμοποιώντας τις τιμές αυτών των γονιδίων ως είσοδο στους ταξινομητές λαμβάνουμε τα αποτελέσματα που παρατίθενται στον Πίνακας 39.

**Πίνακας 39:** Αποτελέσματα που προέκυψαν έπειτα από επιλογή χαρακτηριστικών με τη χρήση του αλγορίθμου CFS.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 86.1 | 83.3 | 88.9 |
| **ANN** | 86.1 | 94.4 | 77.8 |
| **SVM** | 88.9 | 100 | 77.8 |
| **DT** | 75 | 66.7 | 83.3 |
| **RF** | 77.8 | 72.2 | 83.3 |

Τέλος, πραγματοποιήσαμε επιλογή χαρακτηριστικών και με τον αλγόριθμο wrapper, οπότε και για κάθε ταξινομητή διατηρήθηκαν τα πιο κάτω γονίδια ως πιο σημαντικά και με την μεγαλύτερη διακριτική ικανότητα. Για το BN: MSLN και CAPN13; για το ANN: MSLN, CAPN13, C20ORF85, MSMB και OLFM4; για το SVM: LPO, MSLN, CAPN13 και CTAG1A; για το DT: CAPN13 και SCGB2A2; και για RF: MSLN, CAPN13, SLC34A2 και VTCN1. Τα αποτελέσματα για κάθε ταξινομητή παρουσιάζονται στον Πίνακας 40.

**Πίνακας 40:** Αποτελέσματα που προέκυψαν έπειτα από επιλογή χαρακτηριστικών με τη χρήση του αλγορίθμου wrapper.

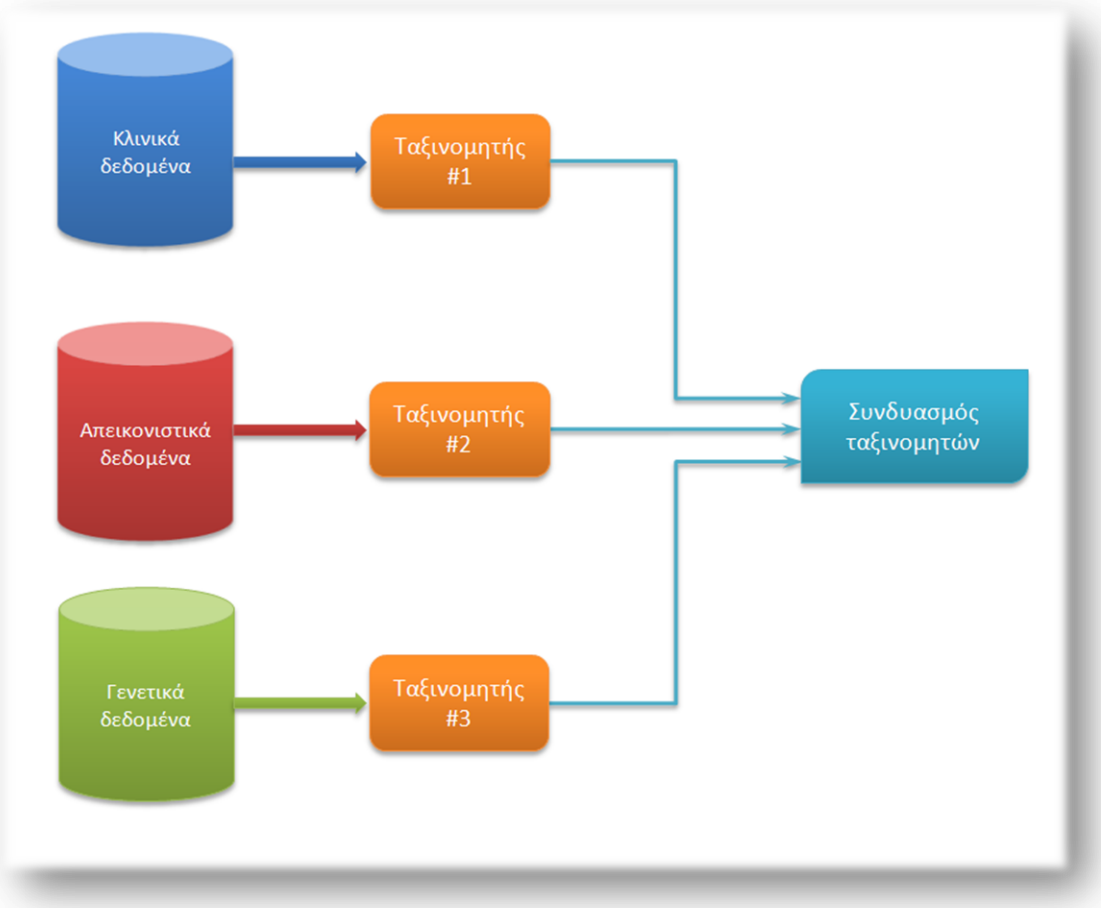
| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 94.4 | 100 | 88.9 |
| **ANN** | 94.4 | 100 | 88.9 |
| **SVM** | 91.7 | 94.4 | 88.9 |
| **DT** | 91.7 | 94.4 | 88.9 |
| **RF** | 91.7 | 94.4 | 88.9 |

Παρατηρούμε ότι για κάθε ταξινομητή, τα αποτελέσματα που προκύπτουν είναι σαφώς υψηλότερα έπειτα από την εφαρμογή του αλγορίθμου wrapper για επιλογή

χαρακτηριστικών. Όσον αφορά τον συγκεκριμένο ταξινομητή, βλέπουμε ότι το BN δίδει πολύ καλά αποτελέσματα όπως και στις περιπτώσεις των κλινικών και απεικονιστικών δεδομένων, και μάλιστα είναι σημαντικό ότι έχει διάφανη αρχιτεκτονική παρέχοντας κατ' αυτόν τον τρόπο επαρκή επεξήγηση για κάθε πρόβλεψη.

Συνδυαστική ανάλυση δεδομένων

Επίσης, πραγματοποιούμε και μια συνδυαστική ανάλυση όπου χρησιμοποιούμε τα ετερογενή δεδομένα κατά τρόπο συμπληρωματικό, ώστε να επωφεληθούμε κατά το δυνατόν περισσότερο από κάθε τύπο δεδομένων και να καλύψουμε πιθανά κενά και ελλείψεις που υπάρχουν στους άλλους τύπους δεδομένων. Η πρώτη αρχιτεκτονική που υλοποιείται προς αυτήν την κατεύθυνση, ουσιαστικά συνδυάζει τις αποφάσεις των ταξινομητών που υλοποιήθηκαν παραπάνω, υπολογίζοντας μια συνολική πιθανότητα για κάθε κλάση βάσει των επιμέρους πιθανοτήτων όπως υπολογίζονται από κάθε ταξινομητή. Η συγκεκριμένη αρχιτεκτονική απεικονίζεται στην Εικόνα 47.




**Εικόνα 47:** Πρώτη αρχιτεκτονική συνδυαστικής ανάλυσης ετερογενών δεδομένων.

Τα αποτελέσματα που προκύπτουν είτε χωρίς επιλογή χαρακτηριστικών, είτε έπειτα από την εφαρμογή των αλγορίθμων CFS και wrapper παρουσιάζονται στους πίνακες που ακολουθούν. Συγκεκριμένα ο Πίνακας 41, περιέχει τα αποτελέσματα χωρίς επιλογή χαρακτηριστικών.

**Πίνακας 41:** Αποτελέσματα που προέκυψαν χωρίς επιλογή χαρακτηριστικών.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 90 | 90 | 90 |
| **ANN** | 80 | 90 | 70 |
| **SVM** | 85 | 90 | 80 |
| **DT** | 75 | 80 | 70 |
| **RF** | 80 | 80 | 80 |

Ακολούθως, αφού πραγματοποιήσουμε επιλογή χαρακτηριστικών σε κάθε τύπο δεδομένων ξεχωριστά με τον αλγόριθμο CFS, και εκπαιδεύσουμε έναν ταξινομητή για κάθε τύπο δεδομένων, συνδυάζουμε τις αποφάσεις των επιμέρους ταξινομητών, βάσει πλειοψηφίας. Τα αποτελέσματα που προκύπτουν παρουσιάζονται στον Πίνακας 42.

**Πίνακας 42:** Αποτελέσματα που προέκυψαν έπειτα από επιλογή χαρακτηριστικών με τη χρήση του αλγορίθμου CFS.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 100 | 100 | 100 |
| **ANN** | 80 | 80 | 80 |
| **SVM** | 85 | 80 | 90 |
| **DT** | 75 | 80 | 70 |
| **RF** | 90 | 90 | 90 |

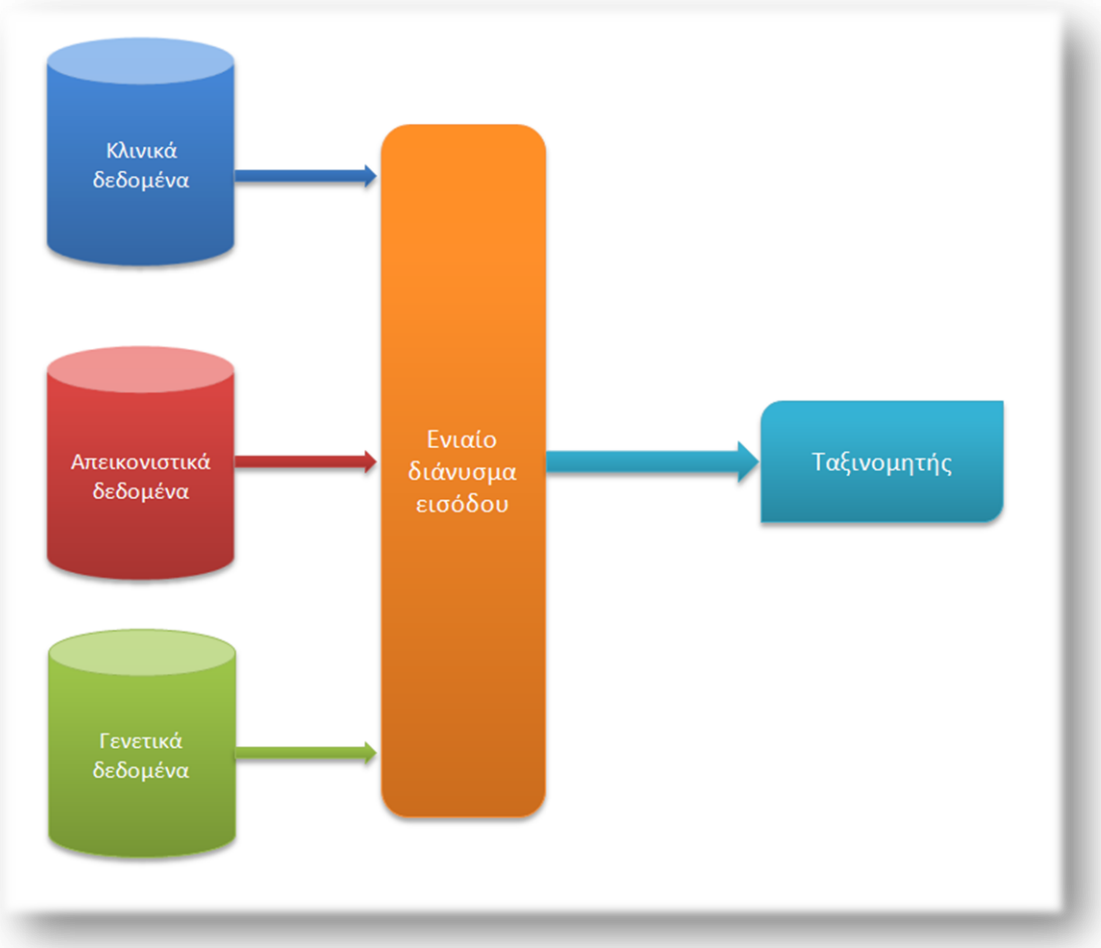
Στον Πίνακας 43, παρουσιάζονται τα αποτελέσματα που προέκυψαν για κάθε ταξινομητή έπειτα από την εφαρμογή του αλγορίθμου wrapper για επιλογή χαρακτηριστικών.

**Πίνακας 43:** Αποτελέσματα που προέκυψαν έπειτα από επιλογή χαρακτηριστικών με τη χρήση του αλγορίθμου wrapper.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 95 | 100 | 90 |
| **ANN** | 100 | 100 | 100 |
| **SVM** | 95 | 90 | 100 |
| **DT** | 95 | 90 | 100 |
| **RF** | 95 | 90 | 100 |

Παρατηρούμε ότι ο συνδυασμός των επιμέρους αποφάσεων των ταξινομητών, σε κάθε περίπτωση οδηγεί σε βελτίωση των αποτελεσμάτων, καθώς επίσης θετική επίδραση έχει και η επιλογή χαρακτηριστικών, και με τους δύο αλγορίθμους. Επίσης, βλέπουμε ότι ο αλγόριθμος wrapper οδηγεί σε υψηλότερα αποτελέσματα με κάθε ταξινομητή με μόνη εξαίρεση το BN, όπου παρουσιάζει τα καλύτερα αποτελέσματα έπειτα από επιλογή χαρακτηριστικών με τον αλγόριθμο CFS.

Στη συνέχεια, υλοποιούμε και μια δεύτερη αρχιτεκτονική για την συνδυαστική ανάλυση των δεδομένων, κατά την οποία όλα τα δεδομένα συνδυάζονται για να σχηματίσουν ένα ενιαίο διάνυσμα εισόδου, που περιλαμβάνει το σύνολο των κλινικών, απεικονιστικών και γενετικών δεδομένων. Το διάνυσμα που προκύπτει, είτε χρησιμοποιείται ως έχει για ταξινόμηση, είτε πραγματοποιούμε σε αυτό επιλογή χαρακτηριστικών με τους αλγορίθμους CFS και wrapper, και εισάγουμε το προκύπτον διάνυσμα σε μια σειρά ταξινομητών. Η δεύτερη αρχιτεκτονική, απεικονίζεται στην Εικόνα 48.




**Εικόνα 48:** Δεύτερη αρχιτεκτονική συνδυαστικής ανάλυσης ετερογενών δεδομένων.

Τα αποτελέσματα που προέκυψαν έπειτα από την χρήση του ενιαίου διανύσματος εισόδου σε κάθε ταξινομητή, χωρίς επιλογή χαρακτηριστικών, παρουσιάζονται στον Πίνακας 44.

**Πίνακας 44:** Αποτελέσματα που προέκυψαν χωρίς επιλογή χαρακτηριστικών.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 90 | 90 | 90 |
| **ANN** | 85 | 100 | 70 |
| **SVM** | 75 | 80 | 70 |
| **DT** | 60 | 70 | 50 |
| **RF** | 75 | 80 | 70 |

Έπειτα, πραγματοποιούμε στο ενιαίο διάνυσμα εισόδου επιλογή χαρακτηριστικών με τον αλγόριθμο CFS, οπότε διατηρούνται τα εξής χαρακτηριστικά: tumor thickness, depth invasion, side, LPO, CB959193, SLC34A2, TMC5, SCGB1D1, UPK1B, CRISP2, C20ORF114, FOXA1, CLDN8, CTAG1A και LOC63928. Τα επιμέρους αποτελέσματα για κάθε ταξινομητή παρουσιάζονται στον Πίνακας 45.

**Πίνακας 45:** Αποτελέσματα που προέκυψαν έπειτα από επιλογή χαρακτηριστικών με τη χρήση του αλγορίθμου CFS.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 90 | 90 | 90 |
| **ANN** | 95 | 100 | 90 |
| **SVM** | 95 | 100 | 90 |
| **DT** | 65 | 70 | 60 |
| **RF** | 90 | 90 | 90 |

Τέλος, πραγματοποιούμε επιλογή χαρακτηριστικών με τον αλγόριθμο wrapper, οπότε για κάθε ταξινομητή διατηρούνται τα παρακάτω χαρακτηριστικά. Για το BN: tumor thickness και TMC5; για ANN: TMC5; για SVM: BCMP11 και UPK1B; για DT: num of mitoses HPF; και για RF: ecog status, weight, cholesterol, eating habits, site και UPK1B. Τα αποτελέσματα σε αυτήν την περίπτωση παρατίθενται στον Πίνακας 46.

**Πίνακας 46:** Αποτελέσματα που προέκυψαν έπειτα από επιλογή χαρακτηριστικών με τη χρήση του αλγορίθμου wrapper.

| Αλγόριθμος ταξινόμησης | Ακρίβεια (%) | Ευαισθησία (%) | Ειδίκευση (%) |
|---|---|---|---|
| **BN** | 100 | 100 | 100 |
| **ANN** | 95 | 100 | 90 |
| **SVM** | 95 | 100 | 90 |
| **DT** | 90 | 90 | 90 |
| **RF** | 90 | 100 | 80 |

Παρατηρούμε λοιπόν ότι και κατά την δεύτερη αρχιτεκτονική, όπως και στις περισσότερες περιπτώσεις που περιγράφηκαν παραπάνω, μια υπεροχή του αλγορίθμου wrapper. Επίσης, βλέπουμε ότι η συνδυαστική ανάλυση των δεδομένων δίνει ως επί το πλείστον υψηλότερα αποτελέσματα σε σχέση με την μεμονωμένη εκπαίδευση ταξινομητών για κάθε τύπο δεδομένων ξεχωριστά.

Στον Πίνακας 47 που ακολουθεί παρουσιάζουμε μια συνολική σύγκριση ανάμεσα στην προτεινόμενη μεθοδολογίες και τις εργασίες που έχουν προταθεί στην βιβλιογραφία με σκοπό την πρόβλεψη επανεμφάνισης της νόσου είτε τον έγκαιρο προσδιορισμό απομακρυσμένης μετάστασης.

**Πίνακας 47:** Σύγκριση της προτεινόμενης μεθοδολογίας με τη βιβλιογραφία.

| Μεθοδολογία | # ασθενών | Ακρίβεια (%) |
|---|---|---|
| [115] | 22 | 86 |
| [116] | 66 | 88 |
| [117] | 79 | 77 |
| [118] | 39 | 76 |
| [119] | 75 | 87 |
| [120] | 25 | 85 |
| **Προτεινόμενη μεθοδολογία** | **41** | **100** |

Παρατηρούμε λοιπόν ότι η προτεινόμενη μεθοδολογία χάρη στην πολυεπίπεδη και πολυπαραμετρική ανάλυση που εφαρμόζει καταδεικνύει τους προεξάρχοντες παράγοντες στην πορεία εξέλιξης της νόσου και κατά συνέπεια εντοπίζει με μεγάλη ακρίβεια μια πιθανή υποτροπή.

### *Παρακολούθηση εξέλιξης της νόσου*

Στην παρούσα ενότητα, θα αναπτύξουμε το δεύτερο σκέλος της προτεινόμενης μεθοδολογικής ανάλυσης, που αφορά στην συλλογή δεδομένων σε τακτά διαστήματα μετά την θεραπεία και την συστηματική επεξεργασία και ανάλυσή τους, με σκοπό να

εντοπίσουμε την πιθανότητα ένας ασθενής να αναπτύξει επανεμφάνιση του καρκίνου, και μάλιστα να προσδιορίσουμε σε αδρές γραμμές το χρονικό πλαίσιο της επανεμφάνισης. Επικεντρωνόμαστε, λοιπόν, στα δεδομένα που είναι δυναμικά και οι τιμές τους ενδέχεται να μεταβάλλονται με το χρόνο, κατά το χρονικό διάστημα μετά την θεραπεία. Τα δεδομένα αυτά αποτελούν την έκφραση των γονιδίων όπως εξάγονται από το αίμα που κυκλοφορεί γύρω από την αρχική εντόπιση του καρκίνου. Σύμφωνα με το κλινικό σενάριο που ακολουθούμε και το πρωτόκολλο συλλογής δεδομένων, αίμα λαμβάνεται από κάθε ασθενή σε τακτά διαστήματα τριών μηνών, και για βάθος χρόνου 18 μηνών, δηλαδή σε 3, 6, 9, 12, 15 και 18 μήνες μετά την θεραπεία. Η συστηματική ανάλυση αυτών των μεταβαλλόμενων στο χρόνο δεδομένων, μπορεί να παρέχει σημαντικές πληροφορίες σχετικά με την εξέλιξη της νόσου, και να καταγράψει πιθανές διακυμάνσεις στην έκφραση συγκεκριμένων γονιδίων που να σηματοδοτούν εγκαίρως μια επικείμενη υποτροπή. Συνεπώς, σε πρώτο στάδιο επιθυμούμε να εντοπίσουμε τα γονίδια αυτά που σε βάθος χρόνου 18 μηνών εκφράζονται κατά διαφορετικό τρόπο ανάμεσα σε ασθενείς με και χωρίς επανεμφάνιση της ασθένειας. Ακολούθως, χρησιμοποιούμε τα γονίδια αυτά που αποτελούν τους πιο σημαντικούς δείκτες υποτροπής και φέρουν την υψηλότερη διακριτική ικανότητα, ώστε να υπολογίσουμε την πιθανότητα που έχει ένας ασθενής να υποτροπιάσει, καθώς και να εντάξουμε αυτήν την πρόβλεψη σε ένα χρονικό περιθώριο.

Τα βήματα που ακολουθούμε για τη συγκεκριμένη μεθοδολογική ανάλυση που αποσκοπεί στην παρακολούθηση της εξέλιξης της νόσου, παρουσιάζονται στην Εικόνα 49. Αρχικά, τα γενετικά δεδομένα υφίστανται κάποια βασικά στάδια προεπεξεργασίας ώστε να εξαλείψουμε εσφαλμένα και πλεονάζοντα δείγματα, στη συνέχεια εντοπίζουμε τα γονίδια με την μεγαλύτερη διακριτική ικανότητα και άρα τα πιο σημαντικά ως προς την εξέλιξη της νόσου. Αντλώντας πληροφορίες από σχετικές βιολογικές βάσεις, εμπλουτίζουμε το σύνολο των γονιδίων προσθέτοντας τις αλληλεπιδράσεις μεταξύ αυτών. Υπολογίζουμε επίσης για κάθε ασθενή έναν προσωποποιημένο δείκτη κινδύνου, που παρέχει επιπρόσθετη πληροφορία, συγκεκριμένη για κάθε ασθενή. Το δίκτυο των γονιδίων μαζί με τον προσωποποιημένο δείκτη κινδύνου συνιστούν την είσοδο σε ένα Δυναμικό Μπεϋζιανό Δίκτυο (DBN: Dynamic Bayesian Network) [134] που προβλέπει την πιθανότητα επανεμφάνισης σε συνάρτηση με τον χρόνο.

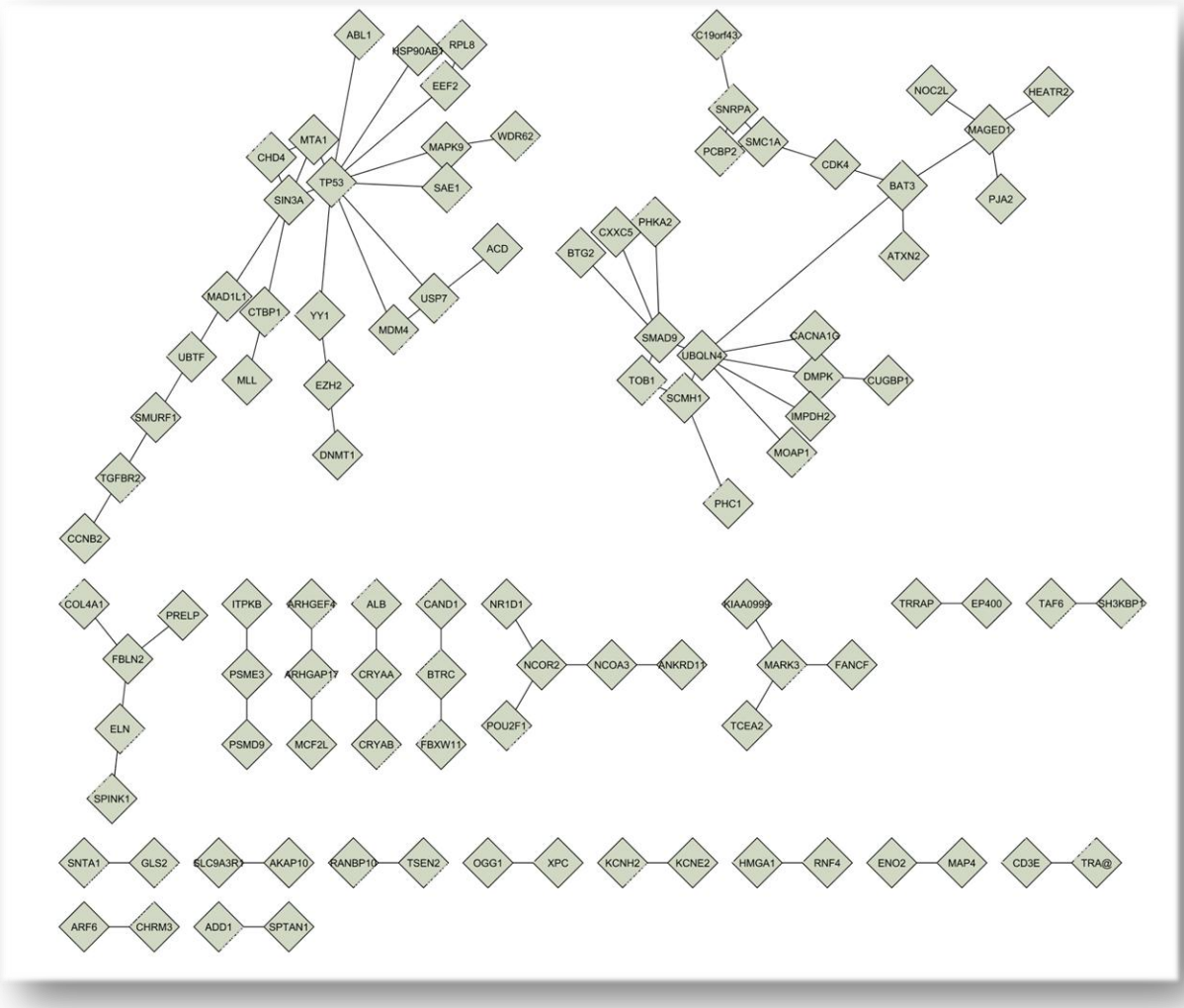

**Εικόνα 49:** Μεθοδολογία για την παρακολούθηση της εξέλιξης της νόσου στο χρόνο.

Από το αίμα που κυκλοφορεί στην περιοχή της αρχικής εντόπισης του καρκινικού ιστού, εξάγουμε το RNA από όπου και υπολογίζουμε την έκφραση 45015 γονιδίων. Αξίζει να σημειωθεί, ότι σε όλους τους ασθενείς χρησιμοποιήθηκε ακριβώς το ίδιο πρωτόκολλο για τον υπολογισμό της γονιδιακής έκφρασης, ώστε να εξαλείψουμε κάθε παράγοντα διακύμανσης πέραν της βιολογικής διαφοροποίησης των δειγμάτων. Κατά την αρχική προεπεξεργασία, αφαιρούμε γονίδια ελέγχου και διπλότυπες εγγραφές, καθώς επίσης και γονίδια με υψηλό αριθμό αγνώστων και ελλιπών τιμών. Προκύπτει λοιπόν ένα σύνολο 33491 γονιδίων, που χρησιμοποιείται στα επόμενα στάδια της προτεινόμενης μεθοδολογικής ανάλυσης.

Στη συνέχεια αναλύουμε την έκφραση των 33491 γονιδίων, ώστε να εντοπίσουμε αυτά που εκφράζονται σε βάθος χρόνου κατά σημαντικά διαφορετικό τρόπο μεταξύ ασθενών με και χωρίς επανεμφάνιση του καρκίνου. Για το σκοπό αυτό εφαρμόζουμε τον αλγόριθμο SAM [130] που αναλύει την διαφορική έκφραση των υπό μελέτη γονιδίων σε διάφορες

χρονικές στιγμές. Δημιουργούμε λοιπόν πολλαπλά σύνολα δεδομένων από αντιμεταθέσεις δειγμάτων και επιτελούμε σε αυτά Wilcoxon τεστ ώστε να εντοπίσουμε τα γονίδια με τις σημαντικότερες διακυμάνσεις. Η έξοδος αποτελείται από 2825 γονίδια στα οποία εφαρμόζουμε έναν επιπλέον περιορισμό, που αφορά το μέγεθος μεταβολής της έκφρασης του κάθε γονιδίου ανάμεσα στις δύο κατηγορίες ασθενών, οπότε λαμβάνουμε τελικά 1149 γονίδια.

Εν συνεχεία, εμπλουτίζουμε το σύνολο των γονιδίων που προέκυψαν με τις αλληλεπιδράσεις μεταξύ αυτών, όπως έχουν επιβεβαιωθεί πειραματικά, ώστε να καταγράψουμε πιο διεξοδικά το μοριακό υπόβαθρο της νόσου. Για το σκοπό αυτό αντλούμε αλληλεπιδράσεις από την βάση MiMI (Michigan Molecular Interactions) [101], που ουσιαστικά συνδυάζει πληροφορίες από ένα πλήθος σχετικών βάσεων όπως HPRD (Human Protein Reference Database) [44], IntAct [46], σύνολα δεδομένων από το Center for Cancer Systems Biology of Harvard [102] καθώς και πλήθος άλλων. Το δίκτυο αλληλεπιδράσεων μεταξύ των γονιδίων απεικονίζεται στην Εικόνα 50.

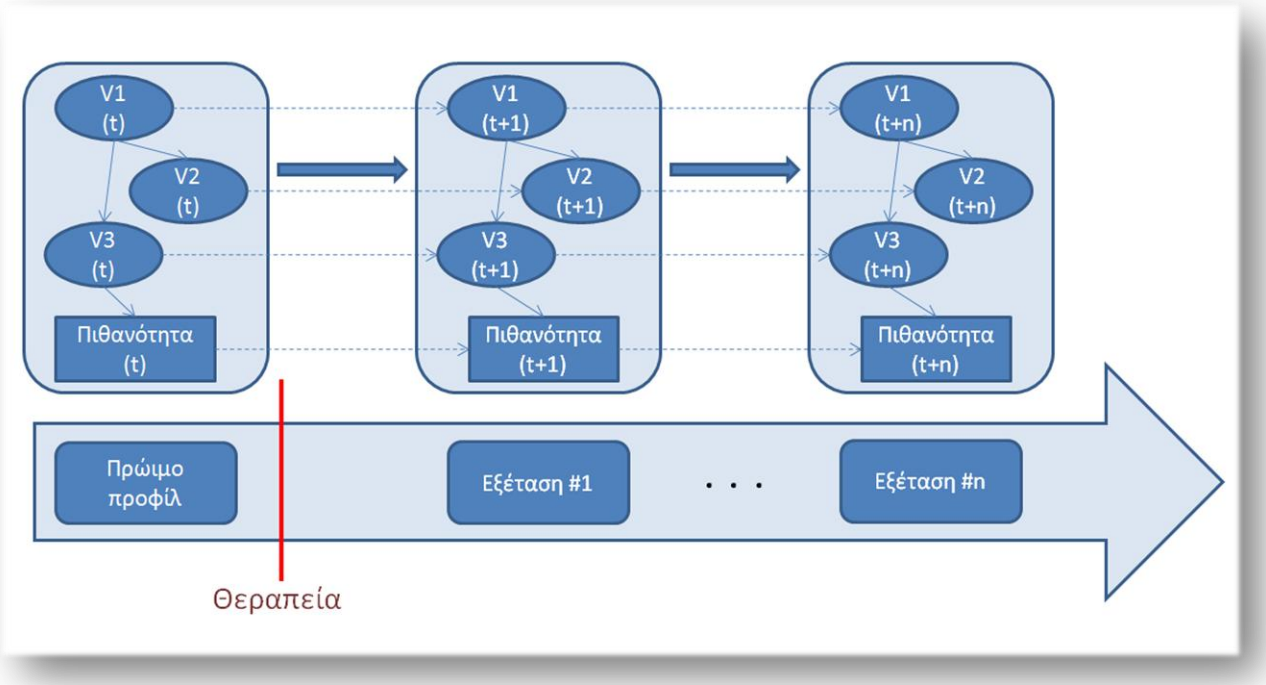



**Εικόνα 50:** Δίκτυο αλληλεπιδράσεων μεταξύ των πιο σημαντικών γονιδίων.

Επίσης, εξάγουμε έναν προσωποποιημένο δείκτη κινδύνου, που είναι συγκεκριμένος για κάθε ασθενή και περιλαμβάνει εκείνα τα γονίδια που για το συγκεκριμένο άτομο φαίνεται να μεταβάλλονται περισσότερο έπειτα από την θεραπεία. Τα γονίδια αυτά εξάγονται και αφορούν μεμονωμένους ασθενείς, σε αντίθεση με την ανάλυση που περιγράφηκε προηγουμένως και εντοπίζει τα γονίδια που εμφανίζουν την μεγαλύτερη διαφορά στην έκφρασή τους συνολικά ανάμεσα σε άτομα με και χωρίς επανεμφάνιση, οπότε και αφορούν το γενικό πληθυσμό που λαμβάνουμε υπόψη. Κατά την κατασκευή του προσωποποιημένου δείκτη για έναν ασθενή, συγκρίνουμε την έκφραση των γονιδίων του πριν τη θεραπεία, με την έκφραση των γονιδίων στους πρώτους μήνες έπειτα από την θεραπεία, οπότε και έχει υποβληθεί σε ενδεχόμενη χημειο/ακτινοθεραπεία που θα μεταβάλει την έκφραση των γονιδίων του. Θεωρώντας ότι η έκφραση των γονιδίων στην

πρώτη περίπτωση συνιστά το "καρκινικό προφίλ", ενώ στη δεύτερη το "μη-καρκινικό προφίλ", η σύγκρισή τους θα αναδείξει εκείνα τα γονίδια που για τον συγκεκριμένο ασθενή έχουν διαφοροποιηθεί σημαντικά ανάμεσα στις δύο αυτές ακραίες καταστάσεις, τα οποία και συνθέτουν τον προσωποποιημένο δείκτη κινδύνου. Εν συνεχεία, σύμφωνα με το κλινικό σενάριο εξάγουμε την έκφραση των γονιδίων του έπειτα από 3, 6, 9, 12, 15 και 18 μήνες και τα συγκρίνουμε εκ περιτροπής με το "καρκινικό" και "μη-καρκινικό προφίλ", μόνο για το υποσύνολο των γονιδίων του προσωποποιημένου δείκτη. Από κάθε σύγκριση υπολογίζουμε την συσχέτιση και την Ευκλείδια απόσταση μεταξύ των δύο διανυσμάτων, όπου αποτελούν μέτρα ποιοτικής και ποσοτικής ομοιότητας αντίστοιχα. Δηλαδή, η συσχέτιση εκφράζει την ομοιότητα στον τρόπο που μεταβάλλονται οι τιμές των γονιδίων και η Ευκλείδια απόσταση την πραγματική διαφορά σε επίπεδο τιμών μεταξύ των διανυσμάτων.

Στο επόμενο βήμα της προτεινόμενης μεθοδολογικής ανάλυσης, εισάγουμε σε ένα DBN τις τιμές των γονιδίων που εκφράζονται κατά τρόπο διαφορετικό ανάμεσα σε ασθενείς με και χωρίς επανεμφάνιση, καθώς επίσης και τις τιμές της συσχέτισης και Ευκλείδιας απόστασης όπως προέκυψαν προηγουμένως. Το DBN αποτελεί ουσιαστικά μια χρονική επέκταση του δικτύου Bayes σε ένα πεπερασμένο πλήθος διακριτών χρονικών στιγμών, όπως απεικονίζεται στην Εικόνα 51.




**Εικόνα 51:** Αρχιτεκτονική ενός DBN.

Ένα BN περιγράφεται ως *B=(G,P)* όπου *G* είναι ένα κατευθυνόμενος ακυκλικός γράφος, του οποίου οι κόμβοι αντιστοιχούν σε ένα σύνολο τυχαίων μεταβλητών $\mathbf{X} = \{x_1, x_2, ..., x_N\}$, και *P* είναι η από κοινού συνάρτηση πυκνότητας πιθανότητας των μεταβλητών στο *X*, όπως φαίνεται στην εξίσωση 17:

$$P(\mathbf{X}) = \prod_{i=1}^{N} P(x_i \mid \pi_G(x_i)) \tag{17}$$

όπου το $\pi_G(x)$ περιέχει τους γονείς της μεταβλητής *x* στο *G*. Ένα DBN ορίζεται ως το ζεύγος *DB=(B₀,B_trans)* όπου $B_0$ είναι ένα BN, με πρότερη κατανομή (prior distribution) $P(X_0)$ και $B_{trans}$ είναι ένα BN με δύο χρονικές στιγμές (2TBN) για το οποίο ισχύει $P(X_t|X_{t-1})$. Με παρόμοια λογική προκύπτει ένα DBN πολλαπλών χρονικών στιγμών, "ξετυλίγοντας" ένα 2TBN για *T* πλήθος επιθυμητών στιγμών. Η προκύπτουσα από κοινού συνάρτηση πυκνότητας πιθανότητας δίνεται από την εξίσωση 18:

$$P(\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X}_T) = \prod_{t=1}^{T} \prod_{i=1}^{N} P(x_i^t \mid \pi(x_i^t)) \tag{18}$$

Πρέπει εν συνεχεία να εκπαιδεύσουμε το DBN που προκύπτει ώστε να καθοριστούν κατάλληλες τιμές για τις αλληλεπιδράσεις/εξαρτήσεις μεταξύ των μεταβλητών που ανήκουν στην ίδια χρονική στιγμή, καθώς και για τις αλληλεπιδράσεις μεταξύ των διαφορετικών χρονικών στιγμών.

Όσον αφορά τις αλληλεπιδράσεις μεταξύ των μεταβλητών της ίδιας χρονικής στιγμής, αυτές λαμβάνονται από το δίκτυο αλληλεπίδρασης γονιδίων όπως καθορίστηκε προηγουμένως, και απεικονίζεται στην Εικόνα 50. Στη συνέχεια χρησιμοποιούμε τον αλγόριθμο Structural EM [135] για να καθορίσουμε περαιτέρω τις εξαρτήσεις μεταξύ των γονιδίων μιας χρονικής στιγμής και σε ποσοτικό επίπεδο. Ενώ για τον καθορισμό και βελτιστοποίηση των εξαρτήσεων μεταξύ των διαδοχικών χρονικών στιγμών χρησιμοποιούμε τον αλγόριθμο REVEAL [136]. Το εκπαιδευμένο DBN που προκύπτει αντικατοπτρίζει και την σχετική βιβλιογραφική γνώση όπως έχει εξαχθεί από βάσεις γονιδιακών αλληλεπιδράσεων αλλά έχει επανακαθοριστεί με βάση τα δεδομένα, και μπορεί να προβλέπει για κάθε ασθενή την πιθανότητα στον χρόνο για ανάπτυξη επανεμφάνισης.

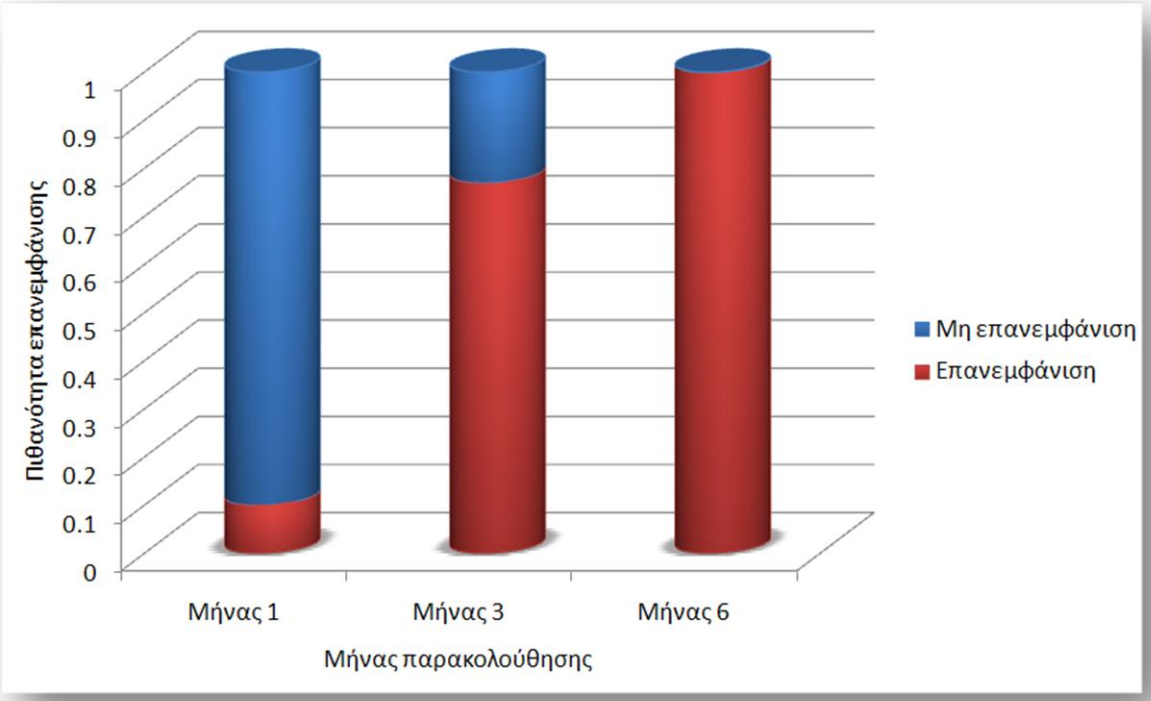
Για την αξιολόγηση του DBN χρησιμοποιούμε μια παραλλαγή της μεθόδου n-fold cross validation, κατά την οποία το n ισούται με το πλήθος των δειγμάτων. Δηλαδή, χρησιμοποιούμε τα n-1 δείγματα - στην περίπτωσή μας ασθενείς - για να εκπαιδεύσουμε το DBN και κατόπιν υπολογίζουμε την απόδοσή του στον εναπομείναντα ασθενή. Η διαδικασία συνεχίζεται κυκλικά μέχρις ότου όλοι οι ασθενείς έχουν βρεθεί ακριβώς μια φορά στο σύνολο ελέγχου, και ενδείκνυται για σύνολα δεδομένων με περιορισμένο αριθμό δειγμάτων. Τα αποτελέσματα που προέκυψαν απεικονίζονται στον Πίνακας 48.

**Πίνακας 48:** Αποτελέσματα που προέκυψαν από το DBN.

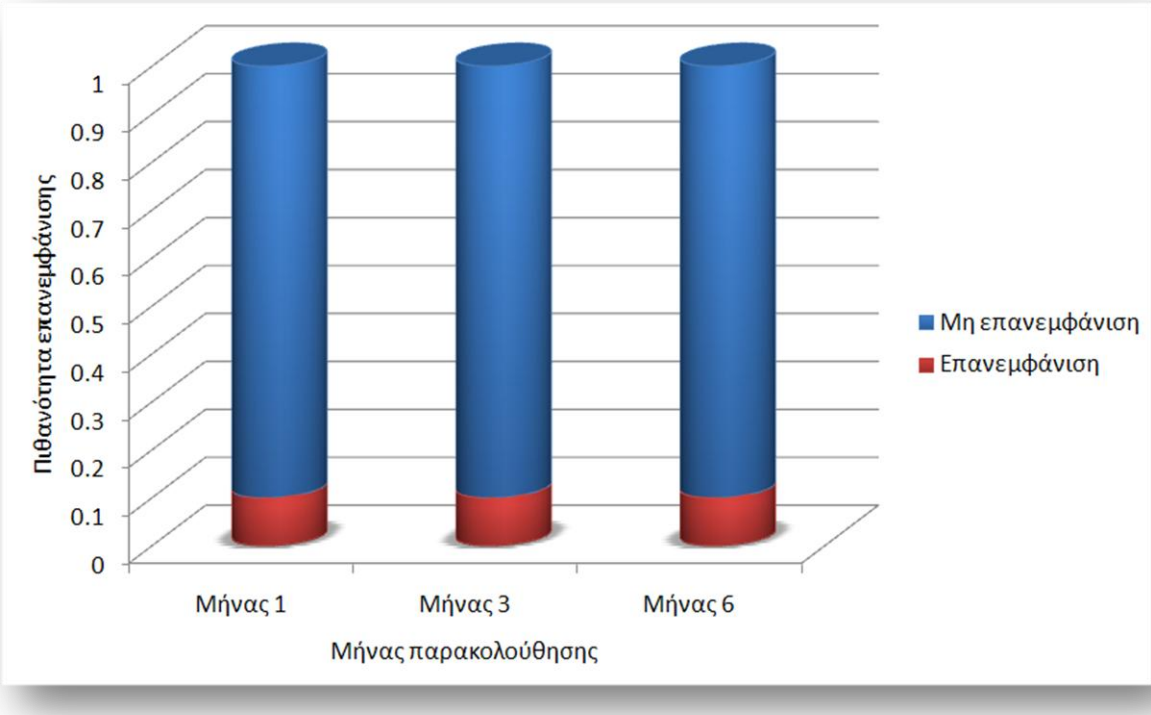
| Ασθενής | Έκβαση | Μήνας παρακολούθησης (M1) | Μήνας παρακολούθησης (M3) | Μήνας παρακολούθησης (M6) | Κατάσταση ασθενούς |
|---|---|---|---|---|---|
| **92** | Επανεμφάνιση | 0.1 | 0.1 | 0.1 | Μη επανεμφάνιση |
| | Μη επανεμφάνιση | 0.9 | 0.9 | 0.9 | |
| **93** | Επανεμφάνιση | 0.1 | 0.77 | 0.999 | Επανεμφάνιση (M6) |
| | Μη επανεμφάνιση | 0.9 | 0.23 | 0.001 | |
| **97** | Επανεμφάνιση | 0.1 | 0.1 | 0.1 | Μη επανεμφάνιση |
| | Μη επανεμφάνιση | 0.9 | 0.9 | 0.9 | |
| **146** | Επανεμφάνιση | 0.1 | 0.1 | 0.1 | Μη επανεμφάνιση |
| | Μη επανεμφάνιση | 0.9 | 0.9 | 0.9 | |
| **147** | Επανεμφάνιση | 0.1 | 0.1 | 0.1 | Μη επανεμφάνιση |
| | Μη επανεμφάνιση | 0.9 | 0.9 | 0.9 | |
| **156** | Επανεμφάνιση | 0.1 | 0.999 | 0.9933 | Επανεμφάνιση (M6) |
| | Μη επανεμφάνιση | 0.9 | 0.001 | 0.0067 | |

Αμέσως πιο κάτω, στην Εικόνα 52 παρατίθενται και εποπτικά με ραβδογράμματα οι πιθανότητες επανεμφάνισης για έναν ασθενή που παρουσίασε επανεμφάνιση (Εικόνα 52-α) και για έναν που παραμένει υγιής (Εικόνα 52-β).

(α)




(β)




**Εικόνα 52:** Εξέλιξη της πιθανότητας επανεμφάνισης στο χρόνο για (α) έναν ασθενή με επανεμφάνιση και (β) για έναν ασθενή χωρίς επανεμφάνιση.

Παρατηρούμε ότι οι προβλεπόμενες πιθανότητες επανεμφάνισης (κόκκινη μπάρα) για τους ασθενείς που πράγματι παρουσίασαν επανεμφάνιση της νόσου, αυξάνονται σταδιακά με την πάροδο των μηνών, ενώ αντίθετα στην περίπτωση των ατόμων χωρίς επανεμφάνιση, οι αντίστοιχες πιθανότητες παραμένουν χαμηλές και σταθερές σε όλη τη διάρκεια παρακολούθησης.

## *5.4 Συμπεράσματα*

Στο παρόν κεφάλαιο παρουσιάσαμε μια ολοκληρωμένη μεθοδολογία για την πρόβλεψη της επανεμφάνισης του καρκίνου. Συγκεκριμένα, εστιάσαμε στον στοματικό καρκίνο, που είναι μέσα στους 10 συχνότερους καρκίνους στον άνθρωπο και μάλιστα είναι και ιδιαιτέρως επιθετικός. Η προτεινόμενη μεθοδολογία, περιλαμβάνει την συγκέντρωση, ανάλυση και επεξεργασία μιας σειράς ετερογενών δεδομένων που καλύπτουν όλες τις εκφάνσεις της νόσου, από το μοριακό υπόβαθρο μέχρι την κλινική εικόνα του ασθενούς. Τα δεδομένα αποτελούνται από την έκφραση των γονιδίων στον καρκινικό ιστό και το κυκλοφορών αίμα, εικόνες αξονικής και μαγνητικής τομογραφίας της στοματοφαρυγγικής οδού καθώς και πληροφορίες από το κλινικό ιστορικό του ασθενούς. Αξίζει να σημειωθεί ότι κάποια από τα παραπάνω δεδομένα λαμβάνονται μάλιστα σε τακτά χρονικά διαστήματα κατά την διάρκεια της αποθεραπείας του ασθενούς οπότε και μοντελοποιούμε κατ' αυτόν τον τρόπο τον μηχανισμό δράσης και εξέλιξης της νόσου. Η ολιστική αυτή προσέγγιση μας επιτρέπει να αναδείξουμε τους παράγοντες που παίζουν καθοριστικό ρόλο στην επανεμφάνιση της νόσου σε μικροσκοπικό καθώς και σε μακροσκοπικό επίπεδο. Η γνώση του μηχανισμού εξέλιξης της νόσου μας βοηθά περαιτέρω ώστε να αποφανθούμε για την πορεία εξέλιξης της ασθένειας καθώς και την πιθανότητα μιας πιθανής υποτροπής και επανεμφάνισης της νόσου. Λόγω της επιθετικότητας και της διεισδυτικής φύσης του στοματικού καρκίνου, η έγκαιρη και κατά το δυνατόν ακριβέστερη γνώση της πιθανής πορείας της νόσου μπορεί να συμβάλλει σημαντικά στην πιο αποτελεσματική αντιμετώπισή της. Συγκεκριμένα, η πιθανότητα επανεμφάνισης που υπολογίζεται για κάθε ασθενή και σε συνάρτηση με την σχετική ιατρική βιβλιογραφία, μπορούμε να επαναπροσδιορίσουμε την θεραπευτική αγωγή κατάλληλα ώστε να ανταποκρίνεται καλύτερα στις ανάγκες του κάθε ασθενούς. Αν δηλαδή κάποιος ασθενής ανήκει στην ομάδα υψηλού κινδύνου όσον αφορά την επανεμφάνιση της νόσου, τότε ενδεχόμενα να

είναι προσφορότερη μια πιο εντατική θεραπεία (π.χ. πιο συχνές χημειοθεραπείες), ενώ αντίστροφα, αν κάποιος ασθενής έχει μικρή πιθανότητα να υποτροπιάσει τότε ο θεράπων ιατρός μπορεί να προτείνει μια πιο συντηρητική θεραπεία κατά τη διάρκεια της ύφεσης, παρεκκλίνοντας από την τυπική θεραπευτική και αποφεύγοντας κατ' αυτόν τον τρόπο τις δυσμενείς επιπτώσεις π.χ. από συστηματικές ακτινο/χημειο-θεραπείες.

# 6ο ΚΕΦΑΛΑΙΟ: Συμπεράσματα διατριβής

Στα πλαίσια της παρούσας διδακτορικής διατριβής αναπτύχθηκαν μια σειρά από μεθοδολογίες για την συστηματική ανάλυση και επεξεργασία βιολογικών δεδομένων που κατά κανόνα χαρακτηρίζονται από εκτεταμένο όγκο και μεγάλη πολυπλοκότητα. Ο αντικειμενικός στόχος της έρευνας είναι η εξαγωγή νέας, μη τετριμμένης γνώσης αναφορικά με λειτουργικά σημαντικές περιοχές των πρωτεϊνικών μορίων και βιολογικούς μηχανισμούς γενετικών ασθενειών.

Αρχικά μελετήσαμε πρωτεϊνικές περιοχές που φέρουν cis πεπτιδικούς δεσμούς και παίζουν ιδιαίτερο δομικό και λειτουργικό ρόλο για το πολυπεπτίδιο καθώς και για το κύτταρο ως σύνολο. Συγκεκριμένα εντοπίσαμε τους παράγοντες που επηρεάζουν σημαντικά την διαμόρφωση του πεπτιδικού δεσμού και μελετήσαμε τη συνεισφορά του κάθε γειτονικού αμινοξέος όσον αφορά τον cis/trans ισομερισμό. Η γνώση αυτή χρησιμοποιήθηκε περαιτέρω για την κατασκευή ενός ταξινομητή που λαμβάνοντας ως είσοδο αμιγώς την πρωτοταγή δομή, δηλαδή την αλληλουχία των αμινοξέων στην πολυπεπτιδική αλυσίδα, εξάγει μια σειρά από χαρακτηριστικά με σκοπό την πρόβλεψη της διαμόρφωσης του πεπτιδικού δεσμού μεταξύ δύο οποιονδήποτε αμινοξέων. Προς την ίδια κατεύθυνση και στην προσπάθεια να εντοπίσουμε και να αναπαραστήσουμε με εύληπτο τρόπο τους παράγοντες που συμβάλλουν σημαντικά στην cis διαμόρφωση, εξαγάγαμε ακολουθιακά πρότυπα που εμφανίζονται με υψηλή συχνότητα στις περιοχές που φέρουν cis πεπτιδικούς δεσμούς. Τα εξαχθέντα πρότυπα στη συνέχεια χρησιμοποιήθηκαν ώστε να ανακαλύψουμε συστηματικά αναδείξουμε συστηματικά τις λειτουργικές συσχετίσεις των cis πεπτιδικών δεσμών που λόγω της σπανιότητας τους έχουν μελετηθεί σποραδικά και μεμονωμένα.

Έπειτα εστιάσαμε το ενδιαφέρον μας στις ρεομορφικές πρωτεΐνες που όμοια με τους cis πεπτιδικούς δεσμούς, παίζουν ιδιαίτερο ρόλο στην λειτουργία του κυττάρου. Συγκεκριμένα ξεκινώντας από ένα σύνολο ρεομορφικών πρωτεϊνών που εκφράζονται στον άνθρωπο, «χτίζουμε» ένα δίκτυο πρωτεϊνικών αλληλεπιδράσεων. Στη συνέχεια εντοπίζουμε ακολουθιακά πρότυπα μέσω των οποίων οι ρεμορφικές πρωτεΐνες

αλληλεπιδρούν με ένα πλήθος ετερόκλητων πρωτεϊνικών μορίων και κατ' επέκταση επάγουν και επιτελούν τις λειτουργίες τους. Αξίζει να σημειωθεί επίσης ο καταλυτικός ρόλος που έχει βρεθεί ότι παίζουν οι ρεομορφικές πρωτεΐνες στην επαγωγή διαφόρων ασθενειών, όπως ο καρκίνος, τα καρδιαγγειακά νοσήματα, ο διαβήτης κ.ά., όπου η γνώση και συστηματική καταγραφή των ακολουθιακών προτύπων μέσω των οποίων οι ρεομορφικές πρωτεΐνες ασκούν τον λειτουργικό τους ρόλο μπορεί να οδηγήσει στην αντιμετώπισή τους.

Στο επόμενο βήμα μελετάμε την διαδικασία εξέλιξης του στοματικού καρκίνου και συγκεκριμένα στοχεύουμε στην πρόβλεψη μιας ενδεχόμενης επανεμφάνισής του. Προς αυτήν την κατεύθυνση, συνδυάζουμε βιολογικά δεδομένα προερχόμενα από τον καρκινικό ιστό, με πληροφορίες που συνιστούν το κλινικό προφίλ του ασθενούς, δηλαδή απεικονιστικά δεδομένα και τυπικές πληροφορίες από τον ιατρικό του φάκελο. Ο σκοπός μας είναι να ανακαλύψουμε τον βιολογικό μηχανισμό που καθοδηγεί την εξέλιξη της νόσου, καθώς και να αναδείξουμε τα κλινικά δεδομένα που συσχετίζονται με τις εκφάνσεις του καρκίνου σε μακροσκοπικό και φαινοτυπικό πλέον επίπεδο. Σε κάθε περίπτωση τα εξαχθέντα αποτελέσματα και συμπεράσματα πρέπει να συγκριθούν με την σχετική βιβλιογραφία αλλά πιο σημαντικά πρέπει να εφαρμοστούν πιλοτικά σε ένα μεγάλο αριθμό ασθενών ώστε να εξακριβωθεί η ακρίβεια και η αξιοπιστία τους. Τέλος, η προτεινόμενη μεθοδολογία που συνδυάζει κατά τρόπο συμπληρωματικό ετερογενή δεδομένα, μπορεί να χρησιμοποιηθεί με τις κατάλληλες τροποποιήσεις και σε άλλους τύπους καρκίνου, με σκοπό την πληρέστερη καταγραφή και παρακολούθηση της εξέλιξης της νόσου.

# Περίληψη διδακτορικής διατριβής

Η παρούσα διδακτορική διατριβή εστιάζεται στην ανάπτυξη και εφαρμογή ευφυών υπολογιστικών μεθόδων για την οργάνωση, επεξεργασία, ανάλυση και κατανόηση μεγάλου όγκου βιολογικών δεδομένων. Απώτερος σκοπός είναι η σταδιακή αποκρυπτογράφηση του γενετικού υποβάθρου ασθενειών και κατ' επέκταση την αποδοτικότερη και στοχευμένη αντιμετώπισή τους. Τα στάδια ανάπτυξης και υλοποίησης της παρούσας μελέτης ανάγονται σε ένα επαγωγικό πλαίσιο έρευνας που δομείται σταδιακά από το ειδικό προς το γενικό. Η έρευνα ξεκινάει από το κατώτερο επίπεδο οργάνωσης των πρωτεϊνών, την πρωτοταγή δομή, όπου μελετάμε τους δεσμούς που αναπτύσσονται μεταξύ των αμινοξέων, που αποτελούν τους δομικούς λίθους των πρωτεϊνών. Στη συνέχεια, εξετάζουμε τις εξαρτήσεις και αλληλεπιδράσεις μεταξύ των πρωτεϊνικών μορίων, εστιάζοντας συγκεκριμένα σε πρωτεΐνες που είτε μερικώς είτε πλήρως είναι ρεομορφικές, δηλαδή δεν φέρουν σαφή τριτοταγή δομή. Οι ρεομορφικές πρωτεΐνες λόγω της εγγενούς ευμεταβλητότητάς τους, έχει βρεθεί ότι συμμετέχουν σε πλήθος κυτταρικών λειτουργιών και συνακόλουθα έχουν συσχετιστεί με την επαγωγή σοβαρών ασθενειών. Στη συνέχεια της διδακτορικής διατριβής επεκτείνουμε και συνδυάζουμε την ανάλυση από το μικροσκοπικό επίπεδο με μακροσκοπικές παρατηρήσεις και δεδομένα. Συγκεκριμένα, συλλέγουμε βιολογικά δεδομένα που αφορούν την κυτταρική και συστημική λειτουργία του οργανισμού, καθώς και κλινικά δεδομένα (ιατρικό ιστορικό και απεικονιστικά δεδομένα) που αφορούν σε φαινοτυπικό πλέον επίπεδο ανατομικές οντότητες και τον οργανισμό ως σύνολο. Εφαρμόζουμε αυτή την πολύπλευρη και πολυπαραγοντική ανάλυση σε μια πολύπλοκη νόσο όπως ο καρκίνος – και πιο συγκεκριμένα στον στοματικό καρκίνο – που φέρει εκφάνσεις σε όλα τα επίπεδα της φυσιολογίας του οργανισμού.

Στο 1ο κεφάλαιο παραθέτουμε εισαγωγικές έννοιες για όλα τα επιμέρους ερευνητικά πεδία που εξετάζονται, είτε άπτονται της παρούσας διδακτορικής διατριβής. Συγκεκριμένα, οι τομείς όπου αναλύονται οι βασικές έννοιες και παρουσιάζεται η σχετική βιβλιογραφία είναι: τα επίπεδα οργάνωσης των πρωτεϊνών, ο ισομερισμός του πεπτιδικού δεσμού, τα δίκτυα πρωτεϊνικών αλληλεπιδράσεων, οι ρεομορφικές πρωτεΐνες και ο

λειτουργικός τους ρόλος, η μικροσκοπική και μακροσκοπική θεώρηση του καρκίνου και ιδιαίτερα ο μηχανισμός εξέλιξης του στοματικού καρκίνου.

Στο 2ο κεφάλαιο παρουσιάζουμε μια μεθοδολογία για την πρόβλεψη της διαμόρφωσης του πεπτιδικού δεσμού, μεταξύ των αμινοξέων μιας πρωτεΐνης. Εξάγοντας ένα πλήθος χαρακτηριστικών με βιολογική σημασία αμιγώς από την πρωτοταγή ακολουθία των αμινοξέων, προβλέπουμε την διαμόρφωση του πεπτιδικού δεσμού. Συνεχίζοντας στο ίδιο ερευνητικό πεδίο, στο 3ο κεφάλαιο εστιάζουμε το ενδιαφέρον μας στην εξαγωγή ακολουθιακών προτύπων, που χαρακτηρίζουν και περιγράφουν με έναν εύληπτο τρόπο τις ισομερείς διαμορφώσεις του πεπτιδικού δεσμού. Στη συνέχεια χρησιμοποιούμε τα εξαχθέντα πρότυπα ώστε να αναδείξουμε τις λειτουργικές συσχετίσεις των cis πεπτιδικών δεσμών.

Στο 4ο κεφάλαιο μελετάμε τα δίκτυα αλληλεπίδρασης των ρεομορφικών πρωτεϊνών και μέσα από αυτά αναδεικνύουμε τον τρόπο με τον οποίο επάγουν τον λειτουργικό τους ρόλο καθώς και την συνακόλουθη συμμετοχή τους σε πλήθος ασθενειών στον άνθρωπο. Εντοπίζουμε κατά τρόπο συστηματικό ακολουθιακά πρότυπα μέσω των οποίων οι ρεομορφικές πρωτεΐνες αλληλεπιδρούν με ένα πλήθος πρωτεϊνικών μορίων και επιτελούν τον ετερόκλητο λειτουργικό τους ρόλο.

Στο 5ο κεφάλαιο αναπτύσσουμε μια ολιστική και συνδυαστική προσέγγιση με σκοπό την έγκαιρη πρόβλεψη της επανεμφάνισης του στοματικού καρκίνου. Συγκεκριμένα i) αναλύουμε γενετικά δεδομένα για να εντοπίσουμε σε μικροσκοπικό επίπεδο τον βιολογικό μηχανισμό που καθοδηγεί την εξέλιξη της νόσου και στη συνέχεια ii) αναλύουμε πληροφορίες σχετικά με τις μακροσκοπικές εκφάνσεις του στοματικού καρκίνου σε απεικονιστικά δεδομένα καθώς και το κλινικό προφίλ του ασθενούς. Η προτεινόμενη πολυπαραγοντική και πολυεπίπεδη ανάλυση μας βοηθά να εντοπίσουμε τους παράγοντες που επιδρούν καθοριστικά στην εξέλιξη της νόσου και κατ' επέκταση να ανιχνεύσουμε έγκαιρα και με ακρίβεια μια ενδεχόμενη υποτροπή.

Η συνεισφορά της παρούσας διδακτορικής διατριβής εντοπίζεται στα ακόλουθα σημεία: (i) στην ανάδειξη του βιολογικού μηχανισμού που καθορίζει την διαμόρφωση του πεπτιδικού δεσμού, (ii) στον συστηματικό εντοπισμό λειτουργικών συσχετίσεων των πρωτεϊνικών περιοχών που φέρουν cis πεπτιδικούς δεσμούς, (iii) στην εξαγωγή ακολουθιακών προτύπων που σηματοδοτούν και επάγουν τις αλληλεπιδράσεις και λειτουργίες των ρεομορφικών πρωτεϊνών, (iv) στην συστηματική καταγραφή του τρόπου με τον οποίο οι ρεομορφικές πρωτεΐνες συμμετέχουν σε γενετικές ασθένειες, (v) στην

μελέτη του γενετικού υποβάθρου πολυπαραγοντικών ασθενειών (στοματικός καρκίνος) και την ανάδειξη γενετικών παραγόντων που συμβάλλουν στην εξέλιξή τους, (vi) στη συνδυαστική ανάλυση κλινικών και βιολογικών δεδομένων για την πολύπλευρη πλαισίωση γενετικών ασθενειών, καταγράφοντας μεταβολές στα κύτταρα, τα συστήματα, τους ιστούς, μέχρι και ολόκληρο τον οργανισμό.

.

# Summary in English

*PhD dissertation of Konstantinos P. Exarchos*

This thesis focuses on the development and implementation of intelligent computational methods for extending the use of biological data as well as the organization, processing, analysis and understanding of these data. The uttermost aim is to build upon the gained conclusions towards deciphering the genetic background of diseases and consequently facilitate treatment in an efficient and targeted manner. The study in centered around an inductive research framework where we proceed gradually from the specific to the general. The research starts from the lowest level of proteins organization, the primary structure, by studying the bonds developed between the amino acids. Next, we analyze the dependencies and interactions among protein molecules, specifically focusing on proteins that are either partially or fully disordered, aka unstructured, which in their native state lack specific 3D structure. Due to their conformational flexibility they are often found to interact with a multitude of protein molecules, hence they are involved with many cellular functions as well as the induction of severe diseases. Afterwards, we aim to extend and combine the analysis from the microscopic level with macroscopic observations and data. Specifically, we collect and analyze biological data on cellular and systemic organism functions, coupled with clinical data (i.e. medical history and imaging modalities) referring to the phenotypic level of anatomical entities and the organism as a whole. We apply this multilevel and multiparametric analysis in a complex disease like cancer - and more specifically oral cancer - which has manifestations in all levels of human physiology.

In the first chapter we describe introductory concepts for all individual research areas as well as relevant concepts in order to facilitate fluency and overall coherency. Specifically, we describe in detail and provide all relevant literature for the following topics: the levels of protein structural organization, the isomerization of the peptide bond, protein interaction networks, disordered proteins and their functional implications, the microscopic and macroscopic view of cancer especially the mechanism of oral cancer progression.

The second chapter presents a methodology for predicting the formation of the peptide bond between the amino acids of a protein. We extract a multitude of features with

biological significance solely from the primary amino acid sequence in order to predict the conformation of the peptide bond. Working on the same research area (i.e. the isomers of the peptide bond), in the third chapter we move on to extract sequence patterns that characterize and capture in a systematic manner the underlying mechanism dictating the configuration of the peptide bond. The extracted patterns are subsequently employed in order to uncover the functional implications of cis peptide bonds.

In the fourth chapter we describe protein interaction networks focusing specifically on disordered proteins, which owed to their inherent flexibility exhibit considerable binding promiscuity and dense interactome. Hence, they have been implicated with a wide range of cellular functions as well as the induction of severe diseases. We propose a methodological analysis in order to extract in a systematic manner sequence patterns that mediate the interaction and hence functioning of disordered proteins. These patterns can serve as a reference point in order to unravel the induction and progression of several diseases where disordered proteins are involved.

In the fifth chapter we present an orchestrated approach towards the timely prediction of oral cancer reoccurrence. More specifically we i) analyze genetic data in order to identify at a microscopic level the molecular mechanism that guides the evolution of the disease, and additionally ii) gather and explore information on the macroscopic manifestations of oral cancer, i.e. data from the patient's clinical profile as well as features extracted from imaging modalities of the head and neck. The proposed multivariate and multilevel analysis pinpoints the most prominent factors affecting the progression of oral cancer and hence facilitates the dully identification of a potential disease relapse.

The contribution of this thesis is identified in the following specific areas: (i) the emergence of the biological mechanism that determines the conformation of the peptide bond, (ii) the systematic identification of functionalities associated with protein regions containing cis peptide bonds, (iii) the extraction of sequence patterns mediating the functional role of disordered proteins, (iv) the systematic unraveling of how disordered proteins are involved with the induction of several genetic diseases, (v) to study and analyze the genetic background of multifactorial diseases (oral cancer) and pinpoint the genetic factors dictating the disease progression, (vi) the orchestrated analysis of heterogeneous clinical and genomic data, in a complementary manner in order to capture multilevel disease perturbations ranging from the cellular level, to the tissue phenotype and the organism as a whole.

# Παράρτημα I

*Ακολουθιακά πρότυπα cis-Pro πεπτιδικών δεσμών*

### Ακριβής εξαγωγή προτύπων

| Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ |
|---|---|---|---|---|---|---|---|
| SP.NP.G | 0.99 | NPTG | 0.95 | RL.P.T | 0.93 | P.LP.F | 0.91 |
| V...EP...H | 0.99 | LC....P | 0.95 | N..P.PE | 0.93 | SPN.P | 0.91 |
| GPY.G | 0.99 | T.P...EG | 0.95 | S...PTG | 0.93 | N....FP | 0.91 |
| G...GPY | 0.98 | SPYP | 0.95 | Y...N.V | 0.93 | P....FK | 0.91 |
| P.NPTG | 0.98 | F..PPF | 0.95 | P.NP.G | 0.92 | TSP....G | 0.91 |
| PNNP.G | 0.98 | PN.P.G | 0.95 | SP.NP | 0.92 | ST.....N | 0.91 |
| PY..SG | 0.98 | T.PY..S | 0.94 | V.....QP | 0.92 | EPF...V | 0.91 |
| NNPT | 0.97 | Y.Y..N | 0.94 | V.....V.Y | 0.92 | NN..G | 0.90 |
| V.....N..T | 0.97 | SP.N.T | 0.94 | LT..P....V | 0.92 | MTL | 0.90 |
| P..YP.K | 0.97 | S.S.....N | 0.94 | WP..P.D | 0.92 | S..WP | 0.90 |
| N...P.PE | 0.97 | T.P.N..G | 0.94 | LK.P...I | 0.92 | RE.....I | 0.90 |
| GPY...G | 0.97 | V..T..F | 0.93 | A..YGP | 0.92 | K...RI | 0.90 |
| SP.N..G | 0.97 | V.....V..H | 0.93 | TK.....F | 0.92 | T..PP.A | 0.90 |
| S..NP.G | 0.96 | G..SP...V | 0.93 | V.G..H | 0.92 | PPAT | 0.90 |
| T...NP.G | 0.96 | L...EP...T | 0.93 | GPY...S | 0.92 | W.NG | 0.90 |
| PYG.S | 0.96 | SY.....N | 0.93 | S..N.T | 0.92 | N....FT | 0.90 |
| QL.......Y | 0.96 | S.N.P.G | 0.93 | G.Y.G | 0.92 | G...P..G.A | 0.90 |
| R...Y.P | 0.96 | G..VGP | 0.93 | P.F...S | 0.91 | PE..P..A | 0.90 |
| N.K..F | 0.96 | AVGP | 0.93 | T...N..G | 0.91 | P..CN | 0.90 |
| PPF...S | 0.96 | REPF | 0.93 | TV....H | 0.91 | I..PGL | 0.90 |
| A..EPF | 0.96 | E.P.L.Y | 0.93 | PAT.E | 0.91 | | |
| G..GP....S | 0.96 | G.QP..V | 0.93 | NN.T | 0.91 | | |
| V....P..V.Y | 0.96 | P..T.P..L | 0.93 | P..Y..K | 0.91 | | |

### Χημική ομαδοποίηση αμινοξέων

| Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ |
|---|---|---|---|---|---|---|---|
| [ST]P.NPTG | 1.00 | T.PY..S | 0.94 | T.P...[DE]G | 0.92 | [AG].[FY].G.[AG] | 0.91 |
| [ST]PNNP.G | 1.00 | [AG].[KR]....P.[ST] | 0.94 | [ST]...PA...N | 0.92 | [ILMV].S.A..[ILMV] | 0.91 |
| [ILMV]..P.NP.G | 1.00 | E.P[FY]..V | 0.94 | T...N[QN] | 0.92 | [FY].S.[AG].[ILMV] | 0.91 |
| [ST]PN..T[AG] | 1.00 | [DE]...AG.[ILMV] | 0.94 | A[ST]..P.P | 0.92 | S.E.P.[ST] | 0.91 |
| [ST]P.NP.G | 0.99 | [AG]..[FY][DE].G | 0.94 | [ST].NPP | 0.92 | S.S.P....[QN] | 0.91 |
| [ILMV][KR]EP[FY] | 0.99 | [FY]....P[FY]..I | 0.94 | [ILMV][AG][FY]P...[KR] | 0.92 | G.[FY].[AG][ILMV] | 0.91 |
| SP.NP.G | 0.99 | [FY]..[ILMV][AG].F | 0.94 | [ILMV]..GPY | 0.92 | [ILMV].V...P..[ILMV] | 0.91 |
| [ILMV][KR]EPF | 0.99 | [AG].[ILMV].....[QN]P | 0.94 | Y.Y..[QN] | 0.92 | [ILMV]...P.[ILMV].Y[ILMV] | 0.91 |
| [ILMV]....[QN]P.G | 0.99 | [FY][QN].GP | 0.94 | P.[FY]PS | 0.92 | T..[ST].PP | 0.91 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [QN]..FV[FY] | 0.99 | [FY][ILMV].S.[AG] | 0.94 | [FY]..KAP | 0.92 | [ST].CN | 0.91 |
| E.GP[FY] | 0.99 | S.S.....N | 0.94 | PC[KR][ILMV] | 0.92 | E....P.[QN][FY] | 0.91 |
| GPY.G | 0.99 | P.[KR].....I | 0.94 | T[ILMV].....[ILMV]R | 0.92 | [ST].[FY]...[AG]G | 0.91 |
| PNNP[ST] | 0.99 | Q[ILMV]..P....Y | 0.94 | [ST]YP....[QN] | 0.92 | [ST].TP..[ILMV][ST] | 0.91 |
| [ILMV].[ST].N...G | 0.99 | [ILMV]...[AG]P..IA | 0.94 | L.[AG].P[AG]..V | 0.92 | K.PF.[DE] | 0.91 |
| GY..[ILMV].K | 0.99 | [ILMV]..[AG].[ILMV]P[ILMV] | 0.94 | [ILMV][KR]E.[FY] | 0.92 | [ILMV][ST].[DE].....V | 0.91 |
| C[ST]..NP | 0.99 | [ILMV]..[DE][AG]P.L | 0.94 | [ILMV][KR]..[FY]..[FY] | 0.92 | [ILMV]...P[ILMV]..S[DE] | 0.91 |
| V...EP...H | 0.99 | [ILMV]RE.[FY] | 0.94 | REP....[ILMV] | 0.92 | G..[KR][ILMV].A | 0.91 |
| [QN][ST]P.N.[ST] | 0.99 | [ILMV].P.P[ILMV].D | 0.94 | [ILMV]...[AG]P[FY]..[ILMV] | 0.92 | P.P[AG]..T | 0.91 |
| [ILMV]...[ST]P.N.T | 0.99 | [ILMV]..[QN].P[ST][FY] | 0.94 | P.NP.G | 0.92 | Y[FY]..Y | 0.91 |
| P[FY]PE.[AG] | 0.99 | [ILMV][FY]E...[FY] | 0.94 | SP.NP | 0.92 | [ILMV].....[ILMV]K.[DE] | 0.91 |
| G...GPY | 0.98 | P[ILMV][KR][ILMV].A | 0.94 | [AG]P[FY]..I | 0.92 | D...[AG]G.[ILMV] | 0.91 |
| C[ST]P...T | 0.98 | [ILMV]P..P..V.[ILMV] | 0.94 | [FY]..[FY]P...[QN] | 0.92 | [FY]P.[DE].N | 0.91 |
| [ST].[QN].P[ST]G | 0.98 | V[ST][QN][FY] | 0.94 | P...PF[ST] | 0.92 | P...[ST]G.[ILMV] | 0.91 |
| IG[FY]....[KR] | 0.98 | [ILMV]G..V[ILMV][ILMV] | 0.94 | P.PH[AG] | 0.92 | G..[KR]L.[AG] | 0.91 |
| [ILMV]I....[FY]..[FY] | 0.98 | [FY]P[KR]K | 0.94 | [KR][ST][ILMV]...D | 0.92 | [KR]....P[QN]..G | 0.91 |
| [ST].FDG | 0.98 | [FY].[AG].N.[ILMV] | 0.94 | [ILMV].[ILMV][ILMV]EP[ILMV] | 0.92 | [AG]P.[DE][ILMV].D | 0.91 |
| [ILMV]T....L..V | 0.98 | NP[ST]G | 0.94 | [ILMV]V....[ST].S | 0.92 | [ILMV]...[AG]P[FY]...[AG] | 0.91 |
| SYP....[QN] | 0.98 | [ILMV].[ILMV]..P.F.[ST] | 0.94 | V.....QP | 0.92 | TA.....[QN] | 0.91 |
| TLA[FY] | 0.98 | [ILMV].[ILMV]..P..V.[FY] | 0.94 | I...GP[FY] | 0.92 | [ST]G[DE].[AG] | 0.91 |
| [KR].[FY].A[AG] | 0.98 | [AG].[FY]P[KR].[ILMV] | 0.94 | L..GP[FY] | 0.92 | [FY]..PP...G | 0.91 |
| PY..SG | 0.98 | [AG].[QN]PP..[ILMV] | 0.94 | [FY][AG][AG].[QN] | 0.92 | [AG][ILMV]Y...[DE] | 0.91 |
| [ILMV][KR]..[FY]P.[FY] | 0.98 | [ILMV][ILMV]....[FY]..Y | 0.94 | A..YGP | 0.92 | [KR]E.....I[ILMV] | 0.91 |
| I[AG][FY]....K | 0.98 | [ST]...P..S.D | 0.94 | [ILMV][ST]..P.[ILMV]..V | 0.92 | I[AG].P..[ILMV][KR] | 0.91 |
| G.[FY]P...T | 0.98 | [ST]..NP.G | 0.94 | LK.P...I | 0.92 | [QN]...FP...[DE] | 0.91 |
| I[AG][FY]....[KR] | 0.98 | [ILMV].GPY | 0.94 | [FY]..SPA | 0.92 | G..[AG]PY | 0.91 |
| L[ST]....L..V | 0.98 | [ST]P[QN]..T | 0.94 | S.....Y..[QN] | 0.92 | L.[AG]PY | 0.91 |
| [ILMV][ST]....L..V | 0.97 | PF[KR][FY] | 0.93 | V..[KR]...[KR][ILMV] | 0.92 | [ILMV]...GP...[ILMV]V | 0.91 |
| [FY][DE].PP[FY] | 0.97 | [QN]..MP[ILMV] | 0.93 | GPY...S | 0.92 | [FY]G.PP | 0.91 |
| PPD..[FY] | 0.97 | [ILMV]I...[AG]F | 0.93 | GPY[AG] | 0.92 | G[FY][FY].....[ILMV] | 0.91 |
| P..YP.K | 0.97 | PPF.[FY] | 0.93 | V.....V.Y | 0.92 | [KR]...P[FY]...F | 0.91 |
| V.....N..T | 0.97 | [DE]PPF | 0.93 | E.....[AG]N | 0.92 | [AG]...[FY]P..Y | 0.91 |
| NNPT | 0.97 | V.....V..H | 0.93 | D.[ST]..P.P | 0.92 | [ILMV]Q.P..I | 0.91 |
| GPY...G | 0.97 | [ILMV][ST]..A...[QN] | 0.93 | V.G..H | 0.92 | [ILMV]Q.P.[FY] | 0.91 |
| [ILMV]....PAT.E | 0.97 | [ST]...GP....D | 0.93 | [ILMV]...[AG].V.[FY] | 0.92 | GP.[ILMV][ILMV]A | 0.91 |
| [ILMV][ILMV]..P...S[FY] | 0.97 | [QN].....WA | 0.93 | [AG].Y.[KR].[ILMV] | 0.92 | ST.....N | 0.91 |
| C.P..P.[AG] | 0.97 | [ILMV][ILMV]..P...N[ST] | 0.93 | [QN].[ILMV]..P...[ILMV][QN] | 0.92 | [ILMV]..P[FY]..[KR][FY] | 0.91 |
| [ILMV]....P.NP.G | 0.97 | [ILMV].G..[KR]..A | 0.93 | P..P.[DE].I | 0.92 | [ILMV][ILMV]V.Y | 0.91 |
| [DE]...FP.N | 0.97 | [AG].[AG][DE]..N | 0.93 | P.[FY]...F | 0.92 | V[ST]..[FY].[DE] | 0.91 |
| L.[ST][ILMV].P....T | 0.97 | [ILMV].E.[FY]..[DE] | 0.93 | [QN].P.[QN][FY] | 0.92 | G[FY]P.[ILMV][ILMV] | 0.91 |
| [ILMV][FY]..[ST]P..P | 0.97 | N....P.N.[ILMV] | 0.93 | [ILMV]T..[FY]..[KR] | 0.92 | G[ILMV]G[FY] | 0.91 |
| [ILMV]...[AG]P.F.[ST] | 0.97 | M..P.I.[ILMV] | 0.93 | G..R.[ILMV][AG] | 0.92 | [ILMV].[ILMV]E.P...[ST] | 0.91 |
| [AG].[QN]PP..L | 0.97 | T...[FY]P..[FY] | 0.93 | [DE]..P.P.[ILMV]L | 0.92 | G[ILMV]Y.....[ILMV] | 0.91 |
| G.P[ILMV][ILMV][ILMV] | 0.97 | S.N.P.G | 0.93 | [AG]..GP..[ILMV].[DE] | 0.92 | N....FP | 0.91 |

| K | | | | | | | |
|---|---|---|---|---|---|---|---|
| [ILMV]...[ST]P.N..G | 0.97 | [QN]..[AG].[FY][DE] | 0.93 | WP..P.D | 0.92 | P....FK | 0.91 |
| [ILMV]...[ST]P.NP | 0.97 | E.....[ILMV].[DE]E | 0.93 | LT..P....V | 0.92 | [ILMV]....I[ILMV]K | 0.91 |
| G[FY]P.[ILMV].[KR] | 0.97 | [AG]...G.[ST]D | 0.93 | S..GP.[KR] | 0.92 | V.[KR][KR]....[ILMV] | 0.91 |
| N...P.PE | 0.97 | [ILMV].[ST]...P.G | 0.93 | T.F.[AG]P | 0.92 | [QN]..P[FY]N | 0.91 |
| V.L[ST]A | 0.97 | [ILMV]...G.V[FY] | 0.93 | A...P.P[KR] | 0.92 | P[ILMV]R[ILMV].[AG] | 0.91 |
| SP.N..G | 0.97 | V..T..F | 0.93 | A.P.P[FY] | 0.92 | N...P[FY].[DE] | 0.91 |
| [ST]LA[FY] | 0.97 | [ST].[ST]....D[KR] | 0.93 | T.[AG][FY].[ST] | 0.92 | [FY].[AG].[QN].[ILMV] | 0.90 |
| Y.[AG].N.[ILMV] | 0.97 | [ILMV][ILMV].TP.[FY] | 0.93 | A..[FY][AG].[DE] | 0.92 | GPY[ILMV] | 0.90 |
| P....P..L[FY] | 0.97 | [ILMV][ILMV].P..[ILMV].Y | 0.93 | [ST][AG].[ST][KR] | 0.92 | RE.....I | 0.90 |
| GPY.[AG] | 0.97 | PY.[AG].G | 0.93 | [FY][QN].[AG]...[ILMV] | 0.92 | NN..G | 0.90 |
| C[ST]P..P | 0.97 | [KR].PF...V | 0.93 | [KR].[QN][AG].[FY] | 0.92 | R..F..[FY] | 0.90 |
| [ST].FD[AG] | 0.97 | L...EP...T | 0.93 | [ILMV].....[FY][ILMV].Y | 0.92 | S..WP | 0.90 |
| [ILMV]...[AG]PF..[ILMV] | 0.97 | [AG].[ST]P[AG]..V | 0.93 | PP.[KR].K | 0.92 | MTL | 0.90 |
| S..NP.G | 0.96 | [ILMV].[ST]P[AG]...[FY] | 0.93 | [ILMV].[AG]P.[KR]..[AG] | 0.92 | G[FY]..[ILMV].[KR] | 0.90 |
| [ILMV][ILMV][ILMV]..P...[ILMV][KR] | 0.96 | [ILMV]....[ILMV]V.P | 0.93 | [AG][FY]P..[ILMV][KR] | 0.92 | V...[FY]...V | 0.90 |
| [ILMV][QN]....A.T | 0.96 | [ILMV].[ILMV]V.PV | 0.93 | WS.......[QN] | 0.92 | [ST]..P.W.[ILMV] | 0.90 |
| [KR]EP[FY]...[ILMV] | 0.96 | [ILMV]..G..F[DE] | 0.93 | [ST]L.....[ILMV]R | 0.92 | GP[FY]..[ST] | 0.90 |
| [FY].[AG][AG][QN].[ILMV] | 0.96 | [ST].....[FY]S[QN] | 0.93 | [QN].P[FY].I | 0.92 | PF.[FY][KR] | 0.90 |
| [DE][QN]...P.P | 0.96 | [ILMV]..[AG]P.P.[AG] | 0.93 | V..P[FY].E | 0.92 | T..PP.A | 0.90 |
| [ILMV][AG][FY].[ILMV]..[KR] | 0.96 | N[KR].P..I | 0.93 | G.[FY]P[KR] | 0.92 | [FY][ILMV].[ST].[AG] | 0.90 |
| [ILMV][AG][FY]...[ILMV][KR] | 0.96 | I...[AG]PF | 0.93 | [KR]F.GP | 0.92 | [ILMV].G..[KR]..[AG] | 0.90 |
| [QN].GP....N | 0.96 | [ST]PA.V[ILMV] | 0.93 | [DE]P[AG]..H | 0.92 | [KR]F[AG].....[ILMV] | 0.90 |
| GPP[FY] | 0.96 | [QN]PP...[FY] | 0.93 | [ILMV]..[AG]PH...[ILMV] | 0.92 | [ST]P[AG].V[ILMV] | 0.90 |
| [ILMV]...GPF[ST] | 0.96 | G.[QN]PP..[ILMV] | 0.93 | [DE]Q.P...[FY] | 0.92 | [FY]...P...A[ST] | 0.90 |
| M.......[QN]T | 0.96 | [KR]FG.....[ILMV] | 0.93 | [AG].[QN]PP.[ILMV] | 0.92 | [DE]....[AG]P.[ILMV] | 0.90 |
| PYG.S | 0.96 | G..SP...V | 0.93 | [KR].[QN]GP | 0.92 | SPN...[AG] | 0.90 |
| GP.R[ILMV].[AG] | 0.96 | V...[FY]P..V | 0.93 | [ILMV].[ILMV][KR]EP | 0.92 | SP[FY]P | 0.90 |
| N.K..F | 0.96 | [ILMV]....P[AG].V.[FY] | 0.93 | [ST]..[FY]P..Y | 0.92 | N..[AG]P[FY] | 0.90 |
| [ILMV].[QN].[AG]..T | 0.96 | [ILMV][KR]E.....I | 0.93 | N....P..[KR][FY] | 0.92 | [ST]SPY | 0.90 |
| T...NP.G | 0.96 | P[DE][KR].....[AG] | 0.93 | L....I[ILMV][KR] | 0.92 | K...RI | 0.90 |
| [ILMV].S.N...[AG] | 0.96 | [ILMV]..P[ILMV].[AG]N | 0.93 | [KR].....Y[AG][ILMV] | 0.92 | GPY..[ILMV] | 0.90 |
| C..[QN].P[ST] | 0.96 | SY.....N | 0.93 | [ILMV][KR]...P[FY]...[ILMV] | 0.92 | [AG]PP[FY] | 0.90 |
| [ILMV].[ILMV]...N..T | 0.96 | [ILMV][AG].P[ILMV].[ILMV][KR] | 0.93 | TK.....F | 0.92 | [AG][ST]GP..[ILMV] | 0.90 |
| P.[KR]....[AG].[ILMV] | 0.96 | P[ST]P[KR].[ILMV] | 0.93 | [ST]....P.P[ILMV].[ILMV] | 0.92 | [ST]....P.[FY]..D | 0.90 |
| [AG][FY].[FY][DE] | 0.96 | [ILMV].[AG].P[AG].[ILMV][ILMV] | 0.93 | V.P[ILMV].[ILMV][ST] | 0.92 | T.P[FY]..[ST] | 0.90 |
| [ILMV].[ST][ILMV].P..[ILMV].T | 0.96 | [AG]P[FY]D[AG] | 0.93 | I.[ST]P.F | 0.92 | [ILMV].P.P[ILMV].[DE] | 0.90 |
| N...[FY]P.[ST] | 0.96 | V.[ILMV]..P[ILMV]G | 0.93 | [ILMV][QN]G..[KR] | 0.92 | N...P[FY]..[DE] | 0.90 |
| QL.......Y | 0.96 | PNN..[AG] | 0.93 | TL...P[ILMV].[ILMV] | 0.92 | G.[FY]..[ILMV][ST] | 0.90 |
| V[ILMV]..[FY]P..[ILMV] | 0.96 | [ILMV]..[AG]P[FY]..[ILMV] | 0.93 | [QN]...[FY]P...E | 0.92 | GP.[ILMV][ILMV][AG] | 0.90 |
| [KR][AG]P[FY]...[ILMV] | 0.96 | PPF...[ST] | 0.93 | P[KR]NL | 0.92 | [ILMV].G.P[AG]..V | 0.90 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [ILMV]..[DE][AG]P[FY] | 0.96 | [ILMV]P.[ILMV].....I | 0.93 | [ST].P.NP | 0.92 | F.P[FY].[AG] | 0.90 |
| T...[AG]P..R.[ILMV] | 0.96 | G.P.[ILMV][ILMV]K | 0.93 | P[FY]D..S | 0.92 | [ST].[ST].....N | 0.90 |
| Y.[AG].[QN]P[ILMV] | 0.96 | [ST][ST][FY]...[ST] | 0.93 | [ILMV].[ILMV].[ST]P...T | 0.92 | [AG][ILMV]....D.K | 0.90 |
| [AG].[FY]GV | 0.96 | [ST].....[FY][ST][QN] | 0.93 | V.P[ILMV]P[ILMV] | 0.92 | G...[FY]P...[QN] | 0.90 |
| [AG].[FY]PK.[ILMV] | 0.96 | P[FY]PE | 0.93 | [KR]EP[FY] | 0.92 | [ILMV][ILMV].....[KR].F | 0.90 |
| R...Y.P | 0.96 | S.[ST].P....N | 0.93 | P[FY]P.[ST] | 0.92 | LP.PL[ILMV] | 0.90 |
| S....P..[ILMV][AG]D | 0.96 | V[ILMV][ILMV]..P....[FY] | 0.93 | [AG]P[FY]N | 0.92 | YP[KR][KR] | 0.90 |
| [ST]..[QN]PT..[ILMV] | 0.96 | [ST].S......N | 0.93 | P.[AG].[ILMV]..L | 0.92 | [ILMV]G...[ILMV].K | 0.90 |
| [ILMV]T..P..[ILMV].V | 0.96 | G..VGP | 0.93 | [KR]....P[FY]...S | 0.92 | [AG].[ILMV]E....[KR] | 0.90 |
| G[FY]P[ILMV]..[KR] | 0.96 | [KR]E...P.[ILMV][ILMV] | 0.93 | [FY]..[ILMV][AG].[FY] | 0.92 | [ILMV].[ILMV].[KR]P....[FY] | 0.90 |
| N...P[FY]P | 0.96 | [ILMV].[ST]P[QN]...[AG] | 0.93 | A...[AG].VV | 0.92 | GP[FY]L | 0.90 |
| PPF...S | 0.96 | [KR]..[AG][FY]P..[ILMV] | 0.93 | [FY].[AG][AG][QN] | 0.92 | [ST].GP[FY] | 0.90 |
| [ILMV].[ST].[AG]...Y | 0.96 | REP[FY] | 0.93 | I...P...H[ILMV] | 0.92 | [ST][FY]...[ST].[AG] | 0.90 |
| [ILMV].[ST].[QN]...G | 0.96 | [ST][ILMV]...P[ILMV].V | 0.93 | E.[AG]P[FY] | 0.92 | LPE[ILMV].P | 0.90 |
| N.[KR].PF | 0.96 | QSP..[ILMV] | 0.93 | K....P[FY].V | 0.92 | KEP[FY] | 0.90 |
| G..GP....S | 0.96 | [ILMV][ILMV]..P.[FY]..[FY] | 0.93 | S.[AG].V[ILMV] | 0.92 | TSP[FY] | 0.90 |
| A..EPF | 0.96 | V[AG][ILMV]...[FY] | 0.93 | S..N.T | 0.92 | MFP...[ILMV] | 0.90 |
| [AG]PY.G | 0.96 | [AG]PY..I | 0.93 | C[ST]....T | 0.92 | W.....[ST]..[QN] | 0.90 |
| V....P..V.Y | 0.96 | [QN]...P.PE | 0.93 | P.YP[ST] | 0.92 | G..[AG]...C | 0.90 |
| [ST]P.N..G | 0.96 | G[FY].[ILMV]..[KR] | 0.93 | [ILMV]...GPF | 0.92 | PPAT | 0.90 |
| NPTG | 0.95 | G[FY]..[ILMV].K | 0.93 | GPY...[ST] | 0.92 | S...QP....[ILMV] | 0.90 |
| Y.[AG].[QN].[ILMV] | 0.95 | [ST].A.V[ILMV] | 0.93 | P[FY]D..[ST] | 0.92 | S..H...[QN] | 0.90 |
| [FY]..PPF | 0.95 | [ILMV].[ST].N...[AG] | 0.93 | [ST]P[QN]...G | 0.92 | I..PGL | 0.90 |
| [AG].[FY]D.[ILMV] | 0.95 | [ILMV].[ST].[AG]...[FY] | 0.93 | [KR]...PPF | 0.92 | P.......[AG]I | 0.90 |
| [ST]PN..T | 0.95 | A....[ILMV][KR][FY] | 0.93 | G.Y.G | 0.92 | [AG]PYD | 0.90 |
| [ILMV]...[ST]..NP | 0.95 | PP[DE]..F | 0.93 | [ST]P.N.T | 0.91 | I..P.[FY]..[FY] | 0.90 |
| [ILMV][ILMV][ILMV]..P....Y | 0.95 | [KR][AG][ST].P..[ILMV] | 0.93 | [ST]..[QN]PT | 0.91 | G....P.P..[DE] | 0.90 |
| [AG]P..PL[QN] | 0.95 | Y.[AG][AG][QN] | 0.93 | [ILMV]...SP.N | 0.91 | [AG].[FY].[AG].G | 0.90 |
| L.L.EP[ILMV] | 0.95 | VP[ST]....[ST] | 0.93 | [ILMV][ST]....[ILMV]..V | 0.91 | [ILMV].[QN]..P..[ILMV]E | 0.90 |
| GP[FY].[KR] | 0.95 | GP[ILMV].[ILMV].A | 0.93 | I...[AG]P[FY] | 0.91 | P.[ILMV].[FY].[AG] | 0.90 |
| F..PPF | 0.95 | GP[ILMV][KR]..A | 0.93 | P.[FY]P[ST] | 0.91 | P.P...E[ST] | 0.90 |
| SPYP | 0.95 | GP[FY]..[QN] | 0.93 | [KR]EPF | 0.91 | [ILMV]...P.F.F | 0.90 |
| [ILMV].VGP.P | 0.95 | [ILMV]..R...[AG].L | 0.93 | GP[FY]...[ST] | 0.91 | [ILMV].P.YP.[ILMV] | 0.90 |
| LC....P | 0.95 | [KR]...PP....S | 0.93 | [ST]P.NP | 0.91 | [ILMV]..E.P.[ST][ILMV] | 0.90 |
| G..S...[ILMV]V | 0.95 | AVGP | 0.93 | GP.D[AG] | 0.91 | YP.[KR]..[QN] | 0.90 |
| [ILMV].GP.R[ILMV] | 0.95 | S...PTG | 0.93 | [KR]....P.[ILMV][ILMV]K | 0.91 | [ILMV]...A.[FY][FY] | 0.90 |
| [ILMV]...GP.T.[ILMV] | 0.95 | REPF | 0.93 | P.F...S | 0.91 | [KR].P.[KR]I | 0.90 |
| K.[ILMV]..P.[ILMV].K | 0.95 | P.......H[AG] | 0.93 | GP[FY]...S | 0.91 | [ST].[DE]...T[ILMV] | 0.90 |
| K.G...A[ILMV] | 0.95 | I.[ST]..P.P | 0.93 | [DE]GP[FY] | 0.91 | [ILMV][AG].LP.[ST] | 0.90 |
| T.P...EG | 0.95 | [AG]...GP.D | 0.93 | [ILMV]V.....[ILMV].[FY] | 0.91 | [ST].[FY]...A[AG] | 0.90 |
| [DE][KR].[FY]..[ST] | 0.95 | [DE]...SP.P | 0.93 | TV....H | 0.91 | [ST].A.[ILMV]V | 0.90 |
| [ILMV]P.[ILMV].P..Q | 0.95 | P..T.P..L | 0.93 | T...N..G | 0.91 | PE..P..A | 0.90 |
| [ILMV]...S..NP | 0.95 | RL.P.T | 0.93 | A..Y[AG]P | 0.91 | [ILMV]...EP..[ST].[ILMV] | 0.90 |
| [AG]I[AG].F | 0.95 | LE.D...[ILMV] | 0.93 | [ST].[ST].P....N | 0.91 | P.[ILMV]....[AG].[FY] | 0.90 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [ILMV]P[ST]....S | 0.95 | [ILMV]K..[AG].[ILMV][AG] | 0.93 | V.....[QN]..T | 0.91 | [ILMV]G..[ILMV].[ILMV][KR] | 0.90 |
| [ILMV].P.[FY]P[ST] | 0.95 | [FY].G.N.[ILMV] | 0.93 | [ILMV]P.[ILMV].P..[QN] | 0.91 | [ILMV]....PV..H | 0.90 |
| [AG]..[DE]L.D | 0.95 | [ST]..P.D.D | 0.93 | PP[FY]...S | 0.91 | [AG][KR].[FY].[AG] | 0.90 |
| [FY]SV..P | 0.95 | I.[FY]...[ILMV][KR] | 0.93 | [ST].N.P.G | 0.91 | PF.[KR].[ST] | 0.90 |
| L..[DE][AG]P.[ILMV] | 0.95 | E.P.L.Y | 0.93 | P..Y..K | 0.91 | [QN]..[ST]P..[ILMV][ST] | 0.90 |
| [ST]PN.P[ST] | 0.95 | [KR].[QN].PY | 0.93 | [ILMV][ILMV].....S[FY] | 0.91 | G..[ST]P..[ILMV][ILMV] | 0.90 |
| [ILMV].[ST]PN...[AG] | 0.95 | [ST].[AG].VV | 0.93 | [AG]..[FY][AG].[DE] | 0.91 | QQ..[FY] | 0.90 |
| E[KR].P.P..[ILMV] | 0.95 | T...GP..[KR] | 0.93 | FPS..[ILMV] | 0.91 | S..P.D[KR] | 0.90 |
| S..[QN]PT | 0.95 | GP.[FY].V | 0.93 | [AG][FY]P[ILMV]..[KR] | 0.91 | S.[ILMV]..P[ILMV]..G | 0.90 |
| G[AG].VG | 0.95 | S.YP..[AG] | 0.93 | [ILMV].[QN].[AG]..[ST] | 0.91 | G.Y.....[FY] | 0.90 |
| C[ST]...PT | 0.95 | S.[ST]....[DE][KR] | 0.93 | [KR][ILMV]S[AG] | 0.91 | EF.[AG]P | 0.90 |
| [ST]...[ILMV]..[AG]H | 0.95 | P.[DE]....[ILMV]L | 0.93 | SPN.P | 0.91 | [KR].N.P[FY] | 0.90 |
| [FY]..P[AG]H | 0.95 | G.QP..V | 0.93 | NN.T | 0.91 | [QN]..[ILMV]..P.[ILMV] | 0.90 |
| G[ST]...P[AG]..[AG] | 0.95 | [QN].PF.[FY] | 0.93 | P.LP.F | 0.91 | [ILMV][ILMV].T.....[FY] | 0.90 |
| [ST]..NPT | 0.95 | V....P[ILMV]..H | 0.93 | T...[AG]P..R | 0.91 | [AG]P.[FY]..P | 0.90 |
| PN.P.G | 0.95 | I[ST]..P.L[ILMV] | 0.93 | WP..P.[DE] | 0.91 | P..CN | 0.90 |
| GP.[KR][ILMV].[AG] | 0.94 | [ILMV]I.[ST]P.[FY] | 0.93 | [ILMV].[ST][ILMV].P....T | 0.91 | W.NG | 0.90 |
| K....P.[ILMV][ILMV]K | 0.94 | [ILMV]..A..PL | 0.93 | PAT.E | 0.91 | [ST].[AG]P..H | 0.90 |
| [QN]...[QN]..W | 0.94 | [AG]P[FY].[AG].[AG] | 0.93 | [ILMV].P.P[ILMV].V | 0.91 | G[ILMV].G.P...[ILMV] | 0.90 |
| GP[FY]..I | 0.94 | [AG]P[FY].[ILMV][AG] | 0.93 | E..P.P[KR].[ILMV] | 0.91 | [ILMV]...E..L[ST] | 0.90 |
| [ILMV]..GP[FY]..[ILMV] | 0.94 | [ST]..PF...D | 0.93 | W.V[DE] | 0.91 | [QN]S[QN]..P | 0.90 |
| [QN]..GP[FY] | 0.94 | K...P[FY].V | 0.93 | L...P[ILMV].[ILMV][ST] | 0.91 | [ILMV]G[FY]....K | 0.90 |
| G..GP[FY] | 0.94 | T[ILMV]..[AG]P...[AG] | 0.93 | E.....[ILMV].[DE][DE] | 0.91 | [FY]..K.PK | 0.90 |
| [ST].F.GP | 0.94 | [ILMV]PT.....[ILMV] | 0.93 | [AG]PF..V | 0.91 | T[ILMV]..[AG]P..[KR] | 0.90 |
| F[DE]GP | 0.94 | [AG]T..P.P | 0.93 | IK[AG]...[ILMV] | 0.91 | P[ILMV].[ILMV][STD] | 0.90 |
| K.PF[ST] | 0.94 | Y...N.V | 0.93 | [ILMV]P[ILMV].EP[ILMV] | 0.91 | FP...A[ST] | 0.90 |
| IK[AG].[AG] | 0.94 | [AG]...[AG]P.D[AG] | 0.93 | [QN]..[ILMV]...P[DE] | 0.91 | Y..I[QN] | 0.90 |
| [AG]..[AG]PY[ILMV] | 0.94 | [ST][ILMV].....IR | 0.93 | E.[KR][ILMV][FY] | 0.91 | [FY]P[ILMV].[ILMV][KR] | 0.90 |
| SP[QN]..T | 0.94 | V...P..V.[FY] | 0.93 | Y...NP[ILMV] | 0.91 | [DE]T[KR]....[AG] | 0.90 |
| [ILMV][AG]PV[KR] | 0.94 | [QN]PP..[ILMV][AG] | 0.93 | [ILMV]GP[ILMV]..[ST] | 0.91 | [ILMV]..[DE]P[AG]..[ST] | 0.90 |
| [ILMV]...GP[FY]..[ILMV] | 0.94 | Y..[ILMV]GP | 0.93 | [FY]GP[FY] | 0.91 | [ST][KR]L.....[ILMV] | 0.90 |
| SPN..[ST] | 0.94 | [ILMV][ILMV]..[FY]P..V | 0.93 | EPF...V | 0.91 | [KR]...[AG]PY | 0.90 |
| IE.P...[ST] | 0.94 | [ST]...P[ST][AG][AG] | 0.93 | TPS.[FY] | 0.91 | [ILMV][AG]PV.[ILMV] | 0.90 |
| [ILMV][ILMV]..P.[FY]..Y | 0.94 | A..P[AG].VV | 0.93 | TSP....G | 0.91 | [QN]..[AG]PF | 0.90 |
| [ILMV]..TP.F | 0.94 | G[ILMV][FY]P....[ILMV] | 0.93 | [ILMV].VGP....[ILMV] | 0.91 | [ILMV]KE.[FY] | 0.90 |
| PYD..[ST] | 0.94 | [ILMV]...[FY]..[KR]F | 0.93 | R[AG]....[FY][ILMV] | 0.91 | [FY]D[AG][AG] | 0.90 |
| [KR]..PYG | 0.94 | [ILMV]S..AP...[AG] | 0.93 | [ILMV][ILMV]VPG | 0.91 | [QN]...[FY]P..G | 0.90 |
| G.[FY]..[ILMV]T | 0.94 | Y..K[AG]P | 0.93 | GPY[DE] | 0.91 | T[FY]...P.[ST] | 0.90 |
| Y.Y..N | 0.94 | L.G.P[AG]..[ILMV] | 0.93 | [AG]..GPY | 0.91 | G...P..G.A | 0.90 |
| SP.N.T | 0.94 | N..P.PE | 0.93 | [ILMV][ILMV]..P..[KR | 0.91 | [AG]...[FY]..C | 0.90 |

| | | | | ].[FY] | | | |
|---|---|---|---|---|---|---|---|
| G.[FY].[KR].[ILMV] | 0.94 | [ILMV][ST][ST].P....V | 0.93 | PFK.[KR] | 0.91 | K.P..V[ILMV][ILMV] | 0.90 |
| T.P.N..G | 0.94 | [ILMV][ST].P[FY].[DE] | 0.93 | [FY].....PP | 0.91 | N.[ILMV][ILMV]....[KR] | 0.90 |
| [AG]....P.P.I | 0.94 | P[ILMV]V.....[ILMV] | 0.93 | R...P.[FY]..[DE] | 0.91 | N....FT | 0.90 |
| I.[ILMV]...P..[ILMV] | 0.94 | S.[ST].....N | 0.92 | GP.F[ST] | 0.91 | C....[ST]G | 0.90 |
| [AG]...[AG].[FY]D | 0.94 | S..[QN]P.G | 0.92 | K.N..[FY] | 0.91 | A...P[FY]P | 0.90 |
| W.[ILMV].[AG]P | 0.94 | [ILMV]S..[AG]P...[AG] | 0.92 | FE.P...[FY] | 0.91 | [ILMV][AG]P[ILMV]K | 0.90 |
| M..[DE].P[FY] | 0.94 | PV[ILMV][ILMV]K | 0.92 | P...[ST].[AG].[ILMV] | 0.91 | [AG]...P.PR | 0.90 |
| [ILMV].Y...[ILMV].[FY] | 0.94 | GP[FY]D | 0.92 | [ILMV].[ST].P..[ILMV].V | 0.91 | GP[FY].[AG] | 0.90 |

## Δομική ομαδοποίηση αμινοξέων

| Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ |
|---|---|---|---|---|---|---|---|
| [DLN].[DLN]...[KMR].[ITV][CS] | 1.00 | G.[FHWY].[DLN].G | 0.94 | GPF[DLN] | 0.92 | [DLN][ITV]...P.N.[ITV] | 0.91 |
| [CS]..NPTG | 1.00 | T.PY..S | 0.94 | [FHWY]GPY | 0.92 | [DLN]...F.[FHWY][DLN] | 0.91 |
| SP.NP.G | 0.99 | [DLN]..[DLN].P.[DLN][DLN][KMR] | 0.94 | G...R[ITV][DLN] | 0.92 | A...P.P[KMR] | 0.91 |
| [CS]...[FHWY].[FHWY].N | 0.99 | [ITV]V[ITV]..P.G | 0.94 | [FHWY]P[FHWY]..[ITV][ITV] | 0.92 | QL..P....[FHWY] | 0.91 |
| [EQ].P[FHWY][ITV].V | 0.99 | T.P.N..G | 0.94 | A..PA.[ITV][ITV] | 0.92 | [CS][ITV].....C | 0.91 |
| [FHWY]P.E[FHWY]I | 0.99 | [ITV][CS]..Y..[DLN] | 0.94 | [CS]...[FHWY].[DLN][EQ] | 0.92 | P[KMR]N..D | 0.91 |
| [ITV]..P.NP.G | 0.99 | [DLN]...[FHWY].[FHWY]N | 0.94 | PC.[ITV].P | 0.92 | G[FHWY]P.R | 0.91 |
| GPY.G | 0.99 | G..[FHWY].N[DLN] | 0.94 | [DLN]...[CS]P..P | 0.92 | K[FHWY]FP | 0.91 |
| [CS].PNNP | 0.99 | [KMR][EQ]...P.[ITV].K | 0.94 | P.NP.G | 0.92 | N[ITV]GP | 0.91 |
| V...EP...H | 0.99 | G.[EQ].PG | 0.94 | SP.NP | 0.92 | A[FHWY]PY | 0.91 |
| [ITV].SP.[DLN]P | 0.99 | [ITV].[ITV]TN | 0.94 | S..[FHWY]P...I | 0.92 | G....PL[ITV][DLN] | 0.91 |
| PY..[CS]G | 0.98 | [FHWY]P[KMR][KMR][KMR] | 0.94 | [FHWY]....[FHWY][FHWY]V | 0.92 | [ITV]E.P...S | 0.91 |
| P.NPTG | 0.98 | [DLN]...[FHWY]P[KMR][KMR] | 0.94 | VV.....[ITV].[FHWY] | 0.92 | WP.P.[ITV] | 0.91 |
| G...GPY | 0.98 | G[ITV]Y.....[DLN] | 0.94 | [DLN][ITV]GP.[DLN] | 0.92 | GPY.[ITV] | 0.91 |
| PNNP.G | 0.98 | [DLN]P[CS]G..[KMR] | 0.94 | [ITV].G..[FHWY].[ITV] | 0.92 | PFD..[FHWY] | 0.91 |
| [ITV][FHWY]E...F | 0.98 | E.GP[FHWY] | 0.94 | [CS].Y..R | 0.92 | TSP....G | 0.91 |
| PY..SG | 0.98 | [EQ]...FP[CS] | 0.94 | E..P..[KMR][DLN]L | 0.92 | EPF...V | 0.91 |
| S..P.D[KMR][DLN] | 0.98 | SP.N.T | 0.94 | W.[ITV]..[CS] | 0.92 | YP..[DLN]E | 0.91 |
| K.P[FHWY]T..[ITV] | 0.98 | [DLN]GPY | 0.94 | V.[ITV]..P.[FHWY]..[FHWY] | 0.92 | [KMR]F.GP | 0.91 |
| [KMR]..GP[FHWY]..[ITV] | 0.98 | [KMR]EPF...[ITV] | 0.94 | V.G..H | 0.92 | [ITV]V.G...[EQ] | 0.91 |
| [DLN]..GP[FHWY]..I | 0.98 | A..PA.V[ITV] | 0.94 | [ITV].[FHWY].[FHWY].N | 0.92 | L[ITV]....[FHWY]..[FHWY] | 0.91 |
| IG[FHWY]P...[KMR] | 0.98 | G.....[FHWY].C | 0.94 | [DLN].G.[FHWY]...[ITV] | 0.92 | P....FK | 0.91 |
| TLA[FHWY] | 0.98 | G.[FHWY].[EQ]P.[ITV] | 0.94 | [DLN]..[EQ]P...[FHWY][ITV] | 0.92 | N....FP | 0.91 |
| [FHWY]..YPY | 0.98 | Q[DLN]..P....Y | 0.94 | [KMR]..P.I.L | 0.92 | N...[EQ]..F | 0.91 |
| [ITV]....PAT.E | 0.98 | [KMR]G.[FHWY]P..[DLN] | 0.94 | [ITV][DLN]..P...[FHWY]L | 0.92 | [CS]T....P | 0.91 |
| S.[DLN]NP.G | 0.98 | V....P...[FHWY]H | 0.94 | LT..P....V | 0.92 | L[DLN]..P..[KMR][ITV] | 0.91 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| R...P[FHWY]...D | 0.98 | S.S.....N | 0.94 | K..[FHWY][DLN][ITV] | 0.92 | [FHWY]..P...DA | 0.91 |
| EPF.[EQ] | 0.98 | [DLN].[CS].YP | 0.94 | WP..P.D | 0.92 | S..F...[CS] | 0.91 |
| P[CS][DLN]P.G | 0.98 | GPY..[ITV] | 0.94 | [ITV][ITV].....IE | 0.92 | [FHWY]PN..F | 0.91 |
| C.P...[ITV]G | 0.97 | GP[FHWY]..[ITV][ITV] | 0.94 | [DLN]TK..P | 0.92 | [EQ][ITV]....P.[ITV] | 0.91 |
| N...P.PE | 0.97 | [KMR].[FHWY][DLN]A | 0.94 | P.P[KMR][DLN]L | 0.92 | Y..Y.[FHWY] | 0.91 |
| [KMR]...P[FHWY][DLN]..[FHWY] | 0.97 | T[DLN]....[DLN]G | 0.94 | [DLN].[DLN]..P[KMR][KMR] | 0.92 | [FHWY]...P..F.[FHWY] | 0.91 |
| [EQ]...FP.N | 0.97 | [FHWY].GP....P | 0.94 | [KMR].[EQ]P[EQ][DLN] | 0.92 | [CS].FP.[ITV] | 0.91 |
| Y[DLN][EQ]G | 0.97 | E..P.P.[DLN]L | 0.94 | S[FHWY]P..T | 0.92 | [FHWY][KMR].N..[FHWY] | 0.91 |
| [DLN]....P[FHWY]N.[ITV] | 0.97 | [ITV]..SYP | 0.94 | [DLN]R.EP | 0.92 | [FHWY].[ITV]..P[FHWY]...[DLN] | 0.91 |
| [FHWY][DLN]..GP[FHWY] | 0.97 | [KMR]EP[FHWY]...[ITV] | 0.94 | [ITV]..P[FHWY]D.[DLN] | 0.92 | ST.....N | 0.91 |
| GPY...G | 0.97 | [DLN]..GPY | 0.94 | PF[DLN]..F | 0.92 | L[KMR].P.[ITV].[DLN] | 0.91 |
| NNPT | 0.97 | GP[FHWY]..I | 0.94 | [KMR]...PP....S | 0.92 | [KMR].[ITV]...[FHWY].V | 0.91 |
| T.P.[DLN]P.G | 0.97 | [DLN].S.YP | 0.94 | [ITV].GP[FHWY][DLN] | 0.92 | [CS]..N..G | 0.91 |
| V.....N..T | 0.97 | [DLN].T...[DLN]P | 0.93 | EP[FHWY][ITV][FHWY] | 0.92 | [KMR]EPF | 0.90 |
| P..YP.K | 0.97 | [ITV].[DLN]..P..V.[FHWY] | 0.93 | [KMR]..PKI | 0.92 | GP[FHWY]L | 0.90 |
| Y[FHWY]...P....N | 0.97 | [EQ]PA...Y | 0.93 | VYP...[EQ] | 0.92 | RE.....I | 0.90 |
| IG.[FHWY].[DLN] | 0.97 | [DLN]F.....[DLN]P | 0.93 | GPY...S | 0.92 | S..WP | 0.90 |
| SY.[FHWY]..[ITV] | 0.97 | [DLN].G.[FHWY].P | 0.93 | A..YGP | 0.92 | [CS]..N.T | 0.90 |
| [ITV].[FHWY][EQ].P....[KMR] | 0.97 | N.[CS].P.G | 0.93 | [KMR]..[DLN]PPF | 0.92 | [CS]YP[FHWY] | 0.90 |
| [DLN]....GP.[ITV] | 0.97 | [ITV].[ITV]PY...[DLN] | 0.93 | [ITV]...EPF | 0.92 | MTL | 0.90 |
| G.QP.[DLN][ITV] | 0.97 | L.[KMR]...[ITV][ITV][DLN] | 0.93 | LK.P...I | 0.92 | NN..G | 0.90 |
| SP.N..G | 0.97 | N..P.T.[FHWY] | 0.93 | AW....[FHWY] | 0.92 | GP[FHWY][DLN] | 0.90 |
| [CS]P.N..G | 0.97 | S.N.P.G | 0.93 | V.....V.Y | 0.92 | [ITV].[ITV]...[KMR]P | 0.90 |
| [DLN]..GP[FHWY]..[ITV] | 0.97 | A....[KMR].[ITV][CS] | 0.93 | T...GP..[KMR] | 0.92 | L.[KMR]....I[DLN] | 0.90 |
| IG[FHWY]....[KMR] | 0.97 | GP..[KMR][ITV][DLN] | 0.93 | V.....QP | 0.92 | G.FP...[ITV] | 0.90 |
| P....P..L[FHWY] | 0.96 | [KMR]...P[FHWY].[ITV].[FHWY] | 0.93 | S[ITV]....R[DLN] | 0.92 | S...[FHWY].[FHWY].[DLN] | 0.90 |
| S..NP.G | 0.96 | [DLN][ITV][CS].....G | 0.93 | [DLN]...YP.S | 0.92 | [ITV].E.[FHWY]...[ITV] | 0.90 |
| [CS]..NP.G | 0.96 | D....P.[KMR]Y | 0.93 | [ITV]...P...HL | 0.92 | V...P..V.[FHWY] | 0.90 |
| N.K..F | 0.96 | [ITV][FHWY]..[FHWY].F | 0.93 | PP.[KMR].K | 0.92 | G..GP[FHWY] | 0.90 |
| [CS]..[DLN][DLN]P.G | 0.96 | [EQ].P[FHWY][DLN].[FHWY] | 0.93 | C..[FHWY]..N | 0.92 | [CS].GP[FHWY] | 0.90 |
| [FHWY]..[FHWY]P[FHWY]..[ITV] | 0.96 | [DLN]...[EQ]P[EQ][DLN] | 0.93 | [DLN][KMR].P[DLN]...[EQ] | 0.92 | [FHWY]..Y.Y | 0.90 |
| C.P.[DLN]..G | 0.96 | S[FHWY]P....N | 0.93 | [DLN][DLN].GF | 0.92 | [FHWY][DLN]....G..[KMR] | 0.90 |
| [CS]...[FHWY].[DLN]Q | 0.96 | P[FHWY]G[ITV].[FHWY] | 0.93 | L.[EQ]P[DLN]..[ITV] | 0.92 | [ITV]GPY | 0.90 |
| G..[ITV][KMR][ITV][DLN] | 0.96 | PC[KMR]..[ITV] | 0.93 | T[FHWY]...PF | 0.92 | [FHWY]......R.I | 0.90 |
| [KMR][EQ]...P.I[ITV] | 0.96 | [DLN][ITV].GP[FHWY] | 0.93 | [DLN][FHWY]....P..[ITV] | 0.92 | [KMR]..G....[ITV]V | 0.90 |
| [EQ]...F.[FHWY][KMR] | 0.96 | G.PN[FHWY] | 0.93 | [DLN]...[CS]..[DLN]P | 0.92 | T..PP.A | 0.90 |
| H.....[DLN][ITV][FHWY] | 0.96 | MP...P.[ITV] | 0.93 | [ITV][FHWY].P.[CS]..[DLN] | 0.92 | K...RI | 0.90 |
| [ITV][DLN][KMR].P.T | 0.96 | [ITV]..SP[FHWY][DLN] | 0.93 | S..P..[KMR]D | 0.92 | [KMR][EQ].....I[ITV] | 0.90 |
| QL.......Y | 0.96 | [FHWY]...PY[CS] | 0.93 | [FHWY]P..P...I | 0.92 | T..[DLN].P...D | 0.90 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [FHWY][EQ]..P[FHWY][KMR] | 0.96 | T..AP..[CS] | 0.93 | P....P..[DLN]F | 0.92 | [KMR].[ITV]...Y..[KMR] | 0.90 |
| G[FHWY].[EQ].P....[FHWY] | 0.96 | [ITV].EP[FHWY]...[ITV] | 0.93 | [ITV][EQ].P..[ITV].[FHWY] | 0.92 | [ITV].[ITV]...N[DLN] | 0.90 |
| [DLN].[FHWY]P[KMR][KMR] | 0.96 | G..SP...V | 0.93 | N....P..R[FHWY] | 0.92 | P.T[ITV].Q | 0.90 |
| [DLN]..G[FHWY]P[KMR] | 0.96 | [ITV][ITV]VPG | 0.93 | [ITV].GP.[KMR][DLN] | 0.92 | [EQ].Y...N | 0.90 |
| T...NP.G | 0.96 | [KMR]..PYG | 0.93 | TK.....F | 0.92 | [FHWY]Y..YP | 0.90 |
| [ITV].PG[FHWY]..[EQ] | 0.96 | [DLN]G.Y...[ITV] | 0.93 | A.PY...[DLN] | 0.92 | [ITV][ITV].[KMR].PA | 0.90 |
| [DLN]...GP[ITV].[DLN] | 0.96 | V.....V..H | 0.93 | [FHWY]..Y..[ITV][DLN] | 0.92 | [EQ].[EQ]..V[DLN] | 0.90 |
| GP[FHWY][DLN]..S | 0.96 | I.YP....[ITV] | 0.93 | N.GP....[DLN] | 0.92 | [EQ].W..[FHWY] | 0.90 |
| PP[FHWY][KMR].[KMR] | 0.96 | [KMR]F.G....[DLN] | 0.93 | [FHWY]..YP..[KMR] | 0.92 | E....P[FHWY].A | 0.90 |
| PPAT[ITV] | 0.96 | L...EP...T | 0.93 | S..N.T | 0.92 | V..P[FHWY].E | 0.90 |
| PYG.S | 0.96 | S[ITV]....S..[DLN] | 0.93 | [ITV][FHWY].N..[FHWY] | 0.92 | [CS].P..PT | 0.90 |
| [ITV][EQ]SP....[FHWY] | 0.96 | [CS]...[FHWY]P.Q | 0.93 | [ITV]G[FHWY]....K | 0.92 | V.[ITV]..P....Y | 0.90 |
| [ITV].PV.V.[ITV] | 0.96 | YP...[FHWY][CS] | 0.93 | PP[DLN]..F | 0.92 | L...[FHWY].[ITV].[FHWY] | 0.90 |
| S..[FHWY]P....N | 0.96 | [DLN]...C...P | 0.93 | GPF..[ITV] | 0.92 | [KMR].[EQ].[DLN].S | 0.90 |
| [KMR]....PF...[CS] | 0.96 | N[ITV]..YP | 0.93 | G.[FHWY].[EQ]..[ITV] | 0.92 | [DLN].G.[FHWY]..[ITV] | 0.90 |
| R...Y.P | 0.96 | S..W.....[DLN] | 0.93 | EGP[FHWY] | 0.92 | [EQ].....[DLN][DLN]R | 0.90 |
| [ITV].[ITV].SP.[DLN] | 0.96 | [ITV].SP.[DLN].[ITV] | 0.93 | REP[FHWY] | 0.92 | [FHWY].G.[FHWY].[ITV] | 0.90 |
| G.[DLN][FHWY].T | 0.96 | [DLN]A...PL.[KMR] | 0.93 | [CS]...[FHWY].[FHWY].[DLN] | 0.92 | T[DLN].P..[DLN][DLN] | 0.90 |
| [EQ]..[FHWY]..P.[ITV] | 0.96 | V..T..F | 0.93 | G.Y.G | 0.92 | [KMR][EQ].....I.[DLN] | 0.90 |
| [ITV][FHWY]PN..[FHWY] | 0.96 | [ITV].[KMR].PA.[ITV] | 0.93 | SP.N.[ITV] | 0.92 | [DLN]....P.N.V | 0.90 |
| D[ITV]..[KMR]...[ITV] | 0.96 | SY.....N | 0.93 | [DLN].G[FHWY]...[ITV] | 0.92 | [ITV]...GP..I | 0.90 |
| PPF...S | 0.96 | [KMR]P...P.T | 0.93 | GPY[DLN] | 0.91 | C..[DLN].P[ITV] | 0.90 |
| G..GP....S | 0.96 | [DLN].I....[DLN].T | 0.93 | G...GP[FHWY] | 0.91 | [ITV]...[DLN].PL | 0.90 |
| A..EPF | 0.96 | [ITV]..[CS]YP | 0.93 | GP[FHWY]..[ITV] | 0.91 | [ITV]....PV..H | 0.90 |
| F.GP...[FHWY] | 0.96 | T[DLN]A[FHWY] | 0.93 | PP[FHWY]...S | 0.91 | [KMR]....P.PV | 0.90 |
| [CS].PNN | 0.96 | F...P.T..[EQ] | 0.93 | [KMR]...PPF | 0.91 | [DLN]....P.NP | 0.90 |
| N.[KMR].PF | 0.96 | G..VGP | 0.93 | P....P..[DLN][FHWY] | 0.91 | W.NG | 0.90 |
| Q...[FHWY]P....[FHWY] | 0.96 | [ITV]....PA[ITV].[EQ] | 0.93 | [CS]P.NP | 0.91 | PE..P..A | 0.90 |
| [ITV]..Y....F | 0.96 | [ITV].GP[FHWY]..[ITV] | 0.93 | [ITV]G[FHWY]....[KMR] | 0.91 | I[ITV][ITV].G | 0.90 |
| V....P..V.Y | 0.96 | P[DLN].....I[DLN] | 0.93 | TV....H | 0.91 | V.P[CS][FHWY] | 0.90 |
| GP[FHWY]...G | 0.96 | [FHWY]..PPF | 0.93 | GP[FHWY]D | 0.91 | P[KMR]NL | 0.90 |
| GP[FHWY][ITV].[ITV] | 0.96 | F[DLN]GP | 0.93 | L..GP[FHWY] | 0.91 | G[DLN]P.....[ITV] | 0.90 |
| NPTG | 0.95 | [FHWY][DLN][EQ]G | 0.93 | T...N..G | 0.91 | F....P[ITV]P | 0.90 |
| [ITV].GPY | 0.95 | [FHWY].[FHWY]...Y[DLN] | 0.93 | [ITV]V[ITV].G | 0.91 | [ITV]..P[FHWY]P.[KMR] | 0.90 |
| [CS]..[FHWY]P....N | 0.95 | M[KMR].P.[ITV] | 0.93 | L[KMR].P...I | 0.91 | DT...P[FHWY] | 0.90 |
| A[FHWY].Y...[DLN] | 0.95 | [ITV]L...P.[FHWY]..[FHWY] | 0.93 | [FHWY]....P[FHWY]..I | 0.91 | [DLN]E.D..[DLN] | 0.90 |
| [ITV].SP.N | 0.95 | N[KMR].P..I | 0.93 | P.F...S | 0.91 | L...GP[FHWY] | 0.90 |
| [ITV].T...[KMR]P | 0.95 | K[EQ].....I[ITV] | 0.93 | [DLN][ITV].....G[FHWY] | 0.91 | I...GP[FHWY] | 0.90 |
| T.P...EG | 0.95 | [EQ][ITV].[FHWY]..P | 0.93 | [KMR].P[FHWY][DLN][ITV] | 0.91 | [ITV][FHWY]..PP[FHWY] | 0.90 |
| S.[DLN]..[FHWY].G | 0.95 | [ITV].I..P.F | 0.93 | Q[DLN].....Y | 0.91 | PY[DLN][DLN]P | 0.90 |
| T[DLN]...P[ITV].V | 0.95 | P..T.P..L | 0.93 | PPF.[FHWY] | 0.91 | YP[ITV][ITV][DLN] | 0.90 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [ITV].E.F...[ITV] | 0.95 | [ITV]...[FHWY].[DLN].V | 0.93 | T[DLN]....[ITV].V | 0.91 | SYP...[ITV] | 0.90 |
| [DLN]T..[FHWY]...G | 0.95 | Y.[FHWY].[DLN].[ITV] | 0.93 | [FHWY].AYP | 0.91 | [FHWY]..GPS | 0.90 |
| I.[FHWY].C | 0.95 | [DLN]...R.[ITV].L | 0.93 | L[FHWY]...P...T | 0.91 | GPN[FHWY] | 0.90 |
| [KMR]....[ITV]RL | 0.95 | [CS]S..[FHWY]..[ITV] | 0.93 | [KMR][EQ]..L..[ITV] | 0.91 | YKP[DLN] | 0.90 |
| [KMR]K...[ITV].L | 0.95 | [ITV][DLN]...P.VV | 0.93 | [ITV][FHWY]....[FHWY]Y | 0.91 | I..PGL | 0.90 |
| [DLN][CS]..G..[KMR] | 0.95 | E.P.L.Y | 0.93 | PAT.E | 0.91 | YPS...[ITV] | 0.90 |
| G[EQ].[EQ]...[DLN] | 0.95 | S...[FHWY].[DLN][EQ] | 0.93 | P..Y..K | 0.91 | PPD..[FHWY] | 0.90 |
| LC....P | 0.95 | [KMR]..P.[ITV].[DLN]P | 0.93 | [ITV]...GPY | 0.91 | [ITV]..PNN | 0.90 |
| [DLN].G[FHWY]P..[ITV] | 0.95 | G.QP..V | 0.93 | T[DLN]...P[ITV].[ITV] | 0.91 | TNP.[DLN] | 0.90 |
| [DLN]V[EQ]S | 0.95 | [EQ]..Y.[ITV][KMR] | 0.93 | SPN.P | 0.91 | [ITV]..TNP | 0.90 |
| N...P[FHWY]P | 0.95 | E.G.[FHWY]P | 0.93 | [FHWY]..YP.[KMR] | 0.91 | [DLN].P.[DLN]P.F | 0.90 |
| [KMR]...PP[FHWY].[ITV] | 0.95 | [KMR]AP..[ITV].[ITV] | 0.93 | SYP..[FHWY] | 0.91 | [DLN].[DLN]D.P..[ITV] | 0.90 |
| [DLN]PP[FHWY]...S | 0.95 | N..P.PE | 0.93 | P.LP.F | 0.91 | [DLN].[KMR]W.P | 0.90 |
| [DLN].[DLN]..PY[DLN] | 0.95 | P[CS].[ITV]....[DLN] | 0.93 | NN.T | 0.91 | P.PF..[ITV] | 0.90 |
| QSP..[ITV] | 0.95 | P[FHWY].[ITV][KMR][FHWY] | 0.93 | [ITV][DLN][KMR].P.[ITV] | 0.91 | I..PF..[ITV] | 0.90 |
| F..PPF | 0.95 | [DLN][KMR]..[EQ].Y | 0.93 | [DLN].....[KMR].[ITV][CS] | 0.91 | N....FT | 0.90 |
| SPYP | 0.95 | [ITV][EQ]...P[ITV][KMR] | 0.93 | [ITV]...P.[ITV][FHWY][ITV] | 0.91 | S..[FHWY]P....D | 0.90 |
| [EQ]SP..I | 0.95 | [DLN].GP.N | 0.93 | I...P...[FHWY]L | 0.91 | [ITV].....[KMR]P[ITV] | 0.90 |
| [FHWY][DLN]EG | 0.95 | SP.[DLN]P[ITV] | 0.93 | E...P.P.[ITV] | 0.91 | [DLN]...HK | 0.90 |
| S[FHWY]..G[FHWY] | 0.95 | N[FHWY]P[DLN]..[DLN] | 0.93 | [DLN][ITV]P.....[FHWY] | 0.91 | [ITV]...SP.[DLN].[ITV] | 0.90 |
| [KMR]E...P.[ITV][ITV] | 0.95 | [ITV].SP.P..[FHWY] | 0.93 | [ITV].G...[DLN].S | 0.91 | [KMR].P[FHWY]..[ITV][ITV] | 0.90 |
| [CS]...PTG | 0.95 | NFP...[DLN] | 0.93 | [ITV]N..P...E | 0.91 | P..[EQ]P....[FHWY] | 0.90 |
| [ITV][FHWY]E...[FHWY] | 0.95 | S...PTG | 0.93 | V[ITV].....N[DLN] | 0.91 | [DLN][DLN]..P[EQ]..[FHWY] | 0.90 |
| PN.P.G | 0.95 | AVGP | 0.93 | [ITV].G..A.[ITV] | 0.91 | P.P[FHWY]...D | 0.90 |
| [KMR][EQ]...P.[ITV][ITV] | 0.94 | [KMR].PF...V | 0.93 | V..P[FHWY].I | 0.91 | [DLN]P....[ITV]..I | 0.90 |
| [FHWY]...P.T..Q | 0.94 | REPF | 0.93 | [KMR]....P[FHWY].V[KMR] | 0.91 | [EQ]P[FHWY].E | 0.90 |
| [CS].S.....N | 0.94 | [EQ][ITV]LH | 0.93 | S[ITV].....[CS].[DLN] | 0.91 | [KMR].[ITV][ITV]....D | 0.90 |
| T.P...[EQ]G | 0.94 | SV....[KMR][DLN] | 0.93 | [ITV].S.[FHWY]P..[FHWY] | 0.91 | [EQ].Q[DLN]..[DLN] | 0.90 |
| [ITV].[ITV][ITV].PV | 0.94 | [ITV].TP.F | 0.93 | [DLN][FHWY]P.[KMR][KMR] | 0.91 | G...P..G.A | 0.90 |
| PYD.[DLN] | 0.94 | RL.P.T | 0.93 | [ITV].[FHWY].P....M | 0.91 | Y..N[CS] | 0.90 |
| GP.F.[ITV] | 0.94 | Y...N.V | 0.93 | A..P.[ITV].F | 0.91 | [ITV]S..P...[FHWY][DLN] | 0.90 |
| S..P.[DLN][KMR][DLN] | 0.94 | [DLN].T[CS]..[DLN] | 0.93 | [FHWY]P[FHWY].N | 0.91 | Y...GP.[DLN] | 0.90 |
| [ITV][FHWY].[EQ]...Y | 0.94 | S..[FHWY]..[ITV][CS] | 0.93 | [DLN]H.....[DLN].[DLN] | 0.91 | [ITV][KMR]....[KMR]P | 0.90 |
| G.[FHWY]P...T | 0.94 | N[FHWY]...[FHWY][DLN] | 0.93 | P.K.P....[ITV] | 0.91 | [ITV]...GP.[FHWY].[ITV] | 0.90 |
| [ITV].S..P...[FHWY][DLN] | 0.94 | [ITV]..GP.[FHWY].[ITV] | 0.93 | [DLN]...P.N.V | 0.91 | P..CN | 0.90 |
| A....P[FHWY].R | 0.94 | V...EP...[FHWY] | 0.92 | [EQ]..[FHWY].[FHWY][FHWY] | 0.91 | [ITV].[KMR].....I[ITV] | 0.90 |
| A.[CS]..[DLN]P | 0.94 | P.[KMR]......I | 0.92 | N.[CS].[FHWY]P | 0.91 | G..P.[ITV][DLN].L | 0.90 |
| [ITV]L...P[ITV].V | 0.94 | [CS].P[FHWY]..N | 0.92 | [ITV][FHWY]..S.A | 0.91 | [KMR]...P.[ITV][DLN].L | 0.90 |
| Y.Y..N | 0.94 | [ITV]...GP..[DLN][DLN] | 0.92 | [KMR]..[EQ].PR | 0.91 | G.[FHWY].....[FHWY][DLN] | 0.90 |
| V.[EQ].[FHWY]...[ITV] | 0.94 | GP[FHWY]I | 0.92 | [ITV]..P[FHWY].[FH | 0.91 | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | WY].[ITV] | | | |

## Υδροφοβικότητα

| Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ |
|---|---|---|---|---|---|---|---|
| 100010.0.00 | 0.88 | 0.1101 | 0.83 | 1001001 | 0.82 | 0.0010.1000 | 0.80 |
| 11010 | 0.86 | 1101011000 | 0.83 | 11110000 | 0.81 | 000110.00.0 | 0.80 |
| 11000111 | 0.86 | 101000000.1 | 0.83 | 0001.0.0.00 | 0.81 | 0.00011 | 0.80 |
| 100010 | 0.85 | 1100011 | 0.83 | 10.000.0011 | 0.81 | 1.01011 | 0.80 |
| 10001000..0 | 0.85 | 1101 | 0.82 | 1110101 | 0.81 | 0.101011 | 0.80 |
| 11101011000 | 0.85 | 101000.0011 | 0.82 | 10.00011 | 0.81 | 1101000 | 0.80 |
| 1000 | 0.84 | 110100 | 0.82 | 11000.11 | 0.81 | | |
| 1101011000 | 0.84 | 10001000 | 0.82 | 0.101 | 0.80 | | |
| 11010000 | 0.83 | 001000.1.00 | 0.82 | 11000.1 | 0.80 | | |

## Μικρό μέγεθος

| Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ |
|---|---|---|---|
| 01011100101 | 0.84 | 10111110100 | 0.81 |
| 11111101001 | 0.83 | 01101101.01 | 0.80 |
| 10111100011 | 0.83 | 0101110.101 | 0.80 |
| 1011100101 | 0.82 | 01101101101 | 0.80 |
| 01101101001 | 0.82 | | |

## Αλειφατικός χαρακτήρας

| Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ |
|---|---|---|---|---|---|---|---|
| 111.000001 | 0.90 | 01..0010101 | 0.85 | 1.1100100.0 | 0.84 | 1..0011100 | 0.81 |
| 11100000010 | 0.89 | 111.00100.0 | 0.85 | 1.10001.1.0 | 0.83 | 1.1.0010.10 | 0.81 |
| 111000.0010 | 0.89 | 0100.001110 | 0.85 | 111.001..00 | 0.83 | 10..0010110 | 0.81 |
| 111.00.0010 | 0.89 | 1..0010101 | 0.85 | 1.11.01.0.0 | 0.83 | 1..001.101 | 0.80 |
| 101000101.0 | 0.88 | 010.001.101 | 0.84 | 01..001.101 | 0.83 | 1.00010101 | 0.80 |
| 1110001.1 | 0.87 | 1.110010000 | 0.84 | 111.001...0 | 0.83 | 101100100.0 | 0.80 |
| 010.0010101 | 0.86 | 1.11.01.000 | 0.84 | 0100001.101 | 0.82 | | |

## Θετικό φορτίο

| Ακολουθιακά πρότυπα | Σκορ |
|---|---|
| 110.10010 | 0.87 |
| 0.00011100 | 0.81 |
| 0000011100 | 0.80 |

## Πολικότητα

| Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ |
|---|---|---|---|
| 00011001000 | 0.87 | 11010000011 | 0.82 |
| 00.11001000 | 0.86 | 110010.0111 | 0.81 |
| 00111001000 | 0.85 | 000110010.0 | 0.81 |
| 11000011000 | 0.84 | 001.1000001 | 0.81 |
| 00000000011 | 0.84 | 11001010000 | 0.81 |
| 00101000001 | 0.83 | 00.110010.0 | 0.81 |
| 0001100.000 | 0.83 | 001110010.0 | 0.80 |
| 11001000111 | 0.82 | 11001010111 | 0.80 |

## Προλίνη

| Ακολουθιακά πρότυπα | Σκορ |
|---|---|
| 000001110.0 | 0.81 |
| 0000.1101 | 0.80 |
| 000001101 | 0.80 |

## Αρωματικός χαρακτήρας

| Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ |
|---|---|---|---|---|---|---|---|
| 11010..100 | 0.93 | 0000101.100 | 0.86 | 100101.000 | 0.84 | 1100.010000 | 0.82 |
| 1101...10 | 0.92 | 010010.0010 | 0.86 | 110.0010000 | 0.84 | 1.01001.000 | 0.82 |
| 1101..010 | 0.91 | 00.01010001 | 0.86 | 0100.010.10 | 0.84 | 011010.0..0 | 0.82 |
| 1101.0.10 | 0.90 | 00001010100 | 0.86 | 0.01010001 | 0.84 | 0001.0.10.1 | 0.82 |
| 01.010.0010 | 0.89 | 0100101.0.0 | 0.86 | 0..01010100 | 0.84 | 010.0010.10 | 0.82 |
| 010010100.0 | 0.88 | 0.10100.010 | 0.86 | 010..010.10 | 0.84 | 01000010010 | 0.82 |
| 10010100.0 | 0.88 | 0.000010110 | 0.86 | 0..0101.100 | 0.84 | 01000010.10 | 0.81 |
| 0.101000010 | 0.88 | 1.01010.00 | 0.86 | 110.001.000 | 0.84 | 000100.10.1 | 0.81 |
| 0.0.0010110 | 0.88 | 01.01010.00 | 0.85 | 1.0101..00 | 0.83 | 01000110.0 | 0.81 |
| 01.010.001 | 0.87 | 100101.0.0 | 0.85 | 110..01.000 | 0.83 | 11000010000 | 0.81 |
| 1.01010..0 | 0.87 | 010010.001 | 0.85 | 01.0101..00 | 0.83 | 1.010010000 | 0.81 |
| 01.01010..0 | 0.87 | 00001010001 | 0.85 | 110100..000 | 0.83 | 000.0001110 | 0.81 |
| 01.01000010 | 0.87 | 01.0101...0 | 0.85 | 0100100001 | 0.83 | 0101.0.010 | 0.80 |
| 00000010110 | 0.87 | 0100.010010 | 0.85 | 01001000010 | 0.83 | 0101.0.01.0 | 0.80 |
| 010100101 | 0.87 | 0000010110 | 0.85 | 1101.0..0.0 | 0.83 | 01101000..0 | 0.80 |
| 1001010000 | 0.86 | 1.0101...0 | 0.84 | 0001010001 | 0.83 | 001000110.0 | 0.80 |
| 01.0100001 | 0.86 | 110..010000 | 0.84 | 0010100.010 | 0.82 | 0001100.001 | 0.80 |
| 01001010000 | 0.86 | 0100101.000 | 0.84 | 0110.0101 | 0.82 | | |

## Αρνητικό φορτίο

| Ακολουθιακά πρότυπα | Σκορ |
|---|---|

| | |
|---|---|
| 01.0.000011 | 0.8 |

# Παράρτημα ΙΙ

*Ακολουθιακά πρότυπα cis-nonPro πεπτιδικών δεσμών*

**Ακριβής εξαγωγή προτύπων**

| Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ |
|---|---|---|---|---|---|---|---|
| KPGKGRRK | 1.00 | Y..GH....S | 1.00 | PTF....D | 1.00 | GDH.V | 0.99 |
| EDGTKEPLL | 1.00 | G...W..NDS | 1.00 | G...WS..D | 1.00 | LI...PQ | 0.99 |
| HAESGEYGL | 1.00 | GR.....LAI | 1.00 | PIYQ | 1.00 | K.I.S..Q | 0.99 |
| LGTVINQL | 1.00 | GG.FV.G | 1.00 | G....SL...T | 1.00 | IL....PL | 0.99 |
| ADEAT | 1.00 | S.S.H..N | 1.00 | N.TKV | 1.00 | H....LLL | 0.99 |
| ALNALKLVT | 1.00 | PT.R.E..A | 1.00 | GSG.A...L | 1.00 | S....L.ID | 0.99 |
| YFT...I | 1.00 | S..F...EP | 1.00 | DC.....DA | 1.00 | L.T.A..Y | 0.99 |
| CLA..VN | 1.00 | I.TG..S.A | 1.00 | I.TN..V | 1.00 | G.G...VW | 0.99 |
| R..DP....VV | 1.00 | PDPA....V | 1.00 | GA...RY | 1.00 | IL..V....G | 0.99 |
| H.YSQ | 1.00 | EPG....H | 1.00 | WRQA | 1.00 | IV..S...K | 0.99 |
| VYL..L...Y | 1.00 | LGG...LA | 1.00 | Q.E.SD | 1.00 | V.....DSV | 0.99 |
| NAW..D | 1.00 | IV..ES.G | 1.00 | YT..N.V | 1.00 | SGV...P | 0.99 |
| A...KHF.G.G | 1.00 | G.GG..RL | 1.00 | L...E....AP | 1.00 | Y.VSG | 0.99 |
| L..SRGF | 1.00 | C.P..DD | 1.00 | SV...G...V | 1.00 | IID.....G | 0.99 |
| REPDP | 1.00 | H..S.E....V | 1.00 | GL...N...S | 1.00 | E....E.E.R | 0.99 |
| G.MFW | 1.00 | S.GC....F | 1.00 | PT....G..V | 1.00 | PG.N.V | 0.99 |
| L.G..VVP..S | 1.00 | P....G.HS | 1.00 | DD.A.G | 1.00 | I..TA..G | 0.99 |
| MDHSNY | 1.00 | IGVK.G | 1.00 | IG.P.F | 1.00 | FKPG | 0.99 |
| VL.G..TNI | 1.00 | S.SGSA | 1.00 | A.WS.D | 1.00 | A..KPG | 0.99 |
| L..A..V.SS | 1.00 | V....S.N..V | 1.00 | N....RG..L | 1.00 | G..TA...K | 0.99 |
| C.AMG.E | 1.00 | N.C..K..A | 1.00 | Y..GHL | 1.00 | R..A...AI | 0.99 |
| VLCG.....GV | 1.00 | VI.GD...A | 1.00 | DKA..Y | 1.00 | E.D.G.A | 0.99 |
| SARIGHSLSL | 1.00 | LP.P....M | 1.00 | YK...GA | 1.00 | T.D...LT | 0.99 |
| LDMLDL | 1.00 | T..DG..AP | 1.00 | V..VN...N | 1.00 | S..FGL | 0.99 |
| CAKHFVGDG | 1.00 | Y..G...E.Y | 1.00 | A.TVN..G | 1.00 | GLS...Y | 0.99 |
| F....AIH | 1.00 | N..F...HD | 1.00 | APV....Q | 1.00 | GTA....P | 0.99 |
| SNSTHETNAL | 1.00 | H...G..GL | 1.00 | L..TE.S | 1.00 | K...G.I.V | 0.99 |
| IAGV.TE.CV | 1.00 | QL...N....L | 1.00 | V...IS.N | 1.00 | PE..DE | 0.99 |
| QGY..RL | 1.00 | I.FDF | 1.00 | KN.H...S | 1.00 | VGA...P | 0.99 |
| DMLQQGYTLR | 1.00 | C.....GDG | 1.00 | G.ALW | 0.99 | A.AG..H | 0.99 |
| NIPFD.Y..P | 1.00 | I.YP...F | 1.00 | ELEKR | 0.99 | G.G..V..E | 0.99 |
| LTTQEAGAVF | 1.00 | V..E.VLE | 1.00 | G.AFF | 0.99 | WW.A | 0.99 |
| LGGDHS | 1.00 | G..F.W.S | 1.00 | GSGTA | 0.99 | I..VTP | 0.99 |
| GGGEVKMMSL | 1.00 | S....I.G..T | 1.00 | G.P...E.S | 0.99 | F...Y.....T | 0.99 |
| E.E.L..PE | 1.00 | P.I...YT | 1.00 | F....GG.R | 0.99 | K.L.S...K | 0.99 |
| TDRAEGRAVL | 1.00 | GV....LS.D | 1.00 | V..G..A.P | 0.99 | LQ..DV | 0.99 |
| EGFFSQ | 1.00 | TD...G....V | 1.00 | L.TT.....P | 0.99 | GK..NF | 0.99 |
| RPQVSEGFF | 1.00 | EEL.A.P | 1.00 | HVER | 0.99 | GA....V..A | 0.99 |
| PEFKKKELE | 1.00 | L....IF..T | 1.00 | E.....KAP | 0.99 | TSP..L | 0.99 |
| RLIG.....I | 1.00 | A.IF..F | 1.00 | S.A.WS | 0.99 | LR.T.R | 0.99 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| LKQQNALDKL | 1.00 | LN..M.F | 1.00 | ILG....L | 0.99 | N.L..DL | 0.98 |
| G...M.C..I | 1.00 | YG...WA | 1.00 | V.Q..K...L | 0.99 | G....V.PG | 0.98 |
| GS.A.GK | 1.00 | Y......P..C | 1.00 | V.G..R..N | 0.99 | R....K....P | 0.98 |
| F..C.F..D | 1.00 | LFE.......P | 1.00 | I...TT.S | 0.99 | P..N...M | 0.98 |
| QC.A...R | 1.00 | LL...Y.Y | 1.00 | LS.G....D | 0.99 | FE.LT | 0.98 |
| GG.F.WG | 1.00 | T..NSV | 1.00 | A.MM.T | 0.99 | Q..P..W | 0.98 |
| GA.D.A...A | 1.00 | PK....HH | 1.00 | FA.K...D | 0.99 | M..Q..T | 0.98 |
| YP..F..E | 1.00 | VI.....NY | 1.00 | A...E...EN | 0.99 | VTEV | 0.98 |
| RG..D.K..S | 1.00 | QL.....LM | 1.00 | KDT...K | 0.99 | H.NY | 0.98 |
| GEF.DI | 1.00 | A...D.GN | 1.00 | T.R.E.W | 0.99 | MM..T | 0.97 |
| HS..G..S.SA | 1.00 | V.I.S....P | 1.00 | W..W..T | 0.99 | R.L.......T | 0.97 |
| PV..GGD | 1.00 | T.D.D....K | 1.00 | L...D...A.A | 0.99 | W.Y...G | 0.97 |
| G...W...D.F | 1.00 | A.I.G.....T | 1.00 | LS..RY | 0.99 | VHW | 0.97 |
| GY..WS...N | 1.00 | M.D.G..K | 1.00 | LI...T.V | 0.99 | C...YG | 0.96 |
| GY...SWS | 1.00 | HY.P.Y | 1.00 | I...S.GF | 0.99 | NK.......I | 0.95 |
| SSG.AP | 1.00 | I.S..Q..L | 1.00 | G...TNG | 0.99 | W..AG | 0.94 |
| E.V.GKP | 1.00 | Y..G...PD | 1.00 | I.N..R.T | 0.99 | II.....K | 0.92 |
| R.L..D...LT | 1.00 | A..W.....W | 1.00 | YT.RS | 0.99 | | |

## Χημική ομαδοποίηση αμινοξέων

| Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ |
|---|---|---|---|---|---|---|---|
| KPGKGRRK | 1.00 | [ILMV]SG.[AG]SA | 1.00 | GEF.DI | 1.00 | G..[ILMV]GS[ILMV][AG] | 1.00 |
| G.[AG][DE].K..SL | 1.00 | GA.D.A...A | 1.00 | [ILMV]GDE.[ILMV].E | 1.00 | H..L.[FY][AG][ILMV] | 1.00 |
| [ILMV][ILMV].[AG].D.AT | 1.00 | [ILMV].G[AG][DE]V...S | 1.00 | PDPA...[ILMV]V | 1.00 | EPG....H | 1.00 |
| G[AG].[DE][ILMV]K.[ILMV]S[ILMV] | 1.00 | [ILMV][ILMV]..ES.GF | 1.00 | [KR][ILMV].VGA...P | 1.00 | [ST].[AG].GSV[AG] | 1.00 |
| G..[FY]W[QN]..D[ST] | 1.00 | [ILMV]F..L[ST][AG].R | 1.00 | YP..F..E | 1.00 | R....K.[KR]..P | 1.00 |
| LGTVINQL | 1.00 | I.[ILMV]D[DE].G[KR][ILMV] | 1.00 | [KR]..VHW[AG] | 1.00 | G.GG..RL | 1.00 |
| EDGTKEPLL | 1.00 | [ILMV]..[AG][ILMV][ST].N[ILMV][AG] | 1.00 | VLCG[ILMV]...[ILMV]GV | 1.00 | [ILMV].A...[KR]..[QN][FY] | 1.00 |
| GA.D[DE]A[ST] | 1.00 | V.[QN]D.D[ST][ILMV] | 1.00 | YF..S.[ILMV].N | 1.00 | ADKG.[ILMV] | 1.00 |
| ADEAT | 1.00 | [ILMV]S.G[QN].V..[ILMV] | 1.00 | C.AMG.E | 1.00 | [QN]V.G.[ILMV]S | 1.00 |
| [ST]..A[DE]G.A | 1.00 | I[DE][QN].FR | 1.00 | LDMLDL | 1.00 | H..S.E....V | 1.00 |
| [AG][ILMV]..L[KR]L..D | 1.00 | [ILMV][ILMV]..G.A..A[ST] | 1.00 | HVER[DE] | 1.00 | C.P..DD | 1.00 |
| T.R.E..A.[ILMV] | 1.00 | V[FY][ILMV].[FY]L..[FY] | 1.00 | FKP.[FY][KR] | 1.00 | [ILMV].W.[ILMV]N[DE] | 1.00 |
| [ST].LN.LK[ILMV] | 1.00 | Q..P[ILMV].W.[AG][DE] | 1.00 | PTF....D.[ILMV] | 1.00 | [ILMV]HY.P.Y | 1.00 |
| [AG].HF[ILMV]GD | 1.00 | V[ST][FY]...R.AP | 1.00 | [DE]I..VTP | 1.00 | N..[DE]K[ILMV]D | 1.00 |
| MLQ[QN]...[ILMV][KR] | 1.00 | V.[ILMV].[FY]..NI | 1.00 | EGFFSQ | 1.00 | R[QN]FY..[QN] | 1.00 |
| [AG]KHF.G.G | 1.00 | H[ILMV].Y[AG][FY]A.[ILMV] | 1.00 | LKQQNALDKL | 1.00 | R[QN]...[AG].[KR][ILMV].[ILMV] | 1.00 |
| HAESGEYGL | 1.00 | L..A..V.SS | 1.00 | L.G..VVP.[ST]S | 1.00 | [ST]N[ST]....N[AG] | 1.00 |
| [KR][ILMV].P.[ILMV][ST]..[FY] | 1.00 | [FY].L[AG]F.[ILMV].S | 1.00 | VYL..L.[ST].Y | 1.00 | [QN].K...[AG][AG]K | 1.00 |
| [AG][ST].D..GP | 1.00 | [ILMV].T.E[ST]S.I | 1.00 | PEFKKKELE | 1.00 | N.C..K..A | 1.00 |
| G...M.C..I | 1.00 | T[ILMV]D[ILMV].[DE]L | 1.00 | H.YSQ | 1.00 | S.GC....F | 1.00 |
| P.[ST][FY].K[DE] | 1.00 | T[AG]G[ST]G.E | 1.00 | YFTG.[DE] | 1.00 | W...[ST]..HK | 1.00 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [ILMV][ST].V.Y...W | 1.00 | S[ILMV]V.GG..[ST][ILMV] | 1.00 | GSGTA[DE] | 1.00 | [QN]L...GR.[ILMV] | 1.00 |
| F..C.F..D | 1.00 | [ILMV][ILMV].G[ILMV].T.[ILMV].V | 1.00 | [ILMV].[AG]YAGL | 1.00 | N..F...HD | 1.00 |
| [FY]..K.[DE]M.V | 1.00 | G...W...D.F | 1.00 | MDHSNY | 1.00 | E[ILMV][DE][ST]E[QN] | 1.00 |
| [ILMV][DE]E.V..[KR]P | 1.00 | S....I.G[ST].T | 1.00 | IGVK.G[AG] | 1.00 | [ST].S.S..[FY][AG] | 1.00 |
| [DE]P.H...I | 1.00 | G[AG]D.S[ILMV]A | 1.00 | S.[ST].[ST][AG]F..[ILMV] | 1.00 | LP.P....M | 1.00 |
| G.[ILMV][FY]W.L | 1.00 | [ILMV]V[AG]L.[ST].[AG][FY]L[FY] | 1.00 | GY...SWS | 1.00 | F[QN][DE]..G..[KR] | 1.00 |
| [ST]G.A[DE]....K | 1.00 | G.[ILMV].W..SG | 1.00 | SARIGHSLSL | 1.00 | P....G.HS | 1.00 |
| [ILMV][KR].G.[ILMV]K.T | 1.00 | G..F[ST]W.S | 1.00 | SNSTHETNAL | 1.00 | I.[FY]P[DE]R | 1.00 |
| [KR]S.G..LA | 1.00 | A.V[QN]TE | 1.00 | LTTQEAGAVF | 1.00 | P.[FY].E.H..[ILMV] | 1.00 |
| GA.[ILMV].RY.[QN] | 1.00 | [ILMV]...F.W.[QN][AG] | 1.00 | IV..[ST][ST]V | 1.00 | L....[AG].LRG | 1.00 |
| II....GS.[AG] | 1.00 | K[AG]I.S..Q | 1.00 | R.L..D...LT | 1.00 | [KR]YG.C | 1.00 |
| [AG].LI...PQ.[ILMV] | 1.00 | H[ILMV]..D.E[AG]L | 1.00 | TD...G[KR]...V | 1.00 | [DE]..[ILMV]..DM..[ILMV] | 1.00 |
| [ILMV]I.TG.[AG]S.A | 1.00 | L.[ILMV][FY]...G[AG]T | 1.00 | [ILMV]M.[ILMV].[KR]...I | 1.00 | [AG].....SSR.[ILMV] | 1.00 |
| R..DP....VV | 1.00 | [ILMV].G[AG]T[ILMV]K | 1.00 | Y..GH....S | 1.00 | [ILMV]...YT[ILMV]R | 1.00 |
| [KR][ILMV]I[DE]..[ILMV]G.[AG] | 1.00 | D[ILMV]D...P[ILMV].[AG] | 1.00 | G...W...[AG]D[KR] | 1.00 | N.H..T...[KR] | 1.00 |
| P..[ILMV]G[AG].H | 1.00 | [FY]..A.[ILMV][ILMV][FY][ST] | 1.00 | E[AG].....[AG]PN | 1.00 | [ILMV]NS[ILMV].SI | 1.00 |
| IL.[AG]..PL | 1.00 | E..D.K.[ST][ILMV] | 1.00 | [ILMV]DC.....DA | 1.00 | [ST]....WPD | 1.00 |
| [AG]R[ILMV].A.[AG].AI | 1.00 | K[ILMV].[AG]G.IV | 1.00 | [KR]..[AG]GN.V.G | 1.00 | Y..[AG]C...Y | 1.00 |
| [QN]A[QN].[ST]V | 1.00 | P[FY].[FY]..[DE]P | 1.00 | [ILMV]...ES..[ILMV]AP | 1.00 | R.[ST]....V..S | 1.00 |
| G...[ILMV]K.[AG]A[AG] | 1.00 | [ILMV][ILMV]..YN[ST]N[ILMV] | 1.00 | G[ILMV][ST]..S[ILMV]..G | 1.00 | [QN]Y.P..L | 1.00 |
| [ILMV].L.[DE]L[ST].[DE][DE] | 1.00 | I[QN].[ILMV]G.E | 1.00 | [ST].[QN]....GS.T | 1.00 | [ST].[ILMV].[ST]..M.K | 1.00 |
| [KR]P....G.[FY][ST] | 1.00 | GL...N[ST]..S | 1.00 | [AG]...GL..HH | 1.00 | L[ILMV][ILMV][FY].....T | 1.00 |
| HI.[FY].F[AG] | 1.00 | [DE]..[ILMV].E[ILMV]..E[QN] | 1.00 | [ILMV]..Q[DE]..A[ILMV][FY] | 1.00 | G...W[ST].T | 1.00 |
| [AG]RL..VT | 1.00 | KV[ILMV]..ES.[AG] | 1.00 | S.S.H..N | 1.00 | Y......P..C | 1.00 |
| EP.[DE][FY].[KR][KR] | 1.00 | GS[ILMV]A.GK | 1.00 | [DE][ST]G.YG | 1.00 | L[ST][ST]..[AG]...F | 1.00 |
| S[AG]..[KR]GP | 1.00 | [KR].[KR].G.G.R.P | 1.00 | E.[AG][ST]...H[AG] | 1.00 | [ILMV]N.L..DL | 1.00 |
| IV.[ILMV]ES.G | 1.00 | E.E.L..PE | 1.00 | G...[AG].[ILMV]PH | 1.00 | K[DE]...[QN].T[ILMV] | 1.00 |
| [ILMV]A.WS.D | 1.00 | GL[DE][ST].[AG].[ILMV][FY] | 1.00 | [ILMV]..PG.N.V | 1.00 | AM.I[DE][AG] | 1.00 |
| [ILMV][ILMV]FI..[ILMV].[KR]F | 1.00 | H[ST].[ST].LLL | 1.00 | [ST]..G..[FY]FG | 1.00 | [ILMV][ILMV].....[ILMV]W.[ST] | 1.00 |
| LS.[ILMV]RY | 1.00 | EEL.A[AG]P | 1.00 | V.[QN]LN.....[ILMV] | 1.00 | G[QN].KG.[ILMV] | 1.00 |
| [ILMV][ST]TN..V | 1.00 | [ILMV][FY].NKF[DE] | 1.00 | L..T[DE]R..[ST] | 1.00 | ER.[AG].AL | 1.00 |
| RQ.Y.[ILMV][QN] | 1.00 | RLIG.....I | 1.00 | [AG].W[ST].[ST]..D | 1.00 | [AG]A....VN[FY] | 1.00 |
| Y[FY].L[AG].I[ILMV] | 1.00 | NAW..D | 1.00 | Y[ILMV].[DE]K...K | 1.00 | E[ILMV]E.[KR][KR] | 1.00 |
| I.[ST]GK.S | 1.00 | CLA..VN[ILMV] | 1.00 | N[ILMV]..AA[ILMV].L | 1.00 | E...[KR][DE]..[QN][KR] | 1.00 |
| [AG][FY]GG.F.WG | 1.00 | ISF.[DE][KR] | 1.00 | Y.GG.[FY].[ST] | 1.00 | PE[ILMV].DE | 1.00 |
| [ST]QC.A...R | 1.00 | [ILMV]NIPFD.Y..P | 1.00 | G.[ILMV].W.....W | 1.00 | V..G..A.P | 0.99 |
| G.[QN]TA...K[ST] | 1.00 | P..GIA[FY] | 1.00 | I...TT[ILMV]S | 1.00 | [AG].CY.[ST] | 0.99 |
| V...D.[ST]NY | 1.00 | HS..G..S.SA | 1.00 | L...[KR].G[QN]..V | 1.00 | | |

## Δομική ομαδοποίηση αμινοξέων

| Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ | Ακολουθιακά πρότυπα | Σκορ |
|---|---|---|---|---|---|---|---|
| KPGKGRRK | 1.00 | [CS]..C....FR | 1.00 | NIPFD.Y..P | 1.00 | [KMR]Q..KA..[FHWY] | 1.00 |
| S.S[ITV]H..N | 1.00 | [ITV].G[KMR]H[FHWY].G | 1.00 | GDN.[FHWY]...[ITV][FHWY] | 1.00 | I.TG..S.A | 1.00 |
| E.V[DLN].[KMR]P | 1.00 | [KMR][FHWY]...G[KMR]F[ITV] | 1.00 | C...[FHWY].GDG.[ITV] | 1.00 | V.I.F[DLN]..[FHWY] | 1.00 |
| MLQ...[ITV].[KMR] | 1.00 | [FHWY]GG.F[ITV]WG | 1.00 | VL.G..TNI | 1.00 | IV....RG.[DLN] | 1.00 |
| [EQ].GYT.R | 1.00 | N[KMR]K.[KMR][ITV].I | 1.00 | PTF....D.[ITV] | 1.00 | I[ITV].GV.[KMR]G | 1.00 |
| [KMR]..QGY..R | 1.00 | [ITV]I[DLN]T.[ITV]...A | 1.00 | F.[EQ][CS].AIH | 1.00 | [ITV]...NK.[EQ].L | 1.00 |
| EDGTKEPLL | 1.00 | A[FHWY]IF..F | 1.00 | [ITV][CS]..[KMR].AAA | 1.00 | [KMR]...GN.V.G | 1.00 |
| LGTVINQL | 1.00 | [CS]..Q[DLN]V[DLN].[ITV] | 1.00 | LFE[KMR]...[CS]..P | 1.00 | S..I.[ITV]K[ITV] | 1.00 |
| HAESGEYGL | 1.00 | [CS]..C[FHWY]T..F | 1.00 | [ITV]LGG...LA | 1.00 | Y....N.[ITV]G[DLN] | 1.00 |
| ADEAT | 1.00 | Q[ITV]A[DLN]..K[ITV] | 1.00 | RLIG.....I | 1.00 | G.G[KMR].D[DLN][ITV] | 1.00 |
| ALNALKLVT | 1.00 | [ITV]L[CS]GV....G | 1.00 | LTTQEAGAVF | 1.00 | V.[FHWY]...R.AP | 1.00 |
| GG...[KMR]M..L | 1.00 | [DLN]G[DLN]IV.[DLN] | 1.00 | L[EQ]G..VVP..S | 1.00 | A[ITV]I[FHWY]..[DLN]..[FHWY] | 1.00 |
| [DLN]L.EL..E[EQ] | 1.00 | P..[DLN].G[DLN]HS | 1.00 | FKP.[FHWY][KMR] | 1.00 | [KMR]....[KMR].K.[ITV]P | 1.00 |
| [EQ]..P..[FHWY]P.E | 1.00 | N.[DLN].G[DLN]H | 1.00 | VYL..L...Y | 1.00 | NT[FHWY]..[ITV].G | 1.00 |
| QL...N.L.[KMR][DLN] | 1.00 | [DLN]VL.G[DLN]H | 1.00 | PEFKKKELE | 1.00 | [ITV]..DEKV[DLN] | 1.00 |
| A[FHWY].[FHWY]E...EN | 1.00 | [DLN].K.R.[FHWY]H | 1.00 | EGFFSQ | 1.00 | [ITV]..N.D.P.[ITV] | 1.00 |
| M[FHWY].[EQ][FHWY].D[ITV] | 1.00 | VV..T[ITV]N | 1.00 | RPQVSEGFF | 1.00 | GL..[ITV]N...S | 1.00 |
| [ITV]..G[ITV].T.[ITV].V | 1.00 | [ITV].G[CS]G..D..T | 1.00 | TDRAEGRAVL | 1.00 | G.GG..RL | 1.00 |
| [KMR]Y...N.V[CS] | 1.00 | [KMR]S[KMR]G..LA | 1.00 | A...KHF.G.G | 1.00 | [CS]R..Y..Q | 1.00 |
| [FHWY]..KG.[ITV].R[ITV] | 1.00 | W.[FHWY].S.[FHWY][KMR] | 1.00 | SARIGHSLSL | 1.00 | V.Q..K...L[ITV] | 1.00 |
| GA.D.A...A | 1.00 | K[FHWY]PV.[FHWY] | 1.00 | S.GS..[ITV]G | 1.00 | [ITV]A.[ITV]..A.I.[FHWY] | 1.00 |
| [FHWY][DLN]...M.F..[ITV] | 1.00 | [FHWY].S.G.PS[FHWY] | 1.00 | [ITV]..DP..G.[KMR][ITV] | 1.00 | [FHWY]E.[KMR]..G..G | 1.00 |
| [DLN].PY.Y.T[EQ] | 1.00 | SV[ITV]..G[FHWY]..V | 1.00 | G.[EQ].M.C..I | 1.00 | GA.[ITV].RY | 1.00 |
| WW.A..[FHWY] | 1.00 | R.[FHWY]..[ITV]WH[CS] | 1.00 | HS..G..S.SA | 1.00 | F.G.D[ITV][FHWY] | 1.00 |
| [CS].A.WS.D | 1.00 | QC[EQ]A...R | 1.00 | YP..F..E | 1.00 | [DLN][DLN]L[FHWY]P.G | 1.00 |
| [ITV][ITV]..ES.GF | 1.00 | SG[ITV]A...G | 1.00 | [KMR].G.G..[DLN]P | 1.00 | T[ITV]E.S.I | 1.00 |
| R..DP....VV | 1.00 | D..[FHWY].YP[ITV] | 1.00 | [KMR][DLN].N.[DLN].[ITV]K | 1.00 | C.P..DD | 1.00 |
| GV....LS.D[DLN] | 1.00 | KA[FHWY].W.....[DLN] | 1.00 | P.[FHWY]R.[FHWY].[FHWY] | 1.00 | E...[KMR]E.E.R | 1.00 |
| G..[FHWY]..LS.D | 1.00 | L..V..KP[ITV] | 1.00 | TD...G[KMR]...V | 1.00 | E[DLN].[DLN]..G.[ITV][ITV] | 1.00 |
| L..T.[CS][DLN][ITV]..K | 1.00 | [EQ]...LD[KMR]LD | 1.00 | L..[ITV]..G[EQ].W | 1.00 | VI.GD...A | 1.00 |
| C.[EQ]C....F.[FHWY] | 1.00 | GS[ITV]A.GK | 1.00 | [KMR][DLN]...Y.P..L | 1.00 | H.[FHWY]..[KMR].D[KMR] | 1.00 |
| [KMR]G....K..SL | 1.00 | [KMR]EP[DLN][DLN].[ITV] | 1.00 | [ITV].[DLN].[DLN]S..[ITV]K[ITV] | 1.00 | L.G.S...G[ITV] | 1.00 |
| [ITV][DLN]Y...[DLN].[CS]W | 1.00 | G[ITV]AL[FHWY][DLN]...D | 1.00 | V[ITV].GD[DLN]G | 1.00 | N.C..K..A | 1.00 |
| [DLN][KMR]L.LG..[KMR] | 1.00 | HL[ITV]....GL | 1.00 | A.V[DLN]...[EQ]A.[ITV] | 1.00 | QL...N....L | 1.00 |
| VT.[DLN][KMR]L.[DLN] | 1.00 | DG[DLN]I[ITV][ITV] | 1.00 | [FHWY][KMR]..KG..K | 1.00 | [EQ].A.V[EQ].E | 1.00 |
| [FHWY][FHWY][ITV]L.F...S | 1.00 | [ITV][ITV]AG[ITV]..EV | 1.00 | [DLN].[DLN]..G[FHWY]T[DLN] | 1.00 | Q[DLN]..[DLN]K.D | 1.00 |
| [ITV][DLN]S.V.[CS]A | 1.00 | R.L..D.[EQ].LT | 1.00 | [FHWY][ITV]..GS.[ITV] | 1.00 | G[DLN].[CS]T..K | 1.00 |

| | | | | .K | | | |
|---|---|---|---|---|---|---|---|
| [ITV][ITV][CS][KMR]N.[FHWY] | 1.00 | VT[EQ].[EQ].[ITV][ITV] | 1.00 | L..PF.L..[FHWY] | 1.00 | [FHWY][KMR].T..[FHWY]K | 1.00 |
| I..N.[FHWY]G[CS] | 1.00 | FA[DLN]K.[DLN].D | 1.00 | [KMR][FHWY][DLN]P...[CS].V | 1.00 | E.[DLN]G.A[DLN] | 1.00 |
| [FHWY][FHWY]G[CS]..I...V | 1.00 | N[ITV]..AA[ITV].L | 1.00 | S..[DLN][FHWY]S..[FHWY][KMR] | 1.00 | YG...WA | 1.00 |
| [ITV][EQ][ITV]E[FHWY].[ITV] | 1.00 | E..[EQ]..KAP | 1.00 | A[ITV].[FHWY]AA...[FHWY] | 1.00 | EEL.A.P | 1.00 |
| Y.[ITV]GH....S | 1.00 | EPG[ITV]...H | 1.00 | GR.....LAI | 1.00 | L....IF..T | 1.00 |
| LS.[ITV]RY | 1.00 | [KMR]..VGA[ITV]..P | 1.00 | [ITV].IDE[ITV]...[EQ] | 1.00 | [ITV]V[FHWY]G.D | 1.00 |
| G.[FHWY][ITV]WS..D | 1.00 | H.YSQ.[DLN] | 1.00 | S..P.I..T[FHWY] | 1.00 | S..N[EQ]..[ITV][DLN] | 1.00 |
| I..[DLN]TT.S | 1.00 | CLA..VN | 1.00 | [KMR].[ITV]Y.[FHWY]..N[ITV] | 1.00 | G....SL...T | 1.00 |
| [ITV]...K.[ITV][DLN]GA | 1.00 | NAW..D | 1.00 | G[FHWY].FS...[KMR] | 1.00 | [FHWY]..K..M.V | 1.00 |
| [KMR]TP...[FHWY].D[ITV] | 1.00 | G[ITV].[FHWY]W..NDS | 1.00 | A.KG.[ITV].[KMR]V | 1.00 | DC.....DA | 1.00 |
| [ITV].ES..[ITV]AP | 1.00 | [FHWY].NKF..[DLN][ITV] | 1.00 | A..T.TG.[DLN].[ITV] | 1.00 | SG[ITV].[DLN].P | 1.00 |
| T..DG.[ITV]AP | 1.00 | P..R.[FHWY].H[FHWY] | 1.00 | PT.R.E..A | 1.00 | GSG.A...L | 1.00 |
| [DLN]P.P..L[KMR][KMR] | 1.00 | WSG[DLN]G | 1.00 | [ITV]..PG.N.V | 1.00 | [KMR]G[EQ][ITV]W | 1.00 |
| QL...[DLN].L[KMR][KMR] | 1.00 | L[EQ]...GKP[ITV][DLN] | 1.00 | L.A.AG....[FHWY] | 1.00 | VG..[DLN]DI | 1.00 |
| [FHWY]N.A[CS][CS] | 1.00 | P.[DLN]GIA[FHWY] | 1.00 | [DLN]V[EQ].Q..[FHWY][DLN] | 1.00 | ELEKR | 0.99 |
| [ITV][FHWY].[KMR]T[DLN][ITV] | 1.00 | L[KMR].SRGF | 1.00 | QL....[KMR]LM | 1.00 | AL...A[FHWY]L | 0.99 |
| L..A..V.SS | 1.00 | V...DHS[DLN] | 1.00 | H...[EQ].[KMR]TK | 1.00 | [FHWY].[KMR]G[ITV]..P | 0.99 |
| M..S..A[FHWY][FHWY].[DLN] | 1.00 | MDHSNY | 1.00 | G[EQ]P..[DLN]..H | 1.00 | | |
| [ITV][DLN]P[DLN]P....[KMR][ITV] | 1.00 | REPDP | 1.00 | I.[CS]..Q.[ITV][DLN] | 1.00 | | |
| [KMR]..[FHWY].[FHWY].P[DLN]T | 1.00 | C.AMG.E | 1.00 | VAL....Y[DLN] | 1.00 | | |

# Δημοσιεύσεις διδακτορικής διατριβής

**Κεφάλαια σε βιβλία**

1. **K.P. Exarchos**, T.P. Exarchos and D.I. Fotiadis, "Genome Informatics: protein interaction networks and regions of protein disorder", submitted.

2. **K.P. Exarchos**, Y. Goletsis and D.I. Fotiadis, "Unification of heterogeneous data towards the prediction of oral cancer reoccurrence", CEUR WS, pp. 24-35, 2009.

3. **K.P. Exarchos**, G. Rigas, Y. Goletsis and D.I. Fotiadis, "Modeling of oral cancer progression using Dynamic Bayesian Networks", Springer, submitted.

**Δημοσιεύσεις σε περιοδικά**

1. **K.P. Exarchos**, C. Papaloukas, T.P. Exarchos, A.N. Troganis and D.I. Fotiadis, "Prediction of Cis/trans isomerization using feature selection and support vector machines", Journal of Biomedical Informatics, 42(1), 140-9, 2009.

2. **K.P. Exarchos**, T.P. Exarchos, C. Papaloukas, A.N. Troganis and D.I. Fotiadis, "PBOND: Web server for the prediction of proline and non-proline cis/trans isomerization", Genomics, Proteomics & Bioinformatics, 2010.

3. **K.P. Exarchos**, T.P. Exarchos, C. Papaloukas, A.N. Troganis and D.I. Fotiadis, "Detection of discriminative sequence patterns in the neighborhood of proline cis peptide bonds and their functional annotation", BMC Bioinformatics, 10(113), 2009.

4. **K.P. Exarchos**, T.P. Exarchos, C. Papaloukas, G. Rigas and D.I. Fotiadis, "Extraction of consensus protein patterns in regions containing non-proline cis

peptide bonds and their functional assessment", BMC Bioinformatics, 12(142), 2011.

5. **K.P. Exarchos**, G. Rigas, C. Papaloukas and D.I. Fotiadis, "pCOMPARE: compare, visualize and annotate protein patterns", (in preparation).

6. **K.P. Exarchos**, T.P. Exarchos, C. Papaloukas and D.I. Fotiadis, "Discovery of sequence patterns mediating functions in disorder-oriented protein interaction networks", submitted.

7. C. Lampros, C. Papaloukas, **K.P. Exarchos** and D.I. Fotiadis, "Two-step improvement in fold recognition accuracy of a reduced state-space Hidden Markov Model", accepted in Computers in Biology and Medicine.

8. C. Lampros, T.P. Exarchos, C. Papaloukas, **K.P. Exarchos** and D.I. Fotiadis, "Protein fold recognition using optimized scores of Markov models", submitted.

9. **K.P. Exarchos**, Y. Goletsis and D.I. Fotiadis, " A Multiscale and Multiparametric Approach for Modeling the Progression of Oral Cancer ", submitted.

10. **K.P. Exarchos**, Y. Goletsis and D.I. Fotiadis, " Multiparametric Decision Support System for the Prediction of Oral Cancer Reoccurrence ", accepted in IEEE Transactions on Information Technology in Biomedicine (IEEE TITB).

11. M. Picone, S. Steger, **K.P. Exarchos**, M. Fazio, Y. Goletsis, D.I. Fotiadis, E. Martinelli, D. Ardigo, "Enabling heterogeneous data integration and biomedical event prediction through ICT: the test case of cancer reoccurrence", Advances in Experimental Medicine and Biology, 2011.

12. T.P. Exarchos, M.G. Tsipouras, **K.P. Exarchos**, C. Papaloukas, D.I. Fotiadis and L. K. Michalis, "A methodology for the automated creation of fuzzy expert systems for ischaemic and arrhythmic beat classification based on a set of rules obtained by a decision tree", Artificial Intelligence in Medicine, 40(3), pp. 187-200, 2007.

**Δημοσιεύσεις σε συνέδρια**

1. **K.P. Exarchos**, T.P. Exarchos, C. Papaloukas, A.N. Troganis and D.I. Fotiadis, "Predicting peptide bond conformation using feature selection and the Naïve Bayes approach", 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBC 2007), Lyon, France.

2. **K.P. Exarchos**, T.P. Exarchos, C. Papaloukas, A.N. Troganis and D.I. Fotiadis, "Identification of sequence patterns associated with proline cis/trans isomerization", 7th European Conference on Computational Biology (ECCB 2008), Cagliari, Sardinia. **(Best paper award - EMBRACE Grid fellowship)**

3. **K.P. Exarchos**, T.P. Exarchos, C. Papaloukas, A.N. Troganis and D.I. Fotiadis, "Systematic elicitation of sequence patterns associated with non-proline cis peptide bonds", 8th International Conference on Bioinformatics and BioEngineering (BIBE 2008), Athens, Greece.

4. **K.P. Exarchos**, G. Rigas and D.I. Fotiadis, "Evolutionarily driven algorithm for the quantification of protein patterns' similarity", 17th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2009) and 8th European Conference on Computational Biology (ECCB 2009), Stockholm, Sweden.

5. **K.P. Exarchos**, G. Rigas and D.I. Fotiadis, "pCOMPARE: an evolutionarily driven algorithm for the quantification of protein patterns' similarity", ISMB Students Council Symposium 2009. **(Best paper award - ISCB fellowship)**

6. C. Lampros, C. Papaloukas, **K.P. Exarchos** and D.I. Fotiadis, "Improvement in fold recognition accuracy of a reduced-state-space hidden markov model by using secondary structure information in scoring", 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBC 2007), Lyon, France.

7. **K.P. Exarchos**, G. Rigas, Y. Goletsis, D. Ardigo and D.I. Fotiadis, “Oral cancer reoccurrence prediction using Dynamic Bayesian Networks”, 4th International Conference on Computational BioEngineering (ICCB 2009), Bertinoro, Italy.

8. **K.P. Exarchos**, Y. Goletsis and D.I. Fotiadis, Unification of heterogeneous data towards the prediction of oral cancer reoccurrence”, 5th IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI 2009) - Workshop on Biomedical Informatics and Intelligent Approaches in the support of Genomic Medicince (BMINT 2009), Thessaloniki, Greece.

9. **K.P. Exarchos**, G. Rigas, Y. Goletsis and D.I. Fotiadis, "Dynamic Bayesian Networks for disease evolution monitoring: the test case of oral cancer reoccurrence", 4th Conference of the Hellenic Society for Computational Biology (HSCBB 2009), Athens, Greece. **(Best paper award - HCSBB prize)**

10. **K.P. Exarchos**, G. Rigas, Y. Goletsis and D.I. Fotiadis, "A multilevel and multiscale approach for the prediction of oral cancer reoccurrence", XII Mediterranean Conference on Medical and Biological Engineering and Computing (Medicon 2010), Chalkidiki, Greece.

11. **K.P. Exarchos**, G. Rigas, Y. Goletsis and D.I. Fotiadis, "Modeling of oral cancer progression using Dynamic Bayesian Networks", International Conference on Biomedical Data and Knowledge Mining: Towards Biomarker Discovery (DMBIO 2010), Chania, Greece.

12. **K.P. Exarchos**, Y. Goletsis, F.G. Kalatzis, N. Giannakeas, V. Oikonomou and D.I. Fotiadis, "NeoMark: ICT platform for the prediction of oral cancer reoccurrence", VPH Network of Excellence (VPH-NoE 2010), Brussels, Belgium.

13. M. Picone, S. Steger, **K.P. Exarchos**, M. Fazio, G. Chiari, D. Ardigo, E. Martinelli, "NeoMark: how to predict oral cancer recurrence through multiscale data analysis", VPH Network of Excellence (VPH-NoE 2010), Brussels, Belgium.

14. **K.P. Exarchos**, G. Rigas, Y. Goletsis and D.I. Fotiadis, "Towards building a Dynamic Bayesian Network for monitoring oral cancer progression using time-course gene expression data", 10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB 2010), Corfu, Greece.

15. T. Poli, D. Ardigo, E. Derlindati, D. Lanfranco, **K.P. Exarchos**, Y. Goletsis and D.I. Fotiadis, "Identification of genetic biomarkers for oral cancer reoccurrence prediction", ASCO-NCI-EORTC Annual Meeting on Molecular Markers in Cancer, Hollywood, Florida, 2010.

16. **K.P. Exarchos**, Y. Goletsis, T. Poli and D.I. Fotiadis, "Gene expression profiling towards the prediction of oral cancer reoccurrence", submitted.

# Βιβλιογραφία


1. Crick FH. On protein synthesis. Symp Soc Exp Biol. 1958;12:138-63.
2. Crick F. Central dogma of molecular biology. Nature. 1970 Aug 8;227(5258):561-3.
3. Anfinsen CB. Principles that govern the folding of protein chains. Science (New York, NY. 1973 Jul 20;181(96):223-30.
4. Baxevanis AD, Ouellette BFF, editors. Bioinformatics: a practical guide to the analysis of genes and proteins. 3rd ed: Hoboken, NJ:. Wiley; 2005.
5. Berg JM, Tymoczko JL, Stryer L. Biochemistry. 6th ed. New York: W. H. Freeman; 2007.
6. Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proceedings of the National Academy of Sciences of the United States of America. 1951 Apr;37(4):205-11.
7. MacArthur MW, Thornton JM. Influence of proline residues on protein conformation. Journal of molecular biology. 1991 Mar 20;218(2):397-412.
8. Weiss MS, Jabs A, Hilgenfeld R. Peptide bonds revisited. Nature structural biology. 1998 Aug;5(8):676.
9. Stewart DE, Sarkar A, Wampler JE. Occurrence and role of cis peptide bonds in protein structures. Journal of molecular biology. 1990 Jul 5;214(1):253-60.
10. Grathwohl C, Wuethrich K. NMR studies of the rates of proline cis-trans isomerization in oligopeptides. Biopolymers. 1981;20(12):2623-33.
11. Pal D, Chakrabarti P. Cis peptide bonds in proteins: residues involved, their conformations, interactions and locations. Journal of molecular biology. 1999 Nov 19;294(1):271-88.
12. Pahlke D, Freund C, Leitner D, Labudde D. Statistically significant dependence of the Xaa-Pro peptide bond conformation on secondary structure and amino acid sequence. BMC structural biology. 2005;5:8.
13. Dugave C, Demange L. Cis-trans isomerization of organic molecules and biomolecules: implications and applications. Chemical reviews. 2003 Jul;103(7):2475-532.

14. Lu KP, Finn G, Lee TH, Nicholson LK. Prolyl cis-trans isomerization as a molecular timer. Nature chemical biology. 2007 Oct;3(10):619-29.
15. Lorenzen S, Peters B, Goede A, Preissner R, Frommel C. Conservation of cis prolyl bonds in proteins during evolution. Proteins. 2005 Feb 15;58(3):589-95.
16. Fischer G, Aumuller T. Regulation of peptide bond cis/trans isomerization by enzyme catalysis and its implication in physiological processes. Reviews of physiology, biochemistry and pharmacology. 2003;148:105-50.
17. Brauer AB, Domingo GJ, Cooke RM, Matthews SJ, Leatherbarrow RJ. A conserved cis peptide bond is necessary for the activity of Bowman-Birk inhibitor protein. Biochemistry. 2002 Aug 27;41(34):10608-15.
18. Stoddard BL, Pietrokovski S. Breaking up is hard to do. Nature structural biology. 1998 Jan;5(1):3-5.
19. Frommel C, Preissner R. Prediction of prolyl residues in cis-conformation in protein structures on the basis of the amino acid sequence. FEBS letters. 1990 Dec 17;277(1-2):159-63.
20. Wang ML, Li WJ, Wang ML, Xu WB. Support vector machines for prediction of peptidyl prolyl cis/trans isomerization. J Pept Res. 2004 Jan;63(1):23-8.
21. Tan P-N, Steinbach M, Kumar V. Introduction to data mining. 1st ed. Boston: Pearson Addison Wesley; 2006.
22. Pahlke D, Leitner D, Wiedemann U, Labudde D. COPS--cis/trans peptide bond conformation prediction of amino acids on the basis of secondary structure information. Bioinformatics (Oxford, England). 2005 Mar 1;21(5):685-6.
23. Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry. 1974 Jan 15;13(2):222-45.
24. Song J, Burrage K, Yuan Z, Huber T. Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information. BMC bioinformatics. 2006;7:124.
25. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997 Sep 1;25(17):3389-402.
26. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology. 1999 Sep 17;292(2):195-202.

27. Rooman MJ, Rodriguez J, Wodak SJ. Relations between protein sequence and structure and their significance. Journal of molecular biology. 1990;213(2):337-50.

28. Rooman MJ, Wodak SJ. Weak Correlation Between Predictive Power Of Individual Sequence Patterns and Overall Prediction Accuracy in Proteins. Proteins: Structure, Function, and Genetics. 1991;9:69-78.

29. Lise S, Jones DT. Sequence patterns associated with disordered regions in proteins. PROTEINS: Structure, Function, and Bioinformatics. 2005;58(1):144-50.

30. Vucetic S, Brown CJ, Dunker AK, Obradovic Z. Flavors of protein disorder. Proteins. 2003 Sep 1;52(4):573-84.

31. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annual review of biophysics. 2008;37:215-46.

32. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, et al. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. Journal of proteome research. 2007 May;6(5):1882-98.

33. Vucetic S, Xie H, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, et al. Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. Journal of proteome research. 2007 May;6(5):1899-916.

34. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, et al. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. Journal of proteome research. 2007 May;6(5):1917-32.

35. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. Journal of molecular biology. 2004 Mar 26;337(3):635-45.

36. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform. 2000;11:161-71.

37. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. Journal of molecular biology. 2002;323(3):573-84.

38. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets. The roles of intrinsic disorder in protein interaction networks. The FEBS journal. 2005 Oct;272(20):5129-48.

39. Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J Mol Recognit. 2005 Sep-Oct;18(5):343-84.

40. Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, et al. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS computational biology. 2006 Aug 4;2(8):e100.

41. Dobson CM. Protein misfolding, evolution and disease. Trends in biochemical sciences. 1999 Sep;24(9):329-32.

42. Russell RB, Gibson TJ. A careful disorderliness in the proteome: sites for interaction and targets for future therapies. FEBS letters. 2008 Apr 9;582(8):1271-5.

43. Ideker T, Sharan R. Protein networks in disease. Genome Res. 2008 Apr;18(4):644-52.

44. Prasad TS, Kandasamy K, Pandey A. Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. Methods Mol Biol. 2009;577:67-79.

45. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic acids research. 2004 Jan 1;32(Database issue):D449-51.

46. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, et al. The IntAct molecular interaction database in 2010. Nucleic acids research. 2010 Jan;38(Database issue):D525-31.

47. Bader GD, Cary MP, Sander C. Pathguide: a pathway resource list. Nucleic acids research. 2006 Jan 1;34(Database issue):D504-6.

48. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. Bioinformatics (Oxford, England). 2006 Sep 15;22(18):2291-7.

49. Wachi S, Yoneda K, Wu R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. Bioinformatics (Oxford, England). 2005 Dec 1;21(23):4205-8.

50. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. Proceedings of the National Academy of Sciences of the United States of America. 2007 May 22;104(21):8685-90.

51. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol. 2007 Mar;25(3):309-16.

52. Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. Molecular systems biology. 2008;4:189.

53. Wu X, Liu Q, Jiang R. Align human interactome with phenome to identify causative genes and networks underlying disease families. Bioinformatics (Oxford, England). 2009 Jan 1;25(1):98-104.

54. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, Cho RJ, et al. A network-based analysis of systemic inflammation in humans. Nature. 2005 Oct 13;437(7061):1032-7.

55. Pujana MA, Han JD, Starita LM, Stevens KN, Tewari M, Ahn JS, et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. Nature genetics. 2007 Nov;39(11):1338-49.

56. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Molecular systems biology. 2007;3:140.

57. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug-target network. Nat Biotechnol. 2007 Oct;25(10):1119-26.

58. Peto J. Cancer epidemiology in the last century and the next decade. Nature. 2001 May 17;411(6835):390-5.

59. Weinberg RA. A molecular basis of cancer. Sci Am. 1983 Nov;249(5):126-42.

60. Druker B. Imatinib (Gleevec) as a paradigm of targeted cancer therapies. Keio J Med. 2010 Mar;59(1):1-3.

61. Bos PD, Zhang XH, Nadal C, Shu W, Gomis RR, Nguyen DX, et al. Genes that mediate breast cancer metastasis to the brain. Nature. 2009 Jun 18;459(7249):1005-9.

62. Boring CC, Squires TS, Tong T, Montgomery S. Cancer statistics, 1994. CA Cancer J Clin. 1994 Jan-Feb;44(1):7-26.

63. Haddad RI, Shin DM. Recent advances in head and neck cancer. The New England journal of medicine. 2008 Sep 11;359(11):1143-54.

64. Mork J, Lie AK, Glattre E, Hallmans G, Jellum E, Koskela P, et al. Human papillomavirus infection as a risk factor for squamous-cell carcinoma of the head and neck. The New England journal of medicine. 2001 Apr 12;344(15):1125-31.

65. Forastiere A, Weber R, Ang K. Treatment of head and neck cancer. The New England journal of medicine. 2008 Mar 6;358(10):1076; author reply 7-8.

66. Godden DR, Ribeiro NF, Hassanein K, Langton SG. Recurrent neck disease in oral cancer. J Oral Maxillofac Surg. 2002 Jul;60(7):748-53; discussion53-5.

67. Sciubba JJ. Oral cancer. The importance of early diagnosis and treatment. American journal of clinical dermatology. 2001;2(4):239-51.

68. D'Silva NJ, Ward BB. Tissue biomarkers for diagnosis & management of oral squamous cell carcinoma. The Alpha omegan. 2007;100(4):182-9.

69. Lippman SM, Hong WK. Molecular markers of the risk of oral cancer. The New England journal of medicine. 2001 Apr 26;344(17):1323-6.

70. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest. 1991 Dec;100(6):1619-36.

71. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. Jama. 1993 Dec 22-29;270(24):2957-63.

72. Cruz JA, Wishart DS. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics. 2006;2:59-78.

73. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine. 2005 Jun;34(2):113-27.

74. Shimizu K, Hirose S, Noguchi T, Muraoka Y. Predicting the protein disordered region using modified position specific scoring matrix. Genome Informatics. 2004;150.

75. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, et al. The RCSB Protein Data Bank: redesigned web site and web services. Nucleic acids research. 2010 Oct 29.

76. Wang G, Dunbrack RL, Jr. PISCES: recent improvements to a PDB sequence culling server. Nucleic acids research. 2005 Jul 1;33(Web Server issue):W94-8.

77. Willard L, Ranjan A, Zhang H, Monzavi H, Boyko RF, Sykes BD, et al. VADAR: a web server for quantitative evaluation of protein structure quality. Nucleic acids research. 2003 Jul 1;31(13):3316-9.

78. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, et al. Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic acids research. 2002 Jan 1;30(1):13-6.

79. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics (Oxford, England). 2000;16(4):404-5.

80. Ahmad S, Gromiha MM, Sarai A. RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. Bioinformatics (Oxford, England). 2003 Sep 22;19(14):1849-51.

81. Argos P, Rao JK, Hargrave PA. Structural prediction of membrane-bound proteins. European journal of biochemistry / FEBS. 1982 Nov 15;128(2-3):565-75.

82. Grantham R. Amino acid difference formula to help explain protein evolution. Science (New York, NY. 1974 Sep 6;185(4154):862-4.

83. Su CT, Chen CY, Ou YY. Protein disorder prediction by condensed PSSM considering propensity for order or disorder. BMC bioinformatics. 2006;7:319.

84. Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence. 1997;97(1-2):273-324.

85. Ohno-Machado L. Research on machine learning issues in biomedical informatics modeling. J Biomed Inform. 2004 Aug;37(4):221-3.

86. Witten IH, Frank E. Data mining : practical machine learning tools and techniques. 2nd ed. Amsterdam ; Boston, MA: Morgan Kaufman; 2005.

87. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861-74.

88. Exarchos KP, Papaloukas C, Exarchos TP, Troganis AN, Fotiadis DI. Prediction of cis/trans isomerization using feature selection and support vector machines. J Biomed Inform. 2009 Feb;42(1):140-9.

89. Rigoutsos I, Floratos A. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics (Oxford, England). 1998;14(1):55-67.

90. Floratos A, Rigoutsos I, Parida L, Stolovitzky G, Gao Y, editors. Sequence homology detection through large scale pattern discovery. RECOMB; 1999: ACM.

91. Rigoutsos I, Floratos A, Ouzounis C, Gao Y, Parida L. Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins. Proteins. 1999 Nov 1;37(2):264-77.

92. Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, de Castro E, et al. The 20 years of PROSITE. Nucleic acids research. 2008 Jan;36(Database issue):D245-9.

93. Exarchos KP, Rigas G, Fotiadis DI, editors. Evolutionarily driven algorithm for the quantification of protein patterns' similarity. ISMB/ECCB; 2009.

94. Gould CM, Diella F, Via A, Puntervoll P, Gemund C, Chabanis-Davidson S, et al. ELM: the status of the 2010 eukaryotic linear motif resource. Nucleic acids research. 2010 Jan;38(Database issue):D167-80.

95. Edwards RJ, Davey NE, Shields DC. CompariMotif: quick and easy comparisons of sequence motifs. Bioinformatics (Oxford, England). 2008 May 15;24(10):1307-9.

96. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A Sequence Logo Generator. Genome Research. 2004;14:1188-90.

97. Jabs A, Weiss MS, Hilgenfeld R. Non-proline cis peptide bonds in proteins. Journal of molecular biology. 1999 Feb 12;286(1):291-304.

98. Herzberg O, Moult J. Analysis of the steric strain in the polypeptide backbone of protein molecules. Proteins. 1991;11(3):223-9.

99. Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, Gibson TJ, et al. Systematic discovery of new recognition peptides mediating protein interaction networks. PLOS Biology. 2005;3(12):e405.

100. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, et al. DisProt: the Database of Disordered Proteins. Nucleic acids research. 2007 Jan;35(Database issue):D786-93.

101. Tarcea VG, Weymouth T, Ade A, Bookvich A, Gao J, Mahavisno V, et al. Michigan molecular interactions r2: from interacting proteins to pathways. Nucleic acids research. 2009 Jan;37(Database issue):D642-6.

102. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 2004 Jul 1;430(6995):88-93.

103. Zotenko E, Mestre J, O'Leary DP, Przytycka TM. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. PLoS computational biology. 2008;4(8):e1000140.

104. Jin G, Zhang S, Zhang XS, Chen L. Hubs with network motifs organize modularity dynamically in the protein-protein interaction network of yeast. PloS one. 2007;2(11):e1207.

105. Mizianty MJ, Stach W, Chen K, Kedarisetti KD, Disfani FM, Kurgan L. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics. 2010 Sep 15;26(18):i489-96.

106. Jones DT, Swindells MB. Getting the most from PSI-BLAST. Trends in biochemical sciences. 2002 Mar;27(3):161-4.

107. Pawson T, Scott JD. Signaling through scaffold, anchoring, and adaptor proteins. Science (New York, NY. 1997 Dec 19;278(5346):2075-80.

108. Sudol M. From Src Homology domains to other signaling modules: proposal of the 'protein recognition code'. Oncogene. 1998 Sep 17;17(11 Reviews):1469-74.

109. Bella J, Hindle KL, McEwan PA, Lovell SC. The leucine-rich repeat structure. Cell Mol Life Sci. 2008 Aug;65(15):2307-33.

110. Jorda J, Xue B, Uversky VN, Kajava AV. Protein tandem repeats - the more perfect, the less structured. The FEBS journal. 2010 Jun;277(12):2673-82.

111. Matsushima N, Tachi N, Kuroki Y, Enkhbayar P, Osaki M, Kamiya M, et al. Structural analysis of leucine-rich-repeat variants in proteins associated with human diseases. Cell Mol Life Sci. 2005 Dec;62(23):2771-91.

112. Batra-Safferling R, Abarca-Heidemann K, Korschen HG, Tziatzios C, Stoldt M, Budyak I, et al. Glutamic acid-rich proteins of rod photoreceptors are natively unfolded. J Biol Chem. 2006 Jan 20;281(3):1449-60.

113. Katti MV, Sami-Subbu R, Ranjekar PK, Gupta VS. Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. Protein Sci. 2000 Jun;9(6):1203-9.

114. Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ. Amino acid runs in eukaryotic proteomes and disease associations. Proceedings of the National Academy of Sciences of the United States of America. 2002 Jan 8;99(1):333-8.

115. Roepman P, Wessels LF, Kettelarij N, Kemmeren P, Miles AJ, Lijnzaad P, et al. An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. Nature genetics. 2005 Feb;37(2):182-6.

116. Roepman P, Kemmeren P, Wessels LF, Slootweg PJ, Holstege FC. Multiple robust signatures for detecting lymph node metastasis in head and neck cancer. Cancer Res. 2006 Feb 15;66(4):2361-6.

117. Rickman DS, Millon R, De Reynies A, Thomas E, Wasylyk C, Muller D, et al. Prediction of future metastasis and molecular characterization of head and neck squamous-cell carcinoma based on transcriptome and genome analysis by microarrays. Oncogene. 2008 Nov 20;27(51):6607-22.

118. Watanabe H, Mogushi K, Miura M, Yoshimura R, Kurabayashi T, Shibuya H, et al. Prediction of lymphatic metastasis based on gene expression profile analysis after brachytherapy for early-stage oral tongue carcinoma. Radiother Oncol. 2008 May;87(2):237-42.

119. Nagata T, Schmelzeisen R, Mattern D, Schwarzer G, Ohishi M. Application of fuzzy inference to European patients to predict cervical lymph node metastasis in carcinoma of the tongue. Int J Oral Maxillofac Surg. 2005 Mar;34(2):138-42.

120. Zhou X, Temam S, Oh M, Pungpravat N, Huang BL, Mao L, et al. Global expression-based classification of lymph node metastasis and extracapsular spread of oral tongue squamous cell carcinoma. Neoplasia (New York, NY. 2006 Nov;8(11):925-32.

121. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol. 2010 Jun;17(6):1471-4.

122. Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002;16(1):321-57.

123. Hall M. Correlation-based feature selection for machine learning: Citeseer; 1999.

124. Hall M, editor. Correlation-based feature selection for discrete and numeric class machine learning2000: Citeseer.

125. Oh JH, Craft J, Al Lozi R, Vaidya M, Meng Y, Deasy JO, et al. A Bayesian network approach for modeling local failure in lung cancer. Phys Med Biol. 2011 Feb 18;56(6):1635-51.

126. Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn CE, Jr., Burnside ES. Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. Cancer. 2010 Jul 15;116(14):3310-21.

127. Yousef M, Ketany M, Manevitz L, Showe LC, Showe MK. Classification and biomarker identification using gene network modules and support vector machines. BMC bioinformatics. 2009;10:337.

128. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.

129. Riddick G, Song H, Ahn S, Walling J, Borges-Rivera D, Zhang W, et al. Predicting in vitro drug sensitivity using Random Forests. Bioinformatics (Oxford, England). 2011 Jan 15;27(2):220-4.

130. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America. 2001 Apr 24;98(9):5116-21.

131. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:Article3.

132. Smyth GK, Michaud J, Scott HS. Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics (Oxford, England). 2005 May 1;21(9):2067-75.

133. Glaab E, Garibaldi JM, Krasnogor N. ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization. BMC bioinformatics. 2009;10:358.

134. Murphy KP. Dynamic Bayesian Networks: Representation, Inference and Learning: University of California; 2002.

135. Friedman N, editor. The Bayesian structural EM algorithm. 14th Conf on Uncertainty in Artificial Intelligence (UAI); 1998: Citeseer.

136. Liang S, Fuhrman S, Somogyi R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pacific Symposium on Biocomputing. 1998:18-29.